

# DataDream: Few-shot Guided Dataset Generation

Jae Myung Kim<sup>1,2,3,\*</sup>   Jessica Bader<sup>2,3,4,\*</sup>   Stephan Alaniz<sup>2,3</sup>  
Cordelia Schmid<sup>5</sup>   Zeynep Akata<sup>2,3,4</sup>

<sup>1</sup>University of Tübingen   <sup>2</sup>Helmholtz Munich   <sup>3</sup>MCML   <sup>4</sup>TUM  
<sup>5</sup>Inria, Ecole normale supérieure, CNRS, PSL Research University

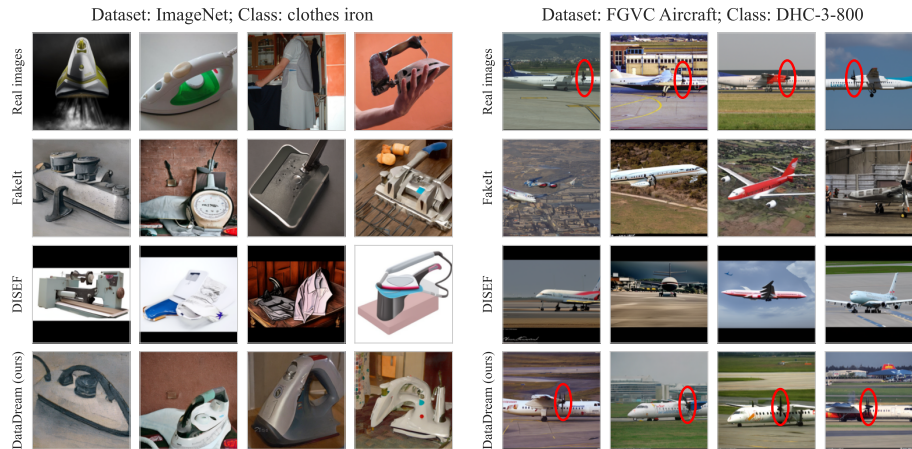
\*equal contribution

**Abstract.** While text-to-image diffusion models have been shown to achieve state-of-the-art results in image synthesis, they have yet to prove their effectiveness in downstream applications. Previous work has proposed to generate data for image classifier training given limited real data access. However, these methods struggle to generate in-distribution images or depict fine-grained features, thereby hindering the generalization of classification models trained on synthetic datasets. We propose DataDream, a framework for synthesizing classification datasets that more faithfully represents the real data distribution when guided by few-shot examples of the target classes. DataDream fine-tunes LoRA weights for the image generation model on the few real images before generating the training data using the adapted model. We then fine-tune LoRA weights for CLIP using the synthetic data to improve downstream image classification over previous approaches on a large variety of datasets. We demonstrate the efficacy of DataDream through extensive experiments, surpassing state-of-the-art classification accuracy with few-shot data across 7 out of 10 datasets, while being competitive on the other 3. Additionally, we provide insights into the impact of various factors, such as the number of real-shot and generated images as well as the fine-tuning compute on model performance. The code is available at <https://github.com/ExplainableML/DataDream>.

## 1 Introduction

The emergence of text-to-image generative models, such as Stable Diffusion [34], not only enables us to create photo-realistic synthetic images, but it also presents opportunities to enhance downstream tasks. One potential application lies in training or fine-tuning task-specific models on synthetic data. This is shown to be particularly useful in domains where access to real data is limited [9, 14, 22, 40], as generative models offer a cost-effective means of generating large amounts of training data. In this paper, we study the impact of synthetic training data on image classification tasks in low-shot settings, i.e. where we have access to a few images per class, but the collection of an entire dataset would be prohibitively expensive.

Previous research has primarily focused on using the class names of a given dataset [14, 38, 39, 44] to inform the data generation process. Concretely, they



**Fig. 1: Synthetic images comparison.** The previous methods for synthesizing training data sometimes misunderstand the class name due to its ambiguity (FakeIt [38] confuses the clothes iron with the metal iron) or fail to capture fine-grained features (DISEF [9] generated images lack the propeller in front of the wings in the DHC-3-800 aircraft, a red circle indicates the propeller). Meanwhile, our method accurately generates images of the class of interest and captures fine-grained details.

generated images with text-to-image diffusion models, using the class names as conditional input. To better guide the model to generate accurate depictions of the target object, they incorporated textual descriptions of each class to the prompt, sourced from language models [14, 44] or human-annotated class descriptions [38]. While intuitive, these methods lead to some generated images lacking the object of interest. For instance, while the real images for the class name "clothes iron" from the ImageNet [36] dataset display the appliance for ironing clothes, the images generated by FakeIt [38] mostly depict iron as the metal or arbitrary objects made thereof (Figure 1, left). This occurs when the generative model misunderstands class name ambiguities or rare classes. Such misalignment between the real and synthetic images limits the generated images' informational value for image classification and hinders performance gains.

To bridge the gap between real and synthetic images, real images can better inform the generative model about the characteristics of the real data distribution [2, 9, 10, 14, 48]. For instance, the concurrently developed DISEF [9] method uses few-shot samples as conditional input to the pre-trained diffusion model by starting from a partially noised real image when generating the synthetic dataset. It additionally uses a pre-trained captioning model to diversify the text-to-image prompt. While this approach improves the alignment of real and synthetic data distributions, it sometimes falls short of capturing fine-grained features. For example, while the real images for the class name "DHC-3-800" in the Aircraft [26] dataset include a propeller in front of the wings, the synthetic images by DISEF

lack this detail (Figure 1, right). Accurately representing class-discriminative features can be critical for classification tasks, particularly in fine-grained datasets.

In this work, we propose a novel approach, called **DataDream**, aimed at adapting generative models using few-shot real data. Motivated by personalized generative modeling methods [11, 35], in which generation models are fine-tuned with a small set of real images depicting an *identical object*, our method focuses on aligning the generative model to a target dataset which has *multiple classes and diverse objects for each class*. This differs from previous few-shot dataset generation methods such as [9, 14], which have not explored fine-tuning the generative model. Concretely, we adapt Stable Diffusion [34] with LoRA [17] in two ways: **DataDream<sub>cls</sub>**, which trains LoRA per class, and **DataDream<sub>dset</sub>**, which trains a single LoRA for all classes. To the best of our knowledge, we are the first to propose using few-shot data to adapt the generative model for synthetic training data, rather than leveraging the frozen, pre-trained generation model. Following training, we generate images with the same prompt used for fine-tuning DataDream, resulting in images depicting the object of interest (e.g. the clothes iron) or fine-grained features (e.g. the propeller of the DHC-3-800 plane) as shown in the last row of Figure 1.

We demonstrate the effectiveness of DataDream through extensive experiments, achieve the state of the art across all datasets when using only synthetic data, and achieve the best performance on 7 out of 10 datasets when training with both real few-shot and synthetic data. To understand the effectiveness of our method, we analyze the alignment between real and synthetic data, revealing that our method shows better alignment with the distribution of real data compared to baseline methods. Finally, we explore the scalability of our method by increasing the number of synthetic data points and real samples, showing the potential benefits of larger datasets. To summarize, the contributions of our work are as follows:

1. We introduce DataDream, a novel few-shot method which adapts Stable Diffusion to generate better in-distribution images for downstream training that outperforms state-of-the-art few-shot classification on 7 out of 10 datasets, with the other 3 comparable.
2. We emphasize the importance of reporting results with only synthetic data. We demonstrate that our method achieves superior performance when training the classifier with solely synthetic data, in some cases outperforming those trained solely with real few-shot images, indicating that our method generates images that glean more insightful information from the few-shot real data.
3. We study the effectiveness of our method by analyzing the distribution alignment between synthetic data and real data. Under few-shot guidance, synthetic data generated by our method aligns the best with real data.

## 2 Related work

Synthetic image generation has made immense progress, now being capable of generating images that even humans may find difficult to distinguish from real images. In the following, we review related work on image generation and training on synthetic data.

**Synthetic Image Generation.** The suite of image generation models is growing, including Variational Auto-Encoders [19], GANs [12], and Diffusion Models [34]. With their recent popularity, diffusion models such as Stable Diffusion [34], SDXL [31], DALL-E [4, 33], Imagen [37], GLIDE [27], and Wuerstchen [30] have revolutionized text-to-image generation. Diffusion models aim to incrementally de-noise data by modeling the reverse process of a Markov chain progressively adding Gaussian noise to the sample conditioned on text. At test-time, this facilitates the generation of synthetic images from specified text and random noise. These large pre-trained models can be adapted to user specific needs [11, 35] or better control generation [7, 24]. Textual inversion [11] uses a small number of images of a specific object to learn a representational language token which can be used to prompt the frozen generation model to create better images of that object (e.g. a photo of *your* cat, rather than *a* cat). On the other hand, DreamBooth [35] achieves personalization by fine-tuning the generation model while providing a unique input token with two losses: one to reconstruct the personalized concept, and the other to preserve the original model generations without the unique token.

**Training with Synthetic Data.** A pool of research has blossomed in its wake, exploring downstream applications; namely: can models be trained on synthetic data? Some works augmented real datasets with synthetic images [3, 6, 10, 48]. Others focused on pre-training on large amounts of synthetic data, followed by fine-tuning on a limited number of real images [13, 40]. Similarly, several works evaluated the effectiveness of training on entirely synthetic datasets [13, 38].

Different tasks have been considered, including classification [3, 6, 14, 38, 39, 48], object detection [22], image generation [1], and representation learning [40]. Attempts have been made to optimize the selection process from large pools of synthetic data, generally by focusing on two primary factors: faithfulness and diversity. Faithfulness has been addressed by CLIP filtering [10, 14, 22], including additional class information [38], and spectral clustering [22]. On the other hand, diversity can be increased by lowering the guidance scale [38], generating a wide variety of natural language prompts with LLMs [13, 14], specifying domains [10, 39] or backgrounds [38], and using multiple text prompt templates [6]. Generally, data collection is considered resource intensive, while generating synthetic data is comparatively inexpensive and can therefore be done at scale; [38] showed that as the number of synthetic images increases, model performance can even surpass that of models trained on a lower fixed number of real images.

Finally, the few-shot setting is seeing increased interest, where the focus lies on leveraging large amounts of synthetic data in conjunction with limited

amounts of real data. More than simply pooling the data sources together, we can guide the generation of better synthetic data with real data. In [14], the authors explored two strategies: 1) generating images by starting from a partially noised few-shot sample and 2) using the similarity of synthetic image features to real ones to remove low-confidence samples. When adapting the CLIP model using Classifier Tuning [42] the first strategy works best. Concurrently to our work, Diversified In-domain Synthesis with Efficient Fine-tuning (DISEF) [9] proposes to create a synthetic augmentation pipeline which leverages few-shots by starting the generation process from a noised real sample (same as [14]), then promotes diversity by denoising it conditioned on the caption from a different real image. The authors apply CLIP filtering to remove synthetic images which would be classified incorrectly and then adapt CLIP as the classifier with LoRA [17] on either the few-real shots alone or the combination of few-shots and synthetic data. In contrast to these methods, we propose to additionally fine-tune the diffusion model with LoRA to obtain a better alignment with the real data distribution.

### 3 Methodology

In this section, we start by describing the preliminaries in §3.1, before introducing DataDream in §3.2. DataDream fine-tunes the text-to-image diffusion model with few-shot data. To measure performance, synthetic images are generated with the adapted model and a classifier trained on both synthetic and real data.

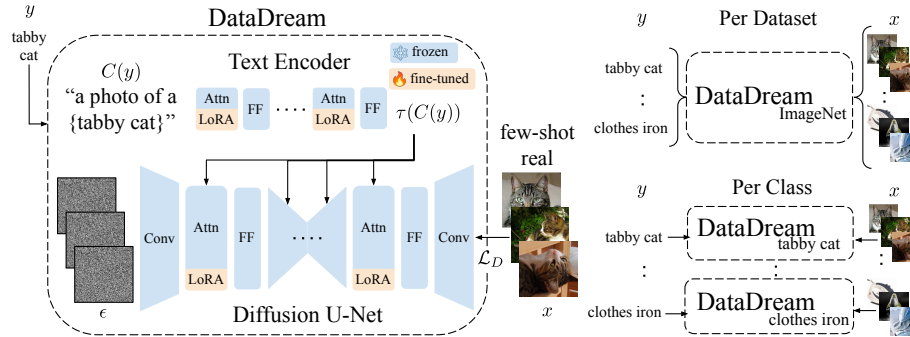
#### 3.1 Preliminaries

**Latent diffusion model.** We implement our method based on Stable Diffusion [34], a probabilistic generative model that learns to generate realistic images using a textual prompt. Given data  $(x, c) \in \mathcal{D}$ , where  $x$  is an image and  $c$  is a caption describing  $x$ , the model learns a conditional distribution  $p(x|c)$  by gradually denoising the Gaussian noise in the latent space. Given a pretrained encoder  $E$  that encodes an image  $x$  to a latent  $z$ , i.e.  $z = E(x)$ , the objective function is defined as:

$$\min_{\theta} \mathbb{E}_{(x,c) \sim \mathcal{D}, \epsilon \sim \mathcal{N}(0,1), t} \left[ \|\epsilon - \epsilon_{\theta}(z_t, \tau(c), t)\|_2^2 \right], \quad (1)$$

where  $t$  is a timestep,  $z_t$  is a latent noised  $t$  steps from the latent  $z$ ,  $\tau$  is a text encoder, and  $\epsilon_{\theta}$  is a latent diffusion model. Intuitively, the parameters  $\theta$  are trained to denoise the latent  $z_t$ , given a text prompt  $c$  as conditional information. In the inference phase, a random noise vector  $z_T$  is passed through the latent diffusion model  $T$  times, along with the caption  $c$ , to get a denoised latent  $z_0$ .  $z_0$  is then fed into a pretrained decoder  $D$  to get an image  $x' = D(z_0)$  for the text-to-image generation.

**Low-rank adaptation.** The Low-Rank Adaption method (LoRA) [17], is a fine-tuning method to adapt a large pre-trained model to downstream tasks in



**Fig. 2:** Overview of DataDream. We fine-tune LoRA weights for the linear weights of the attention layers in both the text-encoder and the diffusion U-net to generate images closer to the few-shot images. We can train one set of DataDream weights for the whole dataset sharing common dataset-specific characteristics between classes, or a separate set of weights for each class to better learn fine-grained details of each classes.

a parameter-efficient manner. Given pre-trained model weights  $\theta \in \mathbb{R}^{d \times k}$ , LoRA introduces a new parameter  $\delta \in \mathbb{R}^{d \times k}$  that is decomposed into two matrices,  $\delta = BA$ , where  $B \in \mathbb{R}^{d \times r}$  and  $A \in \mathbb{R}^{r \times k}$  with small LoRA rank  $r$ ,  $r \ll \min(d, k)$ . The LoRA weights are added to the model weights to obtain the fine-tuned weights, i.e.  $\theta^{(\text{ft})} = \theta + \delta$ , for adaptation to downstream tasks. During training,  $\theta$  remains fixed while only  $\delta$  is updated.

### 3.2 DataDream method

Our goal is to improve classification performance by leveraging synthetic images generated by diffusion models. To this end, it is crucial to align the synthetic image distribution to that of the real images. We achieve alignment by adapting the diffusion model to a few-shot dataset of real images.

We assume access to a few-shot dataset  $\mathcal{D}^{\text{fs}} = \{(x_i, y_i)\}_{i=1}^{KN}$ , where  $x_i$  is an image,  $y_i \in \{1, 2, \dots, N\}$  is its label,  $K$  is the number of samples per class, and  $N$  is the number of classes. To match the real data distribution, we fine-tune it with the few-shot dataset  $\mathcal{D}^{\text{fs}}$ . Concretely, we introduce LoRA weights in both the text-encoder and the U-net of the diffusion model, where we make the parameter-efficient choice of adapting the attention layers. For every attention layer, we consider the query, key, value, and output projection matrices  $W_q, W_k, W_v, W_o$ , where for each matrix, the linear projection is replaced by

$$h_{l,\star} = W_\star h_{l-1} + B_\star A_\star h_{l-1} \quad (2)$$

with  $h$  representing the input/output activations of the projections, and resulting in the trainable LoRA weights  $\delta^{(l)} = \{A_\star, B_\star | \forall \star \in \{q, k, v, o\}\}$  for every attention layer  $l$ . We omit bias weights for notational simplicity. All other model parameters are kept frozen (including  $W_\star$ ) while  $\delta$  weights are optimized with

gradient descent. To start training from the pre-trained diffusion model checkpoint, weight matrices  $B_\star$  are initialized with zeros while  $A_\star$  is initialized randomly. As a result, the combined fine-tuning weights  $B_\star A_\star$  are zero initially and incrementally learn modifications to the original pre-trained weights. At test time, LoRA weights can be integrated into the model by updating the weights with  $W_\star^{(\text{ft})} = W_\star + B_\star A_\star$ , such that inference time is equivalent to the pre-trained model. In contrast to DreamBooth [35], we do not fine-tune all network weights and do not add a preservation loss, as its regularization would prevent a strong alignment with the real images.

We further consider two settings: 1) **DataDream<sub>dset</sub>**, where we train the LoRA weights of the diffusion model on the whole dataset  $\mathcal{D}^{\text{fs}}$ , and 2) **DataDream<sub>cls</sub>**, where we initialize  $N$  sets of LoRA weights  $\{\delta_n | n = 1, \dots, N\}$ , one for each of the dataset classes trained on the subset  $\mathcal{D}_n^{\text{fs}} = \{(x, y) | (x, y) \in \mathcal{D}^{\text{fs}}, y = n\}$ .

In the DataDream<sub>dset</sub> setting, the original model parameters  $\theta$  are kept frozen and only the LoRA weights are trained with the objective function

$$\min_{\delta} \mathcal{L}_D = \min_{\delta} \mathbb{E}_{(x,y) \sim \mathcal{D}^{\text{fs}}, \epsilon \sim \mathcal{N}(0,1), t} \left[ \|\epsilon - \epsilon_{\theta, \delta}(z_t, \tau_{\delta}(C(y)), t)\|_2^2 \right]. \quad (3)$$

In the DataDream<sub>cls</sub> setting,  $\mathcal{D}_n^{\text{fs}}$  and  $\delta_n$  would replace  $\mathcal{D}^{\text{fs}}$  and  $\delta$ , respectively. Since we use a text-to-image diffusion model, we define the text condition through the function  $C$  which maps the label  $y$ , i.e. class name, to a prompt using the standard template, "a photo of a [CLS]" [32, 47]. The prompt is passed through the text encoder and subsequently used during the decoding steps of the diffusion model. We illustrate both DataDream fine-tuning and our two settings in Figure 2.

Both settings have distinct advantages. In DataDream<sub>dset</sub>, LoRA weight sharing between classes allows knowledge transfer about common characteristics within the whole dataset. This would be beneficial in a fine-grained dataset that shares the coarse-grained features across classes. On the other hand, DataDream<sub>cls</sub> allocates more weights to learn about details of each class, which allows the generation model to better align with the per-class data distribution.

After adapting the diffusion model to the few-shot dataset, we generate 500 images per class with the adapted model conditioned on the same textual prompt used for DataDream, forming a synthetic dataset  $\mathcal{D}^{\text{synth}}$ . We train a classifier on either only synthetic images or the combination of synthetic and real few-shot images  $\mathcal{D}^{\text{fs}}$ .

For classifier training, we adapt a CLIP model [32], similar to previous work in few-shot classification [9]. We add LoRA adaptors [17] to both image encoder and text encoder of CLIP ViT-B/16 model [32]. When training with synthetic and real images jointly, we use a weighted average of the losses from real data and synthetic data,

$$\mathcal{L}_C = \lambda \mathbb{E}_{(x,y) \sim \mathcal{D}^{\text{fs}}} \text{CE}(f(x), y) + (1-\lambda) \mathbb{E}_{(x,y) \sim \mathcal{D}^{\text{synth}}} \text{CE}(f(x), y), \quad (4)$$

where  $\lambda$  is the weight assigned to the loss from real data and the function CE is a cross-entropy loss.

## 4 Experiments

In this section, we present our experimental results on DataDream. We present details of the experimental setup in §4.1. In §4.2, we compare our methods to baselines both quantitatively and qualitatively. Furthermore, we analyze the synthetic data of DataDream to understand why it outperforms baselines in §4.3, followed by ablation studies in §4.4.

### 4.1 Experimental setup

**Benchmarks.** We evaluate our method on 10 datasets: ImageNet [36], Oxford Pets [29] containing fine-grained pet classes, FGVC Aircraft [26] containing fine-grained aircraft classes, Food101 [5] containing common food classes, Stanford Cars [20] containing fine-grained car classes, DTD [8] with texture images, EuroSAT [15] with satellite images, Flowers 102 [28] containing fine-grained flower classes, SUN397 [43] with scene images, and Caltech 101 [21] with pictures of common objects.

**Implementation details.** We implement DataDream based on Stable Diffusion [34] version 2.1. For each seed, we randomly sample the few-shot images from the training samples of each dataset. Our method is trained for 200 epochs with a batch size of 8 for all datasets, with the exception of  $\text{DataDream}_{\text{dset}}$  on ImageNet, which is trained for 100 epochs. Hence,  $\text{DataDream}_{\text{dset}}$  and  $\text{DataDream}_{\text{cls}}$  share the same amount of training compute, i.e. each of the  $N$   $\text{DataDream}_{\text{cls}}$  adapter weights (one for each class) performs  $S/N$  update steps where  $S$  is the total number of steps of  $\text{DataDream}_{\text{dset}}$  for the whole dataset. We use AdamW [25] as an optimizer and learning rate  $1e-4$ , with a cosine annealing scheduler. We use LoRA rank  $r = 16$  for all adapted weights in DataDream. For synthetic image generation with DataDream, we use 50 steps and guidance scale 2.0. We generate 500 images per class if not mentioned otherwise. For the classifier, we use CLIP ViT-B/16 [32] as a base model, and fine-tune LoRA applied on both the image encoder and text encoder of CLIP with rank 16. We set the weight assigned to the real loss term to  $\lambda = 0.8$ . DataDream is computed on three random seeds. Additional implementation details can be found in Appendix B.

**Baseline methods.** As all methods adapt CLIP ViT-B/16 as the classifier, we provide CLIP zero-shot performance as a baseline. In our first setting, we update the classifier using only synthetic data. For this, we compare against two alternative data-generation methods: IsSynth [14] and DISEF [9]. In our second setting, classifier adaptation uses the synthetic data in addition to the few-shot real data. We refer to LoRA [17] training with only the real few-shot data as Real-finetune, to signify that this is the baseline version of DataDream, without the benefit of our synthetic data, as done in [9].  $\text{DataDream}_{\text{dset}}$ ,  $\text{DataDream}_{\text{cls}}$ , and DISEF [9] all build upon this foundation. These experiments highlight the benefit of adding synthetic data to the real few-shot images. We also compare against several SOTA few-shot methods. In these, we include two Parameter



Method	R/S	IN	CAL	DTD	EuSAT	AirC	Pets	Cars	SUN	Food	FLO	Avg
CLIP (zero-shot) [32]		70.2	96.1	46.1	38.1	23.8	91.0	63.1	63.8	85.1	71.8	64.1
IsSynth [14]	✓	70.0 $\pm$ 0.6	95.7 $\pm$ 0.7	67.6 $\pm$ 0.5	71.3 $\pm$ 2.9	34.5 $\pm$ 3.4	92.1 $\pm$ 0.6	65.9 $\pm$ 0.0	72.2 $\pm$ 0.0	85.4 $\pm$ 0.1	90.0 $\pm$ 0.2	74.5 $\pm$ 0.9
DISEF [9]	✓	67.1 $\pm$ 0.2	93.4 $\pm$ 1.0	66.1 $\pm$ 0.6	69.2 $\pm$ 2.7	26.8 $\pm$ 2.2	91.0 $\pm$ 0.0	63.2 $\pm$ 0.0	73.5 $\pm$ 0.1	85.1 $\pm$ 0.0	85.4 $\pm$ 0.7	72.1 $\pm$ 0.8
DataDream <sub>cls</sub> (ours)	✓	<b>71.6<math>\pm</math>0.2</b>	<b>96.4<math>\pm</math>0.0</b>	68.6 $\pm$ 2.0	<b>85.4<math>\pm</math>2.9</b>	60.3 $\pm$ 0.9	<b>94.2<math>\pm</math>0.4</b>	90.5 $\pm$ 0.3	74.5 $\pm$ 0.0	<b>86.9<math>\pm</math>0.1</b>	97.2 $\pm$ 0.2	82.6 $\pm$ 0.7
DataDream <sub>dset</sub> (ours)	✓	71.5 $\pm$ 0.0	96.2 $\pm$ 0.1	<b>69.5<math>\pm</math>1.2</b>	80.3 $\pm$ 4.1	<b>71.2<math>\pm</math>0.1</b>	94.0 $\pm$ 0.1	<b>92.2<math>\pm</math>0.1</b>	<b>74.5<math>\pm</math>0.1</b>	86.7 $\pm$ 0.1	<b>98.0<math>\pm</math>0.4</b>	<b>83.4<math>\pm</math>0.7</b>
Real-finetune	✓	73.4 $\pm$ 0.2	96.8 $\pm$ 0.1	78.3 $\pm$ 2.8	93.5 $\pm$ 0.7	59.3 $\pm$ 2.8	94.0 $\pm$ 0.1	87.5 $\pm$ 0.6	77.1 $\pm$ 0.1	<b>87.6<math>\pm</math>0.0</b>	98.7 $\pm$ 0.1	84.6 $\pm$ 0.8
IsSynth [14]	✓	73.9 $\pm$ 0.1	97.4 $\pm$ 0.2	<b>81.6<math>\pm</math>0.4</b>	93.9 $\pm$ 0.1	64.8 $\pm$ 0.8	92.1 $\pm$ 0.1	88.5 $\pm$ 0.3	<b>77.7<math>\pm</math>0.0</b>	86.0 $\pm$ 0.0	99.0 $\pm$ 0.0	85.5 $\pm$ 0.2
DISEF [9]	✓	73.8 $\pm$ 0.2	97.0 $\pm$ 0.1	81.5 $\pm$ 0.6	<b>94.0<math>\pm</math>0.5</b>	64.3 $\pm$ 0.4	92.6 $\pm$ 1.2	87.9 $\pm$ 0.5	77.6 $\pm$ 0.1	86.2 $\pm$ 0.6	99.0 $\pm$ 0.2	85.4 $\pm$ 0.4
DataDream <sub>cls</sub> (ours)	✓	73.8 $\pm$ 0.1	<b>97.6<math>\pm</math>0.2</b>	<b>81.6<math>\pm</math>0.4</b>	93.8 $\pm$ 0.3	68.3 $\pm$ 0.4	94.5 $\pm$ 0.3	91.2 $\pm$ 0.2	77.5 $\pm$ 0.1	87.5 $\pm$ 0.1	<b>99.4<math>\pm</math>0.2</b>	86.5 $\pm$ 0.4
DataDream <sub>dset</sub> (ours)	✓	<b>74.1<math>\pm</math>0.3</b>	96.9 $\pm$ 0.7	<b>81.6<math>\pm</math>0.6</b>	93.4 $\pm$ 0.0	<b>72.3<math>\pm</math>0.2</b>	<b>94.8<math>\pm</math>0.3</b>	<b>92.4<math>\pm</math>0.1</b>	77.5 $\pm$ 0.1	<b>87.6<math>\pm</math>0.1</b>	<b>99.4<math>\pm</math>0.1</b>	<b>87.0<math>\pm</math>0.4</b>

**Table 1: Few-shot classification performance with DataDream using real 16-shot and synthetic images** where the training dataset includes synthetic data only (top), or synthetic data + 16 real shots (bottom). All results use CLIP ViT-B/16 as the base classification model, and 500 synthetic images generated by 16 real shots. Datasets are IN: ImageNet, CAL: Caltech 101, EuSAT: EuroSAT, AirC: FGVC Aircraft, FLO: Flowers 102. R/S means using real/synthetic images for fine-tuning. DataDream and baseline methods are computed on three random seeds.

Efficient Fine-Tuning (PEFT) techniques: VPT [18] and CoOp [47], which only use real few-shot data. We additionally compare to two SOTA image generation techniques, IsSynth [14] and DISEF [9]. For fair comparisons, we use Stable Diffusion v2.1 to generate images for all baselines instead of the originally used GLIDE [27] or Stable Diffusion v1.5. More details are described in Appendix C.

## 4.2 Classification performance with DataDream

**Quantitative results on solely synthetic data.** We refer to the upper portion of Table 1 for the synthetic-only setting, where we show that DataDream-generated data achieves state-of-the-art results on all 10 datasets. For example on FGVC Aircraft [26], DataDream<sub>dset</sub> achieves an impressive 47.4% point increase over the CLIP zero-shot model. In addition, on Stanford Cars [20] DataDream<sub>dset</sub> achieves 92.2%, while IsSynth [14] is at 65.9% and DISEF [9] at 63.2%. On Flowers102 [28], DataDream<sub>dset</sub> obtains 98.0% while IsSynth and DISEF reach only 90.0% and 85.4%, respectively. These boosts signify that DataDream is able to closely follow the real few-shot data distribution in its generated images.

We believe that this evaluation benchmark allows the best assessment of the quality of the synthetic image generations for training image classifiers. While adding real data to the synthetic images at training time generally provides a performance boost, it also makes it harder to quantify the quality of the synthetic images for the task, because most of the improvement still stems from the real images. Hence, issues with synthetic data generation, such as redundancy or class misrepresentation, will be more visible in synthetic-only benchmarks.

Comparing DataDream<sub>cls</sub> and DataDream<sub>dset</sub>, the results are split over method superiority. We hypothesize that this difference comes from inherent dataset properties. For datasets where all classes share high visual similarity, sharing weights becomes beneficial as distribution characteristics generalize across

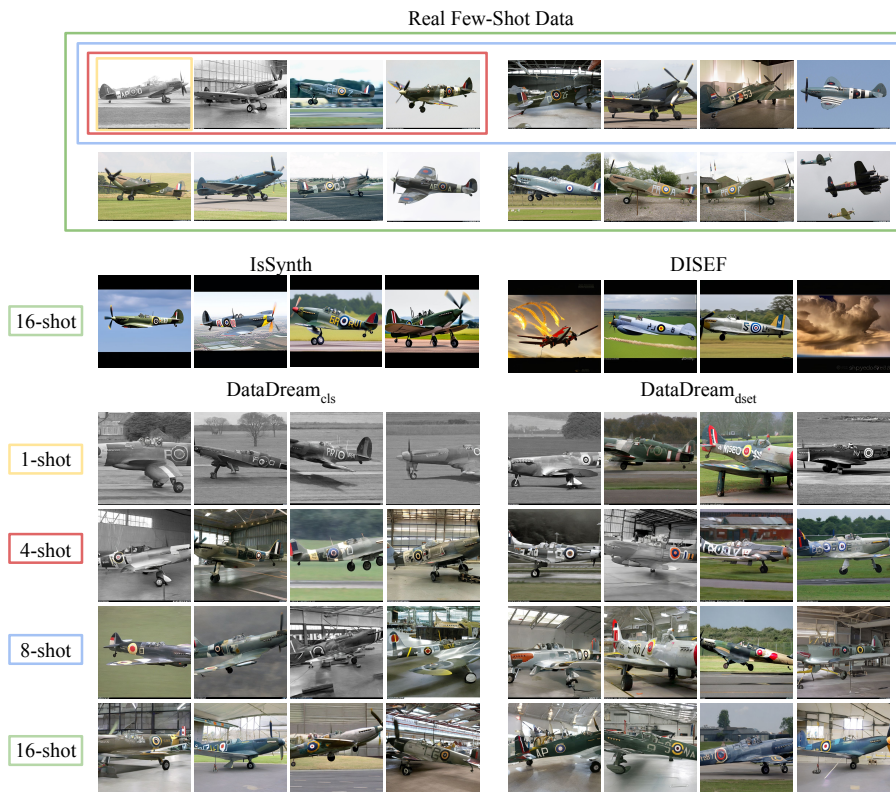
Method	S	IN	CAL	DTD	EuSAT	AirC	Pets	Cars	SUN	Food	FLO	Avg
VPT [18]		69.6	95.4	66.1	92.3	36.2	91.8	69.0	70.5	87.0	91.0	76.9
CoOp [47]		68.0	95.2	70.7	87.1	45.5	89.9	81.4	73.0	83.7	97.6	79.2
IsSynth [14]	✓	73.9	<b>97.4</b>	<b>81.6</b>	93.9	64.8	92.1	88.5	<b>77.7</b>	86.0	99.0	85.5
DISEF [9]	✓	73.8	97.0	81.5	<b>94.0</b>	64.3	92.6	87.9	77.6	86.2	99.0	85.4
DataDream <sub>dset</sub> (ours)	✓	<b>74.1</b>	96.9	<b>81.6</b>	93.4	<b>72.3</b>	<b>94.8</b>	<b>92.4</b>	77.5	<b>87.6</b>	<b>99.4</b>	<b>87.0</b>

**Table 2: Comparing DataDream with few-shot SOTA.** We compare DataDream with SOTA few-shot methods. The base setting, dataset abbreviations, and setting notations match those in Table 1. S indicates methods using synthetic data generation.

classes. For example, we find that FGVC Aircraft [26] and Stanford Cars [20] show a significant advantage of DataDream<sub>dset</sub> over DataDream<sub>cls</sub>. On the other hand, datasets where classes span a wide range benefit from fully specializing to the unique classes, as seen in the results for Caltech101 [21] and Food101 [5].

**Quantitative results on real + synthetic data.** In Table 1 (bottom), we present the results for the synthetic + real setting. Real-finetune provides the foundation for this section, consisting of LoRA applied to CLIP with the real few-shot data. DISEF, DataDream<sub>cls</sub>, and DataDream<sub>dset</sub> build upon this by adding their respective synthetic images to the training data, which is generated from the same few-shot data. Comparing DataDream<sub>dset</sub> and DataDream<sub>cls</sub> to Real-finetune, we observe that our synthetic data improves performance on 9 out of 10 datasets over naive use of real few-shot data. For example, on FGVC Aircraft [26], our synthetic data facilitates an improvement of 13.0% over using the real few-shot data naively. In comparison, DISEF only achieves a 5.0% increase. Furthermore, we improved upon Stanford Cars [20] by 4.9%, where DISEF saw only a 0.4% increase. This shows that generating images with DataDream consistently provides value not only over naive use of the few-shot examples, but also over other data generation techniques. In fact, especially in case of the Stanford Cars dataset, the real images do not provide more information than the synthetic images generated by our model (92.2% on synthetic only vs 92.4% real + synthetic settings), which is an exciting observation.

**Quantitative results comparing with SOTA.** We compare DataDream with SOTA few-shot methods in Table 2. Ours, i.e. DataDream<sub>dset</sub>, improves over the previous SOTA on 7 out of 10 datasets, while being competitive on the other 3. On the FGVC Aircraft [26] dataset, we improve SOTA by 7.5%, on Pets [29] by 2.2%, and on Stanford Cars [20] by 3.9%. This highlights that synthetic images generated by DataDream provide more training value than the previous SOTA generation method. On average, we improve over the next best synthetic augmentation method by 1.5% and over the best method without data generation, CoOp [47], by 7.8%.



**Fig. 3: Qualitative results with increasing number of shots vs 16-shot images generated with SOTA of the class Spitfire from the FGVC Aircraft [26] dataset.** The real few-shot images at the top are used to generate the presented synthetic images at the bottom. We always use a fixed set of 16 samples, i.e. 1-shot image is a subset of 16-shots, to insure fairness in comparing results with the increasing number of shots.

### 4.3 Analysis of DataDream

**Qualitative results.** We provide a qualitative analysis of DataDream in Figure 3 of the Spitfire class in FGVC Aircraft. To support our 1-, 4-, 8-, and 16-shot generated images, we include the real few-shot examples used to generate them. We also show previous SOTA images for comparison, from two other image-generation methods: DISEF [9] and IsSynth [14].

When comparing to the previous SOTA, we notice that DataDream is better able to generate images that match the target domain. For example, they imitate that in the context of the dataset, planes are more likely to be photographed on the ground, in a hangar, or taking off than in the air, unlike both previous methods that generate images that are unlikely to be found in the dataset. We also notice that our models are better able to match the color palette of the real data, as opposed to DISEF, where we find the colors to be too bright compared to



**Fig. 4: Distribution of FID scores per-class.** The FID score is calculated per-class to measure how close the synthetic data distribution is to the real data distribution.

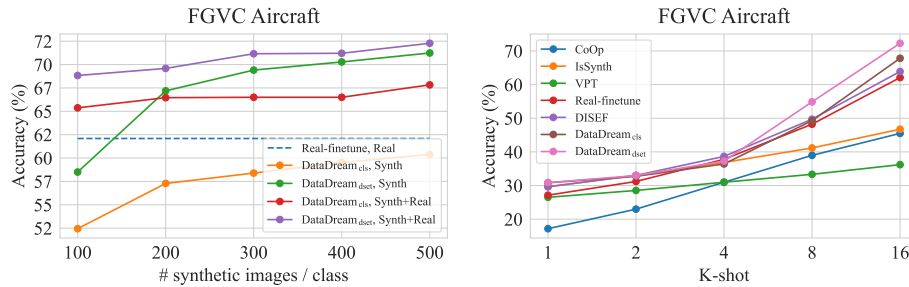
the real data. Furthermore, our model is the only one of all three that replicates the black border at the bottom of all images, even after only a single shot.

Furthermore, we notice that DISEF has a higher tendency to generate out-of-distribution data, sometimes omitting the target class entirely and therefore creating a need for CLIP filtering. We hypothesize this might be due to their use of diverse captions, which may sometimes guide the image generation too far from the core distribution. By focusing on fidelity to the class distribution and keeping our prompts simple, we generate fewer out-of-distribution samples. This allows us to use all samples generated, which is a better use of resources.

We also notice some interesting differences between  $\text{DataDream}_{\text{dset}}$  and  $\text{DataDream}_{\text{cls}}$ . On the one hand, we find that  $\text{DataDream}_{\text{cls}}$  is better able to accurately represent the Spitfire class, especially at a low number of shots. On the other hand, the additional data in  $\text{DataDream}_{\text{dset}}$  allows it to avoid certain overfitting mistakes, such as creating only black and white images after the first shot, which happens to be a monochrome image.

Comparing  $\text{DataDream}$  models trained on different numbers of few-shot examples, we notice an increase in quality with number of images, showing qualitatively the benefit of adding even a few more images. Already at four shots, we obtain images that are not only better quality, but closer to the real data domain. We also note that the lower the number of shots, the more the model benefits from careful selection of a diverse and representative group, so that the model does not pick up on patterns that are not representative of the full data distribution. At only four shots, we notice that the real images contain a specific color palette that is not necessarily representative of the full dataset, as evidenced by the next four images; this led to the 4-shot results lacking diversity of color and brightness. This goes to show that wherever possible, careful selection of representative data is highly beneficial. It also highlights the ability of our method to find and replicate patterns in the few-shot distribution.

**Distribution alignment.** The qualitative analysis shows  $\text{DataDream}$  being able to capture both the presence of objects of interest and the fine-grained features essential for class discrimination. To gain more insights, we examine the alignment between synthetic and real datasets. To quantitatively assess the



**Fig. 5: Ablation study of DataDream.** Left: We vary the number of synthetic images per class to understand the scaling effect. Right: We vary the number of real examples used for training DataDream.

alignment, we use the Fréchet Inception Distance (FID) [16] score, a metric that quantifies the quality of generated images. Concretely, we compute the set of FID scores for each method by evaluating the distance between the distribution of synthetic images and that of real images on a per-class basis. Lower FID indicates synthetic images are closer to the real data distribution. We visualize the FID scores using histograms, as shown in Figure 4.

In the experiment on ImageNet, we observe that the histogram for our method skews left, indicating lower FID scores. Meanwhile, the histogram of DISEF tends to lean towards the right. This is attributed to how DISEF uses LLM-generated prompts for image generation. While it gives the generated images diversity, it also introduces out-of-distribution artifacts, as observed in the qualitative analysis. Compared to DISEF, IsSynth aligns more with the real data, but may have less diversity due to its usage of a standard prompt, generating similar images from the conditioned real image. In contrast, our DataDream<sub>dset</sub> balances fidelity, due to the adaptation of the generative model to match few real shots, and diversity, due to the initial randomness in the generation pipeline. This results in the synthetic images of DataDream<sub>dset</sub> closely matching the real data distribution. We posit that the better alignment contributes to classification performance, as demonstrated in §4.2.

#### 4.4 Ablation study

**Accuracy scaling by number of synthetic images.** Previous literature has shown that as the number of synthetic images increases, model accuracy may also increase [38]. Therefore, we provide Figure 5, where the left part shows the effect on DataDream of increasing the number of images for FGVC Aircraft [26]. We find that as more images are generated, model accuracy increases in all settings: DataDream<sub>dset</sub> and DataDream<sub>cls</sub> and Synth and Synth + Real. Even at 500 images, we observe that performance is not yet saturated. Compared to Real-finetune, we observe that in the Synth + Real setting, DataDream performs better already starting at 100 images per class.

In the synthetic-only setting,  $\text{DataDream}_{\text{cls}}$  out-performs real images entirely after generating only 200 images. Another interesting result is the gap between  $\text{DataDream}_{\text{cls}}$  and  $\text{DataDream}_{\text{dset}}$ : this difference holds between any number of images, showing that  $\text{DataDream}_{\text{dset}}$  is a better fit for this dataset than  $\text{DataDream}_{\text{cls}}$ , regardless of the number of synthetic images. However, remembering from Table 1 that  $\text{DataDream}_{\text{dset}}$  performed better than  $\text{DataDream}_{\text{cls}}$  on 5 out of 10 datasets, we note that this advantage is dataset-dependent rather than a general trend.

**Varying the number of few-shot images  $K$ .** Furthermore, we operate in a  $K$ -shot regime; therefore, we expect that as  $K$  increases, the model accuracy should increase as well. As done in [9], we show the effect of 1-, 2-, 4-, 8-, and 16-shots on FGVC Aircraft [26] dataset, for 500 images in the real + synth setting, compared to previous literature. We observe that as the number of few-shot images increases,  $\text{DataDream}$  consistently shows higher accuracy than previous SOTA. We believe that this behavior is expected; as with any training or fine-tuning regime, at least a small training dataset base is necessary. Too few samples could be prone to overfitting, thus reducing variety and failing to include enough information to successfully understand the overall class distribution. At the same, since we use LoRA on a subset of all model parameters, we limit the amount of overfitting from our fine-tuning as compared to full model fine-tuning. This allows  $\text{DataDream}$  to outperform Real-finetune even when we use only a single shot. As more data becomes available, however,  $\text{DataDream}$  is able to successfully leverage even as few as four or eight shots to noticeably adapt to the data distribution, as was shown in Section 4.3. Hence, we obtain a relative performance boost when compared to other methods.

## 5 Conclusion

In this paper, we studied the efficacy of leveraging the generative models for improving the image classification performance in few-shot scenarios. We proposed  $\text{DataDream}$ , a method to generate synthetic data with the guidance of few-shot samples, which are then used for training the image classifier. We introduced LoRA adaptors on both the text encoder and the diffusion U-Net to efficiently fine-tune the generative model. We proposed two variants:  $\text{DataDream}_{\text{dset}}$ , which trains LoRA on the whole targeted dataset, and  $\text{DataDream}_{\text{cls}}$ , which adapts LoRA per-class. Our experiments demonstrate that our method consistently improves classification performance across benchmarks, both in the synthetic-only and synthetic+real settings. Through qualitative analysis, we observed that images generated by our method more precisely generate objects of interest as well as fine-grained details, contributing to their alignment with real data distributions, as quantitatively examined by FID scores. Furthermore, we investigated the scalability of our method by increasing the number of synthetic data samples and the number of real samples.

## Acknowledgements

Jae Myung Kim thanks the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD programs for support. This work was supported by the ERC (853489 - DEXIM) and the BMBF (Tübingen AI Center, FKZ: 01IS18039A). Cordelia Schmid would like to acknowledge the support by the Körber European Science Prize. Also, the authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. ([www.gauss-centre.eu](http://www.gauss-centre.eu)) for funding this project by providing computing time on the GCS Supercomputer JUWELS at Jülich Supercomputing Centre (JSC).

## References

1. Alemohammad, S., Casco-Rodriguez, J., Luzzi, L., Humayun, A.I., Babaei, H., LeJeune, D., Siahkoochi, A., Baraniuk, R.: Self-consuming generative models go mad. In: ICLR (2023)
2. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023)
3. Bansal, H., Grover, A.: Leaving reality to imagination: Robust classification via generated datasets. arXiv preprint arXiv:2302.02503 (2023)
4. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. <https://cdn.openai.com/papers/dall-e-3.pdf> **2**(3), 8 (2023)
5. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101—mining discriminative components with random forests. In: ECCV. pp. 446–461. Springer (2014)
6. Burg, M.F., Wenzel, F., Zietlow, D., Horn, M., Makansi, O., Locatello, F., Russell, C.: Image retrieval outperforms diffusion models on data augmentation. TMLR (2023)
7. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: Ilvr: Conditioning method for denoising diffusion probabilistic models. arXiv preprint arXiv:2108.02938 (2021)
8. Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., Vedaldi, A.: Describing textures in the wild. In: CVPR (2014)
9. da Costa, V.G.T., Dall’Asen, N., Wang, Y., Sebe, N., Ricci, E.: Diversified in-domain synthesis with efficient fine-tuning for few-shot classification (2023)
10. Dunlap, L., Umino, A., Zhang, H., Yang, J., Gonzalez, J.E., Darrell, T.: Diversify your vision datasets with automatic diffusion-based augmentation (2023)
11. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2022)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. *Advances in neural information processing systems* **27** (2014)
13. Hammoud, H.A.A.K., Itani, H., Pizzati, F., Torr, P., Bibi, A., Ghanem, B.: Synthclip: Are we ready for a fully synthetic clip training? arXiv preprint arXiv:2402.01832 (2024)
14. He, R., Sun, S., Yu, X., Xue, C., Zhang, W., Torr, P., Bai, S., Qi, X.: Is synthetic data from generative models ready for image recognition? In: ICLR (2023)

15. Helber, P., Bischke, B., Dengel, A., Borth, D.: Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *AEROS* (2019)
16. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems* **30** (2017)
17. Hu, E.J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. In: *ICLR* (2021)
18. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning. In: *ECCV* (2022)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013)
20. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: *ICCV workshop*. pp. 554–561 (2013)
21. Li, F.F., Andreeto, M., Ranzato, M., Perona, P.: Caltech 101 (Apr 2022). <https://doi.org/10.22002/D1.20086>
22. Lin, S., Wang, K., Zeng, X., Zhao, R.: Explore the power of synthetic data on few-shot object detection. In: *CVPR* (2023)
23. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: *NeurIPS* (2023)
24. Liu, X., Park, D.H., Azadi, S., Zhang, G., Chopikyan, A., Hu, Y., Shi, H., Rohrbach, A., Darrell, T.: More control for free! image synthesis with semantic diffusion guidance. In: *WACV* (2023)
25. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: *ICLR* (2018)
26. Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A.: Fine-grained visual classification of aircraft. *Tech. rep.* (2013)
27. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741* (2021)
28. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: *Indian Conference on Computer Vision, Graphics and Image Processing* (Dec 2008)
29. Parkhi, O.M., Vedaldi, A., Zisserman, A., Jawahar, C.V.: Cats and dogs. In: *CVPR* (2012)
30. Pernias, P., Rampas, D., Richter, M.L., Pal, C., Aubreville, M.: Würstchen: An efficient architecture for large-scale text-to-image diffusion models. In: *ICLR* (2023)
31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. In: *ICLR* (2023)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *ICML* (2021)
33. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* **1**(2), 3 (2022)
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: *CVPR* (2022)
35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: *CVPR*. pp. 22500–22510 (2023)
36. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *International journal of computer vision* **115**, 211–252 (2015)



37. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS* (2022)
38. Sariyildiz, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: *CVPR* (2023)
39. Shipard, J., Wiliem, A., Thanh, K.N., Xiang, W., Fookes, C.: Diversity is definitely needed: Improving model-agnostic zero-shot classification via stable diffusion (2023)
40. Tian, Y., Fan, L., Isola, P., Chang, H., Krishnan, D.: Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *NeurIPS* (2024)
41. Wang, Z., Liang, J., Sheng, L., He, R., Wang, Z., Tan, T.: A hard-to-beat baseline for training-free CLIP-based adaptation. In: *ICLR* (2024)
42. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: *CVPR*. pp. 7959–7971 (2022)
43. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: *CVPR*. pp. 3485–3492. IEEE (2010)
44. Yu, Z., Zhu, C., Culatana, S., Krishnamoorthi, R., Xiao, F., Lee, Y.J.: Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. *arXiv preprint arXiv:2312.02253* (2023)
45. Yun, S., Han, D., Oh, S.J., Chun, S., Choe, J., Yoo, Y.: Cutmix: Regularization strategy to train strong classifiers with localizable features. In: *ICCV* (2019)
46. Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D.: mixup: Beyond empirical risk minimization. In: *ICLR* (2018)
47. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. *IJCV* (2022)
48. Zhou, Y., Sahak, H., Ba, J.: Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316* (2023)