MFP3D: Monocular Food Portion Estimation Leveraging 3D Point Clouds

Jinge Ma[1], Xiaoyan Zhang[2], Gautham Vinod[1], Siddeshwar Raghavan[1], Jiangpeng He[1], and Fengqing Zhu[1]

[1] Elmore Family School of Electrical and Computer Engineering, Purdue University, West Lafayette, USA
[2] College of Artificial Intelligence, Anhui University, Hefei, China

Abstract. Food portion estimation is crucial for monitoring health and tracking dietary intake. Image-based dietary assessment, which involves analyzing eating occasion images using computer vision techniques, is increasingly replacing traditional methods such as 24-hour recalls. However, accurately estimating the nutritional content from images remains challenging due to the loss of 3D information when projecting to the 2D image plane. Existing portion estimation methods are challenging to deploy in real-world scenarios due to their reliance on specific requirements, such as physical reference objects, high-quality depth information, or multi-view images and videos. In this paper, we introduce MFP3D, a new framework for accurate food portion estimation using only a single monocular image. Specifically, MFP3D consists of three key modules: (1) a 3D Reconstruction Module that generates a 3D point cloud representation of the food from the 2D image, (2) a Feature Extraction Module that extracts and concatenates features from both the 3D point cloud and the 2D RGB image, and (3) a Portion Regression Module that employs a deep regression model to estimate the food's volume and energy content based on the extracted features. Our MFP3D is evaluated on MetaFood3D dataset, demonstrating its significant improvement in accurate portion estimation over existing methods.

Keywords: Food Portion Estimation · 3D Point Cloud · Monocular Image · Multimodality Model.

1 Introduction

The significance of a person's diet on their overall health and well-being is paramount. Chronic diseases such as diabetes are linked to poor dietary habits, therefore understanding one's nutritional intake is of utmost significance [1]. There has been a shift from traditional dietary methods towards image-based dietary assessment due to the ease of usage, fewer measurement or self-reporting errors, and improved accuracy in the estimation of nutritional content from eating occasion images [2,3].

*These authors contributed equally to this work.

However, accurate portion estimation is very challenging even for domain experts such as trained dietitians when they estimate the nutritional content of the food from eating occasion images alone [4]. Directly using monocular image for portion or nutrition estimation is an ill-posed problem due to the loss of 3D information when projecting from the 3D world coordinate to the 2D image plane.

To combat this issue, many existing methods rely on various assumptions such as the availability of a physical reference in the image, such as a checkerboard pattern [5], or the presence of a high-quality depth map with real-world physical units of depth [6]. Methods such as [7,8,9] rely on multiple views, videos, or depth maps, which may be difficult to obtain in real-world applications.

Most existing methods that handle the 3D shape of food typically rely on input images with physical references, and few are able to solely depend on monocular images as input.

In this paper, we propose MFP3D, a new monocular food portion estimation pipeline, reconstructs a point cloud representation of the food, and uses a multimodal approach for 3D and 2D feature adaptation for accurate portion estimation. Our MFP3D consists of three modules: 1) a *3D Reconstruction Module* where the monocular image serves as the input to a depth-estimation network. The estimated depth map is then used to reconstruct a 3D point cloud representation of the food, 2) a *Feature Extraction Module* which comprises a 3D feature extractor network and a 2D feature extractor network, and 3) a *Portion Regression Module* where the extracted features are combined and passed through a deep regression model to estimate the food's volume and energy. Our MFP3D demonstrates significant improvements compared to existing methods on the MetaFood3D dataset The MetaFood3D dataset includes 637 food objects across 108 categories, with diverse modalities and detailed nutritional data. As shown in Figure 1, we have added a sample visualizations of the types of images we used for our experiments.

The main contributions of our paper can be summarized as follows:

- We introduce an end-to-end food portion estimation framework, which uses only a monocular RGB image as input and significantly outperforms existing methods without requiring additional information such as the depth map or physical references.
- We have innovatively utilized 3D point cloud features for food portion estimation.
- We propose to combine the 2D image and corresponding 3D point cloud features in a multimodal approach for accurate portion estimation.

2 Related Works

Food Portion Estimation. Different classes of portion estimation methods use different representations or inputs to recreate the lost 3D information during im-

MetaFood3D - currently under review, the dataset can be accessed at the following link. Please also find the paper of MetaFood3D in the supplementary material.

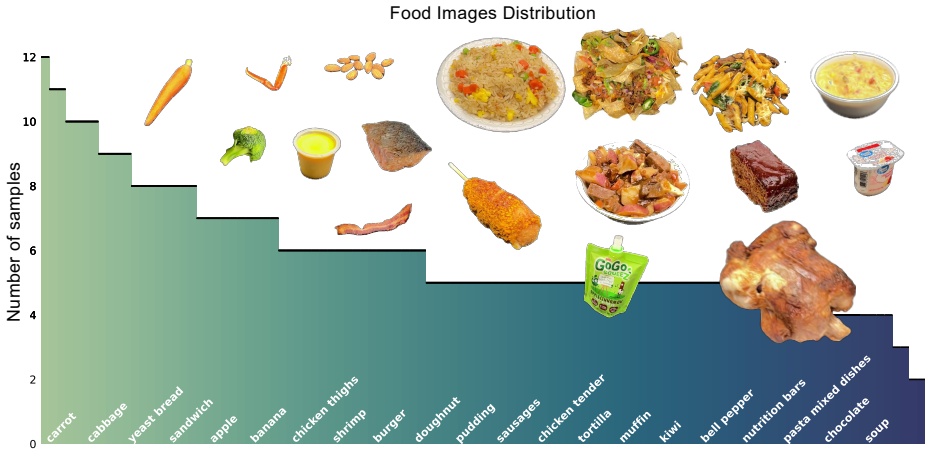

Fig. 1. A sample visualization of the MetaFood3D dataset. The MetaFood3D dataset comprises 637 carefully annotated 3D food objects spanning 108 categories, each accompanied by detailed nutritional information, weight data, and food codes connected to an extensive nutrition database. This dataset highlights intra-class diversity and offers a variety of rich modalities, including textured mesh files, RGB-D videos, and segmentation masks.

age capture. These include multi-view methods [7,8], depth-based methods [10,9], model-based methods [5], and deep-learning based methods [11,6]. The use of 3D food models in [5] shows the efficacy of utilizing such representations of food. The method relies on using predefined 3D models of food and recreating the eating occasion image using the 3D model through object and camera pose estimation. However, the input to this method is constrained by the requirement of a physical reference (checkerboard pattern) in the eating occasion image. Further, food objects that don't fit the geometrical shape of its corresponding 3D model (e.g. whole avocado as compared to sliced avocado) will not achieve reasonable estimates for the food volume. Alternatively, the voxel reconstruction methods require some predefined knowledge of the scene such as in [6] where the distance between the camera and image plane is a known constant. Further, the depth map captured in [6] using a high-quality Intel RealSense RGBD camera makes it easy to capture the distance between the camera and the object in real-world units. However, without this information, there would need to be some scaling between the ground-truth volume and the voxel volume which would require knowledge of ground-truth volume for accurate results [5]. Our method alleviates these concerns by reconstructing the 3D point cloud representation through an estimated depth map while also using its representation for portion estimation.

3D Point Cloud. 3D point clouds can be sampled from real meshes obtained via 3D scanners or reconstructed from 2D images using existing methods such as depth estimation and 3D mesh reconstruction. Zoedepth[12] estimates depth maps for each pixel from monocular images, with depth values representing the coordinates of points in the third dimension. TripoSR[13] is one of the best-

performing models for 3D mesh reconstruction from a single image. It directly reconstructs meshes, which can then be sampled to obtain 3D point clouds.

3D point cloud perception models extract features from a set of three-dimensional coordinates, performing downstream tasks such as classification and segmentation. PointNet[14] was the first model introduced to handle unordered point cloud data. Improving upon its performance, CurveNet[15] introduces continuous sequences of point segments, termed curves, into a ResNet-style network to enhance point cloud geometry learning by effectively aggregating features. Subsequent models introduced many improvements such as using advanced convolution, transformer structures, neighbor clustering, or various pre-training methods. While previous works focus on classification of 3D point clouds, we adapt a 3D point cloud feature extraction model for the regression of food portion.

3 Methodology

Our proposed MFP3D food portion estimation method derives fundamental quantitative attributes of food items such as shape, size, and texture from 3D point clouds and RGB images. The architecture of our three-stage pipeline is illustrated in Figure 2. In **Stage 1**, given an RGB image, $x \in \mathbb{R}^{H \times W \times 3}$, we first separate the each food item from the background using Segment Anything [16] to obtain the mask. Next, we apply the mask to the original image, such that the processed image x_I contains only the food. This processed image is then fed into a point cloud reconstruction model. This model generates a 3D representation x_P of the food object from the single 2D image. In **Stage 2**, the image and its 3D representation are processed by two separate feature extractors: δ^I for the 2D image and δ^P for the 3D point clouds. These extractors produce feature maps f_I and f_P, each with dimensions of $C \times 1$. The feature maps are then concatenated along the second axis to form a comprehensive feature vector $f \in \mathbb{R}^{2C \times 1}$. In **Stage 3**, the concatenated feature vector f is fed into a deep regression module φ, which predicts the food portion $\hat{y}_t$. The attributes of y_t, such as energy content and volume, are defined by the ground truth labels used during training which are provided in the dataset. The pipeline is trained end-to-end in a supervised manner using the $\mathcal{L}1$ loss [17].

3.1 3D Point Cloud Reconstruction

To effectively leverage 3D information, it is essential to acquire accurate 3D representations. In our study, point clouds are chosen as the preferred 3D format due to their lightweight storage requirements and their rich encapsulation of shape and size information. We explore four different types of point clouds to assess their impact on the performance of the portion estimation model.

Ground Truth Point Clouds (GTPCs): GTPCs of food objects provide the most detailed and accurate representation of shape and size, enabling the network to achieve high precision in estimation results. We obtained these real point clouds by using a 3D scanner to capture the food items from multiple

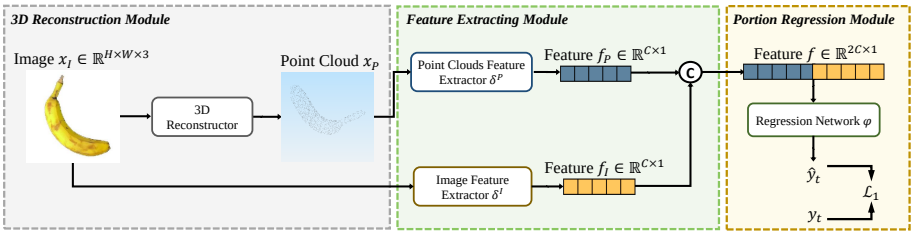

Fig. 2. An overview of the MFP3D framework: The input image x_I goes through a three-stage pipeline for accurate portion estimation. In **Stage 1**, a 3D reconstructor is used to generate the point clouds from the input image. In **Stage 2**, the 3D features (f_P) of the point cloud and the 2D features (f_I) of the input image are extracted using networks δ_P and δ_I, respectively. In **Stage 3**, these features are concatenated and passed through a regression network (φ) to estimate the food portion.

angles. From the original scans, we randomly sampled 1,024 points to derive the GTPCs, seen in Figure 3(a). In contrast, reconstructed point clouds may lose some of this detailed information, leading to less accurate results. Therefore, the performance of models based on GTPCs is considered the upper bound in our experiments.

The true scaling information of 3D point clouds is crucial for accurate portion estimation. However, current 3D reconstruction methods cannot obtain actual size reconstruction from monocular images, thus focusing only on shape. To fairly compare with methods that estimate portions solely from monocular images, as described in Figure 3(b), we **normalize GTPCs** to a range of $[0, 1]$, by rescaling all three dimensions of each point cloud to this range. This removes the true scaling information, allowing us to evaluate performance based on shape alone.

Reconstructed Point Clouds: Acquiring GTPCs requires specialized equipment, making it impractical for many applications. Therefore, we use point clouds reconstructed from monocular RGB images to simulate a more realistic scenario (as shown in Figure 3(c)). Any point cloud reconstruction model that accepts single images as input can be utilized. In our method, we adopt two types of generated point clouds: Depth point clouds and TripoSR point clouds.

For the depth point clouds, we use ZoeDepth [12] to estimate the depth map from a monocular image. Next, we segment the food foreground using masks from MetaFood3D, generated by Segment Anything [16]. To reconstruct the 3D point cloud, we retain the two original dimensions from the 2D image and incorporate the estimated depth as the third dimension. Finally, we randomly sample 1,024 points from the food foreground region to create the depth point cloud reconstruction.

For the TripoSR point clouds, we use the masks from MetaFood3D to generate images that retain only the food foreground. Then, we apply the TripoSR model [13], which can directly reconstruct 3D meshes from monocular images and is widely used for this task. Finally, we randomly sample 1,024 points from the mesh to obtain the TripoSR point cloud reconstruction.

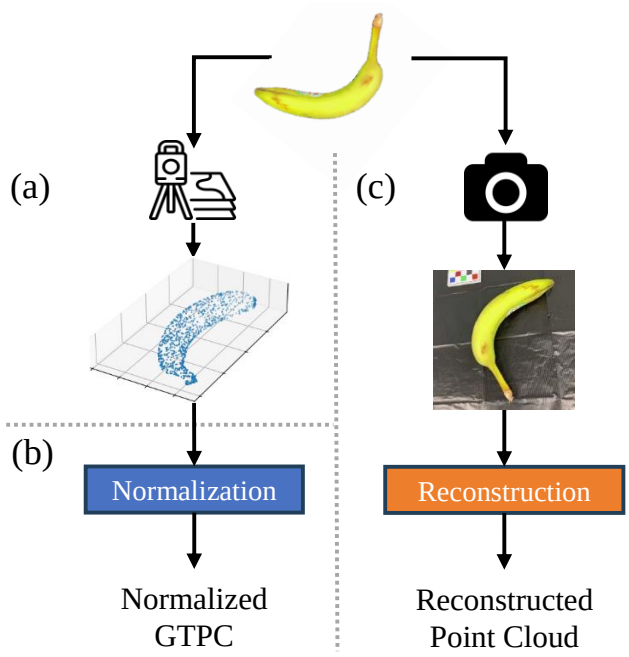

Fig. 3. An overview of (a) Ground Truth Point Clouds (GTPC), (b) Normalized GTPCs and (c) Reconstructed Point Clouds, utilized in our experiments.

3.2 Feature Extraction

The point cloud provides stereoscopic shape and size while the image includes ingredients, edges, and textures. Neither the point cloud modality nor the image modality alone can fully represent the complex information associated with food portion estimation. In this work, we propose to leverage information from both 2D and 3D representations to enhance the understanding of different aspects of the eating occasion image. By concatenating features extracted from the original 2D RGB image and the reconstructed 3D point cloud, our model can capture a more comprehensive view of the food object for better portion estimation.

2D Feature Extraction: We use an image feature extraction model $\delta^I(\cdot)$, built upon ResNet50 [18] pre-trained on the ImageNet [19] dataset. We exclude the last two layers of original ResNet50 but introduce an additional fully connected layer that maps the high-dimensional output to a lower-dimensional feature vector of length 512. This ensures a coherent and efficient feature representation. The overall feature extraction process can be formalized as:

$$f_I^i = \delta^I(x_I^i) \tag{1}$$

where f_P^i represents the 2D feature of the i^{th} sample x_I^i.

3D Feature Extraction: There exists many models designed for extracting features from point clouds. The pioneer network PointNet, known for its simplicity and efficiency, focuses on aggregating global features [14]. On the

other hand, CurveNet's ability to capture local details makes it superior for tasks requiring intricate local feature extraction [15]. Therefore, CurveNet is selected as the backbone of the 3D feature extractor. The architecture of CurveNet consists of a Local Point Feature Aggregation (LPFA) module and a series of CurveNet Inception Convolutions (CIC). Firstly, LPFA aggregates local point features from the input point cloud, which is crucial for capturing fine-grained geometric details. Then CIC layers capture multi-scale features through point cloud down-sampling and feature extraction at various resolutions. After the CIC layers, convolutional and fully connected layers further process the aggregated features and map them to feature vector of the same size as the image features. The 3D feature f_P^i is formulated as:

$$f_P^i = \delta^P(x_P^i) \tag{2}$$

where δ^P is the 3D feature extractor and x_P^i is the reconstruction result of the i^{th} sample x_I^i.

With features f_I^i and f_P^i, we combine them together and form the comprehensive feature f^i. This is achieved by concatenating the two feature vectors, as follows:

$$f^i = f_I^i \oplus f_P^i \tag{3}$$

where $\oplus$ denotes the concatenation of the two vectors along the second axis. In this way, the integrated extractor has the strengths of both modalities by leveraging the geometric details from point clouds and the rich visual features from images.

3.3 Portion Regression

For the portion estimation task, a numerical value is required to represent the final predictive result. To achieve this, we introduce a linear layer, denoted as $\varphi(\cdot)$, which maps the feature f^i to a scalar value. By modifying the ground truth labels in the training data, the model can learn different parameter distributions based on the relationship between inputs and attributes. The model is defined as follows:

$$\hat{y}_t^i = \varphi(f^i) \tag{4}$$

where $\hat{y}_t^i$ represents the estimated value of attribute t for the i^{th} sample.

For the loss function, we use L1 loss to measure the distance between the ground truths and the outputs. The L1 loss is given by:

$$\mathcal{L}_1 = \frac{1}{N'} \sum_{i=1}^{N'} \left| \hat{y}_t^i - y_t^i \right| \tag{5}$$

where y_t^i is the ground truth value for attribute t of the i^{th} sample and N' is the batch size.

4 Experiments

4.1 Experimental setup

Dataset: For our experiments, we utilize the publicly available dataset MetaFood3D. This dataset includes 637 food objects across 108 categories. It is a comprehensive collection featuring 3D object meshes, 2D images, 3D point clouds, segmentation masks, RGBD video captures, nutritional information with weights, and blender renders with camera parameters for all the food items. We randomly select 510 food items for our training set, while 127 food items are reserved for the test set. Since the MetaFood3D dataset is still under review, specially for base experiments, we also train and test our model on SimpleFood45[5] for a more comprehensive evaluation.
Implementation Details: In the base experiments, we take a monocular food image as the input to 3D reconstruction module. It reconstructs a 3D point cloud from the image. The feature extracting module can extract food features solely from the point cloud, or jointly from both point cloud and the image itself. We compared our method with various exiting image-based energy estimation and volume estimation methods.

Our feature extraction network is designed to accommodate relatively flexible input data, such as point clouds reconstructed by different methods (or GTPC), or the option to use images as input. Therefore, in the ablation study, we compared the impact of using different point clouds on the model's performance, and also the effect of incorporating images as the input modality.
Evaluation Metrics: We employ two evaluation metrics to assess the precision of the model's estimation results. The first metric, Mean Absolute Error (MAE) [20], calculates the average of the absolute errors in a set of predictions:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i| \tag{6}$$

where $\hat{y}_i$ is the prediction for the i^{th} input, y_i is the corresponding ground truth, and N is the number of samples in the test batch. The second metric, Mean Absolute Percentage Error (MAPE) [21], expresses errors as a percentage, providing a clear depiction of the prediction error relative to the actual value:

$$\text{MAPE} = \frac{100\%}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i} \tag{7}$$

4.2 Experimental Results

In this subsection, we compare our method MPF3D against existing image-based energy and volume estimation methods. We will also briefly introduce the key idea of each of the previous methods.
Energy Estimation Methods: The *baseline* model always predicts the mean volume and energy values from the dataset. The *RGB only* approach utilize a

ResNet50 backbone and two linear layers to regress the energy estimates from an input image. The *Density Map Only* method employs ground truth "Energy Density Maps" [4] as input to regress the energy estimates. Instead of a regression network, the *Density Map Summing* method sums up the values in the "Energy Density Maps" to estimate the energy. *3D Assisted Portion Estimation* estimates both food volume and energy from 2D images using a physical reference in the eating scene.

Results are shown in Table 1 and Table 2. By comparison, it can be observed that even without relying on the ground truth energy density map or physical reference as additional input or conditions, our method MPF3D still achieves the best results on both datasets, with the lowest MAE of 77.98 kCal and MAPE of 68.05%.

Table 1. Energy Estimation on MetaFood3D

Method	Energy	
	MAE(kCal)↓	MAPE (%)↓
Baseline	221.37	1,287.25
RGB Only [4]	1,932.01	1,124.90
Density Map Only [4]	1100.39	663.43
Density Map Summing [22]	436.12	142.44
3D Assisted Portion Estimation [5]	260.79	102.25
MPF3D (Ours)	**77.98**	**68.05**

Table 2. Energy Estimation on SimpleFood45

Method	Energy	
	MAE(kCal)↓	MAPE (%)↓
Baseline	120.09	547.34
RGB Only [4]	273.56	222.72
Density Map Only [4]	216.73	159.48
Density Map Summing [22]	192.76	93.16
3D Assisted Portion Estimation [5]	32.01	25.13
MPF3D (Ours)	**29.38**	**24.03**

Volume Estimation Methods: For volume estimation, we compare Stereo Reconstruction [7], Voxel Reconstruction [10], baseline method against our MFP3D method as shown in Table 3 and Table 4 . The Voxel Reconstruction method [10]

creates a voxel representation from the input image and corresponding depth maps, translating the number of occupied voxels into physical volume units. A regression network is trained to learn the relationship between voxel volume and ground truth volume, allowing for accurate volume estimation. Conversely, the Stereo Reconstruction method [7] estimates food volume by capturing two images from different angles, using feature matching and triangulation to calculate depth. This depth information is used to reconstruct a 3D model of the food item, which is then analyzed to estimate the volume.

Our method relies **solely on monocular images as the only input**, while other methods depend on additional information, such as binocular images, ground truth depth maps, or physical references. Through comparison, we found that our method can achieve performance close to or even surpassing other methods, despite using less information. On MetaFood3D, our method achieved the lowest MAE of 62.60 ml and MAPE of 41.43%, while on SimpleFood45, our method performed comparably to Voxel Reconstruction and 3D Assisted Portion Estimation.

Table 3. Volume Estimation on MetaFood3D

Method	Volume	
	MAE(ml)↓	MAPE(%)↓
Baseline	151.85	845.69
Stereo Reconstruction [7]	135.96	210.90
Voxel Reconstruction [10]	123.34	104.07
3D Assisted Portion Estimation [5]	195.92	79.33
MPF3D (Ours)	**62.60**	**41.43**

Table 4. Volume Estimation on SimpleFood45

Method	Volume	
	MAE(ml)↓	MAPE(%)↓
Baseline	83.28	170.37
Voxel Reconstruction [10]	**22.35**	24.51
3D Assisted Portion Estimation [5]	24.51	**14.01**
MPF3D (Ours)	25.83	16.15

Our results indicate that the MFP3D method holds significant advantages over existing methods for energy and volume estimation. This is reflected in either lower estimation error or a reduced requirement for input data.

4.3 Ablation Studies

In the ablation studies, we design a series of comparative experiments on Metafood3D to analyze:

1. The impact of using different 3D point clouds as input to the feature extraction module on the model's portion estimation performance.
2. The effect of using RGB images as an additional input modality on the model's performance.
3. The critical information within the point cloud for portion estimation.

The various 3D point clouds used include GTPC (as described in subsection 3.1 and considered to be the upper bound), Normalized GTPC (without true scaling information), TripoSR [13], and Depth Point Clouds [12]. It is worth noting that we used GTPC and Normalized GTPC only as control groups in the ablation studies. We did not use them in the base experiments because they can not be retrieved from monocular images but rather from 3D scanners.

We trained 8 different MFP3D models, as shown in Table 5.

Table 5. Ablation studies on different point clouds and the use of RGB images in **MPF3D**.

Input to Feature Extraction	Energy		Volume	
	MAE(kCal)↓	MAPE (%)↓	MAE(ml)↓	MAPE(%)↓
Point Cloud Only				
Upperbound - GTPCs	114.73	71.00	26.06	19.19
Normalized GTPCs	**135.61**	114.62	**79.93**	68.05
Depth Point Clouds [12]	155.24	**108.53**	80.41	**62.65**
TripoSR Point Clouds [13]	175.45	152.02	121.80	83.47
Point Cloud+RGB Image				
Upperbound - GTPCs	26.16	17.37 (-53.63)	26.68	15.59 (-3.6)
Normalized GTPCs	100.96	**62.65** (-51.97)	**49.26**	42.19 (-25.86)
Depth Point Clouds	**77.98**	68.05 (-40.48)	62.60	41.43 (-21.22)
TripoSR Point Clouds	109.64	98.45 (-53.57)	62.41	**39.45** (-44.02)

The main differences between these MPF3D models lie in: (1) the type of 3D point cloud used, and (2) whether 2D RGB images are also used as input. The top half of Table 5 displays the model performance with portion estimate using only the 3D point cloud as input, while the bottom half shows the model performance when both the 3D point cloud and 2D RGB image are used as

input, as illustrated in Figure 2. In Table 5, excluding the upperbound results from GTPC, the best result for each metric is bolded. In the bottom half of the table, we used small fonts to indicate the changes in MAPE for the models based on point cloud + RGB image compared to those based solely on the same point cloud.

Observations

1. **Different point clouds:** We observed that GTPC achieved upper bound performance in both energy and volume estimation. Depth Point Clouds obtained the lowest Energy MAPE and volume MAPE among the point cloud-only methods. In the point cloud + RGB image methods, Depth Point Clouds achieved the lowest Energy MAE, while TripoSR obtained the lowest MAPE. We can infer that normalized GTPC does not offer a significant advantage over Depth Point Clouds and TripoSR Point Clouds extracted from monocular images.
2. **Multimodality input:** We observed that adding RGB images as supplementary 2D input improved the performance of all models using the same point cloud across the board (as indicated by the small font in the table), though the degree of improvement varied. The percentage decrease in MAPE for volume estimation was less than that for energy estimation. For example, GTPC saw only a 3.6% decrease in volume MAPE after adding RGB images, but a 53.63% decrease in energy MAPE. We believe this may be because the point cloud data includes accurate volume information but lacks the food type, composition, and other energy-related information that might be present in RGB images. This suggests that incorporating multimodal information is crucial for accurate portion estimation.
3. **Important information within the point cloud:** We observed that GTPC performed significantly better than other point clouds reconstructed from monocular images, but normalized GTPC did not show a clear advantage over the above methods. The difference between the two lies in the inclusion of the ground truth scaling factor. Therefore, we can infer that, in addition to the shape of the point cloud, the true scaling factor also contains critical information for portion estimation.

Conclusion

In this paper, we introduce MFP3D for estimating food portions by leveraging the combined power of 3D point clouds and 2D RGB images. This approach enhances the accuracy of volume and energy estimations and simplifies the data acquisition process by utilizing existing 3D point cloud reconstruction methods. These methods reduce dependency on difficult-to-obtain real-world 3D point cloud data and enable the reconstruction of point clouds from monocular images without additional annotations, providing superior performance and demonstrating the practical applicability of our approach. For future work, we plan to improve existing 3D reconstruction algorithms to obtain point clouds that more

accurately represent the actual size of objects and explore additional data modalities such as textual descriptions and videos. Our results demonstrate that our method significantly improves energy and volume estimates, showcasing its great potential for real-world applications deployment.

References

1. A. D. Liese, S. M. Krebs-Smith, A. F. Subar, S. M. George, B. E. Harmon, M. L. Neuhouser, C. J. Boushey, T. E. Schap, and J. Reedy, "The dietary patterns methods project: synthesis of findings across cohorts and relevance to dietary guidance," *The Journal of nutrition*, vol. 145, no. 3, pp. 393–402, 2015.
2. C. Boushey, M. Spoden, F. Zhu, E. Delp, and D. Kerr, "New mobile methods for dietary assessment: review of image-assisted and image-based dietary assessment methods," *Proceedings of the Nutrition Society*, vol. 76, no. 3, pp. 283–294, 2017.
3. K. Poslusna, J. Ruprich, J. H. de Vries, M. Jakubikova, and P. 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*, vol. 101, no. S2, pp. S73–S85, 2009.
4. Z. Shao, S. Fang, R. Mao, J. He, J. L. Wright, D. A. Kerr, C. J. Boushey, and F. Zhu, "Towards learning food portion from monocular images with cross-domain feature adaptation," *Proceedings of 2021 IEEE 23rd International Workshop on Multimedia Signal Processing*, pp. 1–6, 2021.
5. G. Vinod, J. He, Z. Shao, and F. Zhu, "Food portion estimation via 3d object scaling," in *Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3741–3749, 2024.
6. Q. Thames, A. Karpur, W. Norris, F. Xia, L. Panait, T. Weyand, and J. Sim, "Nutrition5k: Towards automatic nutritional understanding of generic food," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8903–8911, 2021.
7. J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3d reconstruction for food volume estimation," *IEEE Transactions on Multimedia*, vol. 19, no. 5, pp. 1090–1099, 2017.
8. F. Konstantakopoulos, E. I. Georga, and D. I. Fotiadis, "3d reconstruction and volume estimation of food using stereo vision techniques," in *Proceedings of the 2021 IEEE 21st International Conference on Bioinformatics and Bioengineering*, pp. 1–4, 2021.
9. F. P.-W. Lo, Y. Sun, J. Qiu, and B. Lo, "Food volume estimation based on deep learning view synthesis from a single depth map," *Nutrients*, vol. 10, no. 12, p. 2005, 2018.
10. Z. Shao, G. Vinod, J. He, and F. Zhu, "An end-to-end food portion estimation framework based on shape reconstruction from monocular image," in *Proceedings of 2023 IEEE International Conference on Multimedia and Expo*, pp. 942–947, 2023.
11. G. Vinod, Z. Shao, and F. Zhu, "Image based food energy estimation with depth domain adaptation," in *Proceedings of 2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval*, pp. 262–267, 2022.
12. S. F. Bhat, R. Birkl, D. Wofk, P. Wonka, and M. Müller, "Zoedepth: Zero-shot transfer by combining relative and metric depth," arXiv preprint arXiv:2302.12288, 2023.

13. D. Tochilkin, D. Pankratz, Z. Liu, Z. Huang, A. Letts, Y. Li, D. Liang, C. Laforte, V. Jampani, and Y.-P. Cao, "Triposr: Fast 3d object reconstruction from a single image," arXiv preprint arXiv:2403.02151, 2024.
14. C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660, 2017.
15. T. Xiang, C. Zhang, Y. Song, J. Yu, and W. Cai, "Walk in the cloud: Learning curves for point clouds shape analysis," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 915–924, 2021.
16. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.
17. J. T. Barron, "A general and adaptive robust loss function," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4331–4339, 2019.
18. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015.
19. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.
20. C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (mae) over the root mean square error (rmse) in assessing average model performance," *Climate research*, vol. 30, no. 1, pp. 79–82, 2005.
21. A. De Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing*, vol. 192, pp. 38–48, 2016.
22. J. Ma, J. He, and F. Zhu, "An improved encoder-decoder framework for food energy estimation," in *Proceedings of the 8th International Workshop on Multimedia Assisted Dietary Management*, pp. 53–59, 2023.
23. Z. Shao, G. Vinod, J. He, and F. Zhu, "An end-to-end food portion estimation framework based on shape reconstruction from monocular image," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 942–947, 2023.