BAMspecies: A Benchmark Dataset for Individual Tree Species Classification from Ultra High-Resolution RGB Imagery and LiDAR Point Cloud

Mareike Weishaupt[1], Wen Fan[2], Youcef Ben Ghorbel[1], Daniel Panangian[1], Jonas Troles[3], Ute Schmid[3], and Jiaojiao Tian[1]*

[1] Remote Sensing Technology Institute, German Aerospace Center (DLR), 82234 Oberpfaffenhofen, Germany.
{mareike.weishaupt, jiaojiao.tian}@dlr.de
[2] Institute of Forest Management, TUM School of Life Sciences Weihenstephan, Technical University of Munich, 85354 Freising, Germany.
[3] University of Bamberg, Cognitive Systems Group, 96047 Bamberg, Germany

Abstract. In forestry individual tree species classification is particularly constrained by the need for expert annotations and high-resolution data capturing complex crown structures. While comprehensive datasets exist for general-purpose tasks, domain-specific applications requiring expert knowledge remain severely underrepresented. To address this gap, we present BAMspecies, a large-scale benchmark dataset for tree species classification in dense forests. BAMspecies contains approximately 25,000 manually annotated individual trees with expert-verified species labels, very high-resolution RGB orthomosaics with 2 cm GSD, and dense LiDAR point clouds with approximately 700 pts/m^2 from unmanned aerial vehicles (UAV) data. The dataset provides species labels, instance annotations, and corresponding 2D and 3D representations, enabling standardized evaluation of tree species recognition from RGB imagery and point clouds. To establish reference performance, we evaluate seven representative 2D models and two 3D models, achieving mean *F*1-*scores* from 47.18% to 59.45% for RGB imagery and from 26.23 % to 53.32 % for point clouds. These results highlight the difficulty of fine-grained species recognition in structurally complex forests. By combining expert annotations, high-resolution RGB data, dense LiDAR, and reproducible benchmark splits, BAMspecies enables systematic evaluation of tree species classification methods across 2D imagery and 3D point clouds. The dataset is available on the BAMspecies website.

Keywords: Tree species classification · Individual tree crown (ITC) · Ultra high-resolution RGB · LiDAR · Benchmark · Remote sensing

1 Introduction

Tree species classification at the individual tree level provides critical information for forest management and serves as a key parameter in forest inventory [3, 28]. Additionally, it influences the analysis of forest biodiversity and accurate quantification of aboveground biomass estimation [44]. With the advancement of very high-resolution remote sensing imagery and light detection and ranging (LiDAR), remote sensing techniques are becoming increasingly widespread [23]. Recent advances in machine learning and deep learning have further enhanced the ability to automatically extract detailed forest attributes from 2D and 3D datasets with improved accuracy and efficiency [11, 20, 42]. However, despite this progress, relatively few studies have successfully applied AI techniques to large-scale tree species classification, mainly due to the limited availability of high-quality labeled training data, required to develop robust and transferable models [49].

Unlike indoor object extraction, or urban features such as roads and buildings, which are distinct and easily identified by non-experts, identifying tree species from RGB imagery or point clouds requires expensive, specialized training for annotators [1]. Especially the high intraclass variability (trees of the same species looking different) and low interclass variance (different species looking similar) make this task more challenging. Therefore, the existing high-quality benchmarks for the classification of tree species are generally restricted to a small number of trees and limited geographical areas [12], or to only annotated points or simple box boundaries [2]. Existing large-scale benchmark data sets are fully or partly converted from

* Corresponding author: jiaojiao.tian@dlr.de

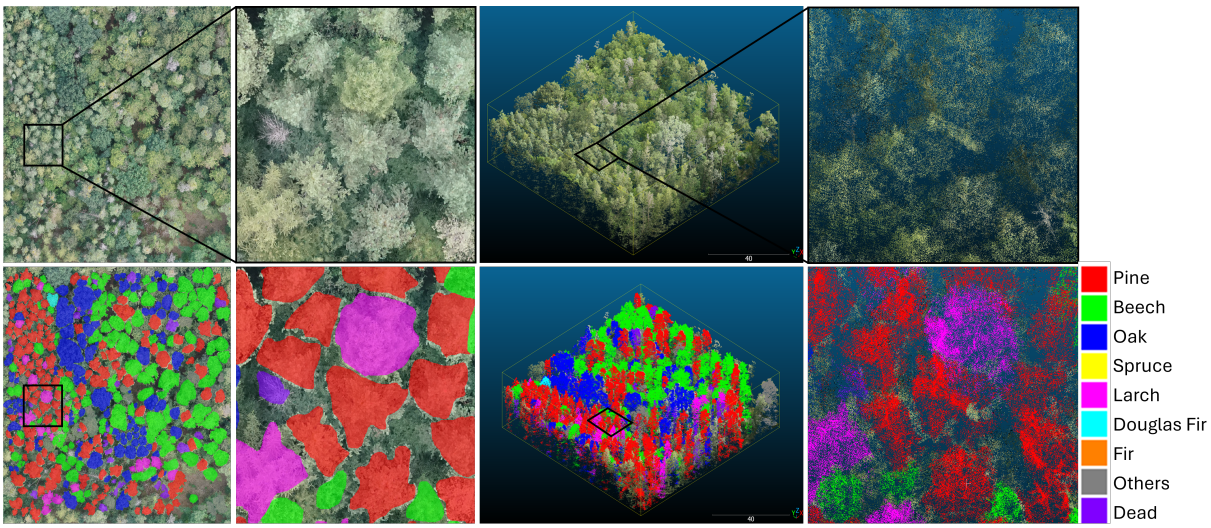


Fig. 1: Example plot from the dataset BAMspecies. The top row shows RGB imagery, while the bottom row displays the corresponding semantic labels. (a) Full 2D aerial plot. (b) Zoomed-in region of the 2D plots. (c) Corresponding 3D point cloud. (d) Zoomed-in view of the point cloud. Classes *Spruce* and *Fir* are not included in this selected plot (a)-(d).

forest inventory plots, and can not present the complex and overlapping crown structures in natural forests. Furthermore, the spatial resolution of the these benchmarks are often relatively low, limiting the level of structural detail that can be captured [1].

To bridge this gap, we introduce BAMspecies, a novel large-scale benchmark for tree species classification (Fig. 1). Our dataset consists of around 25,000 annotated individual trees, featuring co-registered RGB plots (at 2 cm GSD) and high-density LiDAR point clouds (at average 700 pts/m^2) for approximately 22,000 tree instances over a European natural forest. Additional statistics, computational costs, additional experimental qualitative and quantitative results are provided in the supplementary material.

This dataset presents a unique challenge for the computer vision community: learning from sparse, irregular 3D point sets and high-resolution 2D textures for semantic segmentation of similar objects.
Our contributions are summarized as follows:

- We present the largest fully manual annotated large-scale benchmark for detailed individual tree crown (ITC) level tree species classification with around 25,000 trees that integrates centimeter-level RGB imagery with high-density LiDAR point clouds, including seven tree species.
- We provide standardized and spatially aligned 2D image patches and 3D point clouds with expert annotations and predefined data splits. The published dataset is designed to facilitate future research across a wide range of learning tasks, including unimodal and multimodal approaches.
- We establish benchmark baselines by evaluating representative unimodal deep learning models for both RGB imagery and LiDAR point clouds, providing standardized reference results for future comparison of 2D and 3D tree species classification methods.

2 Related work

2.1 Deep Learning in Forest Monitoring

With the increasing availability of annotated remote sensing data, deep learning has become a central tool for forest monitoring, enabling accurate and scalable extraction of forest attributes from both imagery and LiDAR [23, 48]. Early work often focused on canopy height products, which can be derived from LiDAR or aerial photogrammetry and, more recently, also predicted from satellite imagery at large scales [4,5,11,21,37].

ITC segmentation is a key prerequisite for per-tree inventories and species recognition. Modern approaches increasingly replace heuristic pipelines with end-to-end instance segmentation in 2D (e.g. Mask R-CNN variants [16]) or bottom-up grouping and learning-based segmentation in 3D point clouds (e.g. PointGroup [19]). The emergence of larger benchmark datasets, such as NEON and FOR-instance, has substantially improved reproducibility and accelerated progress toward robust methods in structurally complex forests [29, 40, 45]. Tree species recognition has been studied at multiple levels of granularity, ranging from stand-level mapping to crown-level and individual-tree prediction. Studies show that performance can benefit from the combination of complementary sensing modalities, such as optical images and LiDAR [1, 25, 28].

2.2 Tree Species Classification Benchmarks

Public forest remote sensing datasets have grown rapidly, but benchmarks for individual-tree species recognition still vary widely in label granularity, sensing modality, and task formulation. Tab. 1 summarizes representative resources across optical (RGB, multispectral), and LiDAR modalities. At large spatial scales, TreeSatAI [1] provides multi-sensor satellite imagery for forest monitoring and species mapping. While valuable for landscape-level analyses, it operates at meter-scale resolution with pixel- or plot-level labels, and therefore does not represent individual crowns or within-canopy structure required for fine-grained tree-level recognition.

Some work frames species recognition as a detection problem. Beloiu et al. [2] detect and geolocate tree crowns from aerial RGB and jointly predict species using Faster R-CNN. This bounding-box formulation supports crown localization, but it does not provide dense crown masks and does not leverage 3D LiDAR structure. In the 3D domain, FOR-instance [29] and ForestFormer3D [45] establish benchmarks for semantic and instance segmentation of individual trees in point clouds in various types of forests. These datasets enable robust geometric delineation, but do not include species-level annotations.

Species-resolved point-cloud benchmarks have recently emerged. The benchmark FOR-species20K [28] provides >20,000 laser-scanned trees with species labels across multiple European forest types and standardized splits for individual-tree classification, but it relies on proximal scanning (TLS/MLS/ULS) and does not include co-registered centimeter-level aerial RGB imagery. Furthermore, FOR-species20k does not contain patches of natural forest areas, but point clouds of already segmented individual trees. In the boreal domain, BorFIT [33] offers manually segmented UAV–LiDAR plots with >16,000 trees across 14 taxa, but is primarily designed for feature-based classification and provides only partial RGB information. FORTRESS [32] includes very high resolution 2D RGB imagery with very high quality, semantically segmented, tree species labels and covers 51 hectares of independent one-hectare-sized research plots in the south German Schwarzwald.

Comparable datasets for species classification have at least one of the following shortcomings: They offer only one modality [7, 28, 32, 39]; are of much lower resolution [12, 41]; are much smaller [8, 36]; do only offer single trees and not the complete scene [28]; or only contain data of plantations instead of natural forests [22]. BAMspecies combines these traits by offering the combination of: 2D RGB data and overlapping complete LiDAR point clouds; a very high resolution of 2 cm GSD and high point density of 700pts/m^2; being one of the largest datasets with approximately 25,000 trees for RGB and 22,000 with matching LiDAR 3D point cloud; and by containing data of structurally complex forests instead of plantations or individual trees outside their natural context. Tab. 1 compares BAMspecies to the best datasets of each shortcoming-category.

3 BAMspecies Dataset

The BAMspecies dataset is related to the original BAMFORESTS dataset [38] with a special focus on tree species classification. In BAMFORESTS the 2D dataset, including ITC annotation, has already been published and several baseline methods have been developed and evaluated using this dataset [10]. BAMspecies additionally combines very high resolution orthomosaics with LiDAR derived point clouds of a total of three AOIs: Stadtwald, Tretzendorf-1, and Tretzendorf-2. The Hain AOI from the BAMFORESTS dataset is not included in BAMspecies, as EU regulations forbid the use of a larger UAV, capable of carrying LiDAR sensors. The Stadtwald AOI covers 152 hectares and is located approximately 5 kilometres southeast the city center of Bamberg. The area covered by the Tretzendorf-1 AOI is 65 hectares, while the Tretzendorf-2 AOI covers 47 hectares. In addition to the differences in data modality, in comparison to BAMFORESTS,

Table 1: Overview of tree species benchmark datasets. **Modality**: RGB (2D), MS (2D), LDR (LiDAR – 3D). **Task**: IS = instance segmentation, SS = semantic segmentation, Det = detection, CLS = classification. **Classes** counts each dataset's label space. **Trees** is a per-tree count only where the unit is a tree.

Dataset	Year	Modality	Task	Classes	Trees	RGB/MS res. (m)	LDR dens. (pts/m^2)
Quebec Trees [7]	2024	RGB	IS	14	23,000	0.02	–
FOR-species20K [28]	2025	LDR	CLS	33	20,158	–	variable
Dubrovin [8]	2024	RGB & LDR	Det	7	3,600	0.07	37
Lefebvre & Laliberté [22]	2024	RGB & LDR	IS	9	19,564	0.005	1,738
PureForest [12]	2025	MS & LDR	CLS	13	∼23.7 M	0.20	∼40
BAMspecies (ours)	**2026**	**RGB & LDR**	**IS/SS**	**10**	**25,000 22,000**	**0.02**	**∼700**

BAMspecies additionally contains semantic segmented tree species, together with the already published tree instances. All data is provided preprocessed as similar sized plot tiles with ITC IDs and corresponding classes: 0 *Background*, *Pine (Pinus)*, 2 *Beech (Fagus)*, 3 *Oak (Quercus)*, 4 *Spruce (Picea)*, 5 *Larch (Larix)*, 6 *Douglas Fir (Pseudotsuga)*, 7 *Fir (Abies)*, 8 *Others*, 9 *Dead*. The tree instances are published with individual IDs for each plot. In-depth information on pixel-class distributions can be found in Fig. 4. In few plots, the corresponding 3D point cloud data are unavailable. Nevertheless, we still publish the 2D RGB image together with their corresponding annotations for these cases. The exact name of these plots with missing 3D point cloud data are listed in Tab. 2.

3.1 Data Acquisition

For orthomosaics, the Quantum Systems Trinity F90+ UAV with a Sony RX1 RII sensor is used, and data acquisition took place in July 2022. All LiDAR data is acquired with a DJI Matrice 350 RTK quadcopter and a DJI Zenmuse L2 sensor in the month of July in 2024. All flights are conducted in terrain follow mode at 100 - 120 m above ground level. In RGB-flights, images are acquired with 70 % frontal overlap and 90 % sidelap. In LiDAR flights, the sidelap is set to 50 %. Besides, each point in the LiDAR-based point cloud contains the following information: x-coordinate, y-coordinate, z-coordinate, red, green, blue, reflectivity, return number, ground point class, tree species and instance ID.

3.2 Data Processing

The orthomosaics are processed with Agisoft Metashape using Structure from Motion (SfM) algorithms and followed typical processing steps with the aim of improving output quality and reducing processing artifacts. The LiDAR point clouds are processed using the DJI Terra software with the quality settings and point cloud density set to the highest possible levels. While original RGB-GSD is about 1.8 cm and slightly differed between AOIs, we resample all orthomosaics to an RGB-GSD of 2.0 cm.

3.3 Annotation

All labeling of the georeferenced orthomosaics is conducted by one senior forestry expert following a written protocol. Species and vitality labels are subsequently re-checked independently by two trained annotators with access to all plots and to the same field-verification option, with disagreements resolved on site or conservatively assigned to the class Others. Annotators are supported by the Bamberg forestry office, the Bamberg urban green spaces department and the University of Bamberg. For labeling, in addition to the July images, images from September are used as well as on-sight cross checking if needed. Annotators had access to the orthomosaics and the DCHM. Labeling instructions told them to roughly draw the outlines of tree crowns and to prefer slightly smaller crown boundaries to reduce noise from neighboring trees. For

species labels they were instructed to use the "Others" class, if in doubt and explicitly not to guess species or genus.

To validate labeling quality, a second completely independent annotator was commissioned to relabel two hectares in the Stadtwald AOI, one hectare in the Tretzendorf-1 AOI and one hectare in the Tretzendorf-2 AOI. The original labels are referred to as A and the relabeling labels are referred to as B. Over the four plots A labeled 961 tree crowns and B 931. Of those, 751 matched (match = $IoU > 0.5$) with an IoU_{mean} of 0.78. Agreement in crown shapes was higher, the bigger the tree crowns are. Figure 2 gives a comprehensive overview of two of the relabeling plots. All matching tree crowns are visualized in blue shades, tree crowns only labeled by A in orange and tree crowns only labeled by B in green. For genus labels, labeling-agreement

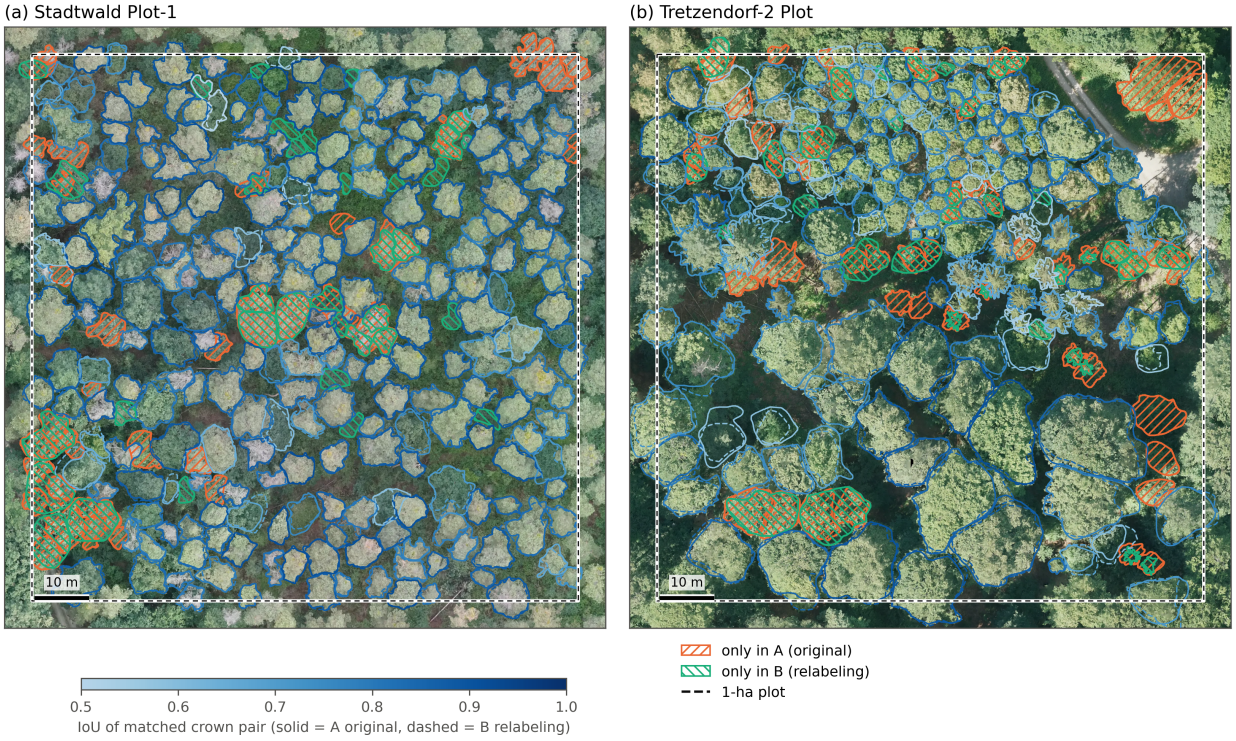


Fig. 2: Original labels (A) with solid outlines and relabeling labels (B) with dashed outlines of the relabeling plots.

is clearly substantial with Cohen's kappa $\kappa = 0.76$, but $F1$-*scores* range from 0.00 for the class Abies with eight matched instances and 1.00 for the class Pinus with 333 matched instances. Figure 3 shows the confusion matrix of the genus labels of all 751 matched tree crowns and lists all $F1$-*scores*.

For the 3D dataset, species labels and instance IDs are transferred from the manually annotated 2D crown polygons to the LiDAR point cloud by assigning each point the label of the polygon containing its horizontal position (x, y). This procedure transfers the expert species and instance annotations available in the 2D data to the corresponding LiDAR point cloud while preserving the spatial correspondence between the two modalities. Similar 2D-to-3D label propagation approaches have previously been used to exploit image-based annotations for point-cloud supervision [13, 27].

The resulting 3D labels should be interpreted as propagated annotations rather than independently annotated 3D ground truth. Since the propagation is performed in the horizontal (x, y) plane, the procedure cannot explicitly resolve vertical overlap between tree crowns or occluded crown structures. In dense forest stands, the top-view crown polygons therefore provide only an approximation of the three-dimensional extent of individual trees. A second source of uncertainty arises from the temporal difference between the two acquisitions: the RGB imagery used for the crown annotations was acquired in 2022, whereas the LiDAR data were acquired in 2024. Consequently, the forest scene is not temporally identical between the two modalities.

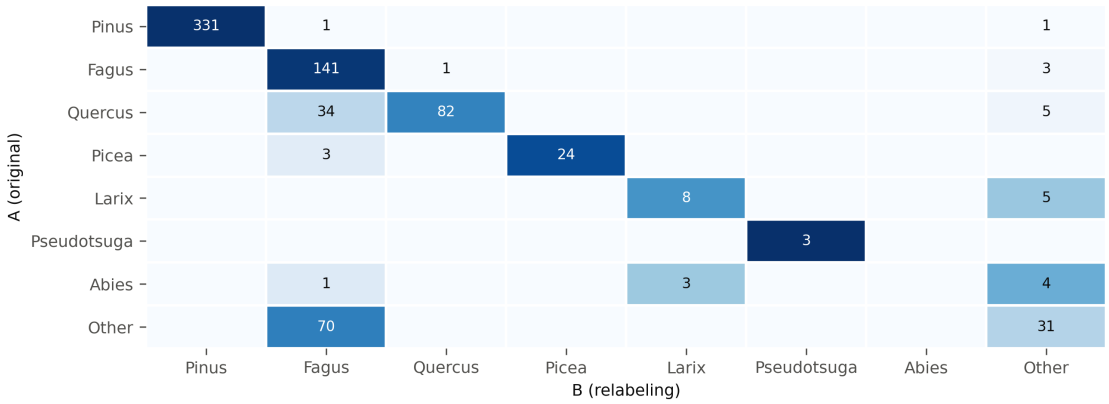


Fig. 3: Confusion matrix of the genus label of all matched tree crowns. F1-scores: Pinus (Pine) = 1.00, Fagus (Beech) = 0.71, Quercus (Oak) = 0.80, Picea (Spruce) = 0.94, Larix (Larch) = 0.67, Pseudotsuga (Douglas Fir) = 1.00, Abies (Fir) = 0.00, Other = 0.41.

Some trees visible in the 2D imagery are absent from the later LiDAR acquisition, including removals near forest edges in the Tretzendorf AOI. These cases cannot be resolved by the geometric propagation itself and contribute to the difference in the number of annotated trees, with 25,000 individual trees in the 2D data and 22,000 in the 3D data.

Despite these limitations, the propagated labels provide a scalable source of species-level supervision for the dense LiDAR point cloud. Manual point-wise annotation of such a high-density 3D dataset would be substantially more labor-intensive than annotating individual tree crowns in high-resolution imagery. The propagation therefore enables the construction of a large-scale 3D species-labeled benchmark while retaining the expert species information available from the 2D annotations. We explicitly consider these propagated labels as an approximation of 3D ground truth and recommend that users account for the associated spatial and temporal uncertainties when evaluating 3D models.

3.4 BAMspecies Classes

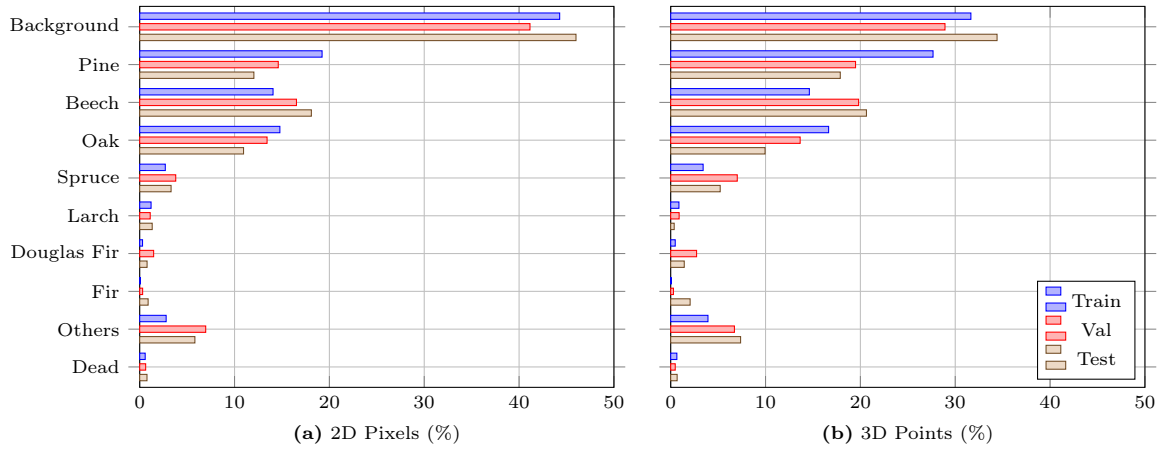
We divide the dataset into training (*train*), validation (*val*), and test (*test*) sets while preserving all classes and including both study regions, Stadtwald and Tretzendorf, in each split to reduce class and region bias. The dataset consists of 10 classes, as summarized in Fig. 4, which shows the distribution of 2D pixels and 3D points across the splits. BAMspecies exhibits a pronounced class imbalance: common species such as Pine, Beech, and Oak dominate, whereas Larch, Douglas Fir, Fir, and Dead are much less frequent. This reflects the natural species composition of the study area but increases the difficulty of learning reliable decision boundaries for rare species. For benchmarking, we evaluate only the tree species classes and the *Dead* class. The *Background* and *Others* classes are included during training but excluded from evaluation because background regions may contain tree-like structures, while *Others* represents a heterogeneous mixture of minority species with very low representation.

3.5 Data Publication and Usability

For the final data publication, geospatial information is removed from all data to ensure exact location protection of the plots. In the prior BAMFORESTS dataset [38] the area's geographic locations are presented, and can be seen fo the major areas around Bamberg. Since the prior published region "Hain" is not released, there is no geographically independent region for this dataset. All data plots are provided same sized for 2D and 3D, being around 100×100 m and non-overlapping. As data is published in plots it has to be tiled for 2D training by the users following the guidelines provided in the next section for both 512×512 and 1024×1024 patches. We publish four folders: Image plots, tree instances (vector data) and tree species (pixel data) masks and the matching LiDAR point cloud plots. We have assigned a synthetic georeferencing

Table 2: Available plot numbers of 2D and 3D across the dataset splits. Named plots denote samples for which only 2D data are available.

Split	3D	2D	Plots with no corresponding 3D data
Train	51	60	Area_47, Area_50, Area_51, Area_52, Area_56, Area_67, Area_74, Area_88, Area_89
Val	14	16	Area_82, Area_85
Test	11	13	Area_42, Area_55
Total	76	89	13



(a) 2D Pixels (%) **(b)** 3D Points (%)

Fig. 4: Class distribution across Train/Val/Test data for (a) image-based data and (b) point cloud data. The subfigures highlight differences in class frequency, because of the two different modality types.

information to ensure the matching of the 2D and 3D data. As some plots lack corresponding point cloud data due to the use of two separate image acquisition methods, only 76 out of 89 plots are available in both 2D RGB and 3D LiDAR point cloud, which is displayed in Tab. 2. As a further characteristic of the 3D data, Area_64 contains regions with incomplete point-cloud coverage, resulting in visible gaps within the plot. This should be considered when using this plot for 3D analysis or model evaluation.

3.6 2D and 3D Data Patch Preparation

For the 2D semantic training we balanced memory efficiency and spatial context with performing preliminary experiments with patch sizes of 256×256, 512×512, and 1024×1024 pixels. Based on performance, the 256×256 patch size was excluded from further experiments. Tiles are generated with a moving-window approach and 50 % horizontal and vertical overlap. No padding or edge handling is applied. In contrast to the 2D workflow, LiDAR data is processed in the original georeferenced coordinate system and prepared as per-point-labeled point clouds. We apply the same spatial split as for the 2D dataset to ensure geographic consistency across modalities and avoid leakage. For 3D, the plots serve as the basic storage and input units for the 3D baselines. All original point attributes are retained, while the number of points per tile varies naturally with canopy density and acquisition conditions.

4 Methodology

In this work, we establish modality-specific baselines by evaluating state-of-the-art models independently on 2D RGB images and 3D LiDAR point clouds. These experiments provide controlled reference points for assessing each modality and for future studies on cross-modal learning. The overall objective of this study is to conduct a through assessment of the quality and reliability of the proposed BAMspecies dataset.

4.1 Existing Baselines

Baselines for ITC segmentation from 2D RGB images [9,10] and ITC-based tree species classification [10,42] have been reported in our previous publications using training and testing splits that differ slightly from BAMspecies. These baseline experiments are not repeated in the present dataset paper. A baseline for ITC segmentation and species classification from 3D point clouds remain to be established.

Further future exploration will evaluate more up-to-date 3D backbones, such as Point Transformer V3 [43], alongside other recent models, using the same benchmark protocol to examine whether contextual modeling improves species recognition in structurally complex forests.

4.2 RGB Image Based Tree Species Classification

The models evaluated in this study encompass CNN-, transformer-, and Mamba-Vision-based feature extractors integrated within different semantic segmentation frameworks. CNN architectures are excellent at extracting local features. However, they struggle to model global dependencies adequately. Transformer architectures have strong capabilities for modeling global context. In contrast, self-attention is computationally intensive, leading to low efficiency when processing high-resolution images. Mamba architectures have low complexity and are efficient for long sequences. Nevertheless, they show clear limitations in understanding spatial structures and tend to produce artifacts in background regions. Given the pros and cons of the various deep learning architectures, seven methods from different development stages are used to evaluate datasets in the 2D domain. Besides, no class has a specific re-weighting, focal loss, or balanced sampling strategy, which ensures a controlled and reproducible benchmark comparison.

Fully Convolutional Networks (FCNs) replace fully connected layers with convolutional layers to enable end-to-end dense prediction [34]. U-Net extends this design with a symmetric encoder–decoder architecture and skip connections to better preserve fine-grained spatial information [31]. In contrast, SegFormer B1 employs a hierarchical Mix Transformer (MiT) encoder to extract multi-scale representations and a lightweight MLP decoder for multi-scale feature fusion [47]. DeepLabV3+ uses atrous convolution and Atrous Spatial Pyramid Pooling (ASPP) to model multi-scale context for semantic segmentation [6]. Based on that, U-Net and DINOv3 are used as a backbone [31,35]. U-Net provides precise spatial features while DINOv3 enhances global semantic representation, improving dense prediction performance. DINOv3 further employs large-scale self-supervised pretraining and high-resolution adaptation to strengthen its dense feature representations. UPerNet integrates a Pyramid Pooling Module (PPM) and a Feature Pyramid Network (FPN) to aggregate multi-scale features and capture both global context and hierarchical features [46]. In this work, we employ ResNet-50, Swin Transformer, and MambaVision backbones, enabling convolutional, self-attention-based, and state-space-enhanced global modeling within a unified framework [15,17,24].

4.3 3D Point Cloud Based Tree Species Classification

The recognition of 3D tree species is considered to be a dense, per-point labeling task on LiDAR tiles. Two widely adopted and complementary methodologies are selected to establish representative 3D baselines, capturing different facets of point cloud learning. One is a point-based method operating directly on sampled point sets, and the other is a voxel-based method leveraging a discretized 3D representation. PointNet is employed as a point-based baseline [30]. As a standard and efficient architecture for point clouds, it processes unordered point sets directly and provides a strong reference for per-point prediction. This baseline quantifies the performance achievable by a lightweight point-based model. PointNet combines point-wise features with a global pooled representation but does not explicitly model local neighborhoods. A sparse convolutional U-Net (SPConv) is used as a complementary baseline [14]. Unlike point-based approaches, this type of model is designed to capture local neighborhood structures and larger spatial context in 3D scenes, which is particularly beneficial in dense forest canopies. By trading some point-level detail for stronger spatial aggregation, it provides a context-aware reference for 3D tree species recognition.

Table 3: Hyperparameter settings for the 2D segmentation models. Abbreviations: DLV3+U (DeepLabV3+ with UNet); FCN+U (FCN with UNet); UPer-R50 (UPerNet with ResNet-50); UPer-Swin (UPerNet with Swin Transformer); UPer-MV (UPerNet with MambaVision); DLV3-DINOv3 (DeepLabV3 with DINOv3).

Model	Best Iteration		Loss	Weight	Optimizer	Augmen-	Batch	Auxiliary
	512	1024				tation	Size	Head
DLV3+U	72k	40k	CE	1.0	SGD	Flip	8	No
FCN+U	64k	40k	CE	1.0	SGD	Flip	8	No
SegFormer	440k	128k	CE	1.0	SGD	Flip	8	No
UPer-R50	56k	32k	CE	0.4	SGD	Flip	8	Yes
UPer-Swin	256k	72k	CE	0.4	SGD	Flip	8	Yes
UPer-MV	152k	148k	CE	0.4	AdamW	No	8	Yes
DLV3-DINOv3	58k	48k	CE	1.0	AdamW	No	8	No

5 Experiments and Results

5.1 Experimental Setup

Evaluation is performed by an element-wise comparison: pixel-wise for 2D masks and point-wise for 3D point clouds. We report Intersection over Union (IoU), mean class accuracy (mAcc), and the $F1$-*scores*. Metrics are computed over the classes of tree species including *Dead*. For class-wise analysis, we report per-class $F1$-*scores*.

The experiments for the 2D semantic segmentation are mainly performed on MMSegmentation [26] and with an own implementation for DeepLabV3 with DINOv3 backbone. All models are trained in both patch sizes. For all other parameters listed in Tab. 3, we adopt the default loss configurations provided by MMSegmentation for all baseline models and modified hyperparameters only for self-implemented architectures. We intentionally use the standard cross-entropy (CE) loss despite the severe class imbalance, as our goal is to provide simple, easily reproducible baselines. The number of training iterations is selected individually per model based on the validation performance. The model selection is performed by choosing the checkpoint with the highest mIoU, evaluated every 2,000 iterations. For reproducibility, all experiments are conducted with a fixed random seed of 42.

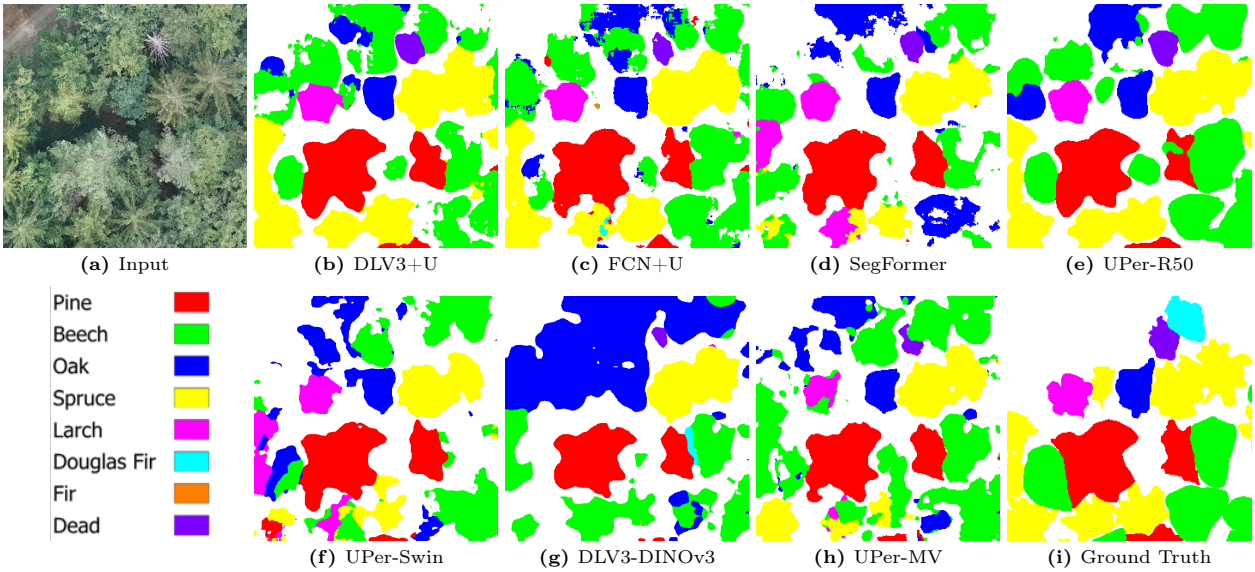
For the 3D experiments, we follow the standard training protocols of the original implementations and only adapt the data format to BAMspecies and the tree species classification task. Both 3D baselines use geometric input features. PointNet uses centered coordinates together with normalized coordinates, while SPConv uses voxelized coordinates augmented with a local verticality feature. For PointNet, we train on each tile by sampling 4096 points. We center the coordinates and append normalized coordinates as additional inputs. To mitigate density variations, we apply a light voxel downsampling with a voxel size of 0.2 m prior to sampling. During inference, we use overlapping windows and aggregate predictions via per-point voting. For the SPConv baseline, we adopt a sparse-convolution architecture, the tiling and sampling pipeline of TreeLearn [18] but adapt it to semantic species labeling by disabling the offset branch and training only the semantic head. Inputs are voxelized at 0.1 m, and we compute a verticality feature within a 0.6 m neighborhood. Training uses random 35 m $\times$ 35 m crops that are filtered by an occupancy criterion to avoid empty samples. At inference time, we follow the TreeLearn tiling and ensembling procedure with inner-region masking to reduce boundary effects, and we fill remaining unlabeled points by propagating predictions from the nearest labeled neighbors within a small radius (kNN-based hole filling) to obtain a complete per-point prediction. Both baselines are optimized with cross-entropy on per-point labels. We select the checkpoint with the best mIoU validation.

5.2 Results on RGB Images

The semantic segmentation results in (Tab. 4) show overall and per-class performance. Among them, UPer-R50 achieves the best overall performance across both evaluated patch sizes. The strong performance of the CNN-based methods suggests that convolutional feature extraction remains well suited for high-resolution

Table 4: Per-class F1-scores (%) for the seven tree species and the *Dead* class (top). Overall metrics are reported at the bottom. Each model is evaluated using patch sizes of 512 × 512 and 1024 × 1024. The best result per class and patch size is highlighted in bold.

	DLV3+U [6]		FCN+U [34]		SegFormer [47]		UPer-R50 [46]		UPer-Swin [46]		UPer-MV [46]		DLV3-DINOv3 [6]	
Class	512	1024	512	1024	512	1024	512	1024	512	1024	512	1024	512	1024
Pine	82.74	82.09	80.79	79.94	82.98	82.68	85.28	85.73	80.76	82.18	80.61	82.57	**87.03**	**86.26**
Beech	67.40	66.30	65.83	62.73	63.61	65.82	**72.11**	**73.72**	61.94	63.98	70.47	66.99	62.90	64.30
Oak	59.20	58.00	58.73	57.50	59.17	57.05	63.96	63.92	56.55	56.23	**67.06**	64.70	66.47	**65.09**
Spruce	74.96	73.45	74.30	73.54	65.08	68.09	**78.47**	**78.79**	66.37	64.45	52.22	68.56	68.39	71.23
Larch	80.84	77.66	**83.06**	78.56	35.63	62.15	82.16	**86.02**	50.63	46.77	45.51	23.30	58.95	43.33
Douglas Fir	18.48	09.35	14.75	12.30	07.41	06.79	21.06	17.64	07.05	02.68	00.02	00.27	**38.13**	**35.08**
Fir	00.00	00.00	01.50	**0.46**	00.00	00.00	00.97	00.25	00.00	00.00	00.00	00.02	**08.09**	0.00
Dead	64.64	66.41	65.32	66.52	66.24	64.81	67.51	69.55	59.42	63.01	**71.58**	**70.99**	66.35	62.71
mIoU	43.64	42.02	43.22	41.51	35.36	38.45	**46.90**	**48.00**	35.19	35.17	36.61	36.13	42.98	40.17
mAcc	57.05	55.67	53.66	54.27	46.91	51.56	**61.32**	**61.63**	46.71	46.71	41.53	41.24	56.68	53.15
mF1	56.03	54.16	55.54	53.94	47.52	50.93	**58.94**	**59.45**	47.84	47.41	48.43	47.18	57.04	53.50



(a) Input **(b)** DLV3+U **(c)** FCN+U **(d)** SegFormer **(e)** UPer-R50

(f) UPer-Swin **(g)** DLV3-DINOv3 **(h)** UPer-MV **(i)** Ground Truth

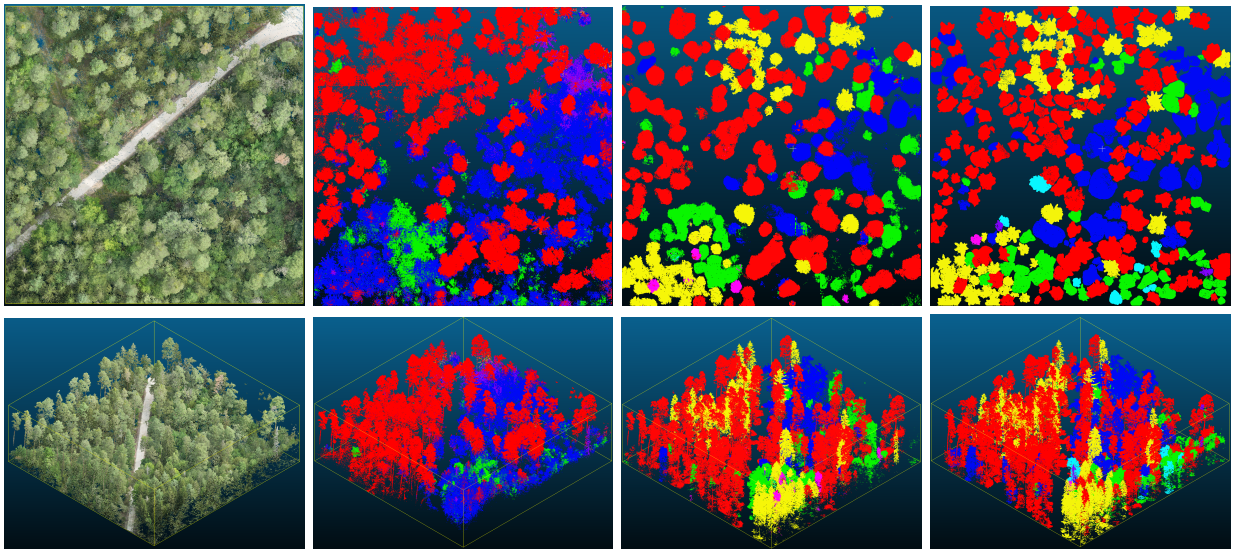
Fig. 5: Semantic segmentation results of one test patch (1024 × 1024) for each method.

aerial tree species segmentation, while the more recent transformer- and state-space-based backbones do not consistently exploit their additional modeling capacity on this dataset.

For a deeper analysis the per-class results give more training insights. For the dominant classes such as *Pine*, *Beech*, and *Oak* are reliably recognized by all models. Frequent classes achieve consistently strong and stable results across all models, while rare classes exhibit substantially larger performance variations. Especially *Douglas Fir* and *Fir*, remain challenging due to the limited representation in the training data. Among the evaluated methods, the DLV3-DINOv3 model shows a comparatively stronger performance on these underrepresented categories, suggesting the benefit of powerful pretrained feature representations. Interestingly, the *Dead* class achieves strong performance despite its relatively small share in the training data, likely due to its distinctive structural and spectral characteristics. A qualitative comparison in Fig. 5 highlights the differences in the spatial consistency of the predictions. UPer-R50, DLV3-DINOv3, and DLV3+U produce more coherent tree crowns with clearer boundaries, whereas models such as UPer-Swin, SegFormer, UPer-MV, and FCN+U tend to generate more fragmented predictions.

Table 5: Per-point metrics and per-species F1-score (in %) for 3D tree species semantic segmentation. Evaluation is performed on the intersection of tiles available for all file-based methods. Classes *Background* (0) and *Others* (8) are excluded. Best shown in bold.

	Per-point metrics (%)			Per-class F1 (%)							
Model	mIoU	mAcc	mF1	Pine	Beech	Oak	Spruce	Larch	Douglas Fir	Fir	Dead
PointNet [30]	17.88	22.46	26.23	65.43	68.19	25.51	24.07	4.92	2.76	0.00	18.93
SPConv [14]	**44.17**	**56.22**	**53.32**	**95.06**	**89.03**	**54.33**	**94.63**	**26.91**	**37.56**	**1.13**	**27.91**



(a) Input **(b)** PointNet **(c)** SPConv **(d)** Ground Truth

Fig. 6: Qualitative comparison of species segmentation results. The top row shows top-view projections, while the bottom row presents the corresponding 3D perspectives (The class colors follow the same legend as in Fig. 5).

5.3 Results on 3D Point Clouds

As presented in Tab. 5, SPConv yields substantially stronger overall performance than PointNet across all mean metrics on the test set. This improvement is also reflected at the class level: SPConv not only performs very well on the dominant species (*Pine* and *Beech*), but also provides markedly better recognition of several minority categories, leading to a more balanced error profile across the full label set. In particular, SPConv consistently recovers the rare conifer classes (*Larch* and *Douglas Fir*) where PointNet struggles, and it improves the discrimination of *Oak* compared to PointNet. At the same time, both methods remain challenging on *Fir*, indicating that this class is still difficult, also due to the strong class imbalance for *Fir* in this dataset. The qualitative results are shown in Fig. 6. For *Spruce*, PointNet demonstrates limited detection capability, whereas SPConv performs very well with a per-class F1 of 94.63 %. In the case of *Oak*, SPConv shows evident confusion with *Fir* in the lower-left area. For *Beech*, however, SPConv achieves superior performance, correctly identifying a larger number of individual trees.

6 Discussion and Conclusions

BAMspecies provides ultra high resolution orthomosaics RGB together with very dense UAV LiDAR point clouds for species-resolved tree mapping in a structurally complex mixed forest. It is the largest publicly released benchmark that jointly offers centimeter-level RGB imagery, high-density LiDAR, and dense semantic labels at scale for ITC level species analysis. With 78% matching tree crowns, a mean IoU of 0.78 of those matched tree crowns, and a Cohen's κ of 0.76, interrater reliability of our data set is good. Especially, when

considering the complexity of this domain. The UAV dataset is designed to support research on both 2D and 3D and multimodal approaches under realistic conditions and mapping, including strong class imbalance and high inter-species similarity. From an operational perspective, relying only on terrestrial LiDAR data restricts large-area forest mapping because of acquisition cost, occlusion, field accessibility, and restricted spatial coverage. UAV-based data acquisition enables a more scalable alternative. RGB imagery provides spectral and crown-surface information, while 3D point clouds capture tree height, vertical structure, and spatial organization. These two types of information are complementary for distinguishing tree species in structurally complex forests.

In the 2D setting, the benchmark behaves as a pronounced long-tail problem. Frequent classes achieve consistently strong results across architectures, while rare classes remain difficult and exhibit large variance between models. For example, *Fir* comprises only 0.09 % of the training data and consequently shows very poor performance across most architectures, whereas species representing more than 13 % of the training data consistently achieve strong and stable results. Classes with low pixel share but high semantic distinctiveness, such as *Larch*, can partially compensate for limited training data, although performance remains comparatively unstable. In general, the 2D results suggest that representation strength and pretraining are key factors for minority-class recognition on BAMspecies and this remains a big challenge for the CV community.

Compared to the RGB setting, LiDAR-only species segmentation shows a stronger dependence on the 3D backbone and a more pronounced long-tail behavior. PointNet remains competitive on the dominant classes (*Pine* and *Beech*) but struggles to recover minority conifers. SPConv improves the mean metrics and delivers a more balanced per-class performance (Tab. 5), with clear gains for *Spruce* and the rare conifer classes (*Larch* and *Douglas Fir*). In addition to the RGB case, also in the 3D domain the *Fir* classes remain challenging on both baselines, suggesting that this category is particularly hard to separate reliably from LiDAR geometry alone. Overall, the gap between PointNet and SPConv highlights the importance of neighborhood-based feature aggregation for 3D species segmentation on BAMspecies.

The baseline experiments evaluate representative 2D and 3D models for dense species-label predictions on BAMspecies under a consistent protocol. The aim is not to compare every recent forest point cloud or new semantic segmentation model exhaustively. Comprehensive comparisons of multimodal fusion architectures and 3D backbones are therefore beyond the scope of this dataset study. Although these results establish strong baselines and highlight the main challenges, BAMspecies still has a few practical limitations that should be considered when interpreting performance. In addition, the 3D point-wise annotations are derived from projected 2D labels rather than manual 3D annotation, they are subject to limitations. Moreover, the class taxonomy includes only seven species classes, while several minority species are grouped into the *Others* class, which limits fine-grained evaluation of the full local species diversity. It has to be noted, the temporal mismatch between RGB labeling data (2022) and LiDAR acquisition (2024) leads to systematic inconsistencies where trees were removed or changed. Nevertheless, these limitations also show as potential to further improve the data quality in the near future.

ITC level tree species classification is a complex and labor-intensive annotation task, particularly for natural dense forest. BAMspecies will serve as a valuable benchmark for developing and evaluating methods using both 2D RGB images and 3D LiDAR point clouds. More importantly, the baseline results suggest clear benchmark tracks for future work that directly target the observed long-tail and modality gaps. This enables BAMspecies to readily serve as a benchmark for multimodal fusion, cross-modal distillation, and instance-based processing workflows.

References

1. Ahlswede, S., Schulz, C., Gava, C., Helber, P., Bischke, B., Förster, M., Arias, F., Hees, J., Demir, B., Kleinschmit, B.: Treesatai benchmark archive: a multi-sensor, multi-label dataset for tree species classification in remote sensing. Earth System Science Data **15**(2), 681–695 (2023). https://doi.org/10.5194/essd-15-681-2023 1, 2, 3
2. Beloiu, M., Heinzmann, L., Rehush, N., Gessler, A., Griess, V.C.: Individual tree-crown detection and species identification in heterogeneous forests using aerial rgb imagery and deep learning. Remote Sensing **15**(5), 1463 (2023) 1, 3
3. BMEL: Ergebnisse der vierten Bundeswaldinventur (4 2024) 1

4. Brandt, M., Chave, J., Li, S., Fensholt, R., Ciais, P., Wigneron, J.P., Gieseke, F., Saatchi, S., Tucker, C., Igel, C.: High-resolution sensors and deep learning models for tree resource monitoring. Nature Reviews Electrical Engineering **2**(1), 13–26 (2025) 2
5. Carcereri, D., Rizzoli, P., Ienco, D., Bruzzone, L.: A deep learning framework for the estimation of forest height from bistatic tandem-x data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **16**, 8334–8352 (2023) 2
6. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV (2018) 8, 10
7. Cloutier, M., Germain, M., Laliberté, E.: Influence of temperate forest autumn leaf phenology on segmentation of tree species from uav imagery using deep learning. Remote Sensing of Environment **311**, 114283 (2024) 3, 4
8. Dubrovin, I., Fortin, C., Kedrov, A.: An open dataset for individual tree detection in uav lidar point clouds and rgb orthophotos in dense mixed forests. Scientific Reports **14**(1), 21938 (2024) 3, 4
9. Fan, W., Tian, J., Troles, J., Döllerer, M., Kindu, M., Knoke, T.: Comparing deep learning and mcwst approaches for individual tree crown segmentation. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **X-1-2024**, 67–73 (2024). https://doi.org/10.5194/isprs-annals-X-1-2024-67-2024, https://isprs-annals.copernicus.org/articles/X-1-2024/67/2024/ 8
10. Fan, W., Knoke, T., Troles, J., Tian, J.: Automatic tree-level based forest inventories retrieval via ultra-high resolution uav images and deep learning. ISPRS Journal of Photogrammetry and Remote Sensing **234**, 261–274 (2026). https://doi.org/10.1016/j.isprsjprs.2026.02.029, https://www.sciencedirect.com/science/article/pii/S0924271626000882 3, 8
11. Fogel, F., Perron, Y., Besic, N., Saint-André, L., Pellissier-Tanon, A., Schwartz, M., Boudras, T., Fayad, I., d'Aspremont, A., Landrieu, L., Ciais, P.: Open-canopy: Towards very high resolution forest monitoring. In: CVPR. pp. 1395–1406 (June 2025) 1, 2
12. Gaydon, C., Roche, F.: Pureforest: A large-scale aerial lidar and aerial imagery dataset for tree species classification in monospecific forests. In: Proceedings of the Winter Conference on Applications of Computer Vision (WACV). pp. 5895–5904 (February 2025) 1, 3, 4
13. Genova, K., Yin, X., Kundu, A., Pantofaru, C., Cole, F., Sud, A., Brewington, B., Shucker, B., Funkhouser, T.: Learning 3d semantic segmentation with only 2d image supervision. In: International Conference on 3D Vision (3DV) (2021) 5
14. Graham, B., Engelcke, M., van der Maaten, L.: 3d semantic segmentation with submanifold sparse convolutional networks. In: CVPR (2018) 8, 11
15. Hatamizadeh, A., Kautz, J.: Mambavision: A hybrid mamba-transformer vision backbone. In: CVPR. pp. 25261–25270 (2025) 8
16. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: ICCV (2017) 3
17. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 8
18. Henrich, J., van Delden, J., Seidel, D., Kneib, T., Ecker, A.S.: Treelearn: A deep learning method for segmenting individual trees from ground-based lidar forest point clouds. Ecological Informatics **84**, 102888 (2024) 9
19. Jiang, L., et al.: Pointgroup: Dual-set point grouping for 3d instance segmentation. In: CVPR (2020) 3
20. Kulicki, M., Cabo, C., Trzciński, T., Będkowski, J., Stereńczak, K.: Artificial intelligence and terrestrial point clouds for forest monitoring. Current Forestry Reports **11**(1), 5 (2024) 1
21. Lang, N., Jetz, W., Schindler, K., Wegner, J.D.: A high-resolution canopy height model of the earth. Nature Ecology & Evolution **7**(11), 1778–1789 (2023) 2
22. Lefebvre, I., Laliberté, E.: Uav lidar, uav imagery, tree segmentations and ground mesurements for estimating tree biomass in canadian (quebec) plantations (2024). https://doi.org/10.20383/103.0979 3, 4
23. Liang, X., Kukko, A., Balenović, I., Saarinen, N., Junttila, S., Kankare, V., Holopainen, M., Mokroš, M., Surovỳ, P., Kaartinen, H., et al.: Close-range remote sensing of forests: The state of the art, challenges, and opportunities for systems and data acquisitions. IEEE Geosci. Remote Sens. Mag. **10**(3), 32–71 (2022) 1, 2
24. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: ICCV (2021) 8
25. Ma, Y., Zhao, Y., Im, J., Zhao, Y., Zhen, Z.: A deep-learning-based tree species classification for natural secondary forests using unmanned aerial vehicle hyperspectral images and lidar. Ecological Indicators **159**, 111608 (2024) 3
26. OpenMMLab Contributors: MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation (2020) 9
27. Pellis, E., Murtiyoso, A., Masiero, A., Tucci, G., Betti, M., Grussenmeyer, P.: 2d to 3d label propagation for the semantic segmentation of heritage building point clouds. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **XLIII-B2-2022**, 861–867 (2022). https://doi.org/10.5194/isprs-archives-XLIII-B2-2022-861-2022 5

28. Puliti, S., Lines, E.R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M.J., Winiwarter, L., Rehush, N., Hristova, H., et al.: Benchmarking tree species classification from proximally sensed laser scanning data: Introducing the for-species20k dataset. Methods in Ecology and Evolution **16**(4), 801–818 (2025) 1, 3, 4
29. Puliti, S., Pearse, G., Surovỳ, P., Wallace, L., Hollaus, M., Wielgosz, M., Astrup, R.: For-instance: a uav laser scanning benchmark dataset for semantic and instance segmentation of individual trees. arXiv preprint arXiv:2309.01279 (2023) 3
30. Qi, C.R., Su, H., Mo, K., Guibas, L.J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In: CVPR (2017) 8, 11
31. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 8
32. Schiefer, F., Frey, J., Kattenborn, T., Schmid, T., Hoffmann, J.: FORTRESS (2022). https://doi.org/10.35097/538, dataset 3
33. Schladebach, J., Heim, B., Enguehard, L., Wieczorek, M., Broers, J., Jackisch, R., Gloy, J., Hao, K., Tretton, J., Gorshunova, A., et al.: Borfit: A novel lidar-based training dataset for individual tree segmentation and species detection in northern boreal forests. Earth System Science Data Discussions **2025**, 1–31 (2025) 3
34. Shelhamer, E., Long, J., Darrell, T.: Fully convolutional networks for semantic segmentation. IEEE TPAMI **39**(4), 640–651 (2017). https://doi.org/10.1109/TPAMI.2016.2572683 8, 10
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: DINOv3 (2025), https://arxiv.org/abs/2508.10104 8
36. Takeshige, R., Htoo, K.K., Onishi, M., Rahman, F.M., Hoshizaki, K., Ida, H., Ishihara, M.I., Itoh, A., Kaneko, T., Katayama, A., et al.: High-resolution digital canopy height models, terrain models, ortho-mosaic photos, and canopy tree crown shapes derived from uav-borne lidar at 22 tree census plots across japanese natural forests. Ecological Research **40**(4), 657–670 (2025) 3
37. Tian, J., Schneider, T., Straub, C., Kugler, F., Reinartz, P.: Exploring digital surface models from nine different sensors for forest monitoring and change detection. Remote Sensing **9**(3), 287 (2017) 2
38. Troles, J., Schmid, U., Fan, W., Tian, J.: Bamforests: Bamberg benchmark forest dataset of individual tree crowns in very-high-resolution uav images. Remote Sensing **16**(11) (2024). https://doi.org/10.3390/rs16111935, https://www.mdpi.com/2072-4292/16/11/1935 3, 6
39. Vasquez, V., Cushman, K., Ramos, P., Williamson, C., Villareal, P., Correa, L.F.G., Muller-Landau, H.C.: Barro Colorado Island 50-ha plot crown maps: manually segmented and instance segmented. (12 2023). https://doi.org/10.25573/data.24784053.v2, https://smithsonian.figshare.com/articles/dataset/Barro_Colorado_Island_50-ha_plot_crown_maps_manually_segmented_and_instance_segmented_/24784053 3
40. Weinstein, B.G., Marconi, S., Bohlman, S.A., Zare, A., Singh, A., Graves, S.J., White, E.P.: A remote sensing derived data set of 100 million individual tree crowns for the national ecological observatory network. Elife **10**, e62922 (2021) 3
41. Weinstein, B.G., Marconi, S., Zare, A., Bohlman, S.A., Singh, A., Graves, S.J., Magee, L., Johnson, D.J., Record, S., Rubio, V.E., et al.: Individual canopy tree species maps for the national ecological observatory network. PLoS Biology **22**(7), e3002700 (2024) 3
42. Weishaupt, M., Fan, W., Troles, J., Tian, J.: Individual tree crown based tree species classification from very high-resolution uav images. In: IGARSS 2025 - 2025 IEEE International Geoscience and Remote Sensing Symposium. pp. 7328–7332 (2025). https://doi.org/10.1109/IGARSS55030.2025.11243070 1, 8
43. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler, faster, stronger. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4840–4851. IEEE (2024) 8
44. Wu, Z., Liu, X., Cheng, S., Yang, C., Wang, Z., Liu, Y., Dong, L., Li, F., Hao, Y.: Evaluating the effectiveness of forest type stratification for aboveground biomass inference. International Journal of Applied Earth Observation and Geoinformation **143**, 104829 (2025) 1
45. Xiang, B., Wielgosz, M., Puliti, S., Král, K., Krůček, M., Missarov, A., Astrup, R.: Forestformer3d: A unified framework for end-to-end segmentation of forest lidar 3d point clouds. In: ICCV. pp. 24717–24727 (October 2025) 3
46. Xiao, T., Liu, Y., Zhou, B., Jiang, Y., Sun, J.: Unified perceptual parsing for scene understanding. In: ECCV. pp. 418–434 (2018) 8, 10
47. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. In: NeurIPS (2021) 8, 10
48. Yun, T., Li, J., Ma, L., Zhou, J., Wang, R., Eichhorn, M.P., Zhang, H.: Status, advancements and prospects of deep learning methods applied in forest studies. International Journal of Applied Earth Observation and Geoinformation **131**, 103938 (2024) 2

49. Zhong, L., Dai, Z., Fang, P., Cao, Y., Wang, L.: A review: Tree species classification based on remote sensing data and classic deep learning-based methods. Forests **15**(5), 852 (2024) 1