INSTANCE SEGMENTATION OF LIDAR DATA WITH VISION TRANSFORMER MODEL IN SUPPORT INUNDATION MAPPING UNDER FOREST CANOPY ENVIRONMENT

Jian Yang[1], Lamiae El Mendili[1], Yasmin Khayer[1], Steven McArdle[1], Leila Hashemi Beni[2]

[1]VeriDaaS Corporation, Greenwood Village, CO, United States

[2]Department of Built Environment, North Carolina Agriculture & Technical State University, Greensboro, NC, United States
smcardle@veridaas.com; lhashemibeni@ncat.edu

KEY WORDS: Point Cloud, Flood, Mask2Former, 3D Modeling, Geiger Mode LiDAR.

ABSTRACT:

Inundation mapping in forest and dense vegetated areas requires the ability to generate well defined Digital Terrain Models (DTM) to derive floodwater extent, depth, and duration. Due to the occlusion caused by overlapping leaves and branch structures of forest canopies, the ability to extract elevation point clouds through UAV and airborne optical imagery and photogrammetry is challenging. LiDAR is an active sensor that acquires direct 3D measurements by transmitting hundreds of thousands of laser measurements per second producing incredibly detailed mapping layers of not only the terrain but also forest attributes such as crown diameter, tree density and height that can support inundation mapping as well as hydrological models and monitoring of floods.

In this research, we propose a methodology to map the inundated areas under canopies by using photon base Geiger Mode LiDAR point cloud dataset and a deep learning model to conduct instance segmentation of tree canopy. The method is to segment the vegetation from water and determine the gap fraction between trees to quantify the penetration through canopy for the detection of water pixels in vegetated areas. To conduct the segmentation Masked-attention Mask Transformer (Mask2Former) for universal segmentation model was implemented and trained to automate the extraction of tree crown segments from the LiDAR data. Furthermore, a semi-automatic experimental approach using a Canopy Height Model and watershed segmentation was applied to develop a rapid tree crown annotation strategy.

1. Introduction

In the hydrological process, forest areas play a vital role in retaining and delaying water flow into drainage networks. They also absorb excess water and release it back into the atmosphere through transpiration. Detection of flood areas by remote sensing has used Synthetic Aperture Radar (SAR) backscatter properties and multispectral indices such as the Normalized Water Difference Index (NWDI) (Gebrehiwot and Hashemi-Beni, 2020). However, the challenges in acquiring reliable mapping of flood area in forests and vegetated areas has been associated with spatial resolution, frequency of cloud cover for optical imagery, reflectance properties, and the ability to detect water pixels through the gap fraction of the canopy (Salem and Hashemi-Beni 2021). Airborne LiDAR technology can be used to compute tree characteristics that include crown area, orientation, and height as well as under canopy terrain that can be combined with imagery (Hashemi-Beni et al. 2021). The advent of Geiger Mode LiDAR data presents an effective solution for forestry applications due to its advantages of high-density data collection, high resolution, accuracy, and its multi-look diversity of oblique overlapping frame measurements. By integrating Geiger Mode LiDAR data with precise individual tree segmentation algorithms, it becomes feasible to accurately calculate tree attributes on a large scale and create high-definition terrain for inundation mapping.

In the past, tree crown segmentation primarily relied on watershed segmentation using a Canopy Height Model (CHM) derived from a 3D point cloud (Zhao and Popescu 2007). Watershed segmentation is a region-based method that is based on mathematical morphology. However, this approach faced challenges such as over-segmentation and vulnerability to noise, limiting its effectiveness in dense forest areas where crown segments boundaries are unclear. Additionally, parameter tuning hinders complete automation. In recent years, deep learning methods have achieved outstanding results in challenging computer vision tasks, including instance segmentation. In the context of forestry applications, Individual Tree Crown (ITC) segmentation is closely related to instance segmentation, which involves the identification and separation of individual objects. Both multispectral images and LiDAR derived CHMs can effectively be leveraged for this task. In terms of ITC detection and segmentation, Convolutional Neural Network (CNN) based architectures play a dominant role, including YOLO (Jiang et al. 2022) and Mask R-CNN (He et al. 2017). In recent years, transformer models have shown remarkable success in natural language processing tasks, which has motivated researchers to explore their application in computer vision problems as well. Through the self-attention mechanism, vision transformers can model and understand the relationship between different patches across the entire image, effectively capturing the global context of the scene. With respect to instance segmentation and object detection, DETR (Carion et al. 2020) was proposed as a transformer-based architecture and has been used for tree crown instance segmentation (Dersch et al. 2023). Although DETR was initially promising, it was still falling behind CNNs in terms of performance as it has not yet fully leveraged the potential of transformers for image instance segmentation. To address this limitation, our research introduces a transformer-based network for ITC segmentation, building upon the state-of-the-art architecture of Mask2Former (Chen et al. 2022).

In our study, we highlight the potential of applying Vision Transformer model Mask2Former to Geiger Mode LiDAR data to obtain accurate forestry analytics to support flood risk mapping. By harnessing the capabilities of Geiger Mode LiDAR

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-W2-2023-203-2023 | © Author(s) 2023. CC BY 4.0 License.

data, we can derive precise and valuable insights for various environmental applications.

2. Materials

2.1 Research data and study area

Geiger Mode LiDAR data was collected over a portion of Payette River, near Crouch Idaho on June 25, 2022, with average acquisition height at 3,787 m above ground level. The sensor collected data within a hemispherical perimeter swath of 27° Field of View using a Palmer Scanner. Data was collected with a 50% overlapping flight line. Elevation measurements are based on laser flashes illuminating a contiguous 2D array Geiger Mode Avalanche Photodiode Detector of 4,096 pixels (Figure 1). The Palmer Scanner rotates the laser light which flashes at frequency of 50 kHz producing overlapping array measurements to collect over 205 million points per second. Based on the Instantaneous Field of View (IFOV) of the individual photodiode detector and the elevation above ground, the measurement resolution which is analogous to linear LiDAR footprint was 12 cm. The internal data processing utilizes a voxel process and produces a uniform point cloud distribution of 50 points per meter squared.

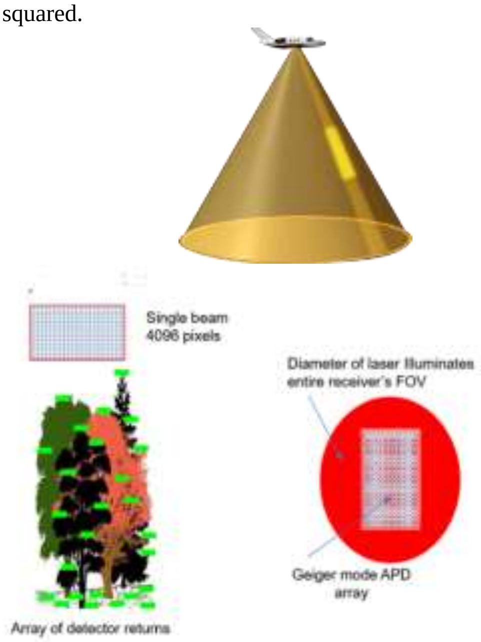


Figure 1. Geiger Mode Palmer Scanner and 2D Geiger Mode Avalanche Photodiode Detector.

Figure 2. 3D visualization of the study area

2.2 Data preparation

The LiDAR data was processed to calibrated point cloud data and projected into NAD_1983_2011_Idaho_West_ft. Then, a classification step was conducted to remove noise and non-vegetation points. This was followed by a data cleaning process and quality review. A 50 cm CHM based on ground and vegetation points was created by normalizing the height to above ground level using the Digital Terrain Model as shown in Figure 3. A total of 4 tiles are used in this study, each of which covers approximately 1,500 x 1,500 sq ft (457 m x 457 m) of area and contains ~ 3,000 trees with over 31 million LiDAR points.

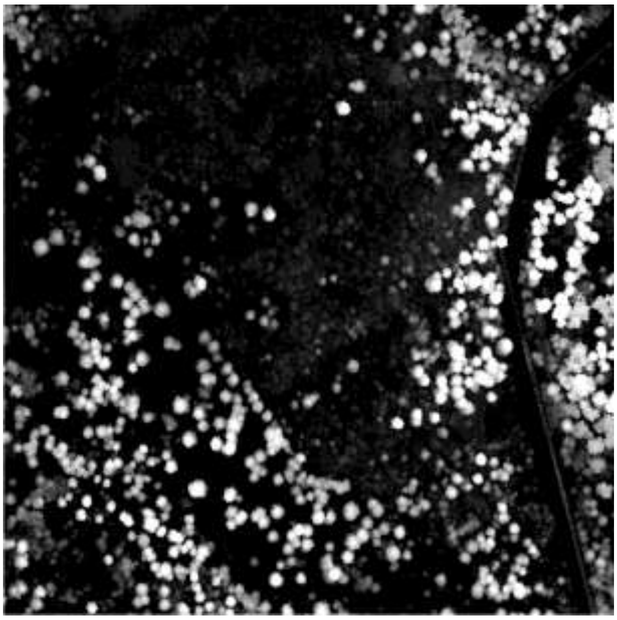


Figure 3. CHMs of the study area

3. Methodology

3.1 Overview

The research proposes a methodology based on conducting ITC segmentation from high-resolution CHM derived from Geiger Mode LiDAR data. Figure 4 provides the workflow for the processing procedures. The method starts with generating a high-resolution CHM from Geiger Mode LiDAR data for the study area. Thereafter, data labeling is conducted in a semi-automated manner. The research uses watershed segmentation to generate crown segments and then refined. The CHM and tree crown polygons are used to construct an instance segmentation dataset and finally, the Mask2Former model is implemented for crown instance segmentation.

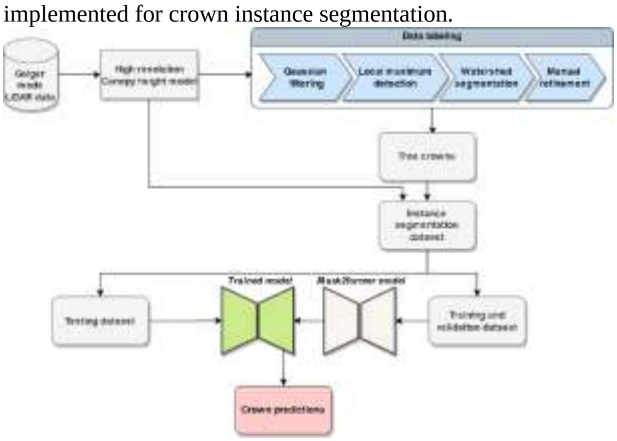


Figure 4. ITC segmentation based on the Mask2Former model.

3.2 Data labeling

Supervised deep learning models need label or annotated data for training and the validation of results. However, manually digitizing tree crown segments is time consuming and expensive especially in dense forest environments. We propose a semi-automated data labeling process incorporating local maxima-based tree detection and watershed-based crown segmentation. This is coupled with a manual refinement process in problematic areas to improve the crown delineation quality. Specifically, we first apply a Gaussian filter of 5 x 5 to smooth the appearance of CHM by minimizing artefact noises. Then, a local maxima filter of 7 x 7 is utilized to detect individual treetops. In this process, treetops below 5 m are removed given the minimum tree height, and two neighbouring treetops are merged when their distance is under 3.5 m. Next, the treetops are used as the markers for watershed segmentation, in which the height difference within a specific crown should not exceed 0.5 m or they are merged. Lastly, these tree crown segments are manually refined to get the final ground truth (Figure 5).

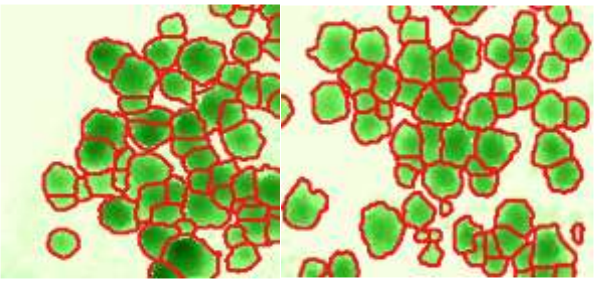


Figure 5. The semi-automated crown segmentation results for data labelling.

The study area was then split into training and validation data. Table 1 summarizes number of trees for each split.

Parameters	Training	Validation
Number of trees	5,705	1,211
Average crown area (m^2)	39.93	41.16

Table 1 Tree parameters for the training and validation split

3.3 Mask2Former

Mask2Former is a mask classification architecture (Cheng et al. 2021) where pixels are grouped into N segments by predicting N binary masks and N class labels. Unlike CNN based segmentation models where the model learns to predict a class for every pixel, mask classification splits the segmentation task into two steps: partitioning the image into N segments/regions represented by binary masks and then associating each segment as a whole to a semantic class. This formulation allows for both semantic and instance segmentation. The Mask2Former model consists of three main components: a backbone, a pixel decoder, and a transformer decoder (Figure 6). The backbone aims at extracting low resolution features from an image. The pixel decoder gradually up-samples these features to generate high-resolution per-pixel embeddings. Finally, the transformer decoder processes these embeddings using learnable object queries to produce binary mask predictions. The Mask2Former model uses the masked attention operation in the transformer decoder which constrains attention only within the foreground region of the predicted mask for each object, instead of attending to the full feature map (Cheng et al. 2022). Furthermore, instead of using use the standard convolution-based ResNet backbones, the Swin Transformer model (Liu et al. 2021) is used in this study, which is a transformer-based backbone. It builds hierarchical feature maps by merging image patches in deeper layers and has linear computation complexity to input image size due to computation of self-attention only within each local and shifted window. This makes it suitable for fine-scale instance segmentation from high-resolution images, such as individual tree crowns.

To train the Mask2Former model, a matching is necessary between the set of predictions and the set of ground truth segments. This is done using a set prediction loss that enforces a one-to-one correspondence between predicted and ground truth instances. Then, the overall model is trained using a cross-entropy classification loss and a binary mask loss. The latter is a linear combination of focal loss (Lin et al. 2018) and dice loss (Sudre et al. 2017).

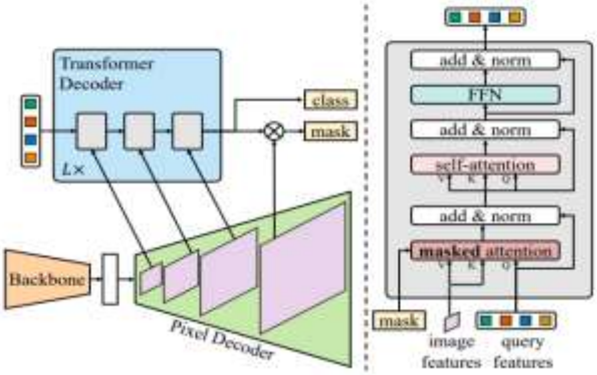


Figure 6. Mask2Former and Transformer decoder architectures (Cheng et al. 2022).

4. Experiment

4.1 Mask2Former training

To implement the model training of Mask2Former, the OpenMMLab MMDetection toolbox (Chen et al. 2019) is set up and used, which is an open-source object detection software package based on PyTorch. Since the LiDAR-derived CHMs is a single-band float image, we must convert them to 8-bit RGB images by duplicating the single band three times. Moreover, the entire CHM images are tiled into 224 x 224 with 20% overlap for both training and validation datasets. The polygons of individual tree crowns should be also converted to MS COCO format for instance segmentation model training of Mask2Former. Only one class is included in this study, i.e., crowns.

We implement our experiments on a virtual workstation that has an Intel(R) Xeon(R) Gold 5218 2.3 GHz CPU, 64 GB RAM, and 21 GB GPU memory provided by NVIDIA GRID P40-24Q. Due to the limited GPU memory, the batch sizes of training and validation are set to 2. The AdamW optimizer, the initial learning rate is set at 1e-4 is used in this study, which is a stochastic optimization method that modifies the typical implementation of weight decay in Adam, by decoupling weight decay from the gradient update. The total iteration of the training process is set to 68,750 for reaching stable validation accuracy. Due to the limited amount of training data and to leverage the capabilities of transfer learning, we initialize the Swin Transformer backbone using the pretrained weights obtained from the imagenet-1k dataset. Figure 7 shows the classification and the mask loss throughout the training process. We also show in Figure 8, the evolution of the segmentation and detection mean average precision (mAP) on the validation set. The mAP is the area under the precision-recall curve averaged for all classes.

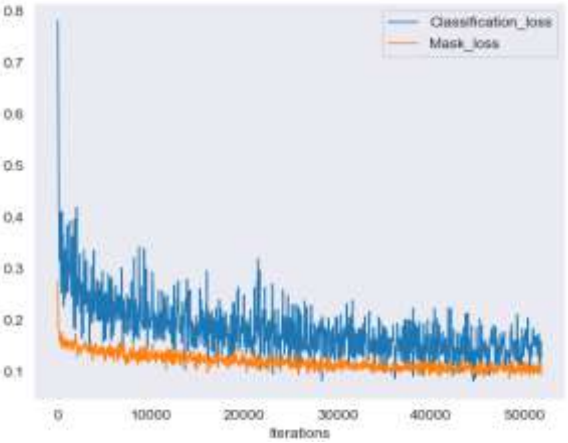


Figure 7. Visualization of the classification loss and the mask loss on the Training Set.

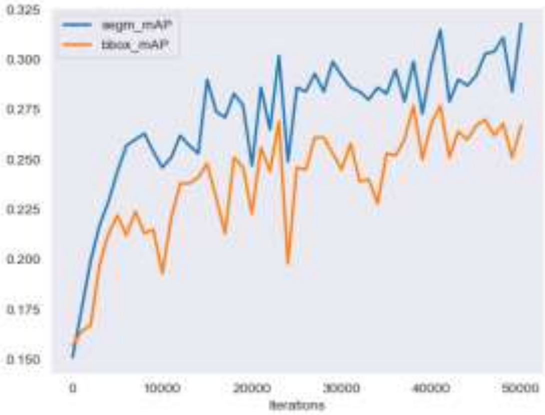


Figure 8. Validation Set Performance: Segmentation mAP and Bounding Box mAP

4.2 Evaluation

To validate the accuracy of ITC segmentation, 300 crowns are manually digitized in the upper right tile of CHMs demonstrated in Figure 3. By visual interpretation, only clearly distinguished crowns are digitized in ArcMap (Figure 9).

Since the Mask2Former segmentation results are tiled images with 224 x 224, we restore them back to the original dimension of CHM with the same georeference coordinates. To fully evaluate the detection and segmentation accuracy of instance segmentation, we define two indices below:

$$Rate\ of\ correct\ detection = \frac{Number\ of\ correctly\ detected\ ITCs}{Number\ of\ digitized\ ITCs} \quad (1)$$

$$Rate\ of\ correct\ segmentation = \frac{Number\ of\ correctly\ segmented\ ITCs}{Number\ of\ correctly\ detected\ ITCs}$$
(2)

Note that the digitized ITC is deemed as correctly detected if it intersects any of the detected ITCs. Meanwhile, the detected ITC is deemed as correctly segmented if the Intersection over Union (IoU) is over 0.5 between the digitized and detected ITCs. For further comparison, the watershed segmentation results are also quantified by the same indices.

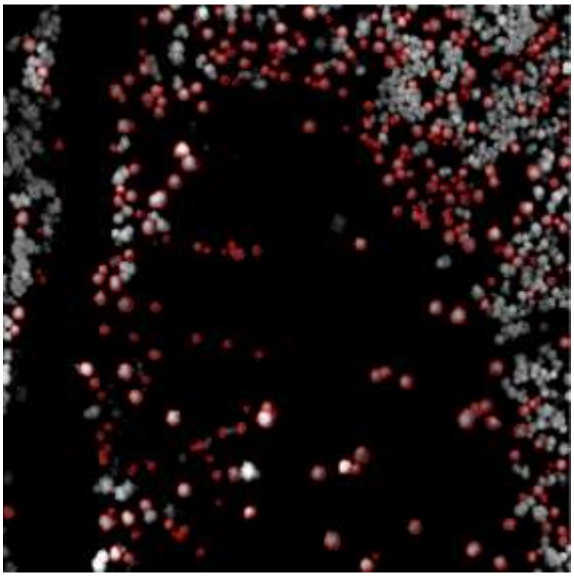


Figure 9. 300 reference ITCs over CHM.

5. Results

According to Table 1, we present different accuracy parameters for both Mask2Former and Watershed segmentation methods, including the number of digitized ITCs, the number of correctly detected ITCs, the number of correctly segmented ITCs, the rate of correct detection, and the rate of correct segmentation. Although the watershed method performs perfectly for the rate of correct detection, the Mask2Former method shows 10% increase in the rate of correct segmentation. This also demonstrates the advantages of deep learning-based instance segmentation model in ITC segmentation.

Description	Mask2Former	Watershed
Number of digitized ITCs	300	300
Number of correctly detected ITCs	261	300
Number of correctly segmented ITCs	167	162
Rate of correct detection	0.87	1.00
Rate of correct segmentation	0.64	0.54

Table 2. Accuracy parameters for ITC segmentation using Mask2Foremer and watershed methods.

Figures 10 and 11 depict an overview of ITC segmentation results by Mask2Former and watershed methods.

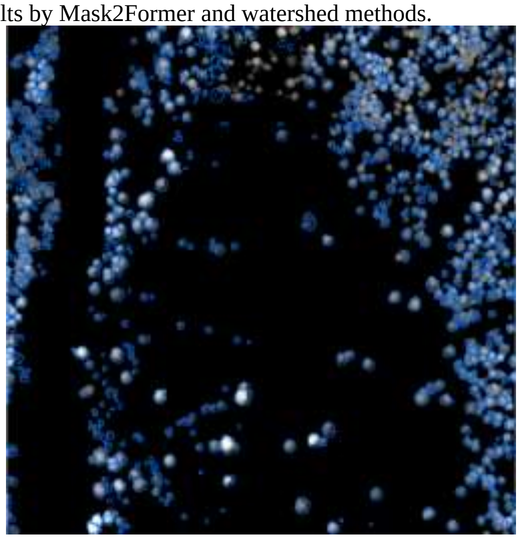


Figure 10. ITC segmentation results by Mask2Former

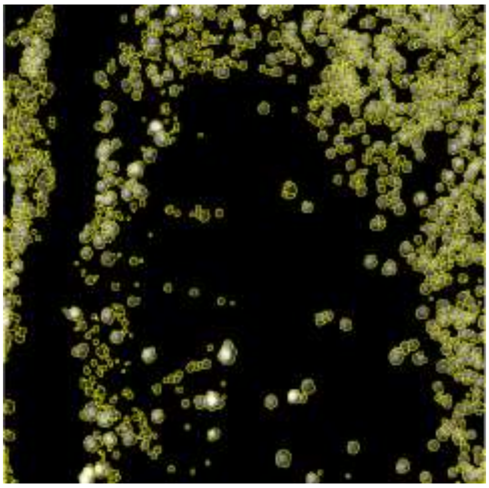
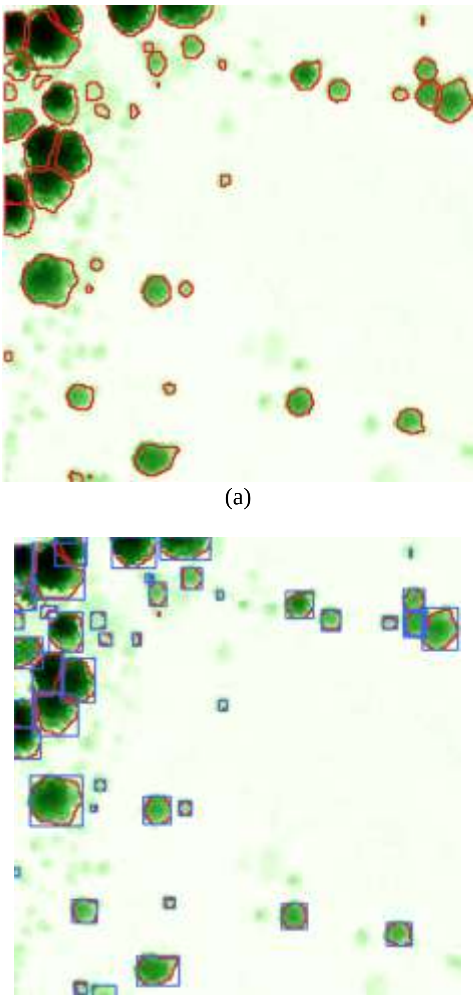


Figure 11. ITC segmentation results by watershed method.

Figure 12 demonstrates two examples of mask and bounding box predictions.



(a)

(b)

Figure 12. Mask and bounding box predictions of the Mask2Former mode.

6. Discussion

Tree instance segmentation is a challenging task that requires detecting and segmenting individual crown segments. Since the Mask2Former model is a mask classification architecture, it can flexibly generate tree instance masks. Although limited training has been applied, the results show that the model can achieve a correct detection rate of 87% and a correct segmentation rate of 64% on the test set. The correct segmentation rate is higher than the watershed baseline which shows the potential of Mask2Former in instance segmentation for challenging objects like trees. On the other hand, although the rate of correct detection is smaller than the watershed, this discrepancy is influenced by the pre-processing step applied to the point clouds, wherein non-tree objects are systematically removed to generate a refined canopy height model. Deep learning models like Mask2Former has the potential to be more robust when deployed without the pre-cleaning step. The Mask2Former can also consistently handle varying sizes and shapes of trees as shown in Figure 9. However, there are still some limitations and challenges that need to be addressed in the future. For example, the Mask2Former model requires a large amount of training data to achieve a good performance, which may not be available or feasible for some regions or scenarios. Data quality and variety is also crucial to improve the robustness of the model. Additionally, the model may benefit from incorporating auxiliary information, such as multispectral imagery, to enhance its discriminative power.

7. Conclusion

Drawing from the results outlined above, the novel approach utilizing Mask2Former displays a noteworthy enhancement in accuracy for ITC segmentation when compared to the traditional watershed technique. This demonstrates the substantial promise of leveraging state-of-the-art instance segmentation models within forestry contexts, notably for intricate attribute extraction at a fine scale, facilitated by Geiger Mode LiDAR data. However, the constrained availability and quality of crown samples, as well as the dense forest canopies curtail the full potential of Mask2Former-based ITC segmentation, underscoring the need for ongoing enhancements in forthcoming endeavors.

ACKNOWLEDGEMENTS

This work was supported in part by NOAA award NA21OAR4590358 and NASA award 80NSSC23M0051, and NSF grant 1800768.

REFERENCES

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S., 2020. End-to-end object detection with transformers. In European conference on computer vision (pp. 213-229). Cham: Springer International Publishing.

Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Lin, D., 2019. MMDetection: Open mmlab detection toolbox and benchmark. arXiv preprint arXiv:1906.07.

Cheng, B., Schwing, A., & Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems, 34, 17864-17875.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., & Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 1290-1299.

Dersch, S., Schöttl, A., Krzystek, P., & Heurich, M., 2023. Towards complete tree crown delineation by instance segmentation with Mask R–CNN and DETR using UAV-based multispectral imagery and lidar data. ISPRS Open Journal of Photogrammetry and Remote Sensing, 8, 100037.

Gebrehiwot, A., & Hashemi-Beni, L., 2020. Automated inundation mapping: comparison of methods. In IGARSS 2020-2020 IEEE International Geoscience and Remote Sensing Symposium, pp. 3265-3268.

Hashemi-Beni, L., Kurkalova, L. A., Mulrooney, T. J., & Azubike, C. S., 2021. Combining Multiple Geospatial Data for Estimating Aboveground Biomass in North Carolina Forests. Remote Sensing, 13(14), 2731.

Hashemi-Beni, L., & Gebrehiwot, A. A., 2021. Flood extent mapping: an integrated method using deep learning and region growing using UAV optical data. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14, 2127-2135.

He, K., Gkioxari, G., Dollár, P., & Girshick, R., 2017. Mask r-cnn. In Proceedings of the IEEE international conference on computer vision, pp. 2961-2969.

Jiang, P., Ergu, D., Liu, F., Cai, Y., & Ma, B., 2022. A Review of Yolo algorithm developments. Procedia Computer Science, 199, 1066-1073.

Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P., 2017. Focal loss for dense object detection. In Proceedings of the IEEE international conference on computer vision (pp. 2980-2988).

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., ... & Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 10012-10022.

Salem, A., & Hashemi-Beni, L., 2022. Inundated Vegetation Mapping Using SAR Data: A Comparison of Polarization Configurations of UAVSAR L-Band and Sentinel C-Band. Remote Sensing, 14(24), 6374.

Salem, A., & Beni, L., 2021. Comparison Between Full Polarized L-Band SAR Data and Dual Polarized C-Band SAR Data for Inundated Vegetation Mapping in Eastern North Carolina. In AGU Fall Meeting Abstracts, Vol. 2021, pp. NH45D-0617.

Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., & Jorge Cardoso, M., 2017. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3, pp. 240 -248.

Zhao, K., &Popescu, S., 2007. Hierarchical Watershed Segmetation of Canopy height model for multi-scale forest inventory. In ISPRS Workshop on Laser Scanning, pp. 436-441.