Depth Any Canopy: Leveraging Depth Foundation Models for Canopy Height Estimation

Depth Any Canopy: Leveraging Depth Foundation Models for Canopy Height Estimation

Daniele Rege Cambrin[1],
Isaac Corley[2], and
Paolo Garza[1]

[1] Politecnico di Torino, Torino Italy
{daniele.regecambrin,paolo.garza}@polito.it
[2] University of Texas at San Antonio, San Antonio, TX USA
isaac.corley@utsa.edu

Abstract. Estimating global tree canopy height is crucial for forest conservation and climate change applications. However, capturing high-resolution ground truth canopy height using LiDAR is expensive and not available globally. An efficient alternative is to train a canopy height estimator to operate on single-view remotely sensed imagery. The primary obstacle to this approach is that these methods require significant training data to generalize well globally and across uncommon edge cases. Recent monocular depth estimation foundation models have show strong zero-shot performance even for complex scenes. In this paper we leverage the representations learned by these models to transfer to the remote sensing domain for measuring canopy height. Our findings suggest that our proposed Depth Any Canopy, the result of fine-tuning the Depth Anything v2 model for canopy height estimation, provides a performant and efficient solution, surpassing the current state-of-the-art with superior or comparable performance using only a fraction of the computational resources and parameters. Furthermore, our approach requires less than $1.30 in compute and results in an estimated carbon footprint of 0.14 kgCO_2. Code, experimental results, and model checkpoints are openly available at github.com/DarthReca/depth-any-canopy.

Keywords: Canopy Height Maps · Remote Sensing · Monocular Depth Estimation

1 Introduction

Measuring tree canopy extent and height accurately is crucial to tracking the health of our world's ecosystem [36]. However, manual in-situ measurements or airborne laser scans (ALS) using LiDAR [1, 9] are expensive and slow, making them unscalable for acquiring global measurements [33]. Furthermore, each acquisition method comes with their own sensitivities and errors [29]. Consequently, there is a growing need for more accessible and cost-effective approaches to canopy height estimation.

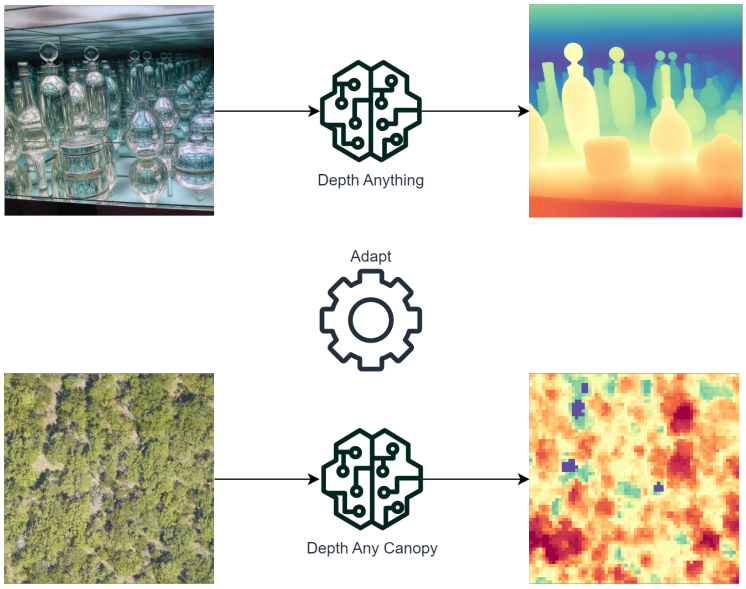


Fig. 1: From Depth Anything [39] to Depth Any Canopy. Depth Anything is a monocular depth estimation foundation model trained on natural imagery. We fine-tune and adapt Depth Anything v2 for the task of estimating tree canopy height in remote sensing imagery, resulting in Depth Any Canopy (DAC).

Remotely sensed satellite data is currently in abundance and an efficient alternative; however, satellite sensors built for measuring canopy height, such as GEDI [6], provide measurements which lack high spatial resolution. With this said, the most reasonable automated solution is to utilize machine learning to extract canopy height learned from satellite imagery geospatially aligned with ALS-derived canopy height maps [8].

One promising alternative is to develop a canopy height estimator that uses single-view remotely sensed image. However, this approach faces the substantial challenge of requiring extensive training data to achieve generalization across diverse global environments. Foundation models have displayed the importance of large-scale pretraining, which allows for the high-quality pretrained weights to be fine-tuned in a low-cost manner for various downstream tasks [4, 42].

The task of canopy height estimation from remote sensing imagery can be posed similarly to the depth estimation task, where the camera view is always fixed overhead and we seek to measure distance above ground, with ground represented as zero. Monocular depth estimation for natural imagery has recently experienced significant zero-shot performance improvements by pretraining on complex synthetic data and then pseudo-labeling a large-scale dataset for further semi-supervised training [38]. These models have shown the capability of being adapted for various applications, suggesting the opportunity for cross-domain transfer learning.

In this work, we explore the adaption of depth foundation models to remote sensing domain, specifically for tree canopy height estimation, without requiring significant amount of pretraining on remote sensing imagery. We propose leveraging Depth Anything v2 [39], a state-of-the-art monocular depth estimation model, to enhance the accuracy and efficiency of canopy height estimation from aerial imagery.

Our findings indicate that a proper finetuning of the model not only surpasses the current state-of-the-art performance but does so with fewer computational resources. Using this new model, Depth Any Canopy (DAC), we aim to provide a scalable and efficient solution for global canopy height estimation, making a step forward in providing high-quality canopy height data more accessible and contributing to better forest management and climate change mitigation efforts.

Our contributions can be summarized as follows:

1. We adapt monocular depth estimation foundation models, Depth Anything v2, fine-tuning them for canopy height estimation derived from remote sensing imagery, resulting in Depth Any Canopy (DAC).
2. We achieve superior or comparable results to a larger baseline with significant advantages of being pretrained on millions of in-domain satellite imagery.
3. Our resulting fine-tuning process and models are efficient and low-cost. They can be reproducible using minimal consumer GPU hardware.

The rest of the article is structured as follows. Section 2 discusses the related works in canopy height estimation and foundational models in computer vision and remote sensing. Section 3 describe the datasets employed in the study. Section 4 presents the models, the metrics, and adopted preprocessing. Section 5 presents the experimental results from both quantitative and qualitative perspectives. Section 6 draws the conclusion and presents future works.

2 Related Works

In this section, we discuss the related works in canopy height estimation and monocular depth estimation.

2.1 Canopy Height Estimation

Various techniques have been employed to estimate canopy height, ranging from traditional field measurements [40] to advanced remote sensing technologies [31]. One common tool is Light Detection and Ranging (LiDAR) due to its high spatial resolution and accuracy, allowing for detailed 3D representations of forest canopies [31]. The integration of LiDAR data with optical imagery and the application of machine learning algorithms have further enhanced canopy height estimation, thanks to the ability to deal with complex and heterogeneous environments [35]. The usage of LiDAR technologies is more expensive than RGB cameras, and generating canopy height maps (CHM) using deep learning models trained on remotely sensed satellite or aerial imagery has become popular

as methods have been able to provide accurate and higher-resolution CHM estimates. Lang et al. [16] trained a model to produce a low-resolution 10m global canopy height map by fusing GEDI and Sentinel-2 satellite imagery. Becker et al. [2] similarly created a low-resolution map by using Bayesian deep learning to estimate canopy height in fused Sentinel-1 SAR and Sentinel-2 optical imagery. Pauls et al. [23] improved upon these works by creating a novel shift resilient loss to adjust their ground truth for the intricacies of the GEDI global height product. Fogal et al. [7] created a 1.5m resolution multitemporal dataset using aligned SPOT satellite imagery and ALS acquisitions in France. Tolan et al. [32] took a self-supervised learning approach by pretraining a Vision Transformer (ViT) [5] using the DINOv2 [22] method on 18 million crops of globally sampled high-resolution WorldView satellite imagery. The authors then fine-tune this model to estimate canopy height using a dataset of U.S. based imagery and 1m resolution ALS-derived CHMs from the NEON catalog. While this approach works in practice, a significant amount of compute and efforts to collect these datasets are required.

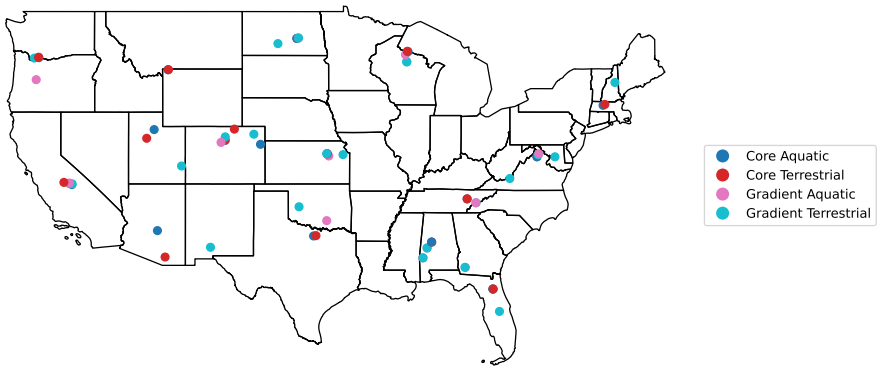


Fig. 2: National Ecological Observatory Network sites across the US. Aquatic sites collect information about aquatic ecosystems, while terrestrial ones collect data about terrestrial ecosystems. Core sites provide long-term support, while Gradient sites are temporary sites to study the ecological response to specific changes.

2.2 Foundation models in Computer Vision and Remote Sensing

Foundation models are becoming one of the prevalent solutions in many fields when there is no sufficient training time or large data availability [42], which is particularly relevant for transformers. In the computer vision field, ViT [5], and subsequent CLIP [24] and DINOv2 [22] have become well-established and robust encoders employed in many architectures and downstream tasks. Models

like Segment Anything [13] and OWL-ViT [21] make use of these advancements for general-purpose usage. In the remote sensing field, following these advancements, many solutions were proposed providing generative models [12], multimodal models [14], or task-specific solutions [19, 41]. With this said, foundation models remain unexplored for the task of canopy height estimation.

2.3 Monocular Depth Estimation

Monocular depth estimation for natural imagery, being a more mainstream area of computer vision research, experiences faster improvements than remote sensing height estimation topics. The MiDaS/DPT family of models [3, 17, 25] has shown that training for relative depth on a mixture of datasets, instead of training on small individual metric depth datasets such as NYU Depth V2 [30], ETH3D [28], and DIODE [34], experiences better transfer learning performance. MegaDepth [18] and Depth Anything [38, 39] have found that pretraining for relative depth estimation on higher quality and large-scale synthetic and pseudo-labeled images further improves zero-shot generalization performance.

Furthermore, Marigold [11] displayed the efficiency of requiring a single consumer NVIDIA RTX 4090 GPU for fine-tuning image generation diffusion models pretrained on large-scale datasets for the task of depth estimation. These advancements in natural imagery provide an interesting and promising alternative to the full training of models for canopy height estimation, and thanks to the pre-training, they provide solid baselines for predicting relative distances.

3 Datasets

In this section, we describe the data sources and the datasets employed for the analysis.

3.1 Data Sources

The primary data source for high-resolution CHM is derived from the National Ecological Observatory Network (NEON) catalog [10]. The network is composed of many sites across the U.S., as shown in Figure 2. Terrestrial sites collect data to monitor changes in climate, surface-atmosphere interactions, biogeochemical processes, organismal populations, and habitat structure over the land, while aquatic sites collect data to monitor changes in freshwater and biogeochemical processes, organismal populations, and habitat structure in water sources. Core sites provide long-term support, while Gradient sites are temporary sites to study the ecological response to specific changes. In our specific case, the data consists of high-resolution, multi-spectral, and hyperspectral aerial imagery captured from aircraft over various ecological sites. Multi-spectral imagery captures data in several specific wavelengths, typically including visible and near-infrared bands. Hyperspectral imagery captures data in hundreds of contiguous spectral bands. This imagery provides detailed spatial data on vegetation, land cover,

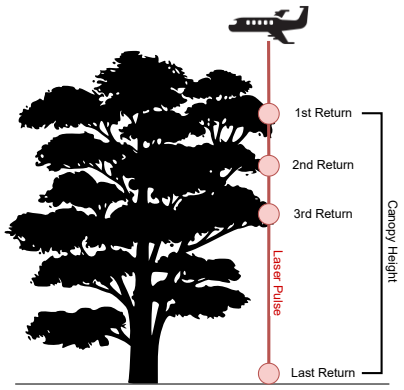


Fig. 3: ALS-derived tree canopy height acquisition process. LiDAR sensors attached to a fixed-wing aircraft, emits a laser pulse that reflects many times until reaching the ground. The canopy height is computed using the time-delta between the first and the last return of the pulse, the pulse obtained when reaching the ground level.

and water bodies, aiding in vegetation health assessment, species distribution, and biomass monitoring.

LiDAR sensors use laser light to measure distances to the Earth's surface and various objects from aircraft or drones. These sensors emit thousands of laser pulses per second toward the ground, capable of penetrating vegetation. These pulses hit various surfaces, such as the canopy, branches, and the ground, reflecting back to the sensor and providing punctual information. Points representing the highest surfaces in the forest canopy are used to create a Digital Surface Model (DSM), including the canopy's top. The Canopy Height Model (CHM) is derived by subtracting the Digital Terrain Model (DTM), which represents the bare earth surface, from the DSM. This provides the height of the canopy above the ground surface [26] as shown in Figure 3.

3.2 EarthView dataset

The Satellogic EarthView Dataset [27] is a comprehensive collection of multispectral earth imagery for general-purpose tasks. The dataset is divided into four distinct subsets sourced from Satellogic, Sentinel-1, Sentinel-2 satellites and aerial imagery and ALS-derived CHMs from the NEON catalog. While the dataset provides satellite imagery provided by one commercial satellite (Satellogic) and two publicly available European Space Agency missions (Sentinel-1 and Sentinel-2), we employed only the aerial imagery, which is the only one with canopy height maps by LiDAR sensor, and it is the common imagery type between the two

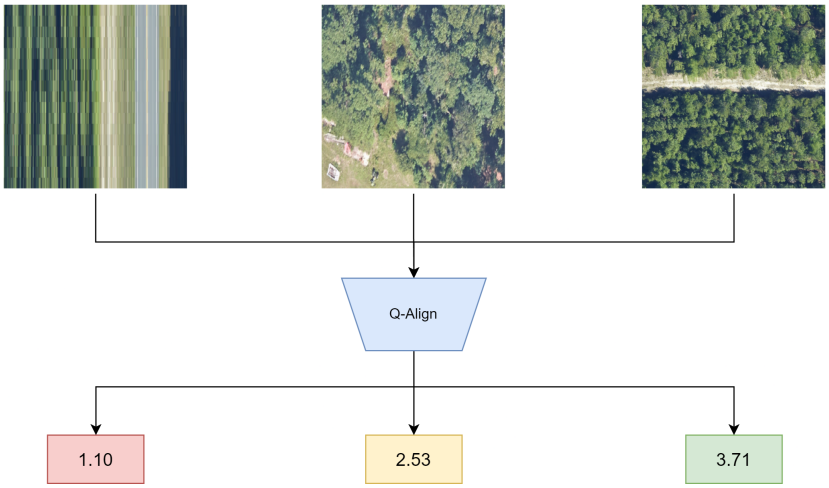


Fig. 4: Example quality scores by Q-Align [37] on NEON RGB images from the EarthView dataset. On the left, a noisy sample scored *1.10*; in the middle, a medium-quality sample scored *2.53*; and on the right, a better-quality sample scored *3.71*. This solution permits the detection and filtering of low-quality samples affected by warping and motion blurs.

analyzed datasets. The NEON subset is composed of very high-resolution RGB images at 0.1m paired with 1m CHMs.

3.3 High Resolution Canopy Height Maps dataset

The High-Resolution Canopy Height Maps (HRCHM) [32] is a collection of approximately 5,800 NEON CHMs with an area of 1 km × 1 km and a resolution of 1 meter per pixel. Samples were selected based on quality, minimal artifacts, and temporal proximity to a collection of Maxar satellite imagery. The CHMs were reprojected and resampled to match the Maxar imagery resolution and paired with a corresponding RGB satellite images to train a canopy height estimation model. The dataset was split into a train and test set, referred to as the *NEON test set*, and preprocessed into 256 × 256 random crops.

4 Methodology

This section outlines the model, preprocessing, and metrics used in the analysis.

4.1 Problem Statement

Let I be an arbitrary RGB image of size $W \times H \times 3$, where W and H are the width and height of the image in pixels, respectively, and the channel depth of

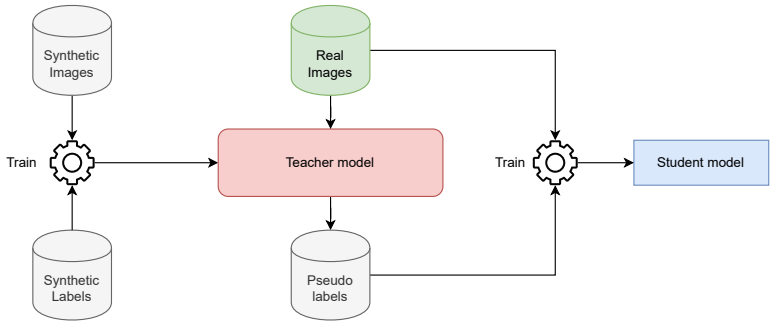


Fig. 5: Depth Anything v2 training procedure. Synthetic images and relative labels are used to train a large teacher model. It is employed to annotate real images to create pseudo-labels. The real images and relative pseudo-labels are used to train a small student model.

the image is 3, corresponding to the red, green, and blue color channels. The objective is to estimate the canopy height map M from this image. M can be represented by a matrix of size $W \times H$, where each element $M_{i,j}$ represents the estimated height of the canopy for the corresponding pixel $I_{i,j}$ in the image I.

4.2 Preprocessing

The NEON imagery in the EarthView dataset contains a significant number of samples with motion blur or warping due to image stitching errors. We find that these samples significantly degrade fine-tuning performance and are likely unuseful for evaluation. Therefore, we seek to filter these samples by employing the Q-Align [37] image quality assessment vision-language model to score the quality of each sample between a range of $[0 - 5]$ where higher values indicate better quality as shown in Figure 4. We then filter low-quality samples using a score threshold of $t = 2.5$ (which indicates above-average quality). The training set contains 45781 samples, validation set 5788, and test set 5682. Every sample contains at least one pixel with the canopy height different from zero. According to KS-test the distribution of heights is similar between train-val ($p \approx 0.68$) and train-test ($p \approx 0.75$)

4.3 Depth Anything for Canopy Height Estimation

In our study, we used Depth Anything v2 in both zero-shot and with a finetuning on EarthView dataset, called Depth Any Canopy, for addressing the canopy height estimation.

Depth Anything [38] is a state-of-the-art monocular depth estimation model. It features a DINOv2 pretrained ViT encoder [22] and DPT decoder [25]. The model is trained using an online student-teacher distillation with a combination

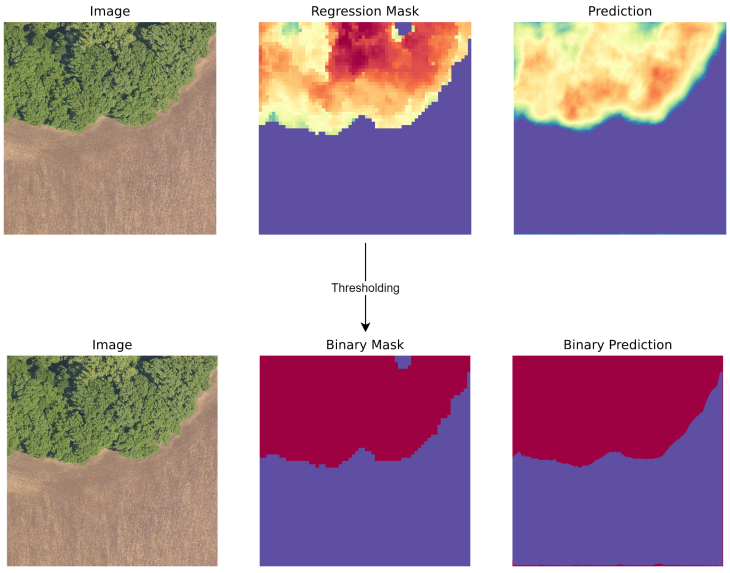


Fig. 6: Tree Canopy Extent evaluation process. We threshold the ground truth and predicted canopy heights to obtain binary masks. We utilize this to evaluate a model's effectiveness in identifying tree extent.

of real depth maps and pseudo-labeled depth maps of 62 million imagery. It excels in zero-shot generalization and improves performance when fine-tuned for metric depth on evaluation benchmark datasets.

The successor, Depth Anything v2 [39], takes a different approach and only pretrains on high-quality synthetic image and depth map pairs due to invalid pixels and low complexity of scenes in canonical metric depth datasets. It then uses the synthetic pretrained model to generate precise pseudo-labels of 62+ million real images in an offline manner. A student model is then trained on these pseudo-labels as shown in Figure 5. The model uses scale-and-shift invariant losses to ensure consistency and gradient matching loss for sharper depth predictions [25]. This model not only outperforms previous state-of-the-art models in accuracy, but also speed, and efficiency primarily due to the large-scale and high-quality pretraining dataset.

4.4 Baseline

As described in Section 2, Tolan et al. [32] proposed to pretrain a model using the DINOv2 on a large-scale dataset of 18 million 256 $\times$ 256 RGB crops from global Maxar WorldView satellite imagery. This model was then fine-tuned on the HRCHM dataset described in Section 3.3.

Table 1: Results on EarthView and HRCHM datasets. The best for each metric is bolded, while the second-best is underlined. *FT* indicates if the model is finetuned on EarthView. *DA* and *DAC* refers to Depth Anything v2 and Depth Any Canopy, while *DAC-S* and *DAC-B* refers to the ViT-S and ViT-B variants, respectively.

Model	FT	# Params	GFLOPs	EarthView [27]			HRCHM [32]		
				MAE ↓	IoU ↑	PC ↑	MAE ↓	IoU ↑	PC ↑
SSL-H [32]	✗	677M	414	0.2236	0.4164	0.1544	**0.0306**	0.485	**0.7441**
DA-S [39]	✗	**24.8M**	**115**	0.4116	0.4164	0.2892	0.5960	**0.6474**	0.1791
DA-B [39]	✗	97.5M	381	0.4607	0.4164	**0.361**	0.5972	**0.6474**	0.1692
DAC-S [39]	✓	**24.8M**	**115**	0.1410	0.5323	0.2740	0.1025	0.5672	0.6102
DAC-B [39]	✓	97.5M	381	**0.1304**	**0.5926**	0.3483	0.1203	0.5494	0.6171

4.5 Metrics

In our experimentation, we analyze the predicted performance on the canopy height estimation regression task using the Mean Absolute Error (MAE) (L1 distance). Furthermore, we evaluate the ability of the model to segment the tree canopy from the background by using the same process by Tolan et al. [32] to mask the predicted and ground truth CHM using a threshold of $t = 1E - 4$ (to include also very small trees which were excluded by the previous work) as shown in Figure 6. We then compute binary segmentation performance using the Intersection-Over-Union (IoU) metric to understand the quality of the predictions of areas considered as trees, tree canopy extent. Lastly we compute the Pearson Correlation (PC) of prediction and ground truth in the tree areas to understand whether relative heights are consistent despite the errors. To compare model size and efficiency, we report the floating-point operations per second (FLOPs) in billions (giga) and the number of parameters in millions.

5 Experimental Results

In this section, we present the experimental settings and an analysis of the results.

5.1 Experimental Settings

The models were finetuned with the AdamW optimizer [20] for 3 epochs (until plateau) with a batch size of 8 on the EarthView training set using an NVIDIA RTX A6000 48GB GPU. We employed a learning rate scheduler with a warmup of 5% of the training steps with linear decay. The maximum learning rate is set to $\alpha = 5E - 6$. We utilize the Mean Squared Error (L2 distance) between the ground truth and predicted canopy height as a loss function. The canopy height maps are min-max normalized (at dataset level) to relative height since EarthView does not provide height in meters, only a relative distance.

We evaluate and fine-tune the Depth Anything v2 ViT-S and ViT-B checkpoints, referred to as DA-S and DA-B, respectively into Depth Any Canopy (DAC-S) and (DAC-B). We do not utilize the best performing ViT-G (Giant) weights because they have not been released. We baseline against the ViT-H (Huge) checkpoint from Tolan et al. [32], referred to as SSL-H, trained using the process described in Section 4.4. We utilize this checkpoint because no other smaller ViT variants of this model were made available.

We note that the SSL-H baseline is comparatively larger than the Depth Anything v2 checkpoints we initialize from. Furthermore, SSL-H is already pretrained in-domain on 18 million remote sensing images and fine-tuned on a large set of NEON aerial RGB imagery for canopy height estimation. Therefore, the Depth Anything v2 weights should be at a significant disadvantage when comparing to this baseline.

5.2 Discussion

Overall Results We tested the SSL-Huge (SSL-H) [32] on both datasets (EarthView and HRCHM) to understand the performance and generalization capabilities. Depth-Anything (DA) was tested on both datasets in zero-shot and after finetuning on EarthView. This approach provides a comparison of both in-domain and out-of-domain data to understand the capability of generalization.

In Table 1, we reported the results obtained by each tested model. The fine-tuning of Depth Anything v2 on EarthView provides good comparative results and consistent performance under both MAE and IoU metrics, while the PC is generally higher on HRCHM for SSL-H.

SSL-H proves to be best in terms of MAE and PC on HRCHM, while the IoU is lower than the one achieved by DA. The performance is worse on EarthView, particularly on MAE, which decreases by 10 times with respect to the performance on HRCHM. Additionally, we can reach good performance with smaller models.

Zero-shot Comparison When comparing Depth Anything v2 in zero-shot to Depth Any Canopy, we can conclude it is necessary to adapt the model to this new unseen task to be an effective solution. For example, shows a reduction in MAE from 0.4116 to 0.1410 on EarthView and from 0.5960 to 0.1025 on HRCHM after fine-tuning DA-S to DAC-S. However, the main advantages in terms of resource efficiency are maintained by providing good results in fewer GFLOPs. In many cases, DAC-S provides the best or second-best performance in around 1/4 of the GFLOPs and 1/27 of the parameters of SSL-H [32]. DAC-S achieves results comparable to DAC-B, indicating it is a preferable solution for the task due to its low resource demands.

Carbon Footprint Analysis As mentioned in Section 5.1, to achieve similar or comparable results to SSL-H by Tolan et al. [32], our DAC fine-tuning requires simply 3 epochs on the EarthView dataset using an NVIDIA RTX A6000 GPU. This requires 1.5 and 2.61 hours at an estimated cost of $1.24 and $2.09 using the

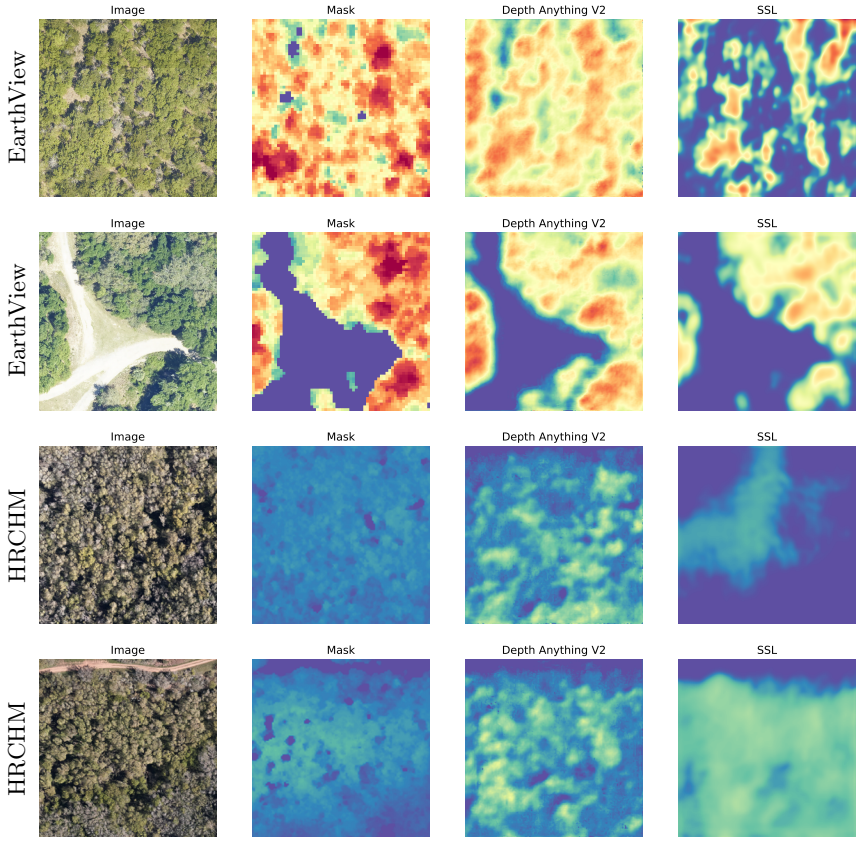
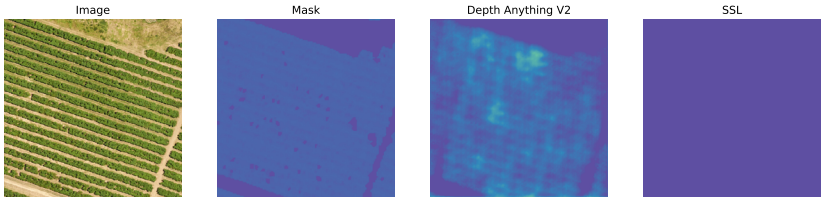


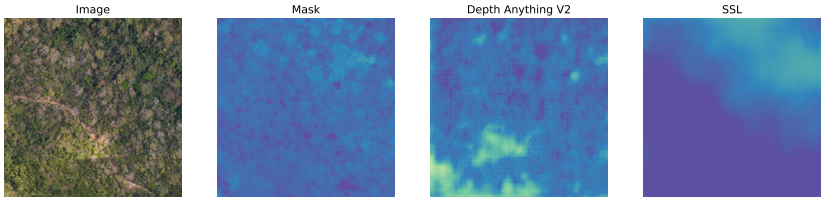
Fig. 7: Example predictions of Depth Any Canopy (DAC-S) and SSL-Huge model from Tolan et al. [32] on NEON imagery from the EarthView and HRCHM datasets. Left to right: NEON RGB Image, Ground Truth Canopy Height Map, DAC-S predicted CHM and SSL-H predicted CHM.

Lambda Labs hourly pricing for the DAC-S and DAC-B variants, respectively. We calculate that our fine-tuning results in an estimated 0.14 and 0.24 kg of CO_2 emissions of using the ML CO_2 Impact calculator [15]. We believe this to be preferable to the 8kg of CO_2 emissions reported by our SSL-H baseline for fine-tuning, not considering the 1.8T of CO_2 emissions for pretraining.

Qualitative Results Figure 7 provides a qualitative example of predictions by DAC-S and SSL-H on both datasets. On EarthView, SSL-H does not recognize many trees as expected, while DAC-S provides better performance. On HRCHM, we can note that SSL-H underestimates or overestimates some areas, although providing good IoU performance. The variations captured by DAC are generally superior to the ones captured by SSL-H.



(a) Example 1



(b) Example 2

Fig. 8: Examples of limitations of SSL-H addressed by Depth Any Canopy. Through our analysis, we find that the SSL-H model by Tolan et al. [32] tends to generate overly smooth predictions or predicts zero height for smaller vegetation. Because Depth Any Canopy is fine-tuned from the Depth Anything v2 weights it is able to recover vegetation heights for complex scenes and edge cases.

Figure 8 provides an overview of other cases. SSL-H avoids prediction on low trees, as in this qualitative example, while DAC-S tries to provide a height map of the area with a good level of detail. This shows the robustness achieved by Depth Any Canopy. The model can deal with complex scenarios, providing a high level of detail for complex scenes.

6 Conclusion

In this work, we have presented a novel approach to canopy height estimation by leveraging Depth Anything v2, a state-of-the-art monocular depth estimation foundation model. Our proposed model, Depth Any Canopy, demonstrates superior or comparable performance to current state-of-the-art methods with fewer computational resources. This is crucial for a scalable and cost-effective global canopy height estimation solution. Our findings show the potential of depth estimation foundation models pre-trained on large-scale natural imagery. They can be adapted for specific tasks in the remote sensing domain with minimal additional training. By fine-tuning Depth Anything v2, we have shown that it is possible to achieve high-quality canopy height maps from single-view images, overcoming the limitations of expensive pre-training on large-scale satellite imagery datasets to achieve comparable performance. In future works, we plan to expand the evaluation to a wider variety of forest biomes and geographical regions to account for more diverse environments and to investigate the usage of

hyperspectral and radiometric imagery, which could enhance the understanding of the area with complex features.

References

1. Andersen, H.E., Reutebuch, S.E., Schreuder, G.F.: Automated individual tree measurement through morphological analysis of a lidar-based canopy surface model. In: Proc. of the 1st International Precision Forestry Symposium. pp. 11–21 (2001)
2. Becker, A., Russo, S., Puliti, S., Lang, N., Schindler, K., Wegner, J.D.: Country-wide retrieval of forest structure from optical and sar satellite imagery with deep ensembles. ISPRS Journal of Photogrammetry and Remote Sensing **195**, 269–286 (2023)
3. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
4. Cha, K., Seo, J., Lee, T.: A billion-scale foundation model for remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing pp. 1–17 (2024). https://doi.org/10.1109/JSTARS.2024.3401772
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2021), https://arxiv.org/abs/2010.11929
6. Dubayah, R., Armston, J., Healey, S.P., Bruening, J.M., Patterson, P.L., Kellner, J.R., Duncanson, L., Saarela, S., Ståhl, G., Yang, Z., et al.: Gedi launches a new era of biomass inference from space. Environmental Research Letters **17**(9), 095001 (2022)
7. Fogel, F., Perron, Y., Besic, N., Saint-André, L., Pellissier-Tanon, A., Schwartz, M., Boudras, T., Fayad, I., d'Aspremont, A., Landrieu, L., et al.: Open-canopy: A country-scale benchmark for canopy height estimation at very high resolution. arXiv preprint arXiv:2407.09392 (2024)
8. Fogel, F., Perron, Y., Besic, N., Saint-André, L., Pellissier-Tanon, A., Schwartz, M., Boudras, T., Fayad, I., d'Aspremont, A., Landrieu, L., Ciais, P.: Open-canopy: A country-scale benchmark for canopy height estimation at very high resolution (2024), https://arxiv.org/abs/2407.09392
9. Gaveau, D.L., Hill, R.A.: Quantifying canopy height underestimation by laser pulse penetration in small-footprint airborne laser scanning data. Canadian Journal of Remote Sensing **29**(5), 650–657 (2003)
10. Kampe, T.U., Johnson, B.R., Kuester, M.A., Keller, M.: Neon: the first continental-scale ecological observatory with airborne remote sensing of vegetation canopy biochemistry and structure. Journal of Applied Remote Sensing **4**(1), 043510 (2010)
11. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9492–9502 (2024)
12. Khanna, S., Liu, P., Zhou, L., Meng, C., Rombach, R., Burke, M., Lobell, D.B., Ermon, S.: Diffusionsat: A generative foundation model for satellite imagery. In: The Twelfth International Conference on Learning Representations (2023)
13. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4015–4026 (2023)

14. Kuckreja, K., Danish, M.S., Naseer, M., Das, A., Khan, S., Khan, F.S.: Geochat: Grounded large vision-language model for remote sensing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27831–27840 (2024)
15. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)
16. Lang, N., Jetz, W., Schindler, K., Wegner, J.D.: A high-resolution canopy height model of the earth. Nature Ecology & Evolution **7**(11), 1778–1789 (2023)
17. Lasinger, K., Ranftl, R., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. arXiv preprint arXiv:1907.01341 (2019)
18. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2041–2050 (2018)
19. Lin, J., Gao, F., Shi, X., Dong, J., Du, Q.: Ss-mae: Spatial–spectral masked autoencoder for multisource remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing **61**, 1–14 (2023). https://doi.org/10.1109/TGRS.2023.3331717
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., et al.: Simple open-vocabulary object detection. In: European Conference on Computer Vision. pp. 728–755. Springer (2022)
22. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: Dinov2: Learning robust visual features without supervision (2024), https://arxiv.org/abs/2304.07193
23. Pauls, J., Zimmer, M., Kelly, U.M., Schwartz, M., Saatchi, S., Ciais, P., Pokutta, S., Brandt, M., Gieseke, F.: Estimating canopy height at scale. arXiv preprint arXiv:2406.01076 (2024)
24. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
25. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
26. Rosette, J., North, P., Suarez, J.: Vegetation height estimates for a mixed temperate forest using satellite laser altimetry. International journal of remote sensing **29**(5), 1475–1493 (2008)
27. Satellogic: Earthview. https://huggingface.co/datasets/satellogic/EarthView, accessed: 06-07-2024
28. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3260–3269 (2017)

29. Sexton, J.O., Bax, T., Siqueira, P., Swenson, J.J., Hensley, S.: A comparison of lidar, radar, and field measurements of canopy height in pine and hardwood forests of southeastern north america. Forest Ecology and Management **257**(3), 1136–1147 (2009)
30. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12. pp. 746–760. Springer (2012)
31. Simard, M., Pinto, N., Fisher, J.B., Baccini, A.: Mapping forest canopy height globally with spaceborne lidar. Journal of Geophysical Research **116**(G4) (Nov 2011). https://doi.org/10.1029/2011jg001708, http://dx.doi.org/10.1029/2011JG001708
32. Tolan, J., Yang, H.I., Nosarzewski, B., Couairon, G., Vo, H.V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., et al.: Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. Remote Sensing of Environment **300**, 113888 (2024)
33. Tomppo, E., Gschwantner, T., Lawrence, M., McRoberts, R.E., Gabler, K., Schadauer, K., Vidal, C., Lanz, A., Ståhl, G., Cienciala, E., et al.: National forest inventories. Pathways for Common Reporting. European Science Foundation **1**, 541–553 (2010)
34. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
35. Wang, S., Liu, C., Li, W., Jia, S., Yue, H.: Hybrid model for estimating forest canopy heights using fused multimodal spaceborne lidar data and optical imagery. International Journal of Applied Earth Observation and Geoinformation **122**, 103431 (Aug 2023). https://doi.org/10.1016/j.jag.2023.103431, http://dx.doi.org/10.1016/j.jag.2023.103431
36. Watch, G.F.: Global forest watch. World Resources Institute, Washington, DC Available from http://www. globalforestwatch. org (accessed March 2002) (2002)
37. Wu, H., Zhang, Z., Zhang, W., Chen, C., Liao, L., Li, C., Gao, Y., Wang, A., Zhang, E., Sun, W., Yan, Q., Min, X., Zhai, G., Lin, W.: Q-align: Teaching lmms for visual scoring via discrete text-defined levels (2023), https://arxiv.org/abs/2312.17090
38. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10371–10381 (2024)
39. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2 (2024), https://arxiv.org/abs/2406.09414
40. Zang, J., Jin, S., Zhang, S., Li, Q., Mu, Y., Li, Z., Li, S., Wang, X., Su, Y., Jiang, D.: Field-measured canopy height may not be as accurate and heritable as believed: evidence from advanced 3d sensing. Plant Methods **19**(1) (Apr 2023). https://doi.org/10.1186/s13007-023-01012-2, http://dx.doi.org/10.1186/s13007-023-01012-2
41. Zheng, Z., Zhong, Y., Zhang, L., Ermon, S.: Segment any change. arXiv preprint arXiv:2402.01188 (2024)
42. Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., Zhang, K., Ji, C., Yan, Q., He, L., Peng, H., Li, J., Wu, J., Liu, Z., Xie, P., Xiong, C., Pei, J., Yu, P.S., Sun, L.: A comprehensive survey on pretrained foundation models: A history from bert to chatgpt (2023)