DiRe: Diversity-promoting Regularization for Dataset Condensation

Saumyaranjan Mohanty[1] Aravind Reddy[2] Konda Reddy Mopuri[1]
[1]Department of Artificial Intelligence, Indian Institute of Technology Hyderabad
[2]Centre for Responsible AI, Wadhwani School of Data Science & AI, Indian Institute of Technology Madras
ai23resch04001@iith.ac.in aravind@cerai.in krmopuri@ai.iith.ac.in

Abstract

*In Dataset Condensation, given an original training dataset, the goal is to synthesize a small dataset that replicates the training utility of the original dataset, when used to train neural networks. Existing condensation methods synthesize datasets that contain significant redundancy, leading to their inefficiency. Thus, there is a dire need to ensure diversity in the synthesized datasets. In this work, we propose an intuitive **Di**versity **Re**gularizer (**DiRe**) composed of cosine similarity and Euclidean distance. Most importantly, the proposed regularizer can be applied off-the-shelf to various state-of-the-art optimization-driven condensation methods. Through extensive experimentation, we demonstrate that our approach improves state-of-the-art condensation methods on various benchmark datasets from CIFAR-10 to ImageNet-1K with respect to generalization and diversity metrics.*

1. Introduction

Training datasets for modern neural networks have grown to large scales, making the learning process cumbersome and expensive. To ameliorate this issue, there has been tremendous recent interest on *Dataset Condensation* (DC), also referred to as *Dataset Distillation* [5, 7, 28, 29, 32, 35, 41, 47, 56]. Here, the goal is to generate a small synthetic dataset from the original large dataset that can be used instead for neural network training and other related tasks, such as neural architecture search [38]. Furthermore, dataset condensation has been shown to have several different applications, such as memory rehearsal in continual learning [6, 13], data privacy [3, 8, 26], and federated learning [20, 44].

Other active approaches for data-efficient deep learning include Data Subset Selection [21, 42], Coreset Selection [15, 43, 48], and Dataset Pruning [46, 51, 53]. A *subset* of the original dataset is selected as a substitute for model training in all these approaches. Since they are constrained to only select *real* data points from the training dataset, they typically need to produce much larger datasets than DC for comparable performance.

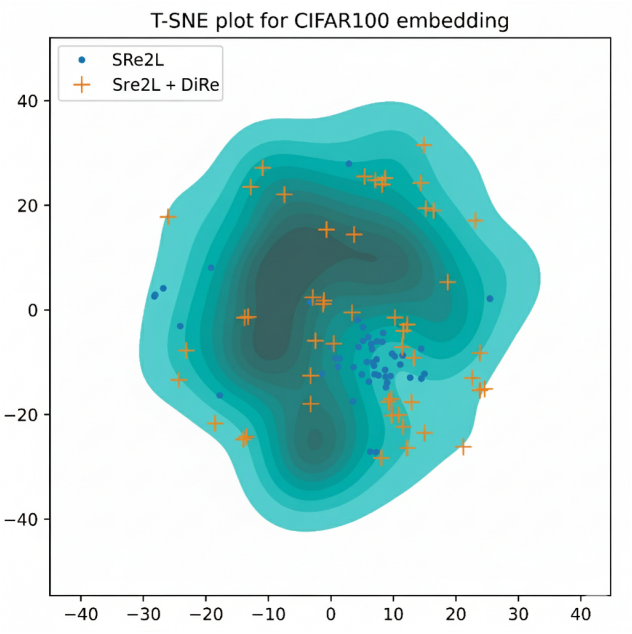

Figure 1. T-SNE plot of embeddings of synthetic images generated from the CIFAR100 dataset with IPC=50 settings. As can be seen, synthetic images generated by adding the **DiRe** regularizer are more diverse than the vanilla SRe^{2}L [50] method. The synthetic images are spread throughout the original dataset's feature space (shown in Cyan color).

Most work on DC has focused on bi-level optimization-based strategies, where the outer optimization task is for synthetic dataset updates, and the inner optimization task is for model updates. Following the taxonomy in [32], we note that there are primarily four different high-level approaches to DC, such as those focusing on meta-model matching [41, 59], gradient matching [54, 56], trajectory matching [1, 14, 25], and distribution matching [55, 58].

Before 2023, due to the bi-level optimization nature of most DC algorithms, none of them could scale to large-scale datasets such as ImageNet-1K, as they require storing the entire original training dataset in memory. To circumvent

this issue, Yin et al. [50] proposed SRe^{2}L, which decouples the outer and inner optimization tasks, enabling DC on large datasets. Recent works have been based on such decoupled optimization-based strategies [10, 34–36, 49, 52].

Given that the primary goal of DC is to reduce redundancy in the training data, it is natural to ask if encouraging *diversity* in the synthesized datasets is helpful. Surprisingly, very little prior work has focused on diversity in dataset condensation [10, 39], despite Dataset Diversity having been widely acknowledged as a crucial aspect for the successful training of machine learning models [12, 40, 45, 57]. The diversity-driven DC approaches [10, 39] currently offer the state-of-the-art generalization performance for dataset condensation. This motivates us to ask the following natural question:

Can we use a simple and intuitive regularizer to enhance diversity in dataset condensation, leading to better generalization?

Surprisingly, we found that the answer is a resounding "Yes!". We initially started using only *cosine similarity* as a regularizer, but found that augmenting it with *Euclidean distance* helps significantly. Our **Di**versity **Re**gularizer (**DiRe**) can be applied off the shelf to any existing optimization-based DC algorithm to improve the diversity of distilled datasets.

Another key issue we noticed is that both [10] and [39] face a pitfall, which is very common to ML research studying diversity. This pitfall is highlighted vividly by the title of [57], which won one of the outstanding paper awards in ICML 2024; "Measure Dataset Diversity, Don't Just Claim It". Although [10] and [39] *claim* that they produce datasets that are more diverse than prior DC algorithms, they do not quantitatively measure diversity using any well-established metrics. We also tackle this by studying the diversity of distilled datasets using established quantitative notions of diversity [11, 27]. In summary, the major contributions of our work can be summarized as follows.

- We propose an intuitive **Di**versity **Re**gularizer (**DiRe**) that can be applied off-the-shelf to any dataset condensation algorithm that has a separate synthesis stage.
- We are the first to quantitatively study the diversity of distilled datasets using well-established Dataset Diversity measures such as Coverage [27] and Vendi Score [11]. We show that DiRe significantly improves SRe^{2}L on both these metrics.
- Additionally, we also demonstrate that the proposed regularizer can be added to other optimization-based DC methods. We have experimented on DWA [10], CDA [49], UFC [52], DELT [36], G-VBSM [34], MTT [2], and DM [55] to improve their generalization and diversity measures. Note that MTT and DM are trajectory matching and distribution matching DC algorithms, which are very different from decoupling-based algorithms such as SRe^{2}L, DWA, CDA, UFC, DELT, and G-VBSM.

We would like to highlight that our proposed regularizer is applicable to optimization-based DC methods. Note that our proposed regularizer is not applicable for optimization-free algorithms such as RDED [39], as they do not incorporate a synthesis stage where optimization occurs.

2. Methodology

In this section, we describe our Diversity Regularizer *DiRe* in detail and provide Algorithm 1 to improve SRe^{2}L. Table 1 provides the notation used throughout the paper.

Table 1. Notation used throughout the paper.

Symbol	Description
V	Original large-scale training dataset
S	Distilled synthetic dataset
θ	Parameters of the deep neural network
f_θ	Pretrained teacher network
h_θ	Feature extractor of f_θ
η	Learning rate for distillation
r_c	Weight for pairwise cosine similarity loss
r_e	Weight for pairwise Euclidean distance loss
$S_{cos}(A, B)$	Pairwise cosine similarity between rows of matrices A and B
$D_{euc}(A, B)$	Pairwise Euclidean distance between rows of matrices A and B
L_{ce}	Cross entropy loss
L_{bn}	Batch-Norm loss
$\langle x, y \rangle$	Inner product of vector x and vector y
$\|\|x\|\|$	L_2 norm of vector x

We have used SRe^{2}L as a case study to explain the methodology and implementation of DiRe while noting that it can be used with other dataset distillation methods that involve an optimization-based synthesis stage.

Dataset condensation aims to condense the original large labelled dataset $V = \{(x_i, y_i)\}_{i=1}^{|V|}$ to a much smaller synthetic dataset $S = \{\tilde{x}_i, \tilde{y}_i\}_{i=1}^{|S|}$, such that a model θ^S trained on S has a generalization performance close to that of a model θ^V trained on V.

Most DC algorithms involve solving a bi-level optimization problem with the outer loop for updating the synthetic dataset and the inner loop for updating the model. Yin et al. [50] decoupled this bi-level optimization and introduced SRe^{2}L, a lightweight tripartite learning paradigm that can scale to large datasets. such as ImageNet-1K. The three stages of SRe^{2}L [50] are:

- **S**queeze, in which a model is trained on the original training dataset, and the Batch-Norm (BN) layer statistics are stored.

- **Re**cover, in which the BN statistics are used to synthesize the condensed dataset.
- **Re**label, in which the synthetic dataset is assigned soft labels through the model trained on the original dataset.

The synthesis stage (Recover) of SRe^{2}L consists of synthesizing the target data starting from independent random Gaussian noise. Yin et al. [50] implicitly assumes that starting from random Gaussian noise provides diversity in this parallelized synthesis process. However, as shown in Figure 1, synthetic datasets generated by SRe^{2}L cover only a small portion of the entire dataset manifold. Du et al. [10] improves the diversity of SRe^{2}L by carefully tailoring their synthesis process using the specific properties of the Batch-Normalization statistics. In contrast to their approach, we show that a simple and intuitive diversity regularizer based on cosine similarity and Euclidean distances is sufficient to ensure diversity, leading to better generalization.

DiRe consists of three components.

1. **Cosine Diversity loss (CD)**: Promotes diversity among the synthetic dataset by minimizing the pairwise cosine similarities between their embeddings, effectively encouraging a more dispersed representation in the embedding space.
2. **Cosine Distribution Matching loss (CDM)**: Enhances diversity in synthetic dataset by maximizing the pairwise cosine similarity between synthetic and real image embeddings. This encourages the synthetic data to align with the real data in the embedding space.
3. **Euclidean Distribution Matching loss (EDM)**: Encourages synthetic embeddings to cluster near real ones in Euclidean space by minimizing the pairwise Euclidean distances between synthetic and real embeddings.

Let X^c_{syn} be the set of all the synthetic images belonging to class c and X^c_{real} be the set of all the real images belonging to class c. Embeddings of the synthetic images computed through the pre-trained feature extractor network are given as $E^c_{syn} = h_\theta(X^c_{syn})$ and embeddings of real images are given as $E^c_{real} = h_\theta(X^c_{real})$.

Pairwise cosine similarity between two matrices A and B with dimensions $N \times D$ and $M \times D$ (i.e., set of N and M vectors of D dimensions respectively; A^i is the i^{th} row in A) is calculated as shown in Equation 1.

$$S_{cos}(A,B) = \sum_{i=1}^{N}\sum_{j=1}^{M} \frac{\sum_{d=1}^{D} A_d^i \cdot B_d^j}{\sqrt{\sum_{d=1}^{D}(A_d^i)^2} \cdot \sqrt{\sum_{d=1}^{D}(B_d^j)^2}} \tag{1}$$

Pairwise Euclidean distance between two matrices A and B with D number of features is calculated as shown in Equation 2.

$$D_{euc}(A,B) = \sum_{i=1}^{N}\sum_{j=1}^{M} \sqrt{\sum_{d=1}^{D}(A_d^i - B_d^j)^2} \tag{2}$$

The three components of DiRe are formulated as follows.

$$\text{CD} = l^c_{cos\_div} = S_{cos}(E^c_{syn}, E^c_{syn}) \tag{3}$$

$$\text{CDM} = l^c_{cos\_dm} = 1 - S_{cos}(E^c_{syn}, E^c_{real}) \tag{4}$$

$$\text{EDM} = l^c_{euc\_dm} = D_{euc}(E^c_{syn}, E^c_{real}) \tag{5}$$

Algorithm 1 presents our approach more formally.

Algorithm 1 SRe^{2}L with Diversity Regularizer (DiRe)

Require: Images per class ipc, feature extractor network h_θ, number of iterations T, learning rate η, cosine distance weight r_c, Euclidean distance weight r_e, Number of classes C

Ensure: Distilled dataset S_T

Initialize S_0 with $C \times ipc$ images drawn from a Gaussian noise distribution
Forward pass real data through h_θ, Store real embeddings E^c_{real} $\forall c \in \{0, \dots C-1\}$
for $t = 1$ to T **do**
    Forward pass synthetic data S_{t-1} through h_θ
    Obtain syn. embeddings E^c_{syn} $\forall c \in \{0, \dots C-1\}$
    Compute $l^c_{\text{cos\_div}}$, $l^c_{\text{cos\_dm}}$, and $l^c_{\text{euc\_dm}}$
    $L_{\text{syn}} = \sum_{c=1}^{C} r_c \cdot (l^c_{\text{cos\_div}} + l^c_{\text{cos\_dm}}) + r_e \cdot l^c_{\text{euc\_dm}}$
    $L_{\text{total}} \leftarrow L_{\text{ce}} + L_{\text{bn}} + L_{\text{syn}}$
    $S_t \leftarrow S_{t-1} - \eta \cdot \nabla_S L_{\text{total}}$
end for

3. Implementation of diversity regularizer

3.1. Embeddings w.r.t. feature extraction layer

We have used outputs of the penultimate layers (for example, the output of the Average Pool layer of ResNet-18) as the feature-rich, low-dimensional representation of the real and synthetic images. These embeddings are used to compute the components of the DiRe regularizer. For the ResNet-18 architecture, these result in 512-dimensional features.

3.2. Pairwise cosine similarity and pairwise Euclidean distance

The complexity of computing pairwise cosine similarity (Equation 1) and pairwise Euclidean distance (Equation 2) between two matrices of K rows (i.e., computation between embeddings of K images) is $\mathcal{O}(K^2)$. Carrying out these computations in a nested loop manner for every epoch and every class would be computationally very expensive. Instead, we utilize these functions from the torchmetrics library[1]. Because of their efficient implementation, they achieve $\approx 30X$

[1] https://lightning.ai/docs/torchmetrics/stable/pairwise/cosine_similarity.html

faster computation. The timing comparison is provided in the subsection 4.9.

4. Experiments and results

4.1. Experimental Setup

Applications. We have evaluated the performance of our method on image classification.

Datasets. For image classification, we evaluate the effectiveness of our regularizer on four popular benchmark image classification datasets, i.e., CIFAR-10, CIFAR-100 [22], Tiny ImageNet [23], and ImageNet-1K [31]. To test the robustness of our method, we consider all these datasets with the number of classes varying from 10 to 1000.

Backbone architecture. Similar to other SOTA works, we use the ResNet-18 [19] architecture to condense the datasets. Unless specified otherwise, we use the same trained model as a teacher model for carrying out knowledge distillation. We also use other CNN architectures, such as ResNet-50 [17], ResNet-101 [18], VGG-16 [37] and MobileNetV2 [33] and transformer architecture such as ViT [24] to carry out our cross-architecture generalization study. The main aim is to study the effect of improved diversity on generalization over the backbone architecture and cross-architecture generalization over various architectures.

Diversity Metrics. We consider Coverage [27], Vendi Score [11], and intra-class cosine similarity as the diversity metrics. **Coverage** measures the fraction of real samples whose neighbourhoods contain at least one synthetic sample. It is calculated as:

$$\text{coverage} = \frac{1}{N}\sum_{i=1}^{N} 1_{\left\{\exists\, j \text{ s.t. } Y_j \in B\left(X_i, \text{NND}_k(X_i)\right)\right\}} \quad (6)$$

where, X is the real dataset, Y is the synthetic dataset, NND_k represents k^{th} Nearest Neighbor Distance and $B(X, r)$ represents a ball of radius r around the data point X. All the calculations are carried out in the feature space, using embedding of the *avgpool* layer of the ResNet-18 architecture. Hence, a higher coverage value will indicate higher diversity among the synthetic images.

Vendi Score is defined as the exponential of the Shannon entropy of the eigenvalues of a similarity matrix. A higher Vendi Score indicates greater diversity in the dataset.

Cosine similarity between two vectors x and y is given as:

$$s_{cos}(x, y) = \frac{< x, y >}{||x|| \cdot ||y||} \quad (7)$$

We compute intra-class pairwise cosine similarity among the embeddings of the synthetic dataset and use the mean as a measure of diversity. Lower intra-class cosine similarity indicates higher diversity.

Baselines We have considered the results reported by the SRe^{2}L, CDA, DWA, UFC, G-VBSM, and DELT methods wherever available. We have reused the publicly available codebase for each method to generate results where they are unavailable. For evaluation purposes, the same hyperparameters are used for all the methods to ensure uniformity.

To showcase that our approach can be used with other optimization-based dataset condensation algorithms, we have used Matching Training Trajectory [1] and Distribution Matching [56]. While we have utilized the original codebase for MTT, the codebase [2] provided by DC-BenchMark [4] is used for DM algorithm.

Codebase We have released the anonymized codebase for easy reproduction of our results.[3]

4.2. Results for CIFAR10

A comparison of the accuracy values and diversity metrics obtained on a randomly initialized ResNet-18 architecture, trained through knowledge distillation from a pre-trained ResNet-18 model, is presented in Tables 2 and 3. It can be clearly seen that, addition of DiRe leads to improvement in generalization accuracy and diversity metrics.

Table 2. Impact on accuracy by addition of DiRe to various DC methods on CIFAR10 Dataset. It can be seen that the addition of DiRe has led to an increase in the accuracy values obtained by each of the methods considered.

Methods	IPC= 10	IPC= 50	IPC= 100
SRe^{2}L	27.2 ± 0.4	47.5 ± 0.5	57.5 ± 0.6
SRe^{2}L + DiRe	**37.4 ± 1.1**	**59.7 ± 1.2**	**71.2 ± 1.2**
DWA	32.6 ± 0.4	53.1 ± 0.3	67.2 ± 0.3
DWA + DiRe	**36.5 ± 0.9**	**62.2 ± 0.7**	**71.0 ± 0.6**
INFER	32.0 ± 0.5	60.4 ± 1.6	-
INFER + DiRe	**45.3 ± 0.8**	**73.9 ± 0.2**	-
INFER (D)	30.7 ± 0.3	60.7 ± 0.9	-
INFER(D) + DiRe	**57.1 ± 0.9**	**85.1 ± 0.3**	-
G-VBSM	53.5 ± 0.6	59.2 ± 0.4	-
G-VBSM + DiRe	**55.8 ± 0.2**	**68.3 ± 0.3**	-
DELT	43.0 ± 0.9	64.9 ± 0.9	-
DELT + DiRe	**49.2 ± 0.5**	**76.3 ± 0.2**	-

4.3. Results for CIFAR100

Tables 4 and 5 compare accuracy values and diversity metrics for various methods at various IPC settings. DiRe is able to imporve both accuracy and diversity metrics for all the DC methods considered.

The plot of classwise intra-class cosine similarity is shown in Figure 2, which further showcases the diversity

[2] https://github.com/justincui03/dc_benchmark

[3] https : / / anonymous . 4open . science / r / DiversifiedDistillation-1426

Table 3. Impact on diversity by the addition of DiRe to various DC methods on CIFAR10 Dataset. Addition of DiRe has resulted in generation of synthetic dataset with higher diversity.

Methods	Coverage ↑	Similarity ↓	Vendi ↑
SRe^{2}L	2.25%	0.90	1.87
SRe^{2}L + DiRe	**3.53%**	**0.86**	**2.25**
DWA	2.43%	0.88	2.20
DWA + DiRe	**2.84%**	**0.78**	**2.34**
INFER	2.97%	**0.74**	2.02
INFER + DiRe	**6.74%**	0.82	**2.24**
INFER (D)	2.97%	**0.74**	2.02
INFER(D) + DiRe	**6.74%**	0.82	**2.24**
G-VBSM	0.04%	0.76	2.07
G-VBSM + DiRe	**0.06%**	**0.69**	**2.67**
DELT	0.9%	**0.84**	1.69
DELT + DiRe	**4.3%**	0.93	**2.28**

Table 4. Impact on accuracy by addition of DiRe to various DC methods on CIFAR100 Dataset. It can be seen that the addition of DiRe has led to an increase in the accuracy values obtained by each of the methods considered.

Methods	IPC= 10	IPC= 50	IPC= 100
SRe^{2}L	31.6 ± 0.5	52.2 ± 0.3	57.5 ± 0.6
SRe^{2}L + DiRe	**41.2 ± 1.1**	**63.4 ± 0.2**	**66.5 ± 0.2**
DWA	39.6 ± 0.6	60.9 ± 0.5	65.2 ± 0.3
DWA + DiRe	**41.4 ± 0.4**	**62.3 ± 0.2**	**65.3 ± 0.2**
CDA	49.8 ± 0.6	64.4 ± 0.5	65.5 ± 0.1
CDA + DiRe	**54.5 ± 0.3**	**66.6 ± 0.1**	**68.0 ± 0.4**
INFER	45.2 ± 0.1	62.8 ± 0.4	66.3 ± 0.1
INFER + DiRe	**53.7 ± 1.5**	**67.6 ± 0.2**	**69.2 ± 0.3**
INFER (D)	53.4 ± 0.6	68.9 ± 0.1	73.3 ± 0.2
INFER(D) + DiRe	**63.9 ± 0.2**	**74.1 ± 0.1**	**76.1 ± 0.2**

introduced by our method. T-SNE plot for embeddings of the CIFAR100 dataset for IPC=50 setting is shown in Figure 1. As we can see, our method can cover diverse regions in the original data manifold compared to the SRe^{2}L and DWA methods.

4.4. Results for Tiny ImageNet

Tables 6 and 7 compares the accuracy values and diversity metrics for various methods on the Tiny ImageNet dataset trained on the ResNet-18 architecture. The proposed regularizer improves both generalization accuracy and diversity metrics for all the three condensation methods considered.

Table 5. Impact on diversity by the addition of DiRe to various DC methods on CIFAR100 Dataset. Addition of DiRe results in improved diversity across all the DC methods considered.

Methods	Coverage ↑	Similarity ↓	Vendi ↑
SRe^{2}L	14.87%	0.81	2.79
SRe^{2}L + DiRe	**23.12%**	**0.65**	**3.08**
DWA	19.32%	0.77	2.99
DWA + DiRe	**32.75%**	**0.61**	**3.11**
CDA	12.31%	0.81	2.43
CDA + DiRe	**15.43%**	**0.67**	**2.85**
INFER	14.30%	0.68	2.66
INFER + DiRe	**28.99%**	**0.60**	**2.70**
INFER (D)	14.30%	0.68	2.66
INFER(D) + DiRe	**28.99%**	**0.60**	**2.70**

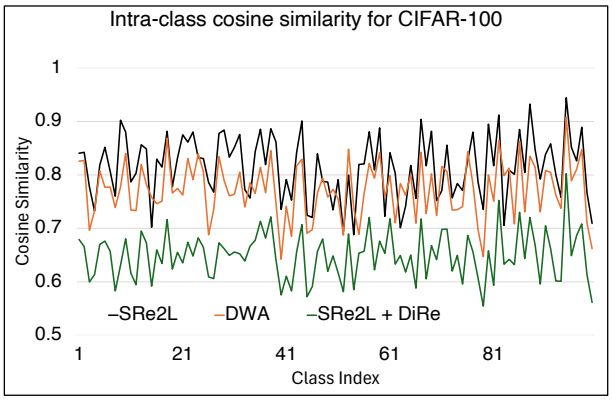

Figure 2. Classwise intra-class cosine similarity for CIFAR100 dataset. Lower cosine similarity indicates higher diversity among the synthetic dataset. SRe^{2}L + DiRe clearly shows lower cosine similarity for all the classes in the dataset as compared vanilla SRe^{2}L and DWA methods.

4.5. Results for ImageNet-1K

A comparison of accuracy values and diversity metrics on a synthetic dataset generated from ImageNet-1K is presented in Tables 8 and 9. The addition of the proposed regularizer results in an improvement in both accuracy and diversity metrics. Figure 3 showcases synthetic images belonging to 'Peacock' class condensed from ImageNet-1K dataset. Diversity among the synthetic images resulting from the addition of the proposed regularizer is clearly noticeable.

4.6. Cross architecture study

To evaluate the generalizability of the synthetic dataset generated by adding our proposed regularizer, we have compared the test set accuracies on various deep learning architec-

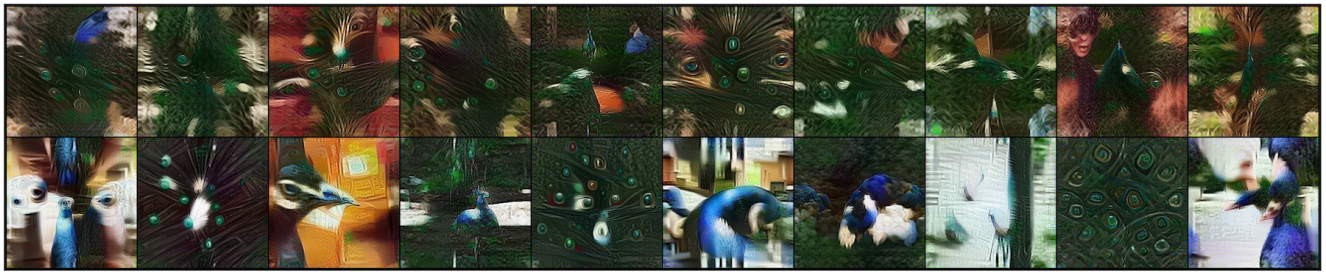

Figure 3. Visualization of synthetic images belonging to class 'Peacock' from ImageNet-1K dataset. The top row images are condensed through SRe^{2}L, and the bottom row images are condensed through SRe^{2}L + DiRe. The diversity among the synthetic images is clearly visible

Table 6. Impact on accuracy by addition of DiRe to various DC methods on Tiny ImageNet Dataset. It can be seen that the addition of DiRe has led to an increase in the accuracy values obtained by each of the methods considered.

Methods	IPC= 10	IPC= 50	IPC= 100
SRe^{2}L	17.7 ± 0.7	41.1 ± 0.4	49.7 ± 0.3
SRe^{2}L + DiRe	**34.7 ± 0.3**	**55.3 ± 0.4**	**57.4 ± 0.1**
DWA	32.1 ± 0.1	52.8 ± 0.2	56.0 ± 0.2
DWA + DiRe	**37.6 ± 0.3**	**55.2 ± 0.1**	**58.5 ± 0.2**
CDA	21.3 ± 0.3	48.7 ± 0.1	53.2 ± 0.1
CDA + DiRe	**34.8 ± 0.5**	**54.5 ± 0.2**	**56.5 ± 0.3**
G-VBSM	-	47.6 ± 0.3	51.0 ± 0.4
G-VBSM + DiRe	-	**50.1 ± 0.2**	**54.2 ± 0.5**
DELT	43.0 ± 0.1	55.7 ± 0.5	-
DELT + DiRe	**45.7 ± 0.5**	**56.8 ± 0.1**	-

Table 7. Impact on diversity by the addition of DiRe to various DC methods on Tiny ImageNet Dataset. Diversity metrics across the three different metrics have been imporved after addition of DiRe.

Methods	Coverage ↑	Similarity ↓	Vendi ↑
SRe^{2}L	30%	**0.66**	3.07
SRe^{2}L + DiRe	**45%**	**0.66**	**3.22**
DWA	36%	0.69	3.04
DWA + DiRe	**52%**	**0.65**	**3.14**
CDA	32%	0.75	6.41
CDA + DiRe	**53%**	**0.69**	**6.96**
G-VBSM	36%	0.71	2.63
G-VBSM + DiRe	**45.5%**	**0.66**	**3.61**
DELT	6.5 %	0.62	2.18
DELT + DiRe	**8.2%**	**0.59**	**2.43**

Table 8. Impact on accuracy by addition of DiRe to various DC methods on ImageNet-1K Dataset. It can be seen that the addition of DiRe has led to an increase in the accuracy values obtained by each of the methods considered.

Methods	IPC= 10	IPC= 50	IPC= 100
SRe^{2}L	21.3 ± 0.6	46.8 ± 0.2	52.8 ± 0.4
SRe^{2}L + DiRe	**38.5 ± 0.1**	**55.6 ± 0.3**	**59.2 ± 0.1**
DWA	37.9 ± 0.2	55.2 ± 0.2	59.2 ± 0.3
DWA + DiRe	**39.1 ± 0.4**	**56.9 ± 0.1**	**61.0 ± 0.1**
CDA	33.5 ± 0.3	52.5 ± 0.3	58.0 ± 0.2
CDA + DiRe	**35.6 ± 0.1**	**56.0 ± 0.1**	**60.3 ± 0.2**
INFER	28.7 ± 0.2	51.8 ± 0.2	-
INFER + DiRe	**38.2 ± 0.3**	**61.2 ± 0.5**	-
G-VBSM	31.4 ± 0.5	51.8 ± 0.4	55.7 ± 0.4
G-VBSM + DiRe	**35.1 ± 0.1**	**55.2 ± 0.2**	**58.7 ± 0.1**
DELT	46.1 ± 0.4	59.2 ± 0.4	-
DELT + DiRe	**47.3 ± 0.1**	**59.6 ± 0.2**	-

tures including CNN architectures from different families, ResNet50 [16], ResNet101 [19], MobileNetV2 [33], and VGG16 [37], and transformer architecture, ViT [9]. Comparisons of test set accuracies are tabulated in Tables 10, 11 and 12 for CIFAR-100, Tiny ImageNet and ImageNet-1K datasets, respectively. Our proposed regularizer is able to improve accuracy across various CNN and transformer architectures.

4.7. Impact of the proposed regularizer on other condensation methods

We have applied DiRe to MTT and DM to demonstrate its effectiveness on DC algorithms that are not decoupling-based. Tables 13 shows the accuracy results obtained on a ConvNet network on the CIFAR-10 and CIFAR-100 datasets for various IPC settings with MTT algorithm. Results for DM are given in the supplementary document.

Table 9. Impact on diversity by the addition of DiRe to various DC methods on ImageNet-1K Dataset. The diversity scores across all three metrics improved with the addition of DiRe.

Methods	Coverage ↑	Similarity ↓	Vendi ↑
SRe^{2}L	2.0%	0.82	4.41
SRe^{2}L + DiRe	**6.4%**	**0.66**	**5.94**
DWA	2.2%	0.78	5.09
DWA + DiRe	**3.7%**	**0.65**	**5.79**
CDA	4.1%	0.80	5.15
CDA + DiRe	**7.9%**	**0.59**	**6.18**
INFER	6.5%	0.71	7.45
INFER + DiRe	**8.3%**	**0.64**	**8.68**
G-VBSM	4.1%	0.81	7.87
G-VBSM + DiRe	**11.6%**	**0.67**	**13.2**
DELT	2.3%	0.83	4.28
DELT + DiRe	**6.5%**	**0.65**	**6.23**

Table 10. Comparison of cross-architecture generalization performance on CIFAR100 dataset generated by ResNet-18.

Target Architecture	SRe^{2}L	SRe^{2}L + DiRe
IPC=50		
ResNet50	52.8 ± 0.7	**63.8 ± 0.6**
ResNet101	51.4 ± 2.6	**64.2 ± 0.3**
VGG16	40.4 ± 1.2	**53.4 ± 0.4**
MobileNetV2	43.2 ± 0.2	**56.7 ± 0.1**
ViT	16.2 ± 0.2	**35.0 ± 0.5**
IPC=100		
ResNet50	59.5 ± 0.5	**67.3 ± 0.2**
ResNet101	59.2 ± 0.9	**67.7 ± 0.1**
VGG16	51.8 ± 0.4	**62.2 ± 0.4**
MobileNetV2	54.6 ± 0.5	**64.1 ± 0.3**
ViT	23.3 ± 0.4	**46.6 ± 0.9**

4.8. Ablation study on impact of various components of DiRe

We analyzed the impact of individual components in the DiRe regularizer and their combinations. Diversity metrics were computed for a synthetic dataset using all the 7 possible combinations. Table 14 compares the normalized impact of various components of DiRe and their combinations on test set accuracy and different diversity metrics for the Tiny ImageNet dataset with IPC = 10. As can be seen, CD mainly affects cosine similarity, CDM primarily influences the Vendi score, and EDM has the strongest impact on the coverage score. Combining all three components together results in balanced performance across all diversity metrics.

Table 11. Comparison of cross-architecture generalization performance on Tiny ImageNet dataset generated by ResNet-18.

METHOD	VGG16	VIT
IPC=50		
SRE^{2}L	32.50 ± 0.45	14.79 ± 0.40
DWA	32.90 ± 0.85	18.79 ± 0.28
CDA	33.50 ± 0.20	16.35 ± 0.51
SRE^{2}L + DIRE	**34.45 ± 0.56**	**19.00 ± 0.10**
IPC=100		
SRE^{2}L	42.80 ± 0.46	22.95 ± 0.75
DWA	43.58 ± 0.12	26.46 ± 0.21
CDA	41.81 ± 0.35	22.65 ± 0.70
SRE^{2}L + DIRE	**45.21 ± 0.84**	**28.20 ± 0.33**

Table 12. Comparison of cross-architecture generalization performance on ImageNet-1K dataset generated by ResNet-18. Addition of DiRe is able to maintain performance gain in terms of accuracy across various architectures.

METHOD	MOBILENETV2	SHUFFLENET
IPC=10		
SRE^{2}L	15.4 ± 0.2	9.0 ± 0.7
DWA	29.1 ± 0.3	**11.4 ± 0.6**
CDA	25.1 ± 0.5	8.5 ± 0.7
SRE^{2}L + DIRE	**30.6 ± 0.3**	10.5 ± 0.3
IPC=50		
SRE^{2}L	48.3 ± 0.5	9.0 ± 0.6
DWA	51.6 ± 0.5	28.5 ± 0.5
CDA	49.4 ± 0.3	25.8 ± 0.4
SRE^{2}L + DIRE	**53.0 ± 0.2**	**30.0 ± 0.3**

4.9. Timing analysis

Table 15 compares the time taken by various methods to synthesize IPC=50 images from CIFAR-100. The increase in time by our method is $\approx 17\%$ higher than SRe^{2}L and $\approx 5\%$ higher than DWA, which, in our opinion, is not very significant compared to the enhancement in generalization and diversity achieved by our method. Comparative analysis of compute requirement is provided in the supplementary.

4.10. Calculation of coverage metric through CLIP embeddings

For completeness and to avoid any bias introduced by a specific backbone, we further evaluated coverage using

Table 13. Comparison of accuracies on ConvNet architecture with MTT Dataset Condensation method. The proposed regularizer is able to improve the performance of the non-decoupled algorithm across various IPC settings.

IPC	MTT	MTT + DiRe
CIFAR-10		
1	46.3 ± 0.8	**49.1 ± 0.2**
10	65.3 ± 0.7	**67.4 ± 0.3**
50	71.6 ± 0.2	**72.4 ± 0.1**
CIFAR-100		
1	24.3 ± 0.3	**27.6 ± 0.7**
10	40.1 ± 0.4	**42.0 ± 0.2**
50	47.7 ± 0.2	**48.9 ± 0.3**

Table 14. Normalized impact of various components of DiRe and their combinations on accuracy and different diversity metrics for Tiny ImageNet dataset on IPC=10 setting. Ven, Sim, and Cov stand for Vendi Score, Cosine Similarity, and Coverage, respectively.

Methods	Ven↑	Sim↓	Cov↑	Acc
CD	0.50	0.00	0.41	33.8 ± 0.4
CDM	0.74	0.30	0.46	33.8 ± 0.2
EDM	0.00	0.98	1.00	34.2 ± 0.3
CD + CDM	0.48	0.17	0.00	34.1 ± 0.6
CD + EDM	0.40	0.93	0.01	33.6 ± 0.1
CDM + EDM	0.36	1.00	0.92	33.6 ± 0.5
CD + CDM + EDM	1.00	0.07	0.80	34.7 ± 0.3

Table 15. Comparison of total time taken for synthesizing IPC=50 images from the CIFAR100 dataset by various methods.

Method	Time (in seconds)
SRe^{2}L	1342.6
DWA	1493.1
SRe^{2}L + DiRe	1579.9

embeddings extracted from a pretrained CLIP model [30] (ViT-B/32, trained on the LAION-2B dataset). Unlike the ResNet-18 features used in our primary experiments, CLIP embeddings are obtained via a vision–language pretraining paradigm, which has been shown to produce semantically aligned and domain-robust representations. Comparative analysis is provided in Table 16.

This additional evaluation confirms that our conclusions remain consistent even when employing a feature extractor with a substantially different training objective and representational bias.

Table 16. Comparison of coverage values obtained by ResNet-18 and CLIP pre-trained models. Results are obtained on the Tiny ImageNet dataset with synthetic data generated through a ResNet-18 architecture.

Method	ResNet-18	CLIP
SRe^{2}L	0.301	0.064
SRe^{2}L + DiRe	0.453	0.088

5. Discussion on optimization-free algorithms

As has been mentioned in sec. 1, our proposed regularizer is not applicable for optimization-free algorithms such as RDED. However, for completeness, we compare the performance of DiRe with SRe^{2}L against RDED in terms of various diversity metrics on ImageNet-1K dataset. As RDED employs different hyperparameters compared to SRe^{2}L, results presented in Table 17 are obtained by the SRe^{2}L + DiRe method with the same hyperparameter settings as RDED. As can be observed, SRe^{2}L + DiRe performs better than RDED across all three diversity metrics.

Table 17. Comparative analysis of diversity metrics for synthetic data generated through RDED and SRe^{2}L + DiRe from ImageNet-1K dataset. Addition of DiRe to SRe^{2}L outperforms RDED across the three diversity metrics.

Method	Coverage↑	Similarity↓	Vendi↑
RDED	0.40	0.71	4.54
SRe^{2}L + DiRe	**0.45**	**0.66**	**6.17**

6. Conclusion

Despite its importance, diversity has not been considered earnestly in the existing works. This paper emphasizes the significance of 'diversity' during dataset condensation. It introduces an intuitive, yet powerful diversity regularizer **DiRe**, which can be added off-the-shelf to existing optimization-based dataset condensation algorithms with a separate synthesis stage. It is the first to study diversity in dataset condensation in a quantitative manner. Through extensive experimentation, we demonstrate that DiRe can improve the accuracy of the SOTA dataset condensation methods while achieving higher diversity across various diversity metrics. We also demonstrate that the improved diversity helps to achieve better cross-architecture generalization on various deep learning architectures, including transformers.

Acknowledgements

Most of this work was done while Aravind Reddy was at the Department of Artificial Intelligence, Indian Institute of Technology Hyderabad.

References

[1] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 4

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

[3] Ming-Yu Chung, Sheng-Yen Chou, Chia-Mu Yu, Pin-Yu Chen, Sy-Yen Kuo, and Tsung-Yi Ho. Rethinking backdoor attacks on dataset distillation: A kernel method perspective. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[4] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. DC-BENCH: Dataset condensation benchmark. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. 4

[5] Justin Cui, Ruochen Wang, Yuanhao Xiong, and Cho-Jui Hsieh. Ameliorate spurious correlations in dataset condensation. In *Forty-first International Conference on Machine Learning*, 2024. 1

[6] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks. In *Advances in Neural Information Processing Systems*, 2022. 1

[7] Jianrong Ding, Zhanyu Liu, Guanjie Zheng, Haiming Jin, and Linghe Kong. CondTSF: One-line plugin of dataset condensation for time series forecasting. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[8] Tian Dong, Bo Zhao, and Lingjuan Lyu. Privacy for free: How does dataset condensation help privacy? In *Proceedings of the 39th International Conference on Machine Learning*, 2022. 1

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[10] Jiawei Du, Xin Zhang, Juncheng Hu, Wenxin Huang, and Joey Tianyi Zhou. Diversity-driven synthesis: Enhancing dataset distillation through directed weight adjustment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 2, 3

[11] Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. 2, 4

[12] Zhiqiang Gong, Ping Zhong, and Weidong Hu. Diversity in machine learning. *IEEE Access*, 7:64323–64350, 2019. 2

[13] Jianyang Gu, Kai Wang, Wei Jiang, and Yang You. Summarizing stream data for memory-constrained online continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. 1

[14] Ziyao Guo, Kai Wang, George Cazenavette, HUI LI, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[15] Sariel Har-Peled and Soham Mazumdar. On coresets for k-means and k-median clustering. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 291–300, 2004. 1

[16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2015. 6

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. 4

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016. 4, 6

[20] Chun-Yin Huang, Kartik Srinivas, Xin Zhang, and Xiaoxiao Li. Overcoming data and model heterogeneities in decentralized federated learning via synthetic anchors. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. 1

[21] Eeshaan Jain, Tushar Nandy, Gaurav Aggarwal, Ashish V. Tendulkar, Rishabh K Iyer, and Abir De. Efficient data subset selection to generalize training across models: Transductive and inductive networks. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[22] Alex Krizhevsky. Learning multiple layers of features from tiny images. pages 32–33, 2009. 4

[23] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. 2015. 4

[24] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets, 2021. 4

[25] Yongmin Lee and Hye Won Chung. Selmatch: Effectively scaling up dataset distillation via selection-based initialization and partial updates by trajectory matching. In *Forty-first International Conference on Machine Learning*, 2024. 1

[26] Noel Loo, Ramin Hasani, Mathias Lechner, Alexander Amini, and Daniela Rus. Understanding reconstruction attacks with the neural tangent kernel and dataset distillation. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[27] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable fidelity and diversity

metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 2, 4

[28] Ding Qi, Jian Li, Jinlong Peng, Bo Zhao, Shuguang Dou, Jialin Li, Jiangning Zhang, Yabiao Wang, Chengjie Wang, and Cairong Zhao. Fetch and forge: Efficient dataset condensation for object detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[29] Tian Qin, Zhiwei Deng, and David Alvarez-Melis. A label is worth a thousand images in dataset distillation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamila Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 8

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge, 2015. 4

[32] Noveen Sachdeva and Julian McAuley. Data distillation: A survey. *Transactions on Machine Learning Research*, 2023. Survey Certification. 1

[33] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. MobileNetV2: Inverted Residuals and Linear Bottlenecks . In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4, 6

[34] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16709–16718, 2024. 2

[35] Shitong Shao, Zikai Zhou, Huanran Chen, and Zhiqiang Shen. Elucidating the design space of dataset condensation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. 1

[36] Zhiqiang Shen, Ammar Sherif, Zeyuan Yin, and Shitong Shao. Delt: A simple diversity-driven earlylate training for dataset distillation. *arXiv preprint arXiv:2411.19946*, 2024. 2

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 4, 6

[38] Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, and Jeffrey Clune. Generative teaching networks: Accelerating neural architecture search by learning to generate synthetic training data. In *Proceedings of the 37th International Conference on Machine Learning*, 2020. 1

[39] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9390–9399, 2024. 2

[40] Grant Van Horn, Elijah Cole, Sara Beery, Kimberly Wilber, Serge Belongie, and Oisin Mac Aodha. Benchmarking representation learning for natural world image collections. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12884–12893, 2021. 2

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

[42] Kai Wei, Rishabh Iyer, and Jeff Bilmes. Submodularity in data subset selection and active learning. In *Proceedings of the 32nd International Conference on Machine Learning*, 2015. 1

[43] Xiaobo Xia, Jiale Liu, Jun Yu, Xu Shen, Bo Han, and Tongliang Liu. Moderate coreset: A universal method of data selection for real-world data-efficient deep learning. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[44] Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. Feddm: Iterative distribution matching for communication-efficient federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16323–16332, 2023. 1

[45] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1686–1697, 2021. 2

[46] Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. In *The Eleventh International Conference on Learning Representations*, 2023. 1

[47] William Yang, Ye Zhu, Zhiwei Deng, and Olga Russakovsky. What is dataset distillation learning? In *Forty-first International Conference on Machine Learning*, 2024. 1

[48] Yu Yang, Hao Kang, and Baharan Mirzasoleiman. Towards sustainable learning: Coresets for data-efficient deep learning. In *Proceedings of the 40th International Conference on Machine Learning*, 2023. 1

[49] Zeyuan Yin and Zhiqiang Shen. Dataset distillation via curriculum data synthesis in large data era. *Transactions on Machine Learning Research*, 2024. 2

[50] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2, 3

[51] Xin Zhang, Jiawei Du, Yunsong Li, Weiying Xie, and Joey Tianyi Zhou. Spanning training progress: Temporal dual-depth scoring (tdds) for enhanced dataset pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26223–26232, 2024. 1

[52] Xin Zhang, Jiawei Du, Ping Liu, and Joey Tianyi Zhou. Breaking class barriers: Efficient dataset distillation via inter-class feature compensator. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[53] Yihua Zhang, Yimeng Zhang, Aochuan Chen, Jinghan Jia, Jiancheng Liu, Gaowen Liu, Mingyi Hong, Shiyu Chang, and Sijia Liu. Selectivity drives productivity: Efficient dataset

pruning for enhanced transfer learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1

[54] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of the 38th International Conference on Machine Learning*, 2021. 1

[55] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023. 1, 2

[56] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2021. 1, 4

[57] Dora Zhao, Jerone Andrews, Orestis Papakyriakopoulos, and Alice Xiang. Position: Measure dataset diversity, don't just claim it. In *Forty-first International Conference on Machine Learning*, 2024. 2

[58] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023. 1

[59] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *Advances in Neural Information Processing Systems*, 2022. 1