Review of: "Comparing YOLOv8 and Mask RCNN for object segmentation in complex orchard environments"

Kinde Anlay Fante[1]

1 Jimma University

Potential competing interests: No potential competing interests to declare.

The manuscript titled “Comparing YOLOv8 and Mask RCNN for object segmentation in complex orchard environments” presents a comparative study of two CNN-based instance segmentation deep learning models (YOLOv8 and Mask RCNN) for agricultural automation applications. The authors evaluated the performance of the two models in single-class and multi-class object segmentation tasks in variable orchard environments, using two datasets of RGB images collected from a commercial apple orchard. The results showed that YOLOv8 outperformed Mask RCNN.

The problem addressed by the authors is challenging and relevant, as instance segmentation is a powerful computer vision technique that can provide valuable information for various automated or robotic tasks in agriculture, such as selective harvesting, precision pruning, and yield estimation. The methods used by the authors are thoroughly described and clear. The results are clearly presented and discussed in detail, with the help of tables, figures, and metrics. The authors have considered the performance of the two models under different conditions, which is interesting and crucial in this domain.

However, I believe that this work can be improved to further advance the applications of the recently proposed deep learning algorithms in this domain. Hence, I have some minor and major comments for the authors, which are listed below.

Minor comments:

- In Table 1, there is inconsistency in the year entries for reference [74]. In the 10th row, it is mentioned correctly that it was published in 2022. However, in the last row, it is mentioned that it was published in 2023, which is incorrect. Please correct it.
- In section 3.2, line 6, “the dataset” is supposed to refer to dataset 2, which is not mentioned before this line. Please replace it with “Dataset 2.”
- The equation numbering formats seem to be incorrect in the PDF version available to me. Please correct the equation numbering format.

Major comments:

- One of the challenges in this work is the accurate annotation of the data (objects in the images). The detailed procedure to ensure the accurate annotation of the objects in the current dataset is missing in this manuscript. How did you verify the quality and consistency of the annotations? How did you handle the cases of partial or occluded objects?

Please provide more details on the data annotation process and its challenges.

- The multiclass dataset (dataset 1) consists of 1,141 annotations for tree trunks and 2,369 objects of primary branches. This shows a huge class imbalance in this data. Are the performance of the two models similar in segmenting the above objects? How did you handle the class imbalance problem? Please provide more analysis and discussion on the impact of class imbalance on the model performance and the possible solutions to mitigate it.
- Comparing the computational complexity of YOLOv8 and Mask RCNN seems to be not important, as there is always one winner in this aspect (as YOLOv8 is a one-stage architecture and Mask RCNN is a two-stage architecture).
- Even though the authors have compared the performance of two deep learning models for object segmentation in complex orchard environments, the reason for choosing these two models is not convincing. The backbone network of the two models is CNN, which uses convolution layers. The locality of convolution layers in these CNN-based models limits their capability of long-range spatial dependencies in images. In computer vision literature, different methods were proposed to improve the deep learning-based instance segmentation to overcome the limitations of CNN-based models. One such option is to use Transformer-based models (such as UnetR). The other option is to combine the strengths of transformer and CNN-based models (such as SymTC). Rather than comparing YOLOv8 (introduced in 2023) and Mask RCNN (introduced in 2017, a relatively old model), which are CNN-based models, comparing YOLOv8 with models that use a different backbone network (e.g., Transformer, or Transformer+CNN) is more important in advancing the research in this domain. Please provide a more convincing justification for choosing these two models.
- The literature review provided should include the successful segmentation algorithms in the computer vision domain in other applications. This would help to show the state-of-the-art methods and the gaps in the current research in the agricultural automation applications and also to provide more insights and inspiration for future work in this domain. Please expand the literature review to cover more related works from other domains.

Overall, I think that this manuscript presents a valuable and interesting contribution to the field of computer vision and agricultural automation, but it needs major revisions and improvements to address the comments and suggestions mentioned above. Therefore, I recommend a major revision for this manuscript. I hope that the authors will find my comments helpful.