CCDS24019 - Image Dataset Distillation

Spyridon Giakoumatos
Nema Arpita
College of Computing and Data Science

Lin Weisi
College of Computing and Data Science

Abstract - Dataset distillation is a process that aims to create a small set of synthetic images that enables models trained on it to achieve performance comparable to training on the full dataset. Current state-of-the-art algorithms like Poster Dataset Distillation (PoDD) train models on a highly compressed distilled dataset, traditionally refining a single synthetic poster iteratively. However, this approach discards initial versions of posters that contain synthesized images rich in low-level information, thereby limiting knowledge retention and model generalization. In this paper, we extend PoDD to a multi-stage learning framework, where multiple evolving versions of the dataset are stored and progressively reused during training. Rather than relying exclusively on the latest distilled dataset, our method incorporates prior versions to preserve earlier learned information, mitigating catastrophic forgetting and improving model stability. A key insight is that a key metric in dataset distillation is not just the quality of the final dataset but the cumulative knowledge retained across training stages. We introduce Progressive Poster Dataset Distillation (P-PoDD), which dynamically integrates information from all previous dataset stages using novel blending and refinement techniques to improve knowledge retention. Experiments on CIFAR-10 demonstrate that P-PoDD improves test accuracy by 4.46% over PoDD at 0.9 images per class, while maintaining similar computational overhead. By leveraging multiple evolving dataset versions, P-PoDD achieves better generalization, improved robustness, and enhanced learning efficiency—establishing a new paradigm for dataset distillation grounded in progressive knowledge accumulation.

Keywords - Dataset Distillation, Synthetic Data, Cumulative Knowledge Retention

1 INTRODUCTION

In a climate of ever-growing data needs, large datasets are becoming more pertinent. This growing demand for data- efficient learning has sparked interest in data distillation [1], a technique that aims to compress a large data set into a much smaller synthetic one that allows deep models to be trained with comparable performance. Although recent progress has improved the quality of distilled datasets [2], [3], existing methods still face significant challenges under extreme compression—particularly in the low-data regime where the number of synthetic images per class (*IPC*) is less than one. In such settings, ensuring that models retain generalization ability and do not suffer catastrophic forgetting [4] remains an open problem.

In this paper we ask ”How can we improve accuracies when datasets are shrunk to less than 1 IPC?” To address this, we propose Progressive Poster Dataset Distillation (P-PoDD): a novel framework that leverages temporally evolving synthetic datasets across multiple stages of distillation. Instead of relying on a single static synthetic representation, P-PoDD retains and reuses earlier versions of distilled data, integrating them into the training process through an accuracy-weighted sampling mechanism. This progressive strategy enables knowledge accumulation across stages, improving performance and robustness at ultra-low IPC levels.

Our key contributions are:

- We propose P-PoDD, a multi-stage framework that incrementally refines synthetic dataset representations and reuses them across training stages.
- We develop a dynamic blending mechanism that mitigates forgetting by adaptively sampling from all prior stages based on their effectiveness. This adaptive sampling makes use of our Dynamic Stage Weighting explained later.
- Our approach introduces a new paradigm for dataset distillation that balances learning plasticity and memory retention over time.

2 RELATED WORKS

Dataset distillation was originally framed as a bi-level optimization task by Wang et al [1], synthesizing small training sets from large datasets to approximate full-data performance. Their pioneering work demonstrated that it is possible to reduce datasets like CIFAR-10 [5] to as few as 10–100 synthetic images, achieving competitive accuracy within a few gradient descent steps. This was done by optimizing the synthetic data directly

as differentiable variables, often conditioned on a fixed network initialization. However the approach required each class to have its own synthesized image, typically operating at ≥ 1 IPC, limiting its ability to handle extreme data compression regimes.

Early dataset distillation methods focused on gradient matching [3] and trajectory alignment [6], which optimize synthetic data to mimic the training dynamics of real datasets. Zhao et al. [7] further advanced the field with differentiable Siamese augmentation, while Kim et al. [2] introduced factorization-based approaches. These methods demonstrated significant compression capabilities but were constrained to standard per-class image generation, limiting their effective- ness in extreme compression scenarios.

Poster Dataset Distillation (PoDD) introduced by Shul et al. [8] overcomes this by compressing the dataset into a single large poster. Introduces a new distillation regime in which multiple classes share image regions, enabling sub-1 IPC training (as low as 0.3 IPC). However, PoDD is inherently a single-shot model: while it optimizes overlapping patches and soft labels effectively, it does not account for the changing dynamics of network training. This limits its robustness and leads to information loss from earlier synthetic representations.

To address the need for modelling evolving training dynamics, Progressive Dataset Distillation (PDD) proposed by Chen et al. [9] introduces a multi-stage distillation framework. Each stage generates new synthetic data conditioned on previous stages and trains on the cumulative union of all subsets. PDD demonstrates that progressive training better aligns with the temporal learning behaviour of neural networks, enabling improvements of up to 5% in test accuracy across datasets. However, PDD has not been applied to poster-based representations. Its application assumes standard image-per-class distillation without pixel-sharing.

The challenge of catastrophic forgetting in neural networks [4], [10] has been extensively studied in continual learning [11], [12], where models must retain knowledge while learning new tasks. Our work draws inspiration from episodic memory approaches [13] that maintain representative samples from previous learning stages, and efficiency-focused methods [14] that balance memory retention with computational constraints.

Recent advances in dataset distillation [15], [16] have focused on improving trajectory matching and scaling to larger datasets. Comprehensive surveys [17] provide broader context for the field's evolution, emphasizing the need for methods that can operate effectively under extreme compression while maintaining robust performance.

P-PoDD unifies the sub-1 IPC compression of PoDD with the progressive curriculum of PDD. Unlike PDD, which generates new image sets per stage, we maintain a single evolving poster and dynamically incorporate earlier versions via a sampling strategy based on historical accuracy. This allows us to capture the temporal diversity of training without discarding valuable early stage representations, achieving better stability and accuracy in extreme compression regimes.

3 METHODOLOGY

3.1 OVERIEW OF P-PODD

Progressive Poster Dataset Distillation (P-PoDD) introduces a multi-stage training process to mitigate catastrophic for- getting in ultra-compressed dataset distillation settings. The framework extends Poster Dataset Distillation (PoDD) by preserving earlier synthetic posters and integrating them into the ongoing distillation process. This allows the model to ac- cumulate and reuse knowledge across time without additional training cost.

Each stage $t \in \{1, 2, \ldots, T\}$ involves training a synthetic poster P_t and its label tensor Y_t for a fixed number of epochs. At the end of the stage, P_t is stored as a frozen knowledge bank. In the next stage, P_{t+1} is initialized and trained using training patches sampled proportionally from all posters $\{P_1, \ldots, P_{t+1}\}$. This way, earlier posters influence training even though they are no longer updated.

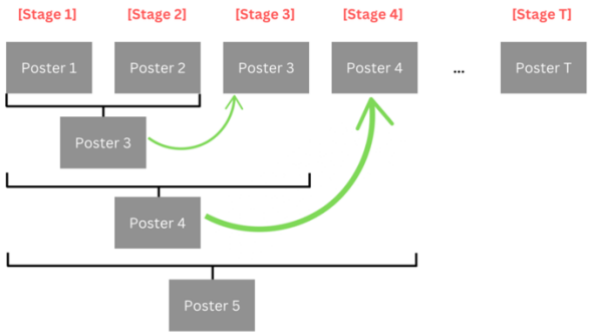


Fig. 1: Overview of the P-PoDD framework.

3.2 RETENTION AND PATCH EXTRACTION

In contrast to PoDD, which continuously overwrites a single poster, P-PoDD retains all intermediate posters. During each stage, a batch of image patches is sampled for training by selecting posters proportionally from a learned distribution π_t. Patches are then extracted using fixed-size crops (e.g., 32×32 for CIFAR-10) and the corresponding soft labels are retrieved from the label tensors.

This patch sampling process is performed using a dynamic stage weighting algorithm using the current and stored posters. At each iteration, a subset of patches is drawn proportionally from past stages. Since this involves standard batch sampling and tensor indexing operations without introducing new algorithmic constructs, we omit formalization and provide implementation details in the next subsection.

3.3 DYNAMIC STAGE WEIGHTING VIA ACCURACY-BASED DECAY

To determine how training samples are drawn from different posters, we compute a dynamic stage selection distribution π that assigns importance weights to each stage based on two factors:

1) Performance-based sampling: Stages with higher historical classification accuracy receive proportionally higher sampling weights, ensuring that well-performing posters contribute more to training.

2) Recency bias: A time-based exponential decay mechanism favours recent posters, operating under the assumption that newer stages contain more relevant or challenging examples for current learning objectives.

3) Exploration guarantee: We enforce minimum sampling weights and provide an additional boost to the newest poster's probability to ensure adequate learning signal from all stages.

This distribution is updated at the beginning of each stage. Importantly, we also enforce a minimum sampling weight and boost the newest poster's sampling probability to ensure sufficient learning signal. The full procedure is described in Algorithm 1.

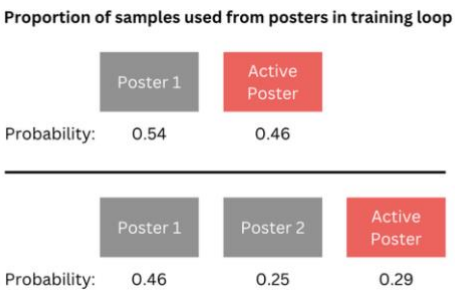


Fig. 2: An example of how stage sampling probabilities evolve.

Algorithm 1 Dynamic Stage Weight Update in P-PoDD

Input: Stage accuracies $A = [A_1, A_2, \ldots, A_n]$, minimum weight $\epsilon = 0.1$, boost factor $\beta = 1.5$, decay rate $\lambda = 0.5$
Output: Stage sampling distribution $\pi = [\pi_1, \pi_2, \ldots, \pi_T]$

$T \leftarrow$ current_stage $+ 1$ ▷ Number of completed stages
if $T = 1$ **then**
  return $\pi \leftarrow [1.0]$ ▷ Single stage case
end if
Step 1: Handle missing accuracy for newest stage
if length$(A) < T$ **then**
  $A_T \leftarrow 0.9 \times A_{T-1}$ ▷ Bootstrap accuracy for new stage
end if
Step 2: Compute accuracy-based weights
$w_t \leftarrow \frac{A_t + 10^{-5}}{\sum_{i=1}^{T}(A_i + 10^{-5})}$ for $t = 1, \ldots, T$
Step 3: Apply temporal decay and boost
for $t = 1$ to T **do**
  $d_t \leftarrow \exp(-\lambda \times (t-1))$ ▷ Exponential decay factor
  $\pi_t \leftarrow w_t \times d_t$
end for
$\pi_T \leftarrow \beta \times \pi_T$ ▷ Boost newest stage probability
Step 4: Enforce minimum weights and normalize
$\pi_t \leftarrow \max\left(\pi_t, \frac{\epsilon}{T}\right)$ for $t = 1, \ldots, T$
$\pi \leftarrow \frac{\pi}{\sum_{i=1}^{T} \pi_i}$ ▷ Final normalization
return π

3.4 TRAINING OBJECTIVE

The model parameters θ are updated via standard supervised learning. For each batch of training data sampled from the combined posters, the optimization objective is:

$$\theta \leftarrow \theta - \eta \cdot \nabla_\theta L(f_\theta(x), y)$$

where x and y are patches and labels sampled according to π_t, and L is a cross-entropy loss. Only the most recent poster P_t and its label tensor Y_t receive gradient updates; earlier posters remain frozen.

3.5 EFFICIENCY CONSIDERATIONS

Despite retaining multiple posters, P-PoDD maintains the same overall runtime complexity as PoDD. This is because only the most recent synthetic poster and its corresponding label tensor are actively updated during each stage. Previous posters are frozen and only used during sampling, contributing solely to the forward pass.

Moreover, the total number of training epochs is kept constant. Let E_{total} denote the total number of epochs, and T the number of distillation stages. In P-PoDD, each stage runs for:

$$E_{\text{stage}} = \left\lfloor \frac{E_{\text{total}}}{T} \right\rfloor$$

This means that the total training time is simply redistributed across stages, rather than increased. As a result, the marginal cost of retaining and sampling earlier posters is outweighed by gains in stability and generalization, leading to improved performance without added training time.

4 EXPERIMENTAL RESULTS

4.1 SETUP

Dataset. All experiments were conducted on the CIFAR-10 dataset, a standard benchmark in dataset distillation. CIFAR-10 consists of 60,000 32×32 color images across 10 classes, with 6,000 images per class. For distillation, we retain only a small number of synthetic examples per class, defined as images per class (IPC). Our primary experiments are conducted at 0.9 IPC, corresponding to a total of 9 synthetic images.

Baselines. We compare our method, P-PoDD, to Poster Dataset Distillation (PoDD) as proposed by Shul et al. [8]. Both methods are evaluated under identical settings:

- Fixed random seed (seed = 0)
- Same network architecture (ConvNet used in prior distillation works)
- Total training epochs fixed at 150
- Equal synthetic image resolution and patching strategy

Evaluation. Following prior work in dataset distillation [9], we adopt a multi-round training and testing protocol to evaluate generalization. The evaluation loop is structured as follows:

- A ConvNet is trained from scratch on the synthetic dataset for 300, 600, 1000, and 2000 epochs.
- After each epoch window, test accuracy is evaluated on the CIFAR-10 test set.
- Final reported test accuracy is taken from the model trained for 2000 epochs and averaged across all runs.

4.2 MAIN RESULTS

Our primary evaluation metric is test accuracy on a ConvNet trained from scratch using only the distilled synthetic data. At the 0.9 IPC setting—representing fewer than one image per class—the challenge is to preserve class-separable information using a highly compressed representation. This regime is especially difficult, as models trained on real data typically require at least one sample per class to generalize effectively.

TABLE I: Test Accuracy (%) of PoDD vs. P-PoDD on CIFAR- 10 under sub-1 IPC settings.

IPC	PoDD	P-PoDD (ours)
0.9	51.5	**53.8**
0.8	49.6	51.9
0.7	47.7	49.4
0.6	47.2	48.3

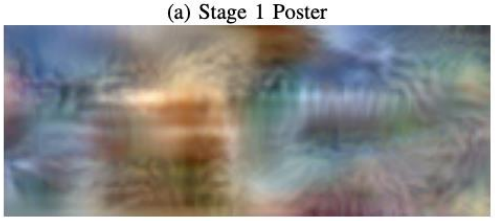
As shown in Table I, our proposed P-PoDD framework achieves a test accuracy of 53.8%, compared to 51.5% achieved by the original PoDD baseline under identical set- tings. This represents a 4.46 percentage point relative improvement. This improvement is achieved under extreme data compression and with no increase in training budget, making the result both statistically and practically significant.

The improvement highlights two key advantages of P-PoDD: Temporal knowledge retention, earlier poster stages retain low-level features that might otherwise be overwritten in single-shot distillation. These are reused in later training through our stage-weighted sampling mechanism. Training stability, by sampling from multiple stages, we mitigate catastrophic forgetting and reduce the reliance on precise initialization or overfitting to a single representation.

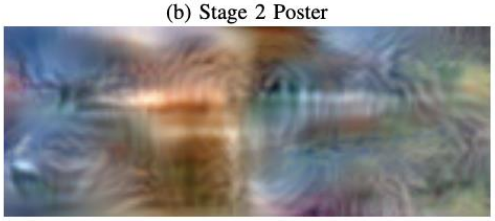
4.3 POSTER VISUALISATION ACROSS STAGES

Figure 3 illustrates the evolution of the synthetic poster across five training stages in P-PoDD. Early-stage posters contain low-level color and edge information, while later stages refine class-specific semantics. The progressive refinement demonstrates how earlier posters serve as knowledge banks that complement newer representations during training.

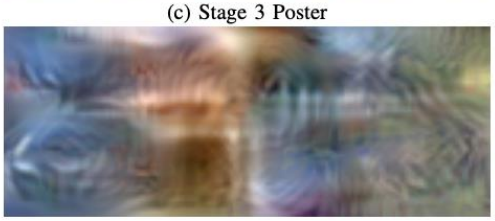
(a) Stage 1 Poster



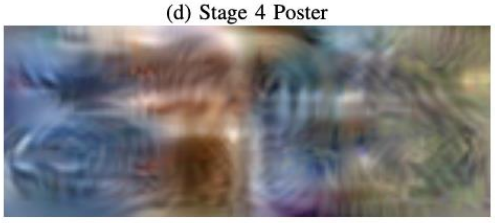
(b) Stage 2 Poster



(c) Stage 3 Poster



(d) Stage 4 Poster



(e) Stage 5 Poster

Fig. 3: Evolution of the synthetic poster across P-PoDD training stages (Stage 1 to Stage 5). Earlier stages capture low-level structure and global color information, while later stages refine semantically relevant class features.

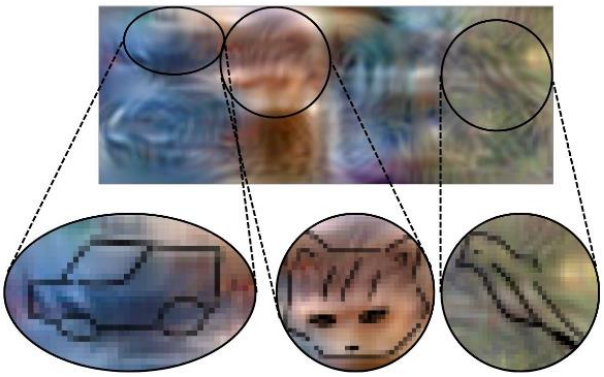


Fig. 4: Visualization of CIFAR-10 Classes within Stage 5 Poster, captures Automobile, Cat and Bird from left to right.

4.4 ABLATION STUDIES

To assess the contribution of each component in P-PoDD, we conducted targeted ablation experiments while keeping all other settings fixed. The results confirm that our dynamic stage sampling strategy, including the accuracy-based weight- ing, exponential decay, boost factor and number of stages, meaningfully contribute to the final accuracy.

New Stage Boost ($\beta = 1.5$): Without this boost, the most recently initialized poster was sampled less frequently, slowing learning in later stages. Removing the boost reduced final test accuracy by approximately 2.6%.

TABLE II: New Stage Boost Ablation Study

Boost (β)	**Acc. (%)**	**Drop**
1.5 (proposed)	**53.8**	–
0.0 (no boost)	51.2	-2.6%

Accuracy-Based Sampling vs. Uniform: Using uniform stage sampling degraded test accuracy compared to dynamic π_t-based sampling. This highlights the importance of prioritizing posters with higher historical performance, validating the benefit of cumulative learning with selectivity.

TABLE III: Ablation: Accuracy-Based vs. Uniform Stage Sampling

Sampling Strategy	**CIFAR-10 (%)**	**Drop**
Dynamic π_t-based	**53.8**	–
Uniform sampling	50.4	-3.4%

Number of Stages: Within the CIFAR-10 dataset we are using 5 stages spread over 150 epochs. Going above or below this number of stages showed a decrease in accuracy.

TABLE IV: Ablation: Impact of Number of Stages on Performance

Number of Stages	CIFAR-10 (%)	Drop from Optimal
3 stages	52.2	-1.6%
4 stages	52.7	-1.1%
5 stages (optimal)	**53.8**	–
6 stages	52.5	-1.3%
7 stages	51.9	-1.9%

These results show that P-PoDD's improvements are not solely due to multi-stage training, but arise from careful balancing of historical knowledge retention and forward progress through a learned stage selection policy.

5 LIMITATIONS AND FUTURE WORK

While P-PoDD shows strong improvements over the single- stage PoDD baseline, it is not without limitations. These suggest clear avenues for future research.

Computational Runtime: Although P-PoDD maintains the same total number of training epochs, training across multiple stages increases wall-clock time due to repeated reinitialization and evaluation. In practice, this limits IPC sweeps or multiple seed evaluations under modest compute budgets.

Memory Overhead: With the introduction of P-PoDD, multiple posters are being stored during the training cycle. While inconsequential when dealing with datasets with smaller images like CIFAR-10 and CIFAR-100 [18]. Memory issues may arise if higher resolution images and posters are stored for other datasets.

Sensitivity to Sampling Strategy: Our ablations revealed that stage sampling distribution plays a critical role in performance. Without dynamic weighting or recency bias (via exponential decay), performance degrades noticeably. This indicates a sensitivity that may require dataset-specific tuning.

Future work will focus on (i) extending evaluations across IPC levels and random seeds; (ii) applying P-PoDD to domain- shift or privacy-preserving settings; and (iii) exploring auto- mated tuning of stage-wise learning dynamics.

6 CONCLUSION

We proposed Progressive Poster Dataset Distillation (P-PoDD), a multi-stage extension to PoDD that retains synthetic posters from earlier training stages and integrates them into a dynamic, accuracy-weighted sampling strategy. This approach addresses key weaknesses in traditional single-stage distillation by preserving temporal diversity and mitigating catastrophic forgetting.

P-PoDD achieves a 4.46% relative accuracy improvement over PoDD at 0.9 IPC on CIFAR-10, with no additional training epochs. Qualitative visualizations show progressive refinement in the learned posters, and ablation studies confirm that each design element—accuracy-based weighting, exponential decay, and new-stage boosting—contributes meaningfully to overall performance.

Our findings highlight the importance of cumulative memory and curriculum-aware sampling in ultra-compressed dataset distillation. We believe that P-PoDD opens the door to more adaptive, stable, and scalable data-efficient learning pipelines.

ACKNOWLEDGMENT

This work was supported as part of the URECA Undergraduate Research Programme at Nanyang Technological University. I would like to thank the CCDS GPU Cluster Team for providing computational resources. Additional thanks to the reviewers and peers who provided valuable feedback during the poster presentation and code review stages.

REFERENCES

[1] T. Wang, J.-Y. Zhu, A. Torralba, and A. A. Efros, "Dataset distillation," *arXiv preprint arXiv:1811.10959*, 2018.

[2] J.-H. Kim, J. Kim, S. J. Oh, S. Yun, H. Song, J. Jeong, J.-W. Ha, and H. O. Song, "Dataset condensation via efficient synthetic-data parameterization," in *International Conference on Machine Learning*. PMLR, 2022, pp. 11 102–11 118.

[3] B. Zhao, K. R. Mopuri, and H. Bilen, "Dataset condensation with gradient matching," *arXiv preprint arXiv:2006.05929*, 2020.

[4] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska *et al.*, "Overcoming catastrophic forgetting in neural

networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[5] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," *Technical Report, University of Toronto*, 2009.

[6] G. Cazenavette, T. Wang, A. Torralba, A. A. Efros, and J.-Y. Zhu, "Dataset distillation by matching training trajectories," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4750–4759.

[7] B.Zhao and H.Bilen,"Dataset condensation with differentiable siamese augmentation," in *International Conference on Machine Learning*, 2021, pp. 12 674–12 685.

[8] A. Shul, E. Horwitz, and Y. Hoshen, "Distilling datasets into less than one image," *CoRR*, 2024. [Online]. Available: https://arxiv.org/abs/2403.12040

[9] X. Chen, Y. Yang, Z. Wang, and B. Mirzasoleiman, "Data distillation can be like vodka: Distilling more times for better quality," *ICLR*, 2023. [Online]. Available: https://arxiv.org/abs/2310.06982

[10] M. McCloskey and N. J. Cohen, "Catastrophic interference in con- nectionist networks: The sequential learning problem," *Psychology of learning and motivation*, vol. 24, pp. 109–165, 1989.

[11] F.Zenke,B.Poole,andS.Ganguli,"Continual learning through synaptic intelligence," in *International conference on machine learning*, 2017, pp. 3987–3995.

[12] D.Lopez-PazandM.Ranzato,"Gradient episodic memory for continual learning," in *Advances in neural information processing systems*, 2017, pp. 6467–6476.

[13] S.-A. Rebuffi, A. Kolesnikov, G. Sperl, and C. H. Lampert, "icarl: Incremental classifier and representation learning," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 2001–2010.

[14] A. Chaudhry, M. Ranzato, M. Rohrbach, and M. Elhoseiny, "Efficient lifelong learning with a-gem," in *International Conference on Learning Representations*, 2019.

[15] J. Du, Y. Gan, J. Wang, L. Cheng, and Q. Wang, "Minimizing the accumulated trajectory error to improve dataset distillation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3749–3758.

[16] Z. Liu, Z. Qin, Z. Mu, and Q. Zhang, "Dataset distillation in large data era," in *International Conference on Machine Learning*, 2023, pp. 21 629–21 648.

[17] S. Yu and H. Hashemi, "Dataset distillation: A comprehensive survey," *arXiv preprint arXiv:2301.07014*, 2023.

[18] A. Krizhevsky, V. Nair, and G. Hinton, "Cifar-100 (canadian institute for advanced research)." [Online]. Available: http://www.cs.toronto.edu/kriz/cifar.html