Food Portion Estimation via 3D Object Scaling

Gautham Vinod   Jiangpeng He   Zeman Shao   Fengqing Zhu

Purdue University, West Lafayette, Indiana, U.S.A.

{gvinod, he416, shao112, zhu0}@purdue.edu

Abstract

Image-based methods to analyze food images have alleviated the user burden and biases associated with traditional methods. However, accurate portion estimation remains a major challenge due to the loss of 3D information in the 2D representation of foods captured by smartphone cameras or wearable devices. In this paper, we propose a new framework to estimate both food volume and energy from 2D images by leveraging the power of 3D food models and physical reference in the eating scene. Our method estimates the pose of the camera and the food object in the input image and recreates the eating occasion by rendering an image of a 3D model of the food with the estimated poses. We also introduce a new dataset, SimpleFood45, which contains 2D images of 45 food items and associated annotations including food volume, weight, and energy. Our method achieves an average error of 31.10 kCal (17.67%) on this dataset, outperforming existing portion estimation methods.

1. Introduction

Dietary assessment is essential for understanding and promoting healthy eating habits, serving as a key indicator of an individual's health [18, 32]. Traditional methods, such as the 24-hour recall, heavily rely on user-reported data, introducing inherent limitations and biases [27]. In recent years, the rise of image-based dietary methods has garnered attention for alleviating the user burden associated with traditional approaches while demonstrating high accuracy [6, 9, 10, 13, 40].

However, a significant challenge emerges for methods relying solely on 2D images - the loss of crucial information when projecting a 3D food object onto a 2D image plane. This limitation has spurred investigations into multi-view images [3, 16, 39] and depth-based methods [4, 19, 35], aiming to capture richer information than a single image can provide.

The emergence of readily available 3D data mitigates the loss of information associated with 2D images. Datasets such as ShapeNet [2], Omni-Object3D [37] among others, have enriched the space of 3D data. The most recent work introduces NutritionVerse3D [34], which provides 3D representations of food objects, offers a crucial tool for addressing the lack of 3D information in publicly available food image datasets. However, there are currently no methods that use this emerging and readily available 3D data for single image portion estimation.

In this work, we introduce a new framework to estimate food portion size by harnessing the power of 3D models while capitalizing on the simplicity and abundance of 2D food images. The key premise of our method is to recreate and render the eating occasion in the 2D input image using the available 3D food models. In 3D space, the key parameters to recreate the 2D input image are the position and orientation of the camera and the food(s) in the image. Our proposed method estimates these parameters and uses them to render an image of a 3D model corresponding to the food(s) in the image. The known volume of the 3D model and a scaling factor, which is the ratio of the area occupied by the food(s) in the rendered image to the input image, are used to resize the 3D model by the scaling factor to produce the estimated volume of the food(s). The area occupied by the food(s) in the input image is obtained from a segmentation mask which is obtained from a neural network segmentation model. From the estimated volume, the USDA Food and Nutrition Database for Dietary Studies (FNDDS) [25] is used to obtain the energy value.

Unlike existing image-based methods reliant on neural networks [7, 31, 35, 36], our proposed method does not rely on complex neural architectures for portion estimation. Only standard tasks such as food classification and segmentation are performed using neural network models while the portion estimation is based solely on the 3D geometry of the food and the estimated camera and object poses.

Many existing food portion estimation methods are evaluated on private datasets (Table 1), making it difficult for comparison. Further, there are no image datasets that contain ground-truth food volume and corresponding 3D models to evaluate our proposed method. Some existing works use the Nutrition5k dataset [35] which is a publicly avail-

able large-scale food dataset with real food images that contain nutritional information. Around 3,000 images in the dataset have an associated depth map which is a pixel-wise mapping of the distance from the camera to the objects in the image. However, the camera position is fixed and it only captures the top view of the foods. This significantly reduces the generalizability of this dataset since typical images of foods captured from smartphone cameras and wearable devices have various camera poses. Therefore, we introduce a new dataset called *SimpleFood45* that comprises images of real foods captured from different camera poses with a checkerboard as the physical reference. The images are captured using a smartphone camera to simulate a typical eating occasion. Furthermore, each image contains the ground-truth volume (milliliters), weight (grams), and energy (kCal). The dataset aims to provide a better baseline for the evaluation of food portion estimation methods.

Key contributions of our paper include the following. 1) We propose a lightweight, geometric-based framework that enables the utilization of existing 3D data for portion estimation from 2D images. 2) We introduce a new food image dataset, *SimpleFood45*, which contains images captured using a smartphone camera with a physical reference and ground-truth volume for portion estimation. 3) Our method outperforms both neural network and 3D representation-based methods on the SimpleFood45 dataset.

2. Related Works

Food portion estimation or volume analysis can be categorized into four main groups [21].

Stereo Based Approach. These methods rely on multiple frames to reconstruct the 3D structure of the food. The food volume is estimated in [28] from multi-view stereo reconstruction relying on epipolar geometry. Two-view dense reconstruction is performed in [29]. Simultaneous Localization And Mapping (SLAM) is used in [5] for continuous and real-time food volume estimation. The main limitation of these methods is the requirement of multiple images for volume estimation which is not practical for real-world deployment.

Model Based Approach. These methods use predefined shapes and templates to estimate target volume. Model-based approaches are used in [38] where certain templates are assigned to foods from a library and transformations based on some physical references are used to estimate the size and location of the food. A similar template matching approach is used in [12] to estimate food volume from a single image. However, exact matching to these predefined templates cannot accommodate the variations of the foods.

Depth Camera Based Approach. In this approach, a depth camera is used which produces a depth map capturing the distance of the camera to the foods in the image. The depth map is used in [20, 33] to form a voxel representation

Table 1. Overview of datasets used in existing portion estimation methods. An asterisk (*) indicates datasets that are not publicly available.

Dataset Description
Stereo Based Approach
Approx. 400 image sets* [28]
Pair of images of six different fruits* [29]
23 video segments with two types of food (iPhone 6+, wearable camera)* [5]
Model Based Approach
35 images per food item of a total of five food items* [38]
30 images per replica of a total of seven replicas* [12]
Depth Camera Based Approach
3,000 images of 14 types of fruit replicas* [20]
Nutrition5k dataset [35]
Deep Learning Approach
192 dietary study images* [31]
909 images from Nutrition5k [36]
Detailed data from the Nutrition5k dataset [35]

of the image which is then used to estimate the volume of the food. The main limitation is the additional requirement of high-quality depth maps and additional post-processing needed for consumer depth sensors.

Deep Learning Approach. Neural network-based methods have utilized the abundance of image data for training complex networks to estimate food portion. Regression networks are used in [31, 36] to estimate the energy value of the food from a single image input and from an "Energy Distribution Map" which is a pixel-to-pixel mapping between the input image and the distribution of energy of the foods in the image. Regression networks using the input image and depth maps are trained in [35] to produce the energy, mass, and macronutrient information of the food(s) in the input image. Deep learning-based methods rely on large amounts of data to train the model and are generally not explainable. Their performance is often degraded when the input image is dissimilar from the training data.

The datasets used by various portion estimation methods detailed in Table 1 highlights the gap in publicly available datasets for portion or energy estimation of foods. Additionally, non deep learning based methods rely on simple foods with rigid geometries. The Nutrition5k dataset [35] only provides top-view images that contain a physical reference (depth map) but do not contain ground-truth portion information. Therefore, there is a need for a public dataset that has a physical reference, ground-truth portion, and rigid foods for validation of portion estimation methods.

In this paper, we introduce a new perspective to the problem of food portion estimation by using 3D food models but relying solely on 2D food images as an input. We also in-
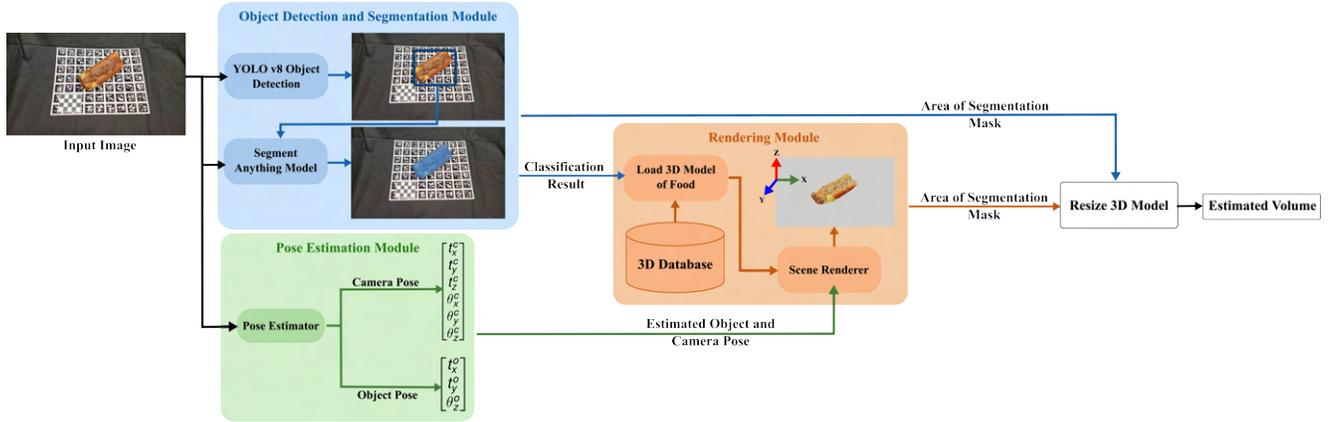
Figure 1. **Overview of proposed method.** The system is divided into 3 modules, the Object Detection and Segmentation Module which uses the 2D image as an input and outputs a segmentation mask. The Pose Estimation Module estimates the camera pose and the orientation and translation of the food. The Rendering Module loads the 3D model based on the class label provided by the Object Detection and Segmentation Module and the pose parameters from the Pose Estimation Module to render an image of the food(s) in the input image. The size of the binary masks of the rendered image and the input image are compared to re-scale the 3D model to obtain the estimated volume.

troduce the SimpleFood45 dataset for the portion estimation benchmark.

3. Method

We propose an end-to-end framework that takes an input 2D image, estimates the camera pose and pose of the food(s) in the image, and renders an image of a 3D model of the food(s) using the estimated poses. The relative difference in the area occupied between the food(s) in the input image and the rendered image is utilized to scale the known volume of the 3D model to obtain the volume estimate. This 3D model is obtained by scanning real food items using a 3D scanner. One 3D model of each food type in the Simple-Food45 is collected and forms our 3D database. Our framework is illustrated in Figure 1.

This proposed method is divided into 3 modules. The *Object Detection and Segmentation Module* is responsible for the classification and segmentation which provides the area occupied by the food(s) in the input image. The *Pose Estimation Module* estimates the location and orientation of both the camera and the food(s) in the 3D world coordinates which are essential for rendering a 3D model. Finally, the *rendering module* uses the 3D model from the 3D database based on the food classification result and the estimated poses for image rendering. The ratio of the area occupied by the food(s) in the rendered image as compared to the area occupied by the food(s) in the input image is used to scale the 3D model and find the estimated volume of the food in the image.

3.1. Object Detection and Segmentation Module

The objective of this module is to obtain a segmentation mask and classification label for the food(s) in the image. The classification label will determine the food type

so that the corresponding 3D model of that food type can be used for image rendering while the segmentation provides information about the area occupied by the food(s) in the image. The Segment Anything Model (SAM) [15] offers zero-shot generalization which eliminates the need to find a suitable dataset with segmentation masks and fine-grained food classes for training. To provide a bounding-box prompt to the SAM for more accurate segmentation, the YOLOv8 [14], which takes an input image and provides a class label with a corresponding bounding box for each food in the image, is used. In this work, we leverage the pre-trained YOLOv8 network on the VIPER-FoodNet (VFN)dataset [11, 23, 24], which contains the most common and frequently consumed food types in the United States of America. Specifically, an input image is passed through the YOLOv8 network to obtain a class label and bounding box which is then used as a prompt to segment the input image using the SAM. Thus, we obtain a segmentation mask associated with each class label.

3.2. Pose Estimation Module

3.2.1 Camera Pose Estimation

The objective of camera pose estimation is to find the orientation of the camera in 3D world coordinates given the intrinsic camera matrix $\mathbf{K}$, which is obtained from camera calibration [41], and the 2D pixel coordinates to 3D world coordinate correspondences of the 12 corner points of the checkerboard.

The mapping from 3D to 2D for the k-th correspondence is described by:

$$\begin{bmatrix} u_k \\ v_k \\ 1 \end{bmatrix} = \mathbf{K}_{3\times3} \left[\mathbf{R} | \vec{\mathbf{t}} \right]_{3\times4} \begin{bmatrix} X_k \\ Y_k \\ Z_k \\ 1 \end{bmatrix} \qquad (1)$$

3743

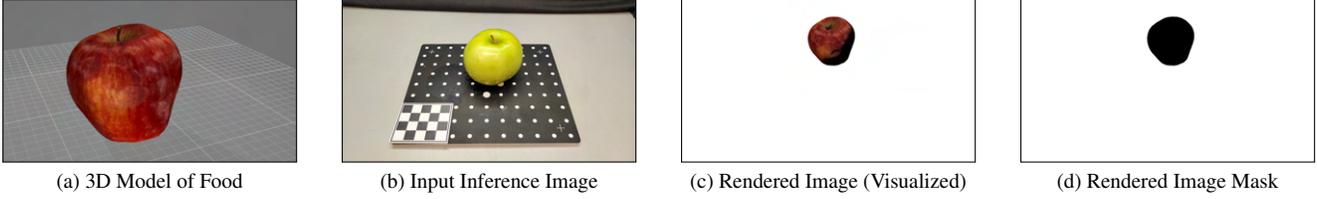| (a) 3D Model of Food | (b) Input Inference Image | (c) Rendered Image (Visualized) | (d) Rendered Image Mask |

Figure 2. The result of rendering an image of an apple based on the estimated pose of the input image using the 3D model of the apple.

where (u_k, v_k) are the pixel coordinates of the k-th corner point and (X_k, Y_k, Z_k) is its corresponding location in 3D world coordinates. The values of (X_k, Y_k, Z_k) are chosen arbitrarily with the constraint that the 12 points of the checkerboard form a uniform grid pattern in 3D space. The rotation and translation of the camera in 3D world coordinates are represented by a rotation matrix $\mathbf{R}_{3\times 3}$ and a translation vector $\vec{\mathbf{t}}_{3\times 1}$.

Equation 1 cannot be solved directly because the correspondences picked out often result in a rank-deficient matrix [8]. Therefore, the Perspective-n-Point algorithm (PnP) [17] is employed which solves for a rotation matrix $\mathbf{R_c}$ and translation vector $\vec{\mathbf{t}_c}$ by making it an optimization problem that minimizes the reprojection loss of Equation 1. However, The extrinsic camera matrix obtained from the PnP, $[\mathbf{R_c}|\vec{\mathbf{t}_c}]$, is a mapping from the world coordinate system to a camera coordinate system where the origin of this camera coordinate system is the camera itself. Given a point $\vec{\mathbf{X}_c}$ in the camera coordinate system and a corresponding point $\vec{\mathbf{X}_w}$ in the 3D world coordinates, their relation is described by:

$$\vec{\mathbf{X}_w} = \left[\mathbf{R_c}|\vec{\mathbf{t}_c}\right]\vec{\mathbf{X}_c} = \mathbf{R_c}\vec{\mathbf{X}_c} + \vec{\mathbf{t}_c} \tag{2}$$

We need to map the camera coordinate system to the 3D world coordinate system to determine the camera's position and angle relative to the origin of the 3D world coordinate system. To obtain this mapping, we determine the origin $\vec{\mathbf{0}}$ of the 3D world coordinate system to the camera coordinate system using Equation 2:

$$\vec{\mathbf{0}} = \mathbf{R_c}\vec{\mathbf{X}_c} + \vec{\mathbf{t}_c}$$
$$\implies \vec{\mathbf{X}_c} = -\mathbf{R_c}^{-1}\vec{\mathbf{t}_c} \tag{3}$$

Here, $\vec{\mathbf{X}_c}$ will be the position of the camera relative to the origin of the 3D world coordinate system, and the rotation relative to the 3D world coordinate system would be the inverse of $\mathbf{R_c}$. Since rotation matrices by definition are orthogonal matrices, it is known that $\mathbf{R_c}^{-1} = \mathbf{R'_c}$. Therefore, we can obtain the camera pose in the 3D world coordinate system through:

$$\mathbf{R} = \mathbf{R_c}^{-1} = \mathbf{R'_c}$$
$$\vec{\mathbf{X}_c} = -\mathbf{R'_c}\vec{\mathbf{t}_c} = \vec{\mathbf{t}} \tag{4}$$

where $\mathbf{R}$ and $\vec{\mathbf{t}}$ are the camera's rotation and translation

in the 3D world coordinates which constitutes the camera pose.

3.2.2 Object Pose Estimation

The position and orientation of the food(s) in the input image in the 3D world coordinate system are estimated using the segmentation mask and checkerboard pattern. To estimate the object orientation, the pixel coordinates of the points in the segmentation mask of the input image are taken as points in 2D space. We perform Principal Component Analysis (PCA) [26] on these points. The eigenvector (principal component) corresponding to the largest eigenvalue gives us the direction of orientation of the major (longer/larger) part of the food. The angle θ_z^o that the largest principal component makes with the horizontal axis is an approximation of the orientation of the object. The 3D world coordinate is defined with the Z-axis pointing upwards so this estimated orientation θ_z^o is the object's rotation about the Z-axis. Our proposed method does not estimate the rotation of the object along the X and Y directions and these values θ_x^o and θ_y^o are assumed to be 0.

Next, we leverage the checkerboard pattern to estimate the object's position. The object needs to be translated only on the XY-plane of the 3D world coordinate system since the food is usually kept on the same surface as the checkerboard (Z translation is 0 relative to the checkerboard). The corners of the checkerboard form a square grid that is 1.2 cm apart, image rectification using DLT [8] is performed. In the rectified image, the scale from pixel to cm is known. The distance between the top-right corner of the checkerboard and the center of the segmentation mask of the food (assumed to be the center of the food) is estimated from this rectified image. This results in the translation along X and Y axes as t_x^o and t_y^o respectively. From the approximated rotation and translation values, the object pose is then $\mathbf{O} = \left[t_x^o, t_y^o, \theta_z^o\right]$.

3.3. Rendering Module

The mapping from the pixel coordinate system to the 3D world coordinate system is obtained from the Pose Estimation Module. The class label obtained from the YOLOv8 network is used to load the corresponding 3D model of the food from the 3D database. In the rendering module, the

3D world coordinate system is mapped back to the image coordinate system to render an image from the 3D model.

The object is translated and oriented in 3D space according to the object pose $\mathbf{O}$. The estimated camera pose (Equation 4) and the camera intrinsic matrix $\mathbf{K}$ together form the camera projection matrix $\mathbf{P}_{3\times4} = \mathbf{K}_{3\times3} \left[\mathbf{R}|\tilde{\mathbf{t}}\right]_{3\times4}$ We obtain the pixel location $\mathbf{x}_i$ in homogeneous coordinates of each 3D point $\mathbf{X}_i$ of the 3D object using this projection matrix

$$\mathbf{x}_i = \mathbf{P}_{3\times4}\mathbf{X}'_i \quad \forall i \tag{5}$$

Finally, a rendered scene using the estimated camera pose of the input image and the estimated object pose is obtained.

Our proposed method requires only a binary mask of the object in the image, therefore the 3D object is rendered without any texture (color). Figure 2 illustrates the input and output of the rendering module using an input image containing an apple (Figure 2b). The 3D model of an apple (Figure 2a) is used to render an image using the estimated camera and object pose from Figure 2b. For visualization, Figure 2c shows the rendered image with its texture, but Figure 2d is the actual rendered binary image mask used for volume estimation. The outputs Figure 2c and Figure 2d may not exactly match the input image Figure 2b because the output camera and object poses may contain pose estimation errors.

3.4. Volume Estimation

The rendered image recreates the input image using the 3D model of the food type in the image. Therefore, we can assume that the ratio of the areas occupied by the food(s) in the rendered image and the input image is proportional to the ratio of the volumes of the objects in these images. The area occupied by the food is estimated by the number of pixels in the segmentation mask of that food.

$$A = \sum_{p\in S} 1 \tag{6}$$

where p denotes each pixel belonging to the segmentation mask S of the input image. Similarly, A' is the area occupied by the object in the rendered image:

$$A' = \sum_{p\in S'} 1 \tag{7}$$

where p denotes each pixel belonging to the segmentation mask S' of the rendered image.

The 3D model is scaled by a factor s:

$$s = \sqrt{A/A'} \tag{8}$$

which is the one-dimensional ratio of the areas occupied by the objects in the input and the rendered images. The estimated volume $\tilde{v}$ of the object in the input image is approximated as the volume of this scaled 3D point cloud. The

energy-to-volume ratio (energy density) is obtained from the FNDDS database [25] by finding the ratio of the energy (kCal) to the volume (mL) of a standard serving size for the corresponding food type. This energy density ρ is multiplied by the estimated volume to obtain the estimated energy of the food(s) in the input image:

$$\tilde{e} = \rho\tilde{v} \tag{9}$$

3.5. SimpleFood45 Dataset Collection

The SimpleFood45[1] contains 12 food types and a total of 513 images of real foods with ground-truth class label, volume(mL), weight(g), and energy(kCal). The images are captured using a Samsung Galaxy S22 Ultra smartphone. Each image contains a checkerboard grid of size 5×4 yielding a total of 4×3 inner corner points. The physical distance between each corner point is 1.2cm. Figure 3 shows some examples of the images in SimpleFood45 dataset.

A Revopoint POP2 [30] 3D Scanner is used to collect the 3D models of the foods. Each of the 12 food types has one 3D scanned model, but 3-4 corresponding food items. Here, the food type refers to a general category of food (Eg. Apple) while the food item refers to an instance within the food type. Each food item has at least 10 images and the camera pose and the pose of the food item are different for each image. The volume occupied by the 3D model in 3D space is taken as the ground-truth volume since the physical dimensions of the 3D model and the actual food are the same. Further, the FNDDS database [25] is used to obtain the energy (kilocalories) for each food item using the ground-truth weight and the food type matching.
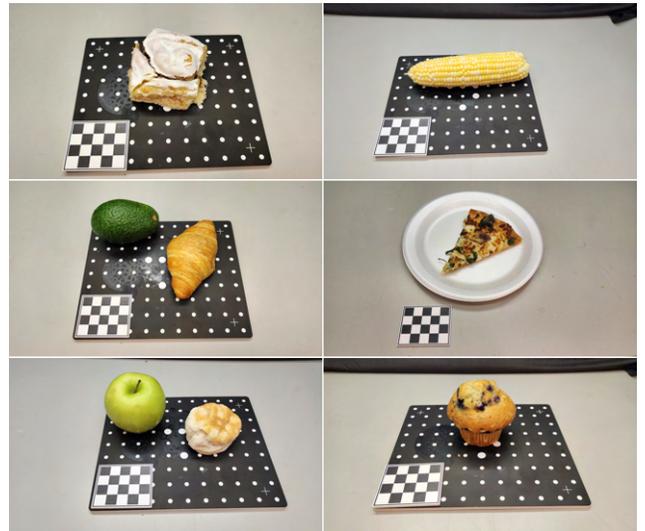

Figure 3. Image samples from the SimpleFood45 dataset. The samples feature different food types and different camera and object poses

.

---

[1]Dataset link - https://lorenz.ecn.purdue.edu/ gvinod/simplefood45/

4. Experimental Results

4.1. Comparison With Other Methods

Our proposed method is validated on the SimpleFood45 dataset. The performance of the proposed method is compared with existing neural network-based methods that have shown high accuracy [22, 31, 35]. Since our proposed method uses 3D representation of food data, we also compare it against a voxel-based method that is based on [33]. The dataset is split into training and testing data using an 80-20 split ensuring equal representation of classes across the training and testing split. All methods are evaluated only on the testing data to ensure fairness of comparison with the neural network based methods.

We use standard metrics including Mean Absolute Error (MAE) and Mean Absolute Percentage Error (MAPE) for comparison, and they are defined as follows:

$$\mathbf{MAE} = \frac{1}{N} \sum_{i=1}^{N} |(\hat{v}_i - v_i)|$$
$$\mathbf{MAPE\,(\%)} = \frac{1}{N} \sum_{i=1}^{N} \frac{|(\hat{v}_i - v_i)|}{v} \tag{10}$$

where v_i is the ground-truth value, $\hat{v}_i$ is the estimated value of the i-th image, and N is the number of images in the dataset. The volume estimation error is denoted by VMAE and VMAPE while the energy estimation error is denoted by EMAE and EMAPE. The units for VMAE and EMAE are mL and kCal, respectively.

The following methods [22, 31, 33, 35] are used for comparison with our proposed method using the SimpleFood45 dataset since these methods have achieved portion estimation results with low error. Results are shown in Table 2.

Baseline. This is a common baseline used in regression tasks and also used in [35]. It is simply a model that predicts the mean ground-truth value of the dataset (volume/energy) for every instance $i = \{1, \ldots, N\}$ of the dataset. For the volume baseline let the ground-truth volume of the i-th image be v_i. The mean ground-truth volume of the dataset is then

$$\bar{v} = \sum_{i=1}^{N} v_i$$

The estimated volume $\tilde{v}_i$ for the i-th image is then given by:

$$\tilde{v}_i = \bar{v}$$

Similarly the energy baseline is the mean energy value of the dataset predicted for each iamge in the dataset.

Deep learning Based Methods. *RGB Only:* a Resnet50 backbone and 2 fully connected layers are trained on the SimpleFood45 to regress the energy value of the input image [31]. *Distribution Map Only:* The ground-truth "Energy Density Map" [31] is used as the input to directly

regress the energy values of the foods. *Distribution Map Summing:* The "Energy Distribution Maps" are summed up directly [22] instead of using a regression network. *2D Direct Prediction:* A regression network that takes the input image and regresses the energy value with the same architecture as [35]. *Depth as 4th Channel:* First, the ZoeDepth Depth Estimation network [1] is used to generate a depth map for each image in the SimpleFood45 dataset. This depth map is appended to the RGB input image as the 4th channel and that is fed to the same backbone as 2D direct prediction [35].

3D Representation Based Methods. *Voxel Representation Estimation Per Food (VREPF):* A voxel representation is created from the input image and the corresponding depth maps [33]. Finally, the number of voxels occupied by a food is translated into actual physical units. Let the number of voxels for the k-th image belonging to the i-th food type be $\tilde{V}_k$ and V_k is its ground-truth volume. A scaling factor

$$\rho_k^i = V_k / \tilde{V}_k$$

is calculated for each image belonging to the i-th food type. The mean scaling factor

$$\bar{\rho}^i = (1/N) \sum_{k=1}^{N} \rho_k^i$$

where N is the number of images in the dataset belonging to the i-th food type, is used to convert the voxel volume to an estimated volume. Therefore, the estimated volume is given by

$$\hat{V}_k = \tilde{V}_k \bar{\rho}^i$$

.

Our proposed method achieves significantly better results than the neural network based methods and outperforms the voxel representation based methods in most cases. Though the neural network models are trained on the SimpleFood45, they fail to accurately estimate the energy values of the food. This could be attributed to the limited data and the large variance in the energy values of the food types, which also speaks to the efficacy of our method.

4.2. Generalization to Other Datasets

The proposed method requires food images with a physical reference and a corresponding 3D model of the food type for volume estimation. Currently, no publicly available dataset contains the required information to be used directly for comparison. Therefore, we select 3 compatible food types from the Nutrition5k dataset [35] which overlaps with food types from the SimpleFood45 dataset - Apple, Bagel, and Pizza. In the Nutrition5k dataset, the camera position is fixed for the top-view images and the known

Table 2. **Comparison with other methods**. The neural network-based methods are made to directly regress the calorific value of the food and hence do not have any metrics for volume output. The 3D Assisted Portion Estimation outperforms all other methods for energy estimation.

Method	VMAE (mL) ↓	VMAPE (%) ↓	EMAE (kCal) ↓	EMAPE (%) ↓
Baseline	83.28	170.37	120.09	547.34
Neural Network Based Methods				
RGB Only [31]	-	-	273.56	222.72
Distribution Map Only [31]	-	-	216.73	159.48
Distribution Map Summing [22]	-	-	192.76	93.16
2D Direct Prediction [35]	-	-	195.28	351.50
Depth as 4th Channel [35]	-	-	173.63	182.29
3D Representation Based Methods				
Voxel Representation Estimation Per Food (VREPF) [33]	**22.35**	24.51	32.01	25.13
3D Assisted Portion Estimation (Ours)	24.51	**14.01**	**31.10**	**17.67**

Table 3. **Results on Other Data**. The table shows results on the Nutrtion5k images using our proposed method.

Food	Method	EMAE (kCal) ↓	EMAPE (%) ↓
Apple	VREPF [33]	9.38	12.36
Apple	Ours	11.5	14.85
Apple	**Ours***	**8.68**	**11.68**
Bagel	VREPF [33]	149.48	62.23
Bagel	**Ours**	**35.74**	**14.79**
Pizza	VREPF [33]	155.34	68.01
Pizza	**Ours**	**97.65**	**35.62**

* indicates that the 3D model source is from the NutritionVerse dataset [34]

distance of the camera to the plate provides a physical reference for our method. The camera parameters are estimated for the camera model in [35].

Table 3 shows the portion estimation results of our proposed method and the Voxel Representation Estimation Per Food (VREPF) on images from the Nutrtion5k dataset. The voxel representation for VREPF is constructed using the depth maps provided in the Nutrtion5k dataset. The only overlap between the food types in NutritionVerse [34] and the SimpleFood45 is Apple. Therefore, for the images of apples, our proposed method is also evaluated on the Nutrition5k images using a 3D model from the NutrtionVerse dataset. In this case, we do not use any of our data for portion estimation. Our proposed method outperforms VREPF in all food types achieving a remarkably low error rate. The VREPF shows poor performance on some food types due to the poor quality of the depth map for these images which leads to errors in the voxel representation. Our proposed method is able to easily adapt to different datasets without any prior knowledge such as training data. While the testing is limited due to the low overlap of food types between the datasets, the lower error rate compared to the VREPF where the actual depth map is used, speaks to the generalizability of our method.

4.3. Ablation Analysis

To understand the importance of Object Pose Estimation on our proposed method, we conducted an ablative study on the

SimpleFood45 dataset. We consider the object pose vector $\mathbf{O} = \begin{bmatrix} t_x^o & t_y^o & \theta_z^o \end{bmatrix}$ and analyze the impact when some or all of these values are set to 0 (the pose is not estimated). Setting $t_x^o = 0$ means the object will only be translated along the Y-axis, while setting $t_y^o = 0$ means the object is only translated along the X-axis. Setting $\theta_z^o = 0$ means the 3D model is not rotated along the Z-axis.

Table 4. **Ablation Analysis.** The impact of the 3D location and orientation of the object in the rendered image is analyzed. The parameter t_y^o plays a vital role in the accuracy of volume estimation.

$\begin{bmatrix} \mathbf{t_x^o} & \mathbf{t_y^o} & \mathbf{\theta_z^o} \end{bmatrix}$	VMAE	VMAPE	EMAE	EMAPE
$\begin{bmatrix} 0 & t_y^o & \theta_z^o \end{bmatrix}$	**24.33**	14.5	**30.24**	20.2
$\begin{bmatrix} t_x^o & 0 & \theta_z^o \end{bmatrix}$	67.37	34.4	84.33	39.66
$\begin{bmatrix} 0 & 0 & \theta_z^o \end{bmatrix}$	75.44	38.14	93.74	43.29
$\begin{bmatrix} t_x^o & t_y^o & 0 \end{bmatrix}$	25.37	14.11	31.24	19.48
$\begin{bmatrix} \mathbf{t_x^o} & \mathbf{t_y^o} & \mathbf{\theta_z^o} \end{bmatrix}$ **(Ours)**	24.51	**14.01**	31.10	**17.67**

From Table 4, we observe that the translation in the Y-axis, *i.e.*, $t_y^o = 0$ plays a major role. This is because translation along the Y-axis moves the 3D object closer to or farther away from the camera and thus influences the object's size on the image. Setting $t_x^o = 0$ means that the object is not translated along the horizontal axis. Our method looks at the area occupied by the food in the rendered image and translating the object along the horizontal axis does not have a significant impact on the area occupied by the food. Finally, $\theta_z^o = 0$ affects only non-symmetrical food items so its impact is restricted to certain food types. Symmetric food items such as apple and bagel can be rotated along the Z-axis but would appear the same in the rendered image due to their symmetry. Therefore, the Object Pose Estimation does play a crucial role in the performance of our method with the lowest EMAPE achieved when all the parameters are taken into consideration.

4.4. Discussion

The major limitation of the proposed method is the requirement of strong prior knowledge (3D model) for each food type. Inconsistencies between the 3D model and the food item will yield poor estimates. For example, the 3D model

for the avocado features a whole avocado, but when the input image contains a slice of an avocado would lead to erroneous results. There needs to be a good similarity between the input object structure and the 3D model structure as is the case for most geometric based models. However, Table 4 shows that even for more general case data, the proposed method is able to fairly estimate the energy content of the foods and even outperforms existing methods. The limitation of 3D model requirement can also be overcome by collecting more 3D models so that the proposed method can be utilized for a lot more food items. The larger the number of models, the more accurate would be the portion estimation.

5. Conclusion

We propose an approach that leverages the simplicity of data availability through 2D images while incorporating 3D data to overcome inherent limitations associated with the missing 3D information. Our method successfully bridges the gap between 2D food images and 3D models for portion size estimation. Notably, our method avoids heavy reliance on neural networks and the availability of training data. We also introduce a SimpleFood45 dataset containing real food images for portion estimation. Experimental results show that our method outperforms both neural network based and 3D representation based methods, yet has the advantage of generalizability without any training data. This work also provides a solid foundation for future research in 3D food analysis.

The major limitation of this work is the need for 3D models specific to each type of food. Future efforts will aim to minimize this dependency by developing methods for 3D reconstruction that use these models primarily as training data, rather than during the inference process.

References

[1] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 6

[2] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 1

[3] Joachim Dehais, Marios Anthimopoulos, Sergey Shevchik, and Stavroula Mougiakakou. Two-view 3d reconstruction for food volume estimation. *IEEE Transactions on Multimedia*, 19(5):1090–1099, 2017. 1

[4] Takumi Ege, Yoshikazu Ando, Ryosuke Tanno, Wataru Shimoda, and Keiji Yanai. Image-based estimation of real food size for accurate food calorie estimation. *Proceedings of the*

[5] Anqi Gao, Frank P.-W. Lo, and Benny Lo. Food volume estimation for quantifying dietary intake with a wearable camera. *Proceedings of the 2018 IEEE 15th International Conference on Wearable and Implantable Body Sensor Networks*, pages 110–113, 2018. 2

[6] Anqi Gao, Frank P-W Lo, and Benny Lo. Food volume estimation for quantifying dietary intake with a wearable camera. *Proceedings of the 15th International Conference on Wearable and Implantable Body Sensor Networks*, pages 110–113, 2018. 1

[7] Yuzhe Han, Qimin Cheng, Wenjin Wu, and Ziyang Huang. Dpf-nutrition: Food nutrition estimation via depth prediction and fusion. *Foods*, 12(23), 2023. 1

[8] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2003. 4

[9] Jiangpeng He, Zeman Shao, Janine Wright, Deborah Kerr, Carol Boushey, and Fengqing Zhu. Multi-task image-based dietary assessment for food recognition and portion size estimation. *2020 IEEE Conference on Multimedia Information Processing and Retrieval*, pages 49–54, 2020. 1

[10] Jiangpeng He, Runyu Mao, Zeman Shao, Janine L Wright, Deborah A Kerr, Carol J Boushey, and Fengqing Zhu. An end-to-end food image analysis system. *Electronic Imaging*, 2021(8):285–1, 2021. 1

[11] Jiangpeng He, Luotao Lin, Heather A Eicher-Miller, and Fengqing Zhu. Long-tailed food classification. *Nutrients*, 15(12):2751, 2023. 3

[12] Wenyan Jia, Yaofeng Yue, John D Fernstrom, Zhengnan Zhang, Yongquan Yang, and Mingui Sun. 3d localization of circular feature in 2d image and application to food volume estimation. *Proceedings of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 4545–4548, 2012. 2

[13] Wenyan Jia, Hsin-Chen Chen, Yaofeng Yue, Zhaoxin Li, John Fernstrom, Yicheng Bai, Chengliu Li, and Mingui Sun. Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. *Public health nutrition*, 17(8):1671–1681, 2014. 1

[14] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, 2023. 3

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 3

[16] Fotis Konstantakopoulos, Eleni I. Georga, and Dimitrios I. Fotiadis. 3d reconstruction and volume estimation of food using stereo vision techniques. *Proceedings of the 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering*, pages 1–4, 2021. 1

[17] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. Ep n p: An accurate o (n) solution to the p n p problem. *International journal of computer vision*, 81:155–166, 2009. 4

[18] Angela D Liese, Susan M Krebs-Smith, Amy F Subar, Stephanie M George, Brook E Harmon, Marian L Neuhouser, Carol J Boushey, TusaRebecca E Schap, and Jill Reedy. The dietary patterns methods project: synthesis of findings across cohorts and relevance to dietary guidance. *The Journal of nutrition*, 145(3):393–402, 2015. 1

[19] Frank P-W Lo, Yingnan Sun, Jianing Qiu, and Benny Lo. Food volume estimation based on deep learning view synthesis from a single depth map. *Nutrients*, 10(12):2005, 2018. 1

[20] Frank P.-W. Lo, Yingnan Sun, and Benny Lo. Depth estimation based on a single close-up image with volumetric annotations in the wild: A pilot study. *Proceedings of the 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics*, pages 513–518, 2019. 2

[21] Frank Po Wen Lo, Yingnan Sun, Jianing Qiu, and Benny Lo. Image-Based Food Classification and Volume Estimation for Dietary Assessment: A Review. *IEEE Journal of Biomedical and Health Informatics*, 24(7):1926–1939, 2020. 2

[22] Jack Ma, Jiangpeng He, and Fengqing Zhu. An improved encoder-decoder framework for food energy estimation. *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, pages 53–59, 2023. 6, 7

[23] Runyu Mao, Jiangpeng He, Luotao Lin, Zeman Shao, Heather A. Eicher-Miller, and Fengqing Zhu. Improving dietary assessment via integrated hierarchy food classification. *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*, pages 1–6, 2021. 3

[24] Runyu Mao, Jiangpeng He, Zeman Shao, Sri Kalyan Yarlagadda, and Fengqing Zhu. Visual aware hierarchy based food recognition. *Proceedings of the International conference on pattern recognition*, pages 571–598, 2021. 3

[25] Janice B. Montville, Jaspreet K.C. Ahuja, Carrie L. Martin, Kaushalya Y. Heendeniya, Grace Omolewa-Tomobi, Lois C. Steinfeldt, Jaswinder Anand, Meghan E. Adler, Randy P. LaComb, and Alanna Moshfegh. Usda food and nutrient database for dietary studies (fndds), 5.0. *Procedia Food Science*, 2:99–112, 2013. 36th National Nutrient Databank Conference. 1, 5

[26] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11):559–572, 1901. 4

[27] Kamila Poslusna, Jiri Ruprich, Jeanne HM de Vries, Marie Jakubikova, and Pieter van't Veer. Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice. *British Journal of Nutrition*, 101(S2):S73–S85, 2009. 1

[28] Manika Puri, Zhiwei Zhu, Qian Yu, Ajay Divakaran, and Harpreet Sawhney. Recognition and volume estimation of food intake using a mobile device. *Proceedings of the 2009 Workshop on Applications of Computer Vision*, pages 1–8, 2009. 2

[29] Md Hafizur Rahman, Qiang Li, Mark Pickering, Michael Frater, Deborah Kerr, Carol Bouchey, and Edward Delp. Food volume estimation in a mobile phone based dietary assessment system. *Proceedings of the 2012 8th International Conference on Signal Image Technology and Internet Based Systems*, pages 988–995, 2012. 2

[30] Revopoint. Pop 2 3d scanner (infrared light — precision 0.05mm). 5

[31] Zeman Shao, Shaobo Fang, Runyu Mao, Jiangpeng He, Janine L. Wright, Deborah A. Kerr, Carol J. Boushey, and Fengqing Zhu. Towards learning food portion from monocular images with cross-domain feature adaptation. *Proceedings of 2021 IEEE 23rd International Workshop on Multimedia Signal Processing*, pages 1–6, 2021. 1, 2, 6, 7

[32] Zeman Shao, Yue Han, Jiangpeng He, Runyu Mao, Janine Wright, Deborah Kerr, Carol Jo Boushey, and Fengqing Zhu. An integrated system for mobile image-based dietary assessment. *Proceedings of the 3rd Workshop on AIxFood*, page 19–23, 2021. 1

[33] Zeman Shao, Gautham Vinod, Jiangpeng He, and Fengqing Zhu. An end-to-end food portion estimation framework based on shape reconstruction from monocular image. *Proceedings of 2023 IEEE International Conference on Multimedia and Expo*, pages 942–947, 2023. 2, 6, 7

[34] Chi-en Amy Tai, Yuhao Chen, Matthew E Keller, Mattie Kerrigan, Saeejith Nair, Xi Pengcheng, and Alexander Wong. Foodverse: A dataset of 3d food models for nutritional intake estimation. *Journal of Computational Vision and Imaging Systems*, 8(1):23–26, 2023. 1, 7

[35] Quin Thames, Arjun Karpur, Wade Norris, Fangting Xia, Liviu Panait, Tobias Weyand, and Jack Sim. Nutrition5k: Towards automatic nutritional understanding of generic food. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8903–8911, 2021. 1, 2, 6, 7

[36] Gautham Vinod, Zeman Shao, and Fengqing Zhu. Image based food energy estimation with depth domain adaptation. *Proceedings of 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval*, pages 262–267, 2022. 1, 2

[37] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023. 1

[38] Chang Xu, Ye He, Nitin Khanna, Carol J. Boushey, and Edward J. Delp. Model-based food volume estimation using 3d pose. *Proceedings of the 2013 IEEE International Conference on Image Processing*, pages 2534–2538, 2013. 2

[39] Chang Xu, Ye He, Nitin Khannan, Albert Parra, Carol Boushey, and Edward Delp. Image-based food volume estimation. *Proceedings of the 5th international workshop on Multimedia for cooking & eating activities*, pages 75–80, 2013. 1

[40] Yifan Yang, Wenyan Jia, Tamara Bucher, Hong Zhang, and Mingui Sun. Image-based food portion size estimation using a smartphone without a fiducial marker. *Public health nutrition*, 22(7):1180–1192, 2019. 1

[41] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000. 3