Can Synthetic Plant Images From Generative Models Facilitate Rare Species Identification and Classification?

Debajyoti Dasgupta[1], Arijit Mondal[2], Partha P. Chakrabarti[1]
[1]Indian Institute of Technology, Kharagpur [2]Indian Institute of Technology, Patna
debajyotidasgupta6@gmail.com, arijit@iitp.ac.in, ppchak@cse.iitkgp.ac.in

Abstract

In the quest to bridge the gap between the burgeoning capabilities of text-to-image generative models and the pragmatic demands of botanical classification, our study delves into the untapped potential of synthetic images for identifying and differentiating rare plant species. By rigorously evaluating the efficacy of cutting-edge generative models, including open-sourced and proprietary frameworks, we illuminate the advantages and inherent challenges of employing synthetic data in zero-shot and few-shot learning scenarios. Our research demonstrates that the zero-shot method sees a marked improvement of 29% over the pretrained weights, with an average increment of 12%. Furthermore, the few-shot method improves the performance by an additional 31%, with an average increment of 19%, achieving new state-of-the-art classification results on rare flora. Through a comprehensive analysis that spans diverse species and models, we unravel the complexities of synthetic data integration, proposing innovative strategies to harness its full potential for the conservation and study of botanical diversity. This investigation stands at the forefront of combining advanced machine learning techniques with environmental science, paving the way for new advancements in accurately identifying and preserving rare plant species.

1. Introduction

In recent years, the rapid advancement of deep learning techniques has revolutionized the field of computer vision, enabling machines to achieve human-like performance in various image recognition tasks [14, 22, 50]. However, the success of these techniques heavily relies on the availability of large-scale, high-quality labeled datasets [4, 26]. In many real-world scenarios, such as identifying and classifying rare plant species, obtaining sufficient labeled data remains a significant challenge [40, 55]. This scarcity of data limits the applicability of deep learning models in these domains, as they often struggle to generalize well to unseen examples [31, 47]. To address this issue, researchers have turned their attention to generating synthetic data using advanced generative models, such as Generative Adversarial Networks (GANs) [13, 20] and Variational Autoencoders (VAEs) [21, 42]. Synthetic data has shown promising results in augmenting limited datasets and improving the performance of classification models in data-scarce settings [28, 63].

Synthetic data generation for rare plant species classification is a relatively under-explored area despite its potential to alleviate the data scarcity problem. Existing works on synthetic data generation for plant classification have primarily focused on common plant species, where obtaining labeled data is less challenging. For instance, [7] used GANs to generate synthetic images of common plant species and demonstrated improved classification performance when augmenting the training data with the generated images. Similarly, Sood et al. [49] employed VAEs to generate synthetic plant images and showed that the generated images could be used to train classifiers with limited real-world data. However, the effectiveness of these approaches in the context of rare plant species classification remains largely unknown, as the unique characteristics and limited availability of rare plant images pose additional challenges for generative models.

In this paper, we present a comprehensive study on using state-of-the-art text-to-image generation models to improve the classification of rare plant species. We focus on five primary rare plant species: **Rafflesia Arnoldii, Encephalartos Woodii (Cycad), Amorphophallus Titanum (Corpse Flower), Ghost Orchid, and Dracaena Cinnabari (Dragon's Blood Tree)**. By leveraging the power of text-to-image generation [39, 45, 46], we aim to generate high-quality labeled data that captures the intricate details necessary to distinguish these rare species from each other and similar-looking common plant species. Our study is conducted using both open-source models, such as OpenJourney [38], Latent Consistency Model [44], and Stable Diffusion [45], as well as closed-source models like

DallE 3 [34] and Midjourney [30]. We investigate two primary questions: 1) Can synthetic data from generative models effectively improve classification models for rare plant species? 2) Can synthetic data be a feasible source for differentiating rare plants from similar-looking ordinary plants? We focus on zero-shot and few-shot settings, as rare plant species often have limited distinct images available, and the impact of synthetic data is most pronounced in these scenarios [48, 54]. Our investigations are built upon state-of-the-art methods, including CLIP [39], MLLM [57], and ViT Classifier [9], with feature extractors initialized using large-scale pre-trained weights and kept frozen.

Our findings: Our first finding demonstrates that synthetic data can significantly improve the classification results for the five diverse rare plant species studied in the zero-shot setting, where no real-world data is available. We observe an average increase of 2% in top-1 accuracy, with the Dracaena Cinnabari (Dragon's Blood Tree) showing an impressive improvement of 73% when using the CLIP model [39]. To further enhance the effectiveness of synthetic data in this setting, we investigate strategies for increasing data diversity, reducing noise, and enhancing reliability by designing diversified text prompts and measuring the correlation between text and synthesized data using CLIP features [16, 39]. Similar improvements are observed for MLLM [57] and ViT Classifier [9], with an average performance increase of 11% and 31%, respectively, and the ViT Classifier achieving the highest accuracy of 83%.

In the few-shot setting, where a limited number of real images are available, synthetic data provides benefits and helps achieve state-of-the-art performance. However, we observe that the domain gap between synthetic data and downstream task data poses a challenge in further improving the effectiveness of synthetic data on classifier learning. To address this issue, we propose using real images to guide the generation process, reducing domain gaps and enhancing effectiveness. Our experiments show that using a distance-based metric for guidance is crucial, as moving too far from or too close to the real image can negatively impact accuracy.

Finally, when differentiating rare plant species from commonly occurring plant species, we observe that creating detailed descriptions of the rare plants, with the assistance of advanced large language models such as Claude-3 [1], GPT-4 [17], LLaMa [51], and Mistral [53], significantly improves the distinguishing performance of the classifiers. This improvement is particularly evident in the few-shot setting, where generating images with descriptive text, rather than just the class name, results in more variations of plant images that are closer to the rare plant species while adding diversity to differentiate from similar-looking flowers. We observe an average increase of 19% in the accuracy of rare plant and ordinary plant classification in the zero-shot setting and a 28% increase in the few-shot setting.

2. Related Works

Synthetic Data for Image Recognition. Synthetic data has recently gained significant attention in image recognition tasks and can be divided into two categories: synthetic datasets created using traditional simulation pipelines and synthetic images generated by generative models. Synthetic datasets [8, 36, 43] are generated using a conventional pipeline that relies on a specific data source, such as 2D renderings of 3D models or scenes from graphics engines. However, this approach has limitations, including a noticeable difference from real-world data, high storage and transfer costs, and limited data diversity. In contrast, generative models offer a more efficient way to represent synthetic data, producing high-quality, photorealistic images that closely resemble real-world data while requiring less storage space and potentially generating unlimited artificial data. Recent studies have explored the use of synthetic data from generative models for image recognition tasks, such as using class-conditional GANs [2], leveraging StyleGAN's latent code [20, 59], and employing GAN-based generators for unsupervised contrastive representation learning [18]. However, these studies focus on traditional GAN-based models, while our work investigates the best-released text-to-image generation model and its customization capabilities for various downstream label spaces.

Researchers have explored various approaches to generate synthetic images, including 3D modeling, domain randomization, and generative models [10, 19, 37, 52]. These approaches have demonstrated the effectiveness of synthetic data in object detection [52], improving model robustness through diverse backgrounds and lighting conditions [37], action recognition [19], and bridging the gap between synthetic and real data [10]. However, most of these works focus on common objects and scenes, while the application of synthetic data for rare and fine-grained categories, such as plant species, still needs to be explored. Sun et al. [15] studied the readiness of synthetic data from generative models for image recognition tasks, highlighting current limitations and future directions. Our work builds upon these findings and focuses explicitly on the challenges and opportunities of synthetic data for classifying rare plant species.

Text-to-Image Diffusion Models. Text-to-image diffusion models have recently emerged as a powerful approach for generating high-quality synthetic images from textual descriptions. Dhariwal and Nichol [6] introduced the concept of diffusion models for image generation, achieving impressive results in image fidelity and diversity. Nichol et al. [32] further improved the performance of diffusion models by introducing a hierarchical architecture and a new objective function. Ramesh et al. [41] proposed the DALL·E model, which combines a transformer-based language model with a diffusion model to generate images from textual prompts. Saharia et al. [46] introduced the Photorealistic Text-to-Image Diffusion Models (Imagen), which generate high-resolution images from textual descriptions well. Other notable text-to-image diffusion models include Stable Diffusion [45], OpenJourney [38], Midjourney [30], and DALL·E 3 [34]. These models have achieved unprecedented synthesis quality, largely facilitating the development of the AI-for-Art community. However, their potential utilization for high-level tasks, particularly in rare plant species classification, remains largely unexplored. Our work aims to bridge this gap by investigating the effectiveness of synthetic images generated from state-of-the-art text-to-image diffusion models for improving plant species classification.

Synthetic Data and Machine Learning in Plant Classification. The application of machine learning techniques for plant species classification has gained significant attention in recent years, with various studies focusing on automated identification methods [55], benchmarking algorithms on large-scale datasets [12], deep learning-based approaches using leaf images [23], and convolutional neural networks for herbarium specimen identification [3]. However, these works primarily focus on common plant species with abundant labeled data, while the classification of rare plant species remains challenging due to the scarcity of labeled examples. To address this challenge, recent works have explored synthetic data, such as generating synthetic leaf images using GANs for data augmentation [5] and employing style transfer techniques to generate synthetic images of rare plant species [62]. While these works demonstrate the potential of synthetic data for plant species classification, they are limited regarding the diversity and realism of the generated images and the range of plant species considered. Our work advances the state-of-the-art by leveraging powerful text-to-image diffusion models to generate high-quality synthetic images for a diverse set of rare plant species and by proposing strategies to effectively utilize these synthetic images for improving classification performance in data-scarce settings.

3. Can Synthetic Data Improve Plant Recognition Performance?

In the following sections, we explore whether synthetic data can benefit recognition tasks and investigate strategies for effectively using synthetic data to address the classification of rare plants. Our exploration is carried out through the lens of two basic settings with three tasks: synthetic data for improving classification models in the data-scarce setting (i.e., zero-shot and few-shot) (see Sec. 3.1 and Sec. 3.2).

Dataset: Due to the scarcity of standard datasets for rare plant species, we curate a custom dataset by collecting images from various sources, including stock photo websites like Pexels and botanical agencies that provide free photographs. The dataset consists of 250 images in total, with 50 images for each of the five rare flora classes: Rafflesia Arnoldii, Encephalartos Woodii, Amorphophallus Titanum, Ghost Orchid, and Dracaena Cinnabari. This dataset serves as the test set for evaluating the performance of our classification models. For the few-shot classification task, we create a separate dataset with 5 images per class, resulting in a total of 25 images. This dataset is used to investigate the effectiveness of our proposed methods in scenarios where only a limited number of real-world examples are available for each rare plant species. Figure 1 presents a sample of the real images from our dataset, showcasing the visual characteristics of each rare flora class. This data is made publicly available along with the code

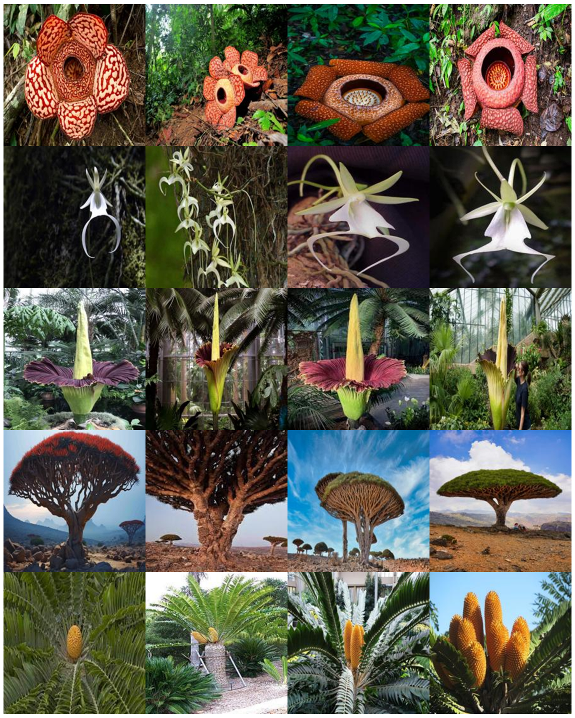

Figure 1. Sample Images of the Real Data collected for Testing.

Common Experimental Setup: For each of the data-scarce setups (zero-shot and few-shot), we present results on three types of classifications: 1) **rare-rare (Task 1)** - in which we classify the five rare plant species shortlisted for this paper from one another. 2) **rare-common (Task 2)** - in which we individually classify each rare plant from its corresponding similar-looking common plant. The list of common plants is selected from the Oxford 102 Flowers dataset [33]. The similarity of the flowers is determined using the average cosine similarity on the features extracted by ResNet-50 [14] from the common flowers and the rare flowers (The final set of flowers is shown in Table 1). 3)

mixed (Task 3) classification is the experiment when we randomly prompt the model with any task from either (1) or (2). Since (1) is a 5-class classification and (2) is a 2-class classification, we exclude the results of this from ViT classifiers which require a fixed set of classes.

Rare Plant	Similar Flower
Rafflesia Arnoldii (RA)	Hibiscus
Encephalartos Woodii (EW)	Sword Lily
Amorphophallus Titanum (AT)	Sweet William
Ghost Orchid (GO)	Petunia
Dracaena Cinnabari (DC)	Cautleya Spicata

Table 1. Rare plants and their similar common flowers from the Oxford-102 dataset.

Model Setup for Data-scarce Image Classification: As CLIP [39] and Multimodal Large Language Models (MLLM) [25, 27, 29, 57] are the state-of-the-art approaches for zero-shot learning, we conduct our study for zero-shot and few-shot settings upon pre-trained CLIP and MLLM models, aiming to understand synthetic data upon strong baselines better. There have been a few attempts at better tuning pre-trained CLIP for data-scarce image classification, such as CoOp [61], CLIP Adapter [11], and Tip Adapter [58], where the image encoder is frozen for better preserving the pre-trained feature space. We present the results with the Tip-Adapter methodology, which provided both easy implementation and the best results in combination with the contrastive training methodology. We use the default qLoRA adaptation script supplied by the model authors of MLLMs with their default values for fine-tuning MLLM models.

CLIP and MLLM allow us to classify the data into any number of unspecified classes; hence, it is easier to perform mixed classification experiments on these models. However, the average performance of the models cannot beat the performance of the classifiers with a fixed number of class labels. Hence, for the test to be complete, we also included the results of the experiments with ViT models [9] after pretraining them on the Oxford 102 Flowers dataset.

Here, we adopt a simple inferencing method for classifiers, a baseline method introduced in Wortsman et al. [56]. Concretely, for a k-way classification, we input the class names C = $\{c_1, ..., c_k\}$ with prompt s_i = “a photo of a $\{c_i\}$” into the text encoder h of CLIP to obtain the text features $h(s_i)$. Then, the text features $h(s_i)$ could be used to construct classifier weights $W \in R^{d\times k}$, where d is the dimension of text features. Finally, we combine the image encoder g with the classifier weights W to obtain a classification model $f(x) = g(x)^T \cdot W$. This is a standard approach where we primarily select the class whose text feature aligns very closely with the image feature. In the case of MLLM, we use prompt engineering where we simply include the image along with all the options of class prompts s_i, and ask the model to select one class. Due to the simplicity of the modeling in MLLM in the form of prompts, we can also include SoTA visual question-answering (VQA) models like BLIP-2 [24] in the loop as well.

3.1. Zero-Shot Image Recognition with Synthetic Images from Generative Models

We aim to investigate how synthetic data benefit zero-shot tasks and explore strategies for better leveraging synthetic data for zero-shot learning.

Zero-shot Image Recognition. We study the inductive zero-shot learning setting where no real training images of the target categories are available. CLIP models are pre-trained with large-scale image-caption pairs, and the similarities between paired image features (from an image-encoder g) and text features (from a text-encoder h) are maximized during pre-training. The pre-trained feature extractor can then be used to solve zero-shot tasks, where given an image, its features from g are compared with text features of different classes from h, and the image is further assigned to the class that has the most significant similarity in the CLIP text-image feature space. In the case of MLLM, we explicitly ask the model to output the name of an individual class. If the name is not returned, we will perform text cleaning and assign it the best matching label. For ViT Classifiers, this class determination is relatively trivial as we already set the number of classes during classifier initialization. Hence, the output of the model is the predicted class.

Synthetic Data for Zero-shot Image Recognition. Though CLIP models exhibit zero-shot solid performance thanks to the large-scale vision-language dataset for pre-training, there are still several shortcomings when the model is deployed for a downstream zero-shot classification task, which may be attributed to unavoidable data noise in CLIP’s pre-training data or the label space mismatch between pre-training and the zero-shot task. Hence, we study whether synthetic data can be used to better adapt CLIP models for zero-shot learning with a given label space for a zero-shot task. Similarly, the noise in the training set of MLLM can also hinder its ability to classify images, and we investigate if its ability to classify rare flora and distinguish it from its common counterparts is enhanced with synthetic data. In the case of the ViT classifier, we tap into the ability of synthetic data to fine-tune the classifier, which has been pre-trained on a similar domain (Oxford Flowers) [33].

Generating Synthetic Data. Given a pre-trained text-to-image generation model to synthesize novel samples, the **primary (P)** strategy is to use the label names of the target categories to build the language input and generate a corresponding image. Then, the paired label names and synthe-

Models	Rare-Rare (Task-1)		Rare-Common (Task-2)		Mixed (Task-3)	
	w/o SYN	+SYN	w/o SYN	+SYN	w/o SYN	+SYN
CLIP Models						
RN50	0.20	0.57 (+0.37)	0.47	0.63 (+0.16)	0.35	0.63 (+0.28)
RN101	0.24	0.60 (+0.36)	0.51	0.69 (+0.18)	0.39	0.74 (+0.25)
RN50x4	0.20	0.58 (+0.38)	0.48	0.65 (+0.17)	0.32	0.64 (+0.32)
RN50x16	0.23	0.61 (+0.38)	0.50	0.66 (+0.16)	0.35	0.66 (+0.31)
RN50x64	0.26	0.64 (+0.38)	0.51	0.69 (+0.18)	0.38	0.72 (+0.34)
ViT-L/14	0.21	0.55 (+0.36)	0.49	0.62 (+0.13)	0.31	0.65 (+0.34)
ViT-B/16	0.22	0.57 (+0.35)	0.49	0.64 (+0.15)	0.33	0.69 (+0.35)
ViT-B/32	0.25	0.60 (+0.35)	0.52	0.72 (+0.20)	0.38	0.75 (+0.37)
MLLM						
llava-v1.6-vicuna-7b	0.24	0.45 (+0.21)	0.47	0.66 (+0.19)	0.31	0.58 (+0.27)
llava-v1.6-mistral-7b	0.25	0.52 (+0.27)	0.53	0.69 (+0.16)	0.35	0.61 (+0.26)
deepseek-vl-7b-chat	0.19	0.34 (+0.15)	0.48	0.62 (+0.14)	0.22	0.49 (+0.27)
blip-2-vqa-base	0.30	0.62 (+0.32)	0.56	0.64 (+0.08)	0.30	0.58 (+0.28)
ViT Classifiers						
vit-base-patch16-224	0.42	0.76 (+0.34)	0.52	0.81 (+0.29)	-	-
vit-base-patch16-384	0.46	0.78 (+0.32)	0.55	0.83 (+0.28)	-	-
vit-base-patch32-384	0.45	0.69 (+0.24)	0.54	0.81 (+0.27)	-	-
vit-large-patch16-224	0.44	0.75 (+0.31)	0.56	0.83 (+0.27)	-	-
vit-large-patch16-384	0.48	0.74 (+0.26)	0.57	0.85 (+0.28)	-	-
vit-large-patch32-384	0.45	0.71 (+0.26)	0.57	0.79 (+0.22)	-	-

Table 2. Performance (Accuracy) of all the models for the tasks defined in experimental setup in zero-shot data-setting. Results presented here include Enhanced Description (**ED**) and Feature Filtering (**FF**) for generating synthetic data.

sized data are employed to train the classifier with 50% of the feature extractor frozen.

Enriching Diversity. Only using the label names as inputs might limit the diversity of synthesized images and cause bottlenecks in validating the effectiveness of synthetic data. Hence, we leverage an off-the-shelf text-generation LLM model **Claude-3** [1] **(OPUS)** to increase the diversity of language prompts and the generated images, namely **Enhanced Description (ED)**, hoping to unleash the potential of synthesized data better. Concretely, we input the label name of each class into the LLM model, which generates diversified sentences containing the class names as language prompts for the text-to-image generation process. For example, if the class label is "Rafflesia Arnoldii," then the enhanced descriptive prompt from the model could be *"Rafflesia flower in full bloom surrounded by huge thick, fleshy oval lobes with pointed tips, entire flower is a deep red-brown, speckled with white spots, giving it a speckled appearance, set in a lush green rainforest, Vibrant colors, Macro lens, f/4 aperture, ISO 200, naturally diffused light"*. The enhanced text descriptions introduce rich context descriptions for more variants of the image.

Reducing Noise and Enhancing Robustness. It's unavoidable that the synthesized data may contain low-quality samples. This is even more severe in the setting with language enhancement as it may introduce undesired items into language prompts. Hence, we introduce a **Feature Filter (FF)** strategy to rule out these samples. Specifically, CLIP extracted feature distance is used to assess the quality of synthesized data, and the low-confidence ones are removed. We define a normalized distance parameter α based on which the images with a distance more than α are filtered out. This α parameter is hyperparameter tuned to find the optimal value. (see Figure 7, 8, 9 in supplementary)

Significant Results: 1) zero-shot classification results on the three tasks mentioned previously; 2) study of synthetic data diversity; 3) study of synthetic data reliability; 4) study of the effect of distance parameter (α) on the accuracy of the model.

Synthetic data can significantly improve the performance of zero-shot learning. Our main studies in zero-shot settings are conducted with CLIP-RN50 (ResNet-50) [14] and CLIP ViT-B/L (Visual Transformer) [9], LLaVa [27], Blip-2 [24], and ViT-B/L-16/32 [9] family of models, and we report results with our best strategy of **ED+FF**. The image generation model used for these results is the Stable Diffusion XL (SDXL) model since it is open-sourced, efficient in image generation (see Supplementary Section 8 for ablation), and the experiment will be easy to replicate. As shown in Table 2, on diverse zero-shot rare flora image classification, we achieve a remarkable average gain of 38% for CLIP, 27% for MLLM, and 34% for ViT-Classifier family in terms of top-1 accuracy.

Flora	w/o SYN	+SYN
Rafflesia Arnoldii	0.45	0.64 (+0.19)
Encephalartos Woodii	0.18	0.35 (+0.17)
Amorphophallus Titanum	0.12	0.54 (+0.42)
Ghost Orchid	0.26	0.38 (+0.12)
Dracaena Cinnabari	0.00	0.70 (+0.70)

Table 3. Zero shot Image classification accuracy improvement of CLIP ViT-B/16 on Task (1) Rare-Rare classification with ED+FF

If we observe the improvement of the performance of the CLIP classifier on different flora from Table 3, we achieve the most significant performance boost of 70% for Dragon's Blood Tree in top-1 accuracy. We notice that the performance gain brought by synthetic data varies differently across flora, which is mainly related to SDXL's training data distribution. The training data distribution of the text-to-image generation model SDXL would exhibit bias and produce different domain gaps with different flora.

Flora	CLIP	P	ED	ED+FF
RA	0.33	0.51 (+0.18)	0.43 (+0.10)	0.66 (+0.33)
EW	0.54	0.59 (+0.05)	0.61 (+0.07)	0.63 (+0.09)
AT	0.57	0.62 (+0.05)	0.61 (+0.04)	0.64 (+0.07)
GO	0.46	0.53 (+0.07)	0.60 (+0.14)	0.69 (+0.23)
DC	0.51	0.55 (+0.04)	0.57 (+0.06)	0.62 (+0.11)

Table 4. Zero shot classification accuracy improvement of CLIP ViT-B/16 on Task (2) (ablation study of Primary Strategy (P), Enhanced Description (ED) and Feature Filtering (FF))

By introducing more linguistic context into the text input, ED helps increase the diversity of synthetic data. As shown in Table 4, **ED** can achieve additional performance gains upon P in most cases, demonstrating ED's efficacy and the importance of synthetic data diversity for zero-shot classification. Enhanced Description (**ED**) achieves performance gain over P in most cases (2% $\uparrow$ for Cycad (**EW**) and 7% $\uparrow$ for Ghost Orchid (**GO**)).

Reliability matters. While ED could help increase the diversity of synthetic data, it also introduces the risks of noisy samples. Observed in Table 4, ED sometimes even brings performance drops compared with P (8% $\downarrow$ for Rafflesia (**RA**) and 1% $\downarrow$ for Corpse Flower (**AT**)), which may be attributed to the noise introduced by enhanced language prompts, e.g., the sentence extended from the class name word may contain other class names or confusing objects. Fortunately, with FF to filter out unreliable samples, **ED+FF** yields consistent improvement.

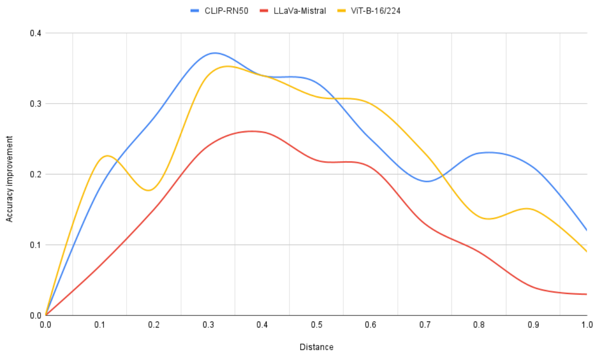

Figure 2. Average Accuracy improvement of different model families with varying filtering distance parameter (P+ED+FF)

The optimal distance for feature filtering is nearly consistent. We prove this from the empirical results in Figure 2 that the optimal distance for filtering the synthetic dataset to achieve optimal accuracy remains nearly consistent. We use the **1 - `cosine-similarity`** as a normalized distance metric. Based on the results, we observe that the value of $\alpha = 0.3$ yields the best average accuracy improvement in nearly most of the family of models (The average accuracy improvement in the image denotes the improvement in accuracy of all the five rare flora averaged weighted on their sample size).

Summary. Current synthetic data from text-to-image generation models could bring significant performance boosts for a wide range of zero-shot image classification tasks and is readily applicable with carefully designed strategies such as large-scale pre-trained models. Diversity and reliability matter for synthetic data when employed for zero-shot tasks. A fixed hyper-parameter tuned distance metric for filtering yields optimal performance.

3.2. Few-Shot Image Recognition with Synthetic Images from Generative Models

Here, we explore the effectiveness of synthetic data for few-shot tasks and investigate how synthetic data impacts performance as more shots are included. Additionally, we design effective strategies to leverage synthetic data better.

Few-shot Image Recognition. We adopt CLIP, MLLM, and ViT as the models for few-shot image recognition due to their state-of-the-art performance [39]. However, for the following experiments, we make use of the CLIP RN50x64, LLaVA-Mistral-7B, BLIP-2, and ViT-L/16@224 models, since these models have the best performance in their family from Table 2. We study how to tune the classifier weights with synthetic data. In an N-way M-shot case, we are given M real images of each test class, where $M \in \{1, 2, 3, 4, 5\}$ in our experiments. With $N \times M$ training samples, we hope to achieve good performance on a test set of the N classes.

Synthetic Data for Few-shot Image Recognition. While there have been a few attempts to study how to better adapt CLIP models for few-shot tasks [58, 60, 61], they all focus on the model optimization level, and very few have explored from the data level. Here, we systematically study whether and how synthetic data can solve few-shot image recognition tasks. With the experience from synthetic data for zero-shot tasks, we adopt the best strategy (i.e., **ED+FF**) as the **primary strategy (P)** in the zero-shot setting. Further, as the few-shot real samples can provide helpful information on the data distribution of the classification task, we propose two new strategies leveraging the in-domain few-shot real data for better using synthetic data: 1) **Real Feature Filtering (RFF):** given synthetic data of one class c, we use the features of few-shot real samples to filter out synthetic images whose features are far from the real sample features by at least distance parameter α (we use the cosine similarity here as well). Since both the real data and synthetic data are in the image format, for feature extraction, we make use of Resnet-101 [14] instead of CLIP; 2) **Real Image Guided Generation (RIGG):** we use the few-shot real samples as guidance to generate synthetic images where the few-shot real samples (added noise) replace the random noise at the beginning of the generation to guide the diffusion process (see Supplementary Section 9 for implementation). We also use the Stable Diffusion XL (**SDXL**) model for the following experiments.

Significant Results: 1) few-shot classification results on the three classification tasks; 2) ablation study of training strategy; 3) ablation study of synthetic data generation strategy; 4) ablation study of distance parameter α.

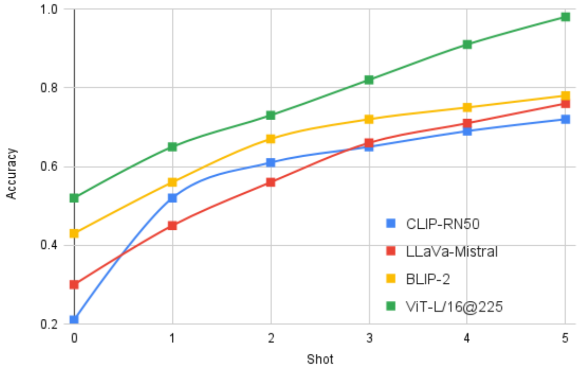

Figure 3. Performance (Accuracy) of the classification models on the Few-shot image recognition Task-1 (Rare-Rare Classification).

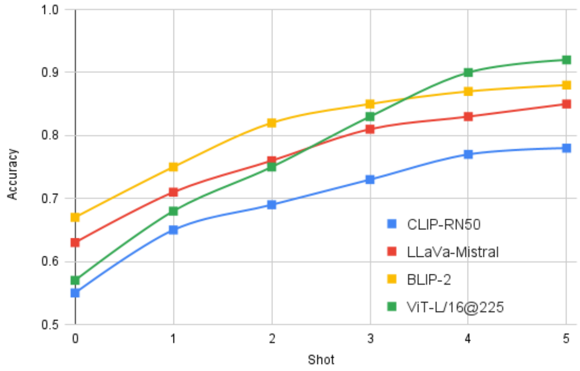

Figure 4. Performance (Accuracy) of the classification models on the Few-shot image recognition Task-2 (Rare-Common Classification).

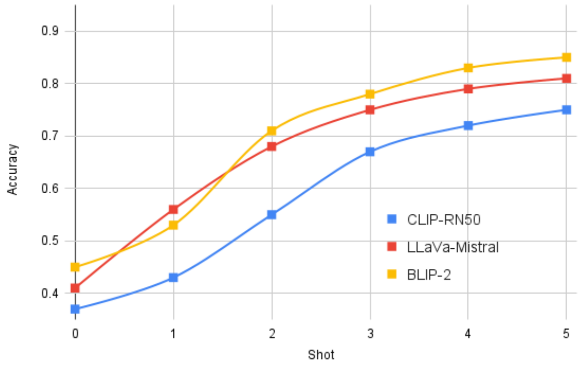

Figure 5. Performance (Accuracy) of the classification models on the Few-shot image recognition Task-3 (Mixed Classification).

Synthetic data can boost few-shot learning, and the positive impact of synthetic data will gradually diminish with the increase of real data shots. Figure 3, 4 and 5 present the result of the model fine-tuned with **P+ED+RFF+RIGG**. As shown in Figure 3, with only a few shots of real images for training, our fine-tuned models have significantly improved performance compared to their zero-shot counterparts. Figures 4 and 5, which denote the results of the fine-tuned models on Task-2 and Task-3, show similar results. With the help of generated synthetic data, models achieve noticeable performance gains upon pre-trained weights, and ViT-L/16@224 achieves a new state-of-the-art few-shot learning performance across different flora with almost perfect prediction accuracy of 98% for Task 1. We argue that synthetic data could help address the problem of insufficient data to boost performance for data-scarce few-shot classification. However, we notice that the boost from synthetic data gradually diminishes as the real shot number increases for other models like CLIP, BLIP, and LLaVa. Due to the relatively low variety of images present in the dataset of rare plant species, the correlation among the real data is high, and it is also transferred to the **RFF+RIGG** generated synthetic data. We thus observed a good boost in performance, which slowly died down with increased shots.

Mix Training fits few-shot learning with synthetic data. Now that we have two parts of data, i.e., few-shot real data and synthetic data, we could either 1) phase-wise train on each part of data with two training phases or 2) adopt

M-shot	Phase-wise		Mix
	syn → real	real → syn	training
1	0.61	0.63	**0.65**
2	0.64	0.67	**0.69**
3	0.68	0.70	**0.73**
4	0.71	0.72	**0.77**
5	0.72	0.75	**0.78**

Table 5. Mix training works better for few-shot classification on Task 2 (rare-common classification) with CLIP model.

mix training that simultaneously utilizes two parts of data to update the model in each iteration. We provide the results in Table 5: we study the Task 2 (Rare-Common Classification) performance and use synthetic data generated from the **RIGG** method; under different shot number settings, mix training performs consistently better than two phase-wise strategies. Mixed training could help learn better classifiers since each part could function as a regularization for the other: synthetic data help alleviate instabilities brought by limited real samples, and real data help address synthetic data's noise and domain gap.

Flora	**P+ED**	**RFF**	**RFF+RIGG**
RA	0.64	0.76 (+0.12)	0.83 (+0.19)
EW	0.68	0.75 (+0.07)	0.81 (+0.13)
AT	0.65	0.66 (+0.01)	0.72 (+0.07)
GO	0.54	0.79 (+0.25)	0.85 (+0.31)
DC	0.61	0.72 (+0.11)	0.78 (+0.17)

Table 6. Few shot classification accuracy improvement of CLIP ViT-B/16 on Task (2) Rare-Common classification (ablation)

Employing real data as guidance can alleviate domain differences and boost performance. We compare three strategies of synthetic data generation for few-shot tasks. As shown in Table 6, both **RFF** and **RIGG** provide performance gains upon P, the primary strategy in the zero-shot setting. This demonstrates the importance of utilizing domain knowledge from few-shot images to prepare synthetic data. Further, **RIGG** significantly outperforms **RFF**, yielding the best performance. This shows utilizing real data as guidance for the diffusion process helps reduce the domain gap. Combining **RFF** and **RIGG** yeilds the best performance as from results in Table 6.

Stability of the distance parameter (α) holds for the few-shot setting. Lastly, we investigate the distance parameter (α) for our few-shot settings with synthetic data. As shown in Figure 6, for mixed data (Real + Synthetic), the α value remains consistent for all the models and is slightly different from the zero-shot value ($\alpha = 0.43$). Thus, we can empirically conclude that the availability of few-shot real data is a characteristic of distance parameter α.

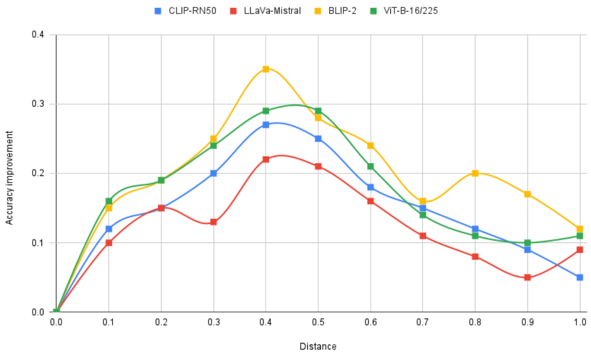

Figure 6. Average Accuracy improvement of different models with varying filtering distance parameter (P+ED+RIFF+RIGG)

Summary. Synthetic data from text-to-image generation models could readily benefit few-shot learning and achieve a new state-of-the-art few-shot classification performance with the strategies we present in this paper. However, the positive impact of synthetic data will diminish as more shots of real data are available. We also show that the distance parameter remains consistent across the model families with the availability of real data.

4. Conclusion

This study explores synthetic images generated from text-to-image diffusion models to improve the classification of rare plant species in both zero-shot and few-shot learning scenarios. We evaluate the efficacy of synthetic data generated using diffusion models, focusing on five rare plant species. In the zero-shot setting, synthetic data significantly improves classification accuracy, with the proposed strategies to increase data diversity, reduce noise, and enhance reliability. Similar improvements are observed for MLLM and ViT classifiers. Synthetic data helps achieve state-of-the-art performance in the few-shot setting, but the domain gap between artificial and real data poses a challenge. We propose using real images to guide the generation process and demonstrate that creating detailed descriptions of rare plants using language models significantly improves classification performance. The study highlights the potential of combining machine learning with environmental science to enhance the conservation and research of botanical diversity, offering a novel approach to improving model robustness when real-world data is limited. Our study demonstrates that synthetic data from text-to-image generation models can readily benefit zero-shot and few-shot learning for rare plant classification. However, the positive impact of synthetic data diminishes as more real data becomes available. We also show that the optimal distance parameter for filtering synthetic data remains consistent across models.

References

[1] Anthropic. Claude 3: A large language model for task-oriented dialogue. https://www.anthropic.com/news/claude-3-family, 2023. 2, 5

[2] Etienne Besnier, Andrei Bursuc, David Picard, and Alexandre Briot. Dataset distillation by matching training trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 1044–1045, 2020. 2

[3] Jose Carranza-Rojas, Herve Goeau, Pierre Bonnet, Erick Mata-Montero, and Alexis Joly. Going deeper in the automated identification of herbarium specimens. *BMC evolutionary biology*, 17(1):1–14, 2018. 3

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

[5] Zehui Deng, Zhengbing Wang, Weikuan Gong, Jianxiong Zhang, and Sheng Li. A gan-based data augmentation strategy for plant leaves classification. *Remote Sensing*, 13(20): 4041, 2021. 3

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[7] M Dileep and PN Pournami. Data augmentation using generative adversarial network for leaf disease detection. *Evolutionary Intelligence*, 14:13–23, 2020. 1

[8] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021. 2, 4, 5

[10] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. On the importance of visual context for data augmentation in scene understanding. In *arXiv preprint arXiv:1809.02492*, 2018. 2

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 4

[12] Herv'e Go"eau, Pierre Bonnet, and Alexis Joly. Plantnet participation at lifeclef2018 plant identification task. *CLEF working notes*, 2125:1–6, 2018. 3

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2014. 1

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 5, 7

[15] 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?, 2023. 2

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clues: A benchmark for open-ended language-guided image understanding. *arXiv preprint arXiv:2112.00327*, 2021. 2

[17] OpenAI Inc. Gpt-4 technical report. *https://arxiv.org/abs/2303.08774*, 2023. 2

[18] Ali Jahanian, Lucy Chai, and Phillip Isola. Generative models as a data source for multiview representation learning. In *International Conference on Learning Representations*, 2021. 2

[19] Amlan Kar, Aayush Prakash, Ming-Yu Liu, Eric Cameracci, Justin Yuan, Matt Rusiniak, David Acuna, Antonio Torralba, and Sanja Fidler. Meta-sim: Learning to generate synthetic datasets. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4551–4560, 2019. 2

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2019. 1, 2

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[22] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012. 1

[23] Sue Han Lee, Chee Seng Chan, Paul Wilkin, and Paolo Remagnino. Deep-plant: Plant identification with convolutional neural networks. *2015 IEEE international conference on image processing (ICIP)*, pages 452–456, 2015. 3

[24] Junnan Li, Ramprasaath R Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1102–1112, 2022. 4, 5

[25] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. 4

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 4, 5

[28] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya. Generative adversarial networks for image and video synthesis: Algorithms and applications. *Proceedings of the IEEE*, 109(5):839–862, 2021. 1

[29] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, Yaofeng Sun, Chengqi Deng, Hanwei Xu, Zhenda Xie,

and Chong Ruan. Deepseek-vl: Towards real-world vision-language understanding, 2024. 4

[30] Inc. Midjourney. Midjourney. https://www.midjourney.com/, 2023. 2, 3

[31] Sharada P Mohanty, David P Hughes, and Marcel Salathe. Using deep learning for image-based plant disease detection. *Frontiers in plant science*, 7:1419, 2016. 1

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021. 2

[33] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 3, 4

[34] Inc. OpenAI. Dall·e 3. https://openai.com/product/dall-e-3, 2023. 2, 3

[35] Luca Paoletti, Lorenzo Lella, Antonio Moro, and Antonio Caputo. Efficient clip: How to improve clip with knowledge distillation and model compression. *arXiv preprint arXiv:2209.10267*, 2022. 6

[36] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. In *arXiv preprint arXiv:1710.06924*, 2017. 2

[37] Aayush Prakash, Shaad Boochoon, Mark Brophy, David Acuna, Eric Cameracci, Gavriel State, Omer Shapira, and Stan Birchfield. Structured domain randomization: Bridging the reality gap by context-aware synthetic data. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7249–7255. IEEE, 2019. 2

[38] Inc. PromptHero. Openjourney. *https://openjourney.art/*, 2023. 1, 3

[39] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1, 2, 4, 6

[40] Sivaramakrishnan Rajaraman, Sameer K Antani, Mahdieh Poostchi, Kamolrat Silamut, Md A Hossain, Richard J Maude, Stefan Jaeger, and George R Thoma. Understanding the learned behavior of customized convolutional neural networks toward malaria parasite detection in thin blood smear images. *Journal of medical imaging*, 5(3):034501, 2018. 1

[41] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2

[42] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014. 1

[43] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *European conference on computer vision*, pages 102–118. Springer, 2016. 2

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022. 1

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2022. 1, 3, 6

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed K Seh Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 1, 3

[47] Rabia Sharif, Farrukh Amin, Aneela Zafar, Farhan Riaz, and Nadeem Anjum. Cnn based heuristic feature selection for hyperspectral image classification. *Journal of applied remote sensing*, 8(1):083646, 2014. 1

[48] Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017. 2

[49] Mohit Sood, Khyati Mohiuddin, and Parvaiz Ahmad. A novel deep learning approach for plant disease detection. *International Journal of Advanced Computer Science and Applications*, 12(12):98–104, 2021. 1

[50] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015. 1

[51] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[52] Jonathan Tremblay, Aayush Prakash, David Acuna, Mark Brophy, Varun Jampani, Cem Anil, Thang To, Eric Cameracci, Shaad Boochoon, and Stan Birchfield. Training deep networks with synthetic data: Bridging the reality gap by domain randomization. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 969–977, 2018. 2

[53] Rahul Varshney, Tianyi Sun, Dilip Arumugam, Kevin Yang, Kaylee Farrell, Yitian Wang, Shibing Hu, Lu Ji, David Gunter, Kaixin Shi, et al. Mistral: An open-source large language model with multi-exit layers for deployment efficiency.
urlhttps://github.com/stanford-crfm/mistral, 2023. 2

[54] Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29, 2016. 2

[55] Jana W"aldchen and Patrick M"ader. Plant species identification using computer vision techniques: A systematic literature review. *Archives of computational methods in engineering*, 25:507–543, 2018. 1, 3

[56] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi,

Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2022. 4

[57] Jiayang Wu, Wensheng Gan, Zefeng Chen, Wan Shicheng, and Philip Yu. Multimodal large language models: A survey. In *IEEE International Conference on Big Data (Big Data)*, 2023. 2, 4

[58] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2022. 4, 6

[59] Yuxuan Zhang, Huan Ling, Jun Gao, Kangxue Yin, Jean-Francois Lafleche, Adela Barriuso, Antonio Torralba, and Sanja Fidler. Datasetgan: Efficient labeled data factory with minimal human effort. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10145–10155, 2021. 2

[60] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Cocoop: Conditional prompt learning with contrastive prompts. *arXiv preprint arXiv:2203.05557*, 2022. 6

[61] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2236–2245, 2022. 4, 6

[62] Huiling Zhu, Xuesong Yao, Yawen Sun, Jiang Zhang, and Dengsheng Li. Few-shot fine-grained plant image classification based on style transfer and multi-attention fusion. *Knowledge-Based Systems*, 241:108244, 2022. 3

[63] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *arXiv preprint arXiv:1703.10593*, 2021. 1