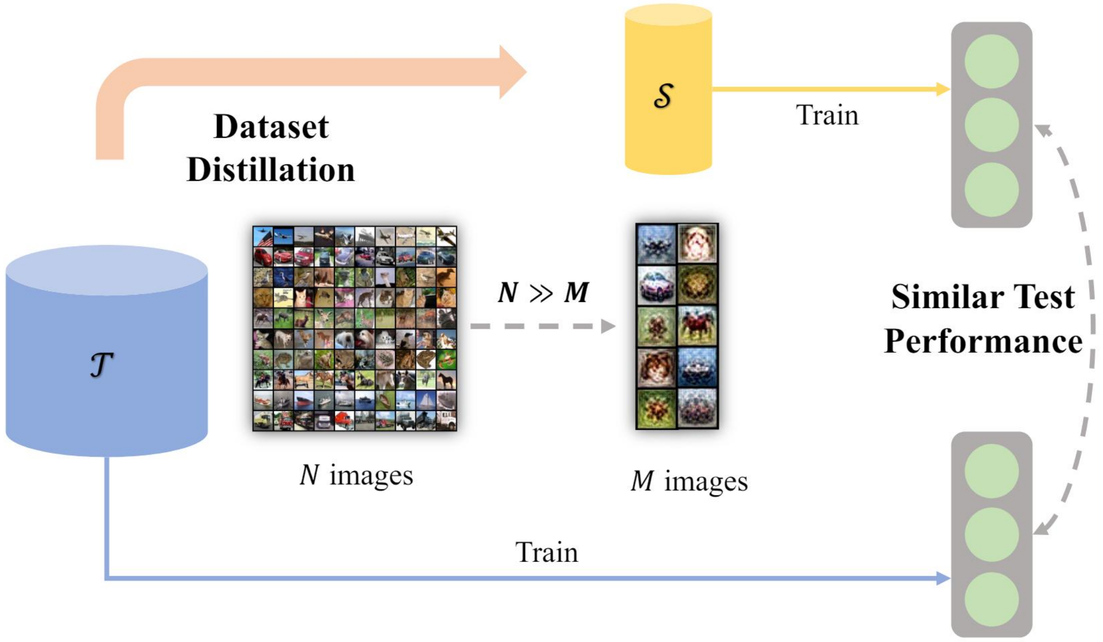
NEURAL INFORMATION PROCESSING SYSTEMS

Elucidating the Design Space of Dataset Condensation

Shitong Shao, Zikai Zhou, Huanran Chen and Zhiqiang Shen

NeurIPS 2024

What is Dataset Distillation/Condensation?



Source: Dataset Distillation: A Comprehensive Review—— https://arxiv.org/pdf/2301.07014

Motivation

- *Some dataset condensation (DC) methods incur high computational costs, which limit scalability to larger datasets*
- *Others are restricted to less optimal design spaces, which could hinder potential improvements*

NEURAL INFORMATION PROCESSING SYSTEMS

Definition

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 test protocol $\mathcal{P}_{\mathcal{D}}$.

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

NEURAL INFORMATION PROCESSING SYSTEMS

Definition

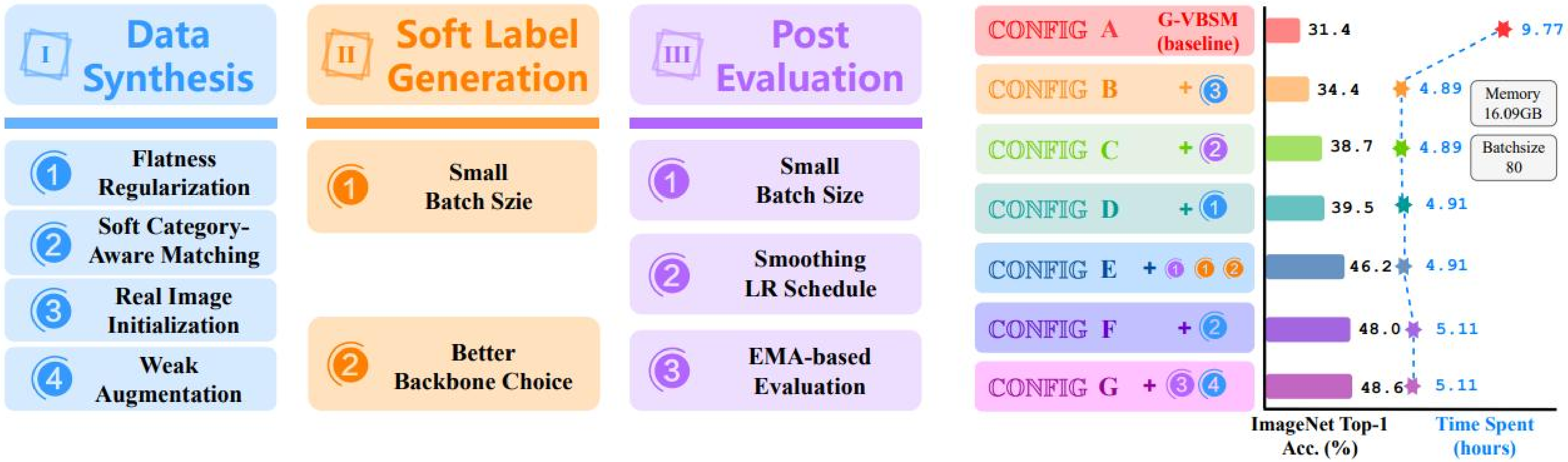
$$\mathcal{L}_{\textbf{syn}} = ||p(\mu|\mathcal{X}^{\mathcal{S}}) - p(\mu|\mathcal{X}^{\mathcal{T}})||_2 + ||p(\sigma^2|\mathcal{X}^{\mathcal{S}}) - p(\sigma^2|\mathcal{X}^{\mathcal{T}})||_2, \; s.t. \; \mathcal{L}_{\textbf{syn}} \sim \mathbb{S}_{\text{match}},$$
$$\mathcal{X}^{\mathcal{S}*} = \arg\min_{\mathcal{X}^{\mathcal{S}}}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

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

where $\mathbf{C}$ denotes the number of classes, 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 $\mathcal{X}^{\mathcal{S}}$.

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

Design Choice



Observation

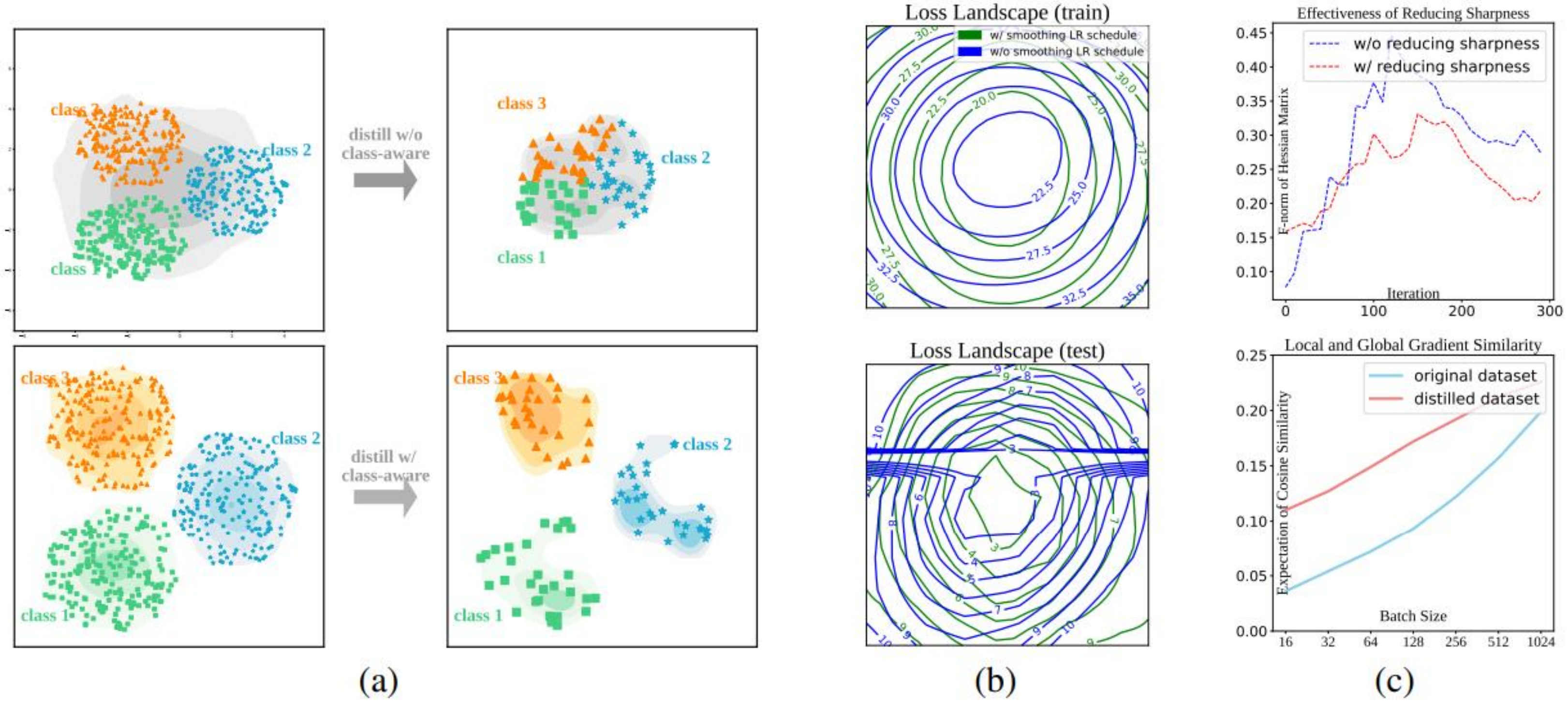
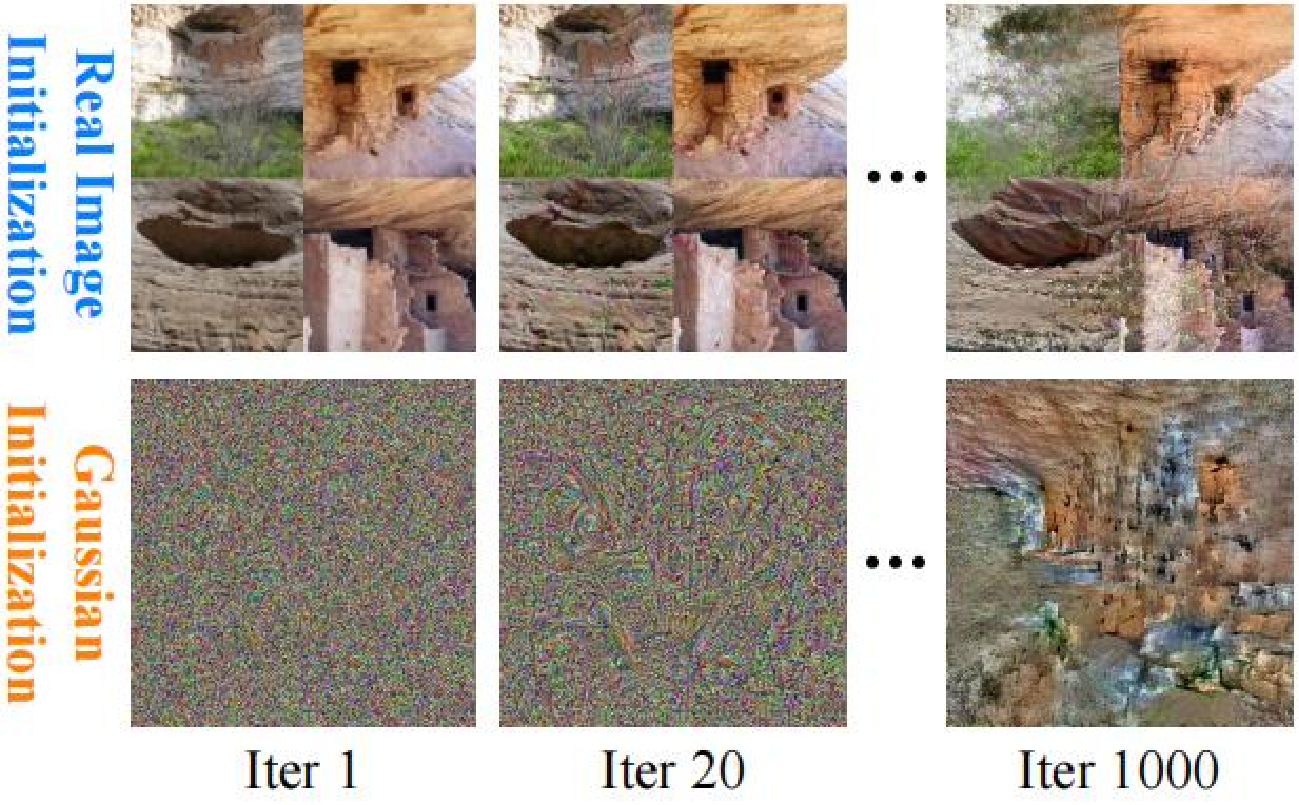


Figure 2: **(a):** Illustration of soft category-aware matching (②) using a Gaussian distribution in $\mathbb{R}^2$. **(b):** The effect of employing smoothing LR schedule (②) on loss landscape sharpness reduction. **(c) top:** The role of flatness regularization (①) in reducing the Frobenius norm of 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).

Real Data Initialization



Soft Category-aware Matching

Sketch Definition 3.1. *(formal definition in Appendix B.2) Given N random samples $\{x_i\}_{i=1}^{N}$ with an unknown distribution $p_{\text{mix}}(x)$, 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 $p_{\text{mix}}(x)$ 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 $\{y_i^j\}_{i=1}^{M_j}$.*

e.g., soft category-aware matching *More details of proofs and theorems can be found in our paper*

$$\mathcal{L}'_{\text{syn}} = \alpha||p(\mu|\mathcal{X}^{\mathcal{S}}) - p(\mu|\mathcal{X}^{\mathcal{T}})||_2 + ||p(\sigma^2|\mathcal{X}^{\mathcal{S}}) - p(\sigma^2|\mathcal{X}^{\mathcal{T}})||_2 \quad \text{\#Form (1)}$$

$$+ (1-\alpha)\sum^{\mathbf{C}} p(c_i)\left[||p(\mu|\mathcal{X}^{\mathcal{S}}, c_i) - p(\mu|\mathcal{X}^{\mathcal{T}}, c_i)||_2 + ||p(\sigma^2|\mathcal{X}^{\mathcal{S}}, c_i) - p(\sigma^2|\mathcal{X}^{\mathcal{T}}, c_i)||_2\right], \quad \text{\#Form (2)}$$

NEURAL INFORMATION PROCESSING SYSTEMS

Soft Category-aware Matching

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* $D_{KL}[p_{mix}||Q] = 0$.

NEURAL INFORMATION PROCESSING SYSTEMS

Flatness Regularization

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

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

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

Experiment

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.

NEURAL INFORMATION PROCESSING SYSTEMS

Experiment

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.

Experiment

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+(① ① ②)	42.3	48.4	47.0
G-VBSM+(③)	34.4	36.8	42.0
G-VBSM+(③ ②)	38.8	45.8	47.9
G-VBSM+(③ ② ① ①)	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 (③), smoothing LR schedule (②) and smaller batch size (① ①) 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	$\mathcal{L}_{\mathrm{FR}}$	0.025	1.5	0.999	4	38.8	43.2	40.3
CONFIG D	$\mathcal{L}_{\mathrm{FR}}$	0.25	1.5	0.999	4	37.9	43.5	40.3
CONFIG D	$\mathcal{L}_{\mathrm{FR}}$	2.5	1.5	0.999	4	31.7	37.0	32.9
CONFIG D	$\mathcal{L}_{\mathrm{FR}}$	0.25	1.5	0.99	4	39.0	43.3	40.2
CONFIG D	$\mathcal{L}'_{\mathrm{FR}}$	0.25	1.5	0.99	4	39.5	44.1	41.9
CONFIG D	$\mathcal{L}'_{\mathrm{FR}}$	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 (①).

Experiment

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).

NEURAL INFORMATION PROCESSING SYSTEMS

Experiment

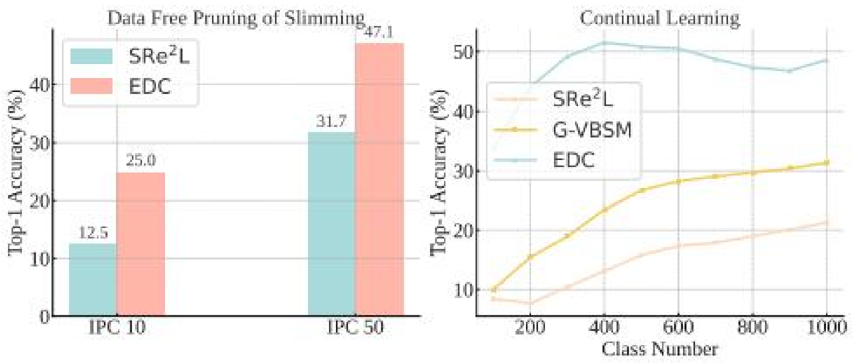


Figure 4: **Application on ImageNet-1k.** We evaluate the effectiveness of data-free network slimming and continual learning using VGG11-BN and ResNet-18, respectively.

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.**

THANK YOU!