UNDERSTANDING REPRESENTATION GAPS ACROSS SCALES IN TROPICAL TREE SPECIES CLASSIFICATION FROM DRONE IMAGERY

Sulagna Saha[1,2], **Arthur Ouaknine**[1,2], **Etienne Laliberté**[3,1], **Carol Altimas**[3,1]
Evan M. Gora[4,5], **Adriane Esquivel Muelbert**[6,7], **Ian R. McGregor**[4]
Cesar Gutierrez[4], **Vanessa Rubio**[4], **David Rolnick**[1,2]

[1]**Mila – Quebec AI Institute** [2]**McGill University**
[3]**Université de Montréal** [4]**Cary Institute of Ecosystem Studies**
[5]**Smithsonian Tropical Research Institute**
[6]**Department of Plant Sciences, University of Cambridge**
[7]**Universidade do Estado do Mato Grosso (UNEMAT)**

ABSTRACT

Accurate classification of tropical tree species from unoccupied aerial vehicle (UAV) imagery remains challenging due to high species diversity and strong visual similarity among species at typical image resolutions (centimeters per pixel). In contrast, models trained on close-up citizen science photographs captured with smartphones achieve strong plant species classification performance. Recent advances in UAV data acquisition now enable the collection of close-up images that are spatially registered with crown-view aerial imagery and approach the level of visual detail found in smartphone photographs, with the trade-off that such high-resolution photos cannot be acquired for many trees. In this work, we evaluate the performance of existing methods using paired crown-view and close-up UAV imagery collected in a species-rich tropical forest. Through fine-tuning experiments, we quantify the performance gap between vision foundation models and in-domain generalist plant recognition models across both image types (high-resolution close-up versus coarser-resolution crown-view imagery). We show that classification performance is consistently higher on close-up images (77.9%) on single date than on crown-view aerial imagery (74.3%) even after aggregating over 16 dates, and that this performance gap widens for rare species. Finally, we propose that self-supervised representation alignment across these two spatial scales offers a promising approach for integrating fine-grained visual information into canopy-level species classification models. Leveraging high-resolution close-up UAV imagery to enhance canopy-level species classification could substantially improve large-scale monitoring of tropical forest biodiversity.

1 INTRODUCTION

Tropical forests are the most biodiverse terrestrial ecosystems on Earth, harboring more than half of all tree species while occupying only about 10% of the global land area (Beech et al., 2017). Large canopy trees are particularly important because of their disproportionate contributions to carbon storage and ecosystem functioning (Slik et al., 2013). Despite their ecological significance, we know remarkably little about tropical canopy tree species. (Esquivel-Muelbert et al., 2019). Individual tree species often exhibit distinct responses to environmental change, making accurate canopy-level biodiversity monitoring both essential and challenging (Araujo et al., 2020). Traditional ground-based surveys are costly, labor-intensive, and difficult to scale across large or remote regions (ForestPlots.net et al., 2021), while satellite imagery frequently lacks the spatial resolution required for reliable species classification (Phillips, 2023). High-resolution crown-view RGB UAV imagery offers a promising, scalable alternative; however, annotating such data requires expert botanical knowledge and typically results in severe class imbalance, particularly for rare species (Schiefer et al., 2020). Earlier work has shown that species-level tree classification from UAV imagery is feasible when high-resolution RGB data and accurate individual crown delineation are available (Kattenborn et al., 2021), with strong performance reported primarily in temperate forests, plantations, and other low-diversity systems where species exhibit pronounced morphological differences and labeled data are more abundant (Ferreira et al., 2023). However, recent studies demonstrate that classification accuracy degrades sharply as species richness increases and inter-species visual

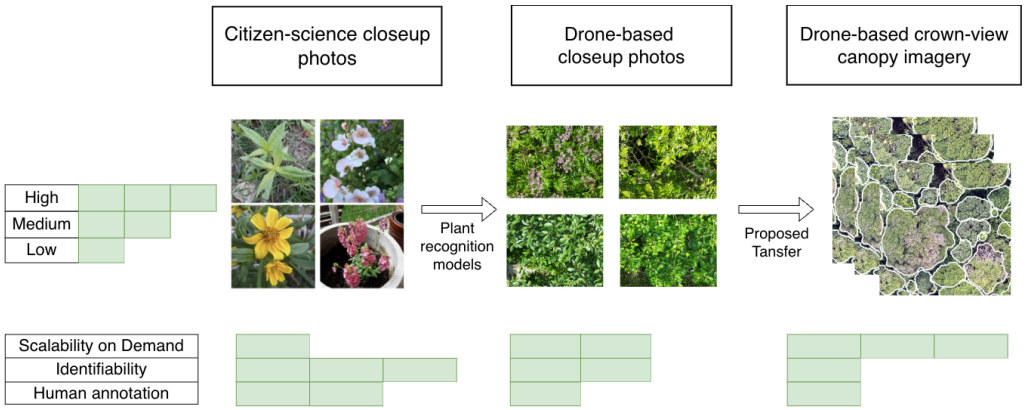


Figure 1: From left to right, we show citizen-science close-up photographs (Affouard et al., 2025), drone-acquired close-up images, and drone-based crown-view canopy imagery. Citizen-science images offer high species identifiability and are easily annotated by humans but are weakly scalable on demand. Drone-based close-ups reduce annotation effort and improve scalability but introduce a domain shift relative to citizen data. Crown-view canopy imagery is the most scalable modality for large-area monitoring, yet lacks fine-grained botanical cues.

differences become more subtle, even with accurate crown segmentation (Teng et al., 2025; Nasiri et al., 2025). This challenge is exacerbated even more in tropical forests, which exhibit extreme species richness. (Cooper et al., 2024). High-resolution close-up imagery, acquired through citizen science platforms (Boone & Basille, 2019; Garcin et al., 2021) or targeted drone flights (Laliberté et al., 2025; Zhang et al., 2016) can help as they capture fine-grained characteristics such as leaf shape and arrangement, or flowers and fruits, that enable botanists to reliably identify species (Fig. 1). Modern plant recognition models such as Pl@ntNet are predominantly trained on diverse citizen science close-up photographs (Lefort et al., 2026a), which differ significantly from crown-view RGB UAV imagery in terms of spatial resolution, acquisition geometry, viewpoint, illumination, and background. Recent studies show how these models can be leveraged to produce meaningful, species-relevant representations when applied to UAV imagery (Soltani et al., 2022), but this has not been explored in species-rich tropical forests.

Species-discriminative visual cues are often ambiguous in crown-view canopy imagery acquired at centimeter-scale resolutions (Schiefer et al., 2020; Cloutier et al., 2024). Recent drone-based workflows now enable the rapid and low-cost acquisition of close-up canopy photographs at sub-millimeter resolution (approximately 0.4 mm) (Laliberté et al., 2025), substantially narrowing the gap between conventional crown-view UAV imagery and close-up citizen science photographs (Fig. 1). In this study, we leverage drone-acquired close-up imagery to transfer fine-grained species information from plant recognition models to canopy-level classification. Our results show that plant recognition models generalize to both close-up and crown-view drone imagery for tree species classification, even under severe label scarcity and long-tailed class distributions. The observed performance gap between these spatial scales highlights opportunities for cross-scale representation alignment as a scalable pathway toward improved biodiversity monitoring in species-rich tropical forests.

2 Dataset

We conduct experiments using high-resolution RGB drone imagery collected over Barro Colorado Island (BCI) during 2024–2025 (Saha et al., 2026). The dataset comprises monthly whole-island orthomosaics captured at around 4 cm GSD. Such high-resolution UAV imagery is critical for reliable individual tree crown detection (Baudchon et al., 2026), delineation (Duguay et al., 2026), and species-level classification in structurally complex tropical forests. To enable individual-based analysis, we segment the RGB orthomosaics into tree-level crown-view polygons using `CanopyRS`, an automated canopy segmentation pipeline designed for high-resolution aerial imagery (Baudchon

Table 1: Performance on crown-view canopy images. Models are fine-tuned on labeled canopy data across 16 acquisition periods. We report both individual-image predictions without taking time into account and crown-level soft voting aggregation.

Evaluation mode	Model	Accuracy (%)			F1 Score		
		Top-1	Top-3	Top-5	Macro	Micro	Weighted
Individual-image	ResNet50	65.5	77.8	81.9	0.33	0.65	0.62
	DINOv3	63.6	76.4	81.0	0.31	0.65	0.61
	BioCLIPv2	52.8	66.1	70.8	0.21	0.51	0.47
	Pl@ntNet	**67.8**	**79.8**	**84.3**	**0.37**	**0.69**	**0.66**
Soft-voting	ResNet50	59.1	68.9	72.6	0.24	0.59	0.54
	DINOv3	72.3	80.5	82.4	0.39	0.72	0.68
	BioCLIPv2	61.1	69.5	74.8	0.24	0.61	0.54
	Pl@ntNet	**74.3**	**81.6**	**83.7**	**0.40**	**0.74**	**0.70**

et al., 2026; Duguay et al., 2026). Using the `geodataset` v0.2.21 [1] Python package, we extract 512 × 512 RGB image tiles centered on each crown polygon. Pixels outside the polygon boundary are masked with black values to prevent background leakage and enforce crown-focused learning. The dataset time series enables observation of each individual tree across up to 16 monthly snapshots. These multi-temporal observations capture phenological and illumination variability while maintaining consistent spatial alignment. We also leverage a limited set of close-up images acquired during targeted drone missions designed to support taxonomic identification (Saha et al., 2026).

In total, close-up imagery is available for 5,302 crown polygons, of which 1,999 polygons have species labels (annotated by expert from tropical regions). We restrict our classification experiments to 84 species that have at least 1 labeled individual for training available across the dataset. This results in a highly imbalanced class distribution dominated by rare species. For labeled data, crown polygons are randomly assigned to training (70%), validation (15%), and test (15%) splits, resulting in 1,385 training, 288 validation, and 326 test labeled polygons (1,999 total). All temporal observations of a given tree are confined to the same split. We adopt a random polygon-level split rather than a geospatial split for two reasons. First, the dataset exhibits a large number of species with highly imbalanced frequencies, and enforcing spatial separation would substantially reduce rare-species representation in validation and test sets, leading to unstable and uninformative performance estimates. Second, all model inputs are derived from individually segmented crown polygons, and we explicitly remove all pixels outside each segmentation mask, removing pixel-level information leakage across splits. Close-up images inherit the split assignment of their corresponding crown polygon when labels are available.

3 Experiments & Results

We evaluate a diverse set of vision models commonly used in ecological image recognition. These include **ResNet-50** (He et al., 2015) as a supervised convolutional baseline, **DINOv3** (Siméoni et al., 2025) as a self-supervised vision transformer pretrained on large-scale image collections, **BioCLIP2** (Gu et al., 2025) as a biologically informed vision-language model, and **Pl@ntNet** (Lefort et al., 2026b) as a plant recognition model based on DINOv2, pre-trained primarily on millions of close-up botanical photographs (using up-to-date weights for the production Pl@ntNet pre-trained model). The model specifications are mentioned in Table 3. For crown-view canopy images, comprising 16 temporal observations per tree, we evaluate two settings (Table 1): *individual-image*, where each date is treated as an independent sample, and *soft-voting*, where predicted class probabilities are averaged across the 16 temporal samples.

We observe performance to be strongly dependent on the representation scale. For crown-view canopy imagery (Table 1), soft voting consistently improves over per-image accuracy across all models. Among baselines, Pl@ntNet consistently performs best suggesting that representations

[1] https://github.com/hugobaudchon/geodataset

Table 2: Performance on close-up images acquired on a single date.

Model	Accuracy (%)			F1 Score		
	Top-1	Top-3	Top-5	Macro	Micro	Weighted
ResNet50	40.8	55.1	63.9	0.14	0.42	0.36
DINOv3	76.8	**86.9**	**89.1**	0.39	0.76	0.71
BioCLIPv2	59.5	66.3	67.3	0.22	0.59	0.54
Pl@ntNet	**77.9**	82.9	83.9	**0.46**	**0.81**	**0.77**

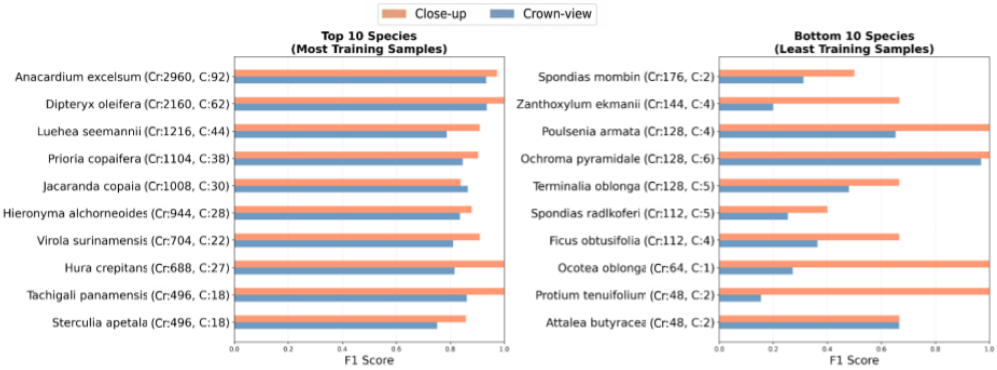


Figure 2: F1 score comparison between crown-view (blue) and close-up (orange) models for top 10 (left) and bottom 10 (right) species by training sample size. Labels show training and test sample counts for crown-view (T) and close-up (C). Bottom 10 filtered for species with both non-zero F1 score with at least 1 training sample.

learned from large-scale plant recognition data partially transfer to canopy imagery. In contrast, on close-up imagery (Table 2), both DINOv3 and Pl@ntNet show a large performance gain, with Pl@ntNet achieving the highest macro- and micro-F1 scores. As shown in Figure 2, for common species, both viewpoints achieve high and relatively stable performance, though close-up images consistently yield equal or higher F1 scores. For rare species, the gap becomes pronounced: crown-view performance degrades sharply with decreasing sample size, while close-up imagery maintains substantially higher F1 scores across most taxa, even in the extreme low-data regime (often fewer than 20 training instances). Notably, this advantage is achieved despite the close-up dataset containing far fewer total samples than the crown-view dataset.

4 Discussion & Future Directions

Our experiments reveal a persistent performance gap between canopy-level species classification from crown-view and close-up UAV imagery. While modern vision foundation models fine-tuned on segmented tree crown polygons achieve strong performance for common species, accuracy degrades substantially for rare taxa. This degradation is consistent across model families and is particularly pronounced in the long tail of the class distribution, where limited labeled data and subtle inter-species visual differences dominate. As future work, we will leverage unlabeled drone-acquired close-up imagery through a teacher–student representation transfer framework. A frozen Pl@ntNet model trained on close-up botanical images will serve as the teacher, while a Pl@ntNet-initialized student will be adapted to operate on crown-view canopy tiles. We will align embeddings via a cosine distillation loss to transfer species-relevant cues from close-up views to canopy-level representations. In conclusion, we systematically demonstrate a persistent representation gap between crown-view canopy imagery and drone based close-up photos for tropical tree species classification, with failures most pronounced under long-tailed, label-scarce regimes. Our results highlight the need for cross-scale representation alignment to transfer identifiable species cues into scalable canopy-level models.

REFERENCES

A. Affouard, A. Joly, J. Lombardo, J. Champ, H. Goeau, M. Chouet, H. Gresse, and P. Bonnet. Pl@ntnet observations, 2025. URL https://doi.org/10.15468/gtebaa. Occurrence dataset accessed via GBIF.org on 2026-01-31.

Raquel Fernandes Araujo, Jeffrey Q. Chambers, Carlos Henrique Souza Celes, Helene C. Muller-Landau, Ana Paula Ferreira Dos Santos, Fabiano Emmert, Gabriel H. P. M. Ribeiro, Bruno Oliva Gimenez, Adriano J. N. Lima, Moacir A. A. Campos, and Niro Higuchi. Integrating high resolution drone imagery and forest inventory to distinguish canopy and understory trees and quantify their contributions to forest structure and dynamics. *PLOS ONE*, 15(12), December 2020. ISSN 1932-6203. URL https://dx.plos.org/10.1371/journal.pone.0243079.

Hugo Baudchon, Arthur Ouaknine, Martin Weiss, Mélisande Teng, Thomas R. Walla, Antoine Caron-Guay, Christopher Pal, and Etienne Laliberte. SelvaBox: A high-resolution dataset for tropical tree crown detection. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GH7z1RURL6.

E. Beech, M. Rivers, S. Oldfield, and P. P. Smith. GlobalTreeSearch: The first complete global database of tree species and country distributions. *Journal of Sustainable Forestry*, 36(5):454–489, July 2017. ISSN 1054-9811. doi: 10.1080/10549811.2017.1310049.

Matthew Earl Boone and Mathieu Basille. Using iNaturalist to Contribute Your Nature Observations to Science. *EDIS*, 2019, August 2019. ISSN 2576-0009. URL https://journals.flvc.org/edis/article/view/107698.

Myriam Cloutier, Mickaël Germain, and Etienne Laliberté. Influence of temperate forest autumn leaf phenology on segmentation of tree species from UAV imagery using deep learning. *Remote Sensing of Environment*, 311:114283, September 2024. ISSN 00344257. URL https://linkinghub.elsevier.com/retrieve/pii/S0034425724003018.

Declan L. M. Cooper, Simon L. Lewis, Martin J. P. Sullivan, Paulo I. Prado, Hans ter Steege, Nicolas Barbier, Ferry Slik, Bonaventure Sonké, Corneille E. N. Ewango, Stephen Adu-Bredu, Kofi Affum-Baffoe, Daniel P. P. de Aguiar, Manuel Augusto Ahuite Reategui, Shin-Ichiro Aiba, Bianca Weiss Albuquerque, de Almeida Matos, et al. Consistent patterns of common species across tropical tree communities. *Nature*, 625(7996):728–734, January 2024. ISSN 1476-4687.

Simon-Olivier Duguay, Hugo Baudchon, Etienne Laliberté, Helene Muller-Landau, Gonzalo Rivas-Torres, and Arthur Ouaknine. SelvaMask: Segmenting Trees in Tropical Forests and Beyond, 2026. URL https://arxiv.org/abs/2602.02426.

Adriane Esquivel-Muelbert, Timothy R. Baker, Kyle G. Dexter, Simon L. Lewis, Roel J. W. Brienen, Ted R. Feldpausch, Jon Lloyd, Abel Monteagudo-Mendoza, Luzmila Arroyo, Esteban Álvarez Dávila, Higuchi, et al. Compositional response of Amazon forests to climate change. *Global Change Biology*, 25(1), January 2019. ISSN 1354-1013, 1365-2486. URL https://onlinelibrary.wiley.com/doi/10.1111/gcb.14413.

Cassiana Alves Ferreira, Janet Gaby Inga Guillen, Raul Huacho Buendia, Osir Daygor Vidal Alanya, Danessa Clarita Reyes Aliaga, Walter Goytendia Centeno, Benji Steve Ascue Miranda, Sthefany Madjory Moya Mateo, Thonny Centeno Utos, Andrés Veléz Echeverry, and Mario Tomazello Filho. Identification of 20 species from the Peruvian Amazon tropical forest by the wood macroscopic features. *CERNE*, 29, 2023. ISSN 2317-6342, 0104-7760. doi: 10.1590/01047760202329013134. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-77602023000100702&tlng=en.

ForestPlots.net, Cecilia Blundo, Julieta Carilla, Ricardo Grau, Agustina Malizia, Lucio Malizia, Oriana Osinaga-Acosta, Michael Bird, Matt Bradford, Damien Catchpole, et al. Taking the pulse of Earth's tropical forests using networks of highly distributed plots. *Biological Conservation*, 260, August 2021. ISSN 00063207. URL https://linkinghub.elsevier.com/retrieve/pii/S0006320720309071.

Camille Garcin, Alexis Joly, Pierre Bonnet, Antoine Affouard, Jean-Christophe Lombardo, Mathias Chouet, Maximilien Servajean, Titouan Lorieul, and Joseph Salmon. Pl@ntNet-300K image dataset, April 2021. URL https://zenodo.org/record/5645731.

Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E. White, et al. BioCLIP 2: Emergent Properties from Scaling Hierarchical Contrastive Learning, 2025. URL https://arxiv.org/abs/2505.23883.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf.

Teja Kattenborn, Jens Leitloff, Felix Schiefer, and Stefan Hinz. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, March 2021. ISSN 09242716. doi: 10.1016/j.isprsjprs.2020.12.010. URL https://linkinghub.elsevier.com/retrieve/pii/S0924271620303488.

Etienne Laliberté, Antoine Caron-Guay, Vincent Le Falher, Guillaume Tougas, Helene C. Muller-Landau, Gonzalo Rivas-Torres, Thomas R. Walla, Hugo Baudchon, Mélvin Hernandez, Adrian Buenaño, Anna Weber, Chambers, et al. Seeing the forest and the trees: a workflow for automatic acquisition of ultra-high resolution drone photos of tropical forest canopies to support botanical and ecological studies, September 2025. URL http://biorxiv.org/lookup/doi/10.1101/2025.09.02.673753.

Tanguy Lefort, Antoine Affouard, Benjamin Charlier, Jean-Christophe Lombardo, Mathias Chouet, Hervé Goëau, Joseph Salmon, Pierre Bonnet, and Alexis Joly. Cooperative learning of pl@ntnet's artificial intelligence algorithm: How does it work and how can we improve it? *Methods in Ecology and Evolution*, 17(2):392–403, 2026a. doi: https://doi.org/10.1111/2041-210X.14486. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.14486.

Tanguy Lefort, Antoine Affouard, Benjamin Charlier, Jean-Christophe Lombardo, Mathias Chouet, Hervé Goëau, Joseph Salmon, Pierre Bonnet, and Alexis Joly. Cooperative learning of Pl@ntNet's Artificial Intelligence algorithm: How does it work and how can we improve it? *Methods in Ecology and Evolution*, 17(2), February 2026b. ISSN 2041-210X, 2041-210X. URL https://besjournals.onlinelibrary.wiley.com/doi/10.1111/2041-210X.14486.

Kamyar Nasiri, William Guimont-Martin, Damien LaRocque, Gabriel Jeanson, Hugo Bellemare-Vallières, Vincent Grondin, Philippe Bournival, Julie Lessard, Guillaume Drolet, Jean-Daniel Sylvain, and Philippe Giguère. Using Citizen Science Data as Pre-Training for Semantic Segmentation of High-Resolution UAV Images for Natural Forests Post-Disturbance Assessment. *Forests*, 16(4):616, March 2025. ISSN 1999-4907. URL https://www.mdpi.com/1999-4907/16/4/616.

Oliver L. Phillips. Sensing Forests Directly: The Power of Permanent Plots. *Plants*, 12(21), October 2023. ISSN 2223-7747. URL https://www.mdpi.com/2223-7747/12/21/3710.

Sulagna Saha, Arthur Ouaknine, Etienne Laliberté, Carol Altimas, Evan M. Gora, Adriane Esquivel Muelbert, Ian R. McGregor, Cesar Gutierrez, Vanessa E. Rubio, and David Rolnick. bci-temporal (revision d222b07), 2026. URL https://huggingface.co/datasets/sulagnasaharasha/bci-temporal.

Felix Schiefer, Teja Kattenborn, Annett Frick, Julian Frey, Peter Schall, Barbara Koch, and Sebastian Schmidtlein. Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS Journal of Photogrammetry and Remote Sensing*, 170, December 2020. ISSN 09242716. URL https://linkinghub.elsevier.com/retrieve/pii/S0924271620302938.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana,

Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. Dinov3, 2025. URL https://arxiv.org/abs/2508.10104.

J. W. Ferry Slik, Gary Paoli, Krista McGuire, Ieda Amaral, Jorcely Barroso, Meredith Bastian, Lilian Blanc, Frans Bongers, Patrick Boundja, Connie Clark, Murray Collins, Gilles Dauby, Yi Ding, Jean-Louis Doucet, Eler, et al. Large trees drive forest aboveground biomass variation in moist lowland forests across the tropics. *Global Ecology and Biogeography*, 22(12), December 2013. ISSN 1466-822X, 1466-8238. URL https://onlinelibrary.wiley.com/doi/10.1111/geb.12092.

Salim Soltani, Hannes Feilhauer, Robbert Duker, and Teja Kattenborn. Transfer learning from citizen science photographs enables plant species identification in uav imagery. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5:100016, 2022. ISSN 2667-3932. URL https://www.sciencedirect.com/science/article/pii/S2667393222000059.

Mélisande Teng, Arthur Ouaknine, Etienne Laliberté, Yoshua Bengio, David Rolnick, and Hugo Larochelle. Bringing SAM to new heights: leveraging elevation data for tree crown segmentation from drone imagery. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=1vSLxdJNq8.

Jian Zhang, Jianbo Hu, Juyu Lian, Zongji Fan, Xuejun Ouyang, and Wanhui Ye. Seeing the forest from drones: Testing the potential of lightweight drones as a tool for long-term forest monitoring. *Biological Conservation*, 198:60–69, 2016.

ACKNOWLEDGMENTS

We would like to thank Alexis Joly, Jean-Christophe Lombardo, and Pierre Bonnet from the Pl@ntNet (Lefort et al. (2026b)) team for providing up-to-date weights for the pre-trained Pl@ntNet model and for their valuable insights. We are grateful to funding from the Canada CIFAR AI Chairs program, the Global Center on AI and Biodiversity Change (NSERC 585136), and the IVADO (R3AI, Postdoc Entrepreneur) program. This research was enabled in part by compute resources provided by Mila - Quebec AI Institute, including material support from NVIDIA Corporation.

A APPENDIX

A.1 SPECIES DISTRIBUTION ACROSS SPLITS

The following figures illustrate the species distribution across our training, validation, and test splits. Our dataset includes 84 species, with a significant portion of the distribution residing in the long tail; specifically, only 26 species have at least 20 labels.

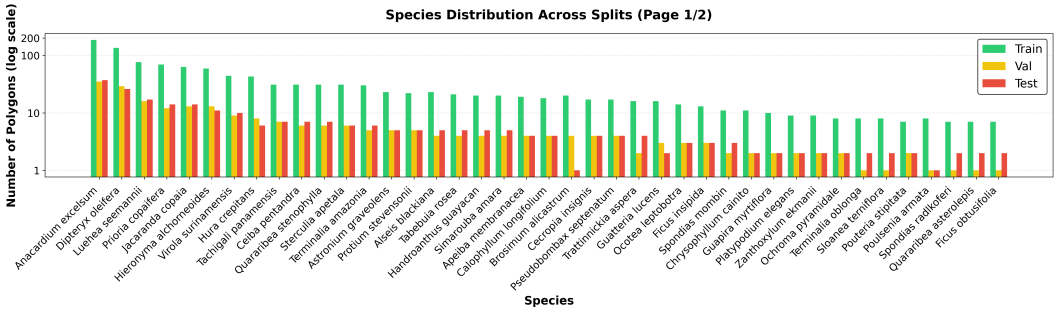


Figure 3: Species Distribution Across Splits (Page 1/2) showing the most common species in the dataset.

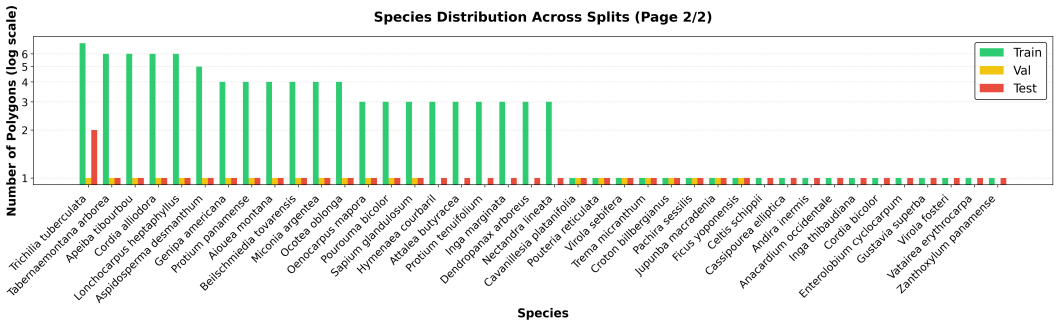


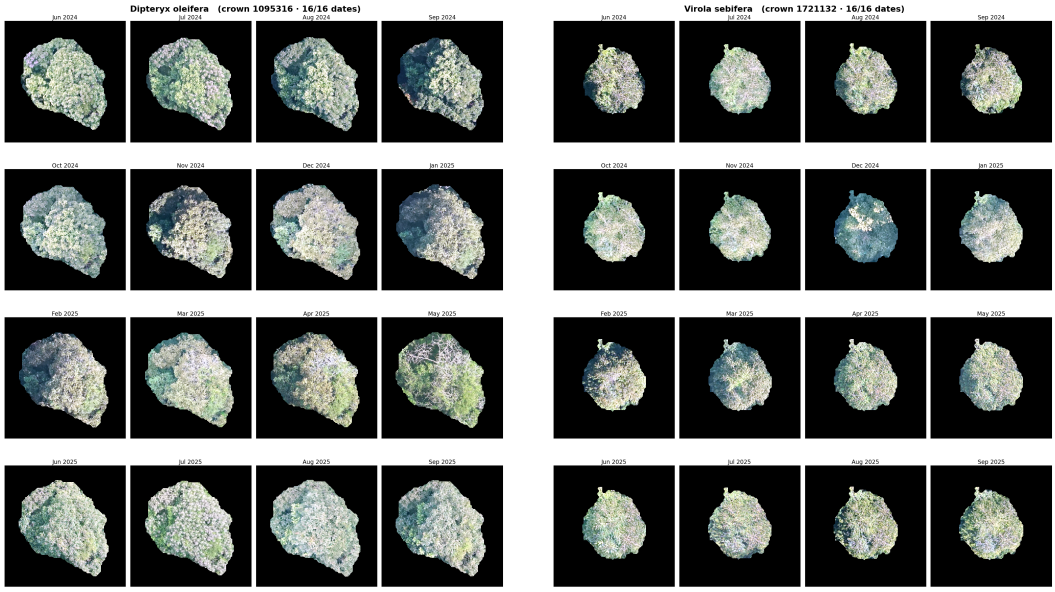
Figure 4: Species Distribution Across Splits (Page 2/2) highlighting the rare species in the long-tail distribution.

A.2 TEMPORAL VARIABILITY

The dataset time series captures phenological and illumination variability across up to 16 monthly snapshots. 5 is a comparison of these temporal changes for a commonly labeled (Dipteryx Oleifera) and a rarely labeled species (Virola Sebifera).

A.3 MODEL SPECIFICATIONS

3 shows the model specifications. During training, we apply a standard set of geometric augmentations consistent across all models. Each image is first randomly cropped and resized to the target input resolution using RandomResizedCrop with a scale range of [0.7, 1.0] and bicubic interpolation,



(a) Commonly labeled species: Dipteryx Oleifera (b) Rarely labeled species: Virola Sebifera

Figure 5: Observation of phenological variability over multiple months.

Table 3: Model specifications and training hyperparameters for all fine-tuned backbones. All models use AdamW, mixed-precision training (fp16), batch size 32, and early stopping on validation loss (patience = 5, min-delta = 0.001) with a maximum of 100 epochs. Training time is total wall-clock time for 3-fold cross-validation on a single A100 GPU. Epochs are reported as mean $\pm$ std across folds.

	ResNet-50	**DINOv3**	**BioCLIP-2**	**Pl@ntNet**
Architecture				
Backbone type	CNN	ViT-B/16	ViT-B/16 (CLIP)	ViT-B/14
Pretraining data	ImageNet-1K	LVD-1.68B	Biological imagery	Citizen science plant imagery
Total parameters	$\sim$25.6M	$\sim$86M	$\sim$149M	$\sim$86M
Input resolution	224×224	512×512	224×224	518×518
Frozen components	None	None	Text encoder	None
Hyperparameters				
Learning rate	1×10^{-4}	1×10^{-4}	5×10^{-5}	6×10^{-6}
Weight decay	1×10^{-4}	1×10^{-4}	0	1×10^{-4}
Classifier dropout	0.0	0.1	0.0	0.1
Training outcomes (3-fold cross-validation)				
Epochs trained	20 ± 7	16 ± 4	34 ± 3	18 ± 4
Total training time	$\sim$49 min	$\sim$2.2 h	$\sim$9.6 h	$\sim$8.1 h

followed by a random rotation of up to ±30°. Random horizontal flipping (p=0.5) is also applied. No color or photometric augmentations are used. All images are normalized using ImageNet statistics (mean=[0.485, 0.456, 0.406], std=[0.229, 0.224, 0.225]).

A.4 Cross-Scale Visual Comparison

We provide visual examples of the two spatial scales used in our experiments: high-resolution close-up imagery (approx. 0.4 mm) and coarser-resolution top-view aerial imagery (4 cm GSD) 6.

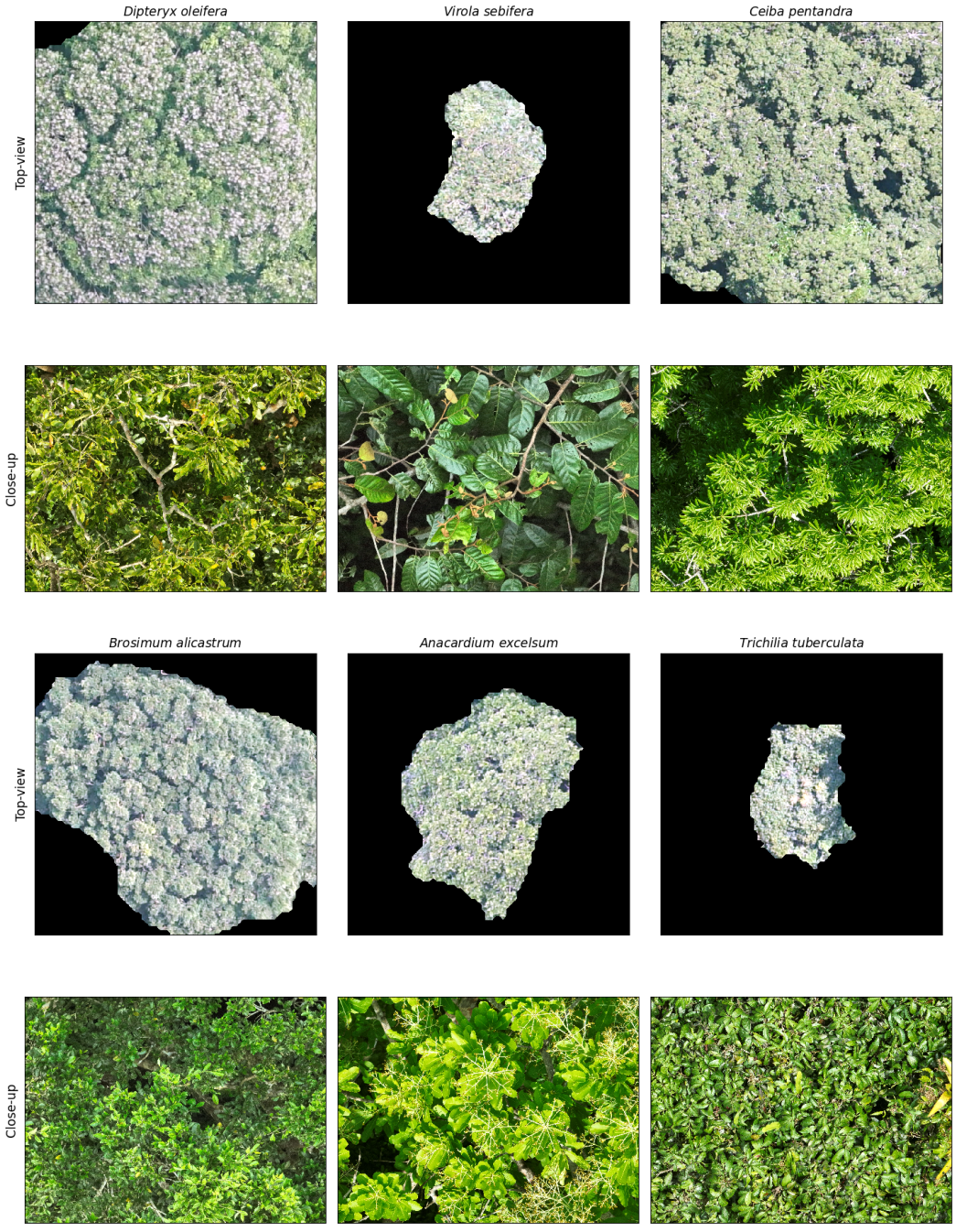


Figure 6: Paired crown-view and close-up drone imagery for six distinct species.

B USE OF LARGE LANGUAGE MODELS (LLMS)

Large Language Models (LLMs) were utilized to assist with minor debugging of the LaTeX and Python code used in the experiments. The LLMs were not used for writing the manuscript, research ideation, data collection, or the generation of novel scientific conclusions. All conceptual, analytical, and experimental contributions remain the work of the authors.