

Curriculum Coarse-to-Fine Selection for High-IPC Dataset Distillation

Yanda Chen* Gongwei Chen* Miao Zhang† Weili Guan Liqiang Nie
 School of Computer Science and Technology, Harbin Institute of Technology, Shenzhen
 cydaaa30@gmail.com
 {chengongwei, zhangmiao, guanweili, nieliqiang}@hit.edu.cn

Abstract

Dataset distillation (DD) excels in synthesizing a small number of images per class (IPC) but struggles to maintain its effectiveness in high-IPC settings. Recent works on dataset distillation demonstrate that combining distilled and real data can mitigate the effectiveness decay. However, our analysis of the combination paradigm reveals that the current one-shot and independent selection mechanism induces an incompatibility issue between distilled and real images. To address this issue, we introduce a novel curriculum coarse-to-fine selection (CCFS) method for efficient high-IPC dataset distillation. CCFS employs a curriculum selection framework for real data selection, where we leverage a coarse-to-fine strategy to select appropriate real data based on the current synthetic dataset in each curriculum. Extensive experiments validate CCFS, surpassing the state-of-the-art by +6.6% on CIFAR-10, +5.8% on CIFAR-100, and +3.4% on Tiny-ImageNet under high-IPC settings. Notably, CCFS achieves 60.2% test accuracy on ResNet-18 with a 20% compression ratio of Tiny-ImageNet, closely matching full-dataset training with only 0.3% degradation. Code: <https://github.com/CYDaaa30/CCFS>.

1. Introduction

Dataset distillation [32, 38] aims to condense the original training dataset into a small but powerful synthetic dataset, which can then be used to train competitive models. Current Dataset Distillation (DD) methods [3, 34, 43] have shown impressive performance at extremely small scales, such as 1 or 5 IPC (images-per-class). Unfortunately, these methods become less effective as IPC increases [4, 8, 46], sometimes even underperforming random sample selection.

Recent studies [8, 19] investigated this phenomenon and attributed it to a key issue in dataset distillation: current approaches tend to distill simple and general features into synthetic images while ignoring rare and complex features. These works attempt to address this issue from perspectives

*Equal contribution

†Corresponding authors

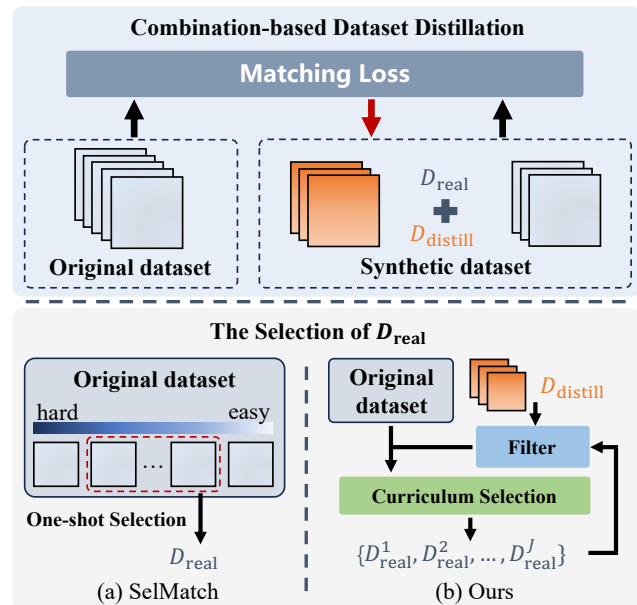


Figure 1. **Comparison of combination-based dataset distillation.** Top: General paradigm. Bottom: (a) SelMatch conducts an independent and one-shot selection of D_{real} . (b) Our method applies curriculum selection, making D_{real} dependent on $D_{distill}$.

of optimization and dataset construction. The former work, DATM [8], attempts to leverage the trajectories from different training stages to generate synthetic images with diverse patterns. It still fails to adequately incorporate the rare features of hard samples [19]. The latter work, SelMatch [19], advocates a combination-based paradigm which merges a distilled image set $D_{distill}$ and a real image set D_{real} to construct the synthetic dataset. By complementing rare and complex features of real data, SelMatch achieves state-of-the-art performance in high-IPC situations.

A vital advantage of the combination-based paradigm is the introduction of a real image set D_{real} . Although SelMatch shows impressive performance, we argue that its selection approach of D_{real} still has two shortcomings. 1) The fixed and one-shot selection of D_{real} is sub-optimal and may produce inappropriate real images. 2) The independence between D_{real} and $D_{distill}$ reduces the complementary effect

of $\mathcal{D}_{\text{real}}$. To verify our point, we compare SelMatch with two naive variants in Section 3.2, “SelMatch w/ two-shot selection” and “SelMatch w/ reverse selection” which only consider a two-shot paradigm and reverse order of selection and distillation. The superiority of proposed variants reveals the underlying incompatibility issue between $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{distill}}$ in SelMatch. This issue reduces the information richness of the generated synthetic dataset, distinctly impacting SelMatch’s performance in high-IPC situations.

To address the incompatibility issue, we propose a novel Curriculum Coarse-to-Fine Selection (CCFS) method for high-IPC dataset distillation. CCFS aims to progressively select suitable real data based on the distilled set. We cast the selection of the real images as a curriculum learning problem. This allows us to merge the real images from easy to difficult through a series of curriculum phases, ensuring comprehensive coverage of the essential patterns. To enhance the connection between the real images and the distilled images, we devise a coarse-to-fine selection strategy that takes into account both the global sample difficulty and the current synthetic dataset. Our selection strategy coarsely filters out correctly-classified samples, then finely chooses a subset from misclassified samples according to their difficulty scores. The coarse stage ensures that the selected samples contain the unlearned patterns of the current synthetic set. The fine stage maximizes the complementary effect of selected samples by avoiding overly complex ones.

Extensive experiments on CIFAR-10/100 [16] and Tiny-ImageNet [18] demonstrate that our approach consistently outperforms current state-of-the-art methods across compression ratios range of 5%-30%. Specifically, we achieved top-1 test accuracies of {92.5%, 71.5%, 60.2%} on {CIFAR-10, CIFAR-100, Tiny-ImageNet} with compression ratios of {10%, 10%, 20%}, surpassing the current SOTA by {6.6%, 5.8%, 3.4%}, separately. We record the state in each curriculum phase, including filter performance, difficulty distribution and visualization of real samples. The analyses of these states effectively illustrate the incremental growth in both performance and difficulty introduction of the synthetic dataset as the curriculum progresses, which aligns with our design principles. Our contributions can be summarized as follows:

1. We advocate the combination-based paradigm for high-IPC dataset distillation, and propose a novel curriculum coarse-to-fine selection method to address the existing incompatibility issue.
2. In each curriculum, we devise a coarse-to-fine selection strategy to yield the optimal real data by inspecting the limitations of the current synthetic dataset.
3. Our method achieves only 0.3% performance loss with a 20% compression ratio on Tiny-ImageNet. The analyses indicate the selected images become more complex and difficult as the curriculum progresses.

2. Related Works

Dataset Distillation. Dataset distillation aims to condense the original training dataset \mathcal{T} into a small but powerful synthetic set \mathcal{S} , which allows models to be trained efficiently and achieve performance comparable to full dataset training. Wang et al. [32] first proposed the concept of dataset distillation with a bi-level framework. Subsequently, a few works have sought to optimize dataset distillation from different perspectives [2, 10, 35, 41, 42]. Meanwhile, various surrogate objectives gradually formed several significant branches, which can be summarized as kernel-based approaches [23, 25, 47], gradient/trajectory-based methods [3, 6, 8, 21, 44], distribution-based techniques [31, 40, 43, 45], distilled dataset parameterization [5, 14, 22, 33] and decoupled-optimization [24, 29, 36, 37].

Recent studies [8, 19] have noticed the issue of dataset distillation failing when synthesizing large number of images per class (IPC). Their consensus is introducing more complex features into the synthetic dataset to alleviate its homogeneous and simplistic nature. DATM [8] employs a flexible trajectory matching to align expert trajectories from later training stages, aligning with the theory that models learn more complex patterns as training progresses. On the other hand, SelMatch [19] emphasizes the initialization of the synthetic dataset by using a sliding window to incorporate real images of appropriate difficulty levels. It divides the synthetic dataset into two subsets, $\mathcal{D}_{\text{distill}}$ and $\mathcal{D}_{\text{real}}$, allowing for updates to $\mathcal{D}_{\text{distill}}$ while keeping $\mathcal{D}_{\text{real}}$ fixed during MTT [3] distillation. This approach serves as the SOTA approach for high-IPC dataset distillation.

Our method CCFS is based on the concept of combining distilled data with real data to solve the failing problem in high-IPC cases. Unlike previous approaches, we design a novel curriculum framework to progressively select suitable real data for the distilled data, and conduct a more targeted selection strategy within each curriculum phase.

Curriculum Learning in Dataset Distillation Curriculum learning [1, 17, 20, 28, 39] is originally defined as a method for progressively training models by strategically arranging the inputting sequence of training data. Some dataset distillation methods leverage the concept of curriculum learning. SeqMatch [7] divides synthetic data into multiple subsets and sequentially optimizes them to learn high-level features. CDA [36] implements a progressive difficulty data augmentation on the synthetic images. CUDD [24] employs curriculum evaluation to gradually expand the distilled dataset. In CCFS, we design a curriculum framework that gradually expands the synthetic dataset by incorporating suitable real samples. This process takes into account both the prior knowledge of sample difficulty and the limitations of the current synthetic dataset. Our curriculum framework effectively enriches the diversity of the synthetic dataset and enhances its performance.

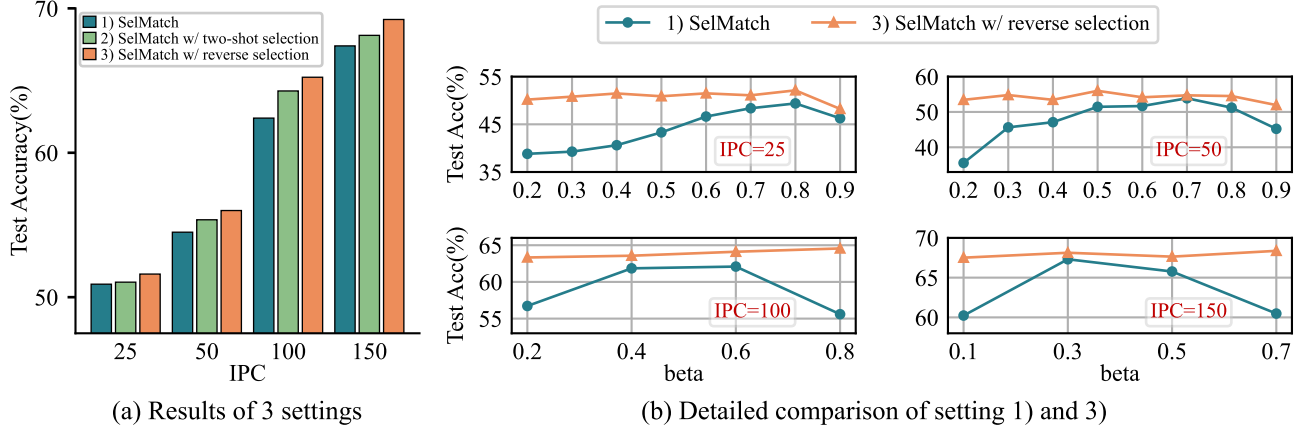


Figure 2. **Results of the analysis experiments on CIFAR-100.** (a) Top-1 accuracy of the 3 settings with IPC=25, 50, 100, 150. In each IPC, setting 2), which modifies only the selection strategy of $\mathcal{D}_{\text{real}}$, outperforms setting 1) with the original SelMatch setup. Setting 3) reverses setting 2)’s process by first distilling $\mathcal{D}_{\text{distill}}$ and then conducting a two-shot selection to obtain $\mathcal{D}_{\text{real}}$, resulting in the best performance among the 3 groups. (b) A detailed comparison between setting 1) and 3) at various window starting point β . In all cases of β , setting 3) outperforms setting 1) and shows more stable performance fluctuations across different β .

3. Preliminary

3.1. Combination-Based Dataset Distillation

Recent studies [4, 8, 46] reveal that traditional dataset distillation methods tend to synthesize simple features of the original dataset. This limits its effectiveness especially when the synthetic dataset contains more images per class (IPC). To address this problem, SelMatch [19] introduces a combination-based framework consisting of selection-based initialization and partial optimization.

SelMatch introduces modifications to the traditional optimization-based method in both the initialization and the updating phase. It begins by arranging the training samples of the original dataset in descending order of difficulty based on pre-calculated difficulty scores. Then it uses a sliding window of size IPC (images-per-class) to select subsets in each class with a window starting point hyperparameter $\beta \in [0, 1]$. It collects these selected subsets of each class as $\mathcal{D}_{\text{initial}}$ to initialize \mathcal{D}_{syn} .

Once the window starting point is determined, SelMatch further partitions samples within the window according to a distillation portion hyperparameter $\alpha \in [0, 1]$. The subset $\mathcal{D}_{\text{real}}$ contains the harder samples of the first $(1 - \alpha) \times |\mathcal{D}_{\text{syn}}|$ portion of the window and keeps unchanged during distillation. The remaining $\alpha \times |\mathcal{D}_{\text{syn}}|$ easier samples serve as the initialization set $\mathcal{D}_{\text{pre-distill}}$ for distillation to produce the subset $\mathcal{D}_{\text{distill}}$. Both β and α are optimal hyperparameters determined through a search process.

During subsequent MTT [3] distillation, the update aims to minimize the matching loss between the entire $\mathcal{D}_{\text{syn}} = \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{distill}}$ and the original dataset \mathcal{T} , i.e.,

$$\mathcal{L}(\mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{distill}}, \mathcal{T}). \quad (1)$$

3.2. Limitations of SelMatch

The combination-based paradigm in SelMatch represents the SOTA approach for addressing the less effective problem of high-IPC dataset distillation. However, we argue that the sliding window selection, as the core of SelMatch, may have shortcomings due to its rigid fixed and one-shot mechanism. To verify the existence of the shortcomings, we design two variants which break the mechanism. Details of the settings are as follows:

- 1) SelMatch:** Sort the original dataset in descending order by difficulty score. Determine \mathcal{D}_{syn} with a sliding window. Partition \mathcal{D}_{syn} into $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{distill}}$. Update $\mathcal{D}_{\text{distill}}$ by MTT’s approach and keep $\mathcal{D}_{\text{real}}$ unchanged.
- 2) SelMatch w/ two-shot selection:** Change the one-shot window selection of $\mathcal{D}_{\text{real}}$ into a two-shot selection. Train a model on the initialization set $\mathcal{D}_{\text{pre-distill}}$ and evaluate it on the full training set. Select the simplest-misclassified samples and add them into $\mathcal{D}_{\text{real}}$. Repeat the process twice, then merge $\mathcal{D}_{\text{real}}$ with $\mathcal{D}_{\text{distill}}$.
- 3) SelMatch w/ reverse selection:** Reverse the process in 2) to first conduct distillation and then make the two-shot selection. At first, conduct dataset distillation on $\mathcal{D}_{\text{pre-distill}}$ to generate $\mathcal{D}_{\text{distill}}$. Then implement the two-shot selection of $\mathcal{D}_{\text{real}}$ based on the distilled set $\mathcal{D}_{\text{distill}}$. Finally merge them to produce the synthetic dataset.

We conduct our analytical experiments on CIFAR-100 and evaluate the final $\mathcal{D}_{\text{syn}} = \mathcal{D}_{\text{real}} \cup \mathcal{D}_{\text{distill}}$ on the test dataset. The results are shown in Figure 2.

Figure 2(a) presents top-1 test accuracy of the 3 settings with the best hyperparameters (such as α and β). Across all IPC cases, setting 3) performs the best, followed by setting 2), with setting 1) falling behind. Specifically, setting 3) achieves an average performance improvement of

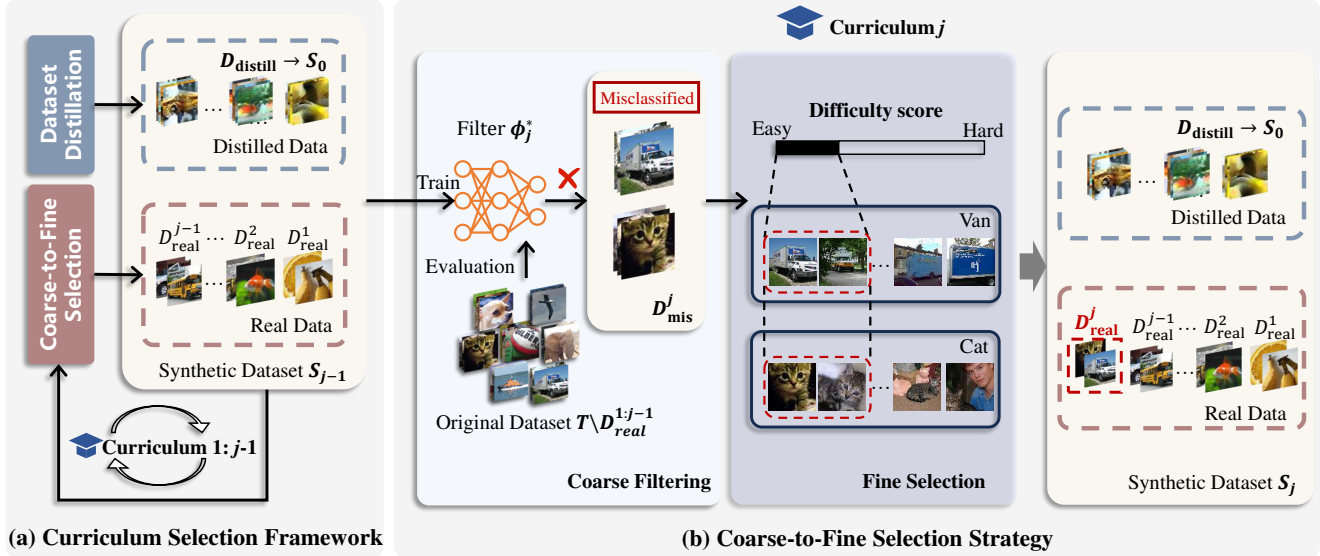


Figure 3. **Architecture of our curriculum coarse-to-fine selection method for high-IPC dataset distillation, CCFS.** CCFS adopts a combination of distilled and real data to construct the final synthetic dataset. We apply a curriculum framework and select the optimal real data for the current synthetic dataset in each curriculum. (a) **Curriculum selection framework:** CCFS begins the curriculum with the already distilled data as the initial synthetic dataset. Then continuously incorporates real data into the current synthetic dataset through the coarse-to-fine selection within each curriculum phase. (b) **Coarse-to-fine selection strategy:** In the coarse stage, CCFS trains a filter model on the current synthetic dataset and evaluates it on the original dataset excluding already selected data to filter out all correctly classified samples. In the fine stage, CCFS selects the simplest misclassified samples and incorporates them into the current synthetic dataset for the next curriculum.

1.7% over setting 1) across the four IPC settings (max: +2.8%), while 2) achieves an average improvement of 0.9% over 1)(max: +1.7%). This improvement becomes more obvious as IPC increases. Figure 2(b) presents a detailed comparison between setting 1) and 3) at various window starting point β . In all cases of β , setting 3) outperforms setting 1) and shows more stable performance fluctuations across different β .

In the first comparison between setting 1) and 2), setting 2) obtains a better $\mathcal{D}_{\text{real}}$ through more targeted, multiple-times selections. This reflects **the rigid limitations of the fixed and one-shot sliding window initialization**. While in the second comparison between setting 2) and 3), the only difference lies in whether the selection is made based on the initialization set $\mathcal{D}_{\text{pre-distill}}$ or the distilled set $\mathcal{D}_{\text{distill}}$. It proves that selecting based on the distilled set is better. In both 1) and 2), although $\mathcal{D}_{\text{real}}$ is suitable for $\mathcal{D}_{\text{pre-distill}}$ at initialization, the updated $\mathcal{D}_{\text{distill}}$ by the distillation prevents the assurance of this compatibility. When we reverse the process as setting 3), the selection process becomes more targeted, further enhancing the connection between $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{distill}}$. **The independence between $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{distill}}$ reduces the complementary effect of $\mathcal{D}_{\text{real}}$.**

These two factors collectively lead to the incompatibility issue between $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{distill}}$ in the current combination paradigm, resulting in a consistent performance gap compared to full datasets. This incompatibility becomes more evident under suboptimal settings shown in Figure 2(b).

Once deviating from the optimal window position, the performance of SelMatch drops rapidly. This incompatibility issue inspired us to explore more effective strategies to select suitable real data based on the distilled data.

4. Method

In this section, we introduce our Curriculum Coarse-to-Fine Selection method (CCFS) for high-IPC dataset distillation. Following the idea of combining distilled and real data to construct the synthetic dataset, CCFS aims to progressively select suitable real data based on the distilled data. We first get distilled data through dataset distillation methods. Then employ a **curriculum selection framework** for real data selection beginning with distilled data. In each curriculum, we conduct our **coarse-to-fine selection strategy** to obtain the optimal real data and integrate it with the current synthetic dataset. Figure 3 illustrates the architecture of CCFS.

Curriculum Selection Framework Our analysis experiments reveal that the fixed and one-shot selection of $\mathcal{D}_{\text{real}}$ hinders obtaining suitable real samples for the synthetic dataset. This motivated us to structure the selection process of $\mathcal{D}_{\text{real}}$ as a curriculum framework for more comprehensive coverage of essential patterns.

To make $\mathcal{D}_{\text{real}}$ a more effective complement to $\mathcal{D}_{\text{distill}}$. We choose to conduct selection after distillation is finished. We begin the curriculum with the already distilled dataset $\mathcal{D}_{\text{distill}}$ as the initial synthetic dataset \mathcal{S}_0 . In each curriculum phase

j , we expect to obtain the optimal $\mathcal{D}_{\text{real}}^j$ from the original dataset for the current synthetic dataset \mathcal{S}_{j-1} :

$$\begin{aligned} \mathcal{D}_{\text{real}}^j &= \text{Select}(\mathcal{S}_{j-1}, \mathcal{T} \setminus \mathcal{D}_{\text{real}}^{1:j-1}) \\ \text{s.t. } \mathcal{S}_0 &= \mathcal{D}_{\text{distill}} \text{ and } j \in \{1, 2, \dots, J\}, \end{aligned} \quad (2)$$

where J is the number of curricula and Select denotes the selection strategy used in each curriculum. Note that previously selected samples are excluded in each curriculum.

Then we incorporate $\mathcal{D}_{\text{real}}^j$ into \mathcal{S}_{j-1} as \mathcal{S}_j for the next curriculum phase. After the last curriculum, we obtain the final synthetic dataset \mathcal{S} with the target size:

$$\begin{aligned} \mathcal{S}_j &= \mathcal{S}_{j-1} \cup \mathcal{D}_{\text{real}}^j, \\ \mathcal{S} &= \mathcal{S}_J. \end{aligned} \quad (3)$$

We expect to gradually incorporate suitable features through this curriculum selection framework. Next, we introduce our strategy of selecting the optimal $\mathcal{D}_{\text{real}}$ in each curriculum phase and explain how this strategy functions within the curriculum framework.

Coarse-to-Fine Selection Strategy The weak connection between $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{distill}}$ in the incompatibility issue calls for a more targeted selection strategy. We design a two-step strategy, coarse-to-fine selection, to get the optimal $\mathcal{D}_{\text{real}}$ in each curriculum phase. Given a synthetic dataset \mathcal{S} , we first use a filter ϕ trained on \mathcal{S} to evaluate on the original training dataset \mathcal{T} and obtain the misclassified samples \mathcal{D}_{mis} :

$$\mathcal{D}_{\text{mis}} = \{(x_i, y_i) \in \mathcal{T} \mid \phi_{\theta_{\mathcal{S}}}^*(x_i) \neq y_i\}, \quad (5)$$

where $\theta_{\mathcal{S}}$ denotes the parameters of ϕ optimized on \mathcal{S} .

This evaluation can coarsely reflect the limitations of current \mathcal{S} . These limitations are primarily concentrated in the misclassified samples \mathcal{D}_{mis} from the original training set. Considering \mathcal{D}_{mis} may include samples with beneficial, harmful, or negligible influence on training [15, 26], we conduct a finer selection next. We arrange the samples of \mathcal{D}_{mis} in ascending order based on pre-computed difficulty scores [12, 30] and then select the easiest samples up to the target size as the optimal complement to current \mathcal{S} :

$$\mathcal{D}_{\text{real}} = \bigcup_{c \in \mathcal{C}} \left\{ (x_i, y_i) \in \mathcal{D}_{\text{mis}}^{(c)} \mid \text{rank}_{\mathcal{D}_{\text{mis}}^{(c)}}(x_i) \leq k \right\}, \quad (6)$$

where \mathcal{C} represents the total number of classes, $\mathcal{D}_{\text{mis}}^{(c)}$ includes samples of class c in \mathcal{D}_{mis} , rank denotes the ascending order arrangement of sample difficulty, and k is the target complement amount for each class. To keep balance, we select an equal number of complement samples per class.

We figure that simple features that haven't been learned are essential for \mathcal{S} . The nature of dataset distillation often leads to synthesizing mainly easy and representative features from the original dataset [8, 19, 24], which provides

the filter with fundamental classification capabilities as evidenced by correctly classified samples. In the first step, we coarsely filter out correctly classified samples to avoid reintroducing these already learned simple features.

Now we have \mathcal{D}_{mis} that reflects the limitations in \mathcal{S} . Among these limitations, simpler features provide greater benefit to model training compared to more difficult and complex features, as they are easier to learn. Pre-calculated difficulty scores effectively measure the relative difficulty of sample features from a global perspective, guiding our fine selection in the next step. By selecting the simplest ones from misclassified samples, we obtain the optimal $\mathcal{D}_{\text{real}}$ while avoiding the introduction of overly complex features that could hinder the performance of \mathcal{S} . Section 5.4 further demonstrates the effectiveness of our selection strategy.

Algorithm 1 describes the entire process of the CCFS algorithm. We employ this coarse-to-fine selection in our curriculum framework. Although we repeatedly select the simplest-misclassified samples across curriculum phases, the continuous enrichment of the synthetic dataset leads to an improvement in the filter's capacity, which in turn raises the lower bound of the difficulty for misclassified samples. Meanwhile, the strategy of selecting the simplest samples maintains a manageable difficulty progression between curriculum phases. We provide more details on the progressive nature of CCFS in Further Analysis 5.5.

Algorithm 1 CCFS: A curriculum coarse-to-fine selection framework for high-IPC dataset distillation

Input: Original full dataset \mathcal{T} , number of classes C , target images per class IPC , distillation portion $\alpha \in [0, 1]$, dataset distillation algorithm \mathcal{A} , number of curricula J , pre-calculated difficulty score *score*.

```

 $\mathcal{D}_{\text{distill}} = \mathcal{A}(\mathcal{T})$ , s.t.  $|\mathcal{D}_{\text{distill}}| = \lceil \alpha \times \text{IPC} \times C \rceil$ 
 $\mathcal{S}_0 \leftarrow \mathcal{D}_{\text{distill}}$ 
for  $j = 1$  to  $J$  do
     $k_j = \lfloor \frac{\text{IPC} \times (1 - \alpha)}{J} \rfloor$ 
    Train the filter model on  $\mathcal{S}_{j-1}$  to get  $\phi_j^*$ 
     $\triangleright$  Coarse Filtering
     $\mathcal{D}_{\text{mis}}^j = \{(x_i, y_i) \in \mathcal{T} \setminus \mathcal{D}_{\text{real}}^{1:j-1} \mid \phi_j^*(x_i) \neq y_i\}$ 
     $\triangleright$  Fine Selection
     $\text{rank}_{\mathcal{D}_{\text{mis}}^{j,c}} \leftarrow \text{sort}_{\text{asc}}(\mathcal{D}_{\text{mis}}^{j,c}, \text{score})$ 
     $\mathcal{D}_{\text{real}}^j = \bigcup_{c \in \mathcal{C}} \{(x_i, y_i) \in \mathcal{D}_{\text{mis}}^{j,c} \mid \text{rank}_{\mathcal{D}_{\text{mis}}^{j,c}}(x_i) \leq k_j\}$ 
     $\mathcal{S}_j \leftarrow \mathcal{S}_{j-1} \cup \mathcal{D}_{\text{real}}^j$ 
end for
Output: The final synthetic dataset  $\mathcal{S} \leftarrow \mathcal{S}_J$ 

```

Table 1. **Performance of CCFS compared to the SOTA dataset distillation and coreset selection baselines.** We report the results of all listed methods with the identical validation model ResNet-18. CCFS achieves state-of-the-art performance across high-IPC settings ranging from 5% to 30% compression ratio. Additionally, the selection-only version of our method, self-evolved selection, beats other coreset selection baselines and exhibits comparable performance to SOTA dataset distillation methods. * denotes results obtained using the official code due to the incomplete results shown in original papers. *IPC*: images per class, *Ratio*: the compression ratio of the synthetic dataset compared to the original dataset.

Dataset	CIFAR-10				CIFAR-100				Tiny-ImageNet	
	IPC	250	500	1000	1500	25	50	100	150	50
Ratio	5%	10%	20%	30%	5%	10%	20%	30%	10%	20%
Random	73.4±1.5	79.3±0.3	85.6±0.4	88.3±0.2	35.8±0.6	40.7±1.0	53.2±0.9	60.3±1.3	30.1±0.6	40.1±0.4
Forgetting [30]	30.7±0.3	41.5±0.7	68.4±1.6	83.5±1.8	9.5±0.3	13.2±0.6	27.0±1.1	42.3±1.0	5.7±0.1	12.4±0.2
Glister [13]	46.6±1.3	56.6±0.5	79.0±0.7	85.0±0.9	21.7±0.8	26.7±1.3	39.9±1.4	52.1±1.3	22.6±0.5	34.0±0.3
Oracle window [19]	79.3±0.7	85.2±0.1	89.9±0.5	90.6±0.3	43.2±1.8	50.0±0.8	59.2±0.8	64.7±0.5	42.5±0.3	49.2±0.3
Self-evolved selection	<u>81.6±0.5</u>	<u>86.4±0.3</u>	<u>90.3±0.5</u>	<u>91.6±0.4</u>	<u>45.6±0.5</u>	<u>50.7±0.7</u>	<u>62.6±0.8</u>	<u>66.5±0.2</u>	<u>43.9±0.6</u>	<u>50.2±0.4</u>
DSA [42]	74.7±1.5	78.7±0.7	84.8±0.5	-	38.4±0.4	43.6±0.7	-	-	27.8±1.4	-
DM [43]	75.3±1.4	79.1±0.6	85.6±0.5	-	37.5±0.6	42.6±0.5	-	-	31.0±0.6	-
MTT [3]	80.7±0.4	82.2±0.4	86.1±0.3	88.6±0.2	49.9±0.7	51.3±0.4	58.7±0.6	63.1±0.3	40.3±0.3	44.2±0.5
SRe ² L* [37]	77.5±0.7	85.1±0.2	86.8±0.3	87.8±0.4	49.7±0.4	51.4±0.4	58.8±0.2	61.9±0.3	41.1±0.4	49.7±0.3
DATM [8]	-	84.8±0.3	87.6±0.3	-	-	51.0±0.5	61.5±0.3	-	42.2±0.2	-
SelMatch [19]	82.8±0.2	85.9±0.2	90.4±0.2	91.3±0.2	50.9±0.3	54.5±0.6	62.4±0.5	67.4±0.2	44.7±0.2	50.4±0.2
CDA* [36]	78.0±0.4	84.4±0.4	86.4±0.2	87.5±0.4	50.6±0.3	59.7±0.2	61.1±0.1	63.4±0.2	45.6±0.2	52.4±0.1
CUDD [24]	-	-	-	-	63.5±0.3	65.7±0.2	-	-	55.6±0.2	56.8±0.2
CCFS (Ours)	87.9±0.4	92.5±0.2	93.2±0.1	93.8±0.1	65.3±0.2	71.5±0.3	73.0±0.2	74.8±0.2	55.8±0.3	60.2±0.2
Full Dataset	95.5±0.2				78.8±0.3				60.5±0.2	

5. Experimental Results

5.1. Experiment Setup

We evaluate the performance of our method CCFS on various datasets including CIFAR-10, CIFAR-100, and Tiny-ImageNet. We compare our method with SOTA dataset distillation and coreset selection methods. For coreset selection baselines, we include Glister [13], Forgetting [30], and the oracle-window selection proposed in SelMatch. For comparison, we also report a selection-only version of CCFS, referred to as self-evolved selection. For dataset distillation baselines, we incorporate DSA [42], DM [43], MTT [3], DATM [8], SelMatch [19], SRe²L [37], CDA [36] and CUDD [24]. We also report the full dataset training performance with 200 training epochs.

Datasets Details.

- CIFAR-10 [16]: 10 classes with 5000 low-resolution (32×32) training images per class, 10000 images for testing.
- CIFAR-100 [16]: 100 classes with 500 low-resolution (32×32) training images per class, 10000 images for testing.
- Tiny-ImageNet [18]: 200 classes with 500 high-resolution (64×64) training images per class, 10000 images for validation.

Evaluation Networks. We choose ResNet-18 [9] as the uniform evaluation network for the main comparison. For cross-architecture generalization, we use ResNet-50/101, DenseNet-121 [11], and RegNet-Y-8GF [27] as the evaluation backbones.

Implement Details of CCFS. We begin with $\mathcal{D}_{\text{distill}}$ synthesized by CDA [36] method, which belongs to SRe²L [37] series and implements a progressive difficulty data augmentation on the synthetic images during distillation. In the next curriculum selection, we set the default number of curriculum phases to 3 and evenly distribute the samples to be selected among them. In each curriculum, we train ResNet-18 from scratch on the current synthetic dataset as the filter, using equal training epochs as those in the final evaluation. We use pre-calculated Forgetting [30] scores and apply our selection strategy on the training set, excluding previously selected samples. We report the results of the optimal distillation portion α at each IPC setting. Our method has excellent scalability and can be adapted to various dataset distillation methods. We present the results of combining CCFS with MTT [3] dataset distillation in the Appendix.

5.2. Main Results

We compare our method with the state-of-the-art dataset distillation and coreset selection methods on CIFAR-10, CIFAR-100, and Tiny-ImageNet under high-IPC settings ranging from 5% to 30% compression ratios. As shown in Table 1, previous distillation methods gradually lose effectiveness as IPC increases, even falling behind random selection. By progressively introducing suitable real samples into the synthetic dataset, CCFS establishes new state-of-the-art performance in high-IPC settings. Notably, our method achieves a performance gain of $\{6.6\%, 5.8\%\}$ on $\{\text{CIFAR-10}, \text{CIFAR-100}\}$ with compression ratio of 10%.

Table 2. Cross-architecture experiment results on Tiny-ImageNet with IPC=100.

Method	Validation Model				
	R18	R50	R101	DenseNet-121	RegNet-Y-8GF
SRe ² L	48.00	51.02	51.92	50.66	54.78
CDA	51.12	54.00	55.04	52.47	57.13
CCFS (Ours)	60.20	60.67	61.17	60.52	62.94

Table 3. Ablation study on the select strategy on CIFAR-100 with IPC=50.

Coarse Stage	Fine Stage		
	Simple	Hard	Random
Classified	66.8	63.5	66.8
Misclassified	71.5	65.0	70.1

For Tiny-ImageNet with IPC=100 (20% compression ratio), we achieve 60.2% top-1 test accuracy, representing a 3.4% improvement over the current state-of-the-art method. This performance comes remarkably close to the 60.5% test accuracy of full dataset training.

Additionally, we report a selection-only version of CCFS, referred to as self-evolved selection. We select real samples of appropriate difficulty with a sliding window as the initial coreset and expand it following the CCFS strategy. Self-evolved selection significantly outperforms other coreset selection methods in high IPC and also exhibits comparable performance to advanced dataset distillation approaches. It indicates that progressively selecting suitable real images is crucial for producing a coreset with the largest coverage of essential patterns in the original dataset.

5.3. Cross-architecture Generalization

To further evaluate CCFS’s effectiveness, we conduct cross-architecture generalization experiments using additional validation models beyond the ResNet-18 in the main table, including ResNet-50/101, DenseNet-121, and RegNet-Y-8GF. We set ResNet-18 as the filter model for curriculum selection and generate the final synthetic dataset, which is then used to train other validation models from scratch. As shown in Table 2, our method demonstrates robust generalization performance.

5.4. Ablation Study

The selection strategy. In our coarse-to-fine selection strategy, we choose the misclassified subset in the coarse stage and select the simplest ones in the fine stage next. This combination has demonstrated excellent performance. Here, we explore other combinations. Specifically, for each curriculum selection, we select the simplest/hardest or just randomly select samples from correctly-classified/misclassified subset by current trained filter model. We conducted experiments on CIFAR-100 with IPC=50. Among all results shown in Table 3, the simplest-misclassified selection strategy outperforms all other combinations, further demonstrating its effectiveness.

Table 4. Ablation study on the difficulty score used in selection strategy with 10% compression ratio on CIFAR-10, CIFAR-100 and Tiny-ImageNet.

Score	CIFAR-10	CIFAR-100	Tiny-ImageNet
Logits	91.8	68.7	52.5
C-score	92.2	71.0	-
Forgetting	92.5	71.5	55.8

Table 5. Ablation study on the number of curricula with 10% compression ratio on CIFAR-10, CIFAR-100 and Tiny-ImageNet.

Number of curricula	CIFAR-10	CIFAR-100	Tiny-ImageNet
1	91.6	67.9	54.4
2	91.8	70.4	55.3
3	92.5	71.5	55.8
4	92.4	71.6	55.7

The difficulty scores. Our method requires difficulty scores to measure the complexity of samples. We explored the impact of using different difficulty scores. We include pre-calculated C-score [12] and Forgetting scores [30]. Additionally, we measure sample difficulty using the predicted values (logits) by the current trained filter model for the actual class of each training sample, with smaller values indicating greater difficulty. We conducted experiments on CIFAR-10, CIFAR-100, and Tiny-ImageNet with a 10% compression ratio using these three scores. Results in Table 4 indicate that the Forgetting score outperformed the others across all three datasets, while the logits approach leads to performance degradation due to its coarse reflection of sample difficulty. The Forgetting score is leveraged for the main results (Table 1).

The number of curricula. Table 5 illustrates the impact of varying the number of curricula. It indicates that moderate increases improve performance, further demonstrating the effectiveness of the curriculum selection framework. However, continuing to increase number of curricula results in only marginal performance gains. Balancing performance and efficiency, we employ 3 curriculum phases in our main results (Table 1).

5.5. Further Analysis

The curriculum framework in CCFS aims to progressively incorporate suitable samples into the synthetic dataset, thereby enhancing its performance incrementally. We expect to observe a continuous improvement in the classification capability of the filter model, along with a gradual increase in the overall difficulty of the selected samples as the curriculum progresses. To verify that CCFS achieves these intended effects, we conduct extensive experiments on Tiny-ImageNet with 3 curriculum phases and record the state in each curriculum phase. Figure 4 illustrates the effectiveness of the curriculum framework in CCFS from both the filter performance and sample difficulty.

In Figure 4(a), we present the performance of the filter model trained on the synthetic dataset in each curriculum

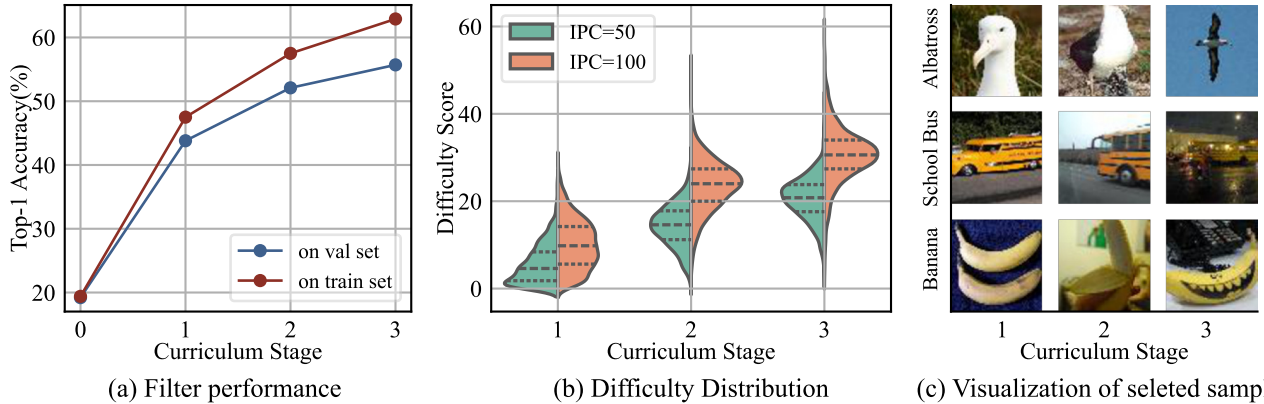


Figure 4. **Further analysis on the curriculum framework.** (a) Performance of the filter model trained on the synthetic dataset in each curriculum phase with IPC=50: The filter’s classification accuracy steadily improves on both the original training set and the validation set. (b) The difficulty distribution of real samples selected in each curriculum phase: As the curriculum progresses, both the average difficulty as well as the upper and lower difficulty bounds of selected samples increase significantly. Moreover, higher IPC tend to include more difficult samples than lower IPC within the same curriculum phase. CCFS effectively guides the synthetic dataset to incorporate more challenging samples. (c) Visualization of the samples selected in each curriculum phase. We present images of median difficulty across several categories in Tiny-ImageNet: Albatross, School Bus and Banana. The visualization effectively illustrates the gradual increase in difficulty (diverse poses, complex backgrounds, other distractions...) facilitated by CCFS.

phase with IPC=50. As the curriculum progresses, the filter’s classification accuracy steadily improves on both the original training set and the validation set. This indirectly reflects the growing representational capacity of the synthetic dataset.

Figure 4(b) illustrates the difficulty distribution of real samples selected in each curriculum phase. We can observe that the selected samples’ average difficulty increases significantly across 3 curriculum phases as intended. Moreover, as the curriculum progresses, both the upper and lower bounds of the difficulty in selected samples are rising. This trend indicates that CCFS effectively enhances the overall difficulty of the synthetic dataset instead of struggling within a similar range. Since we continuously select the simplest-misclassified samples in each curriculum phase, the increase in lower bounds also indirectly reflects the rising threshold for the filter model’s errors, indicating a more concrete improvement in its filtering capability, rather than just a simple performance boost. Consequently, CCFS effectively guides the synthetic dataset to incorporate more challenging and training-valuable samples. Additionally, within the same curriculum phase, synthetic datasets with higher IPC tend to include more difficult samples. This also aligns with our expectations: larger synthetic datasets are supposed to encapsulate more complex information. As IPC increases, CCFS successfully introduces harder and rarer features into the synthetic dataset.

In Figure 4(c), we visualize the samples selected in different curriculum phases, showcasing samples of median difficulty across several categories. In the early stages, CCFS tends to select classic samples that capture the gen-

eral features of the category. These images have simple backgrounds and fully visible objects. As the curriculum progresses, more challenging samples are incorporated into the synthetic dataset, featuring diverse poses (e.g., bird in flight, peeled banana), partial views (e.g., the lower half of bird, the front of school bus), complex backgrounds, and other distractions. This visualization effectively illustrates the gradual increase in difficulty facilitated by CCFS.

6. Conclusion

In this paper, we reveal the incompatibility issue between distilled and real data in the current combination-based dataset distillation method through a series of analysis experiments. We propose CCFS, a novel combination-based framework for high-IPC dataset distillation. We apply a curriculum selection framework for real data and begin the curriculum with distilled data. This ensures suitable features are progressively introduced into the synthetic dataset across curriculum phases. In each curriculum phase, we employ our coarse-to-fine selection strategy to obtain the optimal real data for the current synthetic dataset. This effectively enhances the connection between distilled and real data. CCFS significantly narrows the performance gap between synthetic datasets and full datasets under high-IPC conditions. We achieve state-of-the-art performance in various high-IPC settings on CIFAR-10, CIFAR-100, and Tiny-ImageNet. Further analyses demonstrate the effectiveness of our selection strategy and the expected progressive effect of the curriculum framework. CCFS also exhibits robust cross-architecture generalization and excellent scalability to other distillation approaches.

Acknowledgement

This study is supported by the National Natural Science Foundation of China (Grant No. 62306084, U23B2051, 62476071, and U24A20328), and Shenzhen Science and Technology Program (Grant No. GXWD20231128102243003, KJZD20230923115113026, and ZDSYS20230626091203008), and China Postdoctoral Science Foundation (Grant No. 2024M764192).

References

- [1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48, 2009. 2
- [2] Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Flexible dataset distillation: Learn labels instead of images. *arXiv preprint arXiv:2006.08572*, 2020. 2
- [3] 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, 3, 6
- [4] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. DC-BENCH: Dataset condensation benchmark. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 3
- [5] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. *Advances in Neural Information Processing Systems*, 35:34391–34404, 2022. 2
- [6] 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. 2
- [7] Jiawei Du, Qin Shi, and Joey Tianyi Zhou. Sequential subset matching for dataset distillation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2
- [8] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2, 3, 5, 6
- [9] 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. 6
- [10] Yang He, Lingao Xiao, Joey Tianyi Zhou, and Ivor Tsang. Multisize dataset condensation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 2
- [11] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6
- [12] Ziheng Jiang, Chiyuan Zhang, Kunal Talwar, and Michael C Mozer. Characterizing structural regularities of labeled data in overparameterized models. *arXiv preprint arXiv:2002.03206*, 2020. 5, 7
- [13] Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glisten: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8110–8118, 2021. 6
- [14] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoon Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pages 11102–11118. PMLR, 2022. 2
- [15] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017. 5
- [16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Technical Report, University of Toronto, Department of Computer Science*, 2009. 2, 6
- [17] M Kumar, Benjamin Packer, and Daphne Koller. Self-paced learning for latent variable models. *Advances in neural information processing systems*, 23, 2010. 2
- [18] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 2, 6
- [19] Yongmin Lee and Hye Won Chung. Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. In *Forty-first International Conference on Machine Learning*, 2024. 1, 2, 3, 5, 6
- [20] Yong Jae Lee and Kristen Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR 2011*, pages 1721–1728. IEEE, 2011. 2
- [21] Dai Liu, Jindong Gu, Hu Cao, Carsten Trinitis, and Martin Schulz. Dataset distillation by automatic training trajectories. *arXiv preprint arXiv:2407.14245*, 2024. 2
- [22] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. *Advances in neural information processing systems*, 35:1100–1113, 2022. 2
- [23] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2
- [24] Zhiheng Ma, Anjia Cao, Funing Yang, and Xing Wei. Curriculum dataset distillation. *arXiv preprint arXiv:2405.09150*, 2024. 2, 5, 6
- [25] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2
- [26] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. Estimating training data influence by tracing gradient descent. *Advances in Neural Information Processing Systems*, 33:19920–19930, 2020. 5

- [27] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 6
- [28] Petru Soviany. Curriculum learning with diversity for supervised computer vision tasks. *arXiv preprint arXiv:2009.10625*, 2020. 2
- [29] 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 (CVPR)*, pages 9390–9399, 2024. 2
- [30] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018. 5, 6, 7
- [31] 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. 2
- [32] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1, 2
- [33] Xing Wei, Anjia Cao, Funing Yang, and Zhiheng Ma. Sparse parameterization for epitomic dataset distillation. In *NeurIPS*, 2023. 2
- [34] Yifan Wu, Jiawei Du, Ping Liu, Yuewei Lin, Wenqing Cheng, and Wei Xu. Towards adversarially robust dataset distillation by curvature regularization. *arXiv preprint arXiv:2403.13322*, 2024. 1
- [35] Yue Xu, Yong-Lu Li, Kaitong Cui, Ziyu Wang, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. Distill gold from massive ores: Bi-level data pruning towards efficient dataset distillation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2
- [36] Zeyuan Yin and Zhiqiang Shen. Dataset distillation in large data era. *arXiv preprint arXiv:2311.18838*, 2023. 2, 6
- [37] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2, 6
- [38] Ruonan Yu, Songhua Liu, and Xinchao Wang. A comprehensive survey to dataset distillation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):150–170, 2023. 1
- [39] Dingwen Zhang, Deyu Meng, Chao Li, Lu Jiang, Qian Zhao, and Junwei Han. A self-paced multiple-instance learning framework for co-saliency detection. In *Proceedings of the IEEE international conference on computer vision*, pages 594–602, 2015. 2
- [40] Hansong Zhang, Shikun Li, Fanzhao Lin, Weiping Wang, Zhenxing Qian, and Shiming Ge. DANCE: Dual-view distribution alignment for dataset condensation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2024. 2
- [41] Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Xu Dongkuan. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11950–11959, 2023. 2
- [42] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, 2021. 2, 6
- [43] 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. 1, 2, 6
- [44] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. 2
- [45] 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. 2
- [46] Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. Dataset quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 1, 3
- [47] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022. 2