WHEN SEGMENT ANYTHING MODEL MEETS FOOD INSTANCE SEGMENTATION APPENDIX

Hanyu Jiang *Xing Lan* *Jiayi Lyu* *Kun Dong* *Jian Xue*

University of Chinese Academy of Sciences

A. Experiment details

We conduct experiments on FoodInsSeg using four GeForce RTX 3090Ti GPUs. Model training and evaluation are implemented through the MMDetection framework [1], an open-source toolbox for various computer vision tasks. The backbone networks are initialized with weights pre-trained on ImageNet. For optimization, We use the Adam optimizer with a base learning rate of 0.0001 and a multi-step learning rate scheduler, decaying the learning rate twice. Most models were trained for 50 epochs, decaying at 30 and 40 epochs, while Transformer-based architectures are trained for 100 epochs, decaying at 70 and 90 epochs to account for slower convergence. All experiments use a batch size of 8 and default model hyperparameters provided in MMDetection. In addition, for FoodSAM, we obtain results using their official code without training.

B. Qualitative results

In Fig. 1, we visualize segmentation results from mask2former with the r101 backbone. It can be observed that after training with FoodInsSeg, mask2former exhibits powerful instance segmentation performance on food ingredients. It accurately identifies and segments most of the food ingredients in the images with clear boundaries. This is attributed to the advanced segmentation capabilities of Mask2former and the precision of the ground truth annotation in our dataset. It's worth noting that when the scene contains few ingredients and simple layouts, the model achieves excellent segmentation (first two columns). However, performance slightly decreases for crowded images(last three columns) with more ingredient objects. This indicates challenges remain for complex images in ingredient-level instance segmentation.

C. InsSAM-Tool in different domains

Our InsSAM-Tool exhibits impressive generalization performance across different domains. As shown in Fig. 2, InsSAM-Tool demonstrates powerful annotation capability of diverse scenes, including wild objects [2], remote sensing [3] and food computing [4].

D. Data distribution

Following the practice of FoodSeg103 [4], we group all 103 food ingredient categories into 14 superclass categories, such as Main, Vegetable, and Fruit. We examine the distribution of these superclass categories within our FoodInsSeg. As illustrated in Fig. 3, instance masks of vegetables are the most prevalent.

1. REFERENCES

[1] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[2] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari, "Coco-stuff: Thing and stuff classes in context," 2018.

[3] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong, "Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung, Eds. 2021, vol. 1, Curran.

[4] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun, "A large-scale benchmark for food image segmentation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 506–515.

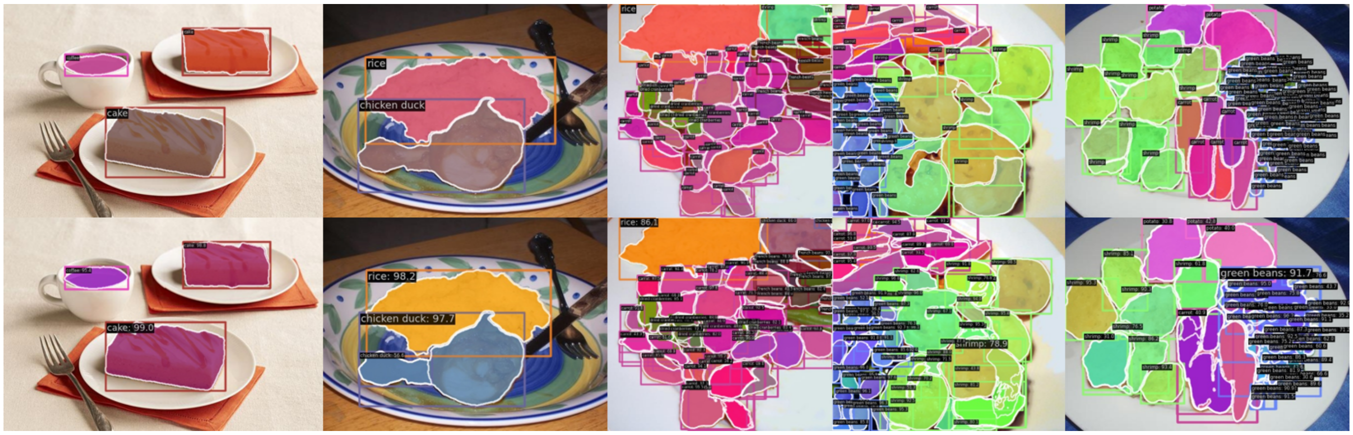
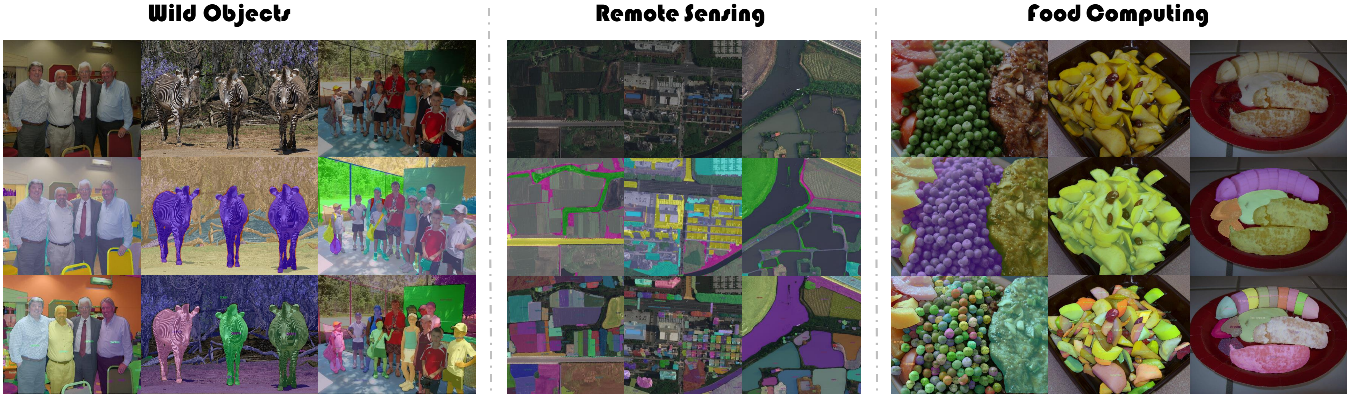


Fig. 1. Examples of prediction masks and annotation masks. The top row displays examples of ground truth annotation. The bottom row shows examples of Mask2former(r101). The two left columns show food image segmentation results on simple scenarios, while the three right columns present more complex food scene segmentation.


Wild Objects

Remote Sensing

Food Computing

Fig. 2. Example annotations of instance segmentation by InsSAM-Tool. The top row shows original images, the middle row displays semantic annotations provided by semantic segmentation datasets, and the bottom row shows instance annotations by our tool. InsSAM-Tool annotates a class-specific mask for each instance, whether it's person, agriculture or green beans, which performs impressive generalization on labeling instance segmentation datasets across different domains, i.e. wild objects [2], remote sensing [3] and food computing [4].

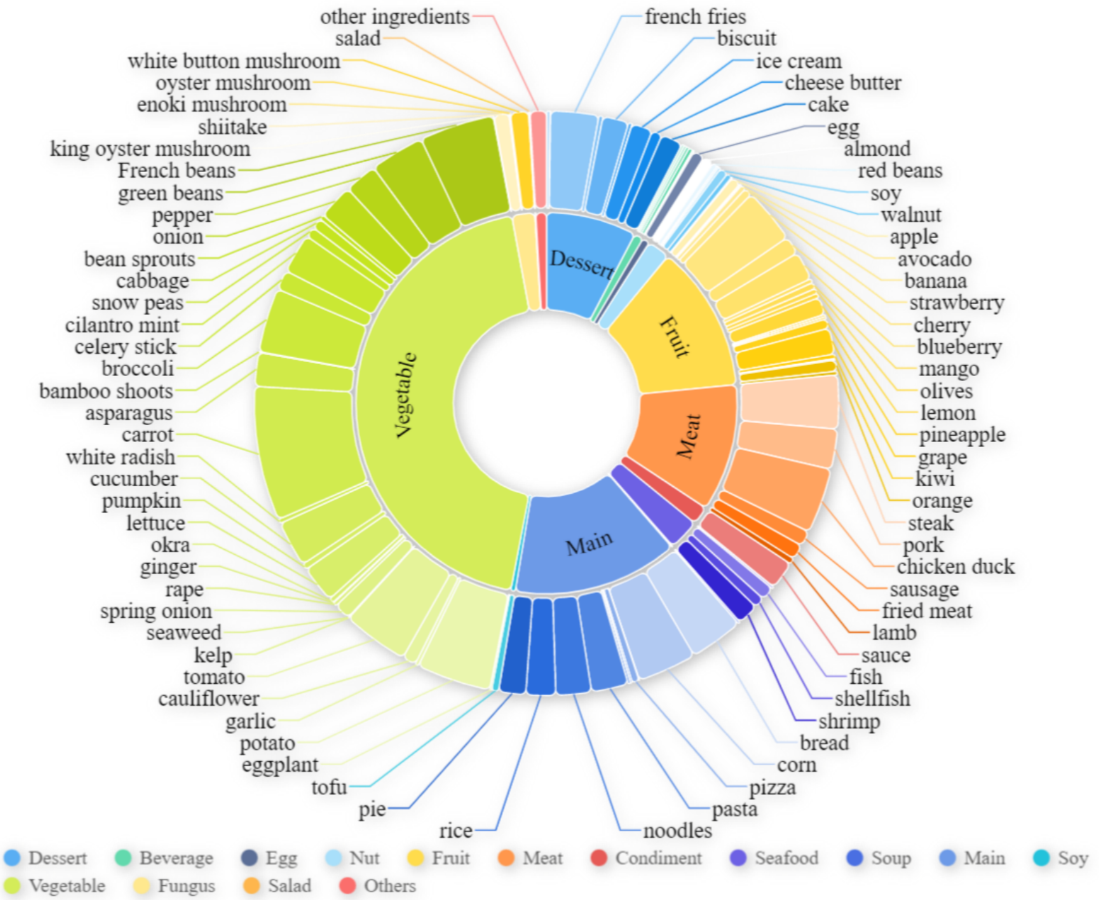


Fig. 3. Number of masks per ingredient category. Ingredients superclass Vegetable has the highest number of occurrences.