*Original Research Article*

Estimating Nutritional Composition from Food Volume via Deep Learning-Based Depth and Segmentation Models

.

ABSTRACT

|  |
| --- |
| Nutrition plays a critical role in human health, with a balanced diet being essential for preventing non-communicable diseases, enhancing immune function, and improving quality of life. However, dietary imbalances contribute to significant global health issues, including obesity and malnutrition, which have far-reaching economic and health consequences. This research aims to address these challenges by developing a method for estimating the nutritional content of food items from a single 2D image. Our approach integrates a U-Net architecture with a ResNet18 encoder for depth prediction and employs FoodSAM for precise food segmentation. These components enable the calculation of food volume and mass, which are then used to estimate nutritional content based on the USDA database. Experimental results show that our model achieves a mean relative error (MRE) ranging from 11.18% to 50.35% for individual food items. Furthermore, our method maintains consistent mass predictions across various scenarios, including complex food combinations. This method demonstrates robustness in handling foods with diverse shapes and colors, providing a solid foundation for practical dietary tracking applications. By enabling nutritional monitoring, our approach has the potential to support public health initiatives and promote healthier lifestyles. |

*Keywords: Computer vision, Nutritional estimation, Volume estimation, Depth estimation, Food segmentation.*

1. INTRODUCTION

Nutrition is an essential factor for human health and development. According to the World Health Organization (WHO), nutrition directly impacts maternal and child health while strengthening the immune system. A well-balanced diet helps reduce the risk of non-communicable diseases such as diabetes and cardiovascular conditions, ultimately contributing to increased life expectancy (World Health Organization, 2018). Maintaining an adequate nutritional intake not only supports physical development but also enhances work productivity and improves overall quality of life (World Health Organization, 2019). Thus, the importance of nutrition is undeniable, playing a central role in improving public health and well-being.

An imbalanced diet can lead to severe health problems, with obesity being one of the most critical concerns. Obesity is a chronic disease that increases the risk of conditions such as type 2 diabetes, cardiovascular disorders, cancer, and skeletal issues. Additionally, it negatively impacts mobility and sleep quality, reducing overall life satisfaction (World Health Organization, 2024b). In 2022, WHO estimated that 2.5 billion adults were overweight, including 890 million classified as obese. The global overweight rate surged from 25% in 1990 to 43% in 2022, with significant prevalence in low- and middle-income regions. The number of obese children under five reached 37 million, with the most noticeable increase in Asia and Africa. Among children aged 5–19, obesity rates rose from 2% in 1990 to 8% in 2022, amounting to 160 million affected children. Obesity also imposes substantial economic burdens, with projected costs reaching 3 trillion by 2030 and 18 trillion by 2060 if preventive measures are not implemented (World Health Organization, 2024b). These figures indicate that obesity is becoming a global health crisis, affecting all age groups and regions, primarily due to unhealthy eating habits and sedentary lifestyles.

Beyond obesity, nutritional imbalance can lead to malnutrition, a condition where the body lacks essential nutrients. Malnutrition severely impacts millions of children, with 149 million under five suffering from stunting, 45 million underweight, and 37 million classified as overweight or obese. It is responsible for 45% of child deaths under five, predominantly in low- and middle-income countries, particularly in Africa and Asia (World Health Organization, 2024a). Malnutrition also hinders cognitive development, reducing IQ and learning capabilities, leading to long-term consequences on workforce productivity. The issue remains critical in South and Southeast Asia (World Health Organization, 2024a). Given its direct impact on societal and economic growth, effective nutrition management is essential for improving public health. This situation presents a challenge in maintaining a balanced diet, ensuring adequate nutrient intake without leading to either malnutrition or obesity. Scientific nutritional monitoring optimizes metabolic processes and ensures each meal provides appropriate energy levels.

The development of computer vision technology has introduced various approaches to nutritional monitoring through food imagery. Recent studies have utilized deep learning models to estimate calorie intake, employing image segmentation and food recognition techniques to determine food volume and mass. For example, Agarwal et al. (2023) applied a deep learning model to estimate calories for 30 food types, requiring a reference object such as a coin to establish actual size. However, this method depends on clear images and the presence of a reference object, posing inconvenience. In another study, Ma et al. (2023) used a Conditional Generative Adversarial Network (cGAN) to extract a tensor density map from images of 21 food items and estimate calorie content, though the data collection process was time- and labor-intensive. Furthermore, Kumar et al. (2021) employed a Support Vector Machine (SVM) and Multilayer Perceptron (MLP) to identify simple foods such as apples and bananas, but this required users to capture images from multiple angles, adding inconvenience. Existing methods are limited in the number of recognizable dishes and often necessitate extensive imaging procedures, reducing their practicality.

In this study, we propose a method for estimating the nutritional value of food based on a single input image. The input image is processed through a depth estimation network to generate a depth map, while simultaneously, a food segmentation network creates labeled segmentation masks for up to 90 distinct food types. These outputs are then merged into a point cloud, and a volume estimation algorithm is utilized to determine the food volume. Subsequently, we convert this volume into mass using a food density dataset that we have assembled. The nutritional value of each identified food type is thereafter obtained from the United States Department of Agriculture (USDA) database. This approach offers potential as a practical tool for users to monitor nutrition in their daily meals.

2. LITERATURE REVIEW

Estimating the volume of food from images is a crucial problem in nutrition tracking and diet management applications. Research in this field can be broadly categorized into two main groups: traditional computer vision methods and deep learning approaches. Traditional methods often require reference objects or multiple viewpoints for calibration, while deep learning methods can predict food volume from a single 2D image without reference objects. Each group has its own advantages and disadvantages, affecting their practical applicability.

2.1 Traditional computer vision methods

Traditional methods typically require users to use objects of known sizes, such as coins, thumbs, or calibration boards, to determine the actual scale from the photo, thereby estimating the food volume (Ege et al., 2019; Kadam et al., 2022). For example, a study by Kasyap and Jayapandian proposed a model combining Convolutional Neural Network with Random Forest and SVM using the ECUSTFD dataset, where each image contains food with a coin as a reference object to determine volume and weight, helping estimate calories (Kasyap & Jayapandian, 2021). In the study conducted by Shi et al. (2024), the nutritional estimation method evaluates tray meal volume through image-based detection. This approach converts pixel coordinates into real-world dimensions by utilizing the tray's actual size as a reference. The method underscores its critical dependence on the tray's real physical dimensions to ensure the accuracy of the estimation.

Other studies are based on 3D reconstruction models, requiring users to take multiple photos from different angles or record videos to extract frames and rebuild the 3D model of the food phẩm (Dehais et al., 2017; Subhi et al., 2018). Additionally, available tools like IOS ARKit can be used to predict depth, thereby calculating the volume of the dish (Ege et al., 2019). A team led by Shao used RGB-D images to estimate food volume, combining image segmentation algorithms with depth data to achieve high accuracy (Shao et al., 2023).

Although these methods can achieve high accuracy, they have significant limitations, particularly requiring users to perform many manual operations such as carrying reference objects or taking photos from multiple angles, which is inconvenient for deploying in daily nutrition tracking applications.

2.2 Deep learning approaches

To address these limitations, recent studies have applied deep learning to estimate food volume from a single 2D image without reference objects or multiple viewpoints. One study used deep learning architectures to predict depth maps from single images, then used this information to reconstruct point clouds to estimate food volume (Graikos et al., 2020). In another study (Oguntimilehin et al., 2024) combined CNN and Google Gemini Pro Vision to classify food types and estimate calories from images, using preprocessing and edge detection for feature enhancement. According to the research conducted by Feng et al. (2025), the authors integrated RegNet Y models pretrained with SWAG weights, combining RegNet Y128, Y32, and Y16 through a gated classifier head. They further enhanced model performance using data augmentation techniques such as cutmix and mixup, along with focal loss to address class imbalance. This approach effectively improved both dish classification accuracy and nutrient prediction.

An improved method using a cGAN model has been proposed to estimate food energy from a single image without depth data or multi-angle images, expanding its applicability on common mobile devices (Jaswanthi et al., 2022; Ma et al., 2023). However, this method still faces challenges in training the model, as it requires a large dataset with high-quality labels, including food images accompanied by volume, composition, and nutritional information, making data collection costly and limiting practical applicability.

Another study proposed using depth sensor cameras to improve accuracy in estimating food volume and calories, exemplified by the Nutrition5k system – a dataset built with over 5,000 food samples including RGB, video, depth images, food weight, and detailed nutritional values (Thames et al., 2021). The study by Shi et al. (2025) utilizes the Intel Realsense depth camera and deep learning to automatically detect calories in vegetables and dishes. Image classification is enhanced through the CBAM attention mechanism and DenseNet264. Volume estimation for calorie calculation employs 3D reconstruction, Principal Component Analysis, and the Monte Carlo algorithm. The approach shows potential for application in robotic catering services. However, both methods depend on depth sensor cameras, which are rare in standard devices, raising hardware costs and reducing their practical usability and accessibility for daily use.

To facilitate deployment across all devices without requiring specialized sensors, recent studies have employed self-supervised deep learning techniques, allowing models to predict depth from a single 2D image without the need for labeled data. Some studies have demonstrated that models can reconstruct depth from single images and self-learn spatial structures of objects through frame reconstruction techniques (Graikos et al., 2020). The results show that this method can estimate food volume when food is placed on a flat surface but faces limitations when food is not on a flat surface or when the distance between the camera and food is not accurately determined.

In summary, traditional methods relying on reference objects or multiple viewpoints can achieve high accuracy but are impractical and less suitable for broad deployment due to their inconvenience. Approaches utilizing depth sensors enhance accuracy yet depend on specialized hardware, limiting their practical applicability. In contrast, the use of self-supervised deep learning with 2D images presents considerable potential for enhancing food recognition and analysis capabilities across all devices, enabling users to conveniently track and manage nutrition without the need for specialized equipment.

3. METHODOLOGY

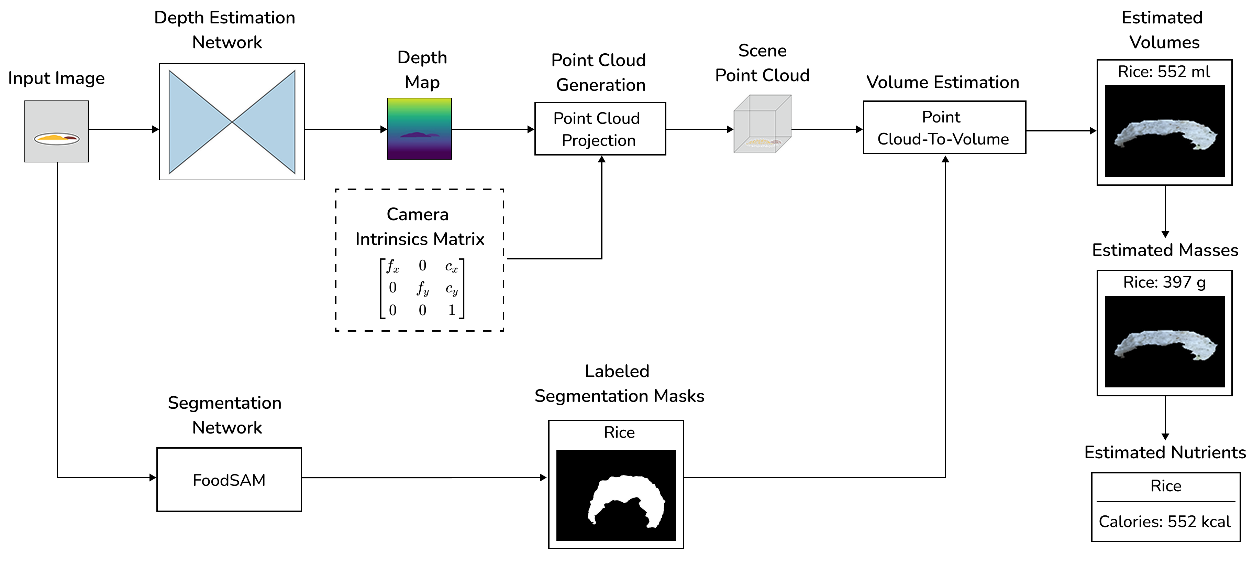
****

Fig. 1. Proposed Model Architecture

We introduce a method for estimating the nutritional value of food from a single 2D image, as illustrated in Fig 1, with the capability to recognize a wide range of globally common food types. This method integrates a depth estimation network (Graikos et al., 2020) to generate a depth map of the food, employs a segmentation network (Lan et al., 2024) to identify individual food items within the image, and subsequently applies a point cloud processing algorithm to calculate volume (Graikos et al., 2020). The system then converts the volume into mass and provides detailed nutritional information, facilitating more convenient nutrition tracking. Our nutritional estimation model comprises four main components: the Depth Estimation Network, the Segmentation Network, the Volume Estimation algorithm, and the nutritional estimation process.

3.1 The Depth Estimation Network

The depth estimation network plays a vital role in providing information about the distance from the camera to objects in the image. We employ the network used in the study by Graikos et al. (2020), pre-trained on the EPIC-KITCHENS dataset, which comprises over 50 hours of video from six kitchens capturing food preparation processes vividly with flexible camera angles and subsequently refined with data collected by the authors. Through a self-supervised approach, this model effectively addresses the challenge of limited real-world depth data. Trained with self-supervised techniques, it surmounts the obstacle of limited ground-truth depth information.

The model adopts the U-Net architecture [18] with ResNet18 [19] as the encoder to predict the depth map from a single input image. In the absence of real-world depth data, the system is trained using a self-supervised approach, incorporating an additional network based on ResNet18 to estimate camera pose. However, this ResNet18 architecture is modified to accept a 6-channel input, consisting of two concatenated RGB images, and outputs a 6-dimensional vector (6 degrees of freedom, 6-DoF), comprising rotation represented in axis-angle form and translation, describing the camera’s positional and orientational changes between frames. The training data consists of monocular video sequences divided into three consecutive frames ​, and . The image ​ is fed into the U-Net depth estimation network to generate the depth map ​. Concurrently, is processed by the camera motion estimation network (ResNet18) to compute camera transformations between frames, from to () and from to (). Using the predicted depth map of the current frame and the camera transformation estimates ( ), the model reconstructs the preceding and following frames ​ and ​. After reconstruction, the system compares the reconstructed frames with the original frames to calculate the loss function. This enables the model to learn in a self-supervised manner without requiring ground-truth depth data, thereby enhancing the U-Net’s ability to produce the depth map from a single input image.

3.2 The Segmentation Network

We utilize a model pre-trained on the FoodSeg103 dataset (Lan et al., 2024), as shown in Fig 2. Specifically, this approach combines the Semantic Segmenter (Zheng et al., 2021) with the Segment Anything Model (SAM) (Kirillov et al., 2023) to generate labeled binary food masks. The method leverages the strengths of both models: the Semantic Segmenter provides classified masks with semantic information but limited precision, while SAM produces high-quality masks without classification details. This integration enables the segmentation network to generate adaptive segmentations that better handle the complex structures of food images.

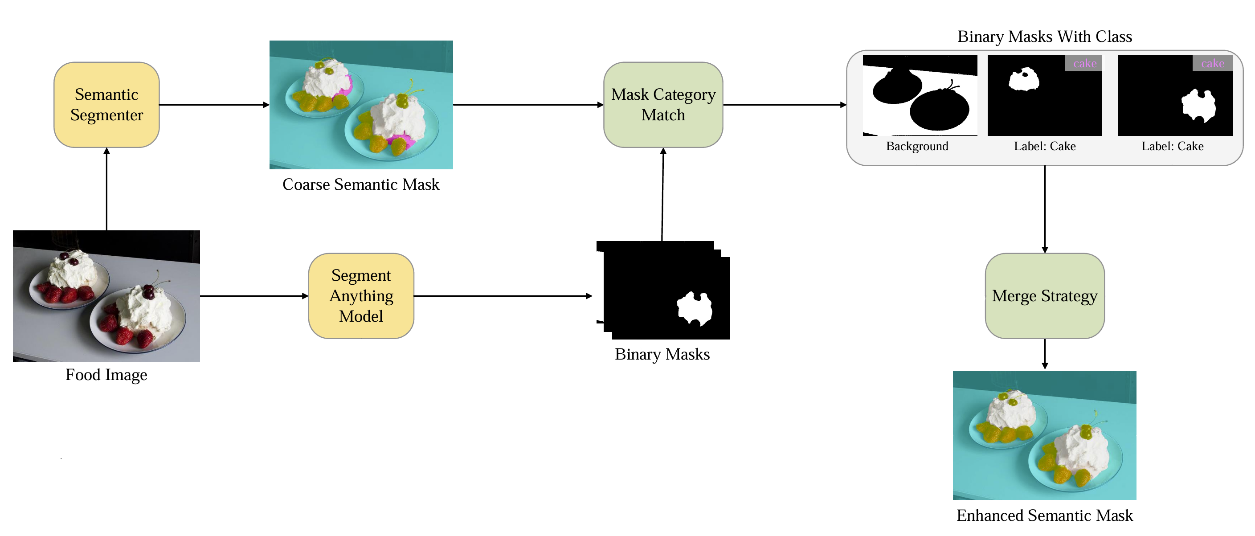


Fig. 2. Food segmentation diagram extracted from FoodSAM (Lan et al., 2024)

The image segmentation process begins with the input image being fed into the Semantic Segmenter, which generates a coarse semantic mask containing food category labels for each pixel but with some noise. Simultaneously, the same image is processed by SAM, producing multiple sharp binary masks lacking category information. To address this, a mask category matching method is applied by comparing the SAM-generated masks with the coarse mask and using a voting mechanism to select the most appropriate semantic label, filtering out noisy labels. Finally, the labeled binary masks are combined using a merge strategy that prioritizes larger regions, resulting in an enhanced semantic mask that integrates clear boundaries from SAM with food category labels from the Semantic Segmenter, thereby improving the overall quality of image segmentation.

3.3 The Volume Estimation algorithm

After obtaining the depth map of the input image, the camera’s intrinsic parameter matrix , and the labeled food segmentation mask, we apply the volume estimation algorithm outlined in the study by Graikos et al. (2020) (Fig 3). Initially, a point cloud is generated by projecting each pixel into a corresponding point in three-dimensional space using its homogeneous coordinates and the inverse projection model, as described by the following equation:

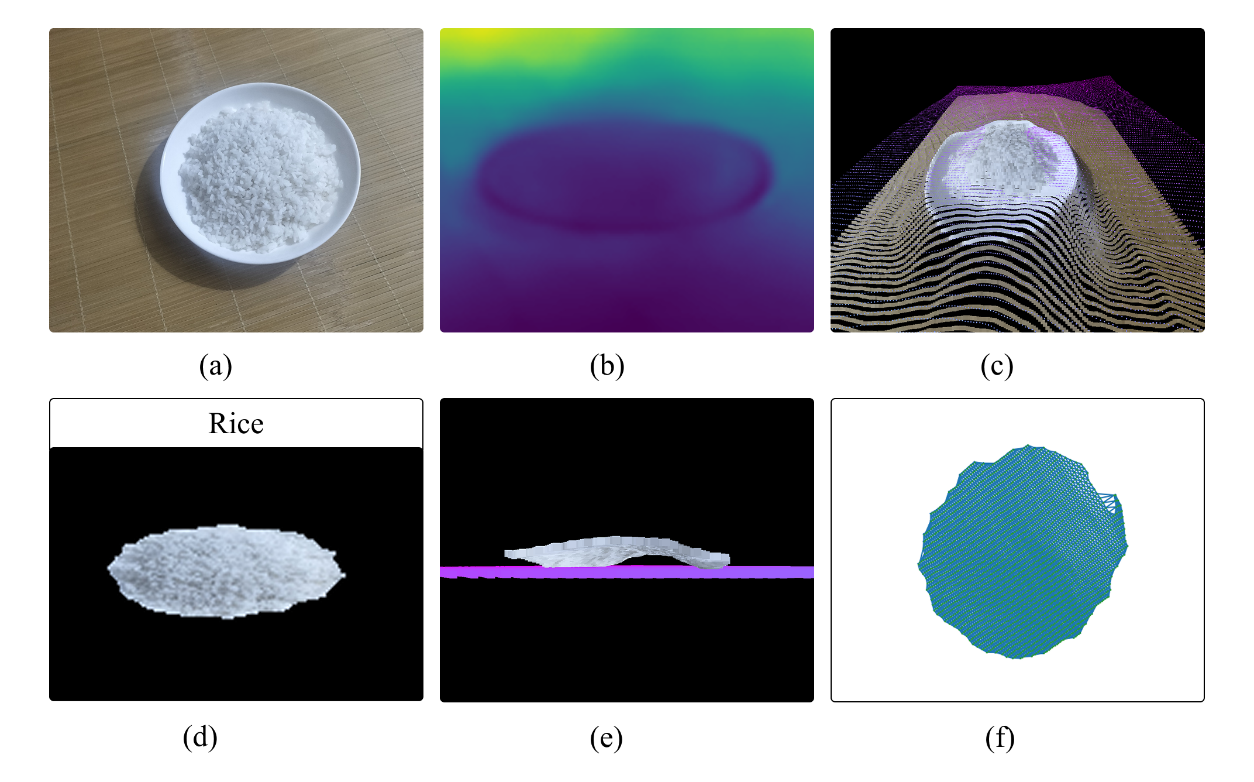


Fig. 3. Food Volume Estimation Process. (a) Input image. (b) Estimated depth map. (c) Point cloud. (d) Labeled food segmentation mask. (e) Estimated base plane. (f) Delaunay triangulation of the food

This set of points forms the point cloud representation . Next, the labeled food segmentation masks, obtained from the food segmentation network, are applied to the image to identify and distinguish different food items. Subsequently, the point cloud is divided into subsets corresponding to each individual food object.

Each point group is processed using a Statistical Outlier Removal (SOR) filter to eliminate inconsistent points. The base plane is determined through Principal Component Analysis (PCA), where the plane’s normal vector corresponds to the least significant eigenvector. This plane is further adjusted to ensure it aligns with the bottom of the object. The final step in volume estimation involves projecting the food points onto the base plane and partitioning the covered region into triangles using an derived from Delaunay triangulation Delaunay (Edelsbrunner & Harer, 2010). The volume is then calculated from triangular prisms, with the triangles serving as bases and the height defined as the average distance from the vertices to the food points.

This method is applicable only when the food is placed on a flat surface. In cases where the food is contained within objects such as bowls, the volume is no longer determined by the food itself but is instead dictated by the container.

3.4 The nutritional estimation process

After determining the volume of the food, the next step in the nutritional estimation process is to calculate the food’s mass. This requires knowledge of the food’s density, as mass density, or food density, is defined as its mass per unit volume. This density can vary depending on temperature and pressure (Charrondiere et al., 2012). Accurately determining the food’s density is essential for converting volume to mass and ensuring the reliability of the nutritional estimation.

In this study, we utilize the database from the Food and Agriculture Organization of the United Nations (FAO) (Charrondiere et al., 2012), which includes density information for 638 food types. This database provides density values gathered from international studies and validated sources, contributing to the reliability of our calculations. Additionally, we incorporate food density data from Aqua-Calc (AquaCalc LLC, n.d.), a reputable online tool offering density information for over 4,000 different foods and ingredients. Aqua-Calc encompasses not only internationally recognized foods but also data on distinctive dishes from various countries, thereby broadening the scope and applicability of the dataset. By integrating data from FAO and Aqua-Calc, we constructed a comprehensive database containing density information for a variety of foods, as illustrated in Table 1.

Table 1. Density of Various Food Types

| Food Item | Density (g/ml) | Source |
| --- | --- | --- |
| Steak | 1.04 | FAO |
| Pork | 0.63 | FAO |
| Egg | 0.37 | FAO |
| Apple | 0.24 | FAO |
| Rice | 0.72 | FAO |
| Avocado | 1.01 | Aqua-Calc |
| Cherry | 0.63 | Aqua-Calc |
| Ginger | 0.41 | Aqua-Calc |
| Walnut | 0.49 | Aqua-Calc |
| Pizza | 0.95 | Aqua-Calc |

After obtaining the density of the food types, the mass of the food is estimated by multiplying the estimated volume of the food by the specific density of the food, as shown in equation:

Where is the food's mass in grams (g), is its volume in milliliters (ml), and is its density in grams per milliliter (g/ml).

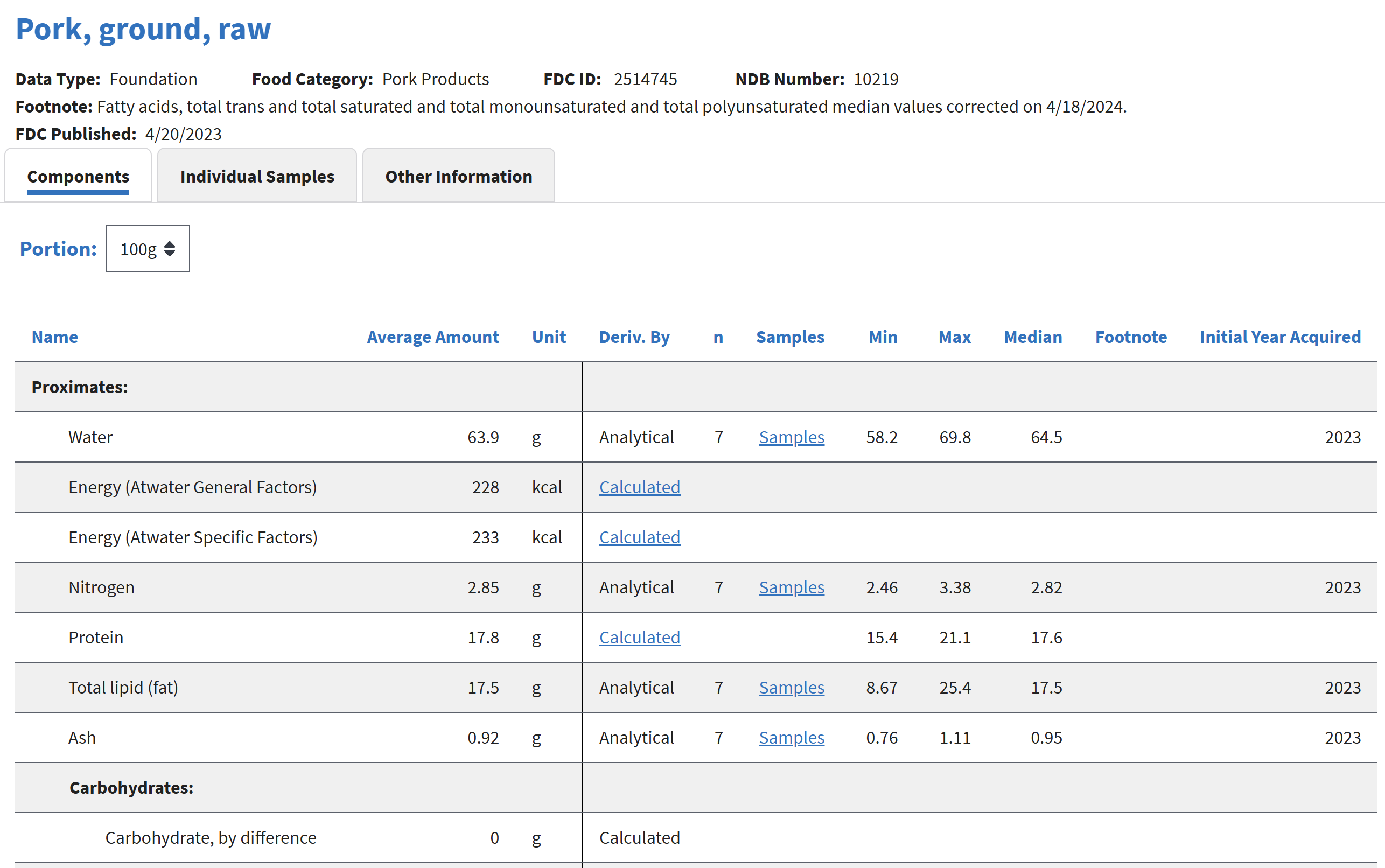


Fig. 4. Selected nutritional information of pork sourced from the USDA (U.S. Department of Agriculture, n.d.)

After estimating the mass of the food, we proceed to retrieve nutritional data from the API provided by the United States Department of Agriculture (USDA), as illustrated in Fig 4. The USDA is a governmental agency responsible for overseeing domains related to agriculture, food, and nutrition (U.S. Department of Agriculture, n.d.). Through the USDA Food Data Central API, we can efficiently compute detailed nutritional information, including metrics such as calories, protein, carbohydrates, and fats, for each food type.

4. RESULTS AND DISCUSSION

The experiments were conducted on a Dell Inspiron 16 Plus 7620 laptop equipped with an Intel Core i7-12700H processor, 40 GB of RAM, a 1024 GB SSD, and an Intel Iris Xe Graphics GPU. To evaluate the performance of our proposed model in estimating nutritional content, which primarily depends on accurate food mass estimation, we constructed a test dataset comprising 11 common food types: apple, steak, bread, cake, chicken, eggs, pork, potato, rice, cucumber, and carrot. Each food type was photographed from three different angles, and the actual mass of each sample was measured using a high-sensitivity precision scale, providing labeled mass data for objective model evaluation. In addition to single-food images, the dataset was expanded to include combinations of two to four food types in a single image, such as steak–potato, rice–chicken, rice–chicken–cucumber, rice–pork, rice–pork–cucumber, and rice–cucumber–pork–carrot. These combinations were designed to simulate real-world meal scenarios, enhancing the dataset's complexity and practical applicability. Our proposed model was compared against the volume estimation model developed by Graikos et al. (2020). Since the dataset focuses on mass estimation, the volume values predicted by the Graikos model were multiplied by a food density table compiled by us to convert them into mass estimates. We evaluate the model using the Mean Relative Error (MRE), where represents the actual value, denotes the predicted value, and is the number of samples. MRE measures the average relative difference between the predicted and actual values, expressed as a percentage, providing an intuitive insight into the model's accuracy.

Table 2. Experimental results comparing food weight estimation methods. GT: Ground Truth (actual weight in grams). Ours: Proposed model.

| Images | Food Item | GT (g) | Ours (g) | | | (Graikos et al., 2020) (g) | | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Apple | 165 | 108 | 106 | 84 | 57 | 37 | 43 |
|  | Steak | 152 | 169 | 168 | 134 | 146 | 145 | 109 |
|  | Bread | 62 | 76 | 72 | 72 | 67 | 65 | 64 |
|  | Cake | 140 | 154 | 204 | 123 | null | null | null |
|  | Chicken | 250 | 286 | 288 | 263 | 251 | 303 | 235 |
|  | Eggs | 284 | 172 | 96 | 155 | null | null | null |
|  | Pork | 110 | 101 | 166 | 104 | 74 | 132 | 100 |
|  | Potato | 408 | 469 | 486 | 567 | 550 | 526 | 561 |
|  | Rice | 300 | 364 | 454 | 372 | 308 | 463 | null |
|  | Steak | 152 | 178 | 113 | 98 | 36 | 111 | 95 |
| Potato | 408 | 830 | 406 | 588 | 104 | 8 | 6 |
|  | Rice | 300 | 730 | 377 | 325 | null | 254 | 167 |
| Chicken | 250 | 199 | 235 | 100 | 168 | 219 | 97 |
|  | Rice | 300 | 862 | 214 | 614 | null | null | null |
| Chicken | 250 | 212 | 102 | 247 | 146 | 117 | 199 |
| Cucumber | 39 | 91 | 81 | 195 | null | null | null |
|  | Rice | 300 | 422 | 270 | 308 | null | 227 | null |
| Pork | 110 | 114 | 162 | 223 | 2 | 4 | 9 |
|  | Rice | 300 | 357 | 392 | 450 | null | null | null |
| Pork | 110 | 207 | 198 | 113 | 8 | 5 | 11 |
| Cucumber | 39 | 148 | 169 | 106 | 25 | 20 | null |
|  | Rice | 300 | 409 | 579 | 353 | null | null | null |
| Pork | 110 | 220 | 143 | 318 | 2 | 6 | 95 |
| Cucumber | 39 | 151 | 108 | 209 | 18 | null | 7 |
| Carrot | 13 | 18 | 12 | 23 | 10 | 7 | 5 |

The results in Table 2 show that our proposed model surpasses the Graikos model in estimating the mass of food items for both single and combined food scenarios. For single food item images, our model achieves higher accuracy on a dataset of 18 images, each containing one of nine different food types, as evidenced by a lower MRE in most cases. For images combining multiple food items, our model successfully estimates the mass of all food components present, whereas the Graikos model frequently produces null values, failing to provide predictions. These findings highlight the robustness and versatility of our approach compared to the baseline method.

4.1 Evaluation on Single Food Item Images

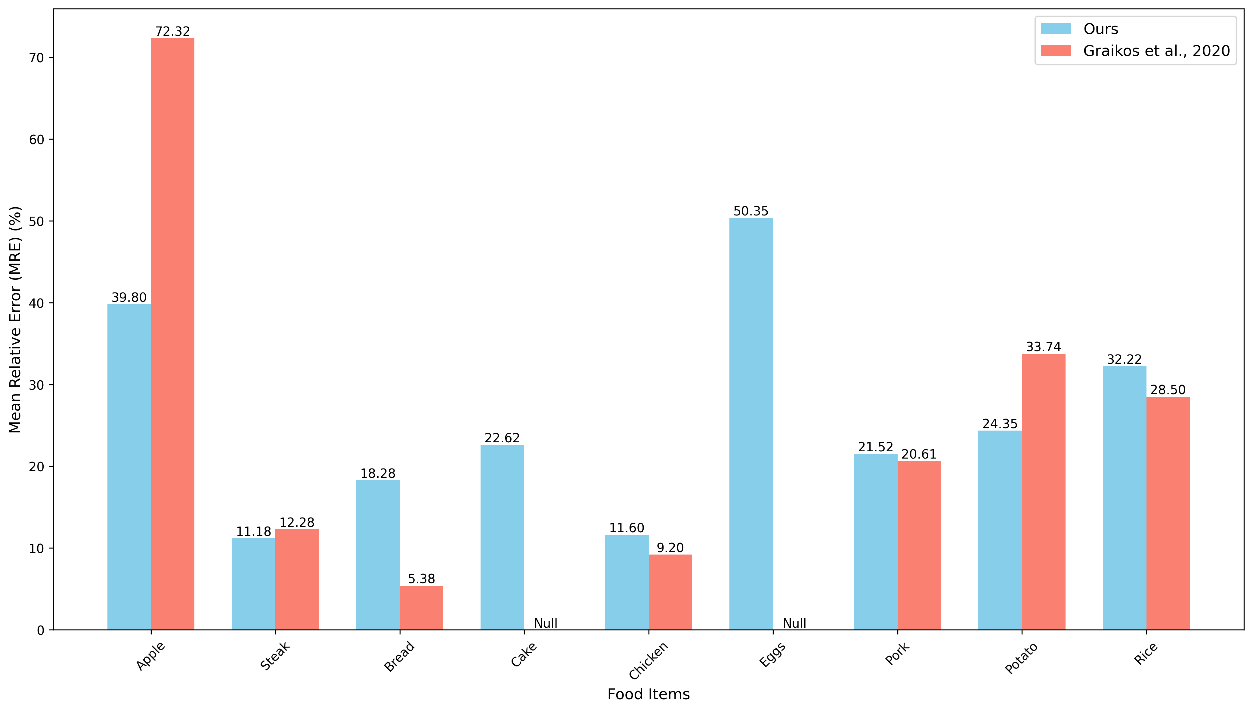


Fig. 5. Mean Relative Error (MRE) Comparison Between Our Model and (Graikos et al., 2020) in Single Food Item

For images of single food items, the Graikos model frequently fails to identify or predict the mass, resulting in null values, especially for challenging items such as cream cakes and eggs. This issue persists across all three viewing angles, as detailed in Table 2. In contrast, our model consistently identifies and delivers mass estimates that closely align with ground truth values for all tested items in this dataset. This superior performance underscores the effectiveness of our proposed method in overcoming the limitations of the Graikos model when processing individual food items.

As shown in Fig 5, the MRE of our model for single food item images ranges from 11.18% (Steak) to 50.35% (Eggs). This indicates stable performance for foods with simple shapes, though limitations remain with items exhibiting complex colors or structures, such as Eggs (50.35%) and Rice (32.22%). Foods with structures featuring significant voids from top to bottom, like eggs, pose a substantial challenge for our model. Specifically, the high error in egg mass estimation stems from their natural shape, which creates occluded voids from the top view to the plate. Consequently, the volume estimation algorithm incorporates air spaces into the food mass calculation, leading to overestimations beyond actual values.

Another influencing factor is food color. For example, cream cakes and eggs, which are white and blend with the plate background, pose difficulties for the Graikos model in mass estimation. In contrast, our proposed model mitigates this limitation by employing FoodSAM for more precise food segmentation, thereby enhancing identification and mass estimation efficiency. This improvement demonstrates the advantage of our approach in addressing challenges related to visual similarity between food items and their backgrounds in single food item scenarios.

4.2 Evaluation on Combined Food Item Images

For images combining two to four different food types, also presented in Table 2, our model exhibits relatively stable identification and mass estimation performance in certain cases. This is particularly evident with components of simple shapes, such as Steak (MRE 26.10%) and Chicken (MRE 25.20% at specific angles). However, significant challenges arise with smaller food items or those with colors similar to the background, such as Cucumber (MRE 213.67% to 300.00%) and Carrot (MRE 41.02%). The high errors in these cases likely result from the algorithm mistakenly identifying occluded voids as actual mass, causing substantial deviations.

In combined food item scenarios, the Graikos model frequently fails to predict masses, producing numerous null values and revealing its inability to handle multi-component data effectively. Conversely, our model consistently estimates the mass of all food items present in the images, demonstrating greater adaptability to diverse food types and complex compositions. Although our model does not yet achieve high accuracy in these combined scenarios, its ability to provide predictions across all tested cases, unlike the Graikos model, establishes a strong foundation for future improvements. This consistent performance highlights the potential of our approach to advance food mass estimation tasks in both single and multi-food contexts.

Despite its advancements, the proposed food mass estimation method faces significant limitations that affect its performance. The inherent variability in food characteristics, such as size, shape, color, and texture, poses challenges for achieving consistent image standardization and precise segmentation. This is particularly evident with items like eggs and rice, which yield mean relative errors (MRE) of 50.35% and 32.22%, respectively, due to their irregular shapes and visual complexity. Furthermore, complex backgrounds with varying lighting and shadow conditions exacerbate difficulties in accurate detection, especially for foods that visually merge with their surroundings, as evidenced by MRE reaching up to 300.00% for cucumbers in multi-food scenarios. The experimental dataset's limited diversity and scale restrict the model's ability to generalize across a broad spectrum of food types and real-world conditions. These constraints highlight the need for more diverse and comprehensive training datasets alongside advanced segmentation techniques, as well as improvements in depth estimation accuracy, to enhance the method’s robustness and precision in practical applications.

5. Conclusion

This study presents a method for estimating the nutritional composition of food from its volume using deep learning with depth and segmentation models. The approach integrates computer vision technology to perform food segmentation, volume estimation, and mass calculation, followed by nutrient inference from the USDA database. Experimental results demonstrate that the model achieves higher accuracy than previous methods, particularly in food recognition and mass estimation from a single image, enhancing convenience and applicability for daily nutritional tracking. However, certain challenges remain, such as volume estimation accuracy for irregularly shaped or small-sized foods. Future research directions may focus on improving depth estimation models to enhance precision, expanding the experimental dataset to include a broader variety of foods, and optimizing processing time for better real-world usability. Additionally, integrating this model into mobile applications or nutrition support systems could be a significant step toward helping users monitor their diet more efficiently, ultimately contributing to public health improvement.

DISCLAIMER (ARTIFICIAL INTELLIGENCE)

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

COMPETING INTERESTS

Authors have declared that no competing interests exist.

ACKNOWLEDGMENTS

This research is funded by Ho Chi Minh City University of Education Foundation for Science and Technology under the student scientific research project for the academic year 2024–2025.

References

Agarwal, R., Choudhury, T., Ahuja, N. J., & Sarkar, T. (2023). Hybrid Deep Learning Algorithm-Based Food Recognition and Calorie Estimation. *Journal of Food Processing and Preservation*, *2023*(1), 6612302. https://doi.org/10.1155/2023/6612302

AquaCalc LLC. (n.d.). *Aqua-Calc*. Aqua-Calc. Retrieved February 12, 2025, from https://www.aqua-calc.com/

Charrondiere, U. R., Haytowitz, D., & Stadlmayr, B. (2012). FAO/INFOODS Density Database version 2. *Food and Agriculture Organization of the United Nations Technical Workshop Report*. https://www.fao.org/4/ap815e/ap815e.pdf

Dehais, J., Anthimopoulos, M., Shevchik, S., & Mougiakakou, S. (2017). Two-View 3D Reconstruction for Food Volume Estimation. *IEEE Transactions on Multimedia*, *19*(5), 1090–1099. IEEE Transactions on Multimedia. https://doi.org/10.1109/TMM.2016.2642792

Edelsbrunner, H., & Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Soc. https://doi.org/10.1090/mbk/069

Ege, T., Ando, Y., Tanno, R., Shimoda, W., & Yanai, K. (2019). Image-Based Estimation of Real Food Size for Accurate Food Calorie Estimation. *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, 274–279. https://doi.org/10.1109/MIPR.2019.00056

Feng, Y., Wang, Y., Wang, X., Bi, J., Xiao, Z., & Luo, Y. (2025). Large-scale image classification and nutrient estimation for Chinese dishes. *Journal of Agriculture and Food Research*, *19*, 101733. https://doi.org/10.1016/j.jafr.2025.101733

Graikos, A., Charisis, V., Iakovakis, D., Hadjidimitriou, S., & Hadjileontiadis, L. (2020). Single Image-Based Food Volume Estimation Using Monocular Depth-Prediction Networks. In M. Antona & C. Stephanidis (Eds.), *Universal Access in Human-Computer Interaction. Applications and Practice* (Vol. 12189, pp. 532–543). Springer International Publishing. https://doi.org/10.1007/978-3-030-49108-6\_38

Jaswanthi, R., Amruthatulasi, E., Bhavyasree, Ch., & Satapathy, A. (2022). A Hybrid Network Based on GAN and CNN for Food Segmentation and Calorie Estimation. *2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS)*, 436–441. https://doi.org/10.1109/ICSCDS53736.2022.9760831

Kadam, P., Pandya, S., Phansalkar, S., Sarangdhar, M., Petkar, N., Kotecha, K., & Garg, D. (2022). FVEstimator: A novel food volume estimator Wellness model for calorie measurement and healthy living. *Measurement*, *198*, 111294. https://doi.org/10.1016/j.measurement.2022.111294

Kasyap, V. B., & Jayapandian, N. (2021). Food Calorie Estimation using Convolutional Neural Network. *2021 3rd International Conference on Signal Processing and Communication (ICPSC)*, 666–670. https://doi.org/10.1109/ICSPC51351.2021.9451812

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., & Girshick, R. (2023). Segment Anything. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 3992–4003. https://doi.org/10.1109/ICCV51070.2023.00371

Kumar, R. D., Julie, E. G., Robinson, Y. H., Vimal, S., & Seo, S. (2021). Recognition of food type and calorie estimation using neural network. *The Journal of Supercomputing*, *77*(8), 8172–8193. https://doi.org/10.1007/s11227-021-03622-w

Lan, X., Lyu, J., Jiang, H., Dong, K., Niu, Z., Zhang, Y., & Xue, J. (2024). FoodSAM: Any Food Segmentation. *IEEE Transactions on Multimedia*, 1–14. https://doi.org/10.1109/TMM.2023.3330047

Ma, J., He, J., & Zhu, F. (2023). An Improved Encoder-Decoder Framework for Food Energy Estimation. *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, 53–59. https://doi.org/10.1145/3607828.3617795

Oguntimilehin, A., Ayirioritse, O., Atachin, A. J., Toyin, O., Tope-Oke, A. M., Abiola, O. B., Jaiyesimi, B. G., Sanya, O. A., & Oluwatoki, T. G. (2024). Convolutional Neural Network Based Food Calorie Estimation System for Dietary Tracking. *2024 IEEE 5th International Conference on Electro-Computing Technologies for Humanity (NIGERCON)*, 1–5. https://doi.org/10.1109/NIGERCON62786.2024.10927328

Shao, W., Min, W., Hou, S., Luo, M., Li, T., Zheng, Y., & Jiang, S. (2023). Vision-based food nutrition estimation via RGB-D fusion network. *Food Chemistry*, *424*, 136309. https://doi.org/10.1016/j.foodchem.2023.136309

Shi, J., Han, Q., Cao, Z., & Wang, Z. (2024). DeepTrayMeal: Automatic dietary assessment for Chinese tray meals based on deep learning. *Food Chemistry*, *434*, 137525. https://doi.org/10.1016/j.foodchem.2023.137525

Shi, Y., Gao, W., Shen, T., Li, W., Li, Z., Huang, X., Li, C., Chen, H., Zou, X., & Shi, J. (2025). Calorie detection in dishes based on deep learning and 3D reconstruction. *Computers and Electronics in Agriculture*, *229*, 109832. https://doi.org/10.1016/j.compag.2024.109832

Subhi, M. A., Md. Ali, S. H., Ismail, A. G., & Othman, M. (2018). Food volume estimation based on stereo image analysis. *IEEE Instrumentation & Measurement Magazine*, *21*(6), 36–43. IEEE Instrumentation & Measurement Magazine. https://doi.org/10.1109/MIM.2018.8573592

Thames, Q., Karpur, A., Norris, W., Xia, F., Panait, L., Weyand, T., & Sim, J. (2021). *Nutrition5k: Towards Automatic Nutritional Understanding of Generic Food*. 8903–8911. https://openaccess.thecvf.com/content/CVPR2021/html/Thames\_Nutrition5k\_Towards\_Automatic\_Nutritional\_Understanding\_of\_Generic\_Food\_CVPR\_2021\_paper.html

U.S. Department of Agriculture. (n.d.). *FoodData Central* [U.S. Department of Agriculture]. Retrieved February 12, 2025, from https://www.usda.gov/

World Health Organization. (2018). *Nutrition*. https://www.who.int/news-room/facts-in-pictures/detail/nutrition

World Health Organization. (2019). *The State of Food Security and Nutrition in the World 2019: Safeguarding Against Economic Slowdowns and Downturns*. Food & Agriculture Org. https://www.who.int/publications/m/item/state-of-food-security-and-nutrition-in-the-world-2019

World Health Organization. (2024a). *Malnutrition*. https://www.who.int/news-room/fact-sheets/detail/malnutrition

World Health Organization. (2024b). *Obesity and overweight*. https://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight

Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P. H. S., & Zhang, L. (2021). *Rethinking Semantic Segmentation From a Sequence-to-Sequence Perspective With Transformers*. 6881–6890. https://openaccess.thecvf.com/content/CVPR2021/html/Zheng\_Rethinking\_Semantic\_Segmentation\_From\_a\_Sequence-to-Sequence\_Perspective\_With\_Transformers\_CVPR\_2021\_paper.html