Comparing DeepLabv3+ and Depth Anything V2 on Canopy Height Model Prediction on a Continental Scale Dataset of Australia

Kevin Qiu[1,*], Rewanth Ravindran[2], Nicolas Pucino[3,4], Dimitri Bulatov[1],
Shaun Levick[5], Martin Brandt[6], Dorota Iwaszczuk[2], Tim R. McVicar[4]

[1]Fraunhofer IOSB, 76725 Ettlingen, Germany - (kevin.qiu, dimitri.bulatov)@iosb.fraunhofer.de
[2] Technical University of Darmstadt, Darmstadt, Germany - (rewanth.ravindran, dorota.iwaszczuk)@tu-darmstadt.de
[3] Fenner School of Environment and Society, Australian National University, Canberra, ACT, Australia
[4] CSIRO Environment, ACT, SA, Australia
[5] CSIRO Environment, Urrbrae, SA, Australia
[6] Department of Geosciences and Natural Resource Management, University of Copenhagen, Copenhagen, Denmark

Keywords: CHM, Foundation Models, PlanetScope, Regression, Monocular Depth Estimation

Abstract

Canopy height models (CHMs) are raster maps representing normalized tree canopy height above ground and are often used as co-products for estimating carbon storage, forest degradation, and biodiversity at regional to global scales. While airborne LiDAR delivers the most accurate canopy height (CH) measurements, its high cost and limited temporal coverage motivate the use of spaceborne (multispectral) imagery combined with machine learning. In this study, we compare two distinct deep-learning approaches for continental-scale CHM estimation from 3 m PlanetScope imagery: (1) a CNN-based regression model (DeepLabv3+), and (2) a monocular depth-estimation model (Depth Anything V2) based on a foundation model. We train/fine-tune both models on a curated dataset of 16,973 pairs of airborne point cloud-derived CHMs and PlanetScope imagery of Australia using a stratified sampling scheme to ensure balanced representation of vegetation structural classes. We then evaluate their generalizability on independent validation sets across Australia, across different heights, and under limited-data scenarios. Through extensive quantitative and qualitative analysis, we show that the DeepLab-based regression model outperforms Depth Anything across all evaluation metrics, partly because it can incorporate additional spectral channels. DeepLab also learns more effectively from less data. On our dataset, the conventional CNN-based regression model performs better than the fine-tuned foundation model.

1. Introduction

Canopy height (CH) is a key structural measure, indicator of carbon storage, and is also in the priority list of biodiversity metrics to observe from space, being recognized as an essential biodiversity variable (Skidmore et al., 2021). Accurate large-scale CH measurements are needed to monitor forest degradation, landscape restoration and above-ground woody biomass estimation for carbon emission and sequestration modeling (Tolan et al., 2024). The most accurate CH estimates over local to regional scales are obtained with airborne light detection and ranging (LiDAR) surveys (Duncanson et al., 2022). At continental to global scales, CHM predictors are needed. Inter-intra regional landscape variations imply varying vegetation types and structures, which challenge the generalizability of the predictors, thus requiring a high degree of flexibility (Kattenborn et al., 2021). Convolutional Neural Networks (CNNs) have large model capacities, as the enormous quantity of their trainable parameters can better adapt to different feature distributions of training data (Lang et al., 2022). As a consequence, since 2019, to derive CHM at continental scales, machine learning and CNN or transformer-based models are being increasingly used, often in combination with optical high or very high spatial resolution (VHR) multispectral imagery (Kattenborn et al., 2021). VHR satellite imagery (spatial resolution under 5 m) such as provided by Planet Lab Doves (1 to 3 m) or DigitalGlobe's WorldView 1-3 (0.5 to 0.3 m), GeoEye-1 (0.4 m) or the WorldView Legion (0.3 m) satellites provide good opportunities for forest monitoring in general (Dupuis et al., 2020), individual tree-crown detection (Brandt et al., 2020; Bulatov et al., 2016; Reiner et al., 2023; Wagner et al., 2018), crop mapping (Bégué et al., 2018) and canopy height modeling (Illarionova et al., 2022; Liu et al., 2023; Tolan et al., 2023).

Methodically, we can interpret canopy height prediction as a regression task, where for each pixel, one continuous value (vertical height above ground) is predicted. Alternatively, CHM prediction can be framed as a monocular depth estimation task, where the output is a dense depth map, the relative or absolute distance from the camera. By changing the depth map to be the distance from the ground, a CHM can be predicted as well. For this work, we chose two very different approaches. First, a well known and commonly used CNN based model DeepLabv3+ (Chen et al., 2018) for regression, and second, recent foundation model for monocular depth estimation on RGB images based on a DINOv2 backbone (Oquab et al., 2024), Depth Anything V2 (DAv2) (Yang et al., 2024). While this model, trained on non-aerial images, can be applied directly to metric depth estimation on ground level images, the CHM prediction from satellite imagery is too specific a task, so the model requires fine-tuning in that domain. DeepLab on the other hand is easily extendable with additional channels (such as infrared) and is trained from the ground up (with ImageNet initialization). We train DeepLab and fine-tune Depth Anything continental-scale models with these approaches using 16,973 aerial laser scanning (ALS) CHM of 1 to 4 km² to predict CHMs from 3 m resolution multispectral Planet imagery. Finally, we validate on independent areas spread across Australia. We aim to answer three questions: (1) Does a fine-tuned depth estimator based on

* Corresponding author

a foundation model return better CH estimation than a traditionally trained regression model with additional channels? (2) Which model works best on independent validation data and different height classes? (3) Finally, for both methods, we look at the effect of limited training data.

2. Related Work

2.1 Large-scale canopy height estimation

Several satellite-derived CHMs have been generated with deep learning-based monocular depth estimation at regional scales, such as in French Guiana (Lahssini et al., 2024), in the Amazon rain forest (Wagner et al., 2025), in California (Dixon et al., 2025), or even at smaller-scale applications such as power line corridor mapping (Almeida et al., 2022). However, given difficulties in grasping variability at scale, only a few key studies attempted to predict CHM at continental or global scales and are reported here. Liao et al. (2020) produced a continental-scale (Australia) 25 m spatial resolution CHM using random forest models, with input data annual Landsat geomedians as well as altimetry training data from ICESat/GLAS and radar data from ALOS PALSAR. As training data, they used LiDAR-derived CHM at 1 m or 2 m resolution, later resampled to 50 m to fit the training pipeline. Potapov et al. (2021) generated a globally available 30 m resolution CHM using pixel-based bagged regression tree ensemble models using 546 spatiotemporal Landsat-based metrics and training data with GEDI-based height metrics as input (Potapov et al., 2021). Lang et al. (2022) produced a global CHM at 10 m resolution by training an ensemble of fully convolutional neural networks with height data derived from the GEDI instrument. Liu et al. (2023) produced a 3 m CHM for Europe with 4-band inputs (blue, green, red, infrared) PlanetsScope imagery and 0.2 to 2.5 m ALS CHMs (resampled to 3 m) as well as Lang et al. (2022) CHM as extra input to give a prior knowledge of height distribution to the model. They trained an U-Net architecture with EfficientNet-B4 backbone.

Tolan et al. (2024) produced a globally available 1 m resolution CHM with a combination of DINOv2 (Oquab et al., 2024) vision transformer self-supervised learning followed by a dense prediction transformer supervised regression, using as inputs 3-bands (red, blue, green) MAXAR (now Vantor) 0.5 m imagery, mainly trained with 5800 1 m resolution ALS-derived CHMs from the National Ecological Observatory Network (NEON) in the USA. Lastly, Weber et al. (2025) recently generated a 10 m resolution global CHM (as well as above-ground biomass and canopy cover) fusing Sentinel-2 optical and thermal bands, Sentinel-1 synthetic aperture VV and VH radar signals, Shuttle Radar Topography Mission (SRTM) elevation, aspect, and slope.The model was a ResNet-50–based CNN with a feature pyramid network decoder and multiple 1×1 convolutional prediction heads which was trained with GEDI-derived canopy height.

2.2 Regression for CHM

Canopy height prediction is an image-to-image regression task, where each (super-)pixel is assigned a regression value. Traditionally, this was done using random forests or support vector machines (Jin et al., 2018). With deep learning, networks for semantic segmentation are used and applied for pixel-wise regression (instead of classification). For CHM prediction, U-Net is very popular. For example, Liu et al. (2023) uses U-Net on 3 m Planet imagery in conjunction with airborne LiDAR data. Kenzhebay et al. (2025) introduces a Semantically-aware Tree Height Estimator (SaTHE), also based on Attention-U-Net. The near infrared channel (NIR) captures vegetation signatures, and Illarionova et al. (2022) show, that the addition of NIR improve the CHM prediction using U-Net.

Vision transformers have shown improved performance in computer vision tasks over CNNs. However, when supervised learning is employed using LiDAR-derived ground truth, the availability of such reference data can be limited or costly to acquire. In contrast, self-supervised approaches leveraging publicly available satellite imagery mitigate this limitation by reducing dependence on LiDAR data. Fogel et al. (2025) found that CNN based models are able to outperform some ViT models in CHM prediction. Furthermore, pretraining on natural images such as ImageNet performs better than foundation models trained on extensive databases. They also found that models pre-trained on other satellite images also do not perform better after fine-tuning. They hypothesized that this is because of the spatial domain shift and differences in the use of IR. The most recently released DINOv3 (Siméoni et al., 2025) offers a pre-trained backbone on 0.6m Vantor (previously MAXAR) satellite imagery, much higher than the 3m resolution of Planet.

2.3 Monocular depth estimation

Monocular depth estimation aims to infer a dense depth map from a single RGB image, enabling metric depth or scale-agnostic depth predictions from monocular sensors. Early approaches relied on CNNs with encoder–decoder architectures (e.g., U-Net or Attention U-Net) to regress depth in a supervised manner, typically under in-domain settings where training and test data share the same distribution. These constraints motivated a shift toward ViT-based architectures and self-supervised learning, where transformer backbones, such as DINOv2, demonstrated improved representational capacity for monocular geometry (Oquab et al., 2024; Yao et al., 2024). Building on these developments, foundation-model approaches such as Marigold (Ke et al., 2025) and Depth Anything (Yang et al., 2024) emphasize cross-dataset robustness through large-scale pseudo-labeled, synthetic, or unlabeled real-image corpora, enabling strong zero-shot relative depth performance across diverse domains (Ranftl et al., 2020). Collectively, these advances illustrate a clear trajectory from CNN-based supervised metric depth prediction toward self-supervised, transformer-driven foundation models capable of producing highly transferable, scale-aware depth estimates across seen and unseen visual domains.

Depth Anything (DAv2) is a transformer-based monocular depth estimation model (Yang et al., 2024) built upon the DINOv2 foundation model and pre-trained on the VKITTI2 and Hypersim datasets for synthetic outdoor and indoor depth understanding, respectively. Recent studies demonstrated good adaptability of DAv2 for various downstream tasks, including those in remote sensing applications. For example, Günaydın et al. (2025) evaluated DAv2 on the task of bathymetric estimation from Sentinel-2 imagery. For tree height estimation, Cambrin et al. (2024) fine-tuned DAv2 for canopy height modeling.

3. Dataset

3.1 Balancing vegetation structure

Given that vegetation types and structures are not homogeneously distributed across Australia, a random selection of training samples from the totality of point cloud datasets available would likely introduce sampling errors and degrade generalizability. Therefore, in this study, the CHMs used as training data have been generated using the pit-free algorithm (Khosravipour et al., 2014) at a resolution of 0.5 m from optimally selected point cloud tiles to statistically represent the vegetation structure distribution of Australia, thus representing a balanced dataset (Figure 1).

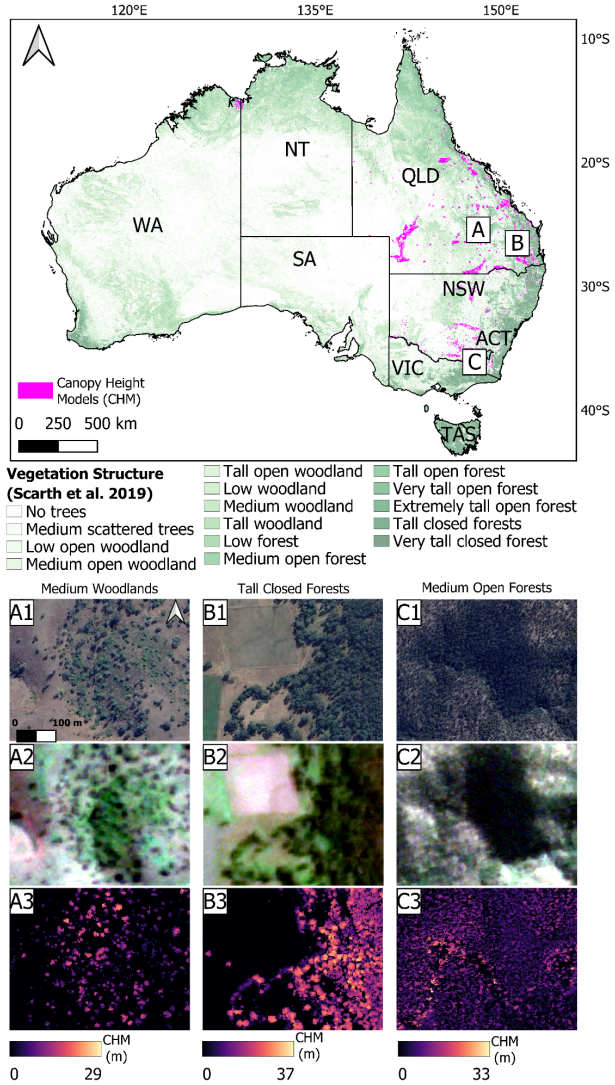


Figure 1. Location of the CHM tiles selected to represent Australia’s vegetation structure variability in our training dataset. Vegetation classes are displayed in shades of greens. Insets A1, B1 and C1 (·1)) represent 0.5 m Vantor imagery of “medium woodlands”, “tall open forests” and “medium open forests” locations, respectively, with their Planet imagery (·2) and point clouds-derived CHMs (·3) shown.

We initially subsampled 5% of all the point cloud data publicly available in Australia through the Elevation and Depth (ELVIS) and Terrestrial Ecosystem Research Network (TERN) databases combined. Given the vegetation-class-specific target coverages needed to obtain a balanced representation, we computed, for each tile, the fractional coverage of vegetation structural classes using the national vegetation structure map from Scarth et al. (2019). We then formulated a mixed-integer linear optimization problem to identify the smallest subset of tiles whose combined area matched the vegetation class target coverages within a 1% tolerance. The optimization was solved with the COIN-OR branch-and-cut mixed-integer programming solver (Boas et al., 2019) implemented in the open-source Python library PuLP, minimizing the number of selected tiles while preserving the proportional representation of all vegetation classes. This approach yielded an initial subset of 26,928 point cloud tiles, which we further refined by selecting only tiles with trees (min. height 2 m) and crowns larger than 10 m^2, resulting in a final 14,384 tiles of 1 to 4 km^2 each. These tiles are part of aerial acquisition missions funded by several different entities across Australia, including LiDAR and photogrammetric surverys spanning from 2004 until 2024.

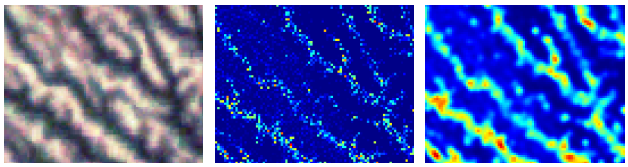
3.2 Planet optimal vegetation mosaic

We used 3 m spatial resolution PlanetScope 4 bands custom mosaics from 2021 that have been used recently in several studies involving individual trees (Brandt et al., 2024) or tree cover (Reiner et al., 2023) mapping at continental scales. These mosaics are composed by the PlanetScope images, consisting of four atmospherically corrected spectral bands (Blue, Green, Red, and Near-Infrared) from the PSScene analytic_sr product, obtained via the Planet API under a research license owned by The University of Copenhagen. Australia was tiled into $1^\circ \times 1^\circ$ grid tiles, and mosaics were created using an automated algorithm that selected scenes within phenologically optimal windows derived from the MODIS/Terra phenology product and the Copernicus Dynamic Land Cover map. Especially at a large geographical scale, it is important to train models from images acquired when only woody plants are photosynthetically active (early dry season) to maximize the crown signal and minimize the ground noise. For each tile, the algorithm prioritized periods when trees were in full foliage and herbaceous vegetation had senesced, dynamically extending the acquisition window in persistently cloudy regions. Selected scenes were filtered by quality metadata (cloud, haze, and shadow indices), reprojected to WGS84, and merged into single-tile mosaics. Radiometric consistency across scenes was achieved via histogram matching to temporally corresponding Landsat reference imagery, producing spatially seamless, phenologically consistent mosaics suitable for tree canopy analysis. For full information, refer to Reiner et al. (2023).

3.3 Image pre-processing

For each CHM, we cropped the CHM and Planet imagery to their common overlapping bounds. We then reprojected both clips to their respective WGS84 Universal Transverse Mercator (UTM) coordinate reference systems (spanning from EPSG 32749 to EPSG 32756). We then resampled the CHM clip to match and align with Planet’s grid and resolution of 3 m via nearest-neighbor interpolation. Then, each matched Planet image clip was resized to a uniform size of 256 × 256 pixels. Smaller tiles were padded with zeros, while larger tiles were randomly cropped to preserve spatial variability. To standardize spatial extents, both rasters were cropped again to their common overlapping bounds. The resulting 256×256 Planet images were subsequently radiometrically normalized to 8-bit

depth percentile-based stretching (0.1 to 99.9-th percentiles per band) using band-specific ranges previously computed over the totality of the dataset. The CHM is processed with a local maximum filter via morphological dilation to preserve tall tree crowns. The resulting map is then smoothed with a Gaussian filter to reduce residual high-frequency noise. Filtering the CHM is necessary to align the target signal with the effective spatial resolution/blurriness and information content of satellite imagery, see Figure 2. Without this step, the model is forced to regress against small peaks and high-frequency height variations that are not observable in the input satellite imagery, leading to systematic underestimation of canopy heights.



RGB image Unfiltered CHM Filtered CHM

Figure 2. A zoomed in patch of the training dataset, with unfiltered and filtered CHM. The latter now matches the satellite image in perceived resolution and detail.

The pipeline was parallelized across 200 compute nodes within the CSIRO HPC system (Petrichor). The CHMs and the Planet images have been partitioned into 70-20-10 splits of respectively 10,063 training, 2,877 validation and 1,439 test tiles.

3.4 Independent validation areas

We selected a total of 39 independent areas of 1×1 km (total validation pixels = 7,001,919), which were withheld from model training in order to perform an unbiased validation of our models. These datasets correspond to the most recent and most representative LiDAR acquisitions for each Australian vegetation class (as defined by Scarth et al. (2019)), ranging from extremely tall open forest to low scattered trees and even no trees. To further assess model generalizability under challenging conditions, we deliberately included tiles containing very tall vegetation—rare in the training data due to the natural distribution of Australian vegetation structures. For instance, 553,730 pixels represent trees above 27 m tall which are classified as ”very tall” in Scarth et al. (2019). Some examples of independent tiles are shown in Figure 7. Here, the CHM remains unfiltered.

4. Methodology

4.1 DeepLabv3+

We modified the DeepLabv3+ (Chen et al., 2018) semantic segmentation model with a ResNet101 backbone pre-trained on ImageNet to become a regressor. The output class number is simply set to 1 and the softmax layer removed so that the network outputs one linear float value. Because the ResNet encoder is initialized with ImageNet weights, we reduce the learning rate of the encoder by a factor of 10 compared to the rest of the network, which has a learning rate of 1×10^{-4}. When using the NIR channel as additional input, the first layer of DeepLab is modified to accept four channels. The initialized NIR channel weights are the same as the red channel. To increase the spatial resolution of DeepLab, we set the output stride s to 8 instead of the default 16. At the cost of higher computational demands, this change produces higher resolution feature maps from the encoder. For robustness of the model, a light color jitter (brightness $= 0.1$, contrast $= 0.1$, saturation $= 0.1$, hue $= 0.02$) is applied. Since the red and NIR channel are closely related, their jitter values are coupled. We set the batch size to 20 and use a scheduler that reduces the learning rate on a plateau with conservative settings (factor $= 0.8$, patience$= 10$). The loss function is a convex combination of the MSE loss and Structural Similarity loss (Wang et al., 2004)

$$\mathcal{L} = \alpha\, \mathcal{L}_{\mathrm{MSE}} + (1 - \alpha)\, \mathcal{L}_{\mathrm{SSIM}} \text{ with } \alpha = 0.86 \qquad (1)$$

to encourage both accurate pixel-level reconstruction and preservation of perceptual image structure, leading to sharper and more visually consistent outputs. We train for 300 epochs, which takes 35h on an Nvidia V100 with 16GB of memory.

4.2 Depth Anything V2

We adopt the DINOv2 (Oquab et al., 2024) based ViT-L variant of the Depth Anything V2 model as the encoder (Yang et al., 2024), using the metric outdoor-only pre-trained checkpoint for initialization. This choice was motivated by its alignment with the data domain of our CHM task. Although pre-trained weights provide useful features out of the box, DAv2 supports fine-tuning, which is essential for CHM prediction because the estimation task is inverted. During fine-tuning, the ViT encoder and the Dense Prediction Transformer-style decoder head are updated on our CHM estimation task. The ViT-based encoder backbone performs patch-embedding by splitting the input image into 14×14 non-overlapping patches and tokenizing them, immediately reducing the resolution. The coarse tokens are subsequently upsampled to a full-resolution height map by the decoder by merging the multi-scale feature maps from the intermediate encoder steps.

We perform fine-tuning for metric-depth by minimizing the MSE loss between predicted and ground truth CHM rasters, with a learning rate of 1×10^{-4}, a batch size of 40, and a maximum height of 30 m; based on suggestions from Cambrin et al. (2024), and the height distribution in the training dataset. Fewer than 0.03% of the training dataset contains values above 30 m, and only 537 pixels in total have CHM values over 50 m. Similar to the method used for training DeepLab, we implement a training learning-rate scheduler (factor = 0.5, patience = 10), and use AdamW optimizer. We fine-tuned the ViT-L model weights for 50 epochs (early stopped at epoch=27 after 47 hours), using an NVIDIA RTX 4090 GPU with 24GB memory. Unlike DeepLab, which requires structural modifications to shift from classification to regression, Depth Anything V2 outputs dense continuous predictions by default, simplifying the pipeline. However, it lacks native multi-modal input support, restricting potential gains from auxiliary data.

5. Results

We first evaluate our models on the test set of the training dataset. However, since the test patches are smoothed, randomly distributed, and many of them sparse, it is very sensible to validate the best performing models on the independent validation tiles described in 3.4 to evaluate performance on interconnected areas of different vegetation types. Contrary to the training examples, the CHMs here are not filtered, which will not allow us to manipulate the data by oversmoothing.

5.1 Training dataset

To evaluate performance, we use standard regression metrics including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Pearson correlation (denoted as r). Table 1 shows the quantitative results on the test set. Configurations of DeepLab with and without NIR were tested. Additionally, one model with a stride of 16 was trained. For DAv2, the inputs are limited to just RGB, and the stride is fixed to 14 due to the DINOv2 encoder. DeepLab outperforms DAv2 even with just RGB input. For example, it reaches an RMSE of 2.40 m, whereas Depth Anything achieves a slightly worse 2.54 m. With the NIR channel DeepLab improves notably to an RMSE of 2.25 m. Increasing the stride of DeepLab results in a dramatic improvement in speed, since the batch size can be tripled, but at the cost of a minimally worse metrics. All qualitative results of DeepLab reported herein are produced by the configuration with NIR and $s = 8$.

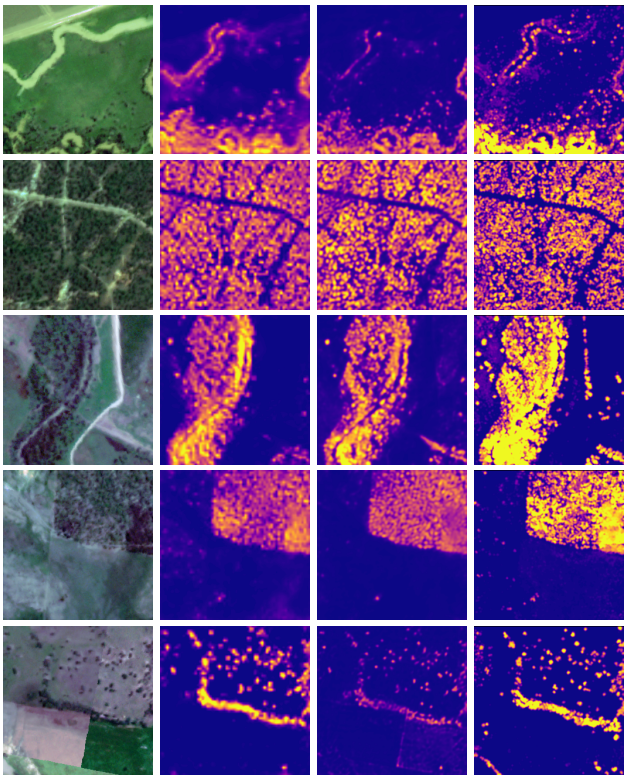
Model		**DeepLabv3+**			**DAv2**
Settings	NIR	Yes	Yes	No	No
	Stride s	16	8	8	14
Metrics	MAE [m] ↓	1.16	**1.15**	1.22	1.25
	RMSE [m] ↓	2.32	**2.25**	2.40	2.54
	r^2 ↑	0.66	**0.68**	0.64	0.60
	Corr. r ↑	0.81	**0.83**	0.80	0.77

Table 1. Quantitative results on the test set of the training dataset. The arrow direction indicates whether higher or lower values correspond to better performance.

Figure 3 shows some patches from the test set, sized 256×256, with predictions from DeepLab and DAv2, using their best configurations, as well as the ground truth (with filtering). While the sharpness and resolution of DAv2 are higher thanks to its decoder reconstructing a full-resolution height map via progressive upsampling of multi-scale feature maps, DeepLab is able to detect more individual trees (first row, along the river), and can better differentiate between lower and taller parts of forests. Depth Anything's sharpness stems from a super-resolution like upscaling of feature maps ($s = 14$ instead of DeepLabs' 8), and thus the additional details may only be inferred. Both DeepLab and DAv2 tend to slightly underestimate the magnitude of canopy heights, with this effect being more pronounced for the latter. This underestimation is particularly visible at the peaks of individual tree crowns, where the models fail to capture the highest CHM values accurately. Such behavior is likely attributable to the use of an MSE-based loss, which inherently favors smoother predictions and penalizes large deviations. Despite this limitation, neither model produces substantial false positives or severe artifacts. The main shortcomings are a tendency to miss smaller trees and to predict generally lower and smoother canopy heights, resulting in outputs of comparatively low spatial detail that match the satellite imagery.

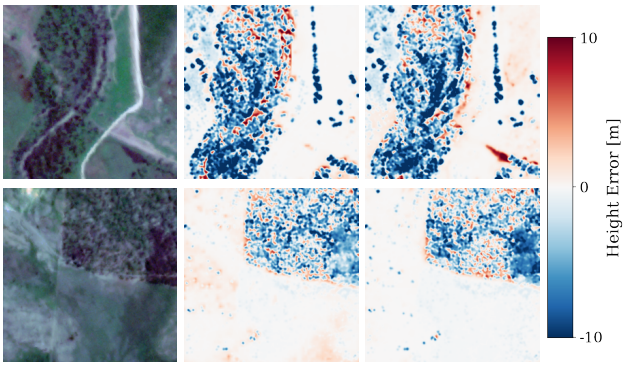
Next, we discuss the error maps in Figure 4. They confirm the quantitative observation that DeepLab is superior to DAv2. For example, in the bottom row, large parts of the forest are light blue in DeepLab while they are dark blue in DAv2, meaning that DAv2 tends to underpredict the height of tree centers. Second, we see that most errors concentrate in the area covered by trees, exhibiting a strong contrast with the largely white tree-less area.

5.1.1 Low-data robustness We trained both the ImageNet-pretrained DeepLab and the DINOv2-based DAv2 using only 10% and 1% of the original training set, while keeping validation and test splits fixed. Although DINO's self-supervised



RGB image | DeepLab with NIR, $s = 8$ | Depth Anything V2 | Ground truth

Figure 3. CHM predictions on the 3 m Planet 256×256 patches of the test set of the training dataset sampled across all over Australia. The scaling (dark blue (0 m) to yellow (15 m)) is consistent across all examples. The ground truth is filtered (see Section 3.3).



RGB | DeepLab | Depth Anything

Figure 4. Error map comparison calculated by simple subtraction of two patches from Figure 3.

pretraining on large, diverse imagery was expected to yield more robust, transferable features under data scarcity, DeepLab nonetheless achieves superior quantitative performance in both low-data regimes (Table 2). Qualitative analysis (not included in this paper) further shows that, with minimal fine-tuning data, DAv2's outputs suffer from stride-induced aliasing from DINO that degrade depth predictions. This improves with more training data, they however still remain ever so slightly when training on the full dataset. Taken together, these findings demonstrate that, in its current form, DAv2 requires substantial train-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-779-2026 | © Author(s) 2026. CC BY 4.0 License.

ing data to produce reliable results, possibly even more than the 10k images we have available.

Training Data	DeepLabv3+ ($s = 8$, NIR)				Depth Anything V2			
	MAE ↓	RMSE ↓	r^2 ↑	Corr ↑	MAE ↓	RMSE ↓	r^2 ↑	Corr ↑
100%	1.15	2.25	0.68	0.83	1.25	2.54	0.60	0.77
10%	1.43	2.80	0.51	0.71	1.52	2.84	0.46	0.68
1%	1.66	3.10	0.39	0.63	1.86	3.42	0.30	0.55

Table 2. Impact of training data amount on model performance. Metrics are computed on the same test set as in Table 1.

5.2 Independent validation areas

We use DeepLab with NIR and $s = 8$ in this section, and evaluate on the unfiltered CHMs. Generally, the results on independent validation areas show similar patterns as on the training dataset. Table 3 shows better metrics for DeepLab over DAv2 across the board. The metrics are noticeably worse than those on the test set, which is largely because the validation tiles contain areas with more—and taller—trees, as well as fewer flat areas, and no smoothing of the CHM, which collectively increase the error metrics. The mean error (ME) metric suggests a general underestimation for both models. At 0-2 m, both models show slight overestimation (Figure 5), while at higher tree heights both models start exhibiting increasing underestimation, with DAv2 slightly more than DeepLab.

Metric	DeepLabv3+	Depth Anything V2
MAE [m] ↓	4.45	5.00
RMSE [m] ↓	6.74	7.67
r^2 ↑	0.45	0.33
ME [m]	−1.17	−2.22

Table 3. Per pixel quantitative results of DeepLab (with NIR and $s = 8$) and DAv2 on the independent validation areas.

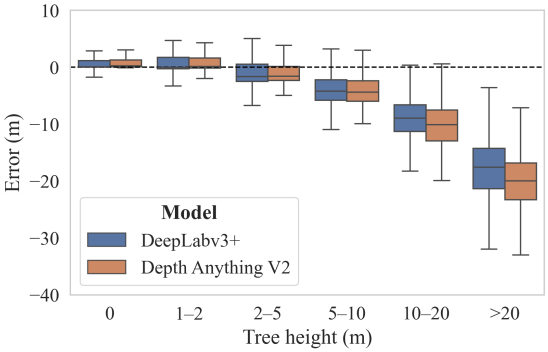


Figure 5. Prediction error across tree heights on the independent validation areas. Underprediction increases with tree height.

The scatter plots in Figure 6 again suggest a general underestimation. Both models exhibit overestimation of ground points and predictions of 0 (underestimation) in the range 0 to 5 m. This L-shaped concentration of points visible in both scatter plots is likely because the CHM around trees exhibits high local variance between adjacent pixels which conflicts with the smoother, more continuous predictions from the 3 m Planet imagery. Both models show a reluctance to predict heights above 25 to 30 m, likely because such tall vegetation is underrepresented in the training data and thus not adequately learned. The validation areas contain a much higher proportion of dense, tall forests than the training dataset—which was designed to represent typical Australian vegetation structure—resulting in a skew in this evaluation.

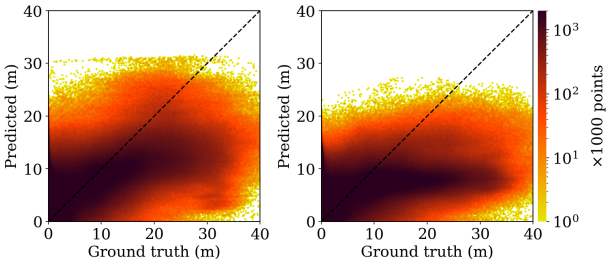


Figure 6. Scatter plot on the independent areas.

Some validation areas are shown in Figure 7. The areas are larger and thus processed in overlapping patches to avoid harsh border artifacts. Similar patterns can be observed as in Figure 3, where the main problem is underestimation.

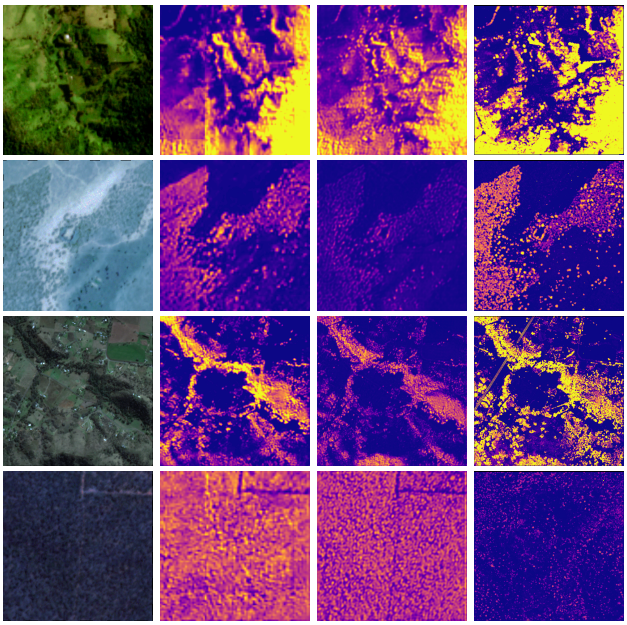


Figure 7. CHM prediction (same scale as Fig. 3) on some larger independent validation areas. From top to bottom, these images are from locations (LiDAR capture year) of Gympie (2020), Moonie South (2023), Tamworth (2022) and Camooweal (2013). The ground truth CHM is unfiltered.

The area in the last row has an outdated CHM (over 10 years older), where the trees are less mature and the road not discernable. Since CHM are difficult to come by at large scale, such large gaps between CHM capture and satellite imagery are challenging to avoid. Although most of the other independent validation tiles are relatively close in time to the Planet imagery (typically 1–2 years apart), the training dataset itself is characterized by optimal vegetation representativity but suboptimal temporal alignment with the Planet imagery, potentially contributing to lowering the r^2 metrics.

6. Discussion and Conclusion

Both models, DeepLab and Depth Anything V2, show generally strong performance; however, underprediction persists when

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-779-2026 | © Author(s) 2026. CC BY 4.0 License.

trained on CHMs filtered to match PlanetScope's effective resolution and then evaluated on unfiltered CHMs. Results are strongly shaped by dataset characteristics: substantial temporal mismatch exists between sources, with many CHMs derived from point clouds over a decade older than the 2021 Planet imagery, while in plantation regions, harvesting and regrowth introduce label inconsistencies. Co-registration errors, the coarse 3 m PlanetScope resolution, and the MSE loss further constrain recovery of sharp canopy structures, yielding smooth predictions lacking fine detail. Despite these limitations, the achieved metrics are encouraging. When evaluated on independent validation areas, both models reproduce canopy structures visually as well as on the training test regions. Filtering CHMs during training was necessary to align reference data with satellite imagery resolution; however, this filtering was deliberately omitted for the independent validation areas to enable fair comparison against unprocessed reference data, introducing a systematic mismatch that contributes to reduced metrics and larger residual spread in Figure 6. Differences in height distributions — fewer flat areas, more tall vegetation — further affect performance. These results highlight that interpreting quantitative metrics is challenging when CHM processing, resolution, and reference data characteristics differ substantially across studies, making direct numerical comparison difficult. In this context, qualitative assessment may be more indicative of model behavior; here, both models produce visually consistent and realistic canopy height patterns.

Within this constrained but realistic setting, the differences between the two model types remain clear. DeepLabv3+ consistently outperforms Depth Anything V2, benefiting from a more efficiently trainable encoder and NIR input. The NIR channel provides a physically meaningful signal closely tied to vegetation density and photosynthetic activity, effectively serving as a spectral proxy for canopy structure that DeepLabv3+ can exploit but which remains inaccessible to Depth Anything V2. The latter relies on a DINOv2 backbone pretrained primarily on natural, RGB only, and front-facing imagery, where depth cues like vanishing points, occlusion, and relative size are fundamentally different from nadir, orthographic satellite imagery. In top-down views, depth/height must be inferred from texture, shadow, and spectral signatures. Although both models are initialized from natural-image pretraining, the substantially larger parameter count of Depth Anything's ViT-L encoder ($\sim$300M vs. $\sim$45M for ResNet101) and the weaker inductive biases inherent to transformers likely require considerably more domain-specific fine-tuning data to fully adapt to the satellite CHM task; a hypothesis supported both by DAv2's steeper performance degradation in our low-data experiments and by DeepLab's superior performance even when restricted to RGB-only input. These findings confirm the results of Fogel et al. (2025), who show that CNNs can outperform large ViTs for CHM prediction, while contrasting with Cambrin et al. (2024), which reported stronger performance of fine-tuned depth foundation models, however, on much higher-resolution imagery. For future work, we plan to train a CHM prediction model on a larger scale, using multi-temporal imagery, and produce a CHM map of the entirety of Australia.

References

Almeida, C. T., Gerente, J., dos Prazeres Campos, J. R., Junior, F. C. G., Providelo, L. A., Marchiori, G., Chen, X., 2022. Canopy height mapping by Sentinel 1 and 2 satellite images, airborne LiDAR data, and machine learning. *Remote Sensing*, 14, 4112.

Boas, M. G. V., Santos, H. G., Merschmann, L. H. d. C., Van den Berghe, G., 2019. Optimal decision trees for the Algorithm Selection Problem: Integer Programming based approaches. *arXiv preprint*.

Brandt, M., Gominski, D., Reiner, F., Kariryaa, A., Guthula, V. B., Ciais, P., Tong, X., Zhang, W., Govindarajulu, D., Ortiz-Gonzalo, D., Fensholt, R., 2024. Severe decline in large farmland trees in India over the past decade. *Nature Sustainability*, 1–9.

Brandt, M., Tucker, C. J., Kariryaa, A., Rasmussen, K., Abel, C., Small, J., Chave, J., Rasmussen, L. V., Hiernaux, P., Diouf, A. A., Kergoat, L., Mertz, O., Igel, C., Gieseke, F., Schöning, J., Li, S., Melocik, K., Meyer, J., Sinno, S., Romero, E., Glennie, E., Montagu, A., Dendoncker, M., Fensholt, R., 2020. An unexpectedly large count of trees in the West African Sahara and Sahel. *Nature*, 587, 78–82.

Bulatov, D., Wayand, I., Schilling, H., 2016. Automatic tree-crown detection in challenging scenarios. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 41, 575–582.

Bégué, A., Arvor, D., Bellon, B., Betbeder, J., De Abelleyra, D., Ferraz, R. P. D., Lebourgeois, V., Lelong, C., Simões, M., Verón, S. R., 2018. Remote sensing and cropping practices: A Review. *Remote Sensing*, 10, 99.

Cambrin, D. R., Corley, I., Garza, P., 2024. Depth any canopy: Leveraging depth foundation models for canopy height estimation. *arXiv preprint arXiv:2408.04523*.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European conference on computer vision (ECCV)*, 801–818.

Dixon, D. J., Zhu, Y., Jin, Y., 2025. Canopy height estimation from PlanetScope time series with spatio-temporal deep learning. *Remote Sensing of Environment*, 318, 114518.

Duncanson, L., Kellner, J. R., Armston, J., Dubayah, R., Minor, D. M., Hancock, S., Healey, S. P., Patterson, P. L., Saarela, S., Marselis, S. et al., 2022. Aboveground biomass density models for NASA's Global Ecosystem Dynamics Investigation (GEDI) LiDAR mission. *Remote Sensing of Environment*, 270, 112845.

Dupuis, C., Lejeune, P., Michez, A., Fayolle, A., 2020. How Can Remote Sensing Help Monitor Tropical Moist Forest Degradation? – A Systematic Review. *Remote Sensing*, 12, 1087.

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., 2025. Open-canopy: Towards very high resolution forest monitoring. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 1395–1406.

Günaydın, E., Yakar, İ., Bakırman, T., Selbesoğlu, M. O., 2025. Evaluation of Depth Anything models for satellite-derived bathymetry. *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 48, 101–106.

Illarionova, S., Shadrin, D., Ignatiev, V., Shayakhmetov, S., Trekin, A., Oseledets, I., 2022. Estimation of the Canopy Height Model From Multispectral Satellite Imagery With Convolutional Neural Networks. *IEEE Access*, 10, 34116-34132.

Jin, S., Su, Y., Gao, S., Hu, T., Liu, J., Guo, Q., 2018. The transferability of Random Forest in canopy height estimation from multi-source remote sensing data. *Remote Sensing*, 10(8).

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 24–49.

Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K., 2025. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Kenzhebay, M., Hanan, A., Khan, M., Gazzea, M., Arghandeh, R., 2025. Semantically aware tree height estimator for infrastructure monitoring using multimodal satellite images. *European Journal of Remote Sensing*, 58(1), 2479010.

Khosravipour, A., Skidmore, A. K., Isenburg, M., Wang, T., Hussin, Y. A., 2014. Generating Pit-free Canopy Height Models from Airborne Lidar. *Photogrammetric Engineering & Remote Sensing*, 80, 863–872.

Lahssini, K., Baghdadi, N., le Maire, G., Fayad, I., Villard, L., 2024. Canopy height mapping in French Guiana using multi-source satellite data and environmental information in a U-Net architecture. *Frontiers in Remote Sensing*, 5.

Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., Wegner, J. D., 2022. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. *Remote Sensing of Environment*, 268, 112760.

Liao, Z., Van Dijk, A. I. J. M., He, B., Larraondo, P. R., Scarth, P. F., 2020. Woody vegetation cover, height and biomass at 25-m resolution across Australia derived from multiple site, airborne and satellite observations. *International Journal of Applied Earth Observation and Geoinformation*, 93, 102209.

Liu, S., Brandt, M., Nord-Larsen, T., Chave, J., Reiner, F., Lang, N., Tong, X., Ciais, P., Igel, C., Pascual, A., Guerra-Hernandez, J., Li, S., Mugabowindekwe, M., Saatchi, S., Yue, Y., Chen, Z., Fensholt, R., 2023. The overlooked contribution of trees outside forests to tree cover and woody biomass across Europe. *Science Advances*, 9(37), eadh4097.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.*, 2024.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M. C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C. E. et al., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment*, 253, 112165.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3), 1623–1637.

Reiner, F., Brandt, M., Tong, X., Skole, D., Kariryaa, A., Ciais, P., Davies, A., Hiernaux, P., Chave, J., Mugabowindekwe, M. et al., 2023. More than one quarter of Africa’s tree cover is found outside areas previously classified as forest. *Nature Communications*, 14, 2258.

Scarth, P., Armston, J., Lucas, R., Bunting, P., 2019. A Structural Classification of Australian Vegetation Using ICESat/GLAS, ALOS PALSAR, and Landsat Sensor Data. *Remote Sensing*, 11, 147.

Siméoni, O., Vo, H. V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M. et al., 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Skidmore, A. K., Coops, N. C., Neinavaz, E., Ali, A., Schaepman, M. E., Paganini, M., Kissling, W. D., Vihervaara, P., Darvishzadeh, R., Feilhauer, H. et al., 2021. Priority list of biodiversity metrics to observe from space. *Nature Ecology & Evolution*, 5, 896–906.

Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J. et al., 2023. Sub-meter resolution canopy height maps using self-supervised learning and a vision transformer trained on Aerial and GEDI LiDAR. *arXiv preprint*.

Tolan, J., Yang, H.-I., Nosarzewski, B., Couairon, G., Vo, H. V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J. et al., 2024. 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.

Wagner, F. H., Dalagnol, R., Carter, G., Hirye, M. C. M., Gill, S., Takougoum, L. B. S., Favrichon, S., Keller, M., Ometto, J. P., Alves, L. et al., 2025. High resolution tree height mapping of the Amazon forest using Planet NICFI images and LiDAR-informed U-Net model. *arXiv preprint*.

Wagner, F. H., Ferreira, M. P., Sanchez, A., Hirye, M. C. M., Zortea, M., Gloor, E., Phillips, O. L., de Souza, C. R., Shimabukuro, Y. E., Aragao, L., 2018. Individual tree crown delineation in a highly diverse tropical forest using very high resolution satellite images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145, 362–377.

Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E., 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4), 600-612.

Weber, M., Beneke, C., Wheeler, C., 2025. Unified deep learning model for global prediction of aboveground biomass, canopy height, and cover from high-resolution, multi-sensor satellite imagery. *Remote Sensing*, 17(9), 1594.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything V2. A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, C. Zhang (eds), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Yao, J., Wu, T., Zhang, X., 2024. Improving depth gradient continuity in transformers: A comparative study on monocular depth estimation with CNN. *35th British Machine Vision Conference, BMVC 2024, Glasgow, UK, November 25-28, 2024*, BMVA Press.