Title

David Szczecina, Hudson Sun , Niloofar Azad, name
{david.szczecina, hudson.sun, n2azad, email}@uwaterloo.ca

Abstract

Tree canopy detection from aerial imagery is an important task for environmental monitoring, urban planning, and ecosystem analysis. The Solafune Tree Canopy Detection [9] challenge provides a small, highly imbalanced dataset of only 150 annotated images, making it difficult to train deep models without severe overfitting. In this work, we evaluate five representative architectures, YOLOv11, Mask R-CNN, DeepLabv3, Swin-UNet, and DINOv2, to determine their suitability for canopy segmentation under extreme data scarcity. Our experiments show that convolution-based models (particularly YOLOv11 and Mask R-CNN) generalize significantly better than transformer-based models. DeeplabV3, Swin-UNet and DINOv2 perform poorly due factors such as the ViT's strong data requirements and lack of inductive biases, as well as differences between semantic and instance segmentation tasks. We confirm that transformer-based architectures struggle in low-data regimes without substantial pretraining or augmentation. We provide a detailed analysis of training strategies, augmentation policies, and model behavior under the small-data constraint and demonstrate that lightweight CNN-based methods remain the most reliable for canopy detection on limited imagery.

1. Introduction

Tree canopy detection from aerial or satellite imagery plays a critical role in a wide range of environmental applications, including biomass estimation, carbon accounting, biodiversity monitoring, and urban forestry. Accurate canopy mapping enables more effective policy decisions and more scalable ecological assessment. Despite its importance, high-quality canopy annotations are expensive to acquire, and many real-world datasets consist of only a few hundred samples. This creates a challenging setting for modern deep learning models, which typically rely on large, diverse datasets to achieve robust performance.

The Solafune Tree Canopy Detection competition provides an example of this setting, offering only 150 labeled images for training. The dataset's limited size, variability in lighting and terrain, and fine-grained boundaries of the canopy regions all contribute to the difficulty of the task. Additionally, individual trees are considered to be a distinct class from tight-knit groups of trees. This competition therefore serves as an ideal testbed for examining how different model classes behave in extreme low-data environments.

In this paper, we investigate four modern deep learning approaches, YOLOv11, Mask R-CNN, DeepLabv3, Swin-UNet, and DINOv2, to evaluate their ability to localize and segment tree canopies. We focus on understanding why some architectures succeed while others fail in limited-data settings, highlighting the importance of inductive bias, pretraining, and model capacity when training from only 150 images. Our findings show that convolution-based models remain considerably more robust than transformer-based models without large-scale pretraining, and we outline practical strategies for applying these methods to small ecological datasets.

2. Background

Semantic Segmentation Semantic segmentation aims to assign a class label to every pixel in an image, making it well-suited for canopy mapping where fine-grained spatial boundaries must be localized. Classical approaches relied on hand-crafted features or region-based methods, but modern segmentation heavily depends on deep neural networks. Two major families exist today: convolutional segmentation models that exploit spatial priors through localized filters, and transformer-based architectures that capture long-range context using global self-attention. In small-dataset settings such as Solafune, differences between these model classes become especially pronounced, motivating a comparative study.

Instance Segmentation TODO: Explain how this differs from semantic. Solafune challenges uses an instance segmentation mAP metric for the evaluation set.

Remote Sensing Remote sensing imagery introduces unique challenges beyond standard computer vision tasks. Variations in spatial resolution, atmospheric conditions,

sensor calibration, shadows, and land-cover diversity all contribute to high intra-class variability. Images may be acquired from satellites, fixed-wing aircraft, drones, or ground-based platforms, each providing different ground-sampling distances (GSD), focal lengths, and viewing geometries. As a result, the apparent size and shape of canopy features can vary dramatically across datasets, tree crowns may span only a few pixels in high-altitude satellite images, yet appear with fine structural detail in low-altitude UAV imagery. These factors create significant scale variation and require models that are robust to changes in feature size, texture, and resolution. Remote sensing datasets are also frequently limited in size due to the cost of expert annotation, making strong inductive biases and resistance to overfitting particularly important.

Convolution-based Approaches Convolutional Neural Networks (CNNs) such as U-Net, Mask-RCNN, DeepLabv3, and the YOLO family introduce useful spatial inductive biases such as locality and translation equivariance, making them naturally well-suited for segmentation when training data is scarce. Models like DeepLabv3 leverage dilated convolutions and multi-scale context through atrous spatial pyramid pooling (ASPP), enabling strong performance even with limited examples. Similarly, YOLOv11 offers a fast, anchor-free object detector with strong generalization due to extensive pretraining on large, diverse datasets. These properties help CNN-based models avoid severe overfitting on the 150-image Solafune dataset.

Transformer-based Approaches Vision Transformers (ViTs), including Swin-UNet and DINOv2, offer high modeling capacity and global receptive fields through self-attention mechanisms. However, ViTs lack the spatial priors inherent to CNNs and therefore require substantially more training data to learn stable visual representations. Swin-UNet introduces hierarchical attention and U-Net-like skip connections, but the core transformer blocks still remain highly data-dependent. DINOv2 provides powerful pretrained representations, yet fine-tuning with only 150 images leads to rapid overfitting and degraded segmentation quality. These limitations highlight the difficulty of applying transformer architectures to specialized remote sensing tasks without large-scale domain-relevant pretraining or heavy augmentation.

Related Works Traditional tree segmentation approaches rely heavily on LiDAR-derived 3D point cloud data and are commonly divided into two categories. The first class consists of methods that convert the point cloud into a canopy height model (CHM) and then apply surface-based analysis to detect individual tree crowns, as demonstrated in Hui et al. [4] and Roussel et al. [8]. The second class comprises full 3D methods that operate directly on the raw point cloud. Among these, the point-based clustering algorithm, as proposed in Williams et al. [14] and Pang et al. [6], is a well-known and widely used method in this category due to its conceptual simplicity and its ability to capture tree-level structure.

The advent of deep learning has substantially expanded the methodological landscape for tree segmentation. For LiDAR-based workflows, Sun et al. [10] apllied YOLOv4 and R-CNN architectures to perform crown segmentation on height maps derived from airborne LiDAR, while Wang et al. [13] used R-CNN to identify and segment tree trunks from 3D point cloud representations. Beyond LiDAR, a growing body of work has explored segmentation directly from aerial imagery, often reporting improved scalability and generalization. High-resolution aerial data, such as RGB or SAR imagery, are particularly attractive because they are easier to acquire at scale and typically provide higher spatial resolution than LiDAR. For example, Velasquez-Camacho et al. [12] employed YOLOv5 and Fast R-CNN for tree segmentation on high-resolution aerial and satellite images, and Tolan et al. [11] introduced a self-supervised vision-transformer approach using DINOv2 to generate large-scale tree height maps from RGB imagery.

Although these methods demonstrate strong segmentation accuracy, most require extensive training datasets to achieve their reported performance. To the best of our knowledge, no existing work explicitly investigates the suitability of modern deep learning architectures for scenarios where only a small training dataset is available. This motivates our study, in which we evaluate which contemporary architectures are most effective under limited-data conditions.

3. Method

In this study, we focus on and compare the performance of five architectures: Yolov11 Seg, Mask R-CNN, DeepLabV3, Swin-UNet, and DINOv2. We randomly split the original training dataset into a training set and a validation set with a 4:1 ratio. To ensure reproducibility, we fix the random seed at 42.

Dataset The dataset provided by the competition [9] contains a training set and an evaluation (test) set, each with 150 aerial images in RGB tif format. Samples of raw images and segmentation from the training dataset, with varying resolutions and scenes, are displayed in Figure 1.

Yolo Ultralytics YOLO offers a diverse suite of models designed to address a wide range of computer vision tasks, including object detection, instance segmentation, image

Figure 1. Examples of Training Images with Different Resolutions and Scenes and their Corresponding Segmentation

classification, and pose estimation. YOLOv11 represents the latest iteration in the YOLO family, introducing architectural and training improvements aimed at enhancing both accuracy and efficiency. In this study, we begin our analysis with the lightweight YOLOv11-nano segmentation model (yolov11n-seg) and subsequently extend our experiments to the small, medium, and large variants to evaluate performance across model scales. All models are initialized from weights pretrained on the COCO dataset, providing a strong general-purpose feature representation. We then fine-tune these pretrained models on the custom canopy tree segmentation dataset to adapt them to the domain-specific characteristics of aerial forest imagery.

Mask R-CNN Mask R-CNN is used as baseline fully-convolutional network. This model was designed specifically for instance segmentation and has been successfully applied across domains. Unlike YOLO, it is a two-stage detector consisting of a region proposal network followed by classification. Weights are initialized from the pretrained model as obtained in the original paper. The pretraining utilized the COCO dataset with images rescaled to several resolutions ranging from 640 to 800 pixels in width and height. The model is fine-tuned on the tree canopy dataset with random cropping and flipping augmentations. Additionally, the input images are resized to a height and width of 640 to match the pretraining data and minimize computational requirements.

DeeplabV3 DeeplabV3 was used to gauge the performance of semantic segmentation with a convolution-based architecture. -This model, developed the same year as mr-cnn, innovated by incorporating atrous convolutions and atrous spatial pyramid pooling into an FCN. -Enables scale-invariant learning, avoiding computational costs associated with high-resolution training. -Pretrained on imagenet. The final classification layer was adjusted to output scores for three classes: background, individual tree, and group of trees. In order to apply instance segmentation metrics, the output semantic segmentation masks were converted polygons using the OpenCV Python library.

DinoV2 For the transformer-based baseline, we implement a lightweight semantic segmentation model using the DINOv2-Base vision transformer as a frozen encoder. Because DINOv2 does not include a built-in decoder for dense prediction, we adapt the model by attaching a 1×1 convolutional segmentation head to the final patch embeddings, followed by bilinear upsampling to produce class logits at a 224×224 resolution. Only this segmentation head is trained; the 86M-parameter DINOv2 backbone remains frozen to avoid overfitting and to stay within the memory constraints

Table 1. Weighted mAP (IoU-based) on the Solafune Tree Canopy Detection test set. Values shown are temp.

Model	Weighted mAP
YOLOv11 Seg Large	0.281
YOLOv11 Seg Medium	0.279
YOLOv11 Seg Small	0.257
YOLOv11 Seg Nano	0.249
Mask R-CNN	0.219
DeepLabv3	0.038
Swin-UNet	0.022
DINOv2	0.021

of the GPU.

All training and evaluation images are resized to 224×224 and normalized using the ImageNet mean and standard deviation expected by DINOv2. During training, we apply standard augmentations, including flips, rotations, Gaussian blur and noise, to counteract the extreme data scarcity of the 150 image dataset. The model is optimized using Adam optimizer with a learning rate of 1e-4, cross-entropy loss, and training is run for 100 epochs.

Swin-UNet For a hierarchical transformer-based encoder–decoder model, we use Swin-UNet, following the original architecture and official implementation provided by Cao et al. [2]. Swin-UNet extends the Swin Transformer to dense prediction tasks using a symmetric U-shaped design with skip connections and shifted window self-attention, allowing the model to capture both local and global context efficiently. In our setup, we adopt the Swin-UNet-Tiny configuration and initialize the backbone with ImageNet-pretrained Swin Transformer weights. The encoder processes features through four hierarchical stages, while the decoder reconstructs spatial resolution through patch-expansion layers and skip connections from corresponding encoder stages.

All images are resized to 224x224 to match the model's native patch size (4x4). We apply standard augmentations, including random flips and rotations, to improve robustness. The model is trained using the AdamW optimizer with a learning rate of 1e-4 and cross-entropy loss for 150 epochs with a batch size of 16. This configuration enables Swin-UNet to effectively capture fine canopy boundaries and small-scale structures in aerial forest imagery.

4. Results

Main results, table comparing val/test results for all our methods

Yolo The performance for four Yolov11 segmentation models is shown in Figure 2 and Figure 3. The small model exhibits early stopping at approximately 80 epochs upon reaching its optimal validation performance, likely due to its moderate capacity, which enables it to capture the key features in the dataset more efficiently than the Nano model. In contrast, the other models continue training for the full 100 epochs, as the Nano model converges more slowly due to limited capacity, and the larger models require more epochs to fully leverage their increased representational capacity. Overall, the experiments show a clear trend of improved mask prediction performance with increasing model capacity. Training and validation losses decreased steadily across all runs, and larger models generally achieved higher mask mAP.

A noticeable gap exists between validation and test mAP, with test performance consistently higher despite similar or slightly lower validation mAP. This discrepancy is likely due to the relative size of the datasets: the validation set is much smaller than the test set, making its mAP estimates more variable and sensitive to the specific samples it contains. With only 30 images in validation versus 150 in testing, the validation mAP may not fully capture the model's generalization ability, whereas the larger test set provides a more stable and representative assessment. Among the four models, the largest variant achieved the highest test mAP, indicating that increased model capacity better captures the complexity of the segmentation task, even if validation performance alone might not fully reflect this advantage.

Mask R-CNN Figures [] show performance of mrcnn during training. Stable/smooth Slightly lower Val, expected The mAP after training was computed to be [] for the validation set and 0.22 for the tr The training and validation performance translated well to the test dataset, achieving a reasonable mAP of 0.22. This is similar to the final mAP achieved on the validation set after training.

DeeplabV3 DeeplabV3 was used to gauge the performance of semantic segmentation with a convolution-based architecture. -This model, developed the same year as mrcnn, innovated by incorporating atrous convolutions and atrous spatial pyramid pooling into an FCN. -Enables scale-invariant learning, avoiding computational costs associated with high-resolution training. -Pretrained on imagenet. The final classification layer was adjusted to output scores for each of the 3 classes. In order to apply instance segmentation metrics, the output semantic segmentation masks were converted polygons using the OpenCV Python library.

DinoV2 As seen in Figure 5 and Figure 4, the DINOv2 segmentation head trains stably, with loss decreasing and pixel accuracy improving consistently over the training

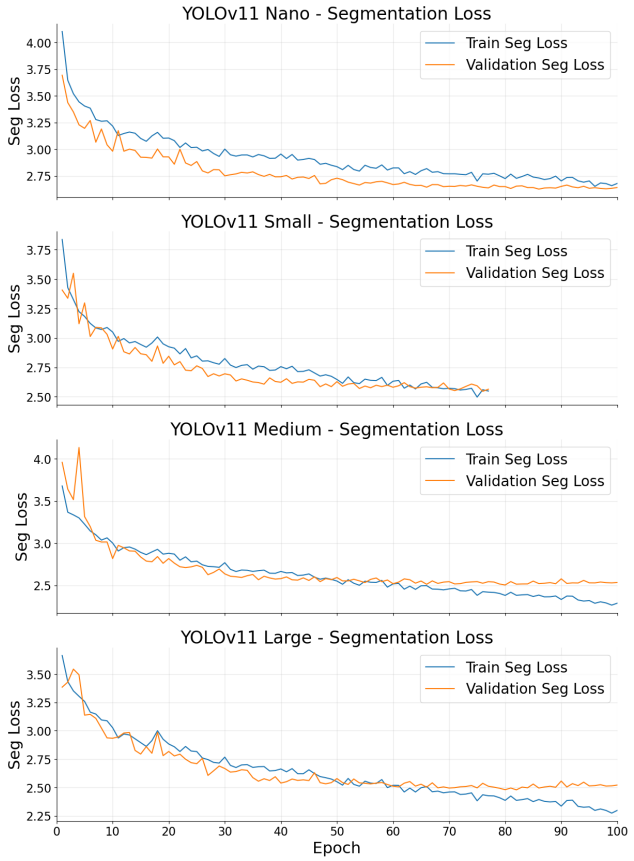

Figure 2. Training and validation loss curves for the YOLOv11 Seg models

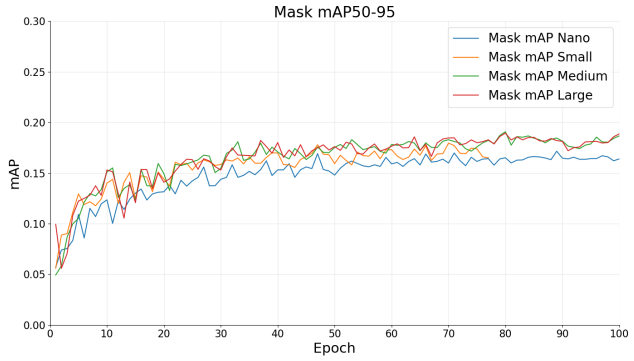

Figure 3. Validation (Mask) mAP for the YOLOv11 Seg models

epochs. This indicates that the model was able to learn a meaningful mapping on the training split despite relying on a frozen backbone and a very small dataset. However, when applied to the evaluation set, the predictions failed to transfer effectively. The final weighted mAP remained extremely low, largely because the evaluation images exhibited distributional differences and greater variability than the 150 training examples could represent. As a result, DINOv2 produced sparse or incomplete canopy predictions that translated poorly into polygon-based instance masks. These findings suggest that, unlike CNN-based methods, transformer models such as DINOv2 require substantially more data or stronger domain-specific pretraining to perform well in small, heterogeneous remote-sensing datasets.

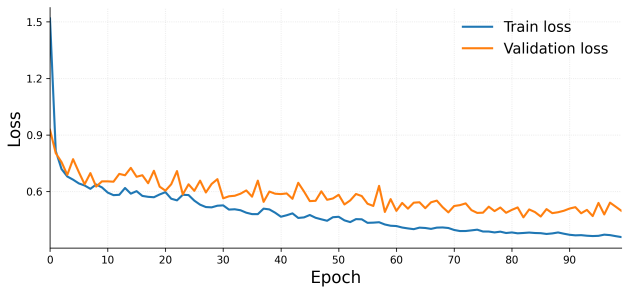

Figure 4. Training and validation loss curves for the DINOv2 model

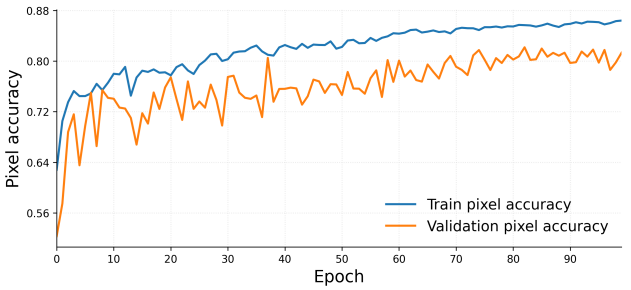

Figure 5. Training and validation pixel accuracy curves for the DINOv2 model.

Swin-UNet To evaluate transformer–based segmentation performance on the tree canopy detection task, we trained a Swin-UNet model for 150 epochs using the same training/validation split and preprocessing strategy applied to the other baselines. The training process was stable, with both training and validation losses decreasing smoothly over time, as shown in Figure 6. Despite the relatively small dataset size, the model converged without oscillation or divergence, indicating that the hybrid CNN–Transformer architecture was able to learn meaningful spatial features.

Pixel accuracy curves (Fig. 7) show that the model achieved steady improvements throughout training: training pixel accuracy increased to roughly 0.87, while validation pixel accuracy stabilized around 0.85. This suggests that Swin-UNet generalized reasonably well to the held-out validation set in terms of pixel-wise correctness.

However, these promising accuracy and loss trends did not translate into strong instance-level performance on the test set. As shown in Table 1, Swin-UNet produced a very

low weighted mAP. The primary reason is that mAP is computed from polygonized instance masks, which penalizes even small inconsistencies at object boundaries. Although the model captured broad canopy regions, the predicted masks tended to be fragmented or overly smooth, causing large errors once converted to polygon instances. This highlights a key limitation of transformer-based architectures in this setting: they require larger, more diverse datasets or domain-specific pretraining to produce crisp and topologically consistent segmentation masks suitable for polygon extraction.

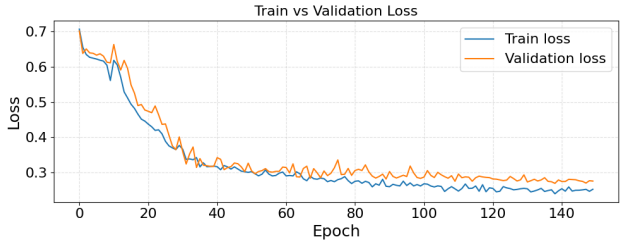

Figure 6. Training and validation loss curves for the Swin-UNet model.

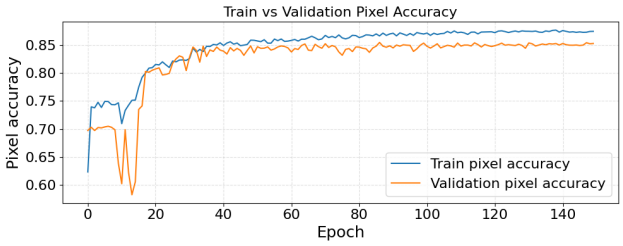

Figure 7. Training and validation pixel accuracy curves for the Swin-UNet model.

5. Discussion

DinoV2 A key reason for DINOv2's poor performance in our setting is the extremely limited amount of training data available. The Solafune dataset contains only 150 annotated images, which is far too small for a high-capacity Vision Transformer to learn meaningful canopy representations. This contrasts sharply with recent work demonstrating DINOv2's success in tree-mapping tasks, such as Tolan et al. [11], where the model was pretrained on 18 million unlabeled aerial images and then refined using 15 thousand of domain-specific canopy masks. In that setting, both the scale of the data and the close alignment between the pretraining imagery and the target task were critical to achieving strong segmentation performance.

In our experiments, neither of these conditions held: the dataset was two orders of magnitude smaller, and the evaluation imagery differed noticeably from the training set in appearance, resolution, and acquisition conditions. As a result, the frozen DINOv2 backbone lacked both sufficient data and domain context to transfer effectively. We therefore expect that, had a substantially larger and more homogeneous set of canopy images been available, similar in scope and structure to the datasets used in prior DINOv2 remote-sensing studies, the transformer would have produced far more accurate and stable canopy predictions on the Solafune competition.

Swin-UNet Beyond DINOv2 specifically, our results show a substantial performance gap between Vision Transformer–based models and conventional CNN architectures. While YOLOv11 segmentation models achieved weighted mAP scores between 0.25 and 0.28, both Swin-UNet and DINOv2 were near 0.02. This gap is consistent with well-established findings that Vision Transformers require significantly more data to achieve competitive performance, whereas CNNs are far more sample-efficient due to their strong inductive biases [3, 7]

CNNs incorporate spatial priors such as locality, translation equivariance, and hierarchical feature extraction, enabling them to generalize reliably even with small or heterogeneous datasets. In contrast, ViTs rely on self-attention without built-in spatial priors and therefore depend heavily on large-scale, domain-aligned pretraining to learn low-level structure [7]. Without such pretraining—and with only 150 labeled images—the transformer models in our study underfit severely.

This explains why Swin-UNet did not perform well despite its hierarchical attention design [5]. The model struggled to learn stable canopy boundaries under data scarcity and domain shift, producing noisy segmentation outputs. Meanwhile, CNN-based models were better able to extract texture, edge, and shape cues from limited imagery, leading to substantially higher accuracy. Overall, our findings align with prior work showing that ViTs become competitive in remote sensing only when supported by large-scale or domain-specific pretraining [1].

(in-progress) interpreting differences between all 5 models Key points: -We have compared models varying in properties: CNN, ViT, instance, semantic. -In general, cnn outperformed ViT. Attribute to small dataset. -Semantic segmentation models have good mIoU/Acc validation performance, but bad mAP test performance. -Yolo outperforming DinoV2 and SWIN-Unet shows how ViT struggles with small dataset. However, this may be biased because dino and swinunet were designed for semantic segmentation, not instance. -MRCNN outperforming DeeplabV3 shows importance of using a model (architecture+loss func) designed for instance segmentation, rather than converting

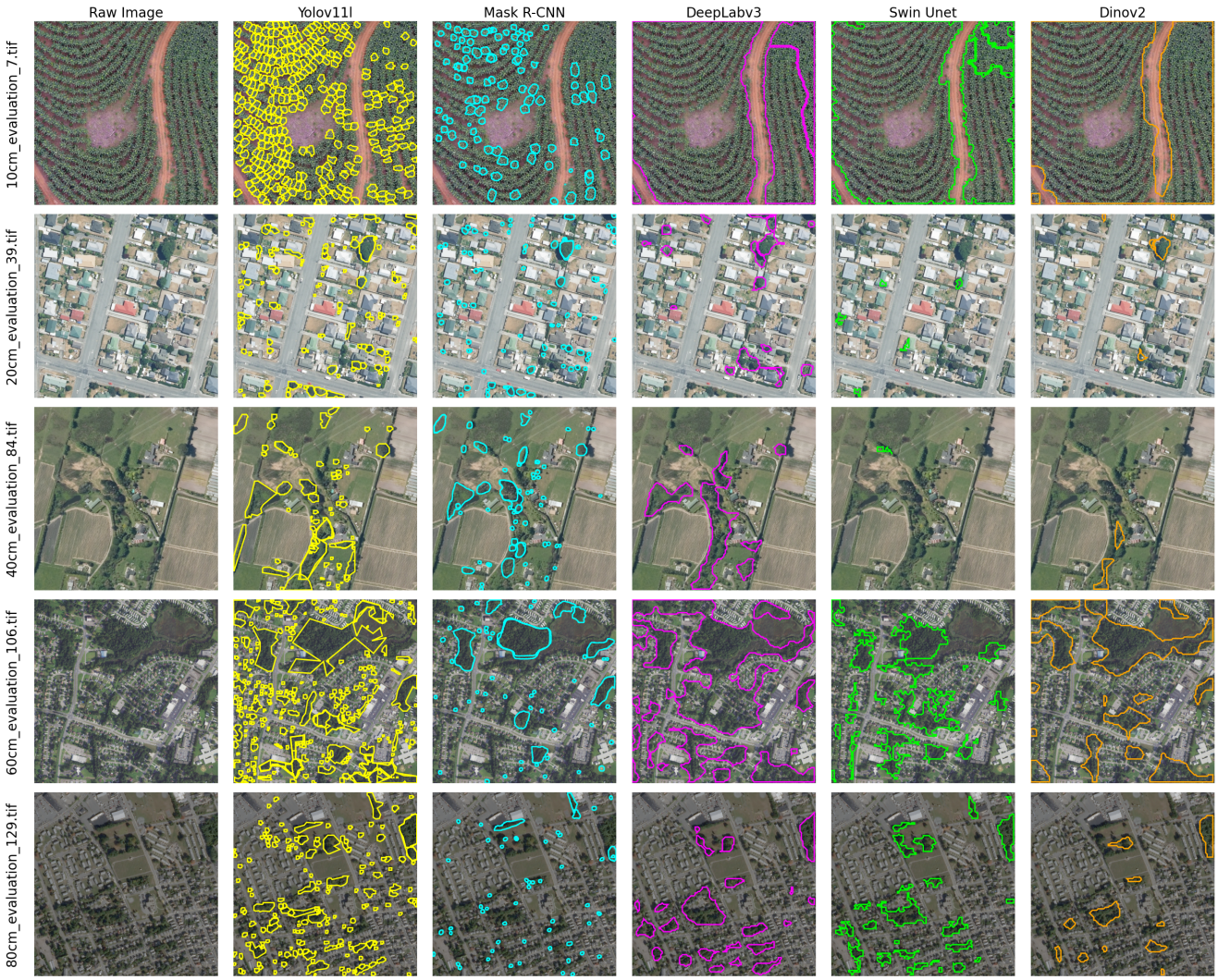

Figure 8. Segmentation Visualization for the Five Architectures

semantic segmentation to polygons. (These were SOTA in the same year, so they should be similar, only difference is the type of segmentation) -Compare mIoU instead of mACC to confirm interpretation in an unbiased way?

6. Conclusion

The convolution-based models, YOLOv11 and DeepLabv3, demonstrated the strongest and most reliable performance on the small 150 image Solafune dataset. Their inherent spatial inductive biases and extensive pretraining allowed them to generalize well despite limited supervision. In contrast, both Vision Transformer approaches, Swin-UNet and DINOv2, performed poorly, consistently overfitting and failing to produce accurate canopy boundaries. These results highlight that, in extreme low-data remote-sensing settings, lightweight CNN architectures remain far more effective than ViT-based models, which require substantially larger datasets or domain-specific pretraining to succeed

References

[1] Chamira Bandara, Vishal Patel, et al. Transformers in remote sensing: A survey. *Remote Sensing*, 14(15):3366, 2022. 6

[2] Hu Cao, Yue Wang, Jiarui Chen, Dongsheng Jiang, Xin Zhang, Qi Tian, and Yunhai Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. *arXiv preprint arXiv:2105.05537*, 2021. 4

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[4] Zhenyang Hui, Penggen Cheng, Bisheng Yang, and Guoqing Zhou. Multi-level self-adaptive individual tree detection for coniferous forest using airborne lidar. *International Jour-*

nal of Applied Earth Observation and Geoinformation, 114: 103028, 2022. 2

[5] Ze Liu, Yutong Lin, Yue Cao, and et al. Swin transformer: Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 6

[6] Yong Pang, Weiwei Wang, Liming Du, Zhongjun Zhang, Xiaojun Liang, Yongning Li, and Zuyuan Wang. Nyström-based spectral clustering using airborne lidar point cloud data for individual tree segmentation. *International Journal of Digital Earth*, 14(10):1452–1476, 2021. 2

[7] Aravindh Raghu, Thomas Unterthiner, Simon Kornblith, and et al. Do vision transformers see like convolutional neural networks? In *Advances in Neural Information Processing Systems*, 2021. 6

[8] Jean-Romain Roussel, David Auty, Nicholas C. Coops, Piotr Tompalski, Tristan R.H. Goodbody, Andrew Sánchez Meador, Jean-François Bourdon, Florian de Boissieu, and Alexis Achim. lidr: An r package for analysis of airborne laser scanning (als) data. *Remote Sensing of Environment*, 251:112061, 2020. 2

[9] Solafune, Inc. Tree canopy detection — competition overview. https://solafune.com/competitions/26ff758c-7422-4cd1-bfe0-daecfc40db70?menu=about&tab=overview, 2025. Accessed: 2025-10-09. 1, 2

[10] Chenxin Sun, Chengwei Huang, Huaiqing Zhang, Bangqian Chen, Feng An, Liwen Wang, and Ting Yun. Individual tree crown segmentation and crown width extraction from a heightmap derived from aerial laser scanning data using a deep learning framework. *Frontiers in Plant Science*, Volume 13 - 2022, 2022. 2

[11] Jamie Tolan, Hung-I Yang, Benjamin Nosarzewski, Guillaume Couairon, Huy V. Vo, John Brandt, Justine Spore, Sayantan Majumdar, Daniel Haziza, Janaki Vamaraju, Theo Moutakanni, Piotr Bojanowski, Tracy Johns, Brian White, Tobias Tiecke, and Camille Couprie. Very high resolution canopy height maps from rgb imagery using self-supervised vision transformer and convolutional decoder trained on aerial lidar. *Remote Sensing of Environment*, 300:113888, 2024. 2, 6

[12] Luisa Velasquez-Camacho, Maddi Etxegarai, and Sergio de Miguel. Implementing deep learning algorithms for urban tree detection and geolocation with high-resolution aerial, satellite, and ground-level images. *Computers, Environment and Urban Systems*, 105:102025, 2023. 2

[13] Jiamin Wang, Xinxin Chen, Lin Cao, Feng An, Bangqian Chen, Lianfeng Xue, and Ting Yun. Individual rubber tree segmentation based on ground-based lidar data and faster r-cnn of deep learning. *Forests*, 10(9), 2019. 2

[14] Jonathan Williams, Carola-Bibiane Schönlieb, Tom Swinfield, Juheon Lee, Xiaohao Cai, Lan Qie, and David A. Coomes. 3d segmentation of trees through a flexible multiclass graph cut algorithm. *IEEE Transactions on Geoscience and Remote Sensing*, 58(2):754–776, 2020. 2