Condensing Large-Scale Datasets Directly with Minimal Information Loss

Xinyi Shang[1*], Peng Sun[2,3*], Bei Shi[4,*], Zixuan Wang[2], and Tao Lin[3,†]

[1] University College London, United Kingdom
[2] Zhejiang University, China
[3] Westlake University, China
[4] University of Macau, China

Abstract. Recent advancements in scaling dataset distillation rely heavily on decoupled information extraction pipelines, comprising SQUEEZE, RECOVER, and RELABEL stages. Despite their scalability to large-scale datasets, these methods suffer from prohibitive computational overhead and poor cross-architecture generalization. In this paper, we reveal the root cause of these bottlenecks: the implicit dual-compression process, from data to model and back to images, inherently induces severe information loss. Crucially, we empirically and theoretically demonstrate that this loss creates a distribution shift that fundamentally compromises the widely adopted RELABEL strategy, transforming the pre-trained model into an unreliable labeler that yields sub-optimal labels. To overcome these critical flaws, we propose CIM, a novel, metric-driven framework that abandons the flawed dual-compression paradigm. Instead, CIM explicitly quantifies and minimizes the information gap between the original and synthetic datasets. By directly aligning the data distributions, our approach ensures high-fidelity information condensation and inherently satisfies the prerequisites for effective relabeling. Extensive experiments demonstrate that CIM establishes a new state-of-the-art. Notably, it distills ImageNet-1K at an `IPC` =10 in merely 80 minutes on a single RTX-4090 GPU, achieving an unprecedented 48.7% Top-1 accuracy on ResNet-18 and significantly outperforming previous SOTA approaches, such as NRR-DD and DELT, by 2.6% and 2.9%, respectively. Our code is available at `https://github.com/LINs-lab/CIM`.

Keywords: Dataset Distillation · Minimal Information Loss · Relabel

1 Introduction

Dataset distillation (DD) [45] aims to condense the knowledge from a massive training dataset into a remarkably small synthetic set, preserving comparable generalization performance. By substituting the original data with this compact proxy, dataset distillation drastically reduces training overhead and storage costs.

* Equal contribution.
† Corresponding author. Email: lintao@westlake.edu.cn

Consequently, it has emerged as an enabling technique for diverse applications, including continual learning [29, 56], neural architecture search [39, 44], and privacy-preserving learning [3, 8, 32, 47].

To scale dataset distillation to large-scale datasets like ImageNet-1K [7] and ImageNet-21K [28], a recent line of research focuses on information extraction pipelines [34, 36, 41, 42, 49]. Most notably, the pioneering SRe^{2}L [49] introduces a decoupled three-stage paradigm: (i) SQUEEZE the dataset information into a pre-trained model, (ii) RECOVER this information back into the image space to form synthetic data, and (iii) RELABEL the distilled images using the pre-trained model to inject label-space knowledge. By avoiding costly unrolled optimization, this paradigm yields strong empirical results.

Despite their success, existing extraction-based pipelines suffer from two critical bottlenecks: poor cross-architecture generalization and prohibitive computational overhead, particularly during the RECOVER stage. We identify the root cause as the severe information loss induced by the implicit *dual-compression* process: first from data to model parameters (SQUEEZE), and then from model back to synthetic images (RECOVER). Furthermore, while the RELABEL stage improves performance, our theoretical and empirical analyses reveal a hidden vulnerability: its efficacy is strictly conditional on distribution alignment. When the dual-compression process causes the synthetic samples to drift away from the original data distribution, the pre-trained model acts as an unreliable labeler, yielding sub-optimal labels. Therefore, ensuring distributional proximity is crucial to fully unlock the benefits of the RELABEL strategy.

Motivated by these insights, we abandon the flawed dual-compression paradigm and propose CIM, a novel framework that explicitly condenses information by minimizing the *information loss* between real and distilled datasets (CIM). Unlike prior extraction-based methods, which rely on complex inversion to recover informative synthetic images from a pre-trained model, CIM is explicitly *metric-driven*. Specifically, we formulate a principled metric to comprehensively quantify the information gap, and consequently the distribution shift, between the synthetic samples and their real counterparts. By directly optimizing the distilled set to minimize this gap within a unified objective, CIM ensures high-fidelity information retention and strict distribution alignment. This design not only eliminates the computationally expensive recovery process but also natively satisfies the conditions required for effective relabeling. **Our contributions are summarized as follows:**

1. We conduct a rigorous revisiting of information extraction-based dataset distillation, revealing that the inherent dual-compression process leads to severe information loss. This loss not only degrades cross-architecture generalization and computational efficiency but also causes a distribution shift that fundamentally compromises the widely adopted RELABEL strategy.
2. We introduce CIM, a novel, metric-driven framework that explicitly quantifies and minimizes the information gap between the original and distilled datasets, achieving more faithful, efficient, and aligned information condensation.

3. Extensive experiments verify that CIM establishes a new state-of-the-art across various scale datasets. Notably, our framework distills ImageNet-1K at $\texttt{IPC} = 10$ in merely 80 minutes on a single RTX-4090 GPU, achieving an unprecedented 48.7% Top-1 accuracy on ResNet-18, outperforming previous SOTA approaches, such as NRR-DD and DELT, by 2.6% and 2.9%, respectively.

2 Related Work

The primary aim of dataset distillation is to condense a large dataset into a smaller, yet highly representative subset, preserving its core semantic and statistical characteristics. Wang et al. [45] first introduce the dataset distillation as a bi-level meta-learning optimization problem. The outer loop aims at optimizing the meta-dataset, while the inner loop focuses on training models using the distilled dataset. Existing methods can be roughly divided into three paradigms for solving this complex bi-level optimization problem.

Uni-level optimization-based paradigm. Tackling this bi-level problem is complex, especially when optimizing proxy models via gradient descent, which involves unraveling an intricate computational graph. studies [24, 58] have proposed approximating model training using kernel ridge regression, which provides a closed-form solution for optimal weights, thereby reducing training costs and improving performance. Despite these advancements, such methods still struggle with extensive computational demands or limitations due to the approximations in convex relaxation.

Matching-based paradigm. Another strategy involves emulating the behaviors of the original dataset in the distilled one. They focus on minimizing disparities between surrogate models trained on both synthetic and original datasets. The key metrics for this are matching gradients [17,23,53,56], features [43], distribution [55, 57], and training trajectories [1, 5, 6, 11, 13, 50]. Trajectory and gradient matching, in particular, has shown impressive results with low IPC. However, these methods often tailor the distilled dataset to specific network architectures, limiting their generalizability. Cazenavette et al. [2] address this by proposing the GLaD that synthesizes more realistic images to enhance generalization. These methods incur substantial computational overhead due to the frequent calculation of discrepancies between the distilled and original datasets, requiring numerous iterations for optimization until convergence. Therefore, computational and memory challenges remain, particularly when scaling to large datasets.

Information extraction-based paradigm. The SRe^{2}L framework [49], as the first work efficiently scalable to ImageNet-1K, introduces a novel decoupled bi-level learning paradigm. This involves three stages: 1) Squeeze relevant information from the original dataset into a pre-trained model, 2) Recover this information into the image space, 3) Relabel the distilled images by using the pre-trained model to further distill knowledge into the label space. Its efficiency and effectiveness have garnered community attention, spurring a series of research efforts. Shao et al. [34] note that SRe^{2}L is limited to specific backbones and layers,

impacting the generalization of the distilled dataset. They advocate for using diverse backbones for more precise and effective distillation. Yin et al. [48] further enhance SRe^{2}L with curriculum data augmentation. [22] use Wasserstein distance to create more representative images. Sun et al. [41] introduce an optimization-free approach RDED that achieves notable diversity and realism in distilled datasets. Recently, DELT [36] mitigates the issue of low within-class diversity of SRe^{2}L by varying the number of iterations for different IPCs during the data synthesis phase. Moreover, NRR-DD [42] addresses the shortcoming of SRe^{2}L in often emphasizing instance-specific features by effectively capturing both class-general and instance-specific features. EDC [35] introduces a unified framework grounded in both empirical and theoretical foundations. Building on the distillation process introduced by SRe^{2}L [49], it incorporates two key enhancements: soft category-aware matching and dynamic adjustments to the learning rate schedule. These additions can further optimize the distillation process.

3 On the Pitfalls of Information Extraction Paradigm

In this section, we revisit the information extraction paradigm for dataset distillation in depth. We highlight two critical aspects: 1) The primary challenge is significant information loss, which hurts the quality and diversity in distilled images. 2) We empirically and theoretically demonstrate the critical attributes necessary for retaining the efficacy of RELABEL and seek an approach to effectively distill information from original images while adhering to these attributes.

3.1 Preliminary

Dataset distillation. Given a large-scale dataset $\mathcal{T} = \{\mathbf{x}_i, y_i\}_{i=1}^{|\mathcal{T}|}$ which consists of $|\mathcal{T}|$ samples, dataset distillation aims to synthesize a smaller set $\mathcal{S} = (\mathcal{S}_X, \mathcal{S}_Y) = \{\widetilde{\mathbf{x}}_j, \widetilde{y}_j\}_{j=1}^{|\mathcal{S}|}$ with $|\mathcal{S}|$ synthetic samples such that models trained on $\mathcal{T}$ will have similar performance as models trained on $\mathcal{S}$:

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}}[\ell\left(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y\right)] \simeq \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}}[\ell\left(\phi_{\boldsymbol{\theta}_{\mathcal{S}}}(\mathbf{x}), y\right)], \tag{1}$$

where $P_{\mathcal{D}}$ is the test real distribution, $\mathbf{x}$ is a data sample, ℓ is the loss function, i.e., cross-entropy loss. Here, $\boldsymbol{\theta}_{\mathcal{T}}$ and $\boldsymbol{\theta}_{\mathcal{S}}$ denotes the parameters of the neural network ϕ trained on $\mathcal{T}$ and $\mathcal{S}$, respectively.

A closer look at information extraction paradigm. As the first effective yet efficient solution to allow the dataset distillation on diverse-scale datasets such as ImageNet-1K [7], information extraction-based methods have attracted attention and inspired many subsequent works. These methods typically employ a three-stage distillation process, *indirectly* transferring information from the original to the distilled images. The first stage involves condensing information from the complete dataset into pre-trained neural network models via SQUEEZE, followed by the extraction of this information into distilled images using RECOVER [22,49]. Upon the distilled data samples, RELABEL applies a pre-trained model to further

distill knowledge into the label space. However, this paradigm encounters several key challenges:

1. The RECOVER stage *necessitates batch normalization* in pre-trained models [16] to align the statistical features between distilled and original images [22, 49].
2. *Significant information loss* during both SQUEEZE and RECOVER stages leads to distilled images with minimal content, adversely affecting performance, particularly in low `IPC` settings [41].
3. Distilled images often *exhibit unrealistic textures or semantics*, tailored to specific networks, thereby limiting their generalization ability [34].
4. Despite outperforming other paradigms (e.g., matching-based methods) in terms of efficiency [49], the RECOVER stage is still *computationally demanding*, requiring numerous optimization iterations [48].

In the meanwhile, the influence of RELABEL is under-explored [34, 49]. These challenges motivate us to explore a method to simultaneously address four key problems by *directly* condensing information from the original dataset for image distillation, abandoning traditional decoupled SQUEEZE and RECOVER stages, and to further explore the role of RELABEL.

3.2 Does RELABEL Always Help Distilled Images?

RELABEL has become a widely adopted and effective technique in dataset distillation, as demonstrated in recent works [33, 41, 49]. The core idea is to use a model pre-trained on the full real dataset to re-assign labels to the distilled images, and then train a student model on these relabeled distilled samples.

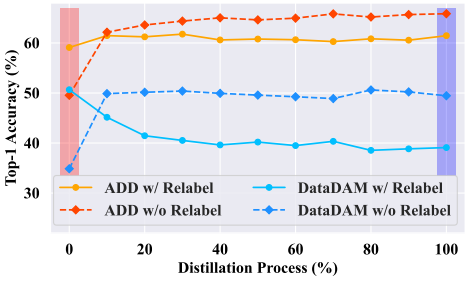


Fig. 1: Applying RELABEL to ADD [53] and DataDAM [30]. We evaluate the distilled images with `IPC` = 10 during the distillation process. The results indicate that RELABEL only assists the early-stage distilled datasets.

To evaluate the effect of RELABEL, we compare state-of-the-art DD methods with and without this strategy. As reported in Tab. 7 (App. E), removing RELABEL leads to a *substantial performance degradation*. For example, on ImageNet-1k, the accuracy of SRe^{2}L and G-VBSM drops to 1.1% and 0.8%, respectively, when RELABEL is disabled.

Given its effectiveness, a natural question is whether RELABEL can be used as a plug-and-play component for DD methods that do not originally rely on it. To answer this, we apply RELABEL to distilled images produced by representative non-RELABEL methods, such as ADD [53] and DataDAM [30]. The results are shown in Fig. 1. We observe that performance gains are mainly limited to the *very early* distillation stage (i.e., when the "distilled" images remain close to the

initial real images used for initialization). In contrast, as distillation proceeds, relabeling the increasingly optimized distilled images provides little benefit and can even hurt performance.

This behavior suggests a mismatch between the relabeling model and the evolved distilled samples. Specifically, the model used for RELABEL is trained exclusively on the original real dataset, whereas distilled images gradually deviate from the real-image distribution during optimization. Such a distribution shift alters semantic and/or textural cues, making the pre-trained relabeler less reliable on distilled samples (see App. K for qualitative evidence). Consequently, inaccurate relabels accumulate, and the downstream student model suffers.

A theoretical perspective. Beyond the above empirical observation, we theoretically reveal that a model well-trained on the original, undisturbed samples (i.e., the model used for RELABEL) may only provide suboptimal labels for shifted distilled samples. Concretely, standard DD typically starts from a subset of real samples $\{\mathbf{x}_j\}_{j=1}^{|\mathcal{S}|}$ drawn from the full dataset $\mathcal{T}$, and iteratively optimizes them into distilled samples $\{\widetilde{\mathbf{x}}_j\}_{j=1}^{|\mathcal{S}|}$ with $\widetilde{\mathbf{x}}_j = \mathbf{x}_j + \epsilon_j$, where ϵ_j denotes the learned perturbation. A shift occurs because the optimization updates ϵ_j may change both semantics and textures of $\mathbf{x}_j$, causing the distilled distribution to deviate from the original one.

Proposition 1. *For two original Gaussian distributions $\mathcal{N}_1(\mu_1, \sigma^2)$ and $\mathcal{N}_2(\mu_2, \sigma^2)$, we first define their shifted versions $\mathcal{N}_1^s(\mu_1 + s_1, \sigma^2)$ and $\mathcal{N}_2^s(\mu_2 + s_2, \sigma^2)$. Then, the optimal classification model f_{opt} for $\mathcal{N}_1^s$ and $\mathcal{N}_2^s$ achieves the following suboptimal classification accuracy for $\mathcal{N}_1$ and $\mathcal{N}_2$:*

$$P_{acc} = \frac{1}{2}\left(\Phi\left(\frac{\mu_2 + s_2 - \mu_1 + s_1}{2\sigma}\right) + 1 - \Phi\left(\frac{\mu_1 + s_1 - \mu_2 + s_2}{2\sigma}\right)\right) \tag{2}$$

$$P_{acc} \leq \frac{1}{2}\left(\Phi\left(\frac{\mu_2 - \mu_1}{2\sigma}\right) + 1 - \Phi\left(\frac{\mu_1 - \mu_2}{2\sigma}\right)\right) \tag{3}$$

Here, $\Phi(\cdot)$ is the cumulative distribution function (CDF) of the standard normal distribution.

That is, a relabeler trained on the original dataset may provide unreliable labels when applied to distribution-shifted distilled samples. Therefore, to maximize the efficacy of RELABEL, it is necessary to ensure the distilled data distribution remains close to that of the original real dataset. Only under such alignment can distilled samples be recognized and labeled correctly by the pre-trained relabeler.

Subset selection from real data. To achieve these goals, a straightforward approach involves directly selecting images from the original dataset to construct the distilled dataset. Numerous works [12, 41, 42, 46] have explored methods for selecting diverse and representative key samples from the original dataset to create a distilled dataset. These methods can ensure that the distilled data can be accurately identified by the pre-trained model and maximize the effectiveness of RELABEL. A notable contribution is RDED [41], which extracts key patches from each image based on *high realism scores* to construct a distilled dataset.

(a) SRe^{2}L [49] **(b)** G-VBSM [34] **(c)** RDED [40]

Fig. 2: Visualization of Feature Distributions for the Original and Distilled Datasets. The distilled datasets are optimized using state-of-the-art distillation techniques: SRe^{2}L [49], G-VBSM [34], and RDED [40]. Orange, green, and blue points depict the first three classes of CIFAR-10, while $\star$ points represent the corresponding distilled datasets with images per class (`IPC`) of 50. The lighter shades represent the original dataset.

To further validate whether the sample selected by RDED can achieve the distribution as the original dataset, we visualize the feature distributions [1] of the distilled dataset alongside those of the original dataset in Fig. 2. Specifically, we compare the state-of-the-art methods, including SRe^{2}L [49], G-VBSM [34], and RDED [41]. Among these, *RDED demonstrates the closest alignment with the feature distribution of the original dataset.* Additionally, as shown in Tab. 1 and Tab. 2, RDED achieves superior performance compared to the other two SOTA methods. Therefore, by default, CIM adopts the selection mechanism of RDED[2]. Note that the subset selection strategy itself is not the focus of this work. Our proposed CIM is selection-method-agnostic and can incorporate various selection strategies, as demonstrated in Tab. 6.

4 Methodology

The following section begins by introducing the formal definitions of effective information for a given sample and the resulting information gap between any two given samples. Furthermore, we propose a method based on minimizing the information gap between the selected sample subsets and the distilled samples to ensure the preservation of information integrity in the distilled set. The distillation process of our CIM is depicted in Fig. 3, with the detailed algorithm outlined in Alg. 1 in Appendix.

On the information gap of two samples. For the selected `IPC` subsets of key images, we aim to compress each of them into a more compact pixel space, i.e., distilled image $\widetilde{\mathbf{x}}$, thus forming $\mathcal{S}_X$. However, given the constraints of limited pixel space storage, using a naive solution—e.g., directly resizing and concatenating multiple images from a subset into one—results in a significant reduction in the fineness and detail of the original images. To effectively capture and condense

[1] Features are extracted using a model trained on the full dataset, followed by visualization with T-SNE [26].
[2] See App. C for details of RDED.

the salient information from the original samples into the distilled ones, we aim to enable each distilled sample $\widetilde{\mathbf{x}}$ to encapsulate the information of a selected subset from $\mathcal{T}_X$. We begin by formally defining the effective information of a data sample as follows.

Definition 1 (Observation-based Effective Information). *Let $\mathbf{x}_i$ represent a sample from any domain (e.g., image). Define an observer group $\mathcal{R} = \{\xi_j\}$, where each observer ξ_j is capable of extracting or interpreting features from $\mathbf{x}_i$, and $|\mathcal{R}| \geq 1$. The effective information of the sample $\mathbf{x}_i$, as observed by the group $\mathcal{R}$, is conceptualized as the distribution $\mathcal{P}_{\mathbf{x}_i|\mathcal{R}}(z)$. This distribution is formulated as:*

$$\mathcal{P}_{\mathbf{x}_i|\mathcal{R}}(z) := \{z|z = \xi_j(\mathbf{x}_i), \forall \xi_j \in \mathcal{R}\}, \tag{4}$$

where z denotes the set of features or interpretations extracted from $\mathbf{x}_i$ by an observer ξ_j within $\mathcal{R}$.

The Def. 1 posits that the effective information of a sample encompasses the set of its features as perceived or extracted by a diverse set of observers. Samples are considered to have similar effective information if they result in comparable feature sets across the observers in $\mathcal{R}$. We consequently define the effective information gap between two given samples $\mathbf{x}_i$ and $\mathbf{x}_j$.

Definition 2 (Pairwise Effective Information Gap). *Let $\mathbf{x}_i$ and $\mathbf{x}_j$ represent two samples from any domain, and given an observer group $\mathcal{R} = \{\xi_j\}$. This effective information gap is formulated as the difference between their effective information:*

$$I_G(\mathbf{x}_i, \mathbf{x}_j; \mathcal{R}) = \mathrm{D}_{\mathrm{KL}}(\mathcal{P}_{\mathbf{x}_i|\mathcal{R}} || \mathcal{P}_{\mathbf{x}_j|\mathcal{R}}). \tag{5}$$

Compressing information into distilled set. However, directly minimizing the information gap between distilled set $\mathcal{S}_X$ and original set $\mathcal{D}_X$ through Def. 2 is intractable, due to Eq. 5 is a sample-level metric that cannot be directly applied to the set. Thanks to the widely adapted data augmentation techniques $\mathcal{A}$ to enhance the diversity of each sample $\widetilde{\mathbf{x}}$ in practice, we relax the estimation of Eq. 5 through calculating the effective information gap between a set of N original samples $\{\mathbf{x}_i\}_{i=1}^N$ and a distilled sample $\widetilde{\mathbf{x}}_j$ using N augmented views of the distilled samples, i.e.,

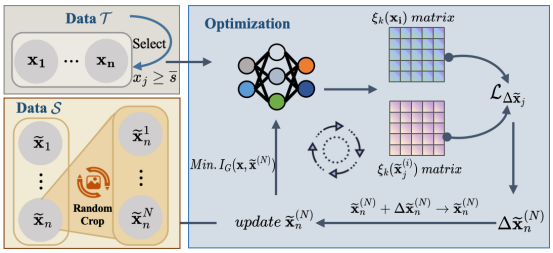


Fig. 3: Distillation Process of Our CIM. First, `IPC` subsets are selected from the original data $\mathcal{T}$, where each subset contains images denoted as $\{\mathbf{x}_j\}_{j=1}^N$; Then, for each image $\widetilde{\mathbf{x}}$ in the initial distilled data $\mathcal{S}$, the `RandomCrop` is applied to generate views $\{\widetilde{\mathbf{x}}^n\}_{n=1}^N$. The information gap $I_G(\mathbf{x}_j, \widetilde{\mathbf{x}}^n)$ is then minimized for each view. This process is iteratively performed for every distilled image $\widetilde{\mathbf{x}}$;

$$\mathbb{E}_{(\mathbf{x}_i, \widetilde{\mathbf{x}}_j^{(i)}) \sim (\{\mathbf{x}_i\}_{i=1}^N, \mathcal{A}(\widetilde{\mathbf{x}}_j))} I_G(\mathbf{x}_i, \widetilde{\mathbf{x}}_j^{(i)}; \mathcal{R}) \text{s.t.} \quad \mathcal{A}(\widetilde{\mathbf{x}}_j) = \{\widetilde{\mathbf{x}}_j^{(1)}, \widetilde{\mathbf{x}}_j^{(2)}, \ldots, \widetilde{\mathbf{x}}_j^{(N)}\}. \tag{6}$$

However, we still cannot directly distill sample $\widetilde{\mathbf{x}}_j$ through Eq. 6 due to the intractable KL divergence estimation in Eq. 5, and thus we derive Thm. 1 that bounds Eq. 5 to enable it computationally tractable.

Theorem 1 (Pairwise Effective Information Gap). *Let $\mathbf{x}_i$ and $\mathbf{x}_j$ represent two samples from any domain, and given an observer group $\mathcal{R} = \{\xi_j\}$. The effective information gap is upper-bounded as*

$$I_G(\mathbf{x}_i, \mathbf{x}_j; \mathcal{R}) \leq \mathbb{E}_{\xi_k \sim \mathcal{R}} \|\xi_k(\mathbf{x}_i) - \xi_k(\mathbf{x}_j)\|^2 \text{s.t.} \quad k \in [1, |\mathcal{R}|]\,. \tag{7}$$

Therefore, we can capture a distilled sample $\widetilde{\mathbf{x}}_j$ through combining Eq. 6 and Thm. 1, namely,

$$\arg\min_{\widetilde{\mathbf{x}}_j} \mathbb{E}_{(\mathbf{x}_i, \widetilde{\mathbf{x}}_j^{(i)}) \sim (\{\mathbf{x}_i\}_{i=1}^N, \mathcal{A}(\widetilde{\mathbf{x}}_j))} \mathbb{E}_{\xi_k \sim \mathcal{R}} \|\xi_k(\mathbf{x}_i) - \xi_k(\widetilde{\mathbf{x}}_j^{(i)})\|^2 \tag{8}$$

To reduce computational complexity, we utilize `RandomCrop` to generate the augmentations $\mathcal{A}(\widetilde{\mathbf{x}}_j)$ and concatenate resized real images to initialize distilled sample $\widetilde{\mathbf{x}}_j$. We further consider a specific scenario where the observer group consists of only one pre-trained model across various transformations[3]. This strategy allows us to employ multiple observers without requiring an excessive number of pre-trained models. Our loss function is defined as follows to achieve Eq. 8:

$$\mathcal{L}_{\Delta\widetilde{\mathbf{x}}_j} = \mathbb{E}_{(\mathbf{x}_i, \widetilde{\mathbf{x}}_j^{(i)}) \sim (\{\mathbf{x}_i\}_{i=1}^N, \mathcal{A}(\widetilde{\mathbf{x}}_j + \Delta\widetilde{\mathbf{x}}_j))} \mathbb{E}_{\zeta_k \sim \mathcal{G}} \left\| \zeta_k \circ \phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}_i) - \zeta_k \circ \phi_{\boldsymbol{\theta}_\mathcal{T}}(\widetilde{\mathbf{x}}_j^{(i)}) \right\|^2 \tag{9}$$

where $\mathcal{R} = \{\xi_k \mid \xi_k = \zeta_k \circ \phi_{\boldsymbol{\theta}_\mathcal{T}}, \forall \zeta_k \sim \mathcal{G}\}$ and $\mathcal{G}$ denotes the transformation group, $\mathcal{A}(\widetilde{\mathbf{x}}_j + \Delta\widetilde{\mathbf{x}}_j) = \{\widetilde{\mathbf{x}}_j^{(1)}, \widetilde{\mathbf{x}}_j^{(2)}, \ldots, \widetilde{\mathbf{x}}_j^{(N)}\}$. By minimizing Eq. 9, we find the optimal $\Delta\widetilde{\mathbf{x}}_j^\star$ and capture the image $\widetilde{\mathbf{x}}_j \leftarrow \widetilde{\mathbf{x}}_j + \Delta\widetilde{\mathbf{x}}_j^\star$ as the distilled image.

Balancing the semantic and textural information. Directly generating the distilled image $\widetilde{\mathbf{x}}_j$ through aligning its effective information with the original image set $\{\mathbf{x}_i\}_{i=1}^N$ in the last layer of model $\phi_{\boldsymbol{\theta}_\mathcal{T}}$ may lead to a substantial loss of texture information. The reason is that the model $\phi_{\boldsymbol{\theta}_\mathcal{T}}$ tends to extract semantic information from input images, which often results in a notable drop in the texture details. To balance between semantic richness and texture preservation in the distilled image $\widetilde{\mathbf{x}}_j$, we leverage intermediate model features instead of the last-layer logits, which helps in retaining more textural details (see Section 5.3 for the validation of its robustness).

Relabel across Various Transformations. We propose an improved relabeling strategy for our informative distilled images $\widetilde{\mathbf{x}}$, which provides more diverse and informative knowledge in the label space compared to the basic one-shot labeling approach. The standard Relabel technique [49], inspired by [52], suggests that a random image crop may contain a different object than the one originally labeled, leading to inaccurate or misleading training data. This

[3] In the context of image data, these transformations encompass a range of nonlinear and linear operations, including techniques for image data augmentation.

highlights the limitations of the one-shot labeling strategy in expressing sufficient knowledge for cropped images.

Our extended RELABEL approach considers a wider range of image transformations beyond random cropping. Similar to the soft labeling approach in [37], we generate transformed-view-level soft label $\widetilde{y}_k = \phi_{\boldsymbol{\theta}_\mathcal{T}}(\zeta_k(\widetilde{\mathbf{x}}))$, where $\zeta_k(\widetilde{\mathbf{x}})$ represents the k-th transformation applied to the distilled image $\widetilde{\mathbf{x}}$.

Therefore, we can train the model $\phi_{\boldsymbol{\theta}_\mathcal{S}}$ on the distilled data by minimizing:

$$\mathcal{L} = -\sum_j \sum_k \|\phi_{\boldsymbol{\theta}_\mathcal{S}}(\zeta_k(\widetilde{\mathbf{x}}_j)) - \widetilde{y}_{(j,k)}\|^2 . \tag{10}$$

The whole distillation process for an entire dataset by using our CIM is illustrated in Alg. 1 in Appendix.

5 Experiment

In this section, we evaluate the performance of our proposed CIM over various datasets and neural architectures. First, we demonstrate the superior results of CIM on real-world datasets, cross-architecture generalization and efficiency. We next perform comprehensive ablation studies to evaluate the impact of each component in our proposed method, as well as to analyze the influence of hyperparameter choices and subset selection strategies. Finally, we demonstrate the superior performance of our approach, CIM, in continual learning applications.

5.1 Experimental Setting

Datasets and neural network architectures. We conduct experiments on varying scales and resolutions of images.

- **Small-scale:** we evaluate on two datasets, including CIFAR-10 (32×32) [19] and CIFAR-100 (32×32) [18].
- **Large-scale:** we also use two large-scale high-resolution datasets including Tiny-ImageNet (64×64) [20] and ImageNet-1K (224×224) [7].

Following prior dataset distillation works [13, 49, 57], we employ ConvNet [13], ResNet-18 [14], and MobileNet-V2 [31], ViT-T/16 [9], ShuffleNet-v2-x2.0 [25], DenseNet-121 [15], as our backbone networks. Specifically, for ConvNet, we use Conv-3 on CIFAR-10/100 and use Conv-4 on Tiny-ImageNet and ImageNet-1K. More details about the used datasets and architectures can be found in App. G.

Baselines. We compare our method with several SOTA distillation methods that can scale to large high-resolution datasets, including G-VBSM [34], SRe^{2}L [49], RDED [41], CDA [48], WMDD [21], Teddy [51], CUDD [10], EDC [35], DWA [10], CV-DD [4], INFER [54], GIFT [33], NRR-DD [42], DELT [36] . The results of additional baseline methods, including DataDAM [30], ADD [53], IDM [57], CDA [48], WMDD [21], DATM [13], DREAM [23], and FreD [38], are presented in App. G. Comprehensive details regarding these approaches are also provided in the same appendix.

Table 1: Comparison with baseline approaches on CIFAR-10 ("CF-10") and CIFAR-100 ("CF-100"). In the table, **bold** indicates the best result. Underline means the second-best result. `IPC` refers to the Images Per Class within distilled datasets.

		ConvNet				ResNet-18				ResNet-50			
Dataset	IPC	G-VBSM	SRe2L	RDED	CIM (Ours)	G-VBSM	SRe2L	RDED	CIM (Ours)	G-VBSM	SRe2L	RDED	CIM (Ours)
CF-10	1	21.2 ± 0.2	22.2 ± 1.1	28.9 ± 0.4	**37.4 ± 0.4**	17.0 ± 1.0	19.9 ± 0.9	27.8 ± 0.4	**32.3 ± 0.2**	17.2 ± 0.9	20.2 ± 0.5	24.8 ± 0.5	**31.8 ± 0.7**
	10	38.6 ± 0.8	39.6 ± 0.4	56.0 ± 0.1	**61.4 ± 0.4**	36.3 ± 0.7	39.4 ± 0.9	47.3 ± 0.5	**66.2 ± 0.9**	33.8 ± 1.1	37.2 ± 0.6	45.1 ± 0.7	**62.9 ± 0.3**
	50	62.7 ± 0.5	57.7 ± 0.4	71.1 ± 0.2	**74.7 ± 0.2**	64.5 ± 0.6	62.8 ± 1.2	76.4 ± 0.4	**85.1 ± 0.3**	61.5 ± 0.6	61.6 ± 0.2	74.1 ± 0.6	**84.2 ± 0.3**
CF-100	1	13.4 ± 0.3	12.9 ± 0.1	21.8 ± 0.4	**27.4 ± 0.4**	13.4 ± 0.5	11.5 ± 0.4	4.6 ± 0.1	**31.1 ± 0.7**	12.6 ± 0.6	10.1 ± 0.1	4.5 ± 0.2	**28.1 ± 0.2**
	10	38.7 ± 0.8	34.2 ± 0.3	47.0 ± 0.3	**49.3 ± 0.3**	47.0 ± 0.4	42.7 ± 0.5	53.4 ± 0.3	**61.7 ± 0.2**	47.5 ± 0.5	44.2 ± 0.5	54.0 ± 0.3	**62.9 ± 0.4**
	50	53.8 ± 0.4	52.2 ± 0.3	**55.3 ± 0.2**	55.1 ± 0.2	60.0 ± 0.1	57.4 ± 0.2	64.0 ± 0.0	**67.0 ± 0.1**	62.2 ± 0.3	60.6 ± 0.2	65.8 ± 0.3	**67.9 ± 0.1**

Table 2: Comparison with the state-of-the-art dataset distillation methods on Tiny-ImageNet ("TN-IN") and ImageNet-1K ("IN-1K") on ResNet-18.

Dataset	IPC	SRe2L	G-VBSM	CDA	WMDD	RDED	Teddy	GIFT	NRR-DD	DELT	CIM (Ours)
TN-IN	1	13.5 ± 0.2	9.0 ± 0.1	-	7.6 ± 0.2	15.4 ± 0.6	-	15.9 ± 0.3	13.5 ± 0.2	9.3 ± 0.5	**25.1 ± 0.4**
	10	43.6 ± 0.5	37.7 ± 0.3	-	41.8 ± 0.1	48.4 ± 0.3	-	49.2 ± 0.1	45.2 ± 0.2	43.0 ± 0.1	**53.3±0.1**
	50	53.4 ± 0.3	52.2 ± 0.0	48.7 ± 0.1	59.4 ± 0.5	57.4 ± 0.2	55.2 ± 0.1	58.1 ± 0.1	**61.2 ± 0.1**	55.7 ± 0.5	58.6 ± 0.1
IN-1K	1	2.8 ± 0.2	2.1 ± 0.1	-	3.2 ± 0.3	5.9 ± 0.1	-	-	**11.6 ± 0.2**	-	7.3±0.3
	10	31.1 ± 0.1	35.7 ± 0.2	-	38.2 ± 0.2	41.1 ± 0.2	34.1 ± 0.1	43.2 ± 0.1	46.1 ± 0.0	45.8 ± 0.0	**48.7±0.2**
	50	49.5 ± 0.2	51.6 ± 0.2	53.5 ± 0.1	57.6 ± 0.5	55.3 ± 0.2	52.5 ± 0.1	56.5 ± 0.1	60.1 ± 0.1	59.2 ± 0.1	**60.4±0.1**
	100	54.3 ± 0.2	56.1 ± 0.1	58.0 ± 0.1	60.7 ± 0.0	58.6 ± 0.1	56.5 ± 0.0	59.3 ± 0.2	-	**62.4 ± 0.1**	**62.4±0.1**

Implementation details. By default, CIM employs the subset selection mechanism of RDED. Notably, the proposed CIM is selection-method-agnostic and can seamlessly integrate alternative selection strategies. All distilled datasets synthesized from these baselines are evaluated using the same post-training process. All the hyper-parameters used in our Alg. 1 are general, insensitive, and easy-implemented for all datasets and network architectures (c.f. Sec. 5.3 and App. I for validation). We employ a generalized configuration for $\mathcal{T}'$ (c.f. App. C for definition), where the size of subset $|\mathcal{T}'|$ is set as 300. We set the number $N = 4$ of images squeezed in a distilled image and number $M = 200$ of compression iteration (c.f. Alg. 1 for definition). More implementation details are provided in App. G.

5.2 Comparison with the SOTA Methods

Small-scale datasets. Following previous research [1, 6, 57], we set `IPC` to 1, 10, and 50 to compare with baselines on varying datasets and networks. As the results reported in Tab. 1, *our method CIM outperforms other methods on varying datasets and neural networks with different `IPC`*. It is noteworthy that prior information extraction-based solutions like SRe^{2}L struggle in scenarios involving small distilled datasets such as CIFAR-100 with `IPC` $= 1$ or CIFAR-10 with all `IPC`, further verifying our claims proposed in Sec. 3.1. Detailed comparison among more datasets and baselines can be found in App. H. Furthermore, we visualize the distilled images of various methods in App. K.

Large-scale datasets. To evaluate the practicality of CIM in real-world settings, we further conduct experiments on the large-scale and high-resolution benchmarks Tiny-ImageNet and ImageNet-1K. The corresponding results are reported in Tab. 2. In addition, we compare CIM with several state-of-the-art methods on ResNet-50, as shown in Tab. 3. Note that we exclude some methods listed in

Table 3: Comparison with baselines on ImageNet-1K using ResNet-50.

IPC	G-VBSM	SRe2L	RDED	Teddy	CUDD	CDA	EDC	DWA	INFER	CV-DD	CIM (Ours)
10	42.6 ± 0.4	38.3 ± 0.5	46.2 ± 0.3	37.2 ± 0.1	46.2 ± 0.3	–	54.1 ± 0.2	43.0 ± 0.1	38.3 ± 0.2	51.3 ± 0.2	**54.3 ± 0.4**
50	60.3 ± 0.1	58.4 ± 0.1	62.5 ± 0.1	58.5 ± 0.1	63.6 ± 0.2	61.3 ± 0.2	64.3 ± 0.2	62.3 ± 0.3	63.4 ± 0.2	63.9 ± 0.1	**65.9 ± 0.1**
100	64.1 ± 0.1	62.9 ± 0.1	65.5 ± 0.1	–	66.7 ± 0.1	65.1 ± 0.1	–	65.7 ± 0.1	–	–	**67.6 ± 0.1**

Table 4: Top-1 accuracy (%) on ImageNet-1k for cross-architecture generalization. We utilize ResNet-18 for distilling the original dataset. Subsequently, the distilled data is transferred to other architectures, with IPC = 10.

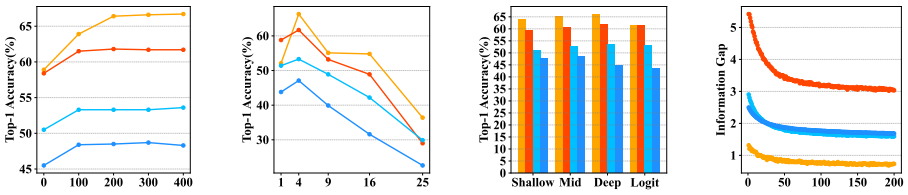
Verifier	EfficientNet-B0	ResNet-18	MobileNet-V2	ViT-T/16	ShuffleNet-v2-x2.0	DenseNet-121
SRe^{2}L	35.2 ± 0.0	35.0 ± 0.3	27.6 ± 0.0	3.2 ± 0.0	38.4 ± 0.3	43.2 ± 0.1
G-VBSM	31.5 ± 0.0	30.2 ± 1.2	26.0 ± 0.4	3.3 ± 0.1	31.8 ± 0.9	39.6 ± 0.1
RDED	41.1 ± 0.0	38.4 ± 0.7	34.8 ± 0.0	8.5 ± 0.1	43.2 ± 0.3	47.4 ± 0.3
Ours	**48.5 ± 0.2**	**48.7 ± 0.3**	**42.3 ± 0.2**	**10.8 ± 0.0**	**50.8 ± 0.2**	**54.4 ± 0.2**

Tab. 2 from the ResNet-50 comparison, since their distilled datasets are not available on larger networks, making a fair reproduction infeasible. Furthermore, beyond the commonly used setting of IPC =50, we also report results under larger budgets (e.g., IPC =100) to assess scalability. It is obvious that our CIM achieves the best performance on most scenarios, demonstrating the effectiveness.

Cross-architecture generalization. An important property of the distilled datasets is their good generalization capability across unseen architectural models. Here we evaluate the generalizability of our distilled datasets when IPC =10. As reported in Tab. 4, *our distilled dataset achieves the best performance on unseen networks*, which reflects the good generalizability of the data and labels distilled by our method. Our success stems from our CIM effectively keeps both textural and semantic information in distilled images.

Efficiency comparison. Efficiency is also a key factor during the distillation process. Here, we use a single RTX-4090 GPU for two methods to conduct experiments on Tiny-ImageNet. The reason why we mainly compare with SRe^{2}L and G-VBSM is based on their outstanding as efficient optimization-based methods currently. We evaluate the distillation efficiency by recording the run-time cost and peak GPU memory usage of distilling the image. As evidenced in Tab. 5, *our CIM achieves superior efficiency in comparison to SOTA methods*, with the exception of RDED, which benefits from an optimization-free paradigm [41], demonstrating a notable advantage of efficacy and efficiency. Significantly, our algorithm can offer a versatile peak memory capacity, enabling adjustments to batch size dynamically without sacrificing performance. This efficiency is attributed to the fact that our Alg. 1 can independently[4] optimize images, allowing us to distill them one by one. More comparisons are in App. H.

[4] Traditional distillation techniques necessitate the concurrent synthesis of a set of images to ensure collective quality [1, 13, 49]. This issue stems from the nature of these methods, which involve using a distilled dataset to approximate the original dataset with a similarity metric.



(a) Number M **(b)** Number N **(c)** Alignment Layer **(d)** Number K

Fig. 4: Ablation study on each component in our CIM. We evaluate our CIM with different number M of compression iterations (4a), number N of images squeezed in one distilled image (4b), feature alignment layer (4c), and the information gap with respect to the number K of iterations(4d). The yellow ●, red ●, blue ● and deep blue ● denote CIFAR-10, CIFAR-100, Tiny-ImageNet and ImageNet-1k respectively.

Table 5: Efficiency comparison on varying networks. Following SRe^{2}L, Time Cost is the consumption (s) when generating 100 images simultaneously, and the peak value of GPU memory usage is measured with a batch size of 100.

Architecture	Time Cost (s)				Peak Memory (GB)			
	SRe^{2}L	G-VBSM	RDED	Ours	SRe^{2}L	G-VBSM	RDED	Ours
Conv-4	51.68	259.84	1.682	13.02	1.36	4.94	0.74	0.65
ResNet-18	191.14	259.84	2.78	25.34	3.62	4.94	0.92	1.56
MobileNet-V2	114.05	259.84	4.07	18.81	1.27	4.94	0.29	0.64

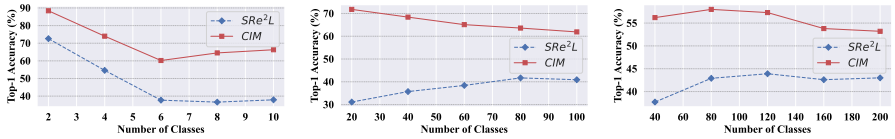
5.3 Ablation Study

In this section, we set the default $\texttt{IPC} = 10$ and employ ConvNet as the network backbone to examine how the components used in our CIM influence the quality of distilled dataset (see App. I for more investigation).

Influence of compression iteration number M. The number of distillation iterations, denoted as M, impacts two aspects: 1) A higher M enhances CIM's ability to generate higher-quality images; 2) A lower M ensures a faster execution of our Alg. 1. Consequently, choosing an optimal iteration number M represents a balance between quality and speed. As illustrated in Fig. 4a, an iteration count of $M = 200$ offers a well-rounded compromise for various datasets. Additionally, it is noteworthy that *our CIM exhibits robustness to variations in M*. Specifically, setting M beyond 200 yields negligible differences in performance.

Influence of size of compressed images N. Though we can compress more original images $\{\mathbf{x}_i\}_{i=1}^{N}$ into each distilled image $\widetilde{\mathbf{x}}$ by increasing N to benefit the feature diversity (c.f. Sec. 4), it also results in less information preserved from each original image $\mathbf{x}_i$. Fig. 4b showcase that *the validation performance rises to the highest on selected four datasets when $N = 4$.*

Why and how to choose the feature alignment layer? The experimental results in Fig. 4c illustrate the impact of the feature alignment layer, alongside the discussion in Sec. 4. As depicted in Fig. 4c, *high performance is often achieved through alignment at the middle layer.* This outcome likely stems from the varied information encoded at different network depths: shallow layers capture more textural details, whereas logit and deep layers are more adept at encoding semantic



(a) CIFAR-10 **(b)** CIFAR-100 **(c)** Tiny-ImageNet

Fig. 5: Application of continual learning on various datasets when IPC =10

Table 6: Comparison of CIM using various selection strategies with IPC = 10. By default, CIM uses the RDED selection method.

	Random	K-means	Herding	Ours
CIFAR-10	64.5 ± 0.2	65.5 ± 1.2	66.2 ± 0.8	**66.6 ± 0.1**
CIFAR-100	60.4 ± 0.3	60.4 ± 0.2	61.0 ± 0.1	**61.8 ± 0.1**
Tiny-ImageNet	51.4 ± 0.5	51.3 ± 0.2	51.3 ± 0.2	**53.2 ± 0.1**
ImageNet-1k	45.3 ± 0.0	44.8 ± 0.3	45.3 ± 0.1	**48.6 ± 0.2**

information. Thus, aligning at the middle layer enables the distilled dataset to achieve an optimal balance of these information types.

Influence of iteration number K on the information gap. Fig. 4d illustrates the effect of iteration count K on the information gap. An increased iteration count K effectively reduces the information gap, enabling the model to capture and retain more features from the original images in the distilled dataset. However, as K nears 200, further iterations offer minimal gains. This trend suggests that an appropriate choice of K balances efficient information retention with computational cost across various datasets.

Influence of subset selection. Importantly, the selection mechanism is not intrinsic to our framework. To assess alternative strategies, we replaced RDED's selection mechanism with approaches such as Random Selection, K-means Clustering [12], and Herding [46]. A ResNet-18 model was trained on datasets distilled using these various selection strategies, with results summarized in Tab. 6. Notably, the results indicate that even with Random Selection, our framework achieves competitive performance. Additional results for the case where IPC = 1 are provided in Tab. 15 in Appendix.

Application: Continual learning. Following prior studies [49,55] that leverage synthetic datasets in continual learning to assess the quality of synthetic data, we employ the GDumb framework [27] for continual learning setup. In comparison to SRe^{2}L, our study implements a 5-step class-incremental learning approach using ResNet-18 across CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets, with IPC = 10. The results of these experiments are depicted in Fig. 5. It is evident that *our results substantially improve upon the baseline methods.*

6 Conclusion

In this paper, we demonstrate that the primary limitation of current state-of-the-art dataset distillation methods at various scales is the issue of huge information

loss. To address this, we introduce the CIM technique, which effectively compresses critical data from images into distilled forms with minimal information loss. Our extensive experiments reveal that CIM markedly surpasses existing SOTA dataset distillation techniques across a range of dataset sizes and network architectures. Additionally, we highlight the efficiency of CIM by showcasing its ability to distill the ImageNet-1k dataset in just 80 minutes.

Acknowledgment

This work was supported in part by the National Science and Technology Major Project (No. 2022ZD0115101), Research Center for Industries of the Future (RCIF) at Westlake University, Westlake Education Foundation, and Westlake University Center for High-performance Computing.

References

1. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Dataset distillation by matching training trajectories. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4750–4759 (2022)
2. Cazenavette, G., Wang, T., Torralba, A., Efros, A.A., Zhu, J.Y.: Generalizing dataset distillation via deep generative prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3739–3748 (2023)
3. Chen, D., Kerkouche, R., Fritz, M.: Private set generation with discriminative information. Advances in Neural Information Processing Systems **35**, 14678–14690 (2022)
4. Cui, J., Li, Z., Ma, X., Bi, X., Luo, Y., Shen, Z.: Dataset distillation via committee voting. arXiv preprint arXiv:2501.07575 (2025)
5. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Dc-bench: Dataset condensation benchmark. Advances in Neural Information Processing Systems **35**, 810–822 (2022)
6. Cui, J., Wang, R., Si, S., Hsieh, C.J.: Scaling up dataset distillation to imagenet-1k with constant memory. In: International Conference on Machine Learning. pp. 6565–6590. PMLR (2023)
7. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
8. Dong, T., Zhao, B., Lyu, L.: Privacy for free: How does dataset condensation help privacy? In: International Conference on Machine Learning. pp. 5378–5396. PMLR (2022)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021)
10. Du, J., Hu, J., Huang, W., Zhou, J.T., et al.: Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. Advances in neural information processing systems **37**, 119443–119465 (2024)
11. Du, J., Jiang, Y., Tan, V.Y., Zhou, J.T., Li, H.: Minimizing the accumulated trajectory error to improve dataset distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3749–3758 (2023)

12. Forgy, E.W.: Cluster analysis of multivariate data: efficiency versus interpretability of classifications. biometrics **21**, 768–769 (1965)
13. Guo, Z., Wang, K., Cazenavette, G., Li, H., Zhang, K., You, Y.: Towards lossless dataset distillation via difficulty-aligned trajectory matching. arXiv preprint arXiv:2310.05773 (2023)
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
15. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks (2018)
16. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. pp. 448–456. pmlr (2015)
17. Kim, J.H., Kim, J., Oh, S.J., Yun, S., Song, H., Jeong, J., Ha, J.W., Song, H.O.: Dataset condensation via efficient synthetic-data parameterization. In: International Conference on Machine Learning. pp. 11102–11118. PMLR (2022)
18. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images (2009)
19. Krizhevsky, A., Nair, V., Hinton, G.: Cifar-10 and cifar-100 datasets. URl: https://www. cs. toronto. edu/kriz/cifar. html **6**(1), 1 (2009)
20. Le, Y., Yang, X.: Tiny imagenet visual recognition challenge. CS 231N **7**(7), 3 (2015)
21. Liu, H., Li, Y., Xing, T., Dalal, V., Li, L., He, J., Wang, H.: Dataset distillation via the wasserstein metric (2024)
22. Liu, H., Xing, T., Li, L., Dalal, V., He, J., Wang, H.: Dataset distillation via the wasserstein metric. arXiv preprint arXiv:2311.18531 (2023)
23. Liu, Y., Gu, J., Wang, K., Zhu, Z., Jiang, W., You, Y.: Dream: Efficient dataset distillation by representative matching. arXiv preprint arXiv:2302.14416 (2023)
24. Loo, N., Hasani, R., Amini, A., Rus, D.: Efficient dataset distillation using random feature approximation. Advances in Neural Information Processing Systems **35**, 13877–13891 (2022)
25. Ma, N., Zhang, X., Zheng, H.T., Sun, J.: Shufflenet v2: Practical guidelines for efficient cnn architecture design (2018)
26. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
27. Prabhu, A., Torr, P.H., Dokania, P.K.: Gdumb: A simple approach that questions our progress in continual learning. In: European conference on computer vision. pp. 524–540. Springer (2020)
28. Ridnik, T., Ben-Baruch, E., Noy, A., Zelnik-Manor, L.: Imagenet-21k pretraining for the masses. arXiv preprint arXiv:2104.10972 (2021)
29. Rosasco, A., Carta, A., Cossu, A., Lomonaco, V., Bacciu, D.: Distilled replay: Overcoming forgetting through synthetic samples. In: International Workshop on Continual Semi-Supervised Learning. pp. 104–117 (2021)
30. Sajedi, A., Khaki, S., Amjadian, E., Liu, L.Z., Lawryshyn, Y.A., Plataniotis, K.N.: Datadam: Efficient dataset distillation with attention matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17097–17107 (2023)
31. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4510–4520 (2018)

32. Shang, X., Lu, Y., Huang, G., Wang, H.: Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. arXiv preprint arXiv:2204.13399 (2022)
33. Shang, X., Sun, P., Lin, T.: Gift: Unlocking full potential of labels in distilled dataset at near-zero cost. In: International Conference on Learning Representations (2025)
34. Shao, S., Yin, Z., Zhou, M., Zhang, X., Shen, Z.: Generalized large-scale data condensation via various backbone and statistical matching. arXiv preprint arXiv:2311.17950 (2023)
35. Shao, S., Zhou, Z., Chen, H., Shen, Z.: Elucidating the design space of dataset condensation. In: Advances in neural information processing systems (2024)
36. Shen, Z., Sherif, A., Yin, Z., Shao, S.: Delt: A simple diversity-driven earlylate training for dataset distillation. CVPR (2025)
37. Shen, Z., Xing, E.: A fast knowledge distillation framework for visual recognition. In: European Conference on Computer Vision. pp. 673–690. Springer (2022)
38. Shin, D., Shin, S., Moon, I.c.: Frequency domain-based dataset distillation. In: Thirty-seventh Conference on Neural Information Processing Systems (2023)
39. Such, F.P., Rawal, A., Lehman, J., Stanley, K., Clune, J.: Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In: International Conference on Machine Learning. pp. 9206–9216 (2020)
40. Sun, P., Shi, B., Yu, D., Lin, T.: On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. arXiv preprint arXiv:2312.03526 (2023)
41. Sun, P., Shi, B., Yu, D., Lin, T.: On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
42. Tran, M.T., Le, T., Le, X.M., Do, T.T., Phung, D.: Enhancing dataset distillation via non-critical region refinement. CVPR (2025)
43. Wang, K., Zhao, B., Peng, X., Zhu, Z., Yang, S., Wang, S., Huang, G., Bilen, H., Wang, X., You, Y.: Cafe: Learning to condense dataset by aligning features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12196–12205 (2022)
44. Wang, R., Cheng, M., Chen, X., Tang, X., Hsieh, C.J.: Rethinking architecture selection in differentiable nas (2021)
45. Wang, T., Zhu, J.Y., Torralba, A., Efros, A.A.: Dataset distillation. arXiv preprint arXiv:1811.10959 (2018)
46. Welling, M.: Herding dynamical weights to learn. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 1121–1128 (2009)
47. Xiong, Y., Wang, R., Cheng, M., Yu, F., Hsieh, C.J.: Feddm: Iterative distribution matching for communication-efficient federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16323–16332 (2023)
48. Yin, Z., Shen, Z.: Dataset distillation in large data era. arXiv preprint arXiv:2311.18838 (2023)
49. Yin, Z., Xing, E., Shen, Z.: Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. arXiv preprint arXiv:2306.13092 (2023)
50. Yu, R., Liu, S., Wang, X.: Dataset distillation: A comprehensive review. arXiv preprint arXiv:2301.07014 (2023)
51. Yu, R., Liu, S., Ye, J., Wang, X.: Teddy: Efficient large-scale dataset distillation via taylor-approximated matching. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)

52. Yun, S., Oh, S.J., Heo, B., Han, D., Choe, J., Chun, S.: 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. pp. 2340–2350 (2021)
53. Zhang, L., Zhang, J., Lei, B., Mukherjee, S., Pan, X., Zhao, B., Ding, C., Li, Y., Xu, D.: Accelerating dataset distillation via model augmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11950–11959 (2023)
54. Zhang, X., Du, J., Liu, P., Zhou, J.T.: Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. ICLR (2025)
55. Zhao, B., Bilen, H.: Dataset condensation with distribution matching. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6514–6523 (2023)
56. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. arXiv preprint arXiv:2006.05929 (2020)
57. Zhao, G., Li, G., Qin, Y., Yu, Y.: Improved distribution matching for dataset condensation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7856–7865 (2023)
58. Zhou, Y., Nezhadarya, E., Ba, J.: Dataset distillation using neural feature regression. Advances in Neural Information Processing Systems **35**, 9813–9827 (2022)