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

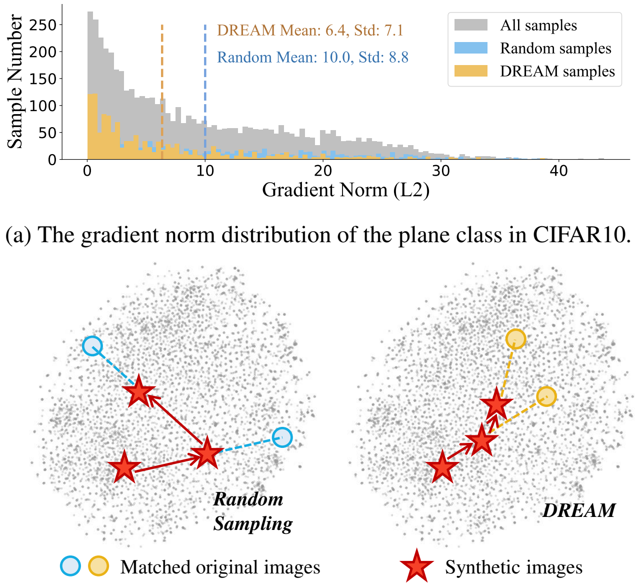
Code: https://github.com/lyq312318224/DREAM

Abstract

*Dataset distillation aims to synthesize small datasets with little information loss from original large-scale ones for reducing storage and training costs. Recent state-of-the-art methods mainly constrain the sample synthesis process by matching synthetic images and the original ones regarding gradients, embedding distributions, or training trajectories. Although there are various matching objectives, currently the strategy for selecting original images is limited to naive random sampling. We argue that random sampling overlooks the evenness of the selected sample distribution, which may result in noisy or biased matching targets. Besides, the sample diversity is also not constrained by random sampling. These factors together lead to optimization instability in the distilling process and degrade the training efficiency. Accordingly, we propose a novel matching strategy named as **D**ataset distillation by **RE**present**A**tive **M**atching (DREAM), where only representative original images are selected for matching. DREAM is able to be easily plugged into popular dataset distillation frameworks and reduce the distilling iterations by more than 8 times without performance drop. Given sufficient training time, DREAM further provides significant improvements and achieves state-of-the-art performances.*

1. Introduction

Deep learning has made remarkable achievements in the computer vision society [20, 11, 36, 29, 43, 16, 7, 55], and the success is closely related to a large amount of efforts in data collection and annotation. But along with the progress of these efforts, the huge amount of data, in turn, becomes a barrier to both storage and training [52, 19]. Many methods are introduced to reduce the scale of datasets [47, 40, 34, 8]. Among these, dataset distillation, aiming at condensing large-scale datasets into smaller ones with little information loss, has become a hot topic to tackle the problem of data burden [3, 46, 24, 5, 12].



(a) The gradient norm distribution of the plane class in CIFAR10.

(b) The migration of synthetic samples during training.

Figure 1: Samples on the decision boundaries usually provide larger gradients, which biases the gradient matching optimization. Random sampling (left) overlooks the evenness of of the selected sample distribution, resulting in unstable optimization process of the synthesized samples. By only matching with proper gradients from representative original samples, our proposed DREAM (right) greatly improves the training efficiency of dataset distillation tasks. Best viewed in color.

Dataset distillation methods are roughly divided into two categories: coreset-based and optimization-based. Coreset-based method employ certain metrics to heuristically select samples for representing the original dataset [26, 41]. How-

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

ever, it is difficult to rely on a small proportion of original samples to contain the information of the whole dataset, resulting in low compression rate. Optimization-based methods alleviate the defect by incorporating image synthesis to introduce more information into single images [47]. Specifically, these methods initialize a small amount of learnable image tensors and update them through matching the training gradients [52, 24], embedding distributions [51, 46] or training trajectories [3, 12] with the original images.

Although the optimization-based methods achieve considerable performance as well as compression ratio, the distillation process itself still requires a large amount of time. We analyze the problem from the strategy of selecting original images for matching, which is mostly set as random sampling in previous works [52]. We argue that random sampling overlooks the evenness of the selected sample distribution. On the one hand, the matching optimization may be overly prone to certain samples with dominant matching targets, such as boundary samples with larger training gradients [46]. On the other hand, the sample diversity inside a mini-batch is also not constrained, leading to potential information insufficiency. These factors together result in optimization instability of the dataset distillation process, and degrade the training efficiency.

Accordingly, we propose a novel matching strategy named as **D**ataset distillation by **RE**present**A**tive **M**atching (DREAM) to address the aforementioned training efficiency issue. Specifically, a clustering process inside each class is conducted at intervals to generate sub-clusters reflecting the sample distribution. The sub-cluster centers, which not only are representative for surrounding samples, but also evenly cover the whole class distribution, are selected for matching. As shown in Fig. 1a, the gradient distribution of the selected samples contains less variation. By only matching with representative samples, DREAM largely reduces the instability during training, and provides a smoother and more robust distillation process. For the synthetic image initialization, we adopt a similar clustering-based strategy, where the center sample is selected from each sub-cluster, which further accelerates the training process.

DREAM can be easily plugged into popular dataset distillation frameworks. Compared with commonly adopted random matching, DREAM significantly improves the training efficiency in the distilling process. We conduct extensive experiments to validate that it only takes less than one eighth of the iterations for DREAM to obtain comparable performance with the baseline methods. In addition, given sufficient training iterations, DREAM further boosts the performance to surpass other state-of-the-art methods.

Our main contributions are summarized as:

- We analyze the training efficiency of optimization-based dataset distillation from the strategy of selecting original samples for matching.
- We propose a Dataset distillation by REpresentAtive Matching (DREAM) strategy. By only matching representative images, DREAM accelerates the training process by more than 8× without performance drop.
- DREAM is able to be easily plugged into a variety of dataset distillation frameworks. Extensive experiments prove that DREAM consistently improves the performance of the distilled dataset.

2. Related Works

2.1. Dataset Distillation

Dataset distillation can be roughly divided into 2 categories: coreset-based and optimization-based.

Coreset-based methods select a certain proportion of data based on certain metrics [17, 4]. Lapedriza *et al.* measure the importance of the sample by the benefits obtained from training the model on the sample [26]. Toneva *et al.* find that samples have different forgetting characteristics and the easily forgotten samples have larger information amount [41]. Coresets are also utilized to solve continual learning [35, 1, 48] and active learning tasks [38]. Besides, Shleifer *et al.* accelerate the search of neural network architecture by selecting a group of "easier" samples [39]. Although coreset-based methods are practical to apply, it is hard to obtain rich information from a small amount of original samples. Therefore, coreset-based methods are restricted from further reducing the compression ratio.

Optimization-based methods implement dataset distillation by synthesizing image samples constrained by various optimization targets. Wang *et al.* raise the dataset distillation concept from the optimization aspect, and update the synthetic images in a meta-learning style [47]. Multiple works are then proposed to constrain the image generation by matching training gradients [52, 50, 22], embedding distributions [51, 46] and training trajectories [3] with original images. IDC injects more information into synthetic samples under the limit of fixed storage size [24]. Nguyen *et al.* build up a distributed meta-learning framework and incorporate the kernel approximation methods [32]. RFAD speeds up the computation by introducing a random feature approximation [30]. HaBa employs data hallucination networks to construct base images and improves the representation capability of distilled datasets [28]. FRePo introduces an efficient meta-gradient computation method and a "model pool" to alleviate the overfitting [56]. DiM [45] transfers knowledge by distilling datasets into generative models. Optimization-based methods largely improve the compression ratio via fusing more information into synthetic images. However, recent state-of-the-art methods require a large number of iterations to obtain desired validation accuracy, indicating low training efficiency. In this work, we focus on designing a novel matching strategy for more efficient dataset distillation training.

2.2. Clustering

Clustering divides samples into groups in an unsupervised manner [37]. K-means [14, 2] specifies the number of target clusters, and optimizes the partition to obtain clusters with similar sizes [18]. DBSCAN, based on density, does not require the number of target clusters in advance. The clusters are formed by gradually adding data points within the tolerance range. [13]. It is applicable to dataset of any shape, yet the size of the generated clusters is unstable, outliers are excluded from clusters, and close clusters may be merged. Hierarchical clustering methods include Agglomerative and Divisive. The former fuses multiple clusters until a certain condition is met, and the latter divides a cluster through segmentation [10].

3. Method

Aiming at addressing the training efficiency problem for dataset distillation tasks, we propose a novel Dataset distillation by REpresentAtive Matching (DREAM) strategy. By only matching the representative original images, DREAM reduces the optimization instability, and achieves a smoother and more robust training process. In this section, we orderly introduce the basic training schemes of dataset distillation, our observations on the training efficiency and the detailed design of DREAM.

3.1. Preliminaries

Given a large-scale dataset $\mathcal{T} = \{(\boldsymbol{x}_t^i, y_t^i)\}_{i=1}^{|\mathcal{T}|}$, the target of dataset distillation is generating a small surrogate dataset $\mathcal{S} = \{(\boldsymbol{x}_s^i, y_s^i)\}_{i=1}^{|\mathcal{S}|}$ with as little information loss as possible, where $|\mathcal{S}| \ll |\mathcal{T}|$. The information loss is usually measured by the performance drop between training a model with the original images $\mathcal{T}$ and the surrogate set $\mathcal{S}$. The commonly adopted optimization-based methods follow a synthetic pipeline. The surrogate set $\mathcal{S}$ is first initialized with random original images from $\mathcal{T}$. Under the constraints of matching objectives $\phi(\cdot)$, the synthetic images are updated to mimicking the distribution of the original images, which is formulated as:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbf{D}\left(\phi(\mathcal{S}), \phi(\mathcal{T})\right), \tag{1}$$

where $\mathbf{D}$ is the matching metric. Typically, we select the training gradients as the matching objective $\phi(\cdot)$. Given a random model $\mathcal{M}_\theta$ with training parameters θ, $\mathcal{S}$ is supposed to give similar gradients to $\mathcal{T}$ throughout the training process of $\mathcal{M}_\theta$, that is:

$$\mathcal{S}^* = \arg\min_{\mathcal{S}} \mathbf{D}\left(\nabla_\theta \mathcal{L}(\mathcal{M}_\theta(\mathcal{A}(\mathcal{S}))), \nabla_\theta \mathcal{L}(\mathcal{M}_\theta(\mathcal{A}(\mathcal{T})))\right), \tag{2}$$

where $\mathcal{L}(\cdot,\cdot)$ is the training loss, and $\mathcal{A}$ is the differentiable augmentation [23, 53, 42, 54]. Practically, the matching objectives are calculated on the synthetic images and a mini-batch of original images $\{(\boldsymbol{x}_t^i, y_t^i)\}_{i=1}^{N}$ sampled from $\mathcal{T}$ with the same class labels. The objective matching and $\mathcal{M}_\theta$ training is conducted alternatively, such that gradients at different training stages are matched, which forms the inner optimization loop. The inner loop is iterated with different random $\mathcal{M}_\theta$ for more varied matching gradients.

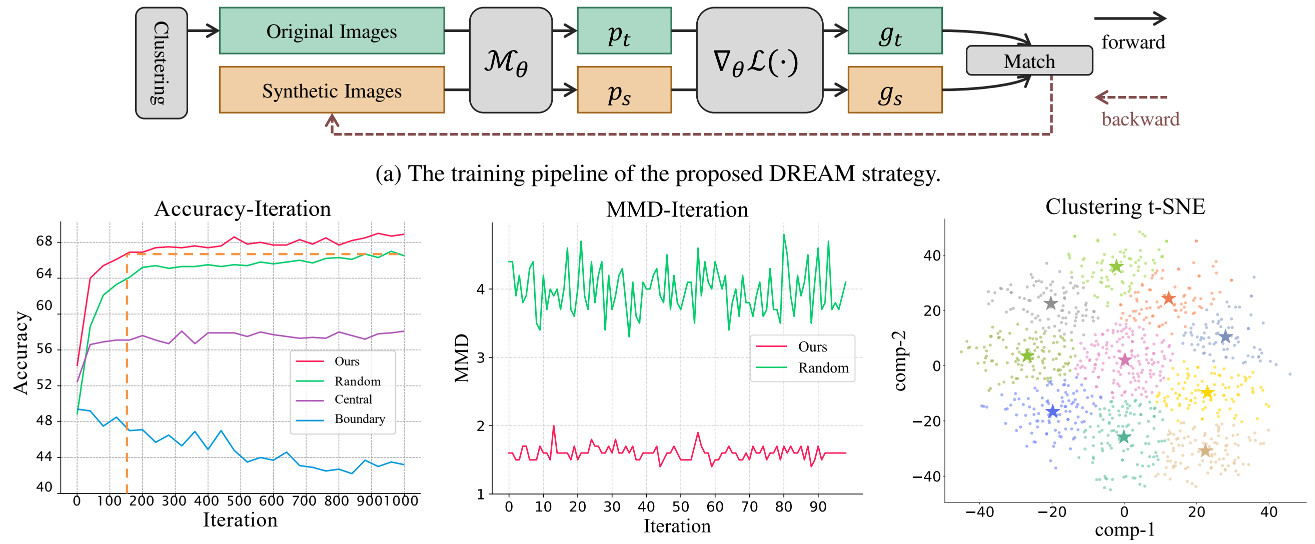
Recent literature offers various matching objectives and achieves significant testing accuracy via training in the small synthetic dataset [46, 3, 24]. However, the distillation process itself still requires a large amount of training time, indicating low training efficiency. We analyze the relationship between the training efficiency and the sampled original images for matching, and accordingly propose a novel matching strategy.

3.2. Observations on Training Efficiency

In the dataset distillation process, the knowledge is distilled from sampled original images by matching certain objectives. The selection of original images thereby has large influences on the training efficiency. Recent literature usually adopts random sampling for selecting the original images [52, 24]. We set gradient matching as an example and carefully illustrate that random sampling disturbs efficient training of the dataset distillation.

Firstly, we analyze the matching effects of samples in different regions. Among all the samples in a class, those near the distribution center have higher prediction accuracy, indicating smaller backward gradients, while those on the decision boundaries have the contrary condition. For gradient matching, the central samples provide less effective supervision, while the gradients of the boundary ones largely dominate the optimization direction. We show the training accuracy curve of matching the synthetic images with only the central or boundary samples in Fig. 2b. The small gradients provided by central samples soon fail to provide effective supervision. On the other hand, although the boundary samples are essential for building decision boundaries, only matching with them brings chaotic matching targets, which degrades the distillation performance.

Secondly, we demonstrate that random sampling cannot guarantee an evenly distributed mini-batch along the training process. We record the Maximum Mean Discrepancy (MMD) between the selected mini-batch and the whole class distribution during training in Fig. 2c. It can be observed that the MMD is kept at a relatively high level, with large fluctuations during the training process. For the gradient matching, as the mini-batch cannot effectively and consistently cover the original class sample distribution, the gradient difference of different samples are not balanced. The matching target of a mini-batch may be biased by



(a) The training pipeline of the proposed DREAM strategy.

(b) The accuracy curve with different strategies for selecting original images.

(c) The MMD curve between the sampled mini-batch and the corresponding class data.

(d) Example clustering and sub-cluster center results of DREAM.

Figure 2: The original images obtained by random sampling have uneven distributions, which may result in noisy or biased matching targets. Besides, the coverage of random sampling on the whole sample space is low and has large fluctuations during training. Comparatively, the centers selected by DREAM (stars) are representative for corresponding sub-clusters, and are evenly distributed over the whole class feature space. Experiments for (b) and (c) are conducted under 10 images-per-class setting on CIFAR-10. Best viewed in color.

boundary samples with larger training gradients, which results in unstable supervision.

Besides, an unevenly distributed mini-batch also indicates relatively poor sample diversity. Information redundancy at dense regions and information lack at sparse regions make the mini-batch insufficient to represent the original data. The above factors result in optimization instability of the distillation process, and hence degrade the training efficiency. Since randomly sampling original images disturbs the training efficiency for dataset distillation training, we propose to design a novel strategy to construct mini-batches with even and diverse distribution for matching.

3.3. Representative Matching

Based on the purpose of achieving stable and fast optimization, only representative original images are selected for gradient matching. The selection of representative images are supposed to obey the following two principles. On the one hand, the selected images should be evenly distributed to avoid biased matching targets. On the other hand, while ensuring diversity, the selected samples should reflect the overall sample distribution of the class as accurately as possible.

Therefore, we employ a clustering process for selecting representative original images. Out of the considerations of uniform sub-cluster sizes and distribution, without loss of generality, we adopt K-Means [14, 2, 33] for dividing sub-clusters. As shown in Fig. 2d, the clustering is conducted inside each class to generate N sub-clusters that reflect the sample density. N is a pre-defined hyper-parameter for the mini-batch size of real images. The sub-cluster centers evenly cover the sample space of the whole class, and simultaneously provide sufficient diversity, which perfectly meets the above principles.

The complete training pipeline is illustrated in Fig. 2a. The clustering-selected original mini-batch and the synthetic images with the same class label are passed through the random model $\mathcal{M}_\theta$ to obtain prediction scores p_t and p_s. Subsequently calculate the classification losses and their corresponding gradients. The gradient differences are backwarded to update synthetic images according to Eq. 2. Considering the brought extra time cost, the clustering process is conducted every I_{int} iterations.

Additionally, at the beginning of the training process, we cluster the data of each class into sub-clusters corresponding to the pre-defined images-per-class number. We select the center samples of each sub-cluster as the initialization of the synthetic images. A more balanced clustering-based initialization better reflects the data distribution, and accelerates the convergence from the very beginning of the training process.

4. Experiments

4.1. Datasets and Implementation Details

We verify the effectiveness of our method on multiple popular dataset distillation benchmarks, including CIFAR10 [25], CIFAR100 [25], SVHN [31], MNIST [27], FashionMNIST [49] and TinyImageNet [9]. For evaluation, we train a model on the distilled synthetic images and test it on the original testing images. Top-1 accuracy is reported to show the performance.

Without specific designation, the experiment is conducted on 3-layer convolutional networks (ConvNet-3) [15] with 128 filters and instance normalization [44]. The matching mini-batch size for original images is set as 128. By default we set IDC [24] as the baseline method. The gradient matching metric $\mathbf{D}$ in Eq. 2 is empirically set as the mean squared error for CIFAR-10, CIFAR-100, TinyImageNet and SVHN. For MNIST and FashionMNIST, $\mathbf{D}$ is set as the mean absolute error [24]. We conduct 1,200 matching iterations in total, inside each of which 100 inner loops are conducted. SGD is set as the optimizer, with a learning rate of 0.005. For clustering, we employ the matching model for feature extraction. The clustering interval I_{int} is set as 10 iterations, whose sensitiveness is analyzed in Sec. 4.3. We also analyze the influence of different sampling strategy from the sub-clusters in Sec. 4.3. For evaluation, we train a network for 1,000 epochs on the distilled images with a learning rate of 0.01. We perform 5 experiments and report the mean and standard deviation of the results.

4.2. Comparison with State-of-the-art Methods

We compare the distilled synthetic dataset performance of DREAM and other state-of-the-art (SOTA) coreset-based and optimization-based methods on multiple datasets with different images-per-class (IPC) settings in Tab. 1. Besides, on TinyImageNet, we compare DREAM with DM [51] and MTT [3] in Tab. 2. Under all experiment circumstances, the proposed DREAM consistently surpasses other SOTA methods. With a small IPC setting especially, under the guidance of proper gradients, DREAM is more robust than other methods, which proves the effectiveness of the representative matching strategy. Further narrowing the performance gap between small-scale distilled datasets and the original ones indicates that the information loss of dataset distillation is reduced. More detailed comparisons are included in the supplementary material.

4.3. Ablation Study and Analysis

Extended experiments are designed to verify the effectiveness of our proposed DREAM strategy. Without specific designation, the experiment is conducted under the 10 IPC setting on CIFAR-10 dataset.

Component Combination Evaluation. Firstly, we verify the isolated effects of each component in our proposed DREAM strategy in Tab. 3. Under the same initialization, our proposed representative matching strategy largely improves the final dataset performance. Comparatively, the clustering-based initialization offers a large performance lead before the training begins, yet eventually brings limited improvements. Nevertheless, it still provides stable boosts and accelerates the training convergence added on the representative matching to form the whole DREAM method. Combining all the components, the full DREAM method cuts the required iterations to achieve the baseline performance by more than 8 times.

Additionally, in Fig. 2 we further illustrate the effectiveness of DREAM. Fig. 2b shows that by simply assigning samples from the sub-clusters as initialization of synthetic images, the validation performance surpasses random initialization by a large margin. Under the joint effect of representative matching and clustering-based initialization, DREAM achieves the final performance of random sampling with less than eighth of training iterations, demonstrating a significant training efficiency improvement. Continuing increasing the training iterations, DREAM further improves the dataset performance by applying proper gradient as supervision.

From the sample distribution perspective, Fig. 2c demonstrates that the original images selected by DREAM consistently show lower MMD scores with the original distribution with less fluctuations, compared with random sampling. The smaller fluctuations validates that sub-cluster centers effectively and stably cover the feature distribution, and reduces the noise at the sample level during the training process. With sufficient sample diversity, distribution evenness and appropriate gradient supervision, DREAM ensures a smoother and more robust optimization process for dataset distillation training.

For better illustration of the universality of DREAM, we apply the representative matching and clustering-based initialization to some other baseline methods and receive similar effects in Tab. 3. The accuracy curve comparisons are presented in the supplementary material. It proves that DREAM is able to be easily plugged into dataset distillation frameworks and help improve the training efficiency.

Cross Architecture Generalization Analysis. It has been a problem for previous optimization-based dataset distillation works to generalize across architectures as the synthetic images would over-fit to the model utilized for gradient matching [52, 24]. In Tab. 4 we demonstrate the cross architecture performance of our proposed DREAM strategy. We distill the dataset with ConvNet-3 and ResNet-10 [20], and validate the performance on ConvNet-3, ResNet-10 and DenseNet-121 [21].

DREAM surpasses the compared methods on both the

Table 1: Top-1 accuracy of test models trained on distilled synthetic images on multiple datasets. The distillation training is conducted with ConvNet-3. † denotes the reported error range is reproduced by us.

	IPC	Ratio %	Coreset Selection Random	 Herding	 DC [52]	 DSA [50]	 DM [51]	Training Set Synthesis CAFE [46]	 MTT [3]	 IDC [24]	 **DREAM**	Whole Dataset
MNIST	1	0.017	$64.9_{\pm 3.5}$	$89.2_{\pm 1.6}$	$91.7_{\pm 0.5}$	$88.7_{\pm 0.6}$	$89.7_{\pm 0.6}$	$93.1_{\pm 0.3}$	-	$94.2_{\pm 0.2}{}^{\dagger}$	**95.7**$_{\pm 0.3}$	$99.6_{\pm 0.0}$
	10	0.17	$95.1_{\pm 0.9}$	$93.7_{\pm 0.3}$	$97.4_{\pm 0.2}$	$97.8_{\pm 0.1}$	$97.5_{\pm 0.1}$	$97.2_{\pm 0.3}$	-	$98.4_{\pm 0.1}{}^{\dagger}$	**98.6**$_{\pm 0.1}$	
	50	0.83	$97.9_{\pm 0.2}$	$94.8_{\pm 0.2}$	$98.8_{\pm 0.2}$	$99.2_{\pm 0.1}$	$98.6_{\pm 0.1}$	$98.6_{\pm 0.2}$	-	$99.1_{\pm 0.1}{}^{\dagger}$	**99.2**$_{\pm 0.1}$	
FashionMNIST	1	0.017	$51.4_{\pm 3.8}$	$67.0_{\pm 1.9}$	$70.5_{\pm 0.6}$	$70.6_{\pm 0.6}$	-	$77.1_{\pm 0.9}$	-	$81.0_{\pm 0.2}{}^{\dagger}$	**81.3**$_{\pm 0.2}$	$93.5_{\pm 0.1}$
	10	0.17	$73.8_{\pm 0.7}$	$71.1_{\pm 0.7}$	$82.3_{\pm 0.4}$	$84.6_{\pm 0.3}$	-	$83.0_{\pm 0.4}$	-	$86.0_{\pm 0.3}{}^{\dagger}$	**86.4**$_{\pm 0.3}$	
	50	0.83	$82.5_{\pm 0.7}$	$71.9_{\pm 0.8}$	$83.6_{\pm 0.4}$	**88.7**$_{\pm 0.2}$	-	$84.8_{\pm 0.4}$	-	$86.2_{\pm 0.2}{}^{\dagger}$	$86.8_{\pm 0.3}$	
SVHN	1	0.014	$14.6_{\pm 1.6}$	$20.9_{\pm 1.3}$	$31.2_{\pm 1.4}$	$27.5_{\pm 1.4}$	-	$42.6_{\pm 3.3}$	-	$68.5_{\pm 0.9}{}^{\dagger}$	**69.8**$_{\pm 0.8}$	$95.4_{\pm 0.1}$
	10	0.14	$35.1_{\pm 4.1}$	$50.5_{\pm 3.3}$	$76.1_{\pm 0.6}$	$79.2_{\pm 0.5}$	-	$75.9_{\pm 0.6}$	-	$87.5_{\pm 0.3}{}^{\dagger}$	**87.9**$_{\pm 0.4}$	
	50	0.7	$70.9_{\pm 0.9}$	$72.6_{\pm 0.8}$	$82.3_{\pm 0.3}$	$84.4_{\pm 0.4}$	-	$81.3_{\pm 0.3}$	-	$90.1_{\pm 0.1}{}^{\dagger}$	**90.5**$_{\pm 0.1}$	
CIFAR10	1	0.02	$14.4_{\pm 2.0}$	$21.5_{\pm 1.2}$	$28.3_{\pm 0.5}$	$28.8_{\pm 0.7}$	$26.0_{\pm 0.8}$	$30.3_{\pm 1.1}$	$46.3_{\pm 0.8}$	$50.6_{\pm 0.4}{}^{\dagger}$	**51.1**$_{\pm 0.3}$	$84.8_{\pm 0.1}$
	10	0.2	$26.0_{\pm 1.2}$	$31.6_{\pm 0.7}$	$44.9_{\pm 0.5}$	$52.1_{\pm 0.5}$	$48.9_{\pm 0.6}$	$46.3_{\pm 0.6}$	$65.3_{\pm 0.7}$	$67.5_{\pm 0.5}$	**69.4**$_{\pm 0.4}$	
	50	1.0	$43.4_{\pm 1.0}$	$40.4_{\pm 0.6}$	$53.9_{\pm 0.5}$	$60.6_{\pm 0.5}$	$63.0_{\pm 0.4}$	$55.5_{\pm 0.6}$	$71.6_{\pm 0.2}$	$74.5_{\pm 0.1}$	**74.8**$_{\pm 0.1}$	
CIFAR100	1	0.2	$4.2_{\pm 0.3}$	$8.4_{\pm 0.3}$	$12.8_{\pm 0.3}$	$13.9_{\pm 0.3}$	$11.4_{\pm 0.3}$	$12.9_{\pm 0.3}$	$24.3_{\pm 0.3}$	-	**29.5**$_{\pm 0.3}$	$56.2_{\pm 0.3}$
	10	2	$14.6_{\pm 0.5}$	$17.3_{\pm 0.3}$	$25.2_{\pm 0.3}$	$32.3_{\pm 0.3}$	$29.7_{\pm 0.3}$	$27.8_{\pm 0.3}$	$40.1_{\pm 0.4}$	$45.1_{\pm 0.4}{}^{\dagger}$	**46.8**$_{\pm 0.7}$	
	50	10	$30.0_{\pm 0.4}$	$33.7_{\pm 0.5}$	-	$42.8_{\pm 0.4}$	$43.6_{\pm 0.4}$	$37.9_{\pm 0.3}$	$47.7_{\pm 0.2}$	-	**52.6**$_{\pm 0.4}$	

Table 2: Top-1 accuracy of test models trained on distilled synthetic images on TinyImageNet. The distillation training is conducted with ConvNet-3.

IPC	Ratio %	DM [51]	MTT [3]	**DREAM**	Whole
1	0.017	$3.9_{\pm 0.2}$	$8.8_{\pm 0.3}$	**10.0**$_{\pm 0.4}$	$37.6_{\pm 0.4}$
50	0.83	$24.1_{\pm 0.3}$	$28.0_{\pm 0.3}$	**29.5**$_{\pm 0.3}$	

Table 3: Ablation study on the components of the proposed DREAM. RM indicates Representative Matching, and Init stands for clustering-based initialization. "Iter" stands for the required iterations to achieve the baseline performance.

	Comp RM	 Init	Top-1	Iter		Comp RM	 Init	Top-1
IDC	-	-	$67.5_{\pm 0.5}$	1000	DC	-	-	$44.9_{\pm 0.5}$
	✓	-	$68.9_{\pm 0.5}$	350		✓	✓	**45.9**$_{\pm 0.3}$
	-	✓	$68.1_{\pm 0.3}$	750	DSA	-	-	$52.1_{\pm 0.5}$
	✓	✓	**69.4**$_{\pm 0.4}$	**150**		✓	✓	**53.1**$_{\pm 0.4}$

Table 4: Ablation study on cross architecture distilled dataset performance of the proposed DREAM strategy. The dataset is first distilled on a model D and then validated on another model T. † denotes the result is reproduced by us.

	$D\backslash T$	Conv-3	Res-10	Dense-121
MTT [3]	Conv-3	$64.3_{\pm 0.7}$	$34.5_{\pm 0.6}{}^{\dagger}$	$41.5_{\pm 0.5}{}^{\dagger}$
	Res-10	$44.2_{\pm 0.3}{}^{\dagger}$	$20.4_{\pm 0.9}{}^{\dagger}$	$24.2_{\pm 1.3}{}^{\dagger}$
IDC [24]	Conv-3	$67.5_{\pm 0.5}$	$63.5_{\pm 0.1}$	$61.6_{\pm 0.6}$
	Res-10	$53.6_{\pm 0.6}{}^{\dagger}$	$50.6_{\pm 0.9}{}^{\dagger}$	$51.7_{\pm 0.6}{}^{\dagger}$
DREAM	Conv-3	**69.4**$_{\pm 0.4}$	**66.3**$_{\pm 0.8}$	**65.9**$_{\pm 0.5}$
	Res-10	**53.7**$_{\pm 0.6}$	**51.0**$_{\pm 0.9}$	**52.8**$_{\pm 0.6}$

Table 5: Ablation study on different sampling strategy to form a mini-batch from sub-clusters.

		Sub-cluster number N 32	 64	 128	 256
Samples per sub-cluster n	1	$67.2_{\pm 0.3}$	$68.5_{\pm 0.1}$	**69.4**$_{\pm 0.4}$	$68.9_{\pm 0.2}$
	2	$67.7_{\pm 0.3}$	$68.6_{\pm 0.3}$	$69.2_{\pm 0.7}$	-
	4	$67.7_{\pm 0.4}$	$68.7_{\pm 0.4}$	-	-
	8	$67.5_{\pm 0.3}$	-	-	-

absolute performance and the performance drop when applying the distilled dataset on an unseen architecture. The strong cross architecture generalization capability verifies that DREAM helps build a more reasonable distilled dataset compared to random sampling.

Sampling Strategy Analysis. Representative matching conducts clustering for each class and samples original images from the sub-clusters to form a mini-batch. We analyze the influence of different sampling strategy on the training results in Tab. 5 and Fig. 3. Among each sub-cluster, the top-n samples closest to the center are selected. By grouping different sub-cluster number and samples per sub-cluster, we are able to obtain original image mini-batches different in scale and diversity. As observed in the results, by representative matching the dataset performance is generally stable, and receives improvements to certain extent over the baseline (67.5).

Compared in more detail, with a small sub-cluster num-

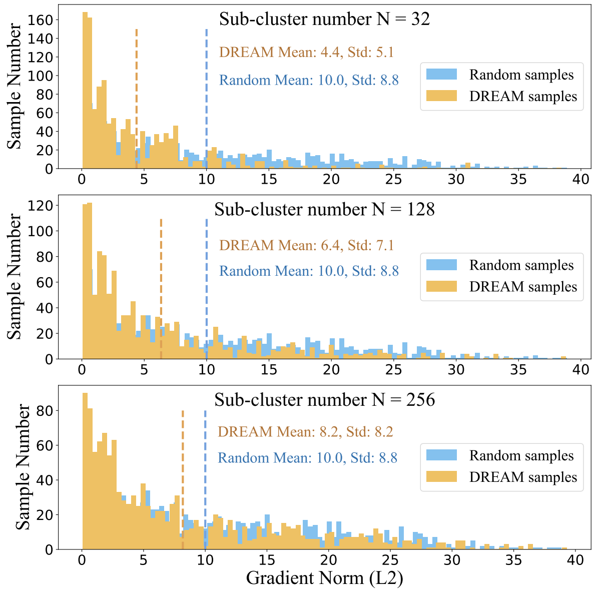


Figure 3: The gradient distribution comparison between random sampling and our proposed DREAM strategy under different sub-cluster sample number N. Best viewed in color.

ber $N = 32$, the sub-cluster centers are more likely to be distributed in areas with smaller gradients, as shown in the first row of Fig. 3. As the random model $\mathcal{M}_\theta$ is trained, these samples gradually fail to provide effective gradients for supervision, resulting in a sub-optimal performance. Oppositely, a larger sub-cluster number $N = 256$ involves a distribution closer to random sampling, which brings a small performance drop, as shown in the last row of Fig. 3. Due to memory limitations, it is not applicable to further increase N, but it is conceivable that the extreme condition should yield similar results to random sampling. On the other hand, the sample number per sub-cluster n has only a slight effect on the results. The group of 1 center sample per sub-cluster and 128 sub-clusters in total is proved to obtain the optimal gradient supervision as in the second row of Fig. 3, and is chosen for mini-batch composition.

Training Stability Analysis. In order to more intuitively demonstrate the effects of the proposed DREAM strategy on the training process, we visualize the feature migration of DREAM and random sampling in Fig. 4a. Specifically, we randomly select a synthetic image as initialization and record its updated version every 10 iterations. We employ a random network to extract the features of all images, and calculate the Euclidean distance between adjacent versions of images. For DREAM, at the beginning of the training process, under the direction of proper gradients, the synthetic image goes through a larger migration. Within 100 iterations, the synthetic image has reached a relatively optimal position, and makes subsequent fine-tuning. On the contrary, there are still large fluctuations for the synthetic image matched with randomly sampled original images in the late training period, partly due to the noisy matching targets generated by uneven mini-batches.

Clustering Interval Sensitivity Analysis. We evaluate the influence of different clustering interval I_{int} on the final dataset performance in Fig. 4b. Conducting clustering at every iteration leads to the best performance, while adding the clustering interval until 10 brings mild influences. As there is an obvious top-1 accuracy degradation when the interval is further increased to 20, we select an interval of 10 to balance the distilled dataset performance and the extra calculation cost. More analysis on the computational cost of clustering is included in the supplementary material.

Experimental results on ImageNet-1K. We compare DREAM with the current state-of-the-art method TESLA [6] on ImageNet-1K in Tab. 6. The experimental setting is the same as in TESLA. DREAM shows excellent performance on large datasets.

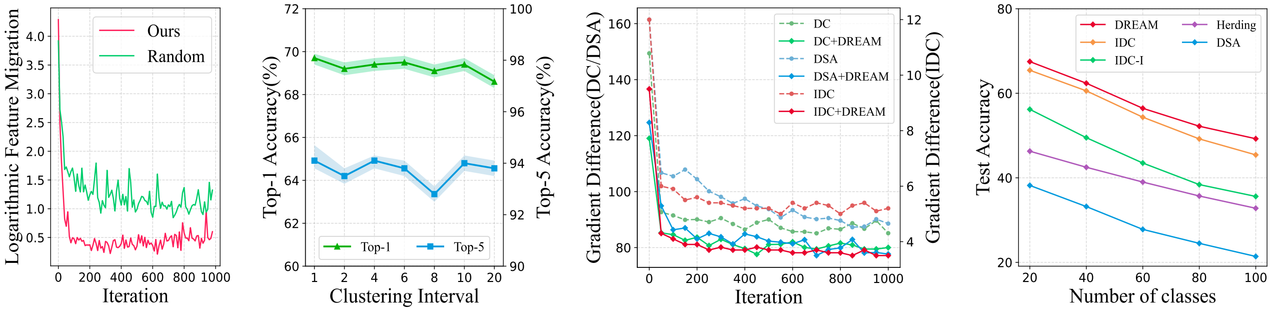
Table 6: Results on ImageNet.

Method	TESLA	DREAM
Acc (IPC=10)	17.8	**18.5**

Differences from DC-BENCH[5]. We provide a detailed comparison between DREAM and DC-BENCH to clarify their distinctions. DC-BENCH solely concentrates on a better initialization and lacks specific designs for the subsequent matching-based optimization, while DREAM selects representative samples for matching and enables the realization of a fully efficient training process for distillation. Furthermore, DREAM conducts extensive experiments to analyze the impact of cluster number, sample number per cluster and clustering interval. DC-BENCH achieves comparable performance using 30% of iterations, whereas DREAM achieves similar results with only 10-20% iterations. Additionally, given sufficient training time, DREAM further achieves up to 3.7% and 5.8% accuracy improvements for gradient matching and distribution matching respectively which surpass DC-BENCH's 1.3%.

4.4. Visualizations

Gradient Difference Curve. As the dataset distillation training is constrained by matching the training gradients, a smaller gradient difference also indicates a better matching effect. Therefore, we also visualize the gradient difference curve of the dataset distillation in Fig. 4c, which is calculated by the training loss Eq. 2. We add the DREAM strategy to DC, DSA and IDC methods. Throughout the training process, DREAM holds a smaller gradient difference compared with the baseline methods. On the one hand, it verifies the effectiveness of DREAM on improving the training



(a) (b) (c) (d)

Figure 4: (a): The feature migration during the training process. (b): Ablation study on different clustering interval. (c): The training loss curve during the training process. (d): The continual learning accuracy curve.

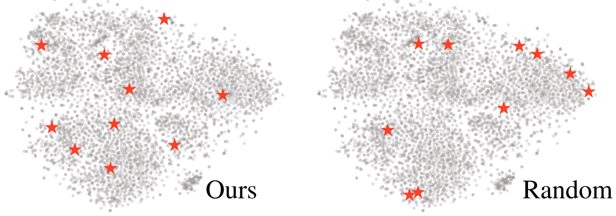


Figure 5: The sample distribution comparison on the final distilled images (marked as red stars) between our proposed DREAM (left) and random sampling (right).

efficiency to reduce the gradient difference in limited iterations. On the other hand, the large fluctuations of the baseline methods also validate the existence of noisy gradients generated by random sampling.

Sample Distribution Visualization. In order to more intuitively demonstrate the effectiveness of our proposed DREAM on generating synthetic sets well covering the original sample distribution, we visualize the t-SNE graphs of the synthetic images for both random sampling and DREAM. As shown in Fig. 5, the final distribution constrained by DREAM strategy evenly cover the whole class, while random sampling generates biased optimization results. Furthermore, a large percentage of samples are pulled to the distribution edge in the random sampling results, which also validates that the matching is biased by boundary samples with larger gradients. By consistently providing proper gradient supervision, DREAM achieves a more diverse and robust distillation result.

Synthetic Image Visualization. In order to more intuitively demonstrate the effects of DREAM on the distilled images, we compare the distillation results of adding the proposed DREAM strategy or not in Fig. 6. 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 helps the distilled datasets to obtain better validation performance.

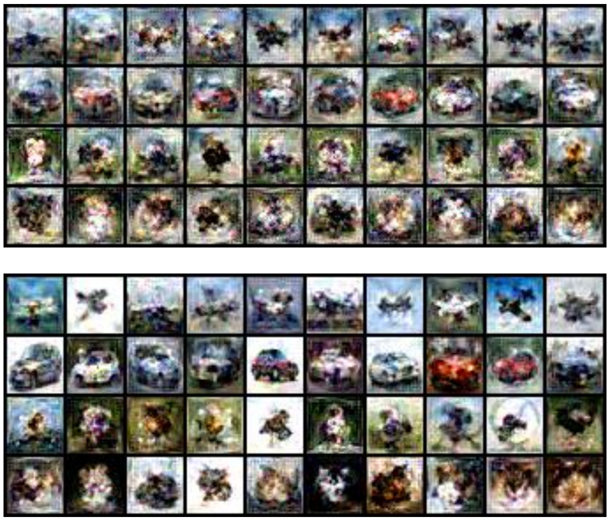


Figure 6: The distilled dataset comparison between DC (Upper row) and DC with DREAM strategy (Bottom row) on CIFAR-10 (plane, car, dog, cat classes). DREAM introduces more obvious categorical characteristics and variety to the distilled image. Best viewed in color. More visualization is provided in supplementary material.

4.5. Application on Continual Learning

Dataset distillation generates compact datasets that are able to represent the original ones, which can thus be applied to continual learning problems [35, 1, 48, 24]. We further validate the effectiveness of the proposed DREAM strategy on the continual learning scenarios in Fig. 4d. Following the settings in [52, 24], we conduct a 5-step class-incremental experiment on CIFAR-100, each step with 20 classes. For better demonstrating the generalization capability of DREAM, the distillation synthesis is conducted on ConvNet-3, and the validation on ResNet-10.

DREAM consistently maintains performance advantages over other approaches throughout the training process, and the performance gap is further enlarged as the learnt class number is gradually increased. It proves that better distillation quality helps the model construct clearer decision boundaries and memorize the discriminative information.

5. Conclusion

In this paper, we propose a novel Dataset distillation by REpresentAtive Matching (DREAM) strategy to address the training efficiency problem for dataset distillation. By only matching with the representative original images, DREAM reduces the optimization instability, and reaches a smoother and more robust training process. It is able to be easily plugged into popular dataset distillation frameworks to reduce the training iterations by more than 8 times without performance drop. The stable optimization also provides higher final performance and generalization capability. The more efficient matching allows future works to design more complicated matching metrics.

6. Limitations and Future Works

Although the proposed DREAM strategy significantly improves the training efficiency of optimization-based dataset distillation methods, the calculation burden is still large when the image size and the class number increases. It is difficult for these methods to handle ultra large-scale datasets like ImageNet [9], even if the training efficiency has been improved by DREAM. We will explore more resource-friendly ways to conduct dataset distillation in future works.

Acknowledgement

This research is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-PhD-2021-08- 008). Yang You's research group is being sponsored by NUS startup grant (Presidential Young Professorship), Singapore MOE Tier-1 grant, ByteDance grant, ARCTIC grant, SMI grant and Alibaba grant. The research is also supported by the National Natural Science Foundation of China (No. 62173302).

References

[1] Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. In *NeurIPS*, pages 11817–11826, 2019. 2, 8

[2] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. Technical report, Stanford, 2006. 3, 4

[3] 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. 1, 2, 3, 5, 6

[4] Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019. 2

[5] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. In *NeurIPS*, 2022. 1, 7

[6] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 7

[7] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, pages 6638–6646, 2017. 1

[8] Zhou Daquan, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. Dataset quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 5, 9

[10] Chris Ding and Xiaofeng He. Cluster merging and splitting in hierarchical clustering algorithms. In *ICDM.*, pages 139–146. IEEE, 2002. 3

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1

[12] Jiawei Du, Yidi Jiang, Vincent Y. F. Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the accumulated trajectory error to improve dataset distillation. In *CVPR*, pages 3749–3758, 2023. 1, 2

[13] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *KDD*, pages 226–231, 1996. 3

[14] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965. 3, 4

[15] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *CVPR*, pages 4367–4375, 2018. 5

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 1

[17] Chengcheng Guo, Bo Zhao, and Yanbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. *arXiv preprint arXiv:2204.08499*, 2022. 2

[18] Greg Hamerly and Charles Elkan. Alternatives to the k-means algorithm that find better clusterings. In *CIKM*, pages 600–607, 2002. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2022. 1

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 5

[21] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017. 5

[22] 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

[23] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *NeurIPS*, 33:12104–12114, 2020. 3

[24] 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. 1, 2, 3, 5, 6, 8

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

[26] Agata Lapedriza, Hamed Pirsiavash, Zoya Bylinskii, and Antonio Torralba. Are all training examples equally valuable? *arXiv preprint arXiv:1311.6510*, 2013. 1, 2

[27] 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. 5

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

[29] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 1

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

[31] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 5

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

[33] Sehban Omer. fast-pytorch-kmeans, 9 2020. 4

[34] Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, Daquan Zhou, and Yang You. Infobatch: Lossless training speed up by unbiased dynamic data pruning. *arXiv preprint arXiv:2303.04947*, 2023. 1

[35] Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, pages 2001–2010, 2017. 2, 8

[36] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, pages 779–788, 2016. 1

[37] Hajar Rehioui, Abdellah Idrissi, Manar Abourezq, and Faouzia Zegrari. Denclue-im: A new approach for big data clustering. *Procedia Computer Science*, 83:560–567, 2016. 3

[38] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017. 2

[39] Sam Shleifer and Eric Prokop. Using small proxy datasets to accelerate hyperparameter search. *arXiv preprint arXiv:1906.04887*, 2019. 2

[40] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022. 1

[41] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018. 1, 2

[42] Ngoc-Trung Tran, Viet-Hung Tran, Ngoc-Bao Nguyen, Trung-Kien Nguyen, and Ngai-Man Cheung. Towards good practices for data augmentation in gan training. *arXiv preprint arXiv:2006.05338*, 2:3, 2020. 3

[43] Yi-Hsuan Tsai, Wei-Chih Hung, Samuel Schulter, Kihyuk Sohn, Ming-Hsuan Yang, and Manmohan Chandraker. Learning to adapt structured output space for semantic segmentation. In *CVPR*, pages 7472–7481, 2018. 1

[44] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[45] Kai Wang, Jianyang Gu, Daquan Zhou, Zheng Zhu, Wei Jiang, and Yang You. Dim: Distilling dataset into generative model. *arXiv preprint arXiv:2303.04707*, 2023. 2

[46] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. Cafe: Learning to condense dataset by aligning features. In *CVPR*, pages 12196–12205, 2022. 1, 2, 3, 6

[47] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1, 2

[48] Felix Wiewel and Bin Yang. Condensed composite memory continual learning. In *IJCNN*, pages 1–8. IEEE, 2021. 2, 8

[49] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 5

[50] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, pages 12674–12685. PMLR, 2021. 2, 6

[51] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. *arXiv preprint arXiv:2110.04181*, 2021. 2, 5, 6

[52] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2020. 1, 2, 3, 5, 6, 8

[53] Shengyu Zhao, Zhijian Liu, Ji Lin, Jun-Yan Zhu, and Song Han. Differentiable augmentation for data-efficient gan training. *NeurIPS*, 33:7559–7570, 2020. 3

[54] Zhengli Zhao, Zizhao Zhang, Ting Chen, Sameer Singh, and Han Zhang. Image augmentations for gan training. *arXiv preprint arXiv:2006.02595*, 2020. 3

[55] Zangwei Zheng, Mingyuan Ma, Kai Wang, Ziheng Qin, Xiangyu Yue, and Yang You. Preventing zero-shot transfer degradation in continual learning of vision-language models. *arXiv preprint arXiv:2303.06628*, 2023. 1

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