DistillScope: Analyzing the Scope and Limits of Diffusion-Based Dataset Distillation

Hannah Carey[2] Kunyang Li[1] Mubarak Shah[1]

[1]University of Central Florida [2]University of Massachusetts Amherst

July 31, 2025

Abstract

Dataset distillation aims to replace large-scale training sets with smaller synthetic datasets while maintaining model performance. While most prior work has focused on optimization-based methods, recent diffusion-based approaches have demonstrated promising empirical results. In this paper, we adapt the evaluation framework of Yang et al. (2024) to diffusion-generated datasets, analyzing their compatibility with real data, learning dynamics, and information content. We perform three experiments: (1) a mixing analysis with real data, (2) loss curvature measurement via Hessian trace estimation, and (3) a transferability test using real-trained models. Our results show that diffusion-distilled data is more compatible with real data and provides continuous learning signals, unlike optimization-based methods which capture only early-stage training information. These findings suggest diffusion models generate higher-quality synthetic datasets and may offer a more robust alternative to prior distillation techniques.

1 Introduction

Deep learning's reliance on large labeled datasets poses increasing challenges in terms of computational cost, storage, and accessibility. Dataset distillation addresses this issue by synthesizing compact datasets that retain the performance characteristics of the original data. Traditionally, most methods have approached this through optimization-based distillation, which iteratively refines synthetic samples to match training trajectories, gradients, or feature distributions.

Recently, diffusion-based approaches have emerged as an alternative, leveraging generative models to synthesize training data that captures a distributional representation of the real dataset. While empirical studies have shown that diffusion-based methods can achieve competitive accuracy with significantly less computational cost, less is known about their theoretical behavior and training dynamics.

Our research builds directly on the evaluation framework introduced by Yang et al. (2024), who studied optimization-based distilled data and concluded that such datasets primarily capture early-stage training information and are often incompatible with real data. In contrast, we analyze diffusion-based distilled data and investigate whether its learning behavior more closely mirrors real data. We adapt their methodology to diffusion-generated data and conduct a handful of experiments, uncovering compatibility testing through real-data mixing, loss curvature analysis via Hessian trace estimation, and transferability of semantic understanding from real to distilled data.

2 Related Work

Recent advances in dataset distillation aim to reduce the cost and storage of training large-scale deep models by synthesizing compact yet information-rich sur-

rogate datasets. Broadly, methods fall into two categories: optimization-based approaches and diffusion-based generative strategies.

Optimization-based methods typically perform iterative updates to optimize synthetic data directly at the pixel or feature level. Zhao and Bilen (2023) proposed Dataset Condensation with Distribution Matching (DC-DM), which avoids expensive bi-level optimization by aligning feature distributions of real and synthetic data in sampled embedding spaces using maximum mean discrepancy (MMD). This approach improves synthesis speed and scalability while preserving test accuracy across larger datasets.

Despite such improvements, many optimization-based methods remain computationally costly and brittle under architectural changes or mixed training regimes.

Yang et al. (2024) conducted a theoretical investigation into what information distilled datasets retain. Their findings indicate that most distilled data — especially those derived via gradient or trajectory matching — encode information similar to what is learned during the early training phase of real data. They further demonstrate that distilled data, while semantically meaningful, may perform poorly when combined with real data during training due to representational mismatch.

More recently, diffusion-based distillation has emerged as an alternative approach. Gu et al. (2023) introduced Minimax Diffusion, leveraging the generative capacity of diffusion models to synthesize datasets while enforcing representativeness and diversity via minimax optimization criteria. Their method minimizes similarity among synthetic data while maximizing similarity to real data, enabling robust distillation even under high-resolution or high-IPC settings with dramatically reduced training time.

Building on this idea, Su et al. (2024) proposed D4M, a disentangled diffusion model that enhances cross-architecture generalization and semantic consistency. D4M decouples the distillation process from the training architecture using a training-time matching (TTM) framework, and employs latent diffusion conditioned on category-level prototypes to synthesize realistic and adaptable data. Unlike bi-level optimization methods, D4M scales efficiently to high-resolution data and achieves competitive accuracy across architectures.

Together, these works highlight a shift toward generative, architecture-agnostic distillation methods that overcome the performance and scalability bottlenecks of traditional approaches. Our research builds on this trajectory by further analyzing how diffusion-distilled data compares with optimization-distilled data in terms of training dynamics, compatibility with real data, and the temporal information it encapsulates.

3 Method

3.1 Overview

To analyze the behavior of diffusion-distilled data, we replicate and adapt the experimental design proposed by Yang et al. (2024). We focus on evaluating whether the synthetic data:

is compatible with real training data,

provides continuous learning signals, and

retains semantic meaning recognizable by real-trained models.

3.2 Real Data Compatibility and Mixing

We train a ConvNet on four data mixtures:

10 distilled images per class (baseline)

10 distilled + 2 real images/class

10 distilled + 10 real images/class

10 distilled + 50 real images/class

10 distilled + 250 real images/class

Each configuration is evaluated on the CIFAR-10 test set to determine how the inclusion of real data influences performance. Prior work showed that mixing optimization-distilled and real data often harms accuracy. We test whether this phenomena is observed with diffusion-based data.

3.3 Information Stored in Images Yang et al made a significant finding in that optimization-based distilled datasets can only distill early-stage information – the information that a model learns very early on, but no information that a mature model learns is stored in the images. Diffusion-generated images *seem* to contain more nuanced

information including more diverse angles and details. We are curious if these images store later-stage training information, and in order to do so, we train a ConvNet on real CIFAR-10 images and evaluate the performance on distilled datasets to analyze if the model is continuously learning information about the dataset.

3.4 Loss Curvature Analysis

To investigate training dynamics, we measure the trace of the Hessian of the loss with respect to model parameters. We compute this using Hutchinson's estimator implemented in PyHessian. Following Yang et al., we track curvature over 300 iterations during model training on real data while evaluating the curvature induced by both real and diffusion-distilled samples.

Higher curvature indicates regions where the model is learning rapidly, while lower curvature suggests convergence or flat loss landscapes. We compare curvature profiles to determine whether diffusion-distilled data continues to offer learning signal throughout training.

3.5 Pixel Intensity Analysis

To evaluate low-level similarity between real and distilled datasets, we analyze the pixel intensity distributions of CIFAR-10 images from three sources: real data, optimization-distilled data (via distribution matching), and diffusion-distilled data. For each dataset, we compute kernel density estimates (KDEs) of normalized pixel values across all channels. Since only the diffusion model is fine-tuned, we compare its distribution both before and after fine-tuning, while optimization-based data remains unchanged. This analysis highlights how the fine-tuning process affects pixel-level realism and potential generalization.

4 Dataset

All experiments are conducted using the CIFAR-10 dataset, which consists of 60,000 32×32 color images across 10 classes, with 50,000 for training and 10,000 for testing. As our baseline to compare the previous analysis of optimization-based data, we used the four distilled datasets provided by Yang et al. These each contain 10 images per class (10 IPC) totaling 100 synthetic images each. For our diffusion-based dataset, we generated two new sets of synthetic images with D4M and Minimax diffusion, each containing 10 IPC, and 100 images total each. As D4M is pre-trained on a larger dataset, we experienced issues with a domain gap in the output images, so we stick to the Minimax data for diffusion-based data analysis. All images are label-conditioned and sampled to match the class distribution of CIFAR-10.

5 Results

4.1 Real Data Mixing

Shown in Figure 1, a model trained on 10 distilled images performs a certain way, and we are interested in how the performance of the model changes when we mix in a certain number of real images. As found previously, all four optimization-based methods suffer a performance decrease initially when 2 real image are included per class in the training set, and it isn't until the real images comprise a majority of the training set that the performance starts to increase. However, unlike the optimization-based methods, our results show that adding even the small number of real images in fact improved performance on diffusion-distilled data. This suggests a unique compatibility between real and diffusion-generated data, which was not observed with optimization-based distilled data.

4.2 Training Stage Info

As seen in Figure 2, a model trained on the real CIFAR-10 training set is evaluated at each training iteration by calculating test accuracy on all distilled datasets. The accuracy of the model on our diffusion-based dataset was continuously increasing, suggesting that these images distill not only early-stage information but later-stage information too.

4.3 Loss Curvature

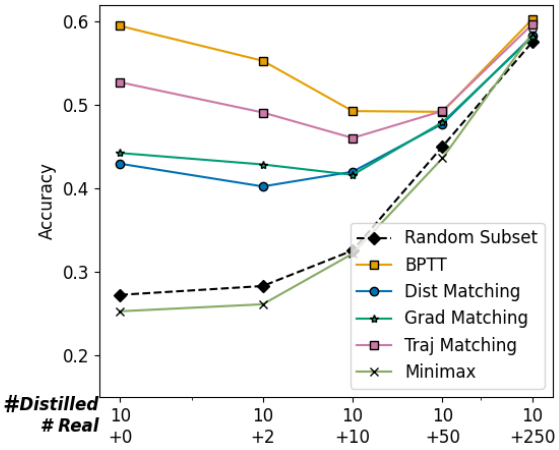


Figure 1: The performance of optimization-based distilled training sets initially *decreases* when real images are mixed into the set. However, shown in green, the performance of our diffusion-based distilled training set *improves* when mixed with real images.

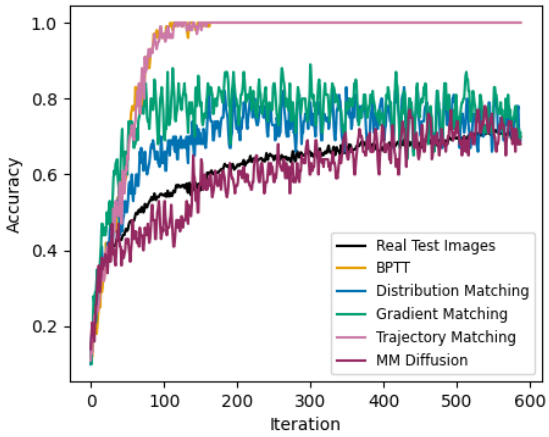


Figure 2: BPTT and MTT (Trajectory Matching) are always recognized by a weak model, since they distill early training dynamics. Distribution and Gradient matching do the same to a degree, but the model actually becomes less effective at classifying them as it learns. The only set of data the model *continuously* improves at classifying is the Minimax diffusion-generated distilled set of data.

We observe that diffusion-distilled data induces high loss curvature early in training, similar to real data. Unlike optimization-based data (e.g., BPTT), which quickly plateaus to a flat region, diffusion-distilled data shows a gradual decline, implying that it provides ongoing learning signal beyond the initial training phase. This suggests diffusion data contains richer, more continuous information than prior synthetic methods.

4.4 Pixel Intensity Distribution

We compare pixel intensity distributions of real data, optimization-based distilled data via distribution matching, and diffusion-distilled data, both before and after fine-tuning of the diffusion model. As shown in Figure 4, fine-tuned diffusion data shifts noticeably away from the real data distribution in pixel space. This suggests that the fine-tuning process may unintentionally degrade low-level alignment with real data, potentially limiting generalization. The result highlights a possible limitation in current diffusion distillation methods or their fine-tuning

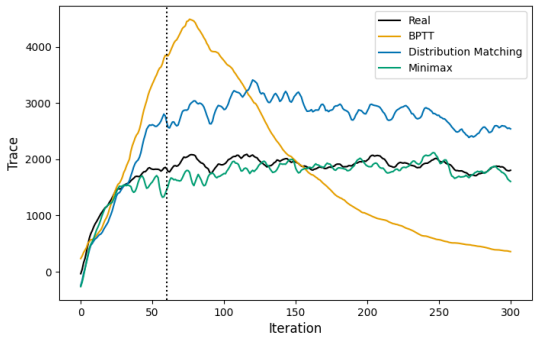


Figure 3: As a model trains on real data, the trace of the Hessian matrix is calculated with respect to a selection of datasets. The higher the trace, the higher the curvature of the loss landscape; indicating the model is changing its state and the high trace is typically seen in periods of learning.

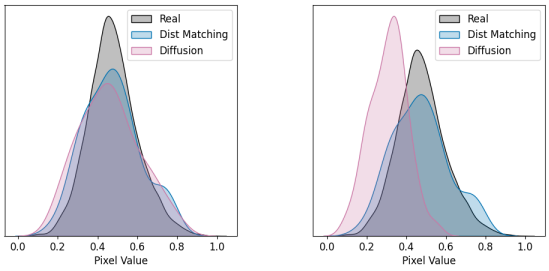


Figure 4: Pixel distributions of real CIFAR-10 images, distribution-matching distilled data, and diffusion-based distilled data. The difference between these two figures is only the diffusion-based data; the figure on the left shows its distribution before fine-tuning on CIFAR-10, and the figure on the right is the distribution after fine-tuning on our dataset.

procedures.

6 Conclusion

This study presents a theoretical and empirical evaluation of diffusion-based dataset distillation by adapting the framework of Yang et al. (2024). We find that diffusion-distilled data behaves differently from optimization-based methods in several key ways; it is compatible with real data in mixed training, it provides continuous learning signals, as indicated by loss curvature, and it retains semantic structure recognizable by real-trained models. These results suggest that diffusion-based distillation offers more faithful and flexible synthetic datasets and may serve as a more robust alternative to traditional distillation approaches. Future work may explore scaling these methods to larger datasets and testing architectural generalization more thoroughly, especially in concern with the distribution mismatch noted in the final figure of our results section.

References

[1] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories, 2022.

[2] Zhiwei Deng and Olga Russakovsky. Remember the past: Distilling datasets into addressable memories for neural networks, 2022.

[3] Jianyang Gu, Saeed Vahidian, Vyacheslav Kungurtsev, Haonan Wang, Wei Jiang, Yang You, and Yiran Chen. Efficient dataset distillation via minimax diffusion, 2024.

[4] Duo Su, Junjie Hou, Weizhi Gao, Yingjie Tian, and Bowen Tang. D^4m: Dataset distillation via disentangled diffusion model, 2024.

[5] William Yang, Ye Zhu, Zhiwei Deng, and Olga Russakovsky. What is dataset distillation learning?, 2024.

[6] Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney. Pyhessian: Neural networks through the lens of the hessian, 2020.

[7] Jingwen Ye, Ruonan Yu, Songhua Liu, and Xinchao Wang. Distilled datamodel with reverse gradient matching, 2024.

[8] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching, 2022.