Information Compensation: A Fix for Any-scale Dataset Distillation

Peng Sun[1,2] **Bei Shi**[2,*] **Xinyi Shang**[3] **Tao Lin**[2,†]
[1]**Zhejiang University** [2]**Westlake University** [3]**University College London**

sunpeng@westlake.edu.cn, shibei0430@gmail.com
xinyi.shang.23@ucl.ac.uk, lintao@westlake.edu.cn

Reviewed on OpenReview: https://openreview.net/forum?id=2SnmKd1JK4

Abstract

Dataset distillation, a recent machine learning paradigm, aims to compress large datasets into smaller, effective versions. In this paper, we introduce a near-Lossless Information Compression (LIC) approach that directly compresses the key information of original datasets into distilled forms with minimal information loss. Our LIC markedly surpasses existing solutions in both efficiency and effectiveness, demonstrating superior performance across a range of dataset sizes, from CIFAR-10 to ImageNet-1K. For instance, using a ResNet-18 backbone with IPC = 10, LIC distills the entire ImageNet-1K dataset in just 80 minutes, achieving a top-1 validation accuracy of 48%, significantly outperforming the SOTA method SRe^{2}L (Yin et al., 2023), which only attains 25% accuracy and requires five times longer to process. We will make our code publicly available.

1 Introduction

Dataset distillation (Wang et al., 2018) aims to distill the knowledge from a large training dataset into a very small set of synthetic training images such that a model trained on it can achieve comparable test performance as one trained on the original dataset. Distilled data provides the significant benefits of accelerating model training and reducing storage costs. As a highly effective paradigm, dataset distillation has shown its potential to facilitate applications in multiple domains, such as continual learning (Zhao et al., 2020; Rosasco et al., 2021), neural architecture search(Wang et al., 2021; Such et al., 2020), and privacy protection (Xiong et al., 2023; Dong et al., 2022; Chen et al., 2022).

Though prior dataset distillation methods (Cazenavette et al., 2022, 2023; Zhao et al., 2020; Wang et al., 2022; Zhao and Bilen, 2023; Zhao et al., 2023) have achieved great success, they mainly focus on small-scale and low-resolution datasets, such as CIFAR (Wang et al., 2018), Tiny-ImageNet (Cazenavette et al., 2022), while struggling with real-world datasets that are larger-scale and higher-resolution, e.g., ImageNet-1K (Deng et al., 2009). An information extraction-based approach, like SRe^{2}L (Yin et al., 2023), has recently been proposed and

1. † Corresponding author.
2. ∗ Equal contribution with the first author.

efficiently generalized on ImageNet-1K. Specifically, it involves a dual-compression process, where it condenses the information from a dataset into a pre-trained model via SQUEEZE, followed by further compressing the information from the model into distilled images and labels respectively using RECOVER and RELABEL. However, this line of research struggles with two issues (Sun et al., 2023): 1) poor cross-architecture generalization, and 2) significant performance drop on the distilled dataset under the setting of low `IPC` or low-resolution.

As our first contribution, we undermine the key factors behind these challenges and conjecture that they are primarily due to the *significant information loss* introduced by the dual-compression process (c.f. Appendix B for detailed exploration). We instead propose to drop the conventional dual-compression approach and resort to leveraging a simple squeezing operation and RELABEL strategy to distill information from both data and label space. Though efficient, its effectiveness is still largely limited by the information-compromised distilled images. As a remedy, we propose to compensate for the information loss in the distilled images by aligning the effective information between them and their original counterparts. Our approach demonstrates impressive adaptability and superior performance across various dataset sizes, model architectures, and image resolutions. Our contributions are summarized as follows:

- We propose LIC, a simple yet effective approach that compensates the information loss to enable near-Lossless Information Compression for dataset distillation.
- We demonstrate through extensive results that LIC notably outperforms all existing SOTA methods for any-scale original datasets. For example, it successfully distills a dataset with `IPC` $= 10$ from the full ImageNet-1K dataset within 1 hour, achieving an impressive 48% top-1 validation accuracy using ResNet-18 (He et al., 2016).

2 Methodology

To compensate for the lost information in compressed images, we introduce a three-stage efficient dataset distillation framework (c.f. Appendix C). This framework begins by selecting a set of key samples (c.f. Section 2.1). Subsequently, it employs a squeezing-based technique to condense these chosen samples into a distilled dataset, followed by compensating information loss (c.f. Section 2.2). Furthemore, the RELABEL method (Shen and Xing, 2022) is employed, innovatively transferring the knowledge from the full original dataset into the more compact label space of the distilled dataset (c.f. Section 2.3).

2.1 Selecting Key Samples

Our goal is to devise an effective method for identifying the crucial samples that are instrumental in benefiting the RELABEL process. Motivated by the assumption in Proposition 1—which suggests that each chosen sample $(\mathbf{x}_i, y_i)$ is supposed to receive an informative and accurate label $\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}_i)$ from the pre-trained model during the RELABEL phase—the goal then relaxes to find the samples that are relabeled most accurately by the pre-trained model $\phi_{\boldsymbol{\theta}_\mathcal{T}}$. Thus, we introduce a loss-based importance score s_i for each sample pair $(\mathbf{x}_i, y_i)$, defined as

$$s_i = -\ell(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}_i), y_i)\,. \tag{1}$$

The key sample selection procedure can be detailed below. Let $\mathcal{T}_c := \{(\mathbf{x}_i, y_i) \mid (\mathbf{x}_i, y_i) \in \mathcal{T}, y_i = c\}$ represents the subset of the dataset $\mathcal{T}$, containing only those samples $(\mathbf{x}_i, y_i)$ that are labeled with the class c. For each class data $\mathcal{T}_c$, we identify the key samples $\mathbf{x}_i$ based on their importance scores s_i. Specifically, we select those samples for which $s_i \geq \bar{s}$, where $\bar{s}$ denotes a predetermined threshold[3].

We further simplify the whole procedure due to the computational overhead and diversity issue, namely 1) computing the importance score for every sample $\mathbf{x}_i$ in an entire class data $\mathcal{T}_c$ presents a significant computational challenge, and 2) focusing solely on samples that closely align with the true label can lead to a lack of diversity within the selected samples. In detail, we utilize a pre-selection strategy inspired by Sun et al. (2023), which involves selecting a subset[4] $\mathcal{T}_c' \subset \mathcal{T}_c$ uniformly at random to serve as a proxy for the entire $\mathcal{T}_c$. Such a pre-selection strategy not only promotes diversity in the data but also lessens the computational load (Sun et al., 2023), thereby laying the groundwork for our subsequent score-based sample selection process.

2.2 Information Compensation for Compressed Data

In this subsection, we first formally define the squeezing operation for images and then detail the case of effective information and the corresponding information loss. A method is further devised based on information compensation to ensure the preservation of information integrity in the squeezed images relative to their original versions.

On the loss of effective information. For the selected key samples (images), we compress them into a more compact pixel space. However, given the constraints of limited pixel space storage, compressing multiple images directly into one can result in a significant reduction in the fineness and detail of the original information (c.f. Appendix B.2). We start with the definition of information compression below.

Definition 1 (Image Squeezing and Expanding) *Let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of N images, where each image $\mathbf{x}_i \in \mathbb{R}^d$ resides in a d-dimensional space. We define a squeezing operation[5] f that transforms this set of images into a single squeezed image $\mathbf{x}_j^* \in \mathbb{R}^d$ as follows:*

$$\mathbf{x}_j^* = \mathcal{F}(\{\mathbf{x}_i\}_{i=1}^N)\,, \tag{2}$$

where $\mathbf{x}_j^*$ shares the same dimensional space as each $\mathbf{x}_i$. Conversely, the expanding operation $\mathcal{F}^{-1}$ reconstructs a set of images $\{\hat{\mathbf{x}}_i\}_{i=1}^N$ from the squeezed image $\mathbf{x}_j^*$, each in the same d-dimensional space:

$$\{\hat{\mathbf{x}}_i\}_{i=1}^N = \mathcal{F}^{-1}(\mathbf{x}_j^*)\,. \tag{3}$$

Note that the expanded set of images, denoted as $\{\hat{\mathbf{x}}_i\}_{i=1}^N$, closely resemble the original set $\{\mathbf{x}_i\}_{i=1}^N$, albeit typically with a certain level of information degradation.

[3] This threshold is determined by the number of samples necessitated for further use, i.e., the `IPC`.

[4] We use the default size of this selected subset in Sun et al. (2023), see details in Section 3.1.

[5] In practical terms, the squeezing operation for images involves resizing and concatenating them. For instance, when handling 4 images, each with a resolution of 224 × 224 pixels, we first downscale each image to 112 × 112 pixels. Then, we concatenate these resized images into a single image, restoring the resolution to 224 × 224 pixels.

We then define the effective information of a data sample.

Definition 2 (Observation-based Effective Information) *Let $\mathbf{x}_i$ represent a sample from any domain (e.g., image). Define an observer group $\mathcal{R} = \{r_j\}$, where each observer r_j is capable of extracting or interpreting features from $\mathbf{x}_i$, and $|\mathcal{R}| \geq 1$. The effective information of the sample $\mathbf{x}_i$, as observed by the group $\mathcal{R}$, is conceptualized as the distribution $p_{\mathbf{x}_i|\mathcal{R}}(z)$. This distribution is formulated as:*

$$p_{\mathbf{x}_i|\mathcal{R}}(z) := \{z | z = r_j(\mathbf{x}_i), \forall r_j \in \mathcal{R}\}\,. \tag{4}$$

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

The Definition 2 posits that the effective information of a sample encompasses the set of its features as perceived or extracted by a diverse set of observers. Samples are considered to have similar effective information if they result in comparable feature sets across the observers in $\mathcal{R}$.

Compensating the loss of effective information. For a given squeezed sample $\mathbf{x}_j^*$, we aim to find $\Delta\mathbf{x}$ that restores the effective information in the compensated squeezed image $(\mathbf{x}_j^* + \Delta\mathbf{x})$ to match that of the original images $\{\mathbf{x}_i\}_{i=1}^N$. Specifically, for a compensated squeezed image $(\mathbf{x}_j^* + \Delta\mathbf{x})$, the corresponding expanded images are given by

$$\{\hat{\mathbf{x}}_i\}_{n=1}^N = \mathcal{F}^{-1}(\mathbf{x}_j^* + \Delta\mathbf{x})\,. \tag{5}$$

To align the effective information in these expanded images $\{\hat{\mathbf{x}}_i\}_{n=1}^N$ with the original images $\{\mathbf{x}_i\}_{i=1}^N$, we seek an optimal $\Delta\mathbf{x}$ through the following minimization:

$$\arg\min_{\Delta\mathbf{x}} \mathbb{E}_{\mathbf{x}_i,\hat{\mathbf{x}}_i} \mathrm{D}_{\mathrm{KL}}(p_{\mathbf{x}_i|\mathcal{R}} || p_{\hat{\mathbf{x}}_i|\mathcal{R}})\,. \tag{6}$$

Given the direct correlation between $\hat{\mathbf{x}}_i$ and $\mathbf{x}_i$, along with the shared observer group $\mathcal{R}$ employed to assess their effective information, (6) can be further simplified to:

$$\arg\min_{\Delta\mathbf{x}} \mathbb{E}_{\mathbf{x}_i,\hat{\mathbf{x}}_i} \mathbb{E}_{r_j} \|r_j(\mathbf{x}_i) - r_j(\hat{\mathbf{x}}_i)\|\,. \tag{7}$$

To reduce computational complexity, we focus on a specific scenario where the observer group consists of only one pre-trained model across various transformations[6]. Let $\mathcal{R} = \{r_j \mid r_j = g_j \circ \phi_{\boldsymbol{\theta}_\mathcal{T}}, \forall g_j \sim \mathcal{G}\}$, where $\mathcal{G}$ denotes the transformation group. Therefore, to achieve (7), our loss function is defined as

$$\mathcal{L}_{\Delta\mathbf{x}} = \mathbb{E}_{\mathbf{x}_i,\hat{\mathbf{x}}_i} \mathbb{E}_{g_j} \|\phi_{\boldsymbol{\theta}_\mathcal{T}}(g_j(\mathbf{x}_i)) - \phi_{\boldsymbol{\theta}_\mathcal{T}}(g_j(\hat{\mathbf{x}}_i))\|_2^2\,. \tag{8}$$

By minimizing (8), we find the optimal compensation $\Delta\mathbf{x}^\star$ and capture the compensated squeezed image $\widetilde{\mathbf{x}}_j = \mathbf{x}_j^* + \Delta\mathbf{x}^\star$ as the distilled image.

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

Balancing the semantic and textural information. Directly generating the compensated squeezed image ($\mathbf{x}_j^* + \Delta\mathbf{x}$) through aligning its effective information with the original image set $\{\mathbf{x}_i\}_{i=1}^N$ may lead to a substantial loss of texture information. The reason is that the model $\phi_{\boldsymbol{\theta}_\mathcal{T}}$ tends to extract semantic information from input images, which often results in a notable drop in the texture details. Therefore, to achieve a balance between semantic richness and texture preservation in the distilled image $\widetilde{\mathbf{x}}_j$, we leverage intermediate model features instead of the last-layer logits, which helps in retaining more textural details (see Section 3.3 for more details).

2.3 Relabel with Observer Model

Instead of using the basic one-shot label, we propose a re-label strategy for our multi-subfigure distilled images $\widetilde{\mathbf{x}}_j$, which provides more informative and diverse knowledge. The proposed module is inspired by Yun et al. (2021) that a random image crop might contain a different object than the one originally labeled and then lead to inaccurate or misleading training data, which illustrates that the one-shot label strategy is challenging to express enough knowledge for cropped images.

Re-label strategy can be implemented by using the soft labeling approach (Shen and Xing, 2022) to generate region-level soft labels $\widetilde{y}_j^k = \ell\left(\phi_{\boldsymbol{\theta}_\mathcal{T}}(\widetilde{\mathbf{x}}_j^k)\right)$, where $\widetilde{\mathbf{x}}_j^k$ is the k-th region in the distilled image $\widetilde{\mathbf{x}}_j$ and $\widetilde{y}_j^k$ is the soft label.

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

$$\mathcal{L} = -\sum_j \sum_k \widetilde{y}_j^k \log \phi_{\boldsymbol{\theta}_\mathcal{S}}(\widetilde{\mathbf{x}}_j^k)\,. \tag{9}$$

3 Experiment

In this section, we evaluate the performance of our proposed LIC over various datasets and neural network architectures. First, we demonstrate the superior results of LIC on real-world datasets, cross-architecture generalization and efficiency. Next, we conduct extensive ablation experiments to investigate the effect of each component of our method.

3.1 Experimental Setting

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

- **Small-scale:** we evaluate on two datasets, including CIFAR-10 (32 × 32) (Krizhevsky et al., 2009b) and CIFAR-100 (32 × 32) (Krizhevsky et al., 2009a).
- **Large-scale:** we also use two large-scale high-resolution datasets including Tiny-ImageNet (64 × 64) (Le and Yang, 2015) and ImageNet-1K (224 × 224) (Deng et al., 2009).

Similar to the prior dataset distillation works (Yin et al., 2023; Zhao et al., 2023; Guo et al., 2023), we employ ConvNet (Guo et al., 2023), ResNet-18 (He et al., 2016), MobileNet-V2 (Sandler et al., 2018), as our backbone networks on all datasets. Specifically, for ConvNet, we

Table 1: **Comparison with baseline models.** In the table, **bold** means the best result, and entries with "-" are absent due to scalability problems. See Appendix D for more details.

Architecture		ConvNet							ResNet-18	
Dataset	IPC	MTT	IDM	TESLA	DATM	DREAM	SRe^{2}L	LIC (Ours)	SRe^{2}L	LIC (Ours)
CIFAR-10	1	46.3 ± 0.8	45.6 ± 0.7	48.5 ± 0.8	46.9 ± 0.5	51.1 ± 0.3	19.2 ± 1.1	**51.7 ± 0.6**	18.8 ± 0.5	**40.7 ± 0.4**
	10	65.3 ± 0.7	58.6 ± 0.1	66.4 ± 0.8	66.8 ± 0.2	69.4 ± 0.4	34.7 ± 1.1	**72.2 ± 0.2**	38.0 ± 0.3	**80.9 ± 0.4**
	50	71.6 ± 0.2	67.5 ± 0.1	72.6 ± 0.7	76.1 ± 0.3	74.8 ± 0.1	54.8 ± 0.6	**78.3 ± 0.1**	65.6 ± 1.0	**89.2 ± 0.0**
CIFAR-100	1	24.3 ± 0.3	20.1 ± 0.3	24.8 ± 0.5	27.9 ± 0.2	29.5 ± 0.3	14.4 ± 0.1	**42.7 ± 0.3**	11.8 ± 0.5	**48.1 ± 0.7**
	10	40.1 ± 0.4	45.1 ± 0.1	41.7 ± 0.3	47.2 ± 0.4	46.8 ± 0.7	40.3 ± 0.3	**54.8 ± 0.2**	46.0 ± 0.4	**64.7 ± 0.1**
	50	47.7 ± 0.2	50.0 ± 0.2	47.9 ± 0.3	55.0 ± 0.2	52.6 ± 0.4	56.3 ± 0.1	**56.6 ± 0.1**	60.2 ± 0.1	**67.6 ± 0.1**
Tiny-ImageNet	1	8.8 ± 0.3	10.1 ± 0.2	-	17.1 ± 0.3	10.0 ± 0.4	4.6 ± 0.3	**31.7 ± 0.4**	5.7 ± 0.2	**36.5 ± 0.2**
	10	23.2 ± 0.2	21.9 ± 0.3	-	31.1 ± 0.3	-	22.5 ± 0.3	**46.3 ± 0.2**	33.8 ± 0.6	**51.7 ± 1.0**
	50	28.0 ± 0.3	27.7 ± 0.3	-	39.7 ± 0.3	29.5 ± 0.3	42.3 ± 0.3	**47.4 ± 0.1**	51.0 ± 0.4	**55.8 ± 0.9**
ImageNet-1K	1	-	-	7.7 ± 0.2	-	-	1.3 ± 0.1	**14.5 ± 0.3**	1.4 ± 0.1	**6.8 ± 0.2**
	10	-	-	17.8 ± 1.3	-	18.4 ± 0.9	9.3 ± 0.4	**24.0 ± 0.5**	25.3 ± 0.4	**48.5 ± 0.2**
	50	-	-	27.9 ± 1.2	-	-	38.8 ± 0.4	**39.1 ± 0.2**	45.7 ± 0.4	**60.0 ± 0.2**

use Conv-3 on CIFAR-10/100, and use Conv-4 on Tiny-ImageNet and ImageNet-1K. More details about the used datasets and architectures can be found in Appendix D.

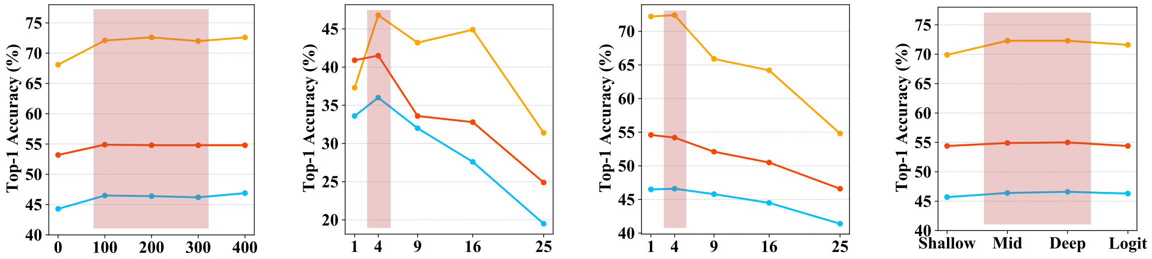
Baselines. We compare our method with several optimization-based distillation methods that can scale to large high-resolution datasets, including MTT (Cazenavette et al., 2022), IDM (Zhao et al., 2023), TESLA (Cui et al., 2023), DATM (Guo et al., 2023), ADD (Zhang et al., 2023), DREAM (Liu et al., 2023b), and SRe^{2}L (Yin et al., 2023). To the best of our knowledge, SRe^{2}L is the only published work that can efficiently scale to any-scale dataset, so we consider it as our closest baseline. More details about these methods can be found in Appendix D.

Implementation details. All the hyper-parameters used in our Algorithm 1 are general, insensitive and easy-implemented for all datasets and network architectures (c.f. Section 3.3 and Appendix F for validation). We employ a generalized configuration for $\mathcal{T}'$ (c.f. Section 2.1 for definition), where the size $|\mathcal{T}'|$ is set as 300. We set the number $N = 4$ of images squeezed in a distilled image (c.f. Section 2.2 for definition) and number $M = 200$ of compression iteration (c.f. Algorithm 1 for definition). More implementation details are provided in Appendix D.

3.2 Comparison with the SOTA Methods

Results on CIFAR and ImageNet. Following previous research(Cazenavette et al., 2022; Cui et al., 2023; Zhao et al., 2023), we set `IPC` to 1, 10, and 50 to compare with baselines on varying datasets and networks. As the results reported in Table 1, *our method LIC outperforms other methods on varying datasets and neural networks with different* `IPC`. It is noteworthy that prior information extraction-based solutions like SRe^{2}L struggle in scenarios involving small distilled datasets such as CIFAR-100 with `IPC` = 1 or CIFAR-10 with all `IPC`, further verifying our claims proposed in Section B. Detailed comparison among more datasets and concurrent baselines, is in Appendix E.

Cross-architecture generalization. An important property of the distilled datasets is their good generalization capability across unseen architectural models. Here we evaluate



(a) iteration number M (b) number N ($\texttt{IPC} = 1$) (c) number N (d) alignment layer

Figure 1: **Ablation study on each component in our LIC.** We evaluate the distilled dataset of our LIC with different number M of compression iterations (1a), number N of images squeezed in one distilled image (1b & 1c), feature alignment layer (1d). The yellow ●, red ●, and blue ● denote CIFAR-10, CIFAR-100, and Tiny-ImageNet respectively.

Table 2: **Tiny-ImageNet top-1 accuracy on cross-architecture generalization.** We use Conv-4, ResNet-18, and MobileNet-V2 to distill the original dataset, and then transfer distilled data to each other architecture.

Verifier \ Observer		Conv-4	ResNet-18	MobileNet-V2
Conv-4	SRe^{2}L	23.0 ± 0.3	14.7 ± 0.3	19.5 ± 0.3
	Ours	**46.6** ± **0.3**	**26.0** ± **1.0**	**25.9** ± **0.3**
ResNet-18	SRe^{2}L	32.5 ± 0.6	33.6 ± 0.3	36.4 ± 0.2
	Ours	**47.4** ± **0.3**	**51.9** ± **0.2**	**44.9** ± **0.2**
MobileNet-V2	SRe^{2}L	8.9 ± 0.4	8.9 ± 0.1	20.3 ± 0.4
	Ours	**36.7** ± **0.5**	**34.6** ± **0.8**	**44.5** ± **0.3**

the generalizability of our distilled datasets when $\texttt{IPC} = 10$. As reported in Table 2, *our distilled dataset performs best on unseen networks*, which reflects the good generalizability of the data and labels distilled by our method. Furthermore, our success stems from that our LIC effectively keeps both textural and semantic information in distilled images, which is evidenced by Cazenavette et al. (2023).

Efficiency comparison. Efficiency is also a key factor during the process of distilling data. Here, We use a single RTX-4090 GPU for two methods to conduct experiments on Tiny-ImageNet. As evidenced in Table 3, *our method LIC achieves superior efficiency in comparison to SOTA methods*, demonstrating a notable advantage of efficacy and efficiency. Significantly, our algorithm offers a versatile peak memory capacity, enabling adjustments to batch size dynamically without sacrificing performance. This efficiency is attributed to the fact that our Algorithm 1 can independently optimize images, allowing us to distill them one by one. More comparisons are in Appendix E.

3.3 Ablation Study

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

Table 3: **Efficiency comparison with SRe^{2}L (Yin et al., 2023) on varying networks on Tiny-ImageNet.** Following SRe^{2}L, Time Cost represents the consumption when generating 100 images simultaneously, and the peak value of GPU memory is measured with a batch size of 100.

Architecture		Time Cost (s)	Peak Memory (GB)
Conv-4	SRe^{2}L	51.68	1.36
	Ours	**13.02**	**0.65**
ResNet-18	SRe^{2}L	191.14	3.62
	Ours	**25.34**	**1.56**
MobileNet-V2	SRe^{2}L	114.05	1.27
	Ours	**18.81**	**0.64**

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

Influence of size of squeezed images N. Specifically, though we can squeeze more images from $\mathcal{T}$ into a distilled dataset $\mathcal{S}$ by increasing N increases to benefit the data diversity, it also results in a lower resolution for the source images (see the definition of squeezing in Footnote 5), thus hurting the textural information. Figure 1c & 1b showcases that *the validation performance rises to the highest on selected three datasets when $N = 4$.*

Why and how to choose the feature alignment layer? The experimental results in Figure 1d intuitively demonstrate the impact of the feature alignment layer, alongside the discussion in Section 2.2. As depicted in Figure 1d, *optimal performance is often achieved through alignment at the deep layer.* A plausible explanation for this observation is the differing information encoded at various network depths: shallow layers tend to capture more textural details, whereas logit layers are more adept at encoding semantic information. Consequently, to effectively infuse the distilled dataset with a suitable balance of information, we harness the feature alignment capabilities of the deep layer.

4 Conclusion

We demonstrate that LIC markedly surpasses existing SOTA techniques in the aspects of effectiveness and efficiency across a range of dataset sizes and network architectures. Additionally, we highlight the efficiency of LIC by showcasing its ability to distill the ImageNet-1k dataset in just 80 minutes. We apply our LIC to a continual learning task, with detailed results presented in Appendix G. We also visualize the distilled images in Appendix H.

References

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4750–4759, 2022.

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

Dingfan Chen, Raouf Kerkouche, and Mario Fritz. Private set generation with discriminative information. *Advances in Neural Information Processing Systems*, 35:14678–14690, 2022.

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

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

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

Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *International Conference on Machine Learning*, pages 5378–5396. PMLR, 2022.

Jiawei Du, Yidi Jiang, Vincent YF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3749–3758, 2023.

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

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

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

Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pages 11102–11118. PMLR, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009a.

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

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

Haoyang Liu, Tiancheng Xing, Luwei Li, Vibhu Dalal, Jingrui He, and Haohan Wang. Dataset distillation via the wasserstein metric. *arXiv preprint arXiv:2311.18531*, 2023a.

Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. Dream: Efficient dataset distillation by representative matching. *arXiv preprint arXiv:2302.14416*, 2023b.

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

Andrea Rosasco, Antonio Carta, Andrea Cossu, Vincenzo Lomonaco, and Davide Bacciu. Distilled replay: Overcoming forgetting through synthetic samples. In *International Workshop on Continual Semi-Supervised Learning*, pages 104–117, 2021.

Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. Datadam: Efficient dataset distillation with attention matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17097–17107, 2023.

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

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

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

Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, and Jeffrey Clune. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In *International Conference on Machine Learning*, pages 9206–9216, 2020.

Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. *arXiv preprint arXiv:2312.03526*, 2023.

Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12196–12205, 2022.

Ruochen Wang, Minhao Cheng, Xiangning Chen, Xiaocheng Tang, and Cho-Jui Hsieh. Rethinking architecture selection in differentiable nas. 2021.

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

Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. Feddm: Iterative distribution matching for communication-efficient federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16323–16332, 2023.

Zeyuan Yin and Zhiqiang Shen. Dataset distillation in large data era. *arXiv preprint arXiv:2311.18838*, 2023.

Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *arXiv preprint arXiv:2306.13092*, 2023.

Ruonan Yu, Songhua Liu, and Xinchao Wang. Dataset distillation: A comprehensive review. *arXiv preprint arXiv:2301.07014*, 2023.

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

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

Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, 2021.

Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023.

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020.

Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023.

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

Appendix A. Related Work

Wang et al. (2018) first introduces the dataset distillation as a bi-level meta-learning optimization problem. The outer loop aims at optimizing the meta-dataset, while the inner loop focuses on training models using the distilled dataset. Existing methods can be roughly divided into three paradigms for solving this bi-level optimization problem.

Uni-level optimization-based paradigm. Tackling this bi-level problem is complex, especially when optimizing proxy models via gradient descent, which involves unraveling an intricate computational graph. Recent studies (Zhou et al., 2022; Loo et al., 2022) have proposed approximating model training using kernel ridge regression, which provides a closed-form solution for optimal weights, thereby reducing training costs and improving performance. Despite these advancements, such methods still struggle with extensive computational demands or limitations due to the approximations in convex relaxation.

Matching-based paradigm. Another strategy involves emulating behaviors of the original dataset in the distilled one. They focus on minimizing disparities between surrogate models trained on both synthetic and original datasets. The key metrics for this are matching gradients (Zhao et al., 2020; Kim et al., 2022; Zhang et al., 2023; Liu et al., 2023b), features (Wang et al., 2022), distribution (Zhao and Bilen, 2023; Zhao et al., 2023), and training trajectories (Cazenavette et al., 2022; Cui et al., 2022; Du et al., 2023; Cui et al., 2023; Yu et al., 2023; Guo et al., 2023). Trajectory and gradient matching, in particular, has shown impressive results with low `IPC`. However, these methods often tailor the distilled dataset to specific network architectures, limiting their generalizability. Cazenavette et al. (2023) address this by proposing the GLaD that synthesizes more realistic images to enhance generalization. Nonetheless, computational and memory challenges remain, particularly when scaling to large, high-resolution datasets like ImageNet-1K (Deng et al., 2009).

Information extraction-based paradigm. The SRe^{2}L framework (Yin et al., 2023), the first work efficiently scaleable to ImageNet-1K, introduces a novel decoupled bi-level learning paradigm. This involves three stages: 1) Squeeze relevant information from the original dataset into a pre-trained model, 2) Recover this information into the image space, 3) Relabel the distilled images by using the pre-trained model to further distill knowledge into the label space (c.f. Section 2.3 for definition). Its efficiency and effectiveness have garnered community attention, spurring a series of research efforts. Shao et al. (2023) note that SRe^{2}L is limited to specific backbones and layers, impacting the generalization of the distilled dataset. They advocate for using diverse backbones for more precise and effective distillation. Yin and Shen (2023) further enhance SRe^{2}L with curriculum data augmentation. Sun et al. (2023) introduce an optimization-free approach that achieves notable diversity and realism in distilled datasets. However, inheriting the information loss issue from SRe^{2}L limits the efficacy of these methods, especially in small-scale datasets.

Appendix B. Motivation and Intuitive Exploration

In this section, we explore the information extraction-based dataset distillation approaches. We highlight two critical aspects: 1) the primary challenge of significant information loss, which hurts the quality of distilled dataset, and 2) the pivotal role of RELABEL in enhancing the effectiveness of these methods. Consequently, we investigate the critical attributes necessary for retaining the efficacy of RELABEL and seek an alternative approach to effectively distill information from datasets while adhering to these attributes.

B.1 Preliminary

We begin by formally defining the task of dataset distillation and subsequently unveil the key challenges in this domain, thereby motivating our exploration.

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

$$\mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}}[\ell\left(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y)\right] \simeq \mathbb{E}_{\mathbf{x}\sim P_{\mathcal{D}}}[\ell\left(\phi_{\boldsymbol{\theta}_{\mathcal{S}}}(\mathbf{x}), y)\right] , \tag{10}$$

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

A closer look at information extraction paradigm. Dataset distillation methods recently introduce the information extraction idea (Yin et al., 2023)—as the first effective yet efficient solution—to allow the dataset distillation on diverse-scale datasets such as ImageNet-1K (Deng et al., 2009). These methods typically employ a three-stage distillation process, *indirectly* transferring information from the original to the distilled images. The first stage involves condensing information from the complete dataset into pre-trained neural network models via SQUEEZE, followed by the extraction of this information into distilled images using RECOVER (Yin et al., 2023; Liu et al., 2023a). Upon the distilled data samples, RELABEL applies a pre-trained model to further distill knowledge into the label space. However, this paradigm encounters several key challenges:

- The RECOVER stage *necessitates batch normalization* in pre-trained models (Ioffe and Szegedy, 2015) to align the statistical features between distilled and original images (Yin et al., 2023; Liu et al., 2023a).
- *Significant information loss* during both SQUEEZE and RECOVER stages leads to distilled images with minimal content, adversely affecting performance, particularly in low `IPC` settings (Sun et al., 2023).
- Distilled images often *exhibit unrealistic textures or semantics*, tailored to specific networks, thereby limiting their generalization ability (Shao et al., 2023).
- Despite outperforming other paradigms in terms of efficiency (Yin et al., 2023), the RECOVER stage is still *computationally demanding*, requiring numerous optimization iterations (Yin and Shen, 2023).

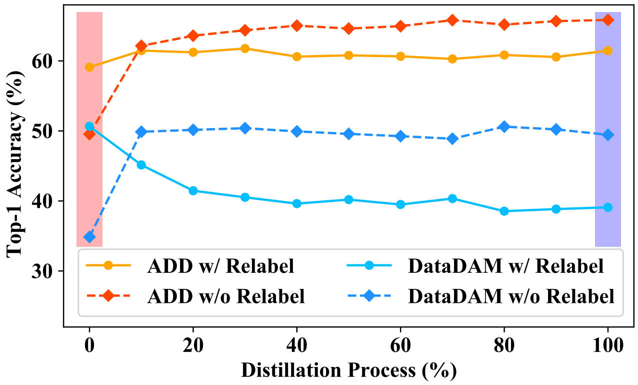


Figure 2: **Apply Relabel with the SOTA ADD (Zhang et al., 2023), DataDAM (Sajedi et al., 2023).** We evaluate the distilled images with $\texttt{IPC} = 10$ during the distillation process. The results indicate that Relabel only assists the early-stage distilled datasets from SOTA methods in achieving improved performance. We use ConvNet as the backbone architecture for CIFAR-10.

In the meanwhile, the importance of Relabel usually is under-estimated (Yin et al., 2023; Shao et al., 2023). Until the recent study in Sun et al. (2023), they suspect the application of Relabel contributes to the primary efficacy of distilled datasets.

Motivation. *To solve the aforementioned four issues simultaneously while maintaining the effectiveness of Relabel, we aim to explore a method to distill images by directly condensing the information from the original dataset, instead of Squeeze and Recover.*

B.2 Exploring Image Distillation with Relabel

To investigate a distillation approach aimed at generating high-quality images and retaining the efficacy of Relabel, this section first unravels the potentials of Relabel by empirically applying it to images distilled from various methods. We then identify the characteristics of images that influence the effectiveness of the Relabel, which motivates us to further examine a method with Relabel that satisfies these constraints in Section B.3.

Applying Relabel to alternative distillation methods. To further explore the potential of Relabel, a natural idea is to directly apply Relabel, a versatile plug-and-play component, to the distilled images from different SOTA approaches. Specifically, by using Relabel, we can substitute the labels in the originally distilled dataset with new ones, effectively creating an updated distilled dataset.

Figure 2 illustrates that, compared to the initial real images[7], namely those with zero distillation iterations, the application of Relabel to distilled and disturbed images yields only a marginal improvement or even harm. The observation demonstrates that the model used for Relabel, which was pre-trained solely on the original full dataset comprising real

[7] Conventional dataset distillation approaches typically commence by selecting a set of initial real images, denoted as $\{\mathbf{x}_j\}_{j=1}^{|\mathcal{S}|}$, from the original full dataset $\mathcal{T}$. These images are then progressively transformed into distilled images, represented as $\{\widetilde{\mathbf{x}}_j = \mathbf{x}_j + \epsilon_j\}_{j=1}^{|\mathcal{S}|}$, through successive distillation iterations. Here, ϵ_j signifies the learned shift applied to each image.

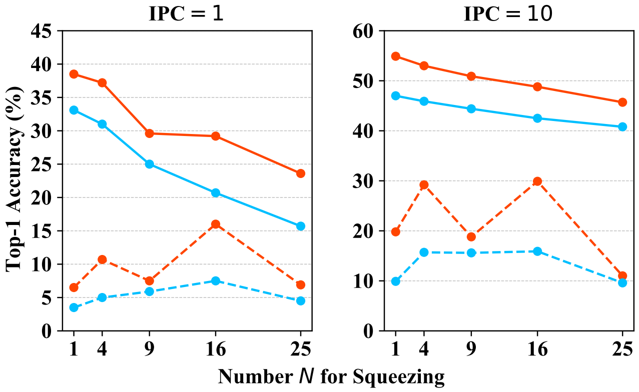


Figure 3: **Apply Relabel into a squeezing-based method.** We squeeze every N real random-sampled original images into one distilled image, thus forming a distilled dataset. We utilize ConvNet as the backbone. The red ● and blue ● respectively denote CIFAR-100 and Tiny-ImageNet. The solid line '—' and dashed line '- -' respectively represent w/ and w/o Relabel.

images, struggles to accurately relabel the distilled data. We conjecture that it might be caused by disturbed distilled images, which have contained altered semantic and textural information and will diverge from the characteristics of real images (see Appendix H for a detailed visualization).

Beyond the above empirical observation, Proposition 1 below further posits that a model well-trained on original and undisturbed samples might only provide sub-optimal labels for disturbed ones (e.g., distilled images).

Proposition 1 (Model Labeling under Perturbation) *Consider a model $\phi_{\boldsymbol{\theta}_\mathcal{T}}$ that has been effectively trained on a dataset $\mathcal{T}$, such that it accurately assigns a label $\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x})$ to each sample $(\mathbf{x}, y)$ drawn from $\mathcal{T}$. It is proposed that for any sample $(\mathbf{x}+\epsilon)$, where ϵ represents a disturbance added to $\mathbf{x}$ (e.g., random noise), the model $\phi_{\boldsymbol{\theta}_\mathcal{T}}$ will yield a label that is less precise than the label assigned to the undisturbed sample $\mathbf{x}$. Formally, this can be expressed as:*

$$\|\phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x}+\epsilon) - \phi_{\boldsymbol{\theta}_\mathcal{T}}(\mathbf{x})\| \geq 0\,, \tag{11}$$

implying that the deviation in the model's output due to the disturbance ϵ is non-negative.

B.3 Examining Squeezing Strategy with Relabel

The analysis in Section B.2 indicates that disturbed images would hurt the performance of Relabel. Therefore, we intuitively shift our attention to the squeezing-based operation (c.f. Footnote 5 for definition), a simple distillation method inspired by Kim et al. (2022) that can approximately retain the semantic and textural details of the original real images[8].

We find that *directly applying image distillation through squeezing negatively impacts the preservation of information, which consequently undermines the efficacy of Relabel* when applied to the distilled images. As shown in Figure 3, integrating Relabel with our squeezing-based method boosts performance. Yet, the benefit diminishes when the

[8] We defer the corresponding visualizations in Appendix H.

number N of images squeezed into a single distilled image grows. This reduction likely stems from the increased loss of semantic and textural information with higher N, undermining both model training and the labeling accuracy of RELABEL for these distilled images. This suggests a trade-off between the quantity of images squeezed and the preservation of essential information within each image for effective distillation.

In the following sections, we justify how to compensate for the lost information in these compressed images, a crucial aspect of our proposed LIC.

Appendix C. Framework

The whole distillation process for an entire dataset by using our proposed LIC is illustrated in Algorithm 1 and Figure 4.

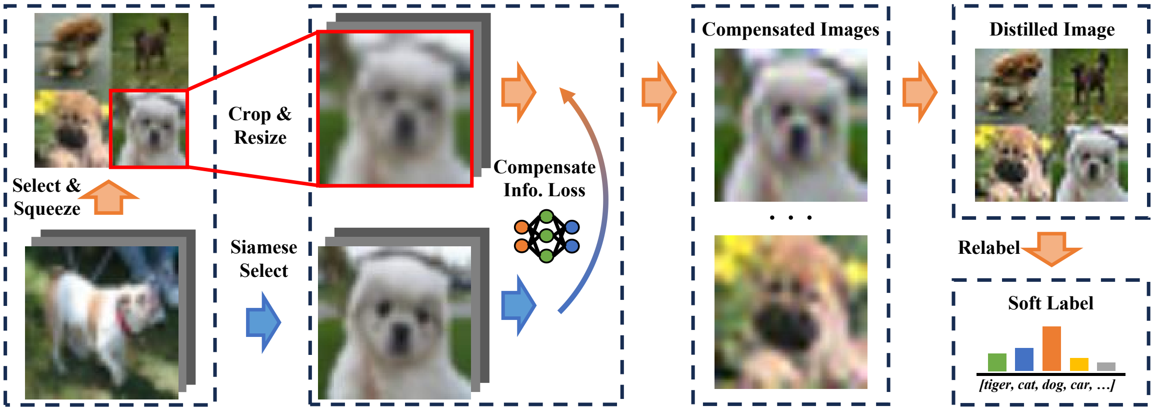


Figure 4: **Overview of our proposed three-stage dataset distillation framework.** Stage 1: select and squeeze the top $N \times$ `IPC` impactful images per class into `IPC` composite image each, where $N = 4$ and `IPC` $= 1$ in this case. Stage 2: refine these composite images to produce distilled images, compensating the information loss. Stage 3: relabel distilled images for accurate, informative annotations.

Appendix D. Experiment Details

Datasets. In addition to the datasets described in Section 3.1, we note that prevalent dataset distillation techniques struggle to scale to large, high-resolution datasets. To this end, we evaluate the baselines and our proposed LIC using two representative subsets of ImageNet-1K, specifically ImageNet-100 and ImageNet-10, as documented in (Kim et al., 2022).

Baselines. We benchmark our proposed LIC against a range of SOTA distillation techniques capable of handling large, high-resolution datasets.

- MTT (Cazenavette et al., 2022) pioneers a trajectory matching-based strategy, effective across *both low and high-resolution datasets.*

Algorithm 1 An efficient framework for dataset distillation

```
Input: Original full dataset $\mathcal{T}$, a corresponding pre-trained observer model $\phi_{\theta_{\mathcal{T}}}$ and initial $\mathcal{S} = \emptyset$.
Parameters: The number $N$ for squeezing images, the number $M$ of compression iterations, the size of $\mathcal{T}_c'$.
for $\mathcal{T}_c' \subset \mathcal{T}_c \subset \mathcal{T}$ do
  {Stage 1. Selecting Key Samples}
  for $(\mathbf{x}_i, y_i) \in \mathcal{T}_c'$ do
    Calculate $s_i = -\ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}_i), y_i)$
  end for
  Select top-($N \times \texttt{IPC}$) images $\{\mathbf{x}_i\}_{i=1}^{N \times \texttt{IPC}}$ via $s_i$
  {Stage 2. Compressing Effective Information}
  for $j = 1$ to IPC do
    Squeeze $N$ selected images as $\mathbf{x}_j^* = \mathcal{F}(\{\mathbf{x}_i\}_{i=1}^{N})$
    for $m = 1$ to $M$ do
      $\Delta\mathbf{x} \leftarrow \Delta\mathbf{x} + \nabla_{\Delta\mathbf{x}} \mathcal{L}_{\Delta\mathbf{x}}$
    end for
    {Stage 3. Relabel with Observer Model}
    Relabel $\widetilde{\mathbf{x}}_j = \mathbf{x}_j^* + \Delta\mathbf{x}^\star$ with $\widetilde{y}_j$
    $\mathcal{S} = \mathcal{S} \cup \{(\widetilde{\mathbf{x}}_j, \widetilde{y}_j)\}$
  end for
end for
Output: Small distilled dataset $\mathcal{S}$
```

- IDM (Zhao et al., 2023) presents an efficient dataset condensation technique utilizing distribution matching, offering a scalable alternative to computationally demanding optimization-focused methods (Zhao et al., 2020; Cazenavette et al., 2022), notably *adapting to ImageNet-100*.
- TESLA (Cui et al., 2023) marks the first distillation approach that *extends to the full ImageNet-1K*, circumventing the extensive memory demands associated with MTT-derived methods through a constant memory footprint.
- ADD (Zhang et al., 2023) demonstrates *effective scalability across varying dataset resolutions*, enhancing distillation speed through model augmentation.
- DataDAM (Sajedi et al., 2023) *efficiently distills images across multiple resolutions and scales* by matching spatial attention maps between real and distilled samples at various layers within families of randomly initialized neural networks.
- DATM (Guo et al., 2023) stands out by occasionally *surpassing the training performance of the full original dataset*, for instance, achieving $\texttt{IPC} = 100$ on CIFAR-100.
- DREAM (Liu et al., 2023b) introduces an *efficient technique while also delivering the most remarkable results.*
- SRe^{2}L (Yin et al., 2023) is a novel entrant that *efficiently handles ImageNet-1K*, significantly outpacing other methods in managing large, high-resolution datasets and serving as our primary comparison point.

Evaluating main results. For both dataset distillation and performance evaluation, we employ identical neural network architectures. Consistent with previous studies (Cazenavette et al., 2022; Cui et al., 2023; Zhao et al., 2023), we use Conv-3 for CIFAR-10 and CIFAR-100 distillation tasks, Conv-4 for Tiny-ImageNet (with the exception of DREAM, which utilizes Conv-3) and ImageNet-1K, Conv-5 for ImageNet-10, and Conv-6 for ImageNet-100 distillation. In line with Cazenavette et al. (2022); Cui et al. (2023), MTT and TESLA apply a reduced resolution for distilling 224×224 images. According to Yin et al. (2023), for retrieving and evaluating distilled datasets, SRe^{2}L and LIC adopt ResNet-18.

Evaluating the distilled dataset. We detail the hyperparameter configurations for our distilled dataset evaluation in Table 4. Consistent with recent works (Yin and Shen, 2023; Yin et al., 2023; Shao et al., 2023), the evaluation on the ImageNet-1K dataset follows the parameters outlined in Table 4a. For other datasets, their assessments are guided by the parameters specified in Table 4b. Furthermore, we implement Differentiable Siamese Augmentation (DSA) as described by Zhao and Bilen (2021) to enhance images during both the distillation and evaluation phases of our experiments.

config	value
epochs	300
optimizer	AdamW
learning rate	0.001
weight decay	1e-4
scheduler	MultiStepLR

(a) ImageNet-1K evaluation setting.

config	value
epochs	1000
optimizer	AdamW
learning rate	0.001
weight decay	1e-4
scheduler	MultiStepLR

(b) Default evaluation setting.

Table 4: Hyperparameter setting.

Appendix E. Experiment Results

Comparison with more datasets and baselines. In addition to the experiments discussed in Section 3.2, we further benchmark our proposed LIC against a broader set of baselines, encompassing recent contributions (Yu et al., 2023; Shao et al., 2023; Sun et al., 2023; Liu et al., 2023a). Our analysis also spans additional subsets of the ImageNet-1K dataset, namely ImageNet-100 and ImageNet-10 (Kim et al., 2022). The outcomes, presented in Table 5, consistently affirm the superior performance of LIC in dataset distillation tasks.

E.1 Efficiency Comparison

Beyond assessing performance in Section 3.2, we expand our evaluation to include additional baselines. Results presented in Table 6 underscore the exceptional efficiency of our proposed LIC, which also requires the least GPU memory.

Architecture		ConvNet					ResNet-18				
Dataset	IPC	DataDAM	ADD	IDM	RDED	LIC (Ours)	G-VBSM	CDA	WMDD	RDED	LIC (Ours)
CIFAR-10	1	32.0 ± 1.2	49.2	45.6 ± 0.7	23.5 ± 0.3	**50.0 ± 0.3**	-	-	-	22.9 ± 0.4	**40.7 ± 0.4**
	10	54.2 ± 0.8	67.1	58.6 ± 0.1	50.2 ± 0.3	**72.2 ± 0.2**	53.5 ± 0.6	-	-	37.1 ± 0.3	**80.9 ± 0.4**
	50	67.0 ± 0.4	73.8	67.5 ± 0.1	68.4 ± 0.1	**78.3 ± 0.1**	59.2 ± 0.4	-	-	62.1 ± 0.1	**89.2 ± 0.0**
CIFAR-100	1	14.5 ± 0.5	29.8	20.1 ± 0.3	19.6 ± 0.3	**42.7 ± 0.3**	25.9 ± 0.5	-	-	11.0 ± 0.3	**48.1 ± 0.7**
	10	34.8 ± 0.5	45.6	45.1 ± 0.1	48.1 ± 0.3	**54.8 ± 0.2**	59.5 ± 0.4	-	-	42.6 ± 0.2	**64.7 ± 0.1**
	50	49.4 ± 0.3	52.6	50.0 ± 0.2	**57.0 ± 0.1**	56.6 ± 0.1	65.0 ± 0.5	-	-	62.6 ± 0.1	**67.6 ± 0.1**
Tiny ImageNet	1	8.3 ± 0.4	-	10.1 ± 0.2	12.0 ± 0.1	**31.7 ± 0.4**	-	-	7.6 ± 0.2	9.7 ± 0.4	**36.5 ± 0.2**
	10	18.7 ± 0.3	-	21.9 ± 0.2	39.6 ± 0.1	**46.3 ± 0.2**	-	-	41.8 ± 0.1	41.9 ± 0.2	**51.7 ± 1.0**
	50	28.7 ± 0.3	-	27.7 ± 0.3	**47.6 ± 0.2**	47.4 ± 0.1	-	48.7	**59.4 ± 0.5**	58.2 ± 0.1	55.8 ± 0.9
ImageNet-10	1	-		-	-	**39.7 ± 0.4**	-	-	-	24.9 ± 1.1	**43.9 ± 0.7**
	10	-	**74.6**	-	-	71.5 ± 0.8	-	-	-	53.3 ± 0.1	**76.3 ± 0.8**
	50	-	-	-	-	**81.1 ± 0.1**	-	-	-	75.5 ± 0.5	**85.8 ± 0.8**
ImageNet-100	1	-	-	11.2 ± 0.5	7.1 ± 0.2	**17.8 ± 0.3**	-	-	-	8.1 ± 0.3	**20.8 ± 0.3**
	10	-	48.4	17.1 ± 0.6	29.6 ± 0.1	**50.2 ± 0.4**	-	-	-	36.0 ± 0.3	**65.0 ± 0.1**
	50	-	-	26.3 ± 0.4	50.2 ± 0.2	**64.3 ± 0.5**	-	-	-	61.6 ± 0.1	**79.8 ± 0.2**
ImageNet-1k	1	2.0 ± 0.1	-	-	-	**14.5 ± 0.3**	-	-	3.2 ± 0.3	6.6 ± 0.2	**6.8 ± 0.2**
	10	6.3 ± 0.0	-	-	-	**24.0 ± 0.5**	31.4 ± 0.5	-	38.2 ± 0.2	42.0 ± 0.1	**48.5 ± 0.2**
	50	15.5 ± 0.2	-	-	-	**37.3 ± 0.1**	51.8 ± 0.4	53.5	57.6 ± 0.5	56.5 ± 0.1	**60.0 ± 0.2**

Table 5: Comparison with SOTA dataset distillation methods.

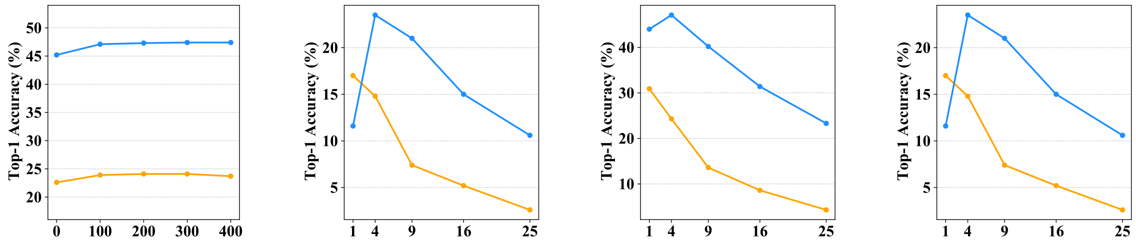
Architecture		Time Cost (s)	Peak Memory (GB)
Conv-4	DREAM	33906.17	15.92
	DATM	12470.90	20.16
	SRe2L	51.68	1.36
	Ours	**13.02**	**0.65**
ResNet-18	SRe2L	191.14	3.62
	Ours	**25.34**	**1.56**
MobileNet-V2	SRe2L	114.05	1.27
	Ours	**18.81**	**0.64**

Table 6: Efficiency comparison with SOTA methods on varying networks on Tiny-ImageNet.

Appendix F. Ablation

Ablation study on ImageNet-1K. The results depicted in Figure 5 demonstrate that, despite exhibiting different behaviors when applied to Conv-4 and ResNet-18 architectures, the parameters of our LIC maintain their robustness against variations. Consequently, the configuration outlined in Section 3.1 is applicable across all datasets.

Effectiveness of each technique. To assess the efficacy of the individual components within LIC, we decompose it into three core techniques. The term "Selection" denotes the synthesis of samples based on their importance scores, "Compensation" pertains to the refinement of data through image alignment, and "Relabel" involves the use of labels furnished by a pre-trained model.



(a) iteration number M (b) number N ($\texttt{IPC} = 1$) (c) number N (d) alignment layer

Figure 5: **Ablation study on each component in our LIC.** We evaluate the distilled dataset of our LIC with different number M of compression iterations (5a), number N of images squeezed in one distilled image (5b & 5c), feature alignment layer (5d). The yellow ●, and blue ● denote ConvNet, and ResNet respectively.

Dataset	Original	+Selection	+Compensation	+Relabel
CIFAR-10	47.0 ± 0.1	52.5 ± 1.1	58.7 ± 0.4	72.2 ± 0.2
CIFAR-100	27.8 ± 0.6	31.8 ± 0.4	40.8 ± 0.5	54.8 ± 0.2
Tiny-ImageNet	7.2 ± 0.5	11.7 ± 0.6	20.4 ± 0.8	46.3 ± 0.2
ImageNet-1K	7.7 ± 0.1	12.3 ± 0.1	10.5 ± 0.1	24.0 ± 0.5

Table 7: **Effectiveness of accumulated techniques in LIC**. The validation accuracy experiences a progressive improvement as we incrementally apply the four techniques within our LIC.

Appendix G. Continual Learning

In comparison to SRe^{2}L, our study implements a five-step class-incremental learning approach using ResNet-18 across CIFAR-10, CIFAR-100, and Tiny-ImageNet datasets, each with an IPC setting of 10. The results of these experiments are depicted in Figures 6, 7, and 8 for CIFAR-10, CIFAR-100, and Tiny-ImageNet, respectively.

Appendix H. Visualization

Baselines. Within the scope of CIFAR-10 distillation under the $\texttt{IPC} = 10$ setting, we illustrate the visual representations of distilled datasets. This includes visualizations for ADD (Zhang et al., 2023) in Figure 11, DataDAM (Sajedi et al., 2023) in Figure 12, SRe^{2}L (Yin et al., 2023) in Figure 14, and DREAM (Liu et al., 2023b) in Figure 13. Distilled images of each method are generated starting from actual images, showcased in Figures 9 and 10.

A simple squeezing-based method. The image squeezing process entails resizing and concatenating images to facilitate dataset distillation. For example, consider the manipulation of 4 images, each originally sized at 224 × 224 pixels. The initial step involves downsizing each image to 112 × 112 pixels. Subsequently, these reduced images are merged into a single composite image, effectively reverting to the original resolution of 224 × 224 pixels. This approach underpins a simplistic, squeezing-based dataset distillation method, whereby N randomly selected original images are compressed into one distilled image to compose a

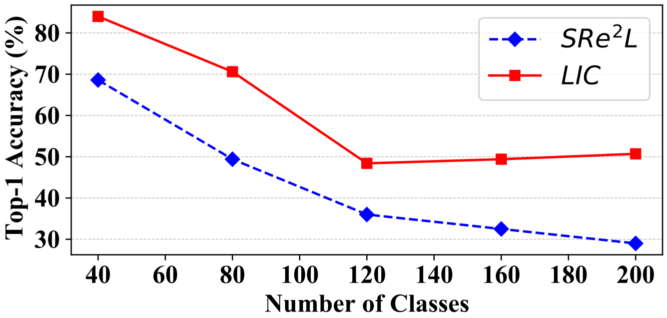


Figure 6: Visualization of continual learning on CIFAR-10 with IPC $= 10$.

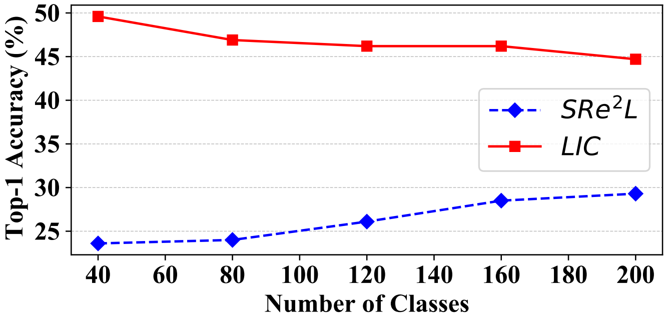


Figure 7: Visualization of continual learning on CIFAR-100 with IPC $= 10$.

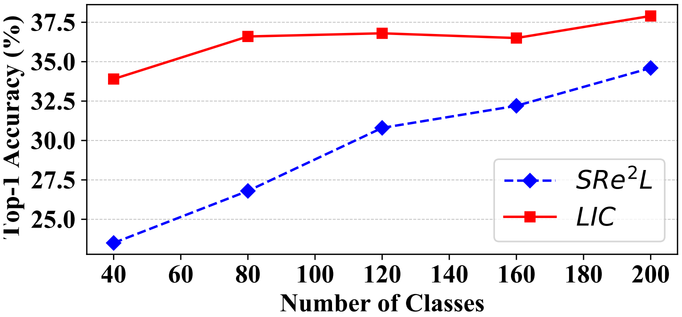


Figure 8: Visualization of continual learning on Tiny-ImageNet with IPC $= 10$.

condensed dataset. In the context of distilling CIFAR-10 with an IPC $= 10$ configuration, we exhibit the visual outcomes of this process for diverse settings: $N = 1$ in Figure 15, $N = 4$ in Figure 16, $N = 9$ in Figure 17, $N = 16$ in Figure 18, and $N = 25$ in Figure 19.

Our proposed LIC. With an IPC setting of 10, we illustrate the distilled datasets generated by our proposed LIC. These include visualizations for CIFAR-10 in Figure 20, CIFAR-100 in Figure 21, Tiny-ImageNet in Figure 22, and ImageNet-1k in Figure 23.

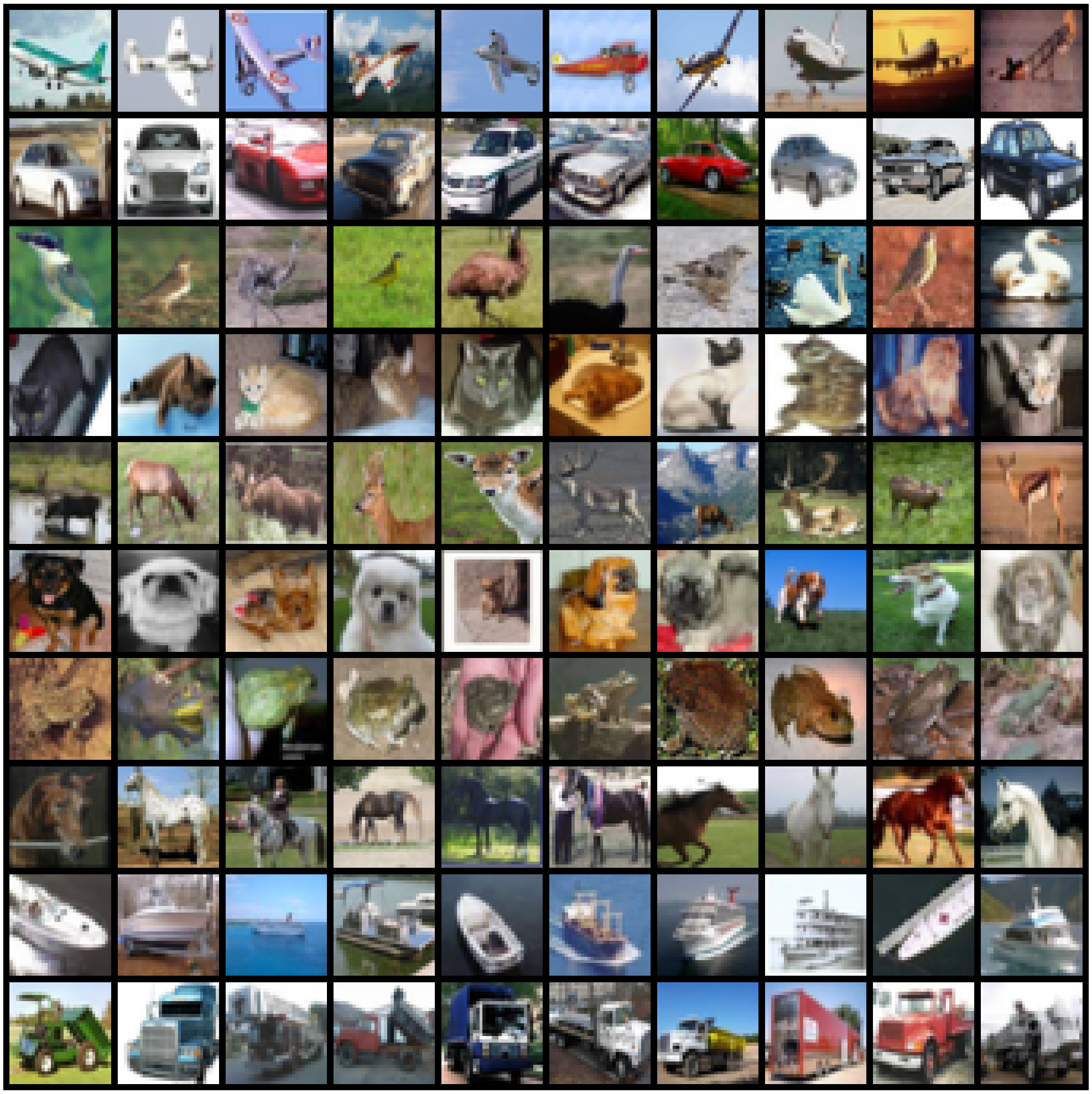


Figure 9: Visualization of initialized images before distilling on CIFAR-10.

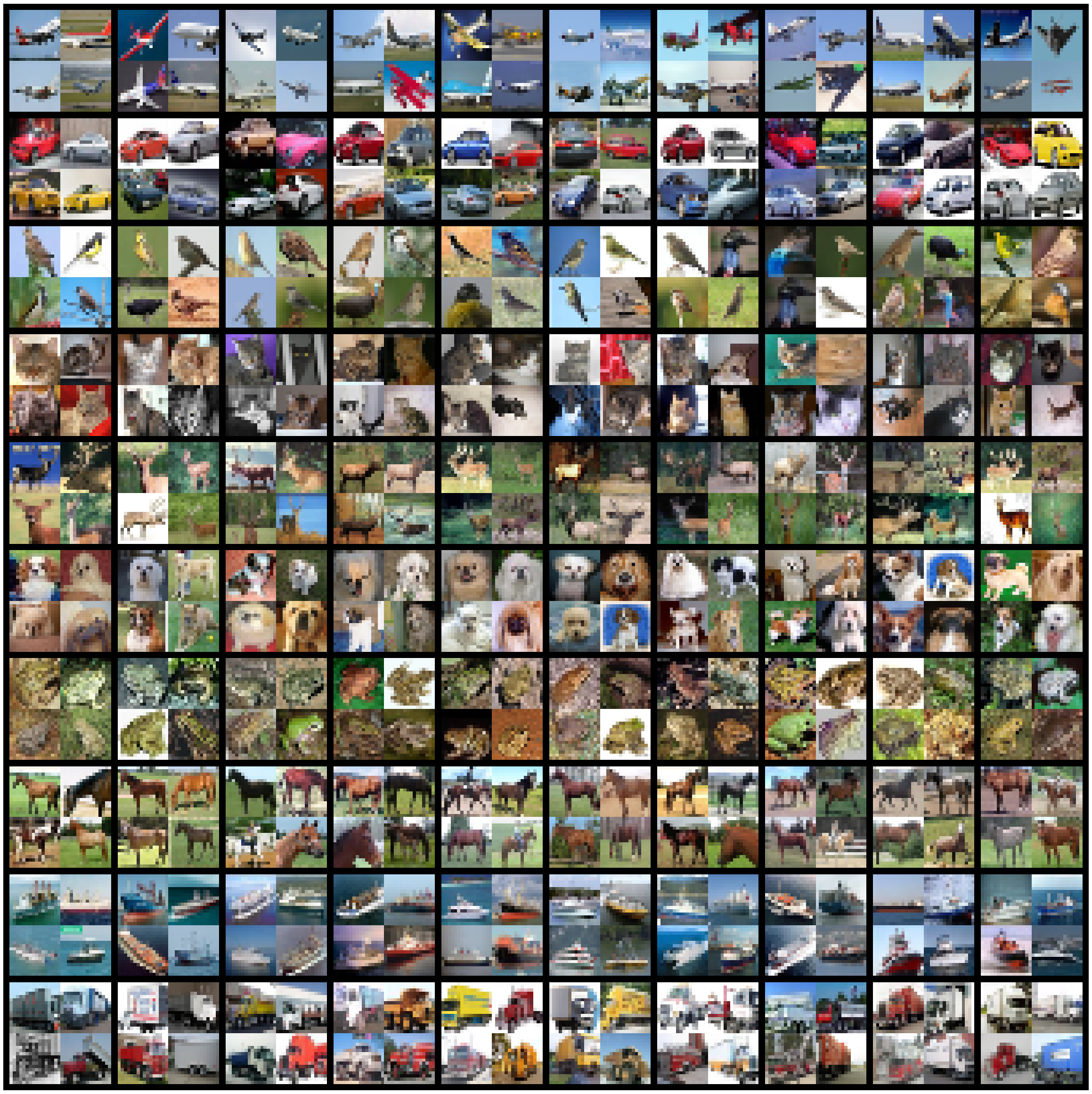


Figure 10: Visualization of initialized data before distilling on CIFAR-10 data showcases the mixture of 4 images per initial instance.

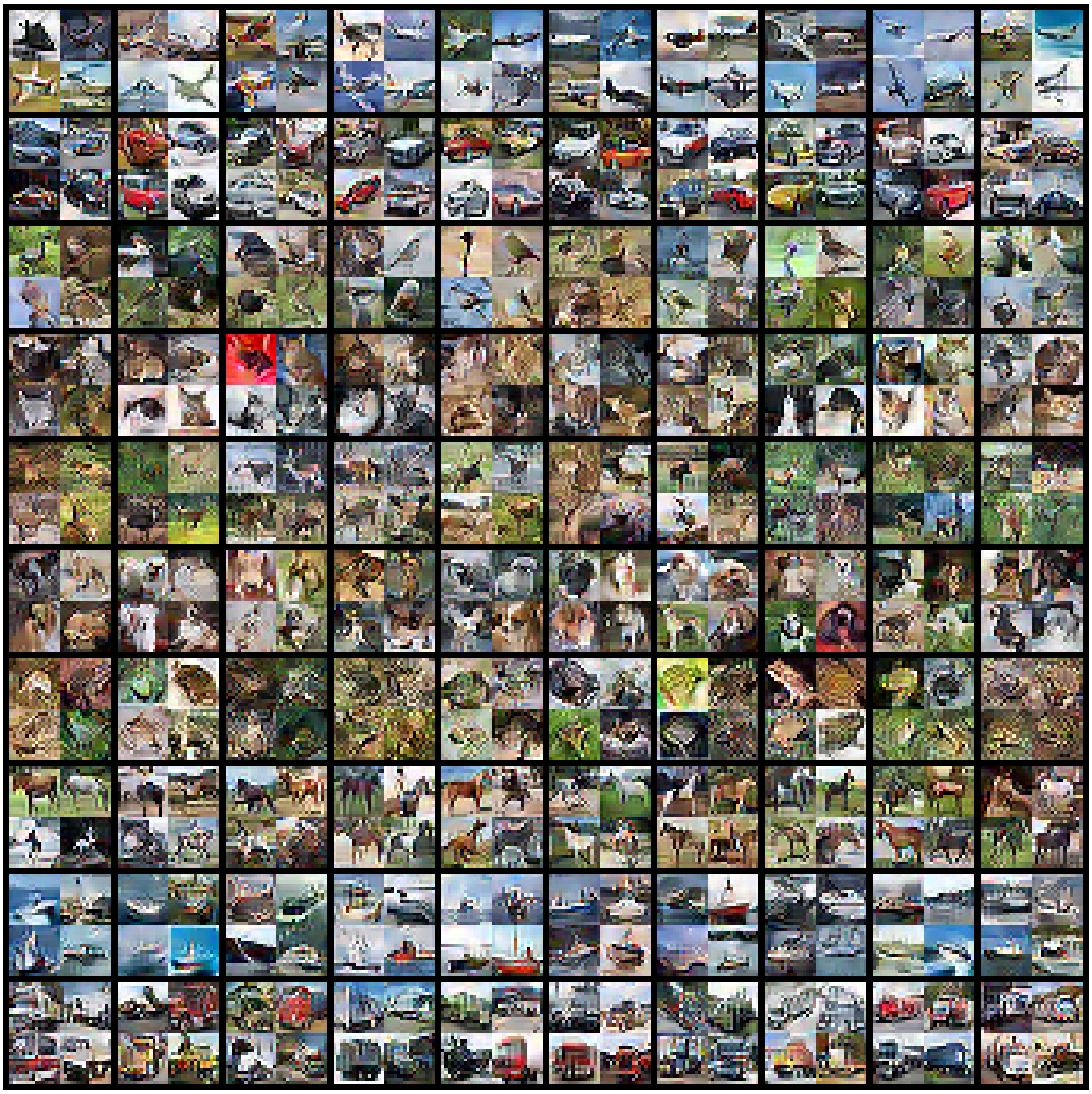


Figure 11: Synthetic data visualization on CIFAR-10 from ADD (Zhang et al., 2023)

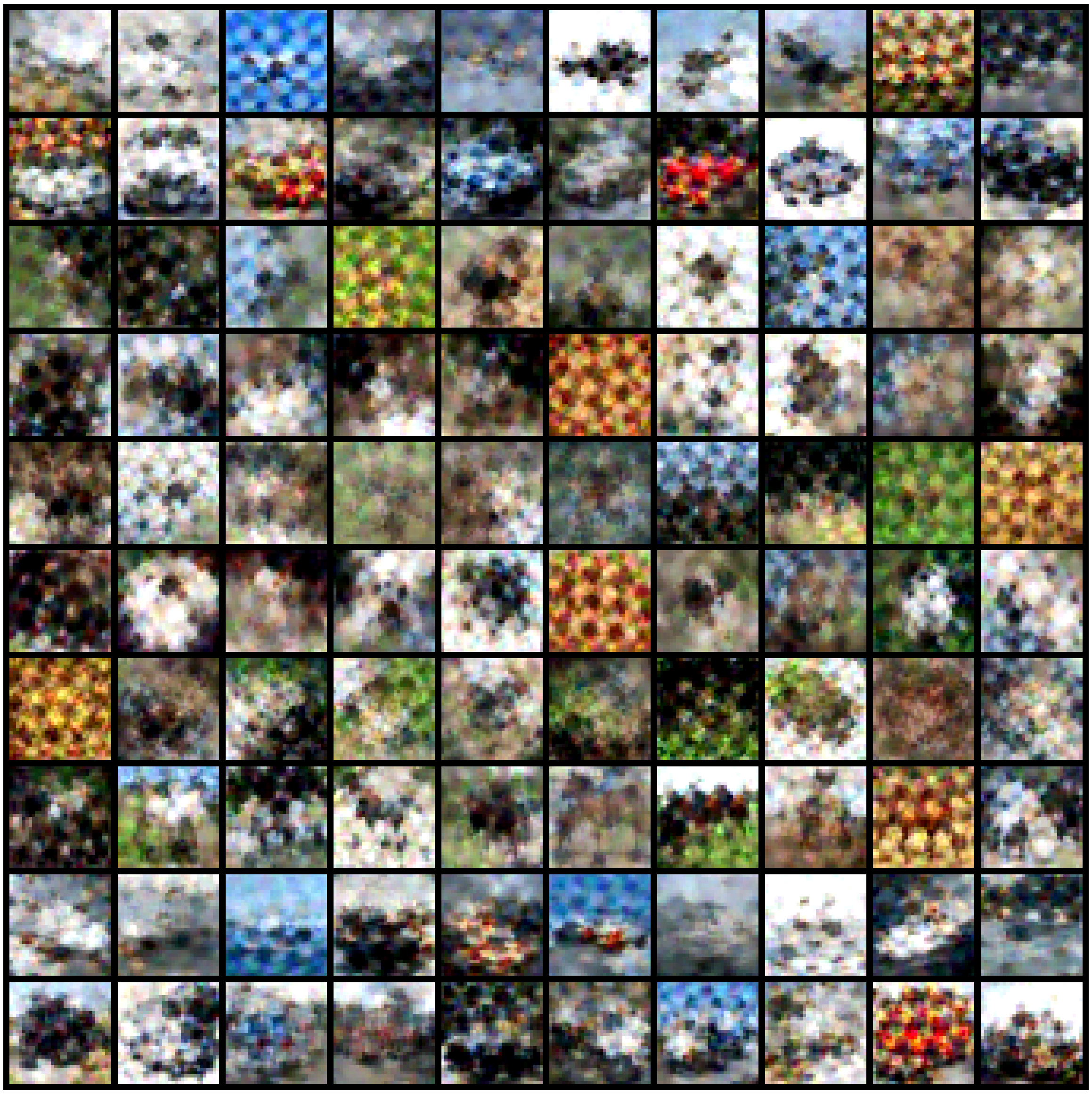


Figure 12: Synthetic data visualization on CIFAR-10 from DataDAM (Sajedi et al., 2023)

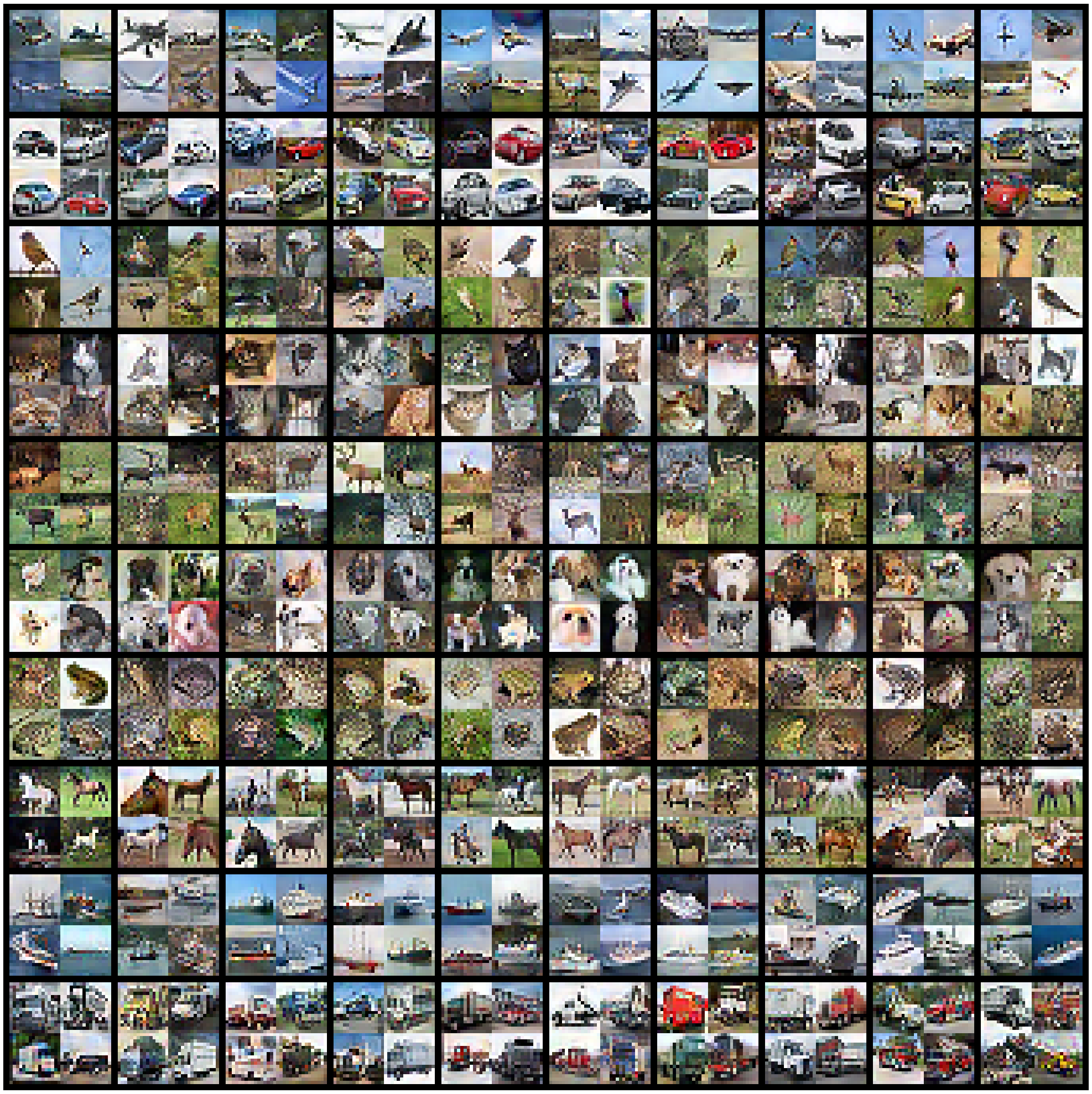


Figure 13: Synthetic data visualization on CIFAR-10 from DREAM (Liu et al., 2023b)

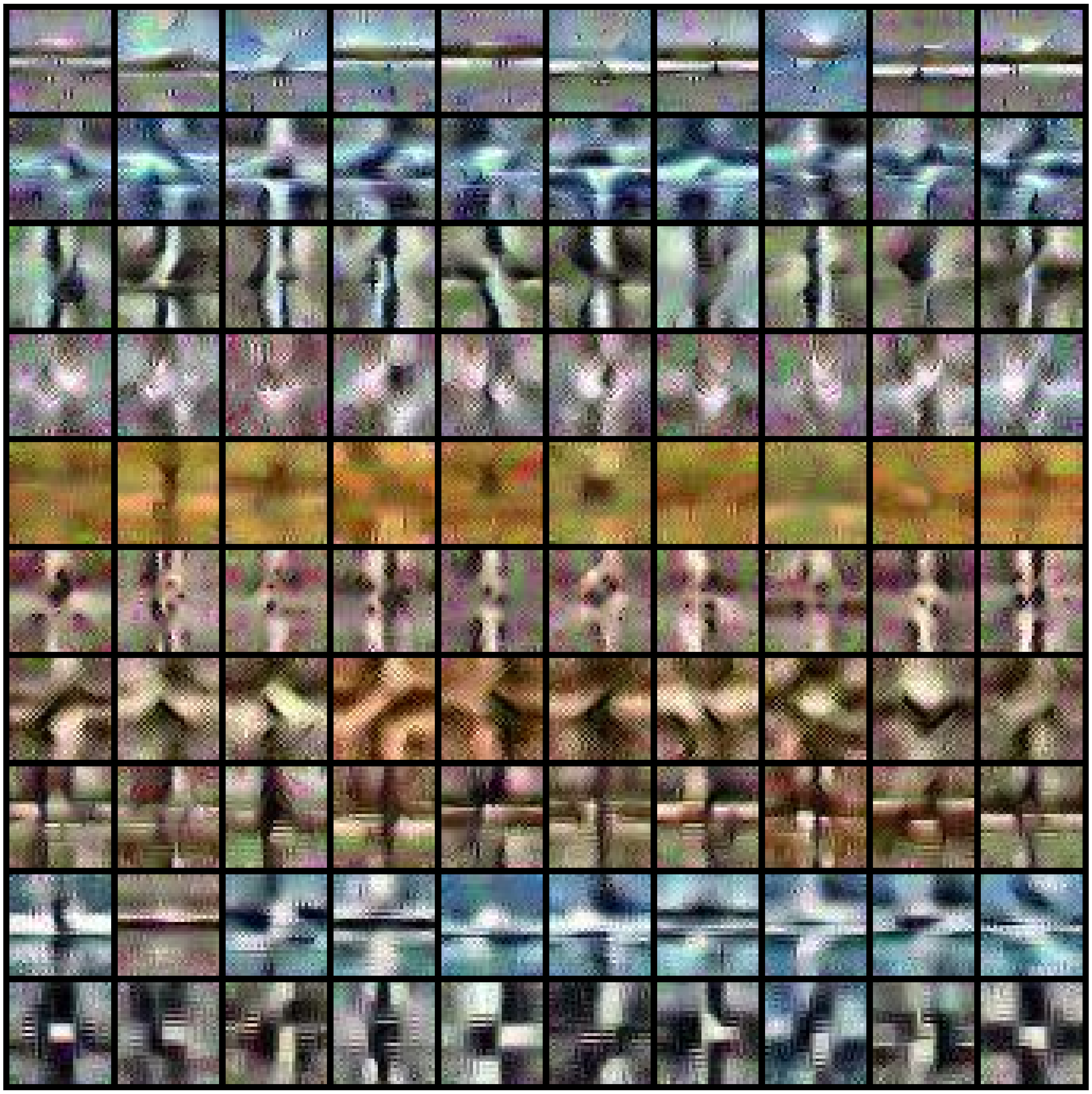


Figure 14: Synthetic data visualization on CIFAR-10 from SRe^{2}L (Yin et al., 2023)

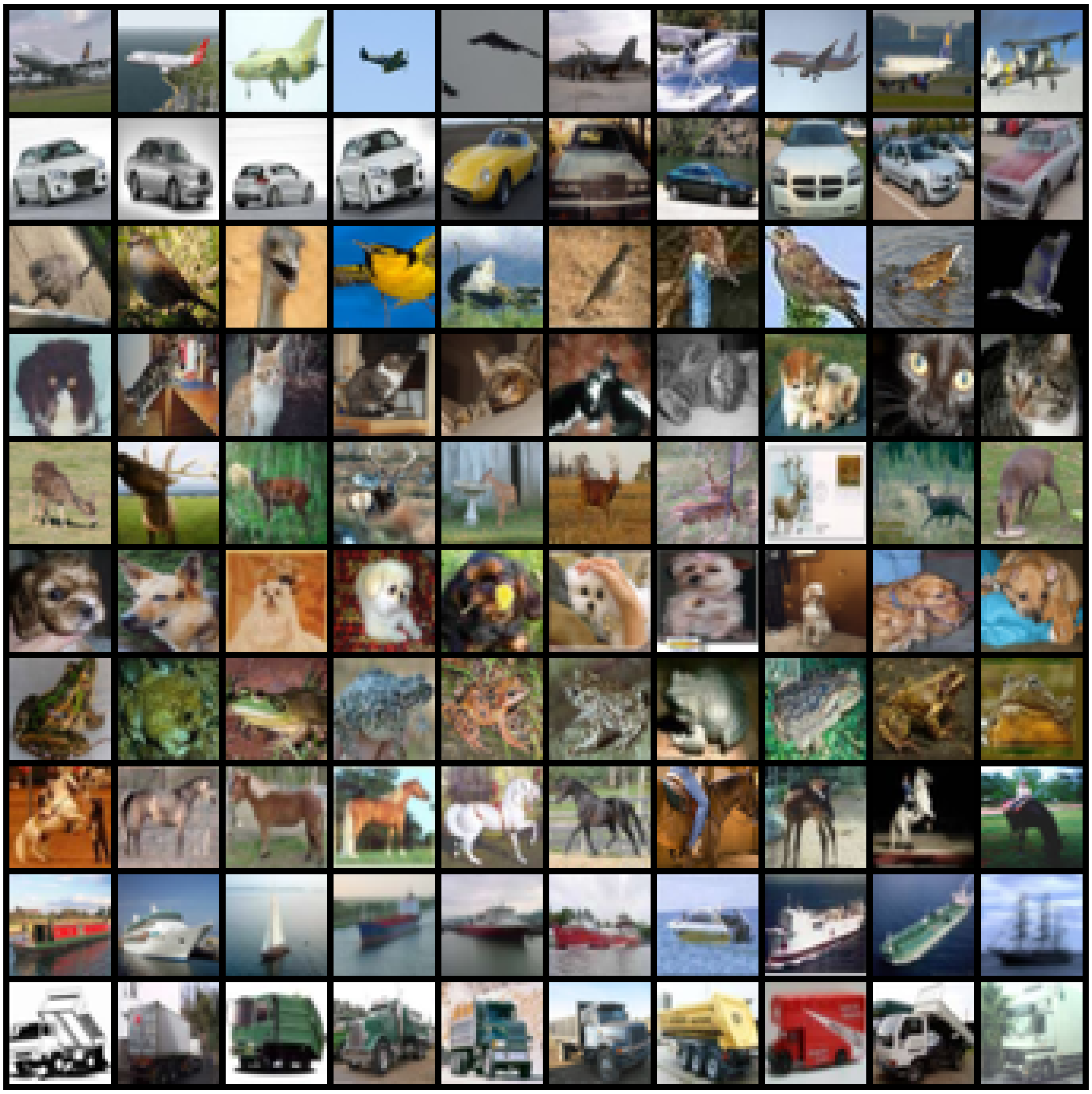


Figure 15: Squeezed real data visualization on CIFAR-10 with $N = 1$

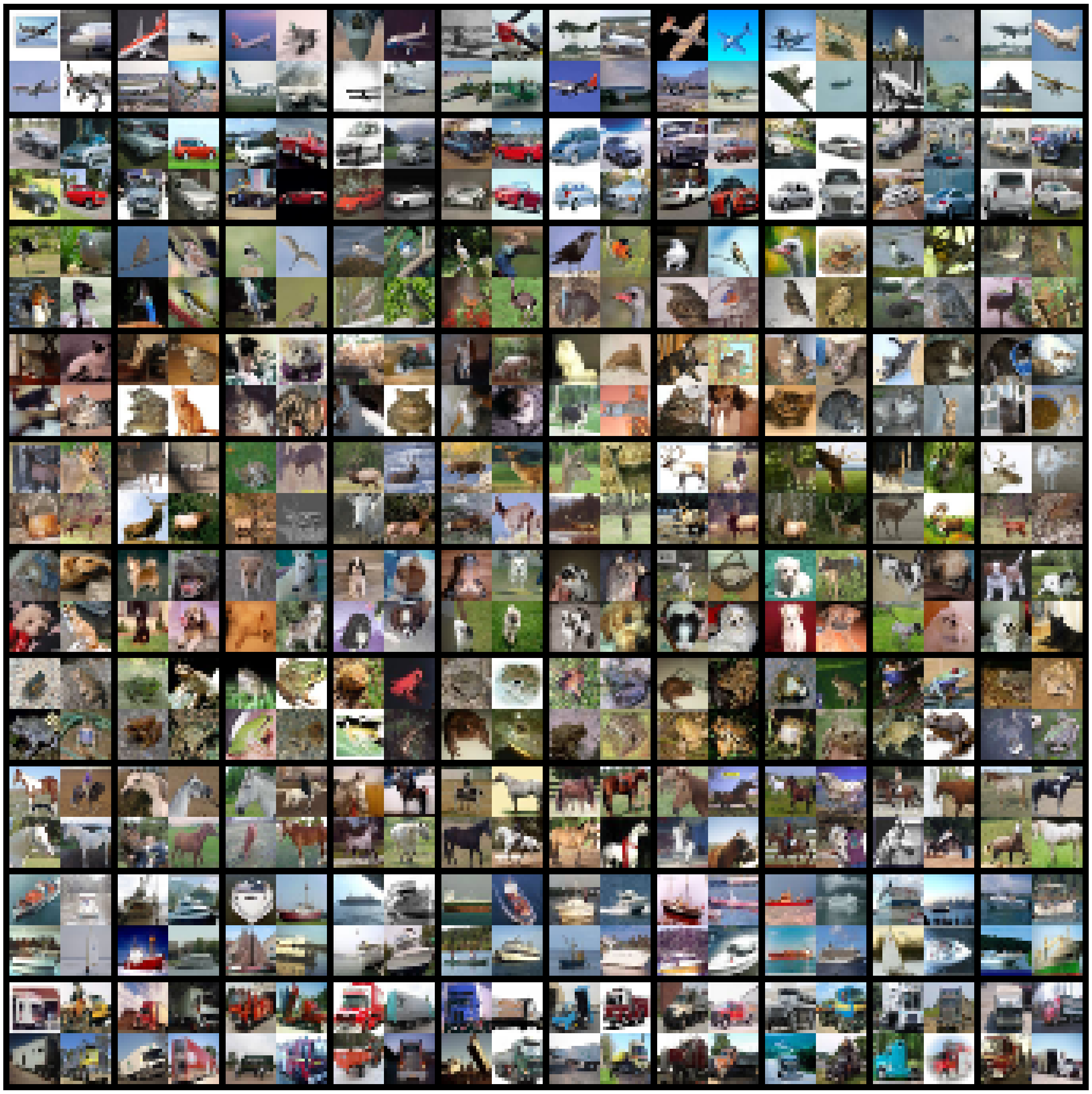


Figure 16: Squeezed real data visualization on CIFAR-10 with $N = 4$

Figure 17: Squeezed real data visualization on CIFAR-10 with $N = 9$

Figure 18: Squeezed real data visualization on CIFAR-10 with $N = 16$

Figure 19: Squeezed real data visualization on CIFAR-10 with $N = 25$

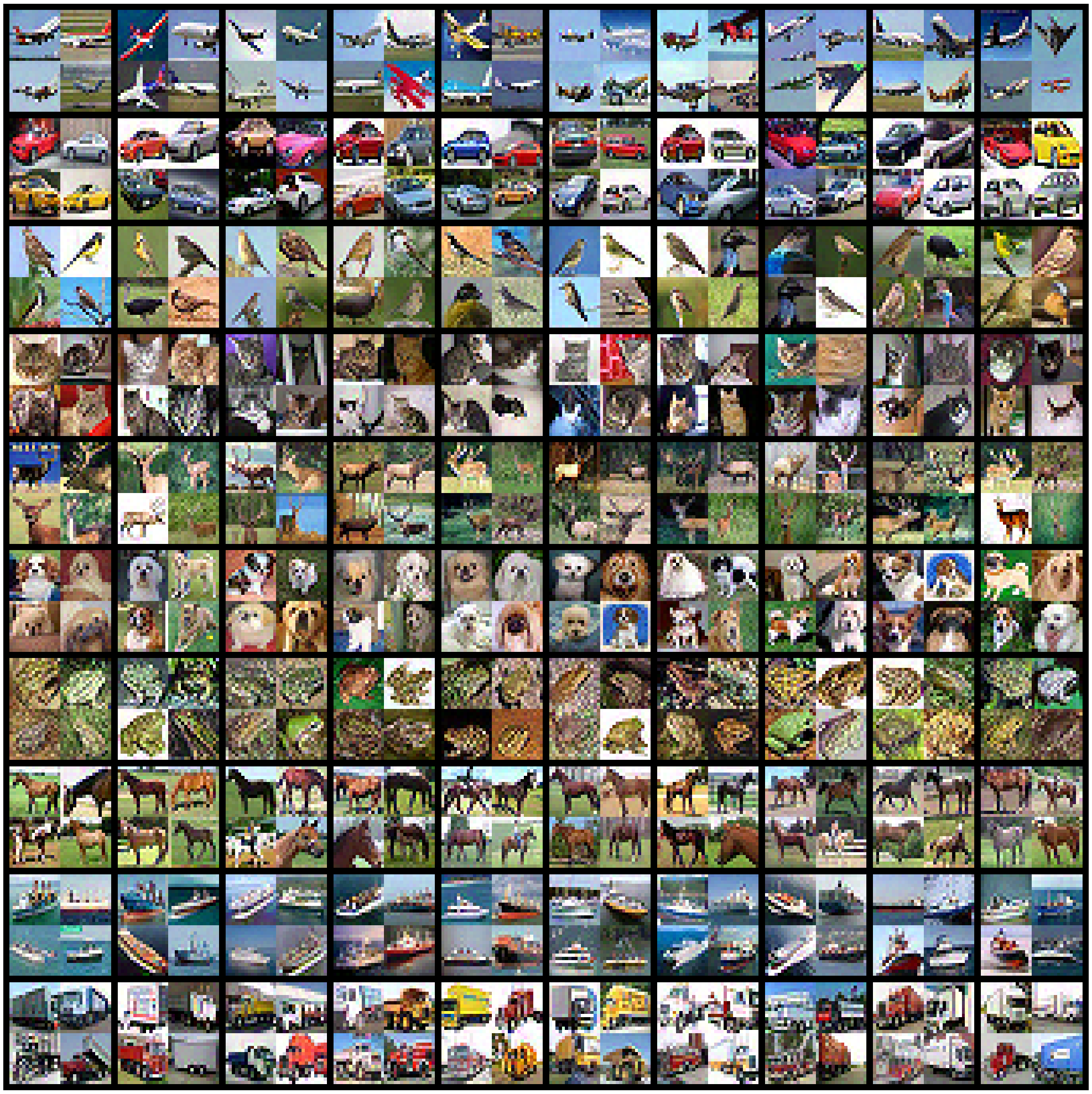


Figure 20: Synthetic data visualization on CIFAR-10 from **LIC(Ours)**

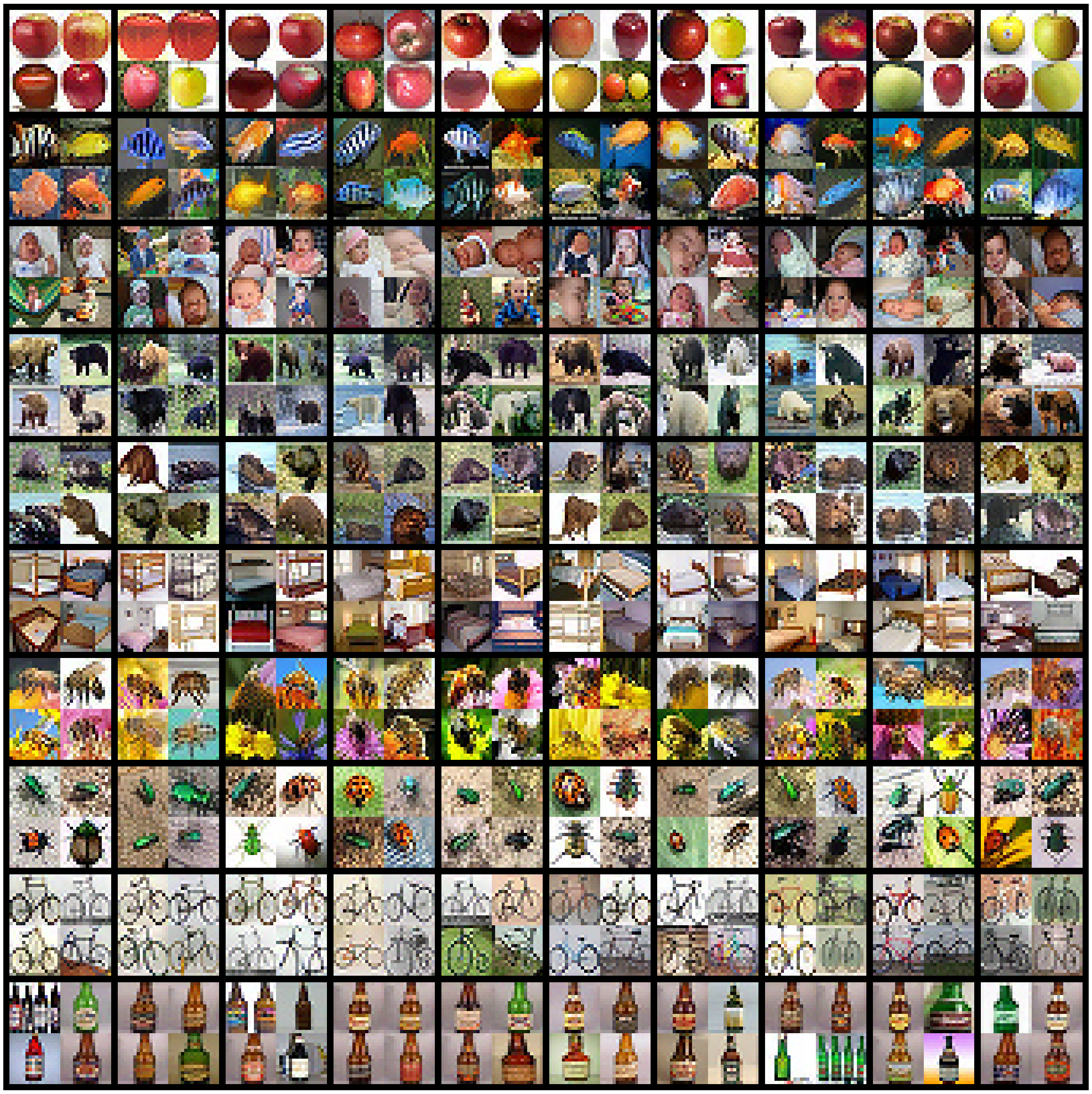


Figure 21: Synthetic data visualization on CIFAR-100 from **LIC(Ours)**

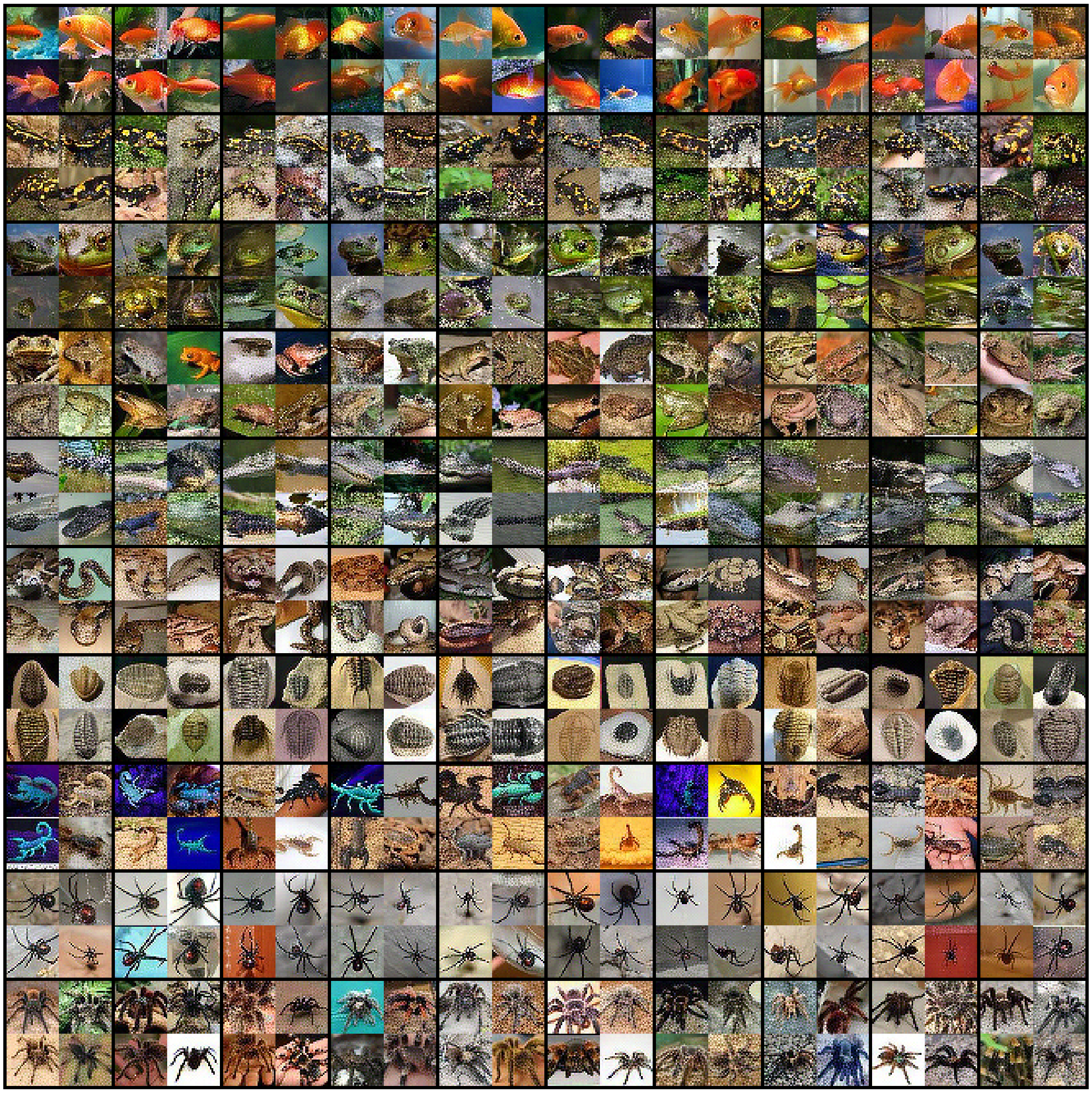


Figure 22: Synthetic data visualization on Tiny-ImageNet from **LIC(Ours)**

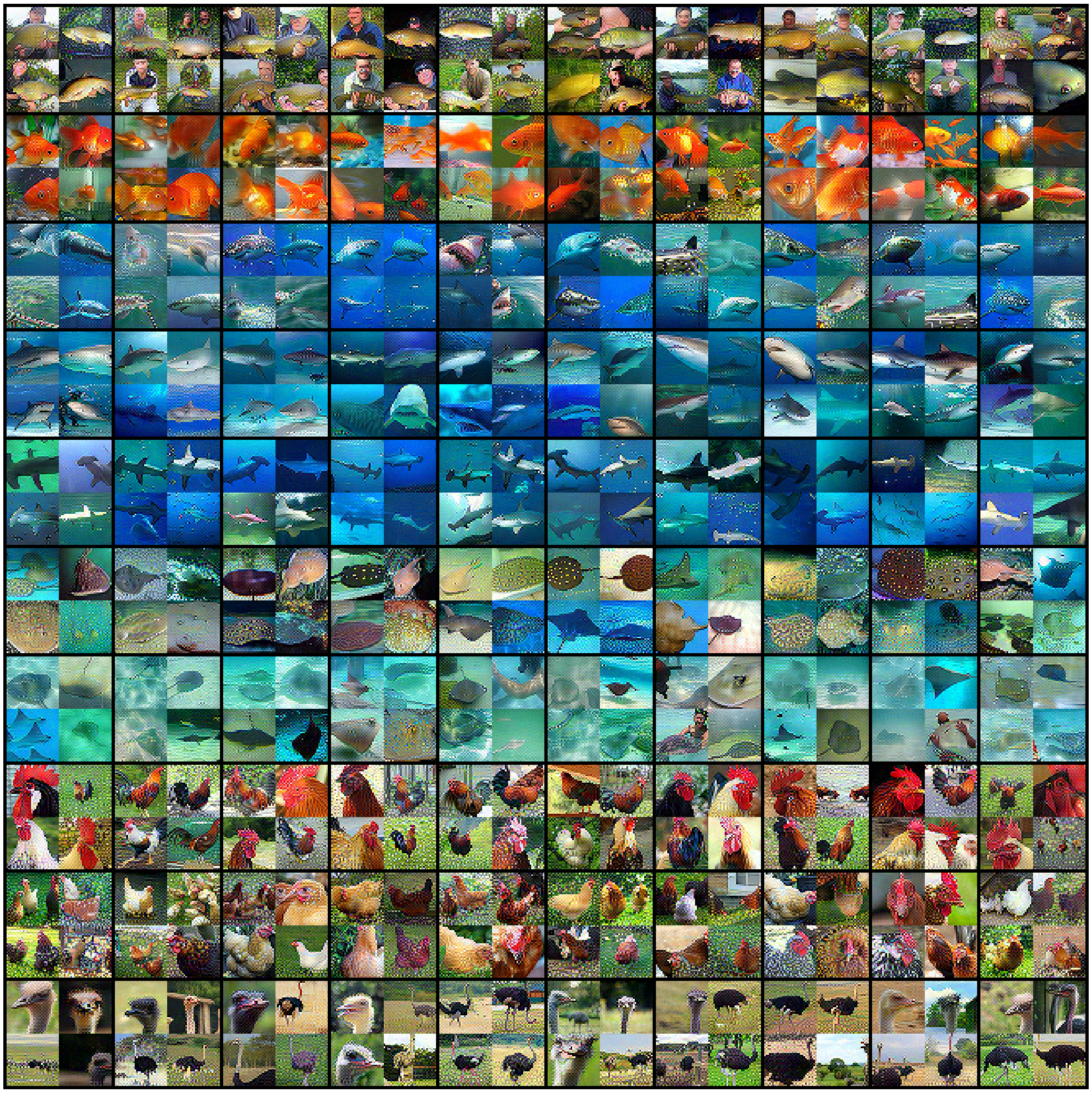


Figure 23: Synthetic data visualization on ImageNet-1k from **LIC(Ours)**