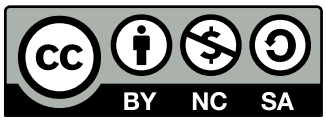
Aalto University
School of Science

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** 20 **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** 20 **Kieli** englanti

Tiivistelmä
Lorem ipsum jne.

Avainsanat pseudolabelöinti, instanssisegmentaatio, metsäily

Contents

Abbreviations

ANN	artificial neural network
CHM	canopy height model
CNN	convolutional neural network
DSM	digital surface model
GSD	ground sample distance
IR	infrared
IoU	intersection over union
MLP	multilayer perceptron
NDVI	normalized difference vegetation index
NIR	near-infrared
R-CNN	region-based convolutional neural network
RPN	region proposal network
SAM	Segment Anything Model
SAM2	Segment Anything Model 2
mIoU	mean intersection over union

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 thesis 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].

In recent years, deep learning based methods have been shown to produce increasingly useful models and predictions in forestry and agriculture [4], especially in the tasks of species detection and land cover prediction. Combined with the recent advances in autonomous data collection methods using unmanned aerial vehicles (UAVs) [5], there is unprecedented potential for massive data collection and mapping with relatively little manual labor. The efficient use of the available manual labor resources is especially important here in Finland where the population in small.

The ability to keep an up to date map of vegetation is essential for monitoring our forest ecosystems, which are increasingly important as global emissions rise and biodiversity decreases [6]. Finland has an especially great resposibility in protecting its nature, as an exceptionally large part of its surface area (75%) is covered by forest. In addition to protecting boreal forests, efficiently managing areas used for by the forest industry is necessary, as it accounts for a significant part of the Finnish economy [7]. Finnish forests face unprecedented risks from bark beetles, drought, and forest fires due to climate change [8].

1.1 Research Questions

1. How much does SAM2 pseudolabeling improve low quality instance segmentation annotations?

2. How much does SAM2 pseudolabeling improve training results in instance segmentaion tasks?

3. How well does SAM2 pseudolabeling perform compared to other zero-shot instance segmentation methods?

1.2 Structure of the Thesis

This thesis 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 Dataset 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

This section goes over the necessary background information for understanding this thesis. Subsections explain the the machine learning task of instance segmentation, methods and model architectures to perform said task, the goals and methods of aerial tree mapping, and differences in remote sensing aerial imaging apparatus.

2.1 Neural Networks

Artificial neural networks (ANNs) are mathematical models, usually written as computer programs, designed to resemble the biological brain by imitating its structure of interconnected neurons. The goal of such models is to be "trained" using example data containing input and target pairs, after which the trained model is able to predict targets for new input data. The training is made possible by the interconnected neuron structure, represented as a graph where each neuron is a node, and each edge between nodes holds a tunable weight and bias parameter. The oldest and simplest type of neural network is the multilayer perceptron (MLP), that consists of an input layer of nodes, an output layer, and fully connected hidden layers in-between (Figure 1). During training, the outputs of the model are compared to the ground truth targets and evaluated using a loss function. The gradients of the weights for each layer are calculated during the prediction phase called the forward pass, and updated based on the resulting loss during the backward pass. By repeating this backpropagation step over multiple iterations, the model's parameters gradually converge to predict the wanted targets for the given domain. [9]

The gradient calculation and backpropagation step utilize the chain rule of differentiation: if $h = f(g(x))$, the derivative of $h(x)$, $h'(x)$ is

$$h'(x) = f'(g(x))g'(x). \tag{1}$$

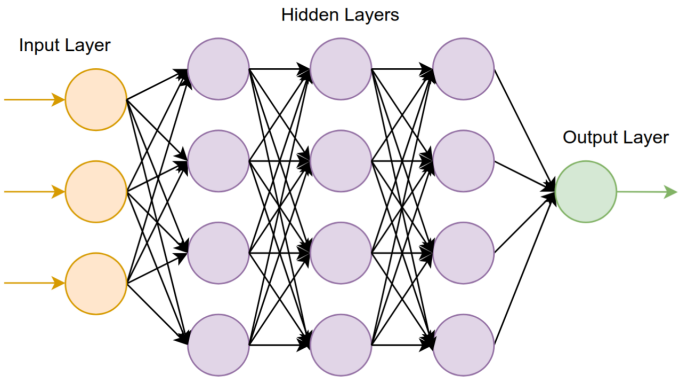


Figure 1: A multilayer perceptron visualized. [10]

2.1.1 Convolutional Nerual Networks

Convolutional neural networks (CNNs) are neural networks that employ convolutional layers. These are commonly used for computer vision tasks due to their efficiency in detecting hierarchies of visual features. A convolutional layer consists of a kernel that is moved over the input as a sliding window, computing an output for each position (Figure 2). The weights of the layer make up the kernel, and optimizing these weights lead to the kernel representing some feature related to the target. [11]

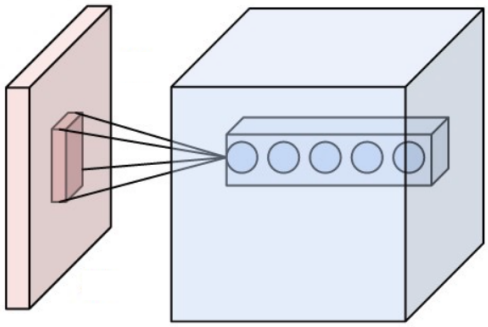


Figure 2: A convolutional layer visualized. [12]

2.1.2 Transformers

The transformer is a more recent neural network architecture that has shown great performance when trained on extremely large amounts of data. The network consists of an encoder that transforms the input into a set of tokens, that are then weighed using a self-attention layer, after which a decoder calculates the output based on the weighted tokens. The main advantage of transformers is the ability to leverage massively parallel compute efficiently, allowing the training of very large models in a reasonable amount of time. [13]

2.2 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. See Figure 4 for a visual comparison of the tasks.

2.2.1 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

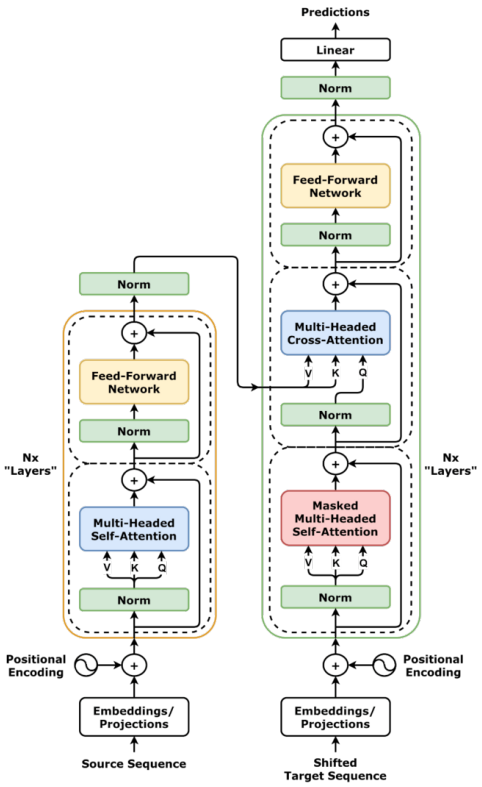
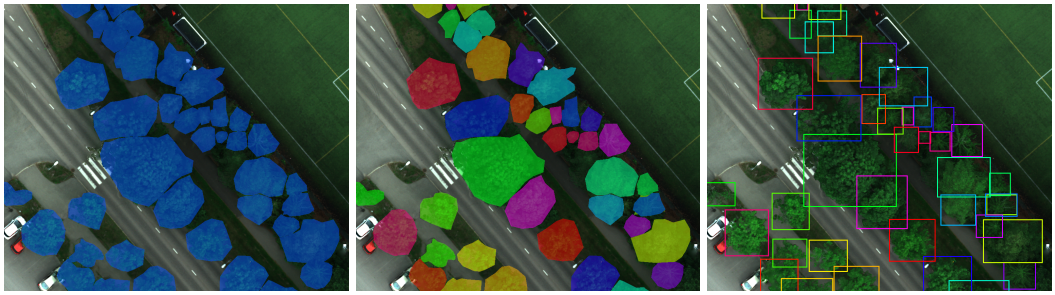


Figure 3: A transformer architecture visualized. Left: encoder, right: decoder. [14]



(a) Semantic segmentation. **(b)** Instance segmentation. **(c)** Object detection.

Figure 4: Comparison of similar computer vision tasks.

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]

2.2.2 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 (Figure 5). The model outputs a binary mask.

(a) Bounding box. **(b)** Foreground point. **(c)** Foreground and background points.

Figure 5: Comparison of SAM prompt types.

SAM2 is a revised version of Segment Anything Model (SAM) [15], 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 Aerial 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 [16], but the advent of aerial photography, laser scanning, and computational analysis has moved this task towards remote sensing.

2.3.1 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.

Aerial remote sensing sensors include laser scanners and multispectral cameras in addition to traditional RGB-cameras that only capture visible wavelengths. Laser

Figure 6: Drone with remote sensing apparatus.

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. In forestry and agriculture, this extended range of bandwidths can provide useful information for biomass and growth stage estimation and species detection. [17]

[hyperspectral graph]

3 Related Work

The previous version of SAM2, SAM has been assessed for tree crown instance segmentation on drone imagery [18]. 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 [19], 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. [18]

FMARS [20] 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 Dataset and Methods

This section covers the technical details of the dataset and methods used to train and evaluate the models.

4.1 Dataset

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:

$$\text{NDVI} = \frac{\text{NIR - Red}}{\text{NIR + Red}}. \tag{2}$$

For each segment, NDVI was calculated using the per-channel mean values. Segments with NDVI < 0.2 were removed. Figure 7 shows the distribution of NDVI values.

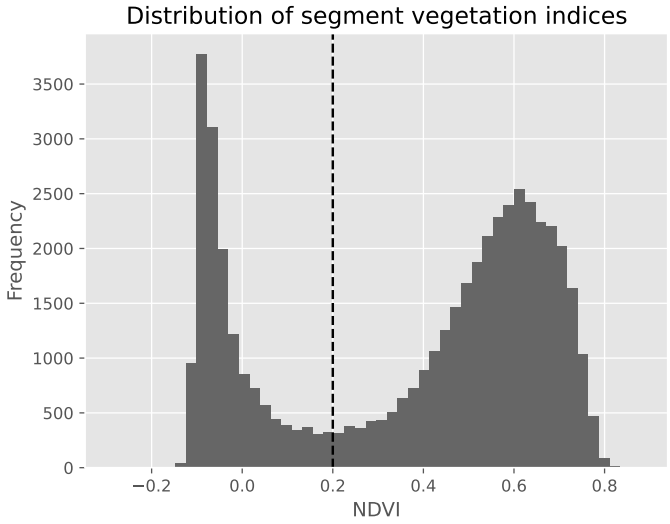


Figure 7: Distribution of NDVI values. The dashed vertical line indicates the threshold below which segments were discarded.

4.2 Models

SAM2 is used for the pseudolabeling task and Mask R-CNN [21] 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 8, and the distribution of tree radii is shown in Figure 9.

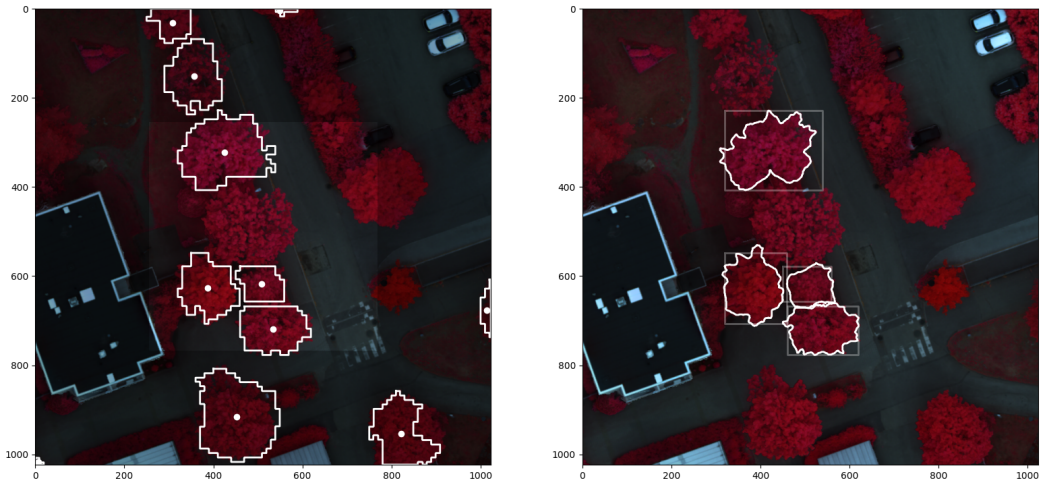


Figure 8: Example of the pseudolabeling process. On the left, CHM-based coarse segments are drawn around tree crowns, and their centroids are displayed as dots. The shaded area around the central 512x512 square is the buffer region. On the right, the box prompts provided to SAM are shown in reduced opacity, 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. Figure 10 shows the visual differences between the original coarse CHM-based segments, pseudolabels produced by SAM, and manually drawn segments.

Then, a Mask R-CNN model with a ResNet-50 [22] backbone was trained separately on both the coarse segments and pseudolabels.

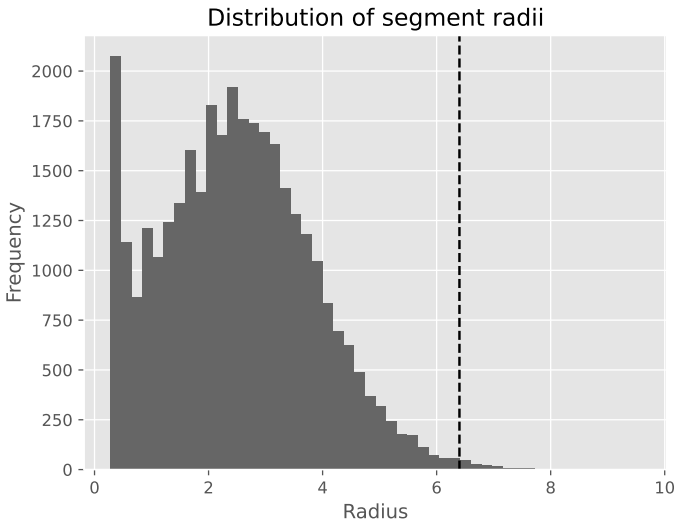
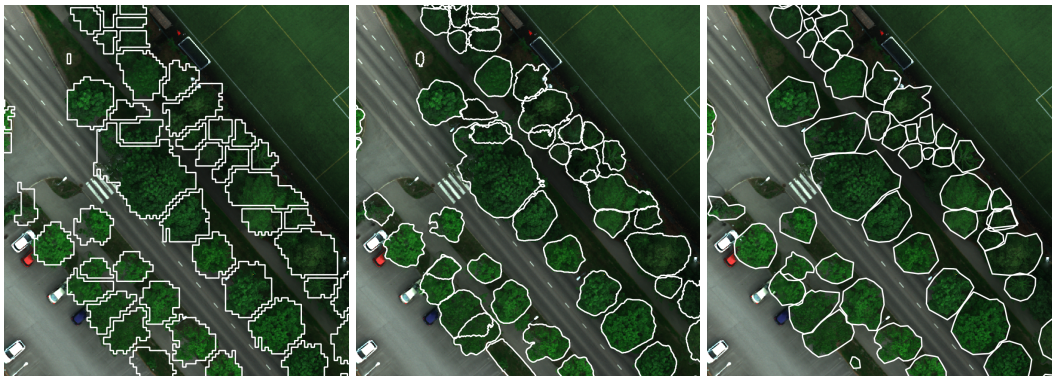


Figure 9: Distribution of tree crown radii. The dashed vertical line indicates the threshold above which segments may span from the central region of interest over the buffer region to the image border, producing stitching artifacts.



(a) Coarse, CHM-based. **(b)** SAM pseudolabels. **(c)** Manual.

Figure 10: Comparison of annotations.

4.4 Performance Measures

4.4.1 Jaccard Index

The Jaccard index[23] is a measure of similarity between two segments:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{3}$$

where A and B are the segments to be compared. This performance measure is also referred to as IoU (intersection over union), and utilized in the mIoU metric, the mean of IoU-scores over all target classes.

4.4.2 Precision and Recall

Precision and recall meaure the quality of a model's predictions. Precision is the fraction of detections that truly contained a target:

$$\text{Precision} = \frac{\text{Correct detections}}{\text{All detections}}, \tag{4}$$

and recall is the fraction of targets that were detected:

$$\text{Recall} = \frac{\text{Detected targets}}{\text{All targets}}. \tag{5}$$

4.4.3 F_1 score

The F_1 score is a performance measure that represents the overall performance of a model as the harmonic mean of precision and recall:

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}}. \tag{6}$$

5 Experiments and 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 ??%.

5.1 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.

6 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] T. Wang, Y. Zuo, T. Manda, D. Hwarari, and L. Yang, "Harnessing artificial intelligence, machine learning and deep learning for sustainable forestry management and conservation: Transformative potential and future perspectives," *Plants*, vol. 14, no. 7, 2025. [Online]. Available: https://www.mdpi.com/2223-7747/14/7/998

[5] A. Jaakkola, J. Hyyppä, X. Yu, A. Kukko, H. Kaartinen, X. Liang, H. Hyyppä, and Y. Wang, "Autonomous collection of forest field reference—the outlook and a first step with uav laser scanning," *Remote Sensing*, vol. 9, no. 8, 2017. [Online]. Available: https://www.mdpi.com/2072-4292/9/8/785

[6] S. Finland, "Finland's natural resources and the environment 2004," *Helsinki: Ministry of the Environment, Environment and Natural Resources*, vol. 2001, p. 3C, 2001.

[7] J. Viitanen, A. Mutanen, M. Kniivilä, E. Haltia, E.-J. Viitala, M. Kallioniemi, K. Härkönen, J. Leppänen, E. Uotila, and J. Routa, "Finnish forest sector economic outlook 2021–2022: Executive summary," 2021.

[8] A. Venäläinen, I. Lehtonen, M. Laapas, K. Ruosteenoja, O.-P. Tikkanen, H. Viiri, V.-P. Ikonen, and H. Peltola, "Climate change induces multiple risks to boreal forests and forestry in finland: A literature review," *Global Change Biology*, vol. 26, no. 8, pp. 4178–4196, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/gcb.15183

[9] S.-C. Wang, "Artificial neural network," in *Interdisciplinary computing in java programming*. Springer, 2003, pp. 81–100.

[10] J. Pesonen, "Pixelwise road surface slipperiness estimation for autonomous driving with weakly supervised learning," 2023.

[11] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: concepts, cnn architectures, challenges, applications, future directions," *Journal of big Data*, vol. 8, no. 1, p. 53, 2021.

[12] Aphex34, "Neurons of a convolutional layer," 2025, [Online; accessed September 1, 2025]. [Online]. Available: https://commons.wikimedia.org/w/index.php?curid=45659236

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[14] dvgodoy, "A standard transformer architecture," 2025, [Online; accessed September 1, 2025]. [Online]. Available: https://commons.wikimedia.org/w/index.php?curid=151216016

[15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," 2023. [Online]. Available: https://arxiv.org/abs/2304.02643

[16] J. L. Morgan, S. E. Gergel, and N. C. Coops, "Aerial photography: A rapidly evolving tool for ecological management," *BioScience*, vol. 60, no. 1, pp. 47–59, 01 2010. [Online]. Available: https://doi.org/10.1525/bio.2010.60.1.9

[17] B. Lu, P. D. Dao, J. Liu, Y. He, and J. Shang, "Recent advances of hyperspectral imaging technology and applications in agriculture," *Remote Sensing*, vol. 12, no. 16, 2020. [Online]. Available: https://www.mdpi.com/2072-4292/12/16/2659

[18] 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

[19] 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

[20] 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

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

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

[23] 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.