Multimodal Deep Learning for Individual Tree Crown Segmentation

Jakob Danel (jakob.danel@eftas.com), Sebastian Mader, Oliver Buck

EFTAS Fernerkundung Technologietransfer GmbH, Oststraße 2, 48145 Münster, Germany

Abstract

Leverage of remote sensing data is an important building block for modern forest inventory, to monitor and manage forest resources effectively. To contribute to this field, we propose a multi-modal deep learning approach that integrates high-resolution aerial imagery and full 3D LiDAR point cloud data for an end-to-end individual tree crown segmentation framework. To complete the task, we extend an existing two-task U-Net architecture to incorporate LiDAR point clouds represented as voxels and append the pipeline with a post-processing step to extract tree crown polygons from the network outputs. We further enhance the model with an ensemble approach to stabilize predictions and provide uncertainty estimates. A systematic comparison of different input modalities reveals that while the fusion of aerial imagery and LiDAR data does not significantly outperform aerial imagery alone, the use of full 3D LiDAR information in voxel format yields better results than traditional 2D rasterized features. The proposed ensemble method effectively improves crown delineation quality and offers interpretable uncertainty metrics, which are valuable for downstream forest monitoring applications. Overall, our work advances the integration of multimodal remote sensing data for individual tree crown segmentation, providing insights into the benefits and limitations of combining 2D and 3D data sources.

Introduction & Challenge

The project investigates how multimodal remote sensing data can improve Individual Tree Crown (ITC) delineation. Specifically, we compare the performance of aerial imagery and LiDAR data as input modalities and explore their fusion within a deep learning framework. Our approach extends the model of Li et al. (2023) with an additional 3D encoder for voxelized LiDAR data and introduces temporal ensembling for enhanced robustness. ITC segmentation represents a crucial step for linking pixel-level predictions to object-level forest metrics, thereby bridging the gap between remote sensing and traditional forest inventory practices.

Different deep learning approaches have been used for ITC delineation on both aerial imagery and LiDAR data. For aerial imagery, U-NETs are widely used, which producing different output maps to fulfill the task (Schindler et al. 2025; Freudenberg, Magdon, and Nölke 2022; Brandt et al. 2020). For LiDAR data, different approaches exist for the full point cloud, ranging from PointNet-based architectures (Xiang et al. 2024) to 3D-CNNs on voxelized data (Wielgosz et al. 2024). However, when it comes to fusing both modalities, most methods choose to leverage only the Canopy Height Model (CHM) derived from LiDAR (Li et al. 2023; Tong and Zhang 2025; Lucas et al. 2024).

In contrast, our work aims to directly integrate the full 3D structure of LiDAR point clouds with high-resolution aerial imagery, thereby preserving more detailed spatial information for improved crown delineation.

Practical Relevance & Stakeholders

The approach facilitates object-level individual tree crown mapping, density estimation, and integration with conventional forest inventory data. This helps determine optimal dataset combinations for future projects and demonstrates the feasibility of LiDAR-imagery fusion for operational forest analytics. This addresses the following stakeholders:

- **Data producers**: Guidance on which sensor combinations are most beneficial for forest mapping tasks.
- **Developers**: Insights into multimodal fusion strategies for DL-based tree crown delineation.
- **Forest owners**: Improved transparency regarding forest structure, density, and composition.
- **Practitioners**: Support for management decisions through spatially explicit stand information.

Methodology & Approach

We trained and evaluated comparable network configurations using different input modalities to generate ITC polygons. The dataset comprises 21,280 hand-annotated tree crowns (16,645 training; 2,870 validation and 1,743 test trees) from the Spessart region in Bavaria. Aerial imagery (0.1 m resolution) was acquired at the end of the 2023 vegetation period, while LiDAR data (100 points/m²) originated from early spring 2023. LiDAR point clouds were voxelized using the mean return intensity and count per voxel. The 3D encoder compresses vertical information, which is then fused with 2D image features within a U-Net-based model (Li et al. 2023). Post-processing transforms segmentation outputs into crown polygons following an adaption of the method by Freudenberg, Magdon, and Nölke (2022). An own approach for watershed-based instance segmentation is proposed and compared to the Freudenberg et al. approach. To enhance prediction stability, we introduce a temporal ensembling strategy, creating a set of model snapshots during training time, by gathering weights at epochs after stabilization of the validation loss. During inference, predictions from these snapshots are averaged to produce stable outputs and uncertainty estimates based on prediction variance. To assess performance, we match predicted crowns to ground truth using a 50% IoU threshold and compute precision, recall, and F1-score. As baselines, we implement the model setup from Li et al. (2023) using only aerial imagery or CHM-derived LiDAR features.

Results

Overall, the configurations using aerial imagery data performed better, than the setups based on LiDAR data. Although the voxelized point clouds yielded acceptable results, they were outperformed based on the texture-information of the aerial imagery. The fusion of both modalities did not lead to a significant performance increase compared to using aerial imagery alone. However, the voxel-based LiDAR representation achieved higher accuracy than traditional 2D rasterized features, indicating the value of preserving 3D structure. The ensemble method improved delineation quality by a F1-Score of +6% and provided uncertainty metrics that can guide decision-making in forest management. Our experiments demonstrate an end-to-end workflow for producing ITC polygons directly from multimodal inputs. The new voxel-based data fusion technique, enables direct comparison of 2D and 3D representations. The integration of LiDAR voxels leads to improved crown boundary precision, while ensemble-based uncertainty metrics enhance interpretability. Figure 1 shows that the output of the workflow can be used to derive stand information of forest areas, on tree level, allowing the transformation from pixel-based remote sensing data to object-based forest inventory data.

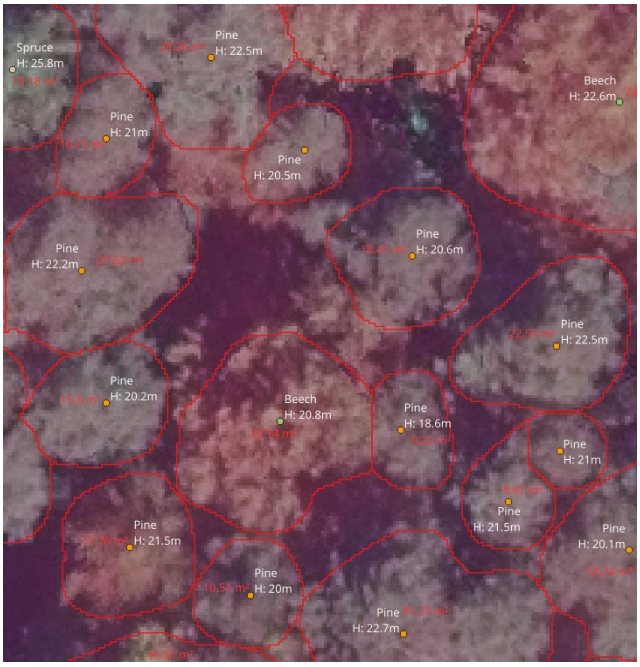

Figure 1: Predicted tree centers and crowns, with stand information height, species and crown area, with underlying CIR-aerial imagery

Conclusion & Outlook

The study contributes to ongoing discussions about integrating 3D information into 2D deep learning architectures. The method's innovation lies in end-to-end ITC processing and uncertainty-aware prediction. From a stakeholder perspective, it supports both research and operational forest management by improving transparency and reliability in tree crown delineation. Future work will investigate the scalability of the approach for large-scale forest monitoring.

Acknowledgment

We greatly acknowledge the support from the Bayerische Staatsforsten AöR in providing the aerial image and LIDAR data for this study.

References

Brandt, Martin et al. (Nov. 5, 2020). "An Unexpectedly Large Count of Trees in the West African Sahara and Sahel". In: Nature 587, pp. 1–5. doi: 10.1038/s41586-020-2824-5.

Freudenberg, Maximilian, Paul Magdon, and Nils Nölke (Dec. 1, 2022). “Individual Tree Crown Delineation in High-Resolution Remote Sensing Images Based on U-Net”. In: Neural Computing and Applications 34.24, pp. 22197–22207. issn: 1433-3058. doi: 10.1007/s00521-022-07640-4.

Li, Sizhuo et al. (Mar. 9, 2023). “Deep Learning Enables Image-Based Tree Counting, Crown Segmentation, and Height Prediction at National Scale”. In: PNAS Nexus 2.4.

Lucas, Moritz et al. (Aug. 2024). “Individual Tree Detection and Crown Delineation in the Harz National Park from 2009 to 2022 Using Mask R–CNN and Aerial Imagery”. In: ISPRS Open Journal of Photogrammetry and Remote Sensing 13, p. 100071. issn: 26673932. doi: 10.1016/j.ophoto.2024.100071.

Schindler, Jan et al. (Aug. 1, 2025). “Efficient Tree Mapping through Deep Distance Transform (DDT) Learning”. In: ISPRS Open Journal of Photogrammetry and Remote Sensing 17, p. 100095. issn: 2667-3932. doi: 10.1016/j.ophoto.2025.100095.

Tong, Fei and Yun Zhang (Mar. 15, 2025). “Individual Tree Crown Delineation in High Resolution Aerial RGB Imagery Using StarDist-based Model”. In: Remote Sensing of Environment 319, p. 114618. issn: 0034-4257. doi: 10.1016/j.rse.2025.114618.

Wielgosz, Maciej et al. (Nov. 1, 2024). “SegmentAnyTree: A Sensor and Platform Agnostic Deep Learning Model for Tree Segmentation Using Laser Scanning Data”. In: Remote Sensing of Environment 313, p. 114367. issn: 0034-4257. doi: 10.1016 / j.rse.2024.114367.

Xiang, Binbin et al. (May 1, 2024). “Automated Forest Inventory: Analysis of High-Density Airborne LiDAR Point Clouds with 3D Deep Learning”. In: Remote Sensing of Environment 305, p. 114078. issn: 0034-4257. doi: 10.1016/j.rse.2024.114078