The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIX-B2-2026
XXV ISPRS Congress 2026 “From Imagery to Understanding”, Commission II, 4–11 July 2026, Toronto, Canada

Todo Fir Crown Instance Segmentation in Dense Plantation Forest Using Polar-FFT and Treetop Queries

Yusuke Iuchi[1], Soki Nishiwaki[2], Fumio Takeuchi[3], Mika Takiya[4], Takanori Emaru[5]

[1] Graduate School of Engineering, Hokkaido University, Sapporo, Hokkaido, Japan - convallariamajalis@eis.hokudai.ac.jp
[2] Graduate School of Engineering, Hokkaido University, Sapporo, Hokkaido, Japan - nishiwaki.soki.j7@elms.hokudai.ac.jp
[3] Forestry Research Institute, Hokkaido Research Organization, Bibai, Hokkaido, Japan – takeuchi-fumio@hro.or.jp
[4] Forestry Research Institute, Hokkaido Research Organization, Bibai, Hokkaido, Japan – takiya-mika@hro.or.jp
[5] Faculty of Engineering, Hokkaido University, Sapporo, Hokkaido, Japan - emaru@eng.hokudai.ac.jp

Keywords: Remote sensing, Forestry, Photogrammetry, Instance Segmentation, Individual Tree Detection

Abstract

Monitoring individual trees from unmanned aerial vehicle (UAV)-derived orthomosaics and digital surface models (DSMs) is important for forest management, but instance segmentation remains difficult in dense planted forests in Japan, where Structure from Motion (SfM)-derived DSMs often exhibit blurred crown boundaries, noise, and substantial variation in quality. Existing approaches can benefit from treetop information, but they are sensitive to threshold selection and may introduce false positives when incorrect treetop candidates are provided. To address this problem, we propose a method based on the polar coordinate transform and the fast Fourier transform (PFFT) that represents the local DSM shape around treetop candidates as compact descriptors and integrates them into Mask2Former. The descriptors are used both to suppress low-reliability candidates and to provide spatially meaningful treetop queries to the decoder, thereby improving the separation of adjacent crowns. We evaluated the method on *Abies sachalinensis* (Todo fir) plantation data acquired at two sites in Hokkaido, Japan, using Mask R-CNN and Mask2Former as baselines. Compared with standard Mask2Former, the proposed method improved mAP_{50} from 90.14% to 90.87%, mAP_{75} from 52.18% to 55.47%, and the F1 score at a confidence threshold of 0.5 from 89.86% to 92.08%. It also reduced the number of false positives by 41% without increasing false negatives. Qualitative results showed fewer over-merged crowns and better crown-boundary delineation. These results indicate that treetop-centered local shape cues are effective for instance segmentation in densely planted forests, although further validation across additional species and regions is needed.

1. Introduction

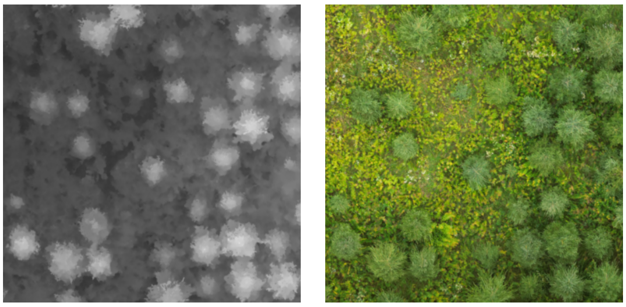
Amid climate change, monitoring forest conditions and tree density across large areas over time has become increasingly important for both ecosystem conservation and a stable timber supply. Recent advances in unmanned aerial vehicles (UAVs) and aircraft have made it possible to collect data efficiently in both natural and planted forests, where ground surveys alone often cannot provide sufficient spatial or temporal coverage. Consequently, the use of such data together with AI methods based on deep learning in forest research and development has expanded, with the expectation that these technologies will improve the efficiency of forest management. Among these approaches, instance segmentation, which detects and delineates objects at the individual level, has become an important technique for planted-forest resource surveys and related applications. In remote sensing, the adaptation of instance-segmentation models developed in computer vision has also enabled the detection of individual trees in situations that were difficult to address using classical methods.

1.1 Background

Instance-segmentation models can achieve very high performance in regions that are well represented in the training data. However, their performance is known to degrade under changes in weather conditions and site characteristics, and their cross-region applicability remains limited. For this reason, improving generalization requires training data with sufficient regional diversity. In practice, however, forest data acquisition with UAVs is highly sensitive to weather and wind. In particular, when orthomosaics and digital surface models (DSMs) are generated using Structure from Motion (SfM), data quality often deteriorates because of feature-matching inconsistencies, wind-driven leaf movement, and occlusion by lower branches. Moreover, SfM outputs vary substantially depending not only on site conditions but also on the equipment used and the acquisition settings. As a result, large differences in data quality can arise between datasets. Figure 1 shows a dataset acquired in Quebec (Lefebvre and Laliberté, 2024), whereas Figure 2 shows a dataset acquired in Hokkaido. Both datasets were resampled to a common ground sampling distance of 5 cm/pixel, and the DSM values in each patch were linearly scaled between the local minimum and maximum for visualization. In the Quebec dataset, individual trees have high contrast, and the boundary between understory vegetation and tree crowns is clear. In contrast, the Tobetsu dataset exhibits blurred, cloud-like DSM patterns caused by wind effects and occlusion, and the ground surrounding tree-covered areas is often obscured.

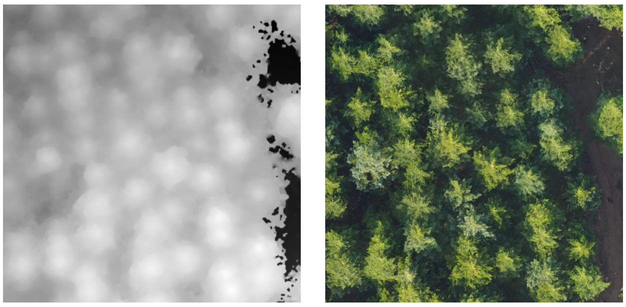
In Japan, coniferous species are often planted at high density, as shown in Figure 2. As a result, the ground surface between neighboring trees is not directly reconstructed by SfM, and crown boundaries become diffuse and blurred. Under such conditions, even human observers may struggle to distinguish individual trees, and manual annotation becomes laborious and time-consuming. This difficulty makes it hard to build large, diverse datasets. Furthermore, because DSMs rarely contain clear edges, height information alone often provides only weak cues for deep neural networks (DNNs). Consequently, reliable

https://doi.org/10.5194/isprs-archives-XLIX-B2-2026-691-2026 |



(a) (b)

Figure 1. Examples of the orthomosaic and linearly scaled DSM derived from the UAV dataset acquired in Quebec, Canada. Adapted from Lefebvre and Laliberté (2024), released under the Creative Commons CC0 1.0 Public Domain Dedication.



(a) (b)

Figure 2. Examples of the orthomosaic and linearly scaled DSM derived from the UAV dataset acquired in Tobetsu, Hokkaido, Japan, illustrating the cloud-like DSM artifacts discussed in the text.

separation of individual trees becomes difficult, and small trees within the same patch may be merged into neighboring crowns, leading to missed detections. Even substantial improvements in model performance do not readily translate into near-perfect accuracy. These rare but challenging cases are often underrepresented in the dataset and cannot easily be resolved by simply adding images from other regions or applying generic data augmentation. Therefore, in Japan, methods capable of detecting targets even from low-quality DSMs are required. Furthermore, increasing robustness to such DSMs may improve applicability to regions with differing data quality and contribute to better cross-regional detection performance.

1.2 Related Works

Early studies related to individual-tree instance segmentation employed image processing techniques such as watershed segmentation and region growing to delineate individual tree crowns. These methods had the advantage of enabling single tree segmentation without requiring large amounts of annotated training data. However, their segmentation accuracy was highly dependent on parameter tuning. In addition, with the increasing use of cost-effective UAV-acquired RGB data, the focus of research in individual tree delineation gradually shifted from conventional image processing approaches to deep neural network (DNN)-based methods.

Given the limitations of these classical approaches, studies increasingly adopted DNN models developed in computer vision, such as Mask R-CNN (He et al., 2017), for individual tree detection. These models made it possible to extract features from RGB images and detect and segment individual tree regions. Early studies mainly relied on fine-tuning with relatively small datasets, but subsequent progress included the development of large-scale individual crown datasets, such as those provided by National Ecological Observatory Network (NEON) sites (Weinstein et al., 2020b), as well as forest remote-sensing models such as DeepForest (Weinstein et al., 2020a). Nevertheless, generalization to regions and acquisition conditions different from those represented in the training data, as well as robustness to low-quality data, remain important challenges.

To improve robustness, recent studies have proposed instance segmentation methods that either constrain the search region in advance or exploit characteristic points, such as treetops, as cues. Teng et al. evaluated, as one of their comparative approaches, an instance-segmentation method that used local maxima in a DSM as point prompts (Teng et al., 2025) for the Segment Anything Model (SAM) (Kirillov et al., 2023). Although this method did not achieve the highest accuracy among the compared models, they reported that local maxima derived from tall grasses and small plants protruding from the ground increased the number of false positives. Lungo Vaschetti et al. improved instance-segmentation accuracy by first estimating peak points likely to correspond to tree centers using a DNN model and then generating masks based on those points (Lungo Vaschetti et al., 2025). In their study, they noted that an excessive number of peak candidates degraded overall performance. Yu et al. improved light detection and ranging (LiDAR)-based individual tree segmentation by detecting treetop seed points from UAV-LiDAR point clouds (Yu et al., 2024). In their method, crown surface points were first extracted from the point cloud, the neighborhood around a candidate treetop was divided into eight angular sectors, and distance-height profiles in each sector were analyzed to locate crown boundaries and detect treetop seed points. Their individual tree segmentation was then performed using a seed-point-based region-growing approach. Unlike detection methods based on local maximum filters, this approach improved detection accuracy without depending on a predefined window size. However, the authors also showed that detection performance depended on parameter settings, and they noted that in mixed forests, stretched branch tops of tall conifers could be misidentified as seed points, reducing precision. Thus, although instance segmentation methods or instance tree segmentation methods that use constrained detection regions or treetop cues are effective for improving accuracy, false positives caused by erroneous peak detection remain a challenge.

To suppress such false positives while also addressing overmerging caused by blurred DSM representations, we propose a network module that focuses on shape information around treetops. Trees, particularly conifers, have radially structured shapes at the instance level, and the shape around treetops may provide a strong clue for tree detection. Therefore, by combining the polar coordinate transform with the fast Fourier transform (PFFT), the local canopy shape around treetop candidates is encoded into descriptors that can be used as network inputs for detection. These descriptors, which encode local shape information as PFFT treetop features, are used in two ways: (1) they are computed for all treetop points, and low-reliability treetop points are removed before network input, and (2) the remaining descriptors are used as treetop queries with explicit spatial meaning to guide the instance segmentation decoder. By

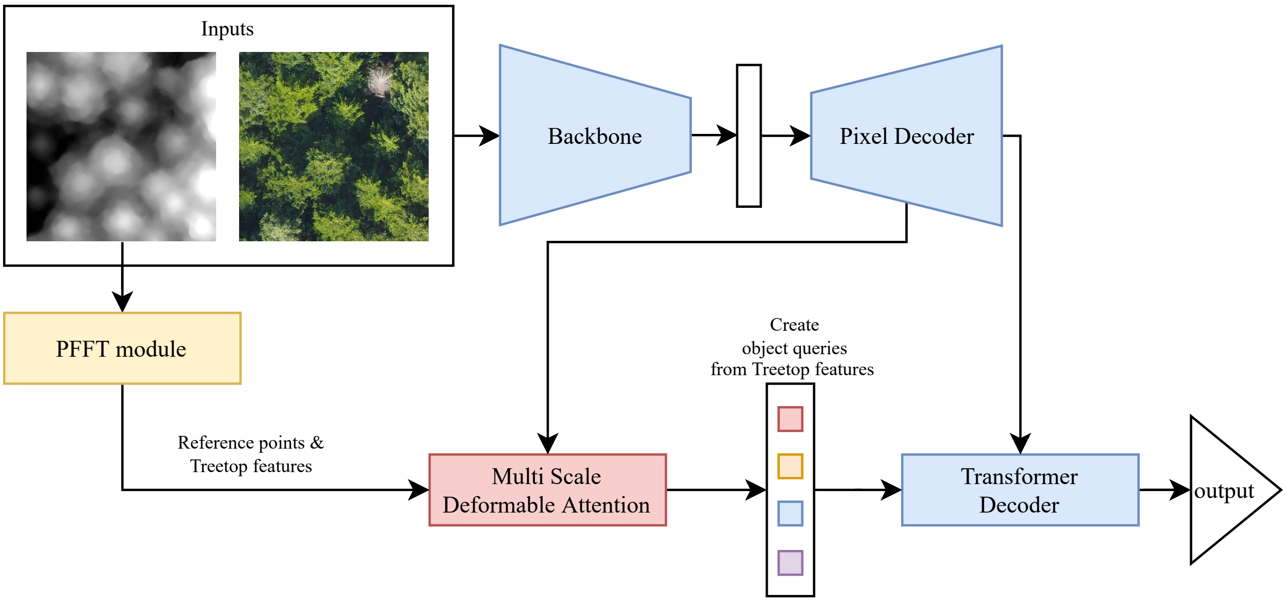


Figure 3. Overall architecture of the proposed framework, showing the extraction of PFFT-based treetop descriptors and their integration into the instance-segmentation decoder.

using these structures, the network can be guided toward treetop locations in a manner that is insensitive to threshold settings. In experiments in planted forests, the proposed method reduced false positives (FPs) by 41% without increasing false negatives (FNs) at a confidence threshold of 0.5 in a test set from a stand different from those in the training data, and improved the F1 score from 90.0% to 92.0% compared with the baseline Mask2Former. In addition, mask quality was improved, and mean Average Precision at an Intersection over Union threshold of 0.75 (mAP75) increased by 3.3%.

The main contributions of this study are as follows. First, we propose PFFT, which represents the local DSM shape around each treetop as a feature associated with that treetop. Second, by integrating PFFT into the deformable attention module, we improve tree-detection accuracy.

2. Method

2.1 Concept

The central idea of the proposed method is to incorporate local shape information around treetops into instance segmentation for individual trees. Figure 3 illustrates the overall architecture. The network is designed to extract treetop candidates from the DSM together with descriptors of candidate quality and local shape, so that the network can automatically suppress the influence of false positive candidates even when such candidates are introduced.

Coniferous trees can be approximated, in an idealized sense, as three-dimensional cones. Under this assumption, the surface profile sampled along a circumference centered on the treetop varies relatively consistently in the radial direction, whereas variation in the tangential direction remains small. If the angular waveform is nearly constant, the direct-current (DC) component dominates in the Fourier domain. Real trees, however, are not perfectly symmetric and may be tilted or distorted, so non-DC frequency components also appear. Based on this observation, we represent the local shape around each treetop candidate in the frequency domain using the polar coordinate transform and the fast Fourier transform. The resulting descriptors, together with the candidate positions, are then incorporated into the Mask2Former head to support instance segmentation.

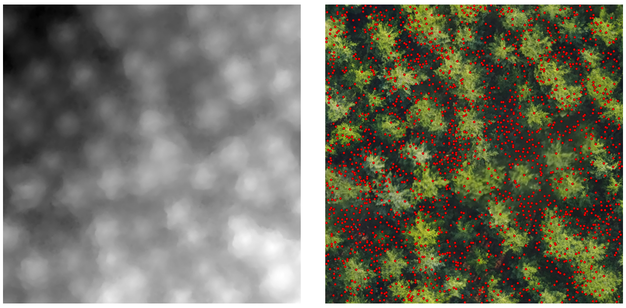
2.2 Extraction of Treetop Candidates

We first extract treetop candidates from the DSM as the basis for the polar coordinate transform. The DSM is locally enhanced by top-hat filtering and robustly normalized so that a common response threshold can be used across scales. In this study, we apply the Laplacian of Gaussian (LoG) to the DSM at multiple scales to detect a broad set of convex structures as candidate points. The LoG response $R(x, y)$ and the candidate set $\mathcal{C}$ are defined by Eq. (1). The threshold for the scale-normalized LoG response is set to 0.1 in the robustly normalized space to retain a broad set of peaks as candidates. As shown in Figure 4, this process detects numerous convex peaks in the DSM.

$$\begin{aligned} R(x,y) &= \max_{\sigma\in\Sigma} \sigma^2 \left|\nabla^2 (G_\sigma * D)(x,y)\right|, \\ \mathcal{C} &= \{(x,y) \mid R(x,y) > \tau,\ (x,y) \text{ is local max}\}, \end{aligned} \tag{1}$$

where
D = DSM
G_σ = Gaussian kernel
Σ = LoG scales
R = LoG response
τ = threshold
$\mathcal{C}$ = candidate set

Next, for each candidate point $\mathbf{p}_i = (x_i, y_i)$, the DSM is sampled on circumferences of multiple radii. Eq. (2) represents the radial component obtained by normalizing, by the radius, the height difference between the candidate center and the circumference. On the other hand, the tangential component



(a) (b)

Figure 4. Detection of treetop candidates by LoG. (a) Source DSM. (b) Candidate peaks extracted from the DSM.

representing change in the circumferential direction is given by Eq. (3).

$$C_{i,b}(\theta_n) = \frac{D(x_i, y_i) - D(x_i + r_b \cos\theta_n,\ y_i + r_b \sin\theta_n)}{r_b}, \tag{2}$$

where $C_{i,b}$ = radial component
(x_i, y_i) = candidate position
r_b = band radius

$$\begin{aligned} S_{i,b}(\theta_n) &= \frac{D^{+}_{i,b}(\theta_n) - D^{-}_{i,b}(\theta_n)}{2 r_b \Delta\theta}, \\ D^{+}_{i,b}(\theta_n) &= D(x_i + r_b \cos(\theta_n + \Delta\theta),\ y_i + r_b \sin(\theta_n + \Delta\theta)), \\ D^{-}_{i,b}(\theta_n) &= D(x_i + r_b \cos(\theta_n - \Delta\theta),\ y_i + r_b \sin(\theta_n - \Delta\theta)), \end{aligned} \tag{3}$$

where $S_{i,b}$ = tangential component
$\Delta\theta$ = angular interval
$D^{+}_{i,b}, D^{-}_{i,b}$ = neighboring samples

Next, the components obtained by Eq. (2) and Eq. (3) are integrated as a complex number, and the shape change in the angular direction is transformed into the frequency domain. The complex signal is defined by Eq. (4), and its Fourier coefficients are defined by Eq. (5).

$$X_{i,b}(\theta_n) = C_{i,b}(\theta_n) + j S_{i,b}(\theta_n), \tag{4}$$

where $X_{i,b}$ = complex signal
j = imaginary unit

$$F_{i,b}(k) = \sum_{n=0}^{N_\theta - 1} X_{i,b}(\theta_n) \exp\left(-j \frac{2\pi k n}{N_\theta}\right), \tag{5}$$

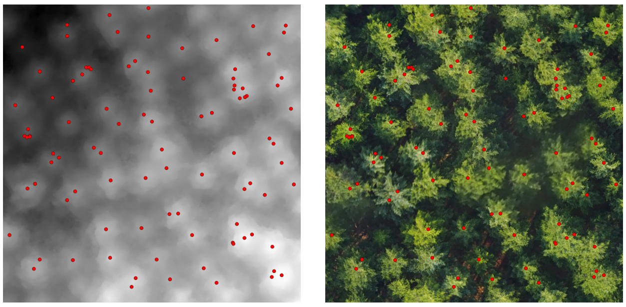
where $F_{i,b}(k)$ = Fourier coefficient
N_θ = the number of angular samples
k = frequency index

By Eq. (5), the shape at each radius is decomposed into frequency components, and the relationship between the DC component and higher-order components can be analyzed. For candidates close to an ideal cone, the DC component dominates, whereas higher-order frequency components increase in candidates with strong distortion or noise. Therefore, the stability of each circumference is defined by Eq. (6).

$$\begin{aligned} A_{i,b}(0) &= |F_{i,b}(0)|^2, \\ A_{i,b}(k) &= |F_{i,b}(k)|^2 + |F_{i,b}(N_\theta - k)|^2, \\ P_{i,b} &= A_{i,b}(0) + \sum_{k=1}^{\lfloor N_\theta/2 \rfloor} A_{i,b}(k), \\ s_{i,b} &= \frac{A_{i,b}(0)}{P_{i,b} + \varepsilon}, \end{aligned} \tag{6}$$

where $A_{i,b}(k)$ = paired spectral energy
$P_{i,b}$ = total energy
$s_{i,b}$ = stability
ε = small constant

By Eq. (6), it is possible to evaluate, for each circumference, how strongly the DC component dominates the total energy. For each candidate point, the stability at multiple radii is calculated, and all points extracted by LoG are ranked according to this stability measure. For each image, the top-ranked points are selected as the final treetop candidates, as shown in Figure 5. At this stage, the treetop candidates still include false positives.



(a) (b)

Figure 5. Top-ranked treetop candidates selected by PFFT according to gradient stability, visualized on (a) the DSM and (b) the RGB image.

2.3 Integration into the Network Structure

Each treetop candidate is associated with circumferential information sampled at multiple radii, together with the amplitude and phase at each radius. These frequency features are rearranged so that they can be handled easily by Mask2Former. The feature vector for each band is given by Eq. (7).

$$\begin{aligned} \mathbf{f}_{i,b} = [&s_{i,b}, \rho_{i,b,1}, \cos\phi_{i,b,1}, \sin\phi_{i,b,1}, \\ &\rho_{i,b,2}, \cos\phi_{i,b,2}, \sin\phi_{i,b,2}, \\ &\rho_{i,b,3}, \cos\phi_{i,b,3}, \sin\phi_{i,b,3}], \end{aligned} \tag{7}$$

where $\mathbf{f}_{i,b}$ = reordered band feature
$\rho_{i,b,k}$ = amplitude ratio
$\phi_{i,b,k}$ = phase

In Eq. (7), in addition to stability, features obtained by converting the amplitude ratios and phases of low-order frequencies into sine and cosine representations are used. Next, all band features for each candidate point are concatenated and fed into a small Multi-Layer Perceptron (MLP), which projects them to the same dimension as the Mask2Former object queries. At the same time, the candidate coordinates are normalized by the image size and used as reference points for multi-scale deformable attention. The projected features, normalized coordinates, and features extracted by multi-scale deformable attention (MSDA) are expressed by Eq. (8).

$$\begin{aligned} \mathbf{t}_i &= \mathrm{MLP}\big([\mathbf{f}_{i,1}; \cdots; \mathbf{f}_{i,B}]\big), \\ \hat{\mathbf{p}}_i &= \left(\frac{x_i}{W}, \frac{y_i}{H}\right), \\ \Delta\mathbf{t}_i &= \mathrm{MSDA}\big(\mathbf{t}_i, \{\mathbf{M}_l\}_{l=1}^{L}, \hat{\mathbf{p}}_i\big), \end{aligned} \tag{8}$$

where $\mathbf{t}_i$ = projected PFFT feature
$\hat{\mathbf{p}}_i$ = normalized position
$\mathbf{M}_l$ = feature map at level l
B = the number of radial bands
L = the number of feature levels
$\Delta\mathbf{t}_i$ = extracted image feature

The frequency features of treetop candidates are converted into query representations, and surrounding image features are further extracted using the candidate positions as reference points. Next, the features obtained by Eq. (8) are used to update $\mathbf{t}_i$, and the updated features are used to initialize the corresponding object queries. The updated query is defined by Eq. (9).

$$\mathbf{q}^{\mathrm{new}}_{\pi(i)} = \mathbf{t}_i + \lambda\Delta\mathbf{t}_i, \tag{9}$$

where $\mathbf{t}_i$ = treetop candidate feature before update
$\Delta\mathbf{t}_i$ = injected feature from Eq. (8)
$\mathbf{q}^{\mathrm{new}}_{\pi(i)}$ = new query after update
λ = learnable scalar
$\pi(i)$ = query assignment

Eq. (9) incorporates local shape information around treetop candidates into the object queries, thereby assisting subsequent instance separation in the Transformer decoder.

3. Experimental Setup

3.1 Model Setup

To evaluate the effectiveness of the proposed method, we used data acquired from two planted-forest sites in the Hokkaido region. To isolate the contribution of the geometric cues targeted by the proposed method, the analysis was restricted to a single tree species. This was intended to separate errors caused by class misclassification from the effects of the proposed feature representation itself. The target species was *Abies sachalinensis*, which has the largest stock volume in Hokkaido and is widely used as a construction material. Because this species is of high practical importance in Hokkaido forestry, it was selected as the target species in the single-species setting. To evaluate generalization under similar plantation conditions, the training and test sets were acquired at sites separated by 104 km. More specifically, the training set consisted of two planted stands in Tobetsu, Hokkaido, whereas the test set was acquired in Shiribeshi, Hokkaido. The training data were collected in November 2017 and July 2018, and the test data were collected in September 2019.

As baselines, we used Mask R-CNN (He et al., 2017) and Mask2Former (Cheng et al., 2022), both of which are widely used as strong baselines in forest remote sensing in Japan. Mask R-CNN is a two-stage object detector, and it predicts an instance mask in parallel with classifying each proposal region and regressing its bounding box. Because it explicitly separates individual targets on the basis of region proposals, it is widely used as a standard baseline for instance segmentation. Mask2Former is a Transformer-based segmentation model that aggregates features for each target region through query-based mask prediction and masked attention. A key advantage of Mask2Former is that it can handle semantic, instance, and panoptic segmentation within a unified framework, and its representational power and flexibility have made it a strong baseline in recent studies. In our implementation of Mask2Former, the Transformer decoder had four stages and used 200 object queries. For the proposed method, training was performed using 50 ordinary object queries and 150 treetop queries, so that the model retained a fallback set of learnable queries even when no reliable treetop candidates were detected.

For each dataset, the DSM and orthomosaic were resized to a ground sampling distance of 5 cm/pixel and then divided into 512×512 tiles. The numbers of instances were 13,415 for the training set, 3,353 for the validation set, and 2,190 for the test set. Data augmentation consisted of random resizing, color jitter, and random flipping, each applied with probability 0.5. The RGB channels were normalized using the mean and variance of the training data. The DSM was preprocessed so that the minimum and maximum values within each patch were 0 and 1, respectively. As the backbone, we used ConvNeXtV2 (base size) pretrained on ImageNet. In the stem layer, the RGB image and DSM were convolved separately and then concatenated along the channel dimension. The overall learning rate was set to 1×10^{-4}, and the learning rate for the backbone was set to 0.1 times the overall learning rate. However, because the stem layer processes a different input modality from that used during pretraining, we did not apply learning-rate decay to this layer, allowing it to adapt more quickly than the pretrained layers. AdamW was used as the optimizer, with weight decay of 0.05. Weight decay was not applied to parameters for which it is unsuitable, such as scalar parameters and batch-normalization parameters. After linear warm-up through epoch 5, training continued to epoch 200 under cosine annealing. The final learning rate was set to 5×10^{-6}. Of the 200 epochs, we selected the model that achieved the best performance on the validation set. All training runs used the same preprocessing pipeline, and evaluation was reported as the average over three runs with random seeds 1, 42, and 2026.

3.2 Evaluation

For evaluation on the test set, the full orthomosaic was divided into patches with 50% overlap. All detections crossing patch boundaries were removed from the tiled outputs. This was done to avoid treating extremely small ground-truth (GT) masks that had been unintentionally split by patching as valid targets. After inference on all patches, hard non-maximum suppression (NMS) was applied to masks with an Intersection over Union (IoU) of at least 0.5. Mean Average Precision (mAP) was then evaluated with Max Dets set to 10000. The mAP calculation followed the Common Objects in Context (COCO) metrics (Lin et al., 2014). Because the number of inferred objects never exceeded Max Dets, this setting did not substantially affect the results.

In addition to mAP, we measured the F1 score, true positives (TPs), FPs, and FNs at a confidence threshold of 0.5. Detections with an IoU greater than 0.5 relative to the GT were regarded as true positives. Because the F1 score, precision, recall, FP, and FN depend on the confidence threshold, these metrics were also computed while varying the confidence threshold from 0.05 to 0.95 in increments of 0.05. By evaluating both overall detection accuracy through mAP and threshold-dependent operating characteristics, we examined not only model performance but also its expected behavior in practical applications.

4. Results

Table 1 summarizes the main quantitative results, and Figure 9 presents the evaluation metrics under varying thresholds. Examples of qualitative improvements and failure cases are shown in Figures 6, 7, and 8. The Mask R-CNN baseline already achieved relatively high mAP at an IoU threshold of 0.50. However, its performance at an IoU of 0.75 and its number of false positives indicate clear limitations in instance separation. Replacing the standard Mask2Former query formulation with the proposed PFFT-based query formulation produced the best overall results. Compared with the standard Mask2Former baseline, PFFT improved mAP50 from 90.14% to 90.87% and mAP75 from 52.18% to 55.47%. At the operating threshold of 0.5, the F1 score also improved from 89.86% to 92.08%. In addition, the proposed method slightly increased the mean TP count from 1907 to 1911, while reducing the mean FP count from 236 to 139 and the mean FN count from 194 to 190.

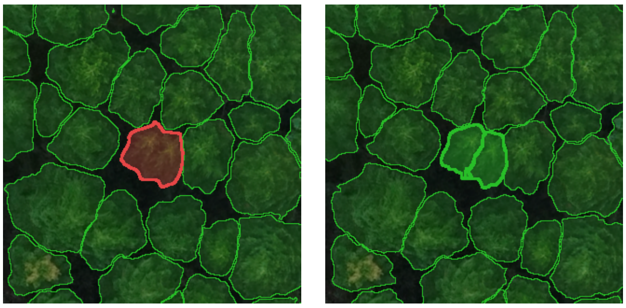
As shown in Figure 9-(a), the proposed method achieved higher accuracy than both Mask2Former and Mask R-CNN across the same IoU thresholds. The improvement was especially noticeable at an IoU threshold of 0.75, where the gain was approximately 3%. Figures 9-(b) to 9-(d) show the F1 score, precision, and recall as functions of the confidence threshold. Recall remained largely unchanged between Mask2Former and the proposed method, whereas precision improved consistently across thresholds, leading to a higher overall F1 score. Figures 9-(e) and 9-(f) show the FP and FN counts as functions of the confidence threshold. The proposed method yielded fewer FPs across all thresholds, whereas FN counts were comparable over most of the threshold range. Near a threshold of 0.7, Mask2Former briefly performed slightly better, but the proposed method outperformed it over the rest of the threshold range.

Method	mAP_{50}	mAP_{75}	F1@0.5
Mask R-CNN	84.25	32.63	83.14
Mask2Former	90.14	52.18	89.86
PFFT	90.87	55.47	92.08

Table 1. Quantitative comparison for large-image evaluation. Values of mAP and F1 are reported in %. The F1 score is reported at a score threshold of 0.5. All values are averaged over three random seeds.

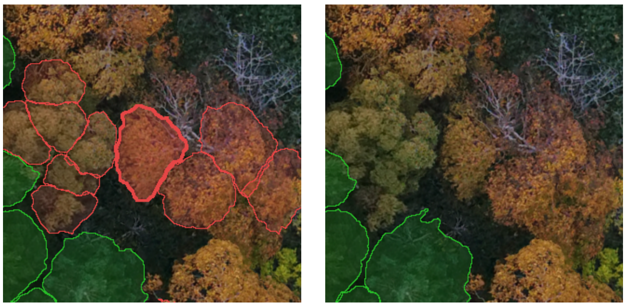
5. Discussion and Limitations

Adding the proposed method to the standard Mask2Former head produced a larger improvement in mAP75 than in mAP50.



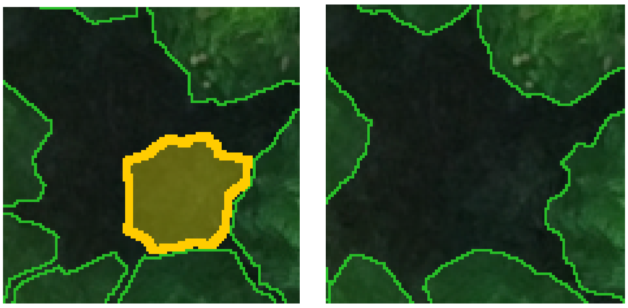
(a) (b)

Figure 6. Example showing suppression of crown over-merging. (a) Mask2Former. (b) Proposed method.



(a) (b)

Figure 7. Example showing suppression of a false positive on an unseen broadleaf tree. (a) Mask2Former. (b) Proposed method.



(a) (b)

Figure 8. Failure case in which a small crown is missed because heavy shadow obscures discriminative features. (a) Ground truth. (b) Proposed method.

This result suggests that PFFT improves the geometric fidelity of the predicted crown masks. That interpretation is consistent with the method design, in which descriptors near treetops are provided to the decoder as spatially meaningful initial hypotheses. However, as the IoU threshold increases, mAP approaches the baseline level. This suggests that the contribution of PFFT is limited for trees that already match the ground truth with high IoU.

A more pronounced effect was observed in the reduction of false positives. As shown in Figure 6, introducing PFFT reduced the number of masks spanning multiple crowns relative to the Mask2Former baseline. In Mask2Former, we frequently observed large over-merged masks that were not adequately removed by hard NMS because they overlapped correct masks. These merged masks therefore remained as false positives. We also observed cases in which over-merged masks suppressed

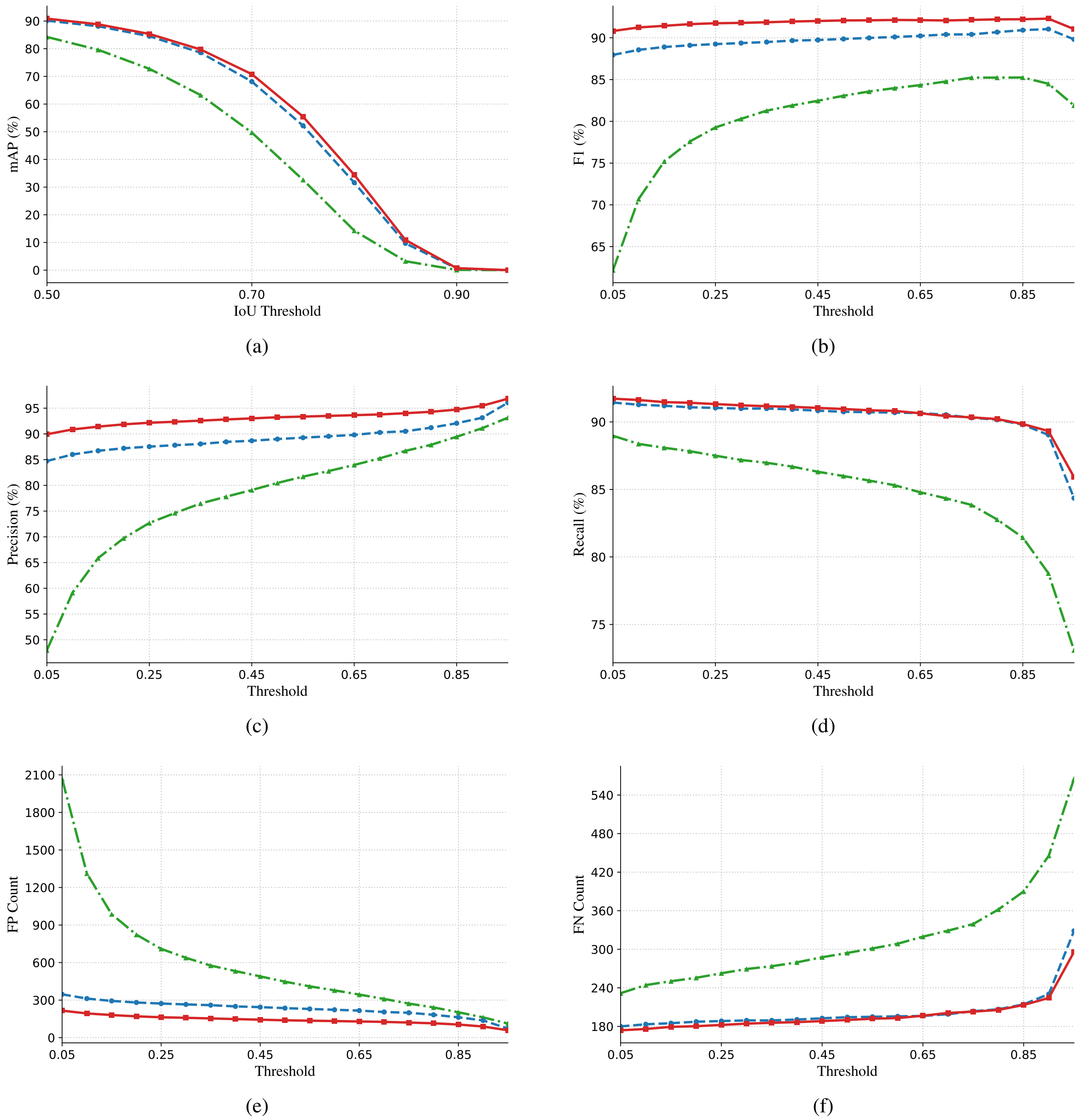


Figure 9. Performance under varying IoU and confidence thresholds. Red squares (■) denote the proposed method, blue circles (●) denote Mask2Former, and green triangles (▲) denote Mask R-CNN. (a) Mean Average Precision (mAP) as the IoU threshold varies from 0.50 to 0.95 in steps of 0.05. (b) F1 score. (c) Precision. (d) Recall. (e) False positives (FP). (f) False negatives (FN). In (b)-(f), the confidence threshold varies from 0.05 to 0.95.

correct masks through NMS. By contrast, when PFFT was introduced, crowns at the same locations were subdivided more appropriately. This improved separation appears to be one major factor behind the observed FP reduction.

The reduction in false positives was also observed in cases beyond the original target scenario of the proposed method. In Figure 7, Mask2Former detected broadleaf trees that were absent from both the training and validation sets. This appears to reflect a common failure mode in instance-segmentation models, in which features from an unseen class are ambiguously interpreted as those of a known class. The proposed method reduced the number of such examples. This may be partly because PFFT updates the queries using features restricted to the vicinity of individual trees, and self-attention among these queries may help the network distinguish one tree from another, even for previously unseen tree types. However, since the proposed method is not designed to enhance feature extraction itself, if an object was correctly detected by the proposed method, it was likely also detectable by Mask2Former. Moreover, because such failure cases may be readily mitigated by using data from mixed-species forests, they should not be regarded as inherent limitations of Mask2Former. Therefore, in this study, the primary effect of PFFT should be interpreted as being limited to improvement in boundary delineation.

Although the effectiveness of PFFT was clearly demonstrated, we also observed failure cases specific to this method. Some small crowns with indistinct peaks were still missed even when learnable queries and PFFT-derived queries were used together. Figure 8 shows one such example. In this image, many regions are heavily shaded by surrounding trees, and the small crown shown in the ground-truth image is almost completely obscured. In addition, when a tree is extremely small relative to its surroundings, the treetop peak itself may not be identifiable. In such cases, PFFT is not effective.

Overall, the principal effect of PFFT appears to be improved crown-boundary delineation and better feature extraction for trees with clearly identifiable treetops. The method is therefore well suited to settings such as dense planted forests, where crown contact and over-merging are major sources of error. Its effect is likely to be more limited in conditions such as natural mixed forests, where failures are driven primarily by misclassification among similar known classes.

6. Conclusions

In this study, we proposed a method based on the polar coordinate transform and the fast Fourier transform to mitigate crown over-merging. Compared with the baseline, the proposed method reduced false positives by 41% without increasing false negatives and improved the F1 score from 90.0% to 92.0%. It also improved mask quality, increasing mAP_{75} by 3.3%. Taken together, these results indicate that the proposed method is effective for reducing over-merging and improving crown delineation in dense planted forests.

At the same time, the dataset used in this study is heavily biased toward the Hokkaido region, and the generalization performance of the proposed method has not yet been verified. Future work should therefore include additional validation using datasets that cover multiple tree species and a wider range of regional conditions. In addition, the current framework does not include a loss function that directly evaluates the validity of the descriptors produced by the proposed method. Future work should therefore examine both loss-function designs that capture descriptor characteristics during learning and network architectures suited to the proposed method.

ACKNOWLEDGEMENTS

This work was supported by JSPS KAKENHI Grant Numbers JP24KJ0262.

References

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R., 2022. Masked-attention Mask Transformer for Universal Image Segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1280–1289. doi.org/10.1109/CVPR52688.2022.00135.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2980–2988. doi.org/10.1109/ICCV.2017.322.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollar, P., Girshick, R., 2023. Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE Computer Society, 3992–4003. doi.org/10.1109/ICCV51070.2023.00371.

Lefebvre, I., Laliberté, E., 2024. UAV LiDAR, UAV Imagery, Tree Segmentations and Ground Mesurements for Estimating Tree Biomass in Canadian (Quebec) Plantations. doi.org/10.20383/103.0979.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C. L., 2014. Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV 2014*, Springer International Publishing, 740–755. 10.1007/978-3-319-10602-1_48.

Lungo Vaschetti, J., Arnaudo, E., Rossi, C., 2025. TreePseCo: Scaling Individual Tree Crown Segmentation using Large Vision Models. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLVIII-M-7-2025, 275–282. doi.org/10.5194/isprs-archives-XLVIII-M-7-2025-275-2025.

Teng, M., Ouaknine, A., Laliberté, E., Bengio, Y., Rolnick, D., Larochelle, H., 2025. Assessing SAM for Tree Crown Instance Segmentation from Drone Imagery. arxiv.org/abs/2503.20199.

Weinstein, B. G., Marconi, S., Aubry-Kientz, M., Vincent, G., Senyondo, H., White, E. P., 2020a. DeepForest: A Python package for RGB deep learning tree crown delineation. *Methods in Ecology and Evolution*, 11(12), 1743-1751. doi.org/10.1111/2041-210X.13472.

Weinstein, B., Marconi, S., Zare, A., Bohlman, S., Graves, S., Singh, A., White, E., 2020b. NEON Tree Crowns Dataset. doi.org/10.5281/zenodo.3765872.

Yu, J., Lei, L., Li, Z., 2024. Individual Tree Segmentation Based on Seed Points Detected by an Adaptive Crown Shaped Algorithm Using UAV-LiDAR Data. *Remote Sensing*, 16(5). doi.org/10.3390/rs16050825.