Mask2Former@PhenoBench

Ronja Güldenring
Technical University of Denmark
2800 Kongens Lyngby
ronjag@dtu.dk

Lazaros Nalpantidis
Technical University of Denmark
2800 Kongens Lyngby
lanalpa@dtu.dk

Abstract

Mask2Former is a transformer architecture that ranks top scores on computer vision benchmark datasets such as COCO, Cityscapes or ADE20K. In this work, we examine the capabilities of Mask2Former on the agricultural dataset PhenoBench, which contains different image characteristics compared to the above mentioned datasets. For the task of "Hierarchical Panoptic Segmentation of Crops and Weeds", we train three different Mask2Former models on the different granularity levels, leading to a final score of PQ+ = 81.06. The code will be available at `https://github.com/DTU-PAS/phenobench_challenge`.

1. Introduction

This is a technical report documenting our followed approach that resulted in our final submission to the "CVPPA@ICCV'23: Hierarchical Panoptic Segmentation of Crops and Weeds" competition[1]. This competition was part of the 8th Workshop on Computer Vision in Plant Phenotyping and Agriculture (CVPPA) at the IEEE/CVF International Conference of Computer Vision (ICCV) 2023. Our final submission achieved the third overall place in the competition.

2. PhenoBench Dataset [5]

The PhenoBench Dataset is an RGB image dataset of sugar beet plants and weeds, including annotations of different granularity on pixel-level: semantics, plant instances and leaf instances. In Figure 1 a sample image with corresponding annotations is shown. It includes 1407/772/693 (train/val/test) images, recorded with a drone at different days during 2020 and 2021. The annotations of the test split are not published, thus, test evaluation can only be performed through the CodaLab Server.

[1] `https://codalab.lisn.upsaclay.fr/competitions/13904`

3. Proposed Approach

In Figure 2, we illustrate our top-down architecture that led to the final score of the challenge *CVPPA@ICCV'23: Hierarchical Panoptic Segmentation of Crops and Weeds* (i.e. PQ+ = 81.06). We exclusively use the Masked-attention Mask Transformer for Universal Image Segmentation (Mask2Former) [1] as our deep learning architecture, because it achieves high ranks in various computer vision segmentation benchmarks such as COCO [3], Cityscapes [2] and ADE20K [6]. Another advantage of Mask2Former is that they provide integration to HuggingFace[2] with an API and powerful pre-trained weights, which serves as an easy entry-point.

We divide the architecture in two stages as further explained in the following subsections. Note, that the second stage expects the predictions of the semantics and plant instances as input.

3.1. Panoptic Plant Segmentation

A Mask2Former with a large SwinTransformer [4] as backbone is trained to perform a panoptic segmentation on the original input image at full size, i.e. 1024×1024. We observed that the mask shape of small plants – crops as well as weeds – has a strong negative impact on the final score. Therefore, a refinement of the small masks in the initial predictions is applied. All masks that have a pixel size < 2500 and are isolated, i.e. not connected to neighboring instances, are cut out and will be processed by another Mask2Former with a base SwinTransformer, that is specifically trained for this task. All small mask shapes will be updated in both the semantic and plant instance map.

3.2. Leaf Instance Segmentation

For the segmentation of leaf instances, yet another Mask2Former model is trained on cut outs of plant instances, where the background (i.e. soil, weeds, and other crop instances) is blacked out. By blacking out the back-

[2] `https://huggingface.co/docs/transformers/main/model_doc/mask2former`

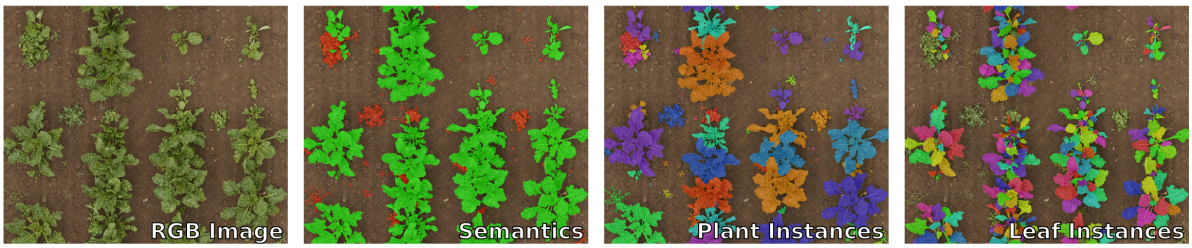

Figure 1. PhenoBench sample image with corresponding hierarchical annotations. Image is taken from [5].

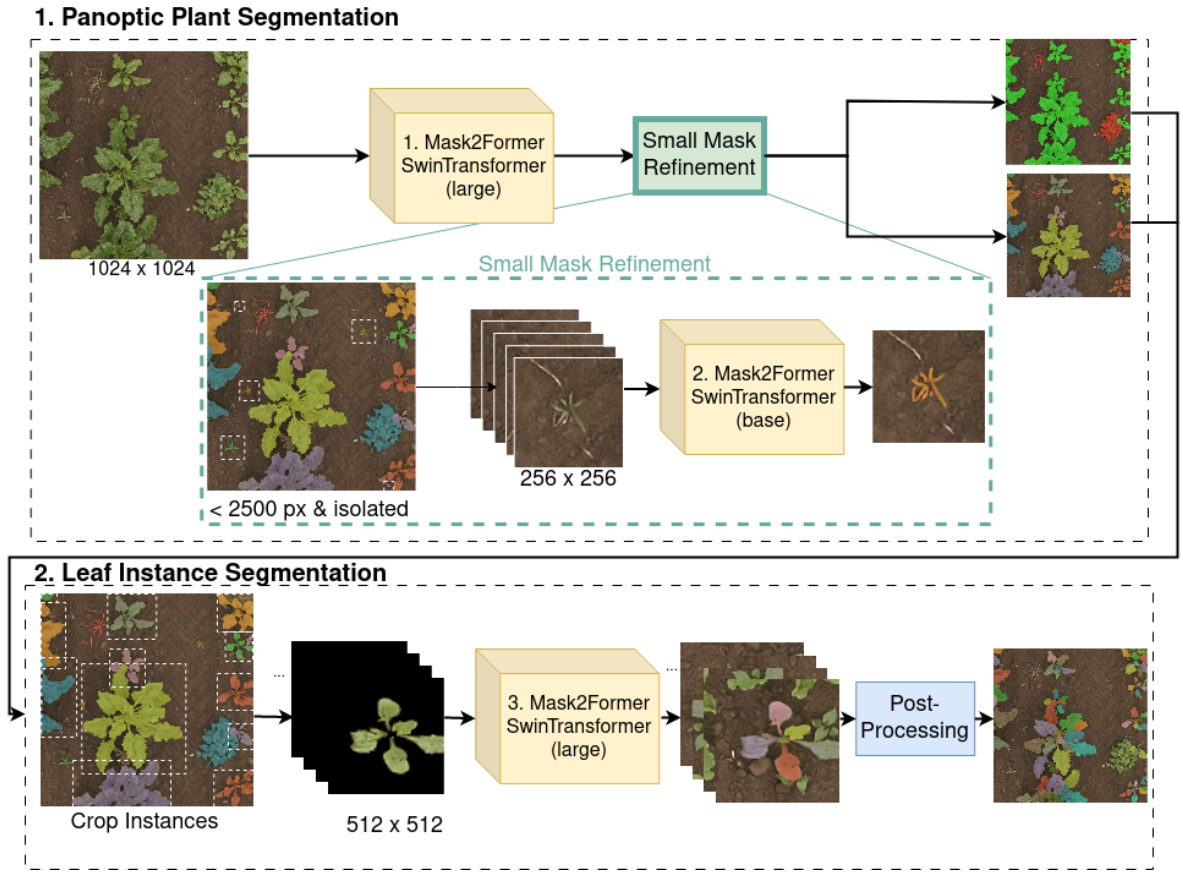

Figure 2. Our top-down architecture contains three separately trained Mask2Former models. *Panoptic Plant Segmenation*: In the first stage, the first Mask2Former model performs panoptic segmentation on the original input image, followed by a mask refinement of small and isolated mask shapes, i.e. of a pixel size < 2500. *Leaf Instance Segmenation*: The semantic and plant instance map serve as input for the second stage, where each plant instance is processed individually by the third Mask2Former model. Note that pixels not belonging to the corresponding instances are blacked out. Finally, all leaf instances are merged and post-processed in order to retrieve the final leaf instance map.

ground, the network can focus on simply one crop at a time and doesn't need to deal with plant overlaps. Finally, we post-process the leaf instance map with some basic filtering in order to remove clutter: (1) for non-connecting instances, we keep the largest component and (2) all leaf instances with a mask size smaller than 10% of the average leaf size of the corresponding plant are removed.

During inference, the semantic and plant instance map from the previous stage is required as input.

4. Experimental Evaluation

In this section, we specify training settings of our final architecture in Seciont 4.1. Furthermore we highlight some intermediate submissions in order to show the path from the baseline of Weyler et al. [5] to the final score in submission 4.

4.1. Training Settings

We used the Mask2Formers integration in HuggingFaces with Pytorch Lightning in order to increase flexibility during model tuning. Our models are trained on a single Nvidia V100. In order to retain a reasonable batch size, while having a large model and large input image, we use accumulative gradients. For all three Mask2Former models, the majority of hyperparameters is identical. In the following, we will summarize the training settings, while deviating parameters are listed in Table 1. The model takes an input of *image size* (see Table 1), which is augmented with the following augmentation methods and probabilies from albumentations[3]: RandomResizedCrop (1.0), HorizontalFlip (0.5), VerticalFlip (0.5), RGBShift (1.0), RandomBrightnessContrast (1.0), Blur (*p_blur*), and Rotate (0.4). Furthermore, the images are normalized. For the first Mask2Former model, we use the normalizing parameters of PhenoBench (mean: [124.574, 97.684, 65.202], std: [51.273, 45.172, 34.660]). Due to a copy-paste mistake, for the other two Mask2Former models, we used the ImageNet normalizing parameters (mean: [123.675, 116.280, 103.530], std: [58.395, 57.120, 57.375]). Each model is initialized with the *pretrained weights* and is trained for *num_steps* timesteps with a batch size of *batch_size*, as defined in Table 1. We use AdamW as optimizer with a learning rate of *lr* and $0.1lr$ for the SwinTransformer backbone. For all models, we use a Polynomial learning rate scheduler with a final learning rate of 0. Furthermore, we use exponential moving average during parameter update with a decay of 0.9998. Like in the Detectron implementation of Mask2Former, we apply gradient clipping at 0.01.

4.2. Results

In the course of the challenge, we have submitted 5 times to the CodaLab server. In Table 2, we want to summarize some of our submissions and shortly highlight the design choices. We also show the Mask2Former scores from the baseline paper by Weyler et al. [5].

Submission 0 In this submission, we applied the official baseline code from phenobench[4], which uses the Detectron integration of Mask2Former. First, we reproduced the result reported by Weyler et al. [5], who apply ImageNet pretrained weights. Secondly, we load the COCO pretrained weights provided by Mask2Former, which are specialized on the targeted segmentation task. This leads to increased detection performance.

Submission 1 In this submission, we embed the Mask2Former integration from HuggingFaces with PytorchLightning. The main reason for the code shift is our limited familiarity with the Detectron framework and the fact that it seems outdated. Having the code in PytorchLightning increases flexibility during model finetuning. We could easily try different optimizers, learning rate schedulers, but also quickly add effective training techniques such as exponential moving average, accumulate gradients and gradient clipping. Furthermore, we added more complex augmentation methods like RandomResizedCrop, HorizontalFlip, VerticalFlip, RGBShift, RandomBrightnessContrast and Blur. Additionally, we normalized the images according to the training split of the PhenoBench dataset. Note, that in this submission, we simply trained two Mask2Former models independently end-to-end with the original input image, one for the panoptic plant segmentation and one for the leaf instance segmentation.

Submission 4 This is the final submission, that was explained in detail in Section 3.

5. Conclusion

We presented our approach, that led to a final score of $PQ+ = 81.06$ in the challenge *CVPPA@ICCV'23: Hierarchical Panoptic Segmentation of Crops and Weeds*. The core of our approach is the state-of-the-art model Mask2Former, which already showed great strength on other segmentation benchmark dataset. We tune two different independent Mask2Former models for each stage of the segmentation hierarchy. Furthermore, we apply minor additional improvements such as *Small Mask Refinement* and post-processing of leaf instances.

Our personal takeaway from this challenge is in accordance with recent advances in the domain of Deep Learning: bigger models, longer training, higher input resolution as well as good pre-trained weights have most impact on the final score. However, available computing capacity is different between participants and certainly a decisive ingredient of how to perform in such challenges.

References

[1] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. 2022.

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. *CoRR*, abs/1604.01685, 2016.

[3] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Doll'a r, and C. Lawrence Zitnick. Microsoft

[3] https://albumentations.ai/

[4] https://github.com/PRBonn/phenobench-baselines

Mask2Former ID	pretrained weights	image size	num_steps	batch_size	lr	p_blur	normalize on
1	swin-large-coco-panoptic	1024	18000	16	0.0001	0.4	PhenoBench
2	swin-base-IN21k-ade-semantic	256	10000	16	0.0005	0.0	ImageNet
3	swin-large-coco-panoptic	512	12000	24	0.0001	0.3	ImageNet

Table 1. Hyperparameters that deviate between the three Mask2Former models.

ID	PQ+	PQ	PQ(crop)	PQ(leaf)	IoU(weed)	IoU(soil)
Weyler et al. [5]	-	64.36	71.21	57.50	-	98.38
Submission 0	74.51	67.56	73.36	61.76	64.33	98.6
Submission 1	78.38	72.36	77.75	66.97	69.62	99.18
Submission 4	**81.06**	**77.4**	**81.82**	**72.98**	**70.1**	**99.35**

Table 2. Our scores are increasing with each submission. The architecture of the final submission is described in detail in Section 3. We also show the results from the baseline paper of Weyler et al. [5]

COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

[4] 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 (ICCV)*, 2021.

[5] 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.

[6] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *CoRR*, abs/1608.05442, 2016.