

Article

Zero-Shot SAM for Food Image Segmentation

Saeed S. Alahmari ^{1,†} , Michael R. Gardner ^{2,*,†}  and Tawfiq Salem ³¹ Department of Computer Science, Najran University, Najran 66462, Saudi Arabia; ssalahmari@nu.edu.sa² Department of Biomedical Engineering, College of Engineering, King Faisal University, Al Ahsa 31982, Saudi Arabia³ Department of Computer and Information Technology, Purdue University, West Lafayette, IN 47906, USA; tsalem@purdue.edu

* Correspondence: mgardner@kfu.edu.sa

† These authors contributed equally to this work.

Abstract

Recent advances in foundation models have enabled strong zero-shot performance across vision tasks, yet their effectiveness for fine-grained domains such as food image segmentation remains underexplored. This study investigates the zero-shot capabilities of the Segment Anything Model (SAM) for segmenting food images. We evaluate two prompting strategies—bounding boxes and coordinate points—and compare SAM's results against established deep learning baselines, U-Net and DeepLabv3+, both trained on the FoodBin-17k dataset with pixel-wise binary annotations. Using bounding box prompts, zero-shot SAM achieved a Dice coefficient of 0.790, closely matching U-Net (0.759), while DeepLabv3+ attained 0.918. Coordinate-based prompting yielded comparable results, with the optimal configuration (75 coordinate points) also reaching a Dice coefficient of 0.790. These results highlight SAM's strong potential for generalizing to novel visual categories without task-specific retraining. This study further provides insights into how prompt type influences segmentation quality, offering practical guidance for applying SAM in food-related and other fine-grained segmentation applications.

Keywords: food computing; foundation models; SAM; deep learning

Academic Editors: Licheng Liu, Yunyi Li and Bing Tu

Received: 15 September 2025

Revised: 23 October 2025

Accepted: 27 October 2025

Published: 3 November 2025

Citation: Alahmari, S.S.; Gardner, M.R.; Salem, T. Zero-Shot SAM for Food Image Segmentation. *Electronics* **2025**, *14*, 4316. <https://doi.org/10.3390/electronics14214316>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

In recent years, the development of deep learning models has significantly advanced the field of image segmentation, achieving performance that rivals human capabilities [1,2]. This remarkable progress has generated excitement across various domains, such as food computing and kitchen robotics automation [3,4]. Nevertheless, as these models continue to improve in terms of accuracy and efficiency, questions have arisen regarding their reliability [5], as well as the substantial expenses involved in data collection, annotation, and model training [6].

One domain where these concerns resonate strongly is in the segmentation of food images [7,8]. Food segmentation presents unique challenges compared to other domains due to its high intra-class variability, frequent occlusions, and complex visual mixtures. The same type of food may appear vastly different depending on preparation, lighting, or viewpoint (e.g., pasta dishes with different sauces or shapes), while meals often include multiple overlapping or blended food items on a single plate. Additionally, reflective surfaces, varied textures, and lack of distinct boundaries make pixel-level segmentation especially difficult. These factors contribute to a lack of generalizability for conventional models trained on limited datasets, underscoring the need for adaptable segmentation approaches.

While deep learning models exhibit remarkable prowess in segmenting images within the scope of their training data and associated datasets, questions persist about their ability to generalize when confronted with previously unseen food image datasets. Can zero-shot segmentation methods demonstrate results reasonably similar to models trained on closely associated datasets?

Zero-shot segmentation refers to the ability to segment objects in new images that have not been encountered during the training of the base model. This approach eliminates the need for task-specific fine-tuning, which can be time-consuming and computationally intensive. Instead, it leverages pre-trained foundation models (FMs) to perform new tasks using prompts that guide the segmentation process. Such zero-shot capability has been increasingly applied in vision foundation models to generate segmentation masks quickly without retraining [9].

The Segment Anything Model (SAM) has demonstrated a remarkable ability to “segment anything” [10]. For example, implementations of SAM have successfully segmented medical images [11–14] and food images [15,16]. Zero-shot SAM has also been explored in medical [17,18] and remote sensing [19] domains. However, to our knowledge, zero-shot SAM has not yet been systematically evaluated for food image segmentation, where the challenges of intra-class diversity, occlusion, and complex visual compositions remain unresolved.

This study’s first aim is to assess the segmentation performance of zero-shot SAM on an external food image dataset (i.e., without fine-tuning the original SAM model) in contrast to two leading segmentation models trained with a subset of images from our FoodBin-17k dataset. These zero-shot SAM implementations are performed using two prompting approaches: (1) bounding boxes and (2) coordinate points.

Moreover, to the best of our knowledge, there has been no published evaluation of the impact of the number of prompt coordinates on SAM’s performance in the food image segmentation domain. It is possible that the number of prompt coordinates may be optimized for food images. Thus, the second aim of this study is to analyze the effect of the number of coordinates on the accuracy of zero-shot SAM.

In this paper, we explore whether the segmentation capabilities demonstrated by state-of-the-art models, such as SAM in a zero-shot setting, can be effectively applied to food image segmentation. We aim to contribute valuable insights to the ongoing discourse surrounding the practicality and real-world applicability of deep learning models in segmentation and analysis tasks, particularly in domains characterized by dynamic and evolving food datasets.

The main contributions of this work are as follows:

- A new dataset, FoodBin-17k, comprising binary pixel-wise labels for food images, providing a new benchmark for food image semantic segmentation research.
- An investigation of zero-shot SAM with two prompting mechanisms: bounding box and coordinate points.
- Analysis of the effect of prompt quantity (number of coordinate points) on the accuracy of zero-shot SAM.
- Quantitative and qualitative comparison of zero-shot SAM against trained segmentation models (U-Net and DeepLabv3+) on FoodBin-17k.

2. Background

With the development of social media and the Internet of Things (IoT), people share images and videos of their food, which has generated large-scale food datasets offering rich information for understanding culture, facilitating automation of dietary monitoring, and automating food preparation and quality check [20–22]. Food computing aims to

study and analyze heterogeneous food data for many purposes and involves various operations including perception, retrieval, detection, and segmentation [23–25]. Food image segmentation aims to separate food image pixels from the background for the purpose of identifying, retrieving, or grasping using robot arms, and nutritional value estimation [26]. A major obstacle for food image segmentation is the lack of enough data with segmentation masks for training supervised artificial deep neural networks and the cost of obtaining the masks manually. To address this issue, different approaches have been proposed such as using synthetic food images and masks [3], unsupervised learning [27], sequential transfer learning [28], zero-shot and few-shot approaches [29], and foundation models such SAM [10].

In this research, we focus on three deep learning models, U-Net, DeepLabv3+, and SAM, for food image segmentation.

U-Net is a widely used convolutional neural network (CNN) architecture that has demonstrated exceptional performance in various image segmentation tasks [30]. Originally introduced for biomedical image segmentation, the U-Net architecture employs an encoder–decoder structure with skip connections. The encoder part captures hierarchical features through down-sampling layers, while the decoder part restores the spatial resolution of the feature maps through up-sampling layers. The skip connections enable the model to fuse high-resolution features from the encoder with low-resolution features from the decoder, enhancing the network’s ability to preserve detailed information during segmentation. U-Net has proven effective in semantic segmentation tasks and has served as a foundational inspiration for subsequent network architectures.

The DeepLabv3+ [31] model employs an encoder–decoder architecture based on a fully convolutional neural network, utilizing its predecessor (DeepLabv3) [32] as its encoder and Xception as its backbone. DeepLabv3+ employs the Atrous Spatial Pyramid Pooling (ASPP) structure [33] in its encoder. The ASPP structure integrates three parallel atrous convolutions with dilation rates of 6, 12, and 18, expanding the receptive field to capture a more comprehensive context of root information. Introducing the concept of depthwise separable convolution, the model reduces parameters, enhancing both speed and classification performance. To unify the multi-scale spatial information from ASPP, the model applies a concatenation of the learned feature maps. A Channel compression is also applied by a 1×1 convolution operation, further reducing network dimensionality and computation time. Ultimately, the encoder produces a root feature map that is 16 times smaller than the input image.

The Segment Anything Model (SAM) [10] was originally designed as a foundational model for segmenting user-defined objects in natural images, leveraging a vast dataset of over 1 billion annotations. While SAM exhibits impressive performance on natural images, its adaptability to diverse domains, such as food images, presents unique challenges.

Zero-shot learning is an approach that aims to use the knowledge that was learned during training the base model from different object categories that are not available as part of the current training set of the target task [34]. This approach has been used in many recent foundation models and has shown good zero-shot performance, including contrastive language image pretraining (CLIP) and SAM [10,35,36]. Although both models are considered foundational vision models, their performance in new images requires investigation. For instance, in food images, images exhibit large variations in textures, shape, color, and compositionality, which pose a challenge for such foundation models.

One paper [17] studied the zero-shot capabilities of SAM on the segmentation of abdominal organs medical images, where they showed that SAM did not illustrate state-of-the-art performance, but it provided a cheap and quick segmentation as a starting point. Another approach proposed to use zero-shot SAM for potato leaves, which showed a

good performance of segmentation with little cost compared to fine-tuning models [37]. More recent work uses SAM with zero-shot capabilities for crowd counting in images [38], object-pose estimation [39], and combined zero-shot capabilities of SAM and CLIP for medical image segmentation [18].

In this study, we conducted a comprehensive assessment of the zero-shot capabilities of SAM in food image segmentation with different sizes and types of prompts. Employing standard methods for simulating interactive segmentation, we generated point and box prompts tailored to the characteristics of food-related objects. The objective was to understand the effectiveness of segmenting these objects in a manner that aligns with the intricacies of food images.

3. Datasets

To establish a large dataset of binary pixel-wise labels of food images (food/not food), we combined two datasets of labeled food images, FoodSeg103 (FS103) and UECFoodPix-Complete (UEC), to create a single dataset. The FS103 dataset [40] contains 7118 images of food(s) with 104 class-type pixel-wise labels (e.g., bread, strawberry, shellfish, ice cream) stored as an 8-bit integer in a grayscale image. The UEC dataset [41] contains 10,000 images of food with 102 class-type pixel-wise labels stored as an 8-bit integer in the red plane of an RGB image. We simplified the pixel-wise labels for each image by converting all labeled food pixel to binary 1 (food) and all background pixels to binary 0 (not food). See Table 1.

Table 1. FoodBin-17k is a combination and simplification of two published datasets.

Dataset	Total Images	Labels
FoodSeg103 [40]	7118	104 classes as intensities (8-bit grayscale image)
UECFoodPixComplete [41]	10,000	102 classes as R-plane intensities (8-bit RGB image)
FoodBin-17k (ours)	17,118	2 classes (food/background) as binary image

The images dataset from FS103 and UEC has assorted dimensions that are first cropped to a uniform size as a preparation step. First, the image is re-scaled such that the shorter dimension becomes 224 pixels. Then, the center 224 pixels are selected. Thus, we center-cropped the images and labels to 224×224 -pixel squares as shown in Figure 1. The resulting image and label dataset contains 17,118 pairs of 8-bit RGB images and pixel-wise binary labels. We refer to this dataset as FoodBin-17k, which can be accessed at <https://github.com/mgardner-lab/FoodRecognition> (accessed on 22 October 2025).

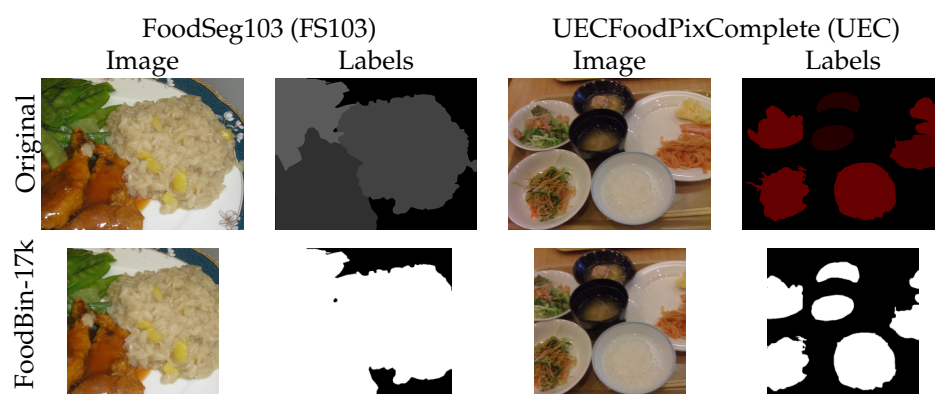


Figure 1. Food images and labels from FS103 and UEC datasets are center-cropped to 224×224 -pixel squares, and the labels are binarized: food (1), not-food (0) for semantic segmentation. The combined dataset has 17,118 8-bit RGB food images and binary pixel-wise labels. The new dataset is called FoodBin-17k.

4. Zero-Shot Learning

Zero-shot-based learning aims to use knowledge accumulated $f : \mathcal{X} \rightarrow \mathcal{Y}$ from large datasets $\mathcal{D}$, where $(x_1, y_1), \dots, (x_n, y_n) \in \mathcal{D}$, and $\mathcal{Y} = \{y_1, y_2, \dots, y_n\}$ used to train the foundation model $f : \mathcal{X} \rightarrow \mathcal{Y}$ to identify new and different objects categories $\mathcal{N} = \{n_1, n_2, n_3, \dots, n_k\}$ that have not been encountered during training of the foundation model f . That is, $\mathcal{N} \not\subseteq \mathcal{Y}$. This approach uses attribute-based classification and learning, which can be divided into two categories: indirect attribute classification and direct attribute classification. The use of semantic per-class attribute s for each class $y \in \mathcal{Y}$ and $n \in \mathcal{N}$ either directly or indirectly enables the use of foundation model f , enables learning attribute representation $a \in \mathcal{A}$, that can be learned using a classifier $\alpha : \mathcal{X} \rightarrow \mathcal{N}$.

With a zero-shot approach, prompts can be used to help guide the foundation model to obtain better results. For instance, for SAM [10], prompts such as points and bounding boxes can be used effectively for improving the segmentation results. This approach provides a segmentation mask in a cheap and effortless way for images where annotations are not available to obtain for training a supervised deep learning algorithm.

5. Method

This study's first aim is to apply zero-shot SAM [10] for food image semantic segmentation. Secondly, we trained deep neural networks, U-Net and DeepLabv3+, to compare their performance to zero-shot SAM. Thirdly, we studied the effect of prompt type and prompt size on zero-shot SAM for food image segmentation. We trained both U-Net and DeepLabv3+ deep neural networks with the same image subset from FoodBin-17k. An ancillary goal of training U-Net and DeepLabv3+ networks was to use the segmentation from the best-performing segmentation model (either U-Net or DeepLabv3+) to generate points and bounding box prompts for the SAM deep neural network. Ultimately, we used DeepLabv3+ segmentation masks to obtain SAM prompts due to DeepLabv3+'s superior performance compared to U-Net. For more details about the training, refer to Section 6.

The SAM deep neural network takes the input color images for food and prompts, where the prompts can be either a set of point coordinates or bounding box coordinates of the region of interest. The SAM deep neural network encodes the input image and the prompt with separate encoders. The input image encoding extracts the image embedding, which is used along with the encoding prompt for the decoder. We have used a SAM pre-trained deep neural network called *Sam_ViT_h*. An illustration of our approach is shown in Figures 2 and 3.

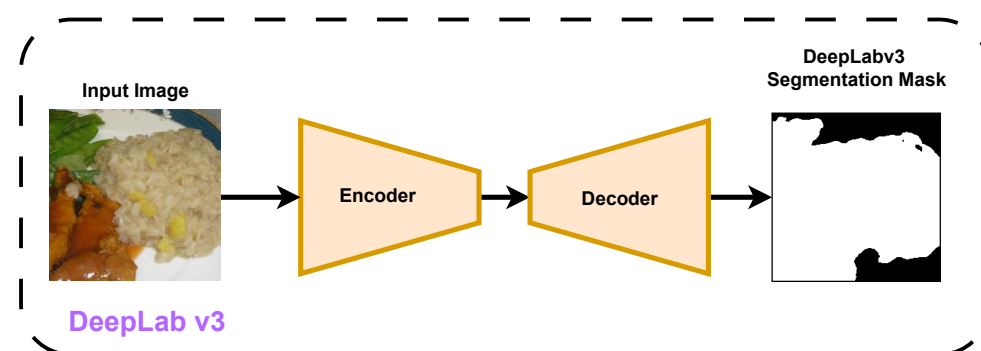


Figure 2. DeepLab v3 model for learning to segment food images. The predicted masks are used for the SAM prompt generation step.

For zero-shot SAM with point coordinate prompting, we used DeepLabv3+ segmentation output to generate coordinates. One hundred five pixel coordinates with binary value '1' were selected with a weighted probability of section—a greater probability of selection

given to pixels with a greater likelihood of being ‘food’ (as determined by DeepLabv3+). These coordinates serve as the prompt encoder for the SAM deep neural network. We performed SAM segmentation using a range of numbers of coordinates, starting at 15, and increasing by 15 up to 105. The goal of selecting different numbers of coordinates is to compare the impact of the number of prompts on SAM segmentation results. This coordinate prompt encoding is the first SAM prompt type we used using a zero-shot approach.

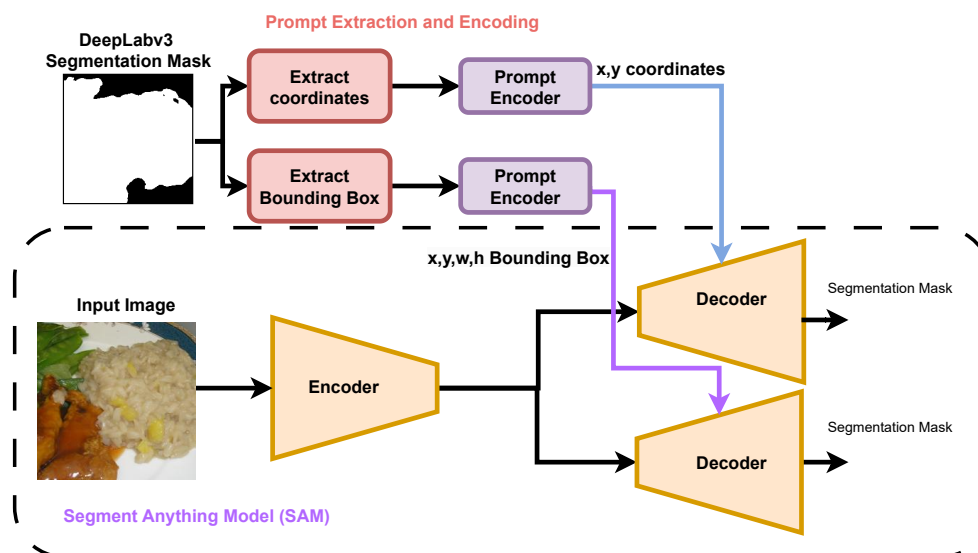


Figure 3. The zero-shot approach for food image segmentation, where the DeepLab v3 food images masks are used for generating the coordinates and bounding box prompt for the SAM model.

The second approach was zero-shot SAM with bounding box prompting. We computed the bounding box coordinates from masks generated by DeepLabv3+ using the OpenCV findContours function. That is, the four extreme Cartesian coordinates of the binary food mask are taken as the four corners of the bounding box.

6. Experimental Setup

The cropped input images (RGB) of size $224 \times 224 \times 3$ were used for training DeepLabv3+ and U-Net to obtain the initial masks. For training, validation, and testing of U-Net and DeepLabv3+, the dataset was split into three parts: the training set contained 10,271 (60%) image/mask pairs from FoodBin-17k, the validation set contained 3424 (20%), and the test set contained the remaining 3423 (20%) image/mask pairs. Both networks were trained in MATLAB (2024b) using stochastic gradient descent with momentum, starting from random weights. The learning rate was initialized at 0.003 and reduced by a factor of 0.3 every 10 epochs. The maximum number of epochs was set to five, with a minibatch size of eight. Training was performed on a single CPU (MacBook Pro 2021, Apple M1 Pro chip, TSMC, Taiwan Semiconductor Manufacturing Company, Hsinchu, Taiwan).

The number of training epochs was limited to five primarily due to hardware constraints and to maintain consistent conditions for both U-Net and DeepLabv3+ on the same CPU-based platform. Preliminary experiments showed that after the third epoch, validation loss and accuracy stabilized, with less than a 1% improvement in subsequent epochs, indicating convergence. Since the primary objective of this study was to compare zero-shot SAM segmentation against conventionally trained networks rather than to maximize absolute accuracy, increasing the number of epochs was unlikely to alter the

comparative conclusions, include extended training with GPU acceleration to further assess the effect of additional epochs on segmentation performance.

For SAM zero-shot prompting, each DeepLabv3+ generated mask was used to extract bounding box coordinates and to extract point coordinates. The bounding box coordinates were generated for the entire segmentation masks, which we used to zero-shot prompt SAM to generate segmentation masks. For the point coordinates-based prompting, we used incremental sets, gradually increasing the number of points coordinates to evaluate the effect of the number of points on the segmentation masks (15, 30, 45, ... 105). We set SAM to return a single image mask for all objects presented in the image rather than a separate image per mask.

7. Evaluation

The proposed method to improve the segmentation results of foods using a zero-shot SAM neural network showed better results than those shown by U-Net, but DeepLabv3+ had the best segmentation results. In the following two sections, we show both quantitative and qualitative results for the four approaches described.

7.1. Quantitative Evaluation

For quantitative evaluation, we used three metrics to assess the performance of each trained deep neural network: (1) pixel-wise accuracy, (2) intersection-over-union (IoU), (3) Dice coefficient, (4) precision, (5) recall, and (6) F1 score. Let G be the ground truth mask and P be the predicted mask. Pixel-wise accuracy is computed by Equation (1), where n and m are the width and height of both ground truth masks G and predicted masks P . IoU is computed by Equation (2), and Dice coefficient is computed by Equation (3). In Equations (4) and (5), we show the precision and recall formulas, where the TP denotes the true positives pixel-wise between the predicted mask and the ground truth mask, the TN is the true negatives pixel-wise, FP is the false positives, and FN is the false negatives.

$$Accuracy(\%) = \frac{\sum_{i=0}^n \sum_{j=0}^m (G_{(i,j)} == P_{(i,j)})}{n \times m} \quad (1)$$

$$IoU = \frac{G \cap P}{G \cup P} \quad (2)$$

$$Dice = 2 \times \frac{G \cap P}{G \cup P} \quad (3)$$

$$Precision = \frac{TP}{TP + FP} \quad (4)$$

$$Recall = \frac{TP}{TP + FN} \quad (5)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (6)$$

Our trained U-Net model's segmentation performance on the unseen test image set showed a pixel-wise accuracy of 77.63%. IoU was 0.647, Dice coefficient was 0.759, precision was 0.826, recall was 0.758, and F1 score was 0.759. The results for DeepLabv3+ segmentation on the same test set were superior to the U-Net segmentation results, where pixel-wise accuracy was 92.54%, IoU was 0.863, Dice coefficient was 0.918, precision was 0.936, recall was 0.908, and F1 score was 0.917.

As for varying the number of point coordinates (15, 30, 45, ... 105), the optimal results were obtained by using 75 point coordinates as prompts for zero-shot SAM, where accuracy reached its maximum at 77.5%, IoU at 0.672, and the Dice coefficient at 0.780. We observed that by increasing the number of point coordinates for prompting zero-shot SAM,

the segmentation generation for food improved by up to 75 points. However, using 90 and 105-point coordinates for prompting, we noted a degradation of segmentation results. The results for different numbers of points as prompts to the zero-shot SAM deep neural network are shown in Table 2.

Table 2. Performance of zero-shot SAM network where coordinates are increased from 15 coordinates to 105 coordinates with an increment of 15 coordinates at a time. The best-observed results for prompting zero-shot SAM with point coordinates are shown in bold font.

# Coords for SAM	Accuracy (%)	IoU	Dice Coefficient	Precision	Recall	F1 Score
15	75.1	0.643	0.748	0.757	0.770	0.748
30	76.3	0.658	0.764	0.759	0.795	0.764
45	77.1	0.668	0.773	0.758	0.815	0.773
60	77.4	0.672	0.778	0.758	0.824	0.778
75	77.5	0.672	0.780	0.760	0.826	0.779
90	77.2	0.669	0.778	0.761	0.824	0.778
105	76.9	0.664	0.777	0.761	0.822	0.776

The best results of zero-shot SAM segmentation using coordinate prompt encoding showed comparable results to U-Net segmentation results in terms of pixel-wise accuracy, and slightly higher IoU, Dice coefficient, precision, recall, and F1-score. For bounding box-based prompts for zero-shot SAM, we observed slightly higher pixel-wise segmentation accuracy, IoU, Dice coefficient, precision, recall, and F1-score compared to coordinate-prompted zero-shot SAM. The pixel-wise segmentation accuracy for zero-shot SAM with bounding box prompts was 78.94%, the IoU was 0.690, the Dice coefficient was 0.790, the precision was 0.769, the recall was 0.848, and the F1-score was 0.790. The comparative results of U-Net, DeepLabv3+, zero-shot SAM with coordinates, and SAM with bounding box are shown in Table 3.

Table 3. The results of trained U-Net, DeepLabv3+, and prompted zero-shot SAM deep neural network with points coordinates (Coords) and bounding box (BB). The best results are shown in bold font. The time cost is shown in the rightmost column, where two asterisks denote train and test time cost and a single asterisk denotes test-only time cost (zero-shot approach).

Network	Accuracy	IoU	Dice Coefficient	Precision	Recall	F1 Score	Time Cost
U-Net	77.63%	0.647	0.759	0.826	0.758	0.759	** 640 min
DeepLabv3+	92.54%	0.863	0.918	0.936	0.908	0.917	** 159 min
SAM + Coords (75 points)	77.5%	0.672	0.780	0.760	0.826	0.779	* 20 min
SAM + BB	78.94%	0.690	0.790	0.769	0.848	0.790	* 24 min

7.2. Qualitative Evaluation

The results of the zero-shot SAM deep neural network with different numbers of prompt coordinates showed improvement in segmentation mask generation. However, a degradation in segmentation mask generation is observed for a larger number of point coordinate-based prompts for zero-shot SAM (90 and 105 coordinates), as shown in Table 2. Figures 4 and 5 show the performance for different numbers of zero-shot SAM prompts using points coordinates.

A visual comparison of segmentation generated by U-Net, DeepLabv3+, zero-shot SAM with points coordinates prompts, and zero-shot SAM with bounding box prompts is shown in Figure 6. The results show that DeepLabv3+ mask generation is superior to all other methods. Zero-shot SAM deep neural network with 75-point prompts and bounding box prompts showed a performance that is superior to U-Net. These results show that even

without training, zero-shot SAM can achieve superior results compared to a U-Net model trained on the dataset.

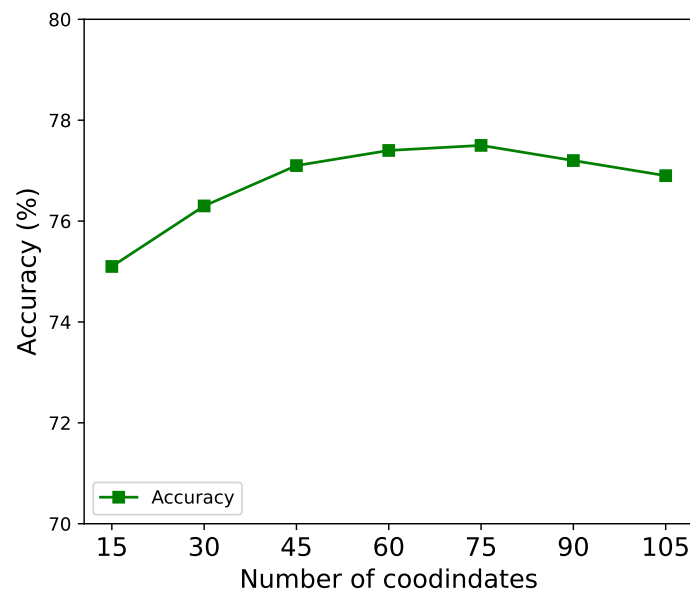


Figure 4. Line plot of accuracy performance for the SAM deep neural network for different sizes of point coordinates.

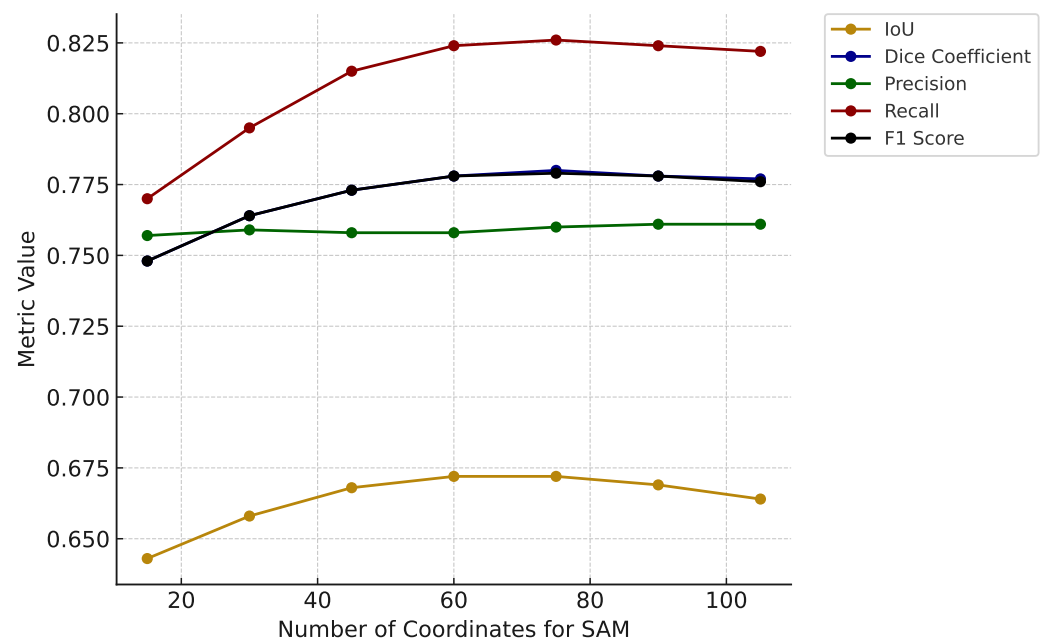


Figure 5. The performance of the SAM zero-shot approach with different prompt sizes in terms of IoU, Dice coefficient, precision, recall, and F1 score.

In Figure 7, we provide multiple images with segmentation masks overlaid on top for different numbers of zero-shot SAM point coordinate prompts, starting with 30 prompts until 90 prompts. We did not show the results for 15 and 105-point coordinate prompts due to space restrictions. The ground truth contour is superimposed on the original image and shown in green. The predicted zero-shot SAM deep neural network segmentation mask counters are superimposed on the original image and shown in red. The masks generated with zero-shot SAM deep neural network with 60 and 75 are optimal for some images and require improvement for other images.

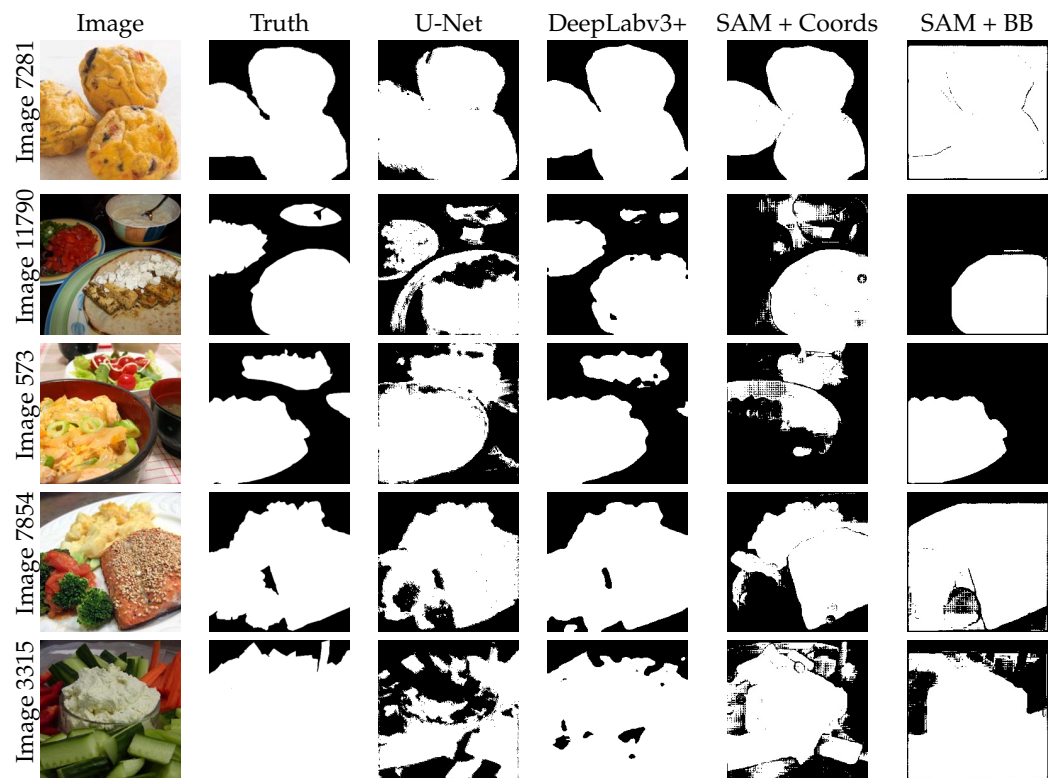


Figure 6. Five test images and their ground truth label from FoodBin-17k accompany the results of the four networks tested in this paper: (1) U-Net, (2) DeepLabv3+, (3) SAM + Coordinates (total coordinates is 75), and (4) SAM + Bounding Box

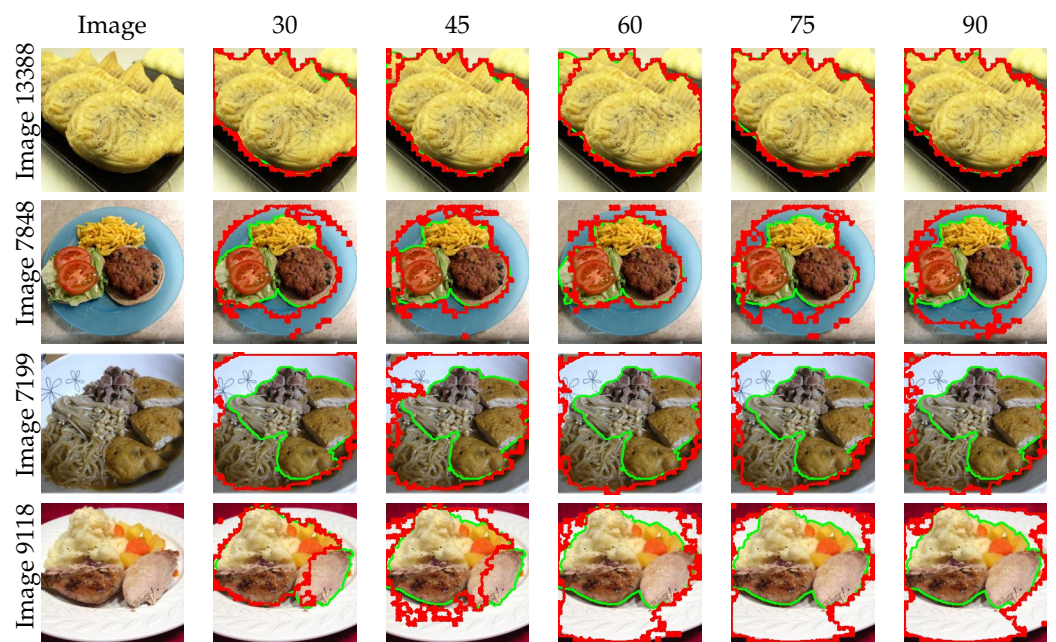


Figure 7. Illustrative examples of the results of zero-shot SAM, where each row represents a different example from the test set, and the columns represent the number of coordinates used as a prompt for SAM. The SAM-generated mask contours are extracted and superimposed on the input images with red color. The ground truth mask contours are extracted and superimposed on the input images with a green color.

An advantage of using zero-shot SAM as a foundation model is to obtain quick and effortless masks for food images without the need for obtaining a large annotated dataset and training supervised deep learning models. This approach is time-consuming;

for instance, in this work training, U-Net took 640 min and 25 s, whereas DeepLabv3+ took 159 min and 32 s. Zero-shot SAM prompting took about 20 min for 75 coordinates prompting, and 24 min for bounding box-based prompting as indicated in Table 3.

8. Discussion

While other segmentation networks are available (e.g., Mask2Former, Swin-UNet, and other vision transformers) and may be examined in future work, DeepLabv3+ and U-Net have been established as leading standard methods and accordingly were the only two examined in this study. DeepLabv3+ showed higher performance than both U-Net and zero-shot SAM with both prompts (point coordinates and bounding box). It is worth noting that DeepLabv3+ and U-Net were trained from scratch for the purpose of segmenting foods in RGB images. Zero-shot SAM deep neural network, on the other hand, was not fine-tuned in this study. Instead, we prompted the zero-shot SAM deep neural network for the purpose of studying the impact of the number of point coordinates on food image segmentation results and to evaluate the zero-shot performance of SAM on food images.

Our study investigates zero-shot SAM performance with two different prompts: bounding box prompting and point coordinate prompting (with systematically increasing numbers of coordinates), and we compared the results to two state-of-the-art trained deep learning segmentation methods: U-Net and DeepLabv3+.

The results of our study show that zero-shot SAM is slightly better than a trained U-Net architecture for segmenting food in RGB images. This demonstrates a compelling performance that may be useful for many applications in which end-users cannot or prefer not to train or fine-tune deep neural network models for segmenting objects. Zero-shot SAM neural network can achieve a similar segmentation result without the need for training or fine-tuning. Such applications include segmenting food for further analysis, feature extraction, and diet intake tracking.

The number of points used as a prompt for zero-shot SAM deep neural network is also investigated, for which we observed the segmentation performance improvement as the number of points increased up to 75 points. However, after 75-point prompts, improved performance abated and slightly decreased through 105 points. This suggests that zero-shot SAM is somehow confused with too many coordinate prompts, diminishing model performance in segmentation tasks. The effect of the number of coordinate prompts should be further explored in other image domains to analyze segmentation errors and establish a larger pattern with the coordinate-prompted zero-shot SAM model.

Our findings demonstrate that the segmentation performance of the SAM model in zero-shot mode is superior when utilizing bounding box prompts compared to coordinate-based prompts. This performance difference is likely attributed to the complex and heterogeneous textures commonly found in food images. When a bounding box is provided as a prompt, it effectively directs the SAM model to focus on and segment the food content contained within the specified region, thereby accommodating the variability in appearance and texture present across different food items. In contrast, when only coordinate points are used as prompts, they may not sufficiently capture the diversity and irregular boundaries of food objects within the images. As a result, the model's ability to accurately delineate the food regions is diminished, leading to a slight decrease in segmentation accuracy. These observations underscore the importance of prompt selection in domain-specific segmentation tasks and suggest that bounding box prompts may offer more robust guidance for complex image content, such as food, than coordinate-based prompts alone.

While the original datasets offer more categories for food image instance segmentation, our work focuses on semantic segmentation, requiring only one mask per food item. We begin with semantic segmentation to evaluate SAM's zero-shot performance across various

prompts versus models like U-Net and DeepLabv3. In future work, we plan to address instance segmentation of food images using SAM.

9. Conclusions

This study examined the zero-shot capabilities of the Segment Anything Model (SAM) for food image segmentation using the FoodBin-17k dataset and compared its performance with U-Net and DeepLabv3+. Results showed that SAM, especially with bounding-box prompts, achieved performance comparable to a trained U-Net while requiring no retraining and far less time, though DeepLabv3+ remained superior. SAM's efficiency makes it valuable for rapid food segmentation in applications such as dietary tracking, robotics grasping, and food quality control, but challenges remain with overlapping foods, glossy surfaces, clutter, and occlusions. These limitations suggest opportunities for future work, including food-specific fine-tuning of SAM, integrating semantic guidance from vision-language models like CLIP or GPT-4V, and developing adaptive prompt strategies to enhance accuracy and robustness. Overall, SAM provides a practical, training-free baseline for real-world food analysis systems and a strong foundation for future multi-modal, domain-adapted segmentation research.

Author Contributions: Conceptualization, S.S.A., M.R.G., and T.S.; methodology, S.S.A., M.R.G., and T.S.; software, S.S.A., M.R.G., and T.S.; validation, S.S.A., M.R.G., and T.S.; formal analysis, S.S.A., M.R.G., and T.S.; investigation, S.S.A., M.R.G., and T.S.; resources, S.S.A., M.R.G. and T.S.; data curation, S.S.A., M.R.G. and T.S.; writing—original draft preparation, S.S.A., M.R.G. and T.S.; writing—review and editing, S.S.A., M.R.G., and T.S.; visualization, S.S.A., M.R.G., and T.S.; supervision, S.S.A.; project administration, S.S.A.; funding acquisition, S.S.A. and M.R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the Deanship of Graduate Studies and Scientific Research at Nahrain University, supported under the Growth Funding Program grant code NU/GP/SERC/13/151-1. Additionally, this work was supported by the Deanship of Scientific Research, Vice Presidency for Graduate Studies and Scientific Research, King Faisal University, Saudi Arabia [Grant No. KFU253638].

Data Availability Statement: The dataset for our proposed approach is based on the following two datasets: FoodSeg103 (FS103) <https://xiongweiwu.github.io/foodseg103.html> (accessed on 22 October 2025) and UECFood-PixComplete (UEC) <https://mm.cs.uec.ac.jp/uecfoodpix/> (accessed on 22 October 2025). Our dataset can be found at <https://github.com/mgardner-lab/FoodRecognition> (accessed on 22 October 2025).

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Yu, Y.; Wang, C.; Fu, Q.; Kou, R.; Huang, F.; Yang, B.; Yang, T.; Gao, M. Techniques and challenges of image segmentation: A review. *Electronics* **2023**, *12*, 1199. [\[CrossRef\]](#)
2. Zhou, L.; Zhang, C.; Liu, F.; Qiu, Z.; He, Y. Application of deep learning in food: A review. *Compr. Rev. Food Sci. Food Saf.* **2019**, *18*, 1793–1811. [\[CrossRef\]](#)
3. Park, D.; Lee, J.; Lee, J.; Lee, K. Deep learning based food instance segmentation using synthetic data. In Proceedings of the 2021 18th International Conference on Ubiquitous Robots (UR), Gangneung-si, Republic of Korea, 12–14 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 499–505.
4. Jang, K.; Park, J.; Min, H.J. Developing a cooking robot system for raw food processing based on instance segmentation. *IEEE Access* **2024**, *12*, 106857–106866. [\[CrossRef\]](#)
5. Alahmari, S.S.; Goldgof, D.B.; Mouton, P.R.; Hall, L.O. Challenges for the repeatability of deep learning models. *IEEE Access* **2020**, *8*, 211860–211868. [\[CrossRef\]](#)

6. Pourpanah, F.; Abdar, M.; Luo, Y.; Zhou, X.; Wang, R.; Lim, C.P.; Wang, X.Z.; Wu, Q.J. A review of generalized zero-shot learning methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 4051–4070. [[CrossRef](#)] [[PubMed](#)]
7. Kaur, R.; Kumar, R.; Gupta, M. Deep neural network for food image classification and nutrient identification: A systematic review. *Rev. Endocr. Metab. Disord.* **2023**, *24*, 633–653. [[CrossRef](#)]
8. Sharma, A.; Kumar, C. Analysis & Numerical Simulation of Indian Food Image Classification Using Convolutional Neural Network. *Int. J. New Pract. Manag. Eng.* **2023**, *12*, 56–70.
9. Shi, P.; Qiu, J.; Abaxi, S.M.D.; Wei, H.; Lo, F.P.W.; Yuan, W. Generalist vision foundation models for medical imaging: A case study of segment anything model on zero-shot medical segmentation. *Diagnostics* **2023**, *13*, 1947. [[CrossRef](#)]
10. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2–6 October 2023; pp. 4015–4026.
11. Wu, J.; Ji, W.; Liu, Y.; Fu, H.; Xu, M.; Xu, Y.; Jin, Y. Medical sam adapter: Adapting segment anything model for medical image segmentation. *arXiv* **2023**, arXiv:2304.12620. [[CrossRef](#)]
12. Zhang, L.; Liu, Z.; Zhang, L.; Wu, Z.; Yu, X.; Holmes, J.; Feng, H.; Dai, H.; Li, X.; Li, Q.; et al. Segment anything model (sam) for radiation oncology. *arXiv* **2023**, arXiv:2306.11730. [[CrossRef](#)]
13. Mazurowski, M.A.; Dong, H.; Gu, H.; Yang, J.; Konz, N.; Zhang, Y. Segment anything model for medical image analysis: An experimental study. *Med Image Anal.* **2023**, *89*, 102918. [[CrossRef](#)] [[PubMed](#)]
14. Huang, Y.; Yang, X.; Liu, L.; Zhou, H.; Chang, A.; Zhou, X.; Chen, R.; Yu, J.; Chen, J.; Chen, C.; et al. Segment anything model for medical images? *Med Image Anal.* **2024**, *92*, 103061. [[CrossRef](#)] [[PubMed](#)]
15. Lan, X.; Lyu, J.; Jiang, H.; Dong, K.; Niu, Z.; Zhang, Y.; Xue, J. Foodsam: Any food segmentation. *IEEE Trans. Multimed.* **2023**, *27*, 2795–2808. [[CrossRef](#)]
16. Alahmari, S.S.; Gardner, M.; Salem, T. Segment Anything in Food Images. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 3715–3720.
17. Roy, S.; Wald, T.; Koehler, G.; Rokuss, M.R.; Disch, N.; Holzschuh, J.; Zimmerer, D.; Maier-Hein, K.H. Sam. md: Zero-shot medical image segmentation capabilities of the segment anything model. *arXiv* **2023**, arXiv:2304.05396. [[CrossRef](#)]
18. Aleem, S.; Wang, F.; Maniparambil, M.; Arazo, E.; Dietlmeier, J.; Curran, K.; Connor, N.E.; Little, S. Test-Time Adaptation with SaLIP: A Cascade of SAM and CLIP for Zero-shot Medical Image Segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 5184–5193.
19. Osco, L.P.; Wu, Q.; de Lemos, E.L.; Gonçalves, W.N.; Ramos, A.P.M.; Li, J.; Junior, J.M. The segment anything model (sam) for remote sensing applications: From zero to one shot. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *124*, 103540. [[CrossRef](#)]
20. Aslan, S.; Ciocca, G.; Schettini, R. Semantic food segmentation for automatic dietary monitoring. In Proceedings of the 2018 IEEE 8th International Conference on Consumer Electronics-Berlin (ICCE-Berlin), Berlin, Germany, 2–5 September 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
21. Xiao, Z.; Wang, J.; Han, L.; Guo, S.; Cui, Q. Application of machine vision system in food detection. *Front. Nutr.* **2022**, *9*, 888245. [[CrossRef](#)]
22. Dalakleidi, K.V.; Papadelli, M.; Kapos, I.; Papadimitriou, K. Applying image-based food-recognition systems on dietary assessment: a systematic review. *Adv. Nutr.* **2022**, *13*, 2590–2619. [[CrossRef](#)]
23. Min, W.; Jiang, S.; Liu, L.; Rui, Y.; Jain, R. A survey on food computing. *ACM Comput. Surv. (CSUR)* **2019**, *52*, 1–36. [[CrossRef](#)]
24. Alahmari, S.S.; Salem, T. Food State Recognition Using Deep Learning. *IEEE Access* **2022**, *10*, 130048–130057. [[CrossRef](#)]
25. Wen, M.; Song, J.; Min, W.; Xiao, W.; Han, L.; Jiang, S. Multi-state Ingredient Recognition via Adaptive Multi-centric Network. *IEEE Trans. Ind. Inform.* **2023**, *20*, 5692–5701. [[CrossRef](#)]
26. Yunus, R.; Arif, O.; Afzal, H.; Amjad, M.F.; Abbas, H.; Bokhari, H.N.; Haider, S.T.; Zafar, N.; Nawaz, R. A framework to estimate the nutritional value of food in real time using deep learning techniques. *IEEE Access* **2018**, *7*, 2643–2652. [[CrossRef](#)]
27. Brufau Vidal, M.; Ferrer Campo, À.; Gavalas, M. Unsupervised Segmentation Using CNNs Applied to Food Analysis. Master's Thesis, Facultat de Matemàtiques, Universitat de Barcelona: Barcelona, Spain, 2018.
28. Siemon, M.S.; Shihavuddin, A.; Ravn-Haren, G. Sequential transfer learning based on hierarchical clustering for improved performance in deep learning based food segmentation. *Sci. Rep.* **2021**, *11*, 813. [[CrossRef](#)]
29. Xian, Y.; Schiele, B.; Akata, Z. Zero-shot learning-the good, the bad and the ugly. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4582–4591.
30. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, 5–9 October 2015; Proceedings, Part III 18; Springer: Berlin/Heidelberg, Germany, 2015; pp. 234–241.
31. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.

32. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587. [[CrossRef](#)]
33. Yang, M.; Yu, K.; Zhang, C.; Li, Z.; Yang, K. Denseaspp for semantic segmentation in street scenes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3684–3692.
34. Lampert, C.H.; Nickisch, H.; Harmeling, S. Learning to detect unseen object classes by between-class attribute transfer. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 951–958.
35. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, PMLR, Online, 18–24 July 2021; pp. 8748–8763.
36. Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 2818–2829.
37. Williams, D.; Macfarlane, F.; Britten, A. Leaf only SAM: A segment anything pipeline for zero-shot automated leaf segmentation. *Smart Agric. Technol.* **2024**, *8*, 100515. [[CrossRef](#)]
38. Wan, J.; Wu, Q.; Lin, W.; Chan, A. Robust Zero-Shot Crowd Counting and Localization With Adaptive Resolution SAM. In Proceedings of the European Conference on Computer Vision—ECCV 2024. Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2025; pp. 478–495.
39. Lin, J.; Liu, L.; Lu, D.; Jia, K. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 17–18 June 2024; pp. 27906–27916.
40. Wu, X.; Fu, X.; Liu, Y.; Lim, E.P.; Hoi, S.C.; Sun, Q. A large-scale benchmark for food image segmentation. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual Event, 20–24 October 2021; pp. 506–515.
41. Okamoto, K.; Yanai, K. UEC-FoodPIX Complete: A large-scale food image segmentation dataset. In Proceedings of the Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, 10–15 January 2021; Proceedings, Part V; Springer: Berlin/Heidelberg, Germany, 2021; pp. 647–659.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.