Slimmable Dataset Condensation
– *Supplementary Materials* –

Songhua Liu, Jingwen Ye, Runpeng Yu, Xinchao Wang[*]
National University of Singapore
songhua.liu@u.nus.edu, jingweny@nus.edu.sg, r.yu@u.nus.edu, xinchao@nus.edu.sg

This document contains the following materials as supplements to the main manuscript:

- Theoretical derivation of Prop. 1 in the main paper.
- Full quantitative comparisons of the main paper and results on more datasets and settings.
- More experimental analysis of the proposed algorithm.
- Results in more settings of continual learning and federated learning.

A. Theoretical Derivations

We explore a significance-aware parameterization for synthetic datasets in this paper. A synthetic dataset (X_s, Y_s) is parameterized by (U, Σ, V_x, V_y):

- $V_x \in \mathbb{R}^{b \times d}$ and $V_y \in \mathbb{R}^{b \times c}$ denote orthogonal bases for constructing samples and labels respectively, where b is the total number of components;
- $\Sigma = \mathrm{diag}(s_1, \dots, s_b)$ with $s_1 \geq \dots \geq s_b$ is a diagonal matrix, where each s_i denotes the significance of the i-th component;
- $U \in \mathbb{R}^{n_s \times b}$ is an orthogonal matrix representing coefficients of different components for constructing each data.

The synthetic samples and corresponding labels are constructed by:

$$X_s = U \Sigma V_x, \quad Y_s = U \Sigma V_y. \tag{1}$$

In implementation, samples of each class share the same U and Σ for memory efficiency. We find that it is possible to simply discard less important components when we need to slim a synthetic dataset, *i.e.*, deleting the entries with least singular values in Σ, the corresponding columns in U, and the corresponding rows in V_x and V_y, which has the potential to serve as either a learning-free slimmable DC strategy or a strong initialization for learning-based slimmable DC. Theoretically, in the case of linear regression, the error on the resultant solution plane satisfies the following proposition:

Proposition 1. *In linear regression, if a synthetic dataset (X_s, Y_s) takes the parameterization in Eq. 1, and rows in V_x and V_y corresponding to the least singular values in Σ, denoted as $\tilde{V_x}$ and $\tilde{V_y}$, are removed for slimmable DC, the first-order parameter distance between parameters before and after slimming is bounded by:*

$$\|w_s'^1 - w_s^1\|_2^2 \leq s_2^2 \|X_s w - Y_s\|_2^2, \tag{2}$$

and the infinity-order parameter distance is bounded by:

$$\|{X_s'}^{\dagger} Y_s' - {X_s}^{\dagger} Y_s\|_2^2 = \|\tilde{V_x}^{\top} \tilde{V_y}\|_2^2 \leq 1. \tag{3}$$

[*]Corresponding Author.

Proof. Let $\hat{\Sigma}$ and $\tilde{\Sigma}$ denote diagonal matrices with the largest and the least singular values in Σ, $\hat{V_x}$ and $\hat{V_y}$ denote rows in V_x and V_y corresponding to the largest singular values in Σ, $\tilde{V_x}$ and $\tilde{V_y}$ denote rows in V_x and V_y corresponding to the least ones, $\hat{U}$ denote columns in U corresponding to the largest singular values in Σ, and $\tilde{U}$ denote columns in U corresponding to the least ones. We have $X_s' = \hat{U}\hat{U}^\top X_s$ and $Y_s' = \hat{U}\hat{U}^\top Y_s$, which can be verified by:

$$\begin{aligned}\hat{U}\hat{U}^\top X_s &= \hat{U}\hat{U}^\top \begin{bmatrix}\hat{U} & \tilde{U}\end{bmatrix} \begin{bmatrix}\hat{\Sigma} & 0\\ 0 & \tilde{\Sigma}\end{bmatrix} \begin{bmatrix}\hat{V_x}\\ \tilde{V_x}\end{bmatrix}\\ &= \begin{bmatrix}\hat{U} & 0\end{bmatrix} \begin{bmatrix}\hat{\Sigma} & 0\\ 0 & \tilde{\Sigma}\end{bmatrix} \begin{bmatrix}\hat{V_x}\\ \tilde{V_x}\end{bmatrix}\\ &= \hat{U}\hat{\Sigma}\hat{V_x} = X_s'.\end{aligned} \tag{4}$$

For Y_s', the verification is similar. Then, for the first-order parameter distance,

$$\begin{aligned}\|w_s'^1 - w_s^1\|_2^2 &= \|{X_s'}^\top (X_s' w - Y_s') - X_s^\top (X_s w - Y_s)\|_2^2\\ &= \|(\hat{U}\hat{U}^\top X_s)^\top (\hat{U}\hat{U}^\top X_s w - \hat{U}\hat{U}^\top Y_s) - X_s^\top (X_s w - Y_s)\|_2^2\\ &= \|X_s^\top \hat{U}\hat{U}^\top \hat{U}\hat{U}^\top (X_s w - Y_s) - X_s^\top (X_s w - Y_s)\|_2^2\\ &= \|(X_s^\top \hat{U}\hat{U}^\top - X_s^\top)(X_s w - Y_s)\|_2^2\\ &\leq \|X_s^\top - X_s^\top \hat{U}\hat{U}^\top\|_2^2 \|X_s w - Y_s\|_2^2\\ &= \|(U\Sigma V_x)^\top - (U\Sigma V_x)^\top \hat{U}\hat{U}^\top\|_2^2 \|X_s w - Y_s\|_2^2\\ &= \|V_x^\top \Sigma \begin{bmatrix}\hat{U}^\top\\ \tilde{U}^\top\end{bmatrix} - V_x^\top \Sigma \begin{bmatrix}\hat{U}^\top\\ \tilde{U}^\top\end{bmatrix} \hat{U}\hat{U}^\top\|_2^2 \|X_s w - Y_s\|_2^2\\ &= \|V_x^\top \Sigma (\begin{bmatrix}\hat{U}^\top\\ \tilde{U}^\top\end{bmatrix} - \begin{bmatrix}\hat{U}^\top\\ 0\end{bmatrix})\|_2^2 \|X_s w - Y_s\|_2^2\\ &= \| \begin{bmatrix}\hat{V_x}^\top & \tilde{V_x}^\top\end{bmatrix} \begin{bmatrix}\hat{\Sigma} & 0\\ 0 & \tilde{\Sigma}\end{bmatrix} \begin{bmatrix}0\\ \tilde{U}^\top\end{bmatrix} \|_2^2 \|X_s w - Y_s\|_2^2\\ &= \|\tilde{V_x}^\top \tilde{\Sigma}\tilde{U}^\top\|_2^2 \|X_s w - Y_s\|_2^2\\ &\leq \sigma_2^2 \|X_s w - Y_s\|_2^2,\end{aligned} \tag{5}$$

where the last inequality holds since the component corresponding to the largest singular value σ_1 would always be kept and singular value of deleted components is σ_2 at most.

For the infinity-order parameter distance,

$$\begin{aligned}\|{X_s'}^\dagger Y_s' - {X_s}^\dagger Y_s\|_2^2 &= \|{X_s'}^\top (X_s' {X_s'}^\top)^{-1} Y_s' - X_s^\top (X_s X_s^\top)^{-1} Y_s\|_2^2\\ &= \|X_s^\top \hat{U}\hat{U}^\top (\hat{U}\hat{U}^\top X_s X_s^\top \hat{U}\hat{U}^\top)^{-1} \hat{U}\hat{U}^\top Y_s - X_s^\top (X_s X_s^\top)^{-1} Y_s\|_2^2\\ &= \|X_s^\top \hat{U}\hat{U}^\top \hat{U}\hat{U}^\top (X_s X_s^\top)^{-1} \hat{U}\hat{U}^\top \hat{U}\hat{U}^\top Y_s - X_s^\top (X_s X_s^\top)^{-1} Y_s\|_2^2\\ &= \|V_x^\top \Sigma U^\top \hat{U}\hat{U}^\top (U\Sigma V_x V_x^\top \Sigma U^\top)^{-1} \hat{U}\hat{U}^\top U\Sigma V_y - V_x^\top \Sigma U^\top (U\Sigma V_x V_x^\top \Sigma U^\top)^{-1} U\Sigma V_y\|_2^2\\ &= \|V_x^\top \Sigma U^\top \hat{U}\hat{U}^\top U(\Sigma^2)^{-1} U^\top \hat{U}\hat{U}^\top U\Sigma V_y - V_x^\top \Sigma U^\top U(\Sigma^2)^{-1} U^\top U\Sigma V_y\|_2^2\\ &= \| \begin{bmatrix}\hat{V_x}^\top & \tilde{V_x}^\top\end{bmatrix} \begin{bmatrix}\hat{\Sigma} & 0\\ 0 & \tilde{\Sigma}\end{bmatrix} \begin{bmatrix}\hat{U}^\top\\ \tilde{U}^\top\end{bmatrix} \hat{U}\hat{U}^\top \begin{bmatrix}\hat{U} & \tilde{U}\end{bmatrix} \begin{bmatrix}\hat{\Sigma}^2 & 0\\ 0 & \tilde{\Sigma}^2\end{bmatrix}^{-1} \begin{bmatrix}\hat{U}^\top\\ \tilde{U}^\top\end{bmatrix} \hat{U}\hat{U}^\top \begin{bmatrix}\hat{U} & \tilde{U}\end{bmatrix} \begin{bmatrix}\hat{\Sigma} & 0\\ 0 & \tilde{\Sigma}\end{bmatrix} \begin{bmatrix}\hat{V_y}\\ \tilde{V_y}\end{bmatrix}\\ &\quad - \begin{bmatrix}\hat{V_x}^\top & \tilde{V_x}^\top\end{bmatrix} \Sigma(\Sigma^2)^{-1}\Sigma \begin{bmatrix}\hat{V_y}\\ \tilde{V_y}\end{bmatrix} \|_2^2\\ &= \|\hat{V_x}^\top \hat{V_y} - \hat{V_x}^\top \hat{V_y} - \tilde{V_x}^\top \tilde{V_y}\|_2^2\\ &= \|\tilde{V_x}^\top \tilde{V_y}\|_2^2 \leq \|\tilde{V_x}\|_2^2 \|\tilde{V_y}\|_2^2 = 1,\end{aligned} \tag{6}$$

□

	IPC	50	20	10	5	2	1
DC [14]	RT	82.89±0.16	84.37±0.23	83.38±0.28	80.54±0.14	76.07±0.31	70.27±0.71
	LBS	-	79.60±0.32	75.42±0.42	67.92±0.39	61.34±0.97	57.35±1.66
	Gap↓	-	4.77	7.96	12.62	14.73	12.92
	LFS	-	72.25±0.53	70.22±0.36	56.39±0.49	54.79±0.74	34.84±0.82
DSA [12]	RT	88.73±0.08	86.68±0.16	85.27±0.13	81.99±0.25	76.66±0.24	70.33±0.72
	LBS	-	86.08±0.16	83.32±0.19	79.17±0.30	70.29±0.65	51.58±1.19
	Gap↓	-	0.60	1.95	2.82	6.37	18.75
	LFS	-	79.86±0.29	74.14±0.18	71.27±0.21	54.63±0.56	43.81±1.46
DM [13]	RT	88.20±0.27	86.21±0.21	83.84±0.16	80.89±0.21	74.42±0.24	71.45±0.49
	LBS	-	85.92±0.14	83.21±0.27	80.21±0.16	73.78±0.25	70.69±0.49
	Gap↓	-	0.29	0.63	0.68	0.64	0.76
	LFS	-	81.05±0.27	**78.56±0.05**	68.04±0.49	59.22±0.53	58.48±0.43
IDC [4]	RT	89.06±0.15	86.81±0.21	85.16±0.35	83.13±0.14	77.96±0.16	70.64±0.37
	LBS	-	84.81±0.29	83.36±0.24	81.16±0.23	76.52±0.49	67.73±1.00
	Gap↓	-	2.00	1.80	1.97	1.44	2.91
	LFS	-	82.57±0.23	77.02±0.28	74.41±0.50	60.86±0.88	52.75±1.41
FRePo [15]	RT	**89.15±0.13**	87.44±0.21	85.54±0.15	**83.80±0.21**	**79.91±0.44**	**75.44±0.45**
	LBS	-	86.60±0.38	81.53±0.38	67.74±0.49	33.44±1.79	29.24±2.25
	Gap↓	-	0.84	4.01	16.06	46.47	46.20
	LFS	-	82.59±0.17	75.65±0.34	71.76±0.28	61.94±0.75	44.00±1.92
Ours	RT	88.68±0.15	**87.50±0.13**	**86.65±0.08**	83.54±0.34	79.63±0.82	74.14±0.31
	LBS	-	**86.81±0.07**	**85.18±0.21**	**83.62±0.20**	**78.58±0.71**	**72.74±0.67**
	Gap↓	-	0.69	1.47	-0.08	1.05	1.40
	LFS	-	**82.96±0.21**	76.71±0.36	**74.72±0.45**	**69.52±0.47**	**66.43±0.76**

Table 1. Comparisons with existing typical DC algorithms on the performance of slimmable DC on FashionMNIST. IPC: number of images per class. RT: retraining using original datasets. LBS: learning-based slimming. LFS: learning-free slimming.

	IPC	50	20	10	5	2	1
DC [14]	RT	53.43±0.28	49.73±0.27	43.74±0.41	39.88±0.50	38.89±0.31	28.20±0.71
	LBS	-	45.59±0.35	39.83±0.53	35.45±0.58	29.65±0.46	23.00±0.58
	Gap↓	-	4.14	3.91	4.43	9.24	5.20
	LFS	-	41.82±0.59	32.66±0.15	25.88±0.36	18.76±0.25	17.37±0.29
DSA [12]	RT	60.58±0.29	57.11±0.22	52.15±0.48	47.31±0.26	34.23±0.26	28.10±0.72
	LBS	-	52.77±0.35	46.55±0.64	39.51±0.37	30.07±0.43	20.48±0.17
	Gap↓	-	4.34	5.60	7.80	4.16	7.62
	LFS	-	41.54±0.43	29.29±0.25	27.56±0.21	20.10±0.32	14.05±0.26
DM [13]	RT	62.94±0.28	55.41±0.55	48.80±0.31	42.89±0.28	33.50±0.50	27.08±0.36
	LBS	-	56.47±0.42	49.89±0.19	43.57±0.38	34.35±0.74	26.67±0.83
	Gap↓	-	-1.06	-1.09	-0.68	-0.85	0.41
	LFS	-	46.76±0.52	35.35±0.75	25.34±0.34	16.05±0.34	13.81±0.45
IDC [4]	RT	69.32±0.30	62.01±0.28	58.50±0.39	52.13±0.66	44.12±0.78	35.34±0.87
	LBS	-	58.77±0.23	54.24±0.35	47.83±0.75	38.61±0.69	29.16±1.41
	Gap↓	-	3.24	4.26	4.30	5.51	6.18
	LFS	-	51.91±0.49	42.17±0.49	30.20±0.36	22.84±0.54	17.68±0.71
FRePo [15]	RT	**71.03±0.34**	**68.63±0.53**	**65.76±0.72**	**61.07±0.31**	53.24±0.37	43.24±0.32
	LBS	-	65.64±0.30	53.76±0.92	38.02±1.03	17.31±0.38	11.01±0.38
	Gap↓	-	2.99	12.00	23.05	35.93	32.23
	LFS	-	59.14±0.73	**50.48±0.19**	38.34±0.88	29.60±0.60	18.22±0.55
Ours	RT	70.33±0.34	67.60±0.22	64.57±0.24	59.49±0.19	52.88±0.73	**43.56±0.43**
	LBS	-	**67.93±0.48**	**63.96±0.59**	**61.05±0.32**	**55.82±0.46**	**47.77±0.35**
	Gap↓	-	-0.33	0.61	-1.56	-2.94	-4.21
	LFS	-	**62.05±0.29**	48.89±0.54	**40.48±0.34**	**36.51±0.16**	**33.09±0.29**

Table 2. Comparisons with existing typical DC algorithms on the performance of slimmable DC on CIFAR10. IPC: number of images per class. RT: retraining using original datasets. LBS: learning-based slimming. LFS: learning-free slimming.

B. Quantitative Comparisons

Here, we provide full quantitative comparison results with previous methods on 5 widely-adopted benchmarks including FashionMNIST [11], CIFAR10, CIFAR100 [5], Tiny-ImageNet [6], and ImageNette [3][1]. The number of classes is 10, 10, 100, 200, and 10 and the resolution is 28, 32, 32, 64, and 128, respectively. The protocol for comparison maintains the same

[1] For experiments on Tiny-ImageNet and ImageNette, we load the publicly-available pre-trained synthetic datasets of FRePo.

	IPC	20	10	5	2	1
DC [14]	RT	28.93±0.26	25.08±0.17	21.29±0.15	16.46±0.39	12.44±0.18
	LBS	-	21.28±0.33	17.07±0.25	12.63±0.21	9.34±0.22
	Gap↓	-	3.80	4.22	3.83	3.10
	LFS	-	21.78±0.32	13.30±0.28	6.74±0.21	4.64±0.05
DSA [12]	RT	36.35±0.18	32.49±0.30	27.35±0.42	20.47±0.06	13.81±0.21
	LBS	-	29.89±0.27	24.34±0.27	17.61±0.41	11.62±0.14
	Gap↓	-	2.60	3.01	2.86	2.19
	LFS	-	23.69±0.34	14.92±0.25	8.06±0.12	4.95±0.12
DM [13]	RT	34.39±0.30	29.33±0.23	23.91±0.23	15.98±0.14	11.51±0.25
	LBS	-	30.84±0.21	24.74±0.34	16.47±0.16	11.62±0.40
	Gap↓	-	-1.51	-0.83	-0.49	-0.11
	LFS	-	26.72±0.20	15.69±0.31	7.95±0.22	5.22±0.11
IDC [4]	RT	41.99±0.23	36.08±0.38	30.68±0.17	23.34±0.19	17.93±0.15
	LBS	-	35.16±0.27	28.29±0.18	18.39±0.17	13.40±0.31
	Gap↓	-	0.92	2.39	4.95	4.53
	LFS	-	30.25±0.16	19.50±0.17	10.96±0.14	7.63±0.11
FRePo [15]	RT	40.57±0.26	39.97±0.32	36.34±0.21	31.63±0.26	**27.07±0.26**
	LBS	-	35.53±0.36	32.08±0.55	26.51±0.36	19.27±0.59
	Gap↓	-	4.44	4.26	5.12	7.80
	LFS	-	35.18±0.32	**30.00±0.59**	19.94±0.28	13.63±0.12
Ours	RT	**42.47±0.20**	**40.29±0.36**	**36.42±0.21**	**32.28±0.14**	26.75±0.34
	LBS	-	**36.23±0.46**	**33.49±0.55**	**29.27±0.36**	**26.04±0.38**
	Gap↓	-	4.06	2.93	3.01	0.71
	LFS	-	**35.39±0.04**	28.58±0.18	**23.69±0.31**	**20.34±0.26**

Table 3. Comparisons with existing typical DC algorithms on the performance of slimmable DC on CIFAR100. IPC: number of images per class. RT: retraining using original datasets. LBS: learning-based slimming. LFS: learning-free slimming.

	IPC	10	5	2	1
FRePo [15]	LBS	26.86±0.05	20.06±0.18	14.12±0.26	10.22±0.17
	LFS	26.86±0.05	20.49±0.29	14.17±0.07	9.82±0.05
Ours	LBS	26.80±0.17	**21.06±0.05**	**18.21±0.21**	**16.21±0.44**
	LFS	26.80±0.17	**20.74±0.15**	**15.30±0.14**	**12.92±0.14**

Table 4. Comparisons with the baseline FRePo on the performance of slimmable DC on Tiny-ImageNet. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming.

	IPC	10	5	2	1
FRePo [15]	LBS	67.23±0.74	42.65±0.24	19.74±0.22	12.49±0.54
	LFS	67.23±0.74	55.37±0.32	33.59±0.41	21.44±0.37
Ours	LBS	67.67±0.35	**60.36±0.64**	**53.57±0.59**	**44.88±1.15**
	LFS	67.67±0.35	**55.76±1.03**	**48.97±0.61**	**39.54±0.39**

Table 5. Comparisons with the baseline FRePo on the performance of slimmable DC on ImageNette. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming.

as that in the main paper: we compress a real dataset to a relatively large synthetic one and then perform successive slimmable condensations. Results for the 5 datasets are shown in Tabs. 1, 2, 3, 4, and 5, respectively. Results for cross-architecture performance can be found in Tabs. 6, 7, 8, 9, and 10. All results are based on 5 repeated evaluations and we report the average results and the standard deviations. The conclusion is consistent with that in the main paper.

Although some works also focus on synthetic dataset parameterization [1,7] to boost the performance of DC, as mentioned in the related work section of the main paper, the highlight of this paper is on the co-design of parameterization and loss terms, which makes it suitable for slimmable DC. Since previous works do not take significance of various components into account, their LFS performances are unsatisfactory as shown in the 3rd, 4th, and 10th cols. of Tab. 11.

The main analysis of this paper is on the state-of-the-art methods, which are based on kernel ridge regression (KRR) [8–10, 15]. In Tab.1 of the main paper, we also present some insights on a wider spectrum of methods including those based on gradient-matching and distribution matching. Here, we further provide experimental results for methods based on back-propagation-through-time (BPTT) and matching-training-trajectory (MTT) in Tab. 11. For **BPTT**, the difference with KRR

	IPC	20	10	5	2	1
ResNet	FRePo	84.00±0.71	76.78±0.78	59.01±2.42	37.26±3.34	31.44±2.56
	Ours	**85.41±0.29**	**82.60±1.06**	**78.07±1.74**	**71.21±0.61**	**56.93±0.43**
AlexNet	FRePo	84.20±0.17	78.69±0.16	58.73±2.39	34.47±3.87	37.53±1.87
	Ours	**85.64±0.09**	**82.32±0.03**	**78.60±0.56**	**70.53±0.84**	**52.59±2.25**
VGG	FRePo	79.60±0.63	71.48±0.48	51.99±1.34	39.80±1.21	28.37±2.02
	Ours	**80.86±0.39**	**76.96±0.42**	**73.19±0.77**	**57.61±3.48**	**39.15±1.96**

Table 6. Comparisons with the baseline FRePo on cross-architecture performance of slimmable DC on FashionMNIST. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming.

	IPC	20	10	5	2	1
ResNet	FRePo	53.22±1.74	39.34±1.57	23.97±1.00	14.09±0.67	11.60±0.28
	Ours	**56.24±1.05**	**49.41±0.89**	**43.73±2.63**	**32.74±1.19**	**25.56±1.02**
AlexNet	FRePo	59.65±0.21	44.03±0.93	29.24±0.39	14.85±1.22	11.76±0.36
	Ours	**63.91±0.29**	**57.33±0.34**	**53.74±1.30**	**45.27±1.10**	**36.27±0.63**
VGG	FRePo	50.72±0.83	34.79±1.40	23.11±1.05	12.93±1.27	10.69±0.89
	Ours	**54.78±1.70**	**43.78±0.39**	**37.43±1.35**	**32.31±1.54**	**28.02±0.89**

Table 7. Comparisons with the baseline FRePo on cross-architecture performance of slimmable DC on CIFAR10. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming.

	IPC	10	5	2	1
ResNet	FRePo	26.14±1.49	18.82±0.82	13.69±0.68	8.63±0.14
	Ours	**26.75±0.99**	**20.60±1.17**	**16.92±0.48**	**12.43±0.22**
AlexNet	FRePo	33.75±0.09	26.97±0.22	19.58±0.11	11.68±0.25
	Ours	**34.85±0.20**	**29.07±0.25**	**24.56±0.35**	**17.27±0.23**
VGG	FRePo	28.88±0.41	21.66±0.63	12.69±0.16	6.62±0.23
	Ours	**29.65±0.48**	**24.23±0.41**	**18.30±0.10**	**13.90±0.42**

Table 8. Comparisons with the baseline FRePo on cross-architecture performance of slimmable DC on CIFAR100. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming.

	IPC	5	2	1
ResNet	FRePo	10.01±0.78	5.91±0.58	4.73±0.48
	Ours	**10.66±0.58**	**8.58±0.72**	**7.15±0.86**
AlexNet	FRePo	15.46±0.28	10.80±0.35	7.50±0.26
	Ours	**17.75±0.28**	**14.83±0.34**	**12.76±0.53**
VGG	FRePo	15.01±0.26	9.22±0.13	5.94±0.27
	Ours	**17.64±0.36**	**15.24±0.39**	**13.49±0.29**

Table 9. Comparisons with the baseline FRePo on cross-architecture performance of slimmable DC on Tiny-ImageNet. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming.

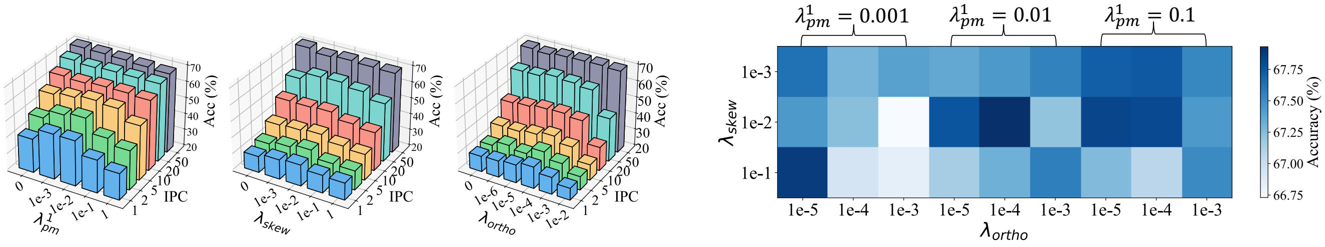
	IPC	5	2	1
ResNet	FRePo	33.97±2.41	16.72±1.23	13.58±0.51
	Ours	**42.29±1.26**	**37.37±2.97**	**28.45±1.86**
AlexNet	FRePo	40.59±1.24	17.82±1.37	12.37±1.09
	Ours	**53.55±0.38**	**48.56±1.42**	**40.43±1.25**
VGG	FRePo	34.11±1.96	17.57±2.98	11.93±0.41
	Ours	**50.13±3.61**	**45.71±1.95**	**35.64±0.51**

Table 10. Comparisons with the baseline FRePo on cross-architecture performance of slimmable DC on ImageNette. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming.

is that the model becomes non-linear. Linear models as analyzed in the paper suffer from the problem of underdetermined solution space. Based on results in the 5th and 6th cols. of Tab. 11, the non-linearity does not get rid of this issue. For **MTT**,

IPC	RS	Deng *et al.* [1]	Liu *et al.* [7]	BPTT		MTT		Ours			
		LFS	LFS	RT	LBS	RT	LBS	RT	LFS	LBS w/o Param.	LBS
10 (Original)	31.60	71.20	69.90	63.94	-	64.10	-	64.57	-	-	-
10 → 5	24.82	39.44	34.98	59.59	53.23	57.92	36.29	59.49	**42.90**	58.51	**60.96**
5 → 1	16.68	27.36	18.83	48.71	34.14	45.69	18.83	43.56	**33.87**	43.21	**46.82**

Table 11. More comparisons with existing methods and more ablation studies in the setting of slimmable DC on CIFAR10. IPC: number of images per class. LBS: learning-based slimming. LFS: learning-free slimming. RS: randomly selected real images.



(a)

Figure 1. Sensitivity analysis of hyper-parameters.

(b)

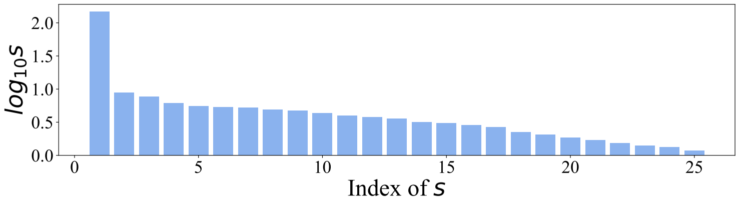


Figure 2. Distribution of top 25 significance scores for 50 IPC of CIFAR10.

LFS	s_1 Only	s_1 and s_2
50 → 2	36.51	37.32
50 → 1	33.09	31.84

Table 12. Impact of including more terms to $\mathcal{L}_{skew}$.

matching gradients for only a limited number of steps may suffer from the error accumulation problem: alignment error can be amplified over successive condensation [2]. It turns out that the one-step and infinity-step matching proposed in this paper can suppress the issue most effectively. Although, as mentioned in the main paper, generating hundreds of teacher training trajectories is inefficient in both time and memory for slimmable DC, we still provide the results in the 7th and 8th cols. of Tab. 11 regardless of efficiency. The large gap reflects significant error accumulation problem.

C. More Analysis

Sensitivity Analysis of Hyper-Parameters: We provide sensitivity analysis of hyper-parameters on CIFAR10: λ_{pm}^1, λ_{skew}, and λ_{ortho} in Fig. 1(a). We observe that the performance is relatively not sensitive to their values when they are set small. Too large weights may decrease the power of infinity-order parameter matching $\mathcal{L}_{pm}^{\infty}$ and impair the final performance. We also conduct joint analysis for these hyper-parameters in Fig. 1(b), for 20 IPC. The results are insensitive to their values (1.2% gap in maximal) and all closed to RT (67.6%).

Studies on Significance Scores: We provide a visualization of the distribution of top 25 significance scores for 50 IPC of CIFAR10 in Fig. 2 and there is a long-tailed effect. Moreover, if we know the minimal IPC, we can add the corresponding number of components to $\mathcal{L}_{skew}$ instead of merely including s_1; if not, we only regulate s_1 by default. In experiments of Tab. 12, we find that the performance of IPC 2 is improved if we further consider s_2 for$\mathcal{L}_{skew}$ while that of IPC 1 is degraded.

Float Number Budgets: The focus of this paper is mainly on integral IPC for slimmable DC. Nevertheless, there are indeed some engineering tricks for float number budgets. For instance, if the budget is 1.4, we can slim to 2 IPC and then downsample images to 0.84× scale. LBS and LFS performances are 50.60 and 35.38 respectively. Performances of IPC 1 and 2 are shown in Tab. 12 as a reference.

Impact of Significance-Aware Parameterization on Typical DC: We find that the proposed significance-aware parameterization for slimmable DC can also impact typical DC like other synthetic dataset parameterization methods [1, 7]. Here, we try disabling it and provide the results in the "Ours" cols. of Tab. 11 to quantify such impact.

D. More Applications

As supplements to the main manuscript, we provide results of more settings for applications of slimmable DC, *i.e.*, continual learning with a fixed synthetic buffer and federated learning with a dynamic number of participants. On CIFAR100,

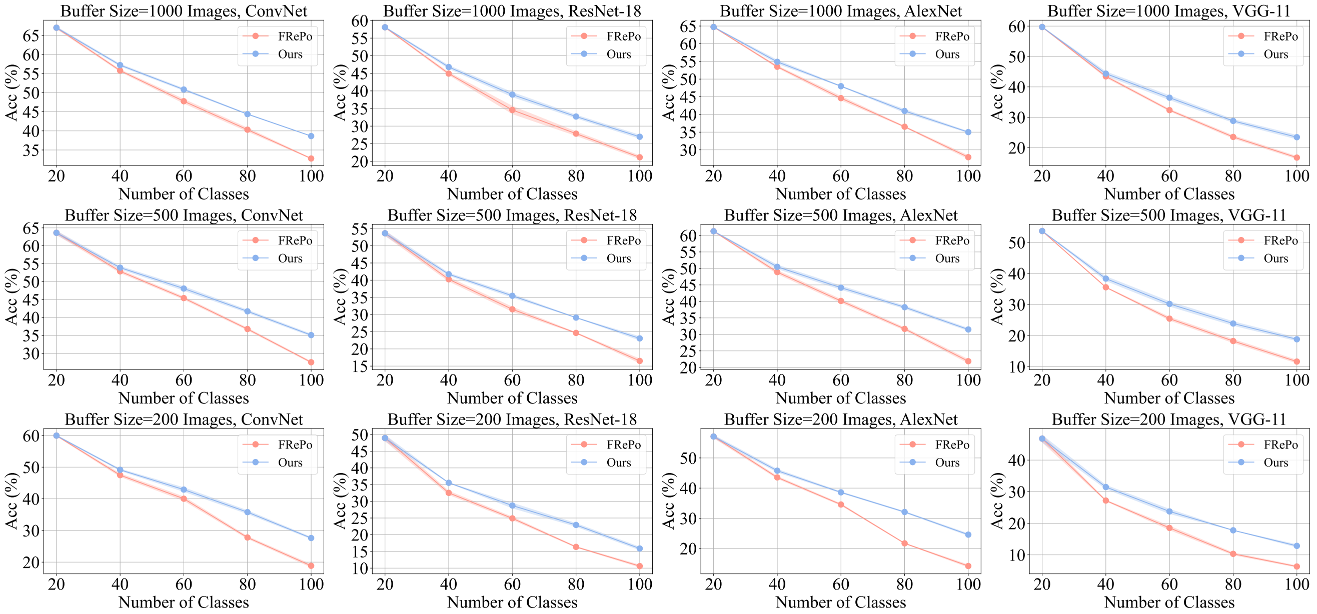


Figure 3. Comparisons with the baseline FRePo on applications of slimmable DC: continual learning using a synthetic buffer with a fixed size, under different buffer sizes and network architectures.

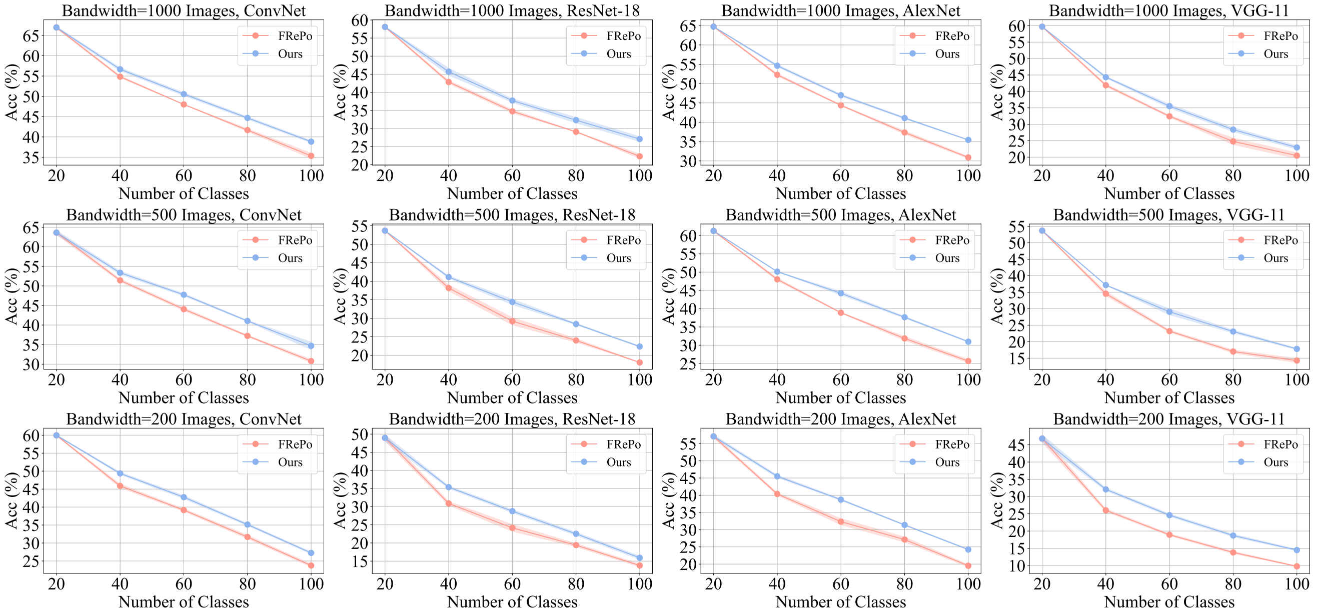


Figure 4. Comparisons with the baseline FRePo on applications of slimmable DC: federated learning with a dynamic number of participant, under different transmission bandwidths and network architectures.

we consider three synthetic buffer sizes / transmission bandwidths: 1,000, 500, and 200 images. Synthetic datasets are trained on CNN with 3 blocks. Beyond the same structure, we also evaluate the performance of synthetic buffers on ResNet-18, AlexNet, and VGG-11. Full results by FRePo [15] and our method are shown in Figs. 3 and 4, with respective to continual learning and federated learning. The conclusion is consistent with that in the main paper.

References

[1] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. *arXiv preprint arXiv:2206.02916*, 2022. 4, 6

[2] Jiawei Du, Yidi Jiang, Vincent TF Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. *arXiv preprint arXiv:2211.11004*, 2022. 6

[3] Fastai. Fastai/imagenette: A smaller subset of 10 easily classified classes from imagenet, and a little more french. 3

[4] 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. *arXiv preprint arXiv:2205.14959*, 2022. 3, 4

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

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

[7] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 4, 6

[8] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 4

[9] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. *arXiv preprint arXiv:2011.00050*, 2020. 4

[10] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *Advances in Neural Information Processing Systems*, 34, 2021. 4

[11] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 3

[12] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *International Conference on Machine Learning*, pages 12674–12685. PMLR, 2021. 3, 4

[13] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. *arXiv preprint arXiv:2110.04181*, 2021. 3, 4

[14] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929*, 2020. 3, 4

[15] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *arXiv preprint arXiv:2206.00719*, 2022. 3, 4, 7