Supplementary Material of DREAM: Efficient Dataset Distillation by Representative Matching

Yanqing Liu [1,2*] Jianyang Gu[1,2*] Kai Wang[1†] Zheng Zhu[3] Wei Jiang[2] Yang You[1‡]
[1]National University of Singapore [2]Zhejiang University [3]Tsinghua University
{yanqing_liu, gu_jianyang, jiangwei_zju}@zju.edu.cn
{kai.wang, youy}@comp.nus.edu.sg zhengzhu@ieee.org
Code: https://github.com/lyq312318224/DREAM

1. Comparisons with More Methods

We compare the distilled dataset performance between DREAM and more methods in Tab. 1. The experiments are conducted under 1, 10, and 50 images-per-class (IPC) settings on MNIST [6], SVHN [10], CIFAR-10, and CIFAR-100 [5] datasets. DREAM achieves SOTA results on most cases. Especially when IPC is small, DREAM has gained significant performance gap over other methods, which also validates the effectiveness of matching representative samples. RFAD [8] employs ConvNet with 1024 convolutional channels, while our results are reported based on 128-channel ConvNet. Except for IPC=1 CIFAR-10, DREAM distills better synthetic images than RFAD. HaBa [7] involves a data hallucination process, which generates more samples from base images. It holds higher performance on IPC=10 CIFAR-10 and IPC=1 CIFAR-100, while in other circumstances, DREAM has superior performances.

2. Accuracy Curve Visualization

We apply the DREAM strategy to more dataset distillation methods, such as DC [18], DSA [16], etc. In addition to stable performance improvements, we visualize the accuracy curve during training in Fig. 1. It can be observed that compared with the original methods, DREAM only requires one fifth and one tenth of the iteration number on the DC and DSA to achieve the original performance, respectively. Further increasing the training time brings continuous performance improvement, which also proves that our method is universal and is able to be easily plugged for popular dataset distillation frameworks. The above experiments are all based on the setting of 10 images-per-class on CIFAR10.

*Equal contribution.
†Project lead.
‡Corresponding author.

3. DREAM on Distribution Matching

In addition to gradient matching, we also explore the applicability of our method in embedding distribution matching. For distribution matching, the optimization is constrained by:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbf{D}\left(\xi(\mathcal{M}_\theta(\mathcal{A}(\mathcal{S}))), \xi(\mathcal{M}_\theta(\mathcal{A}(\mathcal{T})))\right), \quad (1)$$

where ξ represents averaging the features in the channel dimension. For gradient matching, random sampling produces biased matching targets, making the optimization unstable, while for distribution matching, random sampling shifts the average feature. These stochastic factors jointly introduce shifts and reduce training efficiency. DREAM, on the contrary, ensures the evenness and diversity of the original images for matching. It consistently provides appropriate supervision for the optimization. We conduct the experiments on IDC [4] with distribution matching. Under the setting of 10 images-per-class on the CIFAR10 dataset, the original IDC performs much poorer than gradient matching, which is also stated in [4]. Applying DREAM completely reversed this situation by improving the dataset performance by a large margin. Besides, it only takes less than one tenth of the original iteration number to reach the performance, as shown in Fig. 1c.

4. Distilled Dataset Visualization

In order to more intuitively demonstrate the effects on the distilled images, we compare the distillation results of adding the proposed DREAM strategy or not in Fig. 2. DREAM improves the quality of the distilled datasets from two perspectives. Firstly, the images optimized by DREAM show more obvious categorical characteristics. Secondly, DREAM introduces more variety to the distilled images. With these two improvements, dream brings better performance to the distilled datasets.

Table 1: Top-1 accuracy of test models trained on distilled synthetic images on multiple datasets. The distillation training is conducted with ConvNet-3.

Dataset	MNIST			SVHN			CIFAR10			CIFAR100		
	1	10	50	1	10	50	1	10	50	1	10	50
DD [15]	-	79.5	-	-	-	-	-	36.8	-	-	-	-
LD [1]	60.9	87.3	93.3	-	-	-	25.7	38.3	42.5	11.5	-	-
DC [18]	91.7	97.4	98.8	31.2	76.1	82.3	28.3	44.9	53.9	12.8	25.2	-
DSA [16]	88.7	97.8	99.2	27.5	79.2	84.4	28.8	52.1	60.6	13.9	32.3	42.8
DM [17]	89.7	97.5	98.6	-	-	-	26.0	48.9	63.0	11.4	29.7	43.6
CAFE [14]	93.1	97.2	98.6	42.6	75.9	81.3	30.3	46.3	55.5	12.9	27.8	37.9
MTT [2]	-	-	-	-	-	-	46.3	65.3	71.6	24.3	40.1	47.7
IDC [4]	94.2	98.4	99.1	68.5	87.5	90.1	50.6	67.5	74.5	-	45.1	-
KIP [11, 12]	90.1	97.5	98.3	57.3	75.0	80.5	49.9	62.7	68.6	15.7	28.3	-
RFAD [8]	94.4	98.5	98.8	52.2	74.9	80.9	**53.6**	66.3	71.1	26.3	33.0	-
HaBa [7]	92.4	97.4	98.1	69.8	83.2	88.3	48.3	**69.9**	74.0	**33.4**	40.2	47.0
FRePo [19]	93.0	98.6	99.2	-	-	-	46.8	65.5	71.7	28.7	42.5	44.3
DREAM	**95.7**	**98.6**	**99.2**	**69.8**	**87.9**	**90.5**	51.1	69.4	**74.8**	29.5	**46.8**	**52.6**

Table 2: Time cost of adding DREAM strategy (s).

Datasets	Methods	Clustering	Update Images	Inner Loop
CIFAR-10	IDC [4]	-	0.2	0.2
	DREAM	0.1	0.2	0.3
CIFAR-100	IDC [4]	-	2.0	2.0
	DREAM	0.1	2.0	2.1

We provide some extra visualizations of the distilled images on MNIST, FashionMNIST, SVHN, CIFAR-10, CIFAR-100 and TinyImageNet in Fig. 3.

5. Differences from Related Works

There are some recent works focusing on improving the efficiency of dataset distillation. RFAD reduces the calculation of neural tangent kernel matrix in Kernel Inducing Points (KIP) from $O(|S^2|)$ to $O(|S|)$ by using random feature approximation [8]. It takes into account the similarity between Neural Tangent Kernel (NTK) and Neural Network Gaussian Process (NNGP) kernels. RFAD focuses on reducing the calculation complexity in KIP, while our proposed DREAM is aiming at improving the training efficiency through selecting representative original images, which has no contradicts.

Jiang et al. analyze the shortcomings of gradient matching method and propose the idea of matching multi-level gradients from the angle perspective [3]. There are also many other methods [9, 13], which analyze the shortcomings of existing methods from the perspective of two-level optimization and improve the efficiency. Comparatively, DREAM addresses the training efficiency problem from the sampling perspective for both gradient matching and embedding distribution matching. DREAM is able to be easily plugged into other dataset distillation methods to significantly reduce the required training iterations.

6. Clustering Analysis

We further analyze the extra time cost caused by the clustering process in Tab. 2. For CIFAR-10, in each inner loop, the matching process and image updating cost 0.2s. Every 10 inner loops, a clustering process is conducted. The clustering process takes 1s, so by average the clustering time for each inner loop is 0.1s. The total average inner loop time is 0.3s, compared to the original 0.2s. Considering that we only need one tenth to one fifth of the iterations to obtain the original performance, we save more than 70% of the time. For CIFAR-100 with more classes, the extra clustering time is one twentieth of the original image updating time, which is negligible. DREAM significantly improves the training efficiency and reduces the required training time for dataset distillation.

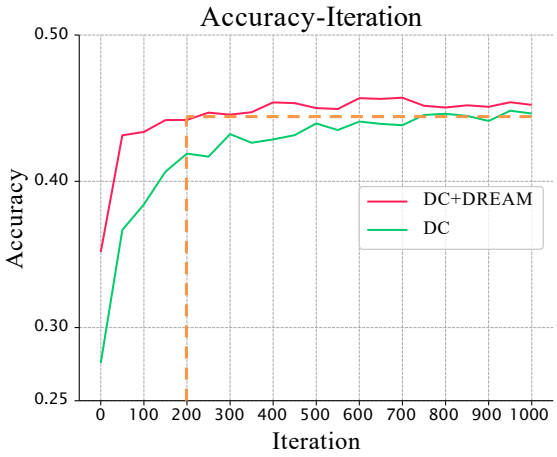
References

[1] Ondrej Bohdal, Yongxin Yang, and Timothy Hospedales. Flexible dataset distillation: Learn labels instead of images. *arXiv preprint arXiv:2006.08572*, 2020. 2

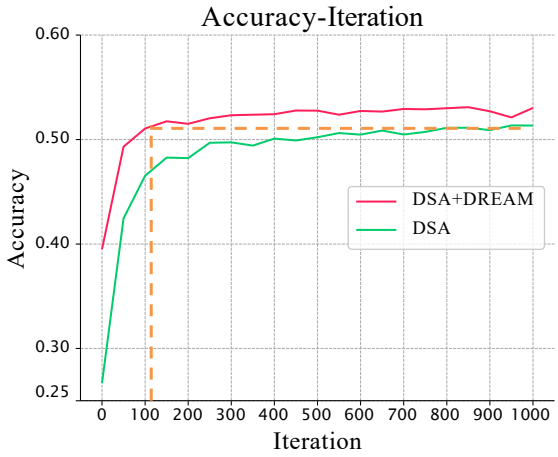
[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, pages 4750–4759, 2022. 2

[3] Zixuan Jiang, Jiaqi Gu, Mingjie Liu, and David Z Pan. Delving into effective gradient matching for dataset condensation. *arXiv preprint arXiv:2208.00311*, 2022. 2

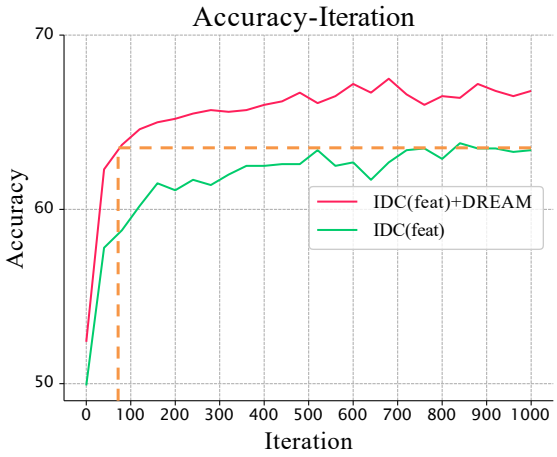
[4] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and



(a) The accuracy curve of adding DREAM to DC.



(b) The accuracy curve of adding DREAM to DSA.



(c) The accuracy curve of adding DREAM to distribution matching.

Figure 1: Applying the DREAM strategy brings stable performance and efficiency improvements.

Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. *arXiv preprint arXiv:2205.14959*, 2022. 1, 2

[5] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 1

[6] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1

[7] Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. Dataset distillation via factorization. In *NeurIPS*, 2022. 1, 2

[8] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *NeurIPS*, 2022. 1, 2

[9] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pages 1540–1552. PMLR, 2020. 2

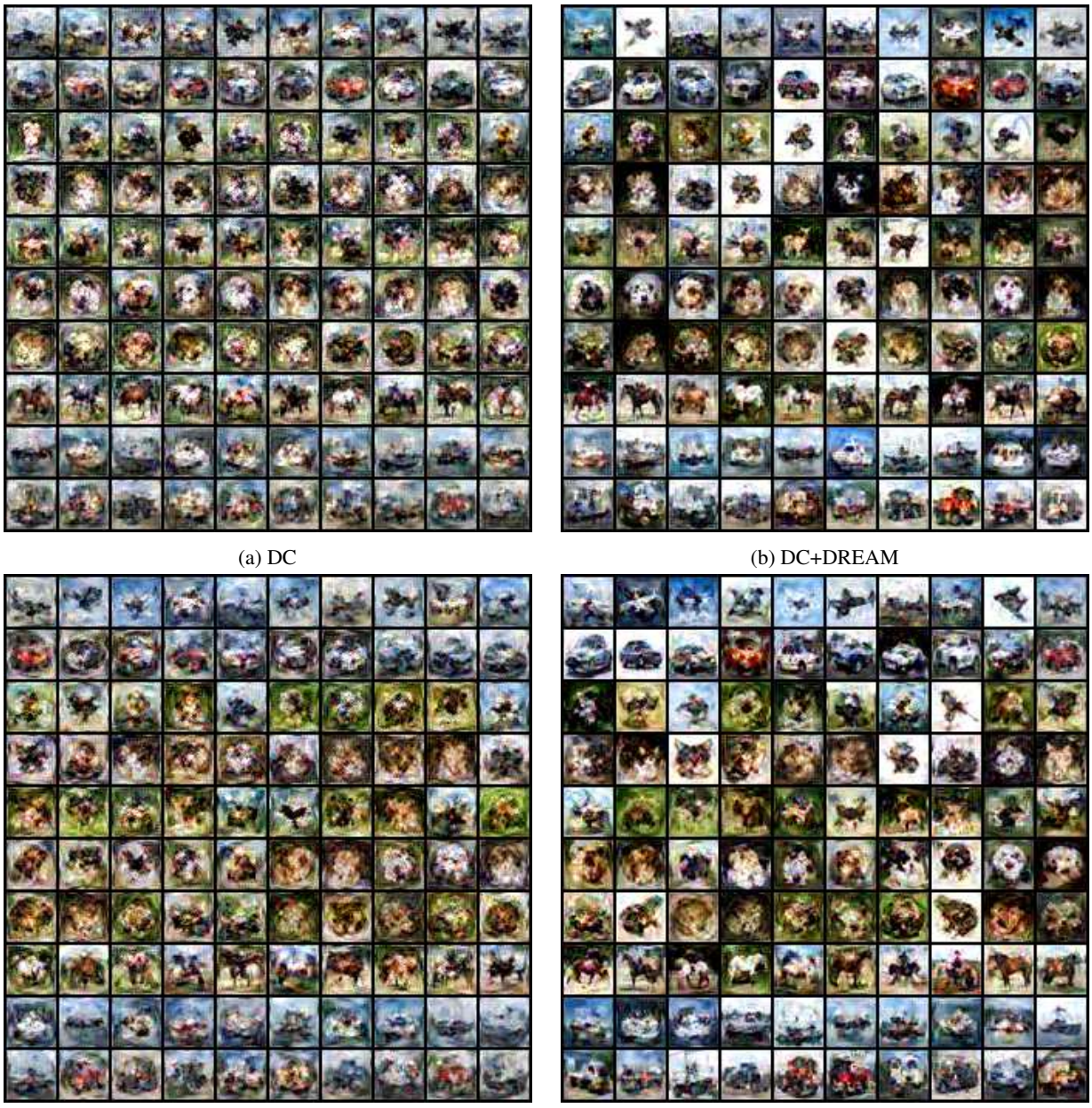
[10] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 1

[11] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *ICLR*, 2020. 2

[12] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *NeurIPS*, 34:5186–5198, 2021. 2

[13] Paul Vicol, Jonathan P Lorraine, Fabian Pedregosa, David Duvenaud, and Roger B Grosse. On implicit bias in overparameterized bilevel optimization. In *International Conference on Machine Learning*, pages 22234–22259. PMLR, 2022. 2

[14] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and



(a) DC (b) DC+DREAM

(c) DSA (d) DSA+DREAM

Figure 2: Applying DREAM improves the image quality and sample diversity.

Yang You. Cafe: Learning to condense dataset by aligning features. In *CVPR*, pages 12196–12205, 2022. 2

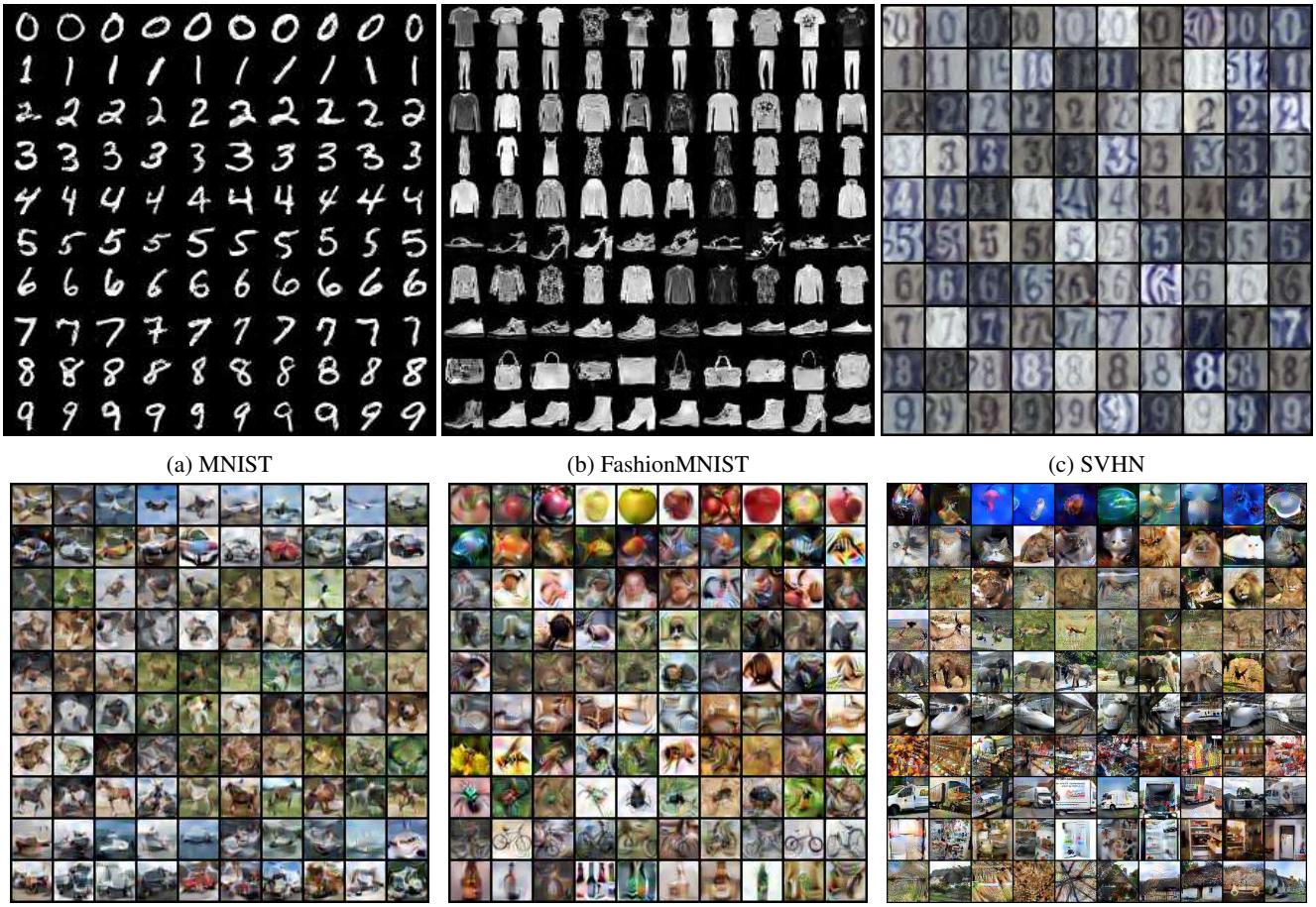
[15] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 2

[16] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, pages 12674–12685. PMLR, 2021. 1, 2

[17] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. *arXiv preprint arXiv:2110.04181*, 2021. 2

[18] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2020. 1, 2

[19] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *NeurIPS*, 2022. 2



(a) MNIST (b) FashionMNIST (c) SVHN

(d) CIFAR-10 (e) CIFAR-100 (f) TinyImageNet

Figure 3: Example visualizations of the distilled images on MNIST, FashionMNIST, SVHN, CIFAR-10, CIFAR-100 and TinyImageNet.