Dataset Distillation Enhancing Long-Tail Learning: A Two-Stage Classification Strategy

BOYANG ZHENG*, Shanghai Jiao Tong University, China
LAI JIANG*, Shanghai Jiao Tong University, China
QUANQUAN PENG*, Shanghai Jiao Tong University, China
XUANCHANG ZHANG*, Shanghai Jiao Tong University, China

This study addresses the challenge of long-tailed distributions in machine learning, where few categories are over-represented, and many are under-represented. We focus on datasets like CIFAR10-LT, CIFAR100-LT, and ImageNet-LT, which mimic these distributions. Traditional methods like re-sampling and re-weighting, though helpful, have limitations such as high computational costs. We propose a novel approach: applying dataset distillation to long-tailed datasets. This technique condenses large datasets into smaller, synthesized versions, ensuring efficient training with less computational demand. Our contributions are threefold. Firstly, we investigate the efficacy of dataset distillation in long-tail learning scenarios, establishing a foundational benchmark by evaluating performance on the CIFAR10-LT dataset. This research serves as a baseline for future explorations in this domain. Secondly, we delve into a comprehensive analysis of dataset distillation in long-tailed learning, proposing the 'grouping hypothesis'. This hypothesis sheds light on the nuances of data representation and model learning in imbalanced data settings, offering a new perspective on data interpretation within deep neural networks. Finally, we introduce a novel two-stage Classification strategy and a corresponding two-stage hierarchical Training method based on dataset distillation principles. Our approach not only demonstrates an improvement in classification accuracy but also significantly mitigates inter-class bias, addressing the challenges posed by long-tail distributions. Importantly, this methodology can be seamlessly integrated as a plugin to enhance future dataset distillation strategies in similar learning environments. The insights and methodologies presented in this paper hold substantial potential for advancing the state-of-the-art in dataset distillation and long-tailed learning.

CCS Concepts: • **Computing methodologies** → **Computer vision**; **Neural networks**; **Supervised learning**.

Additional Key Words and Phrases: Image Classification, Dataset Distillation, Long-tailed Learning

1 INTRODUCTION

In the contemporary landscape of data analysis and machine learning, we encounter the phenomenon of 'long-tailed' distributions frequently [15]. These distributions are characterized by a plethora of occurrences in certain categories (the "head"), followed by a long, tapering tail of rare occurrences in many others. This pattern is not only ubiquitous in real-world data but also poses a significant challenge in machine learning, particularly in the context of data imbalance where common categories are over-represented, and rare ones are under-represented.

To specifically address this challenge, datasets such as CIFAR10-LT, CIFAR100-LT, and ImageNet-LT have been introduced. These are adaptations of their standard counterparts, designed to mimic long-tail distributions. They serve as crucial benchmarks for developing and testing algorithms capable of learning from imbalanced data.

Historically, techniques to tackle the long-tail problem have varied, including methods like re-sampling [11, 20, 26], re-weighting [1, 5, 7] and transfer learning [14, 21]. More recently, knowledge distillation has been introduced to solve long-tailed learning problems. This approach harnesses the knowledge from expert models to guide the learning process in long-tailed distributions[9, 13, 22, 23]. However, the performance of dataset distillation, a techinique closely related to knowledge distillation, remains largely unexplored in the context of long-tailed learning. Dataset distillation involves

*All authors contributed equally to this research. The order of authors is alphabetical.

condensing a large dataset into a smaller, synthesized version, capturing the essential information. A natural question to ask is, does dataset distillation helps long-tailed learning?

Our study addresses this inquiry, exploring the performance of dataset distillation within the realm of long-tail learning. Through detailed analysis, we introduce the 'grouping hypothesis': we postulate that classification models initially categorize images into ambiguous groups composed of minor categories, followed by further subdivision into specific minor classes. Stemming from this observation, we propose a two-stage classification strategy. Building upon this, we introduce a novel method of two-stage hierarchical training based on dataset distillation. Experimental results on CIFAR10-LT substantiate that our methodology not only enhances accuracy but also mitigates the bias introduced by long-tailed distributions. This indicates that our approach is capable of resolving long-tail issues. Furthermore, our method can serve as a plugin, potentially augmenting the performance of future dataset distillation endeavors in long-tail learning.

Our research introduces a novel approach: the application of dataset distillation [18] techniques on long-tailed datasets.

Our contributions in this research are threefold:

- We explore the efficacy of dataset distillation in the context of long-tailed learning and establishd a baseline performance on the CIFAR10-LT dataset. This investigation serves as a foundational benchmark for future research in this area.
- we conduct a detailed analysis of dataset distillation in long-tailed learning scenarios and purpose 'grouping hypothesis'. This hypothesis provides insights into data representation and model learning in data-imbalanced context, offering a novel perspective on how data is interpreted in the aspect of deep neural networks.
- Stemming from our analysis, we propose a two-stage classification strategy and introduce a two-stage Hierarchical Training method based on dataset distillation. Our approach not only improves classification accuracy but also reduces inter-class bias, thereby addressing the challenges inherent in long-tail distributions. Moreover, our method can be employed as a plugin, enhancing the performance of future dataset distillation strategies in long-tail learning environments.

2 BACKGROUND

In this section, we'll introduce two main components of our work: Dataset Distillation and Long-tailed Learning.

2.1 Dataset Distillation

Dataset distillation [18] is a technique in machine learning that focuses on condensing large datasets into smaller, highly informative subsets. This process, akin to the principles outlined in [10], aims to retain essential information for effective model training while reducing computational demands. The evolution of dataset distillation has been marked by significant methodologies, starting from foundational works on knowledge transfer to more advanced techniques. Most dataset distillation approaches can be classified into four categories: Meta-model Matching([18] [16]), Gradient Matching [12, 25]), Trajectory Matching([2, 6]) and distribution matching([17, 24]). Meta-model Matching aims to finds a compact, distilled set from the training set that minimizes the training loss of the model on the distilled dataset. Gradient matching based methods aim to generate a compact set so that the gradients it produces during training are similar to those generated by the full dataset. Trajectory Matching focuses on aligning the learning trajectory of a model trained on a distilled dataset with that of a model trained on the full dataset, thereby ensuring that the distilled

dataset preserves the essential characteristics of the full dataset. Distribution Matching focuses on preserving the underlying data distribution characteristics such as mean, variance, and higher-order moments.The primary objective of this method is to create a condensed dataset where the distribution of features and patterns closely resembles that of the full dataset. In our paper, we utilize Gradient Matching [25] and Matching Training Trajectories [2] as our base method for dataset distillation. The detail of Gradient Matching and Matching Training Trajectories is formulated in 3.1.

2.2 Long-tailed Learning

Modern real-world large-scale datasets often have long-tailed label distributions, which means a few classes account for most of the data, while most classes are under-represented. Such as medical diagnosis and autonomous driving. Most existing algorithms for learning imbalanced datasets can be divided into three categories: re-sampling [11, 20, 26], re-weighting [1, 5, 7] and transfer learning [14, 21]. In re-sampling, the number of examples is directly adjusted by over-sampling (adding repetitive data) for the minor class or under-sampling (removing data) for the major class, or both. Over-sampling adds repeated samples from minor classes, which could cause the model to overfit. To solve this, novel samples can be either interpolated from neighboring samples [3] or synthesized [8, 27] for minor classes. However, the model is still error-prone due to noise in the novel samples. It was argued that even if oversampling incurs risks from removing important samples, under-sampling is still preferred over over-sampling. Re-weighting methods assign weights on different training samples based on the label or instance and have shown significant accuracy increments over few-shot classes.

2.3 Distillation on Long-tailed Learning

Knowledge distillation has been recently introduced to the long-tailed recognition area. LFME [22] divides the entire long-tailed dataset into subsets with a smaller imbalance to train expert models and then distill knowledge into a unified student model. RIDE [19] applies knowledge distillation from a model with more experts to a model with fewer experts for further advancements. SSD [13] and DIVE [9] exploit self-supervision and power mormalization, respectively, to obtain a flatter label distribution as teacher signals.

In this paper, we introduce data distillation to long-tailed recognition area. Data distillation aims to distill the knowledge from a given dataset into a terse data summary. The goal is to reduce model storage size rather than reducing the training time or increasing the sample-fidelity.

3 METHOD

3.1 Preliminaries

In this section we'll give detailed formulation for the distillation methods we utilized i.e. Gradient Matching[25] and Matching Training Trajectories[2].

Gradient Matching(GM) The objective of gradient matching in dataset distillation is to ensure that the gradients of the loss with respect to the model parameters, computed on the distilled dataset, closely match those computed on the original dataset. This is formalized as:

$$\min_{\mathcal{D}_{\text{distill}}} \mathbb{E}_{\theta\in\Theta}\left[\left\|\nabla_\theta \sum_{t=1}^{T}\mathcal{L}(f_\theta(x_t),y_t)-\nabla_\theta \sum_{t=1}^{T}\mathcal{L}(f_\theta(x'_t),y'_t)\right\|^2\right] \tag{1}$$

Here,

- $\mathbb{E}_{\theta \in \Theta}$ denotes the expectation over the distribution of model parameters θ.
- $\mathcal{L}(f_\theta(x_t), y_t)$ and $\mathcal{L}(f_\theta(x'_t), y'_t)$ represent the loss functions computed at each timestep t over the original dataset $\mathcal{D}$ and the distilled dataset $\mathcal{D}_{\text{distill}}$, respectively.
- T is the number of timesteps considered.
- The goal is to minimize the average gradient discrepancy across all possible model parameters θ to ensure that the distilled dataset captures the essential learning characteristics of the original dataset.

Matching Training Trajectories(MTT) The objective of Matching Training Trajectories (MTT) in dataset distillation is to align the learning trajectories of models trained on the distilled dataset $\mathcal{D}_{\text{distill}}$ with that of models trained on the original dataset $\mathcal{D}$. This is formalized as an optimization problem:

$$\min_{\mathcal{D}_{\text{distill}}, \eta} \mathbb{E}_{\theta \in \Theta} \left[\sum_{t=0}^{T-M} \frac{\mathbf{D}(\theta^{\mathcal{D}}_{t+M}, \theta^{\mathcal{D}_{\text{distill}}}_{t+N})}{\mathbf{D}(\theta^{\mathcal{D}}_{t+M}, \theta^{\mathcal{D}}_{t})} \right] \tag{2}$$
$$\text{s.t. } \theta^{\mathcal{D}_{\text{distill}}}_{t+i+1} \leftarrow \theta^{\mathcal{D}_{\text{distill}}}_{t+i} - \eta \cdot \nabla_\theta \mathcal{L}_{\mathcal{D}_{\text{distill}}}(\theta^{\mathcal{D}_{\text{distill}}}_{t+i}); \ \theta^{\mathcal{D}_{\text{distill}}}_{t+1} \leftarrow \theta^{\mathcal{D}}_{t} - \eta \cdot \nabla_\theta \mathcal{L}_{\mathcal{D}_{\text{distill}}}(\theta^{\mathcal{D}}_{t})$$

Here,

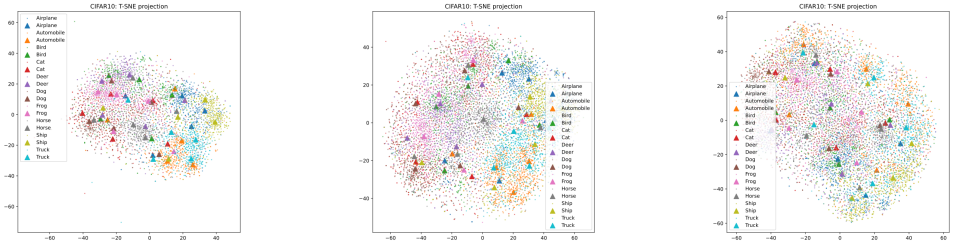
- $\mathbb{E}_{\theta \in \Theta}$ represents the expectation over the distribution of model parameters θ.
- $\mathbf{D} : \mathbb{R}^{|\theta|} \times \mathbb{R}^{|\theta|} \mapsto \mathbb{R}$ is a distance metric of choice (typically L2 distance).
- T is the number of timesteps considered, representing different stages of training.
- The goal is to minimize the discrepancy between the outputs of models trained on the original and distilled datasets at each timestep, thereby ensuring that the distilled dataset effectively replicates the learning trajectory of the original dataset.

3.2 Dataset Distillation on Long-tailed Learning

While knowledge distillation is commonly used in long-tailed learning. Using dataset distillation for long-tailed learning remains unexplored. As dataset distillation condenses information of a specific class in a few images. We could expect a more balanced distribution from a distilled long-tailed dataset(denoted as distilled dataset). We implemented Gradient Matching(GM) and Matching Training Trajectories(MTT) as our baseline methods. More experiment details can be found at Tab. 1. The results in Tab. 1 show that though dataset distillation does not increase the accuracy, it indeed reduces the variance of accuracy among different classes.

3.3 Grouping Hypothesis

To further investigate the impact of dataset distillation on long-tailed learning, we visualize the t-SNE result on the final layer features, following [4]. The visualization can be seen at Fig. 1 . The t-SNE visualization distinctly revealed that certain classes exhibit close proximity in their feature representation, while others are markedly separated. This phenomenon suggests a varying degree of similarity and dissimilarity among the classes based on their intrinsic features. To further elucidate this observation, a k-means clustering experiment was conducted. The clustering algorithm gathers each class of the CIFAR-10 dataset into several discrete centers, based on the final layer feature. We then partioned ten classes into two groups by analyzing the euclidean distance of the clustered centers. This grouping process is formulated as follow.



(a) Balanced Dataset (b) Original Dataset (c) Distilled Dataset

Fig. 1. Data distribution of test images acquired from three distinct datasets—balanced, original, and distilled. The balanced dataset is derived from the test dataset, ensuring an equitable distribution of images across each class. Utilizing a ConvNet3 model, we conduct pretraining on each dataset, generating a 2048-dimensional embedding. Following this, the embedding is projected onto a 2-dimensional plane using t-SNE with a perplexity hyperparameter set to 50, visually represented as dots in the graph. Subsequently, KMeans is applied to obtain 5 cluster centers for each class, visually represented as triangles in the graph.

Consider the CIFAR-10 dataset with $N(N = 10)$ classes. For each class i, let $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ represent the feature vectors corresponding to that class, where n_i is the number of instances in class i.

For each class i, we perform k-means clustering to partition the data X_i into M clusters. Let $C_i = \{c_{i1}, c_{i2}, \ldots, c_{iM}\}$ represent the centroids for each of the M clusters for class i.

The objective is to minimize the within-cluster sum of squares for each class, defined as:

$$\text{WCSS}_i = \sum_{j=1}^{M} \sum_{x \in S_{ij}} \|x - c_{ij}\|^2$$

where S_{ij} is the set of points in the j-th cluster of class i.

After obtaining the M centroids for each class, we proceed to group the classes based on these centroids.

Define a distance measure between the centroids of two different classes. For classes i and k, the distance between their centroids can be calculated as:

$$D_{ik} = \sqrt{\sum_{j=1}^{M} \sum_{l=1}^{M} \|c_{ij} - c_{kl}\|^2}$$

We then partition the N classes into groups. Intuitively, the groups partitioned should be far away from each other. Specifically, we partition the class into two groups $G_1 = \{g_1, \cdots g_k\}, G_2 = [N]/G_1$ that the sum of Euclidean distance between the centroids of the two group is maximized, that is:

$$G_1, G_2 = \max_{G_1, G_2} \sum_{i \in G_1, j \in G_2} D_{ij} \tag{3}$$

We run the grouping process on three version of CIFAR-10: balanced, long-tailed and distilled. The grouping result is stable across those datasets, indicating that the grouping process grasp some inner feature of different classes. Intuitively, two classes in one group indicate that they look "similar" in the aspect of a deep neural classifier. Therefore, it's natural to leverage this observation for better image classification.

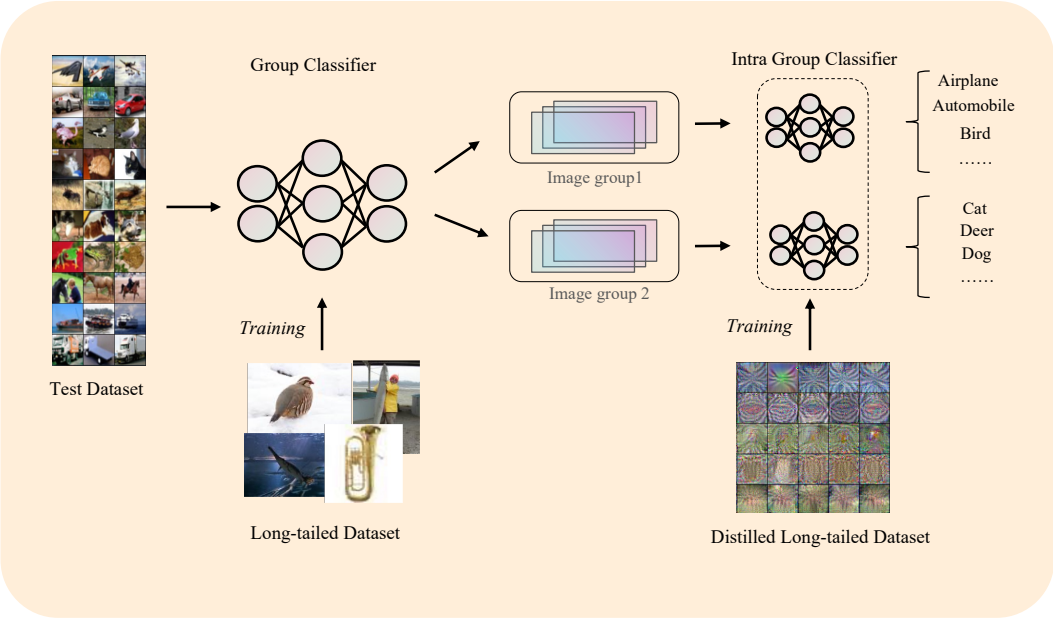


Fig. 2. Pipeline of two-stage training. The image is first classified into groups by the group classifier. Based on the given group, the corresponding intra-group classifier classifies the image into classes.

3.4 A Two-stage Image Classification Pipeline

We approach the classification of images from the CIFAR-10 dataset through a two-stage hierarchical model. Initially, we train a group classifier based on the groups (G_1, G_2) defined in equation 3. This classifier, denoted as $f_{group} : X \rightarrow G$, is tasked with assigning an input image from the set X to one of the predefined groups in the set G.

Following the group classification, we focus on the second stage, where we train specific intra-group class classifiers for each group G_i. These classifiers, represented as $f^i_{class} : G_i \rightarrow C_i$, categorize images within a group into distinct classes, where C_i denotes the set of classes within the ith group. A visualization of our pipeline can be seen at Fig. 2

The classification process follows a two-step pipeline. An input image is first processed by the group classifier f_{group} to determine its group. Subsequently, it is classified into a specific class within this group by the relevant intra-group class classifier f^i_{class}. This structured methodology leverages group-based initial sorting to potentially enhance the efficiency and accuracy of the classification system, particularly in dealing with large and diverse datasets.

3.5 Hierarchical Two-stage Training

To sufficiently utilize information of the long-tailed dataset, we train the group classifier on the full dataset. Thus, the group classifier can learn detailed enough feature for classification. It's worth noticing that training on imbalanced dataset won't brought biases to the group classifier. This observation arises from the understanding that the composition of a group and the distribution of classes in a long-tail context are mutually independent. Consequently, the presence of a long-tail distribution does not necessarily result in an imbalanced distribution within the group. Therefore, the

Table 1. Performance of the ConvNet3 Model directly trained on original and distilled datasets(10 images per class). We use CIFAR10-LT (r=100) for all experiments. Dataset: Type of training dataset, Train Img Num: Total number of images used to train, Total Acc: Overall accuracy on the test dataset(the higher the better), Acc Std: Standard deviation of accuracy across different classes(the lower the better), Cls Acc: Accuracy of each class(the higher the better).

(a) Summary Statistics

Dataset	Train Img Num	Total Acc (%) ↑	Acc Std (%) ↓
Original	2478	**37.6**	32.4
GM Distilled	100	36.2	**11.8**
MTT Distilled	100	22.0	19.2

(b) Class-wise Accuracy

Dataset	Cls Acc (%) ↑									
	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Original	**90.6**	**90.4**	**62.4**	**51.2**	**32.4**	18.2	15.4	11.8	3.7	0.3
GM Distilled	47.3	63.0	32.4	36.9	31.2	**33.1**	**29.8**	**35.4**	**38.1**	**15.1**
MTT Distilled	40.9	64.9	40.2	17.7	16.4	11.2	10.5	13.8	2.8	1.8

distribution of samples for groups is balanced, which is consistent with the distribution in evaluation. This consistency plays a crucial role in mitigating potential biases in the group classifier.

In our approach, we addressed the issue of bias introduction among classes when training intra-group classifiers directly on long-tail datasets. Dataset distillation has been proven effective in mitigating such biases. Therefore, we employed distilled data for training our intra-group classifiers. This strategic choice aimed to reduce the skewness inherent in long-tail distributions, ensuring a more balanced and unbiased learning process.

Results in Tab. 2 show that our classification methodology presents advantages over baseline approaches that train classifiers directly on distilled datasets or long-tailed datasets. We believe the essence of our approach lies in its capacity to harness the comprehensive information available in the entire dataset while concurrently leveraging the benefits of data distillation to eliminate bias.

Contrary to methods that solely focus on distilled or long-tailed datasets, our strategy ensures a more balanced and representative utilization of data. This balanced approach is particularly effective in mitigating the biases that often arise from training on datasets with skewed distributions. By integrating the entirety of the dataset and employing data distillation techniques, we can extract the most relevant and informative features without succumbing to the limitations and potential prejudices inherent in imbalanced datasets.

Furthermore, our methodology is not confined to existing dataset distillation techniques but represents a general two-stage framework for addressing long-tail learning challenges. This framework's modular nature allows for the substitution of our second stage with more advanced dataset distillation approaches on long-tailed problems as they emerge, enhancing its future applicability and performance. In this regard, our approach can also function as a versatile plugin for future methods, offering a means to potentially improve their effectiveness in long-tail learning scenarios.

4 EXPERIMENTS

4.1 Experiment Setup

Data The experiments were carried out using the long-tailed CIFAR-10 dataset [1]. To validate the effectiveness of our approach under extremely skewed long-tail distributions and to mimic real-world long-tailed scenarios, we select the CIFAR10-LT dataset with an imbalanced ratio (r) of 100 to be the dataset for all our experiments. This choice of dataset with a high degree of imbalance enabled us to rigorously test our method in conditions that closely resemble practical challenges encountered in handling long-tailed data distributions.

Backbone Model To create a distilled dataset, we generated distilled images utilizing the default ConvNet3 network as the backbone model. In our training process, we employed the same network architecture as used during the distillation phase. This decision was informed by previous work [25], which highlighted that cross-architecture performance can be significantly impacted by differences in architecture. By maintaining architectural consistency, we aimed to mitigate potential variations in performance attributable to architectural discrepancies, thus ensuring a more accurate assessment of our method.

Hyperparameters When generating the distilled dataset, we apply default hyperparameters and training configurations borrowed from the official implementation of Gradient Matching[2] and Matching Training Trajectories[3]. During the two-stage training, we employ the ConvNet3 model as the classifier for both inter-group and intra-group classification. We set the learning rate empirically to 0.01 and utilize the SGD optimizer with a momentum of 0.9 and a weight decay of 0.0005.

4.2 Baseline Results

To implement a baseline method, we directly applied Gradient Matching and Matching Training Trajectories on the CIFAR10-LT dataset. Subsequently, we trained classification models on distilled datasets derived from this process. This approach represents a baseline performance of dataset distillation techniques applied to long-tailed datasets. We compare this method with directly training classification models on long-tailed dataset. The result is shown in Tab. 1. The results of our experiment indicate that dataset distillation did not enhance the accuracy of the classification models on the CIFAR10-LT dataset, instead yielding performance only comparable to models trained directly on the dataset. Notably, a significant reduction in the standard variance of accuracy among classes was observed, decreasing from 32.4 to 11.8(using Gradient Matching). This stability in performance, despite the reduction in variance, is non-trivial. It is important to highlight that dataset distillation often leads to a substantial decrease in accuracy. Therefore, maintaining comparable accuracy levels, while achieving reduced variance, underscores the potential of dataset distillation methods in managing the challenges posed by long-tailed distributions.

4.3 Overall Results

In our study conducted on the CIFAR10-LT dataset, we implemented our own method, utilizing data distillation techniques namely Matching Training Trajectories and Gradient Matching. The experimental results are detailed in Tab. 2. Our findings indicate that compared to directly training classification models on a long-tailed dataset, our method demonstrated improvements in both accuracy and the variance of accuracy. This suggests that our approach effectively mitigates issues associated with long-tail distributions. Notably, the increase in accuracy was primarily observed in the

[1] https://huggingface.co/datasets/tomas-gajarsky/cifar10-lt
[2] https://github.com/VICO-UoE/DatasetCondensation
[3] https://github.com/GeorgeCazenavette/mtt-distillation

Table 2. The performance of a two-stage trained ConvNet3 model, initially trained on the original dataset and subsequently on the distilled dataset. We use CIFAR10-LT (r=100) for all experiments. Dataset: Type of training dataset, Train Img Num: Total number of images used to train, Total Acc: Overall accuracy on the test dataset(the higher the better), Acc Std: Standard deviation of accuracy across different classes(the lower the better), Cls Acc: Accuracy of each class(the higher the better).

(a) Summary Statistics

Dataset	Train Img Num	Total Acc (%) ↑	Acc Std (%) ↓
Original + Original	2478	37.6	32.4
Original + GM Distilled	2578	**40.4**	**20.1**
Original + MTT Distilled	2578	32.3	20.3

(b) Class-wise Accuracy

Dataset	Cls Acc (%) ↑									
	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Original + Original	**90.6**	**90.4**	**62.4**	**51.2**	**32.4**	18.2	15.4	11.8	3.7	0.3
Original + GM Distilled	75.5	79.7	33.7	47.2	28.2	**32.3**	**25.0**	32.3	**34.9**	**15.1**
Original + MTT Distilled	60.3	65.6	53.0	37.8	20.4	12.8	19.2	**34.8**	10.1	8.3

Table 3. The performance of a two-stage trained ConvNet3 model training on different datasets. A+B represents using A dataset for stage 1 and B dataset for stage 2. We use CIFAR10-LT (r=100) for all experiments. Dataset: Type of training dataset, Train Img Num: Total number of images used to train, Total Acc: Overall accuracy on the test dataset(the higher the better), Acc Std: Standard deviation of accuracy across different classes(the lower the better), Cls Acc: Accuracy of each class(the higher the better). Group Acc: the accuracy of group classification(the higher the better)

(a) Summary Statistics

Dataset	Train Img Num	Total Acc (%) ↑	Acc Std (%) ↓	Group Acc (%) ↑
Original + Original	4756	37.7	32.2	**84.2**
GM Distilled + GM Distilled	200	35.0	**9.9**	75.1
Original + GM Distilled	2578	**40.4**	20.1	**84.2**

(b) Class-wise Accuracy

Dataset	Cls Acc (%) ↑									
	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Original + Original	**89.0**	**88.1**	**65.4**	**51.6**	**35.6**	18.2	12.8	14.5	1.7	0.1
GM Distilled + GM Distilled	43.5	51.6	34.7	46.3	31.8	**34.4**	**25.2**	**34.0**	32.9	**15.1**
Original + GM Distilled	75.5	79.7	33.7	47.2	28.2	32.3	25.0	32.3	**34.9**	**15.1**

head classes, while the rise in variance was also predominantly attributed to these classes. This is intuitive, as more samples are available for head classes, allowing the model to learn more information about them, thereby improving accuracy in these categories. Our method achieved this enhancement in head class accuracy without significant changes in tail class accuracy compared to the direct distillation approach. This underscores the capability of our method in alleviating the long-tail phenomenon while preserving more information from the original dataset.

4.4 Abalation Study

In our ablation study, we investigate the impact of different training approaches on the performance of the two-Stage classifier. Specifically, we examined the effects of training exclusively with distilled data and solely with long-tail data. We use Gradient Matching [25] as our distillation method. The results of these investigations are presented in Tab. 3. Our findings reveal that both training strategies underperformed compared to our hierarchical training recipe. This further substantiates the effectiveness of our proposed method.

5 CONCLUSION

In this work, we explore the impact of dataset distillation on long-tailed learning. Utilizing the CIFAR10-LT dataset, we proposed a two-stage classification scheme based on dataset distillation, which is based on our observation of inner similarities between classes. This approach not only fully leverages the complete information of the dataset but also mitigates the bias introduced by long-tail distributions. Compared to the baseline methods that train directly on long-tailed datasets, our method demonstrated improvements in both average accuracy and the variance of accuracy, emphatically validating its effectiveness. We anticipate that our pioneering research in dataset distillation on long-tailed learning will serve as a valuable reference and source of inspiration for future researchers.

ACKNOWLEDGMENTS

We thank Prof. Yonglu Li, Prof. Cewu Lu for their insight on this idea.

REFERENCES

[1] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett (Eds.). 1565–1576. https://proceedings.neurips.cc/paper/2019/hash/621461af90cadfdaf0e8d4cc25129f91-Abstract.html

[2] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. 2022. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4750–4759.

[3] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research* 16 (June 2002), 321–357. https://doi.org/10.1613/jair.953

[4] Zongxiong Chen, Jiahui Geng, Herbert Woisetschlaeger, Sonja Schimmler, Ruben Mayer, and Chunming Rong. 2023. A Comprehensive Study on Dataset Distillation: Performance, Privacy, Robustness and Fairness. *arXiv preprint arXiv:2305.03355* (2023).

[5] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. 2020. Remix: Rebalanced Mixup. In *Computer Vision - ECCV 2020 Workshops - Glasgow, UK, August 23-28, 2020, Proceedings, Part VI (Lecture Notes in Computer Science, Vol. 12540)*, Adrien Bartoli and Andrea Fusiello (Eds.). Springer, 95–110. https://doi.org/10.1007/978-3-030-65414-6_9

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

[7] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge J. Belongie. 2019. Class-Balanced Loss Based on Effective Number of Samples. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 9268–9277. https://doi.org/10.1109/CVPR.2019.00949

[8] Haibo He, Yang Bai, Edwardo A. Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*. 1322–1328. https://doi.org/10.1109/IJCNN.2008.4633969

[9] Yin-Yin He, Jianxin Wu, and Xiu-Shen Wei. 2021. Distilling Virtual Examples for Long-tailed Recognition. arXiv:2103.15042 [cs.CV]

[10] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (2015).

[11] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. 2020. Decoupling Representation and Classifier for Long-Tailed Recognition. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net. https://openreview.net/forum?id=r1gRTCVFvB

[12] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. 2022. Dataset condensation with contrastive signals. In *International Conference on Machine Learning*. PMLR, 12352–12364.

[13] Tianhao Li, Limin Wang, and Gangshan Wu. 2021. Self Supervision to Distillation for Long-Tailed Visual Recognition. arXiv:2109.04075 [cs.CV]

[14] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. 2020. Deep Representation Learning on Long-Tailed Data: A Learnable Embedding Augmentation Perspective. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2967–2976. https://doi.org/10.1109/CVPR42600.2020.00304

[15] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the Limits of Weakly Supervised Pretraining. arXiv:1805.00932 [cs.CV]

[16] Luke Metz, Niru Maheswaranathan, Jeremy Nixon, Daniel Freeman, and Jascha Sohl-Dickstein. 2019. Understanding and correcting pathologies in the training of learned optimizers. In *International Conference on Machine Learning*. PMLR, 4556–4565.

[17] Kai Wang, Bo Zhao, Xiangyu Peng, Zheng Zhu, Shuo Yang, Shuo Wang, Guan Huang, Hakan Bilen, Xinchao Wang, and Yang You. 2022. Cafe: Learning to condense dataset by aligning features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12196–12205.

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

[19] Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X. Yu. 2022. Long-tailed Recognition by Routing Diverse Distribution-Aware Experts. arXiv:2010.01809 [cs.CV]

[20] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. 2019. Dynamic Curriculum Learning for Imbalanced Data Classification. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 5016–5025. https://doi.org/10.1109/ICCV.2019.00512

[21] Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning From Multiple Experts: Self-paced Knowledge Distillation for Long-Tailed Classification. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part V (Lecture Notes in Computer Science, Vol. 12350)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer, 247–263. https://doi.org/10.1007/978-3-030-58558-7_15

[22] Liuyu Xiang, Guiguang Ding, and Jungong Han. 2020. Learning From Multiple Experts: Self-paced Knowledge Distillation for Long-tailed Classification. arXiv:2001.01536 [cs.CV]

[23] Yue Xu, Yong-Lu Li, Kaitong Cui, Ziyu Wang, Cewu Lu, Yu-Wing Tai, and Chi-Keung Tang. 2023. Distill Gold from Massive Ores: Efficient Dataset Distillation via Critical Samples Selection. *arXiv preprint arXiv:2305.18381* (2023).

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

[25] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. 2020. Dataset condensation with gradient matching. *arXiv preprint arXiv:2006.05929* (2020).

[26] Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min Chen. 2020. BBN: Bilateral-Branch Network With Cumulative Learning for Long-Tailed Visual Recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 9716–9725. https://doi.org/10.1109/CVPR42600.2020.00974

[27] Yang Zou, Zhiding Yu, B. V. K. Vijaya Kumar, and Jinsong Wang. 2018. Domain Adaptation for Semantic Segmentation via Class-Balanced Self-Training. arXiv:1810.07911 [cs.CV]