

Zhao Z

Cardiff University
School of Engineering

Hicks Y

Cardiff University
School of Engineering

Sun X

Cardiff University
School of Computer Science
and Informatics

AI AND DEEP LEARNING

Faster Segmentation Models for Peach Ripeness Determination

To build an in-field fruit harvesting robot, it is important to locate the fruit efficiently and accurately. Instance segmentation is incorporated to evaluate peach ripeness, which enables precise identification of the ripeness level for each peach instance, allowing robots to selectively harvest ripe peaches, thereby maximizing harvesting efficiency. We have proposed a peach-specific instance segmentation model comprising three components: a ResNet50 backbone, a Feature Pyramid Network (FPN), and a Transformer decoder. Demonstrating a mean average precision (mAP) of 66.401, our model outperforms other state-of-the-art models in accurately segmenting peach instances. Notably, it achieves impressive AP of 64.818 for unripe peaches, 62.640 for semi-ripe ones, and 71.745 for ripe peaches, highlighting its effectiveness across varying ripeness levels. The model maintains the rapidest inference time of 91 ms per iteration. This comprehensive summary underscores the model's efficacy in peach instance segmentation, promising significant advancements in automated fruit harvesting and agricultural productivity.

Keywords:
Segmentation, artificial intelligence,
fruit, robot.

Corresponding author:
ZhaoZ60@cardiff.ac.uk



Z. Zhao, Y. Hicks, and X. Sun, 'Faster Segmentation Models for Peach Ripeness Determination', *Proceedings of the Cardiff University School of Engineering Research Conference 2024*, Cardiff, UK, 2024, pp. 33-37.

doi.org/10.18573/conf3.i

INTRODUCTION

The rapid advancement of artificial intelligence has revolutionized the agricultural sector. For example, the development of automated harvesting robots represents a significant trend in the field of agricultural automation, offering a promising solution to challenges such as labor shortages and rising labor costs faced by farmers. However, to achieve efficient fruit harvesting, robots must possess the ability to accurately identify and locate fruits to ensure harvesting efficacy and quality. In this regard, the technology of fruit instance segmentation plays a pivotal role. Fruit instance segmentation, leveraging computer vision and deep learning techniques, facilitates meticulous analysis and processing of fruit images, identifying various parts such as the fruit body, leaves, and branches, and accurately delineating them. This precise segmentation capability furnishes crucial information to automated harvesting robots, enabling them to precisely localize the position and size of fruits, thereby minimizing damage and waste during harvesting operations.

Through fruit instance segmentation technology, automated harvesting robots can better plan harvesting trajectories, optimize harvesting actions, and enhance harvesting efficiency through real-time adjustments. This not only reduces the need for manual intervention and lowers harvesting costs but also improves harvesting speed and quality, maximizing the utilization of orchard yields.

In the context of fruit automation production, the efficient and precise location of the target fruit serves as the foundational prerequisite to building a fruit-picking robot, which senses the working surroundings and guides the robotic arm to detach the fruits. Accurately locating target fruits is fundamental for developing efficient fruit-picking robots. These robots utilize sensors to perceive their environment and guide robotic arms to detach fruits. Over recent years, numerous harvesting robots have emerged for various fruits and crops, including kiwifruit [1], apple [2], green citrus [3] and asparagus spear [4].

Instance segmentation models offer individual identification for each fruit, enabling robots to precisely pinpoint the location of each peach instance. There are several works that have applied segmentation on agriculture. Xu et al. [5] proposed an enhanced instance segmentation model which accepts RGB and depth images for robust visual recognition for cherry tomatoes. Yu et al. [6] introduced Mask RCNN to improve the performance of fruit detection for a strawberry harvesting robot.

However, most of existing works do not consider the ripeness of the target fruit when designing the segmentation model for harvesting robot, thus results in some unnecessary loss when performing automatic in-field fruit picking.

To alleviate this problem, this paper proposed a fast and efficient instance segmentation model to determine in-field peaches at different ripeness stages. This capability facilitates precise harvesting actions, minimizing plant damage, and ensuring selective picking of only ripe fruits.

MATERIALS AND METHODS

The NinePeach Dataset

In this paper, the NinePeach dataset [7] is selected to test the performance of the segmentation models. A total of 3849 images were captured in a peach orchard at

various times throughout a complete picking campaign, representing nine cultivars of peaches that have been classified into three distinct stages of ripeness: unripe, semi-ripe, and ripe. These images are organized into training, and validation subsets, encompassing 2690 and 1159 images, respectively. Samples of images from each kind of peach are presented in Fig. 1.



Fig. 1. The NinePeach Dataset.

To our best knowledge, the NinePeach dataset is the largest and the most varied peach dataset among publicly available peach datasets with individual instance annotation provided. It includes some challenging in-field scenarios like varying natural light intensity, instances of multiple fruit adhesion, and occlusion caused by stems and leaves. The instance category distribution of the dataset can be seen at Table 1.

Category	Train	Validation
unripe	3669	1717
semiripe	3312	1307
ripe	1307	737
Total	8679	3761

Table 1. The instance category distribution of the NinePeach dataset.

The Segmentation Model

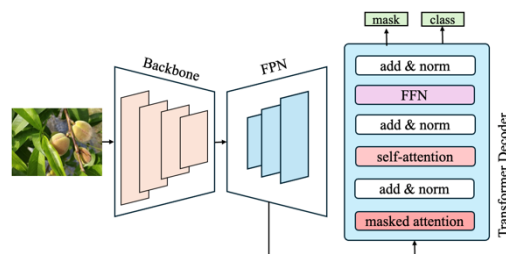


Fig. 2. The proposed segmentation model.

The structure of our segmentation model is shown in Fig. 2. It mainly consists of three parts: a backbone, a FPN and a Transformer decoder.

Backbone. Image feature extraction is the process of identifying and extracting relevant information or features from an image. ResNet [8] was proposed to solve the

image classification task. ResNet's residual blocks enable the learning of residual functions, making it easier for the network to approximate the identity mapping. This property enhances gradient flow during training, enabling smoother optimization and faster convergence. As a result, ResNet's ability to effectively combat degradation and facilitate the training of extremely deep networks makes it a highly desirable choice as a backbone architecture. This makes it easier for ResNet to learn useful features from the input image. ResNet50 is used as backbone in our segmentation model.

FPN. The Feature Pyramid Network (FPN) [9] was introduced to extend the backbone network, which is especially effective for the detection of targets at different scales. FPN works by taking the feature maps produced by backbone at different levels of the network, and building a feature pyramid that includes high-level features with strong semantics, as well as low-level features with strong spatial information. By combining features from various depths of the network, FPN enables the generation of a hierarchical pyramid of features, where each level corresponds to a different scale of the input image. This hierarchical representation facilitates the extraction of rich and diverse information, allowing the network to effectively handle objects of varying sizes and complexities. Additionally, FPN incorporates lateral connections to enhance feature propagation across different scales, enabling seamless information flow and promoting robust feature extraction. The final output of the FPN consists of a set of feature maps at four resolutions. We use FPN to fuse all features at different scales.

Transformer Decoder. The Transformer decoder module serves as a critical role within the model's architecture, acting as a pivotal bridge between the learned features extracted by the Feature Pyramid Network and the generation of final output predictions. Inspired by the foundational design principles of the original Transformer architecture [10], the decoder component orchestrates the transformation of object embeddings into output embeddings. Comprising a series of stacked decoder layers, each layer is meticulously crafted to incorporate essential components such as masked attention mechanisms, self-attention layers, and feed-forward networks (FFN). This intricate architecture allows the decoder to effectively capture intricate context dependencies and semantic relationships within the input data. In each Transformer decoder layer, the model generates predictions for both mask and class, leveraging the rich contextual information encoded within the feature representations. However, to ensure optimal performance and efficiency, only the predictions from the last layer are utilized as the final output. By strategically configuring the number of Transformer decoder layers to 3, we strike a delicate balance between model accuracy and computational efficiency.

RESULTS

Training details

The experiments are meticulously conducted using Python 3.9 and PyTorch 1.13, leveraging the computational power of two Nvidia Tesla P100 GPUs. Employing the efficient AdamW [11] optimizer coupled with a meticulously designed step learning rate schedule, our initial learning rate is set at 0.0001, accompanied by a weight decay of 0.05. To fine-tune the training process, a learning rate multiplier of 0.1 is strategically applied to the backbone architecture.

Additionally, we employ a progressive decay strategy, decreasing the learning rate by a factor of 10 at specific intervals corresponding to 0.9 and 0.95 fractions of the total training iterations. This comprehensive training regimen ensures the robustness and generalization capability of our models. The batch size is 8.

Average Precision

Precision acts as a prevalent and established measure for assessing the network's capacity to precisely detect target objects, mirroring the overall efficacy of the network. We calculated the average precision of proposed model every 5k iterations, as reported in Table 2.

Iteration	mAP	AP _{unripe}	AP _{semiripe}	AP _{ripe}
15k	54.204	51.580	50.693	60.338
20k	61.179	58.281	56.655	68.600
25k	63.051	60.695	57.503	70.956
30k	62.356	63.677	58.413	64.978
35k	64.449	59.082	62.236	72.029
40k	66.401	64.818	62.640	71.745
45k	65.670	64.225	61.469	71.315

Table 2. The average precision of our proposed model across iterations.

We observe a gradual improvement in the mean average precision (mAP) as the number of training iterations increases. From 15k to 45k iterations, the mAP increases from 54.204 to 65.670, indicating enhancements in the model's ability to recognize and segment target objects. Regarding the AP for different classes, we note the performance in identifying various levels of fruit ripeness. With the progression of training, most classes exhibit an upward trend in average precision. Particularly noteworthy is the "ripe" class, where the AP increases from 60.338 to 71.315, signifying significant improvement in identifying ripe fruits. In the "unripe" class, the increase in AP is relatively modest, rising from 51.580 to 64.225, suggesting the need for additional training data or model adjustments to further enhance performance. However, 40k iteration shows slightly better results than 45k, therefore it is selected as our final output model.

Model	mAP	AP _{unripe}	AP _{semiripe}	AP _{ripe}
Mask2Former	61.088	53.926	58.889	70.449
CMT	61.179	58.281	56.655	68.600
PVT	63.051	60.695	57.503	70.956
Ours	66.401	64.818	62.640	71.745

Table 3. The average precision of different models.

We compared the performance of our model and other similar state-of-the-art models Mask2Former [12], CMT [13] and PVT [14]. The results are shown in Table 3, highlight our model's superior performance within the same experimental framework, surpassing all other models across all metrics. With improvements of 5.313, 5.222, and 3.350 in mAP, our model demonstrates a strong performance compared to its counterparts. Notably, the relatively lower AP_{unripe} suggests that the unripe peaches are usually difficult to detect due to their green colour. On the contrary, all models demonstrate higher AP_{ripe} indicates that

ripe peaches are relatively easy to find because they are distinguishable from the leaves and trunks.

Our model outperforms the other state-of-the-art models across the three ripeness classes, demonstrating superior capability in detecting fruit at various ripeness stages. This indicates its potential as a highly effective tool for applications requiring precise ripeness classification.

Model	Inference
Mask2Former	121 ms/iter
CMT	130 ms/iter
PVT	109 ms/iter
Ours	91 ms/iter

Table 4. The inference speed of different models.

In terms of model efficiency, we report the inference speed of our model and the counterparts. The inference time denotes the time taken by the model to process the input data and generate predictions. This encompasses the forward pass through the model architecture, including operations such as feature extraction, feature ablation, and object prediction.

With the shortest inference time of 91 ms per iteration, our model demonstrates its capability to process input data swiftly and accurately. Compared to Mask2Former, CMT, and PVT, our model operates 24.8%, 30.0%, and 16.5% faster, respectively, underscoring its efficiency. Overall, the combination of efficient data loading and swift inference times highlights the effectiveness and practicality of our model for real-world applications, where both speed and accuracy are important.

Visualization

Figure 3 presents a visual representation of our model's performance, showcasing its capability to accurately segment peaches across diverse conditions. Despite encountering challenging scenarios where peaches are positioned at the periphery of the image or partially obscured, our proposed models demonstrate robustness and precision in segmenting these instances. This resilience underscores the model's ability to generalize well to various real-world conditions and effectively delineate peach instances from complex backgrounds.

Moreover, our model exhibits consistent performance across different scenarios, accurately capturing the contours and boundaries of peaches regardless of their location or occlusion level. This reliability is crucial for practical applications where precise fruit segmentation is paramount, such as automated harvesting or quality assessment in agricultural settings.



Fig. 3. Visualization of the segmentation.

CONCLUSION

Integrating instance segmentation for assessing peach ripeness enhances the intelligence and efficiency of automatic fruit-picking robots. By accurately identifying the ripeness level of each peach instance, robots can selectively harvest ripe fruits, maximizing harvesting efficiency while avoiding picking immature ones. To achieve this, we have introduced an instance segmentation model specifically designed for peaches, comprising three key components: a ResNet50 backbone, a FPN, and a Transformer decoder. With the highest mAP of 66.401, our model demonstrates better performance in segmenting peach instances comparing to other state-of-the-art models. Notably, it achieves AP scores of 64.818 for unripe peaches, 62.640 for semi-ripe ones, and 71.745 for ripe peaches, showcasing its effectiveness across different ripeness levels. Despite its high accuracy, the model maintains the shortest inference time of 91 ms per iteration, ensuring swift processing suitable for real-time applications. This comprehensive summary underscores the model's efficacy in peach instance segmentation, promising advancements in automated fruit harvesting and agricultural productivity.

Acknowledgments

The authors appreciate the computational resources provided by Advanced Research Computing at Cardiff (ARCCA).

Conflicts of Interest

The authors declare no conflict of interest.

REFERENCES

- [1] L. Mu, G. Cui, Y. Liu, Y. Cui, L. Fu, and Y. Gejima, 'Design and simulation of an integrated end-effector for picking kiwifruit by robot', *Information Processing in Agriculture*, vol. 7, no. 1, pp. 58–71, Mar. 2020. doi.org/10.1016/j.inpa.2019.05.004
- [2] H. Kang and C. Chen, 'Fruit detection, segmentation and 3D visualisation of environments in apple orchards', *Computers and Electronics in Agriculture*, vol. 171, p. 105302, Apr. 2020. doi.org/10.1016/j.compag.2020.105302
- [3] J. Lu *et al.*, 'Lightweight green citrus fruit detection method for practical environmental applications', *Computers and Electronics in Agriculture*, vol. 215, p. 108205, Dec. 2023. doi.org/10.1016/j.compag.2023.108205
- [4] M. Peebles, S. H. Lim, M. Duke, and B. McGuinness, 'Investigation of Optimal Network Architecture for Asparagus Spear Detection in Robotic Harvesting', *IFAC-PapersOnLine*, vol. 52, no. 30, pp. 283–287, 2019. doi.org/10.1016/j.ifacol.2019.12.535
- [5] P. Xu, N. Fang, N. Liu, F. Lin, S. Yang, and J. Ning, 'Visual recognition of cherry tomatoes in plant factory based on improved deep instance segmentation', *Computers and Electronics in Agriculture*, vol. 197, p. 106991, Jun. 2022. doi.org/10.1016/j.compag.2022.106991
- [6] Y. Yu, K. Zhang, L. Yang, and D. Zhang, 'Fruit detection for strawberry harvesting robot in non-structural environment based on Mask-RCNN', *Computers and Electronics in Agriculture*, vol. 163, p. 104846, Aug. 2019. doi.org/10.1016/j.compag.2019.06.001
- [7] Z. Zhao, Y. Hicks, X. Sun, and C. Luo, 'Peach ripeness classification based on a new one-stage instance segmentation model', *Computers and Electronics in Agriculture*, vol. 214, p. 108369, Nov. 2023. doi.org/10.1016/j.compag.2023.108369
- [8] K. He, X. Zhang, S. Ren, and J. Sun, 'Deep Residual Learning for Image Recognition', in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi.org/10.1109/CVPR.2016.90
- [9] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, 'Feature Pyramid Networks for Object Detection', in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 936–944. doi.org/10.1109/CVPR.2017.106
- [10] A. Vaswani *et al.*, 'Attention Is All You Need', *arXiv*, 1 Aug. 2023. doi.org/10.48550/arXiv.1706.03762
- [11] I. Loshchilov and F. Hutter, 'Decoupled Weight Decay Regularization', *arXiv*, 4 Jan. 2019. doi.org/10.48550/arXiv.1711.05101
- [12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, 'Masked-attention Mask Transformer for Universal Image Segmentation', in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA: IEEE, Jun. 2022, pp. 1280–1289. doi.org/10.1109/CVPR52688.2022.00135
- [13] J. Guo *et al.*, 'CMT: Convolutional Neural Networks Meet Vision Transformers', *arXiv*, 14 Jun. 2022. doi.org/10.48550/arXiv.2107.06263
- [14] W. Wang *et al.*, 'Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction without Convolutions', *arXiv*, 11 Aug. 2021. doi.org/10.48550/arXiv.2102.12122

Proceedings of the Cardiff University School of Engineering Research Conference 2024 is an open access publication from Cardiff University Press, which means that all content is available without charge to the user or his/her institution. You are allowed to read, download, copy, distribute, print, search, or link to the full texts of the articles in this publication without asking prior permission from the publisher or the author.

Original copyright remains with the contributing authors and a citation should be made when all or any part of this publication is quoted, used or referred to in another work.

E. Spezi and M. Bray (eds.) 2024. *Proceedings of the Cardiff University School of Engineering Research Conference 2024*. Cardiff: Cardiff University Press.
doi.org/10.18573/conf3

Cardiff University School of Engineering Research Conference 2024 was held from 12 to 14 June 2024 at Cardiff University.

The work presented in these proceedings has been peer reviewed and approved by the conference organisers and associated scientific committee to ensure high academic standards have been met.

First published 2024

Cardiff University Press
Cardiff University, Trevithick Library
First Floor, Trevithick Building, Newport Road
Cardiff CF24 3AA

cardiffuniversitypress.org

Editorial design and layout by
Academic Visual Communication

ISBN: 978-1-9116-5351-6 (PDF)



This work is licensed under the Creative Commons Attribution - NoCommercial - NoDerivs 4.0 International licence.

This license enables reusers to copy and distribute the material in any medium or format in unadapted form only, for noncommercial purposes only, and only so long as attribution is given to the creator.

<https://creativecommons.org/licenses/by-nc-nd/4.0/>