Improving Tree Species Detection for Operational Forestry: The Role of Dataset Design

Mirela Beloiu[1], Khatereh Meshkini[1], Zhongyu Xia[1], Severin Broch[2], Verena C. Griess[1]

[1] Department of Environmental Systems Science, Institute of Terrestrial Ecosystems, ETH Zurich, 8092 Zurich, Switzerland
[2]Department of Mechanical and Process Engineering, ETH Zurich, 8092 Zurich, Switzerland
mirela.beloiu@usys.ethz.ch, khatereh.meshkini@usys.ethz.ch, zhongyu.xia@usys.ethz.ch, brochs@ethz.ch, verena.griess@usys.ethz.ch

Keywords: Forest monitoring, High-resolution aerial imagery, Dataset balancing, Object detection.

Abstract

Accurate tree species mapping from high-resolution aerial imagery is essential for forest monitoring, yet deep learning approaches remain highly sensitive to training data design and often fail to generalize from curated datasets to operational conditions. This study aims to evaluate how dataset composition and model configuration influence species-level tree detection. Using 10 cm RGB imagery and 18,379 partially labeled annotations across 10 classes, where only a subset of tree crowns is annotated per image, three dataset configurations are compared: unbalanced masked, balanced masked, and mixed masked–unmasked imagery. A YOLO-based object detector is trained with optimized hyperparameters, and an extended architecture incorporating a channel attention mechanism is assessed. Results show that class balancing the dataset yields the highest validation performance (F1 = 0.83, mAP_{50} = 0.87), while mixed datasets reduce performance (F1 = 0.56, mAP_{50} = 0.56). The attention mechanism partially improves results (F1 = 0.60, mAP_{50} = 0.62), particularly for conifer species and dead trees. However, evaluation on full aerial mosaics reveals contrasting behavior: masked models tend to over-detect with overlapping predictions, whereas models trained on unmasked or mixed data exhibit strong omission errors. These findings highlight a trade-off between validation performance and operational behavior, and demonstrate that partially labeled training data limit generalization. For practical applications, fully labeled datasets and evaluation on complete aerial images are critical for reliable and transferable tree species detection.

1. Introduction

Accurate individual tree mapping is important to forest inventory and monitoring, providing insights necessary for maintaining ecosystem resilience and guiding effective forest management practices (Beloiu et al., 2023; Holzinger et al., 2024). For instance, tree-level mapping of *Picea abies* is important for early detection of pest infestations, bark beetles, and understanding species-specific responses to environmental stressors such as drought or climatic fluctuations (Ecke et al., 2022; Pleșoianu et al., 2020; Safonova et al., 2019).

While traditional forest inventories are typically labor-intensive, costly, and spatially limited, remote sensing technologies offer scalable solutions, allowing extensive and multi-temporal forest monitoring. Remote sensing platforms such as uncrewed aerial vehicles (UAVs) have become popular in forest inventory work given their advantages of cost-effectiveness and operational flexibility, and their capacity to provide very-high resolution data. However, individual tree mapping in dense canopies remains technically challenging due to high spectral similarity between species, varying illumination geometries, and the displacement between the visible crown and the geolocated stem (Kattenborn et al., 2021; Liang et al., 2026; Xia et al., 2026).

Recent advancements in deep learning have transformed the analysis of remote sensing data by allowing for more automated and scalable object recognition. Specifically, Convolutional Neural Networks (CNNs) have been effective in extracting complex and hierarchical features and patterns directly from imagery data, fulfilling computer vision tasks such as image classification, object detection, and semantic, instance, and panoptic segmentation (Kattenborn et al., 2021). Despite the high potential of CNNs, they demand large and diverse datasets to achieve good generalization. For applications where high-quality labeled datasets are rare, transfer learning is often applied. This means models are pretrained on large public datasets and further finetuned on specific datasets and are therefore able to process large and complex datasets typically used for forestry applications (Ecke et al., 2024).

High-resolution aerial imagery and CNN models have enabled automated recognition of trees and vegetation patterns at unprecedented scales. One-stage object detectors can simultaneously localize and classify tree crowns in a single forward pass, making them suitable for large-scale forest monitoring. Nevertheless, CNNs remain sensitive to the characteristics of the training data. Class imbalance, masking strategies, and domain shifts between curated training datasets strongly influence generalization, raising questions about how datasets should be constructed for operational use. Several studies have demonstrated individual tree species classification using RGB, hyperspectral, or LiDAR data, often employing object detection or segmentation approaches (Xia et al., 2026). While these works highlight the promise of deep learning for species identification, they frequently optimize on single sites or simplified, masked tiles, rarely testing robustness on unprocessed imagery.

This study aims to evaluate the impact of dataset composition and hyperparameter configurations on species-level tree detection performance in high-resolution aerial imagery. We address the following research questions: RQ1: How do hyperparameter tuning and class balancing affect detection performance on masked datasets? RQ2: To what extent do models trained on masked tiles generalize to unmasked scenes, and which classes are most sensitive to domain shift? RQ3: Which dataset design choices lead to the most consistent and transferable detection performance?

Based on prior findings, we hypothesize that class balancing improves performance for underrepresented classes, while incorporating unmasked imagery introduces background complexity that leads to trade-offs between detection sensitivity and robustness.

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-413-2026

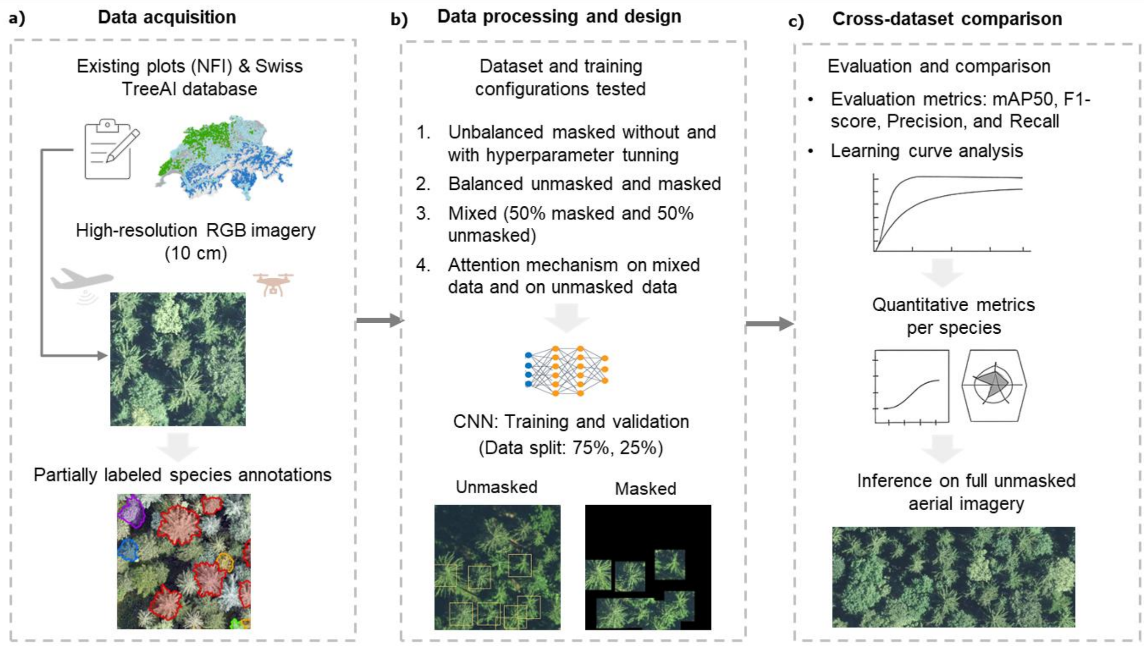

Figure 1. Workflow for evaluating the impact of dataset design on tree species detection accuracy: (a) RGB aerial imagery and species annotations, (b) four dataset configurations, and (c) performance evaluation with full-mosaic inference

2. Materials and Methods

This study evaluates how dataset design influences deep learning performance in tree species detection. High-resolution RGB aerial imagery (10 cm) and corresponding species annotations were obtained from the Swiss TreeAI database and forest inventory plots. Four dataset configurations were constructed: an unbalanced unmasked dataset, an unbalanced masked dataset, a balanced masked dataset, and a mixed masked–unmasked dataset. All models were trained, validated, and tested using 320 × 320 px image tiles. Performance was assessed using precision, recall, F1-score, and mAP_{50}. To evaluate generalization capability, inference was additionally performed on full aerial mosaics. Figure 1 illustrates the overall workflow of this process.

2.1 Data

The dataset used in this study consists of high-resolution RGB aerial image tiles covering selected forest areas in Switzerland, where annotated inventory data of individual trees are available (Figure 2). The imagery was sourced from publicly accessible repositories, including Swisstopo, with a spatial resolution of approximately 10 cm. The primary dataset, *5_RGB_S_320_pL*, originates from the TreeAI Global Initiative database (Beloiu Schwenke et al., 2025c) and contains 320 × 320 px RGB raster tiles. The dataset is partially labelled, meaning that not all tree crowns are annotated on each image tile. In particular, over 78% of all image tiles have fewer than two crowns annotated, indicating that this dataset is sparsely labelled. Annotations are provided in YOLO format, where each line defines the class label, normalized center coordinates (x, y), and bounding box width and height relative to the image dimensions.

Three dataset configurations were tested: masked unbalanced, masked balanced, and mixed balanced. The masked unbalanced version retained the original class distribution and used masked imagery, where only annotated tree crowns remained visible while the background was blacked out, allowing the model to focus solely on labeled objects. The original dataset contained 15 classes but was highly imbalanced, with some species exceeding 500 annotations across the training and validation sets, while others had fewer than 50 samples in the training set. To mitigate this imbalance, a rebalanced version was created, referred to as masked balanced. In this version, the number of dead tree instances was reduced by 50%, and similar classes, such as various *Tilia* and *Acer* species, were merged into aggregated classes (*Tilia* spp. and *Acer* spp.). Rarely represented classes, including *Fraxinus excelsior* and *Ulmus glabra*, were excluded from the balanced dataset.

In addition, a mixed balanced dataset was created by combining masked and unmasked tiles in a 50:50 ratio. This dataset was designed to evaluate model performance under more realistic conditions, where the model is simultaneously exposed to both processed (masked) and original aerial imagery. The hybrid setup enables analysis of dataset design strategies and comparison of model generalization on unprocessed imagery. Specifically, it was constructed by replacing half of the masked tiles with their corresponding unmasked versions from the TreeAI dataset, ensuring an identical species distribution across datasets for fair performance comparison. All datasets were divided into approximately 75% training and 25% validation subsets, ensuring spatial independence between the two (Table 1).

Class	Species/class/genus	Trainings labels	Validation labels
0	*Picea abies*	2285	672
1	*Pinus sylvestris*	636	180
2	*Larix decidua*	652	192
3	*Fagus sylvatica*	1435	338
4	Dead tree	5425	1623
5	*Abies alba*	1390	597
6	*Pseudotsuga menziesii*	441	68
7	*Tilia* spp.	275	86
8	*Acer* spp.	787	310
9	*Quercus* spp.	747	240
Total		**14073**	**4306**

Table 1. Dataset distribution after relabeling and balancing.

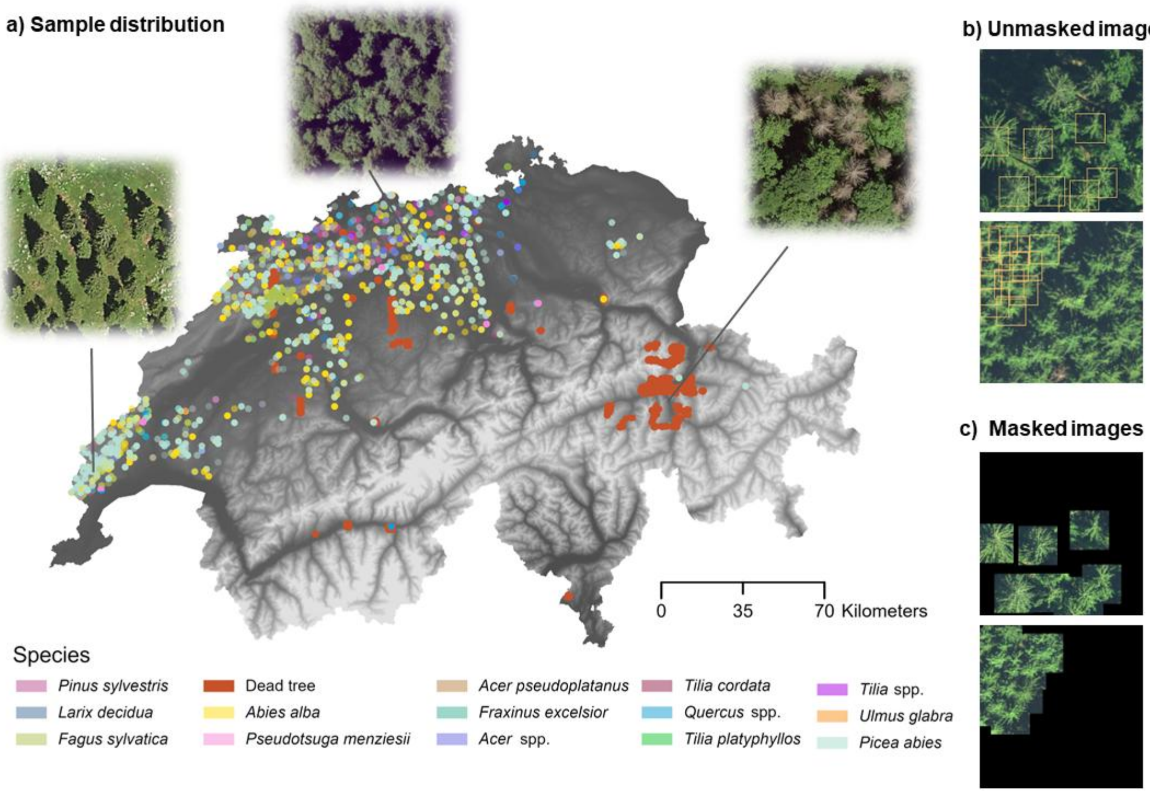

Figure 2. (a) The distribution of ground truth samples. (b) Examples of the unmasked images with partially annotated tree canopies, i.e. only a few trees are annotated in the entire image, where annotated trees are outlined in yellow bounding boxes. (c) Examples of masked images, where unannotated trees are blacked out.

2.2 Model Architecture and Training

The model used in this study was the YOLO11 object detection framework (Ultralytics, 2024). YOLO (You Only Look Once) is a real-time, open-source detector optimized for speed and accuracy, making it suitable for high-resolution aerial imagery analysis.

Our YOLO11 is composed of three main parts: the backbone, neck, and head (see Figure 3). The backbone extracts features from input images, using modules like Conv, a fast version of CSP (Cross Stage Partial Network) bottleneck being C2f, and SPPF to progressively reduce spatial dimensions while increasing channels, producing rich feature maps that capture both low- and high-level information. For some experiments, we modified the C2f module by incorporating a channel attention module into the architecture to enhance the networks ability to focus on the most informative feature channels, improving feature representation and ultimately boosting detection performance, especially for small or hard-to-detect objects. The neck, implemented as a PAN (Path Aggregation Network), sits between the backbone and head to further refine these features by aggregating information across different levels, ensuring more discriminative representations for detection. The head network then performs the final object detection using one detection head generating predictions that include class probabilities, confidence scores, and bounding box coordinates.

Initial experiments were conducted with the lightweight YOLO11n architecture to establish baseline parameters and optimize the training workflow under limited computational resources. Once the workflow was stable, the larger YOLO11x model was trained to maximize detection accuracy. As the dataset contained multiple labeled tree species, a multiclass classification setup was used to enable species-level crown detection. Model training was performed for 300 epochs using standard settings (see Table 2).

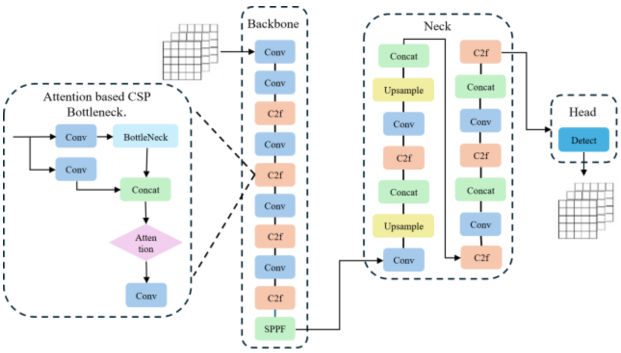

Figure 3. Architecture of the YOLO model with attention mechanism.

Hyperparameter optimization was carried out using Ultralytics' built-in function, which applies a genetic algorithm to search for optimal configurations. The tuning focused on key learning, augmentation, and optimization parameters (e.g., learning rate, momentum, color augmentation, scaling, and translation). Early tuning runs were performed on YOLO11n to accelerate the search, followed by retraining of the YOLO11x model using optimized parameters.

The final trained model was evaluated on the validation dataset using standard object detection metrics (precision, recall, F1-score, mean Average Precision at Intersect over Union (IoU) = 0.5 [mAP_{50}]) and qualitative visual analysis. For the unmasked dataset, predictions that do not overlap with any ground-truth labels are removed by applying a minimum overlap threshold of 0.1.

Argument	Training	Tuning
Model	YOLO11x.pt	YOLO11n.pt
Epochs	300	30, 50
batch	32	32
imgsz	320	320
device	0	0
workers	16	8
optimizer	SGD	SGD
Iou	0.5	0.5
amp	False	False
box	0.05	0.05

Table 2. Baseline parameters for training and tuning of YOLO11

2.3 Evaluation of Training Runs

The performance of the trained YOLO11 models was evaluated using established object detection metrics that provide insight into different aspects of detection quality. These metrics were computed directly by the model itself based on the validation dataset.

The primary evaluation metric in this project is the mean Average Precision at IoU = 0.5 (mAP50), defined in Formula (1). It measures the average precision across all classes using a fixed IoU threshold of 0.5 to determine true positives, which is commonly adopted in remote sensing applications due to annotation uncertainty and overlapping canopies.

Recall (Formula (2)) quantifies the proportion of actual trees that are correctly detected, while precision (Formula (3)) measures the proportion of predicted detections that correspond to true trees. The Intersection over Union (IoU), defined in Formula (4), is used to determine whether a detection is considered correct, with a threshold of 0.5 applied throughout this study.

Finally, the F1-score (Formula (5)) combines precision and recall into a single performance measure and is particularly useful in imbalanced detection scenarios, providing a balanced assessment of model performance across varying forest structures.

$$mAP50 = \frac{1}{N}\sum_{i=1}^{N} AP_{i\ (IoU=0.5)} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{4}$$

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5}$$

2.4 Hyperparameter Tuning

The YOLO11 model was optimized using the built-in model.tune() function from Ultralytics, which applies a genetic algorithm to search for improved hyperparameter configurations. In this process, each generation introduces random variations to the parameter set and selects high-performing configurations based on model performance.

Due to computational constraints, tuning runs were limited to 30–50 epochs per iteration and were conducted using the lightweight yolo11n.pt model to accelerate the search process. Initially, a relatively broad hyperparameter space was defined based on prior experimentation and the default recommendations provided by Ultralytics. This first tuning phase included optimizer parameters and several augmentation settings in order to explore a wide range of potential configurations.

Following this initial search, the tuning results were analyzed using experiment tracking in Weights & Biases. The analysis helped identify which hyperparameters had the strongest influence on model performance. Based on this evaluation, the parameter space was reduced in a second tuning stage, focusing only on the most relevant hyperparameters. This allowed the search process to concentrate on parameters that showed the highest sensitivity to performance improvements.

In the final tuning phase, the parameter space was expanded again to include additional augmentation and optimization parameters such as shear, scaling, vertical flipping, copy–paste augmentation, and warmup scheduling. This extended search space allowed the tuning algorithm to explore a broader range of training dynamics while still focusing on the most influential parameters identified in earlier stages. Complete hyperparameter search space used in this final tuning stage is shown in Table 3. Parameters to tune and ranges.

Hyperparameter	Range
Third tuning	
lr0	1e-5, 1e-1
momentum	0.83, 0.95
hsv_h	0.01, 0.015
hsv_s	0.3, 0.9
translate	0.0, 0.3
cls	0.2, 0.6
dfl	0.8, 1.5
weight_decay	0.0001, 0.001
lrf	0.01, 0.1
scale	0.4, 1.2
shear	0.0, 0.2
fliplr	0.0, 0.5
mosaic	0.0, 1.0
mixup	0.0, 1.0
warmup_epochs	0.0, 5.0
warmup_momentum	0.0, 0.95
hsv_v	0.0, 0.9
degrees	0.0, 45.0
perspective	0.0, 0.001
flipud	0.0, 1.0
copy_paste	0.0, 1.0

Table 3. Parameters to tune and ranges

3. Results

Across the dataset configurations, performance ranged widely, with the unmasked setup performing worst (F1 ≈ 0.27–0.29), while masked training clearly dominated (F1 up to 0.83–0.87); mixed and attention-based approaches showed intermediate gains but did not surpass the fully masked configuration (Table 4).

3.1 Effect of Hyperparameter Tuning on the Unbalanced Masked Dataset

Initial training on the unbalanced masked dataset (300 epochs) yielded an overall F1 score of 0.76 and mAP50 of 0.75. Detection of Picea abies was relatively strong with F1 0.86 and mAP50 0.90, and conifer species such as Larix decidua and Pinus sylvestris also showed high detection quality In contrast, several. broadleaved species, for example Fagus sylvatica, were classified less accurately

https://doi.org/10.5194/isprs-archives-XLIX-B1-2026-413-2026 | © Author(s) 2026. CC BY 4.0 License.

Class	UnM	Masked	Mixed	Mixed + att.	UnM + att.
Macro average	0.27	0.82	0.56	0.55	0.55
Picea abies	0.16	0.87	0.53	0.53	0.50
Pinus sylvestris	0.16	0.82	0.51	0.53	0.54
Larix decidua	0.40	0.86	0.63	0.63	0.67
Fagus sylvatica	0.19	0.83	0.44	0.44	0.42
Dead tree	0.52	0.87	0.64	0.63	0.78
Abies alba	0.18	0.87	0.57	0.54	0.52
Pseudotsuga menziesii	0.22	0.76	0.51	0.42	0.41
Tilia spp.	0.47	0.90	0.73	0.74	0.63
Acer spp.	0.25	0.60	0.50	0.47	0.51
Quercus spp.	0.17	0.81	0.54	0.48	0.41

Table 4. Comparison of species-level F1-scores for masked, mixed, and mixed with attention mechanism configurations. Unmasked = UnM and attention mechanism = att.

After hyperparameter tuning (224 epochs), overall performance increased to F1 0.80 and mAP50 0.86, corresponding to a gain of +0.043 in F1. Improvements were observed for most species. *Picea abies* increased to F1 0.88 and mAP50 0.91, reflecting small but consistent gains in both precision and recall. Additional positive responses to tuning were found for *Larix decidua*, *Abies alba*, and several broadleaved taxa. In contrast, *Pinus sylvestris* remained unchanged, and the dead tree class showed a slight decrease (−0.009).

3.2 Effect of Class Balancing for the Masked Dataset

Training on the balanced masked dataset yielded the highest performance overall (Figure 4 and Figure 5). The F1-score reached 0.83, and mAP50 increased to 0.87. For *Picea abies*, the F1-score was 0.90 and mAP50 0.93, the best result across all experiments. Similarly, *Abies alba* achieved strong results (F1-score 0.91, mAP50 0.94). Balanced sampling reduced class confusion and improved minority class accuracy without sacrificing the performance of dominant species. Misclassification is generally low, with most species showing high diagonal dominance, but notable confusion occurs between ecologically or spectrally similar classes.

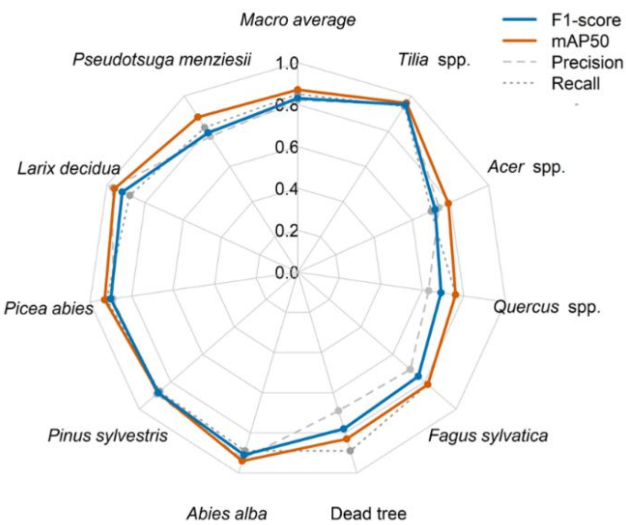

Figure 4. Species-level performance of the DL model trained on the balanced masked dataset after hyperparameter tuning.

In particular, *Quercus* spp. is frequently misclassified as *Tilia* spp. and *Acer* spp., while *Fagus sylvatica* shows confusion with multiple broadleaf species, and background is sometimes predicted as Dead tree, indicating overlap in structural or spectral signatures (Figure 5).

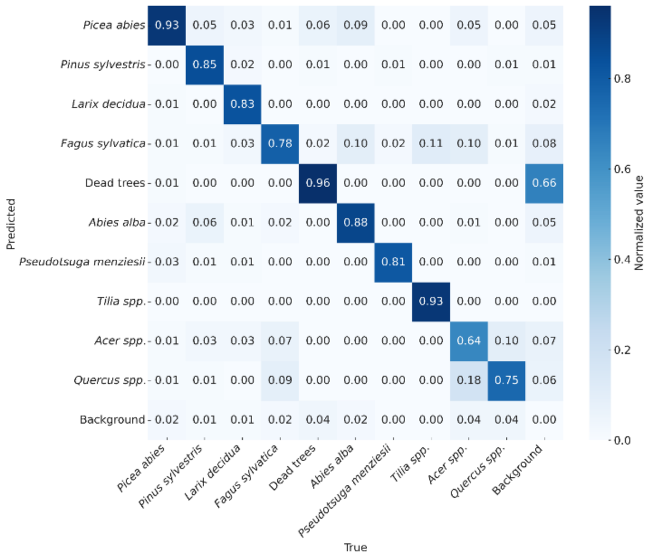

Figure 5. Normalized confusion matrix of the YOLO11x model trained on the balanced masked dataset after hyperparameter tuning (173 epochs).

3.3 Generalization to Mixed Masked and Unmasked Datasets

In the mixed dataset setting (50:50 masked and unmasked tiles), performance declined markedly (Table 5). The overall F1-score decreased to 0.56 and mAP50 to 0.56, indicating reduced robustness compared to fully masked training. While some conifer species maintained moderate accuracy, most broadleaved species were negatively affected by increased background variability.

Class	Precision	Recall	F1	mAP_{50}
Macro average	0.63	0.51	0.56	0.56
Picea abies	0.68	0.43	0.53	0.49
Pinus sylvestris	0.60	0.44	0.51	0.50
Larix decidua	0.73	0.56	0.63	0.69
Fagus sylvatica	0.53	0.38	0.44	0.42
Dead tree	0.63	0.65	0.64	0.64
Abies alba	0.70	0.47	0.57	0.53
Pseudotsuga menziesii	0.53	0.49	0.51	0.53
Tilia spp.	0.70	0.76	0.73	0.79
Acer spp.	0.62	0.42	0.50	0.49
Quercus spp.	0.58	0.51	0.54	0.48

Table 5. Performance metrics of the YOLO11x model trained on the balanced mixed dataset (50:50 masked and unmasked tiles).

3.4 Effect of the Attention Mechanism on Unmasked and Mixed Dataset

Adding an attention mechanism improved performance compared to the mixed dataset without attention (Table 6), increasing the overall F1-score by +0.04 (0.56 → 0.60) and mAP50 by +0.06. Gains were strongest for conifers, particularly Larix decidua (+0.10 F1) and Abies alba (+0.03 F1), and for dead trees (+0.24 F1). In contrast, broadleaved species showed no overall improvement on average (mean ΔF1 $\approx$ −0.01), with gains for Tilia spp. offset by declines or stagnation in other classes. Overall, attention partially reduced background-related errors but remained less effective than fully masked training.

Class	Precision	Recall	F1	mAP$_{50}$
Macro average	0.63	0.49	0.55	0.55
Picea abies	0.7	0.43	0.53	0.52
Pinus sylvestris	0.63	0.46	0.53	0.54
Larix decidua	0.78	0.53	0.63	0.69
Fagus sylvatica	0.5	0.39	0.44	0.42
Dead tree	0.61	0.66	0.63	0.67
Abies alba	0.71	0.44	0.54	0.53
Pseudotsuga menziesii	0.44	0.41	0.42	0.41
Tilia spp.	0.75	0.74	0.74	0.82
Acer spp.	0.63	0.37	0.47	0.46
Quercus spp.	0.54	0.43	0.48	0.42

Table 6. Performance metrics of the Attention based YOLO11x model trained on the balanced mixed dataset.

Applying the attention mechanism to the unmasked dataset improved performance compared to the baseline unmasked setting, increasing the overall F1-score by +0.30 (0.27 to 0.57) and mAP50 by approximately +0.35 (0.22 to 0.57) (Table 7). Gains were most pronounced for conifer species, particularly *Larix decidua* (F1 = 0.70) and *Pseudotsuga menziesii* (+0.35 F1 relative to baseline), as well as for dead trees (F1 = 0.77). In contrast, broadleaved species remained challenging, with low and variable performance (mean F1 $\approx$ 0.28) and some classes such as *Tilia* spp. and *Quercus* spp. not detected at all. Overall, attention substantially improved robustness compared to the unmasked baseline, but performance remained clearly below that achieved with masked training.

Class	Precision	Recall	F1	mAP$_{50}$
Macro average	0.63	0.49	0.55	0.55
Picea abies	0.74	0.38	0.50	0.60
Pinus sylvestris	0.74	0.42	0.54	0.52
Larix decidua	0.69	0.66	0.67	0.72
Fagus sylvatica	0.63	0.32	0.42	0.42
Dead tree	0.75	0.80	0.78	0.82
Abies alba	0.69	0.42	0.52	0.54
Pseudotsuga menziesii	0.44	0.39	0.41	0.38
Tilia spp.	0.56	0.71	0.63	0.63
Acer spp.	0.62	0.44	0.51	0.51
Quercus spp.	0.46	0.37	0.41	0.37

Table 7. Performance metrics of the Attention based YOLO11x model trained on the balanced unmasked.

3.5 Application to Full Aerial Tiles

To assess operational applicability, all models were applied to fully unmasked aerial mosaics, revealing clear differences between training strategies (Figure 6). The balanced masked model (Figure 6b) produced the densest detections with many overlapping boxes and false positives, whereas the unbalanced masked model (Figure 6a) showed sparse detections. In contrast, the balanced unmasked (Figure 6c) and mixed Figure 6d) models yielded fewer detections and higher omission rates but more realistic spatial patterns.

Masked training leads to over-detection and multiple overlapping predictions per crown (Figure 6b), while models trained with unmasked or mixed data show strong under-detection and miss many canopy trees (Figure 6c–d). This highlights that incomplete annotations bias the model and limit generalization to real-world conditions, resulting in an imbalance between over- and under-detection. The validation metrics alone do not fully capture operational performance and exposure to realistic background conditions during training is critical for transferable tree species detection.

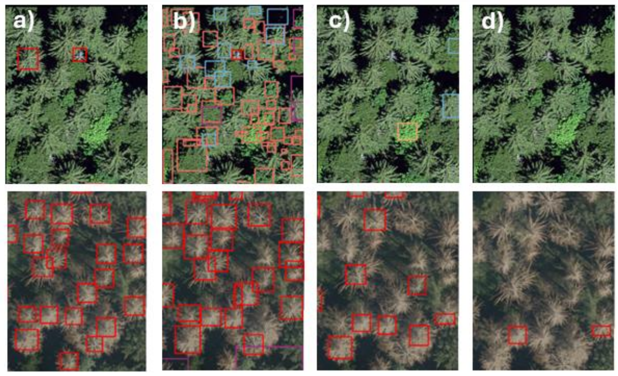

Figure 6. Detections on the same full aerial image for models trained on the following datasets: a) unbalanced masked, b) balanced masked, c) balanced unmasked, and d) mixed masked and unmasked dataset (50:50). For each configuration, results are shown for two example scenes, including a conifer-dominated stand (top rows) and a dead tree–dominated area (bottom rows). Blue colors represent conifers, warm colors represent deciduous trees, whilst red indicates dead trees.

4. Discussion

4.1 Impact of Dataset Composition on Model Performance

The experimental results indicate that dataset architecture is a primary determinant of individual tree detection efficacy. In this study, balancing the dataset increased F1-scores for several underrepresented species without substantially reducing overall performance, suggesting a more even class representation. However, confusion among morphologically similar broadleaved taxa, such as *Acer* and *Quercus*, remained evident. Taxonomic aggregation of classes such as *Tilia spp.* and *Acer spp.* reduced class variability and contributed to more stable predictions, supporting the use of simplified class structures in complex forest conditions. The observed improvement in classification stability through class balancing aligns with previous work on data imbalance (Safonova et al., 2023).

These findings indicate that models trained on partially labeled data, where only a subset of tree crowns is annotated (in our case, 78% of images have fewer than two crowns annotated), are challenging for tree species detection on full aerial images with the current YOLO setup.

4.2 Masking Strategies and Domain Shift

A key finding of this study is a discrepancy between validation performance and behavior on full aerial mosaics. Models trained on masked datasets achieved the highest quantitative scores (e.g., precision, recall, and F1), but their application to unmasked imagery while it showed the best detections of tree crowns, it had multiple overlapping predictions per crown or too big predictions.

This suggests that masking simplifies the feature space by removing background variability (e.g., soil, shadows, understory), which improves performance on curated validation tiles but limits generalization to more complex scenes. The detection of dead trees was the best on big aerial images, similar results have been observed in a previous study, though there the unmasked deep learning approached generalized the best (Beloiu Schwenke et al., 2025a)

4.3 Limitations and Data Constraints

The reliance on partially labeled datasets, where only delineated crowns were annotated and remaining pixels masked, remains a significant constraint. While this approach represented a most practical solution for generating large-scale training samples (Ecke et al., 2024), it inherently biases the feature extractor toward optimized crown geometries. This incomplete annotation problem typically leads to lower recall in dense canopy clusters where crown boundaries are not distinct. Addressing this limitation will require the development of fully labeled unmasked datasets that capture a broader range of canopy and background conditions, potentially supported by semi-automated labeling and active learning approaches. Therefore, the use of big datasets and the contribution to global databases would bring benefits for the entire community aiming to work on forest inventories with deep learning approaches (Beloiu Schwenke et al., 2025b, 2025c).

4.4 Future Work in Tree Species Mapping

To improve transferability from experimental benchmarks to operational inventories, future research should address domain-invariant feature learning. Domain adaptation aims to improve model performance across different data conditions, such as masked tiles and heterogeneous aerial mosaics. One approach is adversarial training, which encourages the model to learn features that are indistinguishable between domains. This may help reduce the discrepancy between high validation performance on curated tiles and unstable behavior on full aerial imagery, although limitations due to partially labeled data are likely to remain.

Moreover, bounding-box detection has limitations in dense forests, where overlapping crowns hinder the delineation of individual trees. Future work should investigate instance or panoptic segmentation approaches, which enable more precise delineation of overlapping crowns.

5. Conclusion

This study highlights dataset design as a key factor for deep learning-based tree species detection. While class balancing improved F1-scores for underrepresented species, a clear trade-off was observed between high validation performance on masked tiles and robust behavior on unmasked orthomosaics. Models trained on masked data achieved the highest quantitative accuracy but showed over-detection, whereas models exposed to unmasked imagery exhibited lower accuracy and increased omission rates. These results indicate that neither approach alone ensures robust generalization.

For practical applications, data quality should be prioritized over data quantity. In particular, fully labeled images, where all trees in an image are annotated, are preferable to partially labeled datasets, as incomplete annotations can bias model learning and limit generalization. Furthermore, model evaluation should not rely solely on validation or test splits but must include inference on complete aerial images to assess operational performance.

These findings emphasize that deep learning workflows for tree species detection should follow an integrated pipeline, including consistent data preparation, appropriate model selection, evaluation on held-out data, and validation through inference on representative full-scale imagery.

Acknowledgements

The authors would like to thank Elias Berger, who helped compile the original dataset.

References

Beloiu, M., Heinzmann, L., Rehush, N., Gessler, A., Griess, V.C., 2023. Individual Tree-Crown Detection and Species Identification in Heterogeneous Forests Using Aerial RGB Imagery and Deep Learning. Remote Sens. 15, 1463. https://doi.org/10.3390/rs15051463

Beloiu Schwenke, M., Berger, E., Ecke, S., Xia, Z., Reichmuth, C., Gessler, A., Klemmt, H.-J., Glatthorn, J., Stillhard, J., Griess, V., Waser, L.T., 2025a. Tree Crown Mortality and Defoliation Assessment in Temperate Forests Using Aerial Imagery and Deep Learning. https://doi.org/10.2139/ssrn.5360617

Beloiu Schwenke, M., Xia, Z., Cheng, Y., Gessler, A., Kattenborn, T., Xinlian, L., Mosig, C., Puliti, S., Rehush, N., Waser, L.T., Gries, V.C., Mokros, M., 2025b. Datenbank für die Annotationen von Baumarten mit hochauflösenden Luftbildern. Swiss For. J. 176, 1.

Beloiu Schwenke, M., Xia, Z., Novoselova, I., Gessler, A., Kattenborn, T., Mosig, C., Puliti, S., Waser, L., Rehush, N., Cheng, Y., Xinliang, L., Griess, V.C., Mokroš, M., 2025c. TreeAI Global Initiative - Advancing tree species identification from aerial images with deep learning. https://doi.org/10.5281/zenodo.15351054

Ecke, S., Dempewolf, J., Frey, J., Schwaller, A., Endres, E., Klemmt, H.-J., Tiede, D., Seifert, T., 2022. UAV-Based Forest Health Monitoring: A Systematic Review. Remote Sens. 14, 3205. https://doi.org/10.3390/rs14133205

Ecke, S., Stehr, F., Frey, J., Tiede, D., Dempewolf, J., Klemmt, H.-J., Endres, E., Seifert, T., 2024. Towards operational UAV-based forest health monitoring: Species identification and crown condition assessment by means of deep learning. Comput. Electron. Agric. 219, 108785. https://doi.org/10.1016/j.compag.2024.108785

Holzinger, A., Schweier, J., Gollob, C., Nothdurft, A., Hasenauer, H., Kirisits, T., Häggström, C., Visser, R., Cavalli, R., Spinelli, R., Stampfer, K., 2024. From Industry 5.0 to Forestry 5.0: Bridging the gap with Human-Centered Artificial Intelligence. Curr. For. Rep. 10, 442–455. https://doi.org/10.1007/s40725-024-00231-7

Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S., 2021. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS J. Photogramm. Remote Sens. 173, 24–49. https://doi.org/10.1016/j.isprsjprs.2020.12.010

Liang, X., Wang, Yinrui, Pan, J., Heiskanen, J., Wang, N., Wu, S., Vuorinne, I., Tian, J., Troles, J., Cloutier, M., Puliti, S., Chandrasekaran, A., Ball, J., Mi, X., Shen, G., Song, K., Shao, G., Astrup, R., Wang, Yunsheng, Pellikka, P., Wang, M., Gong, J., 2026. Empowering tree-scale monitoring over large areas: Individual tree delineation from high-resolution imagery. ISPRS

J. Photogramm. Remote Sens. 232, 974–999. https://doi.org/10.1016/j.isprsjprs.2025.12.022

Pleșoianu, A.-I., Stupariu, M.-S., Șandric, I., Pătru-Stupariu, I., Drăguț, L., 2020. Individual tree-crown detection and species classification in very high-resolution remote sensing imagery using a deep learning ensemble model. Remote Sens. 12, 2426. https://doi.org/10.3390/rs12152426

Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., Ryo, M., 2023. Ten deep learning techniques to address small data problems with remote sensing. Int. J. Appl. Earth Obs. Geoinformation 125, 103569. https://doi.org/10.1016/j.jag.2023.103569

Safonova, A., Tabik, S., Alcaraz-Segura, D., Rubtsov, A., Maglinets, Y., Herrera, F., 2019. Detection of fir trees (Abies sibirica) damaged by the bark beetle in unmanned aerial vehicle images with deep learning. Remote Sens. 11, 643. https://doi.org/10.3390/rs11060643

Ultralytics, 2024. YOLO11 Documentation [WWW Document]. Ultralytics Docs. URL https://docs.ultralytics.com/de/models/yolo11/ (accessed 10.30.25).

Xia, Z., Kattenborn, T., Wegner, J.D., Gessler, A., Griess, V.C., Beloiu, M., 2026. Advances in tree species identification from high-resolution aerial imagery and deep learning.