A Distribution Shift Benchmark for Smallholder Agroforestry: Do Foundation Models Improve Geographic Generalization?

Siddharth Sachdeva[1], **Isabel Lopez**[1], **Chandrashekhar Biradar**[2] **& David Lobell**[1] *
[1]Stanford University, [2]World Agroforestry Centre

Abstract

Recent improvements in deep learning for remote sensing have shown that it is possible to detect individual trees using high resolution satellite remote sensing data. However, there has not been an evaluation of the robustness of individual tree detection methods to distribution shifts across varying geographies, and this limits the applicability of these methods to diverse areas beyond the sites in which they were trained. To address this, we introduce a benchmark dataset comprising varying agro-ecological zones for remote sensing tree detection in agroforestry farms in India. We then use this dataset to conduct a geographic robustness evaluation of out-of-distribution (OOD) performance of different deep learning approaches for remote sensing tree detection. Results indicate strong performance of deep learning in detecting trees under conventional evaluation, yet a significant drop in performance in OOD agro-ecological zones for baseline methods. We report some improvements with foundation model based approaches including SAM and Grounding DINO, but find that they also exhibit similar performance drops OOD. Our study pushes the boundaries of current research by challenging machine learning methods with a dataset and evaluation protocol that better represents real-world variability, shedding light on the robustness and adaptability of different individual tree detection methods.

1 Introduction

Agroforestry, or integrating trees and crops together into agricultural systems, has been promoted as a high-potential solution for scaleable carbon removal (Chapman et al., 2020) with many co-benefits (Nair & Garrity, 2012) (Hendershot et al., 2023). However, there currently exists limited data on the extent and type of different agroforestry practices (Hart et al., 2023). In the past, remote sensing data has been too low resolution to distinguish between different types of agroforestry practices that involve the integration of trees with other crops in farmland (Schnell et al., 2015). As Bégué et al. (2018) note, "Agroforestry is a challenging cropping system to monitor using remote sensing because of its spatial heterogeneity and complexity. . . " Earth scientists have recently reported large accuracy improvements for detection of individual trees using deep learning algorithms and high resolution satellite imagery (Brandt et al., 2020). These advances open up the prospect of monitoring every tree on earth (Hanan & Anchang, 2020) and have enabled entirely new applications like quantifying the total stock of carbon in all of the individual trees continental North Africa (Tucker et al., 2023). However, these large-scale applications require accuracy that is unbiased over large areas, whereas in monitoring, reporting, and verification (MRV) of carbon removal, local accuracy in the area of a particular project is more important (Hart et al., 2023). Brandt et al. (2020) noted that "Owing to the high latitudinal variations in vegetation and soil background and to avoid misclassifications in the very sparsely vegetated Sahara desert, we trained two separate models," indicating that the accuracy of a model clearly varies widely in different areas. Geographic variation means that even if a model shows high accuracy on a random held out test set, local accuracy is often poor. Researchers who are applying machine learning to remote sensing data to detect individual trees find it difficult to evaluate whether different tree detection methods or datasets are robust across varying geographies. The machine learning community has identified this problem under names including

*Correspondence: siddsach@stanford.edu . Data, code: https://github.com/siddsach/distshift_agroforestry/

distribution shifts (Koh et al., 2021), domain adaptation (Ganin & Lempitsky, 2015), and transfer learning (Xie et al., 2020). It has been recognized as a significant bottleneck to realizing the full potential of real-world remote sensing applications (Koh et al., 2021). Inspired by these issues, we start with quantifying how baseline deep learning approaches for object detection perform at individual tree detection using 50 cm satellite imagery (PlanetLabs, 2024) and show that it achieves comparable performance to Brandt et al. (2020). We then compare in-distribution (ID) performance on agro-climatic zones in which the model was trained to OOD performance in other agro-climatic zones, and we find a significant drop in OOD compared to ID performance. Finally, we apply recently released computer vision foundation models including Segment Anything and Grounding DINO to our benchmark. We find that though they slightly improve accuracy, there persists a large difference between ID and OOD performance. Our main contributions are (1) a geographic distribution shift benchmark dataset for detecting individual trees across different agro-climatic zones with evaluations for baseline methods and (2) an empirical investigation demonstrating that though computer vision foundation models slightly improve adaptation to geographic distribution shifts under different fine-tuning and few-shot settings, they remain vulnerable to geographic distribution shifts.

1.1 Related Work

Recently, Ouaknine et al. (2023) published *OpenForest: A data catalogue for machine learning in forest monitoring* which provides a thorough review of publicly available forest remote sensing benchmark datasets. These datasets are highly concentrated in western countries, and they find only two datasets that enable object detection of individual trees: the NEON Tree Evaluation (Weinstein et al., 2019), an individual tree detection aerial imagery dataset published by the US Government network NEON, and Reforestree (Reiersen et al., 2022), a tree object detection dataset collected from aerial images with 2 cm resolution. Though these have been important contributions, their quality and consistency is low upon closer inspection, and they have limited geographic diversity, which is essential to real-world applications. A significant recent contribution was Beery et al. (2022), the first tree species classification dataset to explicitly integrate geographic distribution shift evaluations. However, this dataset focused on species classification and ignores the task of individual tree detection, which is required for monitoring agroforestry systems. There have been many methods proposed for improving robustness to distribution shift. Some of them such as meta learning (Russwurm et al., 2020) and unsupervised learning (Ganin & Lempitsky, 2015) have demonstrated promising results in addressing specific distribution shifts. Other systematic studies of domain adaptation methods find that most adaptation methods provide little performance improvement OOD (Taori et al., 2020). Some have reported large improvements in accuracy OOD from using foundation models (Radford et al., 2021) (Li et al., 2022) (Zhao et al., 2023), but the extent to which foundation models address geographic distribution shift issues in remote sensing remains under-explored.

2 Dataset

We frame the individual tree detection problem as an object detection task where the model predicts bounding boxes for each individual tree. To achieve this, we worked with local partners to construct a ground truth annotated dataset of satellite images from agricultural land in Rajasthan, India, where different forms of agroforestry are a common agricultural practice across the state. We source RGB Skysat images (50cm spatial resolution) from the Planet Labs API. To ensure that we could evaluate performance under geographic distribution shift, we conduct a stratified random sample of imagery from the state of Rajasthan according to a map of 8 agro-climatic zone boundaries based on a classification used by the state government, using some filters we detail in Supplemental Section 1. Then, we break these scenes into 400 x 400 pixel RGB images for annotation. We trained two annotators to use the CVAT platform (CVAT.ai) to annotate individual segmentation masks for each green tree crown in the image. Following Brandt et al. (2020) we require each tree to have an associated shadow to aid visual distinction between trees and other vegetation such as shrubs. To ensure our tree annotations were high quality, we double annotated 400 images and measured inter-annotator agreement of 86 AP, setting a baseline for human performance and ensuring annotation consistency. For ground-truth field data comparison, we also collected field data tree inventories for 38 field plots, and found that 94 percent of trees tagged in the field were correctly detected.

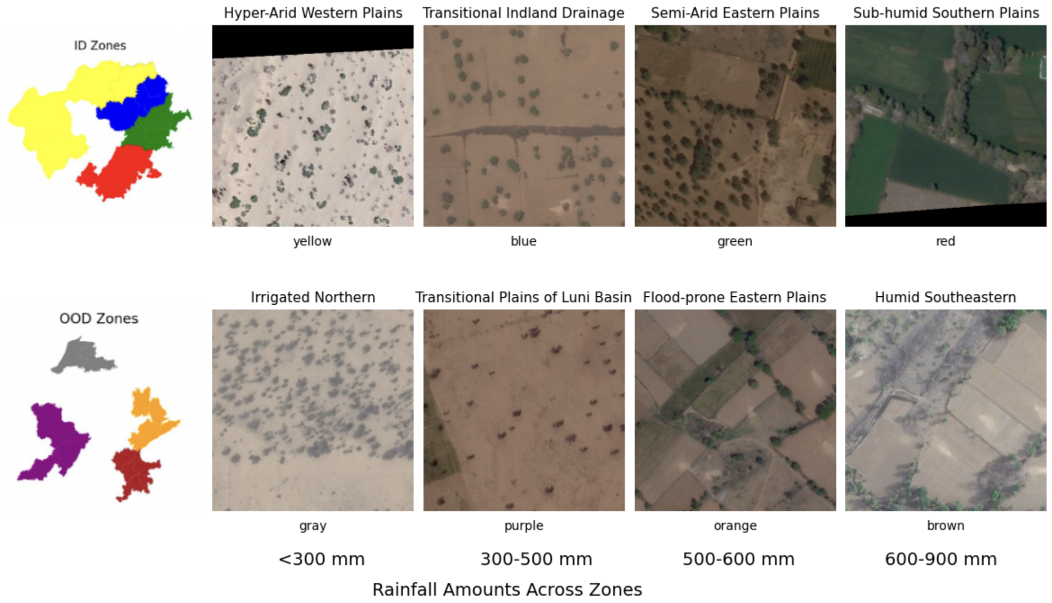
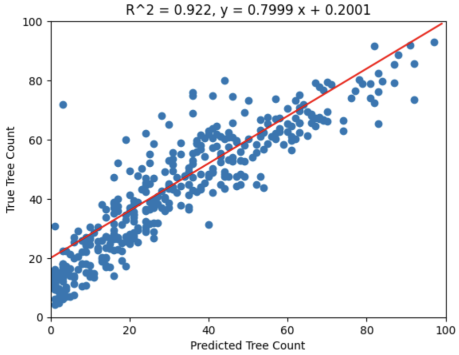

ID Zones
Hyper-Arid Western Plains
Transitional Indland Drainage
Semi-Arid Eastern Plains
Sub-humid Southern Plains
yellow
blue
green
red
OOD Zones
Irrigated Northern
Transitional Plains of Luni Basin
Flood-prone Eastern Plains
Humid Southeastern
gray
purple
orange
brown
<300 mm
300-500 mm
500-600 mm
600-900 mm
Rainfall Amounts Across Zones

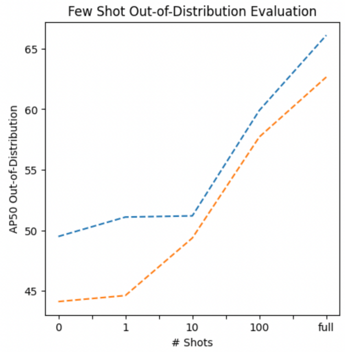
Figure 1: Example Images from different Agro-climatic zones.

3 Methods

To conduct a distribution shift evaluation, we split our dataset of images for each of Rajasthan's 8 agro-climatic zones into train and test images. We then split the zones into four ID and four OOD zones, as shown in figure 1. To split the zones, we bucketed each of them into bins based on the amount of precipitation, with 2 zones in each bin, and then for each bin, we randomly assign zones to ID vs. OOD. We train under 3 evaluation settings. 1) Conventional Model Evaluation: Train on training splits for all zones, evaluate on test splits for all zones. This is similar to the type of cross-validation that is standard in most machine learning for remote sensing research. 2) Distribution shift Evaluation: Train on train splits for ID Zones, Test on test splits for ID Zones vs. OOD Zones. This evaluation will measure the difference between performance on the ID test set sampled from the same agro-climatic zones the model was trained on to with the OOD test set sampled from agro-climatic regions outside the one in which it was trained. 3) Few-shot adaptation evaluation: Train on train splits for ID Zones, plus k={1,10,100,ALL} examples from the training splits of each of the OOD zones. We report Average Precision, a common object detection evaluation metric. We also measure R^2 between predicted and true tree counts on ground truth annotated 400x400 pixel images to directly compare accuracy metrics to the results from Brandt et al. (2020). For our baseline, we train a Faster-RCNN model (we report hyper-parameters in the supplements). For our foundation model experiments, we select two foundation models that have been influential in recent computer vision research: the Segment Anything Model (Kirillov et al., 2023) and Grounding DINO (Zhao et al., 2023). Inspired by recent work demonstrating that fine-tuning foundation models can sometimes reduce their performance OOD, for each model (Kumar et al., 2022), we perform finetuning under three different settings: full fine-tuning, head fine-tuning with a frozen backbone, and head fine-tuning then full fine-tuning. We can directly fine tune Grounding DINO as it is an object detection model, but because the Segment Anything Model requires a prompt embedding, we add a Mask-RCNN head on the SAM encoder to perform automatic object detection.In addition, we use our benchmark to evaluate the accuracy of a tree cover product in India recently released by the Brandt group Brandt et al. (2023) who led the initial work on individual tree detection. As above, we calculate the per-image tree-count R^2 between ground truth tree counts from our annotated dataset and predicted tree counts. We were able to do this by geospatially aligning Brandt's tree cover product in India with our annotations and filtering out areas that the Brandt group did not include in their tree cover product. In addition to reporting the overall tree count R^2, we report the per-zone tree count R^2 for the Brandt data. This will help determine how much variation in accuracy exists between different agro-climatic zones, as well as how much of this variation is specific to a specific model or common to multiple models. In comparing per-zone accuracy of our approach to Brandt et al. (2023), it's important to note that the Brandt tree cover product used 3-meter data to generate its predictions, whereas we used 50cm data, for which predicting tree count is much easier.



(a) Figure 2: Conventional Evaluation of Predicted vs. true tree counts on the ID test set



(b) Figure 3: Few-shot OOD Evaluation of Grounding DINO (blue) and Faster-RCNN (orange)

4 Results

We find that the Faster-RCNN baseline achieves 0.92 Tree Count R^2 on the conventional evaluation ID test set (Figure 2), comparable to Brandt et al. (2020) achieving 0.95 Tree Count R^2. This suggests state-of-the-art tree detection is relatively reproducible with baseline methods as long as a big enough ground truth dataset is present. When we compare the OOD results for the model trained on all zones (conventional evaluation) with the model just trained on ID zones (distribution shift evaluation), we observe more than a 20 percent drop in performance (Table 1). In Figure 4, we take an example image from the OOD test set, and we visualize predictions for the model just trained on ID training sets on the left alongside, and predictions for the model trained on ID + OOD training sets on the right, observing that despite the ID-trained model totally fails on the OOD image. This suggests that although baseline supervised deep learning methods show strong performance when measured under traditional model evaluation, they are highly vulnerable to performance drops under geographic distribution shift. However, even the conventional evaluation performs 10 percent lower on the OOD set, suggesting that different areas might have different inherent levels of difficulty. We report a wide range of accuracy in agro-climatic zones in Supplementary Table 1, giving further evidence that there are differences in inherent difficulty in different areas independent of distribution shift. As shown in table 1, both SAM and Grounding DINO slightly improve over the Faster-RCNN model both ID and OOD, but they still show a similar pattern of large differences between ID and OOD performance. To further explore this issue of varying accuracies in different areas, we report per-zone R^2 metrics of our Faster-RCNN in Supplementary table 1, and we observe large differences in accuracy between zones even when the training data includes all zones. To test whether this is specific to our model, in Supplementary table 1 we also report the per-zone accuracy of 3-meter tree cover product of India recently released by the research team led by Martin Brandt behind the original individual tree detection papers in Nature. We observe even larger variations in accuracy in their product. These results suggest that research on individual tree detection models need to report spatially explicit accuracy results for users to understand where they can expect these models to perform well. When we perform the few-shot evaluation in Figure 3, we find that the grounding DINO model gives larger gains on OOD data when no data from that domain is available, suggesting that foundation models may be slightly more robust than baseline approaches. However, baseline performance comes close to matching grounding DINO with 10 ID examples, suggesting that getting small amounts of ID data can compensate for adaptation to distribution shifts.

5 Discussion

Our results suggest that high overall accuracy metrics using traditional cross-validation for deep learning tree detection models can hide both a wide range in accuracy in different regions and poor

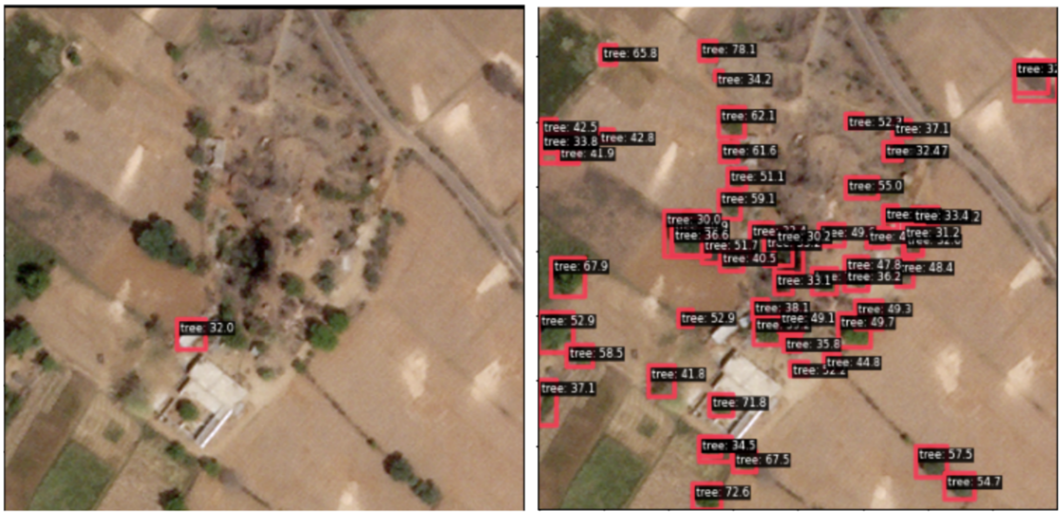


Figure 4: An example image from the OOD test set with predictions from the model just trained on ID data (left), and from the model trained on ID + OOD data (right). We observe a significant drop in tree detection accuracy for OOD examples.

Table 1: Comparison of ID and OOD Average Precision

Method	Eval type	ID AP	OOD AP
Faster-RCNN	Conventional Eval	0.778	0.631
Grounding DINO	Conventional Eval	0.821	0.667
Faster-RCNN	Dist shift eval	0.778	0.441
SAM Finetune Full Finetune	Dist shift eval	0.781	0.485
SAM Finetune Head	Dist shift eval	0.777	0.483
SAM Finetune Head then Full	Dist shift eval	0.592	0.417
Grounding DINO Full Finetune	Dist shift eval	0.814	0.495
Grounding DINO Finetune Head	Dist shift eval	0.810	0.505
Grounding DINO Finetune Head then Full	Dist shift eval	0.808	0.487

generalization. Foundation models did slightly improve performance, especially OOD, even though they were not pre-trained on remote sensing data. This result lends credibility to the argument made in much of the machine learning literature that foundation models pre-trained on large diverse datasets can help address performance drops under distribution shifts. However, there still remains a significant generalization gap even with the most advanced foundation models currently available, especially as these foundation models were not trained on geographically diverse remote sensing data. Moreover, the improvement in OOD performance provided by foundation models over baseline deep learning approaches is much smaller than the improvement in performance by annotating just 10 in-domain samples. This result suggests that until foundation models have much higher robustness to geographic distribution shifts, a more data centric approach that focuses on large-scale high-quality groundtruth dataset collection is perhaps a more practical solution. Foundation models that are trained specifically on diverse remote sensing data may eventually provide the robustness to geographic distribution shifts that they have demonstrated in other distribution shift evaluations, but further work is needed evaluating approaches to solving distribution shift challenges in remote sensing applications in machine learning. At the very least, groundtruth data collection for every geographic region is required to measure the accuracy of different tree detection approaches in different areas, which in turn is a prerequisite to having distributionally robust tree detection approaches. Moreover, given that even baseline approaches can do better with only a small number of ID examples, it is clear that there are multiple benefits to geographically stratified ground truth data collection. We hope that our benchmark will enable researchers who are interested in machine learning for remote sensing of individual trees to measure the extent to which the approaches they develop can improve robustness to geographic distribution shift.

REFERENCES

J. Beery, D. Danescu, A. Avram, D. Shechtman, and L. Zelnik-Manor. The auto-arborist dataset: A large-scale benchmark for multiview urban tree species classification and diameter estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Martin Brandt, Compton J. Tucker, Ankit Kariryaa, Kjeld Rasmussen, Christin Abel, Jennifer Small, Jerome Chave, Laura Vang Rasmussen, Pierre Hiernaux, Abdoul Aziz Diouf, et al. An unexpectedly large count of trees in the west african sahara and sahel. *Nature*, 587:78–82, 2020.

Martin Brandt, Dimitri Gominski, Florian Reiner, Ankit Kariryaa, Venkanna Guthula, Philippe Ciais, Xiaoye Tong, Wenmin Zhang, Dhanapal Govindarajulu, Daniel Ortiz-Gonzalo, Rasmus Fensholt, Sebastian Schnell, Christoph Kleinn, and Göran Ståhl. Severe decline in large agroforestry trees in india over the past decade. *In Review*, 2023.

Agnès Bégué, Damien Arvor, Beatriz Bellon, Julie Betbeder, Diego de Abelleyra, Rodrigo P. D. Ferraz, Valentine Lebourgeois, Camille Lelong, Margareth Simões, and Santiago R. Verón. Remote sensing and cropping practices: A review. *Remote Sensing*, 10(99), 2018.

Melissa Chapman, Wayne S. Walker, Susan C. Cook-Patton, Peter W. Ellis, Mary Farina, Bronson W. Griscom, and Alessandro Baccini. Large climate mitigation potential from adding trees to agricultural lands. *Global Change Biology*, 26(8):4357–4365, 2020.

CVAT.ai. Computer vision annotation tool (cvat).

Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*. PMLR, 2015.

Niall P. Hanan and Julius Y. Anchang. Satellites could soon map every tree on earth. *Nature*, 587: 42–43, 2020.

Drew E. Terasaki Hart, Samantha Yeo, Maya Almaraz, Damien Beillouin, Rémi Cardinael, Edenise Garcia, Sonja Kay, et al. Priority science can accelerate agroforestry as a natural climate solution. *Nature Climate Change*, 13:1179–1190, 2023.

J. Nicholas Hendershot, Alejandra Echeverri, Luke O. Frishkoff, James R. Zook, Tadashi Fukami, and Gretchen C. Daily. Diversified farms bolster forest-bird populations despite ongoing declines in tropical forests. *Proceedings of the National Academy of Sciences*, 120(37):e2303937120, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything: Towards scene parsing with large language models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139. PMLR, 2021.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations (ICLR)*, 2022.

Chunyuan Li, Haotian Liu, Liunian Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Houdong Hu, Zicheng Liu, Yong Jae Lee, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. In *Advances in Neural Information Processing Systems*, volume 35. Advances in Neural Information Processing Systems, 2022.

P.K. Ramachandran Nair and Dennis Garrity (eds.). *Agroforestry - The Future of Global Land Use*, volume 9 of *Advances in Agroforestry*. Springer, Dordrecht, 2012. ISBN 978-94-007-4675-6.

Arthur Ouaknine, Teja Kattenborn, Etienne Laliberté, and David Rolnick. Openforest: A data catalogue for machine learning in forest monitoring. *Cambridge University Press*, 2023.

PlanetLabs. Planet high-resolution monitoring. Technical report, Planet Labs, 2024. Technical description of the high-resolution monitoring service by Planet Labs.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Gyri Reiersen, David Dao, Björn Lütjens, Konstantin Klemmer, Kenza Amara, Attila Steinegger, Ce Zhang, and Xiaoxiang Zhu. Reforestree: A dataset for estimating tropical forest carbon stock with deep learning and aerial imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

Marc Russwurm, Sherrie Wang, Marco Körner, and David Lobell. Meta-learning for few-shot land cover classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2020.

Sebastian Schnell, Christoph Kleinn, and Göran Ståhl. Monitoring trees outside forests: a review. *Environmental Monitoring and Assessment*, 187(600), 2015.

Rohan Taori, Nicholas Carlini, Achal Dave, Benjamin Recht, Vaishaal Shankar, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems*. NeurIPS, 2020.

Compton J. Tucker, Martin Brandt, Pierre Hiernaux, Ankit Kariryaa, et al. Sub-continental-scale carbon stocks of individual trees in african drylands. *Nature*, 615:80–86, 2023.

Ben G. Weinstein, Sergio Marconi, Stephanie Bohlman, Alina Zare, and Ethan White. Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11:1309, 2019.

Michael Xie, Neal Jean, Marshall Burke, David Lobell, and Stefano Ermon. Transfer learning from deep features for remote sensing and poverty mapping. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI, 2020.

Xiangyu Zhao, Yicheng Chen, Xilin Li, Yuchen Liang, Jingbo Wang, Feng Wu, and Wenming Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

6 Supplementary Section 1: The conditions we used to filter images

1. Contained in State of Rajasthan.
2. Published 2021 or later.
3. Captured during the dry season, between March and June.
4. Contains greater than 60 percent cropland or pasture according to ESRI LULC product.
5. Manually filtered out images that are too blurry or low quality to detect trees accurately.

7 Supplementary Section 2: Training details

For our hyperparams, we use a Faster-RCNN with a resnet-101 backbone and feature pyramid network pre-trained on COCO with a batch size of 16, a learning rate of 0.01 that we divide by 2 every 2000 batches. For Grounding DINO and SAM, we use the default hyperparameters in the mm-detection and huggingface repos respectively.

8 Supplementary Table 1: Per-zone accuracies

Table 1: Comparison of R^2 Values for Grounding DINO and Brandt Product

Zone	Grounding DINO R^2	Brandt Product R^2
Hyper-Arid Western Plains	0.862	0.770
Transitional Inland Drainage	0.970	0.806
Semi-Arid Eastern Plains	0.916	0.746
Sub-humid Southern Plains	0.919	0.347
Irrigated Northern	0.818	0.589
Transitional Plains of Luni Basin	0.759	0.619
Flood-prone Eastern Plains	0.771	0.576
Humid Southeastern Plains	0.764	0.326