Supplementary Materials of *OVFoodSeg: Elevating Open-Vocabulary Food Image Segmentation via Image-Informed Textual Representation*

The implementation details of Stage I and Stage II training are presented in Section A. We analyze the failure cases of OVFoodSeg in Section B, and explore the full class training results in Section C. Finally the set of novel classes of these class splits are listed in Section D.

A. Implementation Details

Stage I: We employ the CLIP ViT-L/14 model [1, 3] as the CLIP image encoder, and initialize FoodLearner with weights of qformer pre-trained via BLIP2 [2]. We initialize Q query tokens (Q set 32 in this paper), each with a channel dimension of 768. The model undergoes training for 10 epochs by AdamW optimizer, processing images at a resolution of 224×224 pixels within batches of 100. We utilize cosine learning rate schedules, starting with an initial learning rate of 1e-4 and decaying to an ending rate of 1e-5, coupled with a weight decay set at 0.05. A linear warmup strategy is applied during the initial 5000 iterations, starting at a learning rate of 1e-6.

Stage II: We utilize the CLIP ViT-L/14 model for both image and text encoding, initializing the FoodLearner and query tokens with weights pre-trained from Stage I. Ingredient class names are prompted with the template "`A Photo of {}`" to produce text tokens. The input image resolution is set to 640×640 to generate visual embeddings. For SAN's classification and mask prediction branches, we set the input image resolution as 640×640 and 320×320 respectively. For FoodSeg103, the model undergoes training for 10,000 iterations using the AdamW optimizer, with a batch size of 8. A Poly learning rate schedule is employed, beginning with an initial learning rate of 1e-4, coupled with a weight decay set at 0.0001. For FoodSeg195, the model is trained over 20,000 iterations, maintaining consistency with the other settings used for FoodSeg103. For the remaining implementation details, we adhere to the configurations established in SAN [4].

B. Failure Case Analysis

In this section, we analyze the failure cases of OVFoodSeg, particularly focusing on the classes that perform worse than the baseline SAN model. Among the 20 novel classes of FoodSeg103 Split 1, 3 classes perform worse than the baseline. Notably, the `white button mushroom` class shows the most significant underperformance, with its results being approximately 8.9% lower than the baseline (2.2% vs 11.1%). We visualize the prediction results of OVFoodSeg in Figure 1. The figure illustrates that OVFoodSeg mistakenly classifies the novel class `white button mushroom` into the base class `shiitake`, which bears a high visual resemblance to the target novel class. The confusion between novel and base classes poses a significant challenge in the open-vocabulary setting, and tackling this issue is a key focus for our future research.

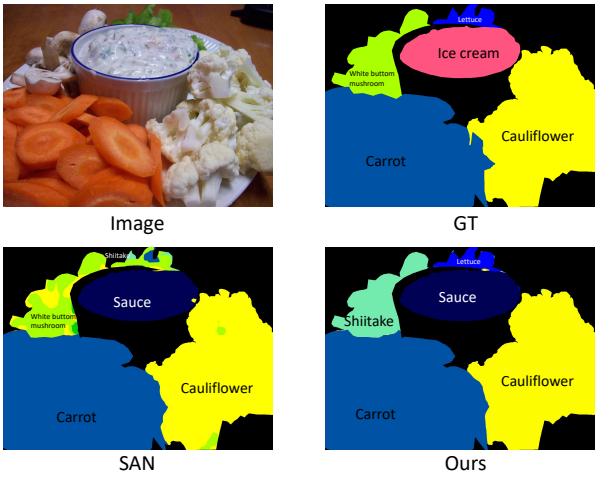

Figure 1. Failure cases of OVFoodSeg on FoodSeg103 (Split 1) where GT means ground-truth. In this example, OVFoodSeg incorrectly classified "white button mushroom", a novel class, as "shiitake", which is a base class.

C. Full Class Training

In this section, we compare OVFoodSeg and SAN trained with both novel and base classes of FoodSeg103. For a fair comparison, we use FoodLearner without Stage I pretraining (OVFoodSeg*). From Table 1, OVFoodSeg* significantly outperforms SAN in the full class training mode,

further proving the effectiveness of our proposed image-informed textual representation mechanism.

	mIoU	mAcc	pAcc
SAN	41.6	56.2	67.1
OVFoodSeg*	45.4	60.4	71.0

Table 1. Performance comparison of OVFoodSeg* and SAN trained with both novel and base classes on FoodSeg103. Here, mIoU, mAcc and pAcc denote mean IoU, mean accuracy and pixel-wise accuracy respectively.

D. Novel Classes in Multiple Class Splits

In this section, we detail the novel ingredient classes utilized in each class split.

D.1. FoodSeg103

Split 1:

candy, french fries, ice cream, wine, coffee, cashew, pineapple, sausage, lamb, crab, pie, seaweed, lettuce, pumpkin, bamboo shoots, celery stick, cilantro mint, cabbage, bean sprouts, white button mushroom

Split 2:

egg tart, coffee, date, blueberry, raspberry, kiwi, chicken or duck, soup, bread, hanamaki baozi, eggplant, kelp, seaweed, ginger, carrot, asparagus, cabbage, onion, green beans, salad

Split 3:

french fries, cake, juice, red beans, apricot, raspberry, melon, watermelon, shellfish, shrimp, kelp, seaweed, spring onion, okra, carrot, cilantro mint, snow peas, king oyster mushroom, shiitake, white button mushroom

D.2. FoodSeg195

Split 1:

popcorn, cheese, cake, milk, date, avocado, raspberries, lemon, pineapple, grape, melon, steak, pork, sausage, bread, pizza, pasta, rice, garlic, kelp, broccoli, celerystick, cilantro mint, pork belly, edamame, ketchup, fish cake, fish balls, rice cake, lotus root, daylily, durian, thosai, tangyuen, idli, spaghetti, nai bai, kangkong, yam, beef

Split 2:

blueberry, melon, steak, shrimp, baozi, pasta, noodle, pie, tomato, ginger, okra, lettuce, others, wolfberry, pig blood curd, meat skewer, meatballs, edamame, curry sauce, salad sauce, garlic sauce, porridge, amaranth, honey dews, papaya, beehoon, lasagna, macaroni, puri, oyster, celery, sweet potato, beef

Split 3:

kiwi, orange, fried meat, baozi, garlic, spring onion, ginger, lettuce, white radish, asparagus, bamboo shoots, bean sprouts, oyster mushroom, beef ribs, minced beef, pork belly, pork skin, pork liver, shredded pork, meat skewer, edamame, barbecued pork sauce, fish tofu, fried banana leaves, bitter gourd, agaric, burger buns, guava, mochi, bun, mussel, celery, beef

References

[1] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[4] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, 2023. 1