Hierarchical Mask2Former: Panoptic Segmentation of Crops, Weeds and Leaves

Madeleine Darbyshire[1] and Elizabeth Sklar[2] and Simon Parsons[1]

Abstract

Advancements in machine vision that enable detailed inferences to be made from images have the potential to transform many sectors. Precision agriculture is one such application area. By detecting weeds and inferring crop growth via leaf counts, interventions can be applied only where they are needed. This enables farmers to maximise their yields while limiting resource use. In this work, we propose a hierarchical panoptic segmentation method to extract crop, weed and leaf instances from images. We adapt Mask2Former, a state-of-the-art architecture for panoptic segmentation, to predict both plant and leaf masks. Our best model achieves a $PQ^{\dagger}$ of 76.78, and a PQ of 71.95 and 66.31 on crop and leaf instances respectively. Additionally, weeds are segmented with an IoU of 69.49.*

1. Introduction

Demand for food is growing as the global population increases. Farmers are now required to meet this demand whilst simultaneously reducing the environmental impact of their operations. All the while, climate changes is making growing conditions more unpredictable leading to new challenges in providing a reliable supply of food.

Precision agriculture aims to leverage data and machine learning to help farmers make more informed decisions. An area where this has notable influence is precision weed management. Farmers can reduce their herbicide usage by first detecting, and then only targeting weeds, rather than spraying the entire field with herbicide. Furthermore, crop monitoring, another subdomain of precision agriculture, can indicate where fertiliser should be targeted for healthy plant growth. Various phenotypic traits can be used as indicators of crop growth but in this paper we use leaf count.

The paper aims to combine crop and weed segmentation as well as leaf segmentation masks in a single hierarchical panoptic segmentation architecture. Our approach employs the latest state-of-the-art panoptic segmentation architecture and improves upon existing baselines.

[1]MD and SP are in the School of Computer Science, University of Lincoln, UK. 25696989@students.lincoln.ac.uk, sparsons@lincoln.ac.uk

[2] ES are with the Lincoln Institute of Agri-food Technology, University of Lincoln, UK. esklar@lincoln.ac.uk

*Our code is published at: https://github.com/madeleinedarbyshire/HierarchicalMask2Former

2. RELATED WORK

Earlier implementations of deep networks for crop and weed semantic segmentation [7] used SegNet [2]. SegNet employs an encoder-decoder structure, where the encoder extracts hierarchical features from input images, and the decoder produces pixel-wise segmentation masks. DeepLabV3+ improves on the performance of encoder-decoder segmentation architectures by adding atrous convolutions to capture larger spatial context [3]. Later approaches utilised U-Net [14], which employs skip connections, outperformed DeepLabv3+ in plant segmentation [18].

Beyond classifying the pixels, instance segmentation using Mask R-CNN made it possible to distinguish between individual crop and weed plants [11]. Mask R-CNN [8] combines pixel-level semantic segmentation with object bounding box predictions to produce instance segmentation. This approach can be applied to many panoptic segmentation problems as well. Panoptic-DeepLab [4] builds on an adapted version of DeepLabV3+, adding instance segmentation segmentation heads, to make it suitable for instance segmentation and panoptic segmentation. Mask2Former [5], and it's predecessor MaskFormer [6], propose an approach to instance segmentation that differs from the per-pixel approaches proposed before. Instead, images are partitioned into a number of regions, represented with binary masks, then each of these is assigned a class.

Adjacent to the problem of identifying individual plant instances is the identification of individual leaf instances within each plant instance. In [16] it was demonstrated that each plant and its leaves could be identified from images, containing multiple plants, taken under real field conditions. Subsequently, the tasks of crop and weed segmentation with leaf instance segmentation were combined in [12]. In [12], a second decoder is added to the ERFNet [13] so that one decoder produces plant masks and classes while the other produces leaf masks.

In this paper we explore whether the quality of segmentation of crops, weeds and leaves can be improved by adapting

state-of-the-art segmentation architecture, Mask2Former. Similar to [12], our approach utilises two decoders: one for generating plant masks and one for generating leaf masks.

3. METHODOLOGY

3.1. Mask2Former

Our approach adapts the Mask2Former [5] architecture. The original Mask2Former architecture is shown in Figure 1 and consists of the following:

Backbone. The backbone extracts low-level image features $F^{C_F \times \frac{H}{S} \times \frac{W}{S}} \in \mathbb{R}$ from an input image of size $H \times W$, where C_F is the number of channels and S is the stride.

Pixel decoder. The pixel decoder gradually upsamples the low-level features to produce a feature pyramid with layers that are of resolution 1/32, 1/16 and 1/8. At each stage in the upsampling process, a per-pixel embedding is created $\epsilon_{pixel} \in \mathbb{R}^{C_\epsilon \times H \times W}$, where C_ϵ is the embedding dimension. In this implementation, the advanced multi-scale deformable attention Transformer, *MSDeformAttn* [17] is used as the pixel decoder.

Transformer decoder. The transformer decoder consists of 3 transformer layers for each layer of the feature pyramid. Therefore, given there are 3 layers in the feature pyramid, there are 9 transformer decoder layers. Each transformer decoder layer consists of a self-attention layer, a cross-attention layer and a feed-forward network. Query features are associating with the positional embeddings produced by the pixel decoder. These query features are learnable and thus updated by each layer of the network. The transformer outputs N per-segment embeddings, $Q^{C_Q \times N} \in \mathbb{R}$, where N is the number of queries and C_Q is the dimension that encodes global information about the segment.

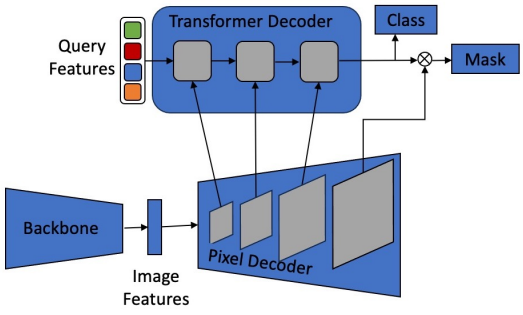

Figure 1: Original Mask2Former architecture.

Segmentation Module The segmentation module transforms the output of the transformer Q into masks and class predictions. To acquire class probability predictions $\{p_i \in \Delta^K\}_{i=1}^N$, a linear classifier and softmax activation are applied to the output. There is an additional no-object class which applies where the embedding does not correspond to any region. To generate the masks, a multi-layer perceptron converts the per-segment embeddings from the transformer into N mask embeddings $\epsilon_{mask} \in \mathbb{R}^{C_\epsilon \times N}$. Lastly, binary masks $m_i \in [0,1]^{H \times W}$ are formed via the dot product of the mask embeddings, ϵ_{mask}, and per-pixel embeddings, ϵ_{pixel}, followed by a sigmoid activation $m_i[h,w] = sigmoid(\epsilon_{mask}[:,i]^T \cdot \epsilon_{pixel}[:,h,w])$.

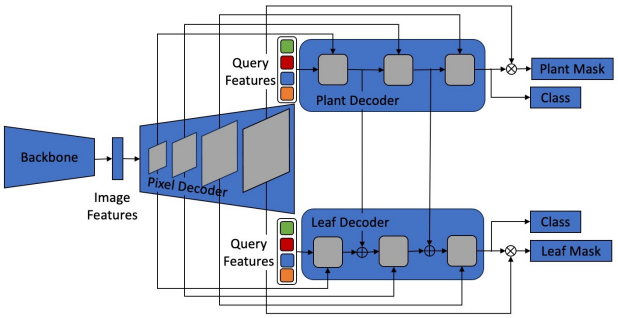

Figure 2: Adapted Mask2Former architecture with separate transformer decoders for plants and leaves.

3.2. Separate Transformer Decoders for Plants and Leaves

We use separate plant and leaf transformer decoders because we thought the network might learn better how to segment at the plant-level versus at the leaf-level. Each transformer decoder takes its own set of learnable query features and each produces N per-segment embeddings Q_{plant} and Q_{leaf}, respectively. Each segmentation module takes the output of each transformer decoder, Q_{plant} and Q_{leaf}, and forms two sets of class predictions $\{p_i^{plant} \in \Delta^{K_{plant}}\}_{i=1}^N$ and $\{p_i^{leaf} \in \Delta^{K_{leaf}}\}_{i=1}^N$ for plants and leaves, respectively. Additionally, separate mask embeddings $\epsilon_{mask}^{plant} \in \mathbb{R}^{C_\epsilon XN}$ and $\epsilon_{mask}^{leaf} \in \mathbb{R}^{C_\epsilon XN}$ are generated via separate multi-layer perceptrons. Each of these, combined with the pixel embedding ϵ_{pixel} using the dot product would produce the two sets of binary mask predictions: $m_i^{plant} \in [0,1]^{H \times W}$ and $m_i^{leaf} \in [0,1]^{H \times W}$ using the same approach described in 3.1.

3.3. Loss Function

As in Mask2Former, the mask loss is calculated using the sum of the binary cross entropy loss and dice loss:

$$L_{\text{mask}} = L_{\text{ce}} + L_{\text{dice}} \quad (1)$$

The mask loss and class loss is calculated for both the plants and leaves, respectively. The total loss is calculates

as a weighted sum of the classification and mask loss of the plants and leaves:

$$L = \lambda_{\text{cls}} L^p_{\text{cls}} + \lambda_{\text{mask}} L^p_{\text{mask}} + \lambda_{\text{cls}} L^l_{\text{cls}} + \lambda_{\text{mask}} L^l_{\text{mask}} \quad (2)$$

where L^p and L^l are the losses for the plants and leaves respectively. The weights for each of the losses were set to $\lambda_{\text{mask}} = 2.5$ and $\lambda_{\text{cls}} = 1.0$.

3.4. Skip Connections

Skip connections are added to connect the output of the first and second feature levels of the plant transformer with the leaf transformer. These skip connections are inspired by those in [12] and aim to share information about the plant location from the plant decoder with the leaf decoder.

4. EXPERIMENTS

4.1. Dataset

Our approach was tested against the PhenoBench dataset [15]. It consists of RGB images of sugarbeet crops and weeds taken from a UAV. These images were annotated on three levels: first plants, weeds and soil was semantically segmented, second plant (crop and weed) instances were segmented, and finally each leaf instance of the sugarbeet crops was segmented. The training set contains 1407 images, the validation set contains 772 images and the test set contains 693 images. Since this work was completed as part of the *CVPPA@ICCV'23: Hierarchical Panoptic Segmentation of Crops and Weeds competition* [1] the test set is hidden at time of publication. Therefore, our work is ablated against the validation set and the only the final results against the test set. Further details about the dataset collection and annotation process can be found in [15].

4.2. Model and Training

As well as ResNet-50 [9], we chose use the Swin transformer [10] as the backbone because this has shown state-of-the-art accuracy in image classification. We trained two sizes of Swin model: a smaller model with Swin-S as the backbone and 100 object queries, and a larger model with Swin-L and 200 object queries. The learning rate for both was set to 0.00001. Each model was trained for a maximum of 85 epochs.

4.3. Evaluation

As in [15], panoptic quality is used to assess the predicted masks of crops PQ_{crop} and leaves PQ_{leaf} . The average over these values is reported as PQ. During evaluation, plant or leaf instances where less than 50% of it's pixels are within the image, do not affect the score, since these are regarded as uninformative. Additionally, the IoU is calculated for the "stuff" categories: weeds IoU_{weed} and soil IoU_{soil}. The metric $PQ^{\dagger}$ is the average over PQ_{crop}, PQ_{leaf}, IoU_{weed} and IoU_{soil}.

5. RESULTS

The results for each model on validation set are presented in Table 1. The best of these results was submitted to be tested against the hidden test set and these results are presented in Table 2.

6. DISCUSSION

Mask2Former combines state-of-the-art semantic segmentation and instance segmentation and since our results show that our model was strong in both semantic segmentation and instance segmentation these demonstrate the benefit of this approach. IoU_{soil} is close to perfect at 99.45,

Table 1: Ablations on the validation set

Backbone	Epochs	Mask Threshold	Skip Connections	$PQ^{\dagger}$	PQ	PQ_{crop}	PQ_{leaf}	IoU_{weed}	IoU_{soil}
ResNet-50	63	0.8	No	76.12	68.53	73.6	63.45	68.05	99.38
ResNet-50	63	0.8	Yes	76.65	69.16	74.36	63.97	68.88	99.38
Swin S	63	0.5	Yes	78.42	70.14	76.92	63.35	73.98	99.42
Swin S	63	0.8	No	78.8	71.21	76.4	66.01	73.37	99.41
Swin S	85	0.5	Yes	79.03	71.36	77.66	65.06	73.98	99.43
Swin S	85	0.8	Yes	78.95	71.2	77.34	65.06	73.98	99.43
Swin S	85	0.8	No	79.18	71.73	77.3	66.17	73.8	99.43
Swin L IN21K	63	0.5	Yes	**80.34**	**73.33**	**78.06**	**68.6**	**75.26**	**99.45**
Swin L IN21K	63	0.8	Yes	80.26	73.16	75.26	68.6	75.26	99.45
Swin L IN21K	85	0.8	Yes	79.89	72.32	77.93	66.71	75.46	99.45

Table 2: Results from our model on hidden test set compared to results published in [16]

Model	$PQ^{\dagger}$	PQ	PQ_{crop}	PQ_{leaf}	IoU_{weed}	IoU_{soil}
Mask2Former	-	-	-	57.50	-	-
HAPT [12]	65.27	50.73	54.61	46.84	61.11	98.50
Ours (Swin S, no skips, MT=0.5)	75.78	67.71	71.56	63.86	68.37	99.35
Ours (Swin L IN21K, with skips, MT=0.5)	**76.78**	**69.13**	**71.95**	**66.31**	**69.49**	**99.36**

however, the IoU_{weed} is much lower at 75.26. PQ_{crop} was 78.06 and PQ_{leaf} was lower at 68.6. Compared to crops, leaves are harder to segment and this could be becuase they smaller and more prone to occlusion. Additionally, weeds, compared with crops, present a greater degree of intra-class variation as well as often being smaller, making them more challenging to segment. The overall $PQ^{\dagger}$ for both models exceeded existing baselines in [15].

The Swin models show a clear advantage compared to ResNet-50 but there only a small difference of around 1% between Swin S and Swin L. Given the hardware requirements to run the larger Swin model, and the small difference in accuracy, the Swin S model might be more appropriate for some applications.

While skip connections do, in some cases, seem to show an improvement, the improvement is not reliable. It might be that there is a more optimal placement of these connections or that they are simply not required.

During inference, mask are only predicted the associated confidence is higher than a given threshold. We found that reducing the mask threshold to 0.5 from 0.8 generally improved performance.

7. CONCLUSIONS

Overall, this work demonstrates the potential of architectures like Mask2Former to solve the visual recognition challenges within precision agriculture. In future work, we would like to investigate how to reduce the size of the proposed model while retaining its accuracy.

References

[1] CVPPA@ICCV'23: Hierarchical panoptic segmentation of crops and weeds. `https://codalab.lisn.upsaclay.fr/competitions/13904`, 2023.

[2] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12), 2017.

[3] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[4] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-DeepLab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022.

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34, 2021.

[7] Maurilio Di Cicco, Ciro Potena, Giorgio Grisetti, and Alberto Pretto. Automatic model based dataset generation for fast and accurate crop and weeds detection. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[10] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2021.

[11] Kavir Osorio, Andrés Puerto, Cesar Pedraza, David Jamaica, and Leonardo Rodríguez. A deep learning approach for weed detection in lettuce crops using multispectral images. *AgriEngineering*, 2(3), 2020.

[12] Gianmarco Roggiolani, Matteo Sodano, Tiziano Guadagnino, Federico Magistri, Jens Behley, and Cyrill Stachniss. Hierarchical approach for joint semantic, plant instance, and leaf instance segmentation in the agricultural domain. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023.

[13] Eduardo Romera, José M Alvarez, Luis M Bergasa, and Roberto Arroyo. ERFnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 19(1), 2017.

[14] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015.

[15] Jan Weyler, Federico Magistri, Elias Marks, Yue Linn Chong, Matteo Sodano, Gianmarco Roggiolani, Nived Chebrolu, Cyrill Stachniss, and Jens Behley. PhenoBench — A Large Dataset and Benchmarks for Semantic Image Interpretation in the Agricultural Domain. *arXiv preprint*, 2023.

[16] Jan Weyler, Federico Magistri, Peter Seitz, Jens Behley, and Cyrill Stachniss. In-field phenotyping based on crop leaf and plant instance segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022.

[17] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.

[18] Kunlin Zou, Xin Chen, Yonglin Wang, Chunlong Zhang, and Fan Zhang. A modified U-Net with a specific data argumentation method for semantic segmentation of weed images in the field. *Computers and Electronics in Agriculture*, 187, 2021.