engineering proceedings

MDPI

Proceeding Paper

A Three-Stage Transformer-Based Approach for Food Mass Estimation †

Sinda Besrour [1], Ghazal Rouhafzay [1,*] and Jalila Jbilou [2,3]

[1] Department of Computer Science, Université de Moncton; email1@email.com
[2] Department of Psychology, Université de Moncton; email2@email.com
[3] Centre de Formation Médicale du Nouveau-Brunswick
* Correspondence: ghazal.rouhafzay@umoncton.ca
† Presented at the 12th International Electronic Conference on Sensors and Applications (ECSA-12), 12–14 November 2025; Available online: https://sciforum.net/event/ECSA-12.

Abstract

Accurate food mass estimation is a key component of automated calorie estimation tools, and there is growing interest in leveraging image analysis for this purpose due to its ease of use and scalability. However, current methods face important limitations. Some rely on 3D sensors for depth estimation, which are not widely accessible to all users, while others depend on camera intrinsic parameters to estimate volume, reducing their adaptability across different devices. Furthermore, AI-based approaches that bypass these parameters often struggle with generalizability when applied to images captured using diverse sensors or camera settings. To overcome these challenges, we introduce a three-stage, transformer-based method for estimating food mass from RGB images, balancing accuracy, computational efficiency, and scalability. The first stage applies the Segment Anything Model (SAM 2) to segment food items in images from the SUECFood dataset. Next, we use the Global-Local Path Network (GLPN) to perform monocular depth estimation (MDE) on the Nutrition5k dataset, inferring depth information from a single image. These outputs are then combined through alpha compositing to generate enhanced composite images with precise object boundaries. Finally, a Vision Transformer (ViT) model processes the composite images to estimate food mass by extracting relevant visual and spatial features. Our method achieves notable improvements in accuracy compared to previous approaches, with a mean squared error (MSE) of 5.61 and a mean absolute error (MAE) of 1.07. Notably, this pipeline does not require specialized hardware like depth sensors or multi-view imaging, making it well-suited for practical deployment. Future work will explore the integration of ingredient recognition to support a more comprehensive dietary assessment system.

Keywords: food mass estimation; calorie estimation; vision transformer; monocular depth estimation; segment anything model

Academic Editor(s): Name

Published: date

Citation: Besrour, S.; Rouhafzay, G.; Jbilou, J. A Three-Stage Transformer-Based Approach for Food Mass Estimation. *Eng. Proc.* **2025**, *volume number*, x. https://doi.org/10.3390/xxxxx

1. Introduction

Food mass estimation (FME) is a fundamental step in determining the nutritional content of meals and is central to calorie estimation (CE) systems. Over the past decade, increasing research efforts have focused on developing image-based and mobile solutions for accurate estimation of food volume and mass as well as analyzing the food nutritional

value. This growing interest is motivated by the need to support individuals monitoring their dietary intake, healthcare professionals overseeing chronic disease management, and researchers advancing digital nutrition tools. Recent scoping reviews have highlighted both the technological evolution of calorie estimation methods [1,2] and their clinical relevance for populations with weight-related chronic conditions [3]. Several studies, ranging from manual methods to deep learning-based (DL) systems, have been conducted on FME and food volume estimation (FVE) as preliminary steps to CE. According to Konstantakopoulos et al. [1] these studies can be grouped into categories, namely, 3D reconstruction, depth camera, perspective transformation, and DL-based methods.

3D reconstruction techniques require multiple images of the food taken from different angles. Using feature point extraction algorithms such as scale-invariant feature transform (SIFT) and speeded-up robust features (SURF), the system constructs a 3D model of the food [4,5]. While effective, this approach necessitates careful image capture and multiple snapshots, as well as high computation power and processing time, making it unsuitable for real-time applications. On the other hand, depth cameras capture the 3D structure of food, allowing for precise FVE with lower computational expense. However, these devices are costly and not commonly used in commercial applications [6,7]. Perspective transformation involves removing projective distortion from a 2D image to reconstruct the shape and size of food. This method requires a reference object within the image for accurate volume calculation. Although perspective transformation can handle irregularly shaped foods, it is sensitive to image capture conditions [8]. DL-based methods utilize advanced models such as generative adversarial networks (GANs) and convolutional neural networks (CNNs) to estimate the depth of food items from RGB images. These methods create depth maps and estimate food volume without the need for specialized cameras or multiple snapshots [9]. While DL-based approaches are more suitable for real-time applications, less costly, and user-friendly, there is room for improvement in estimation performance by using more recent and robust models.

Our study addresses this gap by introducing a cost-effective novel, and robust DL approach for food mass estimation using only RGB images, eliminating the need for reference objects. Designed for real-time deployment, our method provides improved accuracy and adaptability over existing solutions. Our main contributions are as follows:

1. We propose a three-stage, transformer-based approach for enhanced FME. The stages involve image segmentation to extract food regions, MDE to extract depth information, and FME to predict the food mass on plates.
2. We use four variations of the SAM 2 model for image segmentation based on the SUECFood dataset. The GLPN model is employed for MDE using the Nutrition5k dataset, and the ViT model is utilized for FME.
3. We combine the image segmentation and MDE results using the AC method to ensure better contour definition in the depth images and, consequently, more accurate mass estimation namely, an MSE of 5.61 and an MAE of 1.07.
4. We conduct a comparative study with prior research using the same dataset and demonstrate the superiority of our approach.

The remainder of this paper is organized as follows. The literature review is presented in Section 2. In Section 3, we describe our approach. The experimental results are provided in Section 4. Concluding remarks and future work are presented in Section 5.

2. Literature Review

While our study focuses on food mass estimation (FME), we also examine prior research on food volume estimation (FVE), as both applications tackle similar technical

challenges, such as accurate measurement and spatial analysis, and are fundamentally linked by their reliance on density to translate volume into mass.

Konstantakopoulos et al. [10] used the stereo vision technique, which relies on multiple camera views to capture food images from different angles. These images were then processed to reconstruct the 3D object of the food and calculate its volume. Jia et al. [11] introduced a method for estimating food volume by first reconstructing the dining bowl and then isolating the bowl's geometry from the food. By subtracting the bowl's empty volume from the total, FVE could be performed more accurately. The authors of [1] proposed a two-stage approach for reconstructing detailed 3D models of food. The first stage involves capturing images from different angles and using the structure from motion (SfM) method to analyze these images, identify key points, and reconstruct the camera's positions and orientations. The second stage refines the 3D models by adding depth details to each surface point using the multi-view stereo (MVS) method. Naritomi et al. [13] presented an FVE approach that reconstructs 3D meshes from a single image of a dish. After reconstructing the 3D meshes, they separated the actual dish from the plate and used the result to estimate the food volume.

The authors in [14] developed a mobile app for CE based on FVE using depth cameras. They leveraged depth-sensing technology to accurately capture the volume of food items. Thames et al. [15] conducted experiments for CE, micronutrient estimation, and FME using a dataset they collected, which includes RGB images and depth maps produced by a depth camera. Their main contribution is the creation of this comprehensive dataset.

Pouladzadeh et al. [16] proposed a solution that requires the user to place their thumb next to the dish when capturing the image. By knowing the dimensions of the user's thumb, the system can calculate the dimensions of each food item. Similarly, Okamoto and Yanai [17] used the dimensions of a reference object to compute the real size of the food regions. On the other hand, the authors in [18] introduced a method for estimating food portion sizes from smartphone images without requiring a fiducial marker for scale, making the process less intrusive for users.

Lo et al. [19] presented a method that leverages point clouds and view synthesis to estimate food volume. The view synthesis technique creates intermediate views of food items that may not have been directly captured by the camera, using DL models. Han et al. [20] proposed DPF-Nutrition, a two-stage approach that fuses the features of RGB and predicted depth images for CE, FME, and micronutrient estimation. They adopted the depth prediction transformer (DPT) to generate depth maps and designed a cross-modal attention block (CAB) to extract and integrate the complementary features of RGB and predicted depth images. Similarly, Shao, et al. [21] introduced a three-component approach consisting of the backbone network, the feature fusion module, and the nutrition prediction module for CE, FME, and micronutrient estimation.

Although prior research has explored various methods for FME and FVE, contributing with valuable solutions, many gaps remain. For instance, approaches involving 3D reconstruction, while highly accurate, demand extensive computational resources and processing time, making them impractical for real-time or mobile applications. Depth camera-based methods provide precise depth information but are limited by the high cost, restricting their accessibility to the public. Additionally, older techniques relying on traditional image processing and simpler models often lack sufficient accuracy due to their inability to capture the complex textures and shapes of food. Our approach stands out by leveraging RGB images and advanced transformer models, notably SAM 2 for image segmentation, GLPN for MDE, and ViT for FME. We employ the AC method to fuse the segmentation and MDE results, which plays a crucial role in enhancing the performance of

our approach. This strategy enables us to achieve improvements in accuracy, computational efficiency, and cost-effectiveness.

3. Proposed Approach

Our approach consists of three stages. First, we segment the RGB images into regions while generating depth images from them. Second, we combine image segmentation and MDE results using the AC method. Finally, the resulting images are used for FME. The complete pipeline is illustrated in Figure 1.

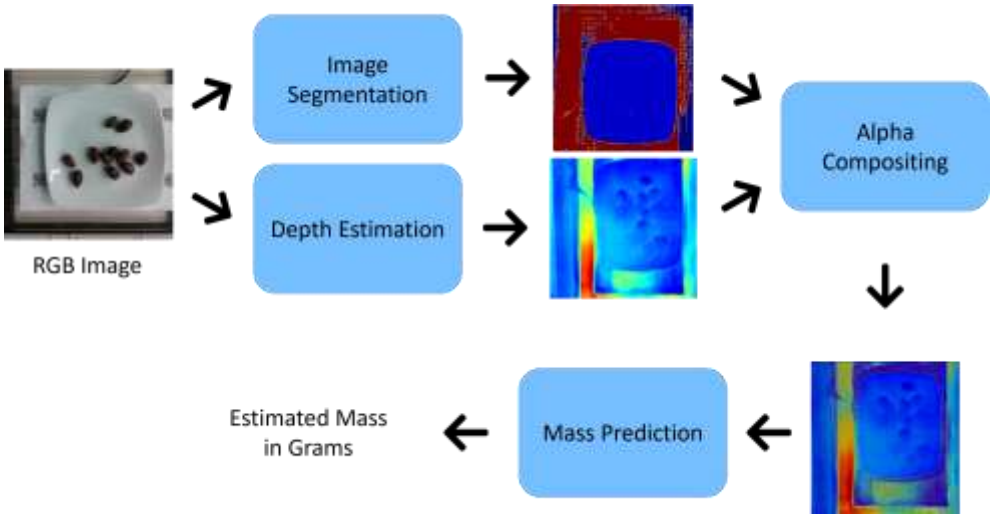


Figure 1. FME pipeline.

Accordingly, the steps to implement our approach are detailed in the subsequent paragraphs.

3.1. Dataset

(1) **MuseFood**: The MuseFood dataset [22] is a large- scale Japanese food image dataset designed primarily for food segmentation tasks. The key features of the MuseFood dataset are as follows:

   i. RGB Images: The dataset contains 31,395 images of various food items, providing a substantial resource for food image segmentation.
   ii. Masks: Each image in the dataset is accompanied by a pixel-level annotation, or mask.

(2) **Nutrition5k:** The Nutrition5k dataset [15] is a large-scale, diverse dataset designed specifically for food image recognition, mass estimation, and calorie estimation tasks in the context of dietary assessment. The key features of the Nutrition5k dataset are as follows:

   i. RGB Images: The dataset includes 5000 unique food dishes, offering substantial diversity in food types and portion sizes. Each dish is captured from multiple angles to enhance the accuracy of 3D reconstruction and volume estimation techniques.
   ii. Depth Images: A depth map is provided for each dish, captured using a depth camera to represent the distance of the food from the camera at each pixel.
   iii. Annotations: The dataset includes detailed annotations, such as dish ingredients (e.g., eggplant, roasted potatoes, cauliflower), to facilitate classification tasks. It also provides mass per ingredient, total dish mass, caloric information per

ingredient, total calories, as well as fat, carbohydrate, and protein content for each dish and ingredient.

3.2. Image Segmentation

For this task, we use SAM 2, a generalized transformer-based model developed by Meta for image and video segmentation tasks [23]. SAM 2 offers improved boundary precision, real-time processing speed, and scalability, making it an ideal choice for handling complex scenes in images and videos. Additionally, SAM 2's few-shot, and zero-shot capabilities reduce annotation costs, making it efficient for deployment. We experiment with four variations: Hiera Base Plus (Hiera-B+), Hiera Large (Hiera-L), Hiera Small (Hiera-S), and Hiera Tiny (Hiera-T). These variations differ primarily in model size and capacity, with larger models (e.g., Hiera-L and Hiera-B+) containing more parameters for enhanced feature extraction, while smaller models (e.g., Hiera-S and Hiera-T) are more lightweight, offering faster processing at the potential cost of segmentation precision. The segmentation process is accomplished through three main steps The segmentation process is thus accomplished through three main steps:

(1) **Fine-Tuning:** Each segmentation model variant (Hiera-B+, Hiera-L, Hiera-S, and Hiera-T) was fine-tuned to partition images into distinct regions such as food ingredients versus background. This training was performed using RGB images and their corresponding segmentation masks from the SUECFood dataset. A total of 80% of the dataset was allocated for training purposes. The models were trained for 1000 epochs with a fixed learning rate of 0.0001.

(2) **Testing:** The evaluation phase was conducted using the remaining 20% of the dataset. The results from this testing phase are summarized in Table 1.

(3) **Inference:** For inference, the best-performing checkpoint of each fine-tuned SAM 2 variant was applied to all RGB images in the Nutrition5k dataset. The goal of this inference step was to generate segmented versions of the images, ultimately facilitating mass estimation based on RGB inputs. This approach is critical because using the depth maps provided by the dataset directly would undermine the objective of the study, which emphasizes leveraging RGB images instead of depth camera data.

Table 1. Libraries and frameworks used in this study.

Library	Version	Usage
Torch	2.4.1	DL Framework
Torchvision	0.19.1	Image Processing
Transformers	4.45.2	Large Language Model Framework
SAM 2	1.11.2	Image Segmentation
Pillow	10.4.0	Image Manipulation
Numpy	1.26.4	Numerical Computing

3.3. Monocular Depth Estimation

For the monocular depth estimation task, the GLPN transformer model was employed, as proposed by [24]. The GLPN model is particularly well-suited for this application due to its ability to capture both global contextual information and fine-grained local details from RGB inputs. The architecture of the model consists of two distinct pathways: a global pathway designed to extract large-scale structural information and a local pathway that focuses on detailed visual cues. These pathways are fused through attention mechanisms and feature aggregation modules, ensuring a balanced representation that maintains spatial coherence while preserving subtle depth variations.

Given the complexity and ingredient diversity in the dish images of the Nutrition5k dataset, GLPN is an appropriate choice for robust depth estimation. The depth estimation process consists of three primary stages:

(1) **Fine-tuning:** The model was trained to generate depth maps using RGB images paired with ground-truth depth images from the Nutrition5k dataset. As with segmentation, 80% of the dataset was used for training. The model was trained over 10 epochs with a learning rate of 0.0001 and a batch size of 512.
(2) **Testing:** The model was evaluated using the remaining 20% of the dataset to assess its generalization performance.
(3) **Inference**: The best checkpoint obtained during fine-tuning was used to infer depth images for the entire Nutrition5k dataset, using RGB inputs.

3.4. Alpha Compositing

In this stage, we use the AC method to combine the inferred images from the best segmentation variation with the inferred depth images [25]. The motivation for combining these results is that we observed the contours in the depth images provided by the Nutrition5k dataset are more defined, as illustrated in Figure 2. Moreover, images resulting from depth cameras are renowned for their superior performance compared to RGB images [1].

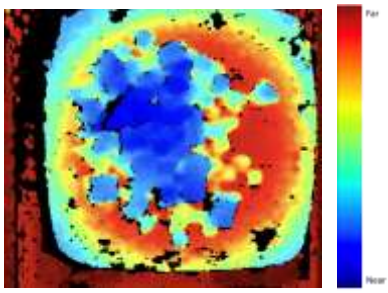
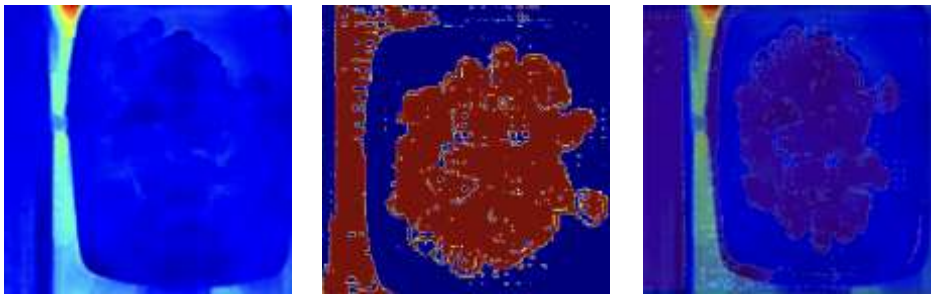


Figure 2. A sample of the depth images provided in the Nutrition5k dataset.

We observed that the contours in our segmented images were well-defined, unlike those in the depth images generated by the GLPN model. Therefore, we applied the AC method to overlay both images, using the depth image as the background with 0% transparency and the segmented image as the foreground with 39% transparency. This approach enhances the contours in the segmented regions while keeping the background depth image visible. The 39% value was determined empirically, as it ultimately delivered the most accurate results for FME during our experiments. The results are illustrated in Figure 3.



(**a**) (**b**) (**c**)

Figure 3. Applying AC to combine the depth and segmented images. (**a**) A sample of the generated depth images. (**b**) A sample of the segmented images. (**c**) The result of AC.

3.5. Food Mass Estimation

For this task, we use the ViT model proposed by [26]. ViT leverages the self-attention mechanism to capture long-range dependencies across an image. It divides images into patches (e.g., 16 × 16 pixels) and processes each patch as a token, enabling efficient parallelization. As a result, ViT is highly scalable for high-resolution images. Consequently, the FME process is carried out in two steps:

(1) **Training:** The model learns to estimate mass from the previously generated AC images. We train the model for 100 epochs, with a learning rate of 0.0001 and a batch size of 8 to avoid overloading the GPU. The data is split, with 80% used for training.
(2) **Testing**: Similar to the previous stages, the model is evaluated on the remaining 20% of the dataset.

3.6. Experimental Settings

In this study, we use Python 3.11 as the programming language. Details on the versions and usage of frameworks and libraries are provided in Table 1. Furthermore, we use PyTorch's CUDA autocast, a feature designed to optimize DL model performance by automatically applying mixed precision during training and inference. We also employ gradient accumulation to reduce the GPU memory load [27].

For computational resources, we utilize Compute Canada, a national high-performance computing system. Dynamic resource allocation is applied based on the submitted job requirements. Each job may execute several Python scripts in parallel, with each script assigned to a separate node. A node is a computational unit comprising 40 CPUs with two cores each (Intel Gold 6148 Skylake @ 2.4 GHz), four GPUs (NVIDIA V100 SXM2 with 16 GB of memory each), and 186 GB of RAM.

4. Experimental Results

In each stage, we conduct evaluations on the test sets to assess model performance. Accordingly, we structure our experimental results as follows:

4.1. Image Segmentation

Table 2 presents the performance of the four variations mentioned in the approach. The reported metrics are MSE and intersection over union (IoU). MSE measures the difference between pixel values of the true and predicted masks, while IoU calculates the overlap between the predicted and the ground-truth region. Lower MSE values indicate better performance, while higher IoU percentages reflect more accurate overlap between the predicted and actual segmentation masks.

Table 2. Image Segmentation Results.

SAM 2 Variation	MSE	IoU(%)
Hiera-T	0.0149	94.35
Hiera-S	0.0162	94.94
Hiera-B+	0.0196	94.97
Hiera-L	**0.0086**	**95.30**

Hiera-L outperforms all other variations, achieving the lowest MSE (0.0086) and highest IoU (95.30%), indicating the most accurate segmentation and best overall performance. This is most likely due to Hiera-L's larger architecture, which encompasses more parameters and allows for more robust feature extraction. Moreover, smaller models like Hiera-T or Hiera-S are more likely to underfit complex data, especially in cases where there is high variability within the images.

4.2. Monocular Depth Estimation

In this phase, we use the scale-invariant logarithmic (Silog) loss, which combines both depth differences and scale-invariance errors to calculate the distance between the logarithm of depth pixel values [28]. Lower Silog loss values indicate better performance.

The MDE Silog value is 1.1339, indicating that the depth maps generated from the RGB images using the GLPN model are moderately accurate. This is likely due to the intricacy of the depth images in the Nutrition5k dataset. Such intricacy makes it challenging for the GLPN model to achieve highly accurate results. Indeed, capturing finer details and subtle depth variations in food images is difficult, which can affect depth estimation accuracy.

Improving this result may require additional refinement, such as further fine-tuning or using other FME datasets, to better handle the complexity present in this dataset.

4.3. Food Mass Estimation

In this stage, we use MSE and MAE to evaluate the distance between the true and predicted masses based on the AC images, which combine depth and segmented images generated using Hiera-L, as it is the best alternative for image segmentation. The evaluation results are 5.61 for MSE and 1.07 for MAE.

The low error values demonstrate that combining depth and Hiera-L-based segmentation images using AC provides a comprehensive input representation for several reasons. First, the robustness of the Hiera-L variation allows it to capture complex patterns in food images due to its larger capacity. Second, complementary insights from MDE and image segmentation contribute to improved accuracy; depth images offer valuable spatial information about the distance and volume of food ingredients, while segmented images delineate the contours of food regions. Lastly, the AC method efficiently merges the depth and segmented images, making contours and boundaries more pronounced while preserving the spatial information provided by MDE.

We further conducted a comparative study with other research papers that used the RGB images of the Nutrition5k dataset. The results are shown in Table 3.

Table 3. Comparative Study Results for FME.

Library	MAE
13	18.8
19	13.7
18	10.6
ours	**1.07**

The results presented in the table further demonstrate the superiority of our approach, as the reported MAE significantly outperforms that of previous studies.

5. Conclusions

In this study, we explored the use of advanced DL techniques to accurately estimate food mass. By leveraging state-of-the-art models such as Meta's SAM 2, the GLPN transformer, and the ViT transformer, alongside the AC technique, the study aimed to address the challenges of FME using RGB images. AC was utilized to effectively combine segmented and depth images, enhancing the visibility of depth contours and improving accuracy in FME. The experimental results highlight the effectiveness of these models and techniques in achieving high accuracy compared to prior research. As a future direction, the next step in this study involves implementing ingredient recognition for each dish, which, when combined with FME, can significantly enhance the accuracy of calorie

estimation. Additionally, ingredient recognition can be used to assess nutritional value in terms of fats, proteins, and carbohydrates, providing users with a more detailed analysis of their food intake.

Author Contributions:

Funding: This work was supported by the ResearchNB Talent Recruitment Fund program, under Application No. TRF-0000000170, awarded to G.R.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Konstantakopoulos, F.S.; Georga, E.I.; Fotiadis, D.I. A Review of Image-Based Food Recognition and Volume Estimation Artificial Intelligence Systems. *IEEE Rev. Biomed. Eng.* **2023**, *17*, 136–152. https://doi.org/10.1109/RBME.2023.3283149.
2. Rouhafzay, A.; Rouhafzay, G.; Jbilou, J. Image-Based Food Monitoring and Dietary Management for Patients Living with Diabetes: A Scoping Review of Calorie Counting Applications. *Front. Nutr.* **2025**, *12*, 1501946. https://doi.org/10.3389/fnut.2025.1501946.
3. Dugas, K.R.; Giroux, M.-A.; Guerroudj, A.; Leger, J.; Rouhafzay, A.; Rouhafzay, G.; Jbilou, J. Calorie Counting Apps for Monitoring and Managing Calorie Intake in Adults with Weight-Related Chronic Diseases: A Decade-long Scoping Review (2013–2024). *JMIR Prepr.* **2024**, 64139. https://doi.org/10.2196/preprints.64139.
4. Dehais, J.; Anthimopoulos, M.; Shevchik, S.; Mougiakakou, S. Two-View 3D Reconstruction for Food Volume Estimation. *IEEE Trans. Multimed.* **2016**, *19*, 1090–1099. https://doi.org/10.1109/TMM.2016.2642792.
5. Hassannejad, H.; Matrella, G.; Ciampolini, P.; De Munari, I.; Mordonini, M.; Cagnoni, S. A New Approach to Image-Based Estimation of Food Volume. *Algorithms* **2017**, *10*, 66.
6. Lo, F.P.-W.; Sun, Y.; Qiu, J.; Lo, B.P.L. Food Volume Estimation Based on Deep Learning View Synthesis from a Single Depth Map. *Nutrients* **2018**, *10*, 2005. https://doi.org/10.3390/nu10122005.
7. Liao, H.-C.; Lim, Z.-Y.; Lin, H.-W. Food Intake Estimation Method Using Short-Range Depth Camera. In Proceedings of the 2016 IEEE International Conference on Signal and Image Processing (ICSIP), Beijing, China, 13–15 August 2016; pp. 198–204.
8. He, Y.; Xu, C.; Khanna, N.; Boushey, C.J.; Delp, E.J. Food Image Analysis: Segmentation, Identification and Weight Estimation. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–19 July 2013; pp. 1–6.
9. Yang, Z.; Yu, H.; Cao, S.; Xu, Q.; Yuan, D.; Zhang, H.; Jia, W.; Mao, Z.-H.; Sun, M. Human-Mimetic Estimation of Food Volume from a Single-View RGB Image Using an AI System. *Electronics* **2021**, *10*, 1556.
10. Konstantakopoulos, F.; Georga, E.I.; Fotiadis, D.I. 3D Reconstruction and Volume Estimation of Food Using Stereo Vision Techniques. In Proceedings of the 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE), Kragujevac, Serbia, 25–27 October 2021; pp. 1–4.
11. Jia, W.; Ren, Y.; Li, B.; Beatrice, B.; Que, J.; Cao, S.; Wu, Z.; Mao, Z.-H.; Lo, B.; Anderson, A.K.; et al. A Novel Approach to Dining Bowl Reconstruction for Image-Based Food Volume Estimation. *Sensors* **2022**, *22*, 1493.
12. Amir, N.; Zainuddin, Z.; Tahir, Z. 3D Reconstruction with SFM-MVS Method for Food Volume Estimation. *Int. J. Comput. Digit. Syst.* **2024**, *16*, 1–11.
13. Naritomi, S.; Yanai, K. Hungry Networks: 3D Mesh Reconstruction of a Dish and a Plate from a Single Dish Image for Estimating Food Volume. In Proceedings of the 2nd ACM International Conference on Multimedia in Asia, Tokyo, Japan, 7–9 March 2021; pp. 1–7.
14. Ando, Y.; Ege, T.; Cho, J.; Yanai, K. DepthCalorieCam: A Mobile Application for Volume-Based Food-Calorie Estimation Using Depth Cameras. In Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, Nice, France, 21 October 2019; pp. 76–81.

15. Thames, Q.; Karpur, A.; Norris, W.; Xia, F.; Panait, L.; Weyand, T.; Sim, J. Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 8903–8911.
16. Pouladzadeh, P.; Shirmohammadi, S.; Al-Maghrabi, R. Measuring Calorie and Nutrition from Food Image. *IEEE Trans. Instrum. Meas.* **2014**, *63*, 1947–1956.
17. Okamoto, K.; Yanai, K. An Automatic Calorie Estimation System of Food Images on a Smartphone. In Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, Amsterdam, The Netherlands, 16 October 2016; pp. 63–70.
18. Yang, Y.; Jia, W.; Bucher, T.; Zhang, H.; Sun, M. Image-Based Food Portion Size Estimation Using a Smartphone without a Fiducial Marker. *Public Health Nutr.* **2019**, *22*, 1180–1192.
19. Lo, F.P.-W.; Sun, Y.; Qiu, J.; Lo, B.P.L. Point2Volume: A Vision-Based Dietary Assessment Approach Using View Synthesis. *IEEE Trans. Ind. Inform.* **2019**, *16*, 577–586.
20. Han, Y.; Cheng, Q.; Wu, W.; Huang, Z. DPF-Nutrition: Food Nutrition Estimation via Depth Prediction and Fusion. *Foods* **2023**, *12*, 4293.
21. Shao, W.; Hou, S.; Jia, W.; Zheng, Y. Rapid Non-Destructive Analysis of Food Nutrient Content Using Swin-Nutrition. *Foods* **2022**, *11*, 3429.
22. Gao, J.; Tan, W.; Ma, L.; Wang, Y.; Tang, W. MuseFood: Multi-Sensor-Based Food Volume Estimation on Smartphones. In Proceedings of the 2019 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), Leicester, UK, 19–23 August 2019; pp. 899–906.
23. Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. SAM 2: Segment Anything in Images and Videos. *arXiv* **2024**, arXiv:2408.00714.
24. Kim, D.; Ka, W.; Ahn, P.; Joo, D.; Chun, S.; Kim, J. Global-Local Path Networks for Monocular Depth Estimation with Vertical CutDepth. *arXiv* **2022**, arXiv:2201.07436.
25. Maji, S.; Nath, A. Scope and Issues in Alpha Compositing Technology. *Int. Issues Alpha Compos. Technol.* **2016**, *2*, 38–43.
26. Dosovitskiy, A. An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
27. Nokhwal, S.; Chilakalapudi, P.; Donekal, P.; Nokhwal, S.; Pahune, S.; Chaudhary, A. Accelerating Neural Network Training: A Brief Review. In Proceedings of the 2024 8th International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence, Dubai, United Arab Emirates, 22–23 February 2024; pp. 31–35.
28. Eigen, D.; Puhrsch, C.; Fergus, R. Depth Map Prediction from a Single Image Using a Multi-Scale Deep Network. *Adv. Neural Inf. Process. Syst.* **2014**, *27*.