VIETNAM ACADEMY OF SCIENCE AND TECHNOLOGY

UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

FINAL REPORT

of

FOOD INGREDIENTS SEGMENTATION BASED ON DEEP LEARNING

Subject: Computer Vision

Lecturer: Assoc. Prof. Dr. Nguyen Duc Dung

Members: Pham Hai Nam – BI12-307

Le Mau Minh Phuc – BI12-351

Nguyen Son – BI12-389

Hanoi, March 10, 2024

ABSTRACT

The final report outlines a study on Food Ingredients Segmentation using Deep Learning, highlighting its significance in addressing obesity through innovative dietary management. The research introduces FoodSAM, an advanced model integrating the Segment Anything Model (SAM) for improved accuracy in food ingredients segmentation, covering semantic, instance, and panoptic segmentation. By using datasets such as UECFoodPixComplete, SA-1B, FoodSeg103, and Objects365, the study explores the effectiveness of FoodSAM for food ingredient segmentation and Yolov8 for reference card detection. The evaluation metrics such as mIOU, aAcc, and mAcc are used to evaluate the model performance, finding optimal thresholds for accurate food segmentation. Results demonstrate FoodSAM's capability in detailed food item identification, with potential applications in health and dietary management. The discussion highlights FoodSAM's promise in combating dietary challenges and suggests future improvements in algorithmic efficiency, user experience, and broader dataset incorporation. The report positions FoodSAM as a valuable tool for advancing dietary management and health applications, emphasizing the need for continuous innovation in this field.

Keywords: Food Ingredients Segmentation, FoodSAM, Yolov8, Reference Card

Table of Contents

1. Project Introduction

Having established the critical context and urgency due to the prevalent dietary health issues, as highlighted by recent statistics, there is a currently indispensable need for innovative solutions like Food Ingredient Segmentation technology. This technology not only promises a new horizon in dietary management but also aligns with the global aim to combat obesity and promote healthier eating habits.

1.1 Inspiration for Project Study

Food Ingredients Segmentation technology stands at the forefront of innovative strategies to combat dietary health challenges and the rising obesity epidemic. According to recent statistics from the National Health and Nutrition Examination Survey in 2018, the situation is dire: approximately 30.7% of adults are overweight, 42.4% are considered obese, and 9.2% suffer from severe obesity. These staggering figures highlight an urgent public health crisis affecting adults and young people, emphasizing the critical necessity for enhanced weight management solutions [1].

In the quest to address this crisis, monitoring and controlling caloric intake emerge as pivotal elements. Traditional methods of tracking food consumption often rely on manual logging and estimation, approaches that are prone to inaccuracies and non-compliance over time. Enter Food Ingredients Segmentation technology: a game-changing solution designed to revolutionize the way individuals interact with their food.

Moreover, Food Ingredients Segmentation technology has broader implications beyond individual dietary management. It can serve as a valuable educational tool, raising awareness about nutritional content and promoting healthier eating habits. In healthcare settings, dietitians and medical professionals can leverage this technology to tailor dietary recommendations and interventions more effectively. Additionally, it holds the potential for addressing global challenges related to food security and nutrition education, providing scalable solutions that can reach diverse populations [3].

1.2 Related Work

1.2.1 SAM model

The landscape of natural language processing has been transformed by the development

of large language models trained on extensive web datasets. These models are particularly notable for their zero-shot generalization capabilities, allowing them to apply their knowledge beyond their initial training and perform well across various tasks and data types. This revolution has now extended into computer vision, with the introduction of the Segment Anything Project (SAM) by Meta AI, marking a significant leap in image segmentation. SAM represents a major step forward, aiming for universal cognitive segmentation and addressing the challenges of interactive segmentation within real-world parameters [4].

SAM has shown impressive performance across different segmentation benchmarks, particularly highlighting its zero-shot capabilities across numerous diverse datasets. This research delves into SAM's application in food image segmentation, a key area within food computing. However, while SAM creates detailed masks, it struggles with providing class-specific information, essential for food segmentation. This issue, compounded by the variety in food appearances and uneven distribution of ingredients, makes distinguishing food types accurately a complex task [5].

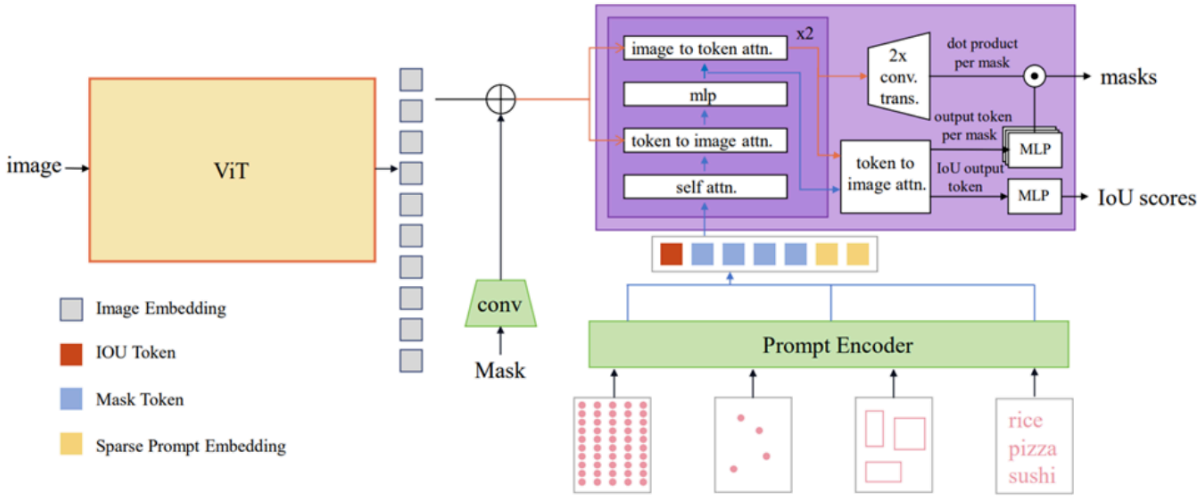

Figure 1: Segment Anything Model (SAM)

Applying SAM to food image segmentation offers new possibilities but also presents unique challenges, mainly due to the lack of specific category information in the generated masks and the intricate nature of food items. To bridge this gap, FoodSAM is a hybrid framework that enhances SAM's detailed mask outputs with categorical data, aligning high-quality segmentation with meaningful semantic information. It accounts for the individuality of food components, treating each as a separate entity for instance segmentation. Additionally, FoodSAM extends its capabilities to detect and segment

non-food elements within the image, achieving comprehensive panoptic segmentation [6].

FoodSAM innovates further by adopting SAM's interactive prompt-based approach, allowing for refined segmentation through user interaction, and addressing the complex variety found in food images. This combination results in a robust, versatile solution for food image analysis, advancing both practical applications and theoretical understanding in the domain [6].

FoodSAM model excels in three key segmentation tasks: instance segmentation, panoptic segmentation, and semantic segmentation. For instance segmentation, FoodSAM accurately identifies and delineates individual food items within images, providing precise insights even amidst multiple similar items. This granular understanding is essential for detailed analysis. In panoptic segmentation, FoodSAM demonstrates a holistic grasp of image content, seamlessly integrating instance-level segmentation with a broader contextual understanding. This comprehensive approach enables accurate labeling of all image elements, including food items and surroundings. In semantic segmentation, FoodSAM effectively categorizes each pixel into meaningful classes, such as different food types and background elements, providing a foundational understanding of image structure for further analysis [6].

1.2.2 Natural Language Processing

Foundation models, which are trained on extensive datasets to adapt to various tasks, have significantly propelled machine learning forward. This approach integrates self-supervised learning, transfer learning, and prompt tuning to enhance performance across a range of applications. In the realm of natural language processing, models like the Generative Pre-trained Transformers have shown remarkable success, leveraging vast amounts of text data to excel in translation, question-answering, and more. Similarly, Contrastive Language-Image Pre-training, which uses image-text pairs, has advanced capabilities in image retrieval and classification based on textual prompts [4].

A unique addition to this domain is the Segment Anything Model (SAM), developed in parallel with model-in-the-loop annotation, creating a vast data engine comprising over 1 billion masks. This model stands out for its strong generalization ability, contributing to its exceptional performance. Foundation models like SAM have reached state-of-the-art levels in various fields, indicating a promising future for advancing machine learning technologies across multiple disciplines [5].

1.2.3 Card Detection

To estimate the true dimensions of a food region, it is necessary to utilize a reference object of known size within the meal photograph. This allows for the simultaneous estimation of food calories. For the purposes of this study, the meal photograph is assumed to be captured from an overhead perspective, eliminating the need for correction of trapezoidal distortion. Consequently, various shaped objects such as wallets and smartphone cases can be utilized alongside rectangular objects like cards. To facilitate the extraction of these objects from the image, GrabCut is employed similarly to food region extraction [7].

1.3 Objective

According to the literature review and inspiration of healthcare problems, the objective of this research is

- Implement a framework for food ingredients segmentation.
- Propose a framework for reference card detection.
- Evaluate models according to evaluation metrics.

2. Methodology

Moving from the introduction to the methodology section, this part explains the steps and tools used to solve the problems mentioned earlier. It covers the specific datasets, methods, and analyses used to improve food ingredients image segmentation. The goal is to clearly show how the study aims to meet its objectives and tackle the challenges of recognizing different foods in images.

2.1 Datasets

2.1.1 UECFoodPixComplete

The UECFoodPixComplete dataset is used to evaluate the semantic segmentation process. The dataset contains 10000 food images with their segmentation masks, covering entire dishes but lacks fine-grained annotation for individual dish components. With a ratio of 9:1 for training and testing (9000 train and 1000 test images), the dataset contains 103 class labels, with 102 labels being food and 1 being the "background" label [7].

Figure 2: Examples of UECFoodPixComplete dataset

2.1.2 SA-1B

SA-1B (Segment Anything 1-Billion mask dataset) was published by Meta AI laboratory, containing 11M licensed and privacy-preserving images and 1.1 billion masks corresponding to the images. In this experiment, SA-1B is used for pretrained SAM (Segment Anything Model) with the ViT-H variation as the image encoder. According to Kirillov et al., 2023 [5] the results of training on 1 million images and training on 11 million images don't yield a significant change in the evaluation metrics value [8].

Figure 3: Examples of SA-1B dataset

2.1.3 FoodSeg103

FoodSeg103 dataset is used to evaluate the semantic segmentation process as well as fine-tune the label prediction model in FoodSAM. The dataset contains 7118 food images with the ingredient segmentation masks, with a ratio of 7:3 for training and testing (4983 training images with 29530 ingredient masks and 2135 test images with 12567 ingredient masks). The dataset contains 104 class labels, with 102 food labels, a “background” label, and an “other ingredients" label. The difference between FoodSeg103 and UECFoodPixComplete is that in FoodSeg103, individual dish components annotations are fine-grained, therefore the dataset is suitable for the semantic and instance segmentation process [9].

Figure 4: Examples of FoodSeg103 dataset

2.1.4 Objects365

Objects365 is a large-scale object detection dataset containing over 2 million images and 365 categories. In this experiment, the object detection model (Unified learned COIM RS200) is pre-trained using 600k images of Objects365 for training. For evaluating the object detection model, 38k images are used for validation and another 100k images for testing. The PASCAL VOC and COCO datasets are subsets of Objects365 and are also used for evaluating the model [10].

Figure 5: Examples of Objects365 dataset

2.1.5 Manual Collection Reference Card Dataset

Since there currently isn't any image data available for various types of Vietnamese cards like bank cards, citizen identity cards, student cards, etc., with the size of 85.6mm*53.98mm, manual collecting data and labeling the cards were carried out to detect and evaluate the card detection model in this project. As a result, a dataset of Vietnamese card images consisting of 310 images with 21 different types of cards are collected, including mostly USTH student cards, followed by citizen identity cards, various bank cards, gym cards, etc., all of the same size; and it is divided in a 90:10 ratio for the training and test sets.

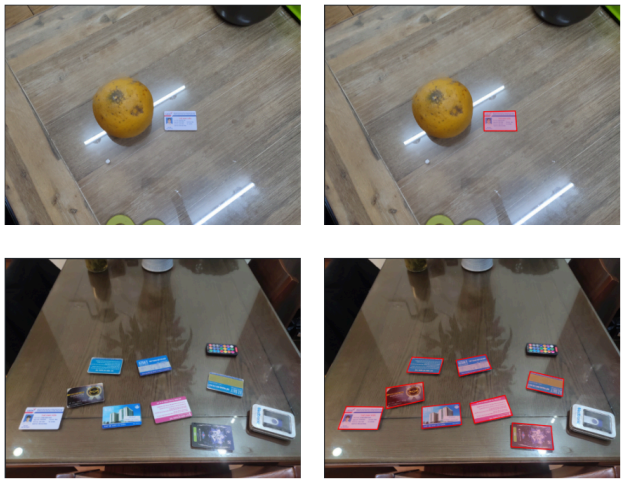

Figure 6: Examples of Manual Collection Reference Card dataset

In the figure 6, the left pictures represent the images without Card labeling. The picture on the right are labeled with Card bounding boxes by using Label Studio

2.2 Implementation Details

The experiments were carried out on a desktop computer using Ubuntu 20.04.6 LTS as the operating system. The system specifications included an i5-10400 CPU running at 2.9GHz with 6 cores, 16GB of RAM operating at a speed of 2133 MHz, and an NVIDIA 3060 GPU with 16GB of memory and 3584 CUDA cores.

2.3 Pipeline of Experiments

2.3.1 Simple Pipeline

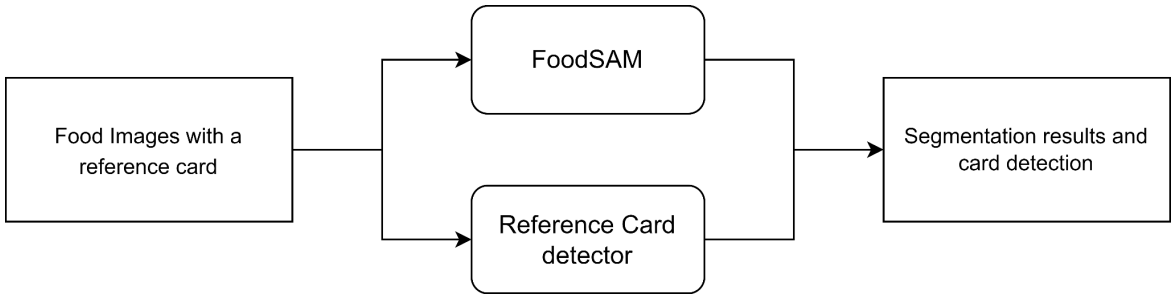

Figure 7: Simple Framework

The diagram in Figures 7 and 8 with the details describes a systematic approach utilized by a system named "FoodSAM" for analyzing food images. The process begins with the

input of food images, each containing a reference card to provide scale or color reference. These images are then fed into the FoodSAM module, which is likely tasked with segmenting and identifying different elements within the images, such as distinguishing various types of food or separating food items from non-food elements. Concurrently, a Reference Card Detector component analyzes the same images to locate and interpret the reference card. This is an essential step to ensure that the subsequent analysis is accurately scaled and calibrated. The final output from this system includes both the segmentation results, which detail the classified elements within the images and the outcomes of the card detection, indicating that the reference card has been successfully identified and utilized.

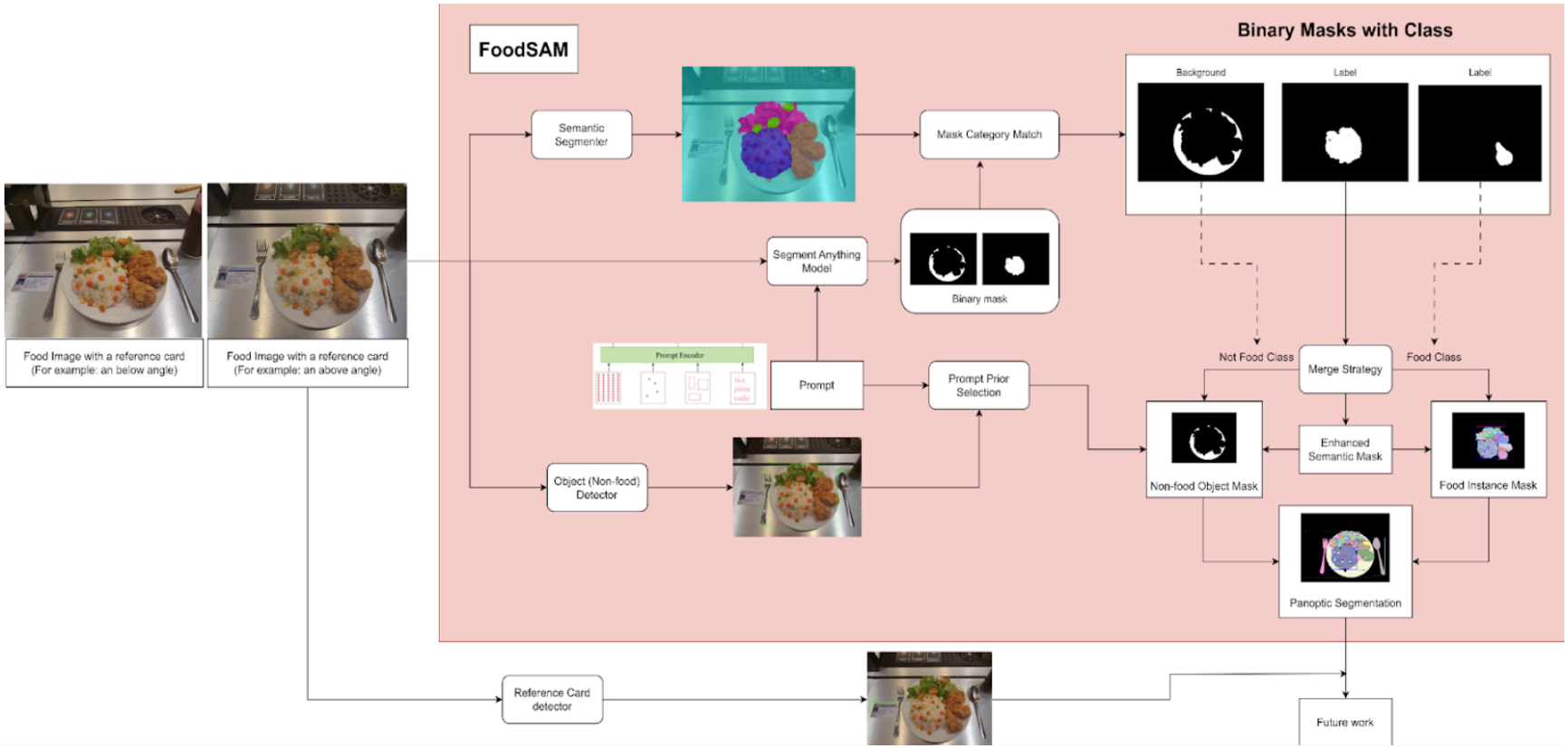

Figure 8: Proposed framework of our study

2.3.2 FoodSAM

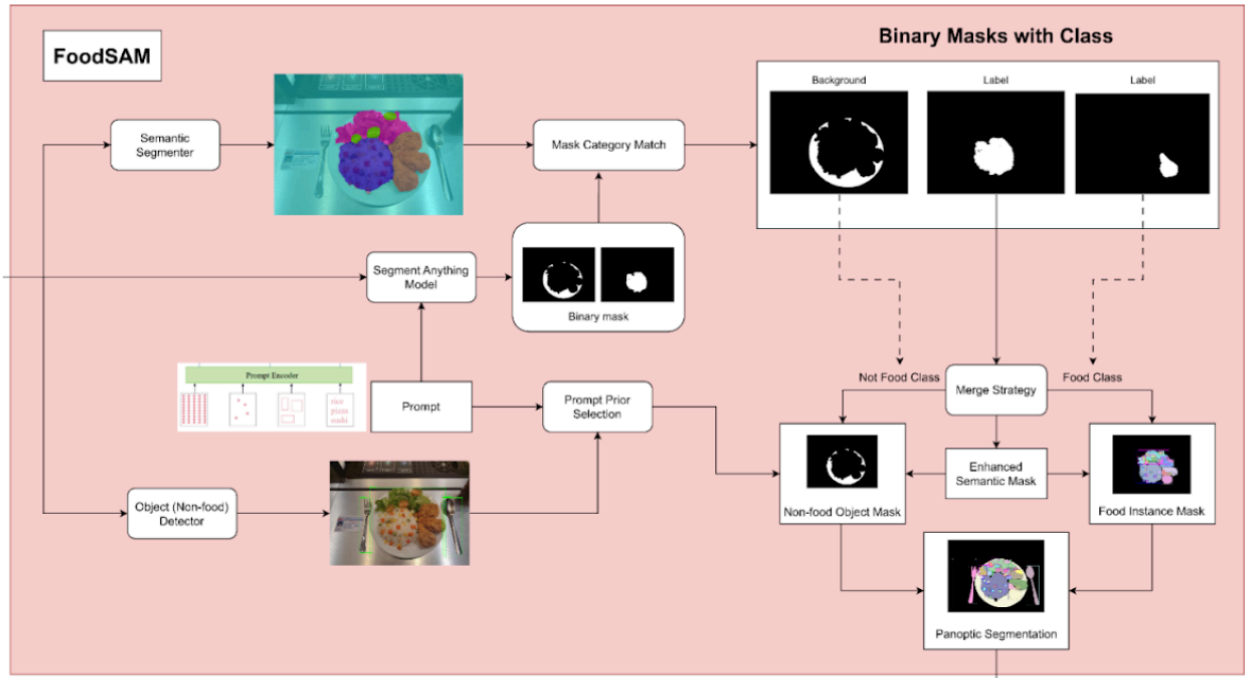

Figure 9: FoodSAM framework

The framework depicted, FoodSAM, integrates multiple advanced techniques for food image segmentation, leveraging three core models: a "Segment Anything Model", a semantic segmentation module, and an object detector. This structured approach aims to refine segmentation results by effectively distinguishing between different food items and non-food objects within images.

Enhance Semantic Segmentation: Initially, food images undergo semantic segmentation through a semantic segmentation module, producing semantic masks that identify categories for each pixel. Simultaneously, the "Segment Anything Model" processes the images to generate binary masks, distinguishing foreground (potential food items) from the background. A mask-category match then aligns each binary mask with semantic labels via a voting scheme based on the highest frequency of semantic values within the mask. This process incorporates a confusion measure to ensure only clear, stable masks are utilized. By integrating high-quality binary masks with semantic labeling, an enhanced semantic mask is created, offering improved detail and accuracy by addressing overlaps with a merge strategy [5].

Semantic-to-Instance Segmentation: This phase transitions from broad semantic labels to precise instance segmentation, treating each food item as an independent entity.

By merging similar small masks and filtering out non-foreground elements, distinct instance masks for each food item are established, acknowledging the random placement and cutting typical in food preparation [5].

Instance-to-Panoptic Segmentation: Extending beyond food, this stage involves detecting non-food objects using Md to recognize items like utensils or plates, integral to understanding the food context. Non-food objects are individually segmented and classified, merging with food instance masks to create a comprehensive panoptic segmentation, which offers a complete overview of both food and related objects within the scene [5].

Promptable Segmentation: This innovative component introduces flexible interaction with the segmentation models through prompts, allowing for refined focus, especially for non-food object segmentation. Utilizing different prompts (point, box, mask), the system can more accurately identify, and segment non-food elements based on predefined criteria, enhancing the granularity and precision of the segmentation process across both food and non-food objects [5].

2.3.3 Reference Card Detection by Yolov8

Figure 10 illustrates the process of object detection within our model. The initial input, typically an image requiring analysis, is represented in the "Input" section. The diagram then outlines the components of the one-stage detector, enclosed within a dotted-line box, which is divided into three main segments: Backbone, Neck, and Dense Prediction. The Backbone serves as the initial segment of the network, tasked with extracting features from the input image, often utilizing a pre-trained deep neural network. Connecting the Backbone to the Dense Prediction layer is the Neck, which enhances feature extraction from the Backbone and prepares the features for precise object detection. Finally, the Dense Prediction stage is where object detection occurs, generating multiple bounding boxes (depicted by small red boxes) around potential objects in the image. Each bounding box is associated with a class label and a confidence score [2].

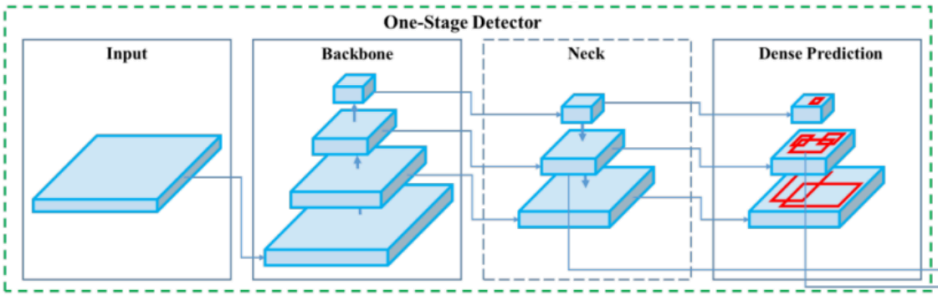

Figure 10: Yolov8 simple architecture

The depicted process showcases a method for analyzing food images using a reference card for accurate scale and color calibration, specifically employing the YOLOv8 algorithm for reference card detection. Initially, the process involves capturing an image of a food plate from an overhead perspective, where a reference card is strategically placed within the frame. This card is essential as it establishes a standardized scale and color reference for the subsequent analysis. Following this, the YOLOv8 model, known for its effectiveness and efficiency in object detection tasks, is applied to identify and delineate the reference card within the image. This detection is typically indicated by a bounding box around the card, providing visual confirmation of the card's location and dimensions. The identification of the reference card is a crucial step, enabling the system to calibrate measurements and color accuracy for a detailed analysis of the food items present, enhancing the precision of size, portion, and color evaluations. This application of YOLOv8 highlights the innovative integration of advanced object detection methodologies into the realm of food image analysis, ensuring more accurate and reliable assessments [2].

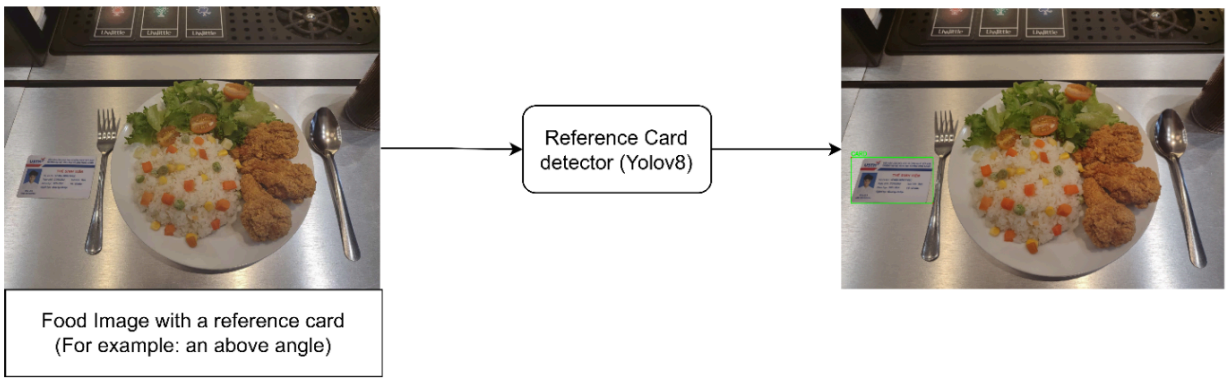

Figure 11: Reference card detection by Yolov8

2.4 Internal Evaluation Metrics

2.4.1 Confusion Matrix

The confusion matrix is a fundamental tool in evaluating the performance of classification models. By providing a tabular representation of actual versus predicted class labels, the confusion matrix shows a clear understanding of how well the model is performing across different classes. In a binary classification scenario, the confusion matrix typically consists of four cells: True Positive, True Negative, False Positives, and

False Negative [11].

In multi-class classification, where there are more than two classes, the confusion matrix extends to capture the performance across all classes. Each row of the matrix represents the instances in the actual class, while each column represents the instances in the predicted class. By examining the values within the matrix, such as precision, recall, and F1-score, one can gain insights into the strengths and weaknesses of the model's classification abilities across various classes, aiding in model refinement and optimization [11].

2.4.2 mIOU, aAcc, and mAcc of Food Ingredients Segmentation

The performance of the model is assessed using various common metrics, including mIoU (mean Intersection over Union), mAcc (mean accuracy over all classes), and aAcc (average accuracy over all pixels). mIoU, a standard indicator in semantic segmentation, is employed to evaluate the overlap and union between inference and ground truth. This assessment provides valuable insights into the model's accuracy in delineating semantic boundaries within the dataset under consideration [6].

$$mIOU = \frac{1}{N}\sum_{i=1}^{n}\frac{TPi}{TPi+FPi+FNi}$$

where N is the number of classes and TPi, FPi, and FNi are described as follows:

- True Positive (TPi) represents the number of pixels that are correctly classified as class i.
- False Positive (FPi) denotes the number of pixels that are wrongly classified as class i.
- False Negative (FNi) is the number of pixels that are wrongly classified as other classes while their true labels are class i.

mAcc is the average accuracy of all categories. For a dataset with N classes, it can be formulated as:

$$mAcc = \frac{1}{N}\sum_{i=1}^{n}\frac{TPi}{TPi+FNi}$$

And aAcc directly calculates the ratio of all pixels that are correctly classified, which can be described as:

$$aAcc = \frac{\sum_{i=1}^{N} TPi}{\sum_{i=1}^{N} TPi+FNi}$$

2.4.3 mAP of Food Region Detection

In the structure of FoodSAM, an object detection model (Unified_learned_COI_RS200_6x) was used to separate food and non-food regions in an image, such as background, tablewares and the food dishes. To evaluate the model, mAP (mean Average Precision) was used to test the performance of object detection models by calculating the average precision across all object classes. mAP quantifies the overall accuracy of an object detection model by considering both the precision and recall of detected objects across different classes and confidence thresholds.

2.4.4 External evaluation metrics – Visualization of Food Ingredients Segmentation

Evaluation on Semantic Segmentation involves testing FoodSAM's proficiency in distinguishing between various food items and background elements within images. This evaluation entails assessing the system's capability to assign the correct category to each pixel in the image, a process that is subsequently visualized for clarity of analysis [6].

Evaluation on Instance Segmentation, FoodSAM's effectiveness is tested in identifying and isolating individual food items within an image. The evaluation scrutinizes whether the system can accurately delineate each item, even when confronted with multiple instances of the same type, thereby gauging its performance in instance-level segmentation [6].

Evaluation on Panoptic Segmentation section examines FoodSAM's comprehensive analysis of both food and non-food items within images. This evaluation encompasses verifying whether the system can accurately label every constituent part of the image, including individual food items and surrounding objects or background elements [6].

3. Results

3.1 Evaluation Metrics of Food Ingredients Segmentation

3.1.1 mIOU, aAcc and mAcc

In Table 1, the outcomes of varying the confidence threshold values on three critical evaluation metrics: mean Intersection over Union (mIoU), mean Accuracy (mAcc), and all Accuracy (aAcc). It is evident from the data that different thresholds impact these metrics in unique ways.

Starting with the mIoU, a key measure of overlap between the model's predictions and the actual data, it is observable that the peak performance is at a 0.5 confidence threshold, indicating this level optimally balances false positives and negatives in segmentation tasks. As the threshold increases or decreases from this midpoint, there is a slight dip in mIoU, suggesting a balanced threshold is crucial for maximizing overlap accuracy.

When considering mAcc, which reflects the precision of class-specific segmentations, the data reveals an optimal performance at a 0.7 confidence threshold. This suggests that a higher threshold, which necessitates greater certainty in pixel classification, enhances the accuracy of specific class identifications within the segments.

Lastly, the aAcc metric, representing the overall pixel accuracy across the entire image, shows a less direct correlation with the confidence thresholds. Interestingly, the table indicates that both the highest and lowest aAcc values correspond to more extreme confidence thresholds (0.3 and 0.9). However, the highest overall accuracy is observed at the lowest threshold (0.3), hinting at a trade-off between overall accuracy and precision in class-specific segmentation.

The choice of an optimal confidence threshold hinges on the specific demands of the segmentation task at hand. If the objective is to maximize general segmentation overlap, a threshold around 0.5 seems advisable. However, for tasks requiring high precision in class-specific identification, a higher threshold like 0.7 may be more appropriate. Finally, if the priority is to ensure overall pixel accuracy, a lower threshold might be beneficial, although it may come at the cost of precision in other metrics. This nuanced performance relationship underscores the importance of tailored threshold selection based on targeted segmentation outcomes.

Table 1: Results of semantic segmentation with dataset FoodSeg103 with different threshold

Conf Threshold	mIoU (%)	mAcc (%)	aAcc (%)
0.3	46.21	57.96	84.48
0.5	46.42	58.27	84.10
0.7	46.19	58.64	84.39
0.9	45.92	57.56	84.02

The table 2 showcases performance results for a panoptic segmentation model across varying confidence thresholds of UECFoodPixComplete. These metrics are pivotal for assessing the model's segmentation effectiveness and precision in classification.

Confidence Threshold (Conf Threshold): The analysis considers four distinct thresholds: 0.3, 0.5, 0.7, and 0.9. These thresholds define the model's confidence level required to assign pixels to specific categories, impacting the overall segmentation performance.

mIoU (mean Intersection over Union): This metric quantifies the average overlap between the model's predicted segments and the actual ground truth, across all classes. The results indicate a peak mIoU at a threshold of 0.5 (66.18%), suggesting that at this point, the balance between sensitivity and specificity in detecting relevant features is optimized. Although there's a slight fluctuation in mIoU values with varying thresholds, they remain relatively high, indicating consistent model performance in segment overlap.

mAcc (mean Accuracy): This reflects the average correctness of the pixel classifications within the identified segments across different classes. The data demonstrates a gradual improvement in mAcc as the threshold increases from 0.3 to 0.7, highlighting that higher thresholds may lead to more precise class-specific segmentations. The peak performance occurs at a 0.7 threshold (78.48%), after which there's a negligible decline, hinting at an optimal balance between excluding false positives and maintaining

true positive rates at this threshold level.

aAcc (all Accuracy): Representing the model's overall pixel accuracy across the entire image, aAcc improves as the threshold increases, reaching its peak at 0.7 (88.20%). This indicates that the model achieves its best overall classification performance at this threshold. Post this peak, there's a minor drop when the threshold is further increased to 0.9, suggesting a marginal overfitting or exclusion of true positives at extremely high confidence levels.

In conclusion, while all metrics show high performance across the board, the model achieves its best balance of segment overlap, class-specific accuracy, and overall accuracy at a confidence threshold of 0.7. This threshold appears to be the sweet spot for optimizing panoptic segmentation performance, providing a guide for setting operational parameters in applications requiring high precision and reliability in image segmentation.

Table 2: Results of semantic segmentation with dataset UECFoodPixComplete with different thresholds.

Conf Threshold	mIoU (%)	mAcc (%)	aAcc (%)
0.3	65.64	77.76	87.56
0.5	66.18	78.11	88.01
0.7	65.88	78.48	88.20
0.9	66.56	77.98	88.04

3.1.2 Evaluation of Reference Card Detection by Yolov8

The confusion matrix is structured around two classes, 'card' and 'background', based on the predictions made by the model versus the actual (true) classifications. The model was tested on a set of 35 images with a total of 69 cards.

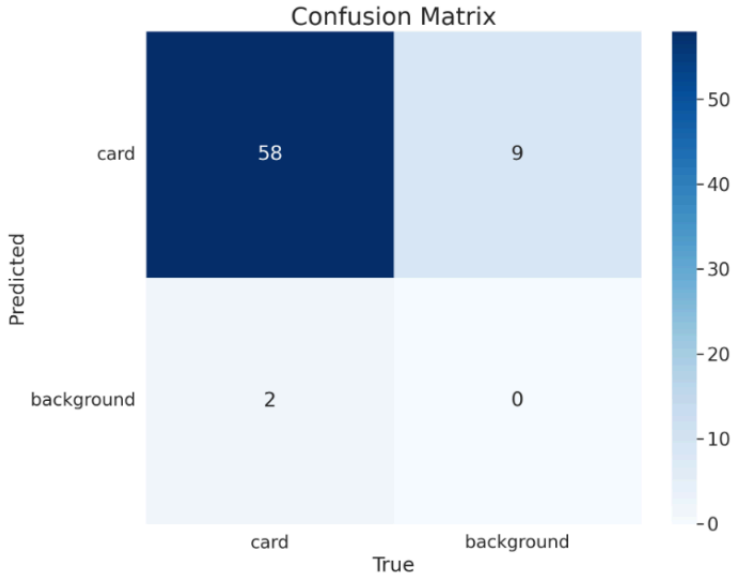

Figure 12: Confusion matrix of Reference Card detection

Table 3: Results of Reference Card detection.

	Precision	Recall	mAP50	mAP50-95	F1 Score	Run-time
Results	0.8775	0.96667	0.90979	0.87548	0.919929	0.1 images per second

According to Table 3, regarding the run-time, the model takes approximately 0.1 images per second. This duration seem acceptable for real-time or rapid processing needs. The explanation provided for this performance is the suboptimal computer configuration, which is a plausible reason. Computing performance, including processing speed, can significantly affect the run-time of machine learning models. Factors such as CPU speed, GPU availability and power, RAM size, and disk speed can all influence how quickly a model processes data.

Precision: This metric, at 0.8775, shows a high level of accuracy in the model's positive predictions. In other words, when the model predicts 'card,' it is correct about 87.75% of the time. This is a good precision rate, suggesting that the model has a low rate of false positives.

Recall: The recall is extremely high at 0.96667, indicating that the model successfully identifies 96.67% of all actual cards. This aligns with the low number of

false negatives in the confusion matrix, demonstrating the model's effectiveness in capturing true positive cases.

mAP50 (mean Average Precision at 50% IoU): This score of 0.90979 is indicative of the model's accuracy and precision in detecting objects (in this case, cards) with an Intersection over Union (IoU) threshold of 50%. A mAP50 score close to 0.91 signifies excellent model performance, especially in tasks requiring binary classification like card detection.

mAP50-95: This is an average of the mean Average Precision calculated at different IoU thresholds, from 50% to 95% (in steps of 5%). The value of 0.87548 here indicates robust model performance across various levels of strictness in evaluating the match between predicted and actual bounding boxes. This high score across a range of IoU thresholds demonstrates the model's reliability and accuracy in localization and classification.

F1 score is 0.9199. This indicates a strong balance between precision and recall, confirming the model's high performance in identifying and classifying cards within the test images. This high F1 score, combined with the other metrics, underscores the model's effectiveness in both detection and precision.

3.2 mAP of Food Region Detection

By using 100k testing images of the food or non-food region detection pre-trained model (Unified_learned_COI_RS200_6x), the mAP was 31.1%. This shows that the object detection model achieved a relatively moderate level of performance in accurately localizing and classifying objects within the dataset's 365 categories, because of some overlapping objects in the test images to each other, different variations of the same object and some occlusions such as lighting and different angles (such as the hand sanitizer in figure 13). This result indicates that there is room for improvement in the model's ability to detect objects across a wide range of classes, given the large-scale nature of the dataset and the diversity of objects it contains.

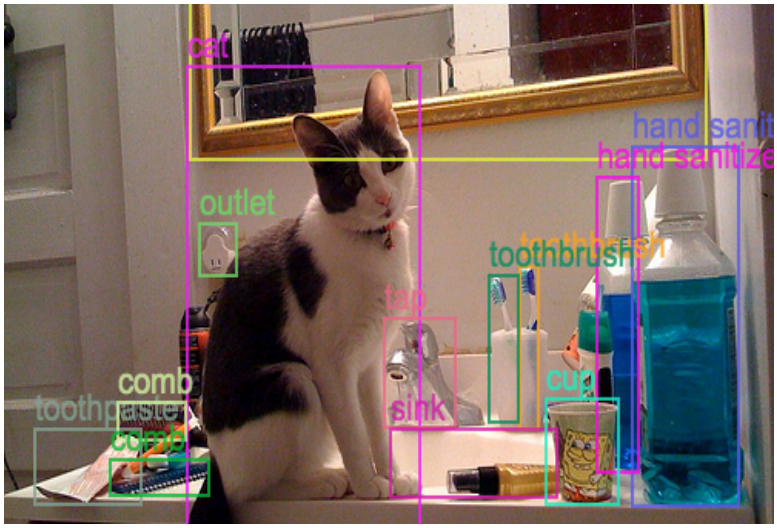

Figure 13: Example of Objects365 testing images with the model's predicted bounding boxes

Though in this project, only tableware labels ("Fork", "Spoon", "Knife") and food labels ("Cake", "Apple", "Grape") are important to detect and extract food regions, tableware regions and the background (figure 14).

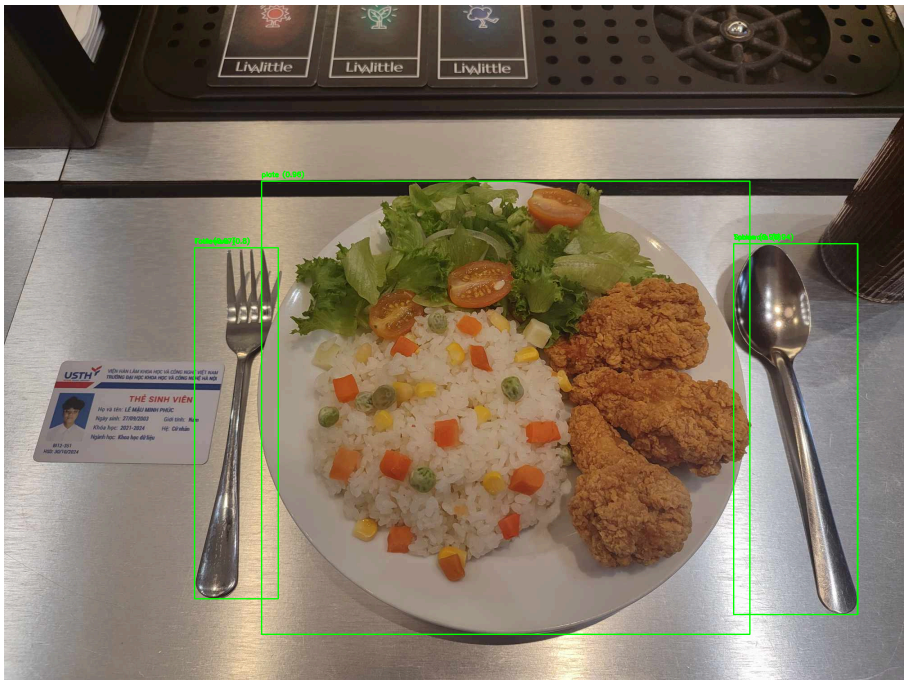

Figure 14: Example of real-life testing image with the model's predicted bounding boxes

3.3 External Evaluation Metrics – Visualization

Input Images

Figure 13: Three example images

3.3.1 Semantic Segmentation (FoodSAM)

First Image: Shows a plate with various food items. Each type of food appears to be segmented with different colors. For example, what might be meat is one color, vegetables another, and possibly sauces or other elements each have their own color. The segmentation allows for distinct identification of each food type based on color coding.

Second Image: Similar to the first, this image displays a container with different food items, each segmented by color. The segmentation seems to distinguish between items like a main dish, side dishes, and condiments. This could be particularly useful for applications like calorie counting, nutritional analysis, or meal planning software.

Third Image: This image again features a segmented plate of food with various components distinctly colored. It shows how semantic segmentation can be applied to complex scenes with multiple food items, which could be overlapping or mixed.

Figure 15: Three example images in Semantic Segmentation

3.3.2 Instance Segmentation

Unlike semantic segmentation, which classifies each pixel into a category, instance

segmentation not only categorizes each pixel but also differentiates between different instances of the same category. Here's what's typically represented in such images:

The first Image: Shows multiple food items on a plate, with each item individually outlined and filled with unique color schemes. Unlike semantic segmentation, here each individual item is distinctly marked, even if they are of the same type. This means, for example, two separate pieces of broccoli would be distinguished from each other, each enclosed within its own contour and color.

Second Image: Displays a more complex scene possibly featuring various containers and food types, again with each instance uniquely identified. You can see bounding boxes around each food item, highlighting the instance segmentation process's ability to not just recognize the item types but to isolate individual entities. This can be particularly useful for detailed analysis, like identifying the count of each type of food or understanding the layout of a meal.

Third Image: Similar to the first two, this image also shows a plate of food with various items. However, each item, even if they are the same type, is individually segmented into distinct entities. This kind of detailed segmentation allows for deeper insights, such as identifying the volume of each food item for nutritional analysis or for advanced culinary applications.

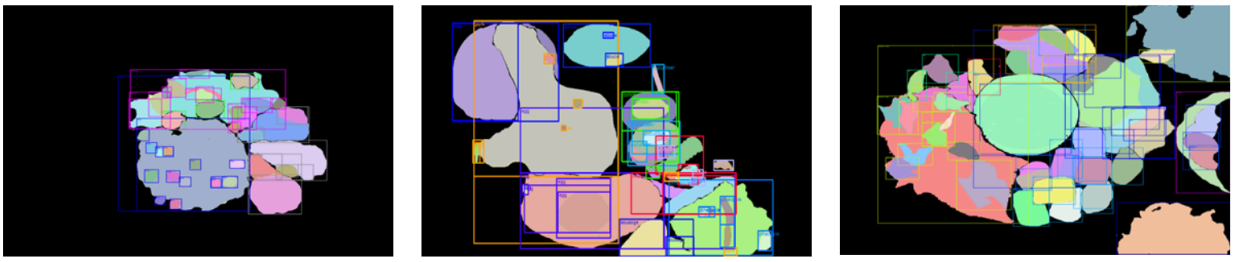

Figure 16: Three example images in Instance Segmentation

3.3.3 Panoptic Segmentation

Panoptic segmentation combines the features of semantic and instance segmentation by classifying every pixel in the image into a category and distinguishing between different instances of the same category.

First Image: This image likely shows a plate with multiple food items, where each item is not only classified (like in semantic segmentation) but also distinctly marked as a

separate instance (as in instance segmentation). The panoptic approach allows for a comprehensive analysis, identifying each piece of food while also understanding the overall composition of the meal. The inclusion of utensils and possibly other non-food elements provides a full picture of the dining setup.

Second Image: This appears to be a more complex scenario, possibly featuring multiple types of containers and food items, each segmented individually. The image demonstrates the panoptic segmentation's capability to provide a detailed breakdown of the scene, marking different food containers, various food items, and possibly other elements of a meal setting. This can be incredibly useful for complex scenes where understanding the context and each item's placement is essential.

Third Image: Similar to the other two, this image provides a detailed view of a meal, with different food items and parts of the meal environment marked and segmented. The diversity of colors and shapes within the segmentation indicates the model's ability to handle varied components and textures, a critical aspect of analyzing real-world meal settings.

Figure 17: Three example images in Panoptic Segmentation

3.3.4 Non-food and Food Segmentation

First Image (Original): This is the original image depicting a plate of food with various items. It's a typical scene one might want to analyze for identifying and differentiating between food and non-food items.

Subsequent Images (Segmentation and Detection): These appear to be binary masks generated as part of the object detection process, specifically aimed at identifying non-food items within the scene.

Second Image: Shows a binary mask of the entire scene. In this mask, objects of interest are likely represented in black against a white background, or vice versa,

depending on the convention used. This could represent an initial segmentation step, highlighting the entire area of interest which includes both food and non-food items.

Third Image: Presents a focused binary mask, likely representing the detection of a non-food item in the scene. Given the context, this could be something like a plate, utensil, or napkin – items that are not edible but typically appear in meal images.

Fourth and Fifth Images: These likely show further detections or a breakdown of different non-food items identified within the original scene. Each mask seems to isolate individual non-food items, separating them from the rest of the scene for individual analysis.

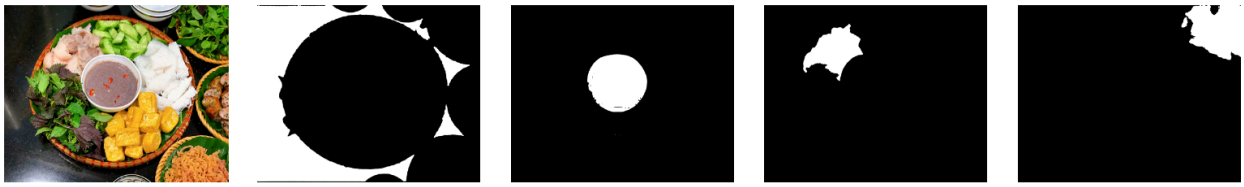

Figure 18: Examples of binary masks in food and non-food segmentation

4. Conclusion

4.1 Discussion

The application of FoodSAM has demonstrated promising results in food ingredient segmentation, addressing both food and non-food segmentation challenges. By integrating the Segment Anything Model (SAM) with additional labeling and segmentation techniques, FoodSAM has enhanced the accuracy of food image analysis. This advancement is particularly significant given the alarming statistics regarding obesity and the necessity for better dietary management tools.

The fusion of high-quality mask generation by SAM and semantic labeling has tackled the inherent challenge of distinguishing complex food items, which vary significantly in appearance and composition. Additionally, the instance and panoptic segmentation capabilities of FoodSAM have provided a more detailed understanding of food images, distinguishing between individual ingredients and contextual items such as utensils and plates. This granularity is crucial for applications like calorie estimation, where the size and type of food directly impact nutritional calculations.

However, while FoodSAM represents a substantial step forward, challenges remain. The technology's performance can vary based on the diversity of food presentations and the quality of the input images. Lighting, shadows, and overlapping ingredients can still pose difficulties for accurate segmentation. Furthermore, the reliance on users to provide images with a reference card for size estimation may limit the spontaneity and convenience of using the application in real-life settings.

4.2 Future Work

To enhance FoodSAM and its applications, there are several key areas for development. Firstly, the refinement of the SAM and FoodSAM algorithms is essential. This involves incorporating advanced deep learning techniques and data augmentation strategies to better handle various food presentations, lighting conditions, and camera angles. Secondly, making the image capture and analysis process more user-friendly is crucial. Future versions could include automatic detection of common objects, like coins or plates, as reference markers, eliminating the need for a physical reference card.

Expanding the dataset used for training is another important step. By including a more diverse range of foods, cooking styles, and presentation methods from different cultures, the model's global relevance will be significantly enhanced. Additionally,

making FoodSAM more interactive and personalized is key. This could mean allowing users to adjust segmentation results, which also helps in refining the model. Tailoring dietary recommendations to individual health profiles and nutritional needs would make the app more user-specific.

Finally, integrating FoodSAM with health and nutrition apps would offer users immediate feedback on their meals and personalized dietary advice. Moreover, researching the impact of non-food items in images on eating habits and dining environments can provide insights into promoting healthier eating behaviors. These advancements collectively aim to make FoodSAM a more robust, user-friendly, and globally applicable tool for dietary management.

5. Reference

[1] Hurt, R. T., Kulisek, C., Buchanan, L. A., & McClave, S. A. (2010). *The obesity epidemic: challenges, health initiatives, and implications for gastroenterologists*. National Library of Medicine. Retrieved from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3033553/

[2] Reis, D., Kupec, J., Hong, J., & Daoudi, A. (2023). *Real-Time Flying Object Detection with YOLOv8*. ArXiv.org. https://arxiv.org/abs/2305.09972

[3] Allegra, D., Battiato, S., Ortis, A., Urso, S., & Riccardo Polosa. (2020). A review on food recognition technology for health applications. *Health Psychology Research*, *8*(3). https://doi.org/10.4081/hpr.2020.9297

[4] Zhang, C., Liu, L., Cui, Y., Huang, G., Lin, W., Yang, Y., & Hu, Y. (2023). *A Comprehensive Survey on Segment Anything Model for Vision and Beyond*. ArXiv. https://arxiv.org/abs/2305.08196

[5] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). *Segment Anything*. ArXiv. https://arxiv.org/abs/2304.02643

[6] Lan, X., Lyu, J., Jiang, H., Dong, K., Niu, Z., Zhang, Y., & Xue, J. (2024). FoodSAM: Any Food Segmentation. *IEEE Transactions on Multimedia*, 1–14. https://doi.org/10.1109/tmm.2023.3330047

[7] Okamoto, K., & Yanai, K. (2016). *An Automatic Calorie Estimation System of Food Images on a Smartphone*. ResearchGate; https://www.researchgate.net/publication/309128551_An_Automatic_Calorie_Estimation_System_of_Food_Images_on_a_Smartphone

[8] *SA-1B Dataset*. (2023). Meta. https://ai.meta.com/datasets/segment-anything/

[9] Wu, X., Fu, X., Liu, Y., Lim, E.-P., Steven, & Sun, Q. (2021). *A Large-Scale Benchmark for Food Image Segmentation*. ArXiv. https://arxiv.org/abs/2105.05409

[10] Shao, S., Li, Z., Zhang, T., Peng, C., Yu, G., Zhang, X., Li, J., & Sun, J. (2019). *Objects365: A Large-Scale, High-Quality Dataset for Object Detection*. https://doi.org/10.1109/iccv.2019.00852

[11] Damir, K., Maja, B., Ljiljana, Š., & Dunja, B.Š. (2020). *Multi-label Classifier Performance Evaluation with Confusion Matrix*. ResearchGate; https://www.researchgate.net/publication/342873309_Multi-label_Classifier_Performance_Evaluation_with_Confusion_Matrix

6. Appendix

For further discussion about FoodSAM evaluation on FoodSeg103, with the confidence threshold of 0.5, these are the result of per class label:

Class	IoU	Acc	Precision	Recall	F1
0	94.83	97.56	97.14	97.35	97.35
1	58.1	80.24	67.8	73.5	73.5
2	0.0	0.0	0.0	0.0	0.0
3	75.04	90.83	81.19	85.74	85.74
4	22.14	49.61	28.56	36.25	36.25
5	51.25	63.1	73.17	67.77	67.77
6	68.07	68.68	98.72	81.0	81.0
7	0.0	0.0	0.0	0.0	0.0
8	47.71	63.37	65.89	64.6	64.6
9	17.2	31.85	27.22	29.36	29.36
10	44.0	58.9	63.5	61.11	61.11
11	29.8	46.25	45.59	45.91	45.91
12	37.57	48.52	62.48	54.62	54.62
13	41.95	81.86	46.25	59.1	59.1
14	41.56	56.03	61.67	58.72	58.72
15	42.56	85.41	45.9	59.71	59.71
16	3.34	54.07	3.44	6.46	6.46
17	48.69	65.35	65.63	65.49	65.49
18	17.97	32.73	28.5	30.47	30.47
19	10.16	13.54	28.94	18.45	18.45
20	25.31	46.31	35.81	40.39	40.39
21	36.46	86.4	38.68	53.44	53.44
22	57.95	84.62	64.77	73.37	73.37
23	0.0	0.0	0.0	0.0	0.0
24	57.25	64.39	83.78	72.81	72.81
25	33.77	50.82	50.17	50.49	50.49
26	38.31	40.46	87.85	55.4	55.4
27	27.94	37.32	52.65	43.68	43.68
28	66.19	87.5	73.1	79.65	79.65
29	88.22	94.16	93.32	93.74	93.74
30	83.01	87.56	94.11	90.72	90.72
31	44.96	73.73	53.54	62.03	62.03
32	72.45	80.52	87.85	84.02	84.02
33	46.65	64.18	63.07	63.62	63.62
34	36.74	56.05	51.6	53.73	53.73

35	49.69	58.45	76.83	66.39	66.39
36	68.43	83.06	79.53	81.25	81.25
37	76.29	84.61	88.58	86.55	86.55
38	22.3	29.63	47.41	36.47	36.47
39	66.96	77.45	83.17	80.21	80.21
40	47.08	60.32	68.2	64.02	64.02
41	75.27	91.6	80.85	85.89	85.89
42	89.92	92.79	96.67	94.69	94.69
43	3.8	9.62	5.92	7.33	7.33
44	54.33	76.15	65.46	70.4	70.4
45	50.65	71.58	63.4	67.25	67.25
46	48.27	65.46	64.77	65.11	65.11
47	20.96	33.33	36.1	34.66	34.66
48	51.17	64.52	71.2	67.7	67.7
49	31.31	41.77	55.57	47.69	47.69
50	28.15	60.61	34.46	43.94	43.94
51	23.68	45.2	33.21	38.29	38.29
52	44.52	59.21	64.22	61.61	61.61
53	41.78	92.47	43.26	58.94	58.94
54	39.66	66.94	49.32	56.79	56.79
55	78.29	81.36	95.41	87.82	87.82
56	74.05	87.12	83.16	85.09	85.09
57	30.5	45.5	48.04	46.74	46.74
58	57.19	75.59	70.15	72.77	72.77
59	87.43	94.6	92.02	93.29	93.29
60	0.0	0.0	0.0	0.0	0.0
61	25.78	49.98	34.75	40.99	40.99
62	3.81	5.66	10.47	7.35	7.35
63	23.09	31.61	46.16	37.52	37.52
64	64.9	76.15	81.47	78.72	78.72
65	62.06	76.1	77.09	76.59	76.59
66	81.82	88.45	91.61	90.0	90.0
67	50.22	67.63	66.12	66.86	66.86
68	52.72	79.21	61.19	69.04	69.04
69	1.32	21.11	1.39	2.61	2.61
70	63.93	77.22	78.8	78.0	78.0
71	43.02	56.71	64.06	60.16	60.16
72	76.64	85.7	87.89	86.78	86.78
73	79.53	87.17	90.07	88.6	88.6
74	0.0	0.0	0.0	0.0	0.0
75	83.62	95.41	87.12	91.08	91.08

76	31.75	62.81	39.11	48.2	48.2
77	24.04	34.5	44.23	38.76	38.76
78	87.62	95.44	91.45	93.4	93.4
79	77.19	92.66	82.22	87.13	87.13
80	52.08	72.59	64.82	68.49	68.49
81	55.61	71.62	71.33	71.48	71.48
82	72.63	83.21	85.11	84.15	84.15
83	28.87	63.58	34.58	44.8	44.8
84	83.93	89.74	92.84	91.26	91.26
85	77.55	86.93	87.78	87.35	87.35
86	0.0	0.0	0.0	0.0	0.0
87	88.92	91.83	96.56	94.13	94.13
88	69.02	79.76	83.67	81.67	81.67
89	58.16	68.39	79.53	73.54	73.54
90	64.84	76.54	80.92	78.67	78.67
91	54.35	74.22	67.0	70.42	70.42
92	54.21	77.2	64.54	70.31	70.31
93	48.08	63.51	66.45	64.94	64.94
94	60.13	74.97	75.24	75.1	75.1
95	86.31	94.41	90.97	92.65	92.65
96	75.49	80.77	92.03	86.03	86.03
97	0.0	0.0	0.0	0.0	0.0
98	32.23	55.24	43.62	48.74	48.74
99	51.79	74.08	63.25	68.24	68.24
100	0.0	0.0	0.0	0.0	0.0
101	26.59	37.48	47.78	42.01	42.01
102	2.84	18.85	3.24	5.52	5.52
103	3.35	9.42	4.95	6.49	6.49

The evaluation metrics' average value of all classes:

Scope	IoU	Acc	Precision	Recall	mAcc	aAcc
global	46.24	60.76	58.17	58.28	58.27	84.10

A closer analysis of the results indicates a high degree of variability across classes. For example, class 0, which represents "background" shows exemplary performance with an IoU of 94.83 and an F1 Score of 97.35, likely due to the distinct and consistent nature of the background in food images. In contrast, classes such as "candy" (class 1) and "chocolate" (class 4) demonstrate moderate performance, which may be attributed to

variations in shape and size, as well as potential similarities in color and texture to other ingredients.

Furthermore, some classes such as “egg tart” (class 2) and “pudding” (class 7) have zero values across all metrics, suggesting that the model failed to recognize these classes within the dataset. This could be due to a lack of representative training data or the complexity of distinguishing these items from similar classes.

On the other end, items like “popcorn” (class 6) and “french fries” (class 3) exhibit strong Recall and Precision, indicating the model's capability to identify and delineate these items accurately. The high performance on 'popcorn' can be especially highlighted with an F1 Score of 81.0, underscoring the model's strength in recognizing texturally distinct food items.

The overall global scores, with an IoU of 46.24 and an Accuracy of 60.76, suggest there is substantial room for improvement. The considerable diversity in the performance of individual classes may reflect the model's sensitivity to the inherent diversity present in food items, such as color, texture, and context within the image.