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

Supplementary Material

A. Distilled Images Comparison

Several additional examples of distilled images are presented in Figure 6. Besides, we conduct a meticulous comparison between our proposed RDED and the closest approach, SRe^{2}L. The distilled images generated by SRe^{2}L are scrutinized in Figure 7, revealing two noteworthy observations:

- SRe^{2}L exhibits a limitation in generating diverse features within each distilled image.
- The diversity and realism of distilled images within each class are notably lacking.

In contrast, our proposed method, RDED, demonstrates a superior capability to achieve high diversity in both the features within individual images and across images within each class, all while maintaining a high level of realism.

B. $\mathcal{V}$-information Theory

B.1. Definitions

The following definitions, as outlined by Xu et al. [42], establish the groundwork for our discussion:

Definition 2 (Predictive Family)**.** *Let* $\Omega = \{f : \mathcal{X} \cup \{\varnothing\} \to \mathcal{P}(\mathcal{Y})\}$*.* $\mathcal{V} \subseteq \Omega$ *is a predictive family if it satisfies*

$$\forall f \in \mathcal{V}, \forall P \in \mathrm{range}(f), \quad \exists f' \in \mathcal{V}, \tag{11}$$

$$s.t. \quad \forall x \in X, f'[x] = P, f'[\varnothing] = P.$$

A predictive family denotes a collection of permissible predictive models (observers) available to an agent, often constrained by computational or statistical limitations. Xu et al. [42] term the supplementary criterion in (2) as *optional ignorance*. In essence, this implies that within the framework of the subsequent prediction game we delineate, the agent possesses the discretion to disregard the provided side information at thier discretion.

Definition 3. *Consider random variables X and Y with corresponding sample spaces $\mathcal{X}$ and $\mathcal{Y}$. Let $\varnothing$ denote a null input that imparts no information about Y. Within the context of a predictive family $\mathcal{V} \subseteq \Omega = \{f : \mathcal{X} \cup \varnothing \to \mathcal{P}(\mathcal{Y})\}$, the* ***predictive $\mathcal{V}$-entropy*** *is defined as:*

$$H_{\mathcal{V}}(Y|\varnothing) = \inf_{f \in \mathcal{V}} \mathbb{E}_{y \sim Y}\left[-\log f[\varnothing](y)\right] . \tag{12}$$

Similarly, the ***conditional $\mathcal{V}$-entropy*** *is expressed as:*

$$H_{\mathcal{V}}(Y|X) = \inf_{f \in \mathcal{V}} \mathbb{E}[-\log f[\mathbf{x}](y)] . \tag{13}$$

Here, log *quantifies the entropies in nats.*

In essence, $f[\mathbf{x}]$ and $f[\varnothing]$ generate probability distributions over the labels. The objective is to identify $f \in \mathcal{V}$ that maximizes the log-likelihood of the label data, both with (13) and without the input (12).

Definition 4. *Consider random variables X and Y with respective sample spaces $\mathcal{X}$ and $\mathcal{Y}$. Within the context of a predictive family $\mathcal{V}$, the* ***$\mathcal{V}$-information*** *is defined as:*

$$I_{\mathcal{V}}(X \to Y) = H_{\mathcal{V}}(Y|\varnothing) - H_{\mathcal{V}}(Y|X) . \tag{14}$$

Given the finite nature of the dataset, the estimated $\mathcal{V}$-information may deviate from its true value. Xu et al. [42] establish PAC bounds for this estimation error, with less complex $\mathcal{V}$ and larger datasets yielding more precise bounds. Besides, several key properties of $\mathcal{V}$-information, enumerated by Xu et al. [42], include:

- *Non-Negativity:* $I_{\mathcal{V}}(X \to Y) \geq 0$
- *Independence:* If X is independent of Y, $I_{\mathcal{V}}(X \to Y) = I_{\mathcal{V}}(Y \to X) = 0$.
- *Monotonicity:* For $\mathcal{V} \subseteq \mathcal{U}$, $H_{\mathcal{V}}(Y|\varnothing) \geq H_{\mathcal{U}}(Y|\varnothing)$ and $H_{\mathcal{V}}(Y|X) \geq H_{\mathcal{U}}(Y|X)$.

B.2. Intuition of $\mathcal{V}$-information on Distilled Dataset

Maximizing the $\mathcal{V}$-information $I_{\mathcal{V}}(X \to Y)$ for real-world datasets proves intractable, primarily attributed to the inherent disparity between the boundless information sources and the constrained capabilities of observers within the predictive family $\mathcal{V}$. A promising avenue arises, however, in the form of distilled datasets, wherein information is derived from a finite original full dataset. This ensures the existence of an optimal predictive family $\mathcal{V} \subseteq \Omega$ exemplified by observer models trained on the original full dataset. Consequently, the realism of the distilled dataset can be precisely assessed by leveraging this (almost) optimal predictive family.

Furthermore, the upper bound of diversity in the distilled dataset can be reliably guaranteed by the finite information (diversity) encapsulated within the original full dataset. This stands in stark contrast to the challenging task of limiting the diversity inherent in real-world datasets.

Data realism and $\mathcal{V}$-information. Consider an observer (predictive) family $\mathcal{V}$ capable of mapping image input X to its corresponding label output Y. If we transform the images X into encrypted versions or introduce additional noisy features beyond their natural background noise, predicting Y given X with the same $\mathcal{V}$ becomes more challenging.

To capture this intuition, a framework termed $\mathcal{V}$-information [42] generalizes Shannon information [5], measuring how much information can be extracted from X about Y when constrained to observers in $\mathcal{V}$, denoted as $I_{\mathcal{V}}(X \to Y)$. When $\mathcal{V}$ encompasses an infinite set of observers, corresponding to unbounded computation, $\mathcal{V}$-information reduces to Shannon information.

Likewise, unrealistic output labels for Y, such as encrypted or noisy labels, or even simplistic one-hot labels, prove inadequate in representing the precise information contained within images X. This inadequacy leads to diminished predictive accuracy, even when employing robust observers from the set $\mathcal{V}$.

Data diversity from the perspective of $\mathcal{V}$-information. $\mathcal{V}$-information $I_{\mathcal{V}}(X \to Y)$ serves as a conceptual tool for gauging the interconnected information between images X and labels Y. Consequently, this measurement is inherently influenced by the overall amount of information within both images X and labels Y. However, in the context of natural image datasets like ImageNet-1K [6], the diversity (information entropy) between images X and labels Y is notably imbalanced. Specifically, the labels Y often encompass considerably less information compared to the images X, thereby constraining $\mathcal{V}$-information $I_{\mathcal{V}}(X \to Y)$.

Summary. *Enhancing the diversity and realism of both the input X and the output Y in a dataset necessitates maximizing the $\mathcal{V}$-information $I_{\mathcal{V}}(X \to Y)$.*

B.3. Maximizing $\mathcal{V}$-information in Practice

Maximizing diversity of distilled data. Consider a predictive family $\mathcal{V} = \{\phi_{\mathrm{h}}, \phi_{\boldsymbol{\theta}_{\mathcal{T}}}\}$ and a distilled dataset $\mathcal{S}_c = (X_c, Y_c)$ for class c dataset $\mathcal{T}_c$, we assume:

$$\forall \mathcal{S}_c,\ \exists h \in \mathcal{H},\ \text{s.t.}\ \mathcal{S}_c = \{(\mathbf{x}_c, y_c) \mid y_c = h(\mathbf{x}_c)\}\,, \tag{15}$$

where $\mathcal{H} = \{h : \mathcal{X} \to \mathcal{Y}\}$. This assumption establishes the upper bound of diversity term for a distilled dataset $\mathcal{S}_c$, defined by the $\mathcal{V}$-entropy as follows:

$$\begin{aligned}
&H_{\mathcal{V}}(Y_c|\varnothing) \\
&= \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\varnothing](y_c)] \\
&= \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\varnothing](h(\mathbf{x}_c))] \\
&\leq \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\varnothing](\mathbf{x}_c)] \\
&= H_{\mathcal{V}}(X_c|\varnothing)\,.
\end{aligned} \tag{16}$$

Given $\mathcal{T}_c = (\hat{X}_c, \hat{Y}_c)$, where $(\hat{X}_c, \hat{Y}_c) := \{(\hat{\mathbf{x}}, \hat{y}) | (\hat{\mathbf{x}}, \hat{y}) \in \mathcal{T}, \hat{y} = c\}$, we have:

$$\begin{aligned}
&H_{\mathcal{V}}(\mathcal{T}_c|\varnothing) \\
&= H_{\mathcal{V}}((\hat{X}_c, \hat{Y}_c)|\varnothing) \\
&= \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\varnothing](\hat{\mathbf{x}}_c, \hat{y}_c)] \\
&= \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\varnothing](\hat{\mathbf{x}}_c, c)] \\
&= \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\varnothing](\hat{\mathbf{x}}_c)] \\
&\geq \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\varnothing](\mathbf{x}_c)] \\
&\geq H_{\mathcal{V}}(Y_c|\varnothing)\,.
\end{aligned} \tag{17}$$

Consequently, the above theoretical analysis can be extended to the entire dataset $\mathcal{T}$ and obtain that: $H_{\mathcal{V}}(Y|\varnothing) \leq H_{\mathcal{V}}(X|\varnothing) \leq H_{\mathcal{V}}(\mathcal{S}|\varnothing) \leq H_{\mathcal{V}}(\mathcal{T}|\varnothing) = C$, where C is a constant for a certain $\mathcal{T}$. Thus, we obtain:

$$H_{\mathcal{V}}(Y|\varnothing) \propto H_{\mathcal{V}}(Y|\varnothing)/H_{\mathcal{V}}(\mathcal{T}|\varnothing) \leq 1\,. \tag{18}$$

If we maximize the diversity term $H_{\mathcal{V}}(Y|\varnothing)$, then the ratio $H_{\mathcal{V}}(Y|\varnothing)/H_{\mathcal{V}}(\mathcal{T}|\varnothing) = 1$ and $H_{\mathcal{V}}(\mathcal{S}|\varnothing) = H_{\mathcal{V}}(\mathcal{T}|\varnothing)$. □

Maximizing realism of distilled data. Given a predictive family $\mathcal{V} = \{\phi_{\mathrm{h}}, \phi_{\boldsymbol{\theta}_{\mathcal{T}}}\}$ and a distilled dataset $\mathcal{S} = (X, Y)$, our objective is to minimize the realism term defined by the conditional $\mathcal{V}$-entropy:

$$\begin{aligned}
&H_{\mathcal{V}}(Y|X) \\
&= \inf_{f\in\mathcal{V}} \mathbb{E}[-\log f[\mathbf{x}](y)] \\
&\leq \mathbb{E}[-\log \phi_{\mathrm{h}}[\mathbf{x}](y)] + \mathbb{E}[-\log \phi_{\boldsymbol{\theta}_{\mathcal{T}}}[\mathbf{x}](y)]\,.
\end{aligned} \tag{19}$$

To estimate the density value $f[\mathbf{x}](y)$, we adopt the approach proposed by Oord et al. [26]:

$$f[\mathbf{x}](y) = \frac{\exp(-\ell(f(\mathbf{x}), y))}{\mathbb{E}_{y'\in Y}[\exp(-\ell(f(\mathbf{x}), y'))]}\,, \tag{20}$$

leading to:

$$\begin{aligned}
&H_{\mathcal{V}}(Y|X) \\
&\leq \mathbb{E}[-\log \frac{\exp(-\ell(\phi_{\mathrm{h}}(\mathbf{x}), y))}{\mathbb{E}_{y'\in Y}[\exp(-\ell(\phi_{\mathrm{h}}(\mathbf{x}), y'))]}] \\
&\quad + \mathbb{E}[-\log \frac{\exp(-\ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y))}{\mathbb{E}_{y'\in Y}[\exp(-\ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y'))]}]\,.
\end{aligned} \tag{21}$$

Assuming the function $\ell(\cdot)$ is symmetric, i.e.,

$$\forall z_1, z_2,\ \text{s.t.}\ \ell(z_1, z_2) = \ell(z_2, z_1)\,, \tag{22}$$

[5] The conventional approach of using Shannon [30]'s mutual information $I(X; Y)$ is not suitable in this context. This metric remains unchanged after the transformation of X, as it permits unbounded computation, including any necessary for the inverse transformation of images.

thus, we derive an alternative objective for minimization:

$$\begin{aligned}
&H_{\mathcal{V}}(Y|X) \\
&\propto \mathbb{E}[-\log \frac{\exp(-\ell(\phi_{\mathrm{h}}(\mathbf{x}), \phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x})))}{\mathbb{E}_{\mathbf{x}\in X}[\exp(-\ell(\phi_{\mathrm{h}}(\mathbf{x}), \phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x})))]}] \\
&\quad + \mathbb{E}[-\log \frac{\exp(-\ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y))}{\mathbb{E}_{y'\in Y}[\exp(-\ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y'))]}] \qquad (23) \\
&\propto \mathbb{E}[-\log \exp(-\ell(\phi_{\mathrm{h}}(\mathbf{x}), \phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x})))] \\
&\quad + \mathbb{E}[-\log \exp(-\ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y))] \\
&= \mathbb{E}[\ell(\phi_{\mathrm{h}}(\mathbf{x}), \phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x})) + \ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}), y)]\,.
\end{aligned}$$

This analysis underpins our strategy to enhance the realism of distilled data by minimizing $H_{\mathcal{V}}(Y|X)$, we focus on samples $\mathbf{x}$ that minimize $\ell(\phi_{\mathrm{h}}(\mathbf{x}), \phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x}))$ and set $y = \phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\mathbf{x})$. □

C. Detailed Implementation

C.1. Pre-training Observer Models

Following prior studies [1, 13, 44, 52], we employ pre-trained observer models to distill the dataset, as illustrated in Table 2: 1) ResNet-18 for ImageNet-10, ImageNette, ImageWoof, ImageNet-100, ImageNet-1K; 2) modified ResNet-18 for CIFAR-10, CIFAR-100 and Tiny-ImageNet; 3) ConvNet-3 for CIFAR-10, CIFAR-100; 4) ConvNet-4 for Tiny-ImageNet; 5) ConvNet-5 for ImageWoof, ImageNette; 6) ConvNet-6 for ImageNet-100.

C.2. Implementing RDED algorithm.

To gain an intuitive understanding the Algorithm 1 of our proposed RDED, we expound on the implementation details in this section. Given a comprehensive real dataset $\mathcal{T}$, such as ImageNet-1K [6], we define three tasks involving distilling this dataset into smaller datasets with distinct `IPC` values, specifically, $\texttt{IPC} = 50$, 10, and 1. Remarkably, our RDED demonstrates the capability to encompass multisize distilled datasets through a single distillation process, effectively handling those with $\texttt{IPC} = 50$, 10, and 1.

Extracting key patches. For each class set $\mathcal{T}_c$ we uniformly pre-select a subset contains 300 images denoted as $\mathcal{T}'_c = \{\hat{\mathbf{x}}_i\}_{i=1}^{300}$. Each pre-selected image $\hat{\mathbf{x}}_i$ undergoes random cropping into $K = 5$ patches[6]. These patches are represented as $\{\xi_{i,k}\}_{k=1}^{K=5}$, and the realism score $s_{i,k} = -\ell(\phi_{\boldsymbol{\theta}_{\mathcal{T}}}(\xi_{i,k}), y_i)$ is calculated for each patch $\xi_{i,k}$, resulting in a set of scores $\{s_{i,k}\}_{k=1}^{K=5}$. Subsequently, the key patch $\xi_{i,\star}$ with the highest realism score $s_{i,\star}$ is selected to represent the corresponding image $\mathbf{x}_i$. This process yields a key patch set with scores $\{\xi_{i,\star}, s_{i,\star}\}_{i=1}^{300}$, which is stored for future use.

Capturing class information. We prioritize key patches, denoted as $\{\xi_{i,\star}\}_{i=1}^{300}$, based on their associated scores $\{s_{i,\star}\}_{i=1}^{300}$ to construct a well-ordered set $\{\xi_{j,\star}\}_{j=1}^{300}$. In addressing the initial task of synthesizing a refined dataset with $\texttt{IPC} = 50$, we strategically choose the top-($200 = \texttt{IPC} \times N$) key patches from the set, denoted as $\{\xi_{j,\star}\}_{j=1}^{200}$. Likewise, for the two subsequent tasks, characterized by $\texttt{IPC} = 10$ and $\texttt{IPC} = 1$, we iteratively refine the selection by opting for the top-40 and top-4 key patches, denoted as $\{\xi_{j,\star}\}_{j=1}^{40}$ and $\{\xi_{j,\star}\}_{j=1}^{4}$, respectively.

Images reconstruction. To construct the ultimate image $\mathbf{x}_j$, we systematically draw $N = 4$ distinct patches $\{\xi_{j,\star}\}_{j=1}^{N=4}$ without replacement and concatenate them. This procedure is iterated times to generate the ultimate distilled image set $\{\mathbf{x}_j\}_{j=1}^{\texttt{IPC}}$.

Labels reconstruction. In accordance with the methodology presented in SRe^{2}L [44], we undertake the process of relabeling the distilled images through the generation and storage of region-level soft labels, denoted as y_j, employing Fast Knowledge Distillation [32]. To achieve this, for each distilled image $\mathbf{x}_j$, we perform random cropping into several patches, concurrently documenting their coordinates on the image $\mathbf{x}_j$. Subsequently, soft labels $y_{j,m}$ are generated and stored for each m-th patch, ultimately culminating in the aggregation of these labels to form the comprehensive y_j.

C.3. Training on Distilled Dataset

Following prior investigations [4, 44, 45], we employ data-augmentation techniques, namely RandomCropResize [41] and CutMix [46]. Further elucidation is available in our publicly accessible code repository at *https://to-be-released*.

D. Experiment

In this section, unless otherwise specified, we adopt ResNet-18 as the default neural network backbone for both the distillation process and subsequent evaluation. The parameters $\texttt{IPC} = 10$ and pre-selected subset size $|\mathcal{T}'_c| = 300$ are consistently applied. For high-resolution datasets, we set the number of patches $N = 4$ within one distilled image, while for datasets with a resolution lower than 64×64, we use $N = 1$. All settings are consistent with those in Section 5.

D.1. Multisize Dataset Distillation

In their recent work, He et al. [15] introduced Multisize Dataset Condensation (MDC), a novel approach that consolidates multiple condensation processes into a unified procedure. This innovative method produces datasets with varying sizes, offering dual advantages:

- DC eliminates the necessity for extra condensation processes when distilling multiple datasets with varying `IPC`.

[6] We empirically set $K = 5$, although smaller values, such as $K = 1$, can be chosen for expedited implementation of our algorithm RDED.

	RDED (Ours)					SRe^{2}L		
Verifier\Observer	ResNet-18	EfficientNet-B0	MobileNet-V2	VGG-11	Swin-V2-Tiny	ResNet-18	EfficientNet-B0	MobileNet-V2
ResNet-18	**42.3 ± 0.6**	**31.0 ± 0.1**	**40.4 ± 0.1**	36.6 ± 0.1	17.2 ± 0.2	21.7 ± 0.6	11.7 ± 0.2	15.4 ± 0.2
EfficientNet-B0	**42.8 ± 0.5**	**33.3 ± 0.9**	**43.6 ± 0.2**	35.8 ± 0.5	14.8 ± 0.1	25.2 ± 0.2	11.4 ± 2.5	20.5 ± 0.2
MobileNet-V2	**34.4 ± 0.2**	**24.1 ± 0.8**	**33.8 ± 0.6**	28.7 ± 0.2	11.8 ± 0.3	19.7 ± 0.1	9.8 ± 0.4	10.2 ± 2.6
VGG-11	**22.7 ± 0.1**	**16.5 ± 0.8**	**21.6 ± 0.2**	23.5 ± 0.3	7.8 ± 0.1	16.5 ± 0.1	9.3 ± 0.1	10.6 ± 0.1
Swin-V2-Tiny	**17.8 ± 0.1**	**19.7 ± 0.3**	**18.1 ± 0.2**	15.3 ± 0.4	12.1 ± 0.2	9.6 ± 0.3	10.2 ± 0.1	7.4 ± 0.1

Table 5. **Evaluating ImageNet-1K top-1 accuracy on cross-architecture generalization.** Distill dataset with VGG-11 [33], Swin-V2-Tiny [22], ResNet-18 [14], EfficientNet-B0 [36], MobileNet-V2 [29], and then versus transfer to other each other architecture.

- It facilitates a reduction in storage requirements by reusing condensed images.

Remarkably, our proposed RDED, also exhibits a mechanism that enables the synthesis of distilled datasets with adaptable IPC without incurring additional computational overhead (c.f. Section C). For a comprehensive comparison, the superior performance of our RDED over MDC on larger distilled datasets is demonstrated in Table 6.

	CIFAR-10			CIFAR-100		
Method \ IPC	1	10	50	1	10	50
MDC	**47.8**	**62.6**	**74.6**	**26.3**	41.4	53.7
Ours	23.5	50.2	68.4	19.6	**50.2**	**57.0**

Table 6. **Comparison with Multisize Dataset Condensation**. The top-1 validation accuracy is evaluated when both MDC and our RDED are targeting at distilling dataset with IPC = 50. The other two distilled datasets with IPC = 10 and IPC = 1 are subsets from the one with IPC = 50. The neural network backbone used for distillation and evaluation is Conv-3.

D.2. CoreSet Selection Baselines

In our investigation, we assess the top-1 validation accuracy resulting from the application of three CoreSet selection strategies for dataset distillation: 1) Random; 2) Herding [40]; 3) K-Means [11]. The outcomes, as depicted in Table 7, indicate catastrophically poor performance when employing these selection methods directly in the context of dataset distillation.

Dataset	Random	Herding	K-Means
ImageNet-10	**36.7 ± 0.1**	33.8 ± 0.4	36.5 ± 0.3
ImageNet-100	10.8 ± 0.2	12.6 ± 0.1	**13.5 ± 0.4**
ImageNet-1K	4.4 ± 0.1	**5.8 ± 0.1**	5.5 ± 0.1
Tiny-ImageNet	7.5 ± 0.1	**9.0 ± 0.3**	8.9 ± 0.2
CIFAR-100	10.9 ± 0.1	**13.3 ± 0.3**	12.9 ± 0.1
CIFAR-10	25.1 ± 0.5	**28.4 ± 0.1**	27.7 ± 0.2

Table 7. **Comparison of different CoreSet selection-based dataset distillation baselines**. Experiments are carried out to evaluate three widely used coreset selection methods.

D.3. Cross-architecture Generalization

We expanded our experimental evaluations by incorporating various neural network architectures that lack batch normalization [16, 44]. This extension aims to thoroughly assess the cross-architecture generalization capabilities of our proposed RDED. The results presented in Table 5 unequivocally demonstrate the superior performance of RDED in comparison to the SOTA method SRe^{2}L. Notably, our algorithm exhibits remarkable effectiveness even in scenarios characterized by substantial architectural disparities, such as knowledge transfer from ResNet-18 to Swin-V2-Tiny.

D.4. Detailed Ablation Study

In addition to the experiments detailed in Section 5.5, we conduct a more comprehensive ablation study, delving into the various approaches and hyperparameters employed in our proposed RDED.

On the impact of $|\mathcal{T}_c'|$ and N. To assess the influence of the pre-selected subset size $|\mathcal{T}_c'|$ and the number of patches within each distilled image N, our experiments are extended to lower-resolution datasets, namely Tiny-ImageNet, CIFAR-10, and CIFAR-100. Figure 5 illustrates that the configurations with $|\mathcal{T}_c'| = 300$ and $N = 1$ are suitable for low-resolution datasets.

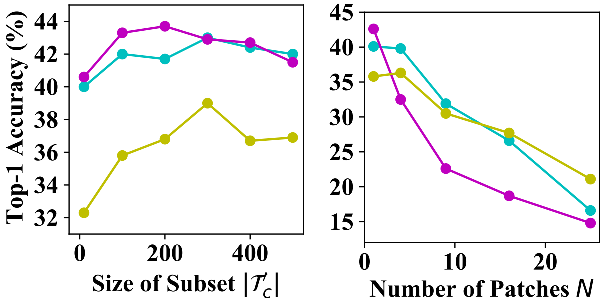


Figure 5. **Ablation study on $|\mathcal{T}_c'|$ and** N, i.e., the pre-selected subset size $\mathcal{T}_c'$ (left), and the number of patches N within each distilled image (right). The lemon ●, purple ●, and turquoise ● denote CIFAR-10, CIFAR-100, and Tiny-ImageNet respectively.

Dataset	Original	+EKP	+CCI	+IR	+LR
ImageNet-10	30.6 ± 0.4	34.5 ± 1.1	39.6 ± 1.6	49.9 ± 1.5	**54.3** ± **2.7**
ImageNet-100	8.2 ± 0.2	9.8 ± 0.1	15.0 ± 0.5	24.1 ± 0.1	**35.9** ± **0.1**
ImageNet-1K	3.2 ± 0.1	3.8 ± 0.1	7.2 ± 0.3	15.2 ± 0.1	**42.1** ± **0.1**
Tiny-ImageNet	6.9 ± 0.1	8.8 ± 0.1	15.7 ± 0.2	-	**41.9** ± **0.2**
CIFAR-100	11.8 ± 0.1	13.2 ± 0.3	18.6 ± 0.3	-	**42.6** ± **0.1**
CIFAR-10	27.7 ± 0.6	26.8 ± 0.2	27.8 ± 0.5	-	**35.8** ± **0.0**

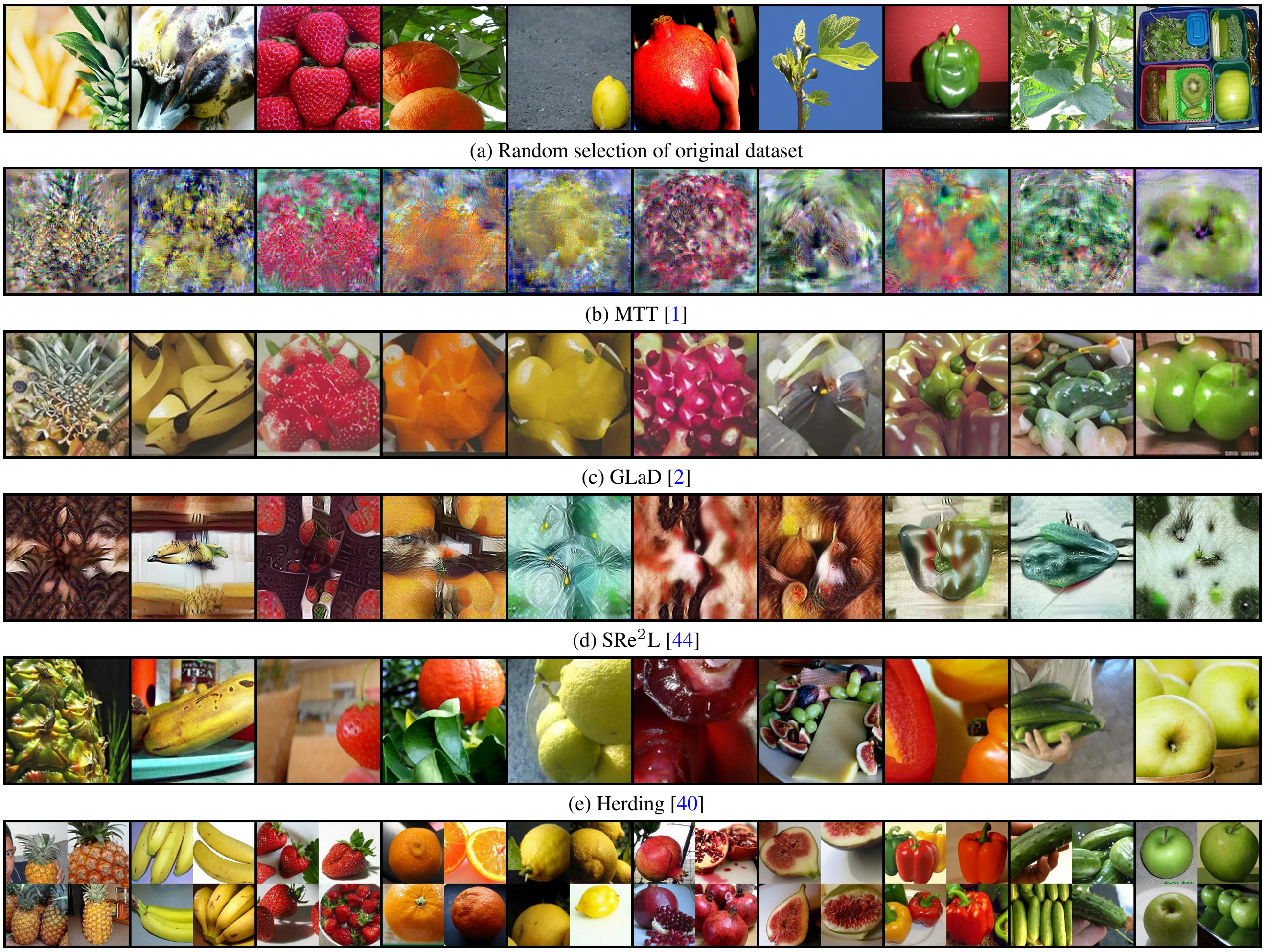
Table 8. **Effectiveness of accumulated techniques in RDED**. The validation accuracy undergoes a gradual evolution as we sequentially apply the four techniques in our RDED. Entries marked with "-" are absent because of the $N = 1$ setting for low-resolution datasets, rendering the Images Reconstruction (IR) step impractical.

Effectiveness of each technique in RDED. To validate the effectiveness of all four components within our RDED, we conduct additional ablation studies for each of them, namely, Extracting Key Patches (EKP), Capturing Class Information (CCI), Images Reconstruction (IR), and Labels Reconstruction (LR), corresponding to the techniques outlined in Sections 4.2 and 4.3. Table 8 illustrates that all four techniques employed in RDED are essential for achieving the remarkable final performance. Furthermore, a plausible hypothesis suggests that LR plays a crucial role in generating more informative (diverse) and aligned (realistic) labels for distilled images, thereby significantly enhancing performance.

Dataset	Random	Herding	K-Means	Realism
ImageNet-10	44.7 ± 2.5	47.9 ± 0.3	49.3 ± 1.1	**53.3** ± **0.1**
ImageNet-100	29.8 ± 0.7	29.7 ± 0.5	28.9 ± 0.1	**36.0** ± **0.3**
ImageNet-1K	37.9 ± 0.5	38.4 ± 0.1	38.2 ± 0.1	**42.0** ± **0.1**
Tiny-ImageNet	40.2 ± 0.0	41.1 ± 0.1	40.1 ± 0.1	**41.9** ± **0.2**
CIFAR-100	41.4 ± 0.5	**42.6** ± **0.1**	41.8 ± 0.1	**42.6** ± **0.1**
CIFAR-10	34.3 ± 0.1	35.5 ± 0.6	**37.9** ± **0.3**	35.8 ± 0.1

Table 9. **Comparison of different patch selection strategies in RDED**. Experiments are conducted to compare our proposed realism-score-based data selection strategy over three widely used coreset selection methods.

Effectiveness of selecting patches through realism socre. Table 9 demonstrates that our realism-score-based selection method, specifically the Capturing Class Information (CCI) technique outlined in Algorithm 1, consistently outperforms alternative approaches, except for CIFAR-10. A plausible inference is that the selection of more realistic images contributes to the observer model's ability to reconstruct correspondingly realistic labels (cf. Section 4.3), thereby optimizing our objective (3).



(a) Random selection of original dataset

(b) MTT [1]

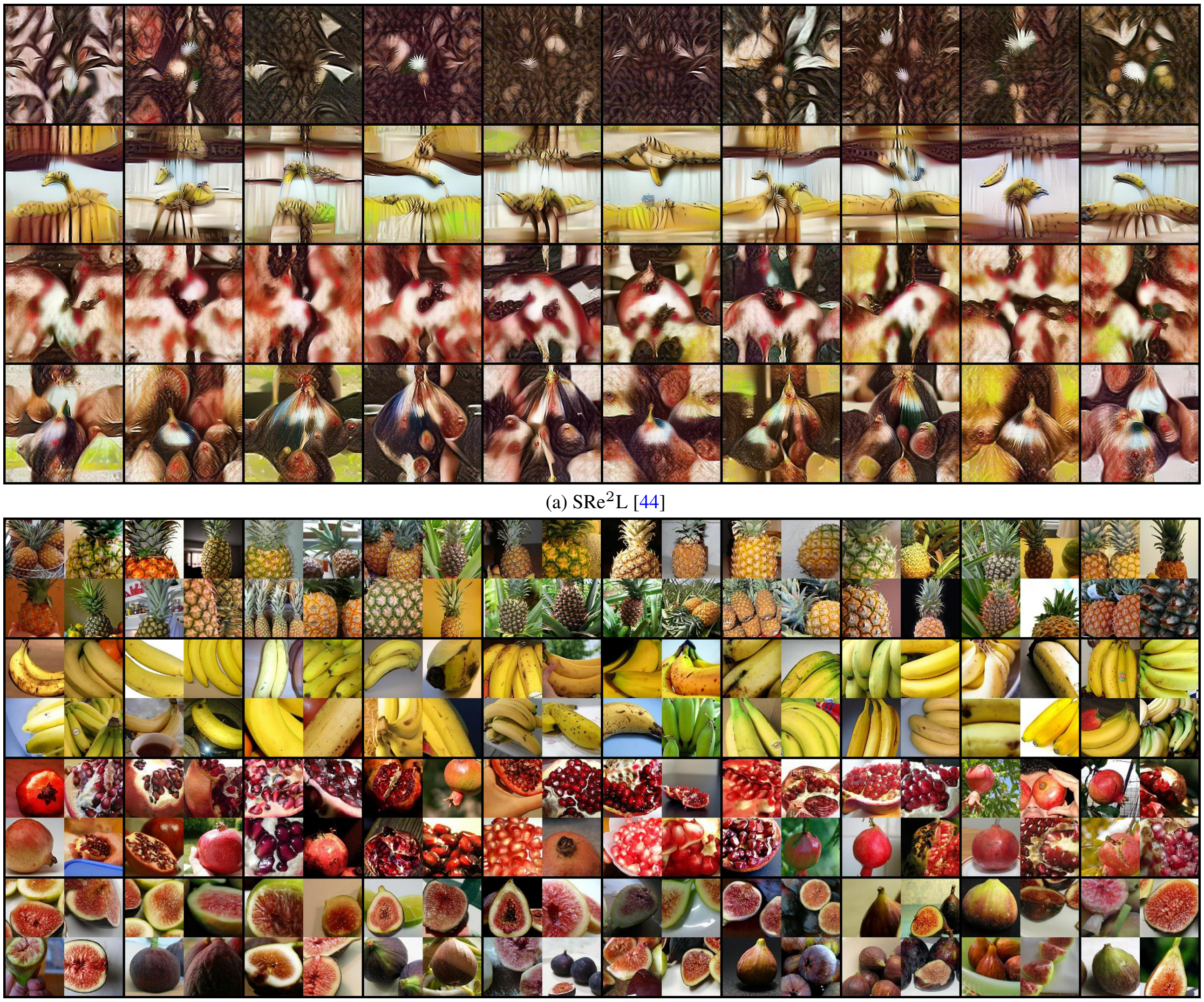
(c) GLaD [2]

(d) SRe^{2}L [44]

(e) Herding [40]

(f) RDED (Ours)

Figure 6. **Visualization of images synthesized using various dataset distillation methods**. We consider the ImageNet-Fruits [1] dataset, comprising a total of 10 distinct fruit types.



(a) SRe^{2}L [44]

(b) RDED (Ours)

Figure 7. **Visualization of images synthesized using two dataset distillation methods**. We consider a subset of the ImageNet-Fruits [1] dataset, comprising a total of 4 distinct fruit types.