Open-Canopy: Towards Very High Resolution Forest Monitoring

Supplementary Material

We present additional results, analyses, and experiments to support our study. First, we detail our validation of the ground truth using terrain measurements and manual verifications in Sec. A. Next, we provide further results in Sec. B, including new analyses, qualitative illustrations, and experimental settings. We then conduct a detailed ablation study in Sec. C to examine the influence of key hyperparameters and design choices. Additionally, we offer a comprehensive description of the dataset and its construction in Sec. D. Finally, we provide the Datasheet for Dataset [1] for our benchmark.

A. Validation with Field Measurements

Ensuring the accuracy of our ground truth data is crucial for the validity of any computer vision benchmark. While the ALS data from LIDAR-HD have been calibrated and validated internally by the French Mapping Agency (IGN) using plots annotated by the National Forest Office (ONF), we performed additional manual verifications to further confirm their reliability, performed at plot-level and tree-level.

Focused Plot-Level Assessment. We sourced measurements from 135 plots, each with a 15 m radius, across the Vosges region. These plots were measured in the field by forestry experts from the National Forest Inventory (INF) within two years of the ALS acquisition. For each plot, we compared the height of the tallest tree measured in situ with the maximum canopy height in the plot as estimated by the ALS data and predicted by our best-performing computer vision model (PVTv2). As shown in Fig. A and detailed in Fig. A, the ALS-derived heights exhibit smaller errors compared to our model's estimates and align closely with the field measurements. This validation confirms the suitability of the ALS data as ground truth for our open-access benchmark.

Country-Scale Plot Assessment. To evaluate performance at a national scale, we extended our assessment to 5,323 plots spanning the entire French metropolitan territory. Because this coverage exceeds the current ALS coverage of France, we only compare the best model's predictions against field measurements. As illustrated in Fig. B, the strong correlation between predictions and measurements confirms the model's accuracy with an alternative reference. We also report metrics comparing the algorithm's outputs against both ALS ground truth and manual field measurements. The agreement between these two ground truths further validates our experimental evaluation protocol.

Tree-Level Assessment. We extended our validation to the individual tree level using data provided by the ONF, consisting of 44 geolocated trees in the Grand Est region. For each tree, we compared the measured height with the highest estimated or predicted height within a 1.5 m radius around the tree's center. The metrics presented in Tab. A corroborate the plot-level findings, further validating the ALS-derived heights. This validation process emphasizes the reliability of our ground truth data, which is essential for advancing computer vision methods in canopy height estimation.

Change Dataset Curation. To ensure the quality and accuracy of the Open-Canopy-Δ benchmark, we conducted a thorough manual validation of the dieback areas constituting its ground truth. As detailed in Section 4.1, each of the 73 change areas was carefully examined and validated by a forest expert. This meticulous process guarantees the reliability of the dataset for challenging computer vision tasks involving canopy height change detection. An example of visual annotation from this validation process is shown in Fig. D. Some false positives were identified, likely due to selective logging activities occurring between the ALS and SPOT acquisitions within the same year.

B. Additional Results

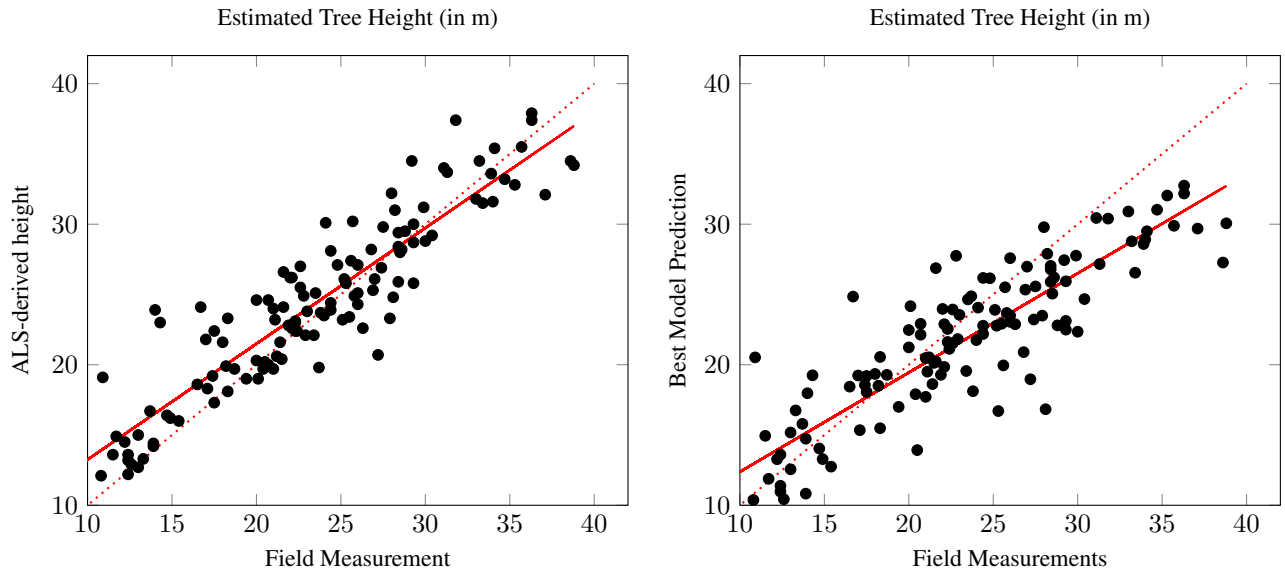
We present several additional analyses of the performance of our models. First, we provide additional qualitative illustrations in Sec. B.1. Then, we offer a detailed analysis of how tree height influences the quality of the results Sec. B.2. Finally, we re-evaluate our models and other products at different resolutions (Sec. B.3), providing a fair comparison in settings more advantageous to coarser predictions.

B.1. Qualitative Illustrations

We provide here additional illustrations for qualitative assessment.

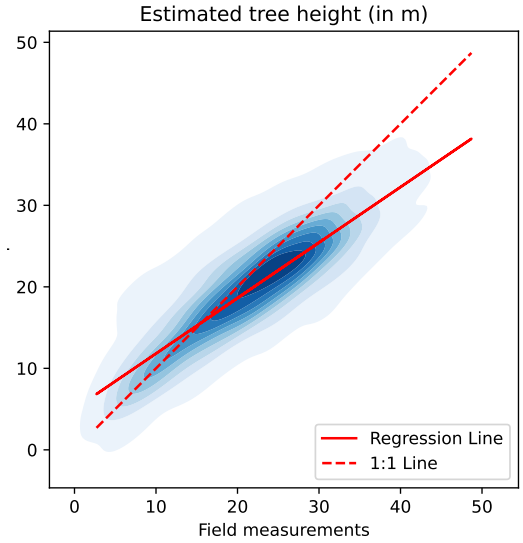
Canopy Height Fig. C showcases a comparison between the ALS-derived canopy height map and the height map predicted by our model using SPOT images. Our model demonstrates the ability to accurately estimate vegetation height across a variety of challenging scenarios:

- **Mountainous Areas** (first row): Capturing complex terrain and varied vegetation.
- **Agricultural Lands** (second row): Detecting small hedges and understory vegetation.



	MAE (m)	nMAE (%)	RMSE (m)	Bias (m)
ALS vs Field measurements	2.3	13.5	3.1	1.2
PVTv2 vs Field measurements	2.9	14.4	3.7	-1.1

Figure A. Plot-Level Quantitative Evaluation. We compare the plot-wise maximum heights as measured through ALS or predicted by a PVTv2 model against the manual field measurements.



	MAE (m)	nMAE (%)	RMSE (m)	Bias (m)
PVTv2 vs ALS	2.52	22.9	4.0	0
PVTv2 vs Field measurements	3.6	16.9	4.7	-1.9

Figure B. Country-Scale Plot-Level Evaluation. Left, we compare the plot-wise performance of our best model compared to field measurement maximum heights. Right, we compare the precision of the model when measured against canopy height derived from ALS or field measurements. This study is performed for 5323 plots spread across France.

- **Dense Forests** (rows 3 and 4): Handling thick canopy cover and shadowed regions.
- **Urban Environments** (row 5): Distinguishing trees amidst buildings and infrastructure.
- **Mixed Scenes** (rows 6 and 7): Managing heterogeneous landscapes with multiple land cover types.

The high spatial resolution of our predictions not only captures fine-grained details but also enables the identification

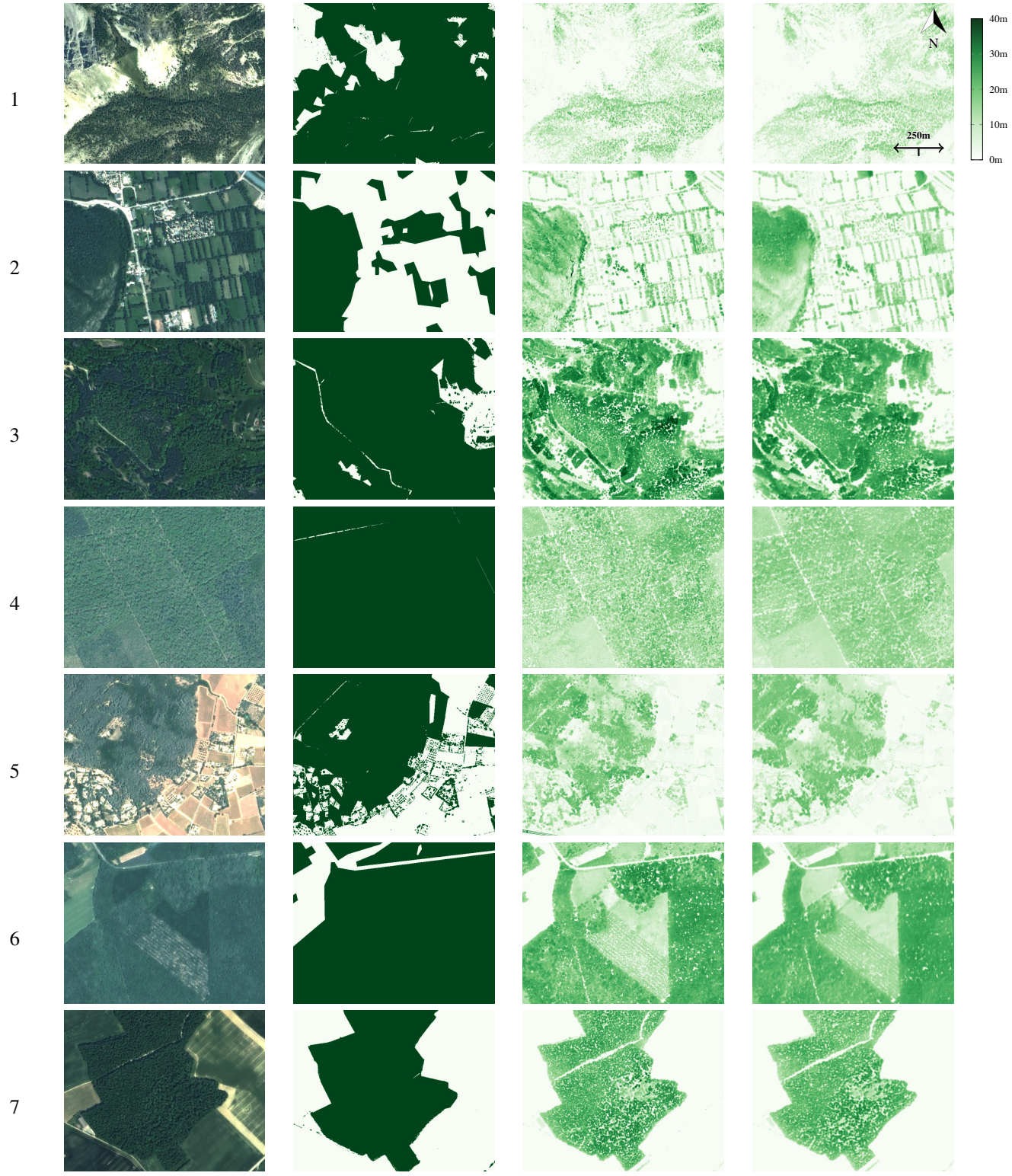
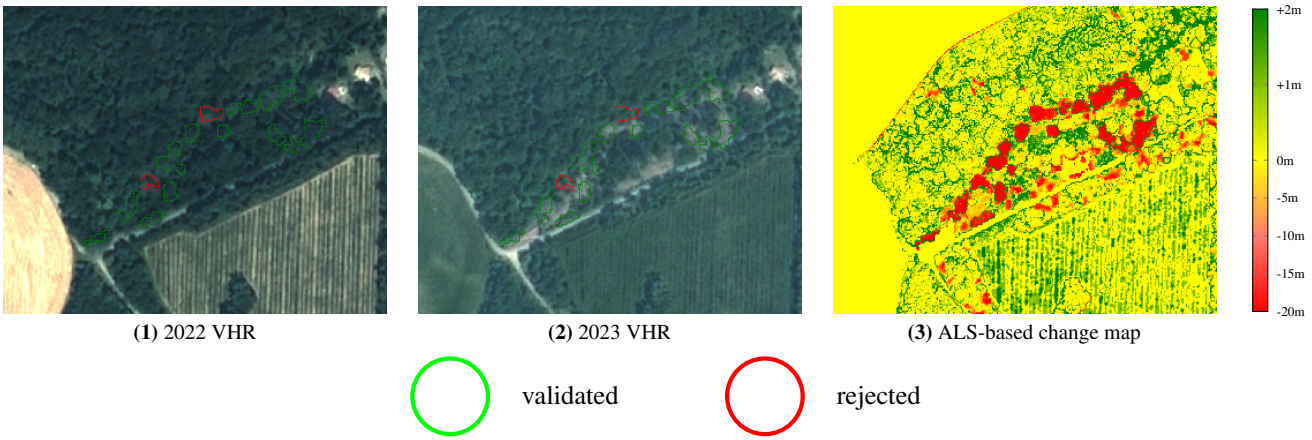


Figure C. Canopy Height Estimation Illustrations. We select seven areas of interest and represent the available VHR image ((1)), the vegetation mask used for evaluation ((2)), the ground truth ALS-derived height map ((3)), and the height map estimated with PVTv2 model from the VHR image ((4)). Scale and orientation are shared across all subfigures.

Table A. Tree-Level Quantitative Evaluation. We compare the precision of ALS and a PVTv2 model when taking the field measurements as ground truth.

	MAE (m)	nMAE (%)	RMSE (m)	Bias (m)
ALS vs Field measurements	1.45	6.7	2.0	0.22
PVTv2 vs Field measurements	4.0	15.4	5.1	-3.2



(1) 2022 VHR **(2)** 2023 VHR **(3)** ALS-based change map

Figure D. Visual Validation of Change Components: Example of a pair of successive VHR images and the corresponding change maps (derived from differences in ALS-based canopy height). We highlight the contours of the change masks validated by forestry experts through visual inspection.

of man-made features such as forest paths, which are crucial for forest management applications.

We further compare the performance of three models in-Fig. E: a standard Vision Transformer (ViT) and two hierarchical models, PVTv2 and SWIN. The hierarchical models exhibit significantly lower errors, which corroborates our quantiative results.

Canopy Height Change We provide additional illustrations of height change detection in Fig. F. While our model tends to over predict small growth or loss of canopy height, the areas of strong disturbances—as denoted by our smoothed and filtered binary change maps—are overall well detected and delineated. Our illustration covers areas of dense forests (first row) and mixed scenes (row 2 and 3). Our method can detect disturbances such as clear and selective cuts.

Note that the Sentinel-derived height maps for 2022 and 2023 were provided by the authors of [2], as only the map for 2020 is available online.

B.2. Influence of Tree Height

We analyzed the performance of our canopy height estimation model across different ranges of true tree heights to understand how tree height influences prediction accuracy. The results are summarized in Tab. B.

- Note that the nMAE (normalized Mean Absolute Error) is computed for all ranges as the average of the pixel-wise normalized absolute error:

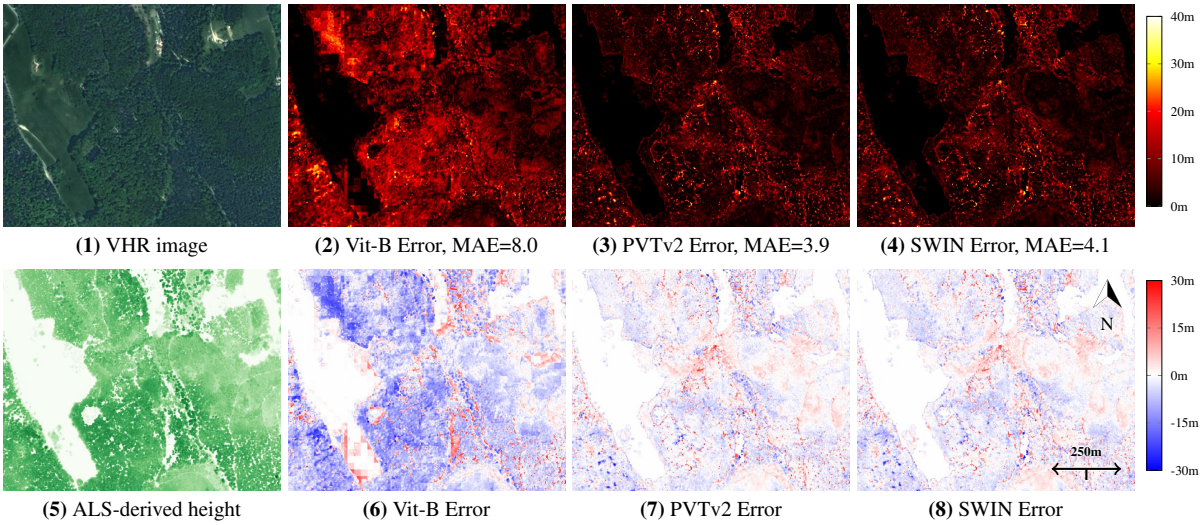
$$\text{nMAE} = \frac{|(z_{\text{true}} - z_{\text{pred}})|}{1 + z_{\text{true}}}, \tag{A}$$

  where z_{true} and z_{pred} are respectively the ALS-derived and predicted height for a given pixel. The additional 1 term in the denominator makes this measure more robust for pixels corresponding to low vegetation.
- When computing the nMAE for the overall range of 0–60 m, we exclude the 0–2 m bin. This exclusion is necessary because values in this range can produce disproportionately large errors due to the normalization, which can dominate the metric and skew the results. Additionally, including this bin may unfairly disadvantage models with lower spatial resolutions that aim to predict the highest value within larger pixels, potentially overlapping with bare soil at higher resolutions.

As shown in Tab. B by the bias of our model for different ranges, our model tends to over-predict the height of small trees and under-predict the height of tall trees. While the average error is higher for larger trees, our model has the lowest nMAE for the 20-30m range, with a value of 12.1%.

B.3. Evaluation at a resolution of 10m

To provide a fair comparison with models predicting canopy height at a 10 m resolution, we resampled both our ground truth and predicted height maps to a 10 m grid and re-evaluated all available models. We performed this by aggre-



(1) VHR image **(2)** Vit-B Error, MAE=8.0 **(3)** PVTv2 Error, MAE=3.9 **(4)** SWIN Error, MAE=4.1

(5) ALS-derived height **(6)** Vit-B Error **(7)** PVTv2 Error **(8)** SWIN Error

Figure E. Difference Maps: Per-pixel absolute (top row) and relative (bottom row) errors for three models: ViT-B, PVTv2, and SWIN. While the differences between PVTv2 and SWIN are subtle (approximately 20cm on average), the advantage of these models over ViT-B is visible.

Table B. Canopy Height Prediction Per Height Bins. We report the metrics for different bins of true tree height for the PVTv2[3] model.

Range in m	0-2	2-5	5-10	10-15	15-20	20-30	30-60	**0-60**
MAE in m	1.67	2.29	2.65	2.70	2.61	3.00	5.52	**2.52**
nMAE in %	138.8	53.6	32.1	20.3	14.3	12.1	16.0	**22.9**
RMSE in m	4.31	3.67	3.69	3.60	3.53	4.19	7.56	**4.02**
Bias in m	1.49	0.87	0.65	0.21	-0.42	-1.90	-5.31	**0.00**
Tree cov. IoU (%)	-	72.6	96.5	99.3	99.7	99.8	99.6	**90.5**

gating the higher-resolution data as follows:
For each 10 m pixel, we took the maximum value from the overlapping 1.5 m pixels. This approach is equivalent to rasterizing the full ALS 3D point cloud directly onto a 10 m grid. Taking the maximum value aligns with models trained to predict metrics like GEDI RH100 or RH95 (relative height at the 100th or 95th percentile), which represent the tallest canopy elements within a pixel.
We report the results in Tab. C, and observe a similar ordering than in Table 3 of the main paper. All methods see improved metrics as the problem is simpler, except for Tolan *et al*. In particular, the tree coverage problem becomes significantly easier at this resolution, with all 10 m-resolution methods nearing 90% IoU. Note that the height map of [8] at a resolution of 3m was provided directly by the authors and is not available online.

C. Ablation Study

We propose an analysis of the influence of several of our hyperparameters and design choices.

C.1. Parameters of the Change Detection

We evaluate how different configurations of the ground truth binary change map affect canopy height change detection. Specifically, we examine: (i) Minimum Height Difference: The threshold for considering a pixel as having a significant change in canopy height; (ii) Minimum Contiguous Change Area: The smallest area of connected changed pixels considered significant.
Tab. D presents the IoU metrics for various combinations of these parameters. Naturally, focusing on larger change areas simplifies the detection problem due to reduced complexity. The influence of the minimum tree height change threshold is less straightforward; higher thresholds require precise detection of significant height reductions, which can be more challenging. Our chosen parameters—15 m minimum height difference and 200 m^2 minimum change area—represent changes that are visually detectable between images (see Fig. F), providing a realistic yet challenging task for computer vision models. 01

C.2. Impact of Initialization Strategy

We provide in Tab. E the results of ablation experiments. We evaluate the impact of omitting the near infrared (NIR) band

Table C. Canopy Height Prediction at 10m resolution. We resample all ground truth and predicted maps on a 10 m grid.

Map	Backbone	Initial res. in m	MAE in m	nMAE in %	RMSE in m	Bias in m	Tree cov. IoU in %
Potapov [4]	UNet	30	6.17	44.6	8.33	-3.31	80.2
Schwartz [2, 5]	UNet	10	4.00	26.9	5.28	-1.38	90.1
Lang [6]	CNN	10	8.64	92.9	29.25	6.27	90.1
Pauls [7]	UNet	10	4.59	32.9	5.96	0.34	90.1
Liu [8]	UNet	3.0	4.58	37.4	10.97	-1.26	88.2
Tolan [9]	ViT-L	1.0	6.10	42.1	7.95	-5.37	81.6
Open-Canopy	UNet	1.5	2.72	19.0	3.95	-2.06	**93.4**
Open-Canopy	PVTv2	1.5	**2.42**	**17.6**	**3.57**	**-1.69**	93.3

Table D. Canopy Height Change Detection We compute the IoU metric (in %) for various minimum height difference (row, in m) and minimum contiguous area of change (column, in m^2). The values chosen in the benchmark are underlined.

min diff \ min surf	10 m^2	25 m^2	100 m^2	200 m^2	300 m^2	400 m^2
-5 m	7.0	7.1	7.2	6.2	5.2	4.2
-10 m	17.1	17.9	22.6	23.6	25.1	28.7
-15 m	22.1	23.4	28.8	37.0	40.6	40.8
-20 m	18.9	20.2	31.4	36.6	31.8	31.5

from input images. We can see in Tab. E that removing the NIR channel from input images decreases the performance for both UNet and PVTv2 backbones. Moreover, we assess various initialization strategies for fine-tuning networks initially trained only on RGB data to accommodate an additional NIR channel. Those include training from scratch, randomizing the first layer, and using LoRa. In Fig. J we show the results for different LoRa ranks and show only the best rank (32) in Tab. E. We see a clear benefit in using our proposed initialization scheme.

D. Dataset description

We describe here in details the dataset used in Open-Canopy and provide information about its constitution.

Table E. Ablation Study. We evaluate the impact of omitting the NIR channel from input images and assess various initialization strategies for fine-tuning networks initially trained only on RGB data to accommodate an additional NIR channel.

			MAE (m)	nMAE (%)	RMSE (m)	Bias (m)
Channels	backbone	pretraining				
RGB	UNet	ImageNet1K	2.77	24.8	4.34	-0.17
RGB+IR	UNet	ImageNet1K	2.67	23.8	4.18	-0.30
RGB	PVTv2	ImageNet1K	3.73	32.6	5.53	-0.50
RGB+IR	PVTv2	ImageNet1K	**2.52**	**22.9**	**4.02**	**0.00**
Initialization	backbone	pretraining				
Fully random	PVTv2	ImageNet1K	11.17	85.77	14.38	-10.94
Rand. 1st layer			2.87	24.3	4.24	-0.04
LoRA (rank 32)			3.64	32.8	5.40	-0.27
Proposed			**2.52**	**22.9**	**4.02**	**0.00**

D.1. Access

- The dataset and model weights are hosted at [URL] with download and usage instructions at [URL].
- The data is governed by the Open License 2.0 of Etalab (https://www.etalab.gouv.fr/wp-content/uploads/2018/11/open-licence.pdf).
- Codes for data preprocessing, training models and evaluation are available at [URL].

D.2. Statistics

We provide here the additional details on the dataset.

- **Compositing:** Our dataset relies on DINAMIS, which provides one SPOT image per location each year. The LiDAR-HD supplies a single airborne LiDAR acquisition per area across the entire country. Consequently, no compositing is needed.
- **Pairing SPOT and ALS:** We pair SPOT and LiDAR data from the same year to create height annotations, resulting in a median difference of 61 days. We provide in Fig. G the temporal distribution of acquisitions.
- **Data Distribution:** Each tile represented in Fig 2 is *entirely featured in our dataset*. The spatial distribution of canopy heights and tree covers per tiles are illustrated in Fig. H and Fig. I.

D.3. Composition

We describe here the organization of the dataset. See Section E for details on how the dataset was prepared.
The dataset is organized in the following way:

- The folder `canopy_height` contains data for canopy height estimation.
- The folder `canopy_height_change` contains data for canopy height change estimation.

The composition of the `canopy_height` folder is the following:

- The file `geometries.geojson` stores a list of 95,429 $1km^2$ square geolocated geometries, giving access to the splits of the dataset. It can be loaded using the python package geopandas [1]. Each geometry designates either a train, validation, test or buffer area. This information is stored in the column `split`. There are 8,046 `buffer` tiles, 66,339 train tiles, 7,369 validation tiles and 13,675 test tiles. Additionally, each geometry is associated to a year (corresponding to the year of the corresponding LiDAR acquisition), stored in the column `lidar_year`.
- The file `forest_mask.parquet` stores geolocated geometries of forests' outlines. It can be loaded using the python package geopandas. The parquet format is used to accelerate loading time.
- Each folder 2021, 2022 and 2023 contains three files:
  - spot.vrt is a geolocalized virtual file that gives access to SPOT 6-7 images stored in the subfolder `spot`. It can be accessed through Qgis software [2] or python rasterio library [3] for instance. It has the same extent as the geometries of the associated year.
  - Similarly `lidar.vrt` gives access to ALS-derived (LiDAR) canopy height maps stored in the subfolder `lidar`.
  - Similarly `lidar_classification.vrt` gives access to classification rasters stored in the subfolder `lidar_classification`.

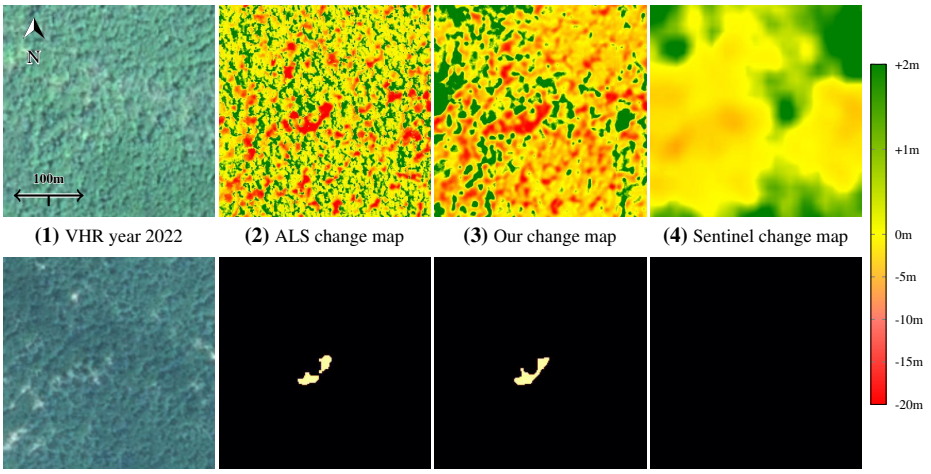
The composition of the `canopy_height_change` folder is the following:

- The file `spot_1.tif` is a geolocalized image extracted from SPOT 6-7 images in the year 2022 in the area of Chantilly, France.
- The file `spot_2.tif` is a geolocalized image extracted from SPOT 6-7 images in the year 2023 in the area of Chantilly (France).
- The file `lidar_1.tif` is a geolocalized ALS-derived height map in the year 2022 in the area of Chantilly (France), derived from LiDAR HD [10].
- The file `lidar_2_m.tif` is a geolocalized ALS-derived height map in the year 2023 in the area of Chantilly (France), provided by [11], at a resolution of 1m, with height in meters, and covering only forests.
- The file `predictions_1_m.tif` is a geolocalized height map predicted by a PVTv2 model in 2022 in the area of Chantilly (France), in meter unit.
- The file `predictions_2_m.tif` is a geolocalized height map predicted by a PVTv2 model in 2023 in the area of Chantilly (France), in meter unit.
- The file `lidar_classification.tif` is an ALS-derived classification raster in 2022 in the area of Chantilly (France).
- Additionally, files that follow the following pattern `*_masked.tif` designate images masked on the extent of the available ALS data for 2023.
- The file `change_mask_delta_15_surface_200_annotated.geojson` can be loaded with geopandas and gives access to geometries detected as "change" for a minimum height difference of 15m and a minimum surface of 200m. We also provide manual annotations of detections in the column "Rating", where "true" indicates a true positive and "false" a false positive.

[1] https://geopandas.org/en/stable/

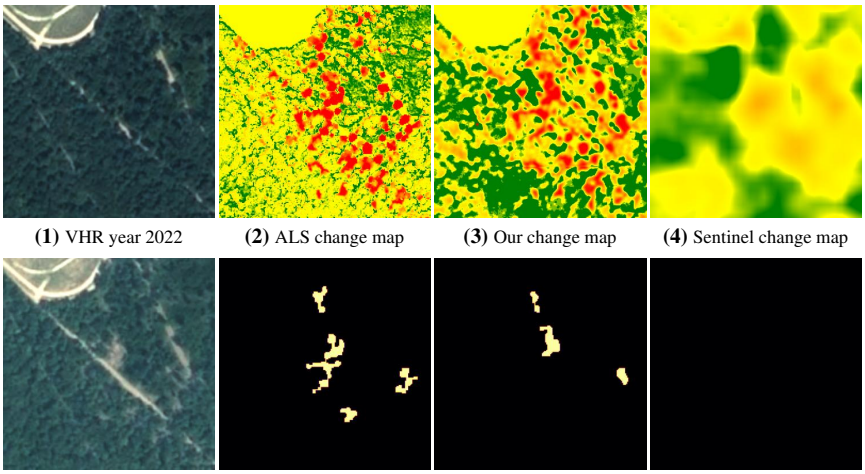
[2] https://www.qgis.org/en/site/3

[3] https://rasterio.readthedocs.io/en/stable/



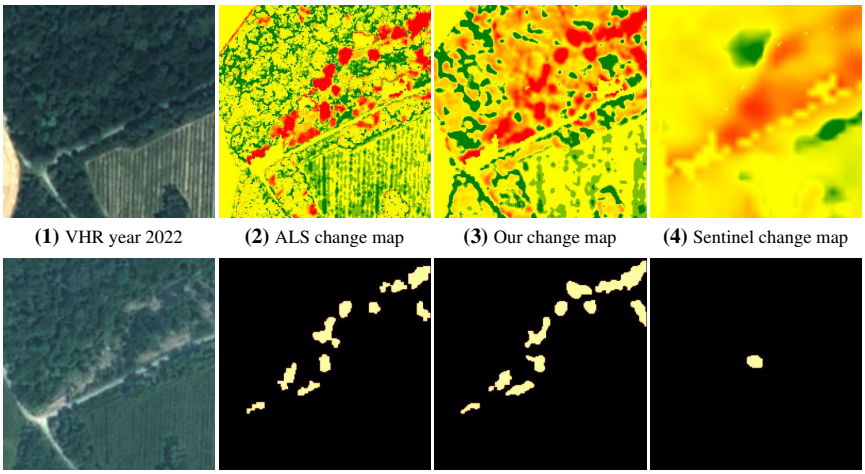
(1) VHR year 2022 **(2)** ALS change map **(3)** Our change map **(4)** Sentinel change map

(5) VHR year 2023 **(6)** ALS change mask **(7)** Our change mask **(8)** Sentinel change mask



(1) VHR year 2022 **(2)** ALS change map **(3)** Our change map **(4)** Sentinel change map

(5) VHR year 2023 **(6)** ALS change mask **(7)** Our change mask **(8)** Sentinel change mask



(1) VHR year 2022 **(2)** ALS change map **(3)** Our change map **(4)** Sentinel change map

(5) VHR year 2023 **(6)** ALS change mask **(7)** Our change mask **(8)** Sentinel change map

Figure F. Canopy Height Change. We consider VHR images taken in 2022 and 2023 in Chantilly Forest: (1) and (5), and use ALS observations of the same years to derive a canopy height change map (2). We compare this map to the ones predicted by a PVTv2 model (3) and by a model from Schwartz *et al*. trained on Sentinel data [2]. We also compare the binary change masks derived from ALS measurements (6) and from predicted change maps: (7) and (8). Scale and orientation are shared across all subfigures.

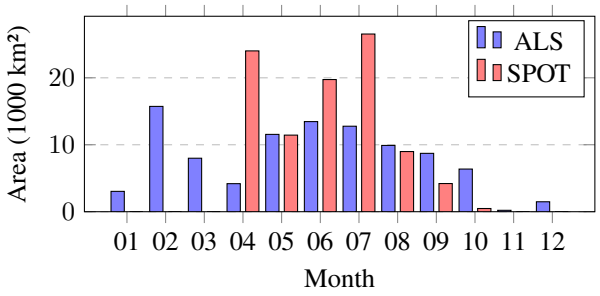


Figure G. Temporal Distribution of Acquisitions.

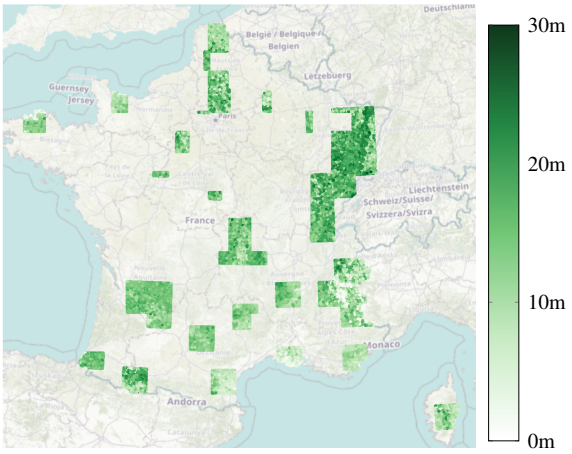


Figure H. Average Canopy Height.

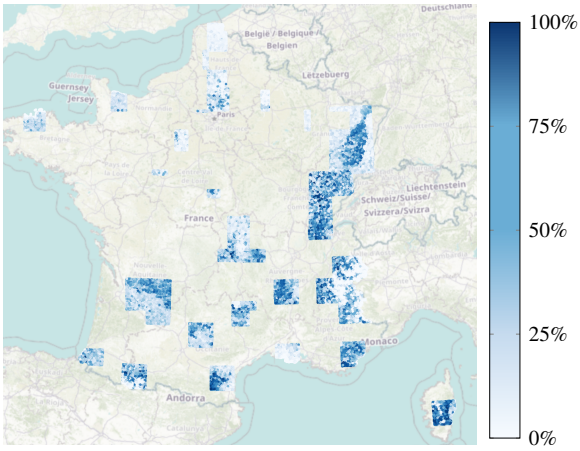


Figure I. Tree Cover Density.

D.4. Characteristics

- We provide SPOT 6-7 images, ALS-derived height maps and classification rasters covering 95,429 km^2 (including a "buffer" area of 8046 km^2, a train area of 66,339 km^2, a validation area of 7,369 km^2 and a test area of 13,675 km^2). Each image has a resolution of 1.5m, with one annotation per pixel, for a total of 42,455,312,381 annotations.
- Additionally, we provide SPOT 6-7 imagery, ALS-derived height maps and a classification raster on the Chantilly forest area for 2022 and 2023 (166 km^2).
- The Open-Canopy dataset is derived from a larger dataset of SPOT 6-7 acquisitions across the full metropolitan French territory between 2013 and 2023 [4], and a larger dataset of ALS acquisitions from the IGN campaign that started in 2021 and aims at covering the full metropolitan French territory (LiDAR HD) [5]. The Open-Canopy dataset focuses on domains that are representative of the diversity of French forests and where LiDAR HD is available at the time of writing, with the goal of limiting the dataset's size to approximately 300 GB, in order to facilitate its usage by the machine learning community.
- Each SPOT image is at a resolution of 1.5 m per pixel, and features 4 spectral channels: red, blue, green, and near-infrared.
- Each height map image is at a resolution of 1.5 m per pixel, and features 1 channel (height in decimeters except if notified in the filename in the following format: "<name>_<unit>.tif").
- Each classification image is at a resolution of 1.5 m per pixel, and features 1 channel (classification [12] for a description of classes). Forests' outlines are stored as geometries in a parquet file. A Python utility is provided to create a vegetation mask from the classification raster and the forests' outlines.

E. Dataset preparation

E.1. Splits

Our sampling strategy is semi-automated and proceeds as follows:

- SPOT images were associated to LiDAR height maps of the same year and geolocation (each LiDAR height map corresponds to a 1km^2 geolocalized square tile, referred to as "geometry" in the following).
- Geometries on overlapping areas between spot full images were removed.
- Geometries that had more than 100 zeros on the first spot band (*e.g.*, on edges of a full spot image) were discarded to avoid tiles with missing data.
- Test geometries of 1km^2 were sampled (with a fixed seed) to form contiguous squares of 7km^2 and to cover 20,000 km^2.
- Test geometries that overlapped each other were dropped.
- Test geometries that covered different years in terms

[4] https://openspot-dinamis.data-terra.org
[5] https://geoservices.ign.fr/lidarhd

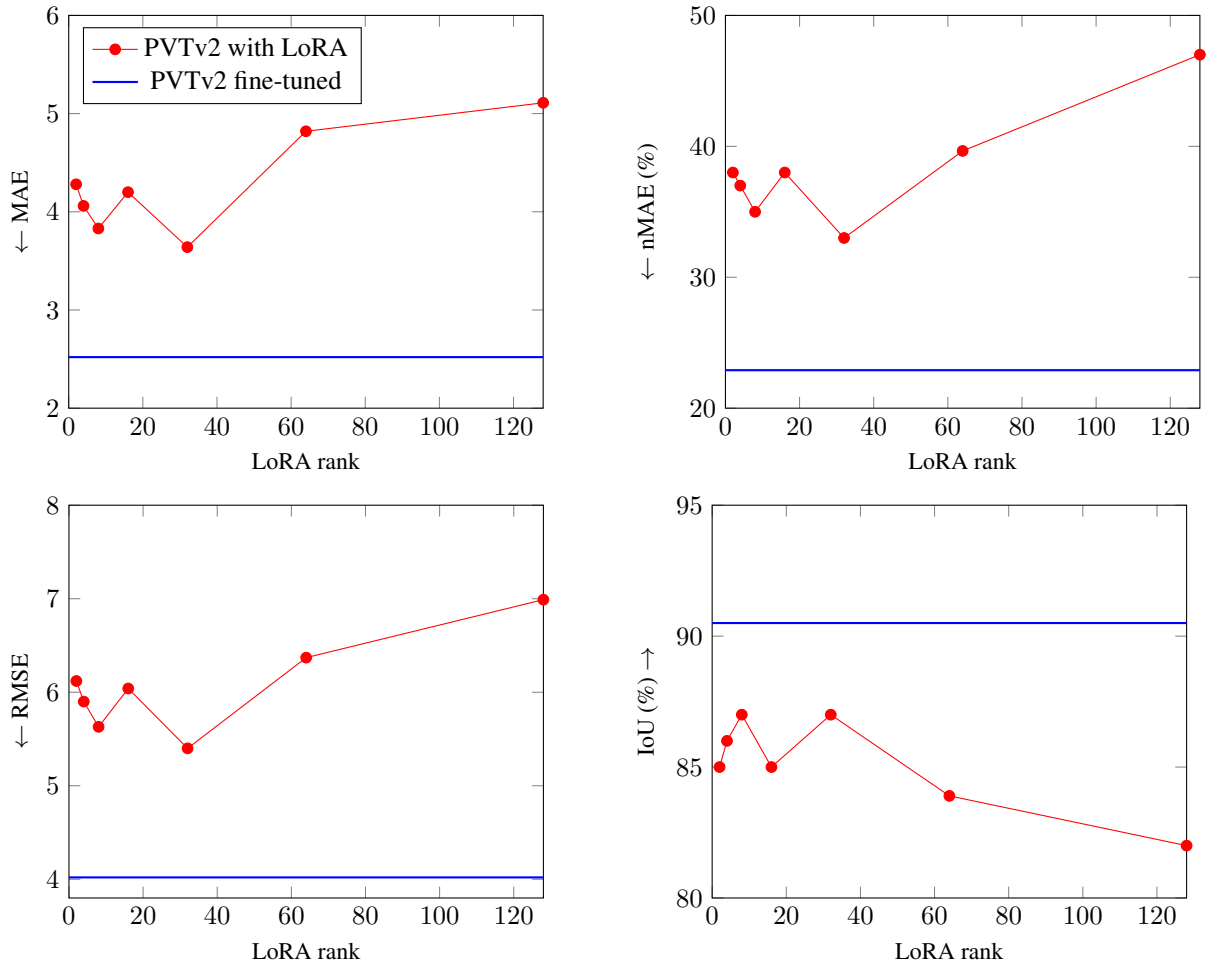


Figure J. LoRa Fine-Tuning of PVTv2. We fine-tun PVTv2 using LoRa for different rank. To allow the network to adapt to the NIR modality, we still train the first layer fully. The best results, obtained with rank $= 32$, are noticeably inferior to a fully fine-tunned PVTv2

of LiDAR acquisitions were dropped.

- This process resulted in a total test area of 13,675 km^2.
- A buffer of $1km$ was applied around each test area of 7km^2.
- Validation and train geometries were randomly sampled (with a fixed seed) among the remaining geometries, with a proportion of 10% for validation and 90% for training.
- This process resulted in a training area of 66,339 km^2 and a validation area of 7,369 km^2.

E.2. SPOT 6-7 satellite imagery

- The aerial images are sampled from the DINAMIS [6] collection. This collection consists of an annual mosaic of selected tiles taken by SPOT 6-7 satellites between March and October of each year between 2013 and 2023, covering the entire French metropolitan territory. All images are orthorectified by IGN and mapped onto a unified cartographic coordinate reference system (Lambert 93). Each tile consists of an image with four spectral bands: red, green, blue, and near-infrared at a resolution of 6m, and an image with one panchromatic band at a resolution of $1.5m$ that can be downloaded separately.
- A total of 52 pairs of spectral and panchromatic images were downloaded from the DINAMIS website, for each year from 2021 to 2023, to cover a very diverse range of forest types in areas where LiDAR HD was available at the time of the creation of the dataset.
- We applied pansharpening with the weighted Brovey algorithm [13] to upsample all four spectral bands to a resolution of 1.5m, resulting in one image with four bands for each tile.
- We cropped each image to the area covered by the ALS acquisitions of the same year.

[6] https://openspot-dinamis.data-terra.org

- Pixels values were clipped to a maximum value of 2000 to avoid outliers (upper bound both quantitatively and qualitatively assessed through histograms and visualization).
- Resulting images were normalized to a 0-255 range and saved as uint8 in a block-tiled compressed tiff format (256×256).
- The pansharpening and normalization procedures were voluntarily kept relatively simple in order to facilitate reproducibility. They may not be optimal for visualization, *e.g.*, lacking harmonization, but we expect deep learning models to be robust to such variations in input data.

E.3. ALS data

- The ALS classified point clouds were downloaded from the LiDAR HD website (IGN). A reference to each download link is saved in the file `geometries.geojson`.
- For each geometry, canopy height images were derived from ALS data by taking the maximum difference between the height of each point and the one of its nearest point classified as ground within its pixel, interpolating values in areas without data.
- LiDAR point clouds were classified by IGN into the main types of land cover (water, ground, high vegetation over 1.5m, buildings...). We use this classification to produce classification rasters at a resolution of 1.5m, where each pixel takes the value of the most frequent class of the corresponding LiDAR points.
- We then create vegetation masks by taking the union of the ALS-derived mask indicating vegetation over 1.5m in height, with the official forest plots outlines (file `forest_mask.parquet`), both provided by IGN. The resulting vegetation masks cover trees and shrubs within forest plots as well as outside, such as hedges and urban trees.
- The official forests' outlines were extracted from "BD foret" [7] and "simplified" using geopandas python library to a precision of 10m, with the goal to limit their size.

F. Datasheet for Open-Canopy dataset

F.1. Motivation

- **For what purpose was the dataset created?** Was there a specific task in mind? Was there a particular gap that needed to be filled? Please provide a description.
  The Open-Canopy dataset was created to train and evaluate models that (i) predict very-high resolution canopy height maps from satellite imagery using LiDAR scans for ground truth, and (ii) detect canopy height changes between images from different years. The main gap we are addressing is the lack of curated open-source datasets with both very high resolution imagery and ALS-based (LiDAR) canopy height maps.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** This dataset was curated by a team of researchers from [TEXT REMOVED FOR ANONYMITY] using data made available by DINAMIS and IGN. DINAMIS [14] is a French platform that provides access to earth observation products for public benefit programs. The IGN is a French public state administrative establishment aiming to produce and maintain geographical information for France.
- **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.
  The funding of the Open-Canopy dataset is 100% public. Open-Canopy benefited from funding by [TEXT REMOVED FOR ANONYMITY]
- **Any other comments?**
  N/A.

F.2. Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?**
  The dataset is split into square areas of width 1.0005 km, rasterized to a 1.5 m resolution (667×667 pixels). Each instance corresponds to an area of 1 km^2 on the French metropolitan territory.
- **How many instances are there in total (of each type, if appropriate)?**
  We provide 95,429 instances of 1km^2: 66,339 train tiles, 7,369 validation tiles, 13,675 test tiles, and 8,046 "buffer" tiles. This corresponds to a total of 42,455,312,381 individual annotated pixels.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?**
  The Open-Canopy dataset covers 17% of the French metropolitan territory. It is derived from a larger dataset of SPOT 6-7 acquisitions across the full metropolitan French territory between 2013 and 2023 (`https://openspot-dinamis.data-terra.org`), and a larger dataset of ALS acquisitions from the campaign that started in 2021 and aims at covering the full metropolitan French territory (LiDAR HD)[10]. The Open-Canopy dataset focuses on domains that are representative of the diversity of French forests and where LiDAR HD is available at the time of submis-

[7] https://geoservices.ign.fr/bdforet#telechargementv2

sion. We also aimed to limit the dataset's size to 300 GB to facilitate its use.

- **What data does each instance consist of?**
  Each instance consists of a GeoJSON geometry ($1km^2$), for which a 667×667 SPOT image, a height map, and a vegetation mask can be extracted from associated .vrt files, in order to associate to each pixel the following values: (i) RGB and near Infrared channels derived from pan-sharpened and ortho-rectified satellite images from SPOT 6-7 acquired between 2021 and 2023; (ii) canopy height derived from LiDAR HD's 3D point clouds [10] acquired in the same year; (iii) label (*e.g.*, vegetation, ground, water, building) derived from LiDAR HD's 3D point clouds [10].
  Additionally, we provide forest outlines obtained from IGN's portal [15] stored as a parquet file.
- **Is there a label or target associated with each instance?**
  Yes. We provide a complete pixel-precise height map and classification raster of the same extent as the satellite images.
- **Is any information missing from individual instances?**
  No. We provide dense information (radiometry, canopy height, class label) for all pixels with the exception of areas that have been selected by the French government as "sensitive" for security reasons (*e.g.*, nuclear plants, military area). We do not provide the 3D point clouds from LiDAR HD, but they are accessible on their platform.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?**
  N/A.
- **Are there recommended data splits (e.g., training, development/validation, testing)?**
  Yes, we provide data splits for reproducing the results of the benchmark. The test split has been explicitly selected to address the complex domain shifts of geospatial data and separated from the train and validation splits by a 1 km^2 buffer to avoid data contamination.
- **Are there any errors, sources of noise, or redundancies in the dataset?**
  The annotations from ALS (LiDAR) data include inherent inaccuracies due to the nature of the acquisition process. Multipath effects from multiple echoes can introduce errors, and outlier points may impact the quality of the canopy height maps. Additionally, variations in tree height due to different acquisition times across seasons can affect consistency between ALS and VHR acquisitions, as trees might be at various stages of their growth cycle. Input images sourced from satellite data pre-processed by IGN and DINAMIS may still exhibit artifacts due to cloud cover or contain small registration errors that can impact the analysis.
  Classification rasters derived from ALS data are also subject to inaccuracies. These can stem from inherent limitations in the ALS technology, including noise in the data which may lead to errors in vegetation classification.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** This dataset is self-contained and will be stored on the Huggingface platform. The dataset is under the Open License 2.0 of Etalab.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?**
  No. The classification raster does not contain any information that would not be available in other open-access sources (DINAMIS, BD-Foret, LiDAR-HD). We have specifically avoided high-risk areas such as military installations or nuclear plants.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*
  No.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?**
  No.
- **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?**
  No. The resolution of 1.5m per pixel and the aerial perspective makes identifying individuals impossible.
- **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?**
  No.
- **Any other comments?**
  No.

F.3. Collection Process

- **How was the data associated with each instance acquired?**
  The satellite images are sampled from the DINAMIS open SPOT collection. This collection consists of an annual mosaic of selected images taken by SPOT 6-7

satellites between March and October of each year between 2013 and 2023, covering the entire French metropolitan territory. All images are preprocessed by IGN and mapped onto a unified cartographic coordinate reference system (Lambert 93).
- The ALS classified point clouds were downloaded from the LiDAR HD website (IGN).
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?**
  The IGN selected several acquisition companies through a call for tender with strict specifications.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**
  The sampling strategy was semi-automated. First a manual selection of spot images was manually chosen and downloaded from DINAMIS website, so as to cover a diverse range of forests types in areas where LiDAR HD was also available. Then training, validation, and test splits were randomly sampled, with constraints such as test tiles having a size of 7 km^2 and being separated from other tiles by a buffer of 1 km^2, and covering an area of about 14,000 km^2. See Section E.1 for more details.
- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**
  The data collection process for the dataset was managed by the European Space Agency (ESA), which provided the Very High Resolution (VHR) Imagery, and the French Mapping Agency (IGN), which provided the LiDAR HD data. The curation of this dataset was overseen by two individuals who were associated with academic institutions as a postdoctoral researcher (ENS) and an intern (LSCE) during the dataset's creation.
- **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?**
  The collection of satellite imagery and ALS data spans from 2021 to 2023, which coincides with the period of availability of LiDAR HD data at the time of the creation of the dataset.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?**
  No.
- **Does the dataset relate to people?**
  No.
- **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**
  N/A.
- **Were the individuals in question notified about the data collection?**
  N/A.
- **Did the individuals in question consent to the collection and use of their data?**
  N/A.
- **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?**
  N/A.
- **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?**
  No. Given the nature of the dataset—which involves high-resolution canopy height data that does not include personal identifiers or directly impact individual privacy—it is unlikely that the dataset poses significant risks to data subjects. The focus is primarily on environmental features rather than personal data.
- **Any other comments?**
  No.

F.4. Preprocessing, Cleaning, and/or Labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?**
  Canopy Height Maps were derived from ALS data by taking the maximum difference between the height of each point and the one of its nearest point classified as ground within its pixel, interpolating values in areas without data.
- For vegetation masks, we take the union of the ALS-derived mask indicating vegetation over 1.5m in height, with the official forest plots outlines, both provided by IGN. The resulting vegetation mask covers trees and shrubs within forest plots as well as outside, such as hedges and urban trees. The official forests' outlines were "simplified" using geopandas python library to a precision of 10m, in order to limit their size.
- SPOT 6-7 images were pansharpenened with the weighted Brovey algorithm to upsample all four spectral bands to a resolution of 1.5m. Then all pixels values were clipped to a maximum value of 2000 to avoid outliers and normalized to a 0-255 range to be saved as uint8, in a block-tiled compressed tiff format.
- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

Yes. The raw data can be downloaded from DINAMIS and LiDAR HD websites.

- **Is the software used to preprocess/clean/label the instances available?**
  Yes. All the codes to preprocess the data are available on the Github of the project [TEXT REMOVED FOR ANONYMITY]
- **Any other comments?**
  No.

F.5. Uses

- **Has the dataset been used for any tasks already?**
  No.
- **What (other) tasks could the dataset be used for?**
  We encourage future researchers to use the Open-Canopy dataset for several tasks. Particularly, the dataset could be used to predict land cover in addition to canopy height, using the classification rasters as complimentary labels. It could also be used for pre-training of models for other tasks such as tree cover segmentation and tree species classification.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?**
  This dataset is geographically limited to metropolitan France. Although France's territory is diverse, featuring oceanic, continental, Mediterranean, and mountainous bioclimatic regions, it does not contain tropical or desert areas.
- The Open-Canopy dataset's reliance on purely optical data may limit the applicability of the models trained on it to regions with pervasive cloud cover.
- **Are there tasks for which the dataset should not be used?**
  No.
- **Any other comments?**
  No.

F.6. Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?**
  Yes. the dataset will be open-source.
- **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?**
  The data will be hosted on Huggingface platform ([TEXT REMOVED FOR ANONYMITY]), with download and usage instructions on the Open-Canopy project page hosted on GitHub ([TEXT REMOVED FOR ANONYMITY]).
- **When will the dataset be distributed?**
  All data is already released under an open-source license, see below.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
  Yes. The data is governed by the Open Licence 2.0 of Etalab (`https://www.etalab.gouv.fr/wp-content/uploads/2018/11/open-licence.pdf`).
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?**
  No.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?**
  No.
- **Any other comments?**
  No.

F.7. Maintenance

- **Who will be supporting/hosting/maintaining the dataset?**
  Hugginface will support hosting of the dataset and metadata. [TEXT REMOVED FOR ANONYMITY] will support maintenance of the dataset in case of revisions.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**
  [TEXT REMOVED FOR ANONYMITY]
- **Is there an erratum?**
  No. There is no erratum for our initial release. Errata will be documented as future releases on the dataset web page.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?**
  Additional satellite imagery and ALS-derived height maps may be added to future versions of the Open-Canopy dataset.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?**
  N/A..
- **Will older versions of the dataset continue to be supported/hosted/maintained?**
  Yes. We are dedicated to providing ongoing support for the Open-Canopy dataset.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?**

Proposed extensions or corrections to the Open-Canopy dataset may be submitted to the providers for consideration. The providers will assess the feasibility of incorporating the suggested modifications, considering factors such as data licensing, maintenance requirements, and relevance.

- **Any other comments?**
  No.

References

[1] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. Datasheets for datasets. *Communications of the ACM*, 2021.

[2] Martin Schwartz, Philippe Ciais, Aurélien De Truchis, Jérôme Chave, Catherine Ottlé, Cedric Vega, Jean-Pierre Wigneron, Manuel Nicolas, Sami Jouaber, Siyu Liu, Martin Brandt, and Ibrahim Fayad. FORMS: Forest multiple source height, wood volume, and biomass maps in France at 10 to 30m resolution based on Sentinel-1, Sentinel-2, and GEDI data with a deep learning approach. *Earth System Science Data*, 2023.

[3] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. PVTv2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 2022.

[4] Peter Potapov, Xinyuan Li, Andres Hernandez-Serna, Alexandra Tyukavina, Matthew C Hansen, Anil Kommareddy, Amy Pickens, Svetlana Turubanova, Hao Tang, Carlos Edibaldo Silva, et al. Mapping global forest canopy height through integration of GEDI and Landsat data. *Remote Sensing of Environment*, 2021.

[5] Martin Schwartz. *Mapping forest height and biomass at high resolution in France with satellite remote sensing and deep learning*. PhD thesis, Université Paris-Saclay, 2023.

[6] Nico Lang, Walter Jetz, Konrad Schindler, and Jan Dirk Wegner. A high-resolution canopy height model of the Earth. *Nature Ecology & Evolution*, 2023.

[7] Jan Pauls, Max Zimmer, Una M Kelly, Martin Schwartz, Sassan Saatchi, Philippe Ciais, Sebastian Pokutta, Martin Brandt, and Fabian Gieseke. Estimating canopy height at scale. In *ICML*, 2024.

[8] Siyu Liu, Martin Brandt, Thomas Nord-Larsen, Jerome Chave, Florian Reiner, Nico Lang, Xiaoye Tong, Philippe Ciais, Christian Igel, Adrian Pascual, et al. The overlooked contribution of trees outside forests to tree cover and woody biomass across Europe. *Science Advances*, 2023.

[9] Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, et al. Very high resolution canopy height maps from RGB imagery using self-supervised vision transformer and convolutional decoder trained on aerial LiDAR. *Remote Sensing of Environment*, 2024.

[10] IGN. LiDAR HD : Vers une nouvelle cartographie 3d du territoire. https://www.ign.fr/institut/lidar-hd-vers-une-nouvelle-cartographie-3d-du-territoire, 2024. [Online; accessed 12-May-2024].

[11] Institut de France. Collectif sauvons la foret de chantilly. https://chateaudechantilly.fr/la-foret/ensemble-sauvons-la-foret-de-chantilly/, 2024. [Online; accessed 12-May-2024].

[12] IGN. Lidar hd technical description. https://geoservices.ign.fr/sites/default/files/2023-10/DC_LiDAR_HD_1-0_PTS.pdf. Online; accessed 2024-02-21.

[13] Alan R Gillespie, Anne B Kahle, and Richard E Walker. Color enhancement of highly correlated images. ii. channel ratio and "chromaticity" transformation techniques. *Remote Sensing of Environment*, 1987.

[14] DINAMIS. French national facility for institutional procurement of vhr satellite imagery. https://openspot-dinamis.data-terra.org, 2024. [Online; accessed 12-May-2024].

[15] IGN. Forest data base. https://geoservices.ign.fr/bdforet, 2024. [Online; accessed 12-May-2024].