Modular Fusion for Individual Tree Crown Delineation from Airborne LiDAR Data

Josué Gourde, Baoxin Hu*, Qian Li

Department of Earth and Space Science and Engineering, York University, Toronto, Ontario, Canada
(jgourde, baoxin, qian2018)@yorku.ca

Keywords: Individual tree crown, LiDAR, canopy height model, detection fusion, Segment Anything, Masked Autoencoder.

Abstract

Accurate delineation of individual tree crowns (ITC) is essential for forest inventory, biomass estimation, and ecological monitoring. Airborne LiDAR data provide an effective basis for this task, but accurate delineation remains challenging due to complex canopy structure and limited labelled training data. In this study, a modular fusion framework is developed that combines bounding box detection, multi-model fusion, and foundation model segmentation to produce tree crown masks from canopy height models (CHMs). Two detection architectures, DINO and Faster R-CNN, are implemented with both ImageNet-pretrained ResNet-50 and domain-specific Masked Autoencoder (MAE) backbones. Due to the small size of the target dataset, supplementary training data from the Finnish Taiga–Tundra Ecotone are incorporated to stabilize transformer training. Predictions from the individual detectors are fused using Weighted Box Fusion, with score normalization applied to account for differing confidence distributions between architectures. The fused detections serve as prompts for the Segment Anything Model (SAM), enabling a decoupled detection and segmentation framework for tree crown delineation. On a mixed-wood forest site in the Great Lakes–St. Lawrence forest region of Ontario, Canada, the framework achieves mask-level F1 scores of 0.79 and 0.61 at IoU thresholds of 0.25 and 0.50 respectively, matching 193 of 233 ground truth trees on the primary test plot. Visual inspection indicates that many delineated crowns align more closely with canopy structure than the reference polygons. The proposed framework demonstrates the potential of modular detector fusion and detection-guided segmentation for ITC delineation from LiDAR-derived canopy height models.

1. INTRODUCTION

Accurate delineation of individual tree crowns (ITC) is essential for forest inventory, above-ground biomass estimation, species classification, and health monitoring (Zhen et al., 2016, Zheng et al., 2024). Airborne LiDAR data provide an effective basis for this task, as canopy height models (CHMs) derived from LiDAR point clouds capture the three-dimensional structure of the forest canopy at sub-metre resolution. While detection produces bounding boxes around individual trees, delineation aims to recover the full crown boundary, a more informative but more difficult output. Accurate crown polygons are needed for downstream tasks such as species-level biomass modelling and canopy gap analysis, yet obtaining them remains labour-intensive: manual annotation of crown boundaries is time-consuming, subjective, and difficult to scale to large forest inventories.

Classical approaches based on local maxima filtering, watershed segmentation, and region growing have been applied to CHMs for over two decades (Popescu et al., 2002, Chen et al., 2006). These methods perform well in open canopy conditions but degrade in structurally complex forests with overlapping, multi-layered canopies where individual crowns are difficult to separate by height alone. More recently, instance segmentation architectures such as Mask R-CNN (He et al., 2017) and Mask-DINO (Li et al., 2023) have been applied to ITC delineation, predicting per-pixel crown masks alongside bounding boxes. However, the mask heads of these models must be trained on the same limited annotated data available for the target forest, and polygon-level annotations are substantially more expensive to produce than bounding boxes. Moreover, because detection and segmentation are tightly coupled within these architectures, it is difficult to fuse outputs from multiple models or to replace individual components without retraining the entire system.

The detection components underlying these instance segmentation models are themselves strong standalone detectors. Faster R-CNN (Ren et al., 2015), the backbone of Mask R-CNN, has been widely adopted for tree detection (Weinstein et al., 2019). DINO (Zhang et al., 2023), on which Mask-DINO is built, is a transformer-based detector offering global attention mechanisms that capture long-range spatial relationships between trees. A common limitation of both CNN and transformer detectors is their dependence on large labelled datasets, which are scarce in forestry. Self-supervised pretraining strategies such as Masked Autoencoders (MAE) (He et al., 2022) offer a pathway to learn domain-specific feature representations from unlabelled CHM data, reducing the labelled data requirement and improving detection in data-scarce settings.

The tight coupling of detection and segmentation within existing architectures limits the ability to combine complementary detectors or to upgrade individual components independently. By decoupling detection from segmentation and operating at the bounding box level, predictions from multiple architecturally diverse detectors can be fused before mask generation, combining the strengths of each at a stage where fusion is straightforward. Different detection architectures exhibit complementary characteristics: CNN-based detectors tend to produce higher-confidence predictions with fewer false positives, while transformer-based detectors often achieve better localization and recall at lower confidence levels. These complementary patterns suggest that fusing predictions could yield more robust detection, but confidence scores across architectures are not directly comparable, making naïve fusion problematic. Weighted Box Fusion (WBF) (Solovyev et al., 2021)

* Corresponding author

addresses this by averaging the coordinates of overlapping boxes weighted by confidence, but requires that scores from different models be on a comparable scale.

A foundation segmentation model such as SAM (Kirillov et al., 2023), trained on over one billion masks from diverse domains, can serve as a downstream segmentation module within a modular framework. When prompted with detection bounding boxes, SAM generates per-object masks without requiring any domain-specific mask training data. This detection-guided approach is central to the proposed framework, as it decouples the segmentation capability from the detection training. SAM has seen rapid adoption in medical imaging and remote sensing, but its application to LiDAR-derived CHM data for ITC delineation has not been systematically evaluated. A key question is whether detection-guided prompts are necessary or if SAM can segment tree crowns autonomously from CHM data.

In this study, we propose a modular framework for individual tree crown delineation that explicitly decouples detection, fusion, and segmentation. Within this framework, we first examine how complementary detection architectures, CNN-based and transformer-based, behave under limited training data when paired with both generic and domain-specific backbones. We then show how operating at the bounding-box level enables principled fusion of heterogeneous detectors using score-normalized Weighted Box Fusion. Finally, we demonstrate how a foundation segmentation model can be incorporated as a downstream segmentation module through detection-guided prompts, allowing high-quality crown delineation without requiring mask-level training data. Together, these contributions illustrate how modular design facilitates model fusion, extensibility, and practical deployment in data-limited forestry applications.

2. METHODS

The proposed framework consists of three functional stages: (1) detection, which locates individual trees using bounding box predictions from multiple architectures; (2) fusion, which reconciles predictions from heterogeneous detectors through score normalization and Weighted Box Fusion; and (3) segmentation, which delineates crown boundaries using detection-guided SAM prompts. Data preparation is described in Section 2.1, followed by each stage in subsequent subsections.

2.1 Study Area and Data

The target site is a mixed-wood forest in the Great Lakes–St. Lawrence (GLSL) forest region near Sault Ste. Marie, Ontario, Canada (Hu et al., 2014). The stand comprises deciduous species including trembling aspen (*Populus tremuloides*) and white birch (*Betula papyrifera*), and coniferous species including jack pine (*Pinus banksiana*) and black spruce (*Picea mariana*). Stand age ranges from 30 to 80 years, producing a closed, multi-layered canopy with substantial crown overlap. Airborne LiDAR was acquired using a Riegl Q-560 scanner at approximately 40 pulses/m^2 from roughly 200 m above ground in August 2009. The native point density of the GLSL data would optimally produce a CHM at approximately 0.5 m resolution; however, to provide sub-crown-level detail for detection, the CHM was generated at both 0.15 m and 0.25 m resolutions by interpolating the point cloud onto a finer grid. The raw point cloud was processed by first extracting ground-classified points (LAS class 2) and interpolating them onto a regular grid using scipy's griddata with linear interpolation to produce a digital terrain model (DTM). The CHM was then derived by subtracting the DTM from the maximum height surface at the target resolution. This upscaling from the native point density introduces interpolation artefacts but enables the resolution of individual crown edges that would be lost at 0.5 m. Individual tree crown annotations were provided as polygon labels (Li et al., 2025).

The Finnish Taiga data was originally provided at 0.5 m resolution. To maintain consistent patch dimensions across datasets, both the Finnish Taiga and GLSL CHMs were generated at 0.15 m and 0.25 m resolutions using bilinear interpolation for the Finnish Taiga upscaling. While upsampling the Finnish Taiga data does not add new structural information, it ensures that all models receive inputs at the same spatial resolution and patch dimensions, avoiding the need for resolution-dependent architecture modifications. Models were trained on both resolutions but tested exclusively at 0.15 m, which provides the finest detail for crown boundary delineation.

The CHM was divided into 128×128 pixel patches (19.2 m × 19.2 m at 0.15 m resolution) using a sliding window with 50% overlap, yielding 808 training, 179 validation, and 55 test patches. Two test plots of 50 m × 50 m each (Plot 1 with 233 ground truth trees and Plot 2 with 107 ground truth trees) were designated for evaluation. Figure 1 shows the CHM of both test plots.

In preliminary experiments, we observed that transformer-based detectors collapsed when trained on GLSL data alone due to insufficient training volume. DINO in particular failed to converge, producing either no detections or a degenerate pattern of uniformly spaced boxes unrelated to actual tree positions. To address this, we incorporated patches from the Finnish Taiga–Tundra Ecotone ITC-samples dataset (Lin et al., 2024a, Lin et al., 2024b), a boreal forest site in northern Finland with well-separated canopies of Scots pine (*Pinus sylvestris*), Norway spruce (*Picea abies*), and mountain birch (*Betula pubescens*). Although the Finnish Taiga site is structurally simpler than the GLSL forest, with lower tree density and less crown overlap, the additional training volume stabilized transformer convergence. The combined training set provided sufficient diversity in crown sizes and canopy configurations to learn generalizable detection features.

2.2 Detection Module

The detection component of the framework is designed to accommodate multiple detection architectures. In this study, two representative detectors, DINO and Faster R-CNN, are implemented to capture complementary detection characteristics.

DINO (Zhang et al., 2023) is a transformer-based detector that employs contrastive denoising training, mixed query selection, and iterative box refinement. Its deformable attention mechanism attends to a learned sparse set of sampling points across multiple feature scales, enabling detection of crowns at varying sizes without the quadratic computational cost of dense global attention. Faster R-CNN (Ren et al., 2015) is a two-stage CNN-based detector with a region proposal network that generates candidate regions, followed by classification and bounding box regression heads that refine each proposal. Its convolutional feature extraction provides strong local pattern recognition but lacks the global context available to transformer architectures.

Each architecture was trained with two backbone feature extractors. The first is ResNet-50 (He et al., 2016) pretrained on

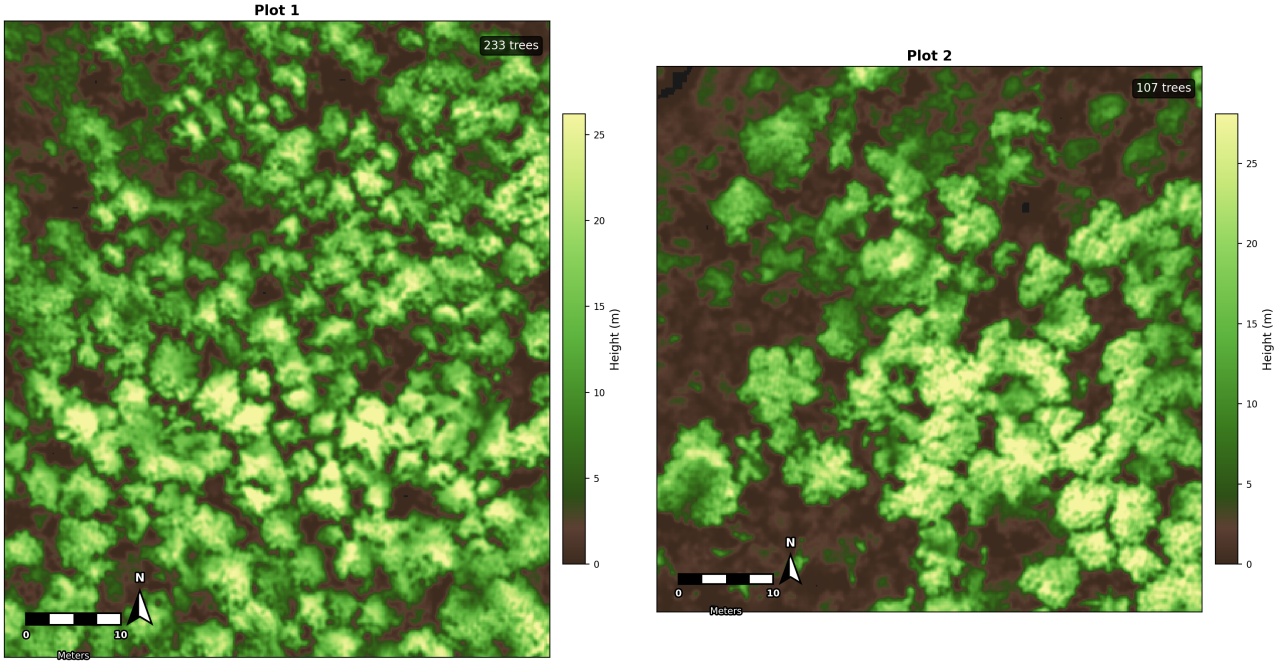

Figure 1. Canopy height models of the two test plots in the GLSL study area. Plot 1 (left) contains 233 ground truth trees in a dense, closed canopy. Plot 2 (right) contains 107 ground truth trees with more visible canopy gaps.

ImageNet, providing general-purpose visual features as a standard baseline. The single-channel CHM input is handled by averaging the pretrained RGB weights into a single input channel. The second backbone is a Masked Autoencoder (MAE) (He et al., 2022) pretrained on unlabeled CHM patches from the study area. The MAE uses a Vision Transformer encoder (384-dimensional, 6 layers, 6 attention heads) trained to reconstruct randomly masked patches (75% masking ratio) from the visible ones. Pretraining was performed on 1,291 CHM patches for 100 epochs. This domain-specific pretraining produces feature representations attuned to canopy height structure rather than natural image statistics, which is particularly relevant for distinguishing overlapping crowns at similar heights.

All four model variants (DINO/ResNet-50, DINO/MAE, FRCNN/ResNet-50, FRCNN/MAE) were trained on the combined GLSL and Finnish Taiga training patches using the AdamW optimizer with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4}. Training ran for 50 epochs with early stopping based on validation F1.

2.3 Fusion Module

Each model produces detection boxes with associated confidence scores, but the score distributions differ substantially across architectures. DINO generates a large number of candidate detections at relatively low confidence, many of which are well-localized. Faster R-CNN produces fewer predictions but with higher confidence values. The MAE backbone further amplifies this difference: DINO/MAE requires a confidence threshold of only 0.15, while FRCNN/MAE requires 0.65, reflecting a four-fold difference in the raw score at which each model achieves its best detection quality.

To select the optimal operating point for each model, we performed a joint sweep of confidence threshold (0.05 to 0.90 in steps of 0.05) and non-maximum suppression (NMS) IoU threshold (0.10 to 0.50 in steps of 0.05) on the 179 validation patches, optimizing for detection F1 at IoU 0.5. NMS removes redundant overlapping detections by suppressing lower-confidence boxes that overlap with a higher-confidence box beyond the IoU threshold. The tuned parameters were: DINO/MAE (confidence 0.15, NMS 0.30), DINO/ResNet-50 (confidence 0.35, NMS 0.20), FRCNN/MAE (confidence 0.65, NMS 0.35), and FRCNN/ResNet-50 (confidence 0.40, NMS 0.45). The tight NMS threshold for DINO/ResNet-50 (0.20) indicates that this model produces more spatially overlapping detections that require aggressive suppression, while the looser threshold for FRCNN/ResNet-50 (0.45) reflects fewer redundant predictions.

To fuse predictions from models with incompatible confidence scales, we apply a threshold-anchored normalization before Weighted Box Fusion (Solovyev et al., 2021). For each model, the tuned confidence threshold τ is mapped to 0.5 using a piecewise linear function:

$$\hat{s} = \begin{cases} 0.5 \cdot s/\tau & \text{if } s \leq \tau \\ 0.5 + 0.5 \cdot (s - \tau)/(1 - \tau) & \text{if } s > \tau \end{cases} \quad (1)$$

where s is the raw confidence score and $\hat{s}$ is the normalized score. This ensures that a prediction at each model's quality threshold receives equal weight during fusion, regardless of the raw score magnitude. For example, a DINO/MAE prediction at confidence 0.15 and a FRCNN/MAE prediction at confidence 0.65 both map to $\hat{s} = 0.5$, reflecting that both are at their respective quality boundaries.

WBF then operates on the normalized predictions by matching overlapping boxes from different models based on IoU and computing a weighted average of the matched box coordinates, with the normalized confidence as the weight. Unmatched boxes are retained as-is. This coordinate averaging tends to produce tighter bounding boxes than either individual model, as localization errors from different architectures partially cancel. The WBF IoU threshold was tuned on the validation set by sweeping values from 0.20 to 0.60; the best threshold was 0.20 for the DINO/MAE and FRCNN/MAE pair.

2.4 Segmentation Module

The segmentation module uses the Segment Anything Model (SAM) (Kirillov et al., 2023) as a detection-guided segmentation component. SAM is particularly attractive in this setting because it enables high-quality instance segmentation without requiring any domain-specific mask annotations, which are expensive and scarce in forestry applications. We use the original SAM (ViT-B) rather than its successor SAM 2, which was optimized for video tracking and produces a more conservative automatic mask generator. In preliminary testing, SAM 2 generated fewer than half the automatic masks of SAM on the same CHM data (124 versus 302 on Plot 1), resulting in fewer gap-filling crowns and lower overall recall. SAM 2 also tended to ignore smaller understory crowns that SAM detected; these crowns are subtle in the CHM but become apparent upon 3D point cloud inspection (Figure 2). The fused bounding boxes serve as spatial prompts to SAM, which generates per-tree segmentation masks.

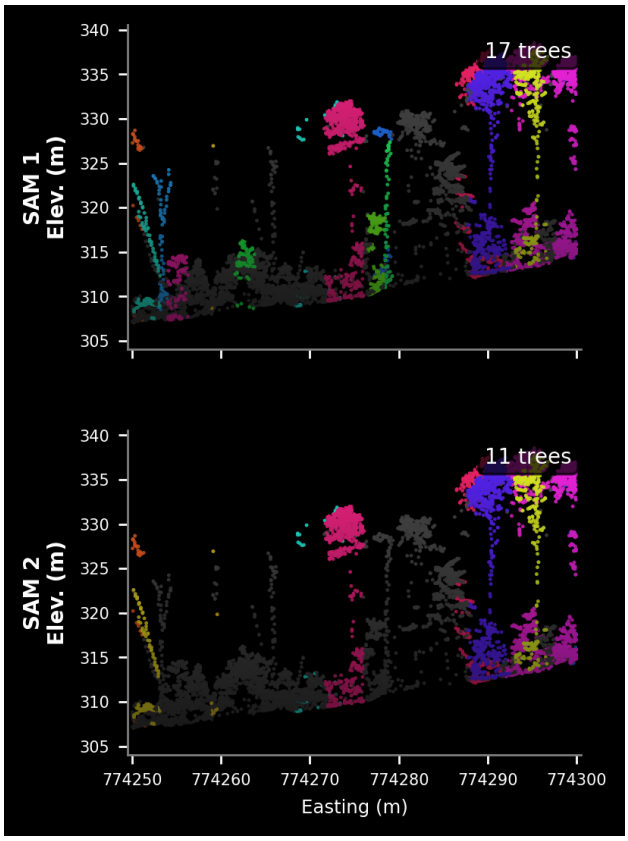

Figure 2. Cross-section through the bottom of Plot 2 comparing SAM 1 (top) and SAM 2 (bottom) segmentation of the same WBF-fused detections. Each colour represents a distinct tree crown. SAM 1 identifies 17 trees in this slice while SAM 2 finds only 11, missing several smaller understory crowns visible on the left side.

Although SAM was trained on natural RGB images, we apply it to the single-channel CHM by replicating the height values across three channels after percentile-based normalization to the 0–255 range.

For each detection box, SAM produces three candidate masks at different granularities; we select the one with the highest predicted IoU score. Masks smaller than 9 pixels (approximately 0.2 m^2) are discarded as noise. Overlapping masks are resolved by giving priority to higher-confidence detections: masks are sorted by descending confidence and placed sequentially, with each mask only writing to pixels not already claimed by a higher-confidence mask. If a mask is more than 70% overlapped by previously placed masks, it is discarded entirely to avoid fragmentation.

This overlap resolution can split a mask into disconnected fragments when a higher-confidence neighbour claims part of its area. To handle this, we identify connected components within each mask using morphological labelling. The largest connected component is retained as the primary mask, and smaller fragments are merged into the neighbouring mask that shares the most border pixels. Fragments with no adjacent neighbour are removed.

Additionally, SAM's automatic mask generator is run on the full mosaic without any box prompts to discover trees that both detection models may have missed. This mode uses a regular grid of point prompts across the image, generating masks for all salient regions. Auto-generated masks are accepted only if they have a mean CHM height above 2 m (to exclude ground-level segments), cover fewer than 25% of the mosaic area (to exclude full-canopy masks), and overlap less than 30% with existing detection-prompted masks (to avoid duplicating already-detected trees). Accepted masks are added as additional tree crowns beyond those from the detection-guided stage.

2.5 Evaluation Metrics

Detection performance is evaluated at the bounding box level using precision, recall, and F1 at IoU 0.5. A predicted box is considered a true positive if it overlaps with a ground truth box with IoU ≥ 0.5; predictions are matched greedily in descending confidence order, and each ground truth box can be matched at most once.

Mask quality is evaluated by rasterizing both predicted SAM masks and ground truth polygons into per-pixel label maps and computing per-instance IoU between matched mask pairs. Masks are matched using greedy highest-IoU assignment with a minimum overlap of 0.1. We report the mean IoU (mIoU) of matched pairs and mask-level F1 at three IoU thresholds: 0.25, 0.50, and 0.75. The 0.25 threshold is included to reflect approximate crown identification accuracy, which is relevant in ITC delineation due to inherent uncertainty in manual polygon boundaries and the difficulty of precisely delineating overlapping canopies. The 0.50 threshold represents standard boundary agreement, and 0.75 captures strict boundary matching. We also report the number of ground truth trees matched by at least one predicted mask.

3. RESULTS

3.1 Detection Performance

Table 1 reports bounding box detection metrics on the two test plots prior to fusion and SAM mask generation. On Plot 1, DINO/ResNet-50 achieves the highest detection precision (0.75) and F1 (0.64), while DINO/MAE achieves the highest recall (0.61). On Plot 2, DINO/ResNet-50 again leads in precision (0.75) and F1 (0.59). The MAE backbone consistently improves Faster R-CNN, increasing detection F1 from 0.55 to 0.64 on Plot 1 and from 0.43 to 0.55 on Plot 2, the largest per-model improvement observed. For DINO, the MAE backbone yields

higher recall but at lower precision than ResNet-50, indicating that domain-specific pretraining increases sensitivity to smaller or partially occluded crowns while also introducing some false positives.

Table 1. Detection performance (bounding box level) on the two test plots. P = Precision, R = Recall.

Plot	Model	P	R	F1
Plot 1	DINO/MAE	0.66	0.61	0.64
	DINO/ResNet-50	0.75	0.56	0.64
	FRCNN/MAE	0.75	0.56	0.64
	FRCNN/ResNet-50	0.57	0.53	0.55
	WBF Fusion	0.64	0.63	0.64
Plot 2	DINO/MAE	0.65	0.52	0.58
	DINO/ResNet-50	0.75	0.49	0.59
	FRCNN/MAE	0.65	0.47	0.55
	FRCNN/ResNet-50	0.47	0.39	0.43
	WBF Fusion	0.59	0.52	0.55

3.2 Fusion Analysis

The WBF fusion of DINO/MAE and FRCNN/MAE matches the best individual detection F1 (0.64 on Plot 1) while achieving the highest recall (0.63) among all configurations. The fusion does not substantially increase detection F1 over the best individual model, which is expected: the primary value of fusion in this framework lies in robustness and calibration rather than maximizing detection-level metrics in isolation. By combining predictions from architecturally diverse detectors, the fusion module produces a more stable and balanced set of bounding boxes that serve as higher-quality prompts for the downstream segmentation module. The threshold-anchored normalization is critical to this process: without it, FRCNN predictions at raw confidence 0.65–1.00 would dominate the WBF weighting over DINO predictions at 0.15–0.99, despite both sets being above their respective quality thresholds.

3.3 Segmentation Results

Table 2 reports mask-level metrics after detection-guided SAM processing with automatic gap-filling. The mean IoU (mIoU) between predicted and ground truth masks ranges from 0.54 to 0.60 across models. At the lenient threshold of 0.25, the best individual model (DINO/ResNet-50 on Plot 1) achieves F1 = 0.82, indicating that the framework correctly identifies the vast majority of trees with at least rough boundary correspondence. The WBF fusion reaches F1 = 0.79 on Plot 1, matching 193 of 233 ground truth trees compared to 192 for the best individual model. At the strict 0.75 threshold, all models drop below F1 = 0.11, reflecting the difficulty of achieving tight pixel-level agreement between SAM’s height-gradient-based boundaries and the manually drawn reference polygons.

Performance on Plot 2 is consistently lower across all models, with mask F1@0.50 ranging from 0.44 to 0.48 compared to 0.60–0.65 on Plot 1. This difference is attributable to the more complex canopy structure in Plot 2, which contains more visible canopy gaps and a higher proportion of smaller, partially occluded crowns that are difficult to delineate from the CHM alone. Mask-level mIoU values are similar across backbones (0.58–0.60), suggesting that SAM’s segmentation quality is largely independent of which detector provides the bounding box prompt.

Table 2. Mask-level metrics after SAM with auto gap-filling. mIoU = mean IoU of matched masks. Match = GT trees matched.

	Model	mIoU	F1@.25	F1@.5	Match
Plot 1	DINO/MAE	0.59	0.78	0.62	191/233
	DINO/RN50	0.60	0.82	0.65	192/233
	FRCNN/MAE	0.59	0.81	0.62	186/233
	FRCNN/RN50	0.58	0.80	0.60	194/233
	WBF Fusion	0.59	0.79	0.61	193/233
Plot 2	DINO/MAE	0.56	0.70	0.45	86/107
	DINO/RN50	0.56	0.65	0.48	83/107
	FRCNN/MAE	0.55	0.70	0.47	86/107
	FRCNN/RN50	0.54	0.65	0.44	88/107
	WBF Fusion	0.55	0.69	0.45	87/107

3.4 Visual Analysis

Figure 3 shows the mask comparison across all four individual models on Plot 1. The ground truth panel (leftmost) shows manually delineated polygon labels, while subsequent panels show SAM-generated masks for each model. The SAM masks exhibit visually smoother and more consistent crown boundaries compared to the hand-drawn ground truth polygons, which contain angular edges and occasional over- or under-estimation of crown extent.

Figure 4 contrasts SAM applied without detection guidance (top) against SAM prompted by the WBF-fused detections (bottom). When run in automatic mode without bounding box prompts, SAM produces 290 masks for 233 ground truth trees, with an F1 of only 0.49 at IoU 0.50. Large areas of canopy are either missed entirely (green regions in the overlap panel) or over-segmented into fragments. With detection-guided prompts from the fusion module, the same SAM model achieves an F1 of 0.61 at IoU 0.50, demonstrating that the detection models provide critical spatial guidance that SAM alone cannot recover from single-band CHM data.

Figure 5 shows vertical cross-sections through both test plots, where each colour corresponds to a distinct tree crown as delineated by the framework. The segmentation cleanly separates adjacent crowns throughout the canopy.

4. DISCUSSION

The results demonstrate that a modular framework consisting of decoupled detection, fusion, and segmentation stages produces a viable approach to ITC delineation from CHM data. Compared to instance segmentation approaches that jointly predict boxes and masks, this decoupled design achieves competitive results without requiring polygon-level training annotations, while enabling each component to be independently evaluated, replaced, or upgraded. The proposed framework achieves 0.64 at the detection level and 0.61–0.65 at the mask level (F1@0.50), with the added benefit of polygon outputs. Several observations merit further discussion.

The MAE backbone provides a clear benefit for Faster R-CNN, where it improves detection F1 by 0.09 on Plot 1 and 0.12 on Plot 2. The improvement is consistent across both test plots, suggesting that domain-specific features provide a genuine advantage over ImageNet features for CHM-based detection. For DINO, the MAE backbone yields the highest recall but at lower precision than ResNet-50. This pattern can be explained by

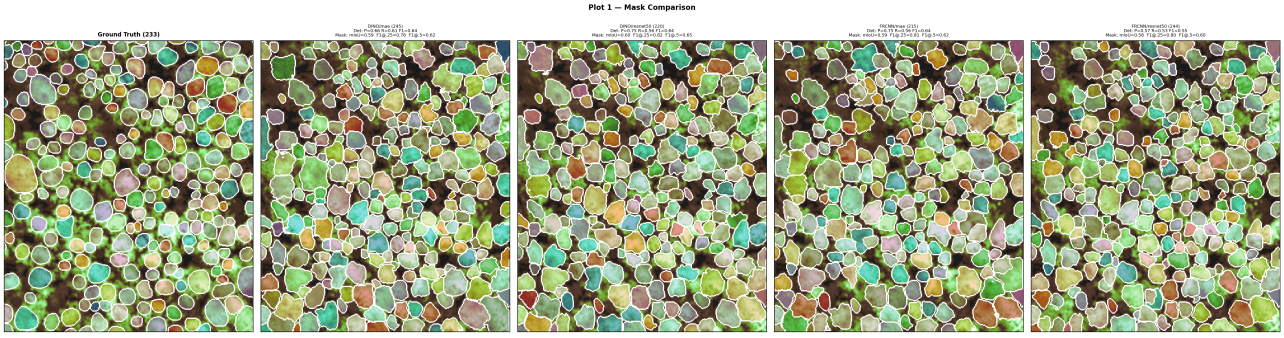

Figure 3. Mask comparison on Plot 1 (233 ground truth trees). Left to right: ground truth polygons, DINO/MAE, DINO/ResNet-50, FRCNN/MAE, FRCNN/ResNet-50. Each panel shows SAM-generated masks overlaid on the CHM with white borders between adjacent crowns. Detection and mask metrics are shown above each panel.

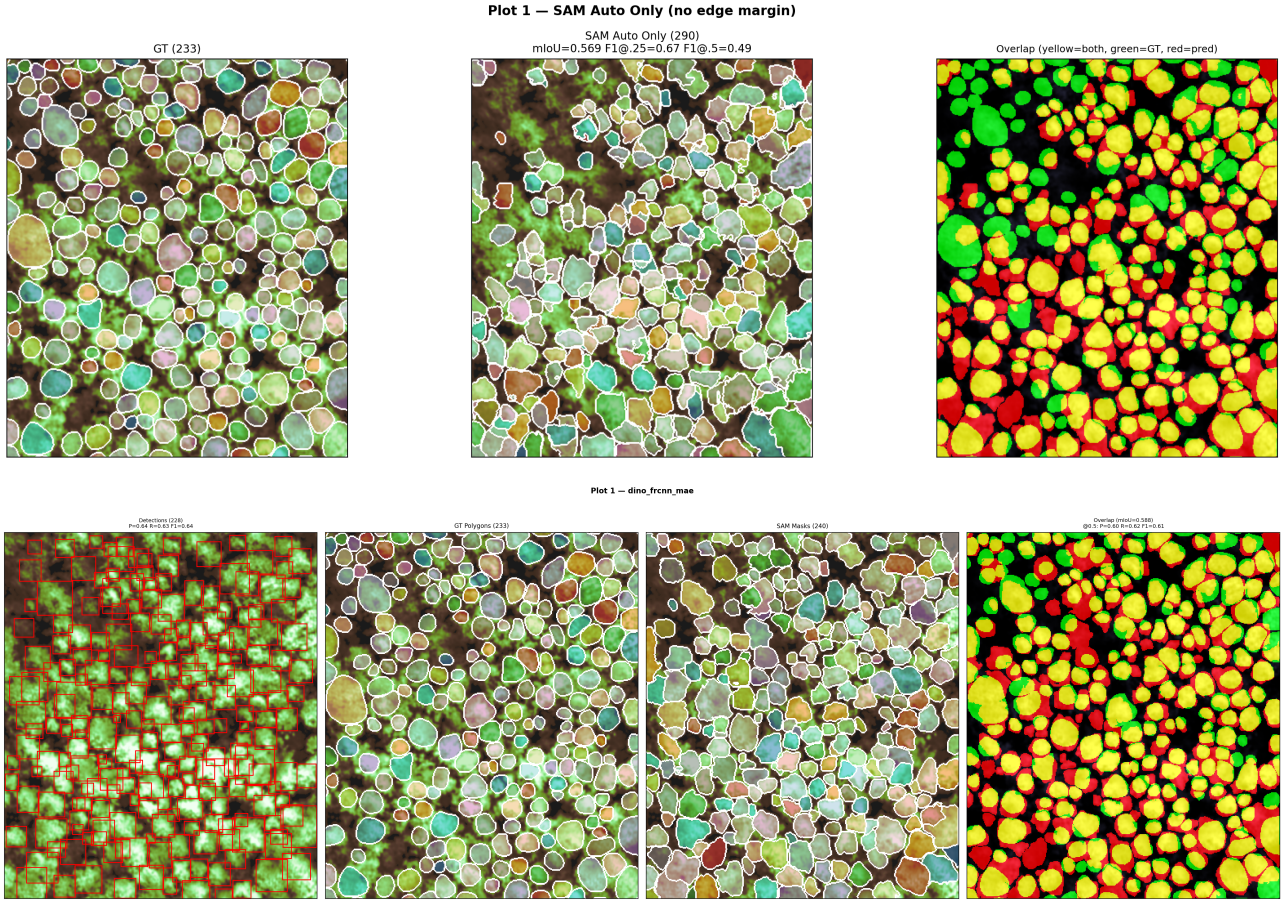

Figure 4. SAM without detection guidance (top) versus SAM with WBF-fused detection prompts (bottom) on Plot 1. The SAM-only overlap panel shows substantial missed canopy (green) and false positives (red). The detection-guided framework recovers these areas, matching 193 of 233 ground truth trees.

the MAE’s pretraining objective: reconstructing masked CHM patches encourages the encoder to learn fine-grained height variation patterns, which improves sensitivity to smaller or partially occluded crowns but may also trigger false detections on height structures that are not trees. The practical consequence is that DINO/MAE requires a much lower confidence threshold (0.15 versus 0.35 for ResNet-50) to achieve its best F1, making the threshold-anchored normalization particularly important when fusing its predictions with those of other models.

The fusion module does not substantially increase detection-level F1 over the best individual detector, which is expected in a setting where individual models are already well-tuned. The value of fusion in this framework lies instead in robustness and calibration: by combining predictions from architecturally diverse detectors with normalized confidence scales, the fusion module produces a more balanced and stable set of bounding boxes. These fused boxes serve as higher-quality prompts for the downstream segmentation module, providing broader coverage than any single detector while maintaining competitive precision. The threshold-anchored score normalization is essential to this process. Without normalization, the WBF would be dominated by whichever architecture produces higher raw confidence, regardless of prediction quality. By anchoring each

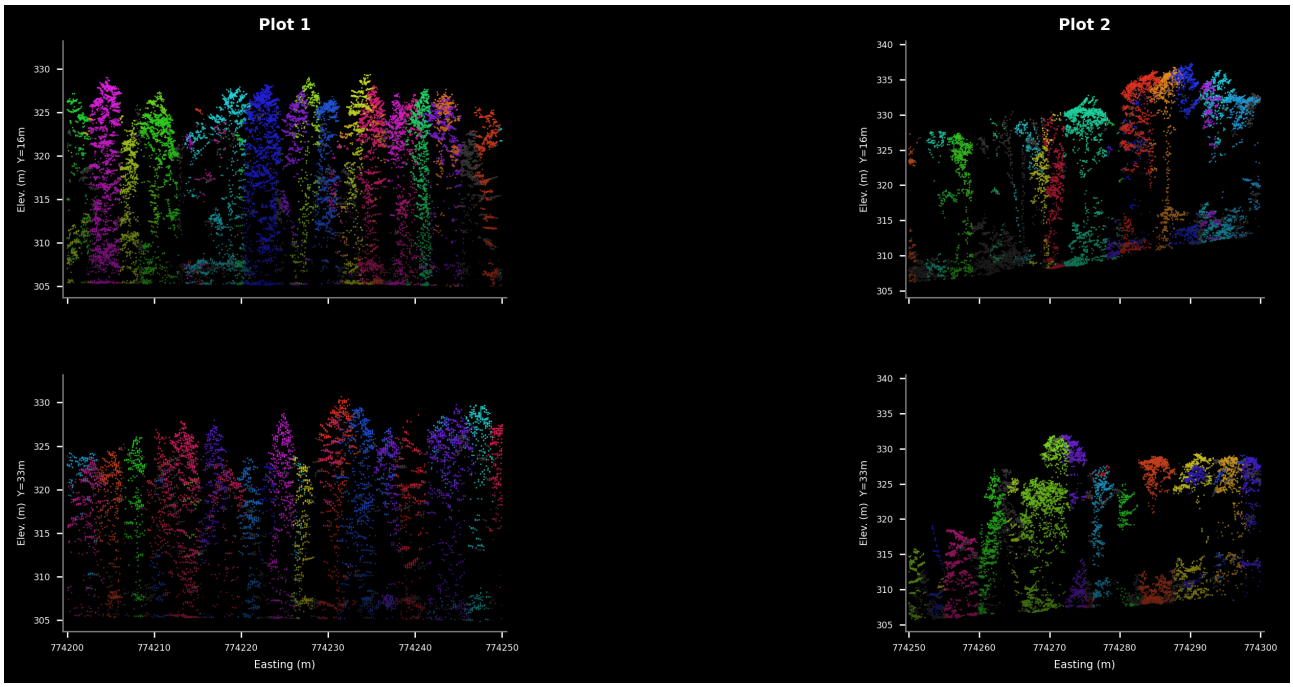

Figure 5. Vertical cross-sections (4 m wide) through both test plots using WBF fusion masks. Each colour represents a distinct tree crown. Left column: Plot 1; right column: Plot 2. Two slices per plot are shown at approximately one-third and two-thirds through the north–south extent.

model’s tuned threshold to 0.5, the fusion treats a “good” prediction from each model equally, enabling the coordinate averaging in WBF to produce tighter bounding boxes.

A notable observation is that SAM-generated crown boundaries are in many cases more consistent with canopy structure than the manually delineated reference polygons. The hand-drawn labels contain angular edges and inconsistencies inherent to manual polygon annotation on CHM data, and manual annotators can miss smaller or partially occluded crowns that the detection models identify. These missed annotations inflate the apparent false positive count and reduce the measured F1, even when the predicted masks are correct. Upon visual inspection, many delineated crowns produced by the framework correspond to real trees with boundaries that follow height gradients more faithfully than the reference polygons. This observation suggests that detection-guided segmentation frameworks could serve as a tool for improving ground truth annotation quality, not only for producing predictions.

The SAM automatic mask generator contributed 28–52 additional masks per plot beyond those prompted by detection boxes. While this improves recall (matching up to 194 of 233 trees), it also introduces false positives in canopy gaps and between trees. A tighter filtering criterion or a second-stage classifier could reduce these errors.

Several limitations and directions for future work should be noted. The evaluation is conducted on a single target site in a temperate mixed-wood forest; the generalizability of the framework to other forest types (e.g., tropical, boreal, plantation) remains to be established. SAM was not trained on CHM data and operates on a three-channel replication of a single height band. Fine-tuning SAM on forestry-specific data, or using SAM 2 with its improved architecture, could yield further improvements in boundary precision. The current framework does not account for the vertical structure of multi-layered canopies, where suppressed trees beneath the dominant canopy are invisible in the CHM. Integrating full 3D point cloud information rather than the 2.5D CHM surface could address this limitation. The modular design of the framework facilitates such extensions: improved detectors, alternative fusion strategies, or more capable segmentation models can be incorporated without redesigning or retraining the entire system.

5. CONCLUSIONS

This paper presented a modular framework for individual tree crown delineation that explicitly decouples detection, fusion, and segmentation into independent, replaceable stages. The key findings are as follows:

1. Training DINO and Faster R-CNN with a domain-specific MAE backbone and supplementary data from the Finnish Taiga addressed the data scarcity problem that causes transformer collapse on small forest datasets. The MAE backbone provided the largest improvement for Faster R-CNN, increasing detection F1 by up to 0.12.

2. The threshold-anchored score normalization enabled principled fusion of architecturally diverse detectors with incompatible confidence distributions. The value of fusion in this framework lies in robustness and calibration, providing more stable prompts for downstream segmentation rather than maximizing detection F1 in isolation.

3. SAM, incorporated as a detection-guided segmentation module, generates crown masks with boundaries that are in many cases more consistent with canopy structure than manually delineated reference polygons. The framework matched 193 of 233 ground truth trees on the primary test plot, achieving a mask F1 of 0.79 at IoU 0.25 and 0.61 at IoU 0.50.

4. Detection guidance is critical for SAM on CHM data. Without bounding box prompts, SAM's automatic mask generator misses large areas of canopy and over-segments others, achieving only F1 = 0.49 at IoU 0.50 compared to 0.61 with detection-guided prompts.

The modular design of the framework provides a practical pathway from LiDAR-derived CHMs to polygon-level crown delineation without requiring per-pixel training annotations. Because detection, fusion, and segmentation are decoupled, improvements in any single component, whether new detection architectures, alternative fusion strategies, or more capable segmentation models, can be incorporated without changes to the overall workflow.

ACKNOWLEDGEMENTS

The authors wish to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Forestry Futures Trust Ontario for their generous financial support of this research. The Finnish Taiga–Tundra Ecotone ITC-samples dataset was provided by (Lin et al., 2024a).

References

Chen, Q., Baldocchi, D., Gong, P., Kelly, M., 2006. Isolating individual trees in a savanna woodland using small footprint lidar data. *Photogrammetric Engineering & Remote Sensing*, 72(8), 923–932.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R., 2022. Masked autoencoders are scalable vision learners. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16000–16009.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, 2961–2969.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.

Hu, B., Li, J., Jing, L., Judah, A., 2014. Improving the efficiency and accuracy of individual tree crown delineation from high-density LiDAR data. *International Journal of Applied Earth Observation and Geoinformation*, 26, 145–155.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., Girshick, R., 2023. Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L. M., Shum, H.-Y., 2023. Mask DINO: Towards a unified transformer-based framework for object detection and segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3041–3050.

Li, Q., Hu, B., Shang, J., Remmel, T. K., 2025. Two-Stage Deep Learning Framework for Individual Tree Crown Detection and Delineation in Mixed-Wood Forests Using High-Resolution Light Detection and Ranging Data. *Remote Sensing*, 17(9), 1578.

Lin, Y., Li, H., Jing, L., Ding, H., Tian, S., 2024a. Individual Tree Crown Delineation Using Airborne LiDAR Data and Aerial Imagery in the Taiga-Tundra Ecotone. *Remote Sensing*, 16(21), 3920.

Lin, Y., Li, H., Jing, L., Ding, H., Tian, S., 2024b. Individual tree crown (ITC) sample set in the Taiga-Tundra ecotone. [dataset].

Popescu, S. C., Wynne, R. H., Nelson, R. F., 2002. Estimating plot-level tree heights with lidar: local filtering with a canopy-height based variable window size. *Computers and Electronics in Agriculture*, 37(1-3), 71–95.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.

Solovyev, R., Wang, W., Gabruseva, T., 2021. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107, 104117.

Weinstein, B. G., Marconi, S., Bohlman, S., Zare, A., White, E., 2019. Individual tree-crown detection in RGB imagery using semi-supervised deep learning neural networks. *Remote Sensing*, 11(11), 1309.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L. M., Shum, H.-Y., 2023. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. *International Conference on Learning Representations (ICLR)*.

Zhen, Z., Quackenbush, L. J., Zhang, L., 2016. Trends in automatic individual tree crown detection and delineation – Evolution of LiDAR data. *Remote Sensing*, 8(4), 333.

Zheng, J., Yuan, S., Li, W., Fu, H., Yu, L., Huang, J., 2024. A review of individual tree crown detection and delineation from optical remote sensing images: Current progress and future. *IEEE Geoscience and Remote Sensing Magazine*, 13(1), 209–236.