Remote Sensing of Environment 300 (2024) 113888

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar

Jamie Tolan [a], Hung-I Yang [a], Benjamin Nosarzewski [a], Guillaume Couairon [b], Huy V. Vo [b], John Brandt [c,*], Justine Spore [c], Sayantan Majumdar [d], Daniel Haziza [b], Janaki Vamaraju [a], Theo Moutakanni [b], Piotr Bojanowski [b], Tracy Johns [a], Brian White [a], Tobias Tiecke [a], Camille Couprie [b]

[a] Meta, 1 Hacker Way, Menlo Park, CA 94025, USA
[b] Fundamental AI Research (FAIR), Meta, 1 Hacker Way, Menlo Park, CA 94025, USA
[c] World Resources Institute, 10 G St NE #800, Washington, DC 20002, USA
[d] Desert Research Institute, 2215 Raggio Pkwy, Reno, NV 89512, USA

ARTICLE INFO

Edited by Jing M. Chen

Keywords:
LIDAR
GEDI
Canopy height
Deep learning
Self-supervised learning
Vision transformers

ABSTRACT

Vegetation structure mapping is critical for understanding the global carbon cycle and monitoring nature-based approaches to climate adaptation and mitigation. Repeated measurements of these data allow for the observation of deforestation or degradation of existing forests, natural forest regeneration, and the implementation of sustainable agricultural practices like agroforestry. Assessments of tree canopy height and crown projected area at a high spatial resolution are also important for monitoring carbon fluxes and assessing tree-based land uses, since forest structures can be highly spatially heterogeneous, especially in agroforestry systems. Very high resolution satellite imagery (less than one meter (1 m) Ground Sample Distance) makes it possible to extract information at the tree level while allowing monitoring at a very large scale. This paper presents the first high-resolution canopy height map concurrently produced for multiple sub-national jurisdictions. Specifically, we produce very high resolution canopy height maps for the states of California and São Paulo, a significant improvement in resolution over the ten meter (10 m) resolution of previous Sentinel / GEDI based worldwide maps of canopy height. The maps are generated by the extraction of features from a self-supervised model trained on Maxar imagery from 2017 to 2020, and the training of a dense prediction decoder against aerial lidar maps. We also introduce a post-processing step using a convolutional network trained on GEDI observations. We evaluate the proposed maps with set-aside validation lidar data as well as by comparing with other remotely sensed maps and field-collected data, and find our model produces an average Mean Absolute Error (MAE) of 2.8 m and Mean Error (ME) of 0.6 m.

1. Introduction

Spatially explicit maps of forest vegetation structure, such as tree canopy height and crown projected area, are powerful tools for assessing forest degradation, forest and landscape restoration (FLR), and estimating above-ground woody biomass for carbon emission and sequestration modeling. Existing assessments of the climate implications of woody vegetation flux, including FLR, deforestation, and natural regrowth, often rely on remotely sensed dynamic vegetation models of deforestation and regrowth (Friedlingstein et al., 2019). Such wall-to-wall data on tree height and canopy structure are used to estimate aboveground woody biomass. However, land-use patterns operate on more granular spatio-temporal scales than those captured by global carbon models, which typically have coarse spatio-temporal resolution. This contributes to the large uncertainty in existing nation-wide and global accounting of carbon stored in forests (Popkin, 2015; Duncanson et al., 2020; Yanai et al., 2020). For instance, Cook-Patton et al. (2020) produce a global 1-km scale map of potential above-ground carbon

* Corresponding author.
E-mail address: john.brandt@wri.org (J. Brandt).

https://doi.org/10.1016/j.rse.2023.113888
Received 19 April 2023; Received in revised form 24 October 2023; Accepted 25 October 2023
Available online 7 November 2023

accumulation rates by developing machine learning models based on more than 13,000 locations derived from literature. Cook-Patton et al. (2020) find significant variability in predicted carbon accumulation rates compared to defaults from the International Panel on Climate Change (IPCC) at the ecozone scale. In the African tropical montane forests, Cuni-Sanchez et al. (2021) model forest carbon density based on 72,336 measurements of height and tree diameter, identifying two-thirds higher carbon stocks than the respective IPCC default values.

The uncertainty of biomass modeling also affects the uncertainty of the carbon implications of deforestation and regrowth. Tree-based FLR, including agroforestry, reforestation, natural regeneration, and enrichment planting, is considered to be a cost-effective natural climate solution for adaptation and mitigation. However, evaluating the effectiveness of FLR interventions at a large scale is difficult due to its highly distributed nature, typically being practiced on individual land parcels by respective land owners (Reytar et al., 2020). While carbon reporting frameworks exist for FLR, for example through verified carbon markets, such data are highly project-specific owing to their reliance on intensive manual field measurements. Utilizing remotely sensed data to assess vegetation structure on areas with FLR interventions such as intercropped agroforestry or natural regeneration is difficult due to the presence of multiple species, multiple canopy strata, and trees of different ages (Viani et al., 2018; Vallauri et al., 2005; Camarretta et al., 2020). For instance, Tesfay et al. (2022) found that 70% of the shade trees in an agroforestry system in Ethiopia were below 3 m in height, while 3% were above 12 m in height, with more than a two-order of magnitude range of per-tree carbon stocks depending on tree size.

Critical to reducing uncertainty in woody carbon models are measurements of forest height and biomass to improve assessments of the spatial variability of carbon removal rates across forest landscapes that have heterogeneous structure (Harris et al., 2021). Tree height is especially critical to accurately assessing carbon removal rates, as growth rate increases continuously with size (Stephenson et al., 2014). Recent earth observation missions from NASA, namely GEDI and ICESat-2, provide repeated vegetation canopy height maps for the first time. Global Ecosystem Dynamics Investigation (GEDI) collects canopy height and relative height at a 25 m resolution (Dubayah et al., 2021). ICESat-2 collects canopy height and relative height at a 13×100 meter native footprint (Markus et al., 2017). Recently, multi-sensor fusion has demonstrated potential to improve aboveground biomass mapping (Silva et al., 2021). To generate wall-to-wall maps of canopy height, researchers commonly combine active optical LiDAR data from ICESat-2 or GEDI with optical imagery from Sentinel-2 (Lang et al., 2022a; Schwartz et al., 2022) or Landsat satellites (Schwartz et al., 2022; Li et al., 2020).

A number of recent studies have utilized spaceborne lidar data from GEDI and ICESat-2 to produce canopy height maps in combination with multispectral optical imagery. Among them, Potapov et al. (2021) combined GEDI RH95 (95^{th} percentile of Relative Height) data with Landsat data to establish a global map at 30 m resolution, using a bagged regression tree ensemble algorithm. More recently, Lang et al. (2022a) produced a global canopy height map at a 10-m resolution, applying an ensemble of convolutional neural network (CNN) models to Sentinel-2 imagery to predict the GEDI RH98 footprint. Other works have produced regional 10-m CHMs utilizing Sentinel-2 and aerial lidar (Astola et al., 2021; Fayad et al., 2023).

Aerial lidar data has also demonstrated utility as training data for high resolution (< 5 m) and very high resolution (< 1 m) canopy height maps. At a national scale, Csillik et al. (2019) generated biomass maps in Peru by applying gradient boosted regression trees between 3.7 m Planet Dove imagery and airborne lidar, with low uncertainty in dense forests but large amounts of uncertainty in transitional landscapes and areas that are hotspots of land use change. Recently, Liu et al. (2023) computed a canopy height map (CHM) map of Europe using 3 m Planet imagery, training two UNets to predict tree extent and CHM using lidar observations and previous CHM predictions from the literature. Utilizing aerial optical imagery, Wagner et al. (2023) generated a submeter CHM of California, USA by training a regression U-Net CNN on 60-cm imagery from the USDA-NAIP program and aerial lidar.

The estimation of canopy height from high resolution optical imagery shares similarities with the computer vision task of monocular depth estimation. Vision transformers, which are a deep learning approach to encoding low-dimensional input into a high dimensional feature space, have established new frontiers in depth estimation compared to convolutional neural networks (Ranftl et al., 2021). While depth estimation models benefit significantly from large receptive fields (Li et al., 2018; Fu et al., 2018; Miangoleh et al., 2021), Luo et al. (2016) demonstrate that the effective receptive fields of CNN models have Gaussian distributions, limiting the ability for CNNs to model long-range spatial dependencies. In contrast to convolutional neural networks (CNNs), which subsequently apply local convolutional operations to enable the modeling of increasingly long-range spatial dependencies, transformers utilize self-attention modules to enable the modeling of global spatial dependencies across the entire image input (Dosovitskiy et al., 2021a).

For dense prediction tasks on high resolution imagery where the context can be sparse, such as ground information in the case of near closed canopies, the ability of transformers to model global information is promising. Among the applications to aerial imagery, the work of Xu et al. (2021) uses a Swin transformer to classify high-resolution land cover. Finding that a baseline transformer model struggled with edge detection, Xu et al. (2021) utilized a self-supervised edge extraction and enhancement method to improve definition of class edges. Wang et al. (2022) utilize the vision transformer architecture as a feature encoder, and apply a feature pyramid decoder to the resulting multi-scale feature maps. Gibril et al. (2023) segment individual date palm trees by applying vision transformers to 5- to 30-cm drone-based imagery, finding that the Segformer architecture improves generalizability to different resolution imagery when compared to CNN-based models. More recently, also leveraging vision transformers, Reed et al. (2022) scale the Masked Auto-Encoder approach of He et al. (2022) and apply it to building segmentation.

A major challenge of applying high resolution, airborne lidar data to the generation of wall-to-wall canopy height maps is the relative scarcity of airborne lidar data to the scientific community. Such scarcity can negatively impact the generalizability of models to unseen geographies, especially data-poor regions where little to no airborne lidar exists (Schacher et al., 2023). Given this context of low annotation, Self-Supervised Learning (SSL) is a promising tool to shape more robust features than traditional deep approaches. In particular, the SSL DINOv2 approach of Oquab et al. (2023) recently led to state-of-the-art performances in several computer vision tasks such as image classification, depth prediction, and segmentation. In the context of satellite image analysis, self-supervised learning has been shown to improve the generalizability of building segmentation models in Africa (Sirko et al., 2021). To mitigate the reliance of vision transformers on self-supervised learning, Fayad et al. (2023) utilized knowledge distillation with a U-Net CNN teacher model to generate 10-m CHM of Ghana using Sentinel-1, Sentinel-2, and aerial lidar.

Understanding the importance of highly spatially explicit vegetation structure mapping to both large-scale carbon modeling and project-specific avoided deforestation and restoration monitoring, the objective of this study is to produce high resolution canopy height maps that are able to scale and generalize to large geographic regions. Our method consists of an image encoder-decoder model, where low spectral dimensional input images are transformed to a high dimensional encoding and subsequently decoded to predict per-pixel canopy height. We employ DINOv2 self-supervised learning to generate universal and generalizable encodings from the input imagery (Oquab et al., 2023), and train a dense vision transformer decoder (Ranftl et al., 2021) to generate canopy height predictions based on aerial lidar data from sites across the USA. To correct a potential bias coming from a geographically limited source of supervision, we finally refine the maps using a

convolutional network trained on spaceborne lidar data. We present canopy height maps for the states of São Paulo, Brazil, and California, USA, and provide qualitative and quantitative error analyses of height estimation and the decomposition of height estimates into tree segmentation maps.

2. Data

2.1. Experimental design

This paper presents canopy height maps for São Paulo State, Brazil, and California State, USA. These geographies were chosen due to their prevalence of timber production, presence of old growth forests, mountainous terrains, and high degree of tree biodiversity (Maioli et al., 2020; Luyssaert et al., 2008; Ribeiro et al., 2011). The dataset was generated with a machine learning model utilizing a transformer encoder and convolutional decoder trained with an input composite of approximately 0.59 m GSD Maxar imagery spanning the years 2018 to 2020 and output labels from 1 m GSD aerial lidar. Our data and methods sections are structured as follows. First, we describe the satellite and aerial lidar data used for model training and map generation. Next, we describe the model training specifics, including self supervised learning and the methods for combining models trained on aerial lidar with models trained on GEDI observations, and the baseline models selected and ablation studies performed. Finally, we present our approach for qualitative and quantitative evaluation of height accuracy and tree segmentation, and discuss the generalization of our model.

2.2. Satellite image data description

Maxar Vivid2 mosaic imagery[1] served as input imagery for model training and inference. This dataset provides global coverage by mosaicing together imagery from multiple instruments (WorldView-2 (WV 2), WorldView-3 (WV 3), Quickbird II) and observation dates. By starting with this mosaiced imagery, we leveraged the extensive data selection pipeline from Maxar, resulting in imagery that had less than 2% percent cloud cover, a global revisit rate predominately (more than 75%) below 36 months (imagery dates from 2017 to 2020 are utilized in this dataset), view angles of less than 30 degrees off nadir, and sun angle of less than 60 degrees from zenith. This imagery consisted of three spectral bands: Red, Green, and Blue (RGB), with approximately a 0.5 m GSD. The imagery was processed in the Web Mercator projection (EPSG:3857) and stored with the Bing tiling scheme.[2] Given the high resolution of the original geotiffs, Bing zoom 15 level tiles, with 2048 × 2048 pixels per tile were used, giving a pixel size of 0.597 m GSD at the equator.

2.3. Satellite image data preparation

2.3.1. Image preparation

For easier training and validation of computer vision models, we extracted small regions from the input satellite imagery. Centered around a given location, a box of fixed ground distance was selected, using a local tangent plane coordinate system. Due to the Web Mercator projection of the image tiles, the extracted images at each position had varying dimensions according to their latitude, which were re-sampled to a fixed number of pixels. We chose a box side length of 152.7 m, which, when re-sampled to 256 × 256 pixel images, provided "thumbnail" images that matches the lowest resolution (0.597 m) of the input imagery described in Section 2.2. Using these thumbnail images both for training and inference ensured that the dataset had constant number of pixels and that the pixel size was the same for all latitudes, preventing potential biases with latitude which may be introduced by variation in pixel size.

2.3.2. Dataset for self-supervised learning

For training the self-supervised encoder, we randomly sampled 18 million 256 × 256 pixel satellite thumbnail images. No labels were used for the SSL stage.

2.3.3. Validation segmentation dataset

We also manually annotated a random selection of 9000 Maxar thumbnail images for segmentation testing. A binary tree / no tree label was applied by human annotators. Pixels estimated to have a canopy height above one meter (1 m) tall and with a canopy diameter of more than three meters (3 m) were labeled as tree.

2.4. Supervised dataset

We gathered approximately 5800 canopy height maps (CHM), selected from the National Ecological Observatory Network (NEON) (2022). Each CHM typically consisted of 1 km × 1 km geotiffs, with a pixel size of one meter (1 m) GSD, in local UTM coordinates. We selected the sites used by Weinstein et al. (2021) and additionally manually filtered for sites that have CHM imagery that was well registered and mostly free from mosaicing artifacts. Additionally, we selected sites with imagery acquired less than two years from the observation date in the associated Maxar satellite imagery. A complete list of NEON sites used for training and validation is contained in Appendix A.

The CHM geotiffs were reprojected to a local tangent plane coordinate system and resized to match the resolution of Maxar images. For each ALS CHM, a corresponding RGB satellite image was linked, and these pairs of imagery served as the training data for our decoder model. The 5800 images in the NEON ALS dataset were split in sets of 80% training images, 10% calibration and 10% set-aside validation images. During the training, validation and testing phases, we sampled 256 × 256 random crops from the RGB - ALS image pairs. Model training was conducted over epochs sampled from the training dataset. At the completion of each epoch, metrics were computed from a 10% calibration dataset to calibrate the hyperparameters of the model training process. The calibration dataset was drawn from the same set of sites as the training datasets, but from separate 1 km × 1 km geotiffs to ensure non overlapping pixels.

We constructed a set-aside validation dataset from a subset of sites in our NEON dataset, which we call "NEON test". None of the sites used in the validation dataset were contained in the training or calibration dataset. A list of NEON sites in the validation set appears in Appendix A. We also prepared two validation datasets from other publicly available ALS Lidar datasets, outside of the NEON collection. These datasets covered different geographic locations and ecosystems: "CA-Brande" (Brande, 2021) covered a coastal ecosystem in CA, and "São Paulo" (Dos-Santos et al., 2019) covered a region in the Brazilian São Paulo State. See Fig. A.18 for a visual breakdown of the Neon dataset splits.

Where these datasets were available as CHMs, we directly used the supplied CHMs. However, for the São Paulo datasets, which only contained point cloud datasets, we processed CHMs following the pit-free algorithm (Khosravipour et al., 2014). The pit-free algorithm was also adopted by the NEON team for generating their CHM product, and we found that different input parameters to the pit-free algorithm had negligible impact on the CHM output.

2.5. Data augmentation

The 256 × 256 pixel image thumbnail images of RGB and CHM imagery were augmented at training time, with random 90 degree rotations, brightness, and contrast jittering. We found that these augmentations improved model prediction stability across the various

[1] https://resources.maxar.com/data-sheets/imagery-basemaps-data-sheet.
[2] https://learn.microsoft.com/en-us/bingmaps/articles/bing-maps-tile-system.

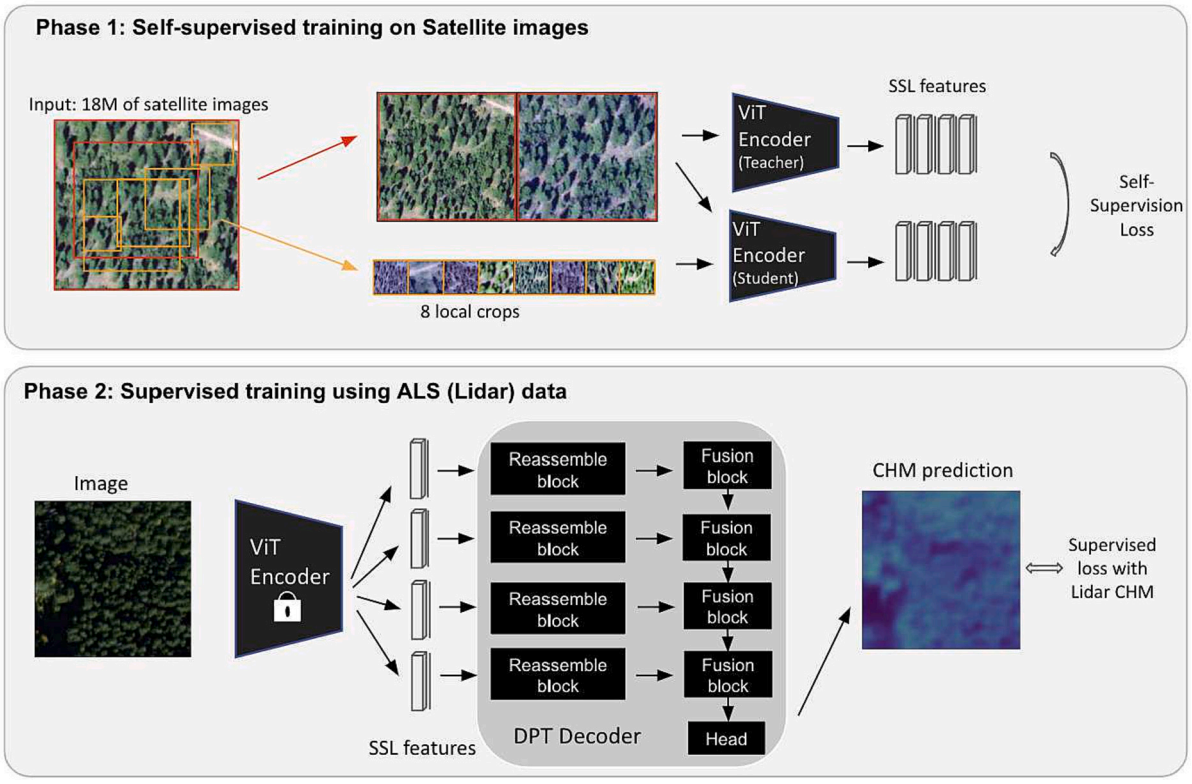


Fig. 1. Overview of our approach for generating ALS-based CHMs. During the first stage, we employed the self-supervised learning approach Oquab et al. (2023) on 18 million 256 × 256 satellite images leading to a set of four spatial feature maps, and four feature vectors, extracted at different layers of the Vision Transformer model (ViT). In the second phase, we trained a convolutional DPT decoder to predict CHMs.

Maxar observations in the input dataset.

3. Model and data generation methods

Our goal was to create a model that produces high resolution canopy height maps and generalizes across large geographic scales. To accomplish that goal, we leveraged the relative strengths of two types of lidar data. Aerial lidar provided high resolution canopy height estimation, but lacks global spatial coverage. In comparison, GEDI has nearly global coverage of transects, but its beam width of approximately 25 m did not allow for the identification of individual trees.

After self-supervised pre-training on satellite images globally, our high-resolution ALS CHM prediction model was trained on images from the NEON dataset, as detailed in Section 3.2 and Fig. 1. As the Neon dataset only has a spatial coverage from sites only within the United States, we expect this ALS CHM model to perform well on ecosystems similar to the training set. To improve generalization of other ecosystems and locations, a low resolution CHM model was independently trained on global GEDI data (Section 3.3). The GEDI model was used to compute a rescaling factor map (Section 3.4), which adjusted the predictions made by the ALS CHM model.

3.1. Self supervised learning

Following the recent success of self-supervised learning on dense prediction tasks from Oquab et al. (2023), we employed a self-supervised learning step on 18 million globally distributed, randomly sampled 256 × 256 pixel Maxar satellite images to obtain an image encoder delivering features specialized to vegetation images. In the training phase, different views of the image were fed to two versions of the encoder: a teacher model receiving global crops, and a student model receiving local and global views where part of the crops were masked (replaced by zero values). We employ a huge ViT architecture, where the inputs are decomposed into 16 × 16 patches. The two networks were trained jointly to output similar feature representations. The procedure is illustrated in the Phase 1 in Fig. 1. In a second phase described in Section 3.2, we freeze the SSL encoder layers using the weights of the teacher model and train the decoder with ALS data to generate high-resolution canopy height maps.

3.2. High resolution canopy height estimation using ALS

We used the reference dataset described in Section 2.4, prepared following the methods described in Section 2.3.1. The output of the ALS model was a raster of predicted canopy heights at the same resolution as the input imagery. For training the supervised decoder, we used the ALS CHM data described in Section 2.4 to create a connection between the SSL features and the full resolution canopy height image. In this second phase, we trained the decoder introduced in Dense Prediction Transformer (DPT) (Ranftl et al., 2021) on top of the obtained features. This approach is described in Fig. 1, phase 2. The DPT paper describes a full model composed of a transformer encoder extracting features at different layers. In the decoder, each output was reassembled and all outputs were fused. In our second phase of ALS training, we replaced the transformer of DPT by our own SSL encoder, and trained the DPT decoder part only, from scratch. Our best results were obtained by freezing all layers from the SSL encoder. We employed a one cycle learning rate schedule with a linear warmup in the encoder training stage and a "Sigloss" loss function. Further architecture and training details are provided in Appendix D.

Sigloss function. We take advantage from the similarity of canopy height mapping to the task of depth estimation and borrow the loss from Eigen et al. (2014). Given a true canopy height map c and our prediction $\hat{c}$, the Sigloss is given by

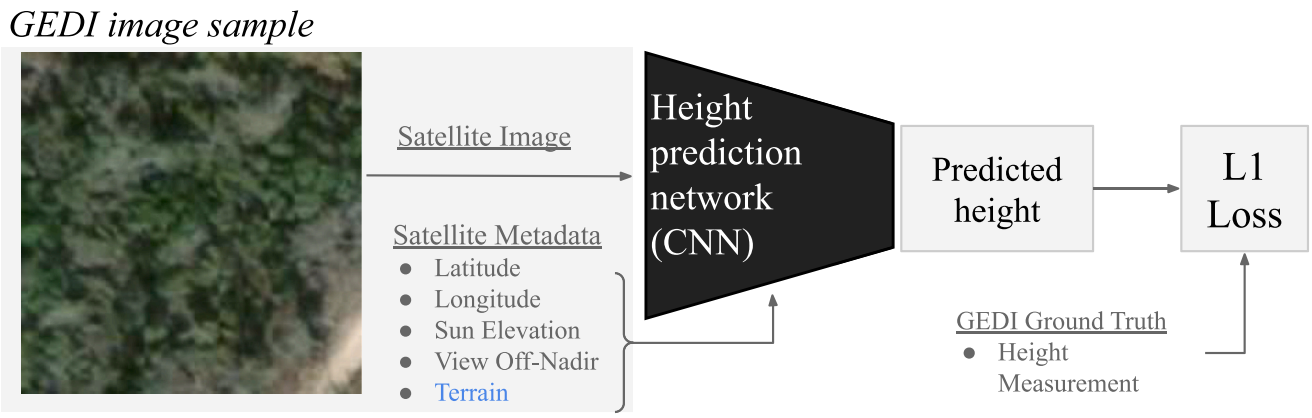


Fig. 2. Overview of our methodology to generate predicted RH95 values using GEDI measurements across the globe. Terrain is used only during the training and set to zero during inference.

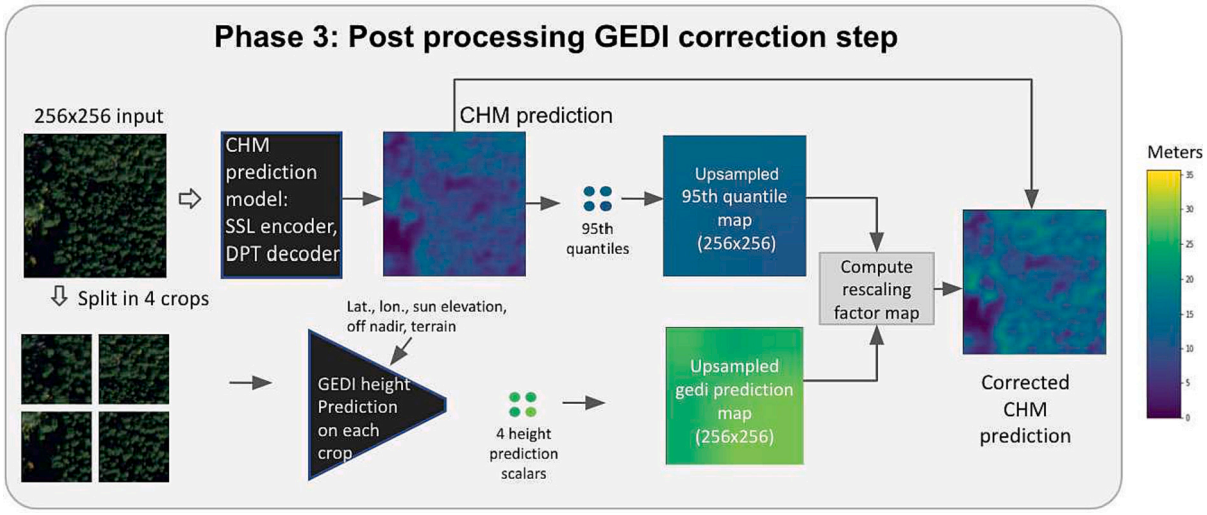


Fig. 3. Post processing step using GEDI predictions during inference. We used the GEDI model to correct our CHM predictions, by computing a dense scaling factor, and multiply it pointwise with the CHM prediction map.

$$\mathscr{L} = \alpha\sqrt{\frac{1}{T}\sum_i \delta_i^2 - \frac{\lambda}{T^2}\left(\sum_i \delta_i\right)^2}, \tag{1}$$

where $\delta_i = log(\hat{c}_i) - log(c_i)$, and T is the number of pixels with valid ground truth values. As in previous works, we fix $\lambda = 0.85$ and $\alpha = 10$.

Classification output. To avoid a bias towards small predicted values, we implemented a classification step first, combined with the Sigloss defined above. The strategy is described by Bhat et al. (2021) as the uniform strategy. Specifically, we modified the output of our decoder to return, instead of one scalar per pixel, a range of B bins. After a normalization on the predictions, we computed the scalar product between the obtained histogram of predicted bins and a linear vector ranging [0,B], with B set to 256.

3.3. Large scale canopy height estimation using GEDI prediction model

To mitigate the effect of the limited geographic distribution of available ALS data, we employed a second regression network trained on GEDI data to rescale the ALS network outputs. The GEDI prediction model was a simple convolutional network, containing five convolutional layers, followed by five fully connected layers. The inputs to the model were 128 × 128 pixel Maxar images containing three RGB bands, in topocentric coordinates, processed as described in Section 2.3.1. The ground truth data consisted of 13 million GEDI measurements, which were randomly sampled from the full GEDI dataset described in Appendix B.1. We trained the GEDI model to output a single scalar value for a 128 × 128 pixel image patch, with a L1 loss on a regression task against the RH95 value from the GEDI instrument. The training details are specified in Appendix B.3.

3.4. Combining ALS and GEDI model outputs

In this section, we describe the process of connecting our GEDI model outputs (Section 3.3) with ALS model outputs (3.2). Conceptually, the ALS model output provides high resolution canopy estimates but lacks the global context to correctly estimate the absolute height of vegetation in different ecosystems. Conversely, the GEDI model is trained on a global dataset and contains position and metadata inputs (Fig. 2). A schematic of the process is shown in Fig. 3.

Correlation between different lidar sources. The first step in making the GEDI/ALS connection is understanding the relationship between the two sets of lidar data: ALS CHM (Section 2.4) and GEDI lidar (Section Appendix B.1). These two datasets make measurements of fundamentally different properties of canopy structure. GEDI measures the relative height of canopy based on the full waveform measurement of the return energy from 25 m diameter beam footprints while aerial lidar constructs higher resolution point clouds. To connect these two, we ran simulations

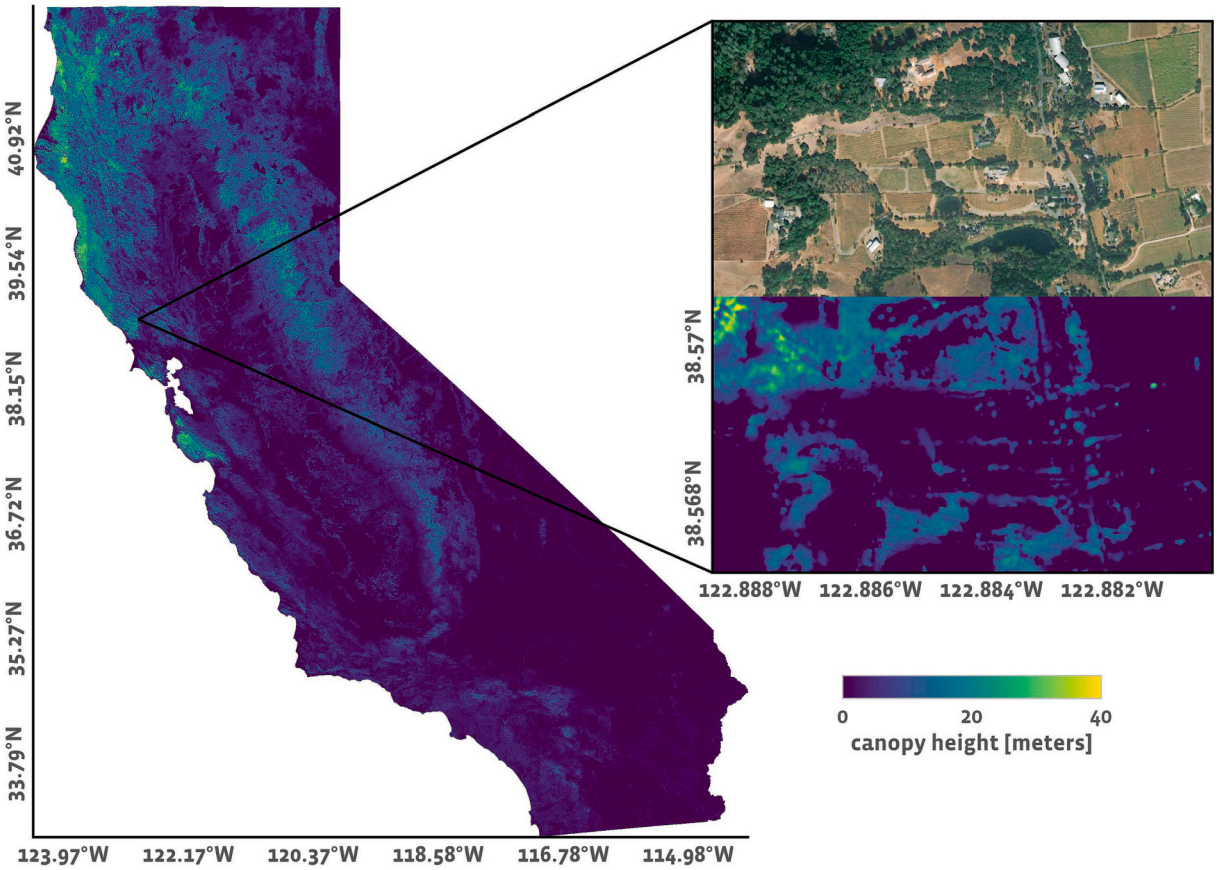


Fig. 4. Canopy Height Map (CHM) for the state of California, inset showing zoomed in region with input RGB imagery.

with the GEDI simulator from Hancock et al. (2019) on the NEON ALS point clouds. We found that there was a strong correlation ($R^2 = 0.88$) between the 95th percentile of ALS canopy height maps and the simulated GEDI RH95 (see Appendix B.2).

GEDI based correction of ALS trained maps. We used this correlation to scale the ALS model canopy height maps by computing a scalar multiplier that match percentiles of the CHM map with the GEDI model predicted value for GEDI RH95. This process works as follows:

Given an input RGB image, x, we combined the outputs of the ALS and GEDI models by computing a dense correction factor $\gamma(x)$, so that the novel prediction, $C^{'}(x)$ was related to the ALS model CHM, $C(x)$:

$$C^{'}(x) = \gamma(x) \odot C(x) \tag{2}$$

where

$$\gamma(x) = \frac{1 + s_\sigma(G(x))}{1 + \left(s_\sigma\left((Q(x)_{95})\right)\right)}. \tag{3}$$

Here $G(x)$ is the output CHM of our GEDI model and $Q(x)_{95}$ is a per block upsampled 95^{th} percentile of the ALS model CHM in meters, computed over the exact same 128×128 pixel input regions as the input to the GEDI model in $G(x)$. More specifically, each input image was divided in four crops, each one independently fed to the height prediction model, leading to four scalars, that were concatenated and upsampled. From the predicted CHM map by our ALS model, we computed four percentiles from the same crops, concatenated and upsampled in the same way.

We used the ratio in Eq. (3) rather than $G(x)/Q(x)_{95}$ to down-weight noisy model estimates near zero canopy height. Since $G(x)$ and $Q(x)_{95}$ are lower resolution than $C(x)$, the correction factor map was upsampled to match the resolution of the ALS CHM, $C(x)$. The ALS and GEDI maps were smoothed with a 20 pixel sigma Gaussian kernel s_σ to prevent sharp transitions, and the correction factor was clipped between 0.5 and 2 to avoid drastic rescaling.

3.5. Baselines

3.5.1. ResUNet-based approach

We utilized a ResUNet-18 architecture for our baseline (Zhang et al., 2017), which is an encoder-decoder architecture predicting a $N \times N$ canopy height map from a $3 \times N \times N$ RGB image, with $N = 256$. The baseline model was trained with the sigloss between the predicted and ground truth CHMs. We also experimented with a classification output, however we did not obtain improvements from this approach.

3.5.2. Supervised transformer-based approach

To assess the benefit of the self supervised training phase on Satellite data, we consider a baseline given the state-of-the-art vision SWAG encoder (Singh et al., 2022). We used the large version of this Vision Transformer (ViT) that was trained to perform hashtag prediction from Instagram images. At the time of writing this manuscript, this model was in the top ten models with highest accuracy on ImageNet, CUB, Places, and iNaturalist datasets, providing a warranty of feature quality. This model contains the same number of parameters as our SSL encoder, allowing a fair comparison in terms of model size.

3.6. Data validation

We evaluated the model performance against a variety of metrics, which we divided into two broad classes: (1.) Metrics which primarily evaluated the accuracy of canopy height maps, which we call canopy height metrics (Section 4.1), and (2.) Metrics which primarily evaluated

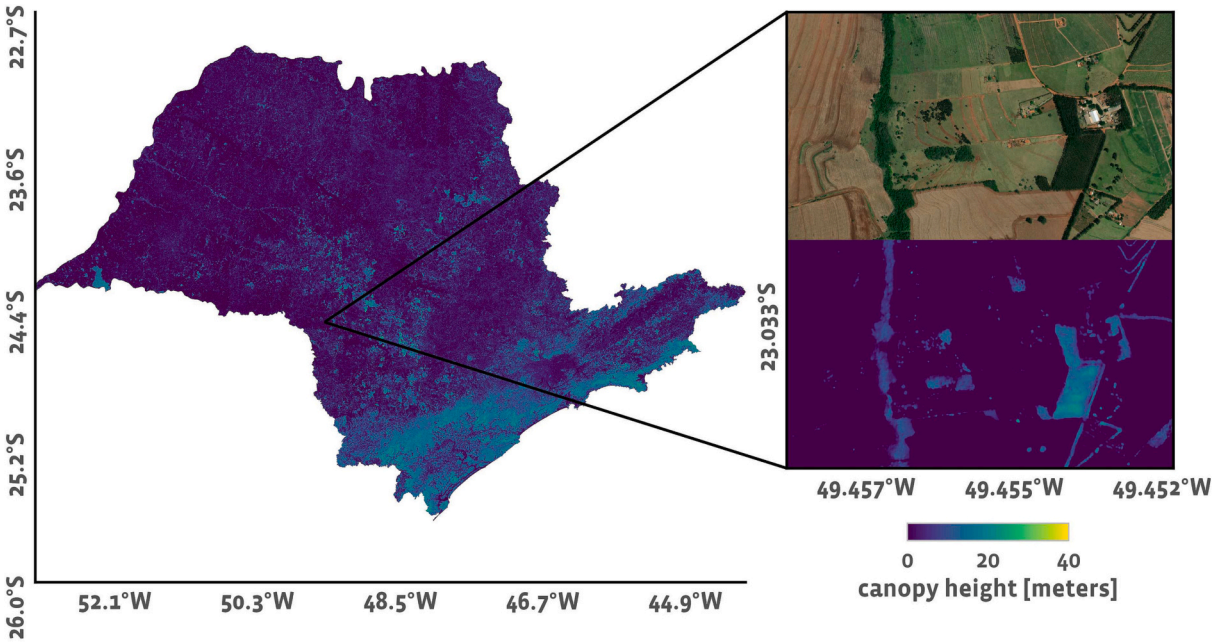


Fig. 5. Canopy Height Map (CHM) for the state of São Paulo, inset showing zoomed in region with input RGB imagery.

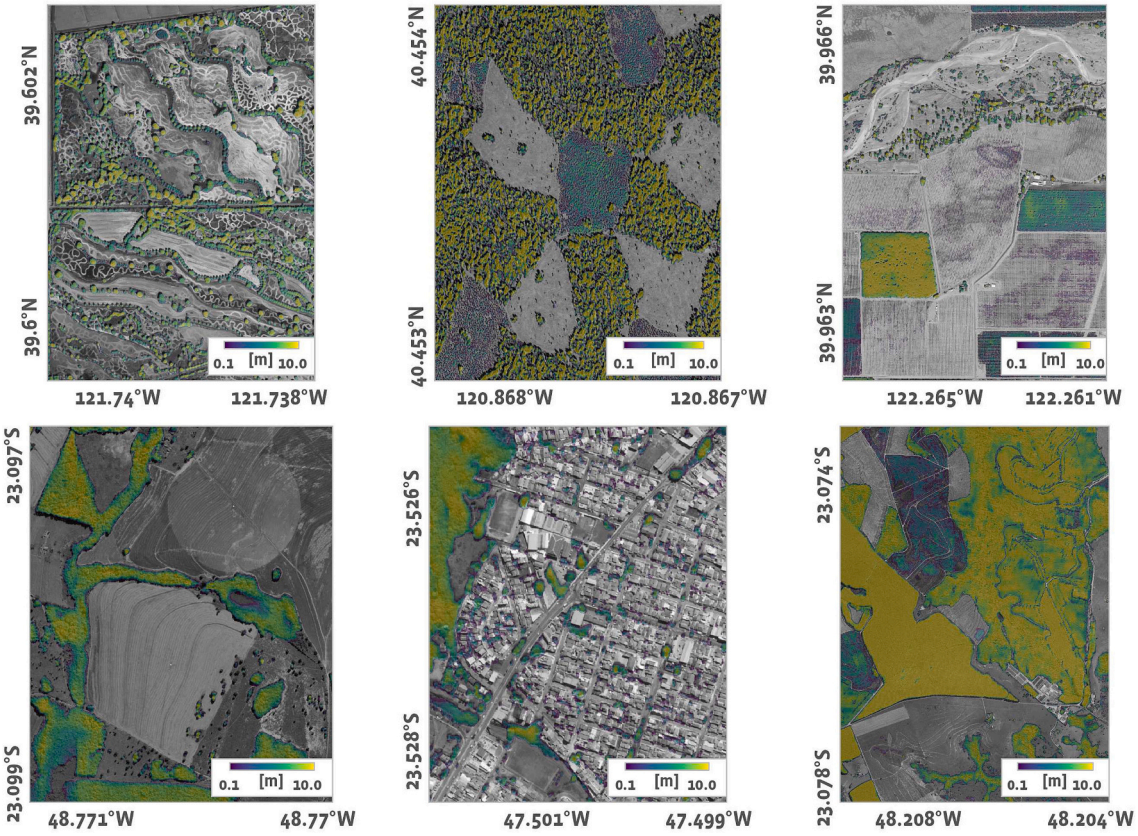


Fig. 6. Selected sample regions from the canopy height predictions (log scale), overlayed on the input Maxar imagery (RGB). Canopy height prediction below 0.1 m is transparent. The top row corresponds to regions in California and the bottom row, São Paulo.

the accuracy of image segmentation into tree or no tree pixels, which we call segmentation metrics (Section 4.2). The set-aside validation dataset of ALS canopy height maps described in Section 2 served as the primary dataset for all types of metrics. For the segmentation metrics, we also evaluated the model predictions against a dataset of human-annotated labels independently labeled by photo-interpretation of Maxar imagery (Section 4.2.1).

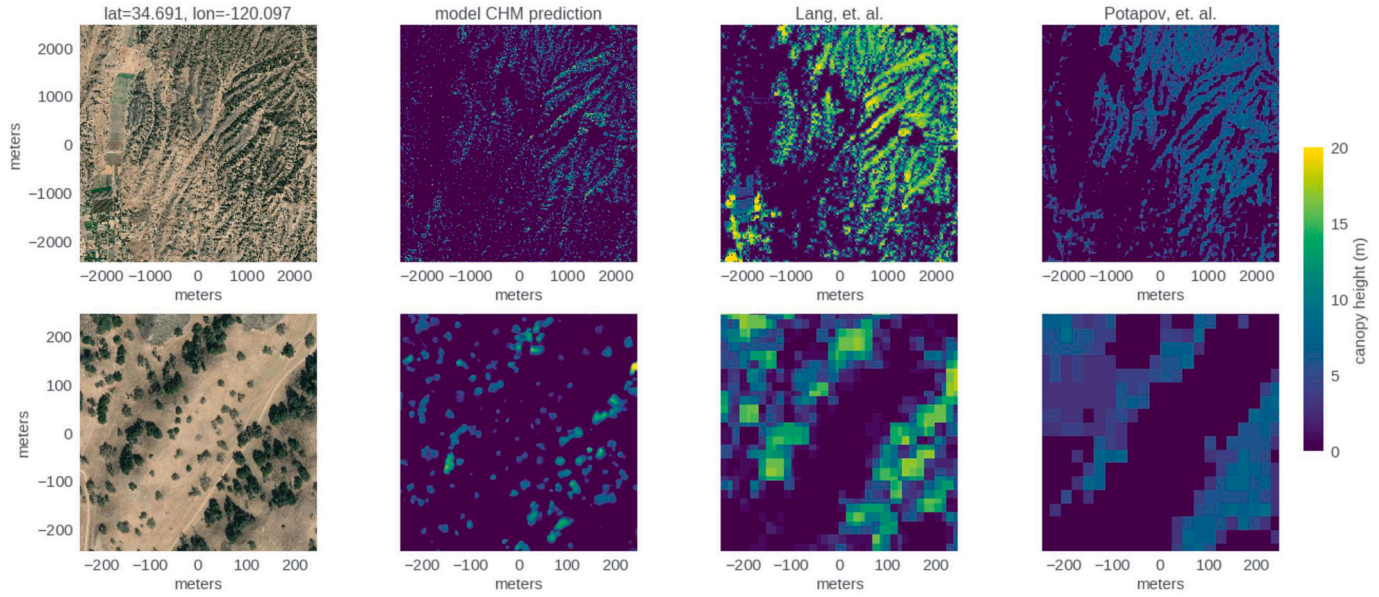


Fig. 7. Comparison of our CHM (second column) with that of Lang et al. (2022a) (third column) and Potapov et al. (2021) (fourth column).

Table 1
Comparison of results with SSL pre-training on different datasets and with other supervised strategies (ResUNet, SWAG). IN: ImageNet. Sat: dataset described in Section 2.3.2. IG: Instagram dataset. R: DPT decoder with a regression (scalar) output. C: DPT decoder with a classification (256 bins) output. ViT L: large, H: huge. Note that the results are non GEDI corrected in this table, and all models were trained with a Sigloss. We later denote the model in the last line as the "SSL" model.

		pre-training	NEON test set			São Paulo			CA Brande		
	model size	dataset	MAE	R^2-block	ME	MAE	R^2-block	ME	MAE	R^2-block	ME
ResUNet	RN18	IN1k	3.1	0.63	**0.0**	5.2	0.42	−2.2	0.6	0.74	**−0.1**
SWAG C	ViT L	IG	3.0	0.63	−1.6	5.8	0.16	−4.3	0.7	0.56	−0.6
DINOv2 R	ViT L	IN1k	3.4	0.52	−1.4	6.8	−0.20	−5.2	0.6	0.67	−0.4
DINOv2 R	ViT H	IN22k+	3.0	0.62	−1.4	5.7	0.27	−2.9	0.6	0.62	−0.4
DINOv2 R	ViT L	Sat 3.5 M	2.8	0.67	−1.2	6.0	0.14	−4.2	0.6	0.70	−0.5
DINOv2 R	ViT L	Sat 18 M	2.9	0.66	−1.4	**4.9**	**0.46**	−1.9	0.6	0.68	−0.5
DINOv2 C	ViT L	Sat 18 M	2.7	**0.70**	−0.9	5.0	**0.46**	−2.1	0.6	0.80	−0.3
DINOv2 C	ViT H	Sat 18 M	**2.6**	**0.70**	−0.9	5.2	0.39	**0.4**	0.6	**0.81**	**−0.1**

4. Results

We generated CHMs for the State of California, USA (Fig. 4) and São Paulo, Brazil (Fig. 5) by running inference on 0.59 m GSD Maxar images with the SSL + GEDI ViT huge model trained with 1 m aerial lidar data. In California, 39% of the area used Maxar imagery observed in 2020, and 90% within the years spanning 2018 to 2020. In São Paulo, 63% of the area was observed in 2019, and 94% within the years spanning 2017–2019. Small regions of the canopy height predictions are visualized in Fig. 6. We compare our maps to the previously available highest resolution, global canopy height maps of Lang et al. (2022a) and Potapov et al. (2021) in Fig. 7. We have added the full resolution dataset to AWS Opendata programs, in the form of cloud optimized geotiffs (COGS) with associated cutlines and image acquisition dates.[3] Additionally, these datasets are visible on a Google Earth Engine public url.[4]

4.1. Canopy height metrics

We compared the predicted canopy height maps with aerial lidar data in terms of mean absolute error (MAE), Root Mean Squared Error (RMSE), and R^2-block (R^2). The R^2-block score is the coefficient of determination, which we computed on cropped images with a resolution of 50 × 50 pixels (∼ 30 × 30 meters). We have chosen the exact size of these blocks somewhat arbitrarily, but were motivated to compute on a scale of 10s of meters due to: a.) georegistration errors in both the Maxar imagery and ALS data, b.) projection differences between the two datasets, with the ALS data being orthorectified and the Maxar imagery have off nadir view angles of up to 30 degrees. As such, the R^2-block score better reflects the local accuracy of CHMs and provides a more direct performance comparison to lower resolution models. However, averages across blocks of this resolution do not provide a good indicator of the edge accuracy of the produced maps, which can be a desirable property for downstream tasks such as segmentation. We separately report the Edge Error (EE) metric we developed to measure the sharpness of the maps, described in Appendix C.3. Finally, to estimate the bias of different models, we report the Mean Error (ME). We provide formulas for the above mentioned metrics in Appendix C.

4.1.1. Canopy height metrics for ALS models

We present in Table 1 an ablation study of different pre-training data, model size and output on the Neon and São Paulo test sets. From this ablation study, we selected the SSL model trained on 18 million images utilizing the classification output, which achieved the highest canopy height accuracy metrics. We also trained a huge model instead of a large one, that significantly reduced the bias of the predictions on the São Paulo dataset. We refer to this model as the SSL model throughout the paper. Table 1 suggests that pre-training on satellite images gives better results compared to pre-training on ImageNet. Compared to the ViTs

[3] https://registry.opendata.aws/dataforgood-fb-forests/.
[4] https://wri-datalab.earthengine.app/view/submeter-canopyheight.

Table 2
Canopy Height Metrics to assess the gedi correction step. R^2 corresponds to $\sim 30 \times 30$ meter block R^2. "Average" is the unweighted average across datasets.

	NEON test			CA-Brande			São Paulo			Average		
	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2	MAE	RMSE	R^2
ResUNet	3.1	4.9	0.63	0.6	1.6	0.75	5.2	7.4	**0.42**	3.0	4.6	0.60
ResUNet + GEDI	3.0	4.8	0.64	0.6	1.6	0.74	5.4	7.7	0.35	3.0	4.7	0.58
SSL	**2.6**	**4.4**	**0.70**	0.6	**1.4**	**0.82**	5.2	7.5	0.39	**2.8**	4.5	**0.64**
SSL + GEDI	2.7	4.5	0.69	0.6	1.5	0.80	**5.1**	**7.3**	0.41	**2.8**	**4.4**	0.63

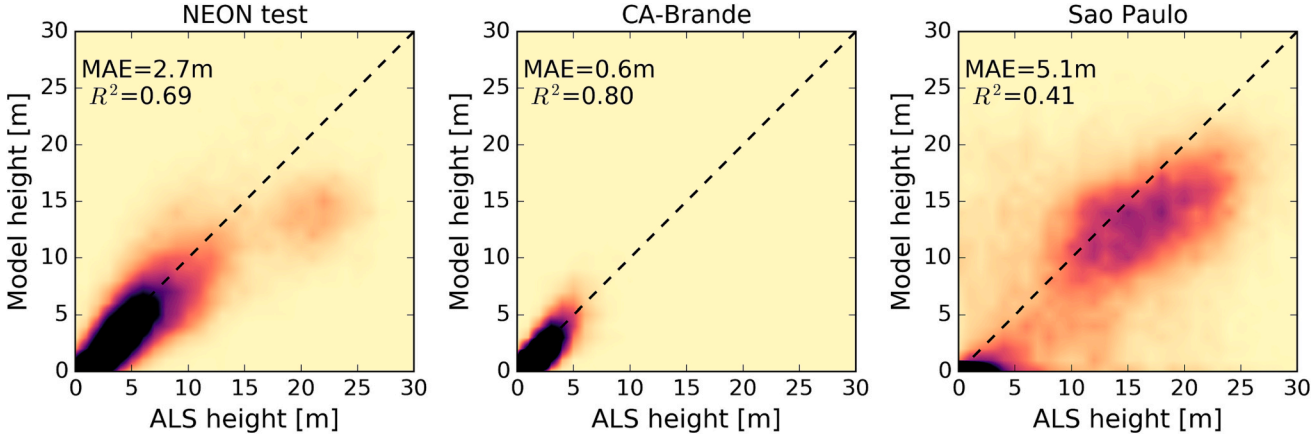
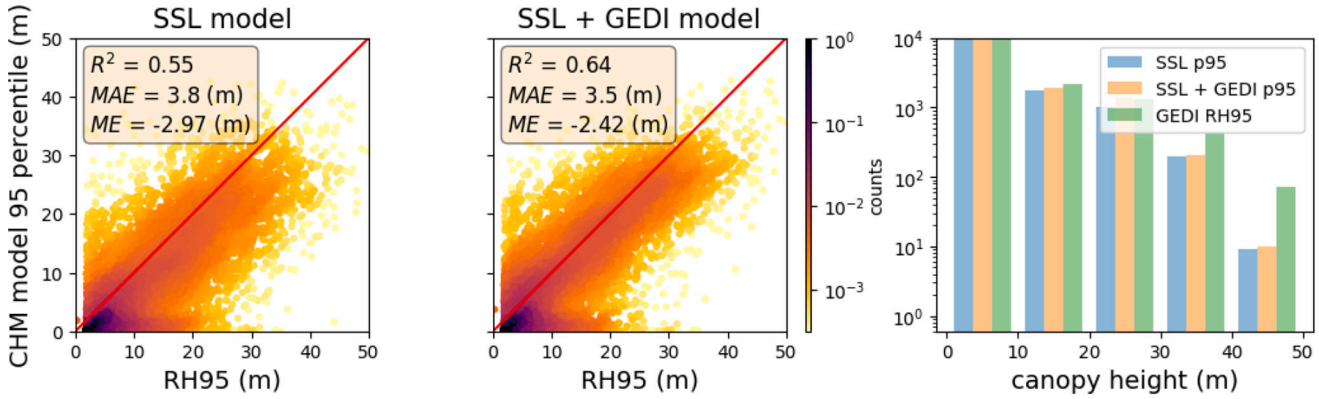
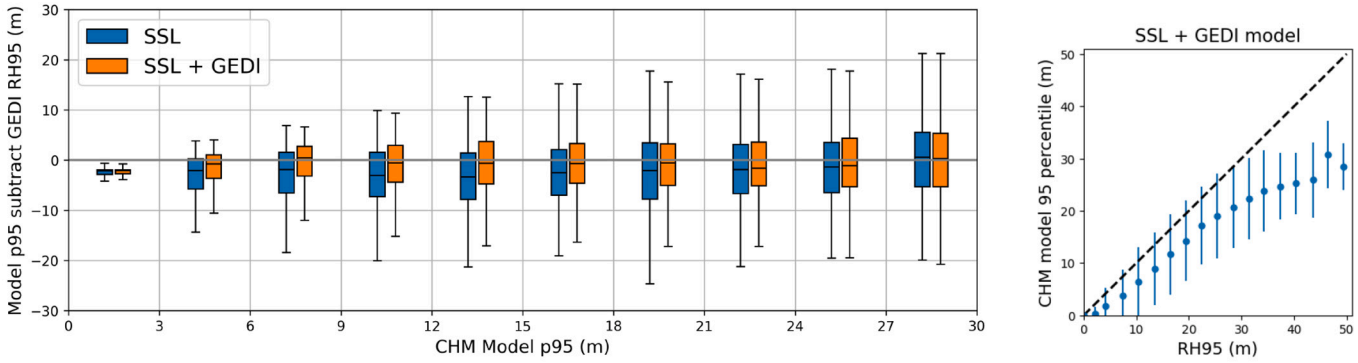


Fig. 8. Block ($\sim 30m \times 30m$) aggregated SSL + GEDI model predictions compared to ALS ground truth measurements for different set-aside validation datasets. Colormap density is normalized to the 99.6th percentile of the heatmaps.



(a) MAE: mean absolute error. ME: mean error. R^2: Coefficient of determination.



(b)

Fig. 9. Global model evaluation on held-out GEDI data. (a) p95 of block ($76m \times 76m$) model CHM predictions compared to the measured GEDI RH95 metrics. (b) Left: Difference between the p95 of block model CHM predictions and the measured GEDI RH95 metrics w.r.t model CHM predictions. Negative values indicate that the model estimates are lower than the GEDI RH95 values. Residuals in function of RH95 appear in Appendix F. Right: CHM p95 in function of RH95.

Table 3
R^2 between predicted CHM p95 and GEDI RH95 by geographic subregion for 20,000 test GEDI observations for models with and without the GEDI calibration model.

Subregion	SSL		ResUNet		SWAG	
GEDI	+	−	+	−	+	−
Central Asia	0.22	0.19	**0.25**	0.23	0.23	0.17
Eastern Asia	**0.50**	0.44	0.47	0.42	0.43	0.38
Eastern Europe	**0.70**	0.66	0.67	0.61	0.67	0.63
Latin America + Caribbean	**0.68**	0.64	0.65	0.56	0.64	0.56
Melanesia	**0.52**	0.48	0.51	0.41	0.44	0.45
Northern Africa	**0.12**	0.11	0.10	0.06	0.06	0.05
Northern America	**0.73**	0.69	0.70	0.65	0.69	0.64
Northern Europe	**0.54**	0.46	0.41	0.30	0.33	0.33
Oceana	**0.68**	0.63	0.66	0.58	0.61	0.54
South East Asia	**0.46**	0.36	0.45	0.34	0.44	0.32
Southern Asia	**0.52**	0.49	**0.52**	0.48	0.47	0.42
Southern Europe	0.46	**0.47**	0.42	0.37	0.46	0.40
Sub-Saharan Africa	**0.68**	0.66	0.58	0.50	0.64	0.59
Western Asia	**0.53**	0.49	**0.53**	0.47	0.47	0.42
Western Europe	**0.64**	0.59	**0.64**	0.55	0.58	0.50
Overall	**0.61**	0.52	0.59	0.44	0.54	0.37

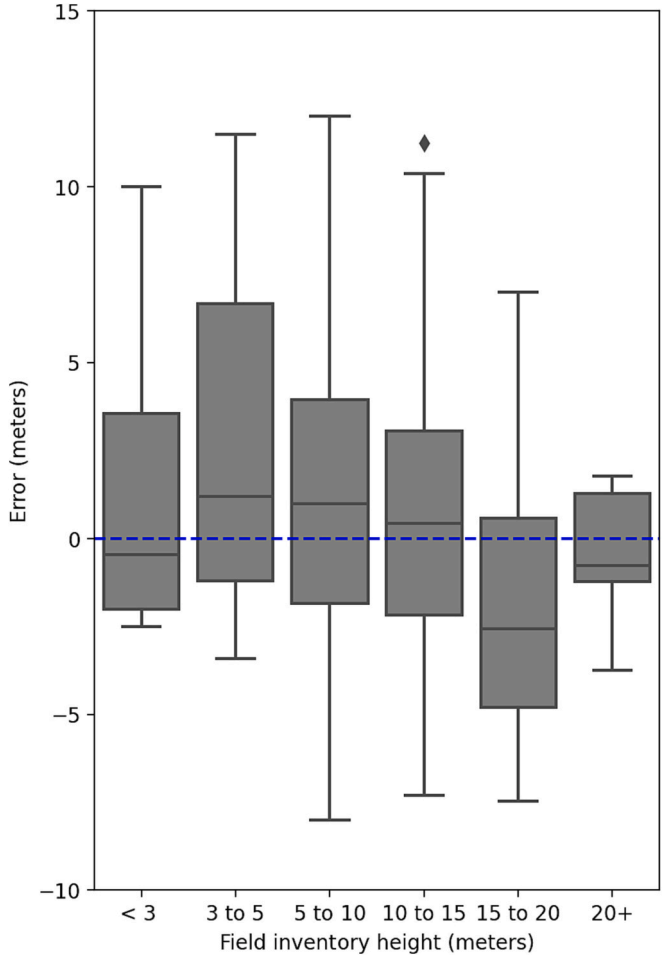


Fig. 10. CHM error compared to reference tree height as indicated in the Brazilian National Forest Inventory for Espirito Santo.

that are pre-trained on ImageNet, including the SWAG approach, the ResUNet remains the strongest baseline. The SSL model clearly outperforms the ResUNet on Neon, reducing the MAE from 3.1 to 2.6 m, is also improving results on CA Brande, and leads to similar results on São Paulo, with a slightly worse R^2 but a much lower ME. We also experimented with different loss functions, and a smaller dataset for self-supervised pre-training. We found that was training on more data was leading to much better results in São Paulo. Comparing L1, L2 and Sigloss, we found that Sigloss and L2 were leading to the best results. Additional discussion of these trials can be found in Appendix E.

4.1.2. Canopy height metrics for ALS + GEDI models

Table 2 presents a quantitative validation of the best performing models, namely the ResUNet and the self-supervised model (SSL), combined with the GEDI correction step (ResUNet+GEDI, SSL + GEDI). We note the improved performance of the SSL model compared to the ResUNet in the NEON test and CA-Brande datasets. Although the SSL model performed the best across the datasets in the USA (NEON test and CA-Brande), it performed worse than the ResUNet and ResUNet + GEDI for São Paulo, possibly due to the large domain shift in ecosystems. In the case of São Paulo, we found that the inclusion of GEDI ("SSL + GEDI") produced the best results, possibly indicating better generalization by including the globally trained GEDI model, which also includes additional metadata such as geographic position (Fig. 2).

Fig. 8 shows 2D-histograms of the SSL + GEDI model predictions vs the set-aside validation ALS-derived canopy height averaged over ∼ 30m blocks and the corresponding pixel MAE and block-R^2 scores.

4.1.3. Quantitative comparison of CHM model with GEDI RH95 data

The ALS set-aside validation datasets used in the previous section are limited in both total area and geographic coverage. In this section, we leverage the global coverage of the GEDI dataset to validate our CHM models. As described in Appendix B.2, CHM maps can be connected to GEDI RH95 metrics by taking the 95th percentile. In this analysis, we draw 2×10^4 GEDI samples globally in the set-aside validation split the same way as in training the GEDI model, i.e., weighted proportional to the square root of the inverse sample size of its RH95 bin. Similarily to Potapov et al. (2021), we removed GEDI observations corresponding to < 0.5 normalized difference vegetation index (NDVI) that had no tree cover in the 2010 data of Hansen et al. (2013), corresponding to 337 samples out of 20,000. In Fig. 9a, we show the scatter plot and histogram of the 128×128 pixels ($76m \times 76m$) block 95^{th} percentile vs. the measured GEDI RH95. In Fig. 9b, we analyze the difference of the CHM p95 and the GEDI RH95 with respect to the referenced GEDI RH95 heights.

We found that the p95 of the CHM model had a small negative bias against the GEDI RH95 values and adding the GEDI correction to the CHM model significantly reduces the bias. There is a slight positive bias in the GEDI RH95 data due to the terrain slope (Lang et al., 2022a). We used terrain slope (Mapzen, 2017) as one of the input metadata when training the GEDI model (see Section 3.3), while setting the terrain slope to zero during inference. With this setup, we were able to calibrate out the positive bias caused by terrain slope in our GEDI model.

To assess the importance of the GEDI calibration model for geographic generalization, and the generalizability of the different baseline models, we computed R^2 on globally distributed GEDI test data (Table 3).

We found that the SSL + GEDI model had the highest agreement with GEDI RH95 data in 13 of 15 geographic regions. In 42 out of 45 combinations of subregions and models, including the GEDI correction model increased R^2.

4.1.4. Correlation with field data

To measure the agreement between our computed CHMs and field-collected tree height data, we utilize the Brazilian National Forest Inventory (NFI) data, which consists of systematic field plot inventories of tree count and height (da Luz et al., 2018). Because the NFI data for São Paulo was not yet available, we additionally generate a CHM of the nearby Espirito Santo state and evaluate its agreement with the NFI data for Espirito Santo. The NFI data analyzed encompassed 1450 10×10 m

Table 4
Segmentation metrics. U/P corresponds to pixel user's / producer's accuracy of the tree class. IOU to the average of tree & no tree IOU class scores. EE: Edge error.

	NEON test		CA-Brande		São Paulo		Average		
	U/P	IOU	U/P	IOU	U/P	IOU	U/P	IOU	EE
ResUNet	0.74/0.75	0.58	0.72/0.64	0.70	**0.91**/0.85	**0.67**	0.79/0.75	0.65	0.50
ResUNet + GEDI	0.77/0.68	0.53	0.73/0.52	0.68	**0.91**/0.84	0.65	0.80/0.68	0.62	0.52
SSL	0.81/**0.76**	**0.65**	0.71/**0.75**	**0.76**	0.90/**0.88**	**0.67**	0.82/**0.81**	**0.68**	0.50
SSL + GEDI	**0.82**/0.71	0.59	**0.74**/0.60	0.74	**0.91**/0.86	0.66	**0.83**/0.76	0.66	**0.49**

Table 5
Segmentation metrics on global, human annotated dataset. U/P corresponds to pixel user's / producer's accuracy. IOU to the average of tree & no tree IOU scores. Since the GEDI correction only adjusts large scale height percentiles, the "+GEDI" rows show only small improvement over the base ALS models.

	Global, Annotated	
	U/P	IOU
ResUNet	0.89/0.86	0.75
ResUNet + GEDI	**0.90**/0.86	0.74
SSL	0.83/0.87	**0.77**
SSL + GEDI	0.82/**0.88**	**0.77**

subplots within 87 plots positioned within a 20 × 20 km grid in Espirito Santo. The field data was collected primarily in November and December 2014, and includes the height of each tree within each subplot having a diameter at breast height (DBH) of at least 10 cm. Of the 1450 initial plots considered, we removed 291 that had tree cover loss since 2014 in the dataset of Hansen et al. (2016). Fig. 10 visualizes box plots of the 95th percentile CHM by reference NFI height bins. The overall ME was 0.72 m while the RMSE was 4.25 m, with a slight positive bias for trees ≤15 m (ME = 1.10 m, RMSE = 4.28 m), and negative bias for trees >15 m (ME = −1.00 m, RMSE = 3.79 m).

4.2. Segmentation metrics

In addition to the canopy height metrics discussed in Section 4.1, we compute a number of metrics that reflect the ability of the model to correctly assign individual pixels as trees. CHMs were converted into binary masks by thresholding height values of at least five meters (5 m) as tree canopy extent. Table 4 shows the pixel user's and producer's accuracy values (also know as precision and recall, respectively) for pixels labeled as trees. Table 4 also shows the Intersection Over Union (IOU) for the binary masks, which was calculated as the average of IOU

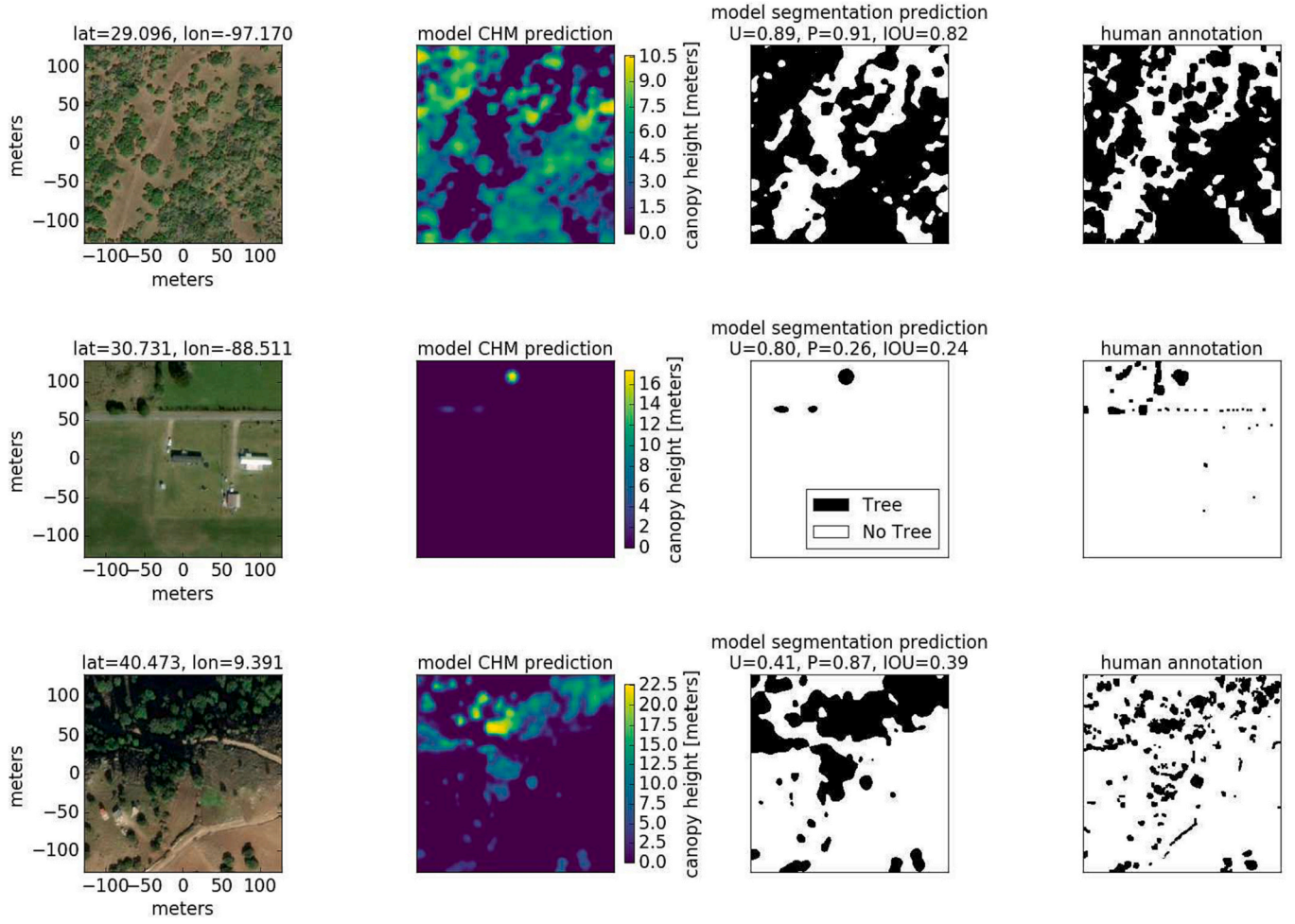


Fig. 11. Tree segmentation predictions from the SSL + GEDI model vs human annotated ground truth. Binary prediction masks were created from the CHM by thresholding at 1 m. U/P corresponds to pixel user's / producer's accuracy of the tree class. The IOU represents the Intersection-Over-Union score for the tree class.

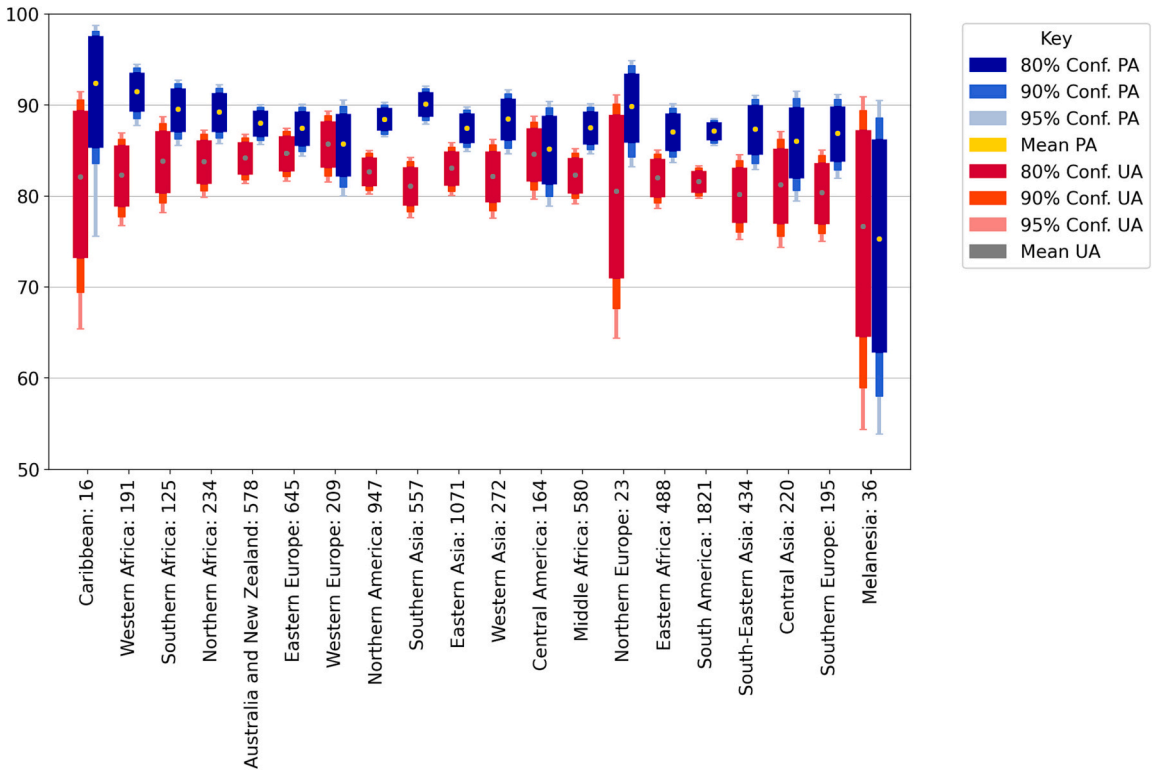


Fig. 12. Pixelwise user's accuracy (UA) and producer's accuracy (PA) for 8903 validation plots stratified by geographic sub-region. Error bars represent the 80, 90, and 95% confidence intervals as derived from 10,000 bootstrap iterations. Numbers in the x-axis tick labels denote sample size.

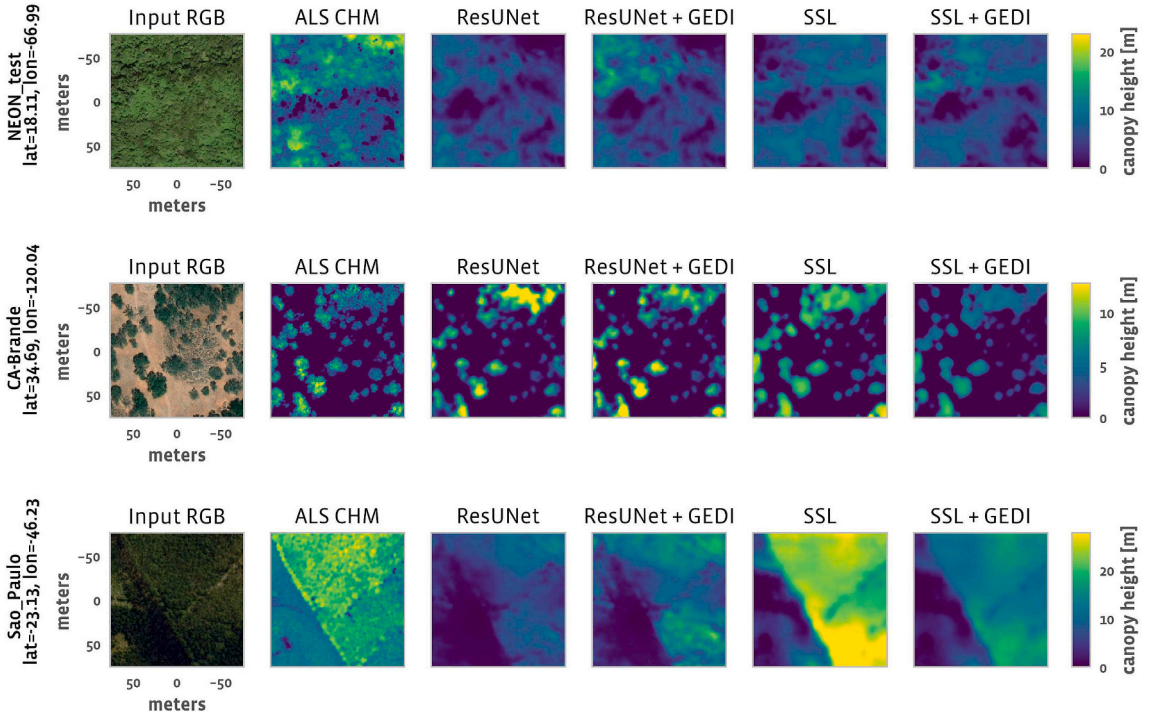


Fig. 13. Qualitative comparison between different models for example imagery. Left: Input Maxar "thumbnail" image, 256 × 256 pixels, in local tangent plane coordinate system. Second from left: ALS CHM image, in same projection and pixelization. Right columns: Model CHMs.

for pixels labeled as tree and the IOU for pixels labeled as ground.

Additionally, we introduce an Edge Error (EE) metric that computes the ratio of the sum of the absolute difference between edges from predicted and ground truth CHM, normalized by the sum of detected edges in both maps. Scores range between 0 and 1, where lower scores indicate improved accuracy along patch edges. In Table 4, the edge error is computed over all set-aside validation datasets. We detail the formula with a figure illustrating the behavior of this metric in Appendix C.3.

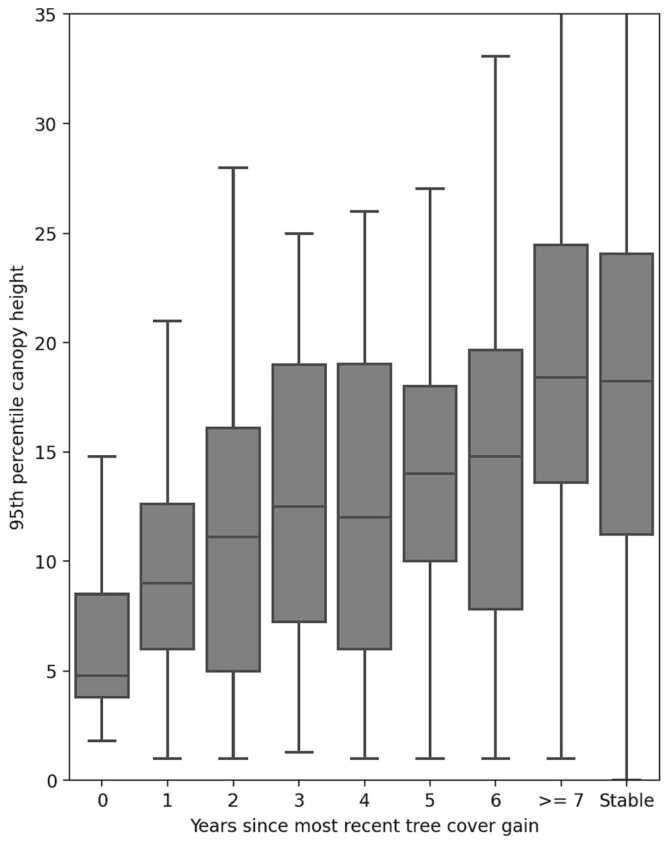


Fig. 14. Canopy height estimates for areas with tree cover gain of various ages in São Paulo relative to the imagery year analyzed.

Table 6
CHM prediction accuracy on NEON test dataset using aerial input images as inputs. Trained on satellite images only, the SSL approaches demonstrates generalization abilities.

	Neon test - aerial					
	Encoder training dataset	Decoder training dataset	MAE	block R^2	ME	EE
ResUNet	INet	Sat. images	3.7	0.34	−2.0	0.77
SSL	Sat. images	Sat. images	3.0	0.55	1.7	0.71
SSL	Sat. images	aerial	**1.8**	**0.86**	**−1.0**	**0.41**

We observe an improvement of the SSL approach over the ResUNet baseline in terms all segmentation metrics. Both approaches leads to maps with the same level of sharpness, and the GEDI correction slightly degrades results.

4.2.1. Tree detection metrics evaluated against human annotated validation data

To assess the ability of the model to generalize to new geographies, we compiled human-annotated validation labels for tree detection (binary classification of tree vs no-tree) across 8,903 Maxar thumbnail images. Human annotators were instructed to label any trees above one meter (1 m) tall and with a canopy diameter of more than three meters (3 m). Annotators were to include standing dead trees and tree-like shrubs, but exclude any grasslands, row crops, lawns, or otherwise vegetative ground cover whose peak height was estimated to be less than 1 m from the ground surface. To create the model binary masks for the annotated dataset, we thresholded the model CHM at 1 m.

The geographic locations for the images in the dataset correspond to randomly sampled GEDI measurement footprints from our global set-aside validation set where the GEDI measurement had RH95 greater than 3 m, which we enforce to bias the dataset towards vegetated areas. The data is independent of the aerial lidar measurements used to train the model. Over the entire dataset, the user's and producer's accuracy was 0.88 ± 0.006 and 0.82 ± 0.008, while the IOU was 0.77 ± 0.006 indicating good agreement with the human annotations, cf. Table 5. Fig. 11 shows examples of model predictions and their corresponding annotations.

We additionally calculated user's and producer's accuracy by geographic subregion according to the United Nations geoscheme. Boostrapping with 10,000 iterations was used to calculate uncertainty for tree segmentation accuracy metrics rather than the methods of Stehman (2014) because the cluster sampling approach was used to generate validation data (Olofsson et al., 2014; Mugabowindekwe et al., 2022; Maxwell et al., 2021). This validation analysis indicated strong generalizability across different geographic regions, without significantly different accuracy metrics in geographic regions where we had training data and where we did not (Fig. 12). This suggests that the use of self supervised learning on global images facilitated the creation of a generalizable segmentation network.

4.3. Qualitative comparison of models

Although we have attempted to capture the performance of each model qualitatively with the included metrics, we note that visual inspection often leads to additional insights. Therefore, we additionally present a few examples of maps produced by our various models. Fig. 13 compares the results with a ResUNet and SSL based strategies.

4.4. Canopy height as a function of plantation age

Densely planted monoculture stands, such as those commonly found in the Atlantic forest, can be many hundreds of hectares large. Assessing the age-height relationship of tree stands with CHMs derived from optical imagery may be challenging due to the relative homogeneity of the canopy structures, the large area to perimeter ratio, and the lack of canopy gaps. To assess the CHMs ability to map the height of planted trees, we utilized the annual 30-m tree cover gain and loss data from MapBiomas in São Paulo (Azevedo et al., 2018). We calculated the number of years since the most recent tree cover gain with no subsequent loss event for each image date analyzed. Fig. 14 shows a positive relationship ($R^2 = 0.59$) between the number of years since the most recent tree cover gain, and our predicted 95th percentile CHM. For areas with gain events older than seven years, there was no significant age-height relationship, as areas with trees with gain events more than seven years prior to the analysis year had similar height distributions to areas with stable (no gain or loss since 2000) trees. For this analysis, it's important to note that the tree cover gain year identified in MapBiomas is a lagging indicator of the tree age, since tree cover gain is not immediately recognizable from Landsat imagery.

4.5. Generalization to aerial imagery

Using a model fully trained on Satellite images. To assess the generalization ability of our approach to other input imagery, we measure model performance using airborne imagery at inference. For inference, we resized the NEON aerial images to match the size of corresponding satellite image, and apply a normalization of the aerial image to match the color histogram of the satellite imagery. Details about image normalization are provided in Appendix G.

The second line of Table 6 shows canopy height metrics computed on predictions made from NEON input RGB imagery. The SSL model almost doubles the R^2 of the ResUNet baseline. Compared to the performance of the SSL model with satellite images as input as reported in Table 1, the MAE is only slightly higher (3.0 instead of 2.7), the R^2 is a bit more

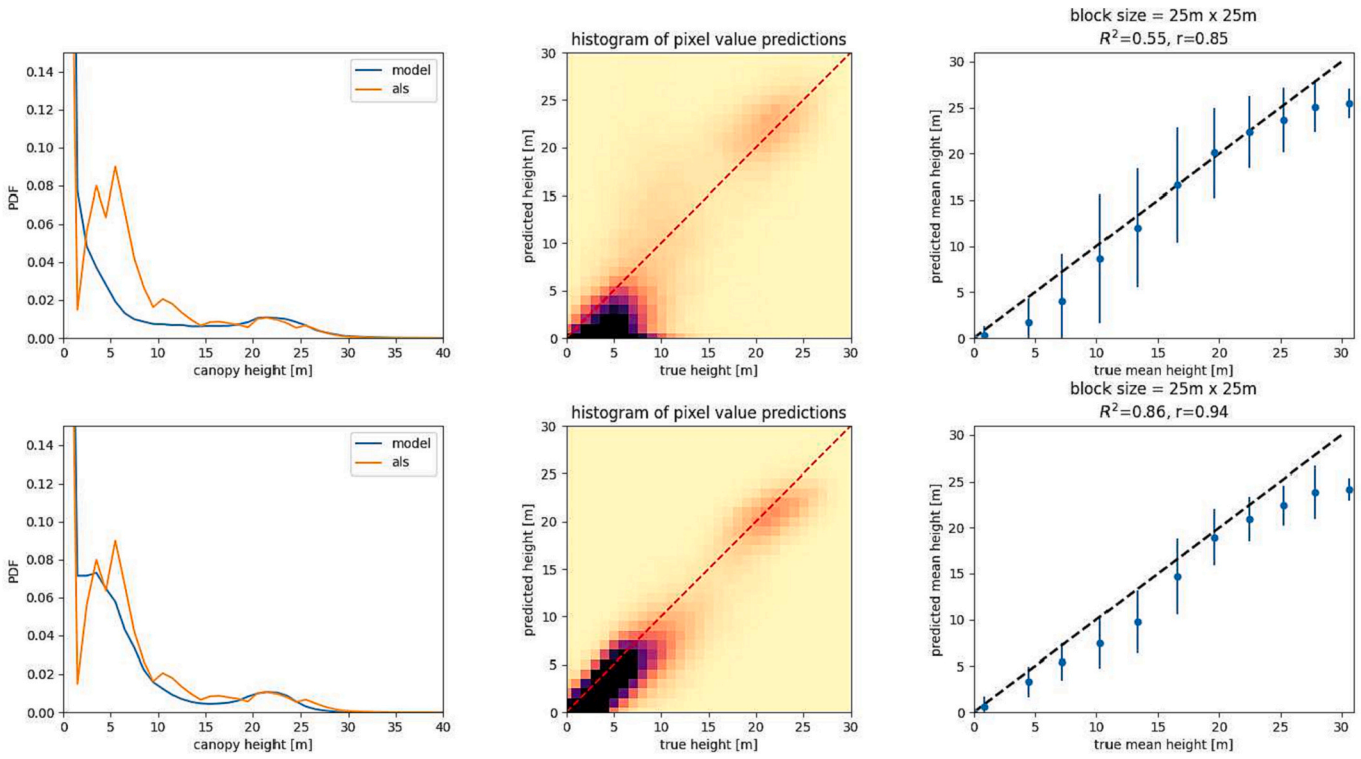


Fig. 15. Performance of models given aerial images inputs. Top: Model fully trained on satellite images; Bottom: Performance of encoder trained on satellite images, decoder trained on aerial images.

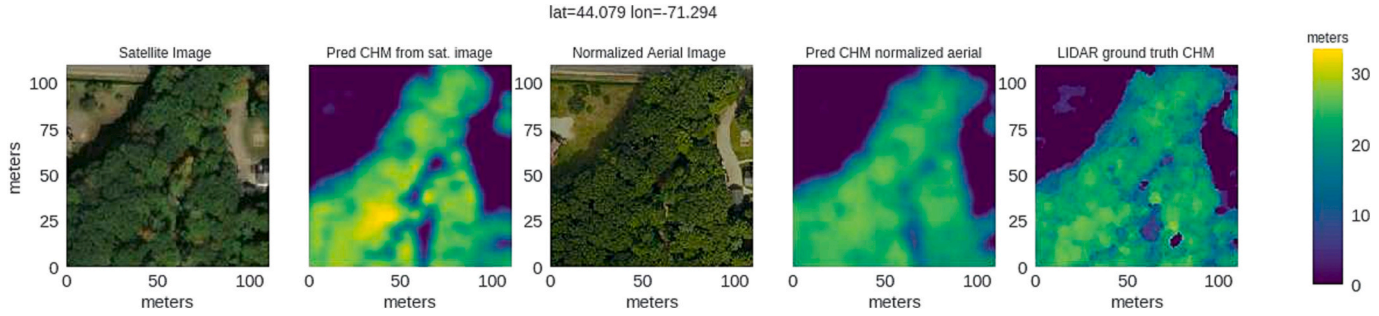


Fig. 16. Generalization of our SSL approach. Even if trained on Satellite images, inference on airborne images does not seem to suffer from a domain shift.

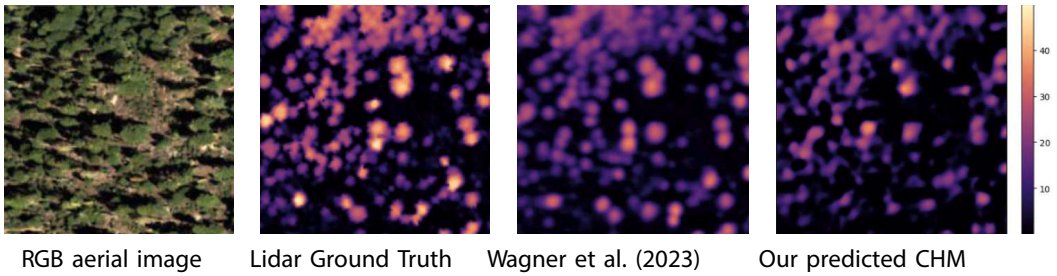
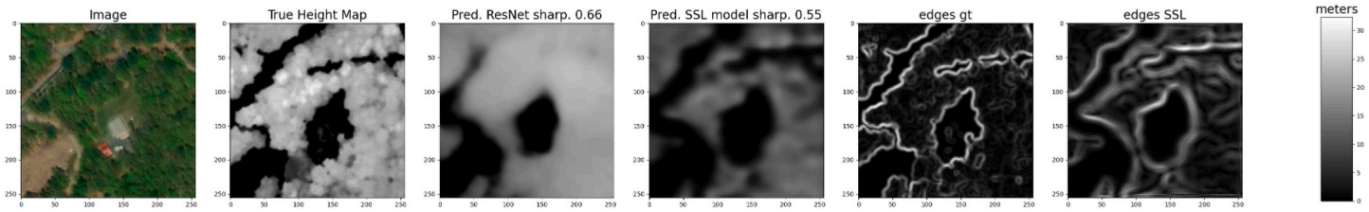


Fig. 17. Comparison of our aerial model, where we trained the DPT decoder on Neon aerial RGB images, with the approach of Wagner et al. (2023). Note that despite a slight change in the scale of the input image, which was zoomed to obtain a 256 × 256 input, and despite the fact we did not use the infra-red input, we obtain a result similar to the one of Wagner et al. (2023). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

impacted (0.55 instead of 0.70), while the bias is much higher, but evenly distributed between different height bins (Fig. 15). Fig. 16 displays a qualitative example, where we observe that despite a blurrier result, the accuracy of the model given an out-of-domain aerial image seems similar to the one obtained using in domain satellite imagery.

Despite changes in color intensity, image angle, and sun angle, our approach manages to generate predictions with consistent visual quality. From an application point of view, the robustness of SSL predictions



without the need to retrain models on new lidar datasets is very interesting.

Training a new decoder on aerial images. We compared these results to another baseline, training a decoder on top of our pretrained SSL features on Neon RGBs (last line of Table 6). Given a better alignment with the ground truth CHMs, and view angles close to Nadir across the Neon dataset, this aerial model performed reasonably well compared to the recent result of Wagner et al. (2023), only using the RGB channels, as illustrated in Fig. 17.

5. Discussion

Our proposed method provides a novel approach to estimating canopy height from VHR satellite imagery. We demonstrate the effectiveness of our approach based on self-supervised learning, dense vision transformers, and introduce an approach to rescale high-resolution canopy height maps from a model trained on Maxar and ALS data with low-resolution canopy height maps from a model trained on Maxar and GEDI data. In contrast to Lang et al. (2022a), which downscales the 25-m GEDI data to generate 10-m canopy height maps by only considering Sentinel-2 pixels at the centroid of each GEDI pixel, our approach uses a GEDI-based canopy height map to rescale an ALS-based model of canopy height map. While both Lang et al. (2022a) and Potapov et al. (2021) only utilize ALS data to validate their generated maps, we directly model the relationship between Maxar imagery and ALS data, enabling the mapping of sub-GEDI scale canopy height variability, some times at a per-tree level outside of dense, closed-canopy forests.

Segmentation. Previous research applying deep learning image segmentation approaches to map trees in high-resolution imagery, such as Brandt et al. (2020) and Mugabowindekwe et al. (2022) have utilized a U-Net model (Ronneberger et al., 2015) and entirely hand-labeled reference data. Focusing in Rwanda, Mugabowindekwe et al. (2022) map carbon stock estimates for individual trees by developing empirical relationships between crown area and carbon, finding that half of the national tree carbon stock is located outside of natural forests. In comparison to these approaches, our results suggest that incorporating SSL can improve model generalizability for vegetation structure mapping, in line with various research demonstrating the effectiveness of SSL in other domains. Additionally, our per-pixel height predictions combine the predictive quality of height for assessing biomass as demonstrated in Lang et al. (2022b) and Potapov et al. (2021) with the predictive quality of crown area for assessing biomass as demonstrated in Mugabowindekwe et al. (2022) and Skole et al. (2021).

Limitations. The production of high-resolution canopy height maps from optical imagery has challenges and associated limitations. Foremost, the availability of recent ALS training data is limited in geographic scope. While the utilization of self-supervised learning and the GEDI-based corrective model improve generalization and reduce test error, increased geographic availability of ALS remains necessary to further validate the proposed maps in new geographies. While we were able to validate error as a function of canopy height for trees under 25 m based on field inventory data in Espirito Santo, Brazil (Fig. 10), we were unable to utilize field data to assess potential height saturation for very tall trees which may affect derived above ground carbon data. However, Fig. 9a does suggest significant saturation of predictions for GEDI RH95 observations above 30 m.

The generated maps are limited by variation in input imagery, particularly by variation in view angle, sun angle, acquisition times and dates, and optical aerosol. As shown in Fig. 17, qualitative data quality improves considerably when processed on VHR aerial optical imagery, as opposed to VHR satellite optical imagery. Additionally, terrain slope appears to influence predicted height, since it affects the length of shadow an individual tree casts. At present, the ability to conduct tree height change detection assessments is limited by the need for improved input image processing to better align these differences between image pairs.

6. Conclusion

This study presents high-resolution canopy height maps at a jurisdictional scale based on VHR (Maxar) optical imagery trained on aerial lidar and calibrated with spaceborn lidar (GEDI) data using latest advances from self-supervised learning and vision transformers. We demonstrate quantitatively and qualitatively the advantages of large-scale self-supervised learning, the versatility of obtained representations allowing generalization to different geographic regions and input imagery. Compared to existing canopy height maps, the presented data better captures tree structure variability at small spatial scales. Such very high resolution maps of canopy height can improve the monitoring of forest degradation, restoration, and forest carbon dynamics. The next steps include (a) developing and validating allometrically-derived high-resolution woody carbon data and (b) testing and validating the utility of the proposed approach for the operation monitoring of tree growth.

CRediT authorship contribution statement

Jamie Tolan: Conceptualization, Supervision, Writing – original draft. **Hung-I Yang:** Investigation, Data curation. **Benjamin Nosarzewski:** Validation, Data curation, Investigation. **Guillaume Couairon:** Methodology, Investigation, Writing – review & editing. **Huy V. Vo:** Methodology, Investigation. **John Brandt:** Writing – original draft, Visualization, Formal analysis. **Justine Spore:** Writing – original draft, Visualization. **Sayantan Majumdar:** Investigation. **Daniel Haziza:** Software, Data curation. **Janaki Vamaraju:** Software. **Theo Moutakanni:** Methodology. **Piotr Bojanowski:** Conceptualization. **Tracy Johns:** Supervision. **Brian White:** Methodology. **Tobias Tiecke:** Supervision, Visualization. **Camille Couprie:** Conceptualization, Methodology, Investigation, Writing – original draft.

Declaration of Competing Interest

None.

Data availability

Input imagery is licensed by Maxar, and not available publicly. We share the derived maps of canopy height under Creative Commons 4.0 and are available for public download.

Acknowledgments

We would like to thank Ben Weinstein for helpful discussions regarding the NEON dataset. We thank Andi Gros and Saikat Basu for their technical advice. We would like to thank Shmulik Eisenmann, Patrick Louis, Lee Francisco, Leah Harwell, Eric Alamillo, Sylvia Lee, Patrick Nease, Alex Pompe and Shawn Mcguire for their project support.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Data used in training / calibration / validation

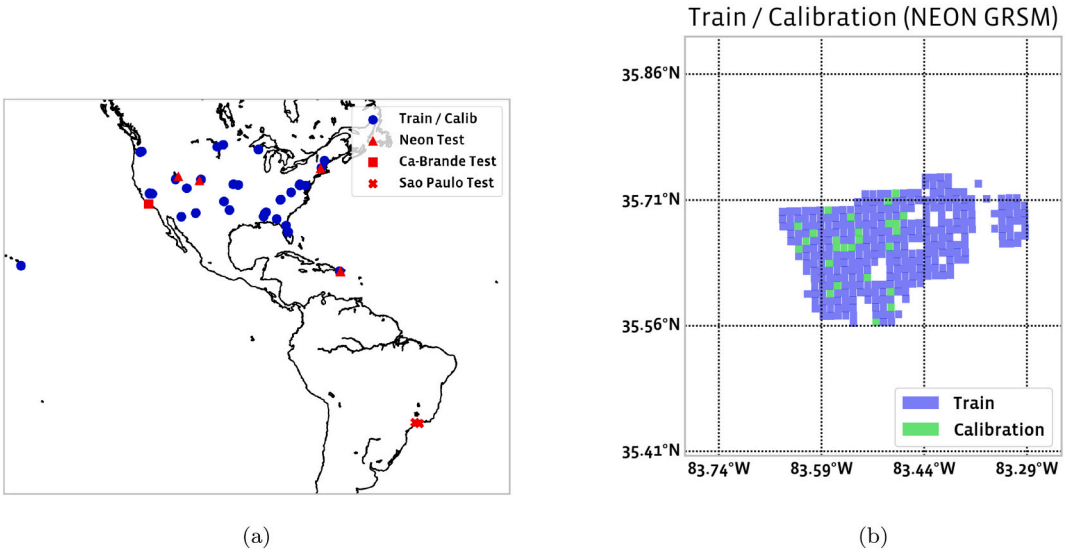


Fig. A.18. Distribution of ALS Datasets: Train/Calibration/set-aside validation (aka Train/Validation/Test): (a) All ALS datasets. Here Train and Calibration points overlap and are shown in blue. Set-aside validation (Test) datasets are from non-overlapping geographic regions. (b) Zooming in on one Train / Calibration site (NEON GRSM) - we have randomly split the data into non spatially overlapping tiles so that calibration data is drawn from the same sites and ecosystems as training data, but separated spatially. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The NEON sites during training / calibration are: SJER, SOAP, TEAK, BART, DSNY, HARV, JERC, OSBS, DELA, GRSM, LENO, MLBS, BLAN, CLBJ, KONZ, NOGP, SCBI, TALL, UKFS, WOOD, ABBY, BONA, DEJU, JORN, MOAB, OAES, ONAQ, SERC, SRER, UNDE, WREF, HEAL, LAJA, RMNP, PUUM.

The set-aside validation dataset, "NEON test", contains the following NEON sites: CUPE, REDB, WLOU, HOPB, GUAN.

To ensure repeatability of our approach, we provide a complete list of CHM files used during training/calibration at: https://dataforgood-fb-data.s3.amazonaws.com/forests/v1/NEON_training_images.csv

Appendix B. GEDI Dataset and model training details

B.1. GEDI dataset

The GEDI instrument is a full waveform lidar instrument aboard the International Space Station which has sampled global regions between 51.6° N & S latitude with a ~25m beam footprint at ground surface. The instrument details are described in Dubayah et al. (2020), and its measurements of canopy height are described in Dubayah et al. (2022). We used the GEDI L2A Version 2 product and filtered the dataset to reduce noise by only including data which had: degrade flag $= 0$, surface flag $= 1$, solar elevation < 0, and sensitivity > 0.95. After this filtering, we were left with a total sample size of 1.3×10^9 measurements. We used the 95th percentile of relative height (RH95) that we paired to 128×128 pixel (76×76 meter) satellite images from Maxar. These images were processed as described in Section 2.3.1, but were smaller to more closely approximate the GEDI footprint. Although these images are still significantly larger than the 25 m GEDI footprint, we have found improved results from our GEDI model using larger areas - potentially due to pointing errors in the GEDI data and a larger spatial context improving the CNN model results.

B.2. Connection between ALS CHM 95th percentiles and GEDI RH95

To leverage the GEDI model output, we made the following assumption: the GEDI model, on a 128×128 pixel sample, approximates the 95th percentile (p95) of the sample's ground truth canopy height map. This is justified by running simulations with the the GEDI simulator from Hancock et al. (2019) on the NEON ALS point clouds. We used simulated values rather than actual GEDI measurements because the GEDI measurements suffer from point errors, and because the simlator allows for denser sampling from with the limited geographic footprint of our ALS dataset.

The GEDI RH95 measurement used for training the GEDI model corresponds to the 95th percentile of the lidar's energy response. We simulated the GEDI RH95 values and find that they have excellent correlation ($R^2 = 0.88$) with the 95th percentile of the canopy height map around the corresponding GEDI footprints. This high correlation between GEDI RH95 and p95 of CHM was consistent across the diverse ecosystems covered in all 40 NEON sites in Appendix A.

B.3. Height prediction network training

The GEDI measurements were split into a 80/10/10% train/calibration/set-aside validation subsets. During training, the samples were drawn with a weight inversely proportional to the natural log of the total number of global samples in its RH95 bin, where each bin has a width of 1 m. We found that this sampling method provided a good number of training sample from higher canopy height locations while not overly biasing the model towards ecosystems with the highest canopy heights. Log inverse sample weighting is a less aggressive re-weighting that typical linear inverse weighting, as done in Lang et al. (2022a), which we choose so as not to overly bias the model towards the relatively few high canopy height samples.

After the convolutional layers, we also input a collection of scalar values, designated as "Satellite Metadata" in Fig. 2. This metadata included: the latitude, longitude position of each image, the off-nadir view angle of the satellite, the angle between zenith and sun position at capture, and the terrain slope (Mapzen, 2017) of the image footprint. Measured terrain slope is used during training, but set to zero height during forward inference, which allows the model to reduce the systematic error resulting from the bias of GEDI measurements towards higher canopy heights when the beam width straddles large surface gradients (see Section 4.1.3, Appendix B.4).

When training the GEDI model, we only used random 90 degree rotations and random horizontal and vertical flips, since the larger volume of data made augmentation less helpful.

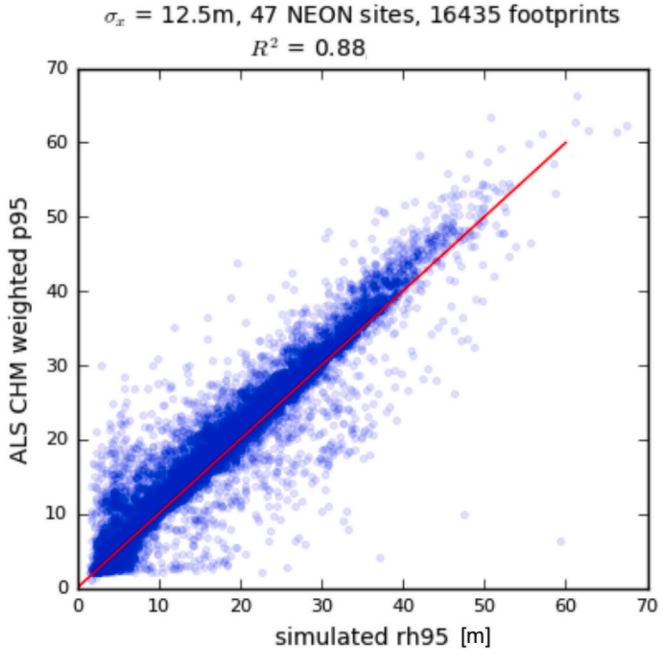


Fig. B.19. Correlation between 95th percentiles of ALS Canopy Height Maps and simulated GEDI RH95 values from the same maps. The 95th percentile is computed within weighted Gaussians with $\sigma = 12.5m$, in order to roughly approximate the GEDI beam width.

B.4. GEDI height and terrain slope correlation

As has been noted in Adam et al. (2020), the GEDI instruments estimate of canopy height is influenced by the terrain slope. We found evidence of this correlation in the data, and due to this have chosen to set the terrain slope to zero during inference to mitigate this systematic.

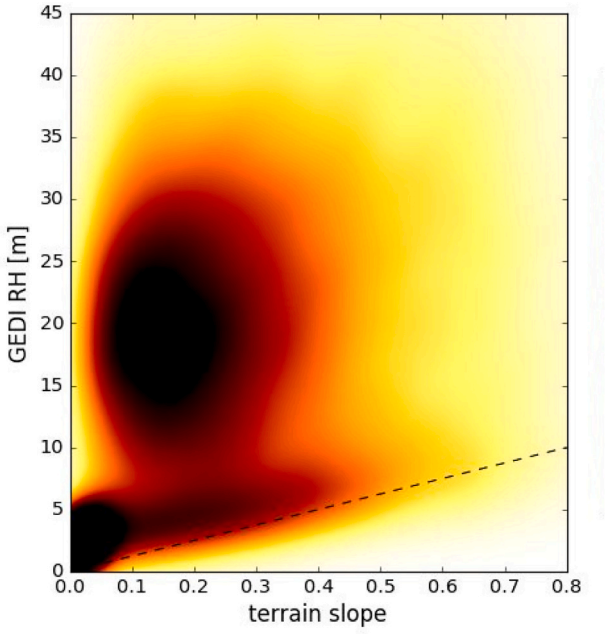


Fig. B.20. Correlation between terrain slope and GEDI RH95 for samples in CA. The dashed line indicates the height of the terrain with the GEDI beam (GEDI beam radius times the terrain slope). The heatmap is predominately above this line, indicating that there are no GEDI height estimates which fall below the terrain change within the beam.

Appendix C. Details on different metrics

C.1. Block R^2

To compute the block R^2 score, we split the ground truth CHM c and the prediction $\widehat{c}$ into 50×50 pixels blocks and average their values, leading to a 5×5 array, reshaped into 1×25 vectors $c^{(b)}$ and $\widehat{c}^{(b)}$. Given the average ground truth CHM average of all $c^{(b)}$ in the test set, the classical R^2 score is then computed:

$$R^2_{block} = 1 - \frac{\sum \left(c_i^{(b)} - \widehat{c}_i^{(b)}\right)^2}{\sum \left(c_i^{(b)} - \overline{c}^{(b)}\right)^2}. \tag{C.1}$$

C.2. Mean error (ME)

We compute the mean error, also referred as bias, as

$$\mathrm{ME} = \frac{1}{|\mathscr{D}|} \sum_{i=1\ldots|\mathscr{D}|} \widehat{c}_i - c_i, \tag{C.2}$$

where $|\mathscr{D}|$ the number of pixels in the test set.

C.3. Edge error metric (EE)

We are interested in measuring the sharpness of the CHM while beeing close to the ground truth. Because a blurry prediction would lead to the same MAE, Block R^2 or PSNR than a sharp one, this metrics is not serving this purpose. Therefore, we established a metric comparing the image gradients of the maps, dubbed "edge error score", given by Algorithm 1. Fig. C.21 illustrates how this metric is computed in an example.

Algorithm 1. Edge error metric.

1. Edge detection
$E(\hat{c})$: Sobel detector on predicted CHM maps $\hat{c}$.
$E(c)$: Sobel detector on GT CHM maps c.
2. Compute normalization factor $d = (\sum_i |E(\hat{c}_i)|) + (\sum_i |E(c_i)|)$.
3. Edge error score
If $d > 0$
  score $= \frac{1}{d} \sum_i |E(\hat{c}_i) - E(c_i)|$.
Else
  score $= 0$.

Fig. C.21. Illustration of edge error metric for two results: the ResUNet (trained with an L1 loss) edge error score is 0.66 in this example, the score of the SSL model is 0.55, computed using difference of the prediction and ground truth edge maps appearing in the two images at the right.

Appendix D. Architecture and training details

Our code uses Pytorch 1.9.0 with Cuda 10.2.

SSL pretraining. We refer the reader to Oquab et al. (2023) for the SSL pretraining phase details. We only changed the image normalization parameters from ImageNet parameters to match the standard deviation and mean color intensities of our Satellite image dataset. The unsupervised pretraining of a large model took a little less than three days on two 8-GPUs Voltas. Instead of the standard ImageNet normalization parameters, we computed the mean and standard deviation on the dataset of 3.5 M images. The large encoder contains 303 million parameters, while the huge one has 606 million.

Decoder training. The training of CHM prediction from SSL features takes 8 h for a large model on 8 GPUs, and 9 h for a huge model. During this step, we kept the weights of the SSL encoder frozen and only train the DPT model. Our DPT decoder for the SSL model was trained for 140 k steps using a Cosine learning rate schedule (from 1×10^{-8} to 1×10^{-4}) with a linear warmup step for 12 k iterations. We used a batch size of 16. The decoder model contains 34.2 M of parameters.

Estimating the carbon footprint of model training. We estimate the carbon footprint of training the ViT huge model using the calculations from Oquab et al. (2023), a Thermal Design Power (TDP) of the V100-32G GPU equal to 250 W, a Power Usage Effectiveness (PUE) of 1.1, a carbon intensity factor of 0.385 kg CO_2 per KWh, a time of 11 days $\times$ 24 h $\times$ 64 GPUs = 16,896 GPU hours. The 4647 kWh used to train the model is approximately equivalent to a CO_2 footprint of 4647 $\times$ 0.385 = 1.8 T of CO_2. The training of the ResUNet baseline took 19 h on 8 V100-32G GPUs, or approximately 16.1 kg of CO_2. The training of the decoder model took 75 GPU hours, generating about 8 kg of CO_2. While the carbon footprint of the ViT huge model, 1.81 T of CO_2, was two orders of magnitude larger than the training of a ResUNet, the model training is a one-shot expense, and the inference time (and thus energy use and emissions) of the ViT huge and ResUNet were within the same order of magnitude.

Architecture details. Our encoder architecture is a ViT architecture as introduced by Dosovitskiy et al. (2021b). It treats an image as a set of patches, called tokens, that are first embedded into a feature space and then processed by a cascade of transformer layers to produce updated representations of the tokens. The transformer layers use multi-head attention and self attention as their fundamental operation. Multihead attention is an operation that relates each token to every token in the image and consequently, has a global receptive field. The ViT does not use down sampling operations in its intermediate stages and thus supports fine-grained feature maps also in the deeper layers of the network. For the huge model, the features consists in outputs from layers (9, 16, 22, 29). At each layer, a $8 \times 8 \times 1280$ feature map and $1 \times 1 \times 1280$ class output is extracted. In the DPT decoder, the set of tokens at various stages of the backbone is first reassembled into image like representations. Then, the decoder iteratively fuses the feature maps from different stages and produces the final dense prediction using an application specific output head. This step involves several residual convolution layers. The code of our backbone is available at https://github.com/facebookresearch/dinov2, with pre-training on natural images.

Appendix E. Alternate loss function ablation

We compare in Table E.7 results of models trained with L1 loss or Sigloss, and using different sizes of pretraining dataset: one with 3.5×10^6 images (referred to as "3.5 M") and one with 18×10^6 images ("18 M"). More pretraining data improves the performance of the SSL models. In terms of loss, we did not notice strong difference between L2 and sigloss, while the L1 results were slightly worse.

Table E.7
CHM prediction accuracy metrics with different loss functions. sl: sigloss. cl: using classification output. Linear: using a linear layer instead of DPT. We do not display CA Brande result to improve visibility but the results are included in the average.

	Neon test				São Paulo				Average			
	MAE	R^2	ME	EE	MAE	R^2	ME	EE	MAE	R^2	ME	EE
3.5 M sl	2.8	0.67	−1.2	0.51	6.0	0.14	−4.2	0.60	3.1	0.56	1.9	0.54
18 M sl	2.9	0.66	−1.3	0.52	4.9	0.46	−2.1	0.59	2.9	0.64	1.3	0.54
18 M linear sl	3.0	0.58	−1.8	0.68	7.1	−0.27	−6.7	0.71	3.6	0.41	2.8	0.67
18 M cl sl	2.6	0.71	−0.9	**0.48**	**4.9**	**0.47**	−1.9	**0.55**	**2.7**	0.70	1.0	**0.51**
18 M cl l1	**2.5**	0.80	**0.0**	0.51	5.2	0.39	−2.6	0.56	2.9	0.72	0.7	0.53
18 M cl l2	2.6	**0.86**	−0.1	0.52	5.1	0.43	**−1.4**	**0.55**	2.8	**0.75**	**0.5**	**0.51**

Appendix F. Residuals with respect to the GEDI RH95

Fig. F.22 displays the difference between the p95 of block model CHM predictions and the measured GEDI RH95 metrics w.r.t the GEDI RH95.

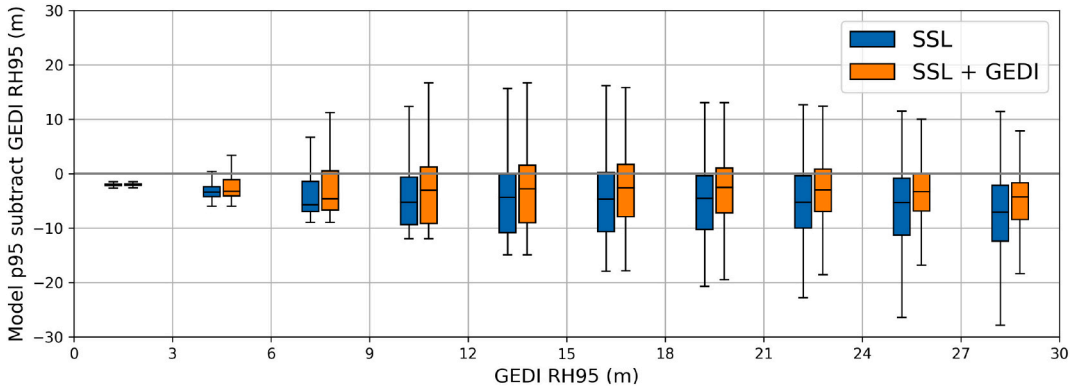


Fig. F.22. Residuals of p95 CHM predictions with GEDI RH95, with respect to the GEDI RH95.

Appendix G. Normalization for inference on aerial imagery

An image normalization step is necessary to improve the SSL inference performance on aerial images, when training only on Satellite imagery. We perform a classical histogram normalization of the aerial images (i.e. normalize the RGB channels of the aerial image to the p5-p95 distribution of the satellite image). This makes the color balance much more similar, leading to better performance for the SSL model. The satellite image is taken through much more atmosphere and we expect it to be less blue on average, because of preferential scattering of shorter wavelengths. Noting I the satellite image, A the original aerial image, we first compute for each color channel i and each image X the 5% percentile $p_5(X)_i$ and 95% percentile $p_{95}(X)_i$. Then, the normalized aerial image is given by

$$A_i = \left(A_i - p_5(A)_i\right) * \frac{p_{95}(I)_i - p_5(I)_i}{p_{95}(A)_i - p_5(A)_i} + p_5(I)_i.$$

We only apply this normalization to the SSL model trained on satellite imagery. Applying this normalization to the ResUNet model deteriorated the results.

References

Adam, M., Urbazaev, M., Dubois, C., Schmullius, C., 2020. Accuracy assessment of gedi terrain elevation and canopy height estimates in european temperate forests: influence of environmental and acquisition parameters. Remote Sens. 12 https://doi.org/10.3390/rs12233948.

Astola, H., Seitsonen, L., Halme, E., Molinier, M., Lönnqvist, A., 2021. Deep neural networks with transfer learning for forest variable estimation using sentinel-2 imagery in boreal forest. Remote Sens. 13 https://doi.org/10.3390/rs13122392.

Azevedo, T., Souza, C., Zanin Shimbo, J., Alencar, A., 2018. Mapbiomas Initiative: Mapping Annual Land Cover and Land Use Changes in Brazil from 1985 to 2017.

Bhat, S.F., Alhashim, I., Wonka, P., 2021. Adabins: depth estimation using adaptive bins. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4009–4018.

Brande, K., 2021. 3d fuel structure in relation to prescribed fire, ca 2020. National center for airborne laser mapping (ncalm). Distributed by opentopography. URL:. https://doi.org/10.5069/G9C53J18. accessed: 2023-02-15.

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.

Camarretta, N., Harrison, P.A., Bailey, T., Potts, B., Lucieer, A., Davidson, N., Hunt, M., 2020. Monitoring forest structure to guide adaptive management of forest restoration: a review of remote sensing approaches. New For. 51, 573–596. https://doi.org/10.1007/s11056-019-09754-5.

Cook-Patton, S.C., Leavitt, S.M., Gibbs, D., Harris, N.L., Lister, K., Anderson-Teixeira, K. J., Briggs, R.D., Chazdon, R.L., Crowther, T.W., Ellis, P.W., Griscom, H.P., Herrmann, V., Holl, K.D., Houghton, R.A., Larrosa, C., Lomax, G., Lucas, R., Madsen, P., Malhi, Y., Paquette, A., Parker, J.D., Paul, K., Routh, D., Roxburgh, S., Saatchi, S., van den Hoogen, J., Walker, W.S., Wheeler, C.E., Wood, S.A., Xu, L., Griscom, B.W., 2020. Mapping carbon accumulation potential from global natural forest regrowth. Nature 585, 545–550. https://doi.org/10.1038/s41586-020-2686-x.

Csillik, O., Kumar, P., Mascaro, J., O'Shea, T., Asner, G.P., 2019. Monitoring tropical forest carbon stocks and emissions using planet satellite data. Sci. Rep. 9, 17831. https://doi.org/10.1038/s41598-019-54386-6.

Cuni-Sanchez, A., Sullivan, M.J.P., Platts, P., et al., 2021. High aboveground carbon stock of African tropical montane forests. Nature 596, 536–542. https://doi.org/10.1038/s41586-021-03728-4.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021a. An image is worth 16x16 words: transformers for image recognition at scale. In: 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. https://doi.org/10.48550/ARXIV.2010.11929.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021b. An image is worth 16x16 words: transformers for image recognition at scale arXiv:2010.11929.

Dos-Santos, M., Keller, M., Morton, D., 2019. Lidar Surveys over Selected Forest Research Sites, Brazilian Amazon, 2008–2018. ORNL DAAC, Oak Ridge, Tennessee, USA. URL: https://daac.ornl.gov/CMS/guides/LiDAR_Forest_Inventory_Brazil.html.

Dubayah, R., Blair, J.B., Goetz, S., Fatoyinbo, L., Hansen, M., Healey, S., Hofton, M., Hurtt, G., Kellner, J., Luthcke, S., Armston, J., Tang, H., Duncanson, L., Hancock, S., Jantz, P., Marselis, S., Patterson, P.L., Qi, W., Silva, C., 2020. The global ecosystem dynamics investigation: high-resolution laser ranging of the earth's forests and topography. Sci. Remote Sens. 1, 100002. URL: https://www.sciencedirect.com/science/article/pii/S2666017220300018 https://doi.org/10.1016/j.srs.2020.100002.

Dubayah, R., Luthcke, S., Sabaka, T., Nicholas, J., Preaux, S., Hofton, M.. Gedi l3 Gridded Land Surface Metrics, Version 1. URL. https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1865.

Dubayah, R., Armston, J., Kellner, J., Duncanson, L., Healey, S., Patterson, P., Hancock, S., Tang, H., Bruening, J., Hofton, M., Blair, J., Luthcke, S., . GEDI L4A footprint level aboveground biomass density, version 2.1. URL: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=2056 https://doi.org/10.3334/ORNLDAAC/2056.

Duncanson, L., Neuenschwander, A., Hancock, S., Thomas, N., Fatoyinbo, T., Simard, M., Silva, C.A., Armston, J., Luthcke, S.B., Hofton, M., Kellner, J.R., Dubayah, R., 2020. Biomass estimation from simulated GEDI, ICESat-2 and NISAR across environmental gradients in Sonoma County, California. Remote Sens. Environ. 242, 111779. URL: https://www.sciencedirect.com/science/article/pii/S0034425720301498 https://doi.org/10.1016/j.rse.2020.111779.

Eigen, D., Puhrsch, C., Fergus, R., 2014. Depth map prediction from a single image using a multi-scale deep network. Adv. Neural Inf. Proces. Syst. 27.

Fayad, I., Ciais, P., Schwartz, M., Wigneron, J.P., Baghdadi, N., de Truchis, A., d'Aspremont, A., Frappart, F., Saatchi, S., Pellissier-Tanon, A., Bazzi, H., 2023. Vision transformers, a new approach for high-resolution and large-scale mapping of canopy heights arXiv:2304.11487.

Friedlingstein, P., Jones, M.W., O'Sullivan, M., Andrew, R.M., Hauck, J., Peters, G.P., Peters, W., Pongratz, J., Sitch, S., Le Quéré, C., Bakker, D.C.E., Canadell, J.G., Ciais, P., Jackson, R.B., Anthoni, P., Barbero, L., Bastos, A., Bastrikov, V., Becker, M., Bopp, L., Buitenhuis, E., Chandra, N., Chevallier, F., Chini, L.P., Currie, K.I., Feely, R. A., Gehlen, M., Gilfillan, D., Gkritzalis, T., Goll, D.S., Gruber, N., Gutekunst, S., Harris, I., Haverd, V., Houghton, R.A., Hurtt, G., Ilyina, T., Jain, A.K., Joetzjer, E., Kaplan, J.O., Kato, E., Klein Goldewijk, K., Korsbakken, J.I., Landschützer, P.,

Lauvset, S.K., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Marland, G., McGuire, P.C., Melton, J.R., Metzl, N., Munro, D.R., Nabel, J.E.M.S., Nakaoka, S.I., Neill, C., Omar, A.M., Ono, T., Peregon, A., Pierrot, D., Poulter, B., Rehder, G., Resplandy, L., Robertson, E., Rödenbeck, C., Séférian, R., Schwinger, J., Smith, N., Tans, P.P., Tian, H., Tilbrook, B., Tubiello, F.N., van der Werf, G.R., Wiltshire, A.J., Zaehle, S., 2019. Global carbon budget 2019. In: Earth System Science Data, 11, pp. 1783–1838. URL. https://essd.copernicus.org/articles/11/1783/2019/. https://doi.org/10.5194/essd-11-1783-2019.

Fu, H., Gong, M., Wang, C., Tao, D., 2018. A compromise principle in deep monocular depth estimation. arXiv:1708.08267.

Gibril, M.B.A., Shafri, H.Z.M., Al-Ruzouq, R., Shanableh, A., Nahas, F., Al Mansoori, S., 2023. Large-scale date palm tree segmentation from multiscale uav-based and aerial images using deep vision transformers. Drones 7. https://doi.org/10.3390/drones7020093.

Hancock, S., Armston, J., Hofton, M., Sun, X., Tang, H., Duncanson, L.I., Kellner, J.R., Dubayah, R., 2019. The GEDI simulator: a large-footprint waveform lidar simulator for calibration and validation of spaceborne missions. Earth Space Sci. 6, 294–310. https://doi.org/10.1029/2018EA000506.

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S.V., Goetz, S.J., Loveland, T.R., Kommareddy, A., Egorov, A., Chini, L., Justice, C.O., Townshend, J.R.G., 2013. High-resolution global maps of 21st-century forest cover change. Science 342, 850–853. https://doi.org/10.1126/science.1244693.

Hansen, M.C., Krylov, A., Tyukavina, A., Potapov, P.V., Turubanova, S., Zutta, B., Ifo, S., Margono, B., Stolle, F., Moore, R., 2016. Humid tropical forest disturbance alerts using landsat data. Environ. Res. Lett. 11, 034008. https://doi.org/10.1088/1748-9326/11/3/034008.

Harris, N.L., Gibbs, D.A., Baccini, A., Birdsey, R.A., de Bruin, S., Farina, M., Fatoyinbo, L., Hansen, M.C., Herold, M., Houghton, R.A., Potapov, P.V., Suarez, D.R., Roman-Cuesta, R.M., Saatchi, S.S., Slay, C.M., Turubanova, S.A., Tyukavina, A., 2021. Global maps of twenty-first century forest carbon fluxes. Nature. Climate Change 11, 234–240. https://doi.org/10.1038/s41558-020-00976-6.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009.

Khosravipour, A., Skidmore, A.K., Isenburg, M., Wang, T., Hussin, Y.A., 2014. Generating pit-free canopy height models from airborne lidar. Photogramm. Eng. Remote. Sens. 80, 863–872. https://doi.org/10.14358/PERS.80.9.863.

Lang, N., Jetz, W., Schindler, K., Wegner, J.D., 2022a. A high-resolution canopy height model of the earth. https://doi.org/10.48550/ARXIV.2204.08322.

Lang, N., Kalischek, N., Armston, J., Schindler, K., Dubayah, R., Wegner, J.D., 2022b. Global canopy height regression and uncertainty estimation from GEDI LIDAR waveforms with deep ensembles. Remote Sens. Environ. 268, 112760 https://doi.org/10.1016/j.rse.2021.112760.

Li, B., Dai, Y., He, M., 2018. Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference. Pattern Recogn. 83, 328–339. https://doi.org/10.1016/j.patcog.2018.05.029.

Li, W., Niu, Z., Shang, R., Qin, Y., Wang, L., Chen, H., 2020. High-resolution mapping of forest canopy height using machine learning by coupling icesat-2 lidar with sentinel-1, sentinel-2 and landsat-8 data. Int. J. Appl. Earth Obs. Geoinf. 92, 102163 https://doi.org/10.1016/j.jag.2020.102163.

Liu, S., Brandt, M., Nord-Larsen, T., Chave, J., Reiner, F., Lang, N., Tong, X., Ciais, P., Igel, C., 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. https://doi.org/10.21203/rs.3.rs-2573442/v1.

Luo, W., Li, Y., Urtasun, R., Zemel, R., 2016. Understanding the effective receptive field in deep convolutional neural networks. Adv. Neural Inf. Proces. Syst. 29 https://doi.org/10.48550/ARXIV.1701.04128.

Luyssaert, S., Schulze, E.D., Börner, A., Knohl, A., Hessenmöller, D., Law, B.E., Ciais, P., Grace, J., 2008. Old-growth forests as global carbon sinks. Nature 455, 213–215. https://doi.org/10.1038/nature07276.

da Luz, N.B., Garrastazu, M.C., Rosot, M.A.D., Maran, J.C., de Oliveira, Y.M.M., Franciscon, L., Cardoso, D.J., de Freitas, J.V., 2018. Inventário florestal nacional do brasil - uma abordagem em escala de paisagem para monitorar e avaliar paisagens florestais. Pesquisa Florestal Bras. 38 https://doi.org/10.4336/2018.pfb.38e201701493.

Maioli, V., Belharte, S., Stuker Kropf, M., Callado, C.H., 2020. Timber exploitation in colonial Brazil: a historical perspective of the Atlantic forest. Hist. Ambient. Latinoamericana Caribeña (HALAC) Rev. Solcha 10, 46–73. https://doi.org/10.32991/2237-2717.2020v10i2.p74-101.

Mapzen, 2017. Amazon. Terrain Tiles on AWS. https://registry.opendata.aws/terrain-tiles.

Markus, T., Neumann, T., Martino, A., Abdalati, W., Brunt, K., Csatho, B., Farrell, S., Fricker, H., Gardner, A., Harding, D., Jasinski, M., Kwok, R., Magruder, L., Lubin, D., Luthcke, S., Morison, J., Nelson, R., Neuenschwander, A., Palm, S., Popescu, S., Shum, C., Schutz, B.E., Smith, B., Yang, Y., Zwally, J., 2017. The ice, cloud, and land elevation satellite-2 (icesat-2): science requirements, concept, and implementation. Remote Sens. Environ. 190, 260–273. https://doi.org/10.1016/j.rse.2016.12.029.

Maxwell, A.E., Warner, T.A., Guillén, L.A., 2021. Accuracy assessment in convolutional neural network-based deep learning remote sensing studies—part 2: recommendations and best practices. Remote Sens. 13 https://doi.org/10.3390/rs13132591.

Miangoleh, S.M.H., Dille, S., Mai, L., Paris, S., Aksoy, Y., 2021. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9685–9694.

Mugabowindekwe, M., Brandt, M., Chave, J., Reiner, F., Skole, D.L., Kariryaa, A., Igel, C., Hiernaux, P., Ciais, P., Mertz, O., et al., 2022. Nation-wide mapping of tree-level aboveground carbon stocks in rwanda. Nat. Clim. Chang. 1–7.

National Ecological Observatory Network (NEON), 2022. Ecosystem Structure (dp3.30015.001). URL: https://data.neonscience.org/data-products/DP3.30015.001.

Olofsson, P., Foody, G.M., Herold, M., Stehman, S.V., Woodcock, C.E., Wulder, M.A., 2014. Good practices for estimating area and assessing accuracy of land change. Remote Sens. Environ. 148, 42–57. https://doi.org/10.1016/j.rse.2014.02.015.

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., 2023. Dinov2: learning robust visual features without supervision arXiv:2304.07193.

Popkin, G., 2015. The hunt for the world's missing carbon. Nature 523, 20–22. https://doi.org/10.1038/523020a.

Potapov, P., Li, X., Hernandez-Serna, A., Tyukavina, A., Hansen, M.C., Kommareddy, A., Pickens, A., Turubanova, S., Tang, H., Silva, C.E., Armston, J., Dubayah, R., Blair, J. B., Hofton, M., 2021. Mapping global forest canopy height through integration of GEDI and Landsat data. Remote Sens. Environ. 253, 112165 https://doi.org/10.1016/j.rse.2020.112165.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. In: International Conference on Computer Vision.

Reed, C.J., Gupta, R., Li, S., Brockman, S., Funk, C., Clipp, B., Candido, S., Uyttendaele, M., Darrell, T., 2022. Scale-MAE: A scale-aware masked autoencoder for multiscale geospatial representation learning arXiv preprint arXiv:2212.14532.

Reytar, K., Buckingham, K., Stolle, F., Brandt, J., Zamora-Cristales, R., Landsberg, F., Singh, R., Streck, C., Saint-Laurent, C., Tucker, C., Henry, M., Walji, K., Finegold, Y., Aga, Y., Rezende, M., 2020. Measuring progress in forest and landscape restoration. Unasylva 71, 62.

Ribeiro, M.C., Martensen, A.C., Metzger, J.P., Tabarelli, M., Scarano, F., Fortin, M.J., 2011. The Brazilian Atlantic Forest: A Shrinking Biodiversity Hotspot. Springer, Berlin Heidelberg, Berlin, Heidelberg, pp. 405–434. https://doi.org/10.1007/978-3-642-20992-5_21.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer, pp. 234–241.

Schacher, A., Roger, E., Williams, K.J., Stenson, M.P., Sparrow, B., Lacey, J., 2023. Use-specific considerations for optimising data quality trade-offs in citizen science: recommendations from a targeted literature review to improve the usability and utility for the calibration and validation of remotely sensed products. Remote Sens. 15 https://doi.org/10.3390/rs15051407.

Schwartz, M., Ciais, P., Ottlé, C., De Truchis, A., Vega, C., Fayad, I., Brandt, M., Fensholt, R., Baghdadi, N., Morneau, F., Morin, D., Guyon, D., Dayau, S., Wigneron, J.P., 2022. High-resolution canopy height map in the Landes forest (France) based on GEDI, Sentinel-1, and Sentinel-2 data with a deep learning approach. arXiv:2212.10265.

Silva, C.A., Duncanson, L., Hancock, S., Neuenschwander, A., Thomas, N., Hofton, M., Fatoyinbo, L., Simard, M., Marshak, C.Z., Armston, J., Lutchke, S., Dubayah, R., 2021. Fusing simulated GEDI, ICESat-2 and NISAR data for regional aboveground biomass mapping. Remote Sens. Environ. 253, 112234 https://doi.org/10.1016/j.rse.2020.112234.

Singh, M., Gustafson, L., Adcock, A., Reis, V.D.F., Gedik, B., Kosaraju, R.P., Mahajan, D., Girshick, R., Dollár, P., van der Maaten, L., 2022. Revisiting Weakly Supervised Pre-Training of Visual Perception Models. CVPR.

Sirko, W., Kashubin, S., Ritter, M., Annkah, A., Bouchareb, Y.S.E., Dauphin, Y.N., Keysers, D., Neumann, M., Cissé, M., Quinn, J., 2021. Continental-scale building detection from high resolution satellite imagery. CoRR abs/2107.12283. URL: https://arxiv.org/abs/2107.12283.

Skole, D.L., Samek, J.H., Dieng, M., Mbow, C., 2021. The contribution of trees outside of forests to landscape carbon and climate change mitigation in West Africa. Forests 12. https://doi.org/10.3390/f12121652.

Stehman, S.V., 2014. Estimating area and map accuracy for stratified random sampling when the strata are different from the map classes. Int. J. Remote Sens. 35, 4923–4939. https://doi.org/10.1080/01431161.2014.930207.

Stephenson, N.L., Das, A.J., Condit, R., Russo, S.E., Baker, P.J., Beckman, N.G., Coomes, D.A., Lines, E.R., Morris, W.K., Rüger, N., Álvarez, E., Blundo, C., Bunyavejchewin, S., Chuyong, G., Davies, S.J., Duque, Á., Ewango, C.N., Flores, O., Franklin, J.F., Grau, H.R., Hao, Z., Harmon, M.E., Hubbell, S.P., Kenfack, D., Lin, Y., Makana, J.R., Malizia, A., Malizia, L.R., Pabst, R.J., Pongpattananurak, N., Su, S.H., Sun, I.F., Tan, S., Thomas, D., van Mantgem, P.J., Wang, X., Wiser, S.K., Zavala, M. A., 2014. Rate of tree carbon accumulation increases continuously with tree size. Nature 507, 90–93. https://doi.org/10.1038/nature12914.

Tesfay, F., Moges, Y., Asfaw, Z., 2022. Woody species composition, structure, and carbon stock of coffee-based agroforestry system along an elevation gradient in the moist mid-highlands of southern Ethiopia. Int. J. Forest. Res. 2022, 1–12. https://doi.org/10.1155/2022/4729336.

Vallauri, D., Aronson, J., Dudley, N., Vallejo, R., 2005. Monitoring and Evaluating Forest Restoration Success. Springer, New York, New York, NY, pp. 150–158. https://doi.org/10.1007/0-387-29112-1_21.

Viani, R.A.G., Barreto, T.E., Farah, F.T., Rodrigues, R.R., Brancalion, P.H.S., 2018. Monitoring young tropical forest restoration sites: how much to measure? Trop. Conserv. Sci. 11 https://doi.org/10.1177/1940082918780916, 1940082918780916.

Wagner, F.H., Roberts, S., Ritz, A.L., Carter, G., Dalagnol, R., Favrichon, S., Hirye, M.C., Brandt, M., Ciais, P., Saatchi, S., 2023. Sub-meter tree height mapping of California using aerial images and lidar-informed u-net model arXiv:2306.01936.

Wang, W., Tang, C., Wang, X., Zheng, B., 2022. A ViT-based multiscale feature fusion approach for remote sensing image segmentation. IEEE Geosci. Remote Sens. Lett. 19, 1–5. https://doi.org/10.1109/LGRS.2022.3187135.

Weinstein, B.G., Graves, S.J., Marconi, S., Singh, A., Zare, A., Stewart, D., Bohlman, S.A., White, E.P., 2021. A benchmark dataset for canopy crown detection and delineation in co-registered airborne RGB, LiDAR and hyperspectral imagery from the National Ecological Observation Network. PLoS Comput. Biol. 17 (7), e1009180.

Xu, Z., Zhang, W., Zhang, T., Yang, Z., Li, J., 2021. Efficient transformer for remote sensing image segmentation. Remote Sens. 13 https://doi.org/10.3390/rs13183585.

Yanai, R.D., Wayson, C., Lee, D., Espejo, A.B., Campbell, J.L., Green, M.B., Zukswert, J. M., Yoffe, S.B., Aukema, J.E., Lister, A.J., Kirchner, J.W., Gamarra, J.G.P., 2020. Improving uncertainty in forest carbon accounting for redd+ mitigation efforts. Environ. Res. Lett. 15, 124002 https://doi.org/10.1088/1748-9326/abb96f.

Zhang, Z., Liu, Q., Wang, Y., 2017. Road extraction by deep residual U-net. CoRR abs/1711.10684. URL: http://arxiv.org/abs/1711.10684.