# FIND THE FOOD

**Cesar Ades**
Northwestern University
cesaradese@gmail.com

**Joshua D'Arcy**
Northwestern University
joshua.darcy@northwestern.edu

**ABSTRACT**

Semantic segmentation of food items in images presents unique challenges due to the high variability and overlapping nature of food items. This study investigates the application of zero-shot object detection models combined with advanced segmentation techniques for food segmentation tasks without additional model training. We benchmarked several models, including OWLv2, OWL-ViT, and OmDet, on the FoodSeg103 dataset. Our findings reveal that a Top-K approach, which assigns each pixel to the most common class label among the top K detections, significantly outperforms traditional threshold-based methods. The OWLv2 models emerged as the best performers, demonstrating robustness and effectiveness in zero-shot food segmentation tasks.

## 1 Introduction

The proliferation of food imagery on social media and dietary tracking applications has increased the need for accurate semantic segmentation of food items in images. Semantic segmentation involves classifying each pixel in an image into predefined categories, which is crucial for applications like nutritional assessment, food recognition, and augmented reality in culinary contexts.

However, food segmentation poses significant challenges. Food items often exhibit high intra-class variability due to differences in preparation styles, presentations, and occlusions. The overlapping nature of food items on plates further complicates the segmentation task. Traditional supervised learning approaches require extensive annotated datasets, which are labor intensive to produce and may not generalize well to novel food items.

Zero-shot learning offers a promising alternative by enabling models to recognize and segment unseen classes without explicit training data for those classes. Recent advancements in zero-shot object detection models, such as OWL-ViT and OWLv2, have shown potential in generalizing to new categories through the use of vision-language models.

In this study, we explore the efficacy of zero-shot object detection models for food segmentation tasks. We propose a methodology that leverages these models in conjunction with the Segment Anything Model (SAM) and introduces a mask combination method that reduces reliance on confidence scores. Our approach aims to assign each pixel to the most common class label among the top K detections, improving segmentation performance.

Our contributions are as follows:

1. We evaluate the performance of several zero-shot object detection models on the FoodSeg103 dataset.

2. We propose a Top-K mask combination method that outperforms traditional threshold-based methods.
3. We analyze the impact of different parameters, such as confidence thresholds, Top-K values, and polygon refinement, on model performance.
4. We identify OWLv2 as superior performers in zero-shot food segmentation tasks.

## 2 Models

In this study, we evaluated several zero-shot object detection models for their performance in food segmentation tasks. The models tested include:

**OWL-ViT (B/L)**

A vision-language model designed for open-world object detection tasks. It leverages the transformer architecture to process images and text jointly, enabling zero-shot detection of objects based on textual descriptions. OWL-ViT utilizes a CLIP backbone with a ViT-like Transformer for visual features and a causal language model for text features. It adapts to detection by removing the final token pooling layer of the vision model and attaching lightweight classification and box prediction heads to each transformer output token.

**OWLv2 (B/L)**

An improved version of OWL-ViT, incorporating advancements in model architecture and training methodologies to enhance zero-shot object detection capabilities.

**OmDet Turbo**

A zero-shot object detection model that focuses on leveraging large-scale pre-training and efficient model architectures for open-vocabulary detection tasks, focusing on real-time applications.

This model introduces an Efficient Fusion Head (EFH) module to alleviate computational bottlenecks found in transformer-based detectors and achieves real-time performance with high frames per second rates, making it suitable for industrial applications. An EFH module enhances the efficiency of object detection by integrating lightweight language-aware encoders and decoders, deformable attention mechanisms, and optimized task decoupling making it ideal for real-time applications.

**Grounding DINO**

Grounding DINO is an open-set object detector that combines the Transformer-based DINO detector with grounded pre-training to detect arbitrary objects using human inputs like category names or referring expressions. It introduces a novel approach to fusing language and vision modalities through a feature enhancer, language-guided query selection, and a cross-modality decoder, enabling open-set concept generalization.

Grounding DINO was excluded from our testing framework despite its promising performance on certain tasks due to several limitations. The model's embedded token limit of 256 was insufficient for the extensive vocabulary of the FoodSeg103 dataset, leading to incomplete detections and an inability to process all necessary class labels. Additionally, the model's output of thresholded prompts, which differed from the original input prompts, caused label inconsistencies that complicated evaluation and hindered accurate performance assessment. This mismatch between detected and ground truth classes further conflicted with our methodology, which relies on consistent input-output mapping. Adapting the model to our framework would have required significant modifications, detracting from the focus of our study. Consequently, we prioritized models that seamlessly integrated with our framework and fulfilled the requirements of the FoodSeg103 dataset.

## 3 Dataset

**Overview**

In this study, we evaluated the performance of zero-shot object detection models for food segmentation tasks using the FoodSeg103 dataset. This dataset is a large-scale collection of food images designed specifically for semantic segmentation, making it an ideal benchmark for our objectives. FoodSeg103 comprises 7,118 images, with 2,135 designated for the test set. As our methodology focuses exclusively on evaluation without training, we utilized only the test set. The dataset is notable for its diverse and rich annotations, making it a valuable resource for assessing segmentation models.

**Annotations**

The dataset includes 104 ingredient classes, each represented by pixel-wise segmentation masks in every image. These masks provide granular details about the location and type of food ingredients present. On average, each image contains six ingredient labels, highlighting the complexity of food segmentation tasks and the variety of elements typically found in food images.

The ground truth masks are grayscale images where pixel values correspond to specific class labels. These integer values directly encode the ingredient class indices, embedding the segmentation information within the masks themselves.

**Challenges**

FoodSeg103 presents several challenges that underline its suitability as a benchmark for segmentation models:

- High Intra-Class Variance: The same ingredient can appear in various forms due to differences in cooking methods, presentations, and styles, creating significant variability within the same class.
- Long-Tail Distribution: The dataset has an uneven distribution of classes, with some ingredients appearing frequently while others are rare. This imbalance tests a model's ability to perform consistently across all classes.
- Complex Contexts: Food items often overlap or are presented in cluttered settings, complicating the identification of individual ingredients.

These challenges make FoodSeg103 a robust and realistic dataset for testing segmentation models under real-world conditions.

## 4 Methodology

Our proposed methodology integrates zero-shot object detection models with advanced segmentation techniques to perform food segmentation without additional model training. The approach consists of the following steps:

**Object Detection**

In this study, we employed pre-trained zero-shot object detection models to identify food items within images. These models leverage their capability to detect objects based on textual class labels without requiring prior training on the specific dataset, making them highly versatile for tasks involving novel categories. For our evaluation, the models were provided with the complete class list from the FoodSeg103 dataset and tuned using specific hyperparameters to optimize detection performance.

The key hyperparameters used were:

- Threshold: The confidence threshold for detections, determining the minimum score required for an object to be considered valid and output by the model.
- NMS (Non-Maximum Suppression): The threshold for suppressing overlapping detections to ensure that each object is detected only once. This parameter was only used in the OmDet Turbo model.
- Top-K: The number of top detections to consider for further processing, focusing on the most confident predictions.

**Segmentation**

The bounding boxes, class labels, and confidence scores generated by the object detection models are passed to the Segment Anything Model (SAM). SAM produces segmentation masks for the detected objects, providing pixel-level delineation of food items.

**Polygon Refinement**

Polygon refinement is the process of enhancing binary object masks by converting them into precise polygonal representations and then reconstructing them back into masks. This process is optimally applied on SAM's output masks. This approach simplifies the contours of detected objects, ensuring accurate localization while reducing unnecessary noise and artifacts, which is especially beneficial in complex or cluttered visual environments. This method can improve the precision of object boundaries and reduce redundant data; however, it may also oversimplify intricate shapes or introduce errors when handling highly complex or irregular objects.

**Mask Combination Method**

We propose a mask combination method to address inconsistencies in confidence scores across models and classes. Instead of relying solely on these scores, the method assigns each pixel to the most common class label among the masks in the top K detections. In the event of a tie, the pixel is assigned to the class with the highest confidence detection. This approach reduces the impact of uneven confidence score distributions and enhances segmentation accuracy by prioritizing frequently detected classes. K is effectively unbounded, allowing for improved performance as more detections become available, ensuring robustness even with a large number of overlapping predictions.

## 5 Results

In this section, we present the performance of the evaluated zero-shot object detection models on the FoodSeg103 dataset. The models were assessed using standard semantic segmentation metrics: Mean Intersection over Union (mIoU), Mean Accuracy (mACC), and overall pixel accuracy (aAcc). We provide an extensive analysis of the results, focusing on the impact of different parameters such as the use of Top-K selection, confidence thresholds, Non-Maximum Suppression (NMS), and polygon refinement.

### 5.1 Performance Comparison Across Models

We compared the performance of various models using the Mean Intersection over Union (mIoU) metric. Figure 1 shows the mIoU scores for all models evaluated on the FoodSeg103 dataset.

Table 1: Statistical Summary of mIoU Metrics

| Model | Mean (%) | Std (%) | Max (%) |
|---|---|---|---|
| OWL-ViT Base | 8.88 | 6.06 | 13.04 |
| OWL-ViT Large | 16.55 | 6.35 | 21.62 |
| OWLv2 Base | **32.13** | 4.18 | 35.80 |
| OWLv2 Large | 31.61 | 7.47 | **37.59** |
| OmDet | 11.59 | 1.13 | 14.46 |

Overall, the **OWLv2 models** outperform the other models in terms of mIoU, with both Base and Large variants demonstrating strong clustering at higher values, while OmDet and OWL-ViT models show weaker performance. The superior performance of OWLv2 models can be attributed to several factors:

1. **Architectural Enhancements Focused on Efficiency**:
   OWLv2 introduces architectural advancements primarily aimed at improving training and inference efficiency without sacrificing performance. Key enhancements include

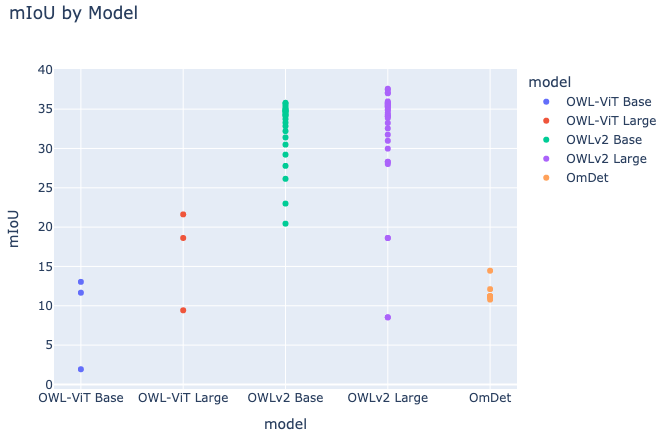



Figure 1: Mean Intersection over Union (mIoU) for all models evaluated on the FoodSeg103 dataset.

token dropping and instance selection, but the most relevant change is in mosaics. OWLv2 employs a more extreme version of mosaics compared to OWL-ViT, combining raw images into grids of up to 6×6. This increases the number of raw images seen during training and improves performance on small objects, which is beneficial for the diverse and intricate food items in the dataset.

2. **Improved Training Procedures**:

   OWLv2 benefits from refined training methodologies that enhance its generalization capabilities in zero-shot settings. The model is initially trained on a massive dataset of pseudo-annotated web image-text pairs. A simple N-grams approach is used for query generation, allowing the model to learn from a vast and diverse set of images without manual annotations. After self-training, OWLv2 is fine-tuned on smaller, human-annotated datasets like $\text{LVIS}_{\text{base}}$. This fine-tuning step refines the model's performance and boosts its ability to generalize to new categories, which is crucial for zero-shot object detection.

3. **Adaptation to High Variability in Food Images**:

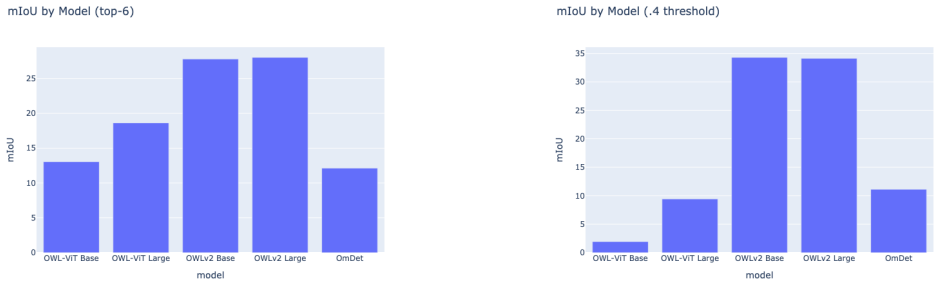
   The OWLv2 models demonstrate robustness to the high intra-class variability and overlapping nature of food items in the FoodSeg103 dataset. The combination of efficient architecture and extensive training allows OWLv2 to capture the diverse appearances of food ingredients, leading to higher mIoU scores.

**OWL-ViT**, despite using similar CLIP backbones, perform less effectively. The lower performance of OWL-ViT models may be due to the lack of architectural enhancements present in OWLv2 that focus on computational efficiency and handling of small objects. Also, OWL-ViT does not benefit from the same extensive self-training on web image-text pairs or the specific fine-tuning strategies employed by OWLv2.

**OmDet** exhibits the lowest mIoU values, generally below 10%, suggesting that it is less suited for the specific challenges presented by the FoodSeg103 dataset. This may be attributed to their design focus. OmDet is designed for real-time detection with an emphasis on inference speed, potentially at the expense of fine-grained segmentation accuracy required for complex food images.

### 5.2 Impact of Confidence Thresholding and Top-K Selection

We further analyzed the models' performance under different control methods: confidence thresholding and Top-K selection. Figure 2 presents the performance of the models when controlled by a confidence threshold of 0.4 and when using Top-K selection with $K = 6$.




(a) Top-K selection with $K = 6$.

(b) Confidence thresholding at 0.4.

Figure 2: Model performance controlled by Top-K selection and confidence thresholding.

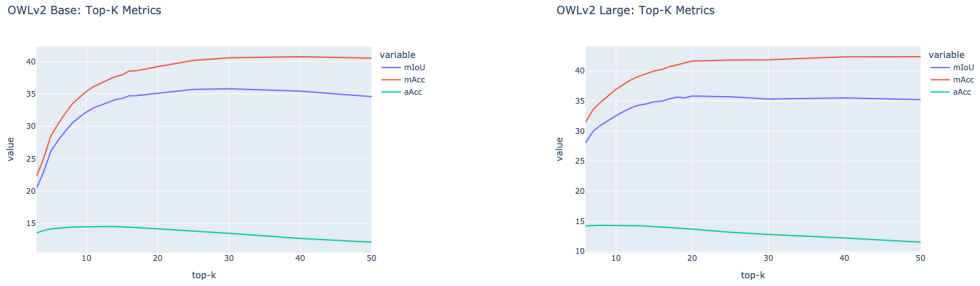
From Figure 2, we notice that when using a confidence threshold of 0.4 (Figure 2b), the OWLv2 models significantly outperform the OWL-ViT models. The OWL-ViT models show lower mIoU scores under thresholding, indicating that their confidence scores may not be as reliable for filtering detections in the food segmentation context.

In contrast, when using Top-K selection with $K = 6$ (Figure 2a), the performance gap between the OWLv2 and other models narrows. The OWL-ViT models perform better under the Top-K approach compared to thresholding, suggesting that selecting the top detections regardless of confidence scores benefits their performance.

The fact that the models are more evenly matched with Top-K selection indicates that this method reduces reliance on confidence scores and focuses on the most prominent detections. However, the value of $K = 6$ is relatively low, which may limit the models' ability to capture all relevant food items in the images. Higher values of $K$ were tested on OWLv2 models to assess the impact of considering more detections.

### 5.3 Analysis of Top-K Selection on OWLv2 Models

We conducted an extensive analysis of the Top-K selection method on the OWLv2 Base and OWLv2 Large models by varying $K$ and observing its effect on the performance metrics. Figure 3 illustrates the results.




(a) OWLv2 Base model.

(b) OWLv2 Large model.

Figure 3: Effect of varying Top-K on OWLv2 models' performance metrics.

From Figure 3, we observe that as $K$ increases, both OWLv2 Base and OWLv2 Large models show a steady improvement in mIoU and mACC, with minimal diminishing returns even at higher values of $K$. This trend persists despite the inclusion of more low-confidence detections when $K$ is increased.

The lack of significant diminishing returns suggests that the OWLv2 models benefit from aggregating more detections, even those with lower confidence levels. By increasing $K$, we include more detections, which allows the model to determine the most common class labels in each area of the image. This frequency-based approach enhances segmentation accuracy because the most frequently detected classes within a region are likely to be correct. Consequently, adding more low-confidence detections does not significantly disrupt the model's performance, provided that the majority of detections are correct. This method effectively captures diverse food items across the image by leveraging the collective information from multiple detections.

### 5.4 Threshold Sensitivity Analysis of OWLv2 Large Model

We analyzed the sensitivity of the OWLv2 Large model to different confidence thresholds to understand how thresholding affects its performance. Figure 4 presents the performance metrics at varying threshold levels.

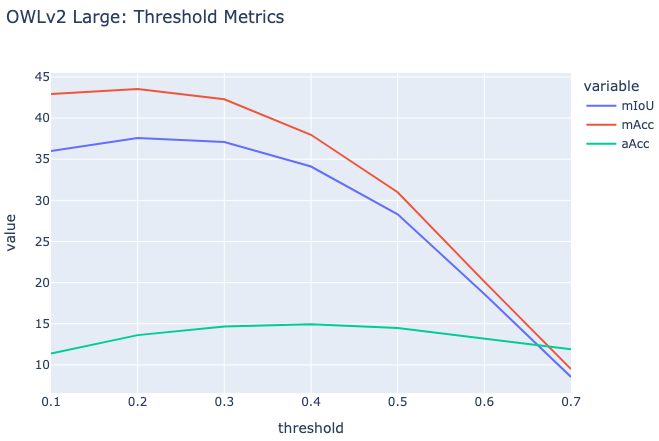



Figure 4: Effect of varying confidence thresholds on OWLv2 Large model's performance metrics.

As shown in Figure 4, the mIoU and mACC metrics for the OWLv2 Large model peak around a confidence threshold between .2 and .3. Beyond this threshold, the performance metrics begin to decline. This indicates that setting an appropriate confidence threshold is crucial for maximizing the model's performance.

At lower thresholds, more detections are included, some of which may be false positives, potentially lowering the precision. At higher thresholds, fewer detections are considered, possibly missing relevant food items, which can reduce recall. The optimal threshold balances these factors to achieve the highest mIoU and mACC. The sensitivity to thresholding underscores the importance of calibrating confidence scores for detection models.

### 5.5 Impact of NMS and Polygon Refinement

We examined the effect of Non-Maximum Suppression (NMS) and polygon refinement on the models' performance. Figure 5 shows the impact of varying the NMS threshold on the OmDet model's performance metrics.

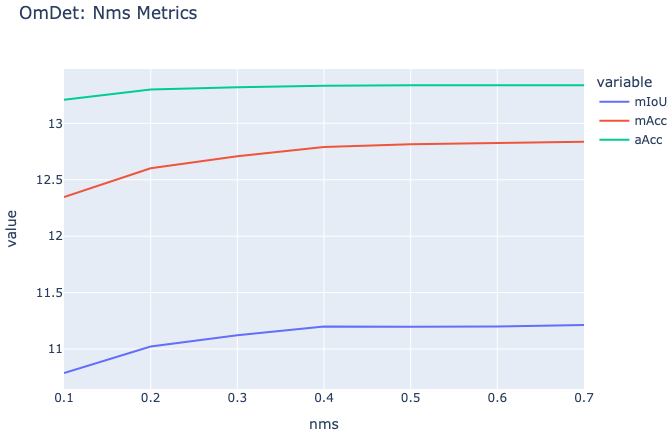



Figure 5: Effect of varying NMS threshold on OmDet model's performance metrics.

From Figure 5, we observe that increasing the NMS threshold from 0.1 to 0.4 leads to improvements in mIoU, mACC, and aAcc. However, beyond an NMS threshold of 0.4, the performance metrics plateau, indicating diminishing returns. This plateau suggests that optimal suppression of overlapping detections is achieved around this threshold, and further increasing it does not significantly affect the model's ability to distinguish between overlapping food items.

We also analyzed the effect of polygon refinement on the OWLv2 Large model under varying confidence thresholds. Figure 6 presents the results.

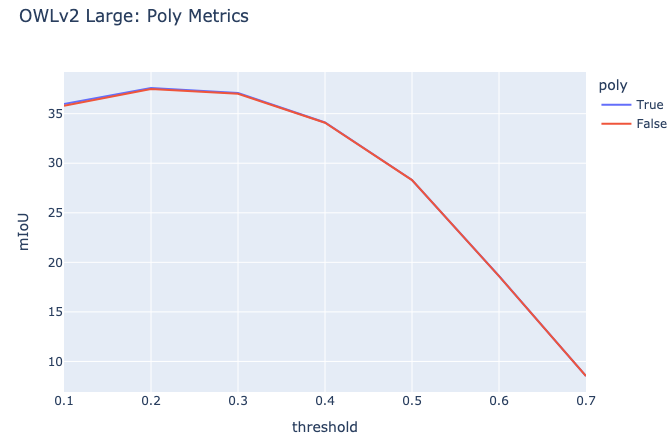



Figure 6: Effect of polygon refinement on OWLv2 Large model's performance under varying thresholds.

Figure 6 shows that polygon refinement has a marginal but positive effect on the performance metrics when tested under varying thresholds. The refinement process improves the accuracy of the segmentation masks by better aligning the predicted boundaries with the actual contours of the food items.

However, the improvements are modest, indicating that the initial segmentation masks produced by the models are already of high quality. The slight enhancement suggests that while polygon

refinement can contribute to better performance, its impact is limited compared to other factors such as the choice of $K$ and confidence thresholding.

## 6 Next Steps

Building upon our findings, several avenues for future research are identified to enhance and validate our methodology:

**Polygon Refinement with Top-K Selection**: While polygon refinement showed marginal improvements under varying thresholds, its impact in conjunction with the Top-K selection method remains unexplored.

**Comparative Analysis with Higher Top-K Values**: Our current analysis primarily focused on relatively low values of $K$. Evaluating all models using a higher Top-K value, such as $K = 30$, could provide insights into the models' ability to handle a larger number of detections.

**Testing Grounding DINO Model**: Although we initially excluded Grounding DINO due to compatibility issues with our framework, revisiting this model with adjustments could be beneficial. Modifying our methodology to accommodate Grounding DINO's output format and token limitations may allow us to assess its performance on the FoodSeg103 dataset, potentially offering alternative insights or competitive results in zero-shot food segmentation.

**Integration with Other Segmentation Models**: Exploring the integration of zero-shot object detection models with other segmentation techniques, such as fully convolutional networks or attention-based models, may yield additional improvements. Combining strengths from different approaches could enhance segmentation accuracy and robustness.

## 7 Conclusion

Our extensive analysis reveals that OWLv2 models outperform it's competitor models in food segmentation tasks, likely due to architectural advancements that enhance feature representation and modality alignment. The Top-K selection method proves effective, as it reduces reliance on confidence scores and leverages the models' ability to produce relevant detections even at lower confidence levels.

The analysis also highlights the importance of calibrating confidence thresholds to optimize model performance. The impact of NMS and polygon refinement on performance is evident but limited. Optimizing NMS thresholds can improve model performance up to a point, after which gains plateau. Polygon refinement offers marginal improvements, suggesting that the models already generate accurate segmentation masks.

Overall, our findings emphasize the significance of model architecture, parameter tuning, and methodological choices in enhancing zero-shot object detection models for food segmentation tasks.