Segment Anything in Food Images

Saeed S. Alahmari[1], Michael Gardner[2], Tawfiq Salem[3]

[1] Najran University, Saudi Arabia [2] King Faisal University, Saudi Arabia
[3] Purdue University, USA

ssalahmari@nu.edu.sa, mgardner@kfu.edu.sa, tsalem@purdue.edu

Abstract

This paper introduces a new approach for food image segmentation utilizing the Segment Anything Model (SAM), with the additional refinement achieved through fine-tuning with Low-Rank Adaptation layers (LoRA). The segmentation task involves generating a binary mask for food in RGB images, with pixels categorized as background or food. We conduct various experiments to assess and compare the performance of our proposed method with previous approaches. Our findings indicate that our method consistently outperforms other techniques, achieving an accuracy of 94.14%. The improved accuracy of our approach highlights its potential for various applications in food image analysis, contributing to the advancement of computer vision techniques in the realm of food recognition and segmentation.

1. Introduction

In recent years, significant advancements have been made in the field of computer vision, particularly in the domain of image segmentation [5, 7, 9, 12]. This progress is driven by the growing importance of accurately segmenting objects within images, given its wide-ranging applications from object recognition to scene understanding [6, 15, 18]. Among these applications, the segmentation of food images stands out as a crucial domain with profound implications for health, nutrition, and sustainability [1, 22, 25, 26]. Food image segmentation involves accurately delineating food items within images, and it holds immense promise for revolutionizing dietary analysis, food recognition, and culinary innovation.

The accurate segmentation of food images presents unique challenges due to the diverse shapes, textures, and colors of food items, as well as variations in lighting and background clutter [17, 18]. Traditional segmentation methods often struggle to capture the intricate details of food items, leading to inaccuracies and inefficiencies [4, 16]. Therefore, there is a pressing need for advanced techniques that can effectively address these challenges and deliver precise segmentation results.

In this paper, we present a novel approach to tackle the challenges of food image segmentation by leveraging recent advancements in deep learning and fine-tuning techniques. Our proposed method integrates state-of-the-art deep learning models with fine-tuning strategies tailored specifically for the task of food image segmentation. By utilizing the power of deep learning and fine-tuning, our goal is to enhance both the efficiency and accuracy of food image segmentation, thereby paving the way for transformative applications in nutrition analysis, culinary innovation, and sustainable food practices.

This paper introduces a novel approach for food image segmentation, leveraging the well-performing Segment Any Thing (SAM) model and fine-tuning through Low-Rank Adaptation layers (LoRA). To improve the performance of our segmentation model, we adopt a strategy of freezing the SAM model and integrating Low-Rank Adaptation layers for fine-tuning. This utilization of the SAM model, combined with Low-Rank Adaptation layers, enhances both the efficiency and accuracy of the segmentation process for food images. Through quantitative and qualitative evaluation, we demonstrate the effectiveness of our approach in generating precise segmentation masks for a wide range of food images.

2. Dataset

To fine-tune the proposed model, we utilized the dataset presented in [4], which combines two existing datasets; FoodSeg103 (FS103)[23] and UECFoodPixComplete (UEC)[19] of labeled food images. Both datasets, FS103 (7,118 images) and UEC (10,000 images) contain 8-bit RGB images and corresponding pixel-wise labels of food(s) stored as an 8-bit integer value. These datasets were combined and simplified by first transforming the 8-bit labels to binary labels (0=not food; 1=food) and then re-scaling and cropping to 224×224-pixel squares. The re-

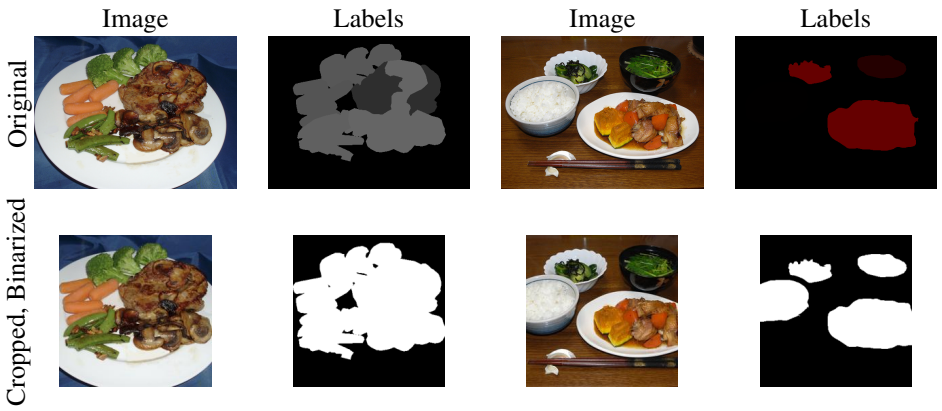

Figure 1. Images and labels from two multi-class datasets (FS103 and UEC) are binarized, cropped, and combined, resulting in a dataset of 17,118 224×224-pixel square 8-bit RGB images and corresponding binary pixel-wise labels (0=not food; 1=food).

sulting dataset contains 17,118 8-bit RGB images with associated binary pixel-wise labels (Figure 1).

3. Previous work

Recent work in food segmentation has focused on various attention mechanisms. Sharma et al. [21] developed GourmetNet architecture that combines channel attention and spatial attention for semantic segmentation. Dong et al. [10] proposed a cross-spatial attention module in which contextual information is combined by cross-calculation and a channel attention module selectively highlights certain features. Another study explored an attention-guided approach for simultaneous food type and food state recognition [4]. Alahmari et al. used DeepLabv3+ [8] CNN architecture to generate binary food masks and direct the attention of a second architecture for indicating food type and state. This approach outperforms the previous non-attention approach [2].

The Segment Anything Method (SAM) [12] has also been employed for food image segmentation. In a study by Lan et al. [13], SAM on its own failed to accurately categorize food items, but the authors demonstrated a zero-shot framework called FoodSAM that combines semantic masks and SAM-generated binary masks to improve semantic segmentation.

In a recent study by Yin et al. [24], SAM was employed alongside a large multi-modal model to address text-based queries related to food images. SAM served as a trainable segmentation decoder. In our methodology, our primary focus is segmenting food pixels from background pixels in food images. To achieve this, we fine-tuned SAM utilizing the LoRA technique outlined in Section 4.

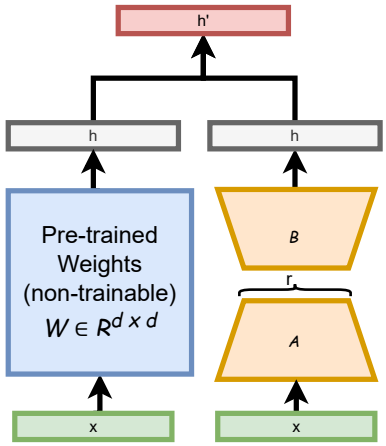

Figure 2. An illustration of LoRA approach for augmenting the original weights with low-rank weight matrices A and B.

4. Method

In this section, we illustrate low-rank adaption (LoRA) for less computation requirement and fast fine-tuning of large foundation models such as SAM. Furthermore, we explain our fine-tuned method for food image segmentation.

4.1. Low-rank Adaption (LoRA)

Since Large Language Models (LLMs) and Vision Transformer (ViT) are large models, fine-tuning these models requires significant computation resources. Accordingly, the Low-rank adaption (LoRA) technique was proposed to accelerate the fine-tuning of LLMs or ViT models by modifying the pre-trained models during fine-tuning [11]. This approach freezes the original model's weights, W, and instead of updating all the weights ΔW during backpropagation, only a set of the weights is updated $A.B$ with a hyper-parameter r for the inner dimension of the two matrices $A.B$. The new weights are concatenated on top of

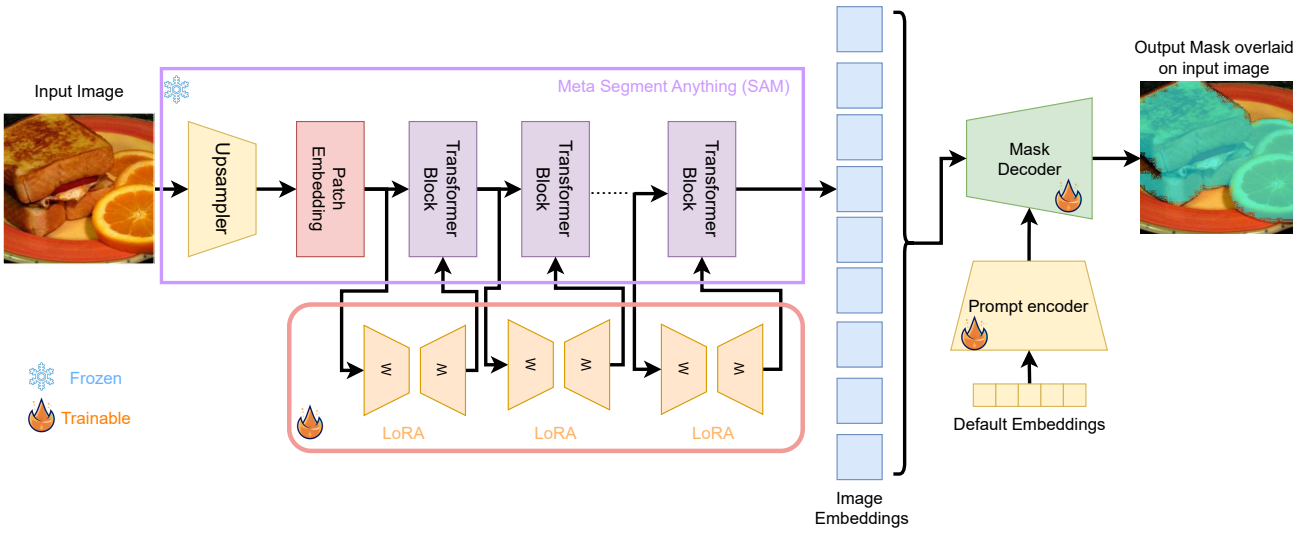

Figure 3. An overview of SAM with LoRA approach for fine-tuning. The input images were fed to a frozen weights SAM, where branches were added for LoRA (trainable weights). The image embeddings and prompts are used as input to the mask decoder which generates the masks.

the original pre-trained model during inference. In Figure 2, we provide an illustrative example of the LoRA approach, where W pre-trained model is frozen, and trainable weight matrices A and B are added and adjusted during fine-tuning.

4.2. Fine-tuning Segment Anything Model (SAM)

Given a food image $x \in \mathbb{R}^{h \times w \times c}$, where h and w are the height and width of food image x, and c is the number of channels. In our dataset $h = 224$, $w = 224$, and $c = 3$ (RGB images). The goal is to generate a binary segmentation mask $\hat{y}$ for the food present in x, where the resolution of $\hat{y}$ is $h \times w$. Each pixel in the segmentation mask y and $\hat{y}$ belongs to predefined categories $y = y_0, y_1$, where y_0 is the background class and y_1 is the food class. The aim is to learn a function $\hat{y} = f(x|\theta, y)$ for food segmentation, where y is the ground truth mask. Each pixel in the y mask is either background or foreground (i.e., binary mask). Furthermore, the weights θ are for the SAM pre-trained model to be fine-tuned using the LoRA approach.

As illustrated in Figure 3, our approach is inspired by [27], where SAM model is adopted (SAM_ViT_b)[1] and was set as non-trainable (frozen), where LoRA approach is used for bypass connection to each transformer block. The LoRA approach transforms the transformer weights representations to a low-rank space and re-projects to align with channels in the frozen transformers block. During training, we fine-tune default prompts embedding and image embedding for learning to segment food in RGB images. The mask decoder is a lightweight transformer that was fine-tuned without LoRA layers, and with default prompts embedding. The fine-tuning process was done for 200 epochs, where the fine-tuning approach took about 206 minutes on the NVIDIA GeForce GTX 1080 card. Fine-tuning SAM using AdamW optimizer [14] was performed, where the learning rate was set to 0.005. Additionally, we have used deterministic settings as in [3]. The loss function for fine-tuning SAM pre-trained model using LoRA is cross entropy and dice coefficient of the binary mask segmentation as shown in Equation 1, where CE is the cross-entropy loss, DC is the dice coefficient loss, $\hat{y}$ is SAM generated mask, y is the ground truth mask, and $\lambda_1 = 0.2$ and $\lambda_2 = 0.8$ are loss weights for balancing both terms of the loss.

After fine-tuning SAM, a weight $A.B$ matrix is used on top of pre-trained model weights W for obtaining segmentation masks in the inference stage. The total time taken for generating the binary masks for food images of 3424 in our test set is 30 minutes using the same graphics card.

$$L = \lambda_1 CE(\hat{y}, D(y)) + \lambda_2 DC(\hat{y}, D(y)) \quad (1)$$

The fine-tuning of SAM using the LoRA approach was done using the dataset described in Section 2 for 200 epochs, where the training set has 10156 images, and the testing set has 3424 images. To evaluate the fine-tuned SAM model, we have used pixel accuracy metric, Intersection over union (IoU) metric, and Dice coefficient as shown in Equations 2, 3, and 4 respectively, where y is the ground truth mask and $\hat{y}$ is deep learning generated mask.

$$accuracy(\%) = \frac{\sum_{i=0}^{n} \sum_{j=0}^{m} (y_{(i,j)} == \hat{y}_{(i,j)})}{n \times m} \quad (2)$$

[1]The SAM pre-trained model can be accessed using this link

Method	Accuracy (%)	(IoU)	Dice coefficient
U-Net (trained from scratch) [20]	77.63	0.647	0.759
DeepLabv3+ (trained from scratch) [8]	92.54	0.863	0.918
SAM (pre-trained) + zero-shot prompt (total of 75 coordinates)*	77.50	0.672	0.780
SAM (pre-trained) + zero-shot prompt (Bounding Box)*	78.94	0.690	0.790
SAM fine-tuned using LoRA approach (ours)	**94.14**	**0.888**	**0.935**

Table 1. Summary of SAM fine-tuned model using LoRA approach compared to SAM model prompted using coordinates and bounding box and segmentation performance of U-Net and DeepLabv3+. The star (*) indicates that the model was not fine-tuned in our dataset. The bolded results represent the highest accuracy achieved on the dataset.

$$IoU = \frac{y \cap \hat{y}}{y \cup \hat{y}} \tag{3}$$

$$Dice = 2 \times \frac{y \cap \hat{y}}{y \cup \hat{y}} \tag{4}$$

5. Results and Discussion

We have compared the results of fine-tuning SAM using the LoRA approach with U-Net and DeepLabv3+ models that we have trained and tested in the same training and testing splits. Furthermore, we have compared the performance of the fine-tuned SAM model in food images with the SAM model that was not fine-tuned but rather prompted using a different number of coordinates and bounding box prompts.

The results of fine-tuning SAM using the LoRA approach showed superior results compared to SAM without fine-tuning (paper under review). The accuracy of fine-tuning SAM using the LoRA approach is 94.14%, the intersection of union (IoU) is 0.888, and the Dice coefficient is 0.935. For comparison with the results of our previous proposed approach (under review), We have compared the fine-tuned SAM model's results with the U-Net and DeepLabv3 on the same split of data. Furthermore, we have compared the fine-tuned SAM model's model's results with the original SAM model (without any fine-tuning) with visual prompting such as coordinates and bounding box prompts. The results of trained U-Net on the same test data is 77.63%, the IoU was 0.647, and the dice coefficient was 0.759. The results of trained DeepLabv3+ on the same test data was 92.54%, the IoU was 0.863, and the dice coefficient was 0.918. The previous work (paper under review) used only visual prompts by using points coordinates or bounding box prompts for obtaining segmentation masks from pre-trained SAM image encoder and mask decoder (without any fine-tuning) which is indicated by star (*) in Table 1.

In Figure 4, we provide a comparison between the fine-tuned SAM model on food images (ours) to previously proposed approaches such as U-Net, DeepLabv3+, SAM with bounding box prompting (without fine-tuning), and SAM with coordinates prompting (without fine-tuning). The segmentation of our fine-tuned SAM model showed the best segmentation results compared to the other approaches. Furthermore, our model is robust and showed good performance compared to DeepLabv3+ in the middle column, where DeepLabv3+ was confused with the flower painted in the dish, however, fine-tuned SAM was not confused by irrelevant features in the images.

6. Conclusion

This paper presents an innovative approach to food image segmentation, integrating the Segment Any Thing (SAM) model with Low-Rank Adaptation layers (LoRA) to enhance accuracy and efficiency. Through extensive experiments, we have demonstrated the effectiveness of our method in addressing the challenges posed by diverse food images. Our approach outperforms existing techniques, offering precise segmentation crucial for applications in dietary analysis, food recognition, and culinary innovation. This work contributes to advancing food image segmentation techniques and holds promise for diverse applications in the nutrition and culinary fields. Future research endeavors could further enhance the method and explore its applicability in real-world scenarios.

References

[1] Palakorn Achananuparp, Ee-Peng Lim, and Vibhanshu Abhishek. Does journaling encourage healthier choices? analyzing healthy eating behaviors of food journalers. In *Proceedings of the 2018 International Conference on Digital Health*, pages 35–44, 2018. 1

[2] Saeed S Alahmari and Tawfiq Salem. Food state recognition using deep learning. *IEEE Access*, 10:130048–130057, 2022. 2

[3] Saeed S Alahmari, Dmitry B Goldgof, Peter R Mouton, and Lawrence O Hall. Challenges for the repeatability of deep learning models. *IEEE Access*, 8:211860–211868, 2020. 3

[4] Saeed S. Alahmari, Michael R. Gardner, and Tawfiq Salem. Attention guided approach for food type and state recognition. *Food and Bioproducts Processing*, 2024. 1, 2

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017. 1

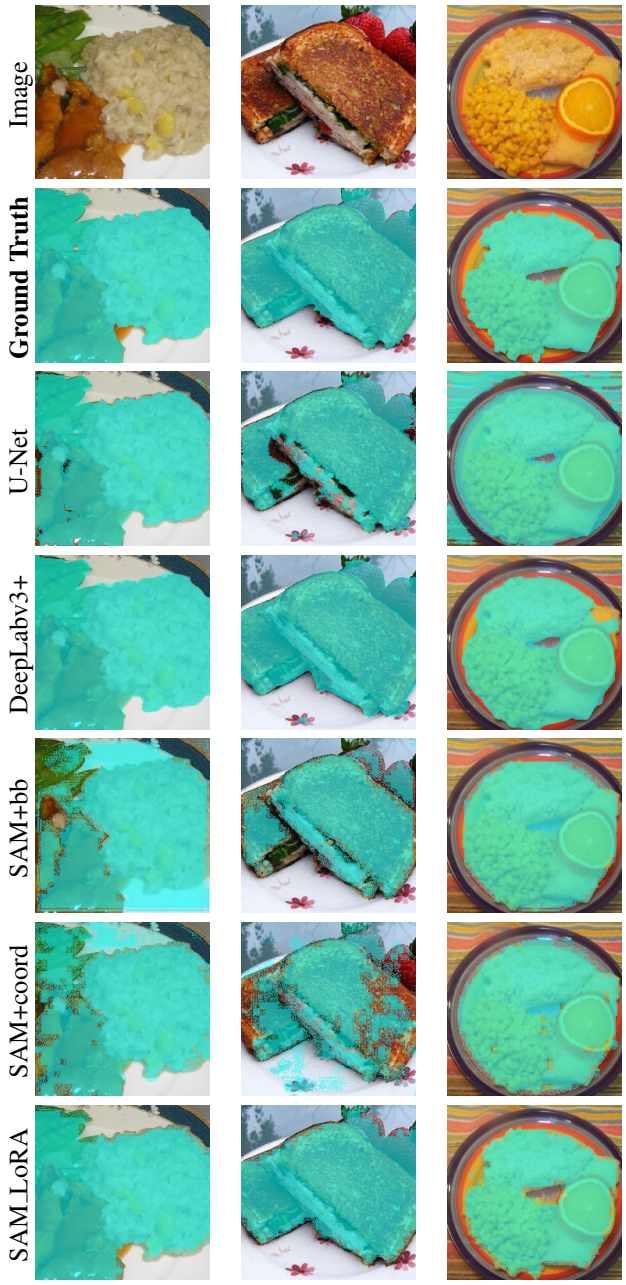

Figure 4. The figure displays sample segmentation masks for our test sets: input images on the top row, ground-truth masks on the second row, and comparisons with previous methods like U-Net, DeepLabv3+, SAM with bounding box prompting, SAM with coordinates-based prompts, and our fine-tuned SAM model on subsequent rows.

[6] Dina Bashkirova, Mohamed Abdelfattah, Ziliang Zhu, James Akl, Fadi Alladkani, Ping Hu, Vitaly Ablavsky, Berk Calli, Sarah Adel Bargal, and Kate Saenko. Zerowaste dataset: Towards deformable object segmentation in cluttered scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21147–21157, 2022. 1

[7] Yu-Ting Chang, Qiaosong Wang, Wei-Chih Hung, Robinson Piramuthu, Yi-Hsuan Tsai, and Ming-Hsuan Yang. Weakly-supervised semantic segmentation via sub-category exploration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8991–9000, 2020. 1

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018. 2, 4

[9] Gabriela Csurka, Riccardo Volpi, Boris Chidlovskii, et al. Semantic image segmentation: Two decades of research. *Foundations and Trends® in Computer Graphics and Vision*, 14(1-2):1–162, 2022. 1

[10] Xiaoxiao Dong, Haisheng Li, Xiaochuan Wang, Wei Wang, and Junping Du. Canet: cross attention network for food image segmentation. *Multimedia Tools and Applications*, pages 1–20, 2023. 2

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 1, 2

[13] Xing Lan, Jiayi Lyu, Hanyu Jiang, Kun Dong, Zehai Niu, Yi Zhang, and Jian Xue. Foodsam: Any food segmentation. *IEEE Transactions on Multimedia*, 2023. 2

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[15] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024. 1

[16] Weiqing Min, Shuqiang Jiang, Linhu Liu, Yong Rui, and Ramesh Jain. A survey on food computing. *ACM Computing Surveys (CSUR)*, 52(5):1–36, 2019. 1

[17] Weiqing Min, Zhiling Wang, Yuxin Liu, Mengjiang Luo, Liping Kang, Xiaoming Wei, Xiaolin Wei, and Shuqiang Jiang. Large scale visual food recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(8): 9932–9949, 2023. 1

[18] Shervin Minaee, Yuri Boykov, Fatih Porikli, Antonio Plaza, Nasser Kehtarnavaz, and Demetri Terzopoulos. Image segmentation using deep learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3523–3542, 2021. 1

[19] Kaimu Okamoto and Keiji Yanai. Uec-foodpix complete: A large-scale food image segmentation dataset. In *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part V*, pages 647–659. Springer, 2021. 1

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 4

[21] Udit Sharma, Bruno Artacho, and Andreas Savakis. Gourmetnet: Food segmentation using multi-scale waterfall features with spatial and channel attention. *Sensors*, 21(22): 7504, 2021. 2

[22] Bonnie A White, Caroline C Horwath, and Tamlin S Conner. Many apples a day keep the blues away–daily experiences of negative and positive affect and food consumption in young adults. *British journal of health psychology*, 18(4):782–798, 2013. 1

[23] Xiongwei Wu, Xin Fu, Ying Liu, Ee-Peng Lim, Steven CH Hoi, and Qianru Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 506–515, 2021. 1

[24] Yuehao Yin, Huiyan Qi, Bin Zhu, Jingjing Chen, Yu-Gang Jiang, and Chong-Wah Ngo. Foodlmm: A versatile food assistant using large multi-modal model. *arXiv preprint arXiv:2312.14991*, 2023. 2

[25] Amir Zarrinpar, Amandine Chaix, and Satchidananda Panda. Daily eating patterns and their impact on health and disease. *Trends in Endocrinology & Metabolism*, 27(2):69–83, 2016. 1

[26] Eliana Zeballos and Jessica E Todd. The effects of skipping a meal on daily energy intake and diet quality. *Public health nutrition*, 23(18):3346–3355, 2020. 1

[27] Kaidong Zhang and Dong Liu. Customized segment anything model for medical image segmentation. *arXiv preprint arXiv:2304.13785*, 2023. 3