Elucidating the Design Space of Dataset Condensation

Shitong Shao[†◇], **Zikai Zhou**[†◇], **Huanran Chen**[†‡], **Zhiqiang Shen**[†*]
[†] Mohamed bin Zayed University of AI, [‡] Tsinghua University
[◇] The Hong Kong University of Science and Technology (Guangzhou)
{1090784053sst,choukai003}@gmail.com, huanran_chen@outlook.com
zhiqiang.shen@mbzuai.ac.ae, *: Corresponding author

Abstract

Dataset condensation, a concept within *data-centric learning*, aims to efficiently transfer critical attributes from an original dataset to a synthetic version, meanwhile maintaining both diversity and realism of syntheses. This approach can significantly improve model training efficiency and is also adaptable for multiple application areas. Previous methods in dataset condensation have faced several challenges: some incur high computational costs which limit scalability to larger datasets (*e.g.,* MTT, DREAM, and TESLA), while others are restricted to less optimal design spaces, which could hinder potential improvements, especially in smaller datasets (*e.g.,* SRe^{2}L, G-VBSM, and RDED). To address these limitations, we propose a comprehensive designing-centric framework that includes specific, effective strategies like implementing soft category-aware matching, adjusting the learning rate schedule and applying small batch-size. These strategies are grounded in both empirical evidence and theoretical backing. Our resulting approach, **E**lucidate **D**ataset **C**ondensation (**EDC**), establishes a benchmark for both small and large-scale dataset condensation. In our testing, EDC achieves state-of-the-art accuracy, reaching 48.6% on ImageNet-1k with a ResNet-18 model at an IPC of 10, which corresponds to a compression ratio of 0.78%. This performance surpasses those of SRe^{2}L, G-VBSM, and RDED by margins of 27.3%, 17.2%, and 6.6%, respectively.

1 Introduction

Dataset condensation, also known as dataset distillation, has emerged in response to the ever-increasing training demands of advanced deep learning models (He et al., 2016a,b; Brown et al., 2020). This task addresses the challenge of requiring massive amount of data to train high-precision models while also being bounded by resource constraints (Dosovitskiy et al., 2020; Shao et al., 2024). In the conventional setup of this problem, the original dataset acts as a "teacher", distilling and preserving essential information into a smaller, surrogate "student" dataset. The ultimate goal of this technique is to achieve comparable performance of models trained on the original and condensed datasets from scratch. This task has become popular in various downstream applications, including continual learning (Masarczyk and Tautkute, 2020; Sangermano et al., 2022; Zhao and Bilen, 2021), neural architecture search (Such et al., 2020; Zhao and Bilen, 2023; Zhao et al., 2021), and training-free network slimming (Liu et al., 2017).

However, the common solution in traditional dataset distillation methods of bi-level optimization requires prohibitively expensive computation, which limits the practical usage, as in prior works (Cazenavette et al., 2022; Sajedi et al., 2023; Liu et al., 2023a). This has become more severe particularly when being applied to large-scale datasets like ImageNet-1k (Russakovsky et al., 2015). In response, the uni-level optimization paradigm has gained significant attention as an alternative solution, with recent contributions from the research community (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2023) highlighting its applicability. These methods primarily leverage the rich

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-3146

and extensive information from static, pre-trained observer models, to facilitate a more streamlined optimization process for synthesizing a condensed dataset without the need to adjust other parameters (*e.g.,* those within the observer models). While uni-level optimization has demonstrated remarkable performance on large datasets, it has yet to achieve the competitive accuracy levels seen with classical methods on small-scale datasets like CIFAR-10/100 (Krizhevsky et al., 2009). Moreover, the recently proposed training-free method RDED (Sun et al., 2024) outperforms training-based methods in efficiency and maintains effectiveness, yet it overlooks the potential information incompleteness due to the lack of optimization on syntheses. Also, some simple but promising skills (*e.g.,* smoothing learning rate schedule) that could enhance performance have not been well-explored in the existing literature. We observe that a performance improvement of 16.2% in RDED comes from these techniques in this paper rather than the proposed data synthesis approach.

These drawbacks show the constraints of previous methods in several respects, highlighting the need for a thorough investigation and assessment of potential limitations in prior frameworks. In contrast to earlier strategies that targeted one or a few specific improvements, our approach systematically examines all possible facets and integrates them into our comprehensive framework. To establish a strong framework, we carefully analyze all potential deficiencies in different stages of the data synthesis, soft label generation, and post-evaluation stages during dataset condensation, resulting in an extensive exploration of the design space on both large-scale and small-scale datasets. As a result, we introduce **E**lucidate **D**ataset **C**ondensation (**EDC**), which includes a range of concrete and effective enhancement skills for dataset condensation (refer to Fig. 1). For instance, *soft category-aware matching (*❷*)* ensures consistent category representation between the original and condensed data batches for more precise matching. Overall, EDC not only achieves state-of-the-art performance on CIFAR-10, CIFAR-100, Tiny-ImageNet, ImageNet-10, and ImageNet-1k, using only half of the computational cost compared to the *baseline* G-VBSM, but it also provides in-depth both empirical and theoretical insights and explanations that affirm the soundness of our design decisions. Our code is available at: https://github.com/shaoshitong/EDC.

2 Dataset Condensation

Preliminary. Dataset condensation involves generating a synthetic dataset $\mathcal{D}^{\mathcal{S}} := \{\mathbf{x}_i^{\mathcal{S}}, \mathbf{y}_i^{\mathcal{S}}\}_{i=1}^{|\mathcal{D}^{\mathcal{S}}|}$ consisting of images $\mathcal{X}^{\mathcal{S}}$ and labels $\mathcal{Y}^{\mathcal{S}}$, designed to be as informative as the original dataset $\mathcal{D}^{\mathcal{T}} := \{\mathbf{x}_i^{\mathcal{T}}, \mathbf{y}_i^{\mathcal{T}}\}_{i=1}^{|\mathcal{D}^{\mathcal{T}}|}$, which includes images $\mathcal{X}^{\mathcal{T}}$ and labels $\mathcal{Y}^{\mathcal{T}}$. The synthetic dataset $\mathcal{D}^{\mathcal{S}}$ is substantially smaller in size than $\mathcal{D}^{\mathcal{T}}$ ($|\mathcal{D}^{\mathcal{S}}| \ll |\mathcal{D}^{\mathcal{T}}|$). The goal of this process is to maintain the critical attributes of $\mathcal{D}^{\mathcal{T}}$ to ensure robust or comparable performance during evaluations on real test protocol $\mathcal{P}_{\mathcal{D}}$.

$$\arg\min \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{P}_{\mathcal{D}}}[\ell_{\textbf{eval}}(\mathbf{x},\mathbf{y},\phi^*)], \quad \text{where } \phi^* = \arg\min_{\phi} \mathbb{E}_{(\mathbf{x}_i^{\mathcal{S}},\mathbf{y}_i^{\mathcal{S}})\sim\mathcal{D}^{\mathcal{S}}}[\ell(\phi(\mathbf{x}_i^{\mathcal{S}}),\mathbf{y}_i^{\mathcal{S}})]. \tag{1}$$

Here, $\ell_{\textbf{eval}}(\cdot,\cdot,\phi^*)$ represents the evaluation loss function, such as cross-entropy loss, which is parameterized by the neural network ϕ^* that has been optimized from the distilled dataset $\mathcal{D}^{\mathcal{S}}$. The data synthesis process primarily determines the quality of the distilled datasets, which transfers desirable knowledge from $\mathcal{D}^{\mathcal{T}}$ to $\mathcal{D}^{\mathcal{S}}$ through various matching mechanisms, such as trajectory matching (Cazenavette et al., 2022), gradient matching (Zhao et al., 2021), distribution matching (Zhao and Bilen, 2023) and generalized matching (Shao et al., 2023).

Small-scale vs. Large-scale Dataset Condensation/Distillation. Traditional dataset condensation algorithms, as referenced in studies such as (Wang et al., 2018; Cazenavette et al., 2022; Cui et al., 2023; Wang et al., 2022; Nguyen et al., 2020), encounter computational challenges and are generally confined to small-scale datasets like CIFAR-10/100 (Krizhevsky et al., 2009), or larger datasets with limited class diversity, such as ImageNette (Cazenavette et al., 2022) and ImageNet-10 (Kim et al., 2022). The primary inefficiency of these methods stems from their reliance on a bi-level optimization framework, which involves alternating updates between the synthetic dataset and the observer model utilized for distillation. This approach not only heavily depends on the model's intrinsic ability but also limits the versatility of the distilled datasets in generalizing across different architectures. In contrast, the uni-level optimization strategy, noted for its efficiency and enhanced performance on the regular 224×224 scale of ImageNet-1k in recent research (Yin et al., 2023; Shao et al., 2023; Yin and Shen, 2024), shows reduced effectiveness in smaller-scale datasets due to the massive optimization-based iterations required in the data synthesis process without a direct connection to actual data. Recent new methods in training-free distillation paradigms, such as in (Sun et al., 2024;

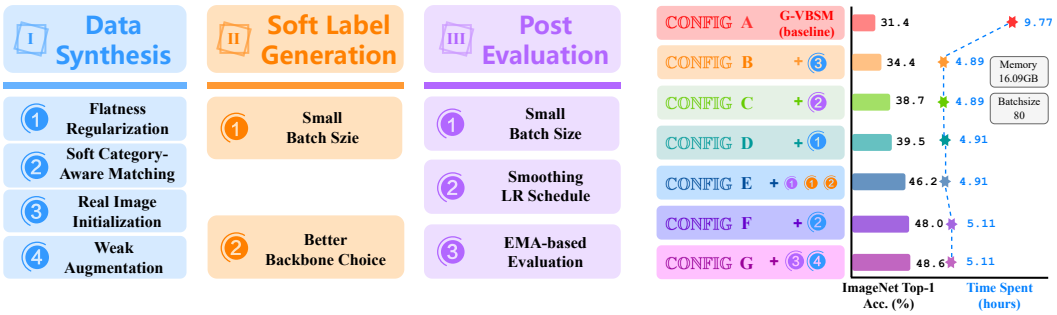


Figure 1: **Illustration of Elucidating Dataset Condensation (EDC). Left:** The overall of our better design choices in dataset condensation on ImageNet-1k. **Right:** The evaluation performance and data synthesis required time of different configurations on ResNet-18 with IPC 10. Our integral EDC refers to CONFIG G.

Zhou et al., 2023), offer advancements in efficiency. However, these methods compromise data privacy by sharing original data and do not leverage statistical information from observer models to enhance the capability of synthetic data, thereby restraining their potential in a real environment.

Generalized Data Synthesis Paradigm. We consistently describe algorithms (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2023; Sun et al., 2024) that efficiently conduct data synthesis on ImageNet-1k as "generalized data synthesis" as these methods are applicable for both small and large-scale datasets. This direction usually avoids the inefficient bi-level optimization and includes both image and label synthesis phases. Note that several recent works (Zhang et al., 2024a,b; Deng et al., 2024), particularly DANCE (Zhang et al., 2024a), can also effectively be applied to ImageNet-1k, but these methods lack enhancements in soft label generation and post-evaluation. Specifically, generalized data synthesis involves first generating highly condensed images followed by acquiring soft labels through predictions from a pre-trained model. The evaluation process resembles knowledge distillation (Hinton et al., 2015), aiming to transfer knowledge from a teacher to a student model (Gou et al., 2021; Hinton et al., 2015). The primary distinction between the training-dependent (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2023) and training-free paradigm (Sun et al., 2024) centers on their approach to data synthesis. In detail, the training-dependent paradigm employs *Statistical Matching (SM)* to extract pertinent information from the entire dataset.

$$\begin{aligned} \mathcal{L}_{\textbf{syn}} &= ||p(\mu|\mathcal{X}^{\mathcal{S}}) - p(\mu|\mathcal{X}^{\mathcal{T}})||_2 + ||p(\sigma^2|\mathcal{X}^{\mathcal{S}}) - p(\sigma^2|\mathcal{X}^{\mathcal{T}})||_2, \ s.t.\ \mathcal{L}_{\textbf{syn}} \sim \mathbb{S}_{\text{match}}, \\ \mathcal{X}^{\mathcal{S}*} &= \arg\min_{\mathcal{X}^{\mathcal{S}}} \mathbb{E}_{\mathcal{L}_{\textbf{syn}} \sim \mathbb{S}_{\text{match}}}[\mathcal{L}_{\textbf{syn}}(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})], \end{aligned} \tag{2}$$

where $\mathbb{S}_{\text{match}}$ represents the extensive collection of statistical matching operators, which operate across a variety of network architectures and layers as described by (Shao et al., 2023). Here, μ and σ^2 are defined as the mean and variance, respectively. For more detailed theoretical insights, please refer to Definition 3.1. The training-free approach, as discussed in (Sun et al., 2024; Zhou et al., 2023), employs a direct reconstruction method for the original dataset, aiming to generate simplified representations of images.

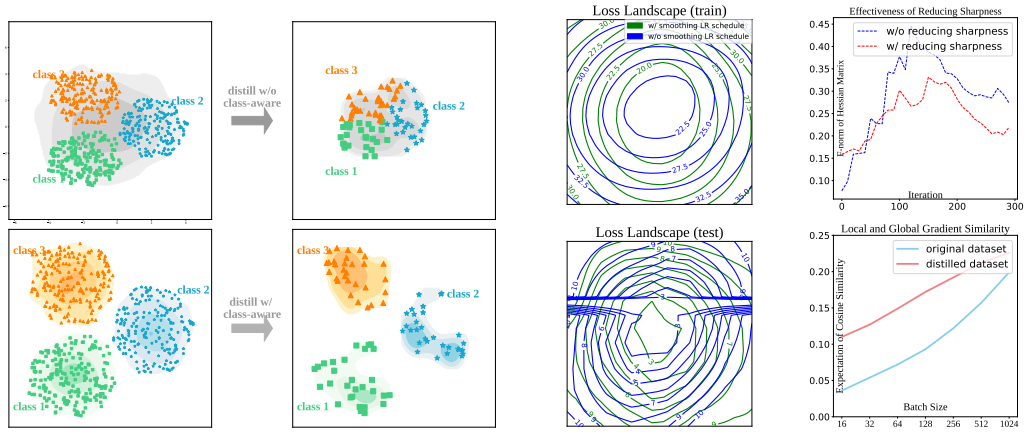
$$\mathcal{X}^{\mathcal{S}} = \bigcup_{i=1}^{\mathbf{C}} \mathcal{X}_i^{\mathcal{S}}, \ \mathcal{X}_i^{\mathcal{S}} = \{\mathbf{x}_j^i = \text{concat}(\{\tilde{\mathbf{x}}_k\}_{k=1}^{N} \subset \mathcal{X}_i^{\mathcal{T}})\}_{j=1}^{\text{IPC}}, \tag{3}$$

where $\mathbf{C}$ denotes the number of classes, $\text{concat}(\cdot)$ represents the concatenation operator, $\mathcal{X}_i^{\mathcal{S}}$ signifies the set of condensed images belonging to the i-th class, and $\mathcal{X}_i^{\mathcal{T}}$ corresponds to the set of original images of the i-th class. It is important to note that the default settings for N are 1 and 4, as specified in the works (Zhou et al., 2023) and (Sun et al., 2024), respectively. Using one or more observer models, denoted as $\{\phi_i\}_{i=1}^{N}$, we then derive the soft labels $\mathcal{Y}^{\mathcal{S}}$ from the condensed image set $\mathcal{X}^{\mathcal{S}}$.

$$\mathcal{Y}^{\mathcal{S}} = \bigcup_{\mathbf{x}_i^{\mathcal{S}} \subset \mathcal{X}^{\mathcal{S}}} \frac{1}{N} \sum_{i=1}^{N} \phi_i(\mathbf{x}_i^{\mathcal{S}}). \tag{4}$$

This plug-and-play component, as outlined in SRe^{2}L (Yin et al., 2023) and IDC (Kim et al., 2022), plays a crucial role for enhancing the generalization ability of the distilled dataset $\mathcal{D}^{\mathcal{S}}$.

https://doi.org/10.52202/079017-3146



(a) (b) (c)

Figure 2: **(a):** Illustration of soft category-aware matching (②) using a Gaussian distribution in $\mathbb{R}^2$. **(b):** The effect of employing smoothing LR schedule (②) on loss landscape sharpness reduction. **(c) top:** The role of flatness regularization (①) in reducing the Frobenius norm of the Hessian matrix driven by data synthesis iteration. **(c) bottom:** Cosine similarity comparison between local gradients (obtained from original and distilled datasets via random batch selection) and the global gradient (obtained from gradient accumulation).

3 Improved Design Choices

Design choices in data synthesis, soft label generation, and post-evaluation significantly influence the generalization capabilities of condensed datasets. Effective strategies for small-scale datasets are well-explored, yet these approaches are less examined for large-scale datasets. We first delineate the limitations of existing algorithms' design choices on ImageNet-1k. We then propose solutions, providing experimental results as shown in Fig. 1. For most design choices, we offer both theoretical analysis and empirical insights to facilitate a thorough understanding, as detailed in Sec. 3.2.

3.1 Limitations of Prior Methods

Lacking Realism (***solved by*** ③**).** Training-dependent condensation algorithms for datasets, particularly those employed for large-scale datasets, typically initiate the optimization process using Gaussian noise inputs (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2023). This initial choice complicates the optimization process and often results in the generation of synthetic images that do not exhibit high levels of realism. The limitations in visualization associated with previous approaches are detailed in Appendix F.

Coarse-grained Matching Mechanism (***solved by*** ②**).** The *Statistical Matching (SM)*-based pipeline (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2023) computes the global mean and variance by aggregating samples across all categories and uses these statistical parameters for matching purposes. However, this strategy exhibits two critical drawbacks: it does not account for the domain discrepancies among different categories, and it fails to preserve the integrity of category-specific information across the original and condensed samples within each batch. These limitations result in a coarse-grained matching approach that diminishes the accuracy of the matching process.

Overly Sharp of Loss Landscape (***solved by*** ① ***and*** ③**).** The optimization objective $\mathcal{L}(\theta)$ can be expanded through a second-order Taylor expansion as $\mathcal{L}(\theta^*)+(\theta-\theta^*)^{\mathrm{T}}\nabla_\theta\mathcal{L}(\theta^*)+(\theta-\theta^*)^{\mathrm{T}}\mathbf{H}(\theta-\theta^*)$, with an upper bound of $\mathcal{L}(\theta^*)+||\mathbf{H}||_{\mathbf{F}}\mathbb{E}[||\theta-\theta^*||_2^2]$ upon model convergence (Chen et al., 2024). However, earlier training-dependent condensation algorithms neglect to minimize the Frobenius norm of the Hessian matrix $\mathbf{H}$ to obtain a flat loss landscape for enhancing its generalization capability through sharpness-aware minimization theory (Foret et al., 2020; Chen et al., 2022). Please see Appendix C for more formal information.

Irrational Hyperparameter Settings (***solved by*** ④**,** ①**,** ②**,** ① ***and*** ②**).** RDED (Sun et al., 2024) adopts a smoothing LR schedule (②) and (Liu et al., 2023b; Yin and Shen, 2024; Sun et al., 2024)

https://doi.org/10.52202/079017-3146

use a reduced batch size (❶❶) for post-evaluation on the full 224×224 ImageNet-1k. These changes, although critical, lack detailed explanations and impact assessments in the existing literature. Our empirical analysis highlights a remarkable impact on performance: absent these modifications, RDED achieves only 25.8% accuracy on ResNet18 with IPC 10. With these modifications, however, accuracy jumps to 42.0%. In contrast, SRe^{2}L and G-VBSM do not incorporate such strategies in their experimental frameworks. This work aims to fill the gap by providing the first comprehensive empirical analysis and ablation study on the effects of these and similar improvements in the field.

3.2 Our Solutions

To address these limitations described above, we explore the design space and elaborately present a range of optimal solutions at both empirical and theoretical levels, as illustrated in Fig. 1.

Real Image Initialization (❸). Intuitively, using real images instead of Gaussian noise for data initialization during the data synthesis phase is a practical and effective strategy. As shown in Fig. 3, this method significantly improves the realism of the condensed dataset and simplifies the optimization process, thus enhancing the synthesized dataset's ability to generalize in post-evaluation tests. Additionally, we incorporate considerations of information density and efficiency by employing a training-free condensed dataset (e.g., RDED) for initialization at the start of the synthesis process. According to Theorem 3.1, based on optimal transport theory, the cost of transporting from a Gaussian distribution to the original data distribution is higher than using the training-free condensed distribution as the initial reference. This advantage also allows us to reduce the number of iterations needed to achieve results to half of those required by our baseline G-VBSM model, significantly boosting synthesis efficiency.

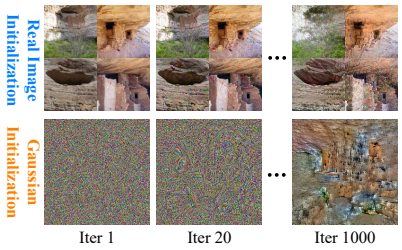


Figure 3: Comparison between real image initialization and random initialization.

Theorem 3.1. *(proof in Appendix B.1) Considering samples $\mathcal{X}^{\mathcal{S}}_{real}$, $\mathcal{X}^{\mathcal{S}}_{free}$, and $\mathcal{X}^{\mathcal{S}}_{random}$ from the original data, training-free condensed (e.g., RDED), and Gaussian distributions, respectively, let us assume a cost function defined in optimal transport theory that satisfies $\mathbb{E}[c(a-b)] \propto 1/I(Law(a), Law(b))$. Under this assumption, it follows that $\mathbb{E}[c(\mathcal{X}^{\mathcal{S}}_{real} - \mathcal{X}^{\mathcal{S}}_{free})] \leq \mathbb{E}[c(\mathcal{X}^{\mathcal{S}}_{real} - \mathcal{X}^{\mathcal{S}}_{random})]$.*

Soft Category-Aware Matching (❷). Previous dataset condensation methods (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2023) based on the *Statistical Matching* (SM) framework have shown satisfactory results predominantly when the data follows a unimodal distribution (*e.g.,* a single Gaussian). This limitation is illustrated with a simple example in Fig. 2 (a). Typically, datasets consist of multiple classes with significant variations among their class distributions. Traditional SM-based methods compress data by collectively processing all samples, thus neglecting the differences between classes. As shown in the top part of Fig. 2 (a), this method enhances information density but also creates a big mismatch between the condensed source distribution $\mathcal{X}^{\mathcal{S}}$ and the target distribution $\mathcal{X}^{\mathcal{T}}$. To tackle this problem, we propose the use of a Gaussian Mixture Model (GMM) to effectively approximate any complex distribution. This solution is theoretically justifiable by the Tauberian Theorem under certain conditions (detailed proof is provided in Appendix B.2). In light of this, we define two specific approaches to *Statistical Matching*:

Sketch Definition 3.1. *(formal definition in Appendix B.2) Given N random samples $\{x_i\}_{i=1}^{N}$ with an unknown distribution $p_{mix}(x)$, we define two forms to statistical matching.* ***Form (1):*** *involves synthesizing M distilled samples $\{y_i\}_{i=1}^{M}$, where $M \ll N$, ensuring that the variances and means of both $\{x_i\}_{i=1}^{N}$ and $\{y_i\}_{i=1}^{M}$ are consistent.* ***Form (2):*** *treats $p_{mix}(x)$ as a GMM with $\mathbf{C}$ components. For random samples $\{x_i^j\}_{i=1}^{N_j}$ ($\sum_j N_j = N$) within each component c_j, we synthesize M_j ($\sum_j M_j = M$) distilled samples $\{y_i^j\}_{i=1}^{M_j}$, where $M_j \ll N_j$, to maintain the consistency of variances and means between $\{x_i^j\}_{i=1}^{N_j}$ and $\{y_i^j\}_{i=1}^{M_j}$.*

In general, SRe^{2}L, CDA, and G-VBSM are all categorized under **Form (1)**, as shown in Fig. 2 (a) at the top, which leads to coarse-grained matching. According to Fig. 2 (a) at the bottom, transitioning to **Form (2)** is identified as a practical and appropriate alternative. However, our empirical result indicates that exclusive reliance on **Form (2)** yields a synthesized dataset that lacks sufficient information density. Consequently, we propose a hybrid method that effectively integrates

https://doi.org/10.52202/079017-3146

Form (1) and **Form (2)** using a weighted average, which we term soft category-aware matching.

$$\begin{aligned}\mathcal{L}'_{\textbf{syn}} &= \alpha||p(\mu|\mathcal{X}^{\mathcal{S}}) - p(\mu|\mathcal{X}^{\mathcal{T}})||_2 + ||p(\sigma^2|\mathcal{X}^{\mathcal{S}}) - p(\sigma^2|\mathcal{X}^{\mathcal{T}})||_2 \quad \#\text{Form (1)} \\ &+ (1-\alpha)\sum_{i}^{\mathbf{C}} p(c_i)\left[||p(\mu|\mathcal{X}^{\mathcal{S}}, c_i) - p(\mu|\mathcal{X}^{\mathcal{T}}, c_i)||_2 + ||p(\sigma^2|\mathcal{X}^{\mathcal{S}}, c_i) - p(\sigma^2|\mathcal{X}^{\mathcal{T}}, c_i)||_2\right], \quad \#\text{Form (2)}\end{aligned} \tag{5}$$

where $\mathbf{C}$ represents the total number of components, c_i indicates the i-th component within a GMM, and α is a coefficient for adjusting the balance. The modified loss function $\mathcal{L}'_{\textbf{syn}}$ is designed to effectively regulate the information density of $\mathcal{X}^{\mathcal{S}}$ and to align the distribution of $\mathcal{X}^{\mathcal{S}}$ with that of $\mathcal{X}^{\mathcal{T}}$. Operationally, each category in the original dataset is mapped to a distinct component in the GMM framework. Particularly, when $\alpha = 1$, the sophisticated category-aware matching described by $\mathcal{L}'_{\textbf{syn}}$ in Eq. 5 simplifies to the basic statistical matching defined by $\mathcal{L}_{\textbf{syn}}$ in Eq. 2.

Theorem 3.2. *(proofs in Theorems B.5, B.7, B.8 and Corollary B.6) Given the original data distribution $p_{mix}(x)$, and define condensed samples as x and y in **Form (1)** and **Form (2)** with their distributions characterized by P and Q. Subsequently, it follows that (i) $\mathbb{E}[x] \equiv \mathbb{E}[y]$, (ii) $\mathbb{D}[x] \equiv \mathbb{D}[y]$, (iii) $\mathcal{H}(P) - \frac{1}{2}\left[\log(\mathbb{E}[\mathbb{D}[y^j]] + \mathbb{D}[\mathbb{E}[y^j]]) - \mathbb{E}[\log(\mathbb{D}[y^j])]\right] \leq \mathcal{H}(Q) \leq \mathcal{H}(P) + \frac{1}{4}\mathbb{E}_{(i,j)\sim\prod[\mathbf{C},\mathbf{C}]}\left[\frac{(\mathbb{E}[y^i]-\mathbb{E}[y^j])^2(\mathbb{D}[y^i]+\mathbb{D}[y^j])}{\mathbb{D}[y^i]\mathbb{D}[y^j]}\right]$ and (iv) $D_{KL}[p_{mix}||P] \leq \mathbb{E}_{i\sim\mathcal{U}[1,\ldots,\mathbf{C}]}\mathbb{E}_{j\sim\mathcal{U}[1,\ldots,\mathbf{C}]}\frac{\mathbb{E}[y^j]^2}{\mathbb{D}[y^i]}$ and $D_{KL}[p_{mix}||Q] = 0$.*

We further analyze the properties of distributions P and Q as in **Form (1)** and **Form (2)**. According to parts (i) and (ii) of Theorem 3.2, Q retains the same variance and mean as P. Regarding diversity, part (iii) of Theorem 3.2 states that the entropy $\mathcal{H}(\cdot)$ of P and Q is equivalent, $\mathcal{H}(P) \equiv \mathcal{H}(Q)$, provided the mean and variance of all components in the GMM are uniform, suggesting a single Gaussian profile. Absent this condition, there is no guarantee that $\mathcal{H}(P)$ and $\mathcal{H}(Q)$ will consistently increase or decrease. These findings underscore the advantages of using GMM, especially when the initial data conforms to an unimodal distribution, thus aligning the mean, variance, and entropy of distributions P and Q in the reduced dataset. Moreover, even in diverse scenarios, the mean, variance, and entropy of Q tend to remain stable. Furthermore, when the original dataset exhibits a more complex bimodal distribution and the parameters of the Gaussian components are precisely estimated, utilizing GMM can effectively reduce the Kullback-Leibler divergence between the mixed original distribution p_{mix} and Q to near zero. In contrast, the divergence $D_{\text{KL}}[p_{\text{mix}}||P]$ always maintains a non-zero upper bound, as noted in part (iv) of Theorem 3.2. Therefore, by modulating the weight α in Eq. 5, we can derive an optimally balanced solution that minimizes loss in data characteristics while maximizing fidelity between the synthesized and original distributions.

Flatness Regularization (1) and EMA-based Evaluation (3). Choices 1 and 3 are utilized to ensure flat loss landscapes during the stages of data synthesis and post-evaluation, respectively.

During the data synthesis phase, the use of sharpness-aware minimization (SAM) algorithms is beneficial for reducing the sharpness of the loss landscape, as presented in prior research (Foret et al., 2020; Du et al., 2022; Bahri et al., 2021). Nonetheless, traditional SAM approaches, as detailed in Eq. 29 in the Appendix, generally double the computational load due to their two-stage parameter update process. This increase in computational demand is often impractical during data synthesis. Inspired by MESA (Du et al., 2022), which achieves sharpness-aware training without additional computational overhead through self-distillation, we introduce a lightweight flatness regularization approach for implementing SAM during data synthesis. This method utilizes a teacher dataset, $\mathcal{X}^{\mathcal{S}}_{\textbf{EMA}}$, maintained via exponential moving average (EMA). The newly formulated optimization goal aims to obtain a flat loss landscape in the following manner:

$$\mathcal{L}_{\textbf{FR}} = \mathbb{E}_{\mathcal{L}_{\text{syn}}\sim\mathbb{S}_{\text{match}}}[\mathcal{L}_{\textbf{syn}}(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{S}}_{\textbf{EMA}})], \ \mathcal{X}^{\mathcal{S}}_{\textbf{EMA}} = \beta\mathcal{X}^{\mathcal{S}}_{\textbf{EMA}} + (1-\beta)\mathcal{X}^{\mathcal{S}}, \tag{6}$$

where β is the weighting coefficient, which is empirically set to 0.99 in our experiments. The detailed derivation of Eq. 7 is in Appendix E. And the critical theoretical result is articulated as follows:

Theorem 3.3. *(proof in Appendix E) The optimization objective $\mathcal{L}_{\textbf{FR}}$ can ensure sharpness-aware minimization within a ρ-ball for each point along a straight path between $\mathcal{X}^{\mathcal{S}}$ and $\mathcal{X}^{\mathcal{S}}_{\textbf{EMA}}$.*

This indicates that the primary optimization goal of $\mathcal{L}_{\textbf{FR}}$ deviates somewhat from that of traditional SAM-based algorithms, which are designed to achieve a flat loss landscape around $\mathcal{X}^{\mathcal{S}}$. The constraint on flatness needs to ensure that the first-order term of the Taylor expansion equals zero, indicating normal model convergence. However, our exploratory experiments found that despite the good

Dataset	IPC	ResNet-18				ResNet-50		ResNet-101		MobileNet-V2
		SRe^{2}L	G-VBSM	RDED	EDC (Ours)	G-VBSM	EDC (Ours)	RDED	EDC (Ours)	EDC (Ours)
CIFAR-10	1	-	-	22.9 ± 0.4	32.6 ± 0.1	-	30.6 ± 0.4	-	26.1 ± 0.2	20.2 ± 0.4
	10	27.2 ± 0.4	53.5 ± 0.6	37.1 ± 0.3	79.1 ± 0.3	-	76.0 ± 0.3	-	67.1 ± 0.5	42.0 ± 0.4
	50	47.5 ± 0.5	59.2 ± 0.4	62.1 ± 0.1	87.0 ± 0.1	-	86.9 ± 0.0	-	85.8 ± 0.1	70.8 ± 0.2
CIFAR-100	1	2.0 ± 0.2	25.9 ± 0.5	11.0 ± 0.3	39.7 ± 0.1	-	36.1 ± 0.5	-	32.3 ± 0.3	10.6 ± 0.3
	10	31.6 ± 0.5	59.5 ± 0.4	42.6 ± 0.2	63.7 ± 0.3	-	62.1 ± 0.1	-	61.7 ± 0.1	44.3 ± 0.4
	50	49.5 ± 0.3	65.0 ± 0.5	62.6 ± 0.1	68.6 ± 0.2	-	69.4 ± 0.3	-	68.5 ± 0.1	59.5 ± 0.1
Tiny-ImageNet	1	-	-	9.7 ± 0.4	39.2 ± 0.4	-	35.9 ± 0.2	3.8 ± 0.1	40.6 ± 0.3	18.8 ± 0.1
	10	-	-	41.9 ± 0.2	51.2 ± 0.5	-	50.2 ± 0.3	22.9 ± 3.3	51.6 ± 0.2	40.6 ± 0.6
	50	41.1 ± 0.4	47.6 ± 0.3	58.2 ± 0.1	57.2 ± 0.2	48.7 ± 0.2	58.8 ± 0.4	41.2 ± 0.4	58.6 ± 0.1	50.7 ± 0.1
ImageNet-10	1	-	-	24.9 ± 0.5	45.2 ± 0.2	-	38.2 ± 0.1	21.7 ± 1.3	36.4 ± 0.1	36.4 ± 0.3
	10	-	-	53.3 ± 0.1	63.4 ± 0.2	-	62.4 ± 0.1	45.5 ± 1.7	59.8 ± 0.1	54.2 ± 0.1
	50	-	-	75.5 ± 0.5	82.2 ± 0.1	-	80.8 ± 0.2	71.4 ± 0.2	80.8 ± 0.0	80.2 ± 0.2
ImageNet-1k	1	-	-	6.6 ± 0.2	12.8 ± 0.1	-	13.3 ± 0.3	5.9 ± 0.4	12.2 ± 0.2	8.4 ± 0.3
	10	21.3 ± 0.6	31.4 ± 0.5	42.0 ± 0.1	48.6 ± 0.3	35.4 ± 0.8	54.1 ± 0.2	48.3 ± 1.0	51.7 ± 0.3	45.0 ± 0.2
	50	46.8 ± 0.2	51.8 ± 0.4	56.5 ± 0.1	58.0 ± 0.2	58.7 ± 0.3	64.3 ± 0.2	61.2 ± 0.4	64.9 ± 0.2	57.8 ± 0.1

Table 1: **Comparison with the SOTA baseline dataset condensation methods.** SRe^{2}L and RDED utilize ResNet-18 for data synthesis, whereas G-VBSM and EDC leverage various backbones for this purpose.

IPC	Method	ResNet-18	ResNet-50	ResNet-101	MobileNet-V2	EfficientNet-B0	DeiT-Tiny	Swin-Tiny	ConvNext-Tiny	ShuffleNet-V2
10	RDED	42.0	46.0	48.3	34.4	42.8	14.0	29.2	48.3	19.4
	EDC (Ours)	48.6	54.1	51.7	45.0	51.1	18.4	38.3	54.4	29.8
	$+\Delta$	6.6	8.1	3.4	10.6	8.3	4.4	9.1	6.1	10.4
20	RDED	45.6	57.6	58.0	41.3	48.1	22.1	44.6	54.0	20.7
	EDC (Ours)	52.0	58.2	60.0	48.6	55.6	24.0	49.6	61.4	33.0
	$+\Delta$	6.4	0.6	2.0	7.3	7.5	1.9	5.0	7.4	12.3
30	RDED	49.9	59.4	58.1	44.9	54.1	30.5	47.7	62.1	23.5
	EDC (Ours)	55.0	61.5	60.3	53.8	58.4	46.5	59.1	63.9	41.1
	$+\Delta$	5.1	2.1	2.2	8.9	4.3	16.0	11.4	1.8	17.6
40	RDED	53.9	61.8	60.1	50.3	56.3	43.7	58.1	63.7	27.7
	EDC (Ours)	56.4	62.2	62.3	54.7	59.7	51.9	61.1	65.2	44.7
	$+\Delta$	2.5	0.4	2.2	4.4	3.4	8.2	3.0	1.5	17.0
50	RDED	56.5	63.7	61.2	53.9	57.6	44.5	56.9	65.4	30.9
	EDC (Ours)	58.0	64.3	64.9	57.8	60.9	55.0	63.3	66.6	45.7
	$+\Delta$	1.5	0.6	3.7	3.9	3.3	10.5	6.4	1.2	14.8

Table 2: **Cross-architecture generalization comparison with different IPCs on ImageNet-1k.** RDED refers to the latest SOTA method on ImageNet-1k and $+\Delta$ stands for the improvement for each architecture.

performance of EDC, the loss of statistical matching at the end of data synthesis still fluctuated significantly and did not reach zero. As a result, we choose to apply flatness regularization exclusively to the logits of the observer model, since the cross-entropy loss for these can more straightforwardly reach zero.

$$\mathcal{L}'_{\textbf{FR}} = D_{\text{KL}}(\text{softmax}(\phi(\mathcal{X}^{\mathcal{S}})/\tau)||\text{softmax}(\phi(\mathcal{X}^{\mathcal{S}}_{\textbf{EMA}})/\tau)),\ \mathcal{X}^{\mathcal{S}}_{\textbf{EMA}} = \beta\mathcal{X}^{\mathcal{S}}_{\textbf{EMA}} + (1-\beta)\mathcal{X}^{\mathcal{S}}, \tag{7}$$

where softmax$(\cdot)$, τ and ϕ represent the softmax operator, the temperature coefficient and the pre-trained observer model, respectively. As illustrated in Fig. 2 (c) top, it is evident that $\mathcal{L}'_{\textbf{FR}}$ significantly lowers the Frobenius norm of the Hessian matrix relative to standard training, thus confirming its efficacy in pushing a flatter loss landscape.

In post-evaluation, we observe that a method analogous to $\mathcal{L}'_{\textbf{FR}}$ employing SAM does not lead to appreciable performance improvements. This result is likely due to the limited sample size of the condensed dataset, which hinders the model's ability to fully converge post-training, thereby undermining the advantages of flatness regularization. Conversely, the integration of an EMA-updated model as the validated model noticeably stabilizes performance variations during evaluations. We term this strategy EMA-based evaluation and apply it across all benchmark experiments.

Smoothing Learning Rate (LR) Schedule (❷) and Smaller Batch Size (❶❶). Here, we introduce two effective strategies for post-evaluation training. Firstly, it is crucial to clarify and distinguish between standard or conventional deep model training and post-evaluation in the context of dataset condensation. Specifically, (1) in dataset condensation, the limited number of samples in $\mathcal{X}^{\mathcal{S}}$ results in fewer training iterations per epoch, typically leading to underfitting; and (2) the gradient of a random batch from $\mathcal{X}^{\mathcal{S}}$ aligns more closely with the global gradient than that from a random batch in $\mathcal{X}^{\mathcal{T}}$. To support the latter observation, we utilize a ResNet-18 model with randomly initialized parameters to calculate the gradient of a random batch and assess the cosine similarity with the global gradient of $\mathcal{X}^{\mathcal{T}}$. After conducting over 100 iterations of this procedure, the average cosine similarity is consistently higher between $\mathcal{X}^{\mathcal{S}}$ and the global gradient than with $\mathcal{X}^{\mathcal{T}}$, indicating a greater similarity and reduced sensitivity to batch size fluctuations. Our findings further illustrate that the gradient from a random batch in $\mathcal{X}^{\mathcal{S}}$ effectively approximates the global gradient, as shown in Fig. 2 (c) bottom. Given this, the inaccurate gradient direction problem introduced by the small batch

https://doi.org/10.52202/079017-3146

Design Choices	ζ	ResNet-18	ResNet-50	ResNet-101
CONFIG C	1.0	34.4	36.8	42.0
CONFIG C	1.5	38.7	42.0	46.3
CONFIG C	2.0	38.8	45.8	47.9
CONFIG C	2.5	39.0	44.6	46.0
CONFIG C	3.0	38.8	45.6	46.2

Design Choices	ResNet-18	ResNet-50	ResNet-101
RDED	25.8	32.7	34.8
RDED+(❶❶❷)	42.3	48.4	47.0
G-VBSM+(❸)	34.4	36.8	42.0
G-VBSM+(❸❷)	38.8	45.8	47.9
G-VBSM+(❸❷❶❶)	45.0	51.6	48.1

Table 3: **Ablation studies on ImageNet-1k with IPC 10. Left:** Explore the influence of the slowdown coefficient ζ with CONFIG C. **Right:** Evaluate the effectiveness of real image initialization (❸), smoothing LR schedule (❷) and smaller batch size (❶❶) with $\zeta = 2$.

Design Choices	Loss Type	Loss Weight	ζ	β	τ	ResNet-18	ResNet-50	DenseNet-121
CONFIG C	-	-	1.5	-	-	38.7	42.0	40.6
CONFIG D	$\mathcal{L}_{\text{FR}}$	0.025	1.5	0.999	4	38.8	43.2	40.3
CONFIG D	$\mathcal{L}_{\text{FR}}$	0.25	1.5	0.999	4	37.9	43.5	40.3
CONFIG D	$\mathcal{L}_{\text{FR}}$	2.5	1.5	0.999	4	31.7	37.0	32.9
CONFIG D	$\mathcal{L}_{\text{FR}}$	0.25	1.5	0.99	4	39.0	43.3	40.2
CONFIG D	$\mathcal{L}'_{\text{FR}}$	0.25	1.5	0.99	4	39.5	44.1	41.9
CONFIG D	$\mathcal{L}'_{\text{FR}}$	0.25	1.5	0.99	1	38.9	43.5	40.7
CONFIG D	vanilla SAM	0.25	1.5	-	-	38.8	44.0	41.2

Table 4: **Ablation studies on ImageNet-1k with IPC 10.** Investigate the potential effects of several factors, including loss type, loss weight, β, and τ, amid flatness regularization (❶).

size becomes less problematic. Instead, using a small batch size effectively increases the number of iterations, thereby helping prevent model under-convergence.

To optimize the training with condensed samples, we implement a smoothed LR schedule that moderates the learning rate reduction throughout the training duration. This approach helps avoid early convergence to suboptimal minima, thereby enhancing the model's generalization capabilities. The mathematical formulation of this schedule is given by $\mu(i) = \frac{1+\cos(i\pi/\zeta N)}{2}$, where i represents the current epoch, N is the total number of epochs, $\mu(i)$ is the learning rate for the i-th epoch, and ζ is the deceleration factor. Notably, a ζ value of 1 corresponds to a typical cosine learning rate schedule, whereas setting ζ to 2 improves performance metrics from 34.4% to 38.7% and effectively moderates loss landscape sharpness during post-evaluation.

Weak Augmentation (❹) and Better Backbone Choice (❷). The principal role of these two design decisions is to address the flawed settings in the *baseline* G-VBSM. The key finding reveals that the minimum area threshold for cropping during data synthesis was overly restrictive, thereby diminishing the quality of the condensed dataset. To rectify this, we implement mild augmentations to increase this minimum cropping threshold, thereby improving the dataset condensation's ability to generalize. Additionally, we substitute the computationally demanding EfficientNet-B0 with more streamlined AlexNet for generating soft labels on ImageNet-1k, a change we refer to as an improved backbone selection. This modification maintains the performance without degradation. More details on the ablation studies for mild augmentation and improved backbone selection are in Appendix G.

4 Experiments

To validate the effectiveness of our proposed EDC, we conduct comparative experiments across various datasets, including ImageNet-1k (Russakovsky et al., 2015), ImageNet-10 (Kim et al., 2022), Tiny-ImageNet (Tavanaei, 2020), CIFAR-100 (Krizhevsky et al., 2009), and CIFAR-10 (Krizhevsky et al., 2009). Additionally, we explore cross-architecture generalization and ablation studies on ImageNet-1k. All experiments are conducted using 4× RTX 4090 GPUs. Due to space constraints, detailed descriptions of the hyperparameter settings, additional ablation studies, and visualizations of synthesized images are provided in the Appendix A.1, G, and H, respectively.

Network Architectures. Following prior dataset condensation work (Yin et al., 2023; Yin and Shen, 2024; Shao et al., 2023; Sun et al., 2024), our comparison uses ResNet-{18, 50, 101} (He et al., 2016a) as our verified models. We also extend our evaluation to include MobileNet-V2 (Sandler et al., 2018) in Table 1 and explore cross-architecture generalization further with recently advanced backbones such as DeiT-Tiny (Touvron et al., 2021) and Swin-Tiny (Liu et al., 2021) (detailed in Table 2).

Baselines. We compare our work with several recent state-of-the-art methods, including SRe^{2}L (Yin et al., 2023), G-VBSM (Shao et al., 2023), and RDED (Sun et al., 2024) to assess broader practical

Design Choices	α	ζ	Weak Augmentation Scale=(0.5,1.0)	EMA-based Evaluation EMA Rate=0.99	ResNet-18	ResNet-50	ResNet-101
CONFIG F	0.00	2.0	✗	✗	46.2	53.2	49.5
CONFIG F	0.00	2.0	✓	✗	46.7	53.7	49.4
CONFIG F	0.00	2.0	✓	✓	46.9	53.8	48.5
CONFIG F	0.25	2.0	✗	✗	46.7	53.4	50.6
CONFIG F	0.25	2.0	✓	✗	46.8	53.6	50.8
CONFIG F	0.25	2.0	✓	✓	47.1	53.7	48.2
CONFIG F	0.50	2.0	✗	✗	48.1	53.9	50.4
CONFIG F	0.50	2.0	✓	✗	48.4	53.9	52.7
CONFIG F	0.50	2.0	✓	✓	48.6	54.1	51.7
CONFIG F	0.75	2.0	✗	✗	46.1	52.7	51.0
CONFIG F	0.75	2.0	✓	✗	46.9	52.8	51.6
CONFIG F	0.75	2.0	✓	✓	47.0	53.2	49.3

Table 5: **Ablation studies on ImageNet-1k with IPC 10.** Evaluate the effectiveness of several design choices, including soft category-aware matching (②), weak augmentation (④) and EMA-based evaluation (③).

impacts. It is important to note that we have omitted several traditional methods (Cazenavette et al., 2022; Liu et al., 2023a; Cui et al., 2023) from our analysis. This exclusion is due to their inadequate performance on the large-scale ImageNet-1k and their lesser effectiveness when applied to practical networks such as ResNet, MobileNet-V2, and Swin-Tiny (Liu et al., 2021). For instance, the MTT method (Cazenavette et al., 2022) encounters an out-of-memory issue on ImageNet-1k, and ResNet-18 achieves only a 46.4% accuracy on CIFAR-10 with IPC 10, which is significantly lower than the 79.1% accuracy reported for our EDC in Table 1.

4.1 Main Results

Experimental Comparison. Our integral EDC, represented as CONFIG G in Fig. 1, provides a versatile solution that outperforms other approaches across various dataset sizes. The results in Table 1 affirm its ability to consistently deliver substantial performance gains across different IPCs, datasets, and model architectures. Particularly notable is the performance leap in the highly compressed IPC 1 scenario using ResNet-18, where EDC markedly outperforms the latest state-of-the-art method, RDED. Performance rises from 22.9%, 11.0%, 7.0%, 24.9%, and 6.6% to 32.6%, 39.7%, 39.2%, 45.2%, and 12.8% for CIFAR-10, CIFAR-100, Tiny-ImageNet, ImageNet-10, and ImageNet-1k, respectively. These improvements clearly highlight EDC's superior information encapsulation and enhanced generalization capability, attributed to the efficiently synthesized condensed dataset.

Cross-Architecture Generalization. To verify the generalization ability of our condensed datasets, it is essential to assess their performance across various architectures such as ResNet-{18, 50, 101} (He et al., 2016a), MobileNet-V2 (Sandler et al., 2018), EfficientNet-B0 (Tan and Le, 2019), DeiT-Tiny (Touvron et al., 2021), Swin-Tiny (Liu et al., 2021), ConvNext-Tiny (Liu et al., 2022) and ShuffleNet-V2 (Zhang et al., 2018). The results of these evaluations are presented in Table 2. During cross-validation that includes all IPCs and the mentioned architectures, our EDC consistently achieves higher accuracy than RDED, demonstrating its strong generalization capabilities. Specifically, EDC surpasses RDED by significant margins of 8.2% and 14.42% on DeiT-Tiny and ShuffleNet-V2, respectively.

Application. Our condensed dataset not only serves as a versatile training resource but also enhances the adaptability of models across various downstream tasks. We demonstrate its effectiveness by employing it in scenarios such as data-free network slimming (Liu et al., 2017) (*w.r.t.,* parameter pruning (Srinivas and Babu, 2015)) and class-incremental continual learning (Prabhu et al., 2020) outlined in DM (Zhao and Bilen, 2023). Fig. 4 shows the wide applicability of our condensed dataset in both data-free network slimming and class-incremental continual learning. It substantially outperforms SRe^{2}L and G-VBSM, achieving significantly better results.

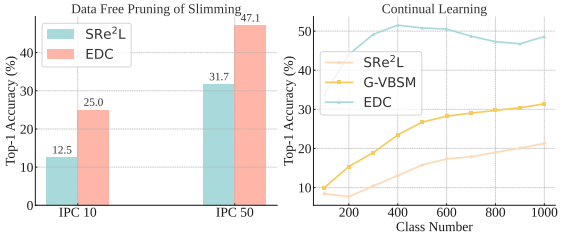


Figure 4: **Application on ImageNet-1k.** We evaluate the effectiveness of data-free network slimming and continual learning using VGG11-BN and ResNet-18, respectively.

https://doi.org/10.52202/079017-3146

4.2 Ablation Studies

Real Image Initialization (3), Smoothing LR Schedule (2) and Smaller Batch Size (1 1). As shown in Table 3 (left), these design choices, with zero additional computational cost, sufficiently enhance the performance of both G-VBSM and RDED. Furthermore, we investigate the influence of ζ within smoothing LR schedule in Table 3 (right), concluding that a smoothing learning rate decay is worthwhile for the condensed dataset's generalization ability and the optimal ζ is model-dependent.

Flatness Regularization (1). The results in Table 4 demonstrate the effectiveness of flatness regularization, while requiring a well-designed setup. Specifically, attempting to minimize sharpness across all statistics (*i.e.,* $\mathcal{L}_{\mathbf{FR}}$) proves ineffective, instead, it is more effective to apply this regularization exclusively to the logit (*i.e.,* $\mathcal{L}'_{\mathbf{FR}}$). Setting the loss weights β and τ at 0.25, 0.99, and 4, respectively, yields the best accuracy of 39.5%, 44.1%, and 45.9% for ResNet-18, ResNet-50, and DenseNet-121. Moreover, our design of $\mathcal{L}'_{\mathbf{FR}}$ surpasses the performance of the vanilla SAM, while requiring only half the computational resources.

Soft Category-Aware Matching (2), Weak Augmentation (4) and EMA-based Evaluation (3). Table 5 illustrates the effectiveness of weak augmentation and EMA-based evaluation, with EMA evaluation also playing a crucial role in minimizing performance fluctuations during assessment. The evaluation of soft category-aware matching primarily involves exploring the effect of parameter α across the range $[0, 1]$. The results in Table 5 suggest that setting α to 0.5 yields the best results based on our empirical analysis. This finding not only confirms the utility of soft category-aware matching but also emphasizes the importance of ensuring that the condensed dataset maintains a high level of information density and bears a distributional resemblance to the original dataset.

5 Conclusion

In this paper, we have conducted an extensive exploration and analysis of the design possibilities for scalable dataset condensation techniques. This comprehensive investigation helped us pinpoint a variety of effective and flexible design options, ultimately leading to the construction of a novel framework, which we call EDC. We have extensively examined EDC across five different datasets, which vary in size and number of classes, effectively proving EDC's robustness and scalability. Our results suggest that previous dataset distillation methods have not yet reached their full potential, largely due to suboptimal design decisions. We aim for our findings to motivate further research into developing algorithms capable of efficiently managing datasets of diverse sizes, thus advancing the field of dataset condensation task.

References

K. He, X. Zhang, and S. Ren, "Deep residual learning for image recognition," in *Computer Vision and Pattern Recognition*. Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. 1, 8, 9

K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*. Amsterdam, North Holland, The Netherlands: Springer, Oct. 2016, pp. 630–645. 1

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020. 1

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations*. Event Virtual: OpenReview.net, May 2020. 1

S. Shao, Z. Shen, L. Gong, H. Chen, and X. Dai, "Precise knowledge transfer via flow matching," *arXiv preprint arXiv:2402.02012*, 2024. 1

W. Masarczyk and I. Tautkute, "Reducing catastrophic forgetting with learning on synthetic data," in *Computer Vision and Pattern Recognition Workshops*. Virtual Event: IEEE, Jun. 2020, pp. 252–253. 1

M. Sangermano, A. Carta, A. Cossu, and D. Bacciu, "Sample condensation in online continual learning," in *International Joint Conference on Neural Networks*. Padua, Italy: IEEE, Jul. 2022, pp. 1–8. 1

B. Zhao and H. Bilen, "Dataset condensation with differentiable siamese augmentation," in *International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., vol. 139. Virtual Event: PMLR, 2021, pp. 12 674–12 685. 1

F. P. Such, A. Rawal, J. Lehman, K. O. Stanley, and J. Clune, "Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data," in *International Conference on Machine Learning*, vol. 119. Virtual Event: PMLR, Jul. 2020, pp. 9206–9216. 1

B. Zhao and H. Bilen, "Dataset condensation with distribution matching," in *Winter Conference on Applications of Computer Vision*. Waikoloa, Hawaii: IEEE, Jan. 2023, pp. 6514–6523. 1, 2, 9, 20

B. Zhao, K. R. Mopuri, and H. Bilen, "Dataset condensation with gradient matching," in *International Conference on Learning Representations*. Virtual Event: OpenReview.net, May 2021. 1, 2, 20

Z. Liu, J. Li, Z. Shen, G. Huang, S. Yan, and C. Zhang, "Learning efficient convolutional networks through network slimming," in *International Conference on Computer Vision*. IEEE, 2017, pp. 2736–2744. 1, 9

G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J. Zhu, "Dataset distillation by matching training trajectories," in *Computer Vision and Pattern Recognition*. New Orleans, LA, USA: IEEE, Jun. 2022. 1, 2, 9, 20, 26

A. Sajedi, S. Khaki, E. Amjadian, L. Z. Liu, Y. A. Lawryshyn, and K. N. Plataniotis, "Datadam: Efficient dataset distillation with attention matching," in *International Conference on Computer Vision*. Paris, France: IEEE, Oct. 2023, pp. 17 097–17 107. 1, 34

Y. Liu, J. Gu, K. Wang, Z. Zhu, W. Jiang, and Y. You, "DREAM: efficient dataset distillation by representative matching," *arXiv preprint arXiv:2302.14416*, 2023. 1, 9

O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015. 1, 8

Z. Yin, E. P. Xing, and Z. Shen, "Squeeze, recover and relabel: Dataset condensation at imagenet scale from A new perspective," in *Neural Information Processing Systems*. NeurIPS, 2023. 1, 2, 3, 4, 5, 8

Z. Yin and Z. Shen, "Dataset distillation in large data era," 2024. [Online]. Available: https://openreview.net/forum?id=kpEz4Bxs6e 1, 2, 3, 4, 5, 8, 23, 33

S. Shao, Z. Yin, X. Zhang, and Z. Shen, "Generalized large-scale data condensation via various backbone and statistical matching," *arXiv preprint arXiv:2311.17950*, 2023. 1, 2, 3, 4, 5, 8, 16, 20, 23

A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009. 2, 8

P. Sun, B. Shi, D. Yu, and T. Lin, "On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm," in *Computer Vision and Pattern Recognition*. IEEE, 2024. 2, 3, 4, 8, 16

T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, 2018. 2

J. Cui, R. Wang, S. Si, and C. Hsieh, "Scaling up dataset distillation to imagenet-1k with constant memory," in *International Conference on Machine Learning*, vol. 202. Honolulu, Hawaii, USA: PMLR, 2023, pp. 6565–6590. 2, 9

K. Wang, B. Zhao, X. Peng, Z. Zhu, S. Yang, S. Wang, G. Huang, H. Bilen, X. Wang, and Y. You, "Cafe: Learning to condense dataset by aligning features," in *Computer Vision and Pattern Recognition*. New Orleans, LA, USA: IEEE, Jun. 2022, pp. 12 196–12 205. 2

T. Nguyen, Z. Chen, and J. Lee, "Dataset meta-learning from kernel ridge-regression," *arXiv preprint arXiv:2011.00050*, 2020. 2, 26

J. Kim, J. Kim, S. J. Oh, S. Yun, H. Song, J. Jeong, J. Ha, and H. O. Song, "Dataset condensation via efficient synthetic-data parameterization," in *International Conference on Machine Learning*, vol. 162. Baltimore, Maryland, USA: PMLR, Jul. 2022, pp. 11 102–11 118. 2, 3, 8

D. Zhou, K. Wang, J. Gu, X. Peng, D. Lian, Y. Zhang, Y. You, and J. Feng, "Dataset quantization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 205–17 216. 3, 16

H. Zhang, S. Li, F. Lin, W. Wang, Z. Qian, and S. Ge, "Dance: Dual-view distribution alignment for dataset condensation," *arXiv preprint arXiv:2406.01063*, 2024. 3, 34

 https://doi.org/10.52202/079017-3146

H. Zhang, S. Li, P. Wang, D. Zeng, and S. Ge, "M3d: Dataset condensation by minimizing maximum mean discrepancy," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 8, 2024, pp. 9314–9322. 3, 34, 35

W. Deng, W. Li, T. Ding, L. Wang, H. Zhang, K. Huang, J. Huo, and Y. Gao, "Exploiting inter-sample and inter-feature relations in dataset distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 17 057–17 066. 3, 34, 35

G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015. [Online]. Available: https://arxiv.org/abs/1503.02531 3

J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, no. 6, pp. 1789–1819, 2021. 3

H. Chen, Y. Zhang, Y. Dong, and J. Zhu, "Rethinking model ensemble in transfer-based adversarial attacks," in *International Conference on Learning Representations*. Vienna, Austria: OpenReview.net, May 2024. 4, 20, 21

P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," in *International Conference on Learning Representations*, 2020. 4, 6, 21, 22

H. Chen, S. Shao, Z. Wang, Z. Shang, J. Chen, X. Ji, and X. Wu, "Bootstrap generalization ability from loss landscape perspective," in *European Conference on Computer Vision*. Springer, 2022, pp. 500–517. 4, 21, 25

H. Liu, T. Xing, L. Li, V. Dalal, J. He, and H. Wang, "Dataset distillation via the wasserstein metric," *arXiv preprint arXiv:2311.18531*, 2023. 4

J. Du, D. Zhou, J. Feng, V. Tan, and J. T. Zhou, "Sharpness-aware training for free," in *Advances in Neural Information Processing Systems*, vol. 35. New Orleans, Louisiana, USA: NeurIPS, Dec. 2022, pp. 23 439–23 451. 6, 21, 22, 23

D. Bahri, H. Mobahi, and Y. Tay, "Sharpness-aware minimization improves language model generalization," *arXiv preprint arXiv:2110.08529*, 2021. 6, 21

A. Tavanaei, "Embedded encoder-decoder in convolutional networks towards explainable AI," vol. abs/2007.06712, 2020. [Online]. Available: https://arxiv.org/abs/2007.06712 8

M. Sandler, A. G. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Computer Vision and Pattern Recognition*. Salt Lake City, UT, USA: IEEE, Jun. 2018, pp. 4510–4520. 8, 9

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., vol. 139. Virtual Event: PMLR, Jul. 2021, pp. 10 347–10 357. 8, 9

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *International Conference on Computer Vision*, 2021, pp. 10 012–10 022. 8, 9

M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114. 9

Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 976–11 986. 9

X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in *Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856. 9

S. Srinivas and R. V. Babu, "Data-free parameter pruning for deep neural networks," *arXiv preprint arXiv:1507.06149*, 2015. 9

A. Prabhu, P. H. S. Torr, and P. K. Dokania, "Gdumb: A simple approach that questions our progress in continual learning," in *European Conference on Computer Vision*. Springer, Jan 2020, p. 524–540. 9

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Neural Information Processing Systems*, Vancouver, BC, Canada, Dec. 2019. 15

B. Ostle *et al.*, "Statistics in research." *Statistics in research.*, no. 2nd Ed, 1963. 17

M. Zhou, Z. Yin, S. Shao, and Z. Shen, "Self-supervised dataset distillation: A good compression is all you need," *arXiv preprint arXiv:2404.07976*, 2024. 23

Y. Wu, J. Du, P. Liu, Y. Lin, W. Cheng, and W. Xu, "Dd-robustbench: An adversarial robustness benchmark for dataset distillation," *arXiv preprint arXiv:2403.13322*, 2024. 33

H. Kim, "Torchattacks: A pytorch repository for adversarial attacks," *arXiv preprint arXiv:2010.01950*, 2020. 33

 https://doi.org/10.52202/079017-3146

Appendix

A Implementation Details

Here, we complement both the hyperparameter settings and the backbone choices utilized for the comparison and ablation experiments in the main paper.

A.1 Hyperparameter Settings

(a) Data Synthesis

Config	Value	Explanation
Iteration	2000	NA
Optimizer	Adam	$\beta_1, \beta_2 = (0.5, 0.9)$
Learning Rate	0.05	NA
Batch Size	80	NA
Initialization	RDED	Initialized using images synthesized from RDED
α, β, τ	0.5, 0.99, 4	Control category-aware matching and flatness regularization

(b) Soft Label Generation and Post-Evaluation

Config	Value	Explanation
Epochs	300	NA
Optimizer	AdamW	NA
Learning Rate	0.001	Only use 1e-4 for Swin-Tiny
Batch Size	100	NA
EMA Rate	0.99	Control EMA-based Evaluation
Scheduler	Smoothing LR Schedule	$\zeta = 2$
Augmentation	RandomResizedCrop RandomHorizontalFlip	NA

Table 6: Hyperparameter setting on ImageNet-1k.

(a) Data Synthesis

Config	Value	Explanation
Iteration	2000	NA
Optimizer	Adam	$\beta_1, \beta_2 = (0.5, 0.9)$
Learning Rate	0.05	NA
Batch Size	100	NA
Initialization	RDED	Initialized using images synthesized from RDED
α, β, τ	0.5, 0.99, 4	Control category-aware matching and flatness regularization

(b) Soft Label Generation and Post-Evaluation

Config	Value	Explanation
Epochs	1000	NA
Optimizer	AdamW	NA
Learning Rate	0.001	NA
Batch Size	50	NA
EMA Rate	0.99	Control EMA-based Evaluation
Scheduler	Smoothing LR Schedule	$\zeta = 2$
Augmentation	RandAugment RandomResizedCrop RandomHorizontalFlip	NA

Table 7: Hyperparameter setting on ImageNet-10.

(a) Data Synthesis

Config	Value	Explanation
Iteration	2000	NA
Optimizer	Adam	$\beta_1, \beta_2 = (0.5, 0.9)$
Learning Rate	0.05	NA
Batch Size	100	NA
Initialization	Original Image	Initialized using images from training dataset
α, β, τ	0.5, 0.99, 4	Control category-aware matching and flatness regularization

(b) Soft Label Generation and Post-Evaluation

Config	Value	Explanation
Epochs	300	Only use 1000 for IPC 1
Optimizer	AdamW	NA
Learning Rate	0.001	NA
Batch Size	100	NA
EMA Rate	0.99	Control EMA-based Evaluation
Scheduler	Smoothing LR Schedule	$\zeta = 2$
Augmentation	RandAugment RandomResizedCrop RandomHorizontalFlip	NA

Table 8: Hyperparameter setting on Tiny-ImageNet.

We detail the hyperparameter settings of EDC for various datasets, including ImageNet-1k, ImageNet-10, Tiny-ImageNet, CIFAR-100, and CIFAR-10, in Tables 6, 7, 8, 9, and 10, respectively. For epochs, a critical factor affecting computational cost, we utilize strategies from SRe^{2}L, G-VBSM, and RDED for ImageNet-1k and follow RDED for the other datasets. In the data synthesis phase, we reduce the iteration count of hyperparameters by half compared to those used in SRe^{2}L and G-VBSM.

A.2 Network Architectures on Different Datasets

This section outlines the specific configurations of the backbones employed in the data synthesis and soft label generation phases, details of which are omitted from the main paper.

(a) Data Synthesis

Config	Value	Explanation
Iteration	2000	NA
Optimizer	Adam	$\beta_1, \beta_2 = (0.5, 0.9)$
Learning Rate	0.05	NA
Batch Size	100	NA
Initialization	Original Image	Initialized using images from training dataset
α, β, τ	0.5, 0.99, 4	Control category-aware matching and flatness regularization

(b) Soft Label Generation and Post-Evaluation

Config	Value	Explanation
Epochs	1000	NA
Optimizer	AdamW	NA
Learning Rate	0.001	NA
Batch Size	50	NA
EMA Rate	0.99	Control EMA-based Evaluation
Scheduler	Smoothing LR Schedule	$\zeta = 2$
Augmentation	RandAugment RandomResizedCrop RandomHorizontalFlip	NA

Table 9: Hyperparameter setting on CIFAR-100.

(a) Data Synthesis

Config	Value	Explanation
Iteration	75	NA
Optimizer	Adam	$\beta_1, \beta_2 = (0.5, 0.9)$
Learning Rate	0.05	NA
Batch Size	All	The number of synthesized images
Initialization	Original Image	Initialized using images from training dataset
α, β, τ	0.5, 0.99, 4	Control category-aware matching and flatness regularization

(b) Soft Label Generation and Post-Evaluation

Config	Value	Explanation
Epochs	1000	NA
Optimizer	AdamW	NA
Learning Rate	0.001	NA
Batch Size	25	NA
EMA Rate	0.99	Control EMA-based Evaluation
Scheduler	MultiStepLR	$\gamma = 0.5$ milestones=[800,900,950]
Augmentation	RandAugment RandomResizedCrop RandomHorizontalFlip	NA

Table 10: Hyperparameter setting on CIFAR-10.

ImageNet-1k. We utilize pre-trained models {ResNet-18, MobileNet-V2, ShuffleNet-V2, EfficientNet-V2, AlexNet} from torchvision (Paszke et al., 2019) as observer models in data synthesis. To reduce computational load, we exclude EfficientNet-V2 from the soft label generation process, a decision in line with our strategy of selecting more efficient backbones, a concept referred to as better backbone choice in the main paper. An extensive ablation analysis is available in Appendix G.

ImageNet-10. Prior to data synthesis, we train {ResNet-18, MobileNet-V2, ShuffleNet-V2, EfficientNet-V2} from scratch for 20 epochs and save their respective checkpoints. Subsequently, these pre-trained models are consistently employed for both data synthesis and soft label generation.

Tiny-ImageNet. We adopt the same backbone configurations as G-VBSM, specifically utilizing {ResNet-18, MobileNet-V2, ShuffleNet-V2, EfficientNet-V2} for both data synthesis and soft label generation. Each of these models has been trained on the original dataset with 50 epochs.

CIFAR-10&CIFAR-100. For small-scale datasets, we enhance the *baseline* G-VBSM model by incorporating three additional lightweight backbones. Consequently, the backbones utilized for data synthesis and soft label generation comprise {ResNet-18, ConvNet-W128, MobileNet-V2, WRN-16-2, ShuffleNet-V2, ConvNet-D1, ConvNet-D2, ConvNet-W32}. To demonstrate the effectiveness of our approach, we conduct comparative experiments and present results in Table 11, which illustrates that G-VBSM achieves improved performance with this enhanced backbone configuration.

	Verified Model	ResNet-18	AlexNet	VGG11-BN
CIFAR-10 (IPC 10)	100 backbones (MTT)	46.4	34.2	50.3
	5 backbones (original setting of G-VBSM)	53.5	31.7	55.2
	8 backbones (new setting of G-VBSM)	58.9	36.2	58.0

Table 11: **Ablation studies on CIFAR-10 with IPC 10.** With the remaining settings are the same as those of G-VBSM, our new backbone setting achieves better performance.

B Theoretical Derivations

Here, we give a detailed statement of the definitions, assumptions, theorems, and corollaries relevant to this paper.

https://doi.org/10.52202/079017-3146

B.1 Random Initialization vs. Real Image Initialization

In the data synthesis phase, random initialization involves using Gaussian noise, while real image initialization uses condensed images derived from training-free algorithms, such as RDED. Specifically, we denote the datasets initialized via random and real image methods as $\mathcal{X}^{\mathcal{S}}_{\textbf{random}}$ and $\mathcal{X}^{\mathcal{S}}_{\textbf{real}}$, respectively. For coupling $(\mathcal{X}^{\mathcal{S}}_{\textbf{random}}, \mathcal{X}^{\mathcal{S}}_{\textbf{real}})$, where $\mathcal{X}^{\mathcal{S}}_{\textbf{random}} \sim \pi_{\textbf{random}}$, $\mathcal{X}^{\mathcal{S}}_{\textbf{real}} \sim \pi_{\textbf{real}}$ and satisfies $p(\pi_{\textbf{random}}, \pi_{\textbf{real}}) = p(\pi_{\textbf{random}})p(\pi_{\textbf{real}})$, we have the mutual information (MI) between $\pi_{\textbf{random}}$ and $\pi_{\textbf{real}}$ is 0, *a.k.a.*, $I(\pi_{\textbf{random}}, \pi_{\textbf{real}}) = 0$. By contrast, training-free algorithms (Sun et al., 2024; Zhou et al., 2023) synthesize the compressed data $\mathcal{X}^{\mathcal{S}}_{\textbf{free}} := \phi(\mathcal{X}^{\mathcal{S}}_{\textbf{real}})$ via $\mathcal{X}^{\mathcal{S}}_{\textbf{real}}$, satisfying $p(\mathcal{X}^{\mathcal{S}}_{\textbf{free}}|\mathcal{X}^{\mathcal{S}}_{\textbf{real}}) > 0$. When the cost function $\mathbb{E}[c(a-b)] \propto 1/I(\text{Law}(a), \text{Law}(b))$, we have $\mathbb{E}[c(\mathcal{X}^{\mathcal{S}}_{\textbf{real}} - \mathcal{X}^{\mathcal{S}}_{\textbf{free}})] \leq \mathbb{E}[c(\mathcal{X}^{\mathcal{S}}_{\textbf{real}} - \mathcal{X}^{\mathcal{S}}_{\textbf{random}})]$.

Proof.

$$
\begin{aligned}
\mathbb{E}[c(\mathcal{X}^{\mathcal{S}}_{\textbf{real}} - \mathcal{X}^{\mathcal{S}}_{\textbf{free}})] &= k/I(\text{Law}(\mathcal{X}^{\mathcal{S}}_{\textbf{real}}), \text{Law}(\mathcal{X}^{\mathcal{S}}_{\textbf{free}})) \\
&= k/\text{D}_{\text{KL}}(p(\pi_{\textbf{real}}, \pi_{\textbf{free}})||p(\pi_{\textbf{real}})p(\pi_{\textbf{free}})) \\
&= k/[H(\pi_{\textbf{real}}) - H(\pi_{\textbf{real}}|\pi_{\textbf{free}})] \\
&\leq k/[H(\pi_{\textbf{real}})] \\
&= k/[H(\pi_{\textbf{real}}) - H(\pi_{\textbf{real}}|\pi_{\textbf{random}})] \\
&= k/I(\text{Law}(\mathcal{X}^{\mathcal{S}}_{\textbf{real}}), \text{Law}(\mathcal{X}^{\mathcal{S}}_{\textbf{random}})) \\
&= \mathbb{E}[c(\mathcal{X}^{\mathcal{S}}_{\textbf{real}} - \mathcal{X}^{\mathcal{S}}_{\textbf{random}})],
\end{aligned}
\tag{8}
$$

where $k \in \mathbb{R}^+$ denotes a constant. And $D_{\text{KL}}(\cdot||\cdot)$ and $H(\cdot)$ stand for Kullback-Leibler divergence and entropy, respectively. □

From the theoretical perspective described, it becomes evident that initializing with real images enhances MI more significantly than random initialization between the distilled and the original datasets at the start of the data synthesis phase. This improvement substantially alleviates the challenges inherent in data synthesis. Furthermore, our exploratory experiments demonstrate that the generalized matching loss (Shao et al., 2023) for real image initialization remains consistently lower compared to that of random initialization throughout the data synthesis phase.

B.2 Theoretical Derivations of Soft Category-Aware Matching

Definition B.1. *(Statistical Matching) Assume that we have N D-dimensional random samples $\{x_i \in \mathcal{R}^D\}_{i=1}^N$ with an unknown distribution $p_{mix}(x)$, we define two forms of statistical matching for dataset distillation:*

Form (1): *Estimate the mean $\mathbb{E}[x]$ and variance $\mathbb{D}[x]$ of samples $\{x_i \in \mathcal{R}^D\}_{i=1}^N$. Then, synthesize M $(M \ll N)$ distilled samples $\{y_i \in \mathcal{R}^D\}_{i=1}^M$ such that the absolute differences between the variances $(|\mathbb{D}[x] - \mathbb{D}[y]|)$ and means $(|\mathbb{E}[x] - \mathbb{E}[y]|)$ of the original and distilled samples are $\leq \epsilon$.*

Form (2): *Consider $p_{mix}(x)$ to be a linear combination of multiple subdistributions, expressed as $p_{mix}(x) = \int_{\mathbf{C}} p(x|c_i)p(c_i)dc_i$, where c_i denotes a component of the original distribution. Given Assumption B.4, we can treat $p_{mix}(x)$ as a GMM, with each component $p(x|c_i)$ following a Gaussian distribution. For each component, estimate the mean $\mathbb{E}[x^j]$ and variance $\mathbb{D}[x^j]$ using N_j samples $\{x_i^j\}_{i=1}^{N_j}$, ensuring that $\sum_{j=1}^{\mathbf{C}} N_j = N$. Subsequently, synthesize M $(M \ll N)$ distilled samples across all components $\bigcup_{j=1}^{\mathbf{C}} \{y_i^j\}_{i=1}^{M_j}$, where $\sum_{j=1}^{\mathbf{C}} M_j = M$. This process aims to ensure that for each component, the absolute differences between the variances $(|\mathbb{D}[x^j] - \mathbb{D}[y^j]|)$ and means $(|\mathbb{E}[x^j] - \mathbb{E}[y^j]|)$ of the original and distilled samples $\leq \epsilon$.*

Based on Definition B.1, here we provide several relevant theoretical conclusion.

Lemma B.2. *Consider a sample set $\mathbb{S}$, where each sample $\mathcal{X}$ within $\mathbb{S}$ belongs to $\mathcal{R}^D$. Assume any two variables x_i and x_j in $\mathbb{S}$ satisfies $p(x_i, x_j) = p(x_i)p(x_j)$. This set $\mathbb{S}$ comprises $\mathbf{C}$ disjoint subsets $\{\mathbb{S}_1, \mathbb{S}_2, \ldots, \mathbb{S}_C\}$, ensuring that for any $1 \leq i < j \leq C$, the intersection $\mathbb{S}_i \cap \mathbb{S}_j = \emptyset$ and the union $\bigcup_{k=1}^{C} \mathbb{S}_k = \mathbb{S}$. Consequently, the expected value over the variance within the subsets, denoted as $\mathbb{E}_{\mathbb{S}_{sub} \sim \{\mathbb{S}_1, \ldots, \mathbb{S}_C\}} \mathbb{D}_{\mathcal{X} \sim \mathbb{S}_{sub}}[\mathcal{X}]$, is smaller than or equal to the variance within the entire set, $\mathbb{D}_{\mathcal{X} \sim \mathbb{S}}[\mathcal{X}]$.*

Proof.

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}\mathbb{D}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}] \\
&= \mathbb{E}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}(\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}\circ\mathcal{X}] - \mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}]\circ\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}]) \\
&= \mathbb{E}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}\circ\mathcal{X}] - \mathbb{E}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}]\circ\mathbb{E}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}] + \mathbb{E}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}]\circ\mathbb{E}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}] \\
&\quad - \mathbb{E}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}]\circ\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}] \\
&= \mathbb{D}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}] - \mathbb{E}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}]\circ\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}] \\
&\quad + \mathbb{E}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}]\circ\mathbb{E}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}] \\
&= \mathbb{D}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}] - \mathbb{E}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}]\circ\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}] \\
&\quad + \mathbb{E}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}]\circ\mathbb{E}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}] \\
&= \mathbb{D}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}] - \mathbb{D}_{\mathbb{S}_{\text{sub}}\sim\{\mathbb{S}_1,\ldots,\mathbb{S}_C\}}\mathbb{E}_{\mathcal{X}\sim\mathbb{S}_{\text{sub}}}[\mathcal{X}] \\
&\leq \mathbb{D}_{\mathcal{X}\sim\mathbb{S}}[\mathcal{X}].
\end{aligned}
\tag{9}
$$

□

Lemma B.3. *Consider a Gaussian Mixture Model (GMM) $p_{mix}(x)$ comprising $\mathbf{C}$ components (i.e., sub-Gaussian distributions). These components are characterized by their means, variances, and weights, denoted as $\{\mu_i\}_{i=1}^{\mathbf{C}}$, $\{\sigma_i^2\}_{i=1}^{\mathbf{C}}$, and $\{\omega_i\}_{i=1}^{\mathbf{C}}$, respectively. The mean $\mathbb{E}[x]$ and variance $\mathbb{D}[x]$ of the distribution are given by $\sum_{i=1}^{\mathbf{C}}\omega_i\mu_i$ and $\sum_{i=1}^{\mathbf{C}}\omega_i(\mu_i^2+\sigma_i^2)-(\sum_{i=1}^{\mathbf{C}}\omega_i\mu_i)^2$, respectively (Ostle et al., 1963).*

Proof.

$$
\begin{aligned}
\mathbb{E}[x] &= \int_\Theta x\sum_{i=1}^{\mathbf{C}}\omega_i\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} \\
&= \sum_{i=1}^{\mathbf{C}}\omega_i\left[\int_\Theta x\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}\right] \\
&= \sum_{i=1}^{\mathbf{C}}\omega_i\mu_i, \\
\mathbb{D}[x] &= \mathbb{E}[x^2] - \mathbb{E}[x]^2 \\
&= \int_\Theta x^2\sum_{i=1}^{\mathbf{C}}\omega_i\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}} - \mathbb{E}[x]^2 \\
&= \sum_{i=1}^{\mathbf{C}}\omega_i\left[\int_\Theta x^2\frac{1}{\sqrt{2\pi}\sigma_i}e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}\right] - \mathbb{E}[x]^2 \\
&= \sum_{i=1}^{\mathbf{C}}\omega_i[\mu_i^2+\sigma_i^2] - (\sum_{i=1}^{\mathbf{C}}\omega_i\mu_i)^2.
\end{aligned}
\tag{10}
$$

□

Assumption B.4. *For any distribution Q, there exists a constant $\mathbf{C}$ enabling the approximation of Q by a Gaussian Mixture Model P with $\mathbf{C}$ components. More generally, this is expressed as the existence of a $\mathbf{C}$ such that the distance between P and Q, denoted by the distance metric function $\ell(P,Q)$, is bounded above by an infinitesimal ϵ.*

Sketch Proof. The Fourier transform of a Gaussian function does not possess true zeros, indicating that such a function, $f(x)$, along with its shifted variant, $f(x+a)$, densely populates the function space through the Tauberian Theorem. In the context of L^2, the space of all square-integrable functions, where Gaussian functions form a subspace denoted as G, any linear functional defined on G—such as convolution operators—can be extended to all of L^2 through the application of the Hahn-Banach Theorem. This extension underscores the completeness of Gaussian Mixture Models (GMM) within L^2 spaces.

Remarks. The proof presents two primary limitations: firstly, it relies solely on shift, which allows the argument to remain valid even when the variances of all components within GMM are identical (a relatively loose condition). Secondly, it imposes an additional constraint by requiring that the coefficients $\omega_i > 0$ and $\sum_i \omega_i = 1$ in GMM. Accordingly, this study proposes, rather than empirically demonstrates, that GMM can approximate any specified distribution.

https://doi.org/10.52202/079017-3146

Theorem B.5. *Given Assumption B.4 and Definition B.1, the variances and means of x and y, estimated through maximum likelihood, remain consistent across scenarios **Form (1)** and **Form (2)**.*

Proof. The maximum likelihood estimation mean $\mathbb{E}[x]$ and variance $\mathbb{D}[x]$ of samples $\{x_i\}_{i=1}^N$ within a Gaussian distribution are calculated as $\frac{\sum_{i=1}^N x_i}{N}$ and $\frac{\sum_{i=1}^N (x_i - \mathbb{E}[x])^2}{N}$, respectively. These estimations enable us to characterize the distribution's behavior across different scenarios as follows:

Form (1): $P(x) \sim \mathcal{N}\left(\frac{\sum_{i=1}^N x_i}{N}, \frac{\sum_{i=1}^N \left(x_i - \frac{\sum_{i=1}^N x_i}{N}\right)^2}{N}\right)$.

Form (2): $Q(y) \sim \sum_i \frac{N_i}{\sum_{j=1}^{\mathbf{C}} N_j} \mathcal{N}\left(\frac{\sum_{k=1}^{N_i} x_k^i}{N_i}, \frac{\sum_{k=1}^{N_i} \left(x_k^i - \frac{\sum_{k=1}^{N_i} x_k^i}{N_i}\right)^2}{N_i}\right)$.

Intuitively, the distilled samples $\{y_i\}_{i=1}^M$ will obey distributions $P(x)$ and $Q(y)$ in scenarios ***Form (1)*** and ***Form (2)***, respectively. Then, the difference of the means between ***Form (1)*** and ***Form (2)*** can be derived as

$$\begin{aligned}\int_\Theta [xP(x)dx - xQ(x)dx] &= \frac{\sum_{i=1}^N x_i}{N} - \sum_i \frac{N_i}{\sum_{j=1}^{\mathbf{C}} N_j} \frac{\sum_{k=1}^{N_i} x_k^i}{N_i} \\ &= 0.\end{aligned} \tag{11}$$

To further enhance the explanation on proving the consistency of the variance, the setup introduces two sample sets, $\{x_i\}_{i=1}^N$ and $\bigcup_{j=1}^{\mathbf{C}} \{y_i^j\}_{i=1}^{N_j}$, each drawn from their respective distributions, $P(x)$ and $Q(y)$. After that, we can acquire:

$$\begin{aligned}\mathbb{D}[x] - \mathbb{D}[y] &= \mathbb{D}[x] - \sum_{i=1}^{\mathbf{C}} \frac{N_i}{\sum_j N_j} (\mathbb{E}[y^j]^2 + \mathbb{D}[y^j]) + \left(\sum_{i=1}^{\mathbf{C}} \frac{N_i}{\sum_j N_j} \mathbb{E}[y^j]\right)^2 \qquad \#\ Lemma\ B.3 \\ &= \mathbb{D}[x] - \mathbb{E}[\mathbb{E}[y^j]^2] - \mathbb{E}[\mathbb{D}[y^j]] + \mathbb{E}[\mathbb{E}[y^j]]^2 \\ &= (\mathbb{D}[x] - \mathbb{E}[\mathbb{D}[y^j]]) - \mathbb{E}[\mathbb{E}[y^j]^2] + \mathbb{E}[\mathbb{E}[y^j]]^2 \\ &= \mathbb{D}[\mathbb{E}[y^j]] - \mathbb{E}[\mathbb{E}[y^j]^2] + \mathbb{E}[\mathbb{E}[y^j]]^2 \qquad \#\ Lemma\ B.2 \\ &= 0.\end{aligned} \tag{12}$$

□

Corollary B.6. *The mean and variance obtained from maximum likelihood for any split form $\{c_1, c_2, \ldots, c_{\mathbf{C}}\}$ in **Form (2)** remain consistent.*

Sketch Proof. According to Theorem B.5 the mean and variance obtained from maximum likelihood for each split form in ***Form (2)*** remain consistent within ***Form (1)***, so that any split form $\{c_1, c_2, \ldots, c_{\mathbf{C}}\}$ in ***Form (2)*** remain consistent.

Theorem B.7. *Based on Definition B.1, the entropy—pertaining to diversity—of the distributions characterized as $\mathcal{H}(P)$ from **Form (1)** and $\mathcal{H}(Q)$ from **Form (2)**, which are estimated through maximum likelihood, exhibits the subsequent relationship: $\mathcal{H}(P) - \frac{1}{2}\left[\log(\mathbb{E}[\mathbb{D}[y^j]] + \mathbb{D}[\mathbb{E}[y^j]]) - \mathbb{E}[\log(\mathbb{D}[y^j])]\right] \leq \mathcal{H}(Q) \leq \mathcal{H}(P) + \frac{1}{4}\mathbb{E}_{(i,j)\sim\prod[\mathbf{C},\mathbf{C}]}\left[\frac{(\mathbb{E}[y^i]-\mathbb{E}[y^j])^2(\mathbb{D}[y^i]+\mathbb{D}[y^j])}{\mathbb{D}[y^i]\mathbb{D}[y^j]}\right]$. The two-sided equality (i.e., $\mathcal{H}(P) \equiv \mathcal{H}(Q)$) holds if and only if both the variance and the mean of each component are consistent.*

Proof.

#Lower bound:

$$
\begin{aligned}
&\mathbb{E}[-\log(P(x))] - \mathbb{E}[-\log(Q(y))] \\
&= \int_{\Theta} -\log(P(x))P(x)dx + \int_{\Theta} \log(P(y))P(y)dy \\
&= \frac{1}{2}\log(2\pi\mathbb{D}[x]) + \frac{1}{2} + \int_{\Theta} \log(\int_j p(y^j)\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}dj)(\int_j p(y^j)\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}dj)dy \\
&= \frac{1}{2}\log(2\pi\mathbb{D}[x]) + \frac{1}{2} + \int_{\Theta} \log(\mathbb{E}[\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}])\mathbb{E}[\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}]dy \\
&\geq \frac{1}{2}\log(2\pi\mathbb{D}[x]) + \frac{1}{2} + \int_{\Theta} \mathbb{E}[\log(\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}})]\mathbb{E}[\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}]dy \\
&= \frac{1}{2}\log(2\pi\mathbb{D}[x]) + \frac{1}{2} + \mathbb{E}_{(i,j)\sim\prod[\mathbf{C},\mathbf{C}]}\left[\int_{\Theta} \log(\frac{1}{\sqrt{2\pi\mathbb{D}[y^i]}}e^{\frac{(y-\mathbb{E}[y^i])^2}{-2\mathbb{D}[y^i]}})(\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}})dy\right] \\
&= \frac{1}{2}\log(2\pi\mathbb{D}[x]) + \frac{1}{2} - \mathbb{E}_{(i,j)\sim\prod[\mathbf{C},\mathbf{C}]}\left[\frac{1}{2}\log(2\pi\mathbb{D}[y^j]) + \frac{\mathbb{D}[y^i] + (\mathbb{E}[y^i] - \mathbb{E}[y^j])^2}{2\mathbb{D}[y^j]}\right] \\
&\geq \frac{1}{2}\log(2\pi\mathbb{D}[x]) - \frac{1}{2}\log(\mathbb{E}[2\pi\mathbb{D}[y^j]]) + \frac{1}{2} - \mathbb{E}_{(i,j)\sim\prod[\mathbf{C},\mathbf{C}]}\left[\frac{\mathbb{D}[y^i] + (\mathbb{E}[y^i] - \mathbb{E}[y^j])^2}{2\mathbb{D}[y^j]}\right] \\
&\geq -\frac{1}{4}\mathbb{E}_{(i,j)\sim\prod[\mathbf{C},\mathbf{C}]}\left[\frac{(\mathbb{E}[y^i] - \mathbb{E}[y^j])^2(\mathbb{D}[y^i] + \mathbb{D}[y^j])}{\mathbb{D}[y^i]\mathbb{D}[y^j]}\right]
\end{aligned}
$$

#Upper bound:

$$
\begin{aligned}
&\mathbb{E}[-\log(P(x))] - \mathbb{E}[-\log(Q(y))] \\
&= \int_{\Theta} -\log(P(x))P(x)dx + \int_{\Theta} \log(P(y))P(y)dy \\
&= \int_{\Theta} -\log(P(x))P(x)dx + \int_{\Theta} \log(\mathbb{E}[\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}])\mathbb{E}[\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}]dy \\
&\leq \int_{\Theta} -\log(P(x))P(x)dx + \mathbb{E}[\int_{\Theta} \log(\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}})\frac{1}{\sqrt{2\pi\mathbb{D}[y^j]}}e^{\frac{(y-\mathbb{E}[y^j])^2}{-2\mathbb{D}[y^j]}}dy] \\
&= \frac{1}{2}\log(2\pi\mathbb{D}[x]) - \mathbb{E}[\frac{1}{2}\log(2\pi\mathbb{D}[y^j])] \\
&= \frac{1}{2}\left[\log(\mathbb{E}[\mathbb{D}[y^j]] + \mathbb{D}[\mathbb{E}[y^j]]) - \mathbb{E}[\log(\mathbb{D}[y^j])]\right]
\end{aligned}
\tag{13}
$$

□

Theorem B.8. *Based on Definition B.1, if the original distribution is p_{mix}, the Kullback-Leibler divergence $D_{KL}[p_{mix}||Q]$ has a upper bound $\mathbb{E}_{i\sim\mathcal{U}[1,\ldots,\mathbf{C}]}\mathbb{E}_{j\sim\mathcal{U}[1,\ldots,\mathbf{C}]}\frac{\mathbb{E}[y^j]^2}{\mathbb{D}[y^i]}$ and $D_{KL}[p_{mix}||P] = 0$.*

Proof.

$$
\begin{aligned}
&D_{\mathrm{KL}}[Q||P] \\
&= D_{\mathrm{KL}}\left[\sum_i \frac{N_i}{\sum_{j=1}^{\mathbf{C}} N_j}\mathcal{N}\left(\frac{\sum_{k=1}^{N_i} x_k^i}{N_i}, \frac{\sum_{k=1}^{N_i}\left(x_k^i - \frac{\sum_{k=1}^{N_i} x_k^i}{N_i}\right)^2}{N_i}\right)\middle\|\mathcal{N}\left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N}\left(x_i - \frac{\sum_{i=1}^{N} x_i}{N}\right)^2}{N}\right)\right] \\
&\leq \sum_i \frac{N_i}{\sum_{j=1}^{\mathbf{C}} N_j} D_{\mathrm{KL}}\left[\mathcal{N}\left(\frac{\sum_{k=1}^{N_i} x_k^i}{N_i}, \frac{\sum_{k=1}^{N_i}\left(x_k^i - \frac{\sum_{k=1}^{N_i} x_k^i}{N_i}\right)^2}{N_i}\right)\middle\|\mathcal{N}\left(\frac{\sum_{i=1}^{N} x_i}{N}, \frac{\sum_{i=1}^{N}\left(x_i - \frac{\sum_{i=1}^{N} x_i}{N}\right)^2}{N}\right)\right].
\end{aligned}
\tag{14}
$$

https://doi.org/10.52202/079017-3146

By applying the notations from Lemma B.3 for convenience, we obtain:

$$
\begin{aligned}
&D_{\mathrm{KL}}[Q||P] \\
&\leq \sum_i \omega_i \left[\frac{1}{2} \log \left(\frac{\sum_{j=1}^{\mathbf{C}} \omega_j[\mu_j^2 + \sigma_j^2] - (\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j)^2}{\sigma_i^2} \right) + \frac{\sum_{j=1}^{\mathbf{C}} \omega_j[\mu_j^2 + \sigma_j^2] - (\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j)^2}{2\sigma_i^2} \right] - \frac{1}{2} \\
&\leq \frac{1}{2} \log \left(\sum_i \omega_i \frac{\sum_{j=1}^{\mathbf{C}} \omega_j[\mu_j^2 + \sigma_j^2] - (\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j)^2}{\sigma_i^2} \right) + \frac{1}{2} \sum_i \omega_i \frac{\sum_{j=1}^{\mathbf{C}} \omega_j[\mu_j^2 + \sigma_j^2] - (\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j)^2}{\sigma_i^2} - \frac{1}{2} \\
&\leq \frac{1}{2} \log \left(1 + \sum_i \omega_i \frac{\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j^2 - (\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j)^2}{\sigma_i^2} \right) + \frac{1}{2} \sum_i \omega_i \frac{\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j^2 - (\sum_{j=1}^{\mathbf{C}} \omega_j \mu_j)^2}{\sigma_i^2} \\
&\leq \frac{1}{2} \log \left(1 + \sum_i \sum_j \omega_i \omega_j \frac{\mu_j^2}{\sigma_i^2} \right) + \frac{1}{2} \sum_i \sum_j \omega_i \omega_j \frac{\mu_j^2}{\sigma_i^2} \\
&\leq \mathbb{E}_{i \sim \mathcal{U}[1,\ldots,\mathbf{C}]} \mathbb{E}_{j \sim \mathcal{U}[1,\ldots,\mathbf{C}]} \frac{\mathbb{E}[y^j]^2}{\mathbb{D}[y^i]}.
\end{aligned}
\tag{15}
$$

□

When the sample size is sufficiently large, the original distribution aligns with Q. Consequently, we obtain $D_{\mathrm{KL}}[p_{\mathrm{mix}}||P] \leq \mathbb{E}_{i \sim \mathcal{U}[1,\ldots,\mathbf{C}]} \mathbb{E}_{j \sim \mathcal{U}[1,\ldots,\mathbf{C}]} \frac{\mathbb{E}[y^j]^2}{\mathbb{D}[y^i]}$ and establish that $D_{\mathrm{KL}}[p_{\mathrm{mix}}||Q] = 0$.

C Decoupled Optimization Objective of Dataset Condensation

In this section, we demonstrate that the training objective, as defined in Eq. 2, can be decoupled into two components—flatness and closeness—using a second-order Taylor expansion, under the assumption that $\mathcal{L}_{\mathbf{syn}} \in \mathbf{C}^2(\mathbf{I}, \mathbb{R})$. We define the closest optimization point $\mathbf{o}_i$ for $\mathcal{X}^{\mathcal{S}}$ in relation to the i-th matching operator $\mathcal{L}^i_{\mathbf{syn}}(\cdot,\cdot)$. This framework can accommodate all matchings related to $f^i(\cdot)$, including gradient matching(Zhao et al., 2021), trajectory matching (Cazenavette et al., 2022), distribution matching (Zhao and Bilen, 2023), and statistical matching (Shao et al., 2023). Consequently, we derive the dual decoupling of flatness and closeness as follows:

$$
\begin{aligned}
\mathcal{L}_{\mathbf{DD}} &= \mathbb{E}_{\mathcal{L}_{\mathbf{syn}}(\cdot,\cdot) \sim \mathbb{S}_{\mathrm{match}}}[\mathcal{L}_{\mathbf{syn}}(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})] = \frac{1}{|\mathbb{S}_{\mathrm{match}}|} \sum_{i=1}^{|\mathbb{S}_{\mathrm{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})] \\
&= \frac{1}{|\mathbb{S}_{\mathrm{match}}|} \sum_{i=1}^{|\mathbb{S}_{\mathrm{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\mathbf{o}_i, \mathcal{X}^{\mathcal{T}}) + (\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i) \nabla_{\mathcal{X}^{\mathcal{S}}} \mathcal{L}^i_{\mathbf{syn}}(\mathbf{o}_i, \mathcal{X}^{\mathcal{T}}) + (\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i)^T \mathrm{H}^i (\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i)] + \mathcal{O}((\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i)^3) \\
&= \frac{1}{|\mathbb{S}_{\mathrm{match}}|} \sum_{i=1}^{|\mathbb{S}_{\mathrm{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\mathbf{o}_i, \mathcal{X}^{\mathcal{T}}) + (\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i)^T \mathrm{H}^i (\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i)],
\end{aligned}
\tag{16}
$$

where H^i refers to the Hessian matrix of $\mathcal{L}^i_{\mathbf{syn}}(\cdot, \mathcal{X}^{\mathcal{T}})$ at the closest optimization point $\mathbf{o}_i$. Note that as the optimization method for deep learning typically involves gradient descent-like approaches (*e.g.,* SGD and AdamW), the first-order derivative $\nabla_{\mathcal{X}^{\mathcal{S}}} \mathcal{L}^i_{\mathbf{syn}}(\mathbf{o}_i, \mathcal{X}^{\mathcal{T}})$ can be directly discarded. After that, scanning the two terms in Eq. 16, the first one necessarily reaches an optimal solution, while the second one allows us to obtain an upper definitive bound on the Hessian matrix and Jacobi matrix through Theorem 3.1 outlined in Chen et al. (2024). Here, we give a special case under the ℓ_2-norm to discard the assumption that H^i and $(\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i)$ are independent:

Theorem C.1. *(improved from Theorem 3.1 in (Chen et al., 2024))* $\frac{1}{|\mathbb{S}_{match}|} \sum_{i=1}^{|\mathbb{S}_{match}|} (\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i)^T \mathrm{H}^i (\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i) \leq |\mathbb{S}_{match}| \cdot \mathbb{E}[||\mathrm{H}^i||_{\mathrm{F}}] \mathbb{E}[||\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i||_2^2]$*, where* $\mathbb{E}[||\mathrm{H}^i||_{\mathrm{F}}]$ *and* $\mathbb{E}[||\mathcal{X}^{\mathcal{S}} - \mathbf{o}_i||_2^2]$ *denote flatness and closeness, respectively.*

Proof.

$$\begin{aligned}
&\frac{1}{|\mathbb{S}_{\text{match}}|}\sum_{i=1}^{|\mathbb{S}_{\text{match}}|}(\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i)^T\mathrm{H}^i(\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i)\leq\frac{1}{|\mathbb{S}_{\text{match}}|}\sum_{i=1}^{|\mathbb{S}_{\text{match}}|}[||(\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i)||_2||\mathrm{H}^i(\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i)||_2] \quad \text{\# Hölder's inequality}\\
&=\frac{1}{|\mathbb{S}_{\text{match}}|}\sum_{i=1}^{|\mathbb{S}_{\text{match}}|}[||(\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i)||_2||\mathrm{H}^i||_{2,2}||(\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i)||_2] \quad \text{\# Definition of matrix norm}\\
&\leq|\mathbb{S}_{\text{match}}|\cdot\mathbb{E}[||\mathrm{H}^i||_{2,2}]\mathbb{E}[||\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i||_2^2]\leq|\mathbb{S}_{\text{match}}|\cdot\mathbb{E}[||\mathrm{H}^i||_{\mathrm{F}}]\mathbb{E}[||\mathcal{X}^{\mathcal{S}}-\mathbf{o}_i||_2^2]
\end{aligned} \tag{17}$$

□

Actually, flatness can be ensured by convergence in a flat region through sharpness-aware minimization (SAM) theory (Foret et al., 2020; Bahri et al., 2021; Du et al., 2022; Chen et al., 2022). Specifically, a body of work on SAM has established a connection between the Hessian matrix and the flatness of the loss landscape (*i.e.,* the curvature of the loss trajectory), with a series of empirical studies demonstrating the theory's reliability. Meanwhile, the specific implementation of flatness is elaborated upon in Sec. E. By contrast, the concept of closeness was first introduced in Chen et al. (2024), where it is observed that utilizing more backbones for ensemble can result in a smaller generalization error during the evaluation phase. In fact, closeness has been implicitly implemented since our *baseline* G-VBSM uses a sequence optimization mechanism akin to the official implementation in Chen et al. (2024). Therefore, this paper will not elucidate on closeness and its specific implementation.

D Traditional Sharpness-Aware Minimization Optimization Approach

For the comprehensive of our paper, let us give a brief yet formal description of sharpness-aware minimization (SAM). The applicable SAM algorithm was first proposed in Foret et al. (2020), which aims to solve the following maximum minimization problem:

$$\min_{\theta}\max_{\epsilon:||\epsilon||\leq\rho}L_{\mathbb{S}}(f_{\theta+\epsilon}), \tag{18}$$

where $L_{\mathbb{S}}(f_\theta)$, ϵ, ρ, and θ refer to the loss $\frac{1}{|\mathbb{S}|}\sum_{x_i,y_i\sim\mathbb{S}}\ell(f_\theta(x_i),y_i)$, the perturbation, the pre-defined flattened region, and the model parameter, respectively. Let us define the final optimized model parameters as θ^*, then the optimization objective can be rewritten as

$$\theta^*=\arg\min_{\theta}R_{\mathbb{S}}(f_\theta)+L_{\mathbb{S}}(f_\theta),\ \text{where}\ R_{\mathbb{S}}(f_\theta)=\max_{\epsilon:||\epsilon||\leq\rho}L_{\mathbb{S}}(f_{\theta+\epsilon})-L_{\mathbb{S}}(f_\theta). \tag{19}$$

By expanding $L_{\mathbb{S}}(f_{\theta+\epsilon})$ at θ and by solving the classical *dual norm* problem, the first maximization objective can be solved as (In the special case of the ℓ_2-norm)

$$\epsilon^*=\arg\max_{\epsilon:||\epsilon||\leq\rho}L_{\mathbb{S}}(f_{\theta+\epsilon})\approx\rho\frac{\nabla_\theta L_{\mathbb{S}}(f_\theta)}{||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_2}. \tag{20}$$

The specific derivation is as follows:

Proof. Subjecting $L_{\mathbb{S}}(f_{\theta+\epsilon})$ to a Taylor expansion and retaining only the first-order derivatives:

$$R_{\mathbb{S}}(f_\theta)=L_{\mathbb{S}}(f_{\theta+\epsilon})-L_{\mathbb{S}}(f_\theta)\approx L_{\mathbb{S}}(f_\theta)+\epsilon^T\nabla_\theta L_{\mathbb{S}}(f_\theta)-L_{\mathbb{S}}(f_\theta)=\epsilon^T\nabla_\theta L_{\mathbb{S}}(f_\theta). \tag{21}$$

Then, we can get

$$\epsilon^*=\arg\max_{\epsilon:||\epsilon||\leq\rho}L_{\mathbb{S}}(f_{\theta+\epsilon})-L_{\mathbb{S}}(f_\theta)=\arg\max_{\epsilon:||\epsilon||\leq\rho}\left[\epsilon^T\nabla_\theta L_{\mathbb{S}}(f_\theta)\right]. \tag{22}$$

Next, we base our solution on the solution of the classical *dual norm* problem, where the above equation can be written as $||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_*$. Firstly, Hölder's inequality gives

$$\begin{aligned}
&\epsilon^T\nabla_\theta L_{\mathbb{S}}(f_\theta)=\sum_{i=1}^{n}\epsilon_i^T\nabla_\theta L_{\mathbb{S}}(f_\theta)_i\leq\sum_{i=1}^{n}|\epsilon_i^T\nabla_\theta L_{\mathbb{S}}(f_\theta)_i|\\
&\leq||\epsilon^T\nabla_\theta L_{\mathbb{S}}(f_\theta)||_1\leq||\epsilon^T||_p||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q\leq\rho||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q.
\end{aligned} \tag{23}$$

https://doi.org/10.52202/079017-3146

So, we just need to find a ϵ that makes all the above inequality signs equal. Define m as $\text{sign}(\nabla_\theta L_{\mathbb{S}}(f_\theta))|\nabla_\theta L_{\mathbb{S}}(f_\theta)|^{q-1}$, then we can rewritten Eq. 23 as

$$\begin{aligned}\epsilon^T \nabla_\theta L_{\mathbb{S}}(f_\theta) &= \sum_{i=1}^{n} \text{sign}(\nabla_\theta L_{\mathbb{S}}(f_\theta)_i)|\nabla_\theta L_{\mathbb{S}}(f_\theta)_i|^{q-1} \nabla_\theta L_{\mathbb{S}}(f_\theta)_i \\ &= \sum_{i=1}^{n} |\nabla_\theta L_{\mathbb{S}}(f_\theta)_i||\nabla_\theta L_{\mathbb{S}}(f_\theta)_i|^{q-1} \\ &= ||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q^q.\end{aligned} \tag{24}$$

And we also get

$$||\epsilon||_p^p = \sum_{i=1}^{n} |\epsilon|^p = \sum_{i=1}^{n} |\text{sign}(\nabla_\theta L_{\mathbb{S}}(f_\theta))|\nabla_\theta L_{\mathbb{S}}(f_\theta)|^{q-1}|^p = ||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q^q, \tag{25}$$

where $1/p + 1/q = 1$. We choose a new ϵ, defined as $y = \rho \frac{\epsilon}{||\epsilon||_p}$, which satisfies: $||y||_p = \rho$, and substitute into $\epsilon^T \nabla_\theta L_{\mathbb{S}}(f_\theta)$:

$$y^T \nabla_\theta L_{\mathbb{S}}(f_\theta) = \sum_{i=1}^{n} y_i \nabla_\theta L_{\mathbb{S}}(f_\theta)_i = \sum_{i=1}^{n} \frac{\rho \nabla_\theta L_{\mathbb{S}}(f_\theta)_i}{||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_p} \nabla_\theta L_{\mathbb{S}}(f_\theta)_i = \frac{\rho}{||\epsilon||_p} \sum_{i=1}^{n} \epsilon_i \nabla_\theta L_{\mathbb{S}}(f_\theta)_i. \tag{26}$$

Due to $||\epsilon||_p = ||\nabla_\theta L_{\mathbb{S}}(f_\theta)_i||_q^{q/p}$ and $\epsilon^T \nabla_\theta L_{\mathbb{S}}(f_\theta) = ||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q^q$, we can further derive and obtain that

$$\frac{\rho}{||\epsilon||_p} \sum_{i=1}^{n} \epsilon_i \nabla_\theta L_{\mathbb{S}}(f_\theta)_i = \frac{\rho}{||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q^{q/p}} \sum_{i=1}^{n} \epsilon_i \nabla_\theta L_{\mathbb{S}}(f_\theta)_i = \rho ||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q. \tag{27}$$

Therefore, y can be rewritten as:

$$y = \rho \frac{\text{sign}(\nabla_\theta L_{\mathbb{S}}(f_\theta))|\nabla_\theta L_{\mathbb{S}}(f_\theta)|^{q-1}}{||\text{sign}(\nabla_\theta L_{\mathbb{S}}(f_\theta))|\nabla_\theta L_{\mathbb{S}}(f_\theta)|^{q-1}||_p} = \rho \frac{\text{sign}(\nabla_\theta L_{\mathbb{S}}(f_\theta))|\nabla_\theta L_{\mathbb{S}}(f_\theta)|^{q-1}}{||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_q^{q-1}}. \tag{28}$$

If $q = 2$, $y = \rho \frac{\nabla_\theta L_{\mathbb{S}}(f_\theta)}{||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_2}$. □

The above derivation is partly derived from Foret et al. (2020), to which we have added another part. To solve the SAM problem in deep learning (Foret et al., 2020), had to require two iterations to complete a single SAM-based gradient update. Another pivotal aspect to note is that within the context of dataset condensation, θ transitions from representing the model parameter f_θ to denoting the synthesized dataset $\mathcal{X}^{\mathcal{S}}$.

E Implementation of Flatness Regularization

As proved in Sec. D, the optimal solution ϵ^* is denoted as $\rho \frac{\nabla_\theta L_{\mathbb{S}}(f_\theta)}{||\nabla_\theta L_{\mathbb{S}}(f_\theta)||_2}$. Analogously, in the dataset condensation scenario, the joint optimization objective is given by $\sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}_{\text{syn}}^i(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})]$. There exists an optimal ϵ^*, which can be written as $\rho \frac{\nabla_{\mathcal{X}^{\mathcal{S}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}_{\text{syn}}^i(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})]}{||\nabla_{\mathcal{X}^{\mathcal{S}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}_{\text{syn}}^i(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})]||_2}$. Thus, a dual-stage approach of flatness regularization is shown below:

$$\begin{aligned}\mathcal{X}_{\text{new}}^{\mathcal{S}} &\leftarrow \mathcal{X}^{\mathcal{S}} + \frac{\rho}{||\nabla_{\mathcal{X}^{\mathcal{S}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}_{\text{syn}}^i(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})]||_2} \left(\nabla_{\mathcal{X}^{\mathcal{S}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}_{\text{syn}}^i(\mathcal{X}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})] \right) \\ \mathcal{X}_{\text{next}}^{\mathcal{S}} &\leftarrow \mathcal{X}_{\text{new}}^{\mathcal{S}} - \eta \left(\nabla_{\mathcal{X}_{\text{new}}^{\mathcal{S}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}_{\text{syn}}^i(\mathcal{X}_{\text{new}}^{\mathcal{S}}, \mathcal{X}^{\mathcal{T}})] \right),\end{aligned} \tag{29}$$

where η and $\mathcal{X}_{\text{next}}^{\mathcal{S}}$ denote the learning rate and the synthesized dataset in the next iteration, respectively. However, this optimization approach significantly increases the computational burden, thus reducing its scalability. Enlightened by Du et al. (2022), we consider a single-stage optimization strategy implemented via exponential moving average (EMA). Given an EMA-updated synthesized dataset $\mathcal{X}_{\text{EMA}}^{\mathcal{S}} = \beta \mathcal{X}_{\text{EMA}}^{\mathcal{S}} + (1-\beta)\mathcal{X}^{\mathcal{S}}$, where β is typically set to 0.99 in our experiments. The trajectories of the synthesized datasets updated via gradient descent (GD) and EMA

can be represented as $\{\theta^0_{\mathbf{GD}}, \theta^1_{\mathbf{GD}}, \cdots, \theta^N_{\mathbf{GD}}\}$ and $\{\theta^0_{\mathbf{EMA}}, \theta^1_{\mathbf{EMA}}, \cdots, \theta^N_{\mathbf{EMA}}\}$, respectively. Assume that $\mathbf{g}_j = \nabla_{\mathcal{X}^\mathcal{S}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\mathcal{X}^\mathcal{S}, \mathcal{X}^\mathcal{T})]$ at the j-th iteration, then $\theta^j_{\mathbf{EMA}} = \theta^j_{\mathbf{GD}} + \sum_{i=1}^{j-1} \beta^{j-i}\mathbf{g}_i$ with the condition $1 \leq j \leq N$[1], as outlined in Du et al. (2022). Consequently, we can provide the EMA-based SAM algorithm and applied to backbone sequential optimization in dataset condensation as follows:

$$\mathcal{L}_{\mathbf{FR}} = \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\mathcal{X}^\mathcal{S}, \mathcal{X}^\mathcal{S}_{\mathbf{EMA}})] = \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}}, \theta^j_{\mathbf{EMA}})], \qquad \text{at the } j\text{-th iteration.} \tag{30}$$

In the vast majority of dataset distillation algorithms (Yin and Shen, 2024; Shao et al., 2023; Zhou et al., 2024), the metric function used in matching is set to mean squared error (MSE) loss. Based on this phenomenon, we can rewrite Eq. 30 to Eq. 31, which guarantees flatness.

$$\begin{aligned}
&\nabla_{\theta^j_{\mathbf{GD}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}}, \theta^j_{\mathbf{EMA}})], \qquad \text{at the } j\text{-th iteration} \\
&= \nabla_{\theta^j_{\mathbf{GD}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}}, \mathcal{X}^\mathcal{T}) - \mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{EMA}}, \mathcal{X}^\mathcal{T})] \\
&= \nabla_{\theta^j_{\mathbf{GD}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}}, \mathcal{X}^\mathcal{T}) - \mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}} + \sum_{k=1}^{j-1} \beta^{j-k}\mathbf{g}_k, \mathcal{X}^\mathcal{T})] \\
&= \nabla_{\theta^j_{\mathbf{GD}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}}, \mathcal{X}^\mathcal{T}) - \mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}} + \beta^{j-1}\mathbf{g}_1, \mathcal{X}^\mathcal{T}) + \cdots \\
&\quad + \mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}} + \sum_{k=1}^{j-2} \beta^{j-k}\mathbf{g}_k, \mathcal{X}^\mathcal{T}) - \mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}} + \sum_{k=1}^{j-1} \beta^{j-k}\mathbf{g}_k, \mathcal{X}^\mathcal{T})] \\
&\approx \nabla_{\theta^j_{\mathbf{GD}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [(\beta^{j-1}\rho)||\nabla_{\theta^j_{\mathbf{GD}}} \mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}}, \mathcal{X}^\mathcal{T})||_2 + \cdots \\
&\quad + (\beta^1\rho)||\nabla_{\theta^j_{\mathbf{GD}} + \sum_{k=1}^{j-2} \beta^{j-k}\mathbf{g}_k} \mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}} + \sum_{k=1}^{j-2} \beta^{j-k}\mathbf{g}_k, \mathcal{X}^\mathcal{T})||_2] \qquad \# \text{ The solution of } \textit{dual norm} \text{ problem} \\
&\approx \nabla_{\theta^j_{\mathbf{GD}}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\sqrt{\mathbb{E}_{(\theta_1,\theta_2)\sim \mathbf{Unif}(\theta^j_{\mathbf{GD}}, \theta^j_{\mathbf{GD}} + \beta^{j-1}\mathbf{g}_1, \cdots, \theta^j_{\mathbf{GD}} + \sum_{k=1}^{j-1} \beta^{j-k}\mathbf{g}_k)} ||\nabla_{\theta_1} \mathcal{L}^i_{\mathbf{syn}}(\theta_1, \mathcal{X}^\mathcal{T})||_2 ||\nabla_{\theta_2} \mathcal{L}^i_{\mathbf{syn}}(\theta_2, \mathcal{X}^\mathcal{T})||_2}].
\end{aligned} \tag{31}$$

Thus, we can further obtain a SAM-like presentation.

$$\begin{aligned}
&\min_{\mathcal{X}^\mathcal{S}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathcal{L}^i_{\mathbf{syn}}(\theta^j_{\mathbf{GD}}, \theta^j_{\mathbf{EMA}})], \qquad \text{at the } j\text{-th iteration} \\
&= \min_{\mathcal{X}^\mathcal{S}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\mathbb{E}_{(\theta_1,\theta_2)\sim \mathbf{Unif}(\theta^j_{\mathbf{GD}}, \theta^j_{\mathbf{GD}} + \beta^{j-1}\mathbf{g}_1, \cdots, \theta^j_{\mathbf{GD}} + \sum_{k=1}^{j-1} \beta^{j-k}\mathbf{g}_k)} ||\nabla_{\theta_1} \mathcal{L}^i_{\mathbf{syn}}(\theta_1, \mathcal{X}^\mathcal{T})||_2 ||\nabla_{\theta_2} \mathcal{L}^i_{\mathbf{syn}}(\theta_2, \mathcal{X}^\mathcal{T})||_2] \\
&= \min_{\mathcal{X}^\mathcal{S}} \sum_{i=1}^{|\mathbb{S}_{\text{match}}|} [\max_{\epsilon:||\epsilon||\leq\rho} \mathbb{E}_{(\theta\sim\beta\theta^j_{\mathbf{GD}} + (1-\beta)\theta^j_{\mathbf{EMA}}, \beta\sim\mathcal{U}[0,1])} \mathcal{L}^i_{\mathbf{syn}}(\theta + \epsilon, \mathcal{X}^\mathcal{T})].
\end{aligned} \tag{32}$$

Consequently, optimizing Eq. 30 effectively addresses the SAM problem during the data synthesis phase, which results in a flat loss landscape. Additionally, Eq. 32 presents a variant of the SAM algorithm that slightly differs from the traditional form. This variant is specifically designed to ensure sharpness-aware minimization within a ρ-ball for each point along a straight path between $\theta^j_{\mathbf{GD}}$ and $\theta^j_{\mathbf{EMA}}$.

[1]Neglecting the learning rate for simplicity does not affect the derivation.

https://doi.org/10.52202/079017-3146

F Visualization of Prior Dataset Condensation Methods

In Fig. 5, we present the visualization results of previous training-dependent dataset condensation methods. These approaches, which optimize starting from Gaussian noise, tend to produce synthetic images that lack realism and fail to convey clear semantics to the naked eye.

SRe2L **CDA** **G-VBSM**

Figure 5: Visualization of the synthetic images of prior training-dependent dataset condensation methods.

G More Ablation Experiments

In this section, we present a series of ablation studies to further validate the design choices outlined in the main paper.

G.1 Backbone Choices of Data Synthesis on ImageNet-1k

Observer Model								Verified Model	
ResNet-18	MobileNet-V2	EfficientNet-B0	ShuffleNet-V2	WRN-40-2	AlexNet	ConvNext-Tiny	DenseNet-121	ResNet-18	ResNet-50
✓	✓	✓	✓					38.7	42.0
✓	✓	✓	✓	✓				36.7	43.3
✓	✓	✓	✓		✓			39.0	43.8
✓	✓	✓	✓	✓	✓			37.4	43.1
✓	✓	✓	✓	✓	✓	✓	✓	34.8	40.6

Table 12: **Ablation studies on ImageNet-1k with IPC 10.** Verify the influence of backbone choices on data synthesis with $\mathbb{CONFIG}$ C ($\zeta = 1.5$).

The results in Table 12 demonstrate the significant impact of backbone architecture selection on the performance of dataset distillation. This study employs the optimal configuration, which includes ResNet-18, MobileNet-V2, EfficientNet-B0, ShuffleNet-V2, and AlexNet.

G.2 Backbone Choices of Soft Label Generation on ImageNet-1k

Observer Model					Cost Time (s)	Verified Model		
ResNet-18	MobileNet-V2	EfficientNet-B0	ShuffleNet-V2	AlexNet		ResNet-18	ResNet-50	ResNet-101
✓	✓	✓	✓		598	9.1	9.5	6.2
✓	✓		✓		519	9.4	8.4	6.5
✓	✓		✓	✓	542	12.8	13.3	8.4

Table 13: **Ablation studies on ImageNet-1k with IPC 1.** Verify the influence of backbone choice on soft label generation with $\mathbb{CONFIG}$ G ($\zeta = 2$).

Our strategy better backbone choice, which focuses on utilizing lighter backbone combinations for soft label generation, significantly enhances the generalization capabilities of the condensed dataset. Empirical studies conducted with IPC 1, and the results detailed in Table 13, show that optimal performance is achieved by using ResNet-18, MobileNet-V2, EfficientNet-B0, ShuffleNet-V2, and AlexNet for data synthesis. For soft label generation, the combination of ResNet-18, MobileNet-V2, ShuffleNet-V2, and AlexNet demonstrates most effective.

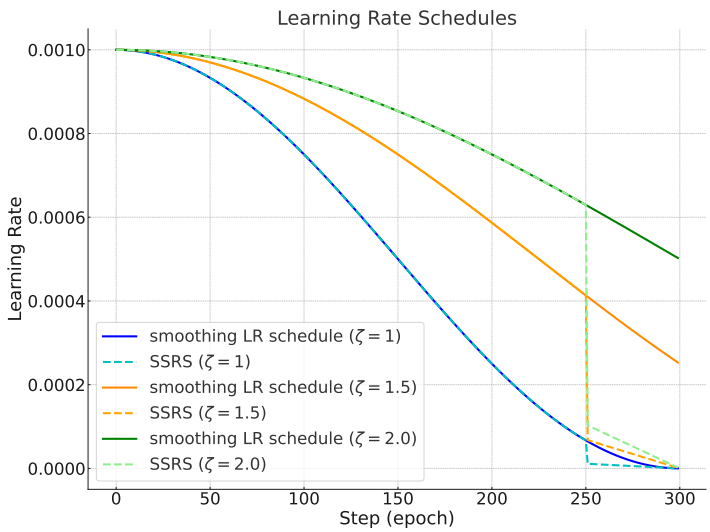


Figure 6: The visualization of SSRS and smoothing LR schedule.

G.3 Smoothing LR Schedule Analysis

Config	Slowdown Coefficient ζ				
	1.0	1.5	2.0	2.5	3.0
CONFIG C	24.5	28.2	30.6	32.4	31.8

Table 14: **Ablation studies on ImageNet-1k with IPC 10.** Additional experimental result of the slowdown coefficient ζ on the verified model MobileNet-V2.

Config	γ	Verified Model		
		ResNet-18	ResNet-50	ResNet-101
CONFIG F	0.997	47.6	53.5	52.0
CONFIG F	0.9975	47.4	54.0	50.9
CONFIG F	0.99775	47.3	53.7	50.3
CONFIG F	0.997875	47.8	53.8	50.7

Table 15: **Ablation studies on ImageNet-1k with IPC 10.** Verify the effectiveness of ALRS in post-evaluation.

Due to space limitations in the main paper, the experimental results for MobileNet-V2, which are not included in Table 3 Left, are presented in Table 14. Additionally, we investigate *Adaptive Learning Rate Scheduler* (ALRS), an algorithm that adjusts the learning rate based on training loss. Although ALRS did not produce effective results, it provides valuable insights for future research. This scheduler was first introduced in (Chen et al., 2022) and is described as follows:

$$\mu(i) = \mu(i-1)\gamma^{\mathbb{1}\left[\frac{|L_i - L_{i-1}|}{|L_i|} \leq h_1 \text{ and } |L_i - L_{i-1}| \leq h_2\right]},$$

Here, γ represents the decay rate, L_i is the training loss at the i-th iteration, and h_1 and h_2 are the first and second thresholds, respectively, both set by default to 0.02. We list several values of γ that demonstrate the best empirical performance in Table 15. These results allow us to conclude that our proposed smoothing LR schedule outperforms ALRS in the dataset condensation task.

Ultimately, we introduce a learning rate scheduler superior to the traditional smoothing LR schedule in scenarios with high IPC. This enhanced strategy, named *early Smoothing-later Steep Learning Rate Schedule* (SSRS), integrates the smoothing LR schedule with MultiStepLR. It intentionally implements a significant reduction in the learning rate during the final epochs of training to accelerate model convergence. The formal definition of SSRS is as follows:

$$\mu(i) = \begin{cases} \frac{1+\cos(i\pi/\zeta N)}{2} & , i \leq \frac{5N}{6}, \\ \frac{1+\cos(5\pi/\zeta 6)}{2} \frac{(6N-6i)}{6N} & , i > \frac{5N}{6}. \end{cases} \tag{33}$$

https://doi.org/10.52202/079017-3146

Config	Scheduler Type	Verified Model			
		ResNet-18	ResNet-50	ResNet-101	MobileNet-V2
CONFIG G	smoothing LR schedule	56.4	62.2	62.3	54.7
CONFIG G	SSRS	57.4	63.0	63.6	56.5

Table 16: **Ablation studies on ImageNet-1k with IPC 40.** Verify the effectiveness of SSRS in post-evaluation.

Note that the visualization of SSRS can be found in Fig. 6. Meanwhile, the comparative experimental results of SSRS and the smoothing LR schedule are detailed in Table 16. Notably, SSRS enhances the verified model's performance without incurring additional overhead.

G.4 Understanding of EMA-based Evaluation

CONFIG F	EMA Rate	0.99	0.999	0.9999	0.999945
	Accuracy	48.2	48.1	22.1	0.45

Table 17: **Ablation studies on ImageNet-1k with IPC 10.** Verify the effect of EMA Rate in EMA-based Evaluation.

The EMA Rate, a crucial hyperparameter governing the EMA update rate during post-evaluation, significantly influences the final results. Additional experimental outcomes, presented in Table 17, reveal that the EMA Rate 0.99 we adopt in the main paper yields optimal performance.

G.5 Ablation Studies on CIFAR-10

This section details the process of deriving hyperparameter configurations for CIFAR-10 through exploratory studies. The demonstrated superiority of our EDC method over traditional approaches, as detailed in our main paper, suggests that conventional dataset condensation techniques like MTT (Cazenavette et al., 2022) and KIP (Nguyen et al., 2020) are not the sole options for achieving superior performance on small-scale datasets.

Iteration	25	50	75	100	125	1000
Accuracy	42.1	42.4	42.7	42.5	42.3	41.8

Table 18: **Ablation studies on CIFAR-10 with IPC 10.** We employ ResNet-18 exclusively for data synthesis and soft label generation, examining the impact of iteration count during post-evaluation and adhering to RDED's consistent hyperparameter settings.

Data Synthesis		Soft Label Generation		Verified Model			
w/ pre-train	w/o pre-train	w/ pre-train	w/o pre-train	ResNet-18	ResNet-50	ResNet-101	MobileNet-V2
✗	✓	✗	✓	77.7	73.0	68.2	38.2
✗	✓	✓	✗	60.5	56.3	52.2	39.9
✓	✗	✓	✗	60.0	56.1	50.7	39.0
✓	✗	✗	✓	74.9	70.9	61.4	38.2

Table 19: **Ablation studies on CIFAR-10 with IPC 10.** Hyperparameter settings follow those in Table 10, excluding the scheduler and batch size, which are set to smoothing LR schedule ($\zeta = 2$) and 50, respectively.

EMA Rate	Batch Size	Verified Model			
		ResNet-18	ResNet-50	ResNet-101	MobileNet-V2
0.99	50	77.7	73.0	68.2	38.2
0.999	50	13.1	11.8	11.6	11.2
0.9999	50	10.0	10.0	10.0	10.0
0.99	25	78.1	76.0	71.8	42.1
0.99	10	76.0	70.0	57.7	42.9

Table 20: **Ablation studies on CIFAR-10 with IPC 10.** Explore the influence of EMA Rate and batch size in post-evaluation. Hyperparameter settings follow those in Table 10, excluding the scheduler, which are set to smoothing LR schedule ($\zeta = 2$).

Our quantitative experiments, detailed in Table 18, pinpoint 75 iterations as the empirically optimal count. This finding highlights that, for smaller datasets with limited samples and fewer categories, fewer iterations are required to achieve superior results.

Scheduler	Option	Verified Model			
		ResNet-18	ResNet-50	ResNet-101	MobileNet-V2
Smoothing LR Schedule	$\zeta = 2$	78.1	76.0	71.8	42.4
Smoothing LR Schedule	$\zeta = 3$	77.3	75.0	68.5	41.1
MultiStepLR	$\gamma = 0.5$, milestones=[800,900,950]	79.1	76.0	67.1	42.0
MultiStepLR	$\gamma = 0.25$, milestones=[800,900,950]	77.7	75.8	67.0	40.3

Table 21: **Ablation studies on CIFAR-10 with IPC 10.** Explore the influence of various scheduler in post-evaluation. Hyperparameter settings follow those in Table 10.

Subsequently, we evaluate the effectiveness of using a pre-trained model on ImageNet-1k for dataset condensation on CIFAR-10. Our study differentiates two training pipelines: the first involves 100 epochs of pre-training followed by 10 epochs of fine-tuning (denoted as 'w/ pre-train'), and the second comprises training from scratch for 10 epochs (denoted as 'w/o pre-train'). The results, presented in Table 19, indicate that pre-training on ImageNet-1k does not significantly enhance dataset distillation performance.

We further explore how batch size and EMA Rate affect the generalization abilities of the condensed dataset. Results in Table 20 show that a reduced batch size of 25 enhances performance on CIFAR-10.

In our final set of experiments, we compare MultiStepLR and smoothing LR schedules. As detailed in Table 21, MultiStepLR is superior for ResNet-18 and ResNet-50, whereas the smoothing LR schedule is more effective for ResNet-101 and MobileNet-V2.

H Synthesized Image Visualization

The visualization of the condensed dataset is showcased across Figs. 7 to 11. Specifically, Figs. 7, 9, 10, and 11 present the datasets synthesized from ImageNet-1k, Tiny-ImageNet, CIFAR-100, and CIFAR-10, respectively.

I Ethics Statement

Our research utilizes synthetic data to avoid the use of actual personal information, thereby addressing privacy and consent issues inherent in datasets with identifiable data. We generate synthetic data using a methodology that distills from real-world data but maintains no direct connection to individual identities. This method aligns with data protection laws and minimizes ethical risks related to confidentiality and data misuse. However, it is important to note that models trained on synthetic data may not achieve the same accuracy levels as those trained on the full original dataset.

J Limitations

The paper offers an extensive examination of the design space for dataset condensation, but it might still miss some potentially valuable strategies due to the broad scope. Additionally, as the IPC count grows, the performance of the described approach converges with that of the *baseline* RDED.

https://doi.org/10.52202/079017-3146

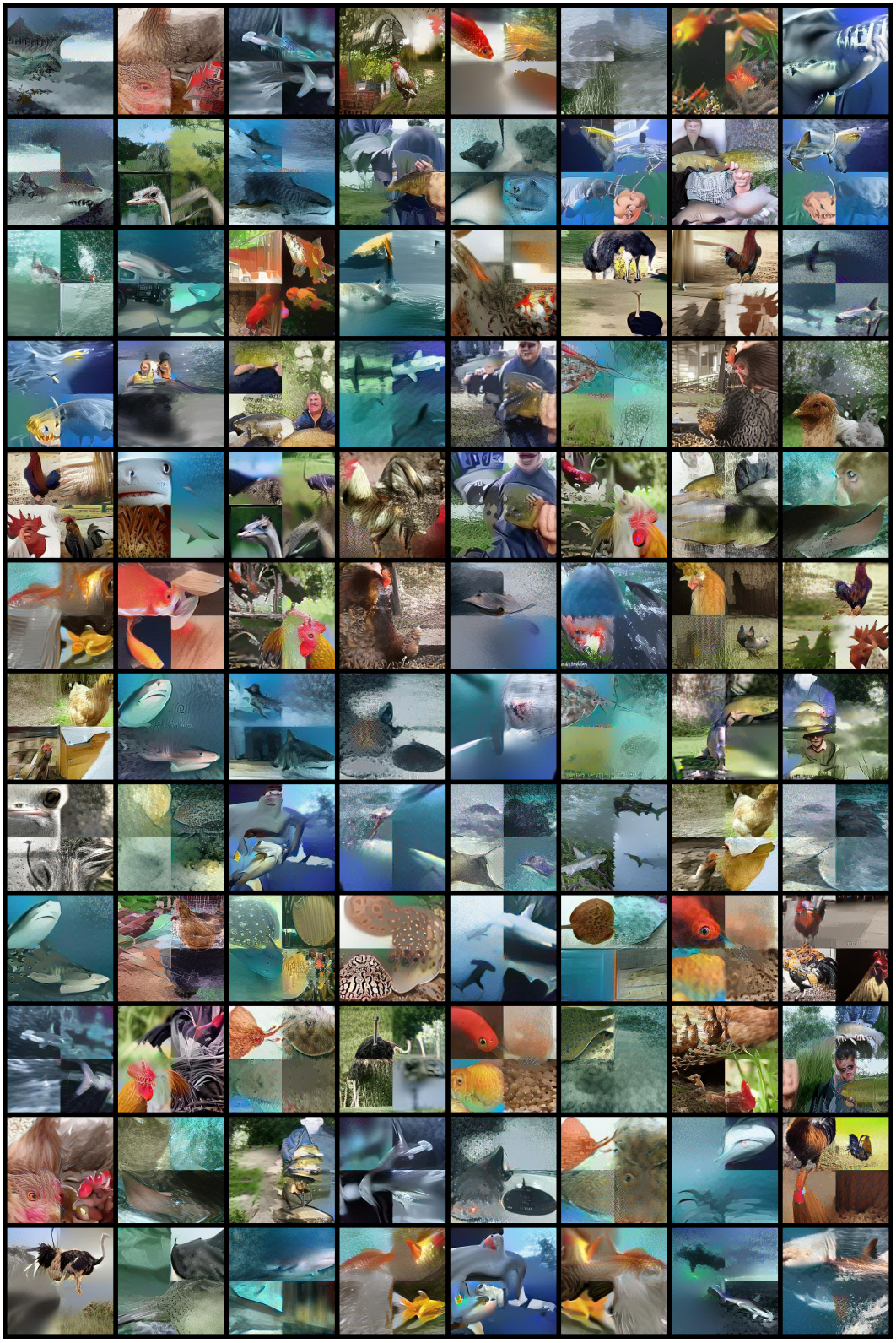


Figure 7: Synthetic data visualization on ImageNet-1k randomly selected from EDC.

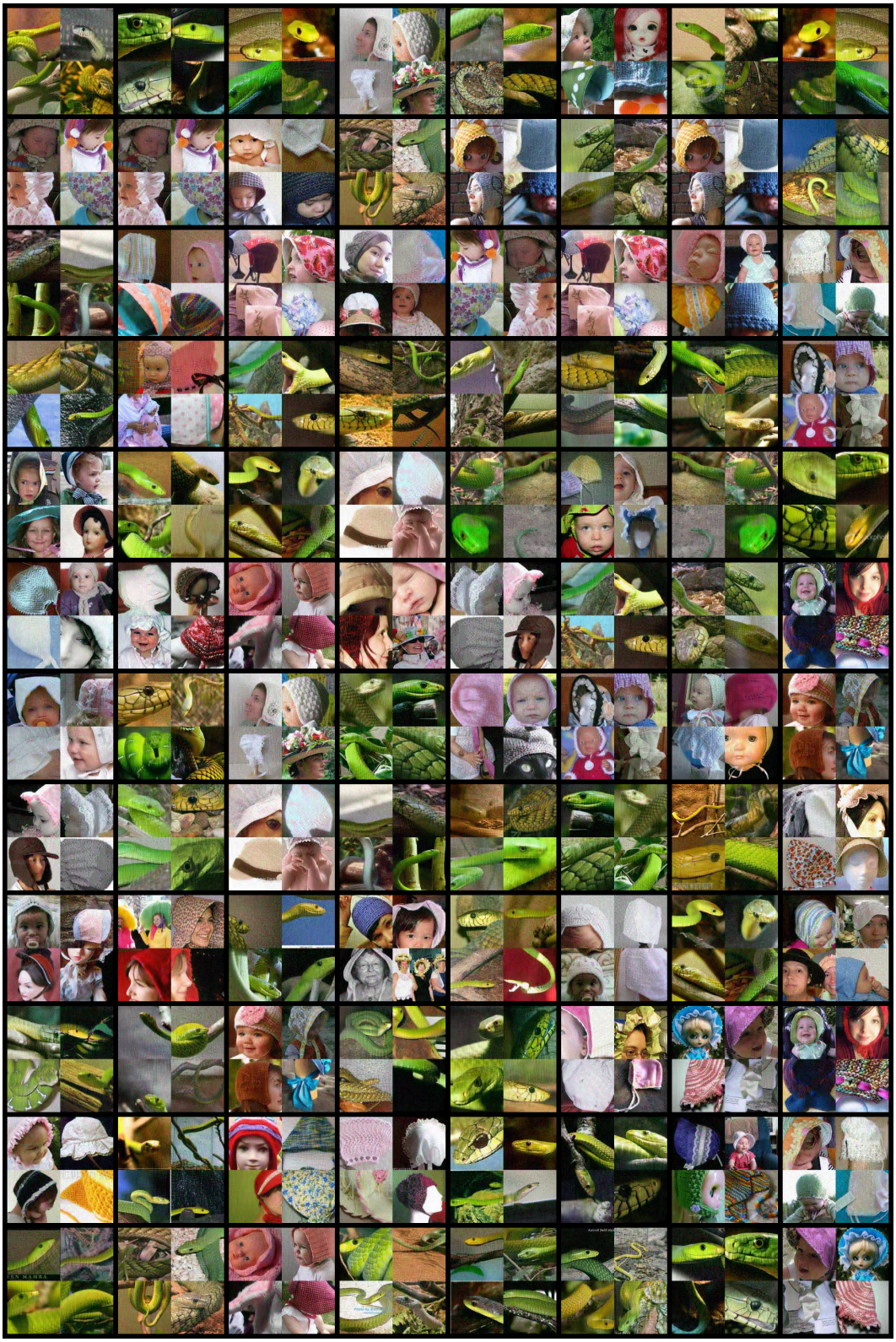


Figure 8: Synthetic data visualization on ImageNet-10 randomly selected from EDC.

https://doi.org/10.52202/079017-3146

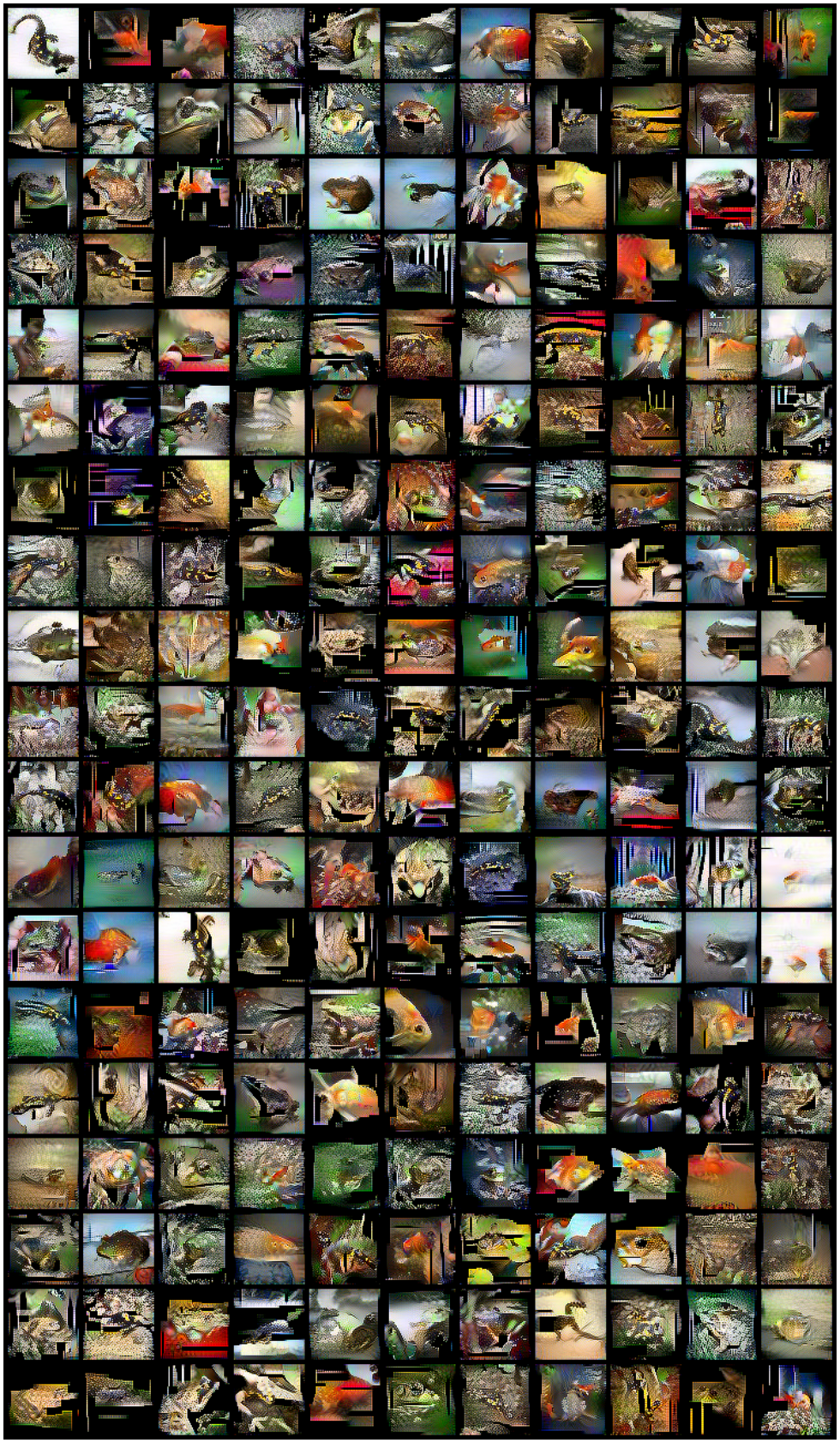


Figure 9: Synthetic data visualization on Tiny-ImageNet randomly selected from EDC.

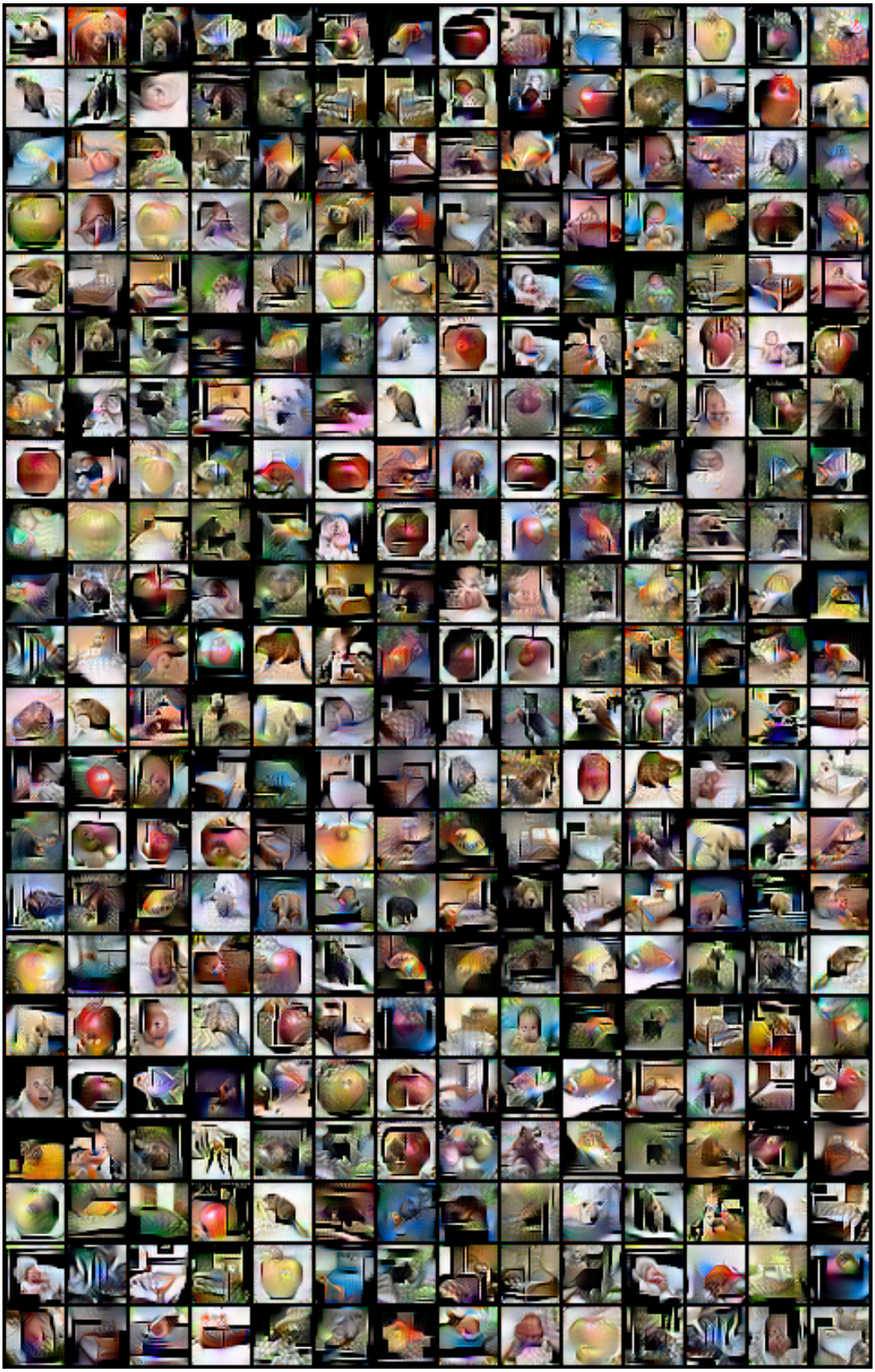


Figure 10: Synthetic data visualization on CIFAR-100 randomly selected from EDC.

https://doi.org/10.52202/079017-3146

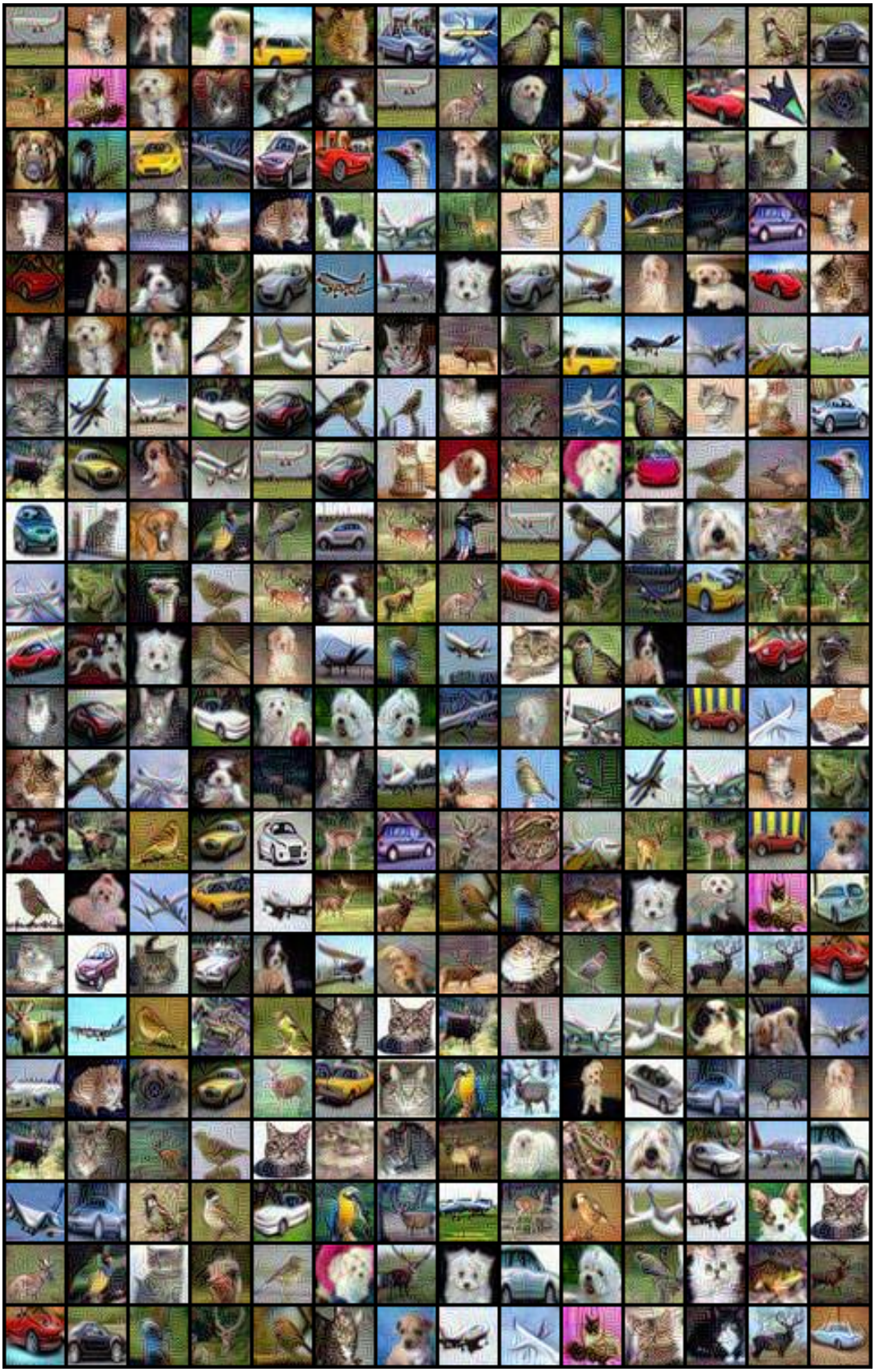


Figure 11: Synthetic data visualization on CIFAR-10 randomly selected from EDC.

K Additional Experiments, Theories and Descriptions (Rebuttal Stage Supplement)

Here we add some experiments, theories and explanations that we think it is necessary to add.

K.1 Scalability on ImageNet-21k

SRe^{2}L	CDA	RDED	EDC	Original Dataset
18.5	22.6	25.6	26.8	38.5

Table 22: **Comparison of Different Methods on ImageNet-21k.**

We conduct experiments on a larger scale dataset ImageNet-21k-P with IPC 10. The results in Table 22 indicate that our method outperforms the state-of-the-art method CDA (Yin and Shen, 2024) on this dataset, demonstrating that EDC can scale to larger datasets.

K.2 Complexity of Implementation

Configuration	GPU Memory (G/per GPU)	Time Spent (hours)	Top-1 Accuracy (%)
CONFIG A	4.616	9.77	31.4
CONFIG B	4.616	4.89	34.4
CONFIG C	4.616	4.89	38.7
CONFIG D	4.616	4.91	39.5
CONFIG E	4.697	4.91	46.2
CONFIG F	4.923	5.11	48.0
CONFIG G	4.923	5.11	48.6

Table 23: **Comparison of computational resources on 4 RTX 4090.**

Here we present Table 23 to complement the computational overhead in Fig. 1 in the main paper. EDC is an efficient algorithm as it reduces the number of iterations by half, compared to the *baseline* G-VBSM. As illustrated in the table above, although transitioning from CONFIG A to CONFIG G adds small GPU memory overhead, it is minor compared to the reduction in time spent. Additionally, introducing EDC to other tasks often requires significant effort for tuning hyper-parameters or even redesigning statistical matching, which is a challenge EDC should address.

K.3 Robustness Evaluation

Attack Methods	MTT	SRe2L	EDC (Ours)
Clean Accuracy	26.16	43.24	57.21
FGSM	1.82	5.73	12.39
PGD	0.41	2.70	10.71
CW	0.36	2.94	5.27
VMI	0.42	2.60	10.73
Jitter	0.40	2.72	10.64
AutoAttack	0.26	1.73	7.94

Table 24: **Comparison on DD-RobustBench.**

We follow the pipeline in Wu et al. (2024) to evaluate the robustness of models trained on condensed datasets, utilizing the well-known adversarial attack library available at Kim (2020). As illustrared in Table 24. Our experiments are conducted on Tiny-ImageNet with IPC 50, with the test accuracy presented in the table above. Evidently, EDC demonstrates significantly higher robustness compared to other methods. We attribute this to improvements in post-evaluation techniques, such as EMA-based evaluation and smoothing LR schedule, which help reduce the sharpness of the loss landscape.

K.4 Theoretical Explanation of Irrational Hyperparameter Setting (Sketch!!)

The smoothing LR schedule is designed to address suboptimal solutions that arise due to the scarcity of sample sizes in condensed datasets. Additionally, the use of small batch size is implemented

https://doi.org/10.52202/079017-3146

because the gradient of the condensed dataset more closely resembles the global gradient of the original dataset, as illustrated at the bottom of Fig. 2. Against the latter, we can propose a complete chain of theoretical derivation:

$$\begin{aligned}
\mathcal{L}_{syn} &= \mathbb{E}_{c_i \sim C} \| p_\theta(\mu|X^S, c_i) - p(\mu|X^T, c_i) \|_2 \\
&\quad + \| p_\theta(\sigma^2|X^S, c_i) - p(\theta^2|X^T, c_i) \|_2 \quad \text{\# (Our statistical matching)} \\
\partial L_{syn}/\partial\theta &= \int_{c_i} (\partial L_{syn}/\partial p_\theta(\cdot|X^S, c_i))(\partial p_\theta(\cdot|X^S, c_i)/\partial\theta) d_{c_i} \\
&\approx \int_{c_i} ([p_\theta(\mu|X^S, c_i) - p(\mu|X^T, c_i)] + [p_\theta(\sigma^2|X^S, c_i) - p(\sigma^2|X^T, c_i)])(\partial p_\theta(\cdot|X^S, c_i)/\partial\theta) d_{c_i}
\end{aligned} \tag{34}$$

where $p_\theta(|X^S, c_i)$ and $p(|X^T, c_i)$ refer to a Gaussian component in the Gaussian Mixture Model. Consider post-evaluation, We can derive the gradient of the MSE loss as:

$$\begin{aligned}
&\partial \mathbb{E}_{x_i \sim X^S} \| f_\theta(x_i) - y_i \|_2^2 / \partial\theta = 2\mathbb{E}_{x_i \sim X^S}[(f_\theta(x_i) - y_i)(\partial f_\theta(x_i)/\partial\theta)] \\
&= 2\mathbb{E}_{x_i \sim X^S}[(f_\theta(x_i) - y_i) \int_{c_i} (\partial f_\theta(x_i)/\partial p_\theta(\cdot|X^S, c_i))(\partial p_\theta(\cdot|X^S, c_i)/\partial\theta) d_{c_i}] \\
&\approx 2\mathbb{E}_{(x_j, x_i) \sim (X^S, X^T)}[(f_\theta(x_j) - y_j) \int_{c_i} (\partial f_\theta(x_i)/\partial p_\theta(\cdot|X^T, c_i))(\partial p_\theta(\cdot|X^T, c_i)/\partial\theta) d_{c_i}] \\
&\approx \partial \mathbb{E}_{x_i \sim X^T} || f_\theta(x_i) - y_i ||_2^2 / \partial\theta,
\end{aligned} \tag{35}$$

where θ stands for the model parameter. The right part of the penultimate row results from the loss $\mathcal{L}_{\text{syn}}$, which ensures the consistency of $p(\cdot|X^T, c_i)$ and $p(\cdot|X^S, c_i)$. If the model initialization during training is the same, the left part of the penultimate row is a scalar and has little influence on the direction of the gradient. Since X^T is the complete original dataset with a global gradient, the gradient of X^S approximates the global gradient of X^T, thus enabling the use of small batch size.

K.5 Additional Related Work

We additionally discuss the differences between published related papers (Sajedi et al., 2023; Zhang et al., 2024b; Deng et al., 2024) and our work.

DataDAM (Sajedi et al., 2023) vs. EDC. Both DataDAM and EDC do not require model parameter updates during training. However, DataDAM struggles to generalize effectively to ImageNet-1k because it relies on randomly initialized models for distribution matching. As noted in SRe^{2}L, models trained for fewer than 50 epochs can experience significant performance degradation. DataDAM does not explore the soft label generation and post-evaluation phases as EDC does, limiting its competitiveness.

DANCE (Zhang et al., 2024a) vs. EDC. DANCE is a DM-based algorithm that, unlike traditional distribution matching, does not require model updates during data synthesis. Instead, it interpolates between pre-trained and randomly initialized models, using this interpolated model for distribution matching. Similarly, EDC also does not need to update the model parameters, but it uses a pre-trained model with a different architecture and does not incorporate random interpolation. The "random interpolation" technique was not adopted because it did not yield performance gains on ImageNet-1k. Although DANCE considers both intra-class and inter-class perspectives, it limits inter-class analysis to the logit level and intra-class analysis to the feature map level. In contrast, EDC performs both intra-class and inter-class matching at the feature map level, where inter-class matching is crucial. To support this, last year, SRe^{2}L focused solely on inter-class matching at the feature map level and still achieved state-of-the-art performance on ImageNet-1k. EDC is the first dataset distillation algorithm to simultaneously improve data synthesis, soft label generation, and post-evaluation stages. In contrast, DANCE only addresses the data synthesis stage. While DANCE can be effectively applied to ImageNet-1k, the introduction of soft label generation and post-evaluation improvements is essential for DANCE to achieve more competitive results.

M3D (Zhang et al., 2024b) vs. EDC. M3D is a DM-based algorithm, but its data synthesis paradigm aligns with DataDAM by relying solely on randomly initialized models, which limits its generalization to ImageNet-1k. M3D, similar to SRe^{2}L, G-VBSM, and EDC, takes into account second-order information (variance), but this is not a unique contribution of EDC. The key contributions of EDC in data synthesis are real image initialization, flatness regularization, and the consideration of both intra-class and inter-class matching.

Deng et al. (Deng et al., 2024) vs. EDC. Deng et al. (Deng et al., 2024) is a DM-based algorithm, but its data synthesis paradigm is consistent with M3D and DataDAM, as it considers only randomly initialized models, which cannot be generalized to ImageNet-1k. Deng et al. (Deng et al., 2024) considers both interclass and intraclass information, similar to EDC. However, while EDC obtains interclass information by traversing the entire training set, Deng et al. (Deng et al., 2024) derives interclass information from only one batch, making its information richness inferior to that of EDC. Deng et al. (Deng et al., 2024) only explores data synthesis and does not explore soft label generation or post-evaluation. Additionally, Deng et al. (Deng et al., 2024) only shares some similarity with Soft Category-Aware Matching among the 10 design choices in EDC.

K.6 Implementation of Cropping

The implementation of this crop operation refers to `torchvision.transforms.RandomResizedCrop`, where the minimum area threshold is controlled by the parameter scale[0]. The default value is 0.08, meaning that the cropped image can be as small as 8% of the original image. Since 0.08 is too small for the model to extract complete semantic information during data synthesis, increasing the value to 0.5 resulted in a significant performance gain.

K.7 Comprehensive Comparison Experiment

Dataset	IPC	MTT	TESLA	SRe^{2}L	G-VBSM	CDA	WMDD	RDED	EDC (Ours)
CIFAR-10	1	-	-	-	-	-	-	22.9 ± 0.4	32.6 ± 0.1
	10	46.1 ± 1.4	48.9 ± 2.2	27.2 ± 0.4	53.5 ± 0.6	-	-	37.1 ± 0.3	79.1 ± 0.3
	50	-	-	47.5 ± 0.5	59.2 ± 0.4	-	-	62.1 ± 0.1	87.0 ± 0.1
CIFAR-100	1	-	-	2.0 ± 0.2	25.9 ± 0.5	-	-	11.0 ± 0.3	39.7 ± 0.1
	10	26.8 ± 0.6	27.1 ± 0.7	31.6 ± 0.5	59.5 ± 0.4	-	-	42.6 ± 0.2	63.7 ± 0.3
	50	-	-	49.5 ± 0.3	65.0 ± 0.5	-	-	62.6 ± 0.1	68.6 ± 0.2
Tiny-ImageNet	1	-	-	-	-	-	7.6 ± 0.2	9.7 ± 0.4	39.2 ± 0.4
	10	-	-	-	-	-	41.8 ± 0.1	41.9 ± 0.2	51.2 ± 0.5
	50	28.0 ± 0.3	-	41.1 ± 0.4	47.6 ± 0.3	48.7	59.4 ± 0.5	58.2 ± 0.1	57.2 ± 0.2
ImageNet-10	1	-	-	-	-	-	-	24.9 ± 0.5	45.2 ± 0.2
	10	-	-	-	-	-	-	53.3 ± 0.1	63.4 ± 0.2
	50	-	-	-	-	-	-	75.5 ± 0.5	82.2 ± 0.1
ImageNet-1k	1	-	-	-	-	-	3.2 ± 0.3	6.6 ± 0.2	12.8 ± 0.1
	10	-	17.8 ± 1.3	21.3 ± 0.6	31.4 ± 0.5	-	38.2 ± 0.2	42.0 ± 0.1	48.6 ± 0.3
	50	-	27.9 ± 1.2	46.8 ± 0.2	51.8 ± 0.4	53.5	57.6 ± 0.5	56.5 ± 0.1	58.0 ± 0.2

Table 25: **Comparison with the SOTA baseline dataset condensation methods.** MTT, TESLA, SRe^{2}L, CDA, WMDD and RDED utilize ResNet-18 for data synthesis, whereas G-VBSM and EDC leverage various backbones for this purpose.

Due to space constraints in the main paper and for aesthetic reasons, we have not fully presented the experimental results of other methods. However, since the benchmark for dataset distillation is uniform and well-recognized, the performance of other algorithms can be found in their respective papers. We present the related experimental results of the popular convolutional architecture ResNet-18 in Table 25.

https://doi.org/10.52202/079017-3146

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the introduction and abstract, we state a comprehensive design framework for dataset condensation, incorporating specific and effective strategies supported by empirical evidence and theoretical foundations.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please see Sec. J.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Please see Sec. B in Appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Our supplemental materials contain the reproducible code.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/079017-3146

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code has been provided in supplemental materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The details have been presented in Appendix A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please see Table 1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments are conducted using 4× RTX 4090 GPUs, as detailed in the experiment section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: Please see Sec. I.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please see Sec. I.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-3146

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: There are no risk factors present here.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In our paper and accompanying code, we have carefully cited and credited the works of G-VBSM and RDED, which form the foundation of our implementation.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We have attached our code and user instructions in the supplementary materials.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not have any experiments or research relevant to human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3146