On the Diversity and Realism of Distilled Dataset: An Efficient Dataset Distillation Paradigm

Peng Sun[1,2] Bei Shi[2] Daiwei Yu[3,*] Tao Lin[2,†]
[1]Zhejiang University [2]Westlake University [3]Independent Researcher
sunpeng@westlake.edu.cn, shibei0430@gmail.com, ydw.ccm@gmail.com, lintao@westlake.edu.cn

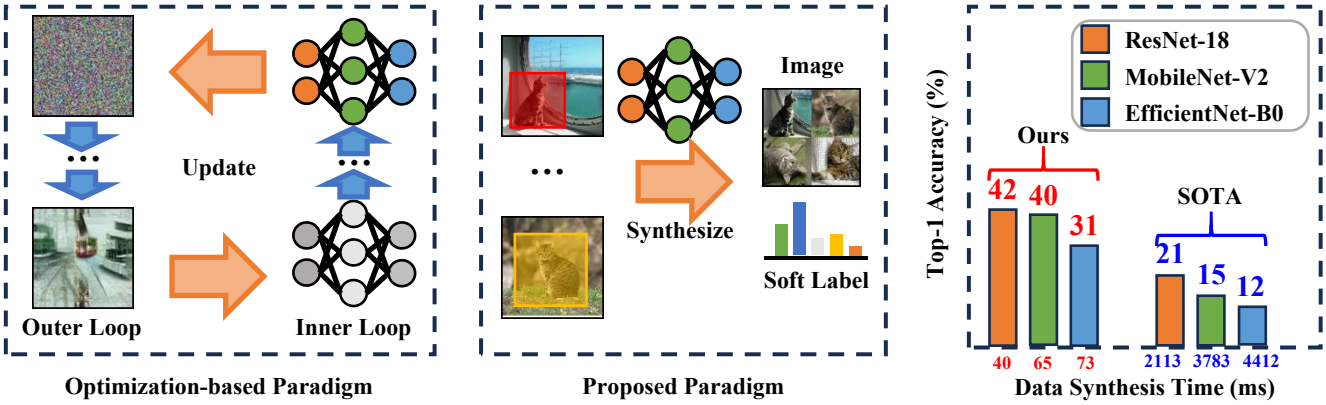


Figure 1. **Proposed paradigm vs. optimization-based paradigm.** Left is the mainstream optimization-based dataset distillation and middle is our proposed non-optimizing paradigm. Right is top-1 validation accuracy vs. synthesis time per image on ImageNet-1K with IPC $= 10$ (10 Images Per Class). Models used for distillation include ResNet-18, EfficientNet-B0, and MobileNet-V2; we use ResNet-18 for evaluation.

Abstract

Contemporary machine learning, which involves training large neural networks on massive datasets, faces significant computational challenges. Dataset distillation, as a recent emerging strategy, aims to compress real-world datasets for efficient training. However, this line of research currently struggles with large-scale and high-resolution datasets, hindering its practicality and feasibility. Thus, we re-examine existing methods and identify three properties essential for real-world applications: realism, diversity, and efficiency. As a remedy, we propose RDED, a novel computationally-efficient yet effective data distillation paradigm, to enable both diversity and realism of the distilled data. Extensive empirical results over various model architectures and datasets demonstrate the advancement of RDED: we can distill a dataset to 10 *images per class from full ImageNet-1K [6] within* 7 *minutes, achieving a notable* 42% *accuracy with ResNet-18 [14] on a single RTX-4090 GPU (while the SOTA only achieves* 21% *but requires* 6 *hours). Code: https://github.com/LINs-lab/RDED.*

1. Introduction

The success of modern deep learning could be largely attributed to the fact of scaling and increasing both neural architectures and training datasets [12, 14, 16, 17]. Though this pattern shows great potential to propel artificial intelligence forward, the challenge of high computational requirements remains a noteworthy concern [4, 39, 45]. Dataset distillation methods have recently emerged [39, 45] and attracted attention for their exceptional performance [1, 2, 38, 50–52]. The key idea is compressing original full datasets by synthesizing and optimizing a small dataset, where training a model using the synthetic dataset can achieve a similar performance to the original.

However, these methods suffer a high computational burden [5, 44] due to the bi-level optimization-based paradigm. Moreover, the synthetic images exhibit certain *non-realistic* features (see Figure 2b and 2d) that have materialized due to overfitting to a specific architecture used during the optimization process, which leads to difficulties in generalizing to other architectures [2, 31].

A notable work [2] investigates the relationship between realism and expressiveness in synthetic datasets. The findings reveal a trade-off: more realistic images come at the

† Corresponding author.
∗ Work done during Daiwei's visit to Westlake University.

sacrifice of expressiveness. While realism aids in generalizing across different architectures, it hurts distillation performance. Conversely, prioritizing expressiveness over realism can enhance distillation performance but may impede cross-architecture generalization.

Inspired by these insights, we introduce an Realistic, Diverse, and Efficient Dataset Distillation (RDED) method. Our goal is to achieve diversity (expressiveness) and realism simultaneously across varying datasets, ranging from CIFAR-10 to ImageNet-1K. Specifically, we directly crop and select realistic patches from the original data to maintain realism. To ensure the greatest possible diversity, we stitch the selected patches into the new images as the synthetic dataset. It is noteworthy that our method is non-optimization-based, so it can also achieve high efficiency, making it well-suited for processing large-scale, high-resolution datasets.

The key contributions of this work can be summarized as:

- We first investigate the limitations of existing dataset distillation methods and define three key properties for effective dataset distillation on large-scale high-resolution datasets: realism, diversity, and efficiency.
- We introduce the definitions of diversity ratio and realism score backed by $\mathcal{V}$-information theory [42], together with an optimization-free efficient paradigm, to enhance diversity and realism of the distilled data.
- Extensive experiments demonstrate the effectiveness of our method: it not only achieves a top-1 validation accuracy that is twice the current SOTA—SRe^{2}L [44], but it also operates at a speed 52 times faster (see Figure 1).

2. Related Work

Dataset distillation, as proposed by Wang et al. [39], condenses large datasets into smaller ones without sacrificing performance. These methods fall into four main categories.

Bi-level optimization-based distillation. A line of work seeks to minimize the surrogate models learned from both synthetic and original datasets, depending on their metrics, namely, the matching gradients [18, 49, 51], features [38], distribution [50, 52], and training trajectories [1, 4, 5, 8, 13, 45]. Notably, trajectory matching-based techniques have demonstrated remarkable performance across various benchmarks with low IPC. However, the synthetic data often overfit to a specific model architecture, struggling to generalize to others.

Distillation with prior regularization. Cazenavette et al. [2] suggest that direct pixel space parameterization is a key factor for the architecture transferability issue, and propose GLaD to integrate a generative prior for dataset distillation to enhance generalization across any distillation method. However, bi-level optimization-based methods, especially those that entail prior regularization, face computational challenges and memory issues [5].

Uni-level optimization-based distillation. Kernel ridge-regression methods [23, 53], with uni-level optimization, effectively reduce training costs [53] and enhancing performance [23]. However, due to the resource-intensive nature of inverting matrix operations, scaling these methods to larger IPC remains challenging. Unlike NTK-based solutions, Yin et al. [44] propose to decouple the bi-level optimization of dataset condensation into two single-level learning procedures, resulting in a more efficient framework.

CoreSet selection-based distillation. CoreSet selection, akin to traditional dataset distillation, focuses on identifying representative samples using provided images and labels. Various difficulty-based metrics are proposed to assess the sample importance, e.g., the forgetting score [37], memorization [10], EL2N score [27], diverse ensembles [25].

3. On the Limits of Dataset Distillation

We start by clearly defining the concept of dataset distillation and then reveal the primary challenges in this field.

3.1. Preliminary

The goal of dataset distillation is to synthesize a smaller distilled dataset, denoted as $\mathcal{S} = (X, Y) = \{\mathbf{x}_j, y_j\}_{j=1}^{|\mathcal{S}|}$, that captures the essential characteristics of a larger dataset $\mathcal{T} = (\hat{X}, \hat{Y}) = \{\hat{\mathbf{x}}_i, \hat{y}_i\}_{i=1}^{|\mathcal{T}|}$. Here, the distilled dataset $\mathcal{S}$ is generated by a distillation algorithm $\mathcal{D}$ such that $\mathcal{S} \in \mathcal{D}(\mathcal{T})$, where the size of $\mathcal{S}$ is considerably smaller than $\mathcal{T}$ (i.e., $|\mathcal{S}| \ll |\mathcal{T}|$). Each $y_j \in Y$ corresponds to the synthetic distilled label for the sample $\mathbf{x}_j \in X$, and similar definitions can be applied to $\hat{y}_i$ and $\hat{\mathbf{x}}_i$. The key motivation for dataset distillation is to create a dataset $\mathcal{S}$ that allows models to achieve performance within an acceptable deviation ϵ from those trained on $\mathcal{T}$. Formally, this is expressed as:

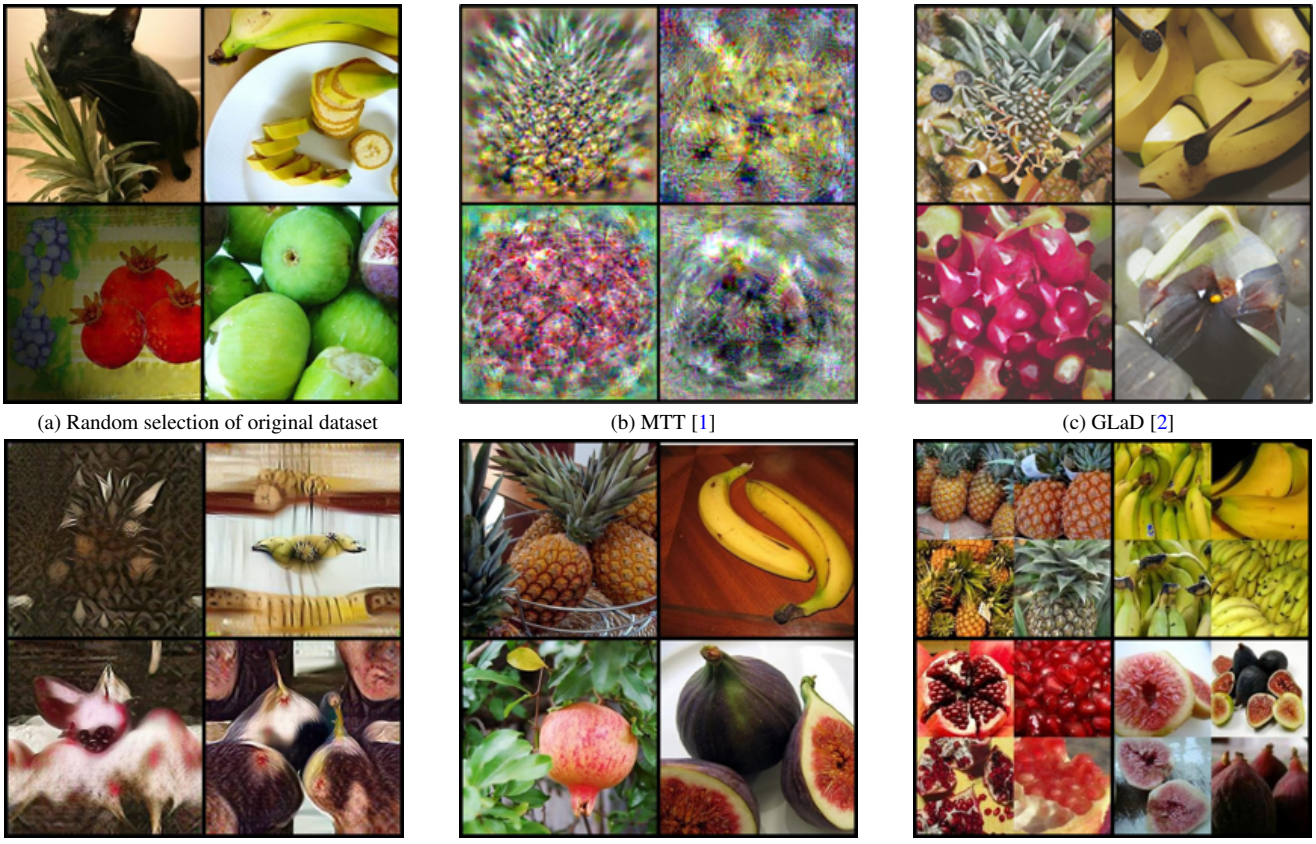
$$\sup \left\{ \left| \ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\hat{\mathbf{x}}), \hat{y}) - \ell(\phi_{\boldsymbol{\theta}_\mathcal{S}}(\hat{\mathbf{x}}), \hat{y}) \right| \right\}_{(\hat{\mathbf{x}}, \hat{y}) \sim \mathcal{T}} \leq \epsilon, \quad (1)$$

where $\boldsymbol{\theta}_\mathcal{T}$ is the parameter set of the neural network ϕ optimized on $\mathcal{T}$:

$$\boldsymbol{\theta}_\mathcal{T} = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_{(\hat{\mathbf{x}}, \hat{y}) \in \mathcal{T}} \left[\ell(\phi_{\boldsymbol{\theta}}(\hat{\mathbf{x}}), \hat{y}) \right], \quad (2)$$

with ℓ representing the loss function. A similar definition applies to $\boldsymbol{\theta}_\mathcal{S}$.

The properties of optimal dataset distillation. The effectiveness and utility of dataset distillation methods rely on key properties outlined in Definition 1. These properties are crucial for creating datasets efficiently, which in turn, enhances model training and generalization.



(a) Random selection of original dataset (b) MTT [1] (c) GLaD [2]

(d) SRe^{2}L [44] (e) Herding [40] (f) RDED (Ours)

Figure 2. **Visualization of images synthesized using various dataset distillation methods**. We consider the ImageNet-Fruits [1] dataset, comprising a total of 10 distinct fruit types, with a resolution of 128 × 128. There are four specific classes for each method, namely, 1) *Pineapple*, 2) *Banana*, 3) *Pomegranate*, and 4) *Fig*. Note that MTT [1], GLaD [2], SRe^{2}L [44], and Herding [40], are four representative methods of conventional dataset distillation paradigms discussed in Section 2 and Section 3.2 (see Appendix A for more visualization). In general, ensuring both superior *realism* and *diversity* simultaneously is challenging for methods other than ours and GLaD.

	Property			Dataset	
Method	Diversity	Realism	Efficiency	Large-scale	High-resolution
MTT	✔	✗	✗	✗	✗
GLaD	✔	✔	✗	✗	✔
SRe^{2}L	✓	✓	✓	✓	✓
Herding	✓	✓	✔	✓	✓
Ours	✔	✔	✔	✔	✔

Table 1. **Properties and performance of various representative SOTA dataset distillation methods.** We give a summary of the properties of different methods and their performance on large-scale or high-resolution datasets, where ✔, ✓, and ✗, denote "Superior", "Satisfactory", and "Bad" respectively.

Definition 1 (Properties of distilled data)**.** *Consider a family of observer models* $\mathcal{V}$[1]*. The core attributes of a distilled dataset* $\mathcal{S} = (X, Y) \in \mathcal{D}(\mathcal{T})$ *are defined as follows:*

1. ***Diversity:*** *Essential for robust learning and generalization, a high-quality dataset should cover a wide range of samples* X *and labels* Y *[17, 28, 34]. This ensures exposure to diverse features and contexts.*
2. ***Realism:*** *Critical for (cross-architecture) generalization, realistic distilled samples* X *and labels* Y *should be accurately predicted and matched by various observer models from* $\mathcal{V}$*. It is important to avoid features or annotations that are overfitted to a specific model [1, 50, 51].*
3. ***Efficiency:*** *A determinant for the feasibility of dataset distillation, addressing the computational and memory challenges is crucial for scaling the distillation algorithm* $\mathcal{D}$ *to large datasets [5, 44].*

[1] $\mathcal{V}$ includes different observer models, for instance, humans ϕ_{h} and pre-trained models $\phi_{\boldsymbol{\theta}_\mathcal{T}}$. Here, ϕ_{h} is an abstraction representing human predictive behavior.

3.2. Pitfalls of Conventional Dataset Distillation

In response to the properties of the optimal dataset distillation, in this section, we conduct a comprehensive examination of four conventional dataset distillation paradigms discussed in Section 2. Limitations are detailed below and summarized in Table 1 (see more details in Appendix A).

- **Bi-level optimization-based distillation.** Conventional dataset distillation methods [1, 50, 51] suffer from noise-like *non-realistic* patterns (see Figure 2b) in distilled high-

resolution images and overfit the specific architecture used in training [2], which hurt its cross-architecture generalization ability [2]. However, these methods suffer a *high computational burden* [5, 44] due to the bi-level optimization-based paradigm.

- **Distillation with prior regularization.** Cazenavette et al. [2] identify the source of the architecture overfitting issue, and thus enhances the realism (see Figure 2c) of synthetic images and the cross-architecture generalization. The current remedy inherits the *low efficiency* of bi-level optimization-based distillation, and thus still cannot generalize to large-scale datasets.
- **Uni-level optimization-based distillation.** As a remedy for the former research, Yin et al. [44]—as the latest progress in the field—alleviate the *efficiency* and *realism* challenges (see Figure 2d) and propose SRe^{2}L to distill large-scale, high-resolution datasets, e.g., ImageNet-1K. Yet, SRe^{2}L is hampered by a limited *diversity* problem arising from its synthesis approach, which involves extracting knowledge from a pre-trained model containing only partial information of the original dataset [43].
- **CoreSet selection-based distillation.** CoreSet selection methods [34, 35, 40] serve to *efficiently* distill datasets by isolating a CoreSet containing *realistic* images (see Figure 2e). However, the advances come at the cost of limited information representation (*data diversity*) [27], leading to a catastrophically degraded performance [35].

4. Methodology

To tackle the remaining concern of distilling high-resolution and large-scale image datasets, in this section, we articulate an novel unified dataset distillation paradigm—RDED—that prioritizes both *diversity* and *realism* within the distilled dataset, yet being *efficient*.

4.1. Enhancing Data Diversity and Realism

Establishing a $\mathcal{V}$-information-based objective for distilled data. Drawing on the artificial intelligence learning principles of *parsimony* and *self-consistency* from [24], we strive to ensure that the models trained on the distilled dataset embody these principles. To achieve this, we aim to construct a representation Y of the input data X that is structured (*parsimony*) and rich in information (*self-consistency*). Consequently, we reinterpret the objective of dataset distillation in (1), as the structured and sufficient information inherent in the original full dataset $\mathcal{T}$:

$$\mathcal{S} = \underset{(X,Y)\in\mathcal{D}(\mathcal{T})}{\arg\max}\ I_{\mathcal{V}}(X \to Y)\,, \tag{3}$$

where $I_{\mathcal{V}}$ denotes the predictive $\mathcal{V}$-information [42] from X to Y, which can be further defined as:

$$I_{\mathcal{V}}(X \to Y) = \underbrace{H_{\mathcal{V}}(Y|\varnothing)}_{\text{diversity}} - \underbrace{H_{\mathcal{V}}(Y|X)}_{\text{realism}}\,, \tag{4}$$

where $H_{\mathcal{V}}(Y|X)$ and $H_{\mathcal{V}}(Y|\varnothing)$ denote, respectively, the predictive conditional $\mathcal{V}$-entropy [42] with observed side information X and no side information $\varnothing$.

Explicitizing diversity and realism via $\mathcal{V}$-information. Building upon Definition 1, maximizing $H_{\mathcal{V}}(Y|\varnothing)$ can enhance the uncertainty/diversity of representations Y measured by the observer models in $\mathcal{V}$. Simultaneously, minimizing $H_{\mathcal{V}}(Y|X)$ aims to improve the predictiveness/realism of the data pairs (X, Y) [9, 42]. Therefore, the objective of (3) is equivalent to maximize the first term $H_{\mathcal{V}}(Y|\varnothing)$ while minimizing the second term $H_{\mathcal{V}}(Y|X)$ in (4), to achieve the improved data diversity and realism.

Approximating and maximizing $\mathcal{V}$-information. For the sake of computational feasibility, we restrict ourselves to the case where the predictive family $\mathcal{V}$ includes only humans and a single pre-trained observer model associated with dataset $\mathcal{T}$, denoted as $\mathcal{V} = \{\phi_{\mathrm{h}}, \phi_{\boldsymbol{\theta}_\mathcal{T}}\}$. Given the computational challenges of solving (3) by maximizing both terms in (4) simultaneously, we decouple the terms in (4), resulting in:

$$\inf_{f\in\mathcal{V}} \begin{cases} \mathbb{E}_{y\sim Y}\left[-\log f[\varnothing](y)\right] \\ -\mathbb{E}_{\mathbf{x},y\sim X,Y}\left[-\log f[\mathbf{x}](y)\right] \end{cases}, \tag{5}$$

where $f[\mathbf{x}]$ is probability measure on Y based on the received information $\mathbf{x}$, and $f[\mathbf{x}](y) \in \mathbb{R}$ is the value of the density evaluated at $y \in Y$. Then, we seek proxies to approximate the decoupled two terms in (5) independently.

Proposition 1 (Proxies on diversity and realism of distilled data)**.** *Given a distilled dataset $\mathcal{S} = (X, Y)$, we derive the following approximations to maximize the diversity term $H_{\mathcal{V}}(Y|\varnothing)$ and the realism term $-H_{\mathcal{V}}(Y|X)$:*

1. *The diversity ratio $H_{\mathcal{V}}(Y|\varnothing)/H_{\mathcal{V}}(\mathcal{T}|\varnothing)$ is posited as a lower bound of the information preservation ratio from the original dataset $\mathcal{T}$ to the distilled one $\mathcal{S}$, justified by:*

$$H_{\mathcal{V}}(Y|\varnothing) \le H_{\mathcal{V}}(\mathcal{S}|\varnothing) \le H_{\mathcal{V}}(\mathcal{T}|\varnothing)\,. \tag{6}$$

   Therefore, we maximize diversity through preserving more information from the original dataset $\mathcal{T}$.

2. *The realism score for a distilled sample $\mathbf{x}$ and label y from a pair $(\mathbf{x}, y)$ is defined as:*

$$-\ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}), \phi_{\mathrm{h}}(\mathbf{x})) - \ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}), y)\,. \tag{7}$$

   To enhance the realism score for each distilled pair $(\mathbf{x}, y)$, we prioritize the distillation of sample $\mathbf{x}$ with higher $-\ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}), \phi_h(\mathbf{x}))$ and assign the label $y = \phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x})$.

Summary. *To bolster the two properties—diversity and realism—of our distilled dataset, we employ two practical proxies, namely 1) the diversity ratio, and 2) the realism score, as the approximation to design distillation algorithm $\mathcal{D}$ (see Appendix B for more (theoretical) analysis).*

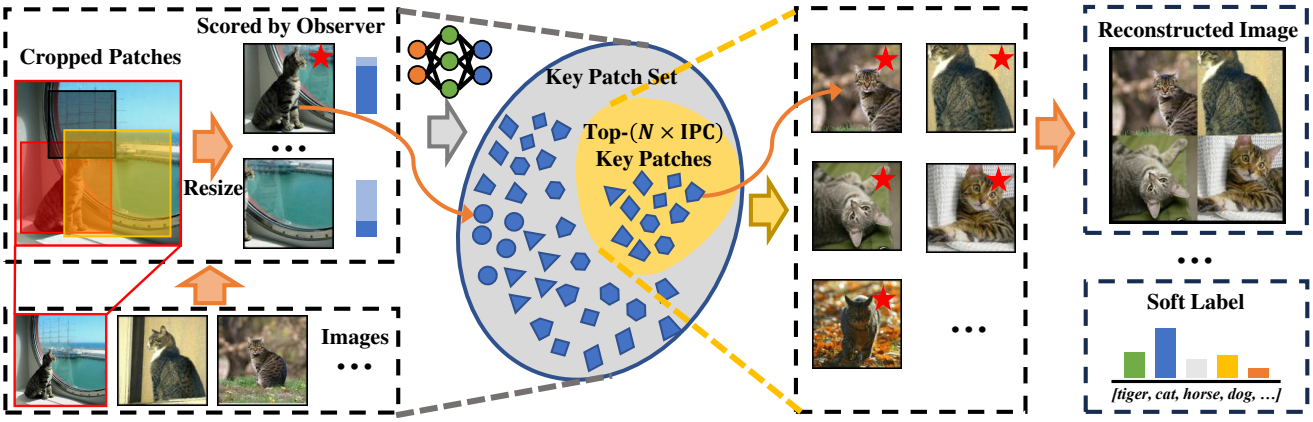


Figure 3. **Visualization of our proposed two-stage dataset distillation framework.** Stage 1: We crop each original image into several patches and rank them using the realism scores calculated by the observer model. Then, we choose the top-1-scored patch as the key patch. For the key patches within a class, we re-select the top-$N \times$ IPC patches based on their scores, where $N = 4$ in this case. Stage 2: We consolidate every N selected patches from Stage 1 into a single new image that shares the same resolution with each original image, resulting in IPC-numbered distilled images per class. These images are then relabeled using the pre-trained observer model.

Overview of our dataset distillation paradigm. To enhance the diversity and realism of our distilled dataset, we introduce a novel two-stage paradigm that practically utilizes the proposed two proxies in Proposition 1 (See Figure 3 and Algorithm 1). In particular, our objective is to preserve the information within a large number of sample pairs exhibiting high realism scores from the original full dataset $\mathcal{T}$ into the distilled dataset $\mathcal{S}$. This process unfolds in two stages:

- First stage in Section 4.2 *extracts major information* (i.e., key sample pairs) with *high realism score* from $\mathcal{T}$.
- In the second stage (see Section 4.3), we aim to *compress the extracted information* from the first stage into finite pixel space to form distilled images and *relabel* them.

4.2. Extracting Key Patches from Original Dataset

To extract the explicit key information from the original full dataset, we capture the key patches with high realism scores at the pixel space level and sample space level respectively.

Extracting key patch per image. Motivated by the common practice in Vision Transformer [7, 48] that *image patches* are sufficient to capture object-related information, we propose to learn the most realistic patch, $\xi_{i,\star}$, from a set of patches $\{\xi_{i,k}\}$, which are extracted from a given image $\hat{\mathbf{x}}_i \in \hat{X}$. The whole procedure can be formulated as:

$$\xi_{i,\star} = \underset{\xi_{i,k} \sim p(\xi_{i,k}|\hat{\mathbf{x}}_i)}{\arg\max} \; -\ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\xi_{i,k}), \phi_{\mathrm{h}}(\xi_{i,k}))\,, \tag{8}$$

where the label $\phi_{\mathrm{h}}(\xi_{i,k})$ annotated by humans is given as y_i. Therefore, $s_{i,k} := -\ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\xi_{i,k}), y_i)$ represents the realism score for the patch $\xi_{i,k}$ and $s_{i,\star}$ denotes the highest one.

Let $\mathcal{T}_c := \{(\hat{\mathbf{x}}, \hat{y}) | (\hat{\mathbf{x}}, \hat{y}) \in \mathcal{T}, \hat{y} = c\}$ denote a sub-dataset comprising all samples associated with class c from the full dataset $\mathcal{T}$. Given the key patches and its corresponding scores $(\xi_{i,\star}, s_{i,\star})$ for $\hat{\mathbf{x}}_i \in \mathcal{T}_c$, we form them into a set $\mathcal{Q}_c$.

Capturing inner-class information. Solely relying on information extraction at the pixel space level is inadequate in averting information redundancy at the sample space level. To further extract key information from the original dataset, we consider a sample space level selection to further scrutiny of the selected patches from the previous stage.

More precisely, certain patches—denoted as $\mathcal{Q}'_c$—are selected based on a given pruning criteria $\bar{s}_\star$ defined over $\mathcal{Q}_c$, aiming to capture the most impactful patches for class c, whose socres are larger than $\bar{s}_\star$. This process is iteratively repeated for all classes of $\mathcal{T}$.

Practical implementation. In practice, extracting all key patches from the entire $\mathcal{T}_c$ and subsequently selecting the top patches based on scoring presents two significant challenges:

- Iterating through each image in $\mathcal{T}_c$ to identify crucial patches incurs a *considerable computational overhead*.
- Utilizing a score-based selection strategy typically introduces *distribution bias* within the chosen subset of the original dataset $\mathcal{T}_c$, which hurts data diversity and adversely affects generalization (see Section 5.5 for more details).

To address the aforementioned issues, we propose the adoption of the random uniform data selection strategy[2] to derive a pre-selected subset $\mathcal{T}'_c \subset \mathcal{T}_c$ (see settings in Section 5.1). The subsequent inner-class information-capturing process is then performed exclusively on this subset $\mathcal{T}'_c$.

4.3. Information Reconstruction of Patches

To effectively save the previously extracted key information in the limited pixel space and label space of the distilled dataset, we propose to reconstruct the information in patches.

[2]This treatment is motivated by the findings [3, 34, 35] on the impact of various data selection strategies, where random uniform selection is a *lightweight* yet *unbiased* data selection strategy.

Images reconstruction. The patch size is typically smaller than the dimensions of an expected distilled image, where directly utilizing the patches selected as distilled images may lead to sparse information in the pixel space.

Therefore, for a given class c with a selected patch set $\mathcal{Q}'_c$, we randomly retrieve N patches[3] without replacement to form a final image $\mathbf{x}_j$ by applying the following operation:

$$\mathbf{x}_j = \text{concatenate}(\{\xi_{i,\star}\}_{i=1}^{N} \subset \mathcal{Q}'_c)\,. \tag{9}$$

Labels reconstruction. The previous investigation [47] highlights a critical limitation associated with single-label annotations, wherein a random crop of an image may encompass an entirely different object than the ground truth, thereby introducing noisy or even erroneous supervision during the training process. Consequently, relying solely on the simplistic one-hot label proves inadequate for representing an informative image, consequently constraining the effectiveness and efficiency of model learning [44].

Inspired by this observation, we propose to re-label the squeezed multi-patches within the distilled images $\mathbf{x}_j$, thereby encapsulating the informative label for the distilled images. It can be achieved by employing the soft labelling approach [32] to generate region-level soft labels $y_{j,m} = \ell\left(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}_{j,m})\right)$, where $\mathbf{x}_{j,m}$ is the m-th region in the distilled image and $y_{j,m}$ is the corresponding soft label.

Training with reconstructed labels. We train the student model $\phi_{\boldsymbol{\theta}_\mathcal{S}}$ on the distilled data using the following objective:

$$\mathcal{L} = -\sum_j \sum_m y_{j,m} \log \phi_{\boldsymbol{\theta}_\mathcal{S}}(\mathbf{x}_{j,m})\,. \tag{10}$$

5. Experiment

This section assesses the efficacy of our proposed method over SOTA methods across diverse datasets and neural architectures, followed by extensive ablation studies.

5.1. Experimental Setting

We list the settings below (see more details in Appendix D).

Datasets. For low-resolution data (32×32), we evaluate our method on two datasets, i.e., CIFAR-10 [20] and CIFAR-100 [19]. For high-resolution data, we conduct experiments on two large-scale datasets including Tiny-ImageNet (64×64) [21] and full ImageNet-1K (224×224) [6]. Moreover, given the fact that most existing dataset distillation methods cannot be extended to large-scale high-resolution datasets, we further consider four widely used ImageNet-1K subsets

[3]We use N and target resolution of distilled images to calculate and set the resolution of the patches, e.g., for an image with 224×224 resolution and set $N = 4$, the resolution of patches is defined as 112×112.

Algorithm 1 RDED: An efficient framework for high-resolution dataset distillation (see Appendix C for more implementation details)

Input: Original full dataset $\mathcal{T}$, a corresponding pre-trained observer model $\phi_{\boldsymbol{\theta}_\mathcal{T}}$ and initial $\mathcal{S} = \varnothing$.
for $\mathcal{T}'_c \subset \mathcal{T}_c \subset \mathcal{T}$ **do**
  for $(\hat{\mathbf{x}}_i, \hat{y}_i) \in \mathcal{T}'_c$ **do** ▷ Stage 1
    Crop $\hat{\mathbf{x}}_i$ into K patches $\{\xi_{i,k}\}_{k=1}^{K}$
    for $k = 1$ to K **do**
      Calculate the score $s_{i,k} = -\ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\xi_{i,k}), \hat{y}_i)$
    Select patch $\xi_{i,\star}$ from $\{\xi_{i,k}\}_{k=1}^{K}$ via $s_{i,\star}$
  Select top-($N \times$ IPC) patches
  for $j = 1$ to IPC **do** ▷ Stage 2
    Squeeze N selected patches into $\mathbf{x}_j$
    Relabel $\mathbf{x}_j$ with y_j
    $\mathcal{S} = \mathcal{S} \cup \{(\mathbf{x}_j, y_j)\}$
Output: Small distilled dataset $\mathcal{S}$

in our evaluation: ImageNet-10 and ImageNet-100 [18], ImageNette and ImageWoof [1].

Network architectures. Similar to the prior dataset distillation works [1, 5, 13, 44, 52], we use ConvNet [13], ResNet-18 and ResNet-101 [14], EfficientNet-B0 [36], MobileNet-V2 [29], as our backbone.

Baselines. We consider SOTA optimization-based dataset distillation methods that can scale to large high-resolution datasets for a broader practical impact:

- MTT [1] is the first work that proposes trajectory matching-based dataset distillation, which can work on *both low and high-resolution datasets*.
- IDM [52] introduces an efficient dataset condensation method based on distribution matching, in contrast to computationally intensive optimization-oriented approaches [1, 51], thus *scaling to ImageNet-100*.
- TESLA [5] is the first dataset distillation method *scales to full ImageNet-1K*, which handles huge memory consumption of the MTT-based method with constant memory.
- DATM [13] is the first to *outperform the original full dataset training performance* with large IPC.
- SRe^{2}L [44] is a recent work to *efficiently scale to ImageNet-1K*, and significantly outperforms existing methods on large high-resolution datasets. We consider it as our closest baseline.

Evaluation. Following previous research, we set IPC to 1, 10, and 50. To evaluate cross-architecture generalization, we use the distilled datasets from one neural architecture to train the other neural architectures from scratch and record the validation accuracy (see Table 4). Furthermore, we evaluate

Dataset	IPC	ConvNet					ResNet-18		ResNet-101	
		MTT	IDM	TESLA	DATM	RDED (Ours)	SRe^{2}L	RDED (Ours)	SRe^{2}L	RDED (Ours)
CIFAR10	1	46.3 ± 0.8	45.6 ± 0.7	**48.5 ± 0.8**	46.9 ± 0.5	23.5 ± 0.3	16.6 ± 0.9	**22.9 ± 0.4**	13.7 ± 0.2	**18.7 ± 0.1**
	10	65.3 ± 0.7	58.6 ± 0.1	66.4 ± 0.8	**66.8 ± 0.2**	50.2 ± 0.3	29.3 ± 0.5	**37.1 ± 0.3**	24.3 ± 0.6	**33.7 ± 0.3**
	50	71.6 ± 0.2	67.5 ± 0.1	72.6 ± 0.7	**76.1 ± 0.3**	68.4 ± 0.1	45.0 ± 0.7	**62.1 ± 0.1**	34.9 ± 0.1	**51.6 ± 0.4**
CIFAR-100	1	24.3 ± 0.3	20.1 ± 0.3	24.8 ± 0.5	**27.9 ± 0.2**	19.6 ± 0.3	6.6 ± 0.2	**11.0 ± 0.3**	6.2 ± 0.0	**10.8 ± 0.1**
	10	40.1 ± 0.4	45.1 ± 0.1	41.7 ± 0.3	47.2 ± 0.4	**48.1 ± 0.3**	27.0 ± 0.4	**42.6 ± 0.2**	30.7 ± 0.3	**41.1 ± 0.2**
	50	47.7 ± 0.2	50.0 ± 0.2	47.9 ± 0.3	55.0 ± 0.2	**57.0 ± 0.1**	50.2 ± 0.4	**62.6 ± 0.1**	56.9 ± 0.1	**63.4 ± 0.3**
ImageNette	1	**47.7 ± 0.9**	-	-	-	28.9 ± 0.1	19.1 ± 0.9	**27.7 ± 1.1**	16.9 ± 0.2	**24.3 ± 0.7**
	10	**63.0 ± 1.3**	-	-	-	59.0 ± 1.0	45.7 ± 1.0	**62.7 ± 0.8**	36.1 ± 2.5	**53.3 ± 2.7**
	50	-	-	-	-	**83.1 ± 0.6**	64.4 ± 0.7	**84.4 ± 0.2**	62.8 ± 0.4	**80.9 ± 0.4**
ImageWoof	1	**28.6 ± 0.8**	-	-	-	18.0 ± 0.3	15.0 ± 0.6	**17.9 ± 1.0**	15.6 ± 0.6	**18.0 ± 0.7**
	10	35.8 ± 1.8	-	-	-	**40.1 ± 0.2**	30.0 ± 0.8	**44.4 ± 1.8**	22.3 ± 0.7	**35.9 ± 2.1**
	50	-	-	-	-	**60.8 ± 0.5**	41.3 ± 1.1	**71.7 ± 0.3**	38.6 ± 0.5	**66.1 ± 0.3**
Tiny-ImageNet	1	8.8 ± 0.3	10.1 ± 0.2	-	**17.1 ± 0.3**	12.0 ± 0.1	2.62 ± 0.1	**9.7 ± 0.4**	1.9 ± 0.1	**3.8 ± 0.1**
	10	23.2 ± 0.2	21.9 ± 0.3	-	31.1 ± 0.3	**39.6 ± 0.1**	16.1 ± 0.2	**41.9 ± 0.2**	14.6 ± 1.1	**22.9 ± 3.3**
	50	28.0 ± 0.3	27.7 ± 0.3	-	39.7 ± 0.3	**47.6 ± 0.2**	41.1 ± 0.4	**58.2 ± 0.1**	**42.5 ± 0.2**	41.2 ± 0.4
ImageNet-100	1	-	**11.2 ± 0.5**	-	-	7.0 ± 0.2	3.1 ± 0.1	**7.7 ± 0.2**	2.9 ± 0.2	**5.7 ± 0.3**
	10	-	17.1 ± 0.6	-	-	**27.2 ± 0.2**	8.2 ± 0.1	**35.4 ± 0.7**	11.0 ± 0.2	**31.3 ± 0.3**
	50	-	26.3 ± 0.4	-	-	**47.8 ± 0.4**	33.2 ± 0.9	**64.8 ± 0.6**	51.7 ± 1.0	**67.3 ± 0.8**
ImageNet-1K	1	-	-	**7.7 ± 0.2**	-	6.4 ± 0.1	0.1 ± 0.1	**6.6 ± 0.2**	0.6 ± 0.1	**5.9 ± 0.4**
	10	-	-	17.8 ± 1.3	-	**20.4 ± 0.1**	21.3 ± 0.6	**42.0 ± 0.1**	30.9 ± 0.1	**48.3 ± 1.0**
	50	-	-	27.9 ± 1.2	-	**38.4 ± 0.2**	46.8 ± 0.2	**56.5 ± 0.1**	60.8 ± 0.5	**61.2 ± 0.4**

Table 2. **Comparison with the SOTA baseline dataset distillation methods.** We use identical neural networks for both dataset distillation and data evaluation. In general, following [1, 5, 52], the ConvNet used for distillation are Conv-3 on CIFAR10 and CIFAR100, Conv-4 on Tiny-ImageNet and ImageNet-1K, Conv-5 on ImageNette and ImageWoof, Conv-6 on ImageNet-100. MTT and TESLA use a down-sampled version of image when distilling 224×224 images [1, 5]. Following [44], SRe^{2}L and RDED use ResNet-18 to retrieve the distilled data, and evaluate on ResNet-18 and ResNet-101. Entries with "-" are absent due to scalability problems. See Appendix C for more details.

the distillation efficiency in Table 3 by estimating the runtime cost of distilling the image, as well as the peak GPU memory usage.

Implementation details of RDED. We employ a generalized configuration for $\mathcal{T}'$ (c.f. Section 4.2 for definition), where the size $|\mathcal{T}'|$ is set as 300. We set $N = 4$ (c.f. Section 4.3 for definition) for high-resolution datasets and set $N = 1$ for datasets with resolution less than 64×64.

5.2. Main Results

High-resolution datasets. To explore the potential of our approach for real-world applications, we first conduct experiments to compare with the SOTA dataset distillation methods on Tiny-ImageNet and ImageNet-1K (including some subsets, e.g., ImageNet-100). Table 2 demonstrates that *our proposed method significantly outperforms existing methods or exhibits comparable results with large* `IPC` $= 10$ *and* 50. However, when `IPC` comes to 1, our approach struggles to effectively retain the information present in the original dataset, consequently leading to suboptimal outcomes.

Low-resolution datasets. To validate the robustness of our method across different-resolution datasets, we conduct more experiments on diminutive datasets such as CIFAR-10 and CIFAR-100 (see Table 2). Our RDED demonstrates superior performance compared to conventional methods, particularly in scenarios involving larger distilled datasets such as CIFAR-100 with `IPC` $= 50$. However, similar to high-resolution scenarios, its efficacy diminishes when confronted with smaller datasets.

Architecture		Time Cost (ms)	Peak Memory (GB)
ResNet-18	SRe^{2}L	2113.23	9.14
	Ours	**39.89**	**1.57**
MobileNet-V2	SRe^{2}L	3783.16	12.93
	Ours	**64.97**	**2.35**
EfficientNet-B0	SRe^{2}L	4412.42	11.92
	Ours	**73.16**	**2.34**

Table 3. **Synthesis time and memory consumption ImageNet-1K.** We use a single RTX-4090 GPU for all methods to conduct experiments on ImageNet-1K. Time Cost represents the consumption (ms) for each image when generating 100 images simultaneously. Following the official implementation of SRe^{2}L [44], the peak value of GPU memory usage is measured with a batch size of 100.

5.3. Efficiency Comparison

Table 3 distinctly showcases the *superior efficiency of our dataset distillation approach in comparison to previous methodologies, demonstrating a significant performance advantage over SOTA methods*. Notably, we present a flexible peak memory scope, allowing dynamic adjustments

Verifier\Observer		ResNet-18	EfficientNet-B0	MobileNet-V2
ResNet-18	SRe^{2}L	21.7 ± 0.6	11.7 ± 0.2	15.4 ± 0.2
	Ours	**42.3 ± 0.6**	**31.0 ± 0.1**	**40.4 ± 0.1**
MobileNet-V2	SRe^{2}L	19.7 ± 0.1	9.8 ± 0.4	10.2 ± 2.6
	Ours	**34.4 ± 0.2**	**24.1 ± 0.8**	**33.8 ± 0.6**
EfficientNet-B0	SRe^{2}L	25.2 ± 0.2	11.4 ± 2.5	20.5 ± 0.2
	Ours	**42.8 ± 0.5**	**33.3 ± 0.9**	**43.6 ± 0.2**

Table 4. **Evaluating ImageNet-1K top-1 accuracy on cross-architecture generalization.** Distill dataset with ResNet-18, EfficientNet-B0, and MobileNet-V2, and then versus transfer to each other architecture. We can not conduct experiments for SRe^{2}L when the model using for distillation without batch normalization, which necessitates [44]. All methods are evaluated with IPC = 10.

to the batch size without compromising performance. This efficiency is attributed to the fact that the primary memory consumption in our distillation procedure occurs exclusively during the scoring process of patches, while this process can be executed in parallel for images within a mini-batch[4]. Furthermore, the optimization-free nature of our RDED ensures that the distillation time for an image is solely dependent on the scoring cost determined by the pre-trained teacher model size.

5.4. Cross-architecture Generalization

To ensure the generalization capability of our distilled datasets, it is imperative to validate their effectiveness across multiple neural architectures not encountered when distilling datasets. Table 4 examines our RDED with the SOTA SRe^{2}L *and underscores the robust generalization ability of our method*. Our success stems from two key aspects:

- it enables high-realism distilled images (evidenced in [2]).
- it exhibits insensitivity to variations in the teacher model.

5.5. Ablation Study

The effectiveness of RDED hinges on two pivotal factors: the size $|\mathcal{T}_c'|$ of pre-selected subset $\mathcal{T}_c'$ (c.f. Section 4.2) and the number of patches N per distilled image (defined in Section 4.3). In this section, we set IPC = 10 and employ ResNet-18 as the network backbone to examine how these factors influence the diversity and realism of the distilled dataset (see Appendix D.4 for investigation on more factors).

On the impact of pre-selected subset size $|\mathcal{T}_c'|$. The experimental results in Figure 4 gives a more intuitive demonstration on the impact of $|\mathcal{T}_c'|$, alongside the discussion in Section 4.2:

- The performance abruptly drops when $|\mathcal{T}_c'|$ is equal to $N \times$ IPC, i.e., the Stage 1 in our Algorithm 1 becomes the simple uniform random sampling. In this case, the diversity is maximized but the realism is poor, thus resulting in catastrophically degraded performance.
- As $|\mathcal{T}_c'|$ continuously increases and exceeds a threshold, our framework collects more realistic images from $\mathcal{T}_c'$ but their patterns may be repeated, thus hurting diversity and consequent performance.

Therefore, a proper $|\mathcal{T}_c'|$ would balance the trade-off between data diversity and realism. In this instance, a value approaching 300 maximizes the total sum for our target in Equation (4), as indicated in Figure 4.

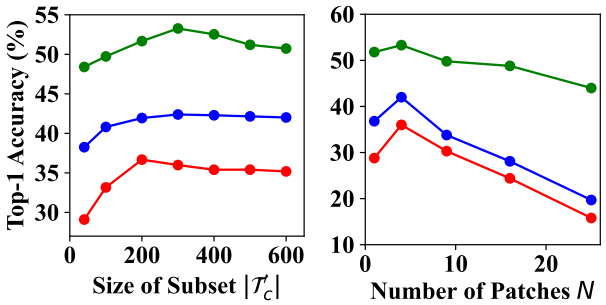


Figure 4. **Ablation study on $|\mathcal{T}_c'|$ and N**, i.e., the pre-selected subset size $\mathcal{T}_c'$ (left), and the number of patches N within each distilled image (right). The emerald ●, red ●, and blue ● denote ImageNet-10, ImageNet-100, and ImageNet-1K respectively.

On the impact of squeezing N patches into one distilled image. The number of patches N shares similar patterns as that of $|\mathcal{T}_c'|$. Specifically, though we can compress more patches from $\mathcal{T}$ into a distilled dataset $\mathcal{S}$ by increasing N increases to benefit the data diversity, it also results in a lower resolution for the source patches (see our explanation in Footnote 3), thus hurting the realism. Therefore, a proper number of patches N is important to achieve our objective in (3). Figure 4 showcases that the validation performance rises to the highest on selected three datasets when $N = 4$.

6. Conclusion

In this work, we introduce an optimization-free and efficient paradigm which successfully distills a dataset with IPC = 10 from the entirety of ImageNet-1K, concurrently achieving 42% top-1 validation accuracy with ResNet-18. Furthermore, our method exhibits robust cross-architecture generalization, surpassing SOTA method by a factor of 2× in performance.

Acknowledgement

We thank Xinyi Shang, Zexi Li and anonymous reviewers for their precious comments and feedback. This work was supported in part by the National Science and Technology Major Project (No. 2022ZD0115101), the Research Center for Industries of the Future (RCIF) at Westlake University, and the Westlake Education Foundation.

[4]Conventional optimization-based dataset distillation methods [1, 13, 44] have to synthesize a batch of images simultaneously to guarantee its overall quality.

References

[1] 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, 7, 8

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3739–3748, 2023. 1, 2, 3, 4, 8, 6

[3] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019. 5

[4] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. *Advances in Neural Information Processing Systems*, 35:810–822, 2022. 1, 2, 3

[5] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 1, 2, 3, 4, 6, 7

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 6, 2, 3

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5

[8] 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

[9] Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In *International Conference on Machine Learning*, pages 5988–6008. PMLR, 2022. 4

[10] Vitaly Feldman and Chiyuan Zhang. What neural networks memorize and why: Discovering the long tail via influence estimation. *Advances in Neural Information Processing Systems*, 33:2881–2891, 2020. 2

[11] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *biometrics*, 21:768–769, 1965. 4

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 1

[13] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*, 2023. 2, 6, 8, 3

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 6, 4

[15] Yang He, Lingao Xiao, Joey Tianyi Zhou, and Ivor Tsang. Multisize dataset condensation. *arXiv preprint arXiv:2403.06075*, 2024. 3

[16] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr, 2015. 1, 4

[17] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 1, 3

[18] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo 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, 6

[19] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[20] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 and cifar-100 datasets. *URl: https://www. cs. toronto. edu/kriz/cifar. html*, 6(1):1, 2009. 6

[21] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[22] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022. 4

[23] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. *Advances in Neural Information Processing Systems*, 35:13877–13891, 2022. 2

[24] Yi Ma, Doris Tsao, and Heung-Yeung Shum. On the principles of parsimony and self-consistency for the emergence of intelligence. *Frontiers of Information Technology & Electronic Engineering*, 23(9):1298–1323, 2022. 4

[25] Kristof Meding, Luca M Schulze Buschoff, Robert Geirhos, and Felix A Wichmann. Trivial or impossible–dichotomous data difficulty masks model differences (on imagenet and beyond). *arXiv preprint arXiv:2110.05922*, 2021. 2

[26] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[27] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021. 2, 4

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning

transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 3

[29] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 6, 4

[30] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948. 2

[31] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. *arXiv preprint arXiv:2311.17950*, 2023. 1

[32] Zhiqiang Shen and Eric Xing. A fast knowledge distillation framework for visual recognition. In *European Conference on Computer Vision*, pages 673–690. Springer, 2022. 6, 3

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[34] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022. 3, 4, 5

[35] Haoru Tan, Sitong Wu, Fei Du, Yukang Chen, Zhibin Wang, Fan Wang, and Xiaojuan Qi. Data pruning via moving-one-sample-out. *arXiv preprint arXiv:2310.14664*, 2023. 4, 5

[36] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 6, 4

[37] 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. 2

[38] 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. 1, 2

[39] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1, 2

[40] Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1121–1128, 2009. 3, 4, 6

[41] Sebastien C Wong, Adam Gatt, Victor Stamatescu, and Mark D McDonnell. Understanding data augmentation for classification: when to warp? In *2016 international conference on digital image computing: techniques and applications (DICTA)*, pages 1–6. IEEE, 2016. 3

[42] Yilun Xu, Shengjia Zhao, Jiaming Song, Russell Stewart, and Stefano Ermon. A theory of usable information under computational constraints. *arXiv preprint arXiv:2002.10689*, 2020. 2, 4, 1

[43] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 4

[44] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *arXiv preprint arXiv:2306.13092*, 2023. 1, 2, 3, 4, 6, 7, 8

[45] Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *arXiv preprint arXiv:2301.07014*, 2023. 1, 2, 3

[46] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 3

[47] Sangdoo Yun, Seong Joon Oh, Byeongho Heo, Dongyoon Han, Junsuk Choe, and Sanghyuk Chun. Re-labeling imagenet: from single to multi-labels, from global to localized labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2340–2350, 2021. 6

[48] Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. *arXiv preprint arXiv:1612.03928*, 2016. 5

[49] Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Dongkuan Xu. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11950–11959, 2023. 2

[50] 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, 3

[51] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020. 2, 3, 6

[52] 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. 1, 2, 6, 7, 3

[53] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *Advances in Neural Information Processing Systems*, 35:9813–9827, 2022. 2