

A Deep Learning Model for Tree Species Instance Segmentation and Classification Using Ground-Level RGB Imagery and Automated Annotations

Hristina Hristova¹*, Clemens Blattert¹, Sunni K. P. Kushwaha¹, Janine Schweier¹

¹ Sustainable Forestry, Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Switzerland
(hristina.hristova, clemens.blattert, sunni.kushwaha, janine.schweier)@wsl.ch

Keywords: Tree species classification, Forestry, Deep learning, Instance segmentation, YOLO.

Abstract

Accurate tree species identification is essential for effective forest management, biodiversity monitoring, and resource estimation. While automated methods relying on aerial and canopy-level remote sensing have become prevalent, they often struggle in dense, multi-layered forest stands, where critical lower-stem and bark features are obscured. To address this limitation, we present a Deep Learning (DL) framework for tree species classification utilizing ground-level RGB imagery. Because manual annotation of terrestrial images in forest environments is labor-intensive and complicated by occlusions, we introduce a new "in-situ" forest image dataset alongside an automated labeling pipeline. This pipeline generates training annotations by projecting tree-species data derived from Mobile Laser Scanning (MLS) onto 2D images based on photogrammetric reconstruction. The proposed DL model leverages these automatically labeled images to effectively recognize tree species based on structural and bark characteristics. The model achieves overall F1-scores of 0.78 and 0.75 for object detection and instance segmentation, respectively. Ultimately, our approach complements existing methods for detecting tree positions and diameters, facilitating the creation of a holistic, cost-effective, and scalable forest inventory dataset.

1. Introduction

Data on tree species in forests is essential for assessing forest health and monitoring species diversity. It also plays a key role in estimating forest resources, such as timber quantity and quality, carbon storage and sequestration, and in planning management activities, such as thinning and harvesting operations (Blattert et al., 2017).

Traditional field inventory surveys for collecting information on forest structure and tree species information are inherently labor-intensive, time-consuming, and limited in scale. Consequently, there is a strong need to adopt automated methods for identifying tree species (Ferretti et al., 2024). Modern approaches have the potential to combine digital tools, deep learning (DL), and advanced remote sensing to facilitate faster, more cost-effective, and standardized data collection across large spatial scales (Kükenbrink et al., 2022; Mokroš et al., 2018).

Nowadays, automated tree species classification relies heavily on aerial remote sensing, primarily using images and point clouds captured from an overhead perspective. A significant body of research focuses on recognizing tree species from a canopy perspective using Unmanned Aerial Vehicle (UAV)-based Red-Green-Blue (RGB) and multispectral imagery (Qin and Zhao, 2025; Wang et al., 2025; Gaydon and Roche, 2024). Such approaches typically develop image-based models using highly optimized Convolutional Neural Networks (CNNs) that leverage advances in computer vision to effectively extract discriminative features (e.g., canopy texture, shape) from high-resolution aerial images (Beloiu et al., 2023).

The release of the FOR-species20K dataset (Puliti et al., 2024a), a diverse benchmark built from individual tree point clouds collected by various platforms, has yielded important

insights into the performance of image-based methods for tree species classification. Such approaches render 3D point clouds into 2D multi-view images and classify them using CNNs. Recent benchmarking on this dataset reveals that multi-view 2D image-based models consistently outperform pure 3D point cloud models, justifying the focus on high-resolution RGB imagery. Despite the richness of 3D data, converting 3D point clouds into 2D multi-view images often yields higher accuracy than processing them as raw 3D points. This is largely due to the maturity and optimization of 2D CNNs (Puliti et al., 2024b).

While aerial methods excel at capturing canopy features, they may miss crucial species-specific information contained in the lower stem and bark, or struggle in multi-layered, dense forest stands. This motivates the exploration of alternative, terrestrial viewpoints. In this context, our research focuses on developing a DL model designed for recognizing tree species directly from RGB images taken from a ground-level viewpoint. Our approach leverages features such as bark texture and stem characteristics and patterns that are not easily visible from the air, offering an alternative solution for tree species identification, especially in dense forests.

Although current terrestrial models achieve high accuracy on curated close-up images of bark (Warner et al., 2024), they do not address cases of unconstrained forest environments where issues like occlusion and the presence of multiple species are common. Additionally, manually annotating ground-level imagery can be tedious and expensive. This paper tackles these challenges by (1) introducing a dataset comprising "in-situ" ground-level 2D images of forests and (2) implementing an automated labeling workflow that projects tree species data derived from Mobile Laser Scanning (MLS) onto the 2D images.

We propose a DL model that recognizes tree species from forest RGB images captured from a terrestrial viewpoint and assess its effectiveness. We introduce a dataset that includes these forest

* Corresponding author

images, along with annotations of the primary tree species depicted in them. Additionally, we propose a pipeline in which the annotations used to train our model are automatically generated from MLS data and photogrammetric reconstructions.

Our approach builds upon previous work that used images from 360-degree cameras to effectively detect tree locations and Diameter at Breast Height (DBH) (Hristova et al., 2025). By incorporating automated tree species identification alongside these established spatial and structural metrics, this research bridges a crucial gap. Ultimately, it enables the creation of a comprehensive forest inventory dataset.

2. Method

2.1 Data acquisition

In July 2025, we conducted data collection across 15 temperate forest plots in the canton of Bern, Switzerland. Local forest experts from the state forest enterprise carefully selected the plots, including both common and rare tree species, to ensure a diverse mix that was easily accessible. Each plot included trees of interest with a DBH ranging from 6.5 cm to 122 cm. The size of each plot was 250 m². In total, the plots contained 340 trees from 14 different species, with *Picea abies* being the most prevalent and *Sorbus aucuparia* the least represented. We marked the trees of interest using AprilTags (Olson, 2011).

Data acquisition was performed using a FARO Focus 3D 120S terrestrial laser scanner and the Insta360 Pro 2 camera. As we followed a walking acquisition path, both devices captured data simultaneously. The 3D scanner generated MLS point clouds of the forest plot areas, while the Insta360 Pro 2 recorded six fisheye videos, providing a 360-degree view of the plots.

In addition to acquiring MLS and Insta360 Pro 2 data, we also conducted a traditional (manual) field inventory of the trees of interest. This inventory included recording each tree's species, DBH, health, and quality. We linked this inventory to the digital data using the AprilTags, as each tag's ID matched the tree ID in the inventory database.

2.2 Dataset

We created a dataset of forest images annotated according to the tree species present in each image. The images were extracted from the fisheye videos recorded with the Insta360 Pro 2 camera, ensuring that only the highest-quality images were retained. The image dataset was dominated by three tree species: *Pseudotsuga menziesii*, *Abies alba*, and *Picea abies*. The least-represented species included *Larix decidua*, *Sorbus aucuparia*, and *Thuja plicata*. The number of images that included a given tree species is shown in Figure 1.

To automatically generate annotations for tree species in images, we first established a connection between the tree species observed in the video frames and those documented in the field inventory. Since the AprilTags were attached to only one side of the tree stem, they were not visible in every video frame, making direct detection in the images inefficient. In addition to identifying tree species, we encountered the challenge of computing the boundaries of tree stems within the images. Therefore, we opted to detect the trees of interest in 3D space and then project them onto the 2D image space. We created a 3D reconstruction of all the plots from videos captured with the Insta360 Pro 2, employing photogrammetric principles to link 3D space to the 2D video frames.

2.2.1 Insta360 Pro 2 reconstruction We applied a multi-view videogrammetry approach, as demonstrated by Hristova et al. (2025), to reconstruct a point cloud for each forest plot from the recorded Insta360 Pro 2 videos. We extracted 4K frames from the camera's six fisheye lenses at 5 frame per second (fps). The lowest-quality images, most of which contained motion blur, were eliminated by applying a Laplacian filter to the grayscale version of each frame. The Laplacian filter measures edge sharpness in a video frame. Images were filtered based on the variance of their Laplacian filter values, computed over the central region of the video frame that contains the most crucial information for our model (the tree stems).

Next, we removed the sky from all images and applied a Structure from Motion (SfM) approach with two spatial constraints to align all video frames and position them in 3D space (Hristova et al., 2025). Afterwards, we generated a point cloud of the forest scene. Hristova et al. (2025) demonstrated that the 3D reconstruction pipeline using Insta360 Pro 2 videos produces point clouds of forest plots that are comparable to those obtained from MLS in terms of tree positioning and DBH.

The point clouds obtained from MLS generally offer a more comprehensive representation of tree stems compared to those generated by the Insta360 Pro 2 camera. In most instances, the MLS point clouds include data points for the entire tree stem, extending up to the canopy. In contrast, the Insta360 Pro 2 pipeline reconstructs a forest stand up to 15 meters in height (Hristova et al., 2025). Therefore, to achieve better, more complete, and accurate annotations of the trees in the video frames, the trees from the MLS point cloud were projected onto the pixels of the recorded video frames.

2.2.2 Alignment and clustering The Insta360 Pro 2 point clouds allow reconstructed 3D points to be associated with corresponding pixels in the video frames. Additionally, the AprilTags and their IDs can be automatically detected in the photogrammetric point clouds, serving as a ground-truth link between the MLS point clouds, the Insta360 Pro 2 point clouds, and the field inventory.

To project MLS trees onto the video frames, we first transformed the MLS point clouds into the coordinate system of the Insta360 Pro 2 point cloud. This process involved aligning the point clouds generated by the Insta360 Pro 2 with their corresponding MLS point clouds for each forest plot. We performed the alignment semi-automatically by first roughly aligning the two point clouds and then applying the Iterative Closest Point (ICP) algorithm of CloudCompare (CloudCompare, 2025).

With the aligned point clouds, we calculated the transformation T_a that converts points from the MLS-derived point cloud to the coordinate system of the Insta360 Pro 2 point cloud. Once the transformation T_a was established, the trees of interest were clustered within the MLS point cloud. We applied an automated tree-clustering algorithm, as proposed in (Hristova et al., 2024). This algorithm cleaned the initial point cloud by removing leaves and small branches, yielding clusters that primarily contained the stems of the trees we focused on. Since tree stems are the most recognizable features in the fisheye videos, we annotated them while minimizing interference from leaves.

Initially, we segmented the entire trees, including the crown. However, when these complete MLS tree masks were projected onto the video frames, the crown areas often included canopy elements and branches from other trees. To streamline the classification, we limited the annotations to the stems.

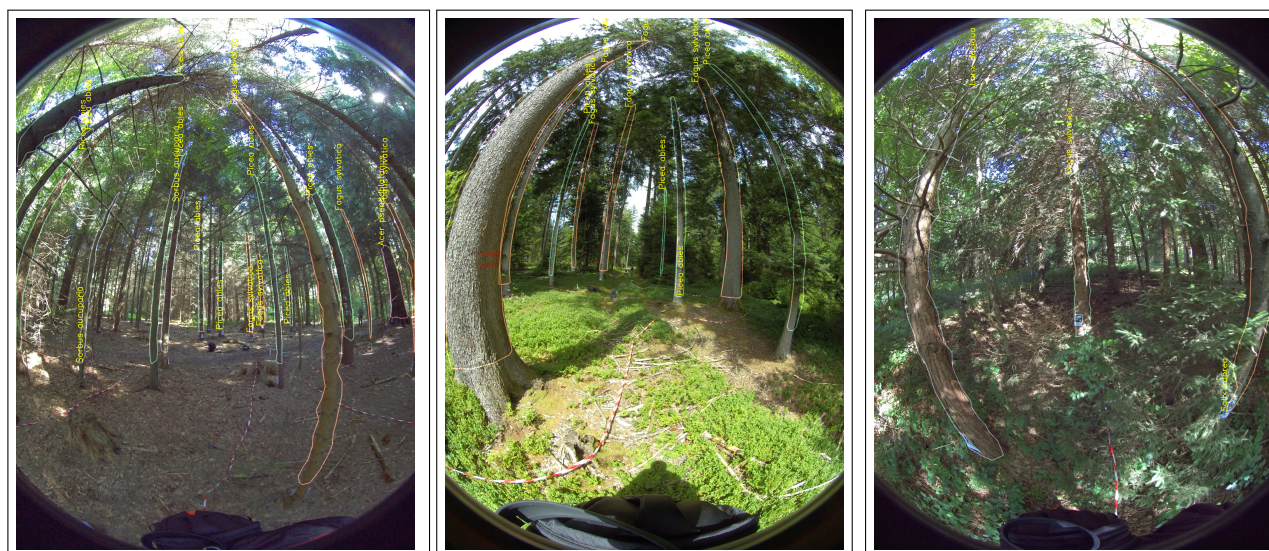


Figure 1. Examples of annotated images. Identical polygon colors within an image indicate masks of the same tree species.

2.2.3 Automated 3D-to-2D annotations To generate 2D annotations, the segmented MLS clusters were mapped into the image coordinate system using two successive transformations. First, the MLS clusters were aligned with the coordinate system of the Insta360 Pro 2 point clouds via the transformation T_a . Second, the transformed MLS points were projected onto the image plane using the camera's projection matrix $T_{3D \rightarrow 2D}$, derived from the SfM reconstruction. The final pixel coordinates $P_{(i,j)}$ were calculated as:

$$P_{(i,j)} = T_{3D \rightarrow 2D} \cdot (T_a \cdot P_{(x,y,z)}) \quad (1)$$

For efficiency, we randomly sub-sampled each MLS tree cluster before projecting it onto the image plane, using a 1 cm grid filter size. Projecting an MLS tree cluster onto a video frame yielded pixels corresponding to that specific tree. Instead of retaining all pixel coordinates from the projection, we computed the concave hull ($\alpha = 0.1$) around the projected points, effectively creating a tight boundary around each visible tree stem. The 2D coordinates of this concave hull were recorded in a file, along with the associated tree ID corresponding to the AprilTag (and the inventory). These records represented the annotations of the trees in the 2D image. Automatically computed tree masks are visualized in Figure 1 for three images in our dataset.

All automatic annotations were thoroughly reviewed by a trained expert who evaluated the accuracy of each bounding box and mask for every tree in the dataset. The review process included confirming that the spatial boundaries captured the proper visible tree stems.

One of the critical aspects of the review was checking for annotations of occluded trees. If any trees were partially obscured and thus difficult to identify with certainty, the expert removed those annotations from the dataset. This decision was made to maintain the integrity of the training data and ensure that the model learns from clear, identifiable examples.

To further enhance the training process, it was ensured that the dataset included frames spaced further apart temporally. This strategic selection was essential to prevent the model from

learning motion patterns in the video sequences, thereby enhancing its ability to generalize effectively across different forest scenes. Moreover, this approach to data preparation improved annotation quality and enhanced the robustness and reliability of the resulting model.

2.3 DL model

The annotated image dataset was split into training, validation, and test sets. For each recorded video, the training set contained 70% of the video frames, the validation set 10%, and the remaining 20% comprised the test set. The data were structured to increase the temporal distance between video frames in the test set relative to the training set, while ensuring that all tree species were represented.

To classify tree species in forest images, we utilized the state-of-the-art YOLOv11 approach (Khanam and Hussain, 2024). To train the YOLOv11 model efficiently with high-resolution images, the original 4K images are usually resized to a fixed resolution (e.g., 640×640 pixels). However, this resizing can reduce small objects to a few pixels, making detection difficult. To mitigate this issue, we sliced each 4K video frame into 48 images, each measuring 640×640 pixels, with a 20% overlap between them. The annotations associated with the 4K image frames were adjusted accordingly, and any sliced images without annotations were excluded from the dataset. This approach increased the size of our training data without significantly downsizing the original video frames. It also enhanced the visibility of trees that may appear small in the captured videos, allowing our model to better explore the features of the tree species, as they would appear larger relative to the image size.

Our method takes as input the sliced images in our training set along with their annotations (2D polygon coordinates of the trees of interest and the tree ID corresponding to their species). The distribution of the training samples per tree species is shown in the last column in Table 1. The model was trained with an image resolution of 640×640 pixels, an initial learning rate of 0.00025, and a batch size of 20. The number of training epochs was set to 1000, but the model converged on epoch 850. The best epoch was chosen based on the evaluation met-

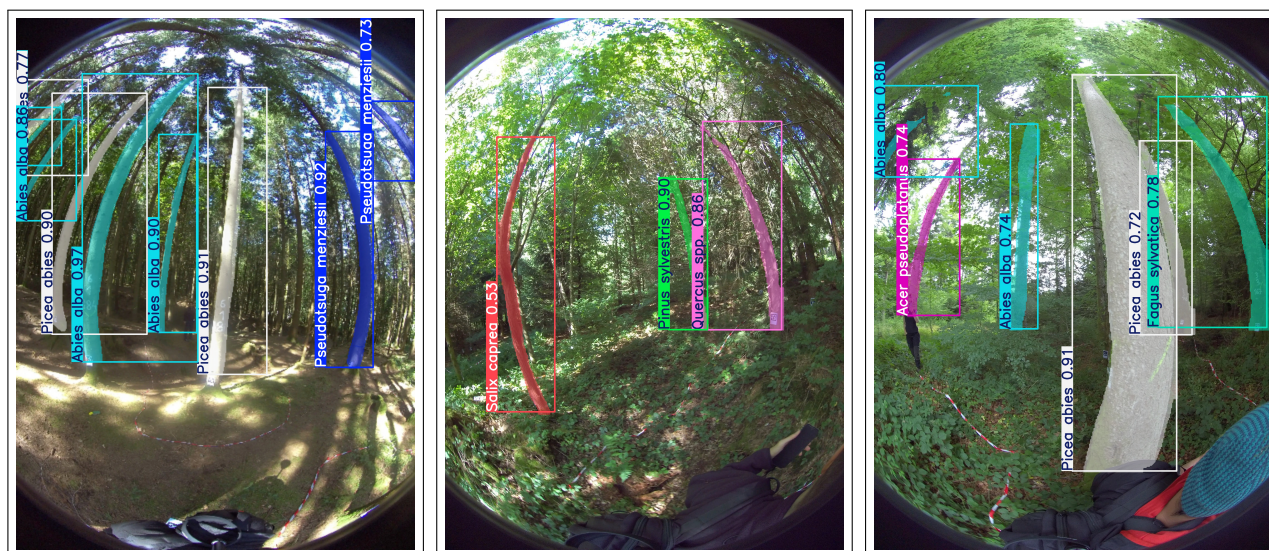


Figure 2. Tree species predicted using our model on three images in the test set.

rics computed on the validation set. For evaluation, we used precision, recall, and F1-score.

2.4 Inference

To address the challenge of identifying tree species at varying distances within a single terrestrial forest scene, we integrated the Slicing Aided Hyper Inference (SAHI) framework (Akyon et al., 2022). Traditional inference may fail to detect distant stems due to significant pixel down-sampling. By employing a slicing-based approach, our model performs high-resolution inference on localized patches, ensuring that fine-grained bark textures and stem characteristics are preserved even for trees located in the background of the image. This is particularly crucial for a multi-species dataset like ours, where overlapping stems and varying depths of field are common.

For inference, each original video frame was divided into 48 slices, each of resolution 640×640 pixels, with a 20% overlap. Our YOLOv11 model was applied to each slice independently, and the predictions were then merged back into the original image space.

3. Results

The model’s overall performance was evaluated on both bounding-box (object detection) and mask (instance segmentation) tasks. Figure 2 visualizes the model’s bounding-box and mask predictions on images from the test set. The precision-recall curves for object detection and instance segmentation for each tree species in the dataset are presented in Figure 3. Generally, the model demonstrated slightly better performance at identifying bounding boxes around trees than at accurately masking their shapes. The overall bounding box mean average precision (mAP@0.5) was 0.819 across all tree species classes. The optimal F1-score for bounding box detection reached 0.78 at a confidence threshold of 0.375. In contrast, the instance segmentation task proved slightly more challenging than bounding box localization. The overall mask mAP@0.5 was 0.771, with a maximum F1-score of 0.75 at a confidence threshold of 0.385.

The precision-recall curves in Figure 3 provide valuable insights into the model’s reliability. For most species, preci-

sion remained close to 1.0 until recall reached about 0.6. This suggests that while the model made reliable predictions, it struggled to detect every tree in a given area. The latter can also be observed in Figure 2, where some trees were not identified. The blue curves in Figure 3, which represent the performance across all classes, are relatively smooth. This indicates that for most species, the proposed model exhibited a predictable trade-off: it achieved approximately 80% recall while maintaining precision between 60% and 70%.

However, the model struggled significantly with specific broad-leaved species, such as *Fraxinus excelsior*, which showed a rapid decline in precision for both object detection and instance segmentation. The “step-like” curve for this species indicated insufficiently representative training samples, as *Fraxinus excelsior* was one of the underrepresented tree species in our dataset. Moreover, the model can detect this tree species with a precision of 0.665, but it struggles to outline its exact shape (mask precision is 0.442).

The confusion matrix in Figure 4 reveals that the model achieved high accuracy for several species, including *Pseudotsuga menziesii* (85%), *Pinus sylvestris* (84%), *Pinus strobus* (84%), *Thuja plicata* (84%), and *Abies alba* (83%). For these species, the model maintained high precision, yielding few false positives, and achieved high recall, identifying nearly all instances, as detailed in Table 1.

Additionally, each species had a proportion of its instances misclassified as background, as shown in the confusion matrix in Figure 4. For example, approximately 36% of *Sorbus aucuparia* and 32% of *Fraxinus excelsior* instances were misclassified, indicating that around a third of these trees were entirely missed by the model (resulting in false negatives). In contrast, Figure 4 also shows that some part of the background was detected as tree species (mainly *Picea abies* and *Fagus sylvatica*).

4. Discussion

4.1 False negatives

The normalized confusion matrix indicates that the model’s main challenge is not confusion between species but rather false

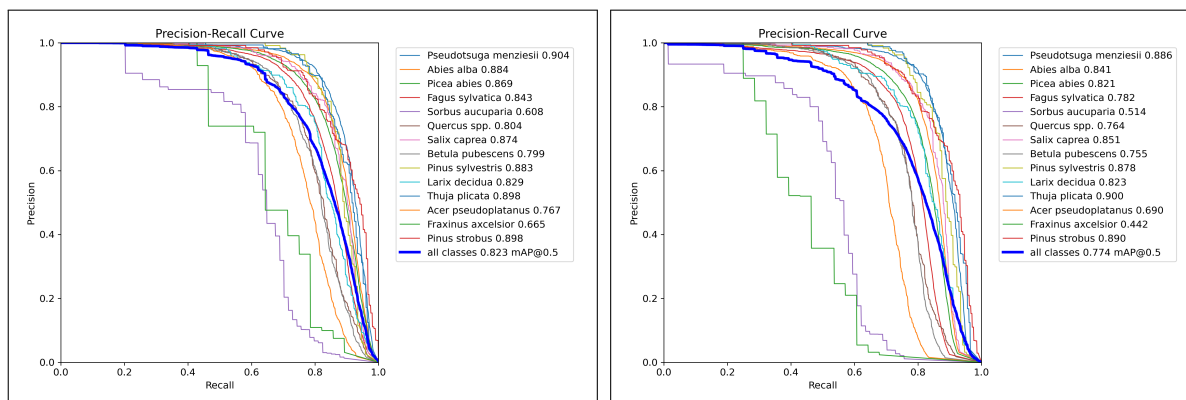


Figure 3. Precision-Recall curves per tree species. The blue curve illustrates the performance across all classes.

Tree species	Box			Mask			Instances
	F1	Precision	Recall	F1	Precision	Recall	
Pseudotsuga menziesii	0.88	0.904	0.85	0.87	0.886	0.85	2094
Abies alba	0.86	0.884	0.83	0.84	0.841	0.83	6599
Picea abies	0.84	0.869	0.81	0.82	0.821	0.81	8844
Fagus sylvatica	0.82	0.843	0.79	0.79	0.782	0.79	6909
Sorbus aucuparia	0.61	0.608	0.61	0.56	0.514	0.61	128
Quercus spp.	0.77	0.804	0.74	0.75	0.764	0.74	718
Salix caprea	0.85	0.874	0.83	0.71	0.728	0.70	623
Betula pubescens	0.76	0.799	0.72	0.74	0.755	0.72	1163
Pinus sylvestris	0.86	0.883	0.84	0.86	0.878	0.84	168
Larix decidua	0.81	0.829	0.79	0.81	0.823	0.79	335
Thuja plicata	0.87	0.898	0.84	0.87	0.900	0.84	260
Acer pseudoplatanus	0.72	0.767	0.68	0.68	0.690	0.68	1721
Fraxinus excelsior	0.47	0.665	0.36	0.40	0.442	0.36	28
Pinus strobus	0.87	0.898	0.84	0.86	0.890	0.84	206

Table 1. Evaluation performed on our test set of forest images. The metrics F1, precision, and recall were computed for bounding boxes (object detection) and object masks (instance detection) for each tree species in our dataset. The last column shows the number of actual instances per species in our training dataset.

negatives. A percentage of each class was misclassified as background, particularly for *Sorbus aucuparia* (36%) and *Fraxinus excelsior* (32%). This confusion with the background may stem from incomplete annotations in the training images. However, slicing the 4K frames into overlapping 640x640-pixel images helped to address some of the issues caused by missing annotations. During our field trip, we collected data on the trees of interest, but frequently encountered additional trees in the captured videos that were outside the perimeter of our plots. As a result, these extra trees remained unannotated. If we had trained our model using the original 4K video frames, it might have struggled with the abundance of background elements. By breaking down the video frames into smaller images and excluding those without annotations, we minimized the impact of the background and missing annotations.

Slicing the original 4K video frames into smaller images enabled a more detailed exploration of tree features and bark's fine textures without reducing the initial resolution. This approach also improved the detection and recognition of small objects, such as distant trees.

4.2 Coniferous vs. deciduous trees

Our model demonstrates high precision and recall for coniferous species, such as *Pseudotsuga menziesii*, *Thuja plicata*, and *Pinus strobus*, and maintains robust performance across both object detection and instance segmentation tasks. In contrast,

performance degrades for deciduous species, such as *Fraxinus excelsior*, *Sorbus aucuparia*, and *Acer pseudoplatanus*. The normalized confusion matrix in Figure 4 reveals that this degradation is not primarily driven by inter-species confusion among broad-leaf trees, but rather by a high rate of false negatives, where deciduous trees were missed and classified as background.

The tree species *Fraxinus excelsior* and *Sorbus aucuparia* were also among the most underrepresented classes in our dataset. To address the class imbalances, we applied data augmentation. However, the model would benefit from additional data for the minority classes representing trees under varied lighting conditions, forest stand types, and young vs. mature trees of the same species. That could help the model learn more robust features for these challenging species. Extending our training dataset could be achieved, for example, by incorporating forest Martelloscope data representing a network of forest training plots for professional and educational purposes (Krumm et al., 2019).

4.3 Automated annotation pipeline

A major bottleneck in developing terrestrial DL models has been the cost and labor associated with manually annotating images, particularly when dealing with occlusions and mixed-species stands. The automated labeling workflow introduced in this paper addresses this challenge directly. By projecting

Pseudotsuga menziesii	85%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	6%	0%
Abies alba	0%	83%	1%	0%	0%	0%	0%	0%	0%	0%	1%	0%	21%	0%
Picea abies	0%	1%	81%	1%	0%	0%	0%	0%	0%	0%	0%	1%	1%	31%
Fagus sylvatica	1%	1%	1%	79%	1%	1%	0%	0%	0%	1%	0%	1%	0%	25%
Sorbus aucuparia	0%	0%	0%	0%	61%	0%	0%	0%	0%	0%	0%	0%	0%	1%
Quercus spp.	0%	0%	0%	0%	0%	74%	0%	1%	1%	1%	0%	0%	0%	3%
Salix caprea	0%	0%	0%	0%	0%	1%	83%	0%	1%	0%	0%	0%	0%	2%
Betula pubescens	0%	0%	0%	0%	0%	1%	0%	72%	0%	1%	0%	0%	0%	3%
Pinus sylvestris	0%	0%	0%	0%	0%	0%	0%	0%	84%	0%	0%	0%	0%	1%
Larix decidua	0%	0%	0%	0%	0%	0%	0%	1%	0%	79%	0%	0%	0%	2%
Thuja plicata	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	84%	0%	0%	1%
Acer pseudoplatanus	0%	0%	0%	1%	0%	0%	0%	0%	0%	0%	0%	68%	0%	4%
Fraxinus excelsior	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	36%	0%
Pinus strobus	0%	0%	0%	0%	0%	0%	0%	0%	0%	1%	0%	0%	0%	84%
Background	13%	15%	17%	19%	36%	23%	16%	27%	14%	17%	15%	30%	32%	11%

Pseudotsuga menziesii
Abies alba
Picea abies
Fagus sylvatica
Sorbus aucuparia
Quercus spp.
Salix caprea
Betula pubescens
Pinus sylvestris
Larix decidua
Thuja plicata
Acer pseudoplatanus
Fraxinus excelsior
Pinus strobus
Background

Figure 4. Normalized confusion matrix computed on the test dataset, showcasing the predictions of our model for each tree species. The y-axis represents the predictions made by our model, whereas the x-axis shows the true classes.

tree species data derived from MLS onto 2D images via photogrammetric reconstruction, we established a scalable method for generating robust ground-truth annotations.

While our automated pipeline mitigates annotation labor, it relies on the initial accuracy of the MLS data and the precision of the photogrammetric projection. Any registration errors during this phase can propagate as noisy labels during model training.

Apart from potential registration errors, our automated labeling pipeline is sensitive to the alignment between the MLS and Insta360 Pro 2 point clouds. As evaluated in (Hristova et al., 2025), the expected error for this alignment ranges from 5 cm to 19 cm. However, during our alignment process, we observed minimal deviation, ranging from 1 cm to 3 cm.

Moreover, forest environments may present extreme variations in lighting conditions and severe occlusions from understory vegetation (e.g., dense shrubs or low branches). While our dataset specifically targeted these “in-situ” conditions, extreme instances of occlusion may influence the model accuracy.

5. Conclusion

This paper showcases the feasibility of a terrestrial-based approach to automated tree species classification. By focusing

on ground-level perspectives, we aimed to overcome the limitations of canopy-only visibility and address the complexities inherent in unconstrained, “in-situ” forest environments.

We have developed and assessed a DL model specifically tailored to identify tree species directly from ground-level RGB images. Our model uses distinct stem and bark characteristics to enhance species recognition across diverse forest settings. We introduced a dataset comprising forest imagery, complete with annotations of primary tree species. To address the time-consuming and expensive process of manual annotation, we proposed a novel workflow that automatically generates training annotations by projecting MLS data onto 2D images, thereby streamlining dataset preparation. Through these contributions, our research advances the accuracy of tree species classification and paves the way for more efficient data collection methods in ecological studies.

6. Funding

This work received funding from the European Research Council (ERC) under the European Union Horizon Europe (HORIZON) Research and Innovation program with grant agreement No. 101135517 (Small4Good), and from the Swiss State Secretariat for Education, Research, and Innovation (SERI).

7. Acknowledgments

The authors express their gratitude to Daniel Kükenbrink and Mauro Marty for their support with processing the MLS point clouds and to Niál Perry for his help in data collection.

References

- Akyon, F. C., Altinuc, S. O., Temizel, A., 2022. Slicing aided hyper inference and fine-tuning for small object detection. *2022 IEEE International Conference on Image Processing (ICIP)*, IEEE, 966–970.
- Beloïu, 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 Sensing*, 15(5), 1463.
- Blattert, C., Lemm, R., Thees, O., Lexer, M. J., Hanewinkel, M., 2017. Management of ecosystem services in mountain forests: Review of indicators and value functions for model based multi-criteria decision analysis. *Ecological Indicators*, 79, 391–409.
- CloudCompare, 2025. CloudCompare: 3D Point Cloud and Mesh Processing Software.
- Ferretti, M., Fischer, C., Gessler, A., Graham, C., Meusburger, K., Abegg, M., Bebi, P., Bergamini, A., Brockerhoff, E. G., Brunner, I. et al., 2024. Advancing forest inventory and monitoring. *Annals of Forest Science*, 81(1), 6.
- Gaydon, C., Roche, F., 2024. PureForest: A Large-Scale Aerial Lidar and Aerial Imagery Dataset for Tree Species Classification in Monospecific Forests. *arXiv 2024. arXiv preprint arXiv:2404.12064*.
- Hristova, H., Blattert, C., Candrea, B., Nita, M., Florea, S., Kushwaha, S. K. P., Schweier, J., 2025. Enhancing forest inventory via a videogrammetry approach for robust 3D reconstruction: A study using Insta 360 Pro 2. *Ecological Informatics*, 103398.
- Hristova, H., Murtiyoso, A., Kükenbrink, D., Marty, M., Abegg, M., Fischer, C., Griess, V. C., Rehush, N., 2024. Viewing the Forest in 3D: How Spherical Stereo Videos Enable Low-Cost Reconstruction of Forest Plots. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.
- Khanam, R., Hussain, M., 2024. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*.
- Krumm, F., Lachat, T., Schuck, A., Büttler, R., Kraus, D., 2019. Marteloskope als Trainingstools zur Erhaltung und Förderung von Habitatbäumen im Wald. *Schweizerische Zeitschrift für Forstwesen*, 170(2), 86–93.
- Kükenbrink, D., Marty, M., Bösch, R., Ginzler, C., 2022. Benchmarking laser scanning and terrestrial photogrammetry to extract forest inventory parameters in a complex temperate forest. *International Journal of Applied Earth Observation and Geoinformation*, 113, 102999.
- Mokroš, M., Liang, X., Surový, P., Valent, P., Čerňava, J., Chudý, F., Tunák, D., Saloň, Š., Merganič, J., 2018. Evaluation of close-range photogrammetry image collection methods for estimating tree diameters. *ISPRS International Journal of Geo-Information*, 7(3), 93.
- Olson, E., 2011. Apriltag: A robust and flexible visual fiducial system. *2011 IEEE International Conference on Robotics and Automation*, 3400–3407.
- Puliti, S., Lines, E., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M. J., Lukas, W., Rehush, N., Hristova, H. et al., 2024a. FOR-species20K dataset.
- Puliti, S., Lines, E. R., Müllerová, J., Frey, J., Schindler, Z., Straker, A., Allen, M. J., Winiwarter, L., Rehush, N., Hristova, H. et al., 2024b. Benchmarking tree species classification from proximally-sensed laser scanning data: introducing the FOR-species20K dataset. *arXiv preprint arXiv:2408.06507*.
- Qin, T., Zhao, Q., 2025. Multi-branch and multi-label tree species classification using deep learning for UAV aerial photography and Sentinel remote sensing images. *Scientific Reports*, 15(1), 32710.
- Wang, S., Liu, C., Wei, S., Tang, H., 2025. Urban tree species classification using multisource satellite remote sensing data and street view imagery. *Geo-spatial Information Science*, 28(1), 164–184.
- Warner, C., Wu, F., Gazo, R., Benes, B., Kong, N., Fei, S., 2024. Centralbark image dataset and tree species classification using deep learning. *Algorithms*, 17(5), 179.