Fine-grained individual tree crown segmentation based on high-resolution images

Yinrui Wang[1], Xintong Dou[1], Xinlian Liang[1]

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
(wangyinrui, douxintong, xinlian.liang) @whu.edu.cn

Keywords: Individual Tree Crown, Instance Segmentation, High-resolution Images, Deep Learning, Forest.

Abstract

The canopy of trees plays an important role in the ecological system of forest. Its cover, distribution, structure are highly relevant to the function of water cycling, carbon storage, and climate modulating of forest. At individual level, accurate tree crown masks are the bases to acquire precise locations, distribution, and structural parameters of canopy. Therefore, accurate individual tree crown (ITC) segmentation has become a key topic of forestry that supports elaborate forest monitoring, biodiversity assessment, and ecological analyses. With the rapid development remote sensing and easy accessibility of the high-resolution earth observation data, fine-grained canopy observation at individual tree level has been feasible in practice. And, deep learning technologies have achieved impressive performances on the tasks of instance segmentation which promote the accuracy of ITC delineation dramatically. This research aims to fully explore the performance of the SOTA instance segmentation networks, i.e., accuracy, generalization, and transferability, on the task of ITC segmentation. Especially, the performance of the large model, e.g., Segment Anything Model (SAM), is estimated as well. Comprehensive datasets for ITC segmentation with considerate quality, quantity, and diversity is adopted for network training and testing. Multiple ITC segmentation methods are developed by training the SOTA instance segmentation networks by datasets. The precision of the ITC segmentation method is evaluated based on standardized metrics. And, the generalization and transferability are estimated by comparing the segmentation results from testing sets that contains data from various forest types and scenarios. The method with the best performance is the network with HTC baseline and CB-ResNet50 backbone that trained by early-stop scheme, and its AP50 and AP75 achieves 40.98% and 21.25%, respectively.

1. Introduction

Forest plays an important role on the earth in carbon cycling (Mo et al., 2023; Pan et al., 2011) , climate adaption (Alkama and Cescatti, 2016), biological diversity, and ecological function (Thomas A. Spies, 1998). As one of the most essential organisms to absorb the anthropogenic CO_2 emission, its structure, distribution, and function impact the global carbon cycle and climate adaptation (Alkama and Cescatti, 2016). And, forest is the largest terrestrial carbon pool, large carbon sinks in tree biomass and forest soil of the forest (Dixon et al., 1994; Pan et al., 2011). Besides, the forest participates in the climate land-atmosphere exchange of energy and water vapor of the forest offers significant biological impacts to the macro-climate and micro-climate (De Frenne et al., 2019). Due to the increasing requirements of the larges-scale spatially continuous models of global forest biomass, the satellite-based remote sensing for earth observation has become an important measure for global forest carbon assessment (Mo et al., 2023)and biophysical climate impact estimation (Alkama and Cescatti, 2016).

The high-resolution earth observation remote sensing is expected to achieve the fine-grained and large-scale forest analysis at individual tree level and replacing the labour-exhausted *in-situ* field investigation (X. Liang et al., 2022). As the crucial part that decide the tree grows, function, and value, accurate assessment of the tree-level canopy characteristics, e.g., cover, density, quantity, horizontal distribution, provides central information for forest understanding. The 2D imagery by aerial- and satellite-based platforms offer the bird-view data of the forest upper canopy in efficient way. And, the ITC segmentation aims to generate masks delineating the canopy of each tree. At the individual tree level, accurate ITC masks is the base to extract accurate canopy characteristics. Therefore, fine-grained ITC segmentation method is one of the most valuable topics for elaborate forest analysis and understanding. The existed method could be grouped as machine-learning- (ML-) and deep-learning- (DL-) based methods.

Selecting an appropriate algorithm is the key point to develop a ML-based ITC segmentation method, e.g., valley following, region growing, watershed segmentation etc. Considering the contours of the outliers of ITCs were similar to the valleys between mountains, the valley following method was applied on the aerial grey-level image to delineate the ITC from the background vegetation (Gougeon, 1995). To improve the accuracy of the ITC segmentation, the lidar-based canopy height model (CHM) and multispectral data were integrated for the valley following method (Leckie et al., 2003). (Hyyppa et al., 2001) utilized the regional-maximum-based region growing method based on the laser-derived airborne CHM. (Wang, 2003) achieved ITC segmentation based on high-resolution aerial images using the watershed segmentation guided by treetops. However, the ML-based methods are poor in robustness and generalization and required priors for manually designed parameters.

The performance of the DL-based instance segmentation methods has surpassed the ML-based method in computer vision. Therefore, several instance segmentation networks have been applied to the task of ITC segmentation based on high resolution images and promoted its development, e.g., Mask R-CNN (He et al., 2017), Cascade Mask R-CNN(Cai and Vasconcelos, 2021), BlendMask (Chen et al., 2020), etc. The

DL-based ITC segmentation methods not only achieves better performance, but also relies on less priority of the manually designed parameters compared with the ML-based methods. (Gan et al., 2023) developed an ITC segmentation method, Dectree2, based on the Mask R-CNN using UAV-acquired RGB imagery. To enhance the texture information extraction, (Zhu et al., 2024) designed a Transformer-based contextual aggregation module to distinguish the texture differences between canopies. (Xie et al., 2024) integrated the RGB and CHM to segment the individual Chinese fir canopy using Mask R-CNN. Due to the high response of the vegetation on NIR, (Sani-Mohammed et al., 2022) utilized the Mask R-CNN to achieve standing dead tree segmentation based on the airborne images with R, G, and NIR. To release the burden of the ITC labelling, (Dersch et al., 2024) proposed a novel semi-supervised learning (SSL) scheme for Mask R-CNN training. And, the SSL-based method achieves ITC segmentation based on airborne images with R, G, B, and CHM. Except for Mask R-CNN, other networks worked well in the task of ITC segmentation. (Sun et al., 2022) adopted Cascade Mask R-CNN for ITC segmentation based on airborne RGB images in urban area. And, BlendMask achieved better performance than Mask R-CNN in ITC segmentation based on UAV images (Zhou et al., 2023). However, most of the existed ITC segmentation methods were applied on only one study site. Hence, their transferability and generalization are unclear.

With the rapid development of the deep learning technology in computer vision, several SOTA networks have achieved better performance in instance segmentation, e.g., Hybrid Task Cascade (HTC) (Chen et al., 2019), Mask DINO (Li et al., 2023), SAM (Kirillov et al., 2023). Their performances in ITC segmentation are worth for a fully exploration, i.e., accuracy, transferability, generalization. However, there is few research achieve it. Therefore, this research aims to:

(1) develop effective ITC segmentation methods by training SOTA instance segmentation networks with comprehensive datasets;

(2) explore the performance of the SOTA instance segmentation networks on the task of ITC segmentation, i.e., accuracy, generalization, and transferability;

(3) conclude the key factors that impact the performance.

2. Methodology

2.1 Study Site and Dataset

The dataset used for this research was released by the ISPRS Individual Tree Crown Segmentation Contest, 2024 (Liang et al., 2024). In general, the rich diversity of study site and sufficient quantity of data are prominent features of this dataset. The detailed information is shown in Table 1.

The study sites cover a wide range around the world. The high-resolution remote sensing images in dataset were collected from 11 study sites located in 9 countries, i.e., Canada, Malaysia, Panama, China, America, Kenya, Norway, German, and Australia. The wide spatial distribution of the study sites enables the dataset cover data from various climate zones and multiple forest type. The climate zones cover tropical, sub-tropical, and temperate zone. And, the forest types include deciduous forest, evergreen broad-leaf forest, mix forest, rain forest, boreal forest, and savanna woodland. Besides, both of the natural and urban forest are covered. Since the imageries from different study are collected under different conditions, e.g., platform, sensor, flight height, date, and light, data with different quality and resolution are included in the dataset. The ground sample distances (GSDs) of the images range from 2 to 10 cm. The diversity of data from multiple study sites is expected to evaluate the generalization and transferability between methods.

Since the deep-learning-based instance segmentation model is data-driven, datasets with adequate images annotation were necessary. Therefore, comprehensive datasets were established and provided to the participants of the contest for model training and performance evaluation, i.e., precision, generalization, transferability. There are more than 1,100 high-resolution remote sensing images of pixels and more than 600,00 ITC masks in the datasets. To maintain the reliability of the annotation, the ITC masks were labelled manually based on visual interpretation and checked by forest experts. Besides, the annotations were stored and organized based on MS COCO Format, which is a standardized data format for detection and instance segmentation. The data from 11 study sites were grouped into 11 sub-datasets. The training, validation, and testing sets contain data from dataset 1-9, 3-5, and 3-11, respectively and without overlap.

No.	Area	Resolution (cm)	Dataset		
			Train	Validate	Test
1	Canada	2.0	1691	−	−
2	Malaysia	10.0	331	−	−
3	Panama	4.5	1200	275	600
4	China	10.0	400	100	200
5	China	2.0	1721	441	786
6	China	3.0	1234	−	346
7	America	5.0	184	−	100
8	Kenya	10.0	300	−	200
9	Norway	2.0-7.0	206	−	100
10	German	2.0	−	−	468
11	Australia	2.0	−	−	200
Total	−	−	7267	816	3000

Table 1. The detailed information of the dataset released by the ISPRS Individual Tree Crown Segmentation Contest, 2024 (Liang et al., 2024).

2.2 Instance Segmentation Networks

The instance segmentation network aims to inference masks that delineating the boundaries of the targeting object. The general architecture of the network is composed of backbone, neck, initial prediction module, and heads. They achieve feature maps extraction, multi-scale feature fusion, initial prediction, and results inference, respectively.

The initial prediction module, which is the key part of detection and instance segmentation networks, locate the targeting object and extract their context feature initially. According to the forms of the initial predictions, the SOTA networks could be divided into 3 categories, i.e., proposal -based, query-based, and prompt-based network. The proposal-based networks, e.g., Masks R-CNN, and Cascade Mask R-CNN generate the masks based on the rectangular proposals that bounding the targets. The query-based networks, e.g., Mask DINO, inference the masks based on the query that embed the positional and content information in implicit representation. The prompt-based networks, e.g., SAM, are compatibility to multiple explicit

prompts indicating the locations and rough area, i.e., points, bounding boxes, and masks.

2.2.1 Proposal-based Network

The proposal-based networks generate the bounding boxes and masks by box and mask heads for the target object in parallel based on the rectangular proposals.

Mask R-CNN, which is one of the most popular baselines of proposal-based instance segmentation network, is the combination of Faster R-CNN and FCN-based mask head. The region proposal network (RPN) generates dense rectangular proposal based on the feature map from backbone and a set of pre-defined anchor boxes. The redundant proposals are removed by non-maximum suppression (NMS), then the box and mask head refine the reminding proposals to bounding boxes and generate a mask within the region of each proposal, respectively. To improve the bounding box refinement, Cascade R-CNN introduce the cascade box head. It is composed of a sequence of box heads that refine the proposals from RPN progressively. The Cascade Mask R-CNN adds mask heads within the box head and generate masks based on the refined proposals. However, there is no interaction between the mask head in the Cascade Mask R-CNN. To refine the mask progressively by the cascade heads, HTC pass the feature map from previous mask head to the succeeding one for feature fusion. Besides, a semantic head is added to provide semantic feature to each mask head.

The accuracy of the results generated by the proposal network is limited by the quality of proposals. Besides, the NMS that removes the redundant proposals is no optimizable by network training. The improper hyper-parameters of NMS will cause reduplicative or omissive segmentation. The setting of anchor and NMS impact the accuracy and generalization of the network.

2.2.2 Query-based Network

The query-based networks generate the bounding boxes and masks based on the initial predictions with implicit representation. The end-to-end object detection with Transformer (DETR) is the first query-based detection network (Carion et al., 2020). It generates the bounding boxes based on learned query that is composed of positional and content embedding. The query-based network is composed with a backbone, and Transformer-based encoder and decoder. The queries are initialized by the noise that follows standardized normal distribution, and updated based on the embedding from encoder. Then, the decoder maps the queries to bounding boxes. To enhance the convergence, the denoise training is proposed by DN-DETR (Li et al., 2022). DETR with improve denoising anchor boxes (DINO) introduces the contrastive denoising training to stabilize the training scheme using both positive and negative queries. And, the mixed query selection enhances both the positional and content embedding learning.

Mask DINO is the combination of DINO and Mask2Former. The feature embedding from encoder is recovered to the pixel-wise feature map based on up-sampling and the feature from backbone. Then, masks are generated based on the dot-production of query and pixel-wise feature map. The query-based requires no prior of pre-defining the initial prediction. However, it must generate queries of specified quantity due to the batch training. The quantity of queries should be higher than the number of the target in input image, and redundant query are filtered by the probability score. Therefore, the hyper-parameter of the query quantity will impact the recall of the results and the generalization of the network. Besides, the training of query-based networks is more difficult to converge than proposal-based networks.

2.2.3 Prompt-based Large Model

Prompt is the general concept of all forms of explicit initial prediction, i.e., point, box, mask. SAM is a generic segmentation network that generates masks for target based on the prompts. It consists of a ViT encoder, prompts encoder, and decoder. The encoder extracts feature from the input images, prompt encoder learns an embedding from for each prompt, and the decoder generate masks based on the feature and prompt embeddings. SAM was trained by a large dataset, SA-1B, which consist of 11M images and 1.1B high-quality segmentation masks. The images are collected from licensed and privacy protecting sources with virous scenes, high resolution, and good quality. The masks are generated by the segmentation anything data engine in three stages, i.e., model-assisted manual, semi-automatic, and fully automatic annotation stage. There are 99.1% of the masks are generated fully automatically. The rich diversity and quantity of dataset enable the SAM with powerful generalization that can be transferred to any new images zero-shot with impressive performance. The initial predictions of the proposal- and query-based network are generated by the module in network; however, the prompts should be input to the prompt-based network. In the data engine, the assisted-manual stage provides point prompts by manually clicking the foreground and background object points; the semi-automatic stage provides box prompts with high confidence generated by a trained detector; the fully automatic stage provides a point-grid as points prompts to the network.

2.3 Individual Tree Crown Segmentation Model

Developing an efficient ITC segmentation method includes 2 main stages, i.e., instance segmentation network establishment, ITC segmentation model training, reliable estimation of the results. This research has made a comprehensive benchmark to explore the performance of the SOTA instance segmentation network on the task of ITC segmentation. And, a proper training scheme is figured out.

According to the results of the benchmark, the network with the baseline of HTC and backbone of CB-ResNet50 (T. Liang et al., 2022) that trained by early-stop scheme is chosen as the ITC segmentation model. The architecture of the network is shown in Figure 2. The CB-ResNet50 composite 2 ResNet50 by dense higher-level composition (DHLC).

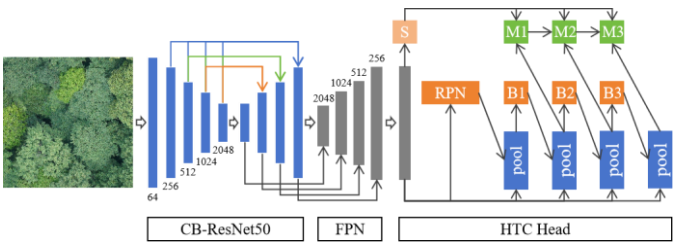


Figure 2. The network architecture of the ITC segmentation model

3. Experiments and Discussions

3.1 Experiment

A comprehensive benchmark was made by this research. There are 4 SOTA baselines of the instance segmentation are evaluated, i.e., Cascade Mask R-CNN, HTC, Mask DINO, and SAM. Besides, the backbone is one of the most important modules that impact the performance of the network. ResNet-50, which is a popular backbone for benchmarking, is adopted as the backbone to execute the experiment to explore the performance of the baselines. Besides, the backbone ResNet50, CB-ResNet50, and CB-Swin-b are compared. Since SAM is a heavy-pretrained network, it is estimated by the fully automatic mask generation function based on the original pre-trained model. The accuracy, generalization, and transferability of the SOTA networks are evaluated based on the standardized metrics and dataset.

3.2 Evaluation Metrics

3.3 ITC Results

The ITC masks generated by the HTC, Mask DINO, and SAM are shown in the Figure 3.

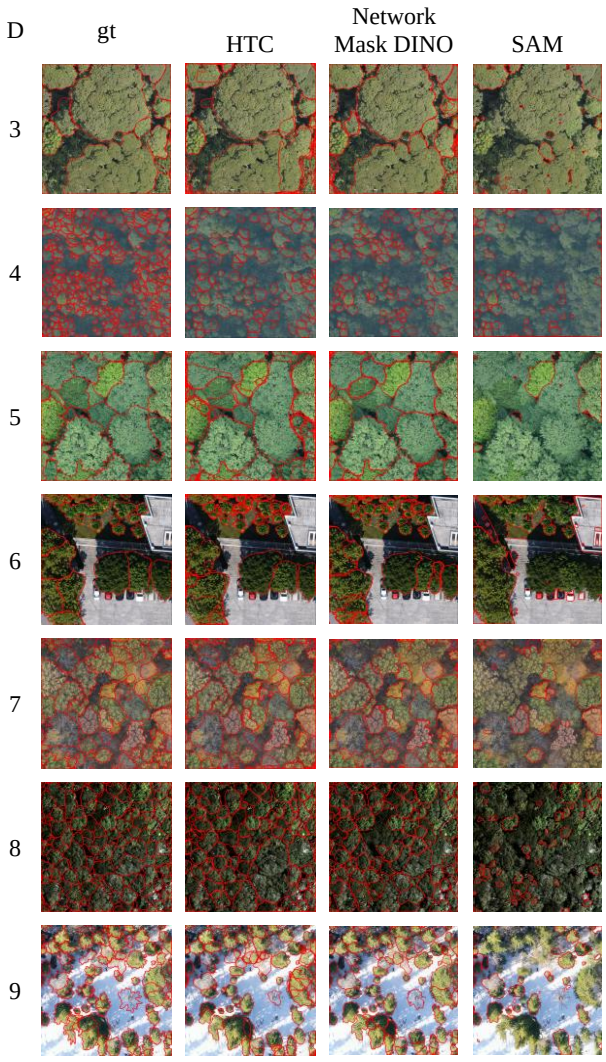


Figure 3. The ground truth and inference ITC masks in testing phase. D represents the dataset, and gt represents the ground truth. The red polygons are the inference masks.

The segmentation errors include duplicate, over-, under-, and miss-segmentation. The duplicate segmentation represents the phenomenon that multiple masks were generated for same ITC. The over- and under-segmentation indicate that one ITC is segmented to multiple masks and multiple ITC are grouped into one mask, respectively. The miss-segmentation represents that no mask is generated for one ITC. According to Figure 3, the quality of the inference masks is impacted by multiple factors.

There is great difference of the inference masks between SAM with the other networks, i.e., HTC, and Mask DINO. SAM suffers from severe problems of miss- and under-segmentation. Most of the masks from SAM are fragments, and it fails to discriminate ITC and other foreground objects. The disparity between the inference masks from HTC and Mask DINO is not evident. The cascade head HTC tends to generate more duplicate predictions slightly.

The approximation and completeness between the inference masks and ground truth in Dataset 3, 5, 6, 7, 9 are better than the other. This phenomenon is probably caused by the disparity between resolution and image quality. The resolution of the images in Dataset 4 and 8 is lower, i.e., 10 cm. While, the resolution of the images in other datasets are all higher than 5 cm, i.e., 4.5, 2, 3, 5, 2-7 cm.

The accuracy of the inference masks in Dataset 10 and 11 is the worst. The data in Dataset 10 and 11 is not included in the training set. Therefore, the performances on these two datasets represent the transferability of the networks. Since the forest scenes of Dataset 10 contain more similarity with the datasets in training set, i.e., Dataset 3, 5, and 9, its transferability is better than Dataset 11.

3.4 Evaluation Metrics

To estimate the performance of the ITC segmentation methods, metrics were adopted to evaluate the accuracy of the ITC masks generated by the methods, i.e., Average Precision (AP), Precision (P), and Recall (R), shown in (1)-(3).

$$P = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

$$P = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$AP = \int_0^1 PdR \quad (3)$$

where TP, FP, FN indicate True Positive, False Positive, and False Negative, respectively.

Intersection over union (IoU) represents the ratio of the union over intersection between inference mask and ground truth. An inference mask whose IoU and probability score exceed specific threshold will be considered as TP. Higher the threshold of probability score is, higher the Precision is and lower the Recall is, and vice versa. The P-R curve represents the relationship between Recall and Precision, and AP is the integration of the P-R curve.

The AP 50 and AP75 are commonly used metrics to evaluate the performance of the detection and instance segmentation network. They represent that the true positives take 0.5 and 0.75 as the IoU threshold, respectively. This research considers both metrics, and the values of the SOTA methods are shown in Table 2.

Network		AP50	AP75
Baseline	Backbone		
SAM	ViT	8.36	4.62
CM R-CNN		34.50	16.51
Mask DINO	ResNet50	34.92	16.92
		36.81	17.46
HTC	**CB-ResNet50**	**40.98**	**21.25**
	CB-Swin-b	40.32	19.14

Table 2. The average precision of inference masks in testing set that generated by SOTA ITC segmentation method. CM R-CNN represents Cascade Mask R-CNN.

The network with HTC baseline and CB-ResNet50 achieve the best performance. And, both of the HTC and CB-ResNet50 outperformed than the other networks in the comparison experiments, respectively. HTC baseline gains the highest AP50 and AP75 compared with Mask DINO and Cascade Mask R-CNN. And, the CB-ResNet50 gains higher AP50 and AP75 than ResNet50 and CB-Swin-b using HTC as network baseline. Although Transformer-based backbone is considered that the performance has surpassed the ConvNet in multiple tasks, e.g., ResNet, the CB-ResNet50 gain higher scores.

To further evaluate the detailed performance of best performed network in different datasets, the values of AP50 and AP75 are shown in Table 3 and Figure 4.

		AP50	AP75
Mean		40.98	21.25
Dataset	3	59.82	35.48
	4	20.22	9.20
	5	49.94	23.44
	6	**65.40**	44.93
	7	64.65	**36.05**
	8	36.03	15.03
	9	47.41	19.92
	10	21.25	5.48
	11	4.07	1.70

Table 3. The average precision of the inference masks of each dataset in testing set that generated by the network with HTC baseline and CB-ResNet50.

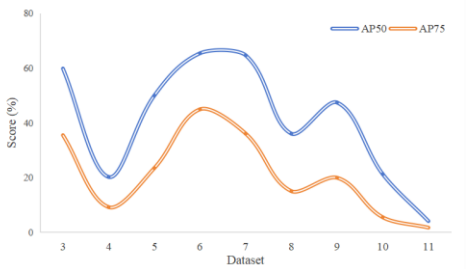
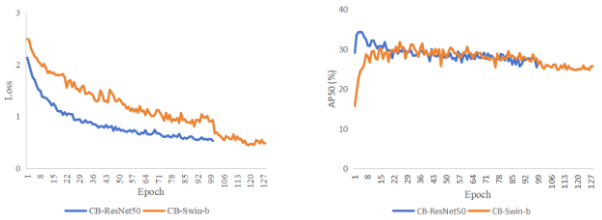


Figure 4. The average precision of the inference masks of each dataset in testing set that generated by the network with HTC baseline and CB-ResNet50.

There is great disparity between AP50 and AP75 of different datasets. The impacts from data and forest scene affect the performance of the ITC segmentation methods are more dramatically than the network architecture. The network gains the highest AP50 in Dataset 6 and 7. The images from these datasets contains high resolution, good quality, sufficient light, low density and clear boundary of ITC.

3.5 The Impact of Training Scheme

The training scheme impact the performance of the ITC segmentation model. The changing tendencies of loss and overall AP50 during the training process are monitored, as shown in Figure 5 and Table 4.



(a) (b)

Figure 5. The changing tendency of the loss and overall AP50 during training. (a) illustrate the tendency of loss, and (b) illustrate the tendency of overall AP50.

Epoch	Backbone	loss	AP50	AP75
100	CB-R50	0.53	31.11	16.01
128	CB-Swin-b	0.47	28.86	14.61
4	**CB-R50**	**1.75**	**40.98**	**21.25**
25	CB-Swin-b	1.57	40.32	19.14

Table 4. The average precision from the model trained by different epoch

The descent loss dose not lead constant accuracy improvement, even, redundant trainings (too many epochs) might damage the performance. This phenomenon represents that the generalization of the networks is required to be improved. Model fine-tuning based on early-stop training scheme helps the networks perform the best. The HTC baseline with CB-ResNet50 and CB-Swin-b achieved the best in epoch 4 and 25, respectively. CB-Swin-b requires more epoch to converge and reach the top because its architecture is more complex than CB-ResNet50.

4. Conclusion

This research made a comprehensive benchmark of SOTA methods in the tasks of ITC segmentation. The network with HTC baseline and CB-ResNet50 backbone achieves the best performance. And, the experiment results show that the resolution, image quality, forest type impact the quality of the inference masks significantly. The images with high resolution, good quality, and clear boundary between ITC are tended to generate better ITC segmentation results. Besides, the training scheme impact the performance of the ITC segmentation model greatly. Model fine-tuning by early-stop training scheme lead better performance, which means the generalization of the SOTA networks is required to be improved. And, the low performance in Dataset 10 and 11 represent the improvement potential of the transferability.

References

Alkama, R., Cescatti, A., 2016. Biophysical climate impacts of recent changes in global forest cover. *Science* 351, 600–604. https://doi.org/10.1126/science.aac8083

Cai, Z., Vasconcelos, N., 2021. Cascade R-CNN: High Quality Object Detection and Instance Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 1483–1498. https://doi.org/10.1109/TPAMI.2019.2956516

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. Springer International Publishing, Cham, pp. 213–229.

Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., Yan, Y., 2020. BlendMask: Top-Down Meets Bottom-Up for Instance Segmentation, *in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8570–8578. https://doi.org/10.1109/CVPR42600.2020.00860

Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., Loy, C.C., Lin, D., 2019. Hybrid Task Cascade for Instance Segmentation, *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).*

De Frenne, P., Zellweger, F., Rodríguez-Sánchez, F., Scheffers, B.R., Hylander, K., Luoto, M., Vellend, M., Verheyen, K., Lenoir, J., 2019. Global buffering of temperatures under forest canopies. *Nature Ecology & Evolution* 3, 744–749. https://doi.org/10.1038/s41559-019-0842-1

Dersch, S., Schöttl, A., Krzystek, P., Heurich, M., 2024. Semi-supervised multi-class tree crown delineation using aerial multispectral imagery and lidar data. *ISPRS Journal of Photogrammetry and Remote Sensing* 216, 154–167. https://doi.org/10.1016/j.isprsjprs.2024.07.032

Dixon, R.K., Solomon, A.M., Brown, S., Houghton, R.A., Trexier, M.C., Wisniewski, J., 1994. Carbon Pools and Flux of Global Forest Ecosystems. *Science* 263, 185–190. https://doi.org/10.1126/science.263.5144.185

Gan, Y., Wang, Q., Iio, A., 2023. Tree Crown Detection and Delineation in a Temperate Deciduous Forest from UAV RGB Imagery Using Deep Learning Approaches: Effects of Spatial Resolution and Species Characteristics. *Remote Sensing* 15. https://doi.org/10.3390/rs15030778

Gougeon, F.A., 1995. A Crown-Following Approach to the Automatic Delineation of Individual Tree Crowns in High Spatial Resolution Aerial Images. *Canadian Journal of Remote Sensing* 21, 274–284. https://doi.org/10.1080/07038992.1995.10874622

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

Hyyppa, J., Kelle, O., Lehikoinen, M., Inkinen, M., 2001. A segmentation-based method to retrieve stem volume estimates from 3-D tree height models produced by laser scanners. *IEEE Transactions on Geoscience and Remote Sensing* 39, 969–975. https://doi.org/10.1109/36.921414

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.-Y., Dollar, P., Girshick, R., 2023. *Segment Anything, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. pp. 4015–4026.

Leckie, D., Gougeon, F., Hill, D., Quinn, R., Armstrong, L., Shreenan, R., 2003. A Crown-Following Approach to the Automatic Delineation of Individual Tree Crowns in High Spatial Resolution Aerial Images. *Canadian Journal of Remote Sensing* 29, 633–649. https://doi.org/10.5589/m03-024

Li, F., Zhang, H., Liu, S., Guo, J., Ni, L.M., Zhang, L., 2022. DN-DETR: Accelerate DETR Training by Introducing Query DeNoising, *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 13619–13627.

Li, F., Zhang, H., Xu, H., Liu, S., Zhang, L., Ni, L.M., Shum, H.-Y., 2023. Mask DINO: Towards a Unified Transformer-Based Framework for Object Detection and Segmentation, *in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 3041–3050.

Liang, T., Chu, X., Liu, Y., Wang, Y., Tang, Z., Chu, W., Chen, J., Ling, H., 2022. CBNet: A Composite Backbone Network Architecture for Object Detection. *IEEE Transactions on Image Processing* 31, 6893–6906. https://doi.org/10.1109/TIP.2022.3216771

Liang, X., Kukko, A., Balenović, I., Saarinen, N., Junttila, S., Kankare, V., Holopainen, M., Mokroš, M., Surový, P., Kaartinen, H., Jurjević, L., Honkavaara, E., Näsi, R., Liu, J., Hollaus, M., Tian, J., Yu, X., Pan, J., Cai, S., Virtanen, J.-P., Wang, Y., Hyyppä, J., 2022. Close-Range Remote Sensing of Forests: The state of the art, challenges, and opportunities for systems and data acquisitions. *IEEE Geoscience and Remote Sensing Magazine* 10, 32–71. https://doi.org/10.1109/MGRS.2022.3168135

Liang, X., Wang, Y., Pan, J., Wang, M., Yang, J., Gong, J., 2024. The ISPRS International Contest on Individual Tree Crown Segmentation using High-Resolution Images and the Initial Findings. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-3–2024, 637–641. https://doi.org/10.5194/isprs-archives-XLVIII-3-2024-637-2024

Mo, L., Zohner, C.M., Reich, P.B., Liang, J., de Miguel, S., Nabuurs, G.-J., Renner, S.S., van den Hoogen, J., Araza, A., Herold, M., Mirzagholi, L., Ma, H., Averill, C., Phillips, O.L., Gamarra, J.G.P., Hordijk, I., Routh, D., Abegg, M., Adou Yao, Y.C., Alberti, G., Almeyda Zambrano, A.M., Alvarado, B.V., Alvarez-Dávila, E., Alvarez-Loayza, P., Alves, L.F., Amaral, I., Ammer, C., Antón-Fernández, C., Araujo-Murakami, A., Arroyo, L., Avitabile, V., Aymard, G.A., Baker, T.R., Bałazy, R., Banki, O., Barroso, J.G., Bastian, M.L., Bastin, J.-F., Birigazzi, L., Birnbaum, P., Bitariho, R., Boeckx, P., Bongers, F., Bouriaud, O., Brancalion, P.H.S., Brandl, S., Brearley, F.Q., Brienen, R., Broadbent, E.N., Bruelheide, H., Bussotti, F., Cazzolla Gatti, R., César, R.G., Cesljar, G., Chazdon, R.L., Chen, H.Y.H., Chisholm, C., Cho, H., Cienciala, E., Clark, C., Clark, D., Colletta, G.D., Coomes, D.A., Cornejo Valverde, F., Corral-Rivas, J.J., Crim, P.M., Cumming, J.R., Dayanandan, S., de Gasper, A.L., Decuyper, M., Derroire, G., DeVries, B., Djordjevic, I., Dolezal, J., Dourdain, A., Engone Obiang, N.L., Enquist, B.J., Eyre, T.J., Fandohan, A.B., Fayle, T.M., Feldpausch, T.R., Ferreira, L.V., Finér, L., Fischer, M., Fletcher, C., Frizzera, L., Gianelle, D., Glick, H.B., Harris, D.J., Hector, A., Hemp, A., Hengeveld, G., Hérault, B., Herbohn, J.L., Hillers, A., Honorio Coronado, E.N., Hui, C., Ibanez, T., Imai, N., Jagodziński, A.M., Jaroszewicz, B., Johannsen, V.K., Joly, C.A., Jucker, T., Jung, I., Karminov, V., Kartawinata, K., Kearsley, E., Kenfack, D., Kennard, D.K., Kepfer-Rojas, S., Keppel, G., Khan, M.L., Killeen, T.J., Kim, H.S., Kitayama, K., Köhl, M., Korjus, H., Kraxner, F., Kucher, D., Laarmann, D., Lang, M., Lu, H., Lukina, N.V., Maitner, B.S., Malhi, Y., Marcon, E., Marimon, B.S., Marimon-Junior, B.H., Marshall, A.R., Martin, E.H., Meave, J.A., Melo-Cruz, O., Mendoza, C., Mendoza-Polo, I., Miscicki, S., Merow, C., Monteagudo Mendoza, A., Moreno, V.S., Mukul, S.A., Mundhenk, P., Nava-Miranda, M.G., Neill, D., Neldner, V.J., Nevenic, R.V., Ngugi, M.R., Niklaus, P.A., Oleksyn, J., Ontikov, P., Ortiz-Malavasi, E., Pan, Y., Paquette, A., Parada-Gutierrez, A., Parfenova, E.I., Park, M., Parren, M., Parthasarathy, N., Peri, P.L., Pfautsch, S., Picard, N., Piedade, M.T.F., Piotto, D., Pitman, N.C.A., Poulsen, A.D., Poulsen, J.R., Pretzsch, H., Ramirez Arevalo, F., Restrepo-Correa, Z., Rodeghiero, M., Rolim, S.G., Roopsind, A., Rovero, F., Rutishauser, E., Saikia, P., Salas-Eljatib, C., Saner, P., Schall, P., Schelhaas, M.-J., Schepaschenko, D., Scherer-Lorenzen, M., Schmid, B., Schöngart, J., Searle, E.B., Seben, V., Serra-Diaz, J.M., Sheil, D., Shvidenko, A.Z., Silva-Espejo, J.E., Silveira, M., Singh, J., Sist, P., Slik, F., Sonké, B., Souza, A.F., Stereńczak, K.J., Svenning, J.-C., Svoboda, M., Swanepoel, B., Targhetta, N., Tchebakova, N., ter Steege, H., Thomas, R., Tikhonova, E., Umunay, P.M., Usoltsev, V.A., Valencia, R., Valladares, F., van der Plas, F., Van Do, T., van Nuland, M.E., Vasquez, R.M., Verbeeck, H., Viana, H., Vibrans, A.C., Vieira, S., von Gadow, K., Wang, H.-F., Watson, J.V., Werner, G.D.A., Wiser, S.K., Wittmann, F., Woell, H., Wortel, V., Zagt, R., Zawiła-Niedźwiecki, T., Zhang, C., Zhao, X., Zhou, M., Zhu, Z.-X., Zo-Bi, I.C., Gann, G.D., Crowther, T.W., 2023. Integrated global assessment of the natural forest carbon potential. Nature 624, 92–101. https://doi.org/10.1038/s41586-023-06723-z

Pan, Y., Birdsey, R.A., Fang, J., Houghton, R., Kauppi, P.E., Kurz, W.A., Phillips, O.L., Shvidenko, A., Lewis, S.L., Canadell, J.G., Ciais, P., Jackson, R.B., Pacala, S.W., McGuire, A.D., Piao, S., Rautiainen, A., Sitch, S., Hayes, D., 2011. A Large and Persistent Carbon Sink in the World’s Forests. Science 333, 988–993. https://doi.org/10.1126/science.1201609

Sani-Mohammed, A., Yao, W., Heurich, M., 2022. Instance segmentation of standing dead trees in dense forest from aerial imagery using deep learning. *ISPRS Open Journal of Photogrammetry and Remote Sensing* 6, 100024. https://doi.org/10.1016/j.ophoto.2022.100024

Sun, Y., Li, Z., He, H., Guo, L., Zhang, X., Xin, Q., 2022. Counting trees in a subtropical mega city using the instance segmentation method. *International Journal of Applied Earth Observation and Geoinformation* 106, 102662. https://doi.org/10.1016/j.jag.2021.102662

Thomas A. Spies, 1998. Forest Structure: A Key to the Ecosystem, *in: S. Proceedings of a Workshop on Structure, Process, and Diversit.* pp. 34–39.

Wang, L., 2003. Object-based methods for individual tree identification and tree species classification from high-spatial resolution imagery (Ph.D.). ProQuest Dissertations and Theses. University of California, Berkeley, United States -- California.

Xie, Y., Wang, Y., Sun, Z., Liang, R., Ding, Z., Wang, B., Huang, S., Sun, Y., 2024. Instance segmentation and stand-scale forest mapping based on UAV images derived RGB and CHM. *Computers and Electronics in Agriculture* 220, 108878. https://doi.org/10.1016/j.compag.2024.108878

Zhou, J., Chen, X., Li, S., Dong, R., Wang, X., Zhang, C., Zhang, L., 2023. Multispecies individual tree crown extraction and classification based on BlendMask and high-resolution UAV images. *Journal of Applied Remote Sensing* 17, 016503. https://doi.org/10.1117/1.JRS.17.016503

Zhu, F., Chen, Z., Li, H., Shi, Q., Liu, X., 2024. CEDAnet: Individual Tree Segmentation in Dense Orchard via Context Enhancement and Density Prior. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 17, 7040–7051. https://doi.org/10.1109/JSTARS.2024.3378167