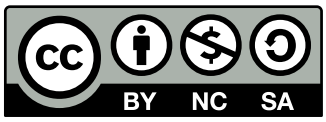
Aalto University
School of Science

Master's Programme in Computer, Communication and Information Sciences

SAM2 pseudolabeling for instance segmentation

Stefan Rua

Master's Thesis
2025



Aalto University
School of Science

Author Stefan Rua

Title SAM2 pseudolabeling for instance segmentation

Degree programme Computer, Communication and Information Sciences

Major Macadamia

Supervisor Jorma Laaksonen

Advisor Julius Pesonen (MSc)

Collaborative partner Finnish Geospatial Research Institute FGI

Date 30 September 2025 **Number of pages** 13 **Language** English

Abstract
Lorem ipsum etc.

Keywords pseudolabeling, instance segmentation, forestry

Aalto-yliopisto
Perustieteiden
korkeakoulu

Tekijä Stefan Rua

Työn nimi SAM2 pseudolabelöinti instanssisegmentaatiokoulutuksessa

Koulutusohjelma Computer, Communication and Information Sciences

Pääaine Macadamia

Työn valvoja Jorma Laaksonen

Työn ohjaaja DI Julius Pesonen

Yhteistyötaho Paikkatietokeskus FGI

Päivämäärä 30.9.2025 **Sivumäärä** 13 **Kieli** englanti

Tiivistelmä
Lorem ipsum jne.

Avainsanat pseudolabelöinti, instanssisegmentaatio, metsäily

Contents

1 Introduction

Airborne remote sensing based tree mapping methods are used for forest health monitoring [1] and city planning [2] due to their efficiency compared to manual methods, but the training process is often bottlenecked by the need for high quality manual annotations. The goal of this study is to test if training results can be improved in instance segmentation tasks by refining coarse segments calculated from a canopy height model (CHM) using Segment Anything Model 2 (SAM2) [3].

This document begins by explaining the necessary background information in the Backgroud section, and exploring the previous works relevant to this study in the Related Work section. The Materials and Methods section contains a detailed explanation of the dataset, models, and training hyperparameters. Finally, the experiment's outcome is displayed and analyzed in the Results and Discussion sections, along with suggestions for future improvement.

2 Background

2.1 Instance Segmentation

Instance segmentation is a machine learning task where the goal is to generate separate masks for individual objects in an image. This differs from semantic segmentation, where each pixel is given a semantic label without separating individual objects, and from object detection where individual objects are given bounding boxes but no masks.

[semantic segmentation vs instance segmentation vs object detection image]

2.2 Segment Anything Model 2

Segment Anything Model 2 (SAM2) is a foundation model that can segment objects very generally, even ones that it hasn't been trained on. There are two ways to use it: only providing it with an image and letting it segment every object on its own, or providing it with an image and a prompt containing the location of the object of interest. The prompts can be of two types: a set of one or more foreground points and optional background points, or a bounding box. The model outputs a binary mask.

[sam bbox to mask example]

SAM2 is a revised version of Segment Anything Model (SAM) [ref], with the main addition being the ability to train and predict on video data. While no video data was used in this thesis, SAM2 was chosen for its higher accuracy and 6x speedup over SAM.

2.3 Mask R-CNN

Mask R-CNN is an instance segmentation model based on a region-based convolutional neural network (R-CNN). It's trained using image-target pairs, where the target contains a binary mask, bounding box, and category label for each object of interest. After training, the model can predict masks in new images for the types of objects seen in training.

A CNN consists of convolutional layers, where the connections to the next layer are formed by sliding a kernel with tunable weights over the previous layer. R-CNNs search over feature maps produced by a CNN called the region proposal network (RPN) to find objects in images.

[cnn image]

2.4 Tree Mapping

Tree mapping is used by the forestry industry to monitor forest health, by urban planners to plan cities, by scientists to model ecological change and assess post-disaster damage.

Traditionally tree mapping has been done by ground-level visual analysis [?], but the advent of aerial photography, laser scanning, and computational analysis has moved this task towards remote sensing.

2.5 Aerial Imaging

Aerial images may be taken from any airborne apparatus, for example drones, helicopters, aeroplanes, or even satellites. For the purpose of mapping individual trees satellite imagery rarely provides a high enough ground sample distance (GSD), and aeroplanes aren't convenient for covering square-ish areas, so drones and helicopters are most often used.

[image of drone with camera]

Aerial remote sensing sensors include laser scanners and multispectral cameras in addition to traditional RGB-cameras that only capture visible wavelengths. Laser scanners produce three-dimensional point could data. Point clouds be used for volumetric segmentation or detection, or for calculating a digital surface model (DSM) or canopy height model (CHM). A DSM models the surface of the ground and protruding objects, and a CHM models the height the tree canopy (how are they different?).

[image of point cloud, image of dsm]

Multispectral cameras can capture additional wavelengths invisible to the human eye. Typically these include infrared (IR) at 1550-1750 nm and near-infrared (NIR) at 750-900 nm. Hyperspectral cameras capture a continuous spectrum instead of discrete channels.

[hyperspectral graph]

3 Related Work

The previous version of SAM2, SAM has been assessed for tree crown instance segmentation on drone imagery [4]. The study uses SAM in the following ways: generating masks without any prompts, generating masks with digital surface model (DSM) maxima as point prompts, and prompting SAM with predictions from a trained Mask R-CNN model. Of these, the last approach is closest to the one examined here. The resulting mean intersection over union (mIoU) scores the single-class detection class were as follows: SAM + no prompts: 35.06%, SAM + DSM prompts: 46.15%, Mask R-CNN + SAM: 78.27%.

Another study examining SAM for remote sensing use proposes RSPrompter [5], a method that learns to generate prompts from the SAM's image encoder, then feeding them to the decoder. The study proposing the method reports the results as AP scores, but the study mentioned above tested RSPrompter as well, achieving 82.58% mIoU. [4]

FMARS [6] is a dataset with annotations generated using GroundingDINO and SAM. GroundingDINO was used to convert text prompts to bounding boxes, and SAM for converting these boxes to segments. A subset of the annotated area was manually annotated, and the generated annotations were compared to the manual ones, resulting in a mIoU score of 50.22%.

4 Materials and Methods

4.1 Datasets

The main dataset consists of a multispectral orthophoto taken by helicopter spanning approximately 2 km^2 and CHM-based coarse tree crown segments. The orthophoto covers both forest and urban area, has a GSD of 2.5 cm, and contains blue, green, red, near-infrared, mid-infrared, and thermal infrared bands.

As the CHM-based coarse tree crown segments had been calculated from point-cloud data from an earlier flight than the multispectral images, many segments no longer contained trees. To eliminate these erroneus segments, they were filtered based on their Normalized Difference Vegetation Index (NDVI) value:

$$\text{NDVI} = \frac{\text{NIR - Red}}{\text{NIR + Red}}.$$

For each segment, NDVI was calculated using the per-channel mean values. Segments with NDVI < 0.2 were removed. Figure 1 shows the distribution of NDVI values.

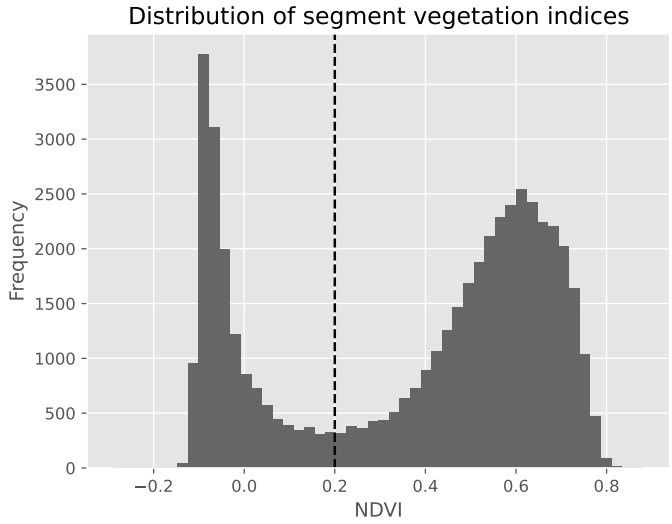


Figure 1: Distribution of NDVI values. The dashed vertical line indicates the threshold below which segments were discarded.

4.2 Models

SAM2 is used for the pseudolabeling task and Mask R-CNN [7] for the supervised instance segmentation.

4.3 Methods

First the coarse segments were pseudolabeled using SAM2, indexing the orthphoto in a grid with a window of size 1024 and a stride of 512. For each window, only the segments whose centroids were located in the central 512x512 square of the window were selected for pseudolabeling. Then, the bounding box of each segment was passed as a prompt to SAM2, and the largest connected component of the output mask was saved as the pseudolabel. These parameters for the window size and stride were selected in order to avoid stitching artifacts and to provide SAM2 with images matching the native input resolution of the model. An example of the pseudolabeling process is shown in Figure 2, and the distribution of tree radii is shown in Figure 3.

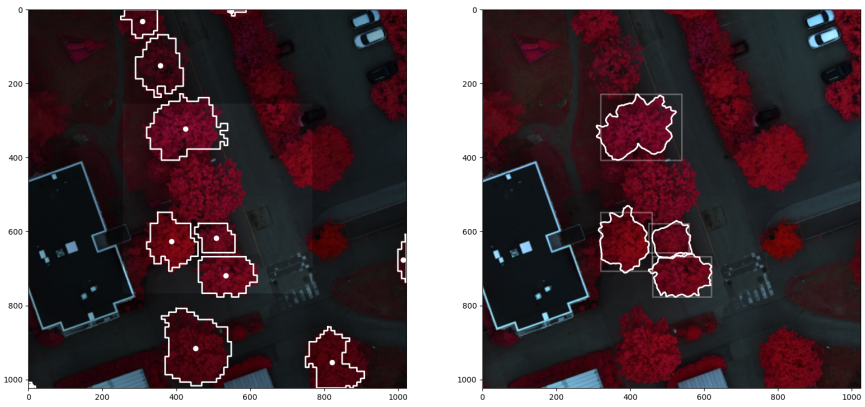


Figure 2: Example of the pseudolabeling process. On the left, CHM-based coarse segments are drawn around tree crowns, and their centroids are displayed as dots. The shaded area around the central 512x512 square is the buffer region. On the right, the box prompts provided to SAM are shown in reduced opacity, and the predicted mask outlines are drawn in white.

A small test area of 362 trees was segmented manually. To evaluate the quality of the coarse segments and quantify the effect of SAM2 pseudolabeling, the coarse segments and pseudolabels were compared to the manual segments using the Jaccard index [8]:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|},$$

where A and B are the segments to be compared. Then, a Mask R-CNN model with a ResNet-50 [9] backbone was trained separately on both the coarse segments and pseudolabels.

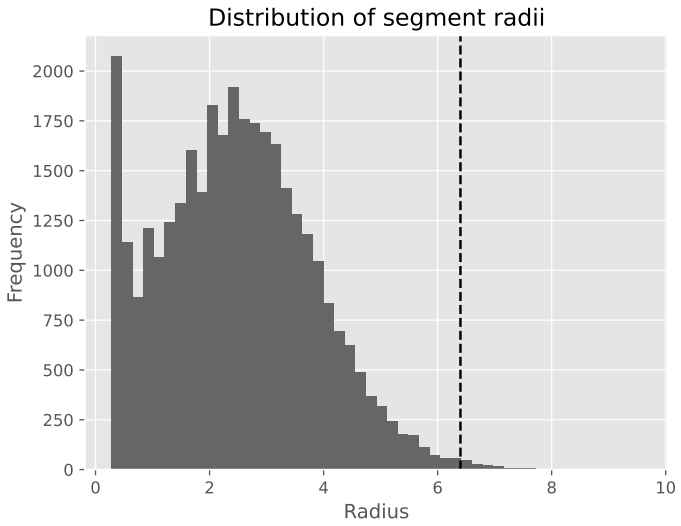


Figure 3: Distribution of tree crown radii. The dashed vertical line indicates the threshold above which segments may span from the central region of interest over the buffer region to the image border, producing stitching artifacts.

5 Results

(Results table with accuracy, recall, f1, and miou for each model here)

The highest scoring model was ???. The addition of the SAM2 pseudolabeling preprocessing step improved accuracy by ??%, recall by ??%, F1 by ??%, and mIoU by ??%.

6 Discussion

The segmentation performance improvement provided by pseudolabeling indicates that it is a worthwhile preprocessing step for datasets, where the original annotations are only approximate.

7 Conclusions

References

[1] S. Ecke, J. Dempewolf, J. Frey, A. Schwaller, E. Endres, H.-J. Klemmt, D. Tiede, and T. Seifert, "Uav-based forest health monitoring: A systematic review," *Remote Sensing*, vol. 14, no. 13, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/13/3205

[2] L. Velasquez-Camacho, A. Cardil, M. Mohan, M. Etxegarai, G. Anzaldi, and S. de Miguel, "Remotely sensed tree characterization in urban areas: A review," *Remote Sensing*, vol. 13, no. 23, 2021. [Online]. Available: https://www.mdpi.com/2072-4292/13/23/4889

[3] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," 2024. [Online]. Available: https://arxiv.org/abs/2408.00714

[4] M. Teng, A. Ouaknine, E. Laliberté, Y. Bengio, D. Rolnick, and H. Larochelle, "Assessing sam for tree crown instance segmentation from drone imagery," 2025. [Online]. Available: https://arxiv.org/abs/2503.20199

[5] K. Chen, C. Liu, H. Chen, H. Zhang, W. Li, Z. Zou, and Z. Shi, "Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," 2023. [Online]. Available: https://arxiv.org/abs/2306.16269

[6] E. Arnaudo, J. L. Vaschetti, L. Innocenti, L. Barco, D. Lisi, V. Fissore, and C. Rossi, "Fmars: Annotating remote sensing images for disaster management using foundation models," 2024. [Online]. Available: https://arxiv.org/abs/2405.20109

[7] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," 2018. [Online]. Available: https://arxiv.org/abs/1703.06870

[8] G. K. Gilbert, "Finley's tornado predictions." *American Meteorological Journal. A Monthly Review of Meteorology and Allied Branches of Study (1884-1896)*, vol. 1, no. 5, p. 166, 1884.

[9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015. [Online]. Available: https://arxiv.org/abs/1512.03385