Validation challenges in large-scale tree crown segmentations from remote sensing imagery using Deep Learning: a case study in Germany

Taimur Khan[1] (✉)[0000−0001−7833−5474], Jasmin Krebs[2][0009−0003−9330−1231], Sharad Kumar Gupta[1,3][0000−0003−3444−1333], Jonathan Renkel[4][0009−0002−1513−9497], Caroline Arnold[5][0000−0002−9458−1517], and Nils Nölke[6][0000−0003−4925−2287]

[1] Helmholtz Centre for Environmental Research (UFZ), Theodor-Lieser-Str. 4, 4109 Halle (Saale), Germany taimur.khan@ufz.de
[2] Leipzig University, Augustusplatz 10, 4109 Leipzig, Germany jk21byxu@studserv.uni-leipzig.de
[3] HZDR - Centre for Advanced System Understanding (CASUS), Untermarkt 20, 2826 Görlitz, Germany sharad.gupta@ufz.de
[4] Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 4, 6120 Halle (Saale), Germany jonathan.renkel@geo.uni-halle.de
[5] Helmholtz-Zentrum Hereon, Max-Planck-Straße 1, 21502 Geesthacht, Germany caroline.arnold@hereon.de
[6] University of Göttingen, Forest Inventory and Remote Sensing, Büsgenweg 5, 37077 Göttingen, Germany nils.noelke@forst.uni-goettingen.de

Abstract. Recent advancements in deep learning have significantly enhanced the extraction of detailed ecological information from multispectral (RGBi) aerial imagery, notably the identification and segmentation of individual tree crowns. Using data from our extensive case study mapping approximately 218.7 million trees across the German federal states of Sachsen and Sachsen-Anhalt, we illustrate how scaling these models beyond controlled settings introduces critical validation challenges. Even minor inaccuracies in tree crown segmentation can substantially impact some practical applications in forestry management and urban planning. Our findings emphasize the influence of tree allometry, seasonal variability, shadow effects, and the characteristics of training dataset annotations on model accuracy. Addressing these issues necessitates careful selection and design of training data and exploration of task-specific pre-training through Foundation Models. We highlight these validation complexities and recommend for rigorous steps to ensure the reliability and practical utility of large-scale deep learning models in ecological and urban management contexts.

Keywords: Deep Learning · Ecology · Forestry · Remote Sensing.

1 Introduction

The urgency of climate change and biodiversity loss significantly heightens the importance of large-scale tree crown segmentation. Forests act as crucial car-

bon sinks and biodiversity reservoirs [18], while urban trees offer vital ecosystem services[20]. To generate comprehensive assessments of forest health, carbon stocks, and urban greenery, we need to accurately map tree crowns over vast areas, including trees both within forests and outside of them. These trees outside forests, often found in agricultural or urban settings, must be part of our inventories, making their detection and mapping a key priority for scaling local observations to a global context. Large-scale segmentation enables consistent, detailed monitoring of these vital resources, informing sustainable forest management, urban planning, and policy development. The ambition to create national and global inventories, such as those envisioned in [25], hinges on the reliability of such foundational segmentation data. In an era marked by rapid environmental change, comprehensive mapping efforts are indispensable. Tree crown segmentation is a cornerstone task for these objectives, as the size and structure of a tree crown—shaped by species-specific branching patterns, site conditions, and competition for light—directly influences primary production.

While recent advances in deep learning offer unprecedented capabilities for automated segmentation, their robust validation at scale presents a critical, unaddressed hurdle.

The integration of deep learning into remote sensing has opened new ways in how we monitor and analyse ecological systems [30,29]. Tree crown segmentation has especially benefited from advances in convolutional neural networks (CNNs) and transformer-based architectures. Latter was used by [8], for single tree detection. Approaches leveraging high-resolution imagery have demonstrated remarkable capabilities in delineating individual tree crowns [26,24,5], illustrating the potential for scalable and automated tree mapping.

However, while these deep learning models demonstrate impressive performance in controlled environments or with limited-scale datasets, their deployment at expansive spatial extents, such as regional or national scales, exposes a distinct spectrum of challenges that remain largely unaddressed. Foremost among these is validation — a critical yet underexplored hurdle in ensuring the reliability and generalizability of model outputs. As recent studies have shown, the performance of tree crown segmentation models is highly sensitive to the characteristics of the training data, such as tree size, species diversity, seasonal variations, vitality conditions, and image quality [16,2], a fact that directly reflects the profound complexity of how tree crowns appear in real-world conditions. To enable scalable validation frameworks, it is therefore necessary to accurately characterize and account for this complexity.

A key metric in this context is the crown projection area (CPA), defined as the vertical projection of the crown onto a horizontal plane. Deriving CPA provides essential insights both at the individual tree level—where it provides predictions of diameter, volume, and growth rates—and at the stand level, supporting models of competition and canopy gap dynamics [21,10]. However, inaccuracies inherent in the segmentation process, particularly when applied at large scales, can lead to biased estimations of these critical tree variables, underscoring the

urgent need to address robust validation challenges in large-scale tree crown segmentation.

Among the available crown metrics, crown spread is particularly valuable as an independent variable for validating segmentation results, as it captures essential information about tree size and structure. However, accurate estimation of crown spread is complex in both field and remote sensing contexts. Errors can arise not only during the segmentation process but also from the methods used to derive crown spread from segmented polygons, especially for irregularly shaped crowns. The choice of calculation method strongly influences the quality of crown spread estimates and, consequently, the reliability of validation metrics. Furthermore, the accuracy of validation is fundamentally dependent on the quality of reference data; coarse or inconsistent ground-truth datasets can substantially limit the precision of assessment.

Building on our extensive experience in applying deep learning models to large-scale tree crown segmentation across Germany at the federal-state level, this paper (1) identifies and characterizes the key challenges involved in assessing segmentation accuracy over expansive, heterogeneous terrains–including variability in canopy structure, imaging conditions, and regional ecological gradients–and (2) provides concrete methodological recommendations to address these challenges. By coupling a rigorous evaluation of terrain-specific segmentation limitations with these targeted recommendations, we aim to foster a dialogue on the methodological innovations required to ensure that deep learning advancements in remote sensing translate into actionable insights at the scales required by global environmental challenges.

2 Case Study: Tree Crown Segmentation in Saxony and Saxony-Anhalt, Germany

To evaluate the performance and limitations of large-scale crown segmentation and its validation, the DeepTrees model [24] was applied in a one-shot prediction approach using pretrained model weights provided by Freudenberg et al [5] as well as model weights trained in DeepTrees [24], applied to high-resolution multispectral digital orthophoto imagery (DOP20) covering the German federal states of Saxony (SN) and Saxony-Anhalt (ST) [12,11]. SN, covering approximately 18,450 km², and Saxony-Anhalt, spanning around 20,452 km², represent diverse ecological and urban landscapes, ideal for assessing large-scale segmentation model performance (Figure 3 in Appendices). The 4-channel (RGBi) DOP20 imagery, with a spatial resolution of 20 cm per pixel , enables precise tree crown delineation. The DeepTrees model identified approximately 218.7 million individual tree crowns—137.3 million in SN and 81.4 million in ST (Figure 1a). The resulting segmentation dataset has been made available upon request on Zenodo (https://doi.org/10.5281/zenodo.15638573).

Despite optimized methods, exact matches between predicted and ground-truth crown spreads were low (32%), though expanding the margin of error to ±5 m increased accuracy significantly (89%), highlighting validation challenges.

The segmentation model systematically overestimated small crowns (<6 m) and underestimated large crowns (>16 m), yet the crown area versus spread relationship remained stable (Figure 1b), indicating consistent segmentations. Validation quality proved crucial; notably, the local inventory ('Baumkataster' Halle/Saale) provided broad intervals (5 m), limiting validation precision but still achieving an overall IOU of ~70%.The segmented tree crowns show substantial regional variability in tree distribution, reflecting ecological, topographical, and land-use gradients (Figure 3). To evaluate the performance and limitations of large-scale crown segmentation and its validation, the DeepTrees model[24] was applied in a one-shot prediction approach using pretrained model weights provided by Freudenberg et al [5] as well as our own model weights, applied to high-resolution multispectral digital orthophoto imagery (DOP20) covering the German federal states of Saxony (SN) and Saxony-Anhalt (ST). SN, covering approximately 18,450 km², and ST, spanning around 20,452 km², represent diverse ecological and urban landscapes, ideal for assessing large-scale segmentation model performance (Figure 3 in Appendix). The 4-channel (RGBi) DOP20 imagery, with a spatial resolution of 20 cm per pixel, enables precise tree crown delineation. The DeepTrees model identified approximately 218.7 million individual tree crowns—137.3 million in SN and 81.4 million in ST (Figure 1a), and derived allometrical traits of crown area and crown spread (Figure 1b). A clear positive correlation between crown spread and crown area is observed, although deviations from the expected linear trend highlight segmentation uncertainties, particularly for small and irregularly shaped crowns. This variation underscores the need for improved methods to capture crown geometry and supports the use of allometric checks for quality assessment.

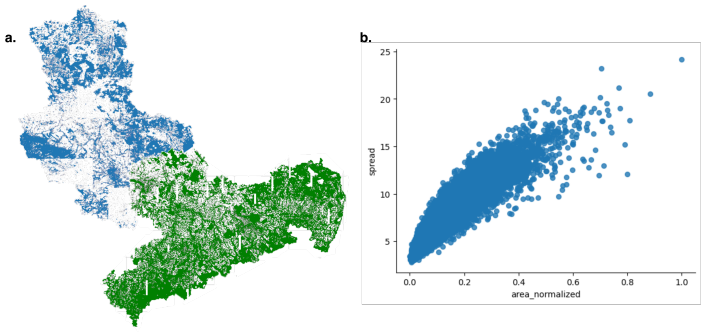

Fig. 1. (a) Spatial distribution of individual tree crown segmentation polygons derived from multispectral DOP20 imagery across the German federal states of Sachsen (SN, green) and Sachsen-Anhalt (ST, blue), totaling 218.7 million predicted crowns. (b) Relationship between normalized crown area (x-axis) and estimated crown spread (y-axis) for a subset of the segmented trees from Halle (Saale), ST.

3 Challenges

Validating large-scale tree crown segmentation models reveals a web of interrelated challenges. Among these challenges is phenology, a moving target: the same forest can look drastically different between leaf-on summer imagery and leaf-off winter scenes. Models trained on a single phenological period often struggle in another, yielding inconsistent segmentation accuracy across the year [19,13,6]. For example, a canopy delineation that performs well on lush summer foliage may under-segment sparse autumn crowns or miss bare branches in winter (Figure 3a). Such season-driven variability not only degrades model performance but also complicates validation – a one-shot model might appear accurate in one season and fail in the next, raising questions about how and when accuracy should be assessed. Incorporating multi-season data including the key phenological stages: leaf emergence, flowering, fruiting, autumn coloring and leaf fall, during both training and validation is essential, as phenological dynamics have been shown to strongly influence model generalizability [16,2].

Spatial and illumination heterogeneity of landscapes, driven by terrain and illumination differences, poses a major hurdle for both segmentation and its validation. An algorithm that segments tree crowns flawlessly in a flat, well-lit park may strugglein a shadow-drenched valley or on a steep mountainside. Variations in ground elevation and slope alter the angle of solar illumination, leading to uneven lighting and shadows that can confuse models. In mountainous or rugged terrains, trees on north-facing slopes might appear darker or partially occluded compared to those on south-facing slopes with direct sun, even if they are the same species and healthy. Such effects result in site-specific performance: accuracy drops when moving to new topographies or sensor angles. Weinstein et al. (2020) observed this kind of cross-site performance gap, where a tree detection model trained in one region underperformed when applied to a different region's imagery without adaptation, underscoring how terrain and context influence outcomes [27]. Fine-tuning can increase accuracy, but requires additional data and computational resources. For validation, this means that accuracy estimates from one area may not transfer to another – a critical issue when assessments at national or global scales are required.

A further fundamental challenge lies in the scarcity of accurate ground-truth data at scale. Reliable validation hinges on high-quality reference data (the "ground truth"), yet collecting detailed crown delineations over large regions is logistically difficult and expensive. For instance, field surveys can map individual trees with great precision, providing highly accurate tree position information (e.g., through GPS measurements of trunk location or detailed canopy spread on the ground) [23]. However, doing this over thousands of square kilometers is infeasible. Conversely, while aerial and satellite imagery offer broad coverage, even at high resolutions, they present significant challenges for unequivocally labeling each tree crown for validation purposes. This is due to factors like canopy overlap, varied lighting conditions, complex tree morphologies, and the sheer scale of the areas to be annotated, which introduce ambiguities and make consistent manual delineation impractical or prohibitively expensive across large regions.

UAV (drone) campaigns can bridge the gap by capturing very high-resolution images or LiDAR of sample areas, but they are limited in flight range and still require extensive human annotation to turn imagery into usable ground truth. The net result is a mismatch of scales: our models aspire to map every tree across entire countries, but our ground truth typically covers only small plots or scattered samples [ˆ]. This mismatch means that validating a "wall-to-wall" tree map often involves extrapolating from a tiny fraction of ground-referenced trees, introducing uncertainty. Moreover, ground-reference datasets may not capture the full diversity of conditions (species, canopy shapes, management regimes, etc.) present in the larger mapping area, biasing the validation. Expanding ground-truth collection – through automated methods or crowdsourcing – is thus not just a recommendation but a necessity to overcome this validation bottleneck (as we discuss later).

Compounding the issue of limited data is the inconsistency in reference annotations and evaluation metrics. Even when ground-reference data exist, their format can differ – sometimes reference trees are marked by a single GPS point (e.g., trunk location), sometimes by a hand-drawn polygon outlining the single crown. This creates a challenge in validation: how do we decide if a predicted crown polygon "matches" a ground-truth point, or how to handle cases where one field-mapped tree corresponds to multiple overlapping crown segments in the image? (Figure 2c) Conversely, field crews might delineate a broad canopy as one crown while an automated model splits it into two segments (or vice versa), especially in dense stands where crowns merge. These ambiguities in one-to-one correspondence make it hard to define what a "correct" segmentation is. Traditional pixel-wise accuracy metrics like Intersection-over-Union (IoU) treat segmentation purely as an image overlap problem, which may not reflect the ecological reality of counting individual trees. IoU penalizes differences in shape or area but doesn't account for whether the count of tree objects is correct. In an extreme case, a model could slightly over-segment every tree (splitting each true crown into two smaller polygons) and still achieve a reasonable IoU, despite doubling the perceived tree count – a significant error for applications. On the other hand, object-centric metrics such as panoptic segmentation quality attempt to consider both detection and delineation of objects [9]. Panoptic metrics combine aspects of object detection (was each tree detected?) with segmentation quality (was each crown correctly outlined?), which can be more appropriate for tree mapping. However, even these require well-defined ground-truth objects to compare against. When the ground truth itself is inconsistent (e.g., how to count a clumped cluster of stems with overlapping crowns), validation metrics struggle to fully capture model performance. The choice of evaluation metric thus becomes non-trivial: depending on whether one prioritizes exact crown shape, tree count, or canopy cover, the "best" metric may differ. Establishing consensus on evaluation protocols is part of the challenge – without it, different studies may report accuracy in incompatible ways.

There is also the issue of scale and resolution in validation reporting. A model's accuracy can appear to vary depending on the spatial scale at which

it is evaluated. For instance, a segmentation model might achieve high overall accuracy when averaged over an entire large region, yet if one zooms into a small test area (say a single city park or forest stand), the error rate might be much higher or lower. This can happen if errors are not evenly distributed: the model could perform very well in one type of landscape (e.g., neat urban street trees) and poorly in another (dense natural forest), and a coarse regional average could mask these extremes. Consequently, a user working on a local conservation project might experience worse performance than the "headline" accuracy suggests, because that headline number was diluted by many easier cases elsewhere. Ensuring that validation is robust across scales is tricky – one must balance broad coverage with local detail. It calls for multi-scale validation approaches, where accuracy is reported at multiple grain sizes or stratified by landscape or habitat type.

The above challenges reflect the multifaceted difficulties of validating tree crown segmentation at scale. For an at-a-glance overview of these issues and recommendations, see Table 1 in the Appendix.

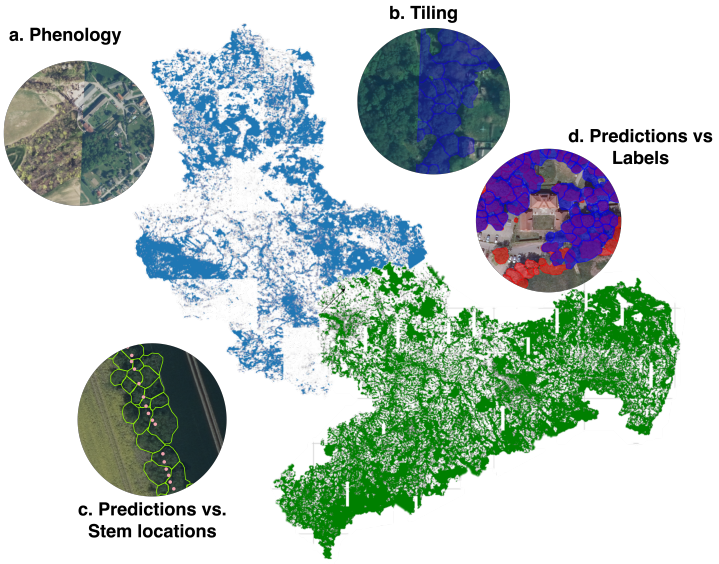

Fig. 2. Illustration of key validation challenges in large-scale tree crown segmentation across two federal states (Sachsen-Anhalt; in blue & Sachsen; in green) in Germany. Insets highlight specific issues: (a) seasonal variation (phenology) affecting appearance, (b) tiling artifacts in image preprocessing, (c) misalignment between predicted crowns and known tree stem locations, and (d) discrepancies between predicted crowns (blue) and labeled ground truth (red), emphasizing model performance gaps. Blue (Sachsen-Anhalt) and green (Sachsen) overlays represent modelled individual tree crowns across different Federal States in Germany.

4 Recommendations

Validating large-scale tree segmentation models is challenging due to seasonal variability, diverse terrain (Figure 3b,c)[3] and illumination conditions, inconsistent ground truth data, ambiguous evaluation criteria, and scale-dependent performance. These factors significantly affect the reliability and practical value of AI-generated tree maps for forestry, ecology, and urban planning.

Addressing these challenges necessitates a multi-faceted approach that integrates advanced modeling techniques, improved validation methods, and robust data collection. Recommended strategies include enhancing model robustness through sophisticated training approaches and developing standardized evaluation frameworks to assess real-world performance accurately. These efforts aim to ensure that advances in deep learning yield reliable, actionable insights for large-scale tree mapping.

Leveraging self-supervised learning (SSL) enhances the generalizability of segmentation models by utilizing abundant unlabeled remote sensing data. Unlike traditional supervised methods constrained by limited annotations, SSL enables models to derive meaningful representations through tasks such as predicting missing image parts or distinguishing augmented views of scenes. Pre-training models with extensive unlabeled geospatial imagery allows them to internalize intrinsic landscape patterns—textures, shapes, and seasonal variations—which can be efficiently fine-tuned with fewer labeled samples for specific tasks like tree crown segmentation. Recent research underscores the effectiveness of SSL-based geospatial foundation models. PhilEO Bench demonstrated improved performance across multiple remote sensing applications, such as building footprint extraction and road mapping, compared to training models from scratch [4]. SSL-trained models inherently recognize basic vegetation structures, shadows, and seasonal dynamics, facilitating accurate segmentation under varying conditions. For example, a model trained on seasonal satellite imagery implicitly distinguishes between a tree's winter and summer appearances, significantly reducing the required fine-tuning. Expanding this approach, Mendieta et al. (2023) employed continual SSL training, integrating new data distributions to develop a robust Geospatial Foundation Model (GFM) proficient across diverse remote sensing tasks [15]. This continual learning paradigm ensures segmentation models remain current with evolving landscapes and sensor technologies. Thus, SSL effectively addresses challenges associated with limited labeled data and dataset biases, improving adaptability and performance of large-scale segmentation models.

Integrating multi-view and multi-temporal data enhances model consistency by training with diverse imagery of the same trees captured from varying angles, sensors, or times. Multi-view data encompass multi-angle (nadir and oblique aerial images), multi-platform (satellite and drone imagery), and multi-temporal (images from different seasons or years) perspectives, enabling models to learn robust invariances for reliable segmentation under varying conditions. Combining leaf-on and leaf-off images, compels the model to recognize structural features beyond mere greenness, thus improving seasonal generalization.

Self-supervised learning (SSL) techniques, such as masked autoencoders and contrastive learning, effectively utilize multi-view datasets by encouraging models to generate consistent representations across different views without manual labels. Studies utilizing masked image modeling and contrastive SSL on multi-view satellite imagery have demonstrated significant performance improvements in segmentation and detection tasks [17,4]. Employing multi-view consistency training directly mitigates challenges associated with phenological and illumination variations, thereby enhancing model robustness and facilitating validation. For example, discrepancies in model predictions between leaf-on and leaf-off imagery could flag areas requiring further examination. Thus, multi-view and multi-temporal training provide models with contextual understanding of the dynamic and three-dimensional nature of trees, stabilizing segmentation results across real-world variability.

Integrating terrain data into tree segmentation models addresses image variability caused by uneven topography. Fusing digital elevation models (DEMs) or LiDAR-derived terrain data with imagery during model training can enhance segmentation accuracy. This can involve providing elevation/slope as an additional input channel or designing models to process terrain context separately. Self-supervised pre-training can leverage elevation data to improve feature representations. By differentiating between bare earth and above-ground structures, models learn to distinguish actual objects from illumination differences caused by slope and aspect. For instance, a terrain-informed model can differentiate shaded hillsides from canopy gaps or recognize a single tree crown on a steep slope despite perspective distortion [7]. Incorporating terrain data directly addresses spatial heterogeneity, providing a reference frame to normalize variability and improve validation. This allows for error analysis stratified by terrain class, ensuring consistent performance across diverse topographies. We recommend that future segmentation models, particularly for regions with varied topography, adopt terrain-aware training strategies. Even with limited DEM availability, approximating slope from imagery or using coarse global elevation data can be beneficial. Ultimately, integrating real-world topography with pixel appearance enhances model understanding and generalizability, leading to more robust deployment in new landscapes.

Focus model attention on domain-specific features. Advancements in model architecture and training objectives, such as feature-guided masked autoencoders, can enhance segmentation reliability by directing learning toward high-level, domain-relevant features rather than pixel-level noise. In remote sensing, specific spectral and textural cues (e.g., high near-infrared reflectance in healthy canopies) distinguish tree crowns from other land covers. Feature-guided methods train models to reconstruct meaningful feature representations—such as vegetation indices or edge maps—instead of raw pixels. For example, FG-MAE (Feature Guided Masked Autoencoder) tasks the model with predicting domain-specific features (e.g., NDVI or engineered representations) for masked image regions, promoting semantic understanding rather than texture replication [17]. Applied to tree segmentation, such pre-training focuses internal representations

on vegetation structure (e.g., crown edges or canopy height), improving delineation of complex canopies and separation from backgrounds. Early studies in multispectral and SAR imagery confirm improved segmentation in challenging environments [1]. Integrating feature-guidance is thus recommended for large-scale tree mapping, particularly in complex landscapes, as it not only increases accuracy but may also yield more interpretable outputs and uncertainties to support validation and error diagnosis.

Expand validation beyond pixel agreement – use ecological consistency checks. Traditional segmentation validation relies on geometric overlap with ground truth, but tree mapping can benefit from leveraging ecological allometric relationships as an additional check. Allometry—well-established links between tree dimensions such as trunk diameter, height, and crown width—provides expected bounds for tree size relationships. Applying allometric equations to segmentation outputs serves as a "sanity check": for example, predictions where crown sizes and tree heights deviate significantly from field-based distributions may indicate model errors or missing trees. Overly large predicted crowns in a region without tall trees, or consistently small crowns in old-growth forests, can reveal segmentation artifacts. Recent studies, such as Song et al. (2023), have used statistical models to connect remote sensing outputs to allometric expectations, highlighting ecological validation as a valuable approach [22]. Implementing such checks requires integrating ancillary data, such as LiDAR-derived heights or species-specific formulas, to cross-validate AI-generated tree maps. This recommendation shifts validation from a purely computer-vision perspective to an application-oriented perspective: after all, if the ultimate goal is to use these maps for carbon accounting, biodiversity, or forestry, then passing an ecological reality check is as important as scoring well on IoU.

Establish community benchmarks and evaluation frameworks. The field of geospatial AI is recognizing the value of standard benchmarks – datasets and metrics on which different methods can be compared in a reproducible way. Standard datasets and metrics like ImageNet and COCO have significantly impacted computer vision, inspiring geospatial counterparts such as GEO-Bench and PANGAEA [14]. We propose developing specific benchmarks for tree crown segmentation and mapping, covering diverse landscapes, seasons, and remote sensing sources. Standardized evaluation metrics, such as IoU for segmentation and detection accuracy, would ensure comparability and transparency in results. Additionally, benchmarks should require multi-scale assessments, evaluating not only overall accuracy but also performance in challenging subsets (e.g., dense forests vs. isolated trees), incorporating auxiliary criteria like allometric consistency. Regular competitive challenges using standardized large-scale tree mapping tasks can accelerate method improvement and robustness. Ultimately, these community-driven benchmarks facilitate transparent, credible model evaluations and iterative progress in geospatial AI.

Innovate in ground-truth data collection and labeling. Advances in modeling alone cannot eliminate the need for enhanced validation datasets. We recommend expanding ground-truth collection through automation, crowdsourc-

ing, and active learning. Automated techniques using drones or AI-equipped aircraft can rapidly generate tree crown annotations, minimizing expert involvement. Crowdsourcing allows non-experts to efficiently label straightforward cases, supported by redundancy and quality controls, proven effective in urban tree mapping. Active learning strategies prioritize uncertain or conflicting model predictions for expert review, significantly optimizing annotation efforts [28]. Additionally, promoting open data practices and sharing existing inventories and LiDAR datasets from public agencies can provide critical validation resources. Integrating these approaches—automation, crowdsourcing, active learning, and open data—will effectively bridge validation data gaps and enhance model robustness and generalizability.

5 Conclusion

Deep learning has opened a new frontier for tree crown segmentation from remote sensing imagery, yet its promise is contingent on our ability to validate these models reliably and at scale. As this paper has outlined, the challenges are not merely technical—they are epistemological. They force us to ask: what does it mean to "know" a tree from above, when the canopy is a moving target shaped by phenology, terrain, land cover and image artifacts?

The comprehensive case study in Sachsen and Sachsen-Anhalt exemplifies how substantial regional variability in ecology, topography, and land use affects model performance. Validation challenges identified—including phenological dynamics, spatial heterogeneity, and scale-dependent accuracy—highlight the need for robust validation frameworks tailored to large spatial extents.

Emerging methods such as self-supervised learning, geospatial foundation models, and multi-view fusion offer a compelling pathway forward. These approaches not only reduce reliance on costly annotations but also capture the underlying structure of complex and dynamic landscapes. Equally, validation must evolve beyond static benchmarks. Indirect metrics—like allometric plausibility checks— must be brought into the fold. Validation, in this context, becomes less about binary correctness and more about probabilistic trust.

Pursuing the recommendations, the field can significantly advance accurate and trustworthy large-scale tree mapping by integrating robust, self-supervised, multi-view, and terrain-aware models with continuous validation against classical metrics and real-world plausibility, using feedback for active learning and expanded training data. The real frontier is integration: aligning spatial, temporal, and ecological knowledge through a fusion of data-driven and domain-aware models. As these systems are deployed across continents, cities, and seasons, the imperative is not just to scale algorithms, but to scale insight. Only then can tree segmentation models become dependable instruments for managing the living infrastructure of our planet.

Acknowledgments. DeepTrees is part of the DeepTrees: Deep-Learning based spatiotemporal tree inventorying and monitoring from public orthoimages project, funded

by the Integration Platform "Sustainable Future Land Use" at Helmholtz-Centre for Environmental Research – UFZ within the Programme oriented Funding (PoF) period IV of the Helmholtz Program "Changing Earth – Sustaining our Future", Topic 5 "Landscapes of the Future. This repository is based on the work described in Freudenberg et al. (2022). This work was supported by Helmholtz Association's Initiative and Networking Fund through Helmholtz AI [grant number: ZT-I-PF-5-01]. This work used resources of the Deutsches Klimarechenzentrum (DKRZ) granted by its Scientific Steering Committee (WLA) under project ID AIM.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Allen, M.J., Owen, H.J., Grieve, S.W., Lines, E.R.: Manual labelling artificially inflates deep learning-based segmentation performance on rgb images of closed canopy: Validation using tls. arXiv preprint arXiv:2503.14273 (2025)
2. Cong, P., Zhou, J., Li, S., Lv, K., Feng, H.: Citrus tree crown segmentation of orchard spraying robot based on rgb-d image and improved mask r-cnn. Applied Sciences **13**(1), 164 (2022)
3. European Space Agency (ESA): Copernicus DEM GLO-30 - Global Digital Elevation Model (30 m). Copernicus Open Access Hub (2020), https://spacedata.copernicus.eu/collections/copernicus-digital-elevation-model, version: GLO-30, Public Release
4. Fibaek, M., et al.: Phileo bench: Evaluating geospatial foundation models. IEEE Transactions on Geoscience and Remote Sensing (2024)
5. Freudenberg, M., Magdon, P., Nölke, N.: Individual tree crown delineation in high-resolution remote sensing images based on u-net. Neural Computing and Applications **34**(24), 22197–22207 (2022)
6. Garnot, V.S.F., Spafford, L., Lever, J., Sigg, C., Pietragalla, B., Vitasse, Y., Gessler, A., Wegner, J.D.: Deep learning meets tree phenology modelling: Phenoformer versus process-based models. Methods in Ecology and Evolution (2025)
7. Han, X., et al.: ms-gfm: Multisensor geospatial foundation models. IEEE Transactions on Geoscience and Remote Sensing (2024)
8. Jiang, T., Freudenberg, M., Kleinn, C., Lüddecke, T., Ecker, A., Nölke, N.: Detection transformer-based approach for mapping trees outside forests on high resolution satellite imagery. Ecological Informatics **87**, 103114 (2025). https://doi.org/https://doi.org/10.1016/j.ecoinf.2025.103114, https://www.sciencedirect.com/science/article/pii/S1574954125001232
9. Kirillov, A., He, K., Girshick, R., Rother, C., Dollár, P.: Panoptic segmentation (2019), https://arxiv.org/abs/1801.00868
10. Krajicek, J.E., Brinkman, K.A., Gingrich, S.F.: Crown competition—a measure of density. Forest science **7**(1), 35–42 (1961)
11. Landesamt für Geobasisinformation Sachsen (GeoSN): DOP20 - Digitale Orthophotos (RGBI, 20cm), Freistaat Sachsen. GeoBasis-DE / GeoSN (2022), https://www.landesvermessung.sachsen.de/digitale-orthophotos-bildflug-2022-8995.html, rGB and Color-Infrared orthophotos, National data license (DL-DE-BY-2.0)

12. Landesamt für Vermessung und Geoinformation Sachsen-Anhalt (LVermGeo): DOP20 - Digitale Orthophotos 20 cm, Sachsen-Anhalt. GeoBasis-DE / LVermGeo ST (2020), https://www.lvermgeo.sachsen-anhalt.de, licensed under Datenlizenz Deutschland - Namensnennung - Version 2.0
13. Liu, G., Migliavacca, M., Reimers, C., Kraft, B., Reichstein, M., Richardson, A.D., Wingate, L., Delpierre, N., Yang, H., Winkler, A.J.: Deepphenomem v1. 0: deep learning modelling of canopy greenness dynamics accounting for multi-variate meteorological memory effects on vegetation phenology. Geoscientific Model Development **17**(17), 6683–6701 (2024)
14. Marsocci, V., Jia, Y., Le Bellier, G., Kerekes, D., Zeng, L., Hafner, S., Gerard, S., Brune, E., Yadav, R., Shibli, A., Fang, H., Ban, Y., Vergauwen, M., Audebert, N., Nascetti, A.: PANGAEA: A global and inclusive benchmark for geospatial foundation models (2024). https://doi.org/10.48550/arXiv.2412.04204, https://arxiv.org/abs/2412.04204
15. Mendieta, M., Han, B., Shi, X., Zhu, Y., Chen, C., Li, M.: GFM: Building geospatial foundation models via continual pretraining. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2023). https://doi.org/10.1109/ICCV.2023.12345, https://arxiv.org/abs/2302.04476
16. Moussaid, A., Fkihi, S.E., Zennayi, Y.: Tree crowns segmentation and classification in overlapping orchards based on satellite images and unsupervised learning algorithms. Journal of imaging **7**(11), 241 (2021)
17. Mukkavilli, K., Lambhate, D., Das, K., Bangalore, R., Oliveira, D., Muszynski, M., Ankur, K., Ramasubramanian, M., Gurung, I., Khallaghi, S., Li, H., Cecil, M., Ahmadi, M., Kordi, F., Alemohammad, H., Maskey, M., Ganti, R., Weldemariam, K., Ramachandran, R.: Foundation models for generalist geospatial artificial intelligence (2023). https://doi.org/10.48550/arXiv.2310.18660, https://arxiv.org/abs/2310.18660
18. Pan, Y., Birdsey, R.A., Phillips, O.L., Houghton, R.A., Fang, J., Kauppi, P.E., Keith, H., Kurz, W.A., Ito, A., Lewis, S.L., et al.: The enduring world forest carbon sink. Nature **631**(8021), 563–569 (2024)
19. Sapkota, R., Karkee, M.: Integrating yolo11 and convolution block attention module for multi-season segmentation of tree trunks and branches in commercial apple orchards. arXiv preprint arXiv:2412.05728 (2024)
20. Sharma, S., Hussain, S., Kumar, P., Singh, A.N.: Urban trees' potential for regulatory services in the urban environment: an exploration of carbon sequestration. Environmental Monitoring and Assessment **196**(6), 1–27 (2024)
21. Shimano, K.: Analysis of the relationship between dbh and crown projection area using a new model. Journal of Forest Research **2**(4), 237–242 (1997)
22. Song, Q., Albrecht, C.M., Xiong, Z., Zhu, X.X.: Biomass estimation and uncertainty quantification from tree height. arXiv preprint arXiv:2305.09555 (2023). https://doi.org/10.48550/arXiv.2305.09555, https://arxiv.org/abs/2305.09555
23. Steier, J., Goebel, M., Iwaszczuk, D.: Is your training data really ground truth? a quality assessment of manual annotation for individual tree crown delineation. Remote Sensing **16**(15), 2786 (2024)
24. Taimur Khan, Arnold, C., Grover, H.: Deeptrees: Tree crown segmentation and analysis in remote sensing imagery with pytorch. Preprint (2025). https://doi.org/10.13140/RG.2.2.32837.36329, https://rgdoi.net/10.13140/RG.2.2.32837.36329
25. Tolan, J., Yang, H.I., Nosarzewski, B., Couairon, G., Vo, H.V., Brandt, J., Spore, J., Majumdar, S., Haziza, D., Vamaraju, J., et al.: Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. Remote Sensing of Environment **300**, 113888 (2024)

26. Weinstein, B.G., Marconi, S., Bohlman, S., Zare, A., White, E.: Individual tree-crown detection in rgb imagery using semi-supervised deep learning neural networks. Remote Sensing **11**(11), 1309 (2019)
27. Weinstein, B.G., Marconi, S., Bohlman, S.A., Zare, A., White, E.P.: Cross-site learning in deep learning rgb tree crown detection. Ecological Informatics **56**, 101061 (2020)
28. Wu, J., Chen, J., Huang, D.: Entropy-based active learning for object detection with progressive diversity constraint. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9397–9406 (June 2022)
29. Zhao, H., Morgenroth, J., Pearse, G., Schindler, J.: A systematic review of individual tree crown detection and delineation with convolutional neural networks (cnn). Current Forestry Reports **9**(3), 149–170 (2023)
30. Zheng, J., Yuan, S., Li, W., Fu, H., Yu, L., Huang, J.: A review of individual tree crown detection and delineation from optical remote sensing images: Current progress and future. IEEE Geoscience and Remote Sensing Magazine (2024)

Appendix

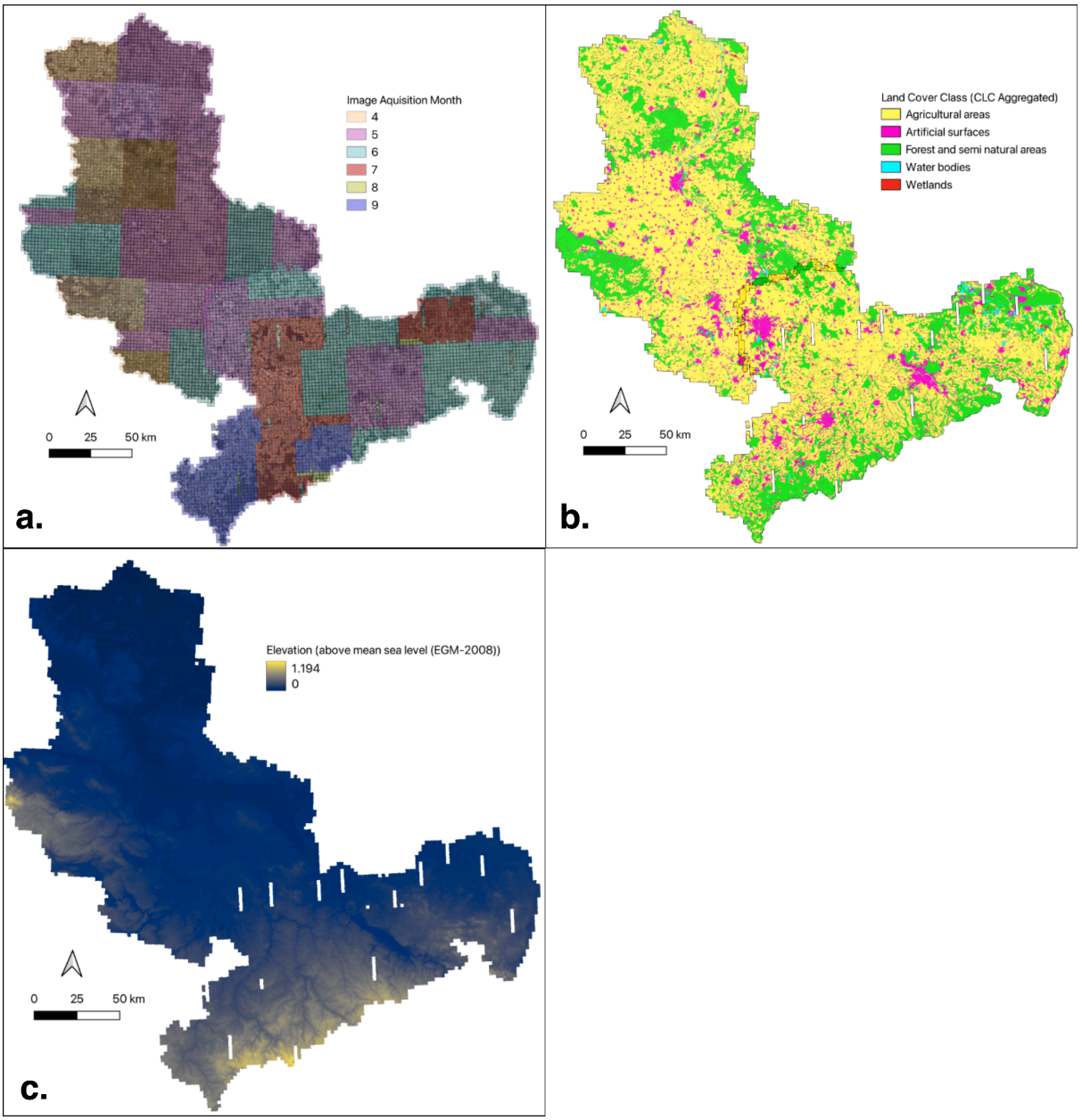

Fig. 3. Environmental heterogeneity across the German federal states of Sachsen and Sachsen-Anhalt, where tree crown segmentation was conducted. (a) Acquisition months of DOP20 imagery (April to September) used in the segmentation, showing strong spatial variation in image phenology [11]; [12]. (Middle) Aggregated land cover classes based on CORINE Land Cover 2020 data, grouped into five categories: agricultural areas, artificial surfaces, forest and semi-natural areas, water bodies, and wetlands [3]. (Bottom) Elevation map derived from the COPDEM GLO-30 digital elevation model at 30m spatial resolution, representing terrain variability used to assess segmentation accuracy across topographic gradients [4].

Table 1: Summary of key validation challenges in large-scale tree crown segmentation, with examples, relevant references, and recommended mitigation strategies.

Challenge	Description	Example/ Implications	References	Recommendation	References
Seasonal Phenology Variability	Tree appearance changes drastically across seasons (leaf-on summer vs. leaf-off winter), causing models trained in one phenological period to underperform in another. This leads to inconsistent segmentation accuracy and complicates validation, as a model may appear accurate in one season but fail in another. Multi-season data covering key phenological stages is essential for generalizable models.	A canopy delineation that performs well on lush summer foliage may under-segment sparse autumn crowns or miss bare branches in winter	[19;13;16;2]	Leverage self-supervised learning (SSL) and integrate multi-view/multi-temporal data. Use SSL to pre-train models on diverse, unlabeled imagery, and incorporate multi-view/multi-season data to build invariance to phenological and illumination changes. This enhances generalizability and reduces the need for extensive labeled data.	[5;15;17]
Spatial and Illumination Heterogeneity	Differences in terrain (elevation, slope) and illumination (shadows, sun angle) affect both segmentation and validation. Models may perform well in flat, well-lit areas but poorly in shadowed valleys or steep slopes, resulting in site-specific performance and complicating transferability and validation at larger scales.	A model that segments tree crowns flawlessly in a flat, well-lit park may stumble in a shadow-drenched valley or on a steep mountainside.	[28]	Integrate terrain data and multi-view consistency. Fuse digital elevation models (DEMs) or LiDAR with imagery to provide terrain context, and train models for multi-view consistency to handle illumination and topographic variability. This improves robustness and allows for stratified validation by terrain class.	[7;17;5]
Scarcity of Accurate Ground Truth Data	Collecting detailed crown delineations over large regions is logistically difficult and expensive. Field surveys are precise but not scalable; aerial/satellite imagery	Ground truth typically covers only small plots, making it difficult	[23]	Innovate in ground-truth data collection and labeling. Combining automation (e.g., AI-equipped drones), crowdsourcing, and active learning to expand and diversify validation	

	lacks resolution for unequivocal labeling due to factors like canopy overlap, varied lighting conditions, complex tree morphologies, and the sheer scale of the areas to be annotated. This leads to a mismatch between model output scale and validation data, introducing uncertainty and bias.	to validate national-scale maps.		datasets. Open data sharing and targeted expert annotation can further close the validation gap.	
Inconsistency in Reference Annotations and Evaluation Metrics	Reference data formats vary (GPS points vs. polygons), leading to ambiguity in matching predictions to ground truth. Dense stands and overlapping crowns complicate one-to-one correspondence. Pixel-wise metrics (e.g., IoU) may not reflect ecological reality, while object-centric metrics (e.g., panoptic segmentation) require well-defined ground-truth objects. Lack of consensus on evaluation protocols results in incompatible accuracy reports across studies.	A model could over-segment every tree and still achieve a good IoU, but double the perceived tree count which is problematic for applications requiring accurate counts such as for managed forest or botanical gardens.	[9;24]	Establish community benchmarks and evaluation frameworks. Develop standardized datasets, protocols, and metrics for tree segmentation. Community benchmarks should cover diverse landscapes and seasons, and include multi-scale and ecological consistency checks for robust, comparable validation.	[14;5]
Scale and Resolution Dependence in Validation Reporting	Model accuracy varies with the spatial scale of evaluation. Regional averages can mask local errors; performance may be high overall but poor in specific challenging areas. Errors may be unevenly distributed across landscape types, making it necessary to use multi-scale validation approaches and stratified reporting.	A model might perform well regionally but poorly in a specific forest stand or urban park, misleading local users.		Expand validation beyond pixel agreement using ecological consistency checks and multi-scale reporting. Integrate allometric relationships and stratified accuracy reporting to ensure ecological plausibility and transparency at multiple spatial scales.	[22;14]