OVFoodSeg: Elevating Open-Vocabulary Food Image Segmentation via Image-Informed Textual Representation

Xiongwei Wu Sicheng Yu Ee-Peng Lim Chong-Wah Ngo
Singapore Management University
{xwwu.2015, scyu.2018}@phdcs.smu.edu.sg, {eplim, cwngo}@smu.edu.sg

Abstract

In the realm of food computing, segmenting ingredients from images poses substantial challenges due to the large intra-class variance among the same ingredients, the emergence of new ingredients, and the high annotation costs associated with large food segmentation datasets. Existing approaches primarily utilize a closed-vocabulary and static text embeddings setting. These methods often fall short in effectively handling the ingredients, particularly new and diverse ones. In response to these limitations, we introduce OVFoodSeg, a framework that adopts an open-vocabulary setting and enhances text embeddings with visual context. By integrating vision-language models (VLMs), our approach enriches text embedding with image-specific information through two innovative modules, e.g.*, an image-to-text learner FoodLearner and an Image-Informed Text Encoder. The training process of OVFoodSeg is divided into two stages: the pre-training of FoodLearner and the subsequent learning phase for segmentation. The pre-training phase equips FoodLearner with the capability to align visual information with corresponding textual representations that are specifically related to food, while the second phase adapts both the FoodLearner and the Image-Informed Text Encoder for the segmentation task. By addressing the deficiencies of previous models, OVFoodSeg demonstrates a significant improvement, achieving an 4.9% increase in mean Intersection over Union (mIoU) on the FoodSeg103 dataset, setting a new milestone for food image segmentation.*

1. Introduction

Food computing has garnered significant attention as a research focus due to its widespread application [1, 5, 20, 28]: it integrates computer science with the analysis of food and dietary patterns that have a direct impact on people's healthy eating habits [21]. At the core of food computing is the task of food image segmentation [8, 24, 31]. It aims to identify the ingredients in a food image and locate them with pixel-level masks. This task can be easily found in many practical applications, *e.g.*, analysis of caloric and nutritional content of food intake [22, 23], hence, making it very important in food computing research.

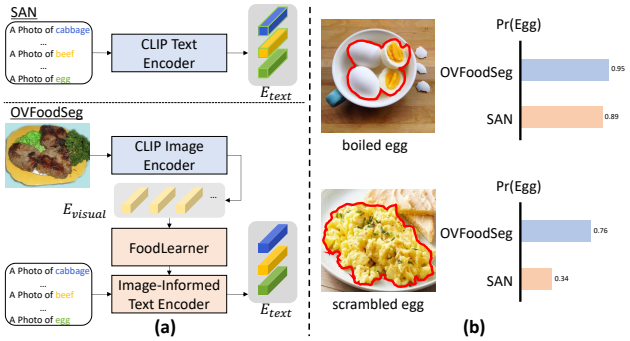

Figure 1. (a) Top: Conventional open-vocabulary segmentation framework Side Adaptive Network (SAN) [34], predicts mask category logits by raw text embeddings from CLIP; Bottom: Proposed OVFoodSeg, constructs image-informed text embeddings through FoodLearner and the Image-Informed Text Encoder for mask classification. (b) This example illustrates the use of both SAN and the proposed OVFoodSeg in identifying egg masks cooked using different methods.

The standard food image segmentation follows a pipeline [31] that begins with the annotation of a large image dataset with pixel-wise masks followed by the training of image segmentation models [3, 35]. However, the pixel mask annotation of images is not only time consuming but also complex. The complexity is further exacerbated in the domain of food as the ingredients can look very small in food images, and they may interact with one another to form obscuring boundaries. While publicly available annotated image datasets for semantic segmentation exist [8, 20], there is only one ingredient-level food segmentation dataset [31] that provides the human annotated masks of 103 ingredient labels across 7,000 images. Even such a small dataset took one year to construct. The set of 103 ingredient labels is also too small to cover all the ingre-

dients that are used in today's food. To tackle the limited ingredient coverage among annotated food image datasets, we propose to address food image segmentation in the open-vocabulary setting. Under this setting, our goal is to learn a segmentation model that adapts to novel ingredients which are not found in the training data without compromising much accuracy performance.

Vision-language models (VLMs) which aligns both vision and language modalities demonstrate remarkable capabilities as a feature extractor in zero-shot classification [26] and open-vocabulary detection tasks [7, 10, 15]. One such example VLMs is CLIP [26] which learns to encode both image and text embeddings via contrastive loss. For image segmentation, CLIP first generates the text embeddings from the ingredient text describing the food image (top of Figure 1(a)) and then contrast with the corresponding image representation. While such methods achieve success with general domain images, it fails to produce similar gains due to intra-class variance of food ingredients: the same ingredient with different cooking methods may look very different visually, *e.g.*, the "egg" class shows considerable variations in appearance between boiled egg and scrambled egg (see Figure 1(b)). The intra-class variance is further exacerbated by unseen classes, where the model lacks prior exposure to these ingredients and the many ways they are prepared and cooked.

To address the aforementioned challenges, we introduce an Open-Vocabulary Food Segmentation framework, named **OVFoodSeg**. The framework effectively integrates the capabilities of CLIP with the image-to-text *FoodLearner*, and replaces CLIP's original fixed text encoder with the proposed *Image-Informed Text Encoder* (see bottom of Figure 1(a)). By harnessing cross-modality capabilities of CLIP, OVFoodSeg effectively transfers knowledge from seen ingredients to novel ingredients. FoodLearner is proposed to mitigate the high intra-class variance for ingredients. Specifically, FoodLearner is a BERT-style text-to-image learner motivated by BLIP2 [17], which is designed to automatically extract visual knowledge from the food images. The extracted visual knowledge is subsequently fed into the Image-Informed Text Encoder, enhancing CLIP's static text representation with image-specific information. This process enables tailored adjustments to the text embeddings, effectively dealing with ingredient classes that exhibit diverse visual appearances. Figure 1 (b) showcases an example of classifying egg prepared in different ways using both the conventional static text embedding and the proposed image-informed text embedding. In more common scenarios (such as boiled egg), both models yield reliably accurate predictions. However, in more ambiguous cases (like scrambled egg), the performance of the conventional static text embedding is poor, while the image-informed text embedding still achieves impressive results.

The training process of OVFoodSeg comprises two stages: the FoodLearner Pre-training stage and the Segmentation Learning stage. In the first stage, the FoodLearner module is trained to align visual information with textual representations. This training utilizes a large-scale dataset of food-related image-text pairs, intended to introduce the FoodLearner to a broad range of food items. Following this, the Segmentation Learning stage involves fine-tuning the FoodLearner for the open-vocabulary food image segmentation task using a specialized segmentation dataset. This stage further hones the model's ability to accurately segment food items in diverse scenarios.

To validate the effectiveness of OVFoodSeg, we conduct experiments under open-vocabulary setting based on the existing publicly available food image segmentation dataset, FoodSeg103 [31]. We specifically design two open-vocabulary scenarios for the experiments: (i) a random splitting of FoodSeg103, allocating 80% of the ingredient classes as base classes for training and the remaining as novel classes for inference, and (ii) the expansion of FoodSeg103 by introducing 92 new classes to create FoodSeg195, followed by a similar random split. The results of these experiments indicate that our OVFoodSeg outperforms the state-of-the-art open-vocabulary segmentation method, *i.e.*, SAN [34], by 4.9% in mean Intersection over Union (mIoU) on novel classes in FoodSeg103 and by 3.5% in FoodSeg195. Here is a summary of our contributions:

- We introduced OVFoodSeg, an innovative framework specifically designed for open-vocabulary food image segmentation. Our OVFoodSeg employs a two-stage training and is the first to address the challenge of segmenting food ingredients at an open-vocabulary level.
- To tackle the issue of large intra-class variance in the visual representation of food ingredients, OVFoodSeg integrates an image-to-text learning module, FoodLearner, along with the Image-Informed Text Encoder. This integration enriches textual representations with visual knowledge derived from food images.
- OVFoodSeg achieves state-of-the-art performance in two open-vocabulary benchmarks for food image segmentation, demonstrating its effectiveness and advancement over existing methods.

2. Related Work

Food-related Segmention Dataset: Food image segmentation is a core problem in food computing, and constructing a large-scale datasets with pixel-wise mask annotation is the fundamentation to address this problem. Myers *et al.* [20] construct Food201, a dataset containing roughly 12,000 images across 201 dish classes. Later UECFoodPix [8] and UECFoodPixComplete [24] are proposed which contain about 10,000 images across 102 dish types for seg-

mentation. However, the annotations of these datasets are all restricted to dish-wise masks, rendering them unsuitable for ingredient segmentation. In response to this limitation, FoodSeg103 [31] is created as the first food image segmentation dataset with ingredient-level annotations which is still far from enough to present the diversity of food. Thus we propose to address food image segmentation in the open-vocabulary setting.

Food Image Segmentation: Concurrently, the exploration of food image segmentation frameworks is advancing. Wu *et al*. [31] introduce ReLeM, a method designed to mitigate the large intra-class variance by incorporating recipe information into the visual representation. Wang *et al*. [29] develope a Swin Transformer-based segmenter called STPPN, which harnesses contextual information from various regions within the food image, thus enriching the global representation. Jaswanthi *et al*. [12] utilize a hybrid approach for food image segmentation, initially applying a GAN model [11] to generate proposal masks for food images, which are then categorized by a CNN-based recognition model. Lan *et al*. [16] introduce FoodSAM, combining the pre-trained SAM model [14] with a segmentation model [35] trained on FoodSeg103 for high-quality mask generation. However, these existing algorithms in food image segmentation predominantly focus on close-set learning, limiting their adaptability to a wider array of ingredient categories not included in the training set.

Open-Vocabulary Segmentation: Open-vocabulary segmentation aims to identify objects with pixel-wise masks beyond the classes seen during training. Early work [2, 32] concentrated on creating joint embeddings that link image pixels with class concepts. Recently, CLIP-based frameworks have made significant strides in this field. Zhou *et al*. [37] present MaskCLIP, which uses a frozen CLIP model to generate pseudo pixel labels for segmentation model training. OpenSeg [9] enhances segment-level visual features with text embeddings through region-word associations. Xu *et al*. [33] develop SimSeg, which produces class-agnostic masks with a mask generator, followed by classification with a CLIP-based classifier. Building upon this, SAN [34] employs CLIP directly for mask generation, foregoing the complex mask generator. Liang *et al*. [18] introduce OVSeg, which fine-tunes the original CLIP image encoder with masked images to improve segmentation understanding. These approaches tend to use static text embeddings from CLIP encoded class names, neglecting the variability of image content, which can be detrimental when handling images with large intra-class variance. Wang *et al*. [30] propose HIPIE, which integrates visual features into text embeddings through attention modules. However, the simple attention modules struggle to handle the misalignment between the visual and text spaces, while our proposed FoodLearner is capable of aligning the visual and text spaces with much less computational cost.

3. OVFoodSeg

To address the significant challenge of large intra-class variance in ingredients for effective food image segmentation in open-vocabulary settings, we present OVFoodSeg. Inspired by the recent success of Vision-Language Models (VLMs) in open-vocabulary settings for various tasks, OVFoodSeg is developed on the foundation of a frozen CLIP model. CLIP integrates both image encoder and text encoder to embed the two modalities. To mitigate the large intra-class variance, OVFoodSeg incorporates two critical components, namely the *FoodLearner* and *Image-Informed Text Encoder*. FoodLearner is responsible for extracting visual information from food images, which is then utilized by the Image-Informed Text Encoder to enhance CLIP's text embeddings. The training procedure of OVFoodSeg is structured into two stages: first stage to pre-train the FoodLearner with a large-scale dataset of food-related image-text pairs, and the second stage to fine-tune the pre-trained FoodLearner to perform segmentation task. In this section, we will introduce two training stages along with corresponding modules.

3.1. Stage I Training: FoodLearner Pre-training

Figure 2 presents the training block for Stage I, focusing on the pre-training of the FoodLearner using pairs of images and texts related to food. The objective is to explicitly pre-train the FoodLearner so that the visual information closely related to the accompanying text will be extracted to enrich the text representation. To achieve this, we simultaneously optimize three types of loss: Image-Text Contrastive loss (ITC Loss), Image-Text Matching loss (ITM Loss), and Language Modeling loss (LM Loss). These losses function in concert, sharing the same model parameters throughout the pre-training process. In the following, we outline the FoodLearner structure and three objective losses utilized during pre-training.

FoodLearner is implemented with the transformer structure [13], *i.e*., multiple transformer blocks, each featuring a self-attention layer, a cross-attention layer, and a feed-forward network. A collection of learnable query tokens T_{query} is introduced to interface with the conditional visual embedding through the cross-attention layer to capture visual knowledge. The primary input of FoodLearner is a sequence of tokens, where these tokens can be either learnable query tokens, text tokens, or a combined sequence of query and text tokens. Additionally, FoodLearner may also receive a conditional input, *i.e*., visual embedding E_{visual}, from the CLIP image encoder. Thus the output of FoodLearner, *i.e*., FoodLearner enriched token, could be denoted as $\mathrm{FL}(E_{visual}, T_{query})$ or $\mathrm{FL}(NULL, T_{text})$, where

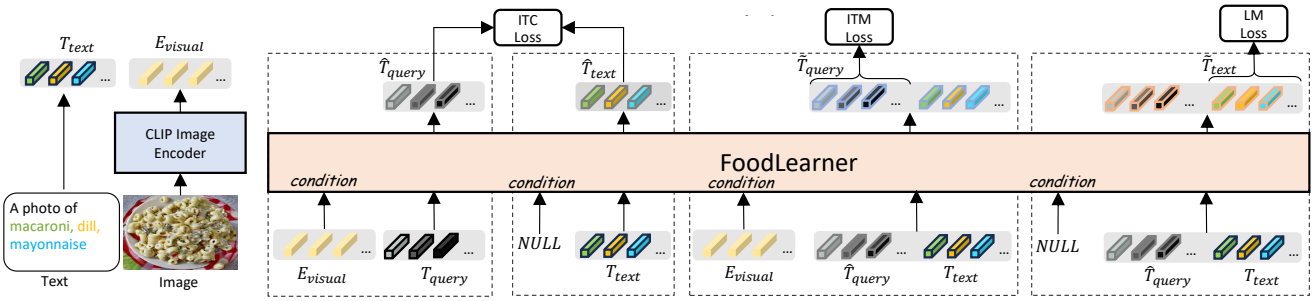

Figure 2. The figure depicts the pipeline of Stage I FoodLearner Pre-training. Stage I is dedicated to pre-training the FoodLearner module with image-text pairs pertinent to food so that the visual information closely related to the accompanying text will be extracted to enrich the text representation.

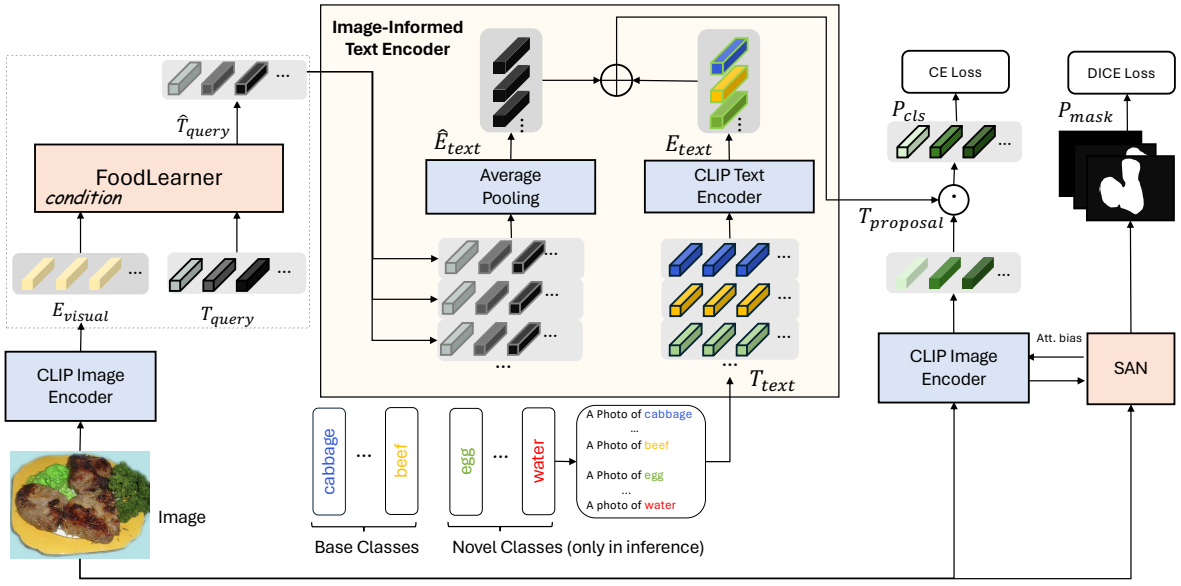

Figure 3. This figure depicts the pipeline of Stage II Segmentation Learning, focusing on training the segmenter using image-informed text embeddings. The FoodLearner extracts image-specific information which are then combined with the text embeddings from Image-Informed Text Encoder to produce the final image-informed text embeddings. Noted that modules with the same name share the parameters, *i.e.*, CLIP image encoder and CLIP text encoder.

the former is a conditional input. Noted that the cross-attention layer is activated only when the conditional input is not $NULL$.

Image-Text Contrastive Loss is designed to promote a higher similarity for positive image-text pairs as opposed to negative pairs. We first generate the FoodLearner enriched tokens for both image and corresponding prompted text, *i.e.*, enriched query tokens $\hat{T}_{query} = \mathrm{FL}(E_{visual}, T_{query})$ and enriched text tokens $\hat{T}_{text} = \mathrm{FL}(NULL, T_{text})$. Following this, we calculate the pairwise similarity between each enriched query token and enriched text token and select the highest value to represent the image-text similarity. For each image and text, we compute the softmax-normalized image-to-text and text-to-image similarities p_i^{t2i} and p_i^{i2t} as:

$$p_i^{t2i} = \frac{max_t(\phi * \left\langle \left[\hat{T}^t_{query}\right]_i, \left[\hat{T}_{text}\right]_i \right\rangle)}{\sum_j max_t(\phi * \left\langle \left[\hat{T}^t_{query}\right]_j, \left[\hat{T}_{text}\right]_i \right\rangle)} \quad (1)$$

$$p_i^{i2t} = \frac{max_t(\phi * \left\langle \left[\hat{T}^t_{query}\right]_i, \left[\hat{T}_{text}\right]_i \right\rangle)}{\sum_j max_t(\phi * \left\langle \left[\hat{T}^t_{query}\right]_i, \left[\hat{T}_{text}\right]_j \right\rangle)} \quad (2)$$

Here $\left[\hat{T}_{query}\right]_i$ and $\left[\hat{T}_{text}\right]_i$ correspond to the i-th image's enriched query tokens and enriched text tokens, respectively. The notation $\langle\ ,\ \rangle$ computes the cosine similarity.

The variable t indicates the index of the query token exhibiting the highest similarity to the text embedding contrasted, and ϕ is the temperature parameter used to scale the similarity (set to 10 in this paper). The ITC Loss is then computed by the Cross-Entropy loss (CE) as:

$$L_{ITC} = \frac{1}{2} * (\text{CE}(p^{t2i}, GT_{t2i}) + \text{CE}(p^{i2t}, GT_{i2t})) \quad (3)$$

where GT_{t2i} and GT_{i2t} indicate the ground truth labels, with positive pairs assigned label 1 and negative pairs label 0.

Image-Text Matching Loss is designed to learn fine-grained alignment between image and the corresponding prompted text. A binary classifier is learned here to predict whether the image-text pair is matching or not. The process begins by concatenating $\hat{T}_{query}$ and T_{text} and feeding them with visual embedding into the FoodLearner, *i.e.*, $\tilde{T}_{query} = \text{FL}(E_{visual}, \hat{T}_{query}; T_{text})$, where ; denotes concatenation. Different from ITC, $\hat{T}_{query}$ and T_{text} could interact within the self-attention layer, thus enabling the query tokens assimilate multimodal information from the text. Subsequently, the resulting output query tokens $\tilde{T}_{query}$ are passed through a fully-connected layer to calculate the match probability. Given that there are Q query tokens, the final probability value p^{itm} is derived by averaging these values.:

$$p^{itm} = \frac{1}{Q} \sum_{t=1}^{Q} \text{FC}(\left[\tilde{T}^{t}_{query}\right]) \quad (4)$$

The ITM Loss is then computed as:

$$L_{ITM} = \text{CE}(p^{itm}, GT_{itm}) \quad (5)$$

where the GT_{itm} indicates the ground truth labels.

Language Modeling Loss aims to facilitate the generation of text informed by image content. The process also begins by concatenating the query tokens, $\hat{T}_{query}$ with the text tokens T_{text}, which are then passed into FoodLearner to generate $\tilde{T}_{text} = \text{FL}(NULL, \hat{T}_{query}; T_{text})$. Within FoodLearner module, we try to reconstruct original input words given the query tokens. Thus we optimize the auto-regressive LM Loss L_{LM} on the W text tokens to maximize the probability of the reconstructed text:

$$p^{word}_{w} = \text{FC}(\tilde{T}^{w}_{text}) \quad (6)$$

$$L_{LM} = \frac{1}{W} \sum_{w} \text{CE}(p^{word}_{w}, GT_{word}) \quad (7)$$

Since T_{text} interacts exclusively with $\hat{T}_{query}$ and not directly with E_{visual}, $\hat{T}_{query}$ are fine-tuned to encapsulate the visual information most pertinent to the text.

Finally, the loss function of Stage I in training FoodLearner is:

$$L_{StageI} = L_{ITC} + L_{ITM} + L_{LM} \quad (8)$$

3.2. Stage II Training: Segmentation Learning

Figure 3 illustrates the training process of Stage II for OVFoodSeg, focusing on the learning of the food image segmentation framework. During this stage, we utilize the FoodLearner, which was pre-trained in Stage I, to extract image-specific information. This information, in conjunction with the text embeddings of the target ingredient classes, is then fed into the Image-Informed Text Encoder to generate image-informed text embeddings, which are subsequently used for the segmentation learning.

Firstly, the FoodLearner utilizes learnable query tokens T_{query} to extract image-specific information from the visual embeddings E_{visual} of the input images: $\hat{T}_{query} = \text{FL}(E_{visual}, T_{query})$. We calculate the average of $\hat{T}_{query}$ across Q query tokens as the visual representation $\hat{E}_{text}$:

$$\hat{E}_{text} = \text{Average-Pooling}(T_{query}) \quad (9)$$

Then we individually pair the name of each ingredient class with a pre-designed template (using "A Photo of {}" in the paper) to create the text token for that specific ingredient class. The text tokens T_{text}, each each comprising L vectors where L denotes the vocabulary size. The text embeddings for the original text tokens are encoded as:

$$E_{text} = \text{CLIP-TE}(T_{text}) \quad (10)$$

We create the image-informed text embeddings by integrating $\hat{E}_{text}$ with E_{text} using an element-wise summation operation:

$$\tilde{E}_{text} = \hat{E}_{text} + E_{text} \quad (11)$$

$\tilde{E}_{text}$ are then input into a standard open-vocabulary segmentation framework to facilitate segmentation learning. In this paper, we follow the pipeline of SAN [34], which is fine-tuned with two objective losses: Cross-Entropy loss for classification and Dice loss [27] for mask prediction. For classification, the CLIP image encoder generates a set of proposal tokens $T_{proposal}$, each conditioned on the attention bias from the SAN module and corresponding to a specific region of the input image. Subsequently, the image-informed text embeddings $\tilde{E}_{text}$ are engaged in a dot product with $T_{proposal}$, yield the class probability distributions:

$$P^{i}_{cls} = \frac{\exp(\tau * < T_{proposal}, \tilde{E}^{i}_{text} >)}{\sum_{j} \exp(\tau * < T_{proposal}, \tilde{E}^{j}_{text} >)} \quad (12)$$

Here P^{i}_{cls} represents the predicted probability values for the i-th class, and τ is the temperature used to re-scale the similarity (set 100 here). For mask prediction, the framework incorporates SAN module, a lightweight Vision Transformer, that calculates binary masks aligned with $T_{proposal}$ from the classification branch. Finally, the classification branch is

optimized by Cross-Entropy loss (CE) while the mask prediction module is optimized by Dice loss (DICE:

$$L_{StageII} = \text{CE}(P_{cls}, GT_{cls}) + \text{Dice}(P_{mask}, GT_{mask}) \tag{13}$$

Here GT_{cls} and GT_{mask} represent the ground truth for class and mask, respectively. During the Stage II training, we maintain the parameters of the CLIP text encoders fixed and focus on training the FoodLearner and the SAN exclusively.

4. Experiment

In this section, we first detail our experimental setup, including dataset configuration and model implementation. We then present the performance of our model on two open-vocabulary food image segmentation benchmarks. Lastly, we provide an ablation study to evaluate the impact of different components and settings within OVFoodSeg based on FoodSeg103 dataset followed by the case study.

4.1. Experimental Setup

Training Dataset of Stage I: For Stage I pre-training, we utilized the Recipe-1M+ dataset [19], which is currently the most extensive dataset of image-recipe pairs available, containing approximately 1 million culinary recipes. Each recipe is detailed with the dish's name, cooking methods, an ingredient list, and about 10 images representing the dish from various restaurants. This dataset, with its extensive collection of food photographs and ingredient lists, is particularly well-suited for Stage I training. However, it's important to note that the ingredient lists in the dataset often include "invisible" ingredients such as sugar, oil, and salt. These ingredients, while integral to the recipes, contribute to data noise since they are invisible in the images. To refine the quality of our ingredient data, we utilize ChatGPT to obtain the most visually evident ingredients for each recipe[1]. We then leveraged the cleaned ingredient lists as text information, paired with the corresponding images, to pre-train the FoodLearner module.

Training Dataset of Stage II: For Stage II segmentation learning, we conduct experiments on two benchmarks: FoodSeg103 and FoodSeg195. FoodSeg103 comprises approximately 7,000 images across 103 ingredient classes. We randomly select 20 classes as novel classes and use the remaining 83 as base classes. Expanding upon FoodSeg103, we add an additional 92 classes to create the larger dataset FoodSeg195[2], which includes around 18k images for training and 16k for testing, totaling 113k annotated masks. From FoodSeg195, we randomly designate 40 classes as novel, with the rest serving as training data. The annotations of novel classes are blocked during training, and the test sets of FoodSeg103 and FoodSeg195 are used for evaluating the model performance. To mitigate the impact of randomness on experimental results, we conducted three random class splits for both the FoodSeg103 and FoodSeg195 datasets in the main experiments. These experiments were carried out over three iterations, and we reported the average values and standard deviations. In the ablation study, we report results from only the first split based on one experiment. For the details of the class splits, please refer to the appendix.

Implementation: We employ the CLIP ViT-L/14 [6, 26] model as VLM, and initialize FoodLearner with weights of qformer pre-trained via BLIP2 [17]. In stage I, we initialize Q query tokens (Q set 32 in this paper), each with a channel dimension of 768 in stage I. In stage II, 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. We detail the specific hyperparameters for the newly proposed module in the Appendix. For the remaining implementation details, we adhere to the configurations established in SAN [34].

Metric: We adopt widely used mean Intersection over Union (mIoU) scores as the metric for segmentation task and further use mIoU_n, mIoU_b, and mIoU_o to represent the mIoU for novel classes, base classes, and all classes, respectively.

4.2. Main Results on FoodSeg103 and FoodSeg195

Results on the FoodSeg103 and FoodSeg195 datasets are detailed in Table 1 and Table 2. We train segmentation models and evaluate their performance on FoodSeg103 and FoodSeg195 using multiple class splits to mitigate the impact of randomness on the experiments. For each split, we randomly generate novel classes following the strategy outlined in Section 4.1. Additionally, we introduce a variant of OVFoodSeg based on FoodLearner without pre-training on the Recipe-1M+ dataset, denoted as OVFoodSeg*.

For both the FoodSeg103 and FoodSeg195 datasets, OVFoodSeg demonstrates a significant improvement over the baseline models for novel classes. For instance, it shows an average improvement of 4.9% on FoodSeg103 and 3.5% on the more challenging FoodSeg195, across multiple class splits, compared to the SOTA (State-Of-The-Art) method, SAN. This strongly validates the effectiveness of the proposed OVFoodSeg framework. It is noted that SimSeg achieves better performance in overall class metrics on FoodSeg195. We attribute this to SimSeg's reliance on a heavily trained mask generator [4], which offers robustness in complex food scenarios with precise annotations. However, SimSeg is markedly slower than OVFoodSeg, almost

[1] We request ChatGPT prompted with "Given a recipe name as {}, generate a short description. Only mention the main ingredients that can be seen in the dish (no more than 5)."

[2] The images were collected from the residents of a country for food logging through a government agency. Due to the confidentiality agreements, FoodSeg195 cannot be made publicaly available at the moment.

Method	Split 1			Split 2			Split 3		
	mIoU_n	mIoU_b	mIoU_o	mIoU_n	mIoU_b	mIoU_o	mIoU_n	mIoU_b	mIoU_o
MaskCLIP [37]	12.3	37.2	32.4	16.8	36.6	32.8	16.4	37.0	33.0
MaskCLIP [36]	15.4	38.5	34.0	20.1	37.3	34.0	18.4	38.7	34.8
OVSeg [18]	17.4	40.0	35.6	25.1	38.7	36.0	22.2	40.2	36.7
SimSeg [33]	21.7	37.5	34.4	26.2	37.7	35.5	22.4	37.2	34.3
FreeSeg [25]	22.1	35.4	32.8	31.6	29.6	30.0	27.2	29.9	29.4
SAN [34]	25.6	39.5	36.8	32.9	39.1	37.9	27.6	42.7	39.8
OVFoodSeg*	28.7±0.8	44.2±0.8	41.2±0.8	37.6±1.6	43.1±0.8	42.1±0.9	31.7±1.7	44.5±1.3	42.0±1.3
OVFoodSeg	**30.0**±1.2	45.5 ±0.9	42.5±1.0	**38.1**±1.0	43.0±0.8	42.0±0.8	**32.8**±1.1	45.8±0.7	43.3±0.8

Table 1. The table presents a comparison between existing open-vocabulary segmentation baselines and the proposed OVFoodSeg on FoodSeg103. All models are trained on the FoodSeg103 training set and evaluated on the FoodSeg103 test set. OVFoodSeg* denotes the model based on the FoodLearner without Recipe-1M+ pre-training.

Method	Split 1			Split 2			Split 3		
	mIoU_n	mIoU_b	mIoU_o	mIoU_n	mIoU_b	mIoU_o	mIoU_n	mIoU_b	mIoU_o
MaskCLIP [37]	11.1	30.4	26.4	11.0	26.8	23.7	9.8	26.2	22.8
MaskCLIP [36]	12.8	30.6	26.9	12.2	27.0	24.0	10.5	27.2	23.8
OVSeg [18]	15.7	24.3	22.5	14.4	21.6	20.1	13.6	21.9	20.2
SimSeg [33]	17.5	35.1	31.5	14.6	33.7	29.8	14.7	32.4	28.8
FreeSeg [25]	16.8	25.1	23.4	15.9	22.4	21.1	15.1	26.5	24.2
SAN [34]	20.0	26.0	24.8	18.0	27.3	25.4	16.6	26.5	24.3
OVFoodSeg*	22.8±0.3	29.9±0.2	28.4±0.2	20.5±0.2	30.4±0.1	28.3±0.1	18.4±0.4	28.7±0.1	26.6±0.2
OVFoodSeg	**24.0**±0.5	29.1±0.3	28.1±0.3	**21.4**±0.4	29.8±0.3	28.1±0.3	**19.5**±0.4	28.0±0.1	26.3±0.2

Table 2. The table presents a comparison between existing open-vocabulary segmentation baselines and the proposed OVFoodSeg on FoodSeg195. All models are trained on the FoodSeg195 training set and evaluated on the FoodSeg195 test set. OVFoodSeg* denotes the model based on the FoodLearner without Recipe-1M+ pre-training.

30 times (0.3fps vs 9.8fps on V-100), and underperforms significantly in the crucial novel class metric (6% in average). Furthermore, the OVFoodSeg* variant, even without Recipe-1M+ pre-training, also achieves notable improvements of 4.0% and 2.4% in novel classes compared with SAN, underscoring the benefits of the image-informed textual representation mechanism.

4.3. Analysis

Objective Losses in Stage I: In this section, we investigate the impacts of three loss functions (ITC loss, ITM loss, and LM loss) employed in the Stage I pre-training of FoodLearner. In our experiments, we employ different combinations of these losses to pre-train FoodLearner, which is initialized using BLIP2 [17]. Our segmentation model coupled with the different pre-trained FoodLearner models are trained and evaluated based on the FoodSeg103 dataset. As shown in Table 3, ITC loss and ITM loss are crucial for achieving competitive results, as they align the representations of food images and text information. Leaving out either ITC or ITM loss leads to a significant deterioration in segmentation performance by 4 to 5%. We observe that LM loss enhances the segmentation results by encouraging the query tokens to learn image knowledge related to the corresponding text input.

ITC	ITM	LM	mIoU_n	mIoU_b	mIoU_o
	✓	✓	26.5	43.1	39.9
✓		✓	27.3	44.2	40.9
✓	✓		30.8	45.5	42.9
✓	✓	✓	31.1	45.3	42.5

Table 3. The impacts of different objective losses used in Stage I FoodLearner Pre-training. The learned FoodLearner of each setting is then used for segmentation learning. We show the results on FoodSeg103.

Prompt Engineering: Prompt engineering is effective for open-vocabulary segmentation and detection [10, 15, 34]. This section examines various prompt strategies, employing multiple templates to prompt ingredient names, with each template's image-informed text embedding subjected to average pooling, as shown in Table 4. Surprisingly, we find that using a default single prompt template, *i.e.*, "`a photo of {}`", or even forgoing the use of templates altogether results in significantly better performance in novel classes compared to existing prompt engineering

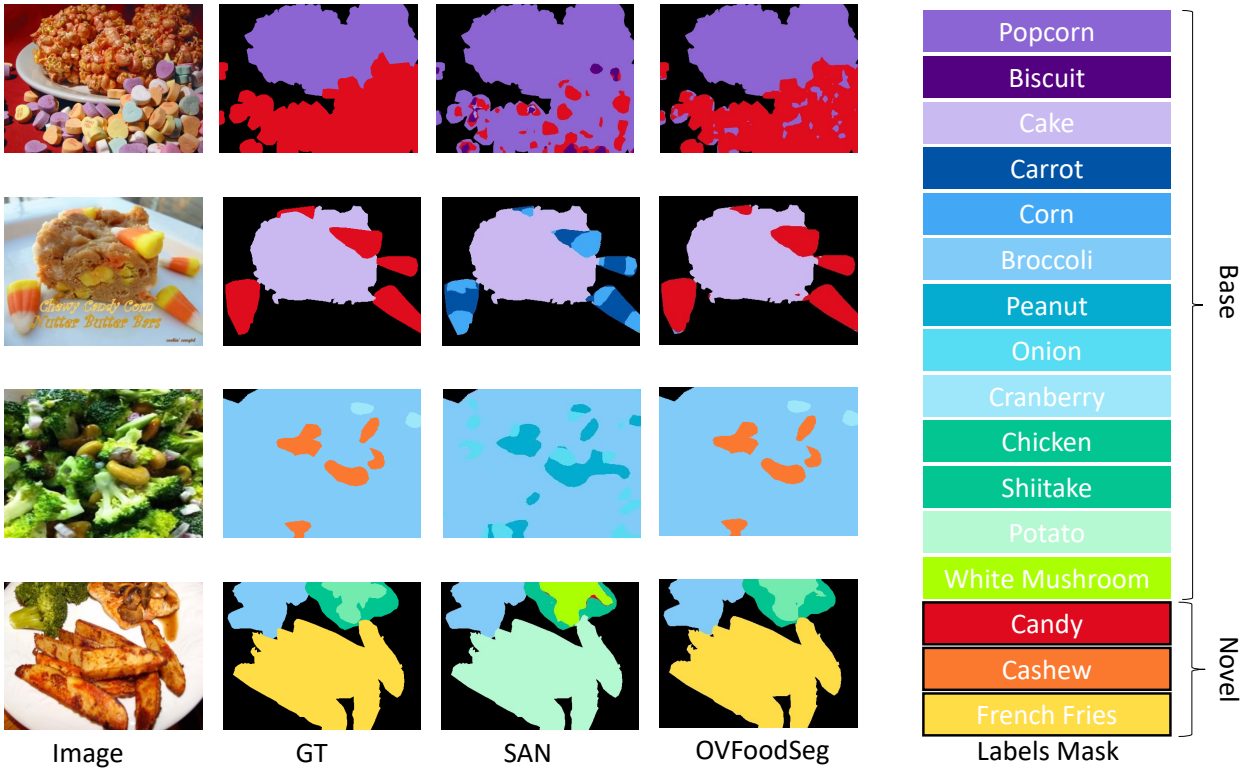

Figure 4. Visulization Results on FoodSeg103 where GT means ground-truth. OVFoodSeg achieves better performance especially for novel classes.

Prompt	mIoU_n	mIoU_b	mIoU_o
None	30.2	44.3	41.6
Default	31.1	45.3	42.5
ViLD [10]	29.4	46.7	43.3
ImageNet [26]	29.3	48.2	44.5

Table 4. Performance comparison of OVFoodSeg using different prompt engineering strategies on FoodSeg103. "Default" denotes the single prompt used in the paper.

strategies. Existing prompt engineering methods, which average embeddings across templates, may not fully exploit the representation capability of text embeddings. Exploring effective ways to combine multiple template text embeddings is a potential direction for future research.

Qualitative Results: In Figure 4, we provide qualitative comparisons between SAN and OVFoodSeg on the FoodSeg103 test set. The results clearly illustrate that OVFoodSeg significantly outperforms SAN, particularly in segmenting novel classes. Specifically, the first two rows highlight OVFoodSeg's ability to accurately segment candy with varying appearances, a task at which the baseline model, SAN, fails. This effectively demonstrates the superior performance of the proposed OVFoodSeg model in addressing the issue of large intra-class variance.

5. Conclusion

In this study, we introduce OVFoodSeg, an innovative open-vocabulary segmentation framework specifically designed for food images. Leveraging the CLIP model, excels in enriching text embeddings with image-specific information, enabled by the innovative FoodLearner and Image-Informed Text Encoder modules. Demonstrating its efficacy on the FoodSeg103 and FoodSeg195 datasets, OVFoodSeg surpasses existing baselines, especially in segmenting novel classes and addressing the substantial intra-class variance prevalent in food imagery. The resultant improvement in segmentation accuracy establishes OVFoodSeg as a new benchmark in the field, paving the way for future advancements in open-vocabulary food image segmentation task.

6. Acknowledgement

This research / project is supported by the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Proposal ID: T2EP20222-0046). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Ministry of Education, Singapore.

References

[1] Rebecca G Boswell, Wendy Sun, Shosuke Suzuki, and Hedy Kober. Training in cognitive strategies reduces eating and improves food choice. *PNAS*, 2018. 1

[2] Maxime Bucher, Tuan-Hung Vu, Matthieu Cord, and Patrick Pérez. Zero-shot semantic segmentation. *NeuriPS*, 2019. 3

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2017. 1

[4] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *NeuriPS*, 2021. 6

[5] Tilman David and Clark Michael. Global diets link environmental sustainability and human health. *Nature*, 2014. 1

[6] 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. 6

[7] Yu Du, Fangyun Wei, Zihe Zhang, Miaojing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 2

[8] Takumi Ege and Keiji Yanai. A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice. In *MADiMa*, 2019. 1, 2

[9] Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 3

[10] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 2, 7, 8

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 3

[12] R Jaswanthi, E Amruthatulasi, Ch Bhavyasree, and Ashutosh Satapathy. A hybrid network based on gan and cnn for food segmentation and calorie estimation. In *ICSCDS*, 2022. 3

[13] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 3

[14] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *ICCV*, 2023. 3

[15] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. Open-vocabulary object detection upon frozen vision and language models. In *ICLR*, 2023. 2, 7

[16] Xing Lan, Jiayi Lyu, Hanyu Jiang, Kun Dong, Zehai Niu, Yi Zhang, and Jian Xue. Foodsam: Any food segmentation. *IEEE Transactions on Multimedia*, 2023. 3

[17] 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. 2, 6, 7

[18] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *CVPR*, 2023. 3, 7

[19] Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. Recipe1m+: a dataset for learning cross-modal embeddings for cooking recipes and food images. *arXiv preprint arXiv:1810.06553*, 2018. 6

[20] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. Im2Calories: towards an automated mobile vision food diary. In *ICCV*, 2015. 1, 2

[21] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys*, 2019. 1

[22] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *arXiv preprint arXiv:2103.16107*, 2021. 1

[23] Weiqing Min, Chunlin Liu, Leyi Xu, and Shuqiang Jiang. Applications of knowledge graphs for food science and industry. *Patterns*, 2022. 1

[24] Kaimu Okamoto and Keiji Yanai. UEC-FoodPIX Complete: A large-scale food image segmentation dataset. In *ICPRW*, 2020. 1, 2

[25] Jie Qin, Jie Wu, Pengxiang Yan, Ming Li, Ren Yuxi, Xuefeng Xiao, Yitong Wang, Rui Wang, Shilei Wen, Xin Pan, et al. Freeseg: Unified, universal and open-vocabulary image segmentation. In *CVPR*, 2023. 7

[26] 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. 2, 6, 8

[27] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *DLMIA*, 2017. 5

[28] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. In *CVPR*, 2021. 1

[29] Qiankun Wang, Xiaoxiao Dong, Ruimin Wang, and Hao Sun. Swin transformer based pyramid pooling network for food segmentation. In *SEAI*, 2022. 3

[30] Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation. In *NeuriPS*, 2023. 3

[31] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *ACMMM*, 2021. 1, 2, 3

[32] Yongqin Xian, Subhabrata Choudhury, Yang He, Bernt Schiele, and Zeynep Akata. Semantic projection network for zero-and few-label semantic segmentation. In *CVPR*, 2019. 3

[33] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *ECCV*, 2022. 3, 7

[34] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, 2023. 1, 2, 3, 5, 6, 7

[35] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 1, 3

[36] Zhuowen Tu Zheng Ding, Jieke Wang. Open-vocabulary universal image segmentation with maskclip. In *ICML*, 2023. 7

[37] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 3, 7