# Parallel Transformer Decoders for Semantic Image Interpretation in the Agricultural Domain

Xiaoqiang Lu, Licheng Jiao*, Lingling Li, Zhongjian Huang,
Yuting Yang, Jiaxuan Zhao, Xu Liu, Fang Liu
Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education
School of Artificial Intelligence, Xidian University
xqlu@stu.xidian.edu.cn, lchjiao@mail.xidian.edu.cn

## Abstract

*Deep learning based methods have shown powerful performance in the area of computer vision tasks. However, previous works focus on human daily life to derive a convenience service, not on the agricultural domain. To this end, the ICCV 2023 CVPPA challenge establishes a large dataset and benchmarks for the semantic interpretation of images of real agricultural fields. Specifically, one has to predict a semantic segmentation of sugar beets and weeds, but also an instance segmentation of crops and an instance segmentation of the leaves. In the report, we introduce our method used in the challenge, achieving the best test set performance. Mask2former, one state-of-the-art universal network for semantic interpretation, is adopted as our baseline model. We make two copies of the original transformer decoder to provide a simultaneous instance segmentation of the whole object and each part. During training, we use rotation augmentation to perform offline data expansion due to the limitation of the labeled data, and we add random resize and random crop to perform online data augmentation to improve the generalization of the model. Furthermore, we design different test-time augmentation for different visual prediction results. Finally, our method achieves a PQ+ of 82.62, surpassing others by a large margin.*

## 1. Introduction

In the coming decades, the agricultural production of food, feed, fiber, and fuel will face a number of difficulties. In order to overcome these difficulties, vision-based perception systems on drones could give tools for making judgments about field management that are better and more sustainable as well as support tools for breeding new types of crops by accurately predicting plant attributes. To this end, the ICCV 2023 CVPPA challenge presents a large dataset for plant segmentation providing accurate instance annotations at the level of plants and leaves [7]. Specifically, one has to predict a semantic segmentation of sugar beets and weeds, but also instance segmentation of crops and leaves.

To solve the challenge, we introduce data-level, model-level, and prediction-level strategies to address the limitation of the labeled data, the ability to get phenotypic information of the whole crop but also more fine-grained information, and better performance improvements, respectively. We first use the rotation of 90 degrees, 180 degrees, and 270 degrees for offline data expansion, deriving four times training data. Then, we adopt random resize, random crop, and random horizontal flip for online data augmentation. For the basic backbone, we adopt an adapter proposed in [1] to inject the image-related inductive bias into the model BEiTv2-Large [5], which can compensate for the shortcomings of the vision transformer and achieve comparable performance to vision-specific models. Mask2former, a universal image segmentation architecture that outperforms specialized architectures across different segmentation tasks, is employed as our baseline model. Additionally, we add a parallel transformer decoder to perform an instance segmentation of the leaf, while the original transformer decoder is utilized to make an instance segmentation of the whole crop. We use the COCO-trained [3] weight to initial our model, improving the basic performance while reducing the training time. The additional leaf transformer decoder uses the copied pre-trained weight to initial, the same as the crop transformer decoder. During inference, we design various test-time augmentation techniques for various visual results. Hard voting, following the rule of majority rule, is used to merge the prediction of semantic segmentation. Large-size testing, using 1.25 times the original training size to derive a better panoptic segmentation of the whole crop. Weighted boxes fusion (WBF) [6] and mask voting based on multi-scale and flipping inference is utilized in the panoptic segmentation of the leaf.

*corresponding author

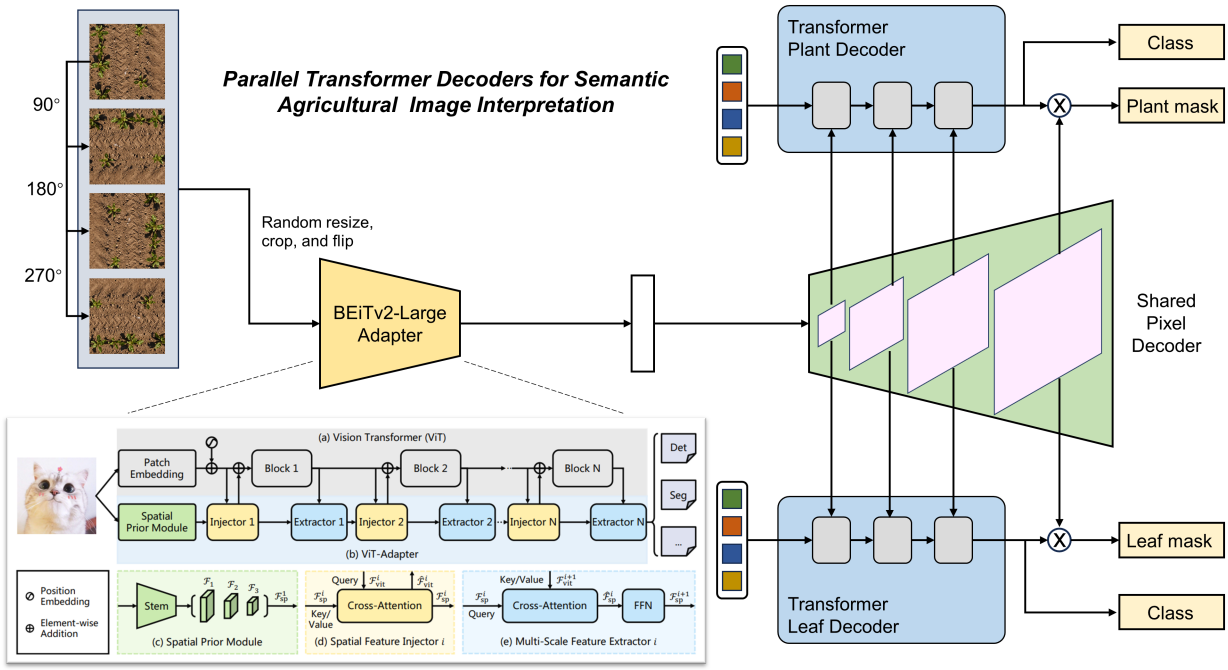


Figure 1. Overview of our method used in the ICCV 2023 CVPPA Challenge-Hierarchical Panoptic Segmentation of Crops and Weeds.

## 2. Method

The overview of our method used in the ICCV 2023 CVPPA Challenge-Hierarchical Panoptic Segmentation of Crops and Weeds is shown in Figure 1.

### 2.1. Data Augmentation

**Offline data augmentation**. The original training set of PhenoBench[7] has 1407 images, and the validation set has 772 images, which is limited. To expand the training data, we first randomly sample 385 images from the validation set to add to the training set, which now contains 1792 images. Then, we employ the rotation of 90 degrees, 180 degrees, and 270 degrees to further expand the expanded training set, deriving 4 times training data as our final training set which has 7168 images. And the rest of the validation set contains 387 images, used to select well-trained weights.

**Online data augmentation**. To improve the generalization of the model, we also introduce random resize between 0.5 and 2.0, random crop to the original training size, and random horizontal flipping with a probability of 0.5. The online data augmentation can also help the performance of multi-scale and flip testing.

### 2.2. Model Architecture

Figure 2 shows the original Mask2former [2] architecture, which consists of a backbone, a pixel decoder, and a transformer decoder. We replace its originally used backbone Swin-Large [4] with BEiTv2-Large[5] with a visual adapter[1], to pursue more representative low-resolution image features. The pixel decoder gradually upsamples low-resolution features from the output of the backbone to generate high-resolution per-pixel embeddings with resolution 1/32, 1/16, and 1/8 of the original image, using the feature pyramid. At the same time, a sinusoidal positional embedding and a learnable scale-level embedding are added for each resolution. The transformer decoder includes a masked attention operator, which extracts localized features by constraining cross attention within the foreground region of the predicted mask for each query, instead of attending to the full feature map.

To make the original Mask2former have the capacity to get a simultaneous instance segmentation of the whole object and each part, we add an additional parallel transformer decoder which is the same as the original one. Then, one of the double transformer decoders is used to perform the whole crop panoptic segmentation, and the else is employed for the leaf panoptic segmentation. Both of them are individual while sharing the same backbone and the same pixel decoder, since the backbone and the pixel decoder are used to extract and fine image features. Before training, we initialize our model using the COCO-trained [3] weight prior, which enhances the fundamental performance while shortening the training period. Similar to the crop transformer decoder, the additional leaf transformer decoder initializes using the copied pre-trained weight. The total loss is the

Table 1. Final performance on the test leaderboard.

| Rank | User | PQ+ | PQ | PQ(crop) | PQ(leaf) | IoU(weed) | IoU(soil) |
|---|---|---|---|---|---|---|---|
| **1** | **IPIU-XDU** | **82.62** | **78.45** | **82.04** | **74.86** | **74.13** | **99.44** |
| 2 | chgiang | 81.33 | 77.73 | 81.66 | 73.81 | 70.66 | 99.18 |
| 3 | RGueldenring | 81.06 | 77.4 | 81.82 | 72.98 | 70.1 | 99.35 |

average of the losses of the whole crop and the leaf.

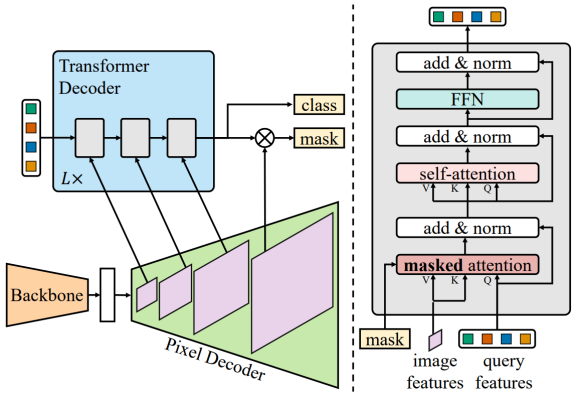


Figure 2. The overall of the original Mask2former [2].

### 2.3. Test-time Augmentation

Before introducing the test-time augmentation (TTA), we observe that the performance of directly using the panoptic segmentation output is worse than converting the instance segmentation output to the panoptic segmentation output. Thus, we design a post-precess based on the confidence threshold $\delta_{conf}$ and the Intersection-over-Union (IoU) threshold $\delta_{iou}$. Specifically, for the whole crop instance segmentation output, we first remove the same predicted object in the same image, and then we sort the object-level predictions according to their box confidence in descending order. After that, we filter the boxes whose confidence scores are lower than $\delta_{conf} = 0.8$. Finally, the IoU between the highest score box and the remaining boxes is calculated from the highest score box downwards. If the value is less than $\delta_{iou} = 0.8$, the box with a low score is deleted. Now, the conversion from instance segmentation output to panoptic segmentation output is completed by simply setting the corresponding mask value to a unique ID according to the sort index of the box. For the conversion of leaf, $\delta_{conf}$ and $\delta_{iou}$ are set to $0.4$ and $0.8$ respectively.

For different tasks of semantic segmentation, panoptic segmentation of the whole crop, and panoptic segmentation of the leaf, we designed three TTA strategies as below.

**Hard voting**. We first use nine scales of x1.0, x1.25, x1.5, x1.75, x2.0, x2.25, x2.5, x2.75, x3.0 times the training size to predict, and then re-scale the results to normal size. Finally, following the principle that the minority is subordinate to the majority, we merge the nine predictions using hard voting.

**Large-size testing**. For the whole crop instance segmentation, we find that only using x1.25 times the training size to predict performs better than using other scales to predict or merge the results based on weighted boxes fusion (WBF) and mask voting.

**WBF and mask voting**. Multi-scale testing and horizontal flipping are two common TTAs to boost the improvement of the model during inference. Benefiting from the online data augmentation, we gradually increase the inference size and find that the model performance continues to improve. We perform 5 scales of x1.0, x1.125, x1.25, x1.375, and x1.5 times the training size to predict. In addition, we apply an additional horizontal flip to these 5 scales for prediction, and the resulting 10 predictions are used in the subsequent WBF and mask voting. During WBF, we additionally average the binary mask of prediction boxes participating in the weighted fusion. Now that the fused binary mask has been normalized to 0 to 1, we set a mask threshold $\delta_{mask} = 0.25$ to export the final binary mask.

## 3. Experimental Results

### 3.1. Implementation Details

MMdetection is employed to implement the proposed method. All experiments are conducted on 8 NVIDIA V100. The model is optimized by AdamW with a base learning rate of $0.0001$, weight decay of $0.05$, and batch size of $8$. The training image base size is set to 1280x1280. $1\times$ standard training schedule is employed. other details please refer to Sec 2.

### 3.2. Experimental results

The final submissions on the test leaderboard are shown in Table 1, our method achieves the best performance with PQ+ $= 82.62$, surpassing others by a large margin. Table 2 shows the effectiveness of each component of our method, indicating that the proposed strategies can benefit from shared information of multi-task learning.

## References

[1] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

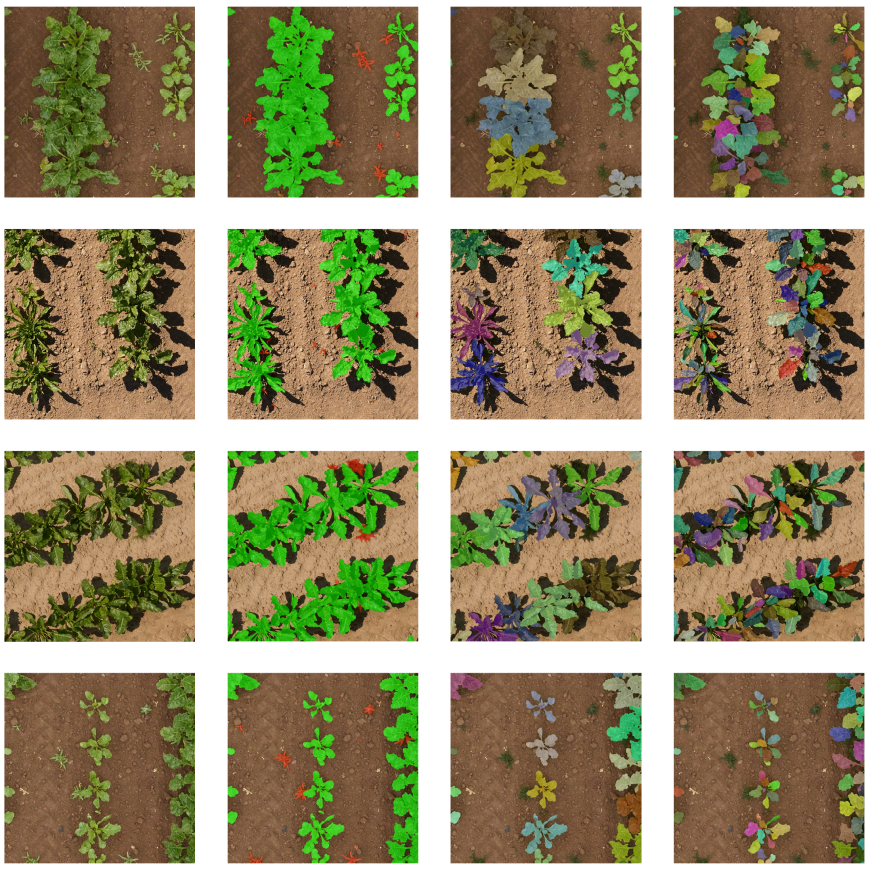

Figure 3. Visual results of our method in the test set. From left to right are the image, semantic segmentation results, plant panoptic segmentation results, and leaf panoptic segmentation results.

Table 2. Ablation study on various components. PTD: parallel transformer decoder. DA: data augmentation. TTA: test-time augmentation.

| Method | Val | | | | | Test | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PQ+ | PQ(crop) | PQ(leaf) | IoU(weed) | IoU(soil) | PQ+ | PQ(crop) | PQ(leaf) | IoU(weed) | IoU(soil) |
| Mask2former+PTD | 82.90 | 81.76 | 71.73 | 78.58 | 99.51 | - | - | - | - | - |
| +Online DA | 83.20 | 82.44 | 72.11 | 78.73 | 99.52 | - | - | - | - | - |
| +TTA | 84.18 | 82.75 | 75.28 | 79.15 | 99.54 | - | - | - | - | - |
| +Offline DA, TTA | - | - | - | - | - | **82.62** | **82.04** | **74.86** | **74.13** | **99.44** |


[2] 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*, pages 1290–1299, 2022.

[3] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 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*, pages 10012–10022, 2021.

[5] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv preprint arXiv:2208.06366*, 2022.

[6] Roman Solovyev, Weimin Wang, and Tatiana Gabruseva. Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107:104117, 2021.

[7] 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 arXiv:2306.04557*, 2023.