Large-scale individual crown tree segmentation across entire white spruce forests using UAV hyperspectral imagery and deep learning

Sylvain Colomer[1,5]*, Aravind Harikumar[1], Leen Al-Sameai[1], Vincent Riedel[1], Vincent Seigner[2], Ian Major[2], Nathalie Isabel[2], Ingo Ensminger[1,3,4]

[1] Department of Biology, University of Toronto Mississauga, Mississauga, ON L5L 1C6, Canada
[2] Laurentian Forestry Centre, Natural Resources Canada, Canada
[3] Graduate Department of Cell and Systems Biology, University of Toronto, Toronto, ON M5S 3B2, Canada
[4] Graduate Department of Ecology and Evolutionary Biology, University of Toronto, Toronto, ON M5S 3B2, Canada
[5] ETIS Laboratory, UMR8051, CY Cergy Paris Université, ENSEA, CNRS, Cergy, France

*Corresponding author: sylvain.colomer@utoronto.ca

Keywords: Remote sensing, UAV, Hyperspectral camera, Instance segmentation, Individual tree crown delineation, Deep learning.

Abstract

The development of high-performance, affordable UAVs has transformed vegetation monitoring, enabling observation of forest canopies at an unprecedented level of detail. UAV-derived datasets now provide high-fidelity structural and physiological information at the individual tree level across entire forest stands, offering novel insights into forest dynamics. In the context of increasing tree mortality, such data are becoming essential for understanding forest resilience and adaptation. However, exploiting this data requires effective individual tree crown segmentation algorithms (ITCS) at the forest scale, capable of tackling large-scale data and variability introduced by the environment. In this paper, we developed a new workflow designed to process UAV hyperspectral imagery at the forest scale, enabling automated ITCS and analysis. Our pipeline integrates hyperspectral-to-RGB conversion, ITCS, and centroid-based mask fusion. To assess the performance of our pipeline, we evaluated the model on two replicated white spruce common gardens in Canada, each comprising approximately 6,000 trees of similar age and structure. The experiments rely on a large multi-temporal dataset of hyperspectral imagery acquired during 60 UAV missions between 2022 and 2024, allowing us to evaluate the robustness of the proposed pipeline across a wide range of seasonal and acquisition conditions. Results show that the proposed pipeline achieves a mean segmentation performance of 0.536 mAP (0.885 mAP50) on the annotated dataset. At the forest scale, the system demonstrates strong detection capability with F1-scores of 0.948 at the Pintendre site and 0.863 at the Pickering site, successfully detecting most trees while maintaining stable performance across varying environmental conditions.

1. Introduction

The development of high-performance, affordable UAVs (Unmanned Aerial Vehicle) has profoundly transformed vegetation monitoring and remote sensing (Syed Hanapi et al., 2019). With the improvement of UAV performance and sensor miniaturization, it's now possible to obtain detailed observations of forest canopies at unprecedented spatial and temporal resolutions. UAV-derived datasets provide high-fidelity structural and physiological data across entire forest stands, offering novel insights into canopy function and forest dynamics (Danilevicz et al., 2022). This new level of detail makes it possible to extract information at the individual-tree level, a major step forward for forest monitoring (Zhen et al., 2016). In the context of accelerating climate change and increasing tree mortality, such data are becoming essential for understanding forest resilience and adaptation (Vacek et al., 2023).

However, transforming raw UAV data into actionable ecological information remains a challenging task. It requires complex algorithms, notably effective methods for doing individual-tree crown segmentation (ITCS), capable of handling large-scale datasets, crown structural complexity, and the variability introduced by illumination conditions (Gu et al., 2022; Harikumar et al., 2022). Although many individual-tree crown segmentation models have been proposed in recent years, most of them were originally designed to operate on small image subsets or limited forest plots (Syed Hanapi et al., 2019). Few have been adapted or optimized for deployment across large areas, a crucial step toward developing efficient and scalable forest monitoring tools. As a result, the generalizability and operational robustness of current approaches remain uncertain, highlighting the need for large-scale, standardized, and repeatable pipelines tailored to real-world forest monitoring applications.

In this paper, we address this problem by proposing a new deep learning pipeline designed for large-scale individual tree crown segmentation from UAV hyperspectral orthomosaics. The proposed pipeline enables the processing of full-forest orthomosaics by combining hyperspectral-to-RGB conversion, tile-based instance segmentation using Mask R-CNN with a ConvNeXt-V2 backbone, and a post-processing strategy to merge overlapping predictions into a consistent tree-level map (Woo et al., 2023). To assess the performance of this new pipeline, a complete evaluation was conducted using UAV hyperspectral imagery acquired over two experimental forest stands in Canada between 2022 and 2024, comprising a total of 60 UAV missions. This unusually large dataset allowed us to evaluate the robustness and consistency of the proposed pipeline across multiple seasons and acquisition conditions. The evaluation included both segmentation accuracy on manually annotated areas and large-scale detection performance by comparing the detected trees with field-measured tree positions.

The remainder of the paper is organized as follows: part 2 describes the materials and methods, including the dataset and the pipeline description. Section 3 presents the experimental results, and Section 4 discusses the results and provides the conclusions of this study.

2. Materials and methods

2.1 Site and data acquisition using a UAV

The tests performed in this study were carried out on hyperspectral images of two replicated white spruce sites in Canada, each comprising approximately 6,000 trees of similar age and structure distributed over an area of about six hectares (see Figure 3)(D’Odorico et al., 2021). The first one is located near the town of Pickering in Ontario (43°58'32"N 79°09'17"W) and the second one near the town of Lévis, in Quebec (46°44'26.0"N 71°08'02.0"W). Due to different climates and topography, the two sites have different tree distributions: the Pickering site, with its warmer climate, is sparser with many dead trees (between 1000 to 1500), unlike the Pintendre site, which is much denser and has few dead trees.

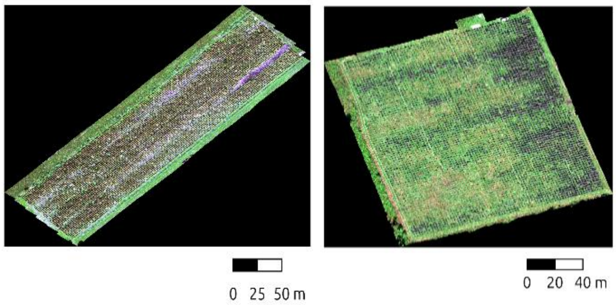

Figure 1: Illustration of the Pintendre site (left) and the Pickering site (right). Although the two sites are replicated experimental stands, significant differences in tree height distribution can be observed.

The images were acquired using two Freefly Alta X multirotor drones (Freefly Systems, USA) equipped with a Nano-Hyperspec® pushbroom hyperspectral camera (Headwall Photonics Inc., USA) and a Velodyne VLP-16 LiDAR sensor (Velodyne Lidar Inc., USA). The geolocation workflow relied on two sources of GNSS information. First, UAV navigation was supported by an RTK-GNSS system using a Freefly RTK GPS Ground Station (Freefly Systems, USA). Second, a Trimble SPS585 GNSS Smart Antenna (Trimble Inc., USA) was used to improve georeferencing accuracy and for validation. (see Table 1 for details of the system).

Table 1: Details of the drone, sensors and GNSS systems used in this study for acquisition.

System Component	Model / Version	Manufacturer (Country)
UAV	Alta X	Freefly Systems, USA
Flight Controller	Pixhawk 4	Holybro, China
Hyperspectral Camera	Nano-Hyperspec® (Nano HP series)	Headwall Photonics Inc., USA
LiDAR Sensor	VLP-16	Velodyne Lidar Inc., USA
RTK GNSS Base Station	Freefly RTK GPS Ground Station	Freefly Systems, USA
GNSS base station	SPS585	Trimble, USA

Flights were conducted regularly in 2022, 2023 and even 2024, with a minimum of 1 flight per month to capture seasonal variability, including differences in illumination and weather. Flight speed ranged between 2 and 5 m/s, depending on lighting and meteorological conditions (see Figure 3). Due to availability issues, the two drones were equipped with different lenses: the Pickering drone used a 12mm lens and the Pintendre drone used a 17 mm lens. This choice resulted in the adoption of different flight altitudes between the two sites, with 35 meters above ground level for Pickering and 75 meters above ground level for Pintendre (see Table 2 details on the conducted flights).

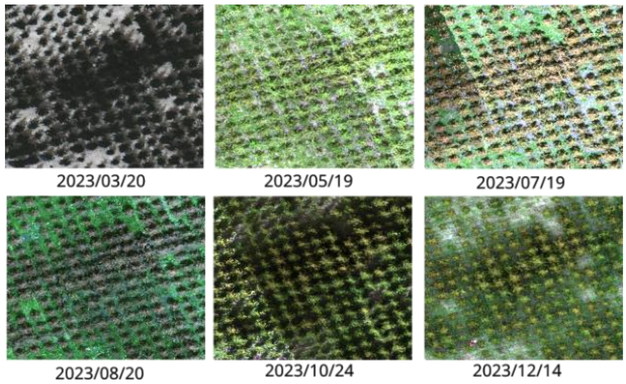

Figure 3: Illustration of some flight conditions encountered at the Pickering site.

Complementary ground-based surveys were conducted in April 2022 and 2023 by field technicians to evaluate the health condition of each individual tree across the two experimental sites. For each tree, a technician evaluates the health state of the tree (alive, dead or missing tree) and measures its size.

Table 2: Overview of flights conducted during the project

Site	Pickering	Pintendre
Dimension (width x height)	180m x 150m	330m x 80m
Quality (width x height)	13794 x 15058	12660 x 12224
Fly altitude	35	75
Duration	8:43	19:23
Fly in 2022	6	7
Fly in 2023	24	20
Fly in 2024	3	0

2.2 Preprocessing and generation of Hyperspectral Images

The drone records three types of data: GNSS coordinates, LiDAR point clouds, and hyperspectral tiles. Four major stages are required to georeference, clean, and merge the data (see Figure 2). First, the GNSS data were cleaned, RTK-corrected, and merged using the PostPac UAV 8.8 software. Second,

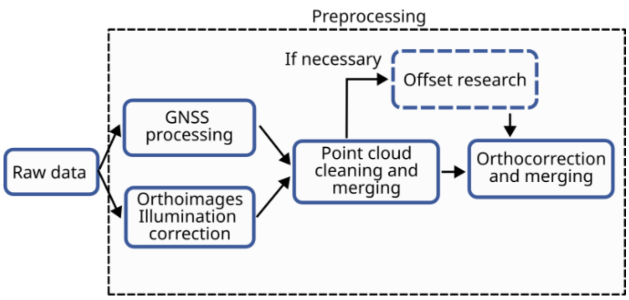

Figure 2: Illustration of the drone data pre-processing pipeline used to transform raw hyperspectral, LiDAR, and GNSS data into a single geolocated hyperspectral image.

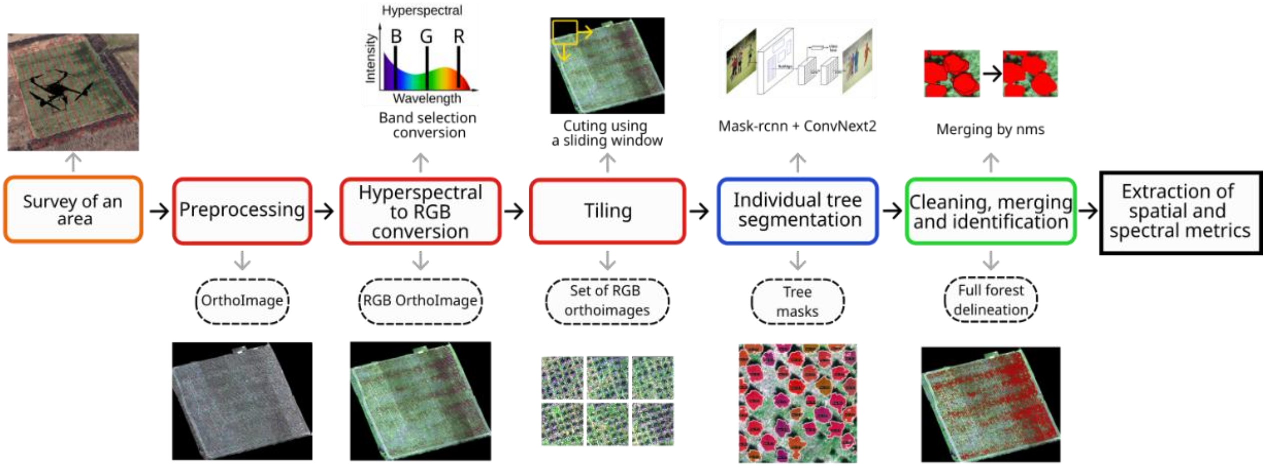

Figure 4: Illustration of the proposed new pipeline. This can be broken down into four processing steps: hyperspectral-to-RGB conversion by band selection, tiling by sliding window, individual tree crown segmentation via the Mask-RCNN model and the ConvNeXt-V2.

illumination effects on the hyperspectral tiles were corrected by adjusting radiance and reflectance values using both black and white references. The black reference was obtained by fully covering the hyperspectral sensor, while the white reference was measured during a dedicated calibration flight over a 3 m × 3 m tarp with known reflectance. Third, LiDAR point clouds acquired during each flight were georeferenced, merged, and manually cleaned using the software CloudCompare.

Finally, the corrected hyperspectral tiles are georeferenced and merged using the corrected GNSS data and the cleaned point cloud. For this step, it is necessary to apply correction offsets (roll, pitch, yaw, altitude, and focal length) to each tile to improve the rendering of the hyperspectral image. Specific to the drone and the flight site, offset calculation is not automatic: to determine them, an incremental trial-and-error procedure was followed, in which the parameters are adjusted little by little and visually evaluated using fixed reference points (buildings, roads). For Pickering, a single set of offsets common to all missions could be determined. For Pintendre, adjustments were necessary on numerous missions, particularly in 2023.

2.3 Orthoimages re-alignment

The preprocessing workflow used to generate the hyperspectral orthomosaics does not ensure a perfect spatial alignment between UAV missions. Small positioning errors can remain between acquisitions, and orthomosaics from different dates may sometimes be shifted by one to two meters relative to each other. Such offsets can affect the temporal consistency of the analysis and make it more difficult to compare the position of individual trees across missions. To correct these discrepancies, an additional realignment step was applied to all missions. This procedure was based on a set of control points that were manually identified on the orthomosaics. These control points correspond to stable and easily recognizable elements of the landscape that remain unchanged over time, such as roads, buildings, or other persistent features visible on the images. For each mission, the position of the control points was compared with the corresponding points from a reference mission. The average displacement between the two sets of points was then estimated and used to determine the horizontal shift affecting the orthomosaic. A simple translation was subsequently applied to the mission in order to realign it with the reference dataset.

2.4 Training data and annotations

To train the tree segmentation model, a 100x100m square containing approximately 350 trees was manually annotated on 8 flights over Pickering and 7 flights over Pintendre. This 100x100 square was chosen in areas representative of the tree distribution in both sites. The dates were chosen to cover a wide range of different brightness conditions, weather, and seasons. Annotations were performed at the instance level, with individual tree canopies delineated using polygons on RGB images derived from the hyperspectral imagery (easier to annotate). Only the canopies of the trees were delineated (no trunks or understory vegetation). The same trees were intentionally annotated across multiple flight dates in order to capture temporal variability in canopy appearance throughout the season.

A three-step annotation protocol was followed: initial annotation by a volunteer, verification and suggestions of correction by an expert, then application of the corrections by the same volunteer. To make the work of the annotator easier, the annotated areas of 100x100 were subdivided into 12 × 12 m patches with a 20% spatial overlap to avoid missing trees. Only the complete trees on the patches were annotated (not those cut by the division) to avoid errors. For the final dataset, all annotations were merged and subsequently re-partitioned into 10 × 10 m patches, allowing the inclusion of partially visible tree canopies for model training. Annotations were made using CVAT.ai (Computer Vision Annotation Tool), a web-based platform that allows collaborative management of image annotation tasks. The final dataset was divided into training, validation, and test sets using a 60/20/20 split to ensure that the evaluation was performed on trees that were not used during model training.

2.5 Model architecture

The proposed pipeline processes hyperspectral images through a sequence of four successive stages (see Figure 4). These steps include (1) hyperspectral-to-RGB conversion, (2) tiling of the converted image using a sliding window approach, (3) individual tree crown segmentation using Mask-Rcnn+ConvNeXt-V2, and (4) post-processing. Each step of the pipeline is described in detail in the following subsections.

2.5.1 Hyperspectral-to-RGB conversion

The pipeline begins the processing of a hyperspectral image by converting the input hyperspectral image into an RGB composite image. This conversion enables the use of transfer learning with deep learning models originally designed for RGB imagery, which constitute most of the pretrained architectures available in the current literature. Multiple methods exist to perform this conversion: the color matching functions approach (Magnusson et al., 2020), which mimic the reception of eye red, green and blue colours to obtain a realistic image rendering; the manual band selection approach, consisting of directly selecting bands corresponding to the blue (450–485 nm), green (500–565 nm), and red (625–740 nm) portions of the electromagnetic spectrum (Wang et al., 2020); or the dimensionality reduction approach that aims to exploit the full spectral information captured by the hyperspectral sensor (Grbovic et al., 2025). In a previous study, these methods were investigated and led to the selection of the manual band selection strategy for the hyperspectral-to-RGB conversion, as it provided better performance than the other approaches.

2.5.2 Tiling of the hyperspectral image

Following the hyperspectral-to-RGB conversion, the model processes the image by subdividing it into smaller tiles. This tiling step is required to handle large images and to ensure compatibility with the fixed input size expected by the deep learning architecture. In this model, the image is divided using a sliding window approach, where square tiles are extracted sequentially across the entire image with a fixed window size and stride. This strategy allows for complete spatial coverage of the scene while preserving local spatial context for each tile.

2.5.3 Individual tree crown segmentation

In the model, the individual tree crown segmentation of the tiles is performed using the Mask-RCNN framework with a ConvNext-V2 backbone (He et al., 2017; Woo et al., 2023). Mask R-CNN is a classic instance segmentation deep network which enables individual objects to be delineated as separate instances within an image. It operates by first generating candidate object regions with a specific network called Region Proposal. Then, for each proposed region, it extracted features and passed them to multiple network heads responsible of a specific operation. A classification and bounding box regression head predicts the object class and refines the bounding box coordinates, and a fully convolutional head produces the pixel-wise masks delineating the trees. ConvNeXt-V2 is a new backbone that implements several important novelties aimed at improving feature representation in convolutional neural networks. It modernizes classical CNN designs by incorporating updated convolutional blocks, improved normalization schemes, and better training strategies. When used as the backbone of Mask R-CNN, ConvNeXt-V2 enhances feature extraction capabilities, getting the model competitive segmentation performances.

Then, for each tile extracted in the previous stage, Mask-RCNN+ ConvNeXt-V2 is used to extract the trees it detects on the images, while assigning them a confidence score. The output of this block is thus many non-georeferenced masks resulting from segmentation, many of which correspond to the same trees (overlapping region of the sliding window).

2.5.4 Cleaning and merging by NMS

The final stage of the pipeline aims to perform three important post-processing tasks: georeferencing the predicted segmentation masks; removing abnormal masks; deduplication of the overlapping masks based on their spatial coordinates; and identifying individual trees. The georeferencing step is realized by combining the spatial coordinates of the tiles, preserved during the tiling process, with the pixel-level positions of the predicted masks within each tile. This step is required because the Mask R-CNN + ConvNeXt-V2 architecture operates in image space and does not natively handle geographic coordinates. The removal of abnormal masks is based on prior knowledge on the general structure of white spruce crowns, which are known to exhibit specific sizes and an approximately circular shape. Three metrics were used to identify abnormal masks: the surface of the mask in m², the solidity of the polygon and its circularity. The surface area corresponds to the area of the polygon representing the tree crown. The solidity is a metric that quantifies how compact a polygon is relative to its convex hull (Gonzalez and Woods, 2008), and is defined as:

$$Solidity = \frac{Area\ of\ the\ region}{Area\ of\ the\ convex\ hull\ of\ the\ region}$$

Circularity measures how close the shape is to a perfect circle hull (Gonzalez and Woods, 2008), and is computed as:

$$Circularity = \frac{4\pi A}{Perimeter^2}$$

In the pipeline, if the polygon is too small, not solid enough or not circular enough, it will be rejected. The deduplication of the overlapping polygon is based on a non-maximum suppression (NMS) algorithm. First all the polygons are sorted according to their confidence score, so the most confident crowns are considered first. Then, by a cartesian product, each polygon is compared to the remaining one using the Intersection over Union (IoU) metric, defined as:

$$IoU(p_i, p_j) = \frac{Area(p_i \cap p_j)}{Area(p \cup p_j)}$$

With p_i the polygon i, and p_j the polygon j, and Area(.) the surface area of a mask. Only the polygon with an $\mathrm{IoU}(p_i, p_j) \leq \tau$ are retained. The last stage, the tree identification, is a primordial stage for future data exploitation and evaluation. To carry out this work, we used the ground truth available in the form of a file containing the approximate coordinates of each tree and their ID. For each known tree coordinate, we selected the polygon that was closest to it, using the position of the polygon's centroid.

2.6 Metrics

The performance of the pipeline was evaluated using two complementary approaches. First, segmentation performance was assessed using the standard evaluation method for segmentation models, based on the Average Precision (AP) computed on representative annotated patches of the site (Segmentation performance). Second, detection performance was evaluated at the scale of the entire site by comparing the detected trees with field measurements using the F1-score (large-scale performance). The first method likely provides the most accurate evaluation of segmentation quality, but it is limited in spatial extent because it requires a large amount of manual annotations. The second method provides less information about the quality of the predicted masks but makes it possible to evaluate whether the model detects the correct number of trees at large scale.

Segmentation performance was assessed by computing the mean Average Precision metric (mAP) on all the 100 × 100 m annotated areas[1]. This metric is computed by determining the area under the model's precision–recall curve, a curve that reflects the balance between the model's ability to correctly identify objects (precision) and its ability to detect all relevant objects (recall). To construct this curve, it is necessary to compute the precision and recall of the model at different confidence thresholds, i.e., the threshold above which the model's prediction is considered a correct tree segmentation, as defined in the following equations:

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

In addition to mAP, we report AP50 and AP75, which correspond to AP computed at IoU thresholds of 0.50 and 0.75, respectively. Finally, AP_s and AP_m evaluate segmentation performance for small and medium-sized objects, providing additional insight into the model's performance depending on object size.

Large-scale performance was evaluated by computing the pipeline F1-score across the entire site. To do this, predicted tree crowns were first matched with field-measured tree positions based on centroid proximity. From this matching, the number of correctly detected trees (true positives), missing trees (false negatives), and incorrect detections (false positives) was determined. These three values were then used to compute the F1-score, a metric that summarizes the balance between correctly detected trees, missed trees, and false detections. The F1-score is defined as follows:

$$F1 = 2 * \frac{precision.recall}{precision + recall}$$

2.7 Hardware and software

All experiments were conducted using the MMDetection framework, an open-source object detection and segmentation toolbox developed by the Multimedia Laboratory of CUHK and SenseTime Research (Chen et al., 2019). The performances presented for each method are those resulting from the best iteration i.e. with the best mAP during the training of the model on the evaluation dataset.

3. Results

3.1 Average segmentation performance of Mask-Rcnn+ ConvNeXt-V2 by site.

Table 3 presents the average segmentation performance of Mask-RCNN+ConvNeXt-V2 on the annotation dataset by site. Results show that the model has good performance with a mean mAP of 0.536 ± 0.129, an mAP50 of 0.885 ± 0.0997 and a mAP75 of 0.610 ± 0.224. mAP_s and mAP_m show that the model performs better with medium-sized object than small size object. The Pintendre and Pickering line show that the model performs better with images coming from the site of Pintendre than the site of Pickering, with a difference of 16.4% in mAP, of 11.2% in AP50 and of 24.2% in mAP75. It's interesting to note that the mAP_s of Pickering is much better than the one of Pintendre with 0.383 ± 0.173 against 0.173 ± 0.185, showing that the model performs better at segmenting small trees of Pickering than the one of Pintendre.

Table 3: Average segmentation performance of the Mask R-CNN + ConvNeXt-V2 model across the Pintendre (PIN) and Pickering (PIK) sites. Values represent mean ± standard deviation.

Site	mAP	AP50	AP75	mAP_s	mAP_m
PIN	0.569 ± 0.0887	0.923 ± 0.0431	0.663 ± 0.186	0.173 ± 0.185	0.621 ± 0.0939
PIK	0.489 ± 0.166	0.830 ± 0.132	0.534 ± 0.263	0.383 ± 0.173	0.593 ± 0.166
Mean	0.536 ± 0.129	0.885 ± 0.0997	0.610 ± 0.224	0.259 ± 0.205	0.610 ± 0.126

3.2 mAP of Mask-Rcnn+ ConvNeXt-V2 by months and sites

Figure 5 shows the temporal evolution of the mAP obtained by the model across Pickering and Pintendre UAV missions. Results show that, regardless of the site, the model generally obtains good performances with an mAP usually ranging between 0.45 and 0.7. Three drops in performance can be observed: 0.29 on 2023/05/19 in Pickering, 0.08 on 2023/09/26 in Pintendre and 0.22 on 2023/11/15 in Pickering. Except for these dates, the model gives similar performances for both the Pickering and Pintendre sites. This suggests that the model performance remains relatively stable over time and across sites, with only a few isolated drops likely related to specific acquisition conditions.

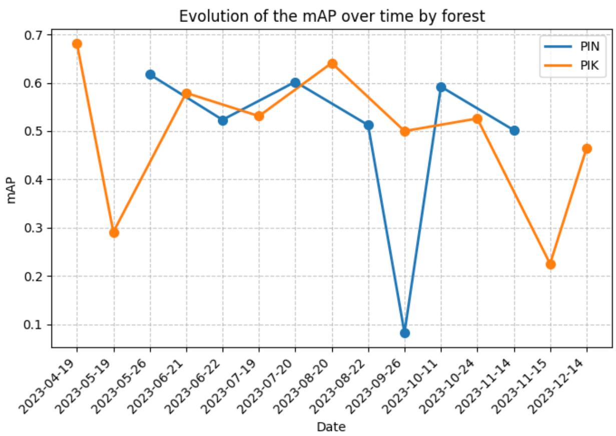

Figure 5: Evolution of the mAP of the Mask-RCNN + ConvNeXt-V2 model across UAV missions for the Pintendre (PIN) and Pickering (PIK) sites. Each point corresponds to the segmentation performance obtained for a given acquisition date.

3.3 Average large-scale detection performances

Table 4 presents the evaluation of the large-scale detection performance of the Mask-RCNN + ConvNeXt-V2 model on all trees in the Pickering and Pintendre sites. Results show that the model achieves better detection performance in Pintendre than in Pickering, with a higher proportion of correctly detected trees

[1] The evaluation presented here was conducted exclusively on the 20% of annotated areas assigned to the test subset.

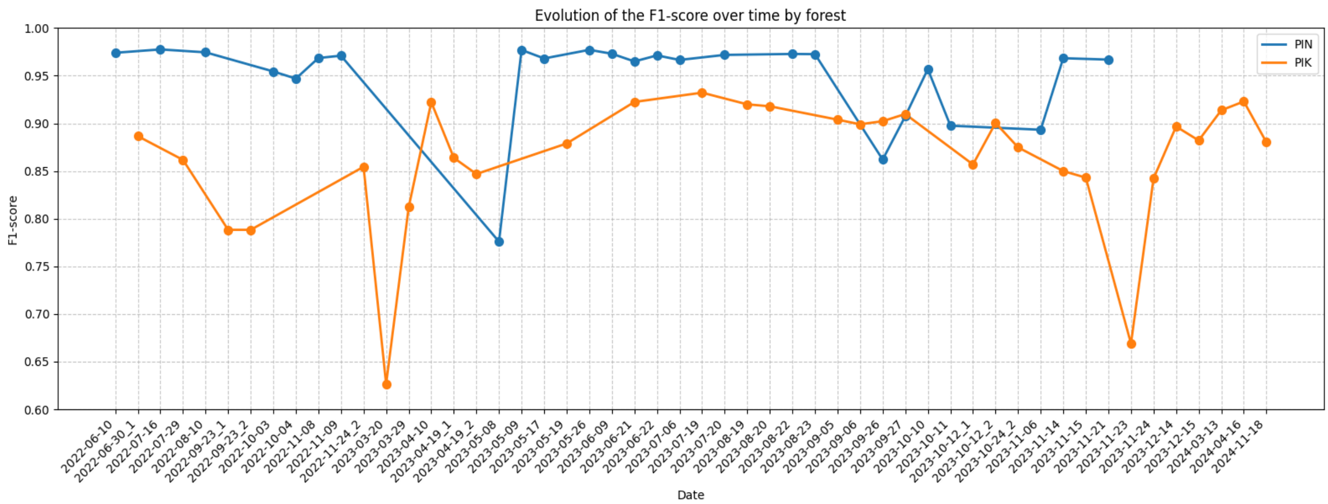

Figure 6: Evolution of the F1-score across UAV missions conducted between 2022 and 2024 in the Pintendre (PIN) and Pickering (PIK) forests. Each point corresponds to the performance obtained for a single UAV mission. The lines connect the missions chronologically to illustrate temporal variations in detection performance

(92.9% vs 78.6%), a lower proportion of missing trees (6.3% vs 13.5%), and fewer incorrect detections (0.8% vs 7.9%). This difference is also reflected in the F1-score, which is higher for Pintendre (0.948) than for Pickering (0.863).

Table 4: Large-scale tree detection performance of Mask-RCNN + ConvNeXt-V2 for the Pintendre and Pickering sites. Values represent mean ± standard deviation across missions.

Metric	**Pintendre (mean ± sd)**	**Pickering (mean ± sd)**
Correctly detected trees (%)	92.9 ± 6.9	78.6 ± 9.3
Missing trees (%)	6.3 ± 6.4	13.5 ± 8.8
Incorrect detections (%)	0.8 ± 0.6	7.9 ± 3.4
Precision	0.962 ± 0.030	0.878 ± 0.049
Recall	0.936 ± 0.066	0.853 ± 0.098
F1-score	0.948 ± 0.047	0.863 ± 0.069

3.4 Temporal evolution of the F1-score across the years

Figure 6 illustrates the temporal evolution of the F1-score for individual tree detection across UAV missions conducted in the Pintendre (PIN) and Pickering (PIK) sites. Each point represents the performance obtained for a specific mission, while the lines connect the missions chronologically to highlight the temporal evolution of the detection performance.

For the Pintendre site, the F1-score remains generally very high throughout the observation period, most of the time exceeding 0.9. Only a few decreases in performance can be observed at the beginning and at the end of the season, where the detection performance temporarily drops to approximately 0.78. Outside these periods, the model maintains consistently high detection accuracy. For the Pickering site, the F1-score typically ranges between 0.85 and 0.95, which is slightly lower than the values observed for Pintendre. Similarly to Pintendre, decreases in performance can be observed at the beginning and the end of the season. However, some of these drops occur during periods when no flights were conducted over the Pintendre site. In particular, the lower performance observed in March and April is mainly explained by the presence of snow on the ground, which makes the segmentation of individual tree crowns more challenging.

Overall, this figure explains the average detection performance observed on the Table 4. The results indicate that the model generally performs slightly better in the Pintendre site than in the Pickering site. However, this difference may also be influenced by the fact that more missions were conducted in Pickering during periods of the year characterized by more challenging environmental conditions.

3.5 Tree height and errors

Figure 7 and Figure 8 present histograms of tree height distributions for two orthomosaics acquired in May. In each figure, the blue histogram represents the overall distribution of tree heights across the site, while the red histogram shows the distribution of living trees that were not detected by the model as a function of their height. Figure 7 shows that most trees in the Pintendre site have heights ranging between approximately 1.5 m and 5 m. The figure also indicates that detection errors are not strongly concentrated around a specific tree height class and appear relatively evenly distributed across the height range. However, very tall trees seem to be more easily detected by the system, as fewer detection errors are observed for this category.

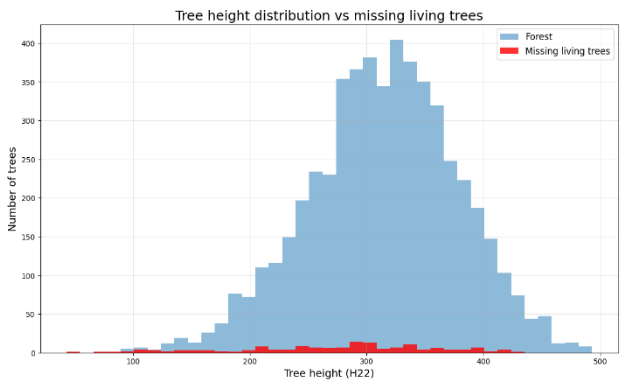

Figure 7: Distribution of tree heights across the Pintendre forest and distribution of missing living trees detected by the pipeline. The blue histogram represents the overall height distribution of trees measured in the field, while the red histogram highlights the subset of living trees that were not detected by the model.

Figure 8, on the other hand, shows that trees in the Pickering site are generally smaller than those in Pintendre, with heights ranging between approximately 0.5 m and 5 m. The figure also reveals that most detection errors occur for trees with heights between 0.5 m and 1.5 m, a size class that is much less represented in the Pintendre site. These results suggest that the smaller average tree size in Pickering may partly explain the lower detection performance observed at this site compared to Pintendre.

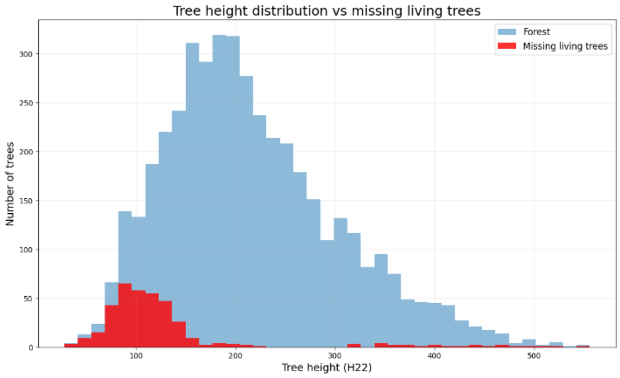

Figure 8: Distribution of tree heights across Pickering and distribution of missing living trees detected by the pipeline.

4. Discussion

In this study, we proposed a new tree segmentation architecture designed to process hyperspectral aerial imagery at the scale of entire forest sites. Composed of four main stages, the pipeline combines relatively conventional methods (conversion of the hyperspectral orthomosaic into RGB imagery, division into tiles suitable for the segmentation model, merging the predicted masks into a consistent representation) with more advanced techniques for the segmentation itself, based on the Mask R-CNN model with a ConvNeXt-V2 backbone, one of the most recent and high-performing deep learning architectures reported in the literature. Two main questions guided this work. First (i), can hyperspectral imagery of entire forests be efficiently segmented using a deep learning model originally designed for RGB images after spectral conversion and transfer learning. Second (ii), to what extent the proposed architecture can maintain stable performance over time despite the natural variability of acquisition conditions encountered during two years of forest monitoring.

Regarding the first question (i), our results show that the proposed architecture performs well and achieves relatively strong performance. From the perspective of small-scale segmentation performance, the model obtains relatively high average values, with a mAP of 0.536, an AP50 of 0.885, and an AP75 of 0.610. These values are comparable to those reported in the literature, where AP50 is typically between 0.80 and 0.90, AP75 between 0.30 and 0.50, and mAP between 0.30 and 0.60 (Tong & Zhang, 2025), although these values may vary depending on forest type and acquisition conditions. The obtained performances are also comparable to those reported for instance segmentation tasks on more conventional datasets (Yang Qingand Peng, 2024).

When comparing the results obtained for each forest, noticeable differences can be observed. Overall, the architecture performs better in the Pintendre forest, with improvements of approximately 16% in mAP, 11.2% in AP50, and 24% in AP75 compared to Pickering. Interestingly, these results also show that the model segments small trees better in Pickering than in Pintendre (about 54% higher performance), likely due to the higher proportion of small trees present in this forest.

From the perspective of large-scale detection performance, the proposed architecture generally achieves relatively high scores. In the Pintendre forest, the system correctly identifies 92.9% of trees, with 6.3% of trees missed and 0.8% incorrectly detected. In contrast, in the Pickering forest, 78.9% of trees are correctly detected, 13.5% are missed, and 7.9% are incorrectly detected. Although the performance is relatively high in both cases, a clear difference between the Pintendre and Pickering forests can be observed. As previously noted, the model detects approximately 8% more trees in Pintendre, while Pickering shows around 50% more missed trees and a higher number of incorrect detections. This difference is explained by Figures 7 and 8, which illustrate how the distribution of tree sizes and detection errors differs between the Pickering and Pintendre forests. These figures show that most errors occur on trees smaller than 1.5 meters, which are highly represented in the Pickering forest but relatively rare in the Pintendre forest. Two main factors may explain this phenomenon: insufficient spatial resolution to clearly distinguish small trees in the imagery, and the presence of understory vegetation that partially covers small trees and complicates their detection.

Regarding the second question (ii), our results show that the proposed architecture produces relatively stable performance over time. For segmentation performance, the average standard deviation remains moderate, with values of 0.129 for mAP and 0.0997 for AP50. The standard deviation for AP75 is higher, with an average value of 0.224. However, this is not unexpected, as achieving a segmentation overlap greater than 75% is particularly challenging. From the perspective of large-scale detection performance, variations of approximately 6% in correctly detected trees are observed for the Pintendre site and 9.3% for the Pickering site. Considering that UAV missions conducted at the beginning and the end of the year were performed under particularly difficult conditions (e.g., snow or cloudy weather), these results remain reasonable and demonstrate the ability of the model to maintain good performance over an entire year. A closer examination of Figure 7 provides further insight into the missions associated with lower performance. For the Pintendre site, only one significant drop in performance can be observed on 2023/05/08. At the Pickering site, two drops are observed: one on 2023/03/20 and another on 2023/11/23. The performance decrease on 2023/03/20 can be explained by the presence of snow on the ground, which makes tree segmentation considerably more difficult. The drops observed on 2023/05/08 and 2023/11/23 occur at the beginning and the end of the season and correspond to missions that were classified by our team as having difficult cloudy conditions. In both cases, an additional flight was conducted the following day to obtain better acquisition conditions. Such weather conditions are particularly challenging because they introduce strong illumination variations in the orthomosaic, which negatively affect model performance. Further work will be conducted to better quantify and analyze the influence of flight conditions on segmentation and detection performance.

These results open several perspectives for future work. First, although the overall performance of the proposed architecture is relatively high, the results show that it remains affected by flight conditions. Improving the robustness of the model to variations in acquisition conditions therefore represents an important direction for future research. One of the main strengths of the dataset used in this study is its large size and temporal diversity, which provides an opportunity to explicitly incorporate these variations during training. Future work could therefore focus on developing models that are more robust to illumination and weather variability, for instance by integrating these variations directly into the training data or by designing dedicated data augmentation strategies. A natural option would be to use the full spectra of the hyperspectral camera or to work on multimodality with LiDAR data.

Second, it would be interesting to evaluate the proposed model on more complex forest environments, particularly deciduous forests, which present significantly greater challenges. Our preliminary investigations suggest that three main factors distinguish such environments from the plantation site considered in this study. First, natural forests generally exhibit much less clearly defined boundaries between individual trees, which makes both the annotation process and the segmentation task more difficult. Second, from the perspective of flight acquisition protocols, obtaining imagery of comparable quality is more challenging due to the more complex spatial structure of natural forests, which often requires adapted flight strategies. Finally, deciduous forests are composed of trees whose appearance changes significantly throughout the year, notably through leaf emergence and leaf fall, introducing additional variability that may complicate the development of stable segmentation models over time.

Third, this work also opens the door to a natural extension of the proposed pipeline: the exploitation of hyperspectral data for forest monitoring applications. While the present study primarily focuses on tree crown detection and segmentation, hyperspectral imagery contains rich spectral information that can potentially be used to characterize physiological and biochemical properties of vegetation. An important next step would therefore be to investigate to what extent the proposed architecture could be extended to accurately assess the health status of individual trees using hyperspectral information. Such an approach could enable the detection of early stress signals related to drought, disease, or nutrient deficiency, and provide valuable insights into forest dynamics at the individual-tree level. Ultimately, combining accurate tree delineation with hyperspectral analysis could contribute to the development of automated monitoring systems for large-scale forest health assessment.

In conclusion, the results presented in this paper demonstrate that the proposed architecture can effectively segment individual trees from UAV hyperspectral imagery across two white spruce experimental site. The pipeline achieves reliable performance both in terms of segmentation accuracy and large-scale tree detection, while maintaining relatively stable results across multiple acquisition dates. The experiments also highlight the strong influence of flight conditions on model performance, as variations in illumination, weather, or seasonal context can significantly affect segmentation results. These findings confirm the potential of the proposed approach for large-scale forest monitoring while emphasizing the importance of carefully controlled acquisition conditions when applying deep learning methods to UAV forest imagery.

.

Acknowledgements

We acknowledge financial support from Ontario Genomics, Genome Quebec, and Genome Canada through the FastPheno projects. We also thank the people who contributed to the annotation of the dataset, particularly X, who played a key role in the pre-processing and annotation of the missions used in this study.

References

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., … Lin, D. (2019). *MMDetection: Open MMLab Detection Toolbox and Benchmark*. http://arxiv.org/abs/1906.07155

Danilevicz, M. F., Gill, M., Anderson, R., Batley, J., Bennamoun, M., Bayer, P. E., & Edwards, D. (2022). Plant Genotype to Phenotype Prediction Using Machine Learning. In *Frontiers in Genetics* (Vol. 13). Frontiers Media S.A. https://doi.org/10.3389/fgene.2022.822173

D’Odorico, P., Schönbeck, L., Vitali, V., Meusburger, K., Schaub, M., Ginzler, C., Zweifel, R., Velasco, V. M. E., Gisler, J., Gessler, A., & Ensminger, I. (2021). Drone-based physiological index reveals long-term acclimation and drought stress responses in trees. *Plant Cell and Environment*, *44*(11), 3552–3570. https://doi.org/10.1111/pce.14177

Grbovic, Z., Panic, M., Filipovic, V., Brdar, S., De Villiers, H., Mensink, M., & Chauhan, A. (2025). An efficient encoding spectral information in hyperspectral images for transfer learning of Mask R-CNN for instance segmentation of tomato sepals. *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3569213

Gu, W., Bai, S., & Kong, L. (2022). A review on 2D instance segmentation based on deep neural networks. In *Image and Vision Computing* (Vol. 120). Elsevier Ltd. https://doi.org/10.1016/j.imavis.2022.104401

Harikumar, A., D’odorico, P., & Ensminger, I. (2022). Combining Spectral, Spatial-Contextual, and Structural Information in Multispectral UAV Data for Spruce Crown Delineation. *Remote Sensing*, *14*(9). https://doi.org/10.3390/rs14092044

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). *Mask R-CNN*. http://arxiv.org/abs/1703.06870

Magnusson, M., Sigurdsson, J., Armansson, S. E., Ulfarsson, M. O., Deborah, H., & Sveinsson, J. R. (2020). Creating RGB Images from Hyperspectral Images Using a Color Matching Function. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2045–2048. https://doi.org/10.1109/IGARSS39084.2020.9323397

Syed Hanapi, S. N. H., Shukor, S. A. A., & Johari, J. (2019). A Review on Remote Sensing-based Method for Tree Detection and Delineation. *IOP Conference Series: Materials Science and Engineering*, *705*(1). https://doi.org/10.1088/1757-899X/705/1/012024

Vacek, Z., Vacek, S., & Cukor, J. (2023). European forests under global climate change: Review of tree growth processes, crises and management strategies. In *Journal of Environmental Management* (Vol. 332). Academic Press. https://doi.org/10.1016/j.jenvman.2023.117353

Wang, Q., Zhang, F., & Li, X. (2020). Hyperspectral Band Selection via Optimal Neighborhood Reconstruction. *IEEE Transactions on Geoscience and Remote Sensing*, *58*(12), 8465–8476. https://doi.org/10.1109/TGRS.2020.2987955

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I. S., & Xie, S. (2023). *ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders*. http://arxiv.org/abs/2301.00808

Zhen, Z., Quackenbush, L. J., & Zhang, L. (2016). Trends in automatic individual tree crown detection and delineation-evolution of LiDAR data. In *Remote Sensing* (Vol. 8, Number 4). MDPI AG. https://doi.org/10.3390/rs8040333