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

Supplementary Material

5. Limitations and Future Research Directions

1. **Limited number of rare plant species:** Our study primarily focuses on five rare plant species. The methods presented in this study should be evaluated using a more extensive set of rare flora to demonstrate broader applicability. However, we theorize and propose initial work on using rare plants for classification. Given the constraint on the number of resources available, we collected data and conducted experiments on these five specific flora, which are well-known rare species.
2. **Small real-world labeled test set:** The real-world labeled dataset used for testing is relatively small, with only 250 images (50 per class). A more extensive test set would provide a more robust evaluation of the proposed techniques. However, each of these 250 samples used in this paper is manually validated for sanity.
3. **Few-shot learning limited to 5 shots:** A maximum of 5 real images per class are used for the few-shot learning experiments. The effectiveness of the methods with more real shots (e.g. 10, 20) has yet to be explored. However, finding distinct images for rare plants that are openly available is already challenging, and in our experiment results, we see that within the five real photos, we start seeing rate performance improvement dropping.
4. **GPU constraints limiting experiment scale:** GPU constraints likely limited the scale of run experiments. Additional model architectures, prompt engineering strategies, and hyperparameter settings could be explored with more computational resources. While additional computational resources could enable a large-scale empirical evaluation, the current experiments demonstrate significant improvements from synthetic data and provide valuable insights to guide future research.
5. **Focus on rare plant classification:** The paper focuses on rare plant classification as a challenging and socially significant fine-grained visual classification problem. While extending the techniques to additional domains would demonstrate a broader impact, the current study provides a meaningful proof-of-concept in a specific application area.
6. **Lack of human evaluation of synthetic images and feedback:** We agree that human assessment of the realism and diversity of the generated synthetic plant images would provide additional insight. However, the significant performance improvements achieved using the synthetic data demonstrate its effectiveness for the classification task. Subjective human judgments, while interesting, are optional to validate the approach. However, human (exceptionally expert botanist) feedback-guided image generation may yield new results.

6. Ethical Considerations

1. **Environmental conservation:** The primary motivation behind this study is to improve the classification of rare plant species, which can ultimately aid in their preservation. By developing techniques that enable accurate identification of rare flora from limited real-world data, our paper contributes positively to environmental conservation efforts. However, it is essential to ensure that the synthetic data generation process does not inadvertently promote the collection of rare plant specimens from the wild, which could harm these vulnerable species.
2. **Data privacy:** This paper uses publicly available data sources for training and evaluation, which mitigates potential concerns around data privacy. However, if the techniques were to be applied to datasets containing sensitive information, such as the location of rare plant populations, it would be crucial to ensure that appropriate data privacy measures are in place to protect this information from misuse.
3. **Bias and fairness:** The paper does not explicitly address bias and fairness issues in the synthetic data generation process. While the focus on rare plant species inherently deals with underrepresented categories, it is essential to consider whether the synthetic data generation process could inadvertently introduce biases, such as overrepresenting certain species or geographic regions. Future work could benefit from a more detailed analysis of the diversity and representativeness of the generated synthetic data.
4. **Potential misuse:** The techniques presented in the paper could be misused to generate misleading or deceptive content related to rare plant species. For example, synthetic images could falsely claim the discovery of new rare species or misrepresent the appearance or distribution of known species. While we, the authors, cannot fully control how others use their techniques, it is essential to raise awareness of these potential misuse cases and to promote responsible use of the technology.
5. **Intellectual property:** The paper builds upon existing open-source models and datasets, which helps to ensure the accessibility and reproducibility of the research. However, if the techniques were applied to proprietary datasets or used in commercial applications, it would be

necessary to consider intellectual property rights carefully and ensure that appropriate licenses and attributions are in place.

6. **Societal impact:** The paper has the potential to positively impact society by contributing to the conservation of rare plant species. However, it is essential to consider the broader societal implications of the technology, such as the potential impact on jobs and expertise in botanical research. It is necessary to ensure that these techniques are developed in collaboration with domain experts and that the technology is used to augment, rather than replace, human expertise.

7. Visualization of Synthetic Images Generated

In this section, we present a visual analysis of the synthetic images generated by our proposed approach. The purpose of this visualization is to provide a qualitative assessment of the generated images and to offer insights into the effectiveness of our synthetic data generation process.

7.1. Visualization of Zero-Shot Synthetic Images

Figure 7. Zero-shot images generated with guidance scale set to 8

Figure 7, 8 and 9 showcases a selection of synthetic images generated by our approach in the zero-shot setting. In this setting, the model generates images of rare plant species without real-world examples to guide the generation process. The images are generated solely based on textual de-

Figure 8. Zero-shot images generated with guidance scale 12

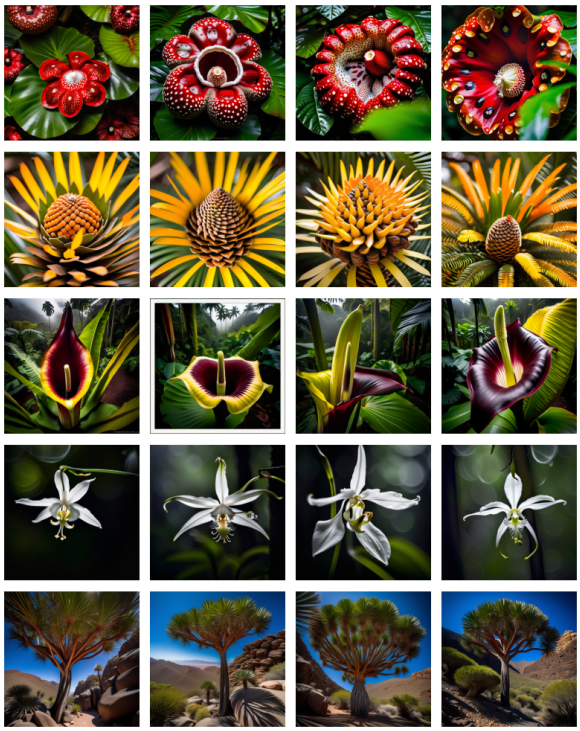

Figure 9. Zero-shot images generated with guidance scale 18

scriptions of the target species, fed into the text-to-image generation model.

As can be seen from the figure, the generated images exhibit a high degree of visual fidelity and realism. The photos accurately capture the distinctive morphological characteristics of each rare plant species, such as the shape and color of the flowers, leaves, and stems. For example, the synthetic images of Rafflesia Arnoldii (first row) depict the large, reddish-brown petals and the central disk characteristic of this species. Similarly, the photos of Amorphophallus titanum (third row) accurately represent the tall, cylindrical spadix and the large, frilled spathe typical of this species.

However, it can also be observed that some of the generated images lack the fine-grained details and textures in real-world images. For instance, the synthetic images of Encephalartos Woodii (second row) capture the overall shape and arrangement of the leaves, but the individual leaflets appear somewhat simplified and need more intricate venation patterns visible in natural specimens. This limitation is likely due to the challenges inherent in generating highly detailed images from textual descriptions alone, without any visual reference.

Despite these limitations, the synthetic images generated by our approach in the zero-shot setting demonstrate the potential of using textual descriptions to guide the generation of realistic images of rare plant species. The generated images capture the essential visual characteristics of each species and provide a valuable resource for training classification models without real-world data. As we refine our approach and explore more advanced text-to-image generation techniques, we expect the quality and diversity of the generated images to improve further, enabling even more effective zero-shot learning for rare plant classification.

7.2. Visualization of Few-Shot Synthetic Images

In the few-shot setting, our approach leverages a small number of real-world images to guide the generation of synthetic images. This is achieved through the Real Image Guided Generation (RIGG) technique, which uses real images as a starting point for the image generation process. By conditioning the generation on real examples, we aim to improve the realism and diversity of the generated images while still benefiting from the ability of the text-to-image model to generate novel variations.

Figure 10 presents a comparison of real images (first row) and synthetic images generated by our RIGG approach (second row) for each of the five rare plant species. The synthetic images shown here are generated without applying the Real Feature Filtering (RFF) technique, which allows us to visualize the types of images filtered out by this process.

As can be seen from the figure, the synthetic images generated by RIGG exhibit a high degree of visual similarity to the real images, capturing the key morphological features and overall appearance of each species. For example, the synthetic images of Encephalartos Woodii (second column) closely resemble the actual images regarding the leaves' shape, size, and arrangement. Similarly, the synthetic images of Amorphophallus titanum (third column) accurately depict the distinctive inflorescence and the mottled patterns on the spathe.

Figure 10. Comparison of real images (first row) and synthetic images generated by our RIGG approach without RFF (second row) for each of the five rare plant species: (a) Rafflesia Arnoldii, (b) Encephalartos Woodii, (c) Amorphophallus Titanum, (d) Ghost Orchid, and (e) Dracaena Cinnabari.

However, it can also be observed that some of the synthetic images contain artifacts or inconsistencies that deviate from the real examples. For instance, the first two synthetic images of Rafflesia Arnoldii (first column) exhibit distortions in the shape and color of the petals, which are not present in the real images. These artifacts are likely due to the challenges of accurately capturing the fine-grained details and textures of the real images in the synthetic generation process.

To address these issues, our approach employs the RFF technique to filter out synthetic images that are too dissimilar from the real examples. By setting a threshold on the distance between the real and synthetic images in feature space, RFF ensures that only the most realistic and consistent synthetic images are retained for training the classifica-

Generative Model	CLIP ViT-B/32	LLaVa-Mistral	BLIP-2	ViT-L/16@384
Stable Diffusion XL	**0.72**	**0.69**	**0.64**	**0.85**
Openjourney-v4	0.71	0.73	0.65	0.81
Latent Consistency Model	0.69	0.65	0.62	0.77
Midjourney	**0.79**	**0.75**	**0.73**	**0.89**
Dall∘E 3	0.70	0.65	0.61	0.83

Table 7. Performance comparison of different generative models (Propreitary + Open Source) and classifiers.

tion models. In the case of Rafflesia Arnoldii, the first two synthetic images shown in Figure 10 would be filtered out by RFF, as their distance from the real images exceeds the specified threshold α. This filtering process helps improve the quality and reliability of the synthetic data while allowing for a diverse range of generated images that capture the essential characteristics of each species.

Overall, the visualization of synthetic images generated by our RIGG approach in the few-shot setting demonstrates the effectiveness of using real images to guide the generation process. By conditioning the generation on real examples and filtering out inconsistent images with RFF, our approach can generate high-quality synthetic data that closely resembles the real-world examples while still providing a diverse range of variations to improve the robustness of the classification models.

8. Ablation Study on Open-Source and Proprietary Text to Image Models

In this section, we present an ablation study to compare the performance of different text-to-image generative models when used in conjunction with various classifiers for rare plant species classification. This study aims to investigate the impact of the choice of generative model on the effectiveness of our proposed approach and to identify the best-performing combinations of generative models and classifiers.

We consider five state-of-the-art text-to-image generative models, including three open-source models (Stable Diffusion XL, Openjourney-v4, and Latent Consistency Model) and two proprietary models (Midjourney and Dall-E 3). These models are selected based on their demonstrated ability to generate high-quality and diverse images from textual descriptions and their popularity in the research community and industry.

We evaluate each generative model's performance when used with four different classifiers: CLIP ViT-B/32, LLaVa-Mistral, BLIP-2, and ViT-L/16@384. These classifiers represent various architectures and training methodologies and have performed strongly on various image classification tasks. We evaluate all these models on a relatively more challenging task of zero-shot rare-common classification (Task-2).

Table 7 presents the results of our ablation study, showing the classification accuracy achieved by each combination of generative model and classifier on the test set of rare plant species images. The results are averaged over multiple runs to ensure robustness and reproducibility.

From the table, the choice of generative model significantly impacts the classifiers' performance. Among the open-source models, Stable Diffusion XL consistently achieves the highest accuracy across all classifiers, with a solid performance when used with ViT-L/16@384 (0.85 accuracy). Openjourney-v4 also performs well, especially with LLaVa-Mistral (0.73 accuracy), while the Latent Consistency Model generally yields lower accuracy than the other open-source models.

When comparing the proprietary models, we find that Midjourney outperforms Dall-E 3 across all classifiers, with a notable performance when used with ViT-L/16@384 (0.89 accuracy). This suggests that the Midjourney model is particularly effective at generating realistic and informative images of rare plant species, which can be effectively leveraged by the classifiers for improved classification performance.

Interestingly, we also observe that the relative performance of the classifiers varies depending on the generative model used. For example, while ViT-L/16@384 consistently achieves the highest accuracy for all generative models, the ranking of the other classifiers differs across models. This highlights the importance of considering the interaction between the generative model and classifier when designing an effective rare plant species classification system.

Overall, our ablation study demonstrates the significant impact of the choice of generative model on the performance of rare plant species classification using synthetic images. The results suggest that Stable Diffusion XL and Midjourney are the most effective generative models for this task and that ViT-L/16@384 is the best-performing classifier across all models. These findings provide valuable insights for researchers and practitioners working on rare plant species classification and can guide the selection of appropriate generative models and classifiers for optimal performance.

9. Real Image Guided Generation (RIGG) Strategy

In this section, we describe our approach to leveraging a small set of real in-domain images to guide the synthetic image generation process in few-shot settings. Our method, called Real Image Guided Generation (RIGG), modifies the initialization step of the standard text-to-image generation pipeline to incorporate information from real reference images.

In a typical text-to-image generation process using the Stable Diffusion XL model, the first step involves sampling a purely noisy latent image $x_T \sim \mathcal{N}(0, I)$, which serves as the starting point for the iterative denoising process. The model then predicts progressively less noisy latent images x_{t-1} $(t = T, T-1, ..., 1)$ by conditioning on the text prompt c and the current noisy latent image x_t.

In contrast, RIGG introduces a reference image x_0^{ref} to guide the generation process. We first add noise to x_0^{ref} to obtain a noisy latent image $x_{t_*}^{ref}$ corresponding to a specific time-step t_*, as shown in Equation 1:

$$x_{t_*}^{ref} = \sqrt{\bar{\alpha}_{t_*}} x_0^{ref} + \sqrt{1 - \bar{\alpha}_{t_*}} \epsilon \tag{1}$$

Instead of starting from a completely noisy latent image at time-step T, we initialize the denoising process with $x_{t_*}^{ref}$ and begin generating less noisy latent images from time-step t_* onwards, as outlined in Algorithm 1. It is worth noting that Stable Diffusion XL employs a two-stage coarse-to-fine generation framework and uses classifier-free guidance. However, for simplicity, we omit these details in Algorithm 1 since our image-guidance strategy only affects the initialization step and leaves the remaining settings unchanged.

By initializing the generation process with a noisy version of a real reference image, RIGG aims to generate synthetic images that share similar in-domain properties, thereby helping to bridge the domain gap between real and synthetic data. The choice of the time-step t_* plays a crucial role in balancing the trade-off between similarity to the reference image and diversity of the generated samples. A small value of t_* results in generated images that closely resemble the reference image but lack diversity, which can hinder the classifier's learning. On the other hand, a large value of t_* retains little information from the reference image, causing the generated images to deviate from the desired domain.

In our experiments, we explore different values of t_* for various few-shot settings to find the optimal balance between similarity and diversity. Empirically, we set t_* to 15, 20, 35, 40, and 50 for 5-shot, 4-shot, 3-shot, 2-shot, and 1-shot settings, respectively.

By incorporating real image guidance into the generation process, RIGG enables the synthesis of diverse yet domain-relevant images, which can significantly improve the performance of classifiers trained on limited real data in few-shot settings.

Algorithm 1 Real Image Guided Generation (RIGG)

Require: Reference image x_0^{ref}, text prompt c and SDXL model (μ_0, Σ_0).
Ensure: Generated image x_0
1: # Noisy variable initialization
2: Select a time-step $t_k \sim 1, 2, 3, \ldots, T$ and random noise $\epsilon \sim \mathcal{N}(0, I)$
3: Obtain initial noisy image $x_{t_k} := x_{t_k}^{ref}$ according to 1
4: # Random Sampling (is be replaced by EulerDiscreteScheduler for speed-up)
5: **for** $s = t_k$ to 1 **do**
6: $\quad \mu, \Sigma \leftarrow \mu_\theta(x_s, s, c), \Sigma_\theta(x_s, s, c)$
7: $\quad x_{s-1} \leftarrow$ sample from $\mathcal{N}(\mu, \Sigma)$
8: **end for**
9: **return** x_0

10. Implementation Details

This section provides the implementation details for our experiments in both zero-shot and few-shot classification settings.

Zero-shot Setting: For the zero-shot classification tasks, we generate a total of 500 images, with 100 images for each of the five classes. These images are generated using the Primary Strategy (P), Enhanced Description (ED), and Feature Filtering (FF) techniques.

Few-shot Setting: In the few-shot settings, we employ three strategies for text-to-image generation: the Primary Strategy (P), Real Feature Filtering (RFF), and Real Image Guided Generation (RIGG). For P and RFF, we follow the same process as in the zero-shot settings. The RIGG strategy is described in detail in Section 9. In total, we generate 800 synthetic images for the few-shot experiments.

Training Procedure: For both zero-shot and few-shot settings, we use the AdamW optimizer with a weight decay of 0.1 and apply the cosine annealing rule for learning rate scheduling. Images are preprocessed by resizing them to 224x224 pixels, regardless of their original aspect ratio. We use a batch size of 32 for few-shot real images and 512 for synthetic images.

In the phase-wise training approach, we train the models for 30 epochs in each stage, using an initial learning rate of 0.002. We train the models for 30 epochs for mixed training with an initial learning rate of 0.001. The loss values from real and synthetic data are added in a 1:1 ratio during each iteration in mix training.

Data Augmentation: Due to the limited number of real images available, we apply various data augmentation techniques to increase the dataset size. These transformations

include RandomResizedCrop, RandomHorizontalFlip, ColorJitter, RandomGrayscale, and GaussianBlur.

Text Description Augmentation: Although we generate Enhanced Descriptions (ED) using the Claude-3 (opus) model, the number of text descriptions remains limited. To prevent overfitting on specific text features during the contrastive learning process for MLLM and CLIP models, we utilize the LLAMA-7B-Chat model in its 4-bit quantized form to rephrase the generated descriptions, thereby increasing the diversity of the text data.

By providing these implementation details, we aim to ensure the reproducibility of our experiments and facilitate a better understanding of our approach to rare plant species classification using synthetic data.

11. Computational Efficiency

In this section, we provide details on the computational efficiency of our proposed approach, including inference time, training time, and deployment-related aspects. We conduct all experiments on a single NVIDIA A100 GPU with 40GB of memory.

11.1. Inference Time

Table 8 presents the average inference time per image for the different classifiers used in our experiments. We observe that the ViT-based classifiers (ViT-B/32 and ViT-L/16@384) have the lowest inference times, making them suitable for real-time applications. The CLIP model also demonstrates fast inference, while the MLLM models (LLaVa-Mistral and BLIP-2) have slightly higher inference times due to their more complex architectures.

Classifier	Inference Time (ms)
CLIP ViT-B/32	12.5
LLaVa-Mistral	28.7
BLIP-2	35.2
ViT-L/16@384	9.8

Table 8. Average inference time per image for different classifiers.

11.2. Training Time

Table 9 shows the training time for the different classifiers in both zero-shot and few-shot settings. In the zero-shot setting, training involves fine-tuning the pre-trained models on the synthetic data generated by our approach. In the few-shot setting, we report the training time for the mix training strategy, which combines real and synthetic data. The training times are measured for 30 epochs in both settings.

11.3. Deployment Considerations

Our approach can be easily deployed on various platforms, including cloud services and edge devices, thanks to the wide availability of pre-trained models and the efficiency of our synthetic data generation pipeline. The Stable Diffusion XL model used for generating synthetic images can be run on consumer-grade GPUs with at least 8GB of memory, making it accessible to a broad range of users [45].

Classifier	Zero-shot (min)	Few-shot (min)
CLIP ViT-B/32	45	60
LLaVa-Mistral	90	120
BLIP-2	105	135
ViT-L/16@384	60	75

Table 9. Training time for different classifiers in zero-shot and few-shot settings.

For deployment on resource-constrained edge devices, we recommend using the ViT-based classifiers due to their low inference time and memory footprint. The CLIP model is also a viable option for edge deployment, as it has been shown to work well with quantization and pruning techniques [35].

In summary, our approach demonstrates strong computational efficiency, with fast inference times and reasonable training times, making it suitable for various real-world applications. The wide availability of pre-trained models and the efficiency of our synthetic data generation pipeline further facilitate the deployment of our approach on diverse platforms.