TreePseCo: Scaling Individual Tree Crown Segmentation using Large Vision Models

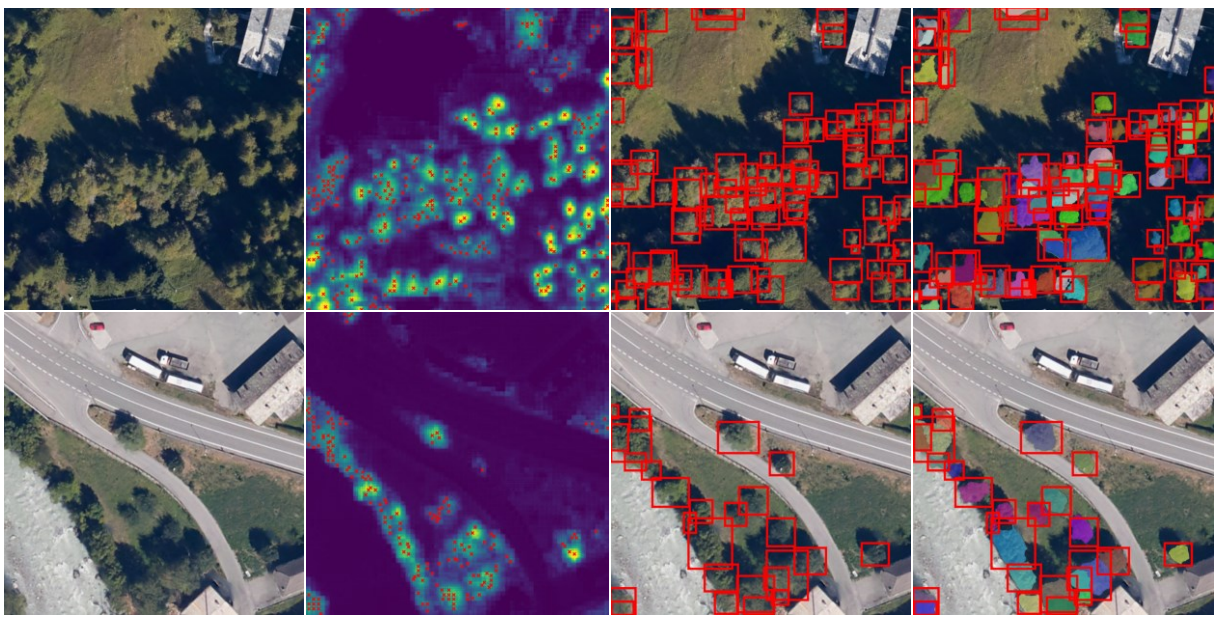
Jacopo Lungo Vaschetti[1], Edoardo Arnaudo[1], Claudio Rossi[1]

[1] Fondazione LINKS, *AI, Data & Space*, Torino, Italy

Keywords: Remote Sensing, Deep Learning, Tree Crown Delineation, Large Vision Models, Forest Monitoring.

Abstract

Forest monitoring through individual tree crown delineation is essential for sustainable management and carbon cycle assessment. This paper presents TreePseCo, an adaptation of the PseCo framework leveraging foundation models for automated tree crown segmentation in aerial imagery. Our approach implements a three-stage pipeline: (1) tree center detection using a modified Segment Anything Model (SAM) decoder that generates probability heatmaps, (2) instance mask generation through prompt-guided segmentation utilizing SAM's visual features, and (3) boundary refinement via specialized classification to eliminate false positives. We validate our method on two datasets: the extensive NEON dataset covering diverse U.S. forest ecosystems and the Valle d'Aosta dataset (VdA), a custom set of high-resolution RGB aerial images from northwestern Italian forests. Experiments against the popular DeepForest demonstrate that, while the baseline maintains excellent performance on its native NEON dataset, TreePseCo exhibits superior generalization capabilities when applied to new geographical contexts, achieving higher mAP scores on the VdA dataset. Our approach shows strength in detecting trees in densely clustered formations and identifying smaller tree instances, areas where existing methods often struggle. Overall, results suggest TreePseCo provides a robust foundation for comprehensive forest inventory applications across diverse environments.



(a) (b) (c) (d)

Figure 1. Visualization of our tree detection and segmentation approach. From left to right: the original satellite image (a), the heatmap indicating tree locations with red crosses marking the peaks (b); the predicted bounding boxes (c); the segmentation masks generated by the TreePseCo model (d). This visualization demonstrates the progressive steps of our detection pipeline, from initial heat-based localization to precise object boundary delineation.

1. Introduction

Forest ecosystems require accurate monitoring for sustainable management, with tree crown delineation serving as a fundamental component in remote sensing applications (Wulder, 2012). Modern technologies, particularly aerial imagery, have in fact emerged as cost-effective alternatives for tree crown delineation, enabling large-scale forest mapping (Huang, 2018).

High-resolution imagery allows for detailed analysis of forest structures, with tree crown dimensions demonstrating strong correlations with above-ground biomass and other biophysical parameters (Puliti, 2019). The accurate delineation of individual tree crowns from such imagery enables numerous forestry applications, including tree counting, species classification, forest health assessment, and growth monitoring. However, achieving precise segmentation of individual crowns presents

significant technical challenges, especially in dense, heterogeneous forest environments (Kandare, 2016).

Traditional methods for tree crown delineation, such as watershed segmentation, or region growing algorithms (Huang, 2018), often struggle with complex forest structures, particularly in dense, multi-layered canopies where crown boundaries overlap and are difficult to distinguish. Furthermore, they require extensive parameter tuning and expert knowledge, limiting their transferability across different forest types and imaging conditions (Shendryk, 2021).

Recently, deep learning approaches have revolutionized image analysis tasks, including object detection and semantic segmentation (Ma, 2019). Latest advancements in foundation models, such as the Segment Anything Model (SAM) (Kirillov, 2023), have further expanded the capabilities of deep learning in computer vision, thanks to their strong generalization abilities and adaptability to specialized domains with relatively small amounts of task-specific data. Despite the advances, scaling these models outside the domain of natural images is often challenging. The variability and the complex structure of forest canopies may create occlusion and shadowing effects, especially when observed from above, hindering the delineation process. Moreover, the limited availability of large-scale, annotated datasets specific to forestry applications further constrains the development of effective models (Davies, 2021).

This paper aims to address these challenges by presenting an adaptation of the PseCo framework (Huang, 2023) for automated individual tree crown segmentation. Our approach leverages the capabilities of SAM in a three-stage pipeline that combines point-based detection, prompt-guided segmentation, and bounding box refinement. First, a heatmap decoder leverages SAM's rich feature representation to accurately localize tree centers. Second, SAM's prompt-guided segmentation utilizes these centers as spatial cues for its mask decoder to generate initial crown boundary proposals. Last, a boundary refinement stage employs a specialized classifier to distinguish true tree crowns from similar vegetation or artifacts, significantly reducing false positives in complex forest environments.

We validate our approach on two datasets, namely NEON (Weinstein, 2021), providing bounding boxes from 22 regions in the U.S., and VdA, a dataset of 80 manually annotated high-resolution RGB aerial images from northwestern Italian forests, demonstrating improved accuracy and robustness compared to existing methods, particularly in challenging canopy environments. Our experimental results show that the proposed method provides more generalized results than DeepForest (Weinstein, 2019), the *de facto* standard for tree detection in remote sensing, especially in terms of recall in areas with high tree density.

The remainder of this paper is organized as follows: Section 2 reviews related works in tree crown delineation and deep learning approaches for forestry applications. Section 3 provides a brief description of the datasets employed, Section 4 describes our TreePseCo methodology, detailing the three-stage pipeline for tree crown segmentation. Section 5 describes the experimental setup, and discusses the results obtained. Section 6 concludes the paper, discussing limitations, and outlining directions for future research.

2. Related Work

Traditional approaches for tree crown segmentation typically rely on image processing techniques such as watershed segmentation, region growing, and valley-following algorithms to identify individual tree crowns (Huang, 2018)(Zhou, 2020)(Zörner, 2018). In particular, local maxima in Canopy Height Models (CHM) derived from LiDAR data have been widely employed for tree mapping (Zörner, 2018), with various methods using small-footprint airborne laser scanning explored for forest inventory data extraction in boreal forests (Hyyppä, 2008). LiDAR has proven particularly valuable for quantifying forest carbon pools and monitoring changes over time (Hudak, 2012)(Wulder, 2012). Combined approaches using hyperspectral and LiDAR data have improved individual tree crown delineation and species classification (Dalponte, 2019). However, LiDAR acquisitions require more specific expensive hardware that is not always available for large-scale forest monitoring.

As a cost-effective alternative, high-resolution RGB aerial imagery has emerged as a promising data source. High-resolution UAV data has been proven valuable for estimating biophysical properties in forest stands, providing advantages for accurate tree density and canopy height estimation (Puliti, 2019). However, traditional image processing approaches often struggle with dense canopies and require extensive tuning based on forest type and image resolution.

Recent advances in deep learning have revolutionized tree detection in remote sensing imagery, enabling for instance the creation of extensive datasets of individual tree crowns, such as the National Ecological Observatory Network (NEON) sites (Weinstein, 2021). The DeepForest framework (Weinstein, 2019) has established itself as a benchmark in forestry remote sensing applications by employing RetinaNet (Lin, 2018) and a Feature Pyramid Network (FPN) (Lin, 2017) as backbones for multi-scale tree detection. Modern object detection frameworks have evolved from region-based approaches like Fast R-CNN (Girshick, 2015) and Faster R-CNN (Ren, 2015) to single-stage detectors like RetinaNet (Lin, 2018), or end-to-end solutions like DETR (Carion, 2020), demonstrating superior performance compared to traditional image processing techniques.

More recently, foundation models such as Segment Anything (SAM) offer prompt-based segmentation capabilities that generalize well across various image types (Kirillov, 2023). Other foundation models like the DINO family (Oquab, 2024) have also demonstrated robust visual feature learning without supervision, allowing effective transfer learning to more specialized domains. These models can be fine-tuned with relatively small domain-specific datasets while maintaining their strong generalization capabilities. An example is the PseCo framework, that introduces a novel approach combining point-based detection with segmentation capabilities of foundation models, demonstrating superior performance in object counting tasks (Huang, 2023).

Our work builds upon these advances by adapting the PseCo framework specifically for tree crown segmentation, leveraging the capabilities of SAM to generate accurate instance masks from point-based detections, and demonstrating improved performance in complex forest environments compared to established methods such as DeepForest.

3. Datasets

We focus on two complementary datasets to develop and evaluate our tree crown segmentation methodology. The first dataset, NEON, provides extensive coverage across diverse forest ecosystems in the United States for model development, while the second offers a focused test case in a forested Alpine environment. Both datasets provide comparable inputs, with aerial VHR RGB imagery in similar conditions and resolutions.

NEON Dataset. In our study, we first utilize data from the National Ecological Observatory Network (NEON), a continental-scale ecological observation facility providing open

data on ecosystem dynamics across the United States (Weinstein, 2021). For the heatmap model training, we used high-resolution RGB aerial imagery (0.15 m/pixel) acquired by NEON's Airborne Observation Platform (AOP) across 22 sites covering diverse forest types. These images are orthorectified and capture 1 km² areas of forest landscapes collected during 2018-2019. The dataset contains more than 30,000 annotated tree crowns with bounding box delineations. Each NEON site represents different forest ecosystems, from temperate deciduous forests of the eastern United States to coniferous forests of the Pacific Northwest, providing essential diversity for generalizable model training. The sites span various ecological domains, elevations, and forest densities, allowing our model to learn robust representations across these variations.

Concerning heatmaps, we made instead use of the corresponding Canopy Height Models (CHMs) derived from co-registered LiDAR data (NEON ID: DP3.30015.001), which provided height information at 1-meter resolution. Similar to DeepForest, we extract estimated tree centers from these LiDAR acquisitions to use as noisy targets for our training.

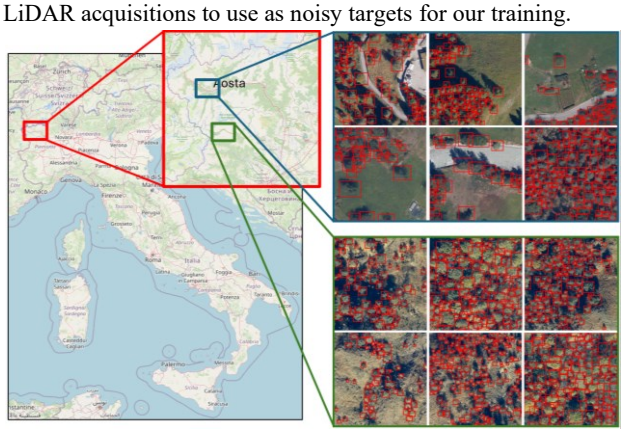


Figure 2. Localization and visual examples of our custom VdA dataset, providing VHR aerial images acquired in Valle d'Aosta, Italy.

VdA Dataset. For our case study implementation, we used a custom dataset of aerial imagery from the Valle d'Aosta region (VdA) in Northwestern Italy (Figure 2). This dataset consists of 22 large-scale RGB orthorectified cm/pixel resolution, each 50.000×50.000 pixels large. Among these areas, we selected 80 high-resolution crops, capturing forest landscapes in a mountainous alpine environment. Each 600×600 pixel image patch was manually annotated with bounding boxes for individual trees, creating a regional and specific ground truth dataset.

These images were collected as part of the NODES (*Nord Ovest Digitale e Sostenibile*) project, with particular emphasis on forest monitoring and carbon cycle assessment. The Valle d'Aosta region represents a challenging environment for tree crown detection due to its varied topography, mixed forest compositions, and altitudinal gradients. Testing our model on this independent regional dataset allowed us to evaluate the transferability of our approach beyond the NEON training sites to European forest systems, which represent different ecological conditions and environments.

4. Methodology

4.1 Problem Statement

Tree crown segmentation involves the delineation of the spatial extent of individual tree crowns from remote sensing imagery. Formally, given an aerial RGB image $I \in \mathbb{R}^{(H\times W\times 3)}$, the task is to predict a set of instance masks $M = \{M_1, M_2, ..., M_n\}$ where each $M_i \in \{0,1\}^{(H\times W)}$ represents the binary segmentation mask of a single tree crown. This task can be effectively decomposed into three distinct phases: (i) tree center estimation, (ii) center-based mask and box generation, (iii) boundary refinement. The first step identifies the approximate centers of individual trees within the image through a point detection task that outputs a probability heatmap $H \in [0,1]^{(H\times W)}$, where peaks correspond to likely tree center locations.

Adopting the standard SAM decoder, the second step uses the detected center points as prompts to generate initial instance segmentation masks for each tree, producing both pixel-wise masks and corresponding axis-aligned bounding boxes $B = \{B_1, B_2, ..., B_n\}$ where each $B_i = (x_1, y_1, x_2, y_2)$ represents the coordinates of the top-left and bottom-right corners of the bounding box.

The last step refines instead the initial mask predictions to better capture the true boundaries of each tree crown, particularly in challenging scenarios with overlapping canopies, through classification and non-maximum suppression techniques that eliminate redundant or low-confidence detections, and box regression to adjust the position of the tree.

This multi-phase approach allows for more robust delineation of individual trees compared to direct methods, especially in different domains and forest environments.

4.2 Segment Anything

SAM (Kirillov, 2023) represents a paradigm shift in computer vision as the first foundation model for image segmentation capable of generalizing across diverse domains. It enables users to specify segmentation targets through various input prompts, including points and bounding boxes. Its architecture comprises three principal components: a Vision Transformer-based (Dosovitskiy, 2021) image encoder that processes input images to create rich feature maps; a prompt encoder that transforms different types of user inputs into standardized representations; and a lightweight mask decoder that integrates these representations to generate final segmentation masks. A notable feature of SAM's architecture is its ability to produce multiple potential segmentations for ambiguous prompts, each representing a different interpretation. Furthermore, SAM's design allows the computationally intensive image encoding to be performed just once per image, regardless of the number of subsequent prompts, significantly enhancing efficiency in interactive scenarios. The development of SAM was made possible by an innovative data collection methodology that progressively leveraged the model's own capabilities—evolving from model-assisted manual annotations to semi-automatic labelling, and ultimately to fully automatic mask generation. This iterative approach yielded a large-scale dataset comprising over one billion masks across 11 million diverse images (SA-1B), enabling the model to generalize effectively to unseen domains without fine-tuning.

4.3 PseCo

PseCo is a generalized framework for both few-shot and zero-shot object counting and detection. The model leverages the complementary strengths of two foundation models: SAM for segmentation capabilities and Contrastive Language-Image Pre-Training (CLIP) (Radford, 2021) for classification. The framework follows three key steps: (1) a class-agnostic object localization that provides accurate point prompts for SAM, reducing computation costs while ensuring small objects aren't missed; (2) SAM-based segmentation to generate mask proposals from these points; and (3) a generalized object classification using CLIP embeddings to identify target objects.

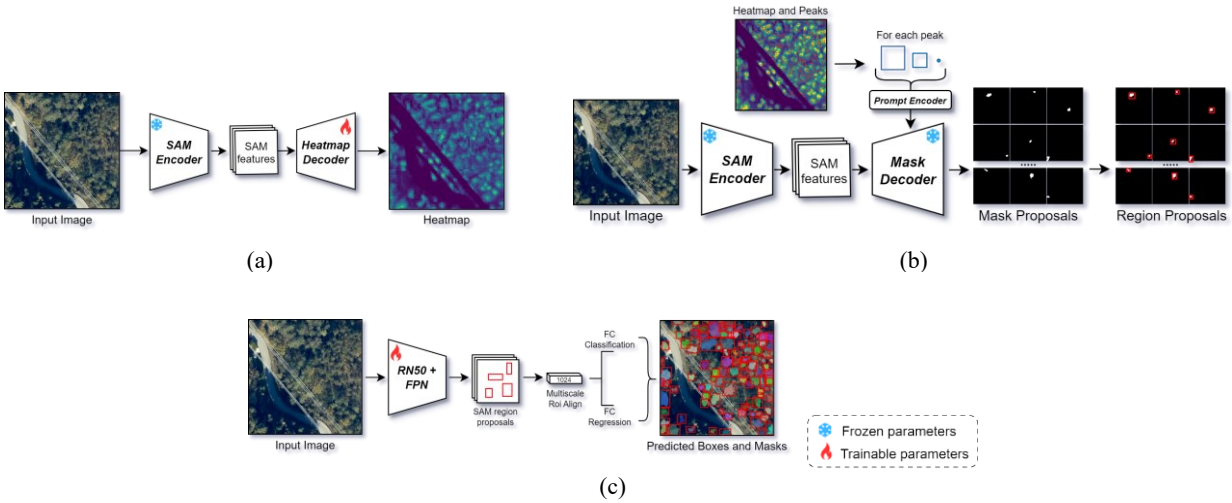


Figure 3. Overview of the TreePseCo framework: heatmap and peak generation to estimate tree centers (a), mask and box generation using SAM (b), classification and box refinement stage using a RCNN-like box regression and classification (c).

PseCo demonstrates state-of-the-art performance across multiple benchmarks including FSC-147 (Ranjan, 2021), COCO (Lin, 2015), and LVIS (Gupta, 2019) datasets. Unlike density-based counting methods, PseCo offers interpretable detection results while maintaining competitive accuracy, even in crowded scenes with small objects that traditional methods struggle to detect. Our architecture's initial stages are inspired by steps 1 and 2 of PseCo, however they have been specifically finetuned to localize tree centers rather than agnostic objects.

4.4 TreePseCo

Our framework consists of two primary components: a modified SAM model for generating bounding box proposals (referred to as point decoder) and a classification network for refinement and final detection. Following the approach established in PseCo, the finetuned mask decoder of SAM generates a heatmap that highlights the presence probability of a tree across the input image. We identify peak values within this heatmap and leverage such locations as prompt points for a standard SAM model, reusing the already computed features. These prompts generate multiple mask proposals per detected peak, from which we extract circumscribing boxes. A modified Faster R-CNN (Ren, 2015) architecture then processes these proposals through a ResNet50-FPN backbone, performing classification and bounding box regression to produce the final tree detections with refined localization.

Heatmap generation and peak detection. As illustrated in Figure 3 (a), an input satellite image is processed through a frozen SAM encoder, which extracts rich visual features. These are fed to a trainable heatmap decoder that has been specifically fine-tuned for tree detection tasks. The decoder produces a heatmap where brighter regions indicate higher probability of tree presence. This design leverages SAM's powerful feature extraction capabilities while allowing task-specific adaptation through the decoder.

Mask and box proposal generation. Leveraging the already computed features, we prompt SAM's original mask decoder to generate proposals for each detected heatmap peak, as shown in Figure 3 (b), without further training. Specifically, each peak location is used as a prompt point, along with a small and a larger surrounding box, resulting in six distinct masks. Using multiple prompts per peak increases robustness, providing additional mask proposals. From each of them, we extract the circumscribing box to create proposals for the subsequent classification stage.

Classification and box refinement. As depicted in Figure 3 (c), our classification component follows a modified Faster R-CNN architecture. Unlike traditional implementations that rely on a Region Proposal Network, our approach leverages the SAM-generated proposals directly, which provides better initial localization for tree instances. The ResNet backbone processes the input image once to extract feature maps, which are then enhanced through a Feature Pyramid Network (FPN) to create multi-scale representations. This multi-scale approach is particularly important for tree detection in aerial imagery, where trees can appear at various sizes depending on species and age. For each proposal, a multiscale RoI Align (He, 2018) operation extracts fixed-size feature map from the appropriate FPN level. Subsequent fully connected heads perform two primary tasks: assigns a classification score indicating tree confidence and generates four bounding box offset values to optimize alignment with ground truth boxes. This refinement stage is crucial for improving the precision of the initial SAM-based proposals. The final output, shown on the right side of Figure 3 (c), consists of predicted boxes and associated SAM's masks.

5. Experiments

5.1 Implementation Details

Considering the point decoder, we implement a two-stage training approach, beginning with pretraining on the NEON dataset. Because of our focus on tree centers for prompting, we did not employ the CHM directly as ground truth map. Given its noisy appearance, this could result in suboptimal model convergence due to the complexity of the raw data. To address this limitation, we adopt a simplified heatmap representation by modelling each tree as a 2D Gaussian distribution. Tree locations within the CHM were identified using the *PyCrown* library[1], which includes raster smoothing, peak detection through 2×2 max pooling, and a minimum height threshold to eliminate peaks below 5 meters. From the original 10,000×10,000-pixel NEON images, we extract random 600×600-pixel crops to maintain consistent resolution with the downstream datasets. At each detected peak location, we

[1] https://github.com/manaakiwhenua/pycrown

generate a 2D Gaussian distribution with fixed extent σ=5, creating an easier target to learn from. We utilize the SAM ViT-H encoder as our foundation model, freezing it during the training phase to preserve its robust representation capabilities while exclusively updating the mask decoder.

Second, we perform two finetuning passes on the NEON dataset annotations and on our VdA annotations respectively. During finetuning, ground truth heatmaps are generated from bounding box annotations by placing a comparable 2D Gaussian (i.e., σ=5) at each box center, maintaining consistency with our pretraining. Throughout the phases, we monitor the component independently from the subsequent refinement steps: we assume to have an *oracle*, a perfect classifier able to maximize our downstream metrics. In practice, this classifier assigned a perfect score of 1.0 to the proposal with the highest IoU (if greater than 0.5) for each ground truth box, allowing us to quantify the theoretical maximum mean Average Precision (mAP) achievable with our proposal generation mechanism. The extraction of the heatmap peaks is instead controlled by three main hyperparameters: smoothing window size, pooling size, and minimum height threshold. Their tuning required careful balance between two competing objectives: (1) generating sufficient point proposals to achieve a high theoretical maximum mAP, and (2) limiting the number of proposals to simplify the classification task. Considering the last refining step, we train the classification head separately for each dataset, initialized from a pretrained RN50 backbone. Following previous works, we maintain the default Faster R-CNN parameters for this phase.

5.2 Baseline

We adopt the DeepForest framework as strong baseline, being the *de facto* standard for individual tree crown localization. The model employs RetinaNet with a ResNet50 backbone and Feature Pyramid Network to generate bounding boxes around individual trees. A key advantage of DeepForest is its pretrained model, which has been trained on over 30 million algorithmically generated crowns from 22 National Ecological Observatory Network (NEON) sites and fine-tuned using 10,000 hand-labeled crowns. This extensive pretraining allows DeepForest to perform well across diverse forest types without additional training, achieving an average recall of 72% and precision of 64% across NEON evaluation sites.

The DeepForest library also supports transfer learning out of the box, enabling users to fine-tune the pretrained model with relatively small amounts of local data (approximately 1,000 annotations) to improve performance for specific forest types.

Despite its ease of use, DeepForest still presents some limitations. Performance degrades in densely forested areas with high crown overlap and in forest types significantly different from the training data. Additionally, the framework struggles with high recall, where closely growing trees are detected as a single entity, particularly in RGB data where height information is unavailable. These limitations create an opportunity for improved approaches that can better handle dense canopy environments and complex forest structures while maintaining the accessibility and scalability of RGB-based detection.

5.3 Results

We test the model on the NEON test set and VdA test annotations comparing DeepForest (DF), DeepForest finetuned (DF FT) and TreePseCo (TP) respectively. We evaluate our models using object detection metrics, specifically mean average precision under different Intersection over Union (IoU) thresholds ($mAP^{@IoU}$), different box sizes (small, medium, large), and mean average recall given 100 instances ($mAR^{@100}$), or small and medium objects.

	NEON		VdA		
Metric	**DF**	**TP**	**DF**	**DF (FT)**	**TP**
mAP	**18,04**	15,43	6,54	9,10	**14,76**
$mAP^{@IoU=50}$	**49,89**	41,68	19,19	26,37	**37,55**
$mAP^{@IoU=75}$	**7,92**	7,07	3,06	5,57	**8,94**
mAP^{small}	**12,08**	3,21	3,56	4,47	**5,42**
mAP^{medium}	**23,67**	16,09	12,07	**17,16**	15,66
mAP^{large}	**32,00**	25,48	0,00	0,00	**23,35**
$mAR^{@100}$	**25,75**	24,50	9,49	14,89	**21,52**
mAR^{small}	**19,08**	6,48	4,13	7,05	**11,34**
mAR^{medium}	**32,14**	26,20	18,78	**28,13**	22,09

Table 1. Overall results in terms of mAP and mAR on the NEON and VdA datasets, comparing DeepForest (DF) and TreePseCo (TP).

Dataset	Extraction	Refinement	mAP
NEON	✓	✓	27.09
NEON	trained	✓	22.03
NEON	trained	trained	15.43
VdA	trained	✓	17.06
VdA	trained	trained	14.76

Table 2. Hyperparameter study results on the point extraction phase.

Loss	Bilinear Ups.	Transposed C.
SSIM	13.48	19.00
MSE	12.91	12.09

Table 3. Comparison between different losses (SSIM, MSE) and upsampling techniques (bilinear, transposed).

As shown in Table 1, on the NEON dataset TreePseCo presents similar performance to DeepForest, which however outperforms our solution numerically. The latter achieves in fact a mAP of 18.04, while our TreePseCo model attains 15.43. This is observable in every metric, such as $mAP^{@IoU=50}$, where DeepForest reaches 49.89 while TreePseco sits at 41.68. However, we note that DeepForest was specifically optimized on the NEON dataset, potentially benefiting from dataset familiarity. Furthermore, our comparison methodology introduces an inherent disadvantage for TreePseCo, as we are forced to evaluate a model designed to generate precise instance masks using bounding box metrics, which naturally favors DeepForest's native detection approach.

In fact, on the VdA dataset, our approach often outperforms both the original DeepForest (6.54 mAP) and its finetuned version (9.10 mAP), reaching a mAP of 14.76. This performance gap widens at $mAP^{@IoU=50}$, where our solution achieves 37.55, compared to 19.19 and 26.37 for the original and finetuned DeepForest models, respectively. Notably, TreePseCo excels in detecting trees across all size categories in the VdA dataset, particularly for large trees (23.35 mAP) where both DeepForest variants failed to detect any instances (0.00 mAP). These results suggest that while DeepForest remains a valid solution on data like its training distribution, TreePseCo demonstrates superior generalization capabilities when applied to new geographical contexts.

Figure 4 illustrates the qualitative improvements of our TreePseCo approach over the baseline. In the top-row example, the baseline model identifies numerous trees while overlooking

smaller crowns, whereas TreePseCo provides more comprehensive detection coverage. The bottom row showcases challenging terrain where DeepForest mainly captures larger, isolated trees but struggles with heterogeneous environments, while TreePseCo demonstrates superior performance in these complex conditions. Ground truth comparisons confirm that while both models occasionally miss smaller instances, TreePseCo substantially reduces false negatives, particularly in complex canopy structures.

5.4 Hyperparameter studies

We further conduct a comprehensive set of experiments on critical architectural components and training parameters to optimize our framework. With the aim of obtaining smoother outputs, we first evaluate upsampling techniques in the SAM heatmap decoder, comparing bilinear interpolation against transposed convolution. We then examine the efficacy of Mean Squared Error (MSE) versus Structural Similarity Index Measure (SSIM) as loss function (Wang, 2004). We restrict our SSIM loss evaluations to the fine-tuning phase exclusively, driven by the observation that pretraining data contained imprecise tree locations derived from CHM, making SSIM's strict structural matching properties counterproductive at this stage.

Concerning upsampling techniques, we find that the latter consistently produced more defined peaks in the heatmap, albeit with some additional noise patterns. This sharper peak definition translates to improved detection performance in our quantitative evaluation. Considering loss functions, SSIM loss demonstrates superior performance by generating more precisely defined maps in the heatmap, leading to more accurate prompt point generation, as visible in Table 3. Given the versatility of the SAM decoder, we also explore several prompt formats, including direct point coordinates, rectangular boxes with negative corner points, and plain bounding boxes. Given the negligible performance differences observed between these approaches, we select the plain box representation for its simplicity and computational efficiency.

In order to optimize the tree peaks extraction phase, we conduct a hyperparameter search focusing on heatmap generation. During this phase, we substitute the refinement component with an oracle classifier to establish theoretical performance bounds. When using ground truth centers as prompts, we obtain a theoretical upper bound on the NEON dataset at 27.09 mAP, as shown in Table 2. Through these experiments, we identify optimal hyperparameter conditions for peaks extraction: window size detection = 2, window size smoothing = 0, and height minimum threshold = 0.3. With these parameters, the maximum theoretical mAP achievable was 22.03 on NEON and 17.06 on VdA (both assuming a perfect classifier). We then integrated the actual refinement module using these optimal parameters, resulting in the final scores displayed in both Table 3 and Table 1. It's worth noting that while retrieving numerous peaks from the heatmap increases the theoretical maximum performance, an excessive number of proposals adversely affects the classification network's discriminative capability, resulting in diminished overall performance.

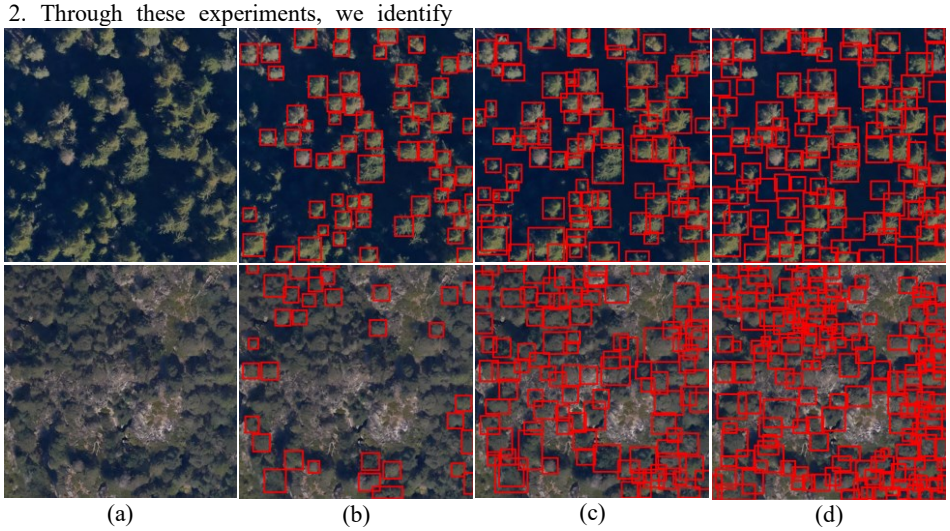
6. Conclusions

This paper introduces TreePseCo, an adaptation of the PseCo framework for individual tree crown segmentation in aerial imagery. By leveraging the SAM foundation model in a three-stage pipeline, our approach demonstrates improved generalization capabilities, especially when tested on geographically diverse datasets, including our custom VdA set. While state of the art approaches such as DeepForest remain a viable alternative, our approach shows superior performance on new environments without extensive retraining. TreePseCo performs particularly well in two challenging scenarios: densely clustered tree formations and detection of smaller tree instances.

Despite these advantages, several limitations should be acknowledged. First, TreePseCo introduces a higher computational overhead compared to frameworks like DeepForest, potentially limiting its deployment in resource-constrained scenarios where processing speed is crucial. Second, the custom VdA dataset remains limited in scope and extension. Future work should focus on these key aspects, providing further validation across more diverse forest types and geographic regions, developing an end-to-end trainable architecture, or reducing model size through distillation or alternative foundation models such as DINO (Oquab, 2024).

Acknowledgements

This work was carried out in the context of the projects FUTUREFOR (HEU GA n. 101180278), and PERFORM, cascade project of NODES through the MUR - M4C2 1.5 of PNRR under Grant ECS00000036.



(a) (b) (c) (d)

Figure 4. Qualitative results : input image (a), DeepForest output (b), TreePseCo outputs (c), and manual ground truth (d).

References

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S., 2020. End-to-End Object Detection with Transformers. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.M. (eds), Computer Vision – ECCV 2020, Vol. 12346, Springer International Publishing, Cham, 213–229. doi.org/10.1007/978-3-030-58452-8_13.

Dalponte, M., Frizzera, L., Gianelle, D., 2019. Individual Tree Crown Delineation and Tree Species Classification with Hyperspectral and LiDAR Data. PeerJ, 6, e6227. doi.org/10.7717/peerj.6227.

Davies, S.J., Abiem, I., Salim, K.A., Aguilar, S., Allen, D., Alonso, A., Anderson-Teixeira, K., Andrade, A., Arellano, G., Ashton, P.S., et al., 2021. ForestGEO: Understanding Forest Diversity and Dynamics through a Global Observatory Network. Biological Conservation, 253, 108907. doi.org/10.1016/j.biocon.2020.108907.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations.

Girshick, R., 2015. Fast R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 1440–1448.
Gupta, A., Dollar, P., Girshick, R., 2019. Lvis: A Dataset for Large Vocabulary Instance Segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 5356–5364.

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.

He, Kaiming, et al. 2015. Mask R-CNN. Proceedings of the IEEE International Conference on Computer Vision, 2961-2969.

Huang, H., Li, X., Chen, C., 2018. Individual Tree Crown Detection and Delineation from Very-High-Resolution UAV Images Based on Bias Field and Marker-Controlled Watershed Segmentation Algorithms. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 11(7), 2253–2262. doi.org/10.1109/JSTARS.2018.2830410.

Huang, Z., Dai, M., Zhang, Y., Zhang, J., Shan, H., 2024. Point Segment and Count: A Generalized Framework for Object Counting. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 17067–17076.

Hudak, A.T., Strand, E.K., Vierling, L.A., Byrne, J.C., Eitel, J.U.H., Martinuzzi, S., Falkowski, M.J., 2012. Quantifying Aboveground Forest Carbon Pools and Fluxes from Repeat LiDAR Surveys. Remote Sensing of Environment, 123, 25–40. doi.org/10.1016/j.rse.2012.02.023.

Hyyppa, J., Hyyppä, H., Leckie, D., Gougeon, F., Yu, X., Maltamo, M., 2008. Review of Methods of Small-footprint Airborne Laser Scanning for Extracting Forest Inventory Data in Boreal Forests. International Journal of Remote Sensing, 29(5), 1339–1366. doi.org/10.1080/01431160701736489.

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

Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S., 2017. Feature Pyramid Networks for Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2117–2125.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2017. Focal Loss for Dense Object Detection. Proceedings of the IEEE International Conference on Computer Vision, 2980–2988.

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. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds), Computer Vision – ECCV 2014, Vol. 8693, Springer International Publishing, Cham, 740–755. doi.org/10.1007/978-3-319-10602-1_48.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al., 2023. DINOv2: Learning Robust Visual Features without Supervision. Transactions on Machine Learning Research.

Puliti, S., Solberg, S., Granhus, A., 2019. Use of UAV Photogrammetric Data for Estimation of Biophysical Properties in Forest Stands Under Regeneration. Remote Sensing, 11(3), 233. doi.org/10.3390/rs11030233.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 2021. Learning Transferable Visual Models from Natural Language Supervision. International Conference on Machine Learning, 8748–8763.

Ranjan, V., Sharma, U., Nguyen, T., Hoai, M., 2021. Learning to Count Everything. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3394–3403.

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.

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. IEEE Transactions on Image Processing, 13(4), 600–612. doi.org/10.1109/TIP.2003.819861.

Wang, Z., Simoncelli, E.P., Bovik, A.C., 2003. Multiscale Structural Similarity for Image Quality Assessment. The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers, 2003, Vol. 2, 1398-1402. doi.org/10.1109/ACSSC.2003.1292216.

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. doi.org/10.3390/rs11111309.

Weinstein, B.G., Marconi, S., Bohlman, S.A., Zare, A., Singh, A., Graves, S.J., White, E.P., 2021. A Remote Sensing Derived Data Set of 100 Million Individual Tree Crowns for the

National Ecological Observatory Network. eLife, 10, e62922. doi.org/10.7554/eLife.62922.

Wulder, M.A., White, J.C., Nelson, R.F., Næsset, E., Ørka, H.O., Coops, N.C., Hilker, T., Bater, C.W., Gobakken, T., 2012. Lidar Sampling for Large-Area Forest Characterization: A Review. Remote Sensing of Environment, 121, 196–209. doi.org/10.1016/j.rse.2012.02.001.

Zhou, Y., Wang, L., Jiang, K., Xue, L., An, F., Chen, B., Yun, T., 2020. Individual Tree Crown Segmentation Based on Aerial Image Using Superpixel and Topological Features. Journal of Applied Remote Sensing, 14(2), 022210. doi.org/10.1117/1.JRS.14.022210.

Zörner, J., Dymond, J.R., Shepherd, J.D., Wiser, S.K., Jolly, B., 2018. LiDAR-Based Regional Inventory of Tall Trees—Wellington, New Zealand. Forests, 9(11), 702. doi.org/10.3390/f9110702.