

FOOD INSTANCE SEGMENTATION

TszChoi Siu

University of Las Vegas, Nevada

ABSTRACT

Accurate calorie estimation from images requires both recognizing the food item and determining its true physical size, something LLMs alone cannot reliably infer. To address this, we combine YOLO11x-seg segmentation with LLM-based analysis. A standard reference card is placed beside the food, allowing YOLO to segment both objects and compute real-world dimensions using the known card size. The cropped food region and its measured size are then passed to an LLM, which identifies the food and estimates calories without guessing portion size. This integrated approach offers more accurate and interpretable calorie estimation than LLM-only methods.

Index Terms— One, two, three, four, five

1. INTRODUCTION

Recognizing food items in images is an essential step toward automatic calorie estimation and personalized nutrition guidance. While recent applications often rely on large language models (LLMs) to describe food and estimate its size, these models can only guess physical dimensions from visual context. As a result, their portion-size predictions are frequently inaccurate—even when the food type is correctly identified. Because calorie estimation depends directly on portion size, a more reliable method must incorporate real-world size measurements, not just semantic understanding.

In this work, we explore a food segmentation system built on YOLO11x-seg architecture. Compared to transformer-based vision models, which are powerful but resource intensive, YOLO offers a lightweight and practical alternative for real-time food analysis.

Our approach adds a simple but highly effective idea: placing a standard-sized reference card next to the food. YOLO simultaneously segments the card and the food, allowing us to compute the food’s real-world dimensions using the known size of the card. After segmentation, the system crops out just the food region and sends it to an LLM. Because we already know the physical size of the food, the LLM no longer needs to guess; instead, it focuses

only on identifying the type of food and estimating calories based on the true measured size.

This combination—precise segmentation, geometric size estimation, and semantic interpretation from an LLM—enables a significantly more accurate and reliable calorie estimation pipeline.

2. METHODS

2.1. Input

Preprocessing. There are two sources: a food dataset and a custom card dataset. From food dataset, we consolidated all food categories into a single class. For the card dataset, we selected only scenes containing exactly one visible card instance to ensure clean supervision. The original 52-class card masks collapsed into a single label.

Label. All masks were converted into YOLO-style polygon annotations to enable efficient training with the YOLO11x-seg architecture. In the final dataset, class labels were defined as: 0 = food and 1 = card. This preprocessing pipeline ensures consistent and high-quality annotations across both domains, enabling a clean two-class segmentation task.

2.2. Segmentation

We employ the YOLO11x-seg model for segmentation tasks. The model performs forward inference on scenes containing both the food item and the reference card. The network outputs object masks together with bounding boxes and class predictions. These masks provide the pixel-accurate boundaries necessary for precise size measurement downstream.

2.3. Measurement and Calorie Estimation

Real-World Measurement. Once food and card regions are segmented, we compute physical dimensions using the known size of the reference card. YOLO’s predicted bounding boxes supply pixel measurements for the card, which are compared to its true height to derive a pixel-to-millimeter scale factor. Applying this scale to the food’s

bounding box yields estimates of the food’s real-world width and height.

Calorie Estimation. After measurement, the segmented food region is cropped and passed to a large language model through an API interface. The LLM identifies the food item, characterizes its state (e.g., cooked or raw), and estimates calorie content using the true geometric dimensions provided by the segmentation module.

2.4. Potential Application

Offloading semantic interpretation to the LLM while retaining measurement accuracy through YOLO segmentation produces a more reliable and interpretable calorie-estimation pipeline. This integration of segmentation, physical measurement, and high-level reasoning supports practical applications such as calorie-tracking and dietary monitoring. Beyond nutrition, the same approach can be extended to any task requiring real-world size estimation from images, including object measurement. Combining YOLO’s pixel-level precision with the semantic capabilities of modern language models provides a flexible foundation for a wide range of vision-based applications.

3. EXPERIMENTAL SETUP

3.1. Datasets and Preprocessing

The dataset used for training is by merging two sources: FoodSeg103 and a custom-built card dataset for reference-object segmentation. All food categories in FoodSeg103 and all card identities in the card dataset were collapsed into two unified labels, forming a clean binary distinction between food and the reference card. Each mask was converted from its original dense segmentation format into YOLO-style polygon annotations to be compatible with the YOLO11x-seg training pipeline. To guarantee reliable supervision from the card dataset, only scenes containing exactly one visible card instance were retained. Both datasets were organized into a consistent directory structure and assigned the same class definitions: class 0 for food and class 1 for cards. The training and validation sets follow a roughly 60:40 ratio, matching the default split of FoodSeg103. For the card dataset, this same ratio was manually enforced. Once combined, the dataset was augmented during training with standard techniques—including horizontal flips, color jittering, and image resizing—to improve generalization and robustness across diverse lighting and scene conditions.

3.2. Food Recognition

We employ the YOLO11x-seg architecture for two-class semantic segmentation, distinguishing food regions and reference card regions. The model is initialized with pretrained YOLO11x-seg weights and fine-tuned on the

merged dataset. All images are resized to 640×640 pixels and annotated using YOLO polygon-based segmentation labels. Training is performed for 10 epochs with a batch size of 32. The network contains 203 layers and 62 millions parameters. Optimization follows the Ultralytics training pipeline, which applies AdamW with built-in learning-rate scheduling and regularization. During validation, the model is evaluated on 2,468 images using standard segmentation metrics.

Class	Images	Box mAP50	Box mAP50-95	Mask mAP50	Mask mAP50-95
all	2468	0.951	0.888	0.949	0.858
food	2135	0.907	0.804	0.903	0.737
card	333	0.995	0.972	0.995	0.979

Fig. 1. Validation metrics of the YOLO11x-seg model on the merged dataset.

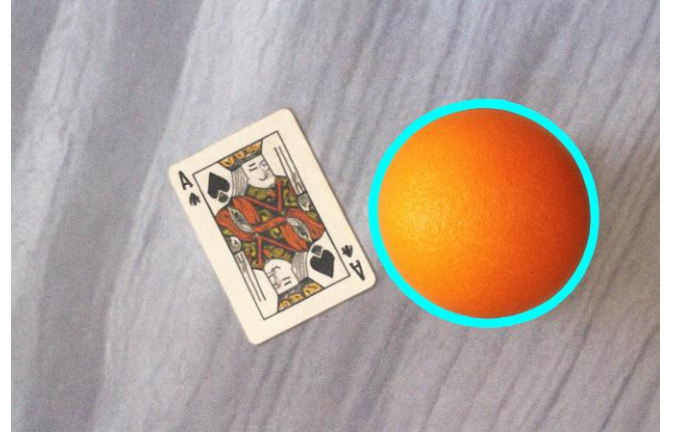


Fig. 2. Input image containing both the reference card and the food item.

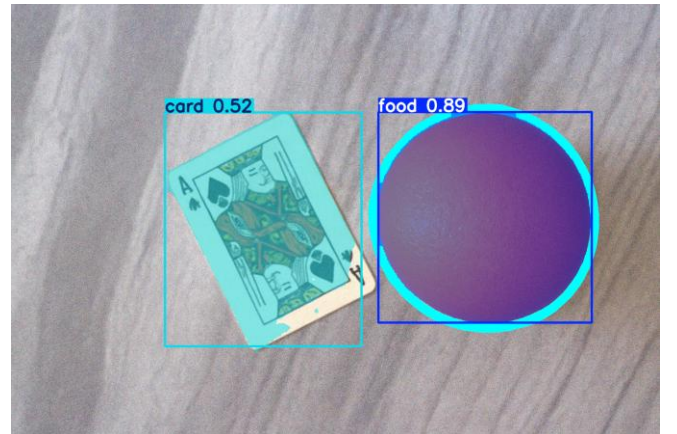


Fig. 3. YOLO11x-seg detection and segmentation output for food and card.



Fig. 4. Cropped food mask used for size estimation and LLM calorie prediction.

4. RESULT

The proposed segmentation system was evaluated on the merged two-class dataset using the YOLO11x-seg validation pipeline. Overall validation metrics indicate strong performance across both classes, with high box- and mask-level mAP scores (Fig. 1).

To demonstrate qualitative behavior, Fig. 2 shows a test image containing a standard playing card placed beside an orange. The model’s segmentation output, illustrated in Fig. 3, correctly identifies both objects, assigning the labels card and food with confidences of 0.52 and 0.89, respectively. The isolated food crop produced from the segmentation mask is shown in Fig. 4, which is subsequently used for LLM-based recognition and caloric estimation.

The geometric measurement procedure was validated using the same example. Based on YOLO’s predicted bounding boxes, the card measures 464.3 px in width and 552.3 px in height, while the food region spans 505.2 px by 497.9 px. Given the known physical height of a standard playing card (88 mm), a pixel-to-millimeter scale factor of approximately 0.159 mm/px is derived. Applying this factor to the food’s bounding box yields estimated dimensions of 80.5 mm \times 79.3 mm. These measurements closely match the expected size range of a medium orange, demonstrating that the segmentation-based scaling method can successfully recover realistic real-world dimensions from a single image.

Finally, the cropped food region and its measured physical dimensions were provided to a large language model for semantic analysis. Because the LLM receives accurate geometric information produced by YOLO—rather than inferring scale from visual context alone—its calorie predictions become more stable and more consistent with standard portion-size references.

5. CONCLUSION

We present a food-measurement framework that integrates segmentation-based size estimation with LLM-powered semantic analysis to improve accuracy in calorie prediction. By jointly detecting food items and a reference card using YOLO11x-seg, the system recovers real-world dimensions directly from a single image, reducing the reliance on visual scale guessing and improving reliability in portion estimation. The measured food region is then analyzed by an LLM, resulting in more stable and interpretable nutritional assessments.