Discovering Galaxy Features via Dataset Distillation

Haowen Guan[1], Xuan Zhao[1], Zishi Wang[1], Zhiyang Li[1], and Julia Kempe[1,2]

[1]Center for Data Science, New York University
[2]Courant Institute of Mathematical Sciences, New York University

Abstract

In many applications, Neural Nets (NNs) have classification performance on par or even exceeding human capacity. Moreover, it is likely that NNs leverage underlying features that might differ from those humans perceive to classify. Can we "reverse-engineer" pertinent features to enhance our scientific understanding? Here, we apply this idea to the notoriously difficult task of galaxy classification: NNs have reached high performance for this task, but what does a neural net (NN) "see" when it classifies galaxies? Are there morphological features that the human eye might overlook that could help with the this task and provide new insights? Can we visualize tracers of early evolution, or additionally incorporated spectral data? We present a novel way to summarize and visualize galaxy morphology through the lens of neural networks, leveraging Dataset Distillation, a recent deep-learning methodology with the primary objective to distill knowledge from a large dataset and condense it into a compact synthetic dataset, such that a model trained on this synthetic dataset achieves performance comparable to a model trained on the full dataset. We curate a class-balanced, medium-size high-confidence version of the Galaxy Zoo 2 dataset, and proceed with dataset distillation from our accurate NN-classifier to create synthesized prototypical images of galaxy morphological features, demonstrating its effectiveness. Of independent interest, we introduce a self-adaptive version of the state-of-the-art Matching Trajectories algorithm to automate the distillation process, and show enhanced performance on computer vision benchmarks.

1 Introduction and Background

The study of galaxy morphology is fundamental in observational cosmology. Morphological features are essential for determining a galaxy's dynamical state and interpreting its evolutionary history. Since Hubble's first classification in 1926, significant efforts have been dedicated to designing morphological classification schemes and data collection methods. For instance, Galaxy Zoo [19, 18], through its crowd-sourcing approach for large-scale analysis, classifies galaxies from the Sloan Digital Sky Survey (SDSS) [31] into three basic types: elliptical (early-type), spiral (late-type), and mergers. Its successor, Galaxy Zoo 2 (GZ2, 30), further expands this classification scheme to include more detailed morphological features, such as bars, bulges, and the shape of edge-on disks. Deep learning techniques, specifically those based on deep convolutional neural networks (CNNs, 17), have emerged as automated approaches for galaxy morphology classification [7, 14], yielding impressive results surpassing previous methods in predicting classifications made by humans. CNN-based galaxy morphology classification has now been applied across multiple different surveys, including SDSS [8, 26, 10], CANDELS [15, 12, 10], and Dark Energy Survey [3, 2] with predicted features included in official surveys such as the new catalogue in SDSS-IV DR17 [9].

Visual morphologies are notoriously hard to classify, given the variability of the data (e.g. sensitivity to red-shift). Prior approaches have aimed to find a set of parameters that correlate with the visual

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

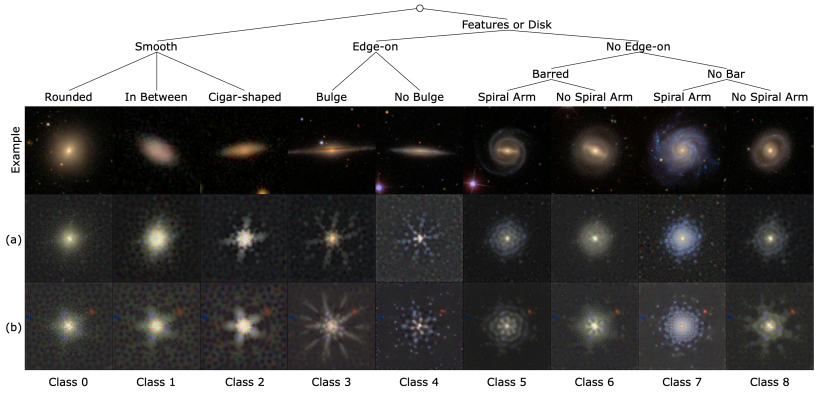

Figure 1: Classification tree based on Galaxy Zoo 2. The bottom two rows are STM-distilled images 128×128 (1 image per class): (a) starting from random real images, without rotational augmentation of the training set; (b) distilled from a rotationally augmented training set with synthetic data initialized from noise.

morphology of a galaxy. These traditionally include concentrations, asymmetries, clumpiness (or smoothness), Gini coefficient, moments of light etc. [25, 13, 23]. Unfortunately, the values of these parameters strongly depend on the data quality and red-shift, as they overlook an enormous amount of information contained in the pixels themselves. Thus, these approaches only provide rough morphological classifications into 2 or 3 classes.

We thus ask: Can the success of deep-learning based classification be leveraged to provide summary representations of morphological information *directly in the shape of galaxy images*? Such a transformation from successful CNN-based classifiers to synthesized summary images could then be extended to classifiers that process additional information (e.g. spectral data) for galaxy classification and generate images that are prototypical of morphology types even when additional non-visual measurement data is incorporated.

In this work we propose to leverage *Dataset Distillation (DD)* as a tool to achieve alternative summarization of galaxy morphologies in image form. DD, originally proposed by [28] in computer vision, can be viewed as a form of dataset curation as a bi-level optimization task involving a neural net classifier. It aims to distill knowledge from a large dataset into a smaller one to reduce the burden of large-scale analysis on images. The dataset distillation optimization performs gradient descent on a synthesized dataset (outer loop) with respect to the loss (on real data) of a network trained on the distilled data (inner loop). Many directions have emerged from the initial bi-level optimization [28], including tractable approximations of the inner loop [21, 22, 20, 35] and new objectives for optimization such as gradient matching [32, 34], trajectory matching [1, 6], distribution matching [27, 33] (see [24] for a recent survey).

Here, we focus on the Matching Training Trajectories (MTT) algorithm [1], reaching recent state-of-the-art for various distillation benchmarks. We propose a new modification to MTT, called Self-Adaptive Trajectory Matching (STM) which allows for enhanced performance and ease on computer vision benchmarks[1]. To apply it to galaxy distillation, we first curate an illustrative customized version of the GZ2 dataset as shown in Figure 1, and train a highly accurate CNN-based classifier on it. By condensing a considerable number of images into one or a few synthetic images for each category of galaxy, we can significantly reduce analysis time while revealing the essential morphological features for these categories. Dataset distillation emphasizes the features in the data that are essential to the classification: for instance, for our galaxy dataset, we will see that it enhances the blue features that are tracers of recent star formation in the galaxy.

2 Methodology and Data

In the Dataset Distillation (DD) framework, the goal is to synthesize a compact dataset $\mathcal{D}_{syn}$ that can replicate the performance of a larger, real dataset $\mathcal{D} \in (\mathcal{X}, \mathcal{Y})$ when used with the same learning algorithm f. The optimal parameters for f estimated on $\mathcal{D}$ and $\mathcal{D}_{syn}$ are represented as $\theta^{\mathcal{D}}$ and $\theta^{\mathcal{D}_{syn}}$, respectively. The objective of DD is to optimize:

[1]Details of the new algorithmic STM approach are relegated to the appendix.

$$\arg\min_{\mathcal{D}_{syn}} \left(|f_{\theta^{\mathcal{D}_{syn}}}(x) - f_{\theta^{\mathcal{D}}}(x)| \quad \forall x \in \mathcal{D}\right) \tag{1}$$

This is an instance of a bilevel optimization problem where the output of one optimization (f trained on $\mathcal{D}_{syn}$) is fed into another optimization problem (the generalization error on $\mathcal{D}$), which is computationally hard.

Matching Training Trajectories (MTT): The MTT method proposed by [1] aims to approximate the optimization in Equation (1) via gradient descent by minimizing the difference between $\theta_t^{\mathcal{D}}$ and $\theta_t^{\mathcal{D}_{syn}}$, for each iteration t of weight updates during training. The objective is to identify a compact dataset $\mathcal{D}_{syn}$ such that when training on it, the model parameters $\theta^{\mathcal{D}_{syn}}$ closely resemble the teacher model parameters $\theta^{\mathcal{D}}$ when trained on the real dataset $\mathcal{D}$ throughout time. If we assume that the learning algorithm f takes T iterations to converge, the MTT objective can be formulated as:

$$\arg\min_{\mathcal{D}_{syn}} \left(\|\theta_{t+N}^{\mathcal{D}_{syn}} - \theta_{t+M}^{\mathcal{D}}\|_2^2 / \|\theta_t^{\mathcal{D}} - \theta_{t+M}^{\mathcal{D}}\|_2^2 \quad \forall t \in [0, \ldots, T) \right) \tag{2}$$

In the equation above, $\theta_t^{\mathcal{D}}$ represents the model parameters for f after t iterations of training on dataset $\mathcal{D}$. The term $\|\theta_t^{\mathcal{D}} - \theta_{t+M}^{\mathcal{D}}\|_2^2$ serves as a normalization factor. While, in theory, the values of N and M should be equal (indicating a comparison after an equal number of gradient descent updates), in practice, especially when mini-batches are utilized for training and given $\mathcal{D} \gg \mathcal{D}_{syn}$, the ratio for $N : M$ becomes a hyperparameter.

MTT, while powerful, exhibits some shortcomings that stand in the way of scalability and practicality. In particular, it has a complex hyperparameter space and lacks a clear stopping criterion. We thus propose **S**elf-adaptive **T**rajectory **M**atching (STM) to achieve two desiderata: eliminate some trajectory hyper-parameters and introduce an early stopping mechanism that can accurately halt the training process upon reaching the optimal result. All further details, as well as benchmarking of STM on standard vision datasets are relegated to Appendix A.1, as here we wish to focus on the topic of the workshop: its connection to physics.

Galaxy Zoo 2 (GZ2) and our curated GZ2: GZ2 [30] is renowned for its vast collection of 243,500 galaxies and the most reliable morphological classifications. In the GZ2 project, human volunteers are presented with galaxy images and are tasked with providing detailed descriptions of their morphologies by answering a series of questions along a classification tree regarding its morphology, such as "Is the galaxy simply smooth and rounded, with no sign of a disk?" The GZ2 tree encompasses 11 classification tasks, with a total of 37 potential responses, leading to a vast number of possible classes with extreme class imbalance, with image counts ranging from 1,761 to 87,139. To partly mediate this, we simplify the classification tree to only 9 leaf nodes (classes), as illustrated in Figure 1 by merging smaller, similar classes. Each classification is labeled with a confidence determined by averaging across responses. To assess the reliability of the dataset, we computed the average confidence level across the 9 classes to lie around 0.53, indicating sub-optimal data quality. To address this and restore class balance, we opted to select the top 600 most confidently classified galaxy images for each class for an average score of 0.79, dividing them into 500 train and 100 test images per class. This "higher-confidence" version of GZ allows for much higher training and test accuracies. For instance, our CNN classifier only achieves 56% accuracy on the entire GZ dataset, while giving 89% test accuracy on our curated dataset. Moreover, the much smaller size of the curated data set enables active learning: astronomers can follow-up on these archetypes with additional spectroscopic and multi-frequency observations. Additionally, noting that data augmentation is often helpful in deep learning applications and given that galaxy imaging does not have a preferred orientation, we also construct an augmented curated GZ2 by rotating each galaxy image 36 degrees for 10 times, resulting in 45,000 train and 900 (non-augmented) test images.

3 Experiments and Results

Experimental Setup: We deploy a simple 3-layer 128-width ConvNet [11], which aligns with the previous DD benchmarks [5]. For data augmentation during the training of the teacher trajectory, we employ DSA [32]. During the distillation process, $\mathcal{D}_{syn}$ can be initialized in two distinct ways:

Table 1: Test accuracy of DD deployed on curated GZ2 dataset (with and without augmentation) with real and noisy initialization, as well as random baseline.

Img/Cls	Random	curated GZ2			curated GZ2-Aug		
		real	noise	Full Dataset	real	noise	Full Dataset
1	27.4 ± 0.5	53.1 ± 1.7	54.6 ± 1.1	84.5 ± 0.6	$\mathbf{54.7 \pm 0.7}$	54.6 ± 2.0	89.0 ± 0.5
10	46.0 ± 0.9	63.9 ± 1.0	62.4 ± 0.9		$\mathbf{68.4 \pm 1.2}$	67.4 ± 1.3	

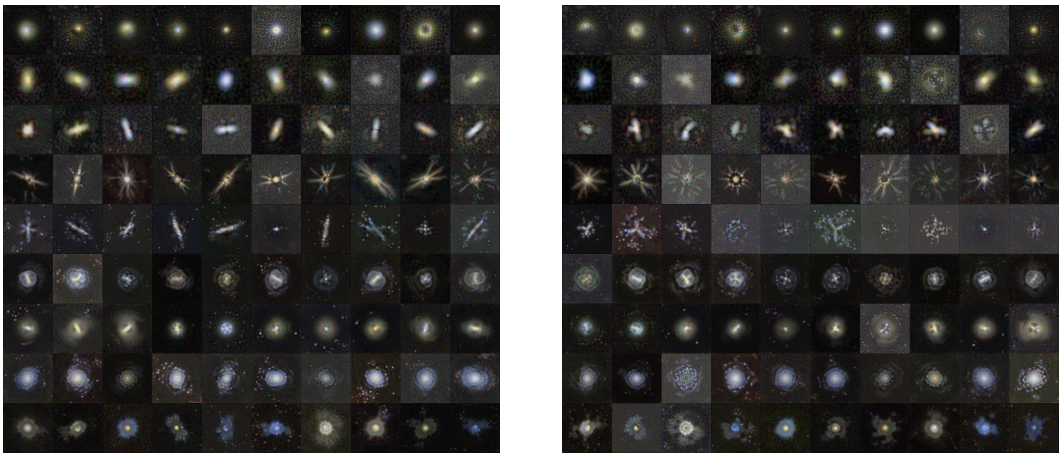

Figure 2: 10 images per class distillation result for curated GZ2 with rotational augmentation; synthetic dataset initialized from real images (left) and noise initial (right). Each row belongs to one class ordered as in Figure 1 (more images in Appendix A.4).

(1) Gaussian noise initialization and (2) initialization with real images, randomly sampled from the original dataset. To assess performance, we train five separate networks from scratch on each distilled dataset and report test accuracy. For comparison, we create a non-synthetic baseline of size $|\mathcal{D}_{syn}|$ by randomly sampling i images from each class in the original dataset and using them to train networks in a manner consistent with the above approach. Detailed hyperparameters setting can be found in Appendix A.3.

Distilling Galaxy Morphologies: Table 1 details the distillation results for both the basic and augmented curated GZ2 datasets, with 1 IPC examples shown in Figure 1. In our experiments, real synthetic data initialization (from random images) typically outperforms initialization from noise. While augmentation does not significantly improve 1 IPC accuracy, it boosts 10 IPC as well as accuracy when trained on the entire data by approximately 5%. Note that our CNN classifier yields a test-accuracy of 89%.

Distilled images capture the learned features of the model, providing insights into galaxy morphology that go beyond predefined characteristics, such as those used in the GZ2 survey questions. By optimizing classification performance, distilled images highlight key attributes for differentiation. The application of data distillation to galaxy imagery signifies the potential for leveraging machine-learned information to provide a fresh perspective and complement existing knowledge about galaxies.

For example, we can look at the 1 IPC distilled images in Figure 1 (a). For Class 0-4 galaxies, key differentiators include core size and arm shape, taking into account the varying orientations of the galaxies. On the other hand, Classes 5&7 exhibit a floral pattern due to asymmetrical arms; a smooth core for Class 7&8 galaxies suggests the absence of a galactic bar. Distilled images provide a better understanding of a galaxy category's overall characteristics. For instance, the blue tint of spiral galaxies indicates the presence of young stars. Compared to 1 IPC images, the 10 IPC versions retain more recognizable features, making them visually closer to real galaxies, as shown in Figure 2.

4 Discussion

We present a novel study that employs dataset distillation for the extraction of galaxy morphology features. We introduce a self-adaptive methodology, STM, which outperforms previous work and provide illustrative results to demonstrate that our approach is capable of distilling knowledge about galaxy morphologies, providing unique insights on key galaxy attributes that are not easily captured by

human-based classification schemes. This opens the possibility of extracting even more informative images when adding spectroscopic or multi-frequency data as additional inputs to the classifier in the future. Our curated GZ2 dataset is of independent interest as an enabler for active learning of additional features on high-quality data.

Software and Data

Our code and the curated GZ2 dataset are available at `https://github.com/HaowenGuan/Galaxy-Dataset-Distillation`.

Acknowledgements

We gratefully acknowledge Adrian Price-Whelan, Marc Huertas-Company, and David Spergel for their pivotal physics insights and guidance regarding galaxy databases and potential insights to be had. Gratitude is also due to Jingtong Su for contributions to the dataset distillation algorithm. Helena Dominguez Sanchez's assistance in preprocessing the galaxy dataset was invaluable. This work was supported by the National Science Foundation under NSF Award 1922658.

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 (CVPR)*, pages 4750–4759, 2022.

[2] Ting-Yun Cheng, Christopher J Conselice, Alfonso Aragón-Salamanca, Michel Aguena, Sahar Allam, F Andrade-Oliveira, J Annis, A F L Bluck, D Brooks, David L Burke, et al. Galaxy morphological classification catalogue of the dark energy survey year 3 data with convolutional neural networks. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 507(3):4425–4444, 2021.

[3] Ting-Yun Cheng, Christopher J Conselice, Alfonso Aragón-Salamanca, Nan Li, Asa F L Bluck, Will G Hartley, James Annis, David Brooks, Peter Doel, Juan García-Bellido, et al. Optimizing automatic morphological classification of galaxies with machine learning and deep learning using dark energy survey imaging. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 493(3):4209–4228, 2020.

[4] Jacob Cohen, Patricia Cohen, Stephen G. West, and Leona S. Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences (3rd ed)*. Routledge, 2002.

[5] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. DC-BENCH: Dataset condensation benchmark. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[6] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.

[7] Sander Dieleman, Kyle W Willett, and Joni Dambre. Rotation-invariant convolutional neural networks for galaxy morphology prediction. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 450(2):1441–1459, 2015.

[8] H Domínguez Sánchez, M Huertas-Company, M Bernardi, D Tuccillo, and J L Fischer. Improving galaxy morphologies for sdss with deep learning. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 476(3):3661–3676, 2018.

[9] Helena Domínguez Sánchez, B Margalef, M Bernardi, and M Huertas-Company. Sdss-iv dr17: final release of manga pymorph photometric and deep-learning morphological catalogues. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 509(3):4024–4036, 2022.

[10] Aritra Ghosh, C Megan Urry, Zhengdong Wang, Kevin Schawinski, Dennis Turp, and Meredith C Powell. Galaxy morphology network: A convolutional neural network used to study morphology and quenching in 100,000 sdss and 20,000 candels galaxies. *The Astrophysical Journal (ApJ)*, 895(2):112, 2020.

[11] Spyros Gidaris and Nikos Komodakis. Dynamic few-shot visual learning without forgetting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4367–4375, 2018.

[12] Ryan Hausen and Brant E Robertson. Morpheus: A deep learning framework for the pixel-level analysis of astronomical image data. *The Astrophysical Journal Supplement Series (ApJS)*, 248(1):20, 2020.

[13] M Huertas-Company, J A L Aguerri, M Bernardi, S Mei, J Sánchez Almeida, et al. Revisiting the hubble sequence in the sdss dr7 spectroscopic sample: a publicly available bayesian automated classification. *Astronomy & Astrophysics (A&A)*, 525:A157, 2011.

[14] M Huertas-Company, R Gravet, Guillermo Cabrera-Vives, Pablo G Pérez-González, J S Kartaltepe, Guillermo Barro, M Bernardi, S Mei, F Shankar, P Dimauro, E F Bell, et al. A catalog of visual-like morphologies in the 5 candels fields using deep learning. *The Astrophysical Journal Supplement Series (ApJS)*, 221(1):8, 2015.

[15] M Huertas-Company, J R Primack, A Dekel, D C Koo, S Lapiner, D Ceverino, R C Simons, G F Snyder, M Bernardi, Z Chen, H Domínguez-Sánchez, et al. Deep learning identifies high-z galaxies in a central blue nugget phase in a characteristic mass range. *The Astrophysical Journal (ApJ)*, 858(2):114, 2018.

[16] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Master's thesis, Dept. of Comp. Sci., University of Toronto, Toronto, ON, Canada, 2009.

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[18] Chris Lintott, Kevin Schawinski, Steven Bamford, Anže Slosar, Kate Land, Daniel Thomas, Edd Edmondson, Karen Masters, Robert C Nichol, M Jordan Raddick, et al. Galaxy zoo 1: data release of morphological classifications for nearly 900 000 galaxies. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 410(1):166–178, 2011.

[19] Chris J Lintott, Kevin Schawinski, Anže Slosar, Kate Land, Steven Bamford, Daniel Thomas, M Jordan Raddick, Robert C Nichol, Alex Szalay, Dan Andreescu, et al. Galaxy zoo: morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 389(3):1179–1189, 2008.

[20] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[21] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[22] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, pages 5186–5198, 2021.

[23] Michael A Peth, Jennifer M Lotz, Peter E Freeman, Conor McPartland, S Alireza Mortazavi, Gregory F Snyder, Guillermo Barro, Norman A Grogin, Yicheng Guo, Shoubaneh Hemmati, et al. Beyond spheroids and discs: classifications of candels galaxy structure at 1.4 <z <2 via principal component analysis. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 458(1):963–987, 2016.

[24] Noveen Sachdeva and Julian McAuley. Data distillation: A survey. *arXiv preprint arXiv:2301.04272*, 2023.

[25] C Scarlata, C Marcella Carollo, S Lilly, M T Sargent, R Feldmann, P Kampczyk, C Porciani, A Koekemoer, N Scoville, J-P Kneib, et al. Cosmos morphological classification with the zurich estimator of structural types (zest) and the evolution since z = 1 of the luminosity function of early, disk, and irregular galaxies. *The Astrophysical Journal Supplement Series (ApJS)*, 172(1):406, 2007.

[26] Mike Walmsley, Lewis Smith, Chris Lintott, Yarin Gal, Steven Bamford, Hugh Dickinson, Lucy Fortson, Sandor Kruk, Karen Masters, Claudia Scarlata, et al. Galaxy zoo: probabilistic morphology through bayesian cnns and active learning. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 491(2):1554–1574, 2020.

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

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

[29] Wikipedia. Statistical hypothesis testing — Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Statistical%20hypothesis%20testing&oldid=1169316066, 2023. [Online; accessed 10-October-2023].

[30] Kyle W Willett, Chris J Lintott, Steven P Bamford, Karen L Masters, Brooke D Simmons, Kevin R V Casteels, Edward M Edmondson, Lucy F Fortson, Sugata Kaviraj, William C Keel, et al. Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society (MNRAS)*, 435(4):2835–2860, 2013.

[31] Donald G York, J Adelman, John E Anderson, Jr., Scott F Anderson, James Annis, Neta A Bahcall, J A Bakken, Robert Barkhouser, Steven Bastian, Eileen Berman, et al. The sloan digital sky survey: Technical summary. *The Astronomical Journal (AJ)*, 120(3):1579, 2000.

[32] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 12674–12685, 2021.

[33] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2023.

[34] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[35] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

A Appendix

A.1 Self-Adaptive Trajectory Matching (STM) - an improved Matching Trajectories Algorithm

MTT [1], while powerful, exhibits some shortcomings that stand in the way of scalability and practicality. i) MTT has a relatively large number of hyper-parameters that interact in a complex way: there are three parameters to control the trajectory matching mode and three different learning rates to configure, necessitating a significant amount of grid searching in hyper-parameter-space to achieve the optimal result. For instance, TESLA [6], a memory-optimized variant of MTT, outperforms MTT by varying hyper-parameter settings. ii) MTT lacks a clear stopping criterion for the distillation; the training process runs through a predetermined iterations. However, different datasets require different optimal iterations, and fixing the maximum iteration can lead to either excessive computation or sub-optimal training.

Algorithm 1: Self-Adaptive Trajectory Matching (STM)

Input: Teacher parameter trajectory set $\Theta^{\mathcal{D}}$, student network matching steps N, initial step size α, threshold value for hypothesis test λ, maximum iterations per stage Max_{iter}

1 **Initialize:** $\mathcal{D}_{syn}$, $iter = 0$, $t = 0$, $T = 1$
2 **while** $iter < Max_{iter}$ **do**
3 Increment $iter$ and t by 1, if t reaches T, reset t back to 0
4 Sample $\theta_t^{\mathcal{D}}$ and $\theta_{t+1}^{\mathcal{D}} \in \Theta^{\mathcal{D}}$, set $\theta_t^{\mathcal{D}_{syn}} = \theta_t^{\mathcal{D}}$
5 **for** $i = 1, ..., N$ **do**
6 Update $\theta_{t+i}^{\mathcal{D}_{syn}} = \theta_{t+i-1}^{\mathcal{D}_{syn}} - \alpha \nabla \ell(\theta_{t+i-1}^{\mathcal{D}_{syn}}; \mathcal{D}_{syn})$
7 Update $\mathcal{D}_{syn}$ and α [2]via gradient descent, minimizing $\frac{\|\theta_{t+N}^{\mathcal{D}_{syn}} - \theta_{t+1}^{\mathcal{D}}\|_2^2}{\|\theta_t^{\mathcal{D}} - \theta_{t+1}^{\mathcal{D}}\|_2^2}$
8 Repeat lines 4-6 replacing t by T to get $\theta_{T+N}^{\mathcal{D}_{syn}}$
9 Collect validation loss $\frac{\|\theta_{T+N}^{\mathcal{D}_{syn}} - \theta_{T+1}^{\mathcal{D}}\|_2^2}{\|\theta_T^{\mathcal{D}} - \theta_{T+1}^{\mathcal{D}}\|_2^2}$ into array ℓ_{val}
 `/* Expand epoch pool if validation loss` ℓ_{val} `decreases fast enough */`
10 **if** $corr(\ell_{val}, time) < -\lambda\sqrt{1/(size(\ell_{val}) - 2)}$ **then**
11 Expand epoch pool by increment T, Reset $iter, \ell_{val}$

Output: $\mathcal{D}_{syn}$

Self-Adaptive Trajectory Matching (STM): To simplify and remedy some of these shortcomings, we propose **S**elf-adaptive **T**rajectory **M**atching (STM) that achieves two desiderata: it eliminates the need for M and T in Equation (2), and introduces an early stopping mechanism that can accurately halt the training process upon reaching the optimal result.

In vanilla trajectory matching algorithms, distillation involves selecting a maximum starting epoch T and randomly sampling a starting point $t \in [0, \ldots, T)$ on the trajectory $\Theta^{\mathcal{D}} := \{\theta_t^{\mathcal{D}}\}_0^{T-1}$ to proceed with parameter matching as in Equation 2. The T can be interpreted as the size of trajectory $\Theta^{\mathcal{D}}$. MTT [1] demonstrates that the trajectory size T and the distillation performance exhibit a parabolic relationship, and the optimal T is positively correlated with the number of images per class (IPC) to distill for the synthetic dataset. An interpretation is that each teacher epoch $\theta_t^{\mathcal{D}}$ carries an amount of knowledge, and the synthetic dataset (student) has a certain "capacity". If we can measure the capacity of a synthetic dataset during distillation, we can decide whether we should feed more teacher epochs to the student or decide to end the distillation.

To achieve this goal, we propose using a validation loss curve as an indicator of the capacity of the synthetic dataset. This validation loss is calculated by matching $\mathcal{D}_{syn}$ on a part of the trajectory outside the training trajectory $\Theta^{\mathcal{D}}$. If the validation loss decreases in a statistically significant manner, we infer that the synthetic dataset possesses "capacity for more knowledge." Consequently, we

[2]As in MTT [1], we make the step size α learnable. This adaptability enables the distillation algorithm to autonomously determine the optimal step size for aligning with the teacher trajectory.

expand $\Theta^{\mathcal{D}}$ by adding $\theta_T^{\mathcal{D}}$ (or equivalently, by incrementing T). To determine statistically significant decreases in validation loss, we employ hypothesis testing [29] on the correlation between validation losses (as time-series data) and time (distillation steps). Specifically, the null hypothesis is that the time-series data display an average zero correlation with time, and the deviation σ is proportional to $\sqrt{1/(size(\ell_{val})-2)}$ (see Appendix A.2). We establish a threshold of $\lambda\sigma$ to convincingly reject the null and expand training trajectory $\Theta^{\mathcal{D}}$ during distillation. If we cannot reject the null hypothesis after a certain maximum distillation steps, denoted as Max_{iter}, we stop. Furthermore, our findings suggest that fixing the teacher matching epoch to $M = 1$ is optimal, and cycling through $[0, \ldots, T)$ is better than randomly sampling from that interval. An ablation study on the parameter λ is presented in Figure 4, and shows that there is minimal variation when the value is sufficiently large. Our approach is summarized in Algorithm 1.

Benchmarking STM: We believe our STM method to be of independent interest and have benchmarked it on CIFAR-10 and CIFAR-100 [16], two key computer vision datasets, using 1/10/50 images per class (IPC) in Table 2. We compare it with the original MTT [1] and a baseline model trained on the same number of randomly selected images from each class (Random). STM shows slight performance gains over MTT as it aims to streamline the data distillation process, enhance the algorithm's robustness, and enable it to consistently achieve optimal results. Figure 3 indeed shows faster convergence and much smaller variance between trials. We see that STM consistently yields improvement, both in performance and in convergence and stability, and we advocate to adapt MTT methods that are currently used to include the modifications brought by STM.

Table 2: Performance (test accuracy %) of MTT and STM, trained on distilled data initialized from random real images.

	Img/Cls	Random	MTT	Ours	Full Dataset
CIFAR-10	1	14.4 ± 2.0	46.3 ± 0.8	**47.7 ± 0.1**	84.8 ± 0.1
	10	26.0 ± 1.2	65.3 ± 0.7	**65.7 ± 0.2**	
	50	43.4 ± 1.0	71.6 ± 0.2	**72.7 ± 0.2**	
CIFAR-100	1	4.2 ± 0.3	24.3 ± 0.3	**24.4 ± 0.4**	56.2 ± 0.3
	10	14.6 ± 0.5	40.1 ± 0.4	**41.6 ± 0.3**	

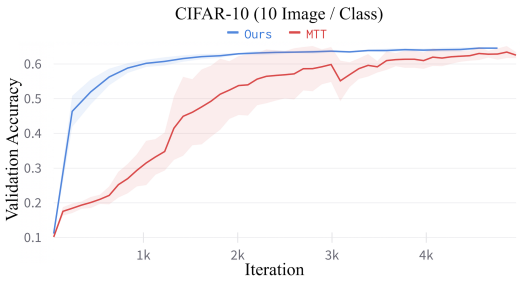

Figure 3: Comparison of MTT and STM distillation process (3 trials each, starting from random real images). STM shows a faster convergence speed and higher final accuracy.

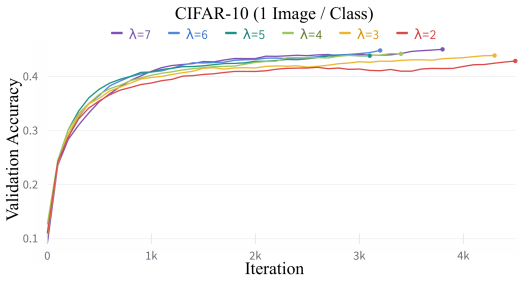

Figure 4: Ablation study for λ-sigma threshold of STM. A higher value implies a stricter requirement for statistical significance of decreasing loss trend. We fix our algorithm to use 5-sigma.

A.2 Standard Error of Correlation Coefficient Between Random Data and Time

The correlation coefficient, denoted as r, quantifies the strength and direction of the linear relationship between two sequences of data with the same length n. The standard error S_r of a correlation coefficient is given by (see e.g. [4]):

$$S_r = \sqrt{\frac{1-r^2}{n-2}}$$

Consider a scenario where we sample n data points from a normal distribution $\mathcal{N}(0, 1)$ and treat them as a time series. When computing the correlation coefficient of this data with respect to time, the true correlation coefficient is $r = 0$, since a randomly sampled dataset is expected to have no correlation with time. In this specific case, the standard error (or deviation σ) is:

$$\sigma = \sqrt{\frac{1}{n-2}}$$

which is the functional form we use in Algorithm 1.

A.3 Experimental Setup and Hyperparameters

We deploy a simple 3-layer 128-width ConvNet [11] following previous DD benchmark [5]. Like MTT, we apply ZCA whitening on all benchmark datasets for image preprocessing, and employ DSA [32] for augmentation during training and evaluation. While most distillation hyper-parameters remain the same as MTT to ensure a fair comparison, certain parameters are adjusted due to modifications in the algorithm.

Table 3 shows the hyperparameters we use for STM for various datasets. We set $\lambda = 5$ in our hypothesis testing step but note that the algorithm is highly insensitive to the value of λ. We also have a parameter for maximum distillation steps, Max_{iter}; its value is fixed to $\text{Max}_{iter} = 1000$ in our experiment.

Table 3: Hyper-parameters used for our best-performing distillation experiments. We adopt the terminology and definitions from MTT [1] (Pixel, Step Size).In the STM algorithm, M, the number of expert epochs is fixed to 1.

dataset	Img/Cls	Synthetic Steps (N)	Learning Rate (Pixels)	Initial Step Size (α)	Learning Rate (Step Size)	ZCA
CIFAR-10	1	50	1000	0.01	0.01	Y
	10	30	1000	0.01	0.01	Y
	50	30	1000	0.01	0.01	Y
CIFAR-100	1	20	1000	0.01	0.01	Y
	10	20	1000	0.01	0.01	Y
GZoo2	1	50	10000	0.0001	0.01	N
	10	20	10000	0.0001	0.01	N

A.4 Additional images

Here we present additional visualizations of the distilled data from GalaxyZoo for various initializations (random or from noise) and distilled images per class.

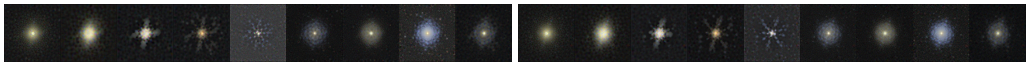

Figure 5: Distilled images 1 img/class without augmentation. Left: initialize from noise; Right: initialize from real

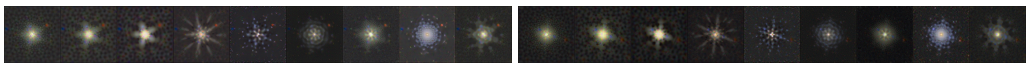

Figure 6: Distilled images 1 img/class with augmentation. Left: initialize from noise; Right: initialize from real

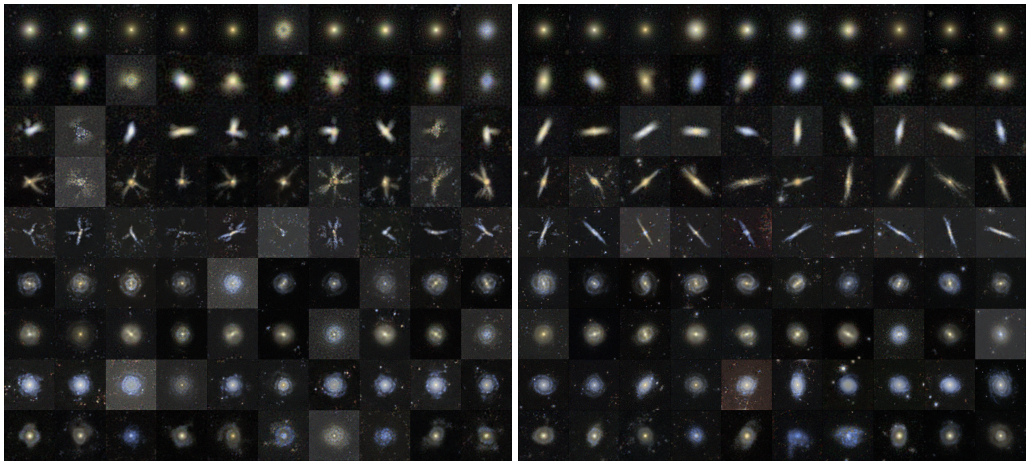

Figure 7: Distilled images 10 img/class without augmentation. Left: initialize from noise; Right: initialize from real