ChatGPT

Datasets for Tree Species Semantic Segmentation in Aerial Imagery

Top 3 Dataset Recommendations

Below are three high-quality datasets (since 2020) that provide top-down aerial/drone imagery with **pixel-level tree species labels**. Each dataset meets the requirement of species-level semantic segmentation (or instance segmentation with species labels) and is suitable for academic research (open license). Key details and pros/cons are summarized for each:

1. FORTRESS – *Forest Tree Species Segmentation (Black Forest, Germany)*

- **Source/Authors:** Karlsruhe Institute of Technology (F. Schiefer, J. Frey, T. Kattenborn et al., 2022) [1] [2]
- **Link:** Hosted on KIT's RADAR repository (DOI: 10.35097/538). Open access download (~40.5 GB) [3] [4].
- **Data Size:** ~40 GB (covers ~47 ha of forest) [5] [3]
- **Species Labeling: Semantic segmentation** – each pixel is labeled with a tree species class. Contiguous canopy regions of the **same species** are masked as one class (no individual tree instances) [6] [7]. (Temperate mixed forest species; number of classes not explicitly stated, but includes common Central European tree species in the study area.)
- **Imagery:** UAV (drone) **RGB orthomosaics** of temperate forest (Southern Black Forest) [8]. Collected 2017–2019, with leaf-on conditions.
- **Resolution (GSD):** Very high (centimeter-level). UAV flights yield ~**2–5 cm** per pixel GSD (typical of low-altitude drone orthos).
- **Format:** Georeferenced **orthoimage(s)** (GeoTIFF) plus **pixel-wise species masks** (either multi-class raster masks or polygon shapefiles per species). The dataset provides aligned orthomosaics and corresponding masks labeling tree species [8].
- **License:** Creative Commons Attribution 4.0 (CC BY 4.0) – free to use with credit [4].
- **Pros:** Provides true *semantic segmentation* of tree species – each species forms a region label, matching the use-case of homogeneous species patch mapping. High spatial resolution (~3 cm) supports fine-grained canopy detail. Focused on natural forest conditions (mix of coniferous & deciduous) with expert-derived labels. Moderate dataset size is manageable, and data is ready to use (no preprocessing needed to get masks). Open license for research and publication.
- **Cons:** Covers a single geographic region (Black Forest) with a limited set of species (native to that region), so species diversity is moderate. Imagery is RGB only (no NIR bands). Since labels are at semantic level (not instance), individual tree crowns are not distinguished – contiguous same-species trees appear as one segment [7]. Documentation is concise (data comes with metadata but without a detailed benchmark paper beyond the archive abstract). Overall, a slightly smaller scale compared to some multi-site datasets.

2. Quebec Trees Dataset – *Multi-season UAV Tree Species Annotations (Quebec, Canada)*

- **Source/Authors:** University of Montréal & Univ. Sherbrooke (M. Cloutier, M. Germain, E. Laliberté, 2023) [9] . Created for a 2023 preprint study on seasonal effects in species segmentation [10] .
- **Link:** Available on **Zenodo** (DOI: 10.5281/zenodo.8148479) – open dataset published Sept 2023 [11] . ~149 GB split across 7 archives (one per date) [12] .
- **Data Size:** ~150 GB total (includes 7 multi-date image sets, each ~20–23 GB) [12] . Represents ~**22,000 labeled trees** over the study area [13] [14] .
- **Species Labeling: Instance segmentation** at the individual tree crown level, with each crown polygon labeled by **species**. Contains **~23k tree crowns** from **14 species classes** (mostly at species level, a few at genus level) [13] [15] . Examples include balsam fir, red maple, sugar maple, yellow birch, spruce spp., aspen spp., etc., reflecting a **mixed temperate forest**. All crowns were manually delineated and tagged by experts; contiguous trees of the same species are separate instances (not merged) in this dataset.
- **Imagery: Drone RGB orthomosaics**, acquired on 7 dates through the 2021 growing season (May–October) to capture phenological variation [16] . Each date covers the *same forest site* (divided into 3 zones) with high-resolution imagery. The repeat coverage allows analysis of seasonal color changes for species. Images are provided as Cloud-Optimized GeoTIFFs (COGs). No multispectral bands (RGB only), but the multi-temporal aspect is a unique feature.
- **Resolution (GSD):** High resolution, roughly **5 cm per pixel or better**. (Exact GSD not explicitly stated; flights were low-altitude UAV. Given the fine detail required for species ID, resolution is on the order of a few centimeters.) Each orthomosaic is detailed enough to distinguish individual crowns and some color/texture differences (e.g., leaf color in autumn).
- **Format:** Geospatial format – for each date and zone, an **RGB orthomosaic** (GeoTIFF) and a vector file of **tree crown polygons** with species labels (GeoPackage `.gpkg`) [17] . Annotations include a tree ID and species code per polygon. Photogrammetric point clouds (COPC .laz) for each date are also included (useful for height, but not required for segmentation tasks) [13] [18] . This organization allows direct use in GIS or for machine learning (polygons can be rasterized to masks, or used as labels for instance segmentation models).
- **License:** Creative Commons Attribution 4.0 (CC BY 4.0) [19] . The data can be used freely for research with attribution to the creators.
- **Pros: Large and richly annotated** dataset – 22k tree instances across 14 species, each with expert-verified labels [20] . Provides **multi-season imagery**, which can improve species discrimination (leaf-on/off differences) or enable seasonal robustness testing. High spatial resolution (~cm-level) and **precise polygon annotations** make it valuable for training segmentation models. Data is in standard GIS formats (COG and GeoPackage) with georeferencing and metadata, ensuring easy integration into workflows. The variety of species (maples, birches, conifers, etc.) and natural mixed forest context offer a realistic, complex scenario for model development. Open license and active use in recent research ensure good documentation and support [10] .
- **Cons: Very large download size (150+ GB)** – requires significant storage and time to download. One can choose a subset (e.g. one date) if full temporal analysis isn't needed, but that still means ~20+ GB. The dataset is from **one geographic location** (a single temperate forest site in Quebec), so while it has diversity of species, it does not cover different landscapes (e.g., no urban trees, no tropical species). All imagery is **RGB only** (no NIR or multispectral bands), so spectral differentiation is limited to visible colors – in some cases species with similar visual appearance might be hard to separate (the authors did merge a few very similar species at the genus level due to identification difficulty [21]). Nonetheless, the multi-date aspect mitigates this by providing phenological cues. Computing instance segmentation from polygons may require

rasterization for some workflows, but the provided data makes this straightforward. Overall, the dataset's richness comes at the cost of size and scope being focused on one region.

3. Northern Australia Savanna Trees – *Tropical Savanna Tree Species Dataset (Kakadu, Australia)*

- **Source/Authors:** James Cook University and collaborators (A. J. Jansen et al.). Released via an MDPI Data descriptor (2023) [22] [23]. Sometimes referred to as the "Savanna Tree AI Dataset."
- **Link:** Published on Zenodo (DOI: 10.5281/zenodo.7094916) with CC BY 4.0 license [24] [23]. Also associated with GeoNadir for raw imagery. The dataset (~8 GB for COCO tiles) is easily accessible.
- **Data Size:** Relatively small – consists of 7 orthomosaic images (each ~1 ha area) and annotation files. The processed machine-learning-ready version (COCO formatted tiles) is around **8 GB** (7 images cropped into many 1024×1024 px tiles plus JSON). Original orthomosaics (GeoTIFF) and shapefiles are also provided (a few GB). In total, **2,547 tree instances** are labeled [24].
- **Species Labeling: Instance segmentation** – **2,547 polygons delineating individual tree crowns**, each labeled with one of **36 tree species** [24]. This is a multi-species dataset covering tropical savanna taxa (e.g., *Eucalyptus tetrodonta*, *E. miniata*, *Acacia* spp., etc.). Species labels were confirmed via on-ground surveys and botanist verification [25] [26], ensuring high label accuracy. All polygons are separate tree crowns; there is no merged region – the focus is on individual tree species identification.
- **Imagery: Drone-based RGB imagery** collected over 7 distinct 1-hectare study plots in Northern Australia's savanna (within Kakadu National Park) [27]. Imagery was captured via a DJI drone at ~80 m altitude, resulting in **high-resolution orthomosaics (~2 cm GSD)** [28] [29]. Each orthomosaic corresponds to a field plot where trees were mapped. The environment is an open woodland (scattered trees over grass), so individual crowns are generally distinct. No multispectral data – only RGB bands are available.
- **Resolution (GSD):** ~**1.8–2.0 cm** per pixel (very high detail) [28] [29]. The fine resolution helps differentiate species by crown shape and color. Even small trees are captured with several pixels across the crown.
- **Format:** Two forms are provided: (1) **Georeferenced orthomosaics + Shapefiles** – Seven GeoTIFF orthoimages (each ~100m × 100m) with accompanying ESRI Shapefile (or GeoJSON) containing polygon annotations for tree crowns (with species names as an attribute) [27] [26]. This is useful for GIS analysis or custom processing. (2) **COCO-format dataset** – the orthomosaics were tiled into 1024×1024 pixel images, and a COCO annotation JSON is provided for easy use in deep learning frameworks [24]. This COCO set has the 2,547 instances with species labels (the species name acts as the "category" of each instance). This dual format means one can either work in geospatial context or directly train models using the COCO data.
- **License:** Creative Commons Attribution 4.0 (CC BY 4.0) [23] – completely open for research and publication with attribution.
- **Pros:** The dataset addresses a **unique ecosystem and species set** (tropical savanna) that is complementary to temperate forest datasets. It has **36 species**, providing high taxonomic diversity [24] – useful for testing multi-class discrimination. All labels are very **accurate** (ground-truthed in the field) [26]. The imagery's resolution (~2 cm) is excellent, and the relatively simple background (grassland) means crowns are well-defined against surroundings. The data is already in an ML-ready format (COCO), saving preprocessing time – one can plug it into instance segmentation training pipelines out-of-the-box. The size is manageable (a few GB) – quick to download and start experimenting. Documentation is clear (the Data journal article explains data collection and provides baseline results [22] [24]).
- **Cons:** The **scale is smaller** – only 7 hectares surveyed, with 2,547 trees. For deep learning, this is on the lower end for training data (though augmentation and tile overlaps effectively increase sample count). Species distribution is **imbalanced** (some dominant species have a few hundred

instances while many others have only a handful) [30], which can make training challenging for the rare classes. The **use-case is specialized**: savanna trees in dry tropical Australia – models trained here may not directly generalize to dense temperate forests or urban trees. Also, only RGB data is available; some species differences might be subtle without NIR or other spectra (although the authors note that spectral data (hyperspectral/LiDAR) was historically needed for savanna species, this dataset attempts RGB-only solutions [31] [32]). In summary, this dataset is extremely useful for its niche (and as a test of model generalization to new regions), but has fewer samples and a narrower context than larger temperate datasets.

Comparison of Recommended Datasets

The table below compares key attributes of the three recommended datasets:

Dataset	Species Labels	Imagery Type	GSD	Format	License
FORTRESS (2022)	Semantic segmentation (pixel-level); tree species per pixel (contiguous same-species areas) [6] [7]. *~5 species (temperate forest)*	UAV RGB orthomosaics (Southern Black Forest, DE) [8]	~3–5 cm (very high-res UAV)	GeoTIFF orthomosaic + labeled mask images (per species) [8]	CC BY 4.0 [4]
Quebec Trees (2023)	Instance segmentation (crown-level polygons); **14 species** classes (mixed hardwood/conifer forest) [13] [15]	UAV RGB orthomosaics – 7 dates (May–Oct) in one Quebec forest site [16]	~5 cm (high-res UAV)	GeoTIFF orthos (COG) + vector crown polygons with species (GeoPackage) [17]	CC BY 4.0 [19]
Savanna Trees (2023)	Instance segmentation (crown polygons); **36 species** (tropical savanna woodland) [24]	UAV RGB orthomosaics – 7 plots (1 ha each) in N. Australia [27]	~2 cm (ultra high-res UAV) [28] [29]	GeoTIFF orthos + shapefile polygons; also COCO 1024×1024 image tiles + JSON [24] [33]	CC BY 4.0 [23]

Citations for table data: FORTRESS from dataset abstract [8]; Quebec from dataset description [13] [15]; Savanna from data descriptor [24]. GSD for Savanna from Jansen et al. [28] [29]. All licenses CC BY 4.0.

Download & Access Instructions – *Quebec Trees Dataset* (Best Option)

The **Quebec Trees Dataset** is an excellent choice due to its large scale and detailed species annotations. Here are step-by-step instructions to access and use this dataset:

1. **Access the Repository:** Visit the dataset's page on Zenodo: **Quebec Trees Dataset on Zenodo** [11]. This page provides the dataset files, description, and metadata. No sign-up is required; the data is openly accessible.

2. **Download the Data:** You can download files individually or use the "Download all" option (which provides a ZIP of all files). The data is organized by date: e.g., `quebec_trees_dataset_2021-05-28.zip`, `2021-06-17.zip`, … up to `2021-10-28.zip` (7 files) [34] [35]. Each ZIP contains the orthomosaic and annotations for that date (split into 3 zone sub-areas). **Select the files** you need – for full use, download all seven archives (~149 GB total) [12]. If storage or bandwidth is a concern, you might start with one file (e.g., a summer month) to test. *(Tip: Using a download manager or Zenodo's command-line* `wget` *script (available via the Zenodo API) is recommended for large files.)*

3. **Verify and Extract:** Each ZIP has an MD5 checksum listed on Zenodo [36]. After downloading, verify the file integrity (optional but recommended given size). Then extract the ZIPs – you will get **Cloud-Optimized GeoTIFF (COG) images** and a **GeoPackage (** `.gpkg` **)** for annotations in each. For example, `2021-07-21.zip` might contain `2021-07-21_zone1.tif`, `2021-07-21_zone2.tif`, etc., and a `2021-07-21_annotations.gpkg` (plus metadata like README).

4. **Data Contents:** Inside each GeoPackage are polygon features for individual tree crowns with attributes for species. You can open the `.gpkg` in GIS software (QGIS, ArcGIS) or in Python (with GeoPandas, Fiona, etc.) to inspect the attributes. The species are labeled with codes (as per Table 1 in the documentation, e.g., "ACRU" for red maple, "Picea" for spruce genus, etc.) [15]. The README provided in the dataset explains these codes and the class list.

5. **Using the Data:** Because the imagery is geo-referenced (likely in UTM projection, and as COGs), you can load the TIFFs in any GIS or remote sensing software. To create **semantic segmentation masks** from the vector crowns, you can rasterize the GeoPackage polygons to the same raster grid as the orthomosaic. Each species could be assigned a unique integer ID in the output mask. Alternatively, for instance segmentation tasks, you can use the polygons directly (many deep learning frameworks accept GeoJSON/Shapefile annotations or you can convert them into COCO format). The dataset's multi-date nature means you have several sets of training data – you can use all dates for a large training set or perhaps use one date for training and another for testing model generalization to different seasons.

6. **Citation and License:** Remember to cite the dataset and the associated preprint if you use this data in publications. The Zenodo page provides a citation (BibTeX) and notes the CC BY 4.0 license [19], which means you should credit *Cloutier et al. (2023)* when publishing results. Academic use and even derived publications (figures, etc.) are allowed under this license, provided attribution is given.

By following these steps, you will have the Quebec Trees Dataset ready for use. As a quick example, you could open a zone's TIFF and overlay the species polygons in QGIS to visually verify the labels – you should see each crown outlined and colored by species, matching the underlying imagery (e.g., conifers vs. deciduous visible in the orthophoto).

Fallback Plan if No Ideal Dataset is Found

In case none of the above datasets perfectly suits your needs (e.g., specific locale, different species, or format requirements), here are some fallback strategies:

- **Utilize TreeSatAI Benchmark Archive:** The *TreeSatAI* dataset (2022) provides multi-source imagery and forest species labels for a large area in Germany [37]. It includes 20 tree species derived from forest inventory data, with aerial RGB imagery at 0.2 m (20 cm) resolution and corresponding Sentinel-1/2 data [37] [38]. While TreeSatAI is geared toward patch-level classification (each sample covers an area that may contain multiple species, hence **multi-label** classification), you can repurpose it for segmentation in creative ways. For example, one could take the high-resolution aerial imagery and overlay known tree positions or stand polygons from the inventory to assign species labels to pixels (essentially creating a segmentation mask from the inventory data). This may involve using the forest administration data that links coordinates or compartment maps to species. TreeSatAI is openly shared (on Zenodo) [39]; you could extract smaller regions where a single species dominates and treat those as labeled segments, or use it to train a classifier to predict species per pixel, then refine with segmentation algorithms. This is admittedly a complex approach, but TreeSatAI offers a rich source of labeled imagery if a ready-made segmentation mask dataset isn't available. It's a valuable resource with broad coverage (Lower Saxony state forests) and could complement the above datasets.

- **Synthetic Labeling from Inventory or Custom Data:** If you have access to **forest inventory data or tree surveys** for your area of interest, you can create a pseudo-dataset for segmentation. For instance, many cities and forest management agencies have GIS data of trees or stands with species names (e.g., a city tree inventory with GPS points and species, or forest stands labeled by dominant species). You could acquire high-resolution aerial imagery (such as UAV flights or public aerial photos at 10–30 cm GSD) and then use the inventory coordinates to mark tree locations. By applying a crown delineation algorithm (or even manual digitization for a small sample), you can generate polygons around tree canopies and assign the inventory species label to those polygons. This effectively produces a labeled segmentation dataset. For example, the PASADENA Urban Trees inventory (80k trees of 18 species) combines city tree records with aerial imagery [40] – one could use those records to cut out or label patches of the aerial image by species. Another example is the New York City street tree map (NYC Parks) with >680k trees and species [41]; paired with high-res aerial imagery (e.g., NYC orthophotos), it could provide training data if segmented appropriately. This approach requires effort in data fusion and perhaps cleaning of position errors (and note that crowns can overlap, complicating polygon drawing). But as a fallback, **it allows you to generate a dataset tailored to your region and species**, using available open data. Ensure any imagery you use is licensed for use and that your labeling process is consistent. The upside is academic projects can often leverage local data to create novel datasets if nothing off-the-shelf exists.

- **Contact Data Owners / Ongoing Projects:** Some high-quality datasets are under development or partially available, and reaching out can be fruitful. For example, the **BAMFORESTS** dataset (described above) provided tree crown shapes and noted species were recorded, but initially withheld in the public release [42]. The authors have stated plans to release the species labels in

a future version (BAMFORESTS-2) [43]. Proactively contacting the creators (Jonas Troles and team) might give you access to the species annotation data or at least early insight. In general, many forestry research groups are actively mapping tree species; even if their datasets are not fully public, they may share data for academic collaboration. Check for recent publications or data papers (the **OpenForest** data catalogue (2023) [44] is a great reference listing many forest datasets) and consider emailing the authors for access. Similarly, government or NGO programs (for instance, regional ecological monitoring) might have unpublished species segmentation data – a polite inquiry explaining your research intentions can sometimes open doors to use data under specific agreements. As a fallback plan, this networking approach can help you obtain or create a dataset that exactly meets your needs when an ideal open dataset is not found.

In summary, while there are only a few ready-made datasets with **pixel-level tree species labels**, the options above (FORTRESS, Quebec, Savanna) cover a range of environments and scales. If those do not suffice, TreeSatAI and other resources can be leveraged to assemble a workable dataset. Combining these strategies – using existing data and augmenting with your own labeled samples – will ensure you have a robust dataset for your tree species segmentation research. Good luck with your project!

---

[1] FORTRESS
https://www.radar-service.eu/radar/en/dataset/AwwREVscwqlcTVSw.FORTRESS

[2] [3] [4] [8] FORTRESS
https://radar.kit.edu/radar/en/dataset/AwwREVscwqlcTVSw

[5] [6] [7] [42] [43] BAMFORESTS: Bamberg Benchmark Forest Dataset of Individual Tree Crowns in Very-High-Resolution UAV Images
https://www.mdpi.com/2072-4292/16/11/1935

[9] [10] [11] [12] [13] [14] [15] [16] [17] [18] [19] [20] [21] [34] [35] [36] Quebec Trees Dataset
https://zenodo.org/records/8148479

[22] [23] [24] [25] [26] [27] [28] [29] [30] [31] [32] [33] Deep Learning with Northern Australian Savanna Tree Species: A Novel Dataset
https://www.mdpi.com/2306-5729/8/2/44

[37] [38] [39] ESSD - TreeSatAI Benchmark Archive: a multi-sensor, multi-label dataset for tree species classification in remote sensing
https://essd.copernicus.org/articles/15/681/2023/

[40] [41] GitHub - blutjens/awesome-forests: A curated list of ground-truth forest datasets for the machine learning and forestry community.
https://github.com/blutjens/awesome-forests

[44] SilvaScenes: Tree Segmentation and Species Classification from Under-Canopy Images in Natural Forests
https://arxiv.org/html/2510.09458v1