Efficient Dataset Distillation via Minimax Diffusion

Jianyang Gu[1] Saeed Vahidian[2] Vyacheslav Kungurtsev[3]
Haonan Wang[4] Wei Jiang[1*] Yang You[4] Yiran Chen[2]
[1]Zhejiang University [2]Duke University [3]Czech Technical University [†]
[4]National University of Singapore
{gu_jianyang, jiangwei_zju}@zju.edu.cn

Abstract

Dataset distillation reduces the storage and computational consumption of training a network by generating a small surrogate dataset that encapsulates rich information of the original large-scale one. However, previous distillation methods heavily rely on the sample-wise iterative optimization scheme. As the images-per-class (IPC) setting or image resolution grows larger, the necessary computation will demand overwhelming time and resources. In this work, we intend to incorporate generative diffusion techniques for computing the surrogate dataset. Observing that key factors for constructing an effective surrogate dataset are representativeness and diversity, we design additional minimax criteria in the generative training to enhance these facets for the generated images of diffusion models. We present a theoretical model of the process as hierarchical diffusion control demonstrating the flexibility of the diffusion process to target these criteria without jeopardizing the faithfulness of the sample to the desired distribution. The proposed method achieves state-of-the-art validation performance while demanding much less computational resources. Under the 100-IPC setting on ImageWoof, our method requires less than one-twentieth the distillation time of previous methods, yet yields even better performance. Source code and generated data are available in https://github.com/vimar-gu/MinimaxDiffusion.

1. Introduction

Data, as a necessary resource for deep learning, has concurrently promoted algorithmic advancements while imposing challenges on researchers due to heavy demands on storage and computational resources [6, 10, 18, 47]. Confronted with the conflict between the requirement for high-precision

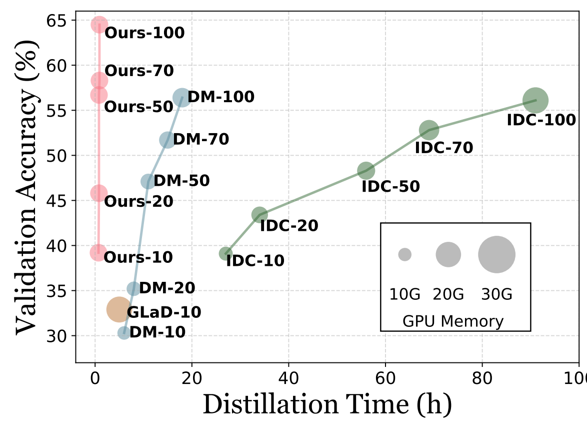

Figure 1. The validation accuracy and distillation time of different methods on ImageWoof [15], with a number following each method denoting the Image-Per-Class (IPC) setting. Previous methods are restricted by the heavier running time and memory consumption as IPC grows larger. In comparison, our proposed method notably reduces the demanding computational resources and also achieves state-of-the-art validation performance.

models and overwhelming resource demands, dataset distillation is proposed to condense the rich information of a large-scale dataset into a small surrogate one [5, 21, 47, 57]. Such a surrogate dataset is expected to achieve training performance comparable to that attained with the original one.

Previous dataset distillation methods mostly engage in iterative optimization on fixed-number samples at the pixel level [21, 26, 30, 31, 44, 54, 57] or embedding level [4, 55]. However, the sample-wise iterative optimization scheme suffers from problems of two perspectives. (1) The parameter space of optimization is positively correlated with the size of the target surrogate dataset and the image resolution [3, 57]. Consequently, substantial time and computational resources are required for distilling larger datasets. As shown in Fig. 1, IDC-1 [21] takes over 90 hours to distill a 100-image-per-class (IPC) set from ImageWoof [15], while training on ImageWoof itself only requires a matter of

*Corresponding author
†This work has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101084642.

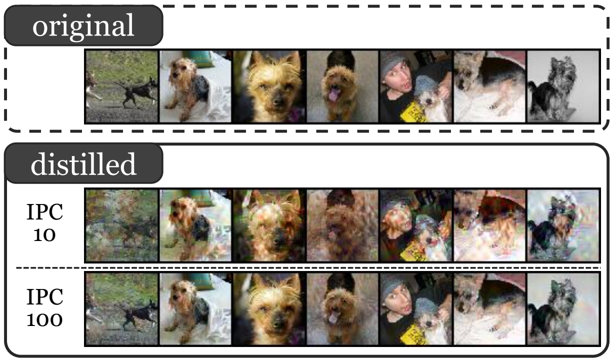

Figure 2. Sample images distilled by the pixel-level sample-wise optimization method DM [56] on ImageWoof. As the parameter space increases along with the Image-Per-Class (IPC) setting, with the same initialization, the appearance disparity between original and distilled images is smaller.

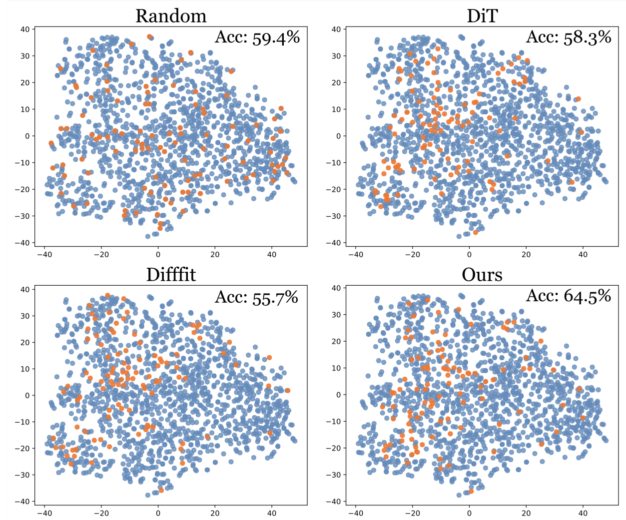

Figure 3. The feature distribution comparison of different image generation methods with the original set. The validation performance of each surrogate set is listed in the upper-right corner.

hours. (2) The expanded parameter space also increases the optimization complexity. As shown in Fig. 2, while distillation yields significant information condensation under small IPC settings, the pixel modification diminishes when distilling larger-IPC datasets. The reduced disparity also leads to smaller performance gain compared with original images, with instances where the distilled set even performs worse. Especially when distilling data of fine-grained classes, the sample-wise optimization scheme fails to provide adequate discriminative information. These constraints severely hinder individual researchers from distilling personalized data. A more practical training scheme is urgently needed to facilitate the broader application of dataset distillation.

In this work, we explore the possibility of incorporating generative diffusion techniques [20, 22, 32] to efficiently compute effective surrogate datasets. We first conduct empirical analysis on the suitability of data generated by raw diffusion models for training networks. Based on the observations, we conclude that constructing an effective surrogate dataset hinges on two key factors: representativeness and diversity. Accordingly, we design extra minimax criteria for the generative training to enhance the capability of generating more effective surrogate datasets without explicit prompt designs. The minimax criteria involve two aspects: enforcing the generated sample to be close to the farthest real sample, while being far away from the most similar generated one. We provide theoretical analysis to support that the proposed minimax scheme aims to solve a well defined problem with all the criteria, including the generative accuracy and the minimax criteria, can be targeted simultaneously without detriment to the others.

Compared with the astronomical training time consumption of the sample-wise iterative optimization schemes, the proposed method takes less than 1 hour to distill a 100-IPC surrogate dataset for a 10-class ImageNet subset, including the fine-tuning and image generation processes. Remarkably, the GPU consumption remains consistent across all IPC settings. Furthermore, the distilled surrogate dataset attains superior validation performance compared with other state-of-the-art methods. Especially on the challenging fine-grained ImageWoof subset, the proposed method outperforms the second-best DD method by 5.5% and 8.1% under the IPC settings of 70 and 100, respectively. The source code is provided in the supplementary material.

The contributions of this work are summarized into:

- We analyze the data generated by diffusion models, and emphasize the importance of representativeness and diversity for constructing effective surrogate datasets.
- We propose a novel dataset distillation scheme based on extra minimax criteria for diffusion models targeting the representativeness and diversity of generated data.
- We theoretically justify the proposed minimax criteria as enforceable without trade-offs in the generation quality of the individual data points.
- We conduct extensive experiments to validate that our proposed method achieves state-of-the-art performance while demanding significantly reduced training time in comparison to previous dataset distillation methods.

2. Method

2.1. Problem Definition

The general purpose of dataset distillation is to generate a small surrogate dataset $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_S}$ from a large-scale one $\mathcal{T} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N_T}$ [47, 57]. Here each $\mathbf{x}_i$ denotes an image with a corresponding class label y_i, and

$N_S \ll N_T$. The surrogate dataset $\mathcal{S}$ is expected to encapsulate substantial information from the original $\mathcal{T}$, such that training a model on $\mathcal{S}$ achieves performance comparable with that on $\mathcal{T}$. After distilling, we train network models on $\mathcal{S}$ and validate the performance on the original test set.

2.2. Diffusion for Distillation

Diffusion models learn a dataset distribution by gradually adding Gaussian noise to images and reversing back. Taking the latent diffusion model (LDM) as an example, given a training image $\mathbf{x}$, the training process is separated into two parts. An encoder E transforms the image into the latent space $\mathbf{z} = E(\mathbf{x})$ and a decoder D reconstructs a latent code back to the image space $\hat{\mathbf{x}} = D(\mathbf{z})$. The forward noising process gradually adds noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to the original latent code $\mathbf{z}_0$: $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t$ is a hyper-parameter. Provided with a conditioning vector $\mathbf{c}$ encoded with class labels, the diffusion models are trained by the squared error between the predicted noise $\epsilon_\theta(\mathbf{z}_t, \mathbf{c})$ and the ground truth ϵ:

$$\mathcal{L}_{simple} = ||\epsilon_\theta(\mathbf{z}_t, \mathbf{c}) - \epsilon||_2^2, \tag{1}$$

where ϵ_θ is a noise prediction network parameterized by θ. Diffusion models are proven to generate images of higher quality compared with GANs [8]. There are also some Parameter Efficient Fine-Tuning (PEFT) methods updating a small number of model parameters in order for the model to be better adapted to specific data domains [34, 50]. We adopt DiT [33] as the baseline and Difffit [50] as the naive fine-tuning method for image generation. The generated images are compared with the original data from the perspective of embedding distribution in Fig. 3.

The samples of random selection and pre-trained diffusion models present two extreme ends of the distribution. Random selection faithfully reflects the original distribution, yet fails to emphasize some high-density regions. In contrast, diffusion models are over-fitted to those dense areas, leaving a large part of the original distribution uncovered. We attribute these two distributions to two properties, respectively. The randomly selected data holds extraordinary *diversity*, and the diffusion-generated data shows *representativeness* to the original distribution. We claim that both properties are essential for constructing an effective surrogate dataset. By naive fine-tuning, Difffit better captures the representative regions but leaves more regions uncovered. To this end, we propose extra minimax criteria for the diffusion model to enhance both of the properties.

2.3. Minimax Diffusion Criteria

Based on the observation that *representativeness* and *diversity* are two key factors to construct an effective surrogate dataset, we accordingly design extra minimax criteria to enhance these two essential properties for the diffusion model.

Algorithm 1: Minimax Diffusion Fine-tuning

Input: initialized model parameter θ, original dataset $\mathcal{T} = \{(\mathbf{x}, y)\}$, encoder E, class encoder E_c, time step t, variance schedule $\bar{\alpha}_t$, real embedding memory $\mathcal{M}$, predicted embedding memory $\mathcal{D}$
Output: optimized model parameter θ^*
for *each step* **do**
- Obtain the original embedding: $\mathbf{z}_0 = E(\mathbf{x})$
- Obtain the class embedding: $\mathbf{c} = E_c(y)$
- Sample random noise: $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
- Add noise to the embedding: $\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$
- Predict the noise $\epsilon_\theta(\mathbf{z}_t, \mathbf{c})$ and recovered embedding $\hat{\mathbf{z}}_\theta(\mathbf{z}_t, \mathbf{c}) = \mathbf{z}_t - \epsilon_\theta(\mathbf{z}_t, \mathbf{c})$
- Update the model parameter with Eq. (5)
- Enqueue the real embedding: $\mathcal{M}_r \leftarrow \mathbf{z}_0$
- Enqueue the predicted embedding: $\mathcal{M}_d \leftarrow \hat{\mathbf{z}}_\theta(\mathbf{z}_t, \mathbf{c})$

end

Representativeness It is essential for the small surrogate dataset to sufficiently represent the original data. A naive approach to improve the representativeness is aligning the embedding distribution between synthetic and real samples:

$$\mathcal{L}_r = \arg\max_\theta \sigma\left(\hat{\mathbf{z}}_\theta(\mathbf{z}_t, \mathbf{c}), \frac{1}{N_B}\sum_{i=0}^{N_B}\mathbf{z}_i\right), \tag{2}$$

where $\sigma(\cdot,\cdot)$ is the cosine similarity, $\hat{\mathbf{z}}_\theta(\mathbf{z}_t, \mathbf{c})$ is the predicted original embedding by subtracting the noise from the noisy embedding $\hat{\mathbf{z}}_\theta(\mathbf{z}_t, \mathbf{c}) = \mathbf{z}_t - \epsilon_\theta(\mathbf{z}_t, \mathbf{c})$, and N_B is the size of the sampled real sample mini-batch. However, the naive alignment tends to draw the predicted embedding towards the center of the real distribution, which severely limits the diversity. Therefore, we propose to maintain an auxiliary memory $\mathcal{M} = \{\mathbf{z}_m\}_{m=1}^{N_M}$ to store the real samples utilized in adjacent iterations, and design a minimax optimization objective as:

$$\mathcal{L}_r = \arg\max_\theta \min_{m\in[N_M]} \sigma\left(\hat{\mathbf{z}}_\theta(\mathbf{z}_t, \mathbf{c}), \mathbf{z}_m\right). \tag{3}$$

By pulling close the least similar sample pairs, the diffusion model is encouraged to generate images that better cover the original distribution. It is notable that the diffusion training objective $\mathcal{L}_{simple}$ itself encourages the generated images to resemble the original ones. Thus, the minimax criterion allows the preservation of diversity to the maximum extent.

Diversity Although the pre-trained diffusion models already achieve satisfactory generation quality, the remaining defect is limited diversity compared with the original data, as shown in Fig. 3. We expect the data generated by the diffusion model can accurately reflect the original distribution, while simultaneously being different from each other. Hence, we maintain another auxiliary memory $\mathcal{D} = \{\mathbf{z}_d\}_{d=1}^{N_D}$ for the predicted embeddings of adjacent iterations and design another minimax objective to explicitly enhance the sample diversity as:

$$\mathcal{L}_d = \arg\min_{\theta} \max_{d \in [N_D]} \sigma\left(\hat{\mathbf{z}}_\theta(\mathbf{z}_t, \mathbf{c}), \mathbf{z}_d\right). \tag{4}$$

The diversity term has an opposite optimization target compared with the representativeness term, where the predicted embedding is pushed away from the most similar one stored in the memory bank. Although diversity is essential for an effective surrogate set, too much of it will cause the generated data to lose representativeness. The proposed minimax optimization enhances the diversity in a gentle way, with less influence on the class-related features.

Combining all the components, we summarize the training process in Algorithm 1. The complete training objective can be formulated as:

$$\mathcal{L} = \mathcal{L}_{simple} + \lambda_r \mathcal{L}_r + \lambda_d \mathcal{L}_d, \tag{5}$$

where λ_r and λ_d are weighting hyper-parameters.

3. Theoretical Analysis

Assume that μ is the real distribution of the latent variables $\mathbf{z}$ associated with the target dataset $\mathcal{T}$. We rewrite the optimization problem presented in Eq. (5) in a modified form:

$$\begin{aligned} \min_{\{\theta^{(i)}\}_{i\in[N_D]}} \quad & \lambda_d \max_{i,j=1,..,N_D} \sigma\left(\hat{\mathbf{z}}(\theta^{(i)}), \hat{\mathbf{z}}(\theta^{(j)})\right) \\ & + \sum_{i=1}^{N_D} \left\{ -\lambda_r Q_{\tilde{q}, w\sim\mu}\left[\sigma\left(\hat{\mathbf{z}}(\theta^{(i)}), w\right)\right] + \|\hat{\mathbf{z}}(\theta^{(i)}) - \mathbf{z}_0^{(i)}\|^2 \right\}, \end{aligned} \tag{6}$$

where $Q_{\tilde{q}}[\cdot]$ denotes the quantile function with $\tilde{q}$ as the quantile percentage. Note that here we consider a theoretical idealized variant of our algorithm wherein we perform simultaneous generation of all the embeddings $\{\hat{\mathbf{z}}(\theta^{(i)})\}$, rather than sample by sample. Hence the objectives turn to the sum of pairwise similarities rather than the form in Eq. (4). And we minimize the negative to aim for maximal representativeness, as in Eq. (5).

It can be considered as a scalarized solution to a multi-objective optimization problem, wherein multiple criteria are weighed (see, *e.g.* [13]). This perspective aligns with a Pareto front with trade-offs. It means that one objective decreasing will by necessity result in another increasing.

However, consider that any solution to the following tri-level optimization problem is also a solution for Eq. (6):

$$\begin{aligned} \min_{\{\theta^{(i)}\}_{i\in[N_D]}} \quad & \max_{i,j=1,..,N_D} \sigma\left(\hat{\mathbf{z}}(\theta^{(i)}), \hat{\mathbf{z}}(\theta^{(j)})\right) \\ \text{subj. to} \quad & \{\theta^{(i)}\} \in \arg\max \sum_{i=1}^{N_D} Q_{\tilde{q}, w\sim\mu}\left[\sigma\left(\hat{\mathbf{z}}(\theta^{(i)}), w\right)\right] \\ \text{subj. to} \quad & \theta^{(i)} \in \arg\min \|\hat{\mathbf{z}}(\theta) - \mathbf{z}_0^{(i)}\|^2, \forall i \in [N_D]. \end{aligned} \tag{7}$$

If a solution to Eq. (7) is discovered, either incidentally through solving Eq. (6) or by careful tuning of step sizes, the set of minimizers will be sufficiently large at both levels, with no trade-offs involved. However, can we justify the presumption that there exists a meaningful set of potential minimizers?

Diffusion Process Model One popular framework for the mathematical analysis of diffusion involves analyzing the convergence and asymptotic properties of, appropriately homonymous, diffusion processes. These processes are characterized by the standard stochastic differential equation with a drift and diffusion term. For a time-dependent random variable Z_t,

$$dZ_t = V(Z_t)dt + dW_t \tag{8}$$

where V is a drift function dependent on the current Z_t and dW_t is a Wiener (Brownian noise) process. This equation serves as an appropriate continuous approximation of generative diffusion, given that Brownian noise is a continuous limit of adding normal random variables. Consequently, we aim for any realization $\mathbf{z} \sim Z_t$ to have certain desired properties that reflect generative modeling with high probability.

The work [43] established a theoretical model utilizing concepts in the control of these diffusions, demonstrating how it can result in sampling from the distribution of a desired data set. In the supplementary material we present a description of their framework and present an argument supporting the well-defined nature of the following problem, indicating that it has non-trivial solutions.

When sampling optimally from the population dataset, we consider a stochastic control problem wherein V depends also on some chosen control $u(\mathbf{z}, t)$. This control aims to find the most representative samples and, among the possible collection of such samples, to obtain the most diverse one while sampling from the desired dataset μ. This involves solving:

$$\begin{aligned} \min_{u(x,t)} \quad & \max_{i,j=1,..,N_D} \sigma\left(Z_1^{u,(i)}, Z_1^{u,(j)}\right) \\ \text{subj. to} \quad & u \in \arg\max \sum_{i=1}^{N_D} \int_0^1 \mathbb{E}_{Z_t^{(i)}} Q_{\tilde{q}, w\sim\mu}\left[\sigma\left(Z_t^{(i)}, w\right)\right] ds \\ & Z_1 \sim \mu, \\ & dZ_t^{u,(i)} = u(Z_t^{u,(i)}, t)dt + dW_t, \; t \in [0,1]; \\ & Z_0 = \mathbf{z}_0. \end{aligned} \tag{9}$$

This problem poses a bi-level stochastic control challenge where employing a layered dynamic programming is far from tractable. Additionally, a multi-stage stochastic programming approximation would also be infeasible given the scale of the datasets involved. Instead, we opt for parameterization with a neural network, forego exact sampling, discretize the problem, redefine the criteria to be time independent, and seek to solve an approximate solution for the tri-level optimization problem Eq. (7).

Table 1. Performance comparison with pre-trained diffusion models and other state-of-the-art methods on ImageWoof. All the results are reproduced by us on the 256×256 resolution. The missing results are due to out-of-memory. The best results are marked as **bold**.

IPC (Ratio)	Test Model	Random	K-Center [37]	Herding [48]	DiT [33]	DM [56]	IDC-1 [21]	GLaD [4]	Ours	Full
10 (0.8%)	ConvNet-6	$24.3_{\pm 1.1}$	$19.4_{\pm 0.9}$	$26.7_{\pm 0.5}$	$34.2_{\pm 1.1}$	$26.9_{\pm 1.2}$	$33.3_{\pm 1.1}$	$33.8_{\pm 0.9}$	$\mathbf{37.0}_{\pm 1.0}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$29.4_{\pm 0.8}$	$22.1_{\pm 0.1}$	$32.0_{\pm 0.3}$	$34.7_{\pm 0.5}$	$30.3_{\pm 1.2}$	$39.1_{\pm 0.5}$	$32.9_{\pm 0.9}$	$\mathbf{39.2}_{\pm 1.3}$	$87.5_{\pm 0.5}$
	ResNet-18	$27.7_{\pm 0.9}$	$21.1_{\pm 0.4}$	$30.2_{\pm 1.2}$	$34.7_{\pm 0.4}$	$33.4_{\pm 0.7}$	$37.3_{\pm 0.2}$	$31.7_{\pm 0.8}$	$\mathbf{37.6}_{\pm 0.9}$	$89.3_{\pm 1.2}$
20 (1.6%)	ConvNet-6	$29.1_{\pm 0.7}$	$21.5_{\pm 0.8}$	$29.5_{\pm 0.3}$	$36.1_{\pm 0.8}$	$29.9_{\pm 1.0}$	$35.5_{\pm 0.8}$	-	$\mathbf{37.6}_{\pm 0.2}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$32.7_{\pm 0.4}$	$25.1_{\pm 0.7}$	$34.9_{\pm 0.1}$	$41.1_{\pm 0.8}$	$35.2_{\pm 0.6}$	$43.4_{\pm 0.3}$	-	$\mathbf{45.8}_{\pm 0.5}$	$87.5_{\pm 0.5}$
	ResNet-18	$29.7_{\pm 0.5}$	$23.6_{\pm 0.3}$	$32.2_{\pm 0.6}$	$40.5_{\pm 0.5}$	$29.8_{\pm 1.7}$	$38.6_{\pm 0.2}$	-	$\mathbf{42.5}_{\pm 0.6}$	$89.3_{\pm 1.2}$
50 (3.8%)	ConvNet-6	$41.3_{\pm 0.6}$	$36.5_{\pm 1.0}$	$40.3_{\pm 0.7}$	$46.5_{\pm 0.8}$	$44.4_{\pm 1.0}$	$43.9_{\pm 1.2}$	-	$\mathbf{53.9}_{\pm 0.6}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$47.2_{\pm 1.3}$	$40.6_{\pm 0.4}$	$49.1_{\pm 0.7}$	$49.3_{\pm 0.2}$	$47.1_{\pm 1.1}$	$48.3_{\pm 1.0}$	-	$\mathbf{56.3}_{\pm 1.0}$	$87.5_{\pm 0.5}$
	ResNet-18	$47.9_{\pm 1.8}$	$39.6_{\pm 1.0}$	$48.3_{\pm 1.2}$	$50.1_{\pm 0.5}$	$46.2_{\pm 0.6}$	$48.3_{\pm 0.8}$	-	$\mathbf{57.1}_{\pm 0.6}$	$89.3_{\pm 1.2}$
70 (5.4%)	ConvNet-6	$46.3_{\pm 0.6}$	$38.6_{\pm 0.7}$	$46.2_{\pm 0.6}$	$50.1_{\pm 1.2}$	$47.5_{\pm 0.8}$	$48.9_{\pm 0.7}$	-	$\mathbf{55.7}_{\pm 0.9}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$50.8_{\pm 0.6}$	$45.9_{\pm 1.5}$	$53.4_{\pm 1.4}$	$54.3_{\pm 0.9}$	$51.7_{\pm 0.8}$	$52.8_{\pm 1.8}$	-	$\mathbf{58.3}_{\pm 0.2}$	$87.5_{\pm 0.5}$
	ResNet-18	$52.1_{\pm 1.0}$	$44.6_{\pm 1.1}$	$49.7_{\pm 0.8}$	$51.5_{\pm 1.0}$	$51.9_{\pm 0.8}$	$51.1_{\pm 1.7}$	-	$\mathbf{58.8}_{\pm 0.7}$	$89.3_{\pm 1.2}$
100 (7.7%)	ConvNet-6	$52.2_{\pm 0.4}$	$45.1_{\pm 0.5}$	$54.4_{\pm 1.1}$	$53.4_{\pm 0.3}$	$55.0_{\pm 1.3}$	$53.2_{\pm 0.9}$	-	$\mathbf{61.1}_{\pm 0.7}$	$86.4_{\pm 0.2}$
	ResNetAP-10	$59.4_{\pm 1.0}$	$54.8_{\pm 0.2}$	$61.7_{\pm 0.9}$	$58.3_{\pm 0.8}$	$56.4_{\pm 0.8}$	$56.1_{\pm 0.9}$	-	$\mathbf{64.5}_{\pm 0.2}$	$87.5_{\pm 0.5}$
	ResNet-18	$61.5_{\pm 1.3}$	$50.4_{\pm 0.4}$	$59.3_{\pm 0.7}$	$58.9_{\pm 1.3}$	$60.2_{\pm 1.0}$	$58.3_{\pm 1.2}$	-	$\mathbf{65.7}_{\pm 0.4}$	$89.3_{\pm 1.2}$

In the supplementary material we provide a rationale for the meaningfulness of the problem in Eq. (9) based on the model of generative diffusion [43]. Specifically, we argue that the set of controls that leads to the desired final distribution and the set of minimizers, is sufficiently large for a low value of the objective at the top layer.

Table 2. The Maximum Mean Discrepancy (MMD) between the extracted features of surrogate dataset and the original one.

IPC	DiT [33]	Difffit [50]	DM [56]	IDC-1 [21]	Ours
50	5.4	5.4	4.8	6.7	4.0
100	5.5	5.3	4.0	6.4	4.3

4. Experiments

4.1. Implementation Details

For the diffusion model, we adopt pre-trained DiT [33] as the baseline and conduct PEFT with Difffit [50]. λ_r and λ_d are set as 0.002 and 0.008 for Eq. (5), respectively. The image size for the diffusion fine-tuning and sample generation is set as 256×256. The fine-tuning mini-batch size is set as 8, and the fine-tuning lasts 8 epochs. The learning rate is set as 1e-3 for an AdamW optimizer. After fine-tuning, the images are generated by 50 denoising steps on a pre-defined number of random noise, according to the IPC setting. All the experiments are conducted on a single RTX 4090 GPU.

4.2. Datasets and Evaluation Metric

For practical applicability, the experiments are exclusively conducted on full-sized ImageNet [6] subsets in this work. The selected subsets include ImageWoof, ImageNette [15] and the 10-class split adopted in [21, 42], denoted as ImageIDC afterward. ImageWoof is a challenging subset, containing only classes of dog breeds, while ImageNette and ImageIDC contain classes with less similarity, and hence are easier to discriminate. For evaluation, we adopt the same setting as in [21]. The surrogate dataset is trained on different model architectures, with a learning rate of 0.01, and a scheduler decaying the learning rate at 2/3 and 5/6 of the whole training iterations. The top-1 accuracy on the original testing set is reported to illustrate the performance.

4.3. Comparison with State-of-the-art Methods

We compare our method with other state-of-the-art methods across different IPC settings and model architectures. For a fair comparison, the results are all reproduced by us under the same evaluation protocol. ResNet-10 [18] with average pooling is adopted for matching the feature distribution (DM [56], GLaD [4]) and training gradients (IDC-1 [21]). DM is implemented on IDC-1 by only modifying the matching objective from training gradients to feature distribution, such that better performance is achieved. Each experiment is conducted 3 times, with the mean value and standard variance reported. Firstly, we present the validation results on the challenging ImageWoof subset [15] in Tab. 1.

With the target of distilling surrogate datasets of small IPCs (*e.g.*, 10 and 20), the pixel-level optimization method IDC-1 [21] demonstrates outstanding performance gain over random original images. However, as the IPC increases, the performance gain drastically drops. Especially under the 100-IPC setting, the distilled dataset even performs worse than random original images. This observation

Table 3. Performance comparison with pre-trained diffusion models and state-of-the-art methods on more ImageNet subsets. The results are obtained on ResNet-10 with average pooling. The best results are marked as **bold**.

	IPC	Random	DiT [33]	DM [56]	Ours
Nette	10	$54.2_{\pm 1.6}$	$59.1_{\pm 0.7}$	$60.8_{\pm 0.6}$	$\mathbf{62.0}_{\pm 0.2}$
	20	$63.5_{\pm 0.5}$	$64.8_{\pm 1.2}$	$66.5_{\pm 1.1}$	$\mathbf{66.8}_{\pm 0.4}$
	50	$76.1_{\pm 1.1}$	$73.3_{\pm 0.9}$	$76.2_{\pm 0.4}$	$\mathbf{76.6}_{\pm 0.2}$
IDC	10	$48.1_{\pm 0.8}$	$\mathbf{54.1}_{\pm 0.4}$	$52.8_{\pm 0.5}$	$53.1_{\pm 0.2}$
	20	$52.5_{\pm 0.9}$	$58.9_{\pm 0.2}$	$58.5_{\pm 0.4}$	$\mathbf{59.0}_{\pm 0.4}$
	50	$68.1_{\pm 0.7}$	$64.3_{\pm 0.6}$	$69.1_{\pm 0.8}$	$\mathbf{69.6}_{\pm 0.2}$

aligns with the empirical findings in Fig. 2, where pixel-level methods struggle to optimize the expanded parameter space of large IPCs. The embedding-level optimization method GLaD [4] yields good performance under the 10-IPC setting. However, it requires overwhelming GPU resources for larger IPC settings, which is inapplicable for resource-restricted scenarios. It is also notable that on large IPCs, the coreset method Herding [48] surpasses previous DD methods with far less computational cost.

The pre-trained DiT [33] here serves as the baseline for generative diffusion techniques. Under the 50-IPC setting, DiT outperforms both random original images and IDC-1. However, the insufficiency of representativeness and diversity restricts its performance on smaller and larger IPCs, respectively. In contrast, our proposed minimax diffusion consistently provides superior performance across all the IPCs over both original images and Herding. Besides, the proposed method eliminates the need of specific network architectures for matching training metrics. Consequently, the cross-architecture generalization is significantly improved. Under most IPC settings, the performance gap between ConvNet-6 and ResNetAP-10 is even smaller than that of the original images. It validates the universality of the rich information learned by the minimax fine-tuning process.

Furthermore, we extensively assess the Maximum Mean Discrepancy (MMD) between the embedded features of the selected/generated surrogate dataset and the original one in Tab. 2. The features are extracted by a ResNet-10 network pre-trained on the full original dataset. Our method achieves the lowest discrepancy by average, where DM [56] directly sets MMD as the optimization target, proving the validity of extra minimax criteria in fitting distributions.

Moreover, we show the performance comparison on ImageNette [15] and ImageIDC [21] in Tab. 3. The performance trend generally aligns with that on ImageWoof. More specifically, on these two easier subsets, DiT quickly loses the advantage over original images as IPC increases. Conversely, our proposed minimax diffusion method consistently demonstrates state-of-the-art performance.

Table 4. Performance comparison on ImageNet-1K.

IPC	SRe^{2}L [53]	RDED [41]	DiT	Ours
10	$21.3_{\pm 0.6}$	$42.0_{\pm 0.1}$	$39.6_{\pm 0.4}$	$\mathbf{44.3}_{\pm 0.5}$
50	$46.8_{\pm 0.2}$	$56.5_{\pm 0.1}$	$52.9_{\pm 0.6}$	$\mathbf{58.6}_{\pm 0.3}$

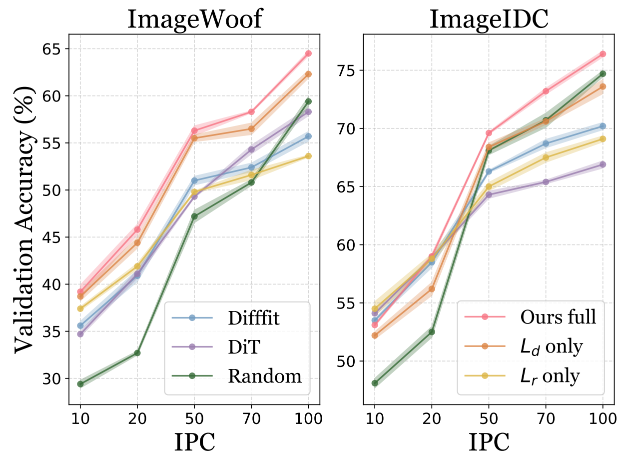

Figure 4. With the help of the minimax diffusion, the proposed method significantly enhances the representativeness and diversity of the generated images. Thereby it consistently provides superior performance compared with random selection and baseline diffusion models by a large margin across different IPC settings.

Experiments on ImageNet-1K. We further conduct experiments on the full ImageNet-1K with the validation protocol of RDED [41] and present the results in Tab. 4. The synthetic images are resized to 224×224 for evaluation. The significant performance advantage over the compared works validates the scalability of the proposed method.

4.4. Ablation Study

Component Analysis. We compare the performance with baseline diffusion models to validate the effectiveness of proposed minimax criteria in Fig. 4. The experiments are conducted on ImageWoof and ImageIDC to evaluate the effect on challenging and easy tasks, respectively. Under the IPC of 10 and 20, the raw diffusion models (DiT) generate informative images, with validation performance much higher than randomly selected original samples. However, as the IPC is continuously increased, the performance gap diminishes for ImageWoof, and random original images surpass the DiT-generated ones at the IPC of 100. On ImageIDC the intersection occurs even earlier at the IPC of 50. The main reason is reflected in Fig. 3, where the sample diversity remains limited without external guidance. The naive Difffit fine-tuning adapts the model to specific domains, yet on large IPCs, the over-fitted generative model still yields inferior performance than the original images.

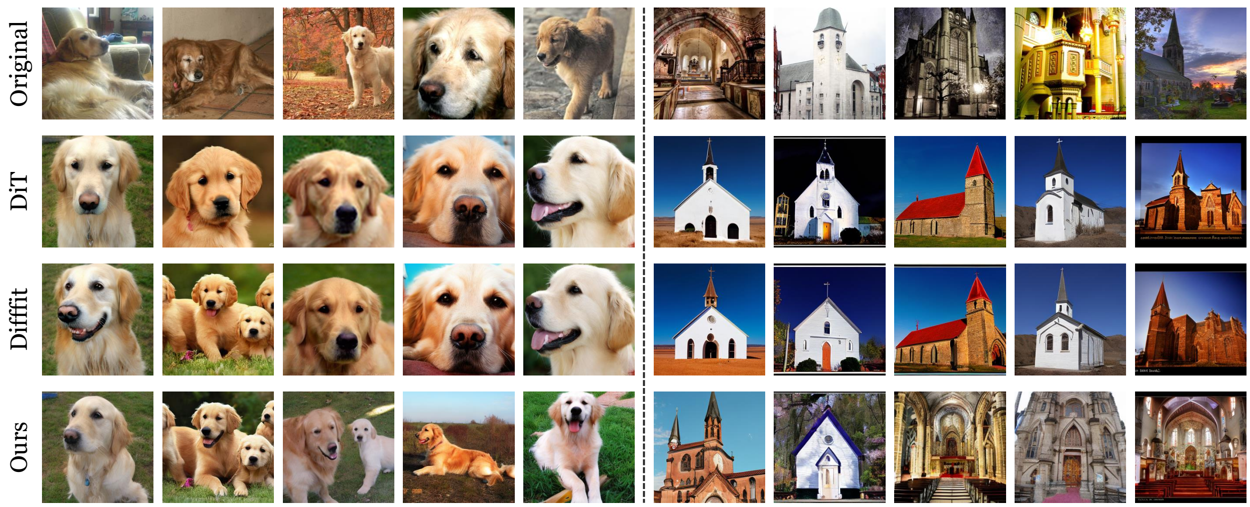

Figure 5. Visualization of random original images, images generated by baseline diffusion models (DiT [33] and Difffit [50]) and our proposed method. For each column, the generated images are based on the same random seed. Comparatively, our method significantly enhances the coverage of original data distribution and the diversity of the surrogate dataset.

Table 5. The ablation study of the proposed minimax scheme. The result are obtained with ResNet-10 on ImageWoof and ImageIDC. $\mathbf{m}$ denotes the proposed minimax optimization scheme.

$\mathcal{L}_r$	$\mathcal{L}_r$ w\m	$\mathcal{L}_d$	$\mathcal{L}_d$ w\m	ImageWoof 10-IPC	ImageWoof 50-IPC	ImageIDC 10-IPC	ImageIDC 50-IPC
-	-	-	-	$35.6_{\pm 0.9}$	$51.0_{\pm 0.9}$	$53.5_{\pm 0.2}$	$66.3_{\pm 0.2}$
✓	-	-	-	$34.4_{\pm 1.1}$	$47.1_{\pm 0.5}$	$49.6_{\pm 0.7}$	$60.2_{\pm 1.2}$
-	✓	-	-	$37.4_{\pm 0.4}$	$49.5_{\pm 1.0}$	$54.5_{\pm 1.2}$	$65.0_{\pm 0.8}$
-	-	✓	-	$35.7_{\pm 0.8}$	$48.3_{\pm 0.6}$	$51.5_{\pm 0.6}$	$64.8_{\pm 0.8}$
-	-	-	✓	$38.7_{\pm 0.9}$	$54.9_{\pm 0.7}$	$52.2_{\pm 0.6}$	$68.4_{\pm 0.7}$
✓	-	✓	-	$38.3_{\pm 0.5}$	$54.9_{\pm 0.4}$	$53.3_{\pm 0.5}$	$66.8_{\pm 0.5}$
-	✓	-	✓	$39.2_{\pm 1.3}$	$56.3_{\pm 1.0}$	$53.1_{\pm 0.2}$	$69.6_{\pm 0.2}$

The addition of representativeness constraint to the training process further enhances the effect of distribution fitting. At small IPCs, the generated images contain richer information, yet for larger IPCs, the lack of diversity brings a negative influence. The diversity constraint, in contrast, significantly boosts the information contained in the generated surrogate dataset. Despite the performance advantage of $\mathcal{L}_d$ over $\mathcal{L}_r$, combining them still brings stable improvement as our full method. Especially on the easier ImageIDC task, grouping these two constraints together contributes to a consistent performance margin over random original images. The experimental results validate that both representativeness and diversity play essential parts in constructing effective surrogate datasets.

Minimax Scheme. In this work, we propose to enhance the representativeness and diversity each with a minimax objective. We compare the distillation result with or without the minimax operation in Tab. 5. The first row presents the performance of naive Difffit fine-tuning. Matching the embeddings to the distribution center as in Eq. (2) severely degrades the validation performance across all IPCs. In contrast, the minimax version constraint as in Eq. (3) encourages better coverage, where the performance on small IPCs is improved. The effects of diversity constraint and the full method show similar trends. The superior performance suggests the effectiveness in enhancing the essential properties of the generative diffusion techniques.

4.5. Visualization

Sample Distribution Visualization. The target of our proposed method is to construct a surrogate dataset with both representativeness and diversity. We visualize the t-SNE distribution of samples generated by our proposed method in Fig. 3. In comparison with random original images and baseline diffusion models, our method demonstrates a more thorough coverage over the entire data distribution while maintaining consistency in sample density. At the original high-density region, the generated images also form a dense sub-cluster, which is not reflected by random sampling. On the other hand, at the original sparse regions, our method exhibits better coverage than baseline diffusion models. By simultaneously enhancing the representativeness and diversity in the generative model, the proposed method manages to significantly improve the validation performance of the generated surrogate dataset.

Generated Sample Comparison. The proposed method notably enhances the representativeness and diversity of the generated surrogate dataset. We compare the samples generated with the same random noise (for each column) of

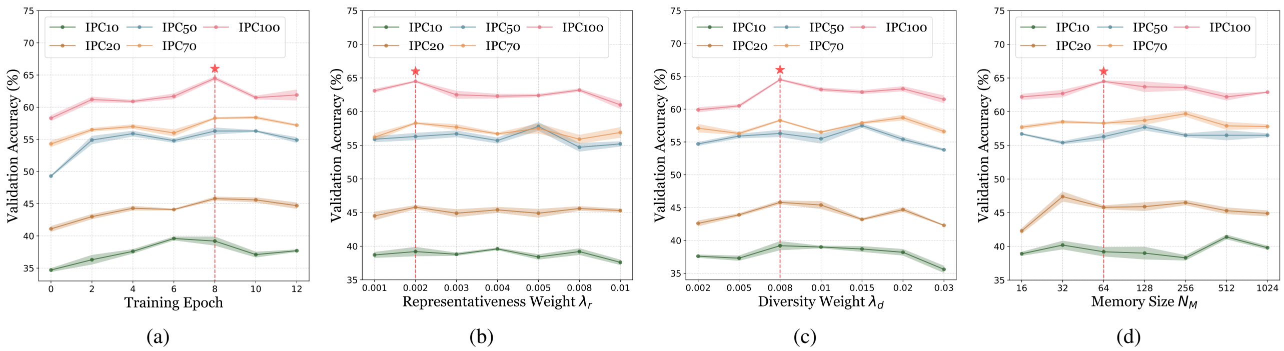

Figure 6. Hyper-parameter analysis on (a) the training epochs; (b) the representativeness weight λ_r; (c) the diversity weight λ_d; (d) the memory size N_M. The results are obtained with ResNetAP-10 on ImageWoof. The dashed line indicates the value adopted in this work.

different generative methods in Fig. 5 to explicitly demonstrate the improved properties.

The images generated by baseline DiT exhibit a realistic high-quality appearance. However, the images tend to share similar poses and only present the most prominent features of the objects. In the golden retriever case, the generated images mostly present the head part, while for the churches the exterior appearance. Difffit fine-tuning further fits the model to the distribution, but in most cases, the differences only lie in small details. Comparatively, the proposed minimax criteria significantly enhance both the representativeness and diversity of the generated images. On the one hand, there occurs more class-related content in the generated images. The golden retriever images include more body parts and the church images encompass the interior layout. The minimax optimization leads to better coverage over the entire original distribution, with more related features encapsulated. On the other hand, the diversity is significantly enhanced, including variations in pose, background, and appearance styles. In such a way the surrogate dataset better represents the original large-scale one, leading to superior validation performance.

Training Curve Visualization. We visualize the accuracy curve during the training process in Fig. 6a. The validation performance is rapidly improved as the fine-tuning process starts. After four epochs, the model tends to converge and reaches the highest performance at the 8th epoch. Further extending the training epochs injects excessive diversity into the model, leading to performance degradation. We demonstrate the influence of training epochs on the generated images in supplementary material.

4.6. Parameter Analysis

Objective Weight λ_r λ_d. We show the influence of representativeness weight λ_r and diversity weight λ_d in Fig. 6b and Fig. 6c, respectively. The λ_r variation only produces negligible performance fluctuation on small IPCs, while on large IPCs the performance is also relatively stable. For λ_d, at a proper variation range, the performance is stable. However, continuously increasing the diversity of the generated dataset leads to a lack of representativeness, which results in a negative impact. The negative impact of over-diversity can also validated by the poor performance of K-Center in Tab. 1. A uniform performance decrease is observed as λ_d reaches 0.03. Based on the performance of 100 IPC, we set λ_r as 0.002 and λ_d as 0.008.

Memory Size N_M. The memory size N_M influences the number of samples involved in the objective calculation. We investigate its influence in Fig. 6d. When the memory is extremely small (N_M=16), the provided supervision is also limited, yet the performance is already higher than naive fine-tuning. As the memory size is increased in a proper range, the model yields stable performance improvement. It is notable that with a larger memory, the performance under the IPC of 10 is better. It can be explained by that a larger memory contains more representative information. Out of the consideration of performance as well as storage burden, we select the memory of 64 in the other experiments.

5. Conclusion

In this work, we propose a novel dataset distillation method based on generative diffusion techniques. Through extra minimax criteria, the proposed method significantly enhances the representativeness and diversity of the generated surrogate dataset. With much less computational time consumption, the proposed method achieves state-of-the-art validation performance on challenging ImageNet subsets. It reduces the resource dependency of previous dataset distillation methods and opens up new possibilities for more practical applications for distilling personalized data.

Limitations and Future Works. In addition to object classification, we will explore the possibility of incorporating generative techniques for more specific data domains.

Acknowledgement. We would like to specially thank Olga Russakovsky for invaluable advice on the manuscript.

References

[1] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*, 2023. 2

[2] Victor Besnier, Himalaya Jain, Andrei Bursuc, Matthieu Cord, and Patrick Pérez. This Dataset Does Not Exist: Training Models from Generated Images. In *ICASSP*, pages 1–5, 2020. 2

[3] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Dataset distillation by matching training trajectories. In *CVPR*, pages 4750–4759, 2022. 1

[4] George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A Efros, and Jun-Yan Zhu. Generalizing dataset distillation via deep generative prior. In *CVPR*, pages 3739–3748, 2023. 1, 5, 6

[5] Justin Cui, Ruochen Wang, Si Si, and Cho-Jui Hsieh. Dc-bench: Dataset condensation benchmark. *NeurIPS*, 35:810–822, 2022. 1

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1, 5

[7] Zhiwei Deng and Olga Russakovsky. Remember the Past: Distilling Datasets into Addressable Memories for Neural Networks. In *NeurIPS*, 2022. 1

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, pages 8780–8794, 2021. 3, 2

[9] Thinh T Doan. Nonlinear two-time-scale stochastic approximation convergence and finite-time performance. *IEEE Transactions on Automatic Control*, 2022. 1

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2022. 1

[11] Jiawei Du, Yidi Jiang, Vincent Y.F. Tan, Joey Tianyi Zhou, and Haizhou Li. Minimizing the Accumulated Trajectory Error to Improve Dataset Distillation. In *CVPR*, pages 3749–3758, 2023. 1

[12] Lisa Dunlap, Alyssa Umino, Han Zhang, Jiezhi Yang, Joseph E Gonzalez, and Trevor Darrell. Diversify your vision datasets with automatic diffusion-based augmentation. *arXiv preprint arXiv:2305.16289*, 2023. 2

[13] Gabriele Eichfelder. Scalarizations for adaptively solving multi-objective optimization problems. *Computational Optimization and Applications*, 44:249–273, 2009. 4

[14] Ronen Eldan and James R Lee. Regularization under diffusion and anticoncentration of the information content. *Duke Mathematical Journal*, 167(5):969–993, 2018. 1

[15] Fastai. Fastai/imagenette: A smaller subset of 10 easily classified classes from imagenet, and a little more french. 1, 5, 6

[16] Jianyang Gu, Kai Wang, Wei Jiang, and Yang You. Summarizing stream data for memory-restricted online continual learning. *arXiv preprint arXiv:2305.16645*, 2023. 1

[17] Swaminathan Gurumurthy, Ravi Kiran Sarvadevabhatla, and R. Venkatesh Babu. DeLiGAN: Generative Adversarial Networks for Diverse and Limited Data. In *CVPR*, pages 4941–4949, 2017. 2

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1, 5, 2

[19] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is Synthetic Data from Generative Models Ready for Image Recognition? In *ICLR*, 2022. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, pages 6840–6851, 2020. 2

[21] 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 *ICML*, pages 11102–11118, 2022. 1, 5, 6, 2, 3

[22] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational Diffusion Models. In *NeurIPS*, pages 21696–21707, 2021. 2

[23] Saehyung Lee, Sanghyuk Chun, Sangwon Jung, Sangdoo Yun, and Sungroh Yoon. Dataset condensation with contrastive signals. In *ICML*, pages 12352–12364, 2022. 1

[24] Joseph Lehec. Representation formula for the entropy and functional inequalities. In *Annales de l'IHP Probabilités et statistiques*, pages 885–899, 2013. 1

[25] Ping Liu, Xin Yu, and Joey Tianyi Zhou. Meta Knowledge Condensation for Federated Learning. In *ICLR*, 2022. 1

[26] Yanqing Liu, Jianyang Gu, Kai Wang, Zheng Zhu, Wei Jiang, and Yang You. Dream: Efficient dataset distillation by representative matching. In *ICCV*, 2023. 1, 2

[27] Noel Loo, Ramin Hasani, Alexander Amini, and Daniela Rus. Efficient dataset distillation using random feature approximation. *NeurIPS*, 35:13877–13891, 2022. 1

[28] Noel Loo, Ramin Hasani, Mathias Lechner, and Daniela Rus. Dataset distillation with convexified implicit gradients. *arXiv preprint arXiv:2302.06755*, 2023. 1

[29] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode Seeking Generative Adversarial Networks for Diverse Image Synthesis. In *CVPR*, pages 1429–1437, 2019. 2

[30] Timothy Nguyen, Zhourong Chen, and Jaehoon Lee. Dataset meta-learning from kernel ridge-regression. In *ICLR*, 2021. 1

[31] Timothy Nguyen, Roman Novak, Lechao Xiao, and Jaehoon Lee. Dataset distillation with infinitely wide convolutional networks. *NeurIPS*, 34:5186–5198, 2021. 1

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models. In *ICML*, pages 8162–8171, 2021. 2

[33] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 3, 5, 6, 7

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 3

[35] Ahmad Sajedi, Samir Khaki, Ehsan Amjadian, Lucy Z Liu, Yuri A Lawryshyn, and Konstantinos N Plataniotis. DataDAM: Efficient Dataset Distillation with Attention Matching. In *ICCV*, 2023. 1

[36] Mert Bulent Sariyildiz, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it Till You Make it: Learning Transferable Representations from Synthetic ImageNet Clones. In *CVPR*, pages 8011–8021, 2023. 2

[37] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. In *International Conference on Learning Representations*, 2018. 5

[38] Allahkaram Shafiei, Vyacheslav Kungurtsev, and Jakub Marecek. Trilevel and multilevel optimization using monotone operator theory. *arXiv preprint arXiv:2105.09407*, 2021. 1

[39] Jordan Shipard, Arnold Wiliem, Kien Nguyen Thanh, Wei Xiang, and Clinton Fookes. Diversity Is Definitely Needed: Improving Model-Agnostic Zero-Shot Classification via Stable Diffusion. In *CVPR*, pages 769–778, 2023. 2

[40] Felipe Petroski Such, Aditya Rawal, Joel Lehman, Kenneth Stanley, and Jeffrey Clune. Generative Teaching Networks: Accelerating Neural Architecture Search by Learning to Generate Synthetic Training Data. In *ICML*, pages 9206–9216, 2020. 1

[41] Peng Sun, Bei Shi, Daiwei Yu, and Tao Lin. On the diversity and realism of distilled dataset: An efficient dataset distillation paradigm. In *CVPR*, 2024. 6

[42] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, pages 776–794, 2020. 5

[43] Belinda Tzen and Maxim Raginsky. Theoretical guarantees for sampling and inference in generative models with latent diffusions. In *COLT*, pages 3084–3114, 2019. 4, 5, 1

[44] Saeed Vahidian, Mingyu Wang, Jianyang Gu, Vyacheslav Kungurtsev, Wei Jiang, and Yiran Chen. Group distributionally robust dataset distillation with risk minimization. *arXiv preprint arXiv:2402.04676*, 2024. 1

[45] 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 *CVPR*, pages 12196–12205, 2022. 1

[46] Kai Wang, Jianyang Gu, Daquan Zhou, Zheng Zhu, Wei Jiang, and Yang You. Dim: Distilling dataset into generative model. *arXiv preprint arXiv:2303.04707*, 2023. 1

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

[48] Max Welling. Herding dynamical weights to learn. In *ICML*, pages 1121–1128, 2009. 5, 6, 3

[49] Xindi Wu, Zhiwei Deng, and Olga Russakovsky. Multimodal dataset distillation for image-text retrieval. *arXiv preprint arXiv:2308.07545*, 2023. 1

[50] Enze Xie, Lewei Yao, Han Shi, Zhili Liu, Daquan Zhou, Zhaoqiang Liu, Jiawei Li, and Zhenguo Li. Difffit: Unlocking transferability of large diffusion models via simple parameter-efficient fine-tuning. *arXiv preprint arXiv:2304.06648*, 2023. 3, 5, 7, 2

[51] Yuanhao Xiong, Ruochen Wang, Minhao Cheng, Felix Yu, and Cho-Jui Hsieh. FedDM: Iterative Distribution Matching for Communication-Efficient Federated Learning. In *CVPR*, pages 16323–16332, 2023. 1

[52] Dingdong Yang, Seunghoon Hong, Yunseok Jang, Tianchen Zhao, and Honglak Lee. Diversity-Sensitive Conditional Generative Adversarial Networks. In *ICLR*, 2018. 2

[53] Zeyuan Yin, Eric Xing, and Zhiqiang Shen. Squeeze, recover and relabel: Dataset condensation at imagenet scale from a new perspective. In *NeurIPS*, 2023. 6

[54] Bo Zhao and Hakan Bilen. Dataset condensation with differentiable siamese augmentation. In *ICML*, pages 12674–12685, 2021. 1, 2

[55] Bo Zhao and Hakan Bilen. Synthesizing informative training samples with gan. *arXiv preprint arXiv:2204.07513*, 2022. 1

[56] Bo Zhao and Hakan Bilen. Dataset condensation with distribution matching. In *WACV*, pages 6514–6523, 2023. 2, 5, 6, 1

[57] Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. Dataset condensation with gradient matching. In *ICLR*, 2021. 1, 2

[58] Ganlong Zhao, Guanbin Li, Yipeng Qin, and Yizhou Yu. Improved Distribution Matching for Dataset Condensation. In *CVPR*, pages 7856–7865, 2023. 1

[59] Yongchao Zhou, Ehsan Nezhadarya, and Jimmy Ba. Dataset distillation using neural feature regression. *NeurIPS*, 35:9813–9827, 2022. 1

[60] Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023. 2