One Category One Prompt: Dataset Distillation using Diffusion Models

Ali Abbasi
Vanderbilt University
ali.abbasi@vanderbilt.edu

Ashkan Shahbazi
Vanderbilt University
ashkan.shahbazi@vanderbilt.edu

Hamed Pirsiavash
University of California, Davis
hpirsiav@ucdavis.edu

Soheil Kolouri
Vanderbilt University
soheil.kolouri@vanderbilt.edu

Abstract

The extensive amounts of data required for training deep neural networks pose significant challenges on storage and transmission fronts. To address this, dataset distillation has emerged as a promising technique, aiming to condense the information of massive datasets into a much smaller yet representative set of synthetic samples. However, traditional dataset distillation approaches often struggle to scale effectively with high-resolution images and more complex architectures due to the inherent bi-level optimization. Recently, several works have proposed exploiting knowledge distillation with decoupled optimization schemes to scale up dataset distillation. Although these methods effectively address the scalability issue, they rely on extensive image augmentations requiring the storage of soft labels for augmented images. In this paper, we introduce Dataset Distillation using Diffusion Models (D3M) as a novel paradigm for dataset distillation, leveraging recent advancements in generative text-to-image foundation models. Our approach utilizes textual inversion, a technique for fine-tuning text-to-image generative models, to create concise and informative representations for large datasets. By employing these learned text prompts, we can efficiently store and infer new samples for introducing data variability within a fixed memory budget. We show the effectiveness of our method through extensive experiments across various computer vision benchmark datasets with different memory budgets.

1. Introduction

The substantial data required for training deep neural networks [16] presents challenges in storage, transmission, and handling, which complicate tasks like hyperparameter optimization [21, 23]. Additionally, these datasets raise copyright and privacy concerns, as noted in [37], prompting questions about dataset compression: can a small synthetic subset achieve competitive model performance? This concept, known as Dataset Distillation [37] or Dataset Condensation [43], has led to extensive research [3, 5, 7, 12, 19, 33, 34, 40, 46].

Dataset distillation is often framed as a bi-level optimization problem: the inner optimization trains a model on a distilled set, while the outer optimization refines the distilled samples [37]. This approach requires significant computational resources, as the outer loop must backpropagate through the inner training. Some methods tackle this by using surrogate objectives, such as gradient matching [43] or training trajectory matching [2]. However, scaling these methods to larger datasets remains challenging [40]. To address this, recent works decouple the bi-level process into single-level procedures [34, 40] for scalability on datasets like ImageNet-1K [8].

Another issue is that synthetic datasets are often optimized for specific architectures, which limits generalization [3]. Generating realistic images has shown promise for improving cross-architecture generalizability by better matching the training data manifold [3, 34, 40]. Approaches include generative models [3], pretrained batch statistics [40], or collaging patches from training data [34].

This work, inspired by foundation models [1, 29, 30] and the question raised by Wang et al. [37], asks: **how compressible is a dataset with access to a foundation model?** Such compression could reduce storage costs and facilitate data accessibility via universal APIs. **We demonstrate that ImageNet categories can be compressed into single prompts with latent diffusion models [30]**, achieving state-of-the-art dataset distillation performance. Our proposed framework, Dataset Distillation using Diffusion Models (D3M), builds on recent advances [3, 11, 30, 34, 40] to enable unprecedented compression of large datasets with competitive cross-architecture generalization.

Contributions. Our contributions in this paper are:

1. Achieving unprecedented condensation rates for dataset

distillation via diffusion models [30].

2. Adapting textual inversion [11] to dataset distillation, and demonstrating the possibility of generating realistic and diverse collages of images representing an image category via a single prompt.
3. offering a solution for the memory overhead associated with storing soft-labels and discussing its trade-offs in large-scale dataset distillation.

2. Related Work

Current dataset distillation research can be divided into four main approaches: 1) bi-level optimization, 2) uni-level relaxation, 3) coreset selection, and 4) generative priors for better alignment with the data manifold.

Bi-level optimization-based dataset distillation. This approach frames dataset distillation as a bi-level optimization problem [37], with an outer loop that optimizes the distilled dataset and an inner loop that trains a model on this dataset. To reduce computational costs, various surrogate objectives are proposed: gradient matching [15, 41, 43], feature alignment [36, 42], and trajectory alignment [2, 6, 12]. Although effective for small datasets, bi-level methods struggle with scaling to larger datasets and generalizing across architectures. Our work avoids bi-level optimization to facilitate scaling and improve generalization.

Uni-level relaxation of dataset distillation. A promising line of research uses Neural Tangent Kernels (NTKs) [14, 26], yielding a closed-form solution that converts bi-level optimization to uni-level [20, 25]. While effective on small datasets, NTK methods face challenges with high-resolution datasets. Recent alternatives [34, 40] decouple the bi-level optimization into two uni-level problems, achieving scalability for large datasets like ImageNet. Similarly, we also leverage a uni-level optimization approach.

Coreset selection for efficient machine learning. Unlike dataset distillation, which generates synthetic samples [37], coreset selection [22, 24, 28] focuses on identifying a small subset of the original training set that provides competitive model performance. A key benefit of coreset selection is that the samples remain on the data manifold, supporting stronger generalization across architectures. However, these methods often rely on bi-level optimization or difficulty-based metrics like forgetting scores [35] and EL2N scores [27]. Our distilled dataset images are synthetic but appear realistic due to diffusion models.

Generative priors for dataset distillation. Generative priors have proven beneficial for dataset distillation. Cazenavette et al. [3] demonstrate that bi-level optimization in a generative model's latent space can enhance performance [2, 42]. A balanced generative prior supports generalization, whereas overly strong priors may limit expressivity [3]. Yin et al. [40] use a teacher model's batch statistics to guide distillation towards more realistic images, aligning with techniques from model inversion [10, 38]. Sun et al. [34] emphasize diversity and realism, recommending synthetic collages that capture valuable foregrounds while discarding irrelevant backgrounds. Building on these insights, our method uses latent diffusion models [30] and textual inversion [11] to create diverse, foreground-focused distilled samples.

3. Method

We strive to distill large-scale datasets into condensed representations that maintain high accuracy when employed for training classification or regression models. In line with recent advancements in large-scale data distillation [34, 40], we steer clear of bi-level optimization. The pivotal components of our proposed framework are outlined as follows:

1. Prioritizing informative patches from training images and optimizing the utilization of the number of images per category by generating collage images composed of these patches, akin to the approach outlined in [34].
2. Instead of directly storing collages of important patches for each category, we utilize a text-to-image diffusion model and employ textual inversion techniques [11] to generate prompts that directly create the collage images, enabling on demand generation of collage images.

During classifier training, the diffusion model can be applied to the stored prompts using either in-house or API-based services. This allows for the efficient generation of collage images, which can then be effectively utilized for classifier training. Our approach is motivated by the enhanced dataset condensation enabled by the low dimensionality of textual prompts in text-to-image diffusion models. Additionally, we leverage the remarkable capability of generative models to concentrate the most discriminative details of an entire category using just a single low-dimensional prompt vector (e.g., 768 scalars). Our proposed framework, denoted as Dataset Distillation using Diffusion Models (D3M), is shown in Figure 1, and it consists of three main steps for condensing the dataset, and a fourth step for training a classifier/regressor. Below, we delve into a detailed explanation of these steps.

3.1. Step 1: Collage of Important Patches

Let us denote the training data as $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^{N}$, where $x_i \in \mathbb{R}^{H\times W\times 3}$ represents the i'th training image, and $y_i \in \mathbb{R}^{K}$ is the corresponding label. Moreover, we denote the data belonging to category c as $\mathcal{D}_c := \{x_i | y_i = c\}_{i=1}^{N_c}$. Additionally, let $f : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^{K}$ denote a pretrained teacher model on dataset $\mathcal{D}$. Our goal in collage generation is to first identify an informative patch for each image, denoted as $x'_i \in \mathbb{R}^{H_p\times W_p\times 3}$. To achieve this, we follow the methodology of Sun et al. [34] and solve the following optimization problem:

$$x_i^* = \text{argmax}_{x'\sim p(x'_i|x_i)}\, \ell(f(x'), y_i) \tag{1}$$

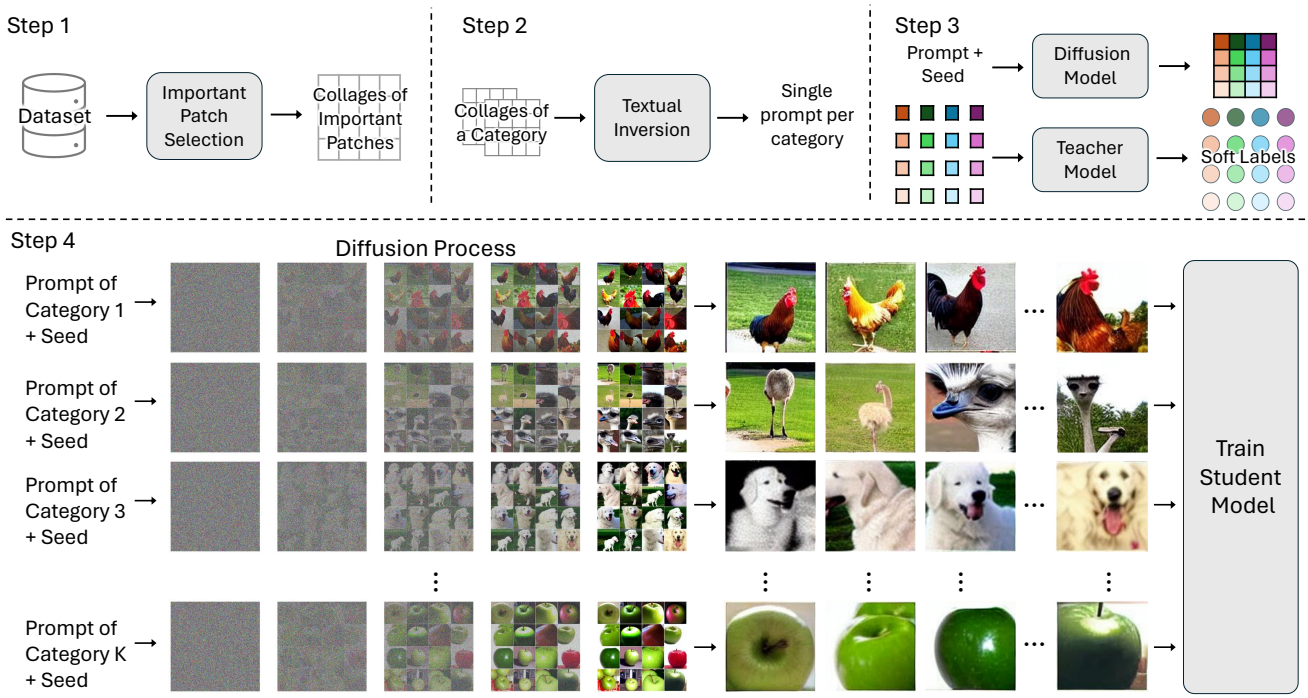

Figure 1. Illustrating the core steps in our proposed framework, Dataset Distillation using Diffusion Models (D3M). Step 1 follows the work of [34] and utilizes a teacher network to identify important patches of the training data and create collages of these patches. Step 2 employs textual inversion [11] to optimize a single prompt per category, resulting in the creation of collage images through stable diffusion [30]. Regarding labels, we consider two different settings, namely, one-hot and soft-labels. To generate soft-labels for synthetic images of each category, in Step 4, a random seed is fixed, and stable diffusion is utilized to generate collage images, which are then fed to the teacher network to obtain the soft-labels. Finally, in Step 4, the categorical prompts and random seeds are employed to create the distilled dataset and train the student.

where $\ell(\cdot,\cdot)$ denotes the cross-entropy loss, and $p(x_i'|x_i)$ is the probability density function of $H_p \times W_p$ patches of the $H \times W$ image. We approximate this optimization problem by first randomly sampling a set of patches from the input image, and then feeding them to the pretrained and frozen teacher model, f, and selecting the patch with the minimum cross-entropy loss. Alternatively, visual explainability methods like class activation maps (CAM) [45] and its variations [4, 32] could be used to identify the important patches and guide the collage generation.

Having found the important patches, $\mathcal{D}_c^* = \{x_i^* | y_i = c\}_{i=1}^{N_c}$ we generate grids of these patches to create collage images. We denote these collage images for each category via $\{X_j^c \in \mathbb{R}^{H \times W \times 3}\}_{j=1}^{L_c}$. Figure 2 demonstrates this concept, when $H = W = 512$ and $H_p = W_p = 128$, and X_j^c is a 4×4 collage image of important patches.

3.2. Step 2: Textual Inversion

In the second step of our framework, our goal is to reduce the generated collage images into textual prompts for a text-to-image diffusion model, e.g., the Latent Diffusion Model (LDM) [30]. We aim to find an optimal prompt (per category), which in turn leads to the generation of realistic-looking collage images for each category. We adopt the textual inversion framework proposed by Gal et al. [11] to optimize such prompts. Throughout the remainder of this subsection, we present the diffusion equations without distinction between whether diffusion is applied in the raw pixel space or in the latent space of an auto-encoder, as our discussion applies to both settings.

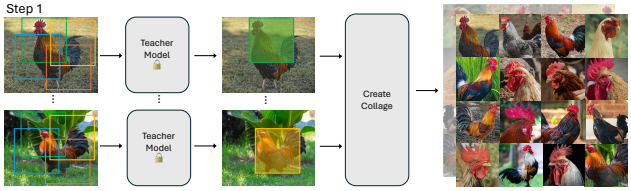

Figure 2. Following the work of Sun et al. [34], we first identify a patch per input image that results in the lowest cross-entropy loss for a pre-trained and frozen teacher model. Then, we construct collage images of these important patches.

Let $p_X(c)$ denote the distribution of collage images of category c from Step 1, $X(t)$ denote a collage image X noised to time t, and $\epsilon \sim \mathcal{N}(0,1)$ denote an unscaled noise sample used to create $X(t)$. Moreover, let ϕ denote the conditional generator (i.e., the denoiser), and ρ denote a text encoder/transformer (e.g., BERT [9]) that encodes the

textual prompt into a conditioning vector for the denoiser. Following the approach in [11], we designate a placeholder string, S_*, to represent the new concept we wish to learn, and utilize the following prompt: "A photo of S_*." The word embedding for the placeholder S_* is what we refer to as the "prompt," denoted by v. Then, for category c, we optimize this prompt via the following optimization problem (see Figure 4):

$$v_*^c = \text{argmin}_{v \in \mathbb{R}^d} \mathbb{E}_{X \sim p_X(c), \epsilon \sim \mathcal{N}(0,1), t} \left[\|\epsilon - \phi(X(t), t, \rho(v))\|^2 \right] \quad (2)$$

The beauty of this framework lies in the fact that the text encoder and the diffusion model remain frozen or unchanged, while a single textual token (i.e., word embedding) is optimized to enable the diffusion model to generate collages of important patches. This significantly increases the rate of dataset compression, as it allows us to represent an entire category of images with a d-dimensional vector, v_*^c, where, for instance, $d = 768$ (see Figure 4).

Note that, with abuse of notation, we denote the entire inverse diffusion process as $X^c = \Phi(\epsilon, \rho(v_*^c))$. Moreover, to demonstrate the effectiveness of the textual inversion framework for generating collage images, we provide a qualitative comparison between the generated collage images using the optimized prompt $\Phi(\epsilon, \rho(v_*^c))$, versus using an engineered prompt like "A 4×4 natural collage of 'name_of_class' images," in Figure 3. We can clearly see that the textual inversion recovers images that are closer to realistic collages both visually and semantically. The quantitative results of these experiments are included in Appendix A1.

Next, we describe the soft labeling process used in our framework.

3.3. Step 3: One Hot vs. Soft Labeling

For dataset distillation utilizing one-hot labels, storing $[v_*^c, c]$ is sufficient to represent category c. However, recent studies [34, 40] highlight the advantages of employing soft-labeling techniques, leveraging a pre-trained teacher model. To incorporate soft labeling into our framework, we initially recognize that for a fixed $\rho(v)$, the stochasticity of the diffusion model can be encapsulated within the random generator seed. This implies that possessing the random generator seed alongside $\rho(v)$ uniquely identifies a collage image. Given a fixed seed, after generating a collage image for category c via $\Phi(\epsilon, \rho(v_*^c))$, we partition it into its constituent patches and feed these patches through the pre-trained and frozen teacher model to derive the soft labels. Additionally, alongside $[v_*^c, c]$, we store the random generator seed and the computed soft labels to facilitate training the classifier with these images and soft labels. Notably, for higher IPCs, it suffices to store more seeds (a scalar per image) and the corresponding soft labels, which proves to be significantly more economical than preserving entire images.

4. Experiments

To demonstrate D3M's effectiveness, we conducted extensive experiments on multiple benchmark datasets.

4.1. Datasets

Following previous protocols [3, 34, 40], we used:

1. **CIFAR-10 [17]:**A 10-class dataset with 32 × 32 images. We used a 128-width, 3-layer CNN (ConvNet-3) for distillation and evaluation.
2. **CIFAR-100 [17]:** A 100-class dataset with 32 × 32 images, also using ConvNet-3.
3. **Tiny-Imagenet [18]:** A 200-class subset of ImageNet with 64 × 64 images. ConvNet-4 was used for small-scale experiments, while ResNet-18 [13] was used for large-scale.
4. **ImageNet-100 [31]:** A 100-class subset of ImageNet-1k with 224 × 224 images, evaluated with ResNet-18.
5. **ImageNet-1k [8]:** The full 1,000-class dataset with 224× 224 images, also evaluated with ResNet-18.

4.2. Baselines

We divided the methods into two main categories: 1) Bi-level-optimization-based and 2) Knowledge-distillation-based methods. Here, we briefly introduce our selected baselines and their categories. For bilevel-optimization approaches, we used the following baselines:

- **MTT** [2] proposes to first generate a dataset of expert trajectories and then poses a bilevel optimization problem to match the trajectory of the student to those of the teachers.
- **IDM** [44] proposes to match the output distribution of the synthetic and the real images while improving the computational cost of the previous works.
- **TESLA** [7] solves a very similar objective to MTT's [2] by calculating the exact unrolled gradients with a constant memory complexity.
- **DATM** [12] distinguishes the early and late stages of the expert trajectories and proposes a difficulty-aware solution for the trajectory matching.

For the knowledge-distillation-based approaches we used these recent baselines:

- **SRe2L** [40] is the first knowledge-distillation-based dataset distillation or condensation method. They propose to synthesize the distilled images using ideas from model inversion and, more particularly, utilizing the batch norm statistics of the teacher network.
- **CDA** [39] is another knowledge-distillation-based method that proposes a curriculum for the parameters of RandomResizedCrop augmentation while synthesizing the data.
- **RDED** [34] is a nascent method that introduces collages as an effective technique for dataset condensation. They showed that the important patches within each collage help with incorporating the pixel budget more efficiently.

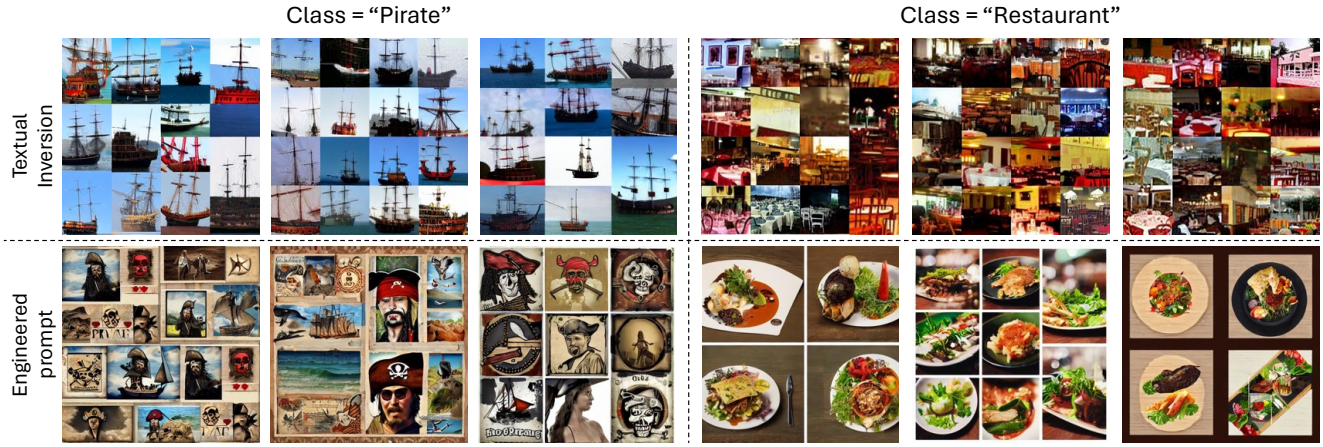

Figure 3. Comparison of collage images generated by textual inversion versus the engineered prompt, “A 4×4 natural collage of ‘name_of_class’ images,” for classes ‘pirate’ and ‘restaurant’ in ImageNet-1k dataset.

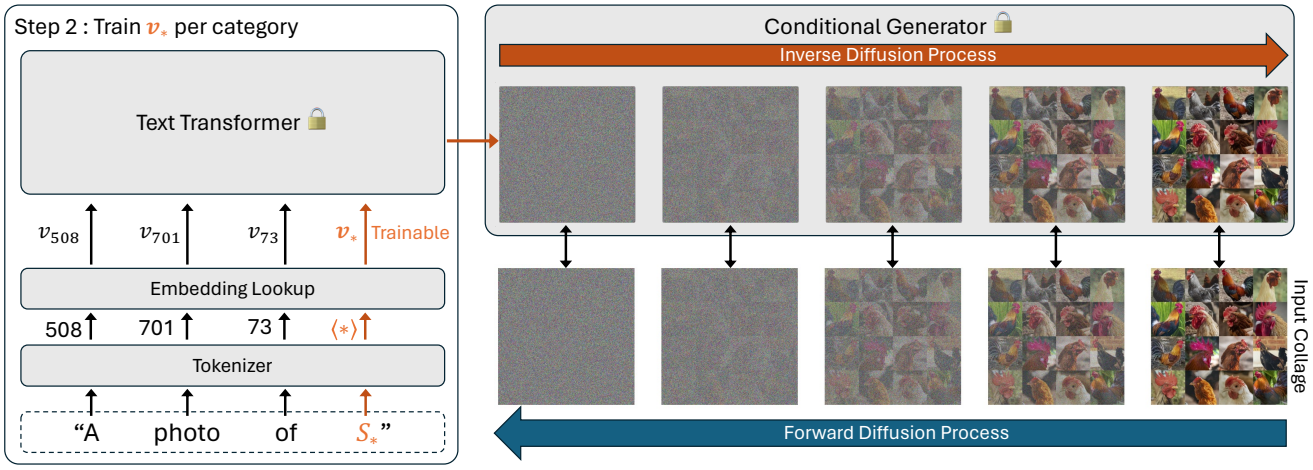

Figure 4. Given a text-to-image diffusion model, such as Latent Diffusion Model (LDM) [30], for each category of the training data, we employ ‘textual inversion’ [11] to optimize a token (i.e., a prompt), v_*, resulting in the generation of collage images that are similar to the ones constructed in Step 1 (Figure 2).

Experimental setup: We evaluated the knowledge-distillation-based methods as well as ours in the following two scenarios: 1) utilizing one-hot labels and performing augmentations on the distilled data, 2) storing the soft labels per image but without augmentations. We emphasize that in the learning-with-soft-labels setting, one cannot utilize data augmentation. Doing so would render the stored soft labels obsolete and would require access to the teacher model to update the soft labels, which defeats the purpose of dataset distillation. We provide a fair comparison of the baselines under the specified settings. We repeat each experiment three times and report the mean and standard deviation for three different image-per-category (IPC) values, namely IPC $\in \{1, 10, 50\}$.

The large-scale and small-scale results are summarized in Tables 1 and 2, respectively. Importantly, some bilevel-optimization-based methods, such as [7], also learn the soft labels along with the synthetic data. We evaluated these approaches as suggested in their respective papers. Please note that each collage will be cut and resized into the allowed IPC budget, following the approach outlined in [34]. Specifically, in CIFAR-10, CIFAR-100, and Tiny-ImageNet, the collage used for training consists of just one patch of the class instance, resized to dimensions of 32×32, 32×32, and 64×64, respectively. For ImageNet-100 and ImageNet-1K, 2×2 collage images are resized to a spatial resolution of 224×224, containing four patches in total. Lastly, we investigate the effect of the number of patches in the collage images on the classification accuracy of the trained classifier in our ablation studies.

4.3. Accuracy vs. Compression

While Tables 1 and 2 showcase the remarkable performance of our proposed single-prompt dataset distillation framework, D3M, they do not entirely capture the compression benefits inherent in the proposed method. It is crucial to highlight that higher compression rates would translate to

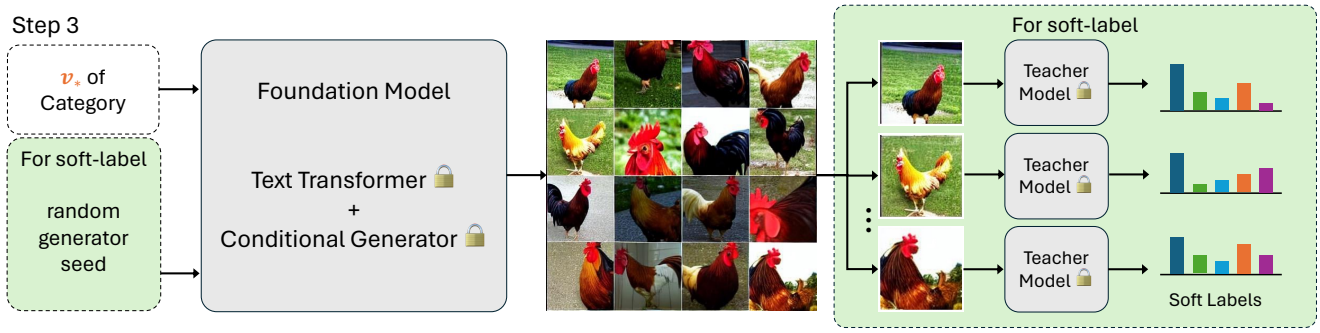

Figure 5. We conduct dataset distillation under two distinct settings. In the one-hot setting, each image category is solely represented by the prompt, v_*, derived from the textual inversion process illustrated in Figure 4. Conversely, in the soft-label setting, alongside the prompt v_*, we retain the random generator seed and the soft labels corresponding to the generated patches. It's noteworthy that the soft labels and seed will scale linearly with IPC, while the prompt remains fixed for the entire category, thus remaining independent of IPC.

		Tiny ImageNet			ImageNet 100			ImageNet 1k		
	Method	IPC=1	IPC=10	IPC=50	IPC=1	IPC=10	IPC=50	IPC=1	IPC=10	IPC=50
Hot Labels	SRe2L	1.64(.06)	4.04(.06)	18.39(.49)	1.75(.02)	4.34(.18)	10.31(.39)	0.26(.03)	1.81(.06)	4.09(.09)
	CDA	1.16(.03)	2.74(.03)	14.98(.37)	1.53(.02)	4.13(.25)	10.33(.53)	0.22(.02)	1.48(.01)	5.83(.14)
	RDED	3.00(.07)	9.99(.25)	24.59(.12)	5.23(.10)	14.64(.73)	35.91(.41)	1.12(.03)	8.61(.10)	26.28(.15)
	Ours	4.98(.06)	14.16(.64)	18.46 (.65)	7.57(.24)	16.42(1.41)	16.96(1.08)	6.22(.10)	12.27 (.57)	12.38 (.67)
Soft Labels	SRe2L	2.36(.06)	9.50(.30)	27.1(.29)	2.41(.12)	6.85(.31)	16.42(.81)	0.44(.02)	1.45(.03)	5.56(.10)
	CDA	2.42(.19)	9.36(.27)	26.40(.41)	1.80(.31)	7.22(.01)	15.56(.55)	0.51(.01)	1.44(.05)	5.83(.14)
	RDED	5.06(.18)	19.67(.45)	41.45(.07)	6.85(.50)	24.13(.10)	49.51(.18)	1.66(.04)	6.40(.13)	19.60(.10)
	Ours	11.36(.51)	38.76(.30)	51.43 (.24)	13.1(1.02)	42.27(.54)	45.25(.43)	5.04 (.15)	23.57(.05)	32.23 (.12)

Table 1. Accuracy of knowledge-based distillation approaches in large-scale experiments on Tiny ImageNet, ImageNet 100, and ImageNet 1K, with IPC values of 1, 10, and 50. Results are presented for both one-hot labels and soft labels settings. The experiments are repeated three times, and the means and standard deviations (stds) are reported, with stds shown in parentheses. For these experiments, we used a ResNet-18 both for generating the soft labels and for the evaluation stage.

more efficient communication and storage, particularly in bandwidth-constrained environments. In this context, we present the accuracy of various dataset distillation methods relative to the size of compressed data utilized for training the classifier network.

We observe that an advantage of storing the prompts as a distributional representation of the data, rather than the individual samples, is the flexibility to introduce variations to the data within a fixed running-memory and computation budget, with minimal overhead of only storing the random seed and patch-level soft labels or collage-level one-hot labels. Consequently, in our experiments, these new variations of the data replace the old ones without consuming additional memory, aside from potential extra soft labels. Moreover, the total number of iterations remains constant to ensure a fair comparison. Figure 6 illustrates the performance of dataset distillation methods for the three IPCs as a function of the size of compressed data used for training the classifier network. It is evident that D3M exhibits greater efficiency in compressing the datasets, particularly as IPC increases in large-scale experiments while maintaining high accuracy.

4.4. Cross-architecture analysis

In order to demonstrate the generalizability of our synthetic data across various architectures, we performed a cross-architectural analysis while fixing the $IPC = 10$ in ImageNet-1k. We first generated our collages using the learned prompts. Then, we assigned soft labels to the synthetic data using various pre-trained models on ImageNet-1k. We used the ResNet-18, MobileNet-v2, and DenseNet-121 in our experiments. In the final stage, we trained several student models from scratch on the synthetic images and their corresponding soft labels. The results are cross-examined and shown in Table 3. Moreover, in order to better understand the generalizability power of our distilled data, unlike [34], we do not perform data augmentation nor replace our data with their variations during the training stage. Table 3 shows that depending on the different teacher/student pairs, using a different model for the evaluation can be an effective strategy. For instance, going from ResNet-18 with 11.5 million parameters to DenseNet-121 with 8 million parameters can lead to performance improvement while reducing the model size. We speculate that by synthesizing images close to the

		CIFAR-10			CIFAR-100			Tiny ImageNet		
	Method	IPC=1	IPC=10	IPC=50	IPC=1	IPC=10	IPC=50	IPC=1	IPC=10	IPC=50
BiLevel Opt.	MTT	46.3(.8)	65.3(.7)	71.6(.3)	24.3(.3)	40.1(.4)	47.7(.2)	8.8(.3)	23.2(.2)	28.0(.3)
	IDM	45.6(.7)	58.6(.1)	67.5(.1)	20.1(.3)	45.1(.1)	50.0(.2)	10.1(.2)	21.9(.3)	27.7(.3)
	TESLA	48.5(.8)	66.4(.8)	72.6(.7)	24.8(.5)	41.7(.3)	47.9(.3)	-	-	-
	DATM	46.9(.5)	66.8(.3)	76.1(.3)	27.9(.2)	47.2(.4)	55.0(.2)	17.1(.3)	31.1(.3)	39.7(.3)
Hot Labels	SRe2L	14.12(.97)	20.51(.45)	32.03(.31)	3.11(.08)	7.44(.21)	13.81(.17)	1.64(.11)	7.23(.13)	13.62(.30)
	CDA	14.71(.68)	22.88(.08)	32.41(.35)	3.24(.06)	7.52(.21)	14.07(.41)	2.23(.07)	4.00 (.09)	6.03(.04)
	RDED	20.05(.14)	31.78(.12)	46.53(.18)	5.21(.10)	12.71(.06)	28.35(.69)	1.88(.02)	6.08(.11)	16.12(.14)
	Ours	28.43(.09)	37.69(.49)	47.09(.30)	9.26(.25)	22.98(.13)	29.06(.23)	4.44(.19)	10.80(.28)	14.03(.16)
Soft Labels	SRe2L	14.23(.30)	26.07(.31)	37.26(.92)	5.08(.06)	24.54(.09)	39.24(.38)	2.14(.07)	10.29(.34)	25.10(.12)
	CDA	16.21(.28)	25.64(.59)	36.24(.52)	5.10(.07)	24.15(.44)	38.99(.18)	2.56(.11)	7.66(.19)	15.98(.48)
	RDED	15.50(.66)	38.64(1.29)	56.28(.33)	10.77(.16)	35.68(.42)	47.29(.10)	4.68(.15)	21.11(.40)	37.81(.31)
	Ours	35.88(.07)	58.58(.12)	70.52(.29)	30.78(.37)	49.09(.04)	54.51(.04)	12.87(.05)	37.60(.25)	47.77(.13)

Table 2. Performance of various dataset condensation techniques on small-scale benchmarks. Standard ConvNet-3 architecture was employed for CIFAR-10 and CIFAR-100, while the ConvNet-4 architecture was used for Tiny-ImageNet, following prior literature. TESLA [7] did not report accuracies for Tiny-ImageNet, and consistent with previous studies [12, 34], its entries are left blank.

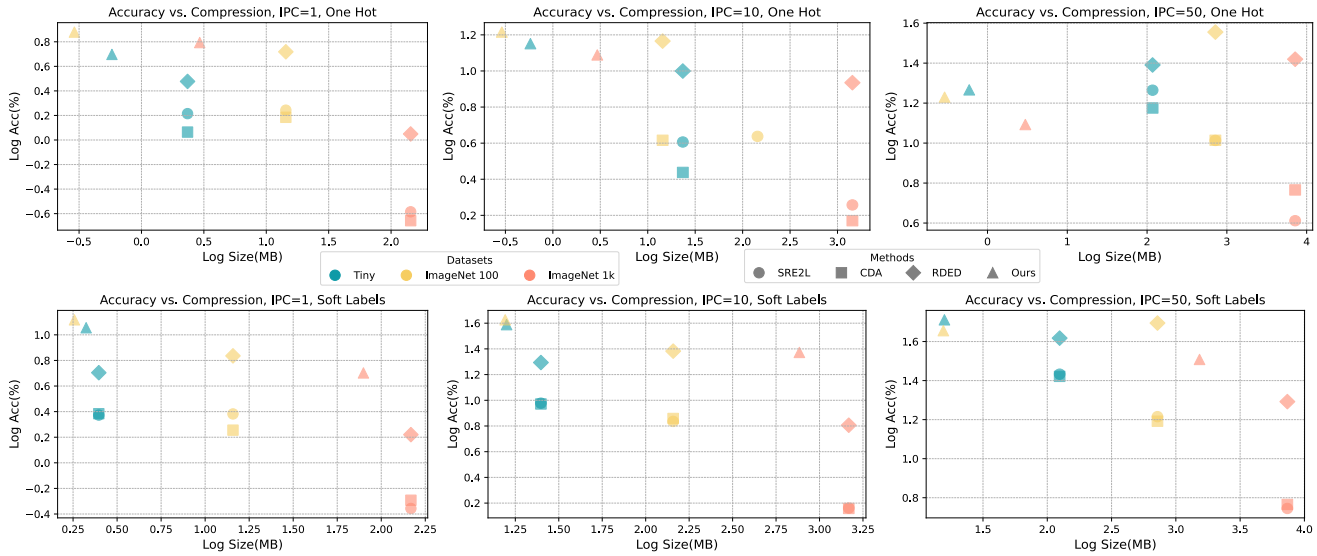

Figure 6. Log-log plot illustrating the performance of dataset distillation methods on large-scale datasets, including Tiny ImageNet, ImageNet 100, and ImageNet 1K, across IPC values of 1, 10, and 50. The performance is depicted as a function of the size of compressed data utilized for training the classifier network. Methods closer to the top left demonstrate superior performance.

		ResNet-18	MobileNet-v2	DenseNet-121
SRe2L	ResNet-18	1.45	0.84	5.60
Ours	ResNet-18 (11.5M)	12.81	9.59	20.14
	MobileNet-v2 (3.4M)	12.12	9.66	19.45
	DenseNet-121 (7.97M)	10.64	10.31	17.24

Table 3. Cross-architecture evaluation of the synthetic data across ResNet-18, MobileNet-v2, and DenseNet-121.

distribution of natural images, we can significantly reduce the architecture-specific biases in the synthesis process.

4.5. Ablation: Patch Size

In this ablation study, we studied the effect of different patch sizes in the generated collages. We experimented with 1×1, 2×2, 3×3, and 4×4 patches in $IPC = 1$ setting of ImageNet-1k. Here, we report our results in the soft-label, no augmentation and variation scenario. The results are shown in Figure 7. We can observe that for high patch sizes, performance diminishes, especially in the low-resolution setting of CIFAR-100. This is due to the fact that too much details are squeezed inside high patch sizes of a collage and we will loose the information of the important patches. Please note that although increasing the patch sizes beyond 2×2 in ImageNet-1k leads to better performance, the memory requirement for storing the soft labels will be more than twice as much.

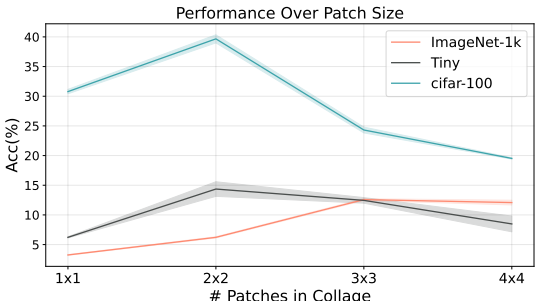

Figure 7. Accuracy over different collage patch sizes, across various datasets. The results demonstrate that there is an optimal value of patch sizes depending on the original resolution of the dataset. Nevertheless, for high patch sizes, the performance deteriorates due to the loss of information in the squeezed images.

4.6. Ablation: Steps 1 and 2

We also conducted additional ablation studies to illustrate the necessity of steps 1 and 2 in our framework. For step 1, instead of selecting the important patches, we simply resized and center-cropped the input images to construct patches. This alternative approach exhibited inferior performance compared to selecting the important patches when training the classifier. Regarding step 2, alongside the qualitative results presented in Section 3.2, we quantitatively demonstrated that merely generating training images using the class label and an engineered prompt for the diffusion model resulted in a significant drop in performance compared to textual inversion. These findings are detailed in Appendix A1 due to space constraints.

4.7. Ablation: GLaD+MTT, GLaD+DC, and GLaD+DM vs our method

GLaD uses a ConvNet to distill the samples in 128×128 resolutions and reports the average over 4 architectures (ViT, VGG11, AlexNet, ResNet18). Despite the fundamental differences in the experimental setting of GLaD and our framework, we refactored our code and repeated GLaD's experiments using our method with both a ConvNet and a ResNet teacher. Table 4 compares GLaD+MTT, GLaD+DC, and GLaD+DM with our method using different teacher architectures in IPC=1 on ImageWoof, and ImNet-Cats.

	MTT	DC	DM	Ours(Conv)	Ours(Res)
ImageWoof	17.1	17.8	15.2	**19.4**	18.6
ImNet-Cats	19.6	21.2	16.1	21.2	**23.1**

Table 4. GLaD+(MTT, DC, and DM) vs. our method.

4.8. Ablation: Generated images' class dist.

Our main claim of "one prompt per category" inherently assumes homogeneous categories. We evaluate how effectively we can generate samples from multiple sub-classes with a single prompt (e.g., 0.1 prompts per subclass!). We combined 10 classes from ImageWoof into a single category and performed textual inversion for this superclass. We generated 5000 images and used a pre-trained model (ResNet50 finetuned on ImageWoof, with 92.6%) to classify the generated images and visualize the class distributions in Figure 8. In particular, a single prompt can generate images for all subclasses, with some imbalance in the sampled data.

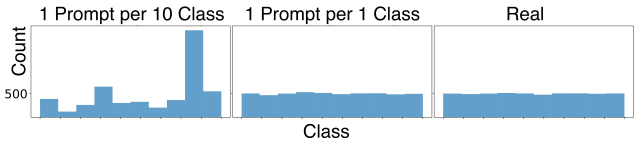

Figure 8. Class distribution of generated images.

5. Conclusion

This paper explores the compressibility of large-scale image datasets using foundational text-to-image models. We introduced Dataset Distillation using Diffusion Models (D3M), leveraging recent advances in dataset distillation. D3M combines collage images of essential patches with textual inversion and diffusion models, achieving significant compression of ImageNet-scale datasets. We show that D3M can condense an entire image category into a single textual prompt, demonstrating robust condensation capabilities. Our experiments reveal that D3M achieves high dataset compression rates while maintaining strong classifier performance, thanks to the realistic and diverse images generated by the diffusion model.

References

[1] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 1

[2] 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*, pages 4750–4759, 2022. 1, 2, 4

[3] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3739–3748, 2023. 1, 2, 4

[4] Aditya Chattopadhay, Anirban Sarkar, Prantik Howlader, and Vineeth N Balasubramanian. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pages 839–847. IEEE, 2018. 3

[5] Xuxi Chen, Yu Yang, Zhangyang Wang, and Baharan Mirzasoleiman. Data distillation can be like vodka: Distilling more times for better quality. In *The Twelfth International Conference on Learning Representations*, 2024. 1

[6] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. *Advances in Neural Information Processing Systems*, 35:810–822, 2022. 2

[7] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Scaling up dataset distillation to imagenet-1k with constant memory. In *International Conference on Machine Learning*, pages 6565–6590. PMLR, 2023. 1, 4, 5, 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 4

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3

[10] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015. 2

[11] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 1, 2, 3, 4, 5

[12] Ziyao Guo, Kai Wang, George Cazenavette, Hui Li, Kaipeng Zhang, and Yang You. Towards lossless dataset distillation via difficulty-aligned trajectory matching. *arXiv preprint arXiv:2310.05773*, 2023. 1, 2, 4, 7

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4

[14] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018. 2

[15] Jang-Hyun Kim, Jinuk Kim, Seong Joon Oh, Sangdoo Yun, Hwanjun Song, Joonhyun Jeong, Jung-Woo Ha, and Hyun Oh Song. Dataset condensation via efficient synthetic-data parameterization. In *International Conference on Machine Learning*, pages 11102–11118. PMLR, 2022. 2

[16] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 1

[17] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Toronto, ON, Canada, 2009. 4

[18] Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 4

[19] Songhua Liu, Jingwen Ye, Runpeng Yu, and Xinchao Wang. Slimmable dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3759–3768, 2023. 1

[20] Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. Dataset distillation with convexified implicit gradients. *arXiv preprint arXiv:2302.06755*, 2023. 2

[21] Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International conference on artificial intelligence and statistics*, pages 1540–1552. PMLR, 2020. 1

[22] Alaa Maalouf, Gilad Eini, Ben Mussay, Dan Feldman, and Margarita Osadchy. A unified approach to coreset learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2

[23] Dougal Maclaurin, David Duvenaud, and Ryan Adams. Gradient-based hyperparameter optimization through reversible learning. In *International conference on machine learning*, pages 2113–2122. PMLR, 2015. 1

[24] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960. PMLR, 2020. 2

[25] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. In *Advances in Neural Information Processing Systems*, 2021. 2

[26] Roman Novak, Lechao Xiao, Jiri Hron, Jaehoon Lee, Alexander A Alemi, Jascha Sohl-Dickstein, and Samuel S Schoenholz. Neural tangents: Fast and easy infinite neural networks in python. In *International Conference on Learning Representations*, 2019. 2

[27] Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in Neural Information Processing Systems*, 34:20596–20607, 2021. 2

[28] Omead Pooladzandi, David Davini, and Baharan Mirzasoleiman. Adaptive second order coresets for data-efficient machine learning. In *International Conference on Machine Learning*, pages 17848–17869. PMLR, 2022. 2

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 1, 2, 3, 5

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 4

[32] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3

[33] Shitong Shao, Zeyuan Yin, Muxin Zhou, Xindong Zhang, and Zhiqiang Shen. Generalized large-scale data condensation via various backbone and statistical matching. *arXiv preprint arXiv:2311.17950*, 2023. 1

[34] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. *arXiv preprint arXiv:2312.03526*, 2023. 1, 2, 3, 4, 5, 6, 7

[35] Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2018. 2

[36] 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*, pages 12196–12205, 2022. 2

[37] Tongzhou Wang, Jun-Yan Zhu, Antonio Torralba, and Alexei A Efros. Dataset distillation. *arXiv preprint arXiv:1811.10959*, 2018. 1, 2

[38] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8715–8724, 2020. 2

[39] Zeyuan Yin and Zhiqiang Shen. Dataset distillation in large data era. *arXiv preprint arXiv:2311.18838*, 2023. 4

[40] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. *Advances in Neural Information Processing Systems*, 36, 2024. 1, 2, 4

[41] Lei Zhang, Jie Zhang, Bowen Lei, Subhabrata Mukherjee, Xiang Pan, Bo Zhao, Caiwen Ding, Yao Li, and Dongkuan Xu. Accelerating dataset distillation via model augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11950–11959, 2023. 2

[42] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6514–6523, 2023. 2

[43] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *International Conference on Learning Representations*, 2020. 1, 2

[44] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved distribution matching for dataset condensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7856–7865, 2023. 4

[45] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3

[46] Daquan Zhou, Kai Wang, Jianyang Gu, Xiangyu Peng, Dongze Lian, Yifan Zhang, Yang You, and Jiashi Feng. Dataset quantization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17205–17216, 2023. 1

Appendix

A1. Ablation

We investigated the significance of prompt optimization in the diffusion model's performance for generating collage images across different categories. To gauge its importance, we conducted an experiment where we generated collages without fine-tuning the prompt. Specifically, we utilized a basic prompt structure: "A 4×4 natural collage of 'name_of_class' images." Moreover, acknowledging potential limitations in the diffusion model's ability to produce collage images effectively, we utilized another comparable baseline prompt structure: "A photo of ¡category¿." The results, presented in Table 5, illustrate the superiority of our optimized method over this baseline approach. This ablation study was conducted on the ImageNet-1k dataset in various IPC scenarios, with and without having access to soft labels. Please note that in order to better isolate the effect of different components of our method, we did not replace the images with their variations throughout the training. The results underscore the crucial role of prompt customization in achieving optimal outcomes for the task at hand. This is qualitatively shown in Figure 3.

In a separate ablation study, we assessed the significance of selecting important patches in constructing collage images. In this experiment, we opted to exclude the importance selection step and instead cropped the central portions of the images, resizing them to match the dimensions of a cell in the collage. The experiments were conducted on the ImageWoof dataset (a subset of ImageNet) with 10 images per class. Our findings, detailed in Table 6, indicate marginal improvement while using the process of selecting important patches. Generally, extracting the important patches will be increasingly important if the photos are of small objects, especially when a higher number of patch sizes is used in collages.

		ImageNet 1k		
	Method	IPC=1	IPC=10	IPC=50
One Hot	No Collage + No Text Inv.	0.56	1.93	3.71
	Collage + No Text Inv.	0.45	1.57	3.05
	Collage + Text Inv.	0.77	6.05	10.66
Soft Labels	No Collage + No Text Inv.	0.86	3.09	9.66
	Collage + No Text Inv.	0.96	5.73	19.03
	Collage + Text Inv.	1.38	9.47	23.69

Table 5. Ablation study of using the learned textual embeddings vs. not using them.

	With Imp Patches	Without Imp Patches
Soft Labels	26.8	24.8
One Hot Patch	9.8	9.0

Table 6. Selecting vs. not selecting the important patches, IPC=10.