From Optimization to Diffusion: Rethinking What Dataset Distillation Means

CRCV REU Week 6

Hannah Carey

Project: DistillScope: Analyzing the Scope and Limits of Diffusion-Based Dataset Distillation with Kunyang Li and Dr. Shah

Introduction on Dataset Distillation

Goal: Derive a small dataset from a much larger one.

Traditional focus: Optimization-based approaches
(gradient, trajectory , distribution matching).

Key insights (from the paper “What is Dataset Distillation Learning?” [1]**):**

- Distilled data ≠ Real data
- Compress early training dynamics
- Individual samples contain semantic meaning

Motivation for Expansion

Why Look Beyond Optimization?

- Modern generative models (Diffusion) dominate image and video synthesis.
- Diffusion models naturally support scalable resolution and modality (images, videos, text).
- What if dataset distillation itself could leverage these strengths?

"What Is Dataset Distillation?" Expanded

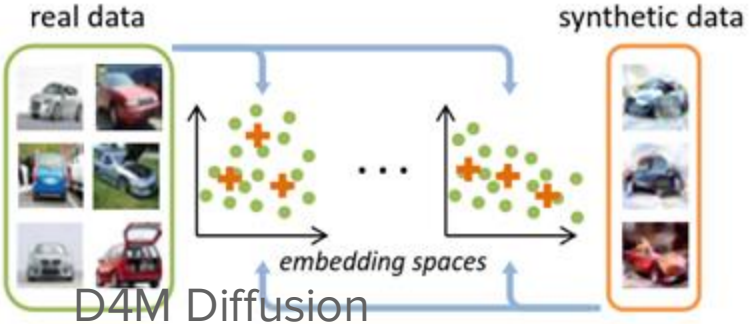
- What behaviors, information, and representation do distilled datasets contain (especially diffusion-based method)?
- How can generative processes simulate or amplify such information to scale distillation?
- …

Diffusion-Based Distillation – The Paradigm Shift

Aspect	Optimization-Based	Diffusion-Based
Mechanism	Bi-level, gradient/trajectory/distribution	Forward sampling
Representation	Explicitly learned points	Implicit in generative model’s prior
Resolution/Modality	Fixed, low-resolution	Scalable (e.g., high-res, ImageNet-1K)
Generalization	Often model-specific	Strong transfer potential

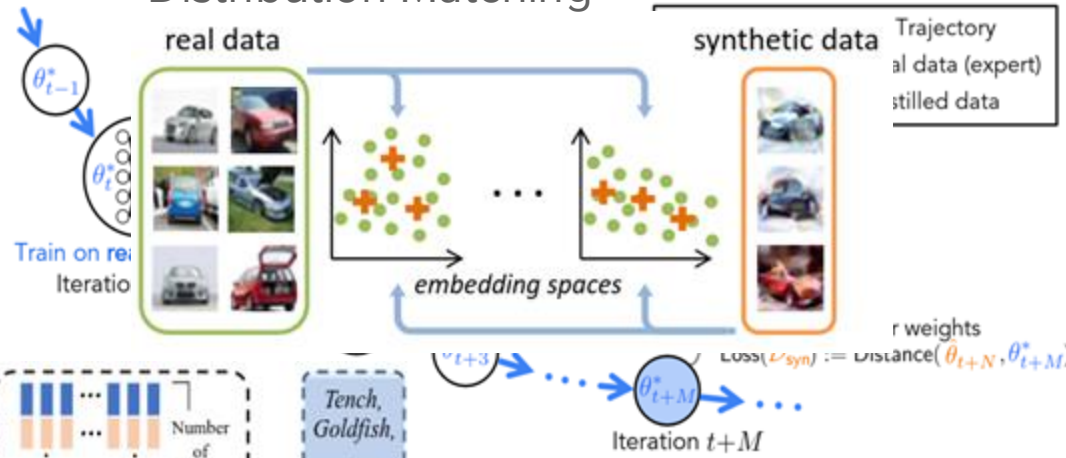
Distillation Architectures

Distribution Matching

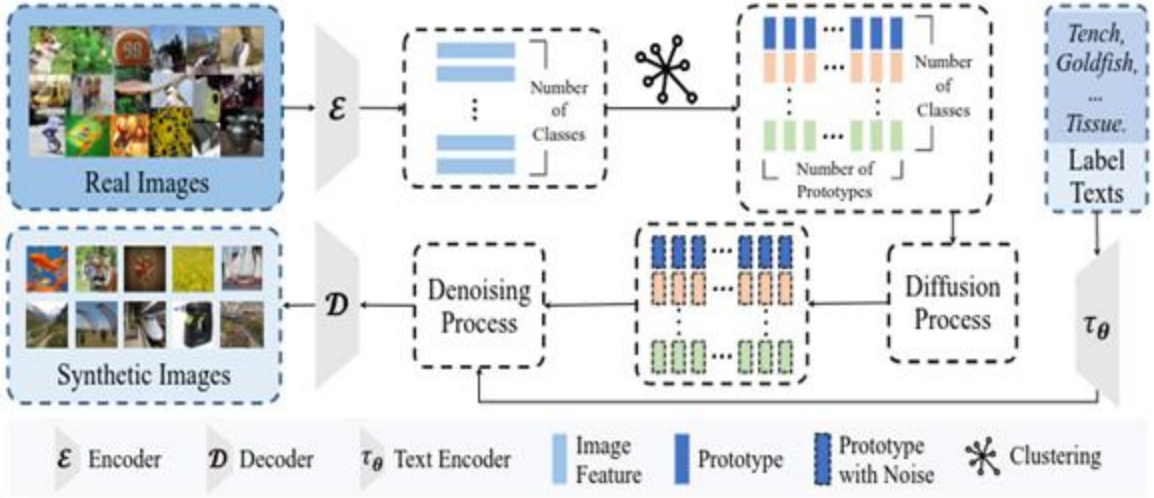


Trajectory Matching

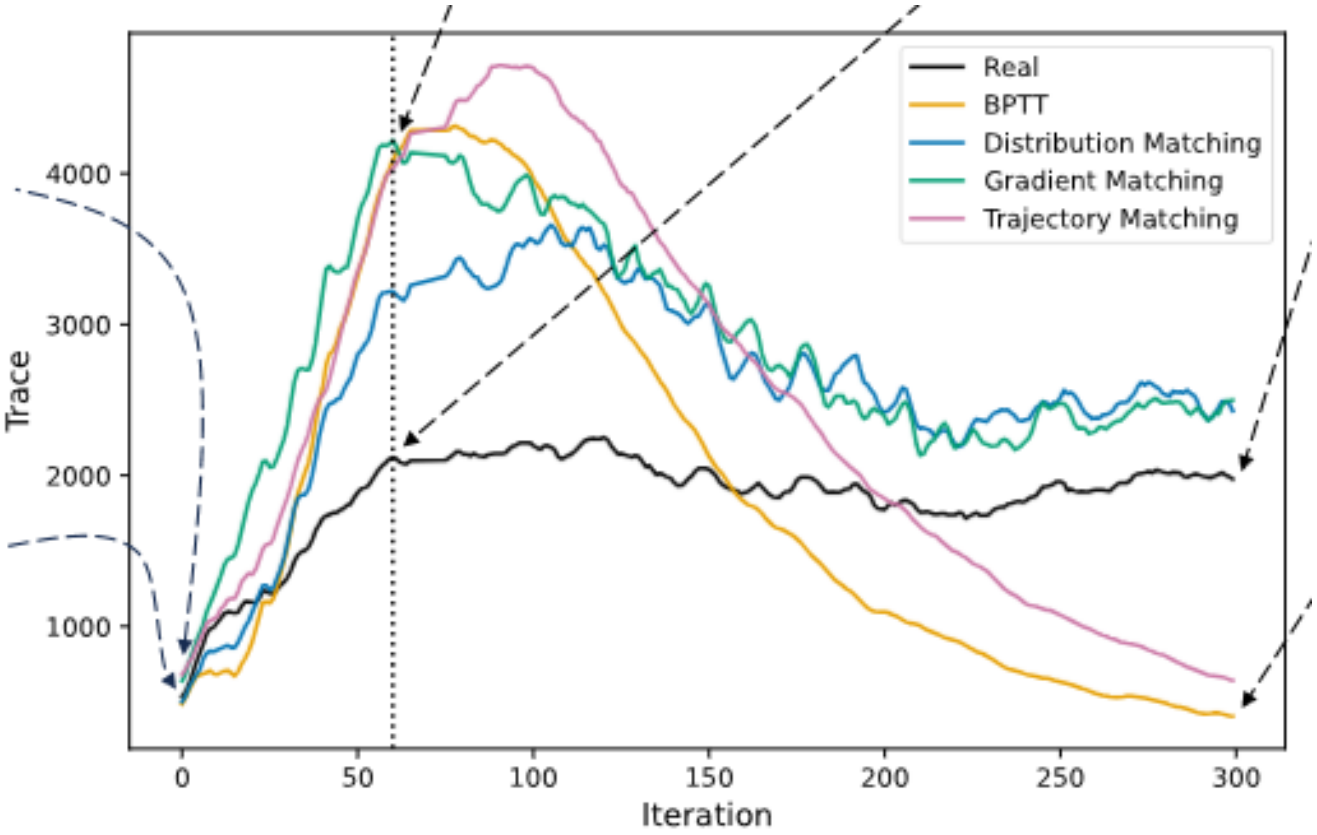
Distribution Matching



D4M Diffusion



Hessian Matrix Analysis including D4M

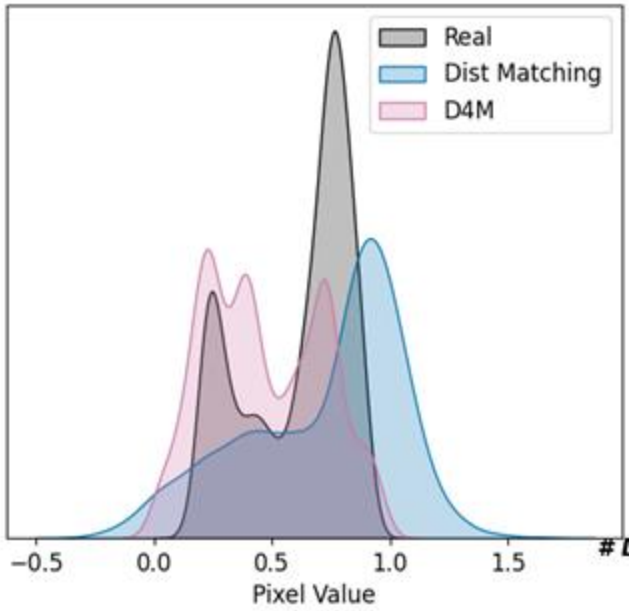


D4M behaves similar to distribution

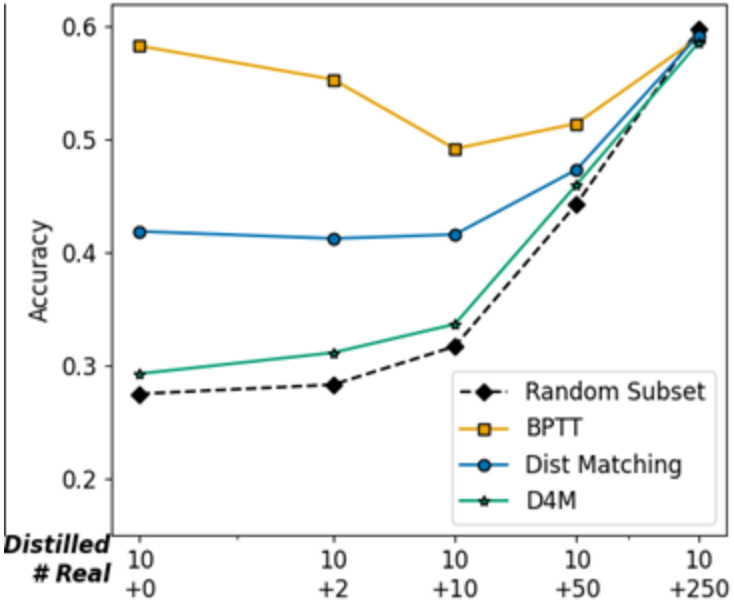
Trace is tumultuous, possibly due to domain gap issues

[1]

Dynamics of D4M data



D4M pixel intensity distribution is comparable to that of real data



D4M data properly mixes with real data, unlike optimization-based methods.

Next Steps

Deal with domain gap issue for CIFAR-10 diffusion models and re-evaluate metrics like the Hessian figure

Expansion into Imagenet-1K distillation datasets and investigating with these metrics

Deeper understanding of diffusion based distillation and investigating questions that come up

References

[1] Yang, William et al. (2024) What is Dataset Distillation Learning?
https://arxiv.org/pdf/2406.04284v1