

A SAM Based Tool for Semi-Automatic Food Annotation

Lubnaa Abdur Rahman, Ioannis Papathanail, Lorenzo Brigato and Stavroula Mougiakakou

ARTORG Center, Graduate School for Cellular and Biomedical Sciences, University of Bern, Switzerland

{name}. {lastname}@unibe.ch

https://github.com/lubnaa25/MealSAM_food_annotator

Abstract. The advancement of artificial intelligence (AI) in food and nutrition research is hindered by a critical bottleneck: the lack of annotated food data. Despite the rise of highly efficient AI models designed for tasks such as food segmentation and classification, their practical application might necessitate proficiency in AI and machine learning principles, which can act as a challenge for non-AI experts in the field of nutritional sciences. Alternatively, it highlights the need to translate AI models into user-friendly tools that are accessible to all. To address this, we present a demo of a semi-automatic food image annotation tool leveraging the Segment Anything Model (SAM) [15]. The tool enables prompt-based food segmentation via user interactions, promoting user engagement and allowing them to further categorise food items within meal images and specify weight/volume if necessary. Additionally, we release a fine-tuned version of SAM's mask decoder, dubbed MealSAM, with the ViT-B backbone tailored specifically for food image segmentation. Our objective is not only to contribute to the field by encouraging participation, collaboration, and the gathering of more annotated food data but also to make AI technology available for a broader audience by translating AI into practical tools.

1 Introduction

The realm of artificial intelligence (AI) has witnessed extraordinary growth [34], especially in healthcare, demonstrating its potential in disease prevention, diagnosis, and personalised management [12, 24]. Nutrition and diet-related diseases, crucial yet often overlooked, stand as key areas where AI can make a significant impact. This is especially pertinent amid the global increase in non-communicable diseases like obesity, diabetes, and cardiovascular diseases, highlighting the necessity for unbiased, accurate, and cost-efficient dietary intake monitoring and assessment [20, 31, 32]. While traditional dietary assessment methods have been the norm, they are time-consuming, expensive, and error-prone due to reliance on people's memory instead of objective measurements [29, 28]. Consequently, exploring alternatives such as visual estimations through computational image analysis is essential for improving the accuracy of dietary assessment across various socioeconomic contexts [13, 5, 4, 14, 21].

Significant strides have been made towards leveraging AI and computer vision to automate dietary assessment through the analysis of meal images, achieving levels of accuracy comparable to those of dietitians [17, 23, 30]. As efforts continue to automate the dietary assessment process using AI, diverse models have emerged and undergone further exploration in food image analysis tasks of segmen-

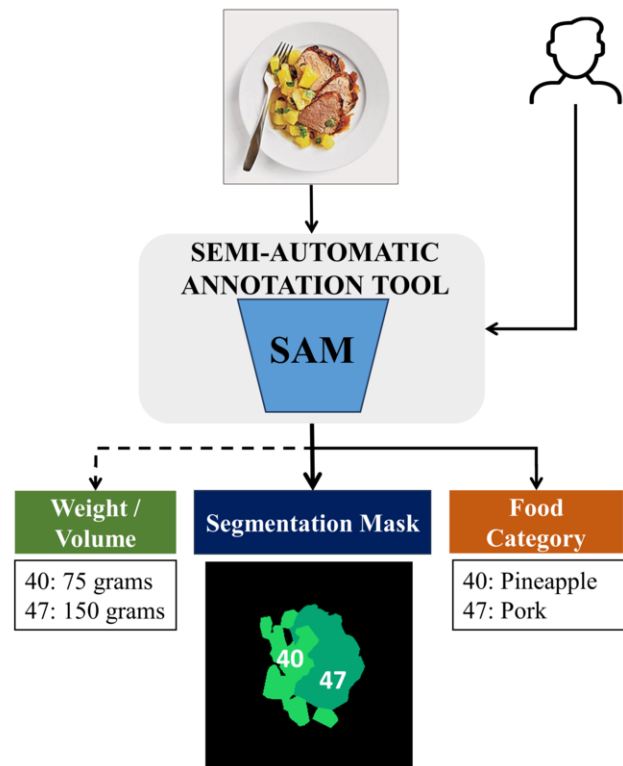


Figure 1: Inputs and outputs of the semi-automatic tool

tation, classification, and volume estimation, opening up novel avenues for improved accuracy [15, 16, 22, 10, 35, 25, 8, 1]. However, the challenge of scarcity of annotated food data essential for model training and evaluation persists. Even though open-source food image datasets [3, 27, 2, 1] serve as valuable resources for food segmentation, classification, and volume estimation, they often lack thorough aggregated annotations. Moreover, there is an ongoing need to transform AI models into practical tools accessible to a wider audience, especially to non-AI experts [26, 9]. Collaboration between AI and non-AI experts, particularly nutrition specialists like dietitians, is crucial for advancing the field.

We acknowledge the labor-intensive nature of manually segmenting or annotating food items on a plate and, therefore, leverage AI's capabilities. Previous research has developed AI tools for image-based automatic segmentation, focusing primarily on medical images, as described in [19]. It has also been seen that past works have ventured into semi-automatic food segmentation tools [21], but none,

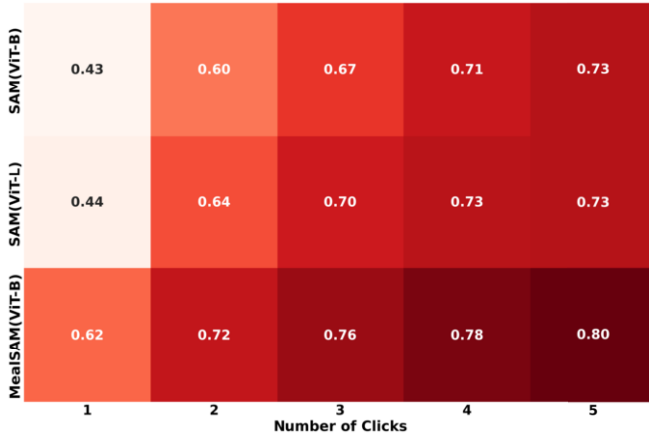


Figure 2: MealSAM vs SAM-variants on FoodSeg103 test set in terms of IoU (≥ 0.5) as a function of the number of clicks - average of three runs

to our best knowledge, have been made publicly accessible. Therefore, this work presents a practical food annotation tool to simplify the food annotation tasks, covering segmentation, classification of food items, and, optionally, volume/weight annotation. To promote effective user interaction, we leverage the capabilities of an open-source prompt-driven segmentation model known as the Segment Anything Model (SAM) [15]. SAM receives image inputs alongside prompts, which, in our tool, are user interactions via clicks within the meal image.

Our primary contribution is an open-source tool, supporting not only pretrained versions of SAM but also MealSAM, which is a fine-tuned version of the SAM's mask decoder with the vision transformer base (ViT-B) backbone. As such, the tool is specifically designed to facilitate AI-driven food annotation. In this demo, we show the tool's capabilities by annotating meal images from the FoodSeg103 test set [33], covering segmentation, classification, and, optionally, weight/volume estimation. Our broader aim is to make AI accessible to a wider audience, including non-AI specialists, which is essential for advancing technology in nutrition research.

2 Materials and Methods

2.1 Prompt-based semi-automatic segmentation

SAM [15] is a notable advancement in image segmentation by supporting user prompts, showcasing remarkable zero-shot transfer skills across several domains [11, 18]. SAM is structured around three core components: an image encoder built on the vision transformer (ViT) architecture [7], a prompt encoder that accepts different input types such as points, boxes, text, and pre-existing masks, and a mask decoder that merges these inputs and produces a segmentation mask. The image encoder is available in different model scales, precisely, base (ViT-B), large (ViT-L), and huge (ViT-H).

In our annotation tool, prompts are equivalent to user clicks, thereby enhancing user engagement in image annotation. Our tool also allows users to select MealSAM or one of the three different pre-trained SAM model scales. Even though SAM can simultaneously generate multiple masks, we opt not to use this feature as we aim to create a singular mask for each food item for further annotation per food item. Each user interaction, through clicks, must target a specific food item, which is then processed to produce a single mask output. Users can mark desired areas (foreground) for masking with a left click, indicating the regions to be included. When using

pre-trained variants of SAM, users can also denote unwanted areas (background) with a right click to mark exclusion from the mask. This approach is beneficial for accurately segmenting complex, layered food items. Upon a mask prediction, users are then prompted to categorise the segmented food/container item within the mask to complete the annotation process.

2.2 MealSAM

To enhance the segmentation mask quality, we opted to fine-tune the mask decoder specifically for the food domain. The training was done, with a batch size of 4, on a total of 18,320 binary food masks of the FoodSeg103 training set [33]. We used the Adam optimizer, with an initial learning rate of $1e-4$ set to change every 10 epochs with loss as in [6]. To obtain prompts, we randomly sampled 1-5 points from the ground truth masks, which served as input points. For the evaluation of the test set, which contains 7,697 binary food masks, we again randomly sample between 1-5 points based on the ground truth masks, which are fixed while evaluating. This evaluation compares MealSAM against pre-trained SAM variants: ViT-B and ViT-L. The average intersection over union (IoU) across three different runs with 1-5 clicks, using a threshold of 0.5, as a function of the number of clicks, is presented in Figure 2. For both MealSAM and ViT-B, the average time taken for the generation of 1 mask on GPU acceleration (RTX A6000) was 0.23 seconds, while for ViT-L, it was 0.37 seconds. MealSAM not only achieves a higher average IoU than ViT-B but also surpasses ViT-L for food image segmentation while being faster, making it the preferred choice.

2.3 Tool design and implementation

The tool, developed in Python and utilising libraries and frameworks for image processing, is structured around three key components: segmentation, category selection, and weight or volume assignment to streamline the annotation process. We employ the Tkinter library for the GUI, PIL for image manipulation, OpenCV for image processing, and PyTorch for leveraging the deep learning back-end. We show an overview of the code architecture in the following snippet.

```

Tool Pseudocode
def ImageAnnotation(image, cat_file):
    # Input: Image file and category
    # Output: Annotated image,
    #         volume/weight of items (optional)
    UserUploadsImage(image)
    UserSelectsModel("SAM") # MealSAM as
    # default, user can choose
    while True:
        UserClicksOnFoodItem() #
        # Left-click to include in mask,
        # Right-click to exclude
        MaskPredictionbySAM()
        if MaskNotAccurate():
            if UserClearsClickedPoints():
                continue
            else:
                CorrectMaskWithBrush()
        UserSelectsFoodItem()
        Optional: UserInputsWeightVolume()
        UserValidatesAnnotation()
        if not MoreFoodItems():
            break
    UserClicksSave()
    SaveAnnotatedImage()

```

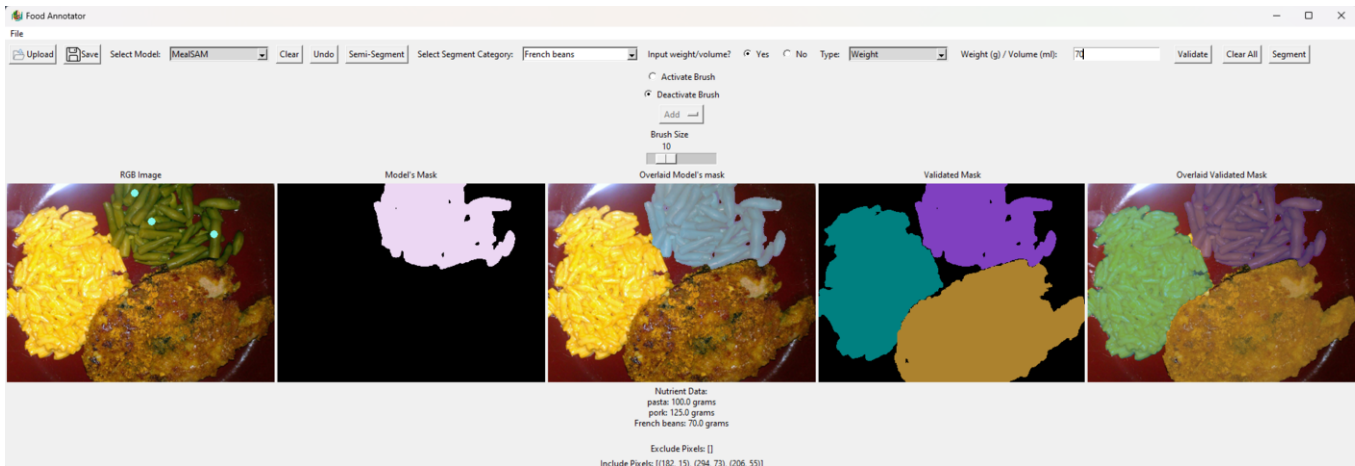


Figure 3: Illustration of meal image segmentation with annotated labels and volume estimation via semi-automatic tool

2.4 Dataset and food categories

For the demo, we make use of the food categories within Foodseg103. In the demo video, we annotate 2 meal images from the test set of Foodseg103 to demonstrate the capabilities of our segmentation tool. Annotations were made according to the ground truth classes specified by the dataset annotators.

2.5 Weight/Volume annotation

The tool incorporates an optional feature for annotating the weight or volume of segmented items motivated by the importance of quantitative analysis in food studies. After the segmentation step, the user optionally performs the annotation of weight/volume for a specific item, and a category is assigned. This part of the tool is specifically designed for cases where users are aware of ground truth weight/volume either through food weighing or volume computation.

3 Demo

In Figure 3, we show a comprehensively annotated meal image featuring segmentation, food labeling, and food weight/volume annotations. For this example, the operating machine specifics and segmentation tool configuration are as follows:

- Machine: Lenovo Legion 5 (15" AMD)
- Operating System: Windows
- Acceleration - GPU: NVIDIA RTX 3060
- Model: MealSAM
- Time Taken for complete annotation: 32.93 seconds

To begin using the segmentation tool, users upload their meal image. Based on the users' use case and accuracy needs, as previously mentioned, we allow them to choose between MealSAM and the different model sizes of the SAM image encoder. To perform the segmentation, the user first selects specific image areas by left-clicking, which the tool marks with blue dots. Conversely, to remove areas from SAM's segmentation prediction, users right-click to mark these exclusions with pink dots, as shown in Figure 3. The tool also displays the exact coordinates of both included and excluded clicks. Users have the option to undo their last action in case any mistakes have occurred. Moving forward, by selecting *Semi-Segment*, the tool generates a predictive mask by utilising the specified inclusions

and exclusions and shows it in the *Model's Mask* window. For enhanced visualisation, the generated mask is overlaid on the original image in *Overlaid Model's Mask*. In case the predicted mask is unsatisfactory, an additional brush feature is provided to correct the mask through the "Activate Brush" feature. If the user approves of the outcome, they can then categorise the segment. In case the item is not included in the dataset, the user has additionally the option to add a new category. We assign each category a specific color code to facilitate distinct visual differentiation. For those wishing to annotate volume or weight, in case of presence of ground-truth weight/volume, the tool allows users to assign a value and further validate the mask. The user may repeat this annotation cycle as needed until all food items have been annotated. An example of the meal annotation process can be viewed in the demo video [here](#).

4 Conclusion

In this work, we present food image annotation tool that leverages a powerful prompt-based segmentation model (SAM) for seamless, user-friendly segmentation and categorisation, with optional weight/volume annotation of food items. We also release the fine-tuned version of the mask decoder for the base architecture of SAM, MealSAM, specifically designed for food image segmentation. By enabling prompt-based interactions, our tool not only facilitates the generation of accurately annotated food data while reducing the time involved in fully manually annotating food items in images but also encourages the involvement of non-AI specialist users in the process. Our results showcase the tool's efficacy in producing high-quality, detailed annotations, contributing significantly to the field of dietary assessment and offering a solution to the ongoing challenge of data scarcity. Future plans include supporting larger versions of MealSAM and allowing annotation at different levels of granularity depending on use cases. By making this tool open-source, we aim to foster a community of collaboration, inviting contributions that will further enhance its capabilities and increase the amount of data available in the food domain for training and evaluating AI models.

Acknowledgements

This work was partly supported by the European Commission and the Swiss Confederation - State Secretariat for Education, Research and Innovation (SERI) within the project 101057730 Mobile Artificial Intelligence Solution for Diabetes Adaptive Care (MELISSA).

References

- [1] L. Abdur Rahman, I. Papathanail, L. Brigato, and S. Mougiakakou. A comparative analysis of sensor-, geometry-, and neural-based methods for food volume estimation. In *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, pages 21–29, 2023. [1](#)
- [2] D. Allegra, M. Anthimopoulos, J. Dehais, Y. Lu, F. Stanco, G. M. Farinella, and S. Mougiakakou. A multimedia database for automatic meal assessment systems. In *New Trends in Image Analysis and Processing—ICIAP 2017: ICIAP International Workshops, WBICV, SS-PandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11–15, 2017, Revised Selected Papers 19*, pages 471–478. Springer, 2017. [1](#)
- [3] L. Bossard, M. Guillaumin, and L. Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. [1](#)
- [4] C. Boushey, M. Spoden, F. Zhu, E. Delp, and D. Kerr. New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods. *Proceedings of the Nutrition Society*, 76(3):283–294, 2017. [1](#)
- [5] B. C. Braga, P. H. Nguyen, L. M. Tran, N. T. Hoang, B. Bannerman, F. Doyle, G. Folson, R. Gangupantulu, N. Karachiwalla, B. Kolt, et al. Feasibility of using an artificial intelligence-based phone application for dietary assessment and nudging to improve the quality of food choices of female adolescents in vietnam: Evidence from a randomized pilot study. *Current Developments in Nutrition*, 2023. [1](#)
- [6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020. [2](#)
- [7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. [2](#)
- [8] C. N. Freitas, F. R. Cordeiro, and V. Macario. Myfood: A food segmentation and classification system to aid nutritional monitoring. In *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 234–239. IEEE, 2020. [1](#)
- [9] F. Gama, D. Tyskbo, J. Nygren, J. Barlow, J. Reed, and P. Svedberg. Implementation frameworks for artificial intelligence translation into health care practice: scoping review. *Journal of Medical Internet Research*, 24(1):e32215, 2022. [1](#)
- [10] J. He and F. Zhu. Online continual learning for visual food classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2337–2346, 2021. [1](#)
- [11] S. He, R. Bao, J. Li, J. Stout, A. Bjornerud, P. E. Grant, and Y. Ou. Computer-vision benchmark segment-anything model (sam) in medical images: Accuracy in 12 datasets. *arXiv preprint arXiv:2304.09324*, 3, 2023. [2](#)
- [12] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang. Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology*, 2(4), 2017. [1](#)
- [13] M. L. Jobarteh, M. A. McCrory, B. Lo, M. Sun, E. Sazonov, A. K. Anderson, W. Jia, K. Maitland, J. Qiu, M. Steiner-Asiedu, et al. Development and validation of an objective, passive dietary assessment method for estimating food and nutrient intake in households in low-and middle-income countries: A study protocol. *Current developments in nutrition*, 4(2):nzaa020, 2020. [1](#)
- [14] Y. Kawano and K. Yanai. Foodcam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications*, 74:5263–5287, 2015. [1](#)
- [15] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. [1](#), [2](#)
- [16] X. Lan, J. Lyu, H. Jiang, K. Dong, Z. Niu, Y. Zhang, and J. Xue. Foodsam: Any food segmentation. *IEEE Transactions on Multimedia*, 2023. [1](#)
- [17] Y. Lu, T. Stathopoulou, M. F. Vasiloglou, L. F. Pinault, C. Kiley, E. K. Spanakis, and S. Mougiakakou. gofoodtm: an artificial intelligence system for dietary assessment. *Sensors*, 20(15):4283, 2020. [1](#)
- [18] M. A. Mazurowski, H. Dong, H. Gu, J. Yang, N. Konz, and Y. Zhang. Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89:102918, 2023. [2](#)
- [19] M. Muñoz-Aguirre, V. F. Ntasis, S. Rojas, and R. Guigó. Pyhist: a histological image segmentation tool. *PLoS computational biology*, 16(10):e1008349, 2020. [1](#)
- [20] W. H. Organization et al. Noncommunicable diseases: progress monitor 2022, 2022. [1](#)
- [21] I. Papathanail, J. Brühlmann, M. F. Vasiloglou, T. Stathopoulou, A. K. Exadaktylos, Z. Stanga, T. Münzer, and S. Mougiakakou. Evaluation of a novel artificial intelligence system to monitor and assess energy and macronutrient intake in hospitalised older patients. *Nutrients*, 13(12):4539, 2021. [1](#)
- [22] I. Papathanail, M. F. Vasiloglou, T. Stathopoulou, A. Ghosh, M. Baumann, D. Faeh, and S. Mougiakakou. A feasibility study to assess mediterranean diet adherence using an ai-powered system. *Scientific reports*, 12(1):17008, 2022. [1](#)
- [23] I. Papathanail, L. Abdur Rahman, L. Brigato, N. S. Bez, M. F. Vasiloglou, K. van der Horst, and S. Mougiakakou. The nutritional content of meal images in free-living conditions—automatic assessment with gofoodtm. *Nutrients*, 15(17):3835, 2023. [1](#)
- [24] P. Rajpurkar, E. Chen, O. Banerjee, and E. J. Topol. Ai in health and medicine. *Nature medicine*, 28(1):31–38, 2022. [1](#)
- [25] Z. Shen, A. Shehzad, S. Chen, H. Sun, and J. Liu. Machine learning based approach on food recognition and nutrition estimation. *Procedia Computer Science*, 174:448–453, 2020. [1](#)
- [26] D. F. Steiner, P.-H. C. Chen, and C. H. Mermel. Closing the translation gap: Ai applications in digital pathology. *Biochimica et Biophysica Acta (BBA)—Reviews on Cancer*, 1875(1):188452, 2021. [1](#)
- [27] Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8903–8911, 2021. [1](#)
- [28] M. F. Vasiloglou, S. Christodoulidis, E. Reber, T. Stathopoulou, Y. Lu, Z. Stanga, and S. Mougiakakou. What healthcare professionals think of “nutrition & diet” apps: an international survey. *Nutrients*, 12(8):2214, 2020. [1](#)
- [29] M. F. Vasiloglou, I. Marcano, S. Lizama, I. Papathanail, E. K. Spanakis, and S. Mougiakakou. Multimedia data-based mobile applications for dietary assessment. *Journal of diabetes science and technology*, 17(4):1056–1065, 2023. [1](#)
- [30] W. Wang, W. Min, T. Li, X. Dong, H. Li, and S. Jiang. A review on vision-based analysis for automatic dietary assessment. *Trends in Food Science & Technology*, 122:223–237, 2022. [1](#)
- [31] WHO. Obesity and overweight — who.int. <https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight>, 2021. [Accessed 13-02-2024]. [1](#)
- [32] WHO. Diabetes — who.int. <https://www.who.int/news-room/fact-sheets/detail/diabetes>, 2023. [Accessed 13-02-2024]. [1](#)
- [33] X. Wu, X. Fu, Y. Liu, E.-P. Lim, S. C. Hoi, and Q. Sun. A large-scale benchmark for food image segmentation. In *Proceedings of the 29th ACM international conference on multimedia*, pages 506–515, 2021. [2](#)
- [34] D. Zhang, S. Mishra, E. Brynjolfsson, J. Etchemendy, D. Ganguli, B. Grosz, T. Lyons, J. Manyika, J. C. Niebles, M. Sellitto, et al. The ai index 2021 annual report. *arXiv preprint arXiv:2103.06312*, 2021. [1](#)
- [35] L. Zhu, P. Spachos, E. Pensini, and K. N. Plataniotis. Deep learning and machine vision for food processing: A survey. *Current Research in Food Science*, 4:233–249, 2021. [1](#)