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

Taimur Khan, Jasmin Krebs, Nils Nölke, Sharad Kumar Gupta, Jonathan Renkel, Caroline Arnold (order tbd)

Abstract. Advances in deep learning models have opened up new avenues for computer vision tasks, like tree crown segmentations from remote sensing images. As more and more models become available, the challenges of applying the models to really large spatial scales (beyond the forest/plot level) become visible. One of these challenges is the validation of the tree crown detection and segmentation. Imprecise detections/segmentations can lead to inaccuracies in downstream tasks in forestry and urban planning, creating a barrier for the application of these models to real-life use cases. Attributes of training data, e.g tree sizes, phenological stages, number of polygon vertices, shadows, impact the prediction accuracy of the segmentation models, hence training data should be chosen with due diligence. Here we identify some of the key validation challenges based on the collective experiences of not just training deep learning models, but also applying the models to very large spatial scales in Germany (federal state level).

Keywords: First Keyword · Second Keyword · Another Keyword.

1 Introduction

Accurately mapping tree crowns over large areas is critical for scaling local observations to regional and global assessments of forest health, carbon stocks, and urban greenery. 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. 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. 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. While recent advances in deep learning offer unprecedented capabilities for automated segmentation, their robust validation at scale presents a critical, unaddressed hurdle.

Among the various tree crown variables, crown spread presents unique estimation challenges, both in field measurements and, critically, when derived from remote sensing imagery. Its accurate derivation is not only susceptible to inaccuracies in the initial segmentation process but also relies heavily on the methods employed to extract crown spread from segmented polygons, a process fraught with its own complexities. Selecting an appropriate crown spread calculation method is crucial, especially for irregularly shaped tree crown polygons. Even with optimized calculation methods, direct matches with ground-truth data proved low (e.g., 32% in our experience), although a broader acceptable margin of error (±5 m) captured 89% of cases. These results immediately point to difficulties in establishing precise validation metrics. Furthermore, model performance exhibits size-dependent biases: trees with small crowns (<6 m) are often overestimated, while large crowns (>16 m) tend to be underestimated. These discrepancies underscore the inherent limitations of deep learning segmentation models at the extremes of the size spectrum. Critically, the quality of validation data profoundly influences the assessment of crown spread accuracy. For instance, our experience with the local tree inventory ('Baumkataster' Halle/Saale), which provides only broad crown spread ranges (5 m intervals), severely constrains the precision of validation, highlighting the need for more granular and reliable ground-truth datasets.

The integration of deep learning into remote sensing has opened new ways in how we monitor and analyse ecological systems [[**?**], [**?**]]. Tree crown segmentation, a task central to assessing forest health, carbon stocks, and urban greenery, has especially benefited from advances in convolutional neural networks (CNNs) and transformer-based architectures. Approaches leveraging high-resolution imagery have demonstrated remarkable capabilities in delineating individual tree crowns, as evidenced by works such as [**?**], [**?**] and [**?**], illustrating the potential for scalable and automated tree mapping. The urgency of climate change and biodiversity loss further amplifies the importance of large-scale tree crown segmentation. Forests serve as critical carbon sinks and biodiversity reservoirs [[**?**]], while urban trees provide essential ecosystem services [[**?**]]. 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 [**?**], hinges on the reliability of such foundational segmentation data.

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, and image quality [[**?**]; [**?**]]. These factors, when coupled with the com-

plexities of real-world landscapes, underscore the need for robust and scalable validation frameworks.

This Perspective builds on our extensive experience in applying deep learning models to large-scale tree crown segmentation across Germany, addressing federal state-level landscapes. By synthesizing insights from model training, deployment, and validation, we highlight the nuanced challenges associated with assessing segmentation accuracy over expansive and heterogeneous terrains. These challenges are not mere technicalities—they bear significant implications for downstream applications in forestry, conservation, and urban planning. Through this work, we aim to not only highlight these critical validation hurdles but also 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 demanded by global environmental challenges.

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

In this case study, the DeepTrees model [[**?**]] was applied in a one-shot prediction approach using pretrained model weights provided by DeepTrees, applied to high-resolution multispectral digital orthophoto imagery (DOP20) covering the German federal states of Saxony (SN) and Saxony-Anhalt (ST). Sachsen, covering approximately 18,450 km^2, and Saxony-Anhalt, spanning around 20,452 km^2, represent diverse ecological and urban landscapes, ideal for assessing large-scale segmentation model performance.

DOP20 imagery provides detailed spatial information at 20 cm resolution per pixel, capturing visual and near infrared spectral regions essential for accurate tree crown delineation. Using these pretrained weights, the DeepTrees model identified approximately 218.7 million individual tree crowns—137.3 million in Saxony and 81.4 million in Saxony-Anhalt. The resulting segmentation dataset has been made publicly available on Zenodo (https://zenodo.org/record/exampledoi), facilitating further research and validation.

This comprehensive segmentation highlights substantial regional variability in tree distribution, reflecting ecological, topographical, and land-use gradients. The case study thus forms a practical foundation for addressing the validation challenges detailed subsequently, emphasizing the necessity for robust validation frameworks to ensure reliable ecological monitoring and informed decision-making based on large-scale automated analyses.

Figure 1: 6 panels (1 for each state). From left to right -> 1) tiles with date classification 2) Land use (CORINE) + isolines. 3) the tree segmentations.

3 Challenges

Validating large-scale tree crown segmentation models reveals a web of interrelated challenges that go beyond those encountered in small, controlled datasets.

Seasonal phenology is a moving target: the same forest can look drastically different between leaf-on summer imagery and leaf-off winter scenes. Models trained on one season often struggle in another, yielding inconsistent segmentation accuracy across the year. For example, a canopy delineation that performs well on lush summer foliage may under-segment sparse autumn crowns or miss bare branches in winter. 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 during both training and validation is essential, as phenological dynamics have been shown to strongly influence model generalizability [[**?**]; [**?**]].

Spatial heterogeneity of landscapes, including terrain and illumination differences, poses another major hurdle for both segmentation and its validation. An algorithm that segments tree crowns flawlessly in a flat, well-lit park may stumble in 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: models often need fine-tuning or can suffer accuracy drops when moved 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 [[**?**]]. 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. Field surveys can map individual trees with great precision (e.g., measuring trunks and canopy spread on the ground), but doing this over thousands of square kilometers is infeasible. Aerial and satellite imagery provide broader coverage for reference data, but even these often lack the resolution or fidelity to unequivocally label each tree crown for validation purposes. 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 val-

idation. 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 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 ref) 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 [[**?**]]. 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's evaluated. For instance, a segmentation model might achieve high overall accuracy when averaged over an entire 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 type. Highlighting this, one could imagine a figure plotting model accuracy as a function of spatial extent or across different habitat types, illustrating how performance can drop off in specific challenging subsets despite looking good overall. This emphasizes that heterogeneous performance is itself a challenge to acknowledge in validation: we need methods to detect where and why a model fails, not just an aggregate score.

To summarize, validating large-scale tree segmentation models is far from a straightforward task. Seasonal changes, diverse terrain and illumination, limited and inconsistent ground truth, ambiguous evaluation criteria, and scale-dependent performance all intertwine to create a demanding setting. These challenges are not just academic – they directly impact how much trust we can place in AI-generated tree maps for real-world decisions in forestry, ecology, and urban planning. Recognizing these pain points is the first step; the next is devising strategies to overcome them. We now turn to several recommendations aimed at improving both models and validation practices, with an eye toward bridging the gap between controlled experiments and the complexity of continental-scale deployments.

Table 1: Potential validation challenges and corresponding mitigation strategies. This table could list key challenges in validating large-scale tree segmentations alongside proposed approaches to address them. For example: (Challenge) Phenological variation between seasons – (Mitigation) use multi-temporal training data and season-specific validation sets; (Challenge) Heterogeneous terrain and shadowing – (Mitigation) integrate digital elevation models and terrain-aware modeling; (Challenge) Scarcity of ground-truth annotations – (Mitigation) leverage semi-automated labeling, crowdsourcing, and active learning to expand validation data; (Challenge) Ambiguity in crown delineation – (Mitigation) adopt object-based accuracy metrics and consensus protocols for what counts as a single tree; (Challenge) Scale-dependent performance – (Mitigation) evaluate models at multiple spatial scales and stratify results by landscape type. This overview would visually reinforce the narrative that each challenge has an identifiable path forward, linking the "Challenges" and "Recommendations" sections.

![][image1]

Figure 2: Incomplete

4 Recommendations

Addressing the above challenges requires a multi-pronged approach, combining improvements in modeling techniques with innovations in validation methodology and data collection. We outline here several complementary strategies that, together, can bolster the reliability of large-scale tree segmentation efforts. These recommendations emphasize building more robust models through advanced training paradigms, as well as creating better frameworks to evaluate and support those models in real-world conditions. The overarching goal is to

ensure that deep learning advancements in tree mapping translate into trustworthy, actionable insights at regional to global scales.

Leverage self-supervised learning for robust representations. A key step towards more generalizable segmentation models is tapping into unannotated data via self-supervised learning (SSL). Unlike traditional supervised training which is bottlenecked by limited labeled examples, SSL allows models to learn from the abundant pool of unlabeled remote sensing images – for example, by predicting missing pieces of an image or distinguishing augmented views of the same scene. By pre-training a model on large geospatial datasets without any human-provided labels, the model can absorb intrinsic patterns of the landscape: textures, shapes, seasonal changes, and other context that are common across images. These rich foundational representations can then be fine-tuned for tree crown segmentation with far fewer annotated samples than would otherwise be needed. Recent efforts like PhilEO Bench demonstrate the promise of this approach, evaluating geospatial foundation models that were pre-trained using SSL techniques [[**?**]]. The results show improved performance on a range of tasks (e.g. building footprint extraction, road mapping) after such pre-training, compared to models trained from scratch. In the context of trees, a model with an SSL-pretrained backbone may already "know" about basic vegetation structures, shadows, and seasonal appearances, making it more adept at delineating crowns under varied conditions. For instance, a foundation model trained on year-round satellite imagery might implicitly understand the difference between a leafless oak in winter and the same oak in summer, and thus require only a light fine-tuning to accurately segment each. Mendieta et al. (2023) take this further by continually pre-training on new data distributions (a form of continual learning), which helped build a Geospatial Foundation Model (GFM) that excelled across multiple remote sensing tasks [[**?**]]. Such continual SSL training could allow segmentation models to keep improving as more unlabeled data (e.g., new satellite images over time or from new regions) become available, staying up-to-date with changes in landscapes and sensor characteristics. Exploiting SSL is a powerful way to tackle the twin issues of limited labels and dataset bias, yielding a model that is better equipped for the diverse scenarios encountered in large-scale mapping.

Integrate multi-view and multi-temporal data for consistency. Another promising avenue is to train and evaluate models with multi-view inputs – that is, imagery of the same trees captured from different angles, sensors, or times. Multi-view here can mean different things: multi-angle (such as combining nadir and oblique aerial images), multi-platform (satellite imagery plus drone photos), or multi-temporal (images taken in different seasons or years). By exposing models to such spatiotemporal diversity, we can help them learn invariances that make segmentation more reliable. For example, a tree crown seen from directly above might have a certain shape, but from a side angle the outline might blend with neighbors; a model trained with both perspectives could learn to robustly identify the crown in either view. Likewise, combining leaf-on and leaf-off images of the same forest during training can force a model to rely

on structural features (branches, trunk hints, relative spacing) in addition to just greenness, thereby improving its ability to generalize across seasons. Self-supervised methods are particularly well-suited to capitalize on multi-view data because they can be set up to encourage the model to produce similar latent representations for different views of the same object. Techniques such as masked autoencoders or contrastive learning can be used on multi-view datasets to make a model predict or align one view from another, without any manual labels needed. Such approaches have already shown success: for instance, researchers have used masked image modeling and contrastive SSL on multi-view satellite imagery to boost performance on segmentation and detection tasks, essentially teaching the model that "these two different-looking images actually contain the same trees" [[**?**]; [**?**]]. By training with multi-view consistency, the model becomes more robust to viewpoint and temporal changes, which directly addresses the phenology and illumination challenges discussed earlier. Importantly, this not only helps the model's predictions but also strengthens validation, because a model that is consistent across views makes it easier to compare predicted and true crowns even when the reference data comes from a slightly different perspective or date. In practice, one could imagine a validation scheme where the agreement of a model's output between leaf-on and leaf-off images serves as an indicator of reliability: large discrepancies might flag areas for further human inspection. Overall, multi-view and multi-temporal training imbue models with a form of contextual intelligence about the 3D and time-varying nature of trees, making segmentation outcomes more stable across the real-world variability that large-scale applications inevitably encompass.

Incorporate terrain information into the modeling pipeline. As noted, uneven terrain can cause substantial variability in how trees appear in images, so bringing explicit knowledge of terrain into the segmentation process is a logical remedy. One recommendation is to fuse digital elevation models (DEMs) or LiDAR-derived terrain data with the imagery during model training. This could be as simple as providing the model with an extra channel of input encoding elevation/slope, or as complex as designing the model to separately process terrain context. Self-supervised pre-training can be extended here too: recent work on multisensor geospatial foundation models has shown that including elevation data in SSL (e.g., tasking the model to distinguish imagery of flat vs. mountainous areas) yields better feature representations for downstream tasks [[**?**]]. By differentiating between bare earth and above-ground structures in pre-training, the model learns to discount illumination differences that are purely due to slope and aspect, focusing instead on actual objects like trees. In a tree segmentation scenario, a terrain-informed model could recognize that a dark region in an image is a shaded hillside rather than a non-existent gap in canopy cover, or that an elongated shape on a steep slope is still a single tree crown albeit skewed by perspective. Incorporating terrain data directly addresses the challenge of spatial heterogeneity: it provides a reference frame to normalize out some variability. This can improve validation as well – for example, error analysis can be stratified by terrain class to ensure a model

works not just on average, but on hilltops and valleys alike. We recommend that future segmentation models, especially for large regions with varied topography, adopt terrain-aware training strategies. Even if a full DEM is not available everywhere, approximating slope from the imagery or using coarse global elevation data could still offer benefits. Ultimately, bridging the gap between pixel appearance and real-world topography makes the model's understanding more physical and generalizable, reducing surprises when it's deployed on a new landscape.

Focus model attention on domain-specific features. Beyond data augmentation and multi-source inputs, improvements in the model architecture and training objectives themselves can yield more reliable segmentation. One intriguing direction is the use of feature-guided masked autoencoders or similar techniques that encourage the model to learn high-level features rather than getting bogged down in pixel-level noise. In remote sensing, not all pixels are equal – the spectral signature of a healthy tree canopy, for instance, is characterized by certain reflectance patterns (like high near-infrared reflectance for foliage), and textural cues can differentiate a tree from grass. A vanilla model might or might not latch onto these subtle cues. However, a feature-guided approach explicitly trains the model to reconstruct or predict meaningful feature representations (such as vegetation indices or edge maps) instead of raw pixel intensity. For example, a recent approach called FG-MAE (Feature Guided Masked Autoencoder) masks out parts of an image and tasks the model with predicting domain-relevant features (like a NDVI – Normalized Difference Vegetation Index image, or other engineered representations) for the masked region rather than the raw pixels. This forces the model to infer what type of object should be there, not just to copy textures, thereby learning a more semantic understanding of the scene [[**?**]]. Applying this idea to tree segmentation, we could pre-train models to predict features that highlight vegetation structure (perhaps oriented gradients that capture tree crown edges, or canopy height estimations from stereo images) so that the model's internal representations become highly attuned to "tree-ness." When fine-tuned for segmentation, such a model may be better at delineating tangled canopies or differentiating trees from confusing background elements, because it has learned to focus on the attributes that define a tree in imagery. Early studies in multispectral and SAR domains have found that this approach yields improved performance in complex environments [[**?**]]. We recommend integrating feature-guidance in training for large-scale tree mapping, especially in areas with complex backgrounds (e.g., urban environments where trees mingle with buildings). Not only does this likely boost accuracy, it could also produce more interpretable model outputs or uncertainties – a model that knows what features it's looking for might provide reasoning (explicit or implicit) for its segmentation, which in turn aids validation and error diagnosis.

Expand validation beyond pixel agreement – use ecological consistency checks. Traditional validation of segmentation focuses on geometric overlap with ground truth shapes, but for tree mapping we can also exploit well-established ecological relationships as an additional form of validation. Trees

have allometric relationships – mathematical links between dimensions like trunk diameter, height, crown diameter, and biomass. These relationships have been measured in forestry for decades. For example, a tree of a given height typically has a crown of roughly proportional diameter, and there are known bounds on how big a crown can get for a given trunk size. We propose using such allometric equations as a sanity check for segmentation outputs. If a model's predicted tree crowns violate basic allometry (say, a tiny tree height but a huge crown width, or a cluster of crowns whose total basal area implies an impossible biomass for that area), it might indicate errors in the segmentation or missing trees. One could quantitatively compare the distribution of predicted crown sizes and tree heights (if height data or estimates are available) against expected distributions from field data. Significant deviations could flag problems: for instance, if the model frequently delineates very large crowns that in reality would correspond to 80-meter tall trees (which don't exist in the region), those are likely over-segmentation artifacts. Conversely, if predicted crowns are all very small in an old-growth forest where trees should be large, the model might be under-segmenting (splitting one crown into many). Researchers have indeed used this kind of approach in related contexts; for example, Song et al. (2023) employed Gaussian process regression on tree height to estimate biomass, demonstrating how linking remote sensing outputs to allometric models can validate whether the outputs make ecological sense [[**?**]]. In practice, implementing allometric checks means bringing in additional data or models (e.g., a LiDAR-derived tree height map, or species-specific allometric formulas from forestry literature) and cross-verifying the plausibility of the AI-generated tree map. 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. For computer vision in general, ImageNet and COCO served this role, and now geospatial analogues are emerging. Efforts like GEO-Bench and PANGAEA have started assembling diverse geospatial tasks into comprehensive evaluation suites, underscoring the importance of benchmarking models across varying conditions [[**?**]]. We recommend developing similar benchmark datasets and challenges specifically for tree crown segmentation and mapping. These benchmarks should cover a wide range of landscapes (different forest types, urban trees, plantations, etc.), seasons, and remote sensing data sources, reflecting the real-world diversity highlighted in our Challenges section. Importantly, a common benchmark would promote consensus on validation metrics – if the community agrees to evaluate on, say, a mix of IoU for segmentation quality and perhaps a detection metric for counting accuracy, it would standardize how results are reported. This comparability accelerates progress: researchers can pinpoint which innovations truly lead to better performance and robustness. Moreover, a shared

evaluation framework can include protocols for multi-scale assessment (e.g., requiring methods to report not just overall accuracy but also performance on hard subsets like "dense forest canopy" vs "isolated trees") and encourage inclusion of auxiliary checks like the allometric consistency mentioned above. In short, a benchmarking initiative for large-scale tree segmentation would provide an invaluable feedback loop, where challenges discovered by one team (e.g., failure cases in a certain region or condition) become part of the test set that everyone then tries to solve. This kind of iterative improvement cycle has been crucial in other domains of AI. We envision something like an annual challenge where models compete on a standard large-scale tree mapping task – pushing methods to be not only accurate but also general and validation-friendly. By establishing these community standards, we move toward a future where claims of model performance are transparent and believable, because they've been vetted on a broad, agreed-upon spectrum of scenarios.

Innovate in ground-truth data collection and labeling. Lastly, none of the above model improvements obviates the need for better and more plentiful validation data. We therefore recommend parallel efforts to enhance ground-truth collection through automation and crowdsourcing. On the automation side, advances in high-resolution imaging and onboard AI mean that drones or small aircraft could be deployed to automatically detect and delineate trees in sample areas, providing semi-automated annotations that experts only need to lightly verify. For instance, a drone flying over a forest could use its own simpler model to suggest crown boundaries, which are then corrected by a human specialist – vastly speeding up the annotation process compared to drawing polygons from scratch. Similarly, leveraging crowdsourcing and citizen science could dramatically expand validation datasets: non-experts can be asked to label trees on accessible platforms (especially for easy tasks like clicking the center of a visible tree), and with enough redundancy and quality control, these contributions can rival expert labels. Projects for mapping trees in cities (e.g., through apps that engage the public to identify street trees) hint at this potential. Of course, when using crowd data, one must account for variability in skill and ensure rigorous cleaning and validation of the contributed labels – but as a complementary source of truth data, it could fill gaps in areas professionals have not covered. Another approach is active learning, where the model itself guides where more data is needed: the model can flag areas or examples where it is most uncertain or makes conflicting predictions (say, an area of weirdly segmented crowns), and those areas would be prioritized for human labeling. This way, instead of randomly sampling locations for ground truth, we focus effort on the most informative samples – those likely to teach the model something new or reveal a blind spot. Active learning strategies have been shown to greatly reduce the amount of data needed for training in other remote sensing tasks by smartly selecting the right samples to label [^]. Integrating an active learning loop into large-scale validation means we continually refine the ground truth in the areas that matter most, thereby improving the model in a targeted fashion. Finally, we should not overlook the role of policy and open data initiatives: governments

and organizations conducting tree inventories or LiDAR scans should be encouraged to share these as open benchmarks. Even if such datasets are not perfectly aligned with satellite imagery, they can often be matched or used to validate parts of a map. In summary, closing the validation data gap will require creativity – using machines to help label, using the crowd to scale up, and using the model's own intelligence to guide where to look next. These efforts complement the technical recommendations above: better models make use of the new data more effectively, and better data enables more powerful models.

By pursuing these recommendations in concert, the field can make significant strides toward accurate and trustworthy large-scale tree mapping. What we envision is an end-to-end pipeline where robust models (forged through self-supervised, multi-view, terrain-aware training) produce segmentation maps that are continuously checked against both classical metrics and real-world plausibility (via benchmarks and allometric/field validations), and where the feedback from validation drives further improvement of the models through active learning and expanded training data.