DataDAM: Efficient Dataset Distillation with Attention Matching

Ahmad Sajedi *[1] Samir Khaki *[1] Ehsan Amjadian [2,3] Lucy Z. Liu [2] Yuri A. Lawryshyn [1] Konstantinos N. Plataniotis [1]

[1]University of Toronto, Canada [2]Royal Bank of Canada (RBC), Canada [3]University of Waterloo, Canada

* Equal Contribution

ICCV23 PARIS
International Conference on Computer Vision
October 2 - 6, 2023

Introduction and Motivation

Dataset Distillation

Synthesizing a small but informative dataset $\mathcal{S}$ that has competitive performance to the original large training dataset $\mathcal{T}$.

Applications

- ❖ Replay exemplars in continual learning (CL)
- ❖ Accelerating neural architecture search (NAS)
- ❖ Privacy protection in federated learning
- ❖ Membership inference defense

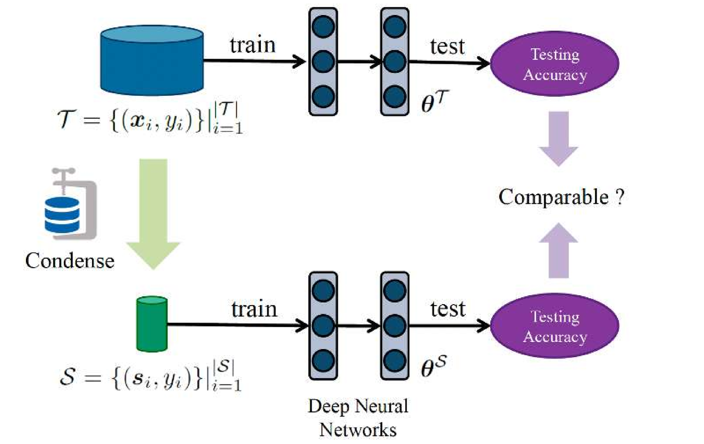


Figure 1. Figure by Zhao et al., Dataset Condensation with Gradient Matching, ICLR, 2021.

Motivation

- ❖ Dataset distillation algorithms typically suffer from expensive computational costs.
- ❖ Large time consumption hinders scalability to high-quality and large-scale datasets.
- ❖ Overfitting to biased samples during dataset distillation procedures (biased data distribution).
- ❖ There still exists a substantial performance gap between models trained on condensed synthetic sets and those trained on the whole dataset.

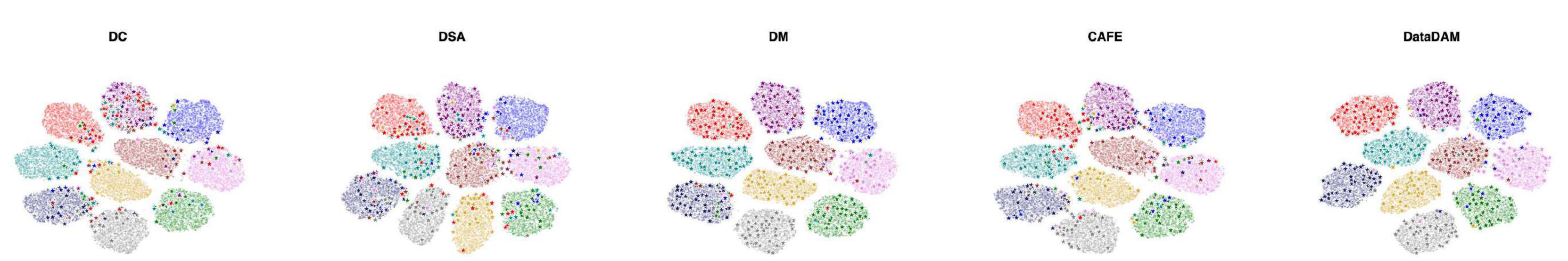


Figure 2. Data distributions of the synthetic images learned by prior methods on the CIFAR10 dataset with IPC 50. The stars represent the synthetic data dispersed amongst the original dataset.

Research Question

This study aims to answer the following research question: *Can we develop a simple yet effective data distillation algorithm capable of learning unbiased samples for any training datasets, regardless of their resolution and scale?*

Contributions

Keeping the research question in mind, we introduce a novel dataset distillation framework designed to overcome existing limitations, facilitating fast and data-efficient learning for visual classification tasks. Our contributions can be summarized as follows:

- ❖ We propose a simple method, **dataset distillation with attention matching (DataDAM)** to effectively approximate the distribution of the real dataset. This is achieved by matching the **spatial attention maps** of real and synthetic data generated by different layers within a family of randomly initialized neural networks.
- ❖ We evaluate **DataDAM** on computer vision datasets with **low, medium, and high resolutions**, where it achieves state-of-the-art results across multiple benchmark settings. Our approach also enables cross-architecture generalizations.
- ❖ We illustrate that **DataDAM** offers up to a **100x** reduction in run time costs while maintaining the lowest GPU memory consumption. Our approach also enables cross-architecture generalizations.
- ❖ We show that **DataDAM** can enhance downstream applications by improving memory efficiency for **continual learning** and accelerating **neural architecture search** through a more representative proxy dataset.

Dataset Distillation with Attention Matching (DataDAM)

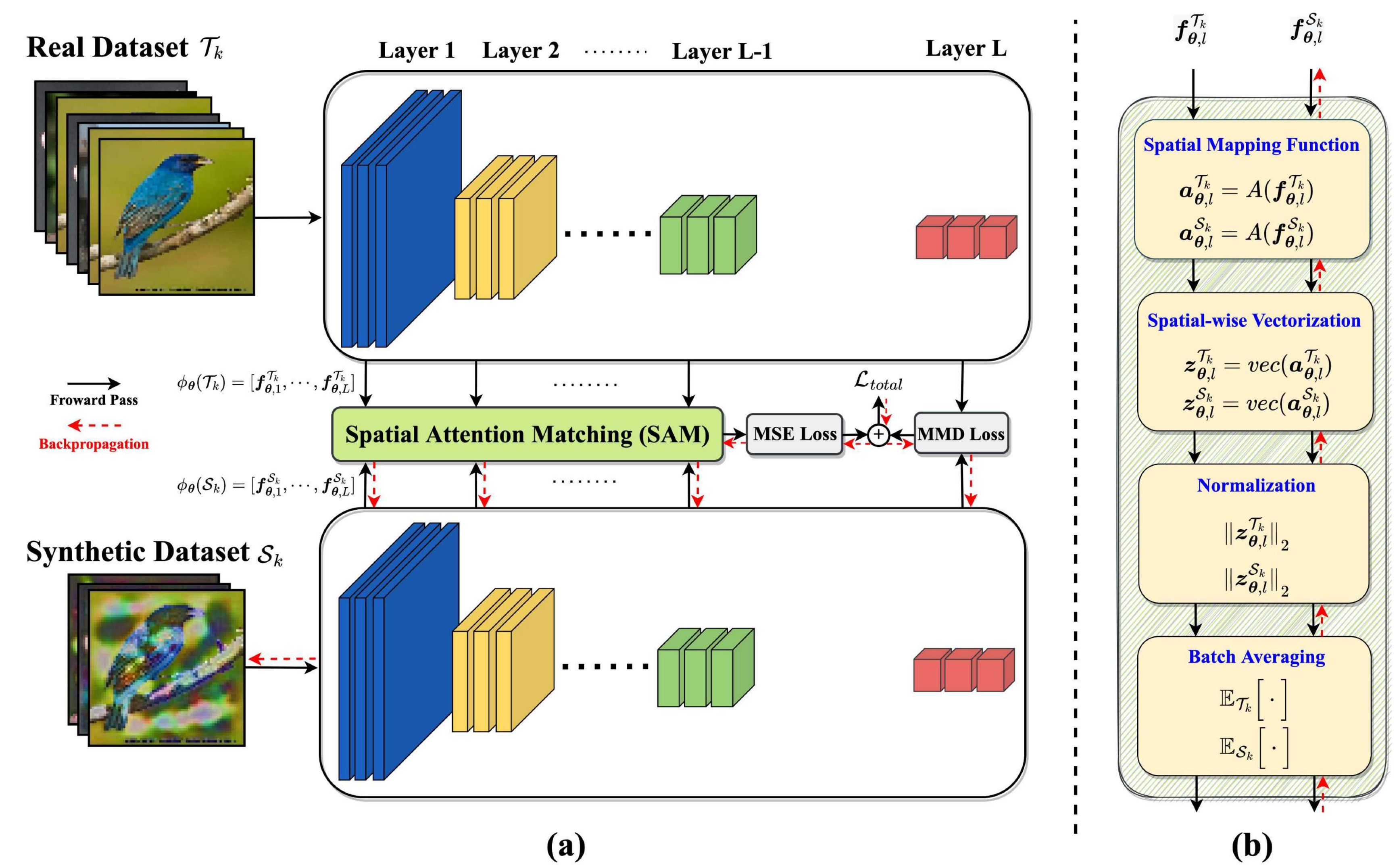


Figure 3. (a) Illustration of the DataDAM method. DataDAM includes a Spatial Attention Matching (SAM) module and a complementary MMD loss to capture the dataset's distribution. (b) The internal architecture of the SAM module.

Overall Performances on Benchmark Datasets

	IPC	Coreset Selection		Training Set Synthesis								Whole Dataset
		Random	K-Center	DD	DC	DSA	DM	CAFE	KIP	MTT	DataDAM	
CIFAR-10	1	$14.4_{\pm2.0}$	$21.5_{\pm1.3}$	-	$28.3_{\pm0.5}$	$28.8_{\pm0.7}$	$26.0_{\pm0.8}$	$31.6_{\pm0.8}$	$29.8_{\pm1.0}$	$31.9_{\pm1.2}$	$\mathbf{32.0_{\pm1.2}}$	$84.8_{\pm0.1}$
	10	$26.0_{\pm1.2}$	$14.7_{\pm0.9}$	$36.8_{\pm1.2}$	$44.9_{\pm0.5}$	$52.1_{\pm0.5}$	$48.9_{\pm0.6}$	$50.9_{\pm0.5}$	$46.1_{\pm0.7}$	$\mathbf{56.4_{\pm0.7}}$	$54.2_{\pm0.8}$	
	50	$43.4_{\pm1.0}$	$27.0_{\pm1.4}$	-	$53.9_{\pm0.5}$	$60.6_{\pm0.5}$	$63.0_{\pm0.4}$	$62.3_{\pm0.4}$	$53.2_{\pm0.7}$	$65.9_{\pm0.6}$	$\mathbf{67.0_{\pm0.4}}$	
CIFAR-100	1	$4.2_{\pm0.3}$	$8.4_{\pm0.3}$	-	$12.8_{\pm0.3}$	$13.9_{\pm0.3}$	$11.4_{\pm0.3}$	$14.0_{\pm0.3}$	$12.0_{\pm0.2}$	$13.8_{\pm0.6}$	$\mathbf{14.5_{\pm0.5}}$	$56.2_{\pm0.3}$
	10	$14.6_{\pm0.5}$	$17.3_{\pm0.3}$	-	$25.2_{\pm0.3}$	$32.3_{\pm0.3}$	$29.7_{\pm0.3}$	$31.5_{\pm0.2}$	$29.0_{\pm0.3}$	$33.1_{\pm0.4}$	$\mathbf{34.8_{\pm0.5}}$	
	50	$30.0_{\pm0.4}$	$30.5_{\pm0.3}$	-	$30.6_{\pm0.6}$	$42.8_{\pm0.4}$	$43.6_{\pm0.4}$	$42.9_{\pm0.2}$	-	$42.9_{\pm0.3}$	$\mathbf{49.4_{\pm0.3}}$	
Tiny ImageNet	1	$1.4_{\pm0.1}$	$1.6_{\pm0.1}$	-	$5.3_{\pm0.1}$	$5.7_{\pm0.1}$	$3.9_{\pm0.2}$	-	-	$6.2_{\pm0.4}$	$\mathbf{8.3_{\pm0.4}}$	$37.6_{\pm0.4}$
	10	$5.0_{\pm0.2}$	$5.1_{\pm0.2}$	-	$12.9_{\pm0.1}$	$16.3_{\pm0.2}$	$12.9_{\pm0.4}$	-	-	$17.3_{\pm0.2}$	$\mathbf{18.7_{\pm0.3}}$	
	50	$15.0_{\pm0.4}$	$15.0_{\pm0.3}$	-	$12.7_{\pm0.4}$	$5.1_{\pm0.2}$	$25.3_{\pm0.2}$	-	-	$26.5_{\pm0.3}$	$\mathbf{28.7_{\pm0.3}}$	

Table 1. The testing accuracy % comparison to state-of-the-art methods for low- and medium-resolution datasets.

	IPC	Random	DM	DataDAM	Whole Dataset
ImageNet-1K	1	$0.5_{\pm0.1}$	$1.3_{\pm0.1}$	$\mathbf{2.0_{\pm0.1}}$	$33.8_{\pm0.3}$
	2	$0.9_{\pm0.1}$	$1.6_{\pm0.1}$	$\mathbf{2.2_{\pm0.1}}$	
	10	$3.1_{\pm0.2}$	$5.7_{\pm0.1}$	$\mathbf{6.3_{\pm0.0}}$	
	50	$7.6_{\pm1.2}$	$11.4_{\pm0.9}$	$\mathbf{15.5_{\pm0.2}}$	
ImageNette	1	$23.5_{\pm4.8}$	$32.8_{\pm0.5}$	$\mathbf{34.7_{\pm0.9}}$	$87.4_{\pm1.0}$
	10	$47.7_{\pm2.4}$	$58.1_{\pm0.3}$	$\mathbf{59.4_{\pm0.4}}$	
ImageWoof	1	$14.2_{\pm0.9}$	$21.1_{\pm1.2}$	$\mathbf{24.2_{\pm0.5}}$	$67.0_{\pm1.3}$
	10	$27.0_{\pm1.9}$	$31.4_{\pm0.5}$	$\mathbf{34.4_{\pm0.4}}$	
ImageSquawk	1	$21.8_{\pm0.5}$	$31.2_{\pm0.7}$	$\mathbf{36.4_{\pm0.8}}$	$87.5_{\pm0.3}$
	10	$40.2_{\pm0.4}$	$50.4_{\pm1.2}$	$\mathbf{55.4_{\pm0.9}}$	

Table 2. The performance (testing accuracy %) comparison to state-of-the-art methods for large-scale and high-resolution computer vision datasets.

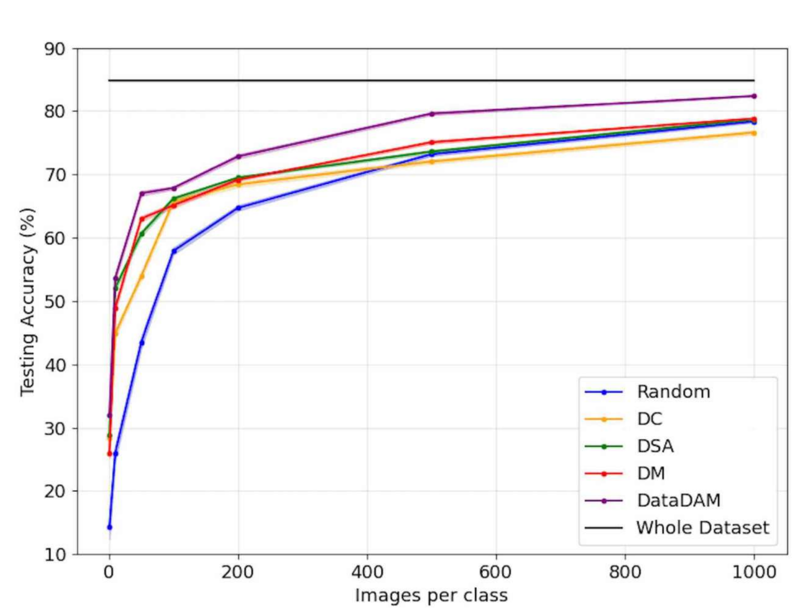


Figure 4. The testing accuracy % comparison with state-of-the-art methods on the CIFAR10 dataset for varying numbers of images per class (IPCs).

Cross-architecture Performances & Computational Cost

	T\E	AlexNet	VGG-11	ResNet-18
DC	ConvNet	$28.8_{\pm0.7}$	$38.8_{\pm1.1}$	$20.9_{\pm1.0}$
CAFE	ConvNet	$43.2_{\pm0.4}$	$48.8_{\pm0.5}$	$43.3_{\pm0.7}$
DSA	ConvNet	$53.7_{\pm0.6}$	$51.4_{\pm1.0}$	$47.8_{\pm0.9}$
DM	ConvNet	$60.1_{\pm0.5}$	$57.4_{\pm0.8}$	$52.9_{\pm0.4}$
MTT	ConvNet	$43.9_{\pm0.9}$	$48.7_{\pm1.3}$	$60.0_{\pm0.7}$
DataDAM	ConvNet	$\mathbf{63.9_{\pm0.9}}$	$\mathbf{64.8_{\pm0.5}}$	$\mathbf{60.2_{\pm0.7}}$

Table 3. Cross-architecture testing performance (%) on CIFAR10 with 50 images per class.

Method	run time(sec)			GPU memory(MB)		
	IPC1	IPC10	IPC50	IPC1	IPC10	IPC50
DC	$0.16_{\pm0.0}$	$3.31_{\pm0.0}$	$15.74_{\pm0.1}$	3515	3621	4527
DSA	$0.22_{\pm0.0}$	$4.47_{\pm0.1}$	$20.13_{\pm0.6}$	3513	3639	4539
DM	$\mathbf{0.08_{\pm0.0}}$	$\mathbf{0.08_{\pm0.0}}$	$\mathbf{0.08_{\pm0.0}}$	3323	3455	**3605**
MTT	$0.36_{\pm0.2}$	$0.40_{\pm0.2}$	OOM	**2711**	8049	OOM
DataDAM	$\mathbf{0.09_{\pm0.0}}$	$\mathbf{0.08_{\pm0.0}}$	$0.16_{\pm0.0}$	3452	**3561**	3724

Table 4. Training time and GPU memory comparisons for state-of-the-art synthesis methods. Run time is expressed per step, averaged over 100 iterations.

Applications

DataDAM has the potential to significantly boost several downstream applications, such as enhancing memory efficiency for *continual learning* and expediting *neural architecture search* by utilizing a more representative proxy dataset.

❖ Continual Learning:

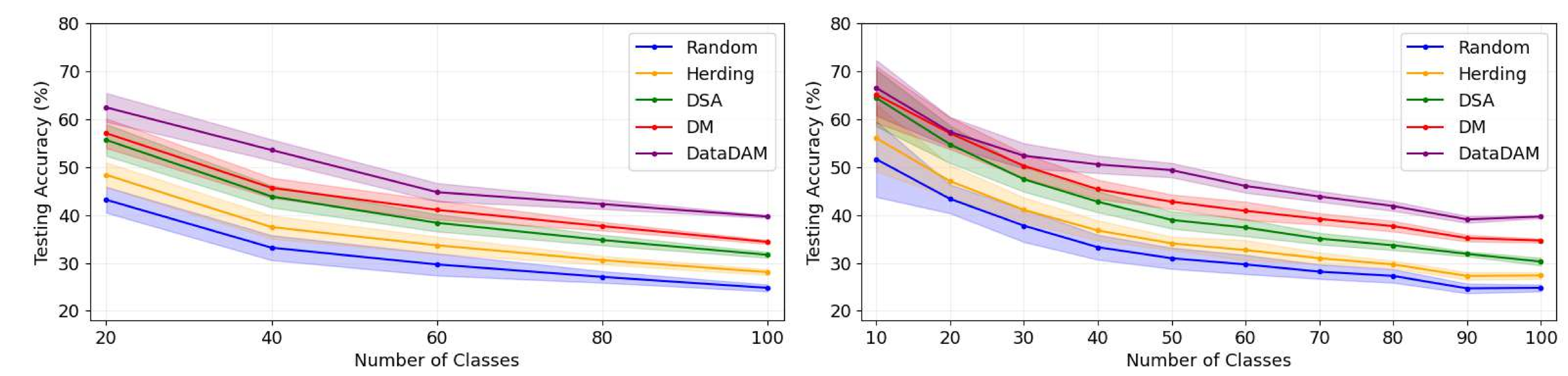


Figure 5. (Left): Showcases 5-step and (Right): Showcases 10-step continual learning with tolerance region.

❖ Neural Architecture Search:

	Random	DM	CAFE	Ours	Early-stopping	Whole Dataset
Performance	88.9	87.2	83.6	**89.0**	88.9	89.2
Correlation	0.70	0.71	0.59	**0.72**	0.69	1.00
Time cost (min)	206.4	206.6	206.4	206.4	206.2	5168.9
Storage (imgs)	**500**	**500**	**500**	**500**	5×10^4	5×10^4

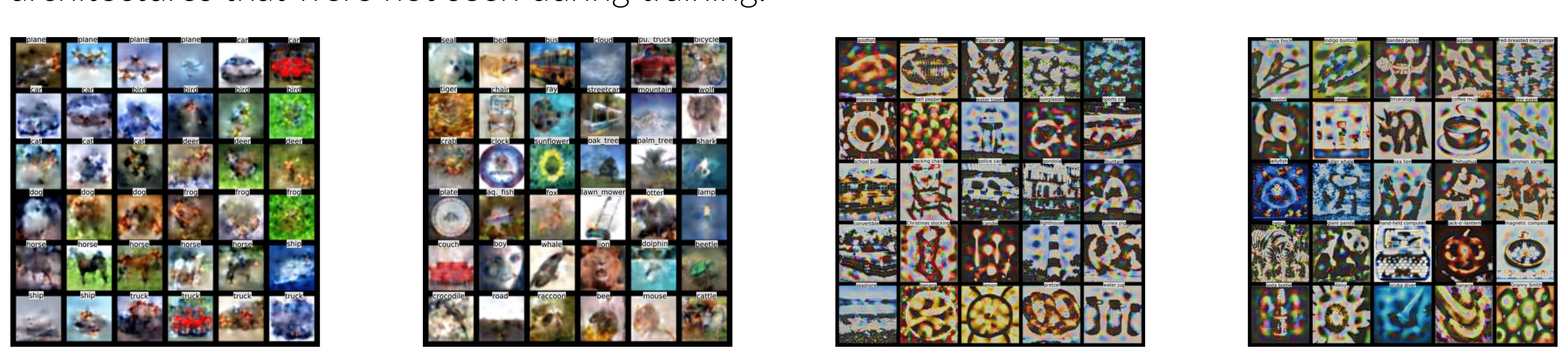
Table 5. Neural architecture search on CIFAR10 dataset with a search space of the whole sample space.

	Random	DM	CAFE	Ours	Early-stopping	Whole Dataset
Performance	88.9	87.2	83.6	**89.0**	88.9	89.2
Correlation	0.44	0.51	0.36	**0.69**	0.64	1.00
Time cost (min)	33.0	32.2	**30.7**	34.8	37.1	5168.9
Storage (imgs)	**500**	**500**	**500**	**500**	5×10^4	5×10^4

Table 6. Neural architecture search on CIFAR10 with a search space of the top 20% of the sample space.

Distilled Image Visulization

The distilled images generated by DataDAM look real and are well-suited to be used with a variety of architectures that were not seen during training.



((a)) CIFAR10 ((b)) CIFAR100 ((c)) Tiny ImageNet ((d)) ImageNet-1K

Figure 6. Example distilled images from 32x32 CIFAR10/100 (IPC10), 64x64 Tiny ImageNet (IPC1), and 64x64 ImageNet-1K (IPC1) datasets.