Point2Volume: A Vision-Based Dietary Assessment Approach Using View Synthesis

Frank P.-W. Lo, *Student Member, IEEE*, Yingnan Sun, *Student Member, IEEE*, Jianing Qiu, *Student Member, IEEE*, and Benny P. L. Lo, *Senior Member, IEEE*

***Abstract*—Dietary assessment is an important tool for nutritional epidemiology studies. To assess the dietary intake, the common approach is to carry out 24-h dietary recall (24HR), a structured interview conducted by experienced dietitians. Due to the unconscious biases in such self-reporting methods, many research works have proposed the use of vision-based approaches to provide accurate and objective assessments. In this article, a novel vision-based method based on real-time three-dimensional (3-D) reconstruction and deep learning view synthesis is proposed to enable accurate portion size estimation of food items consumed. A point completion neural network is developed to complete partial point cloud of food items based on a single depth image or video captured from any convenient viewing position. Once 3-D models of food items are reconstructed, the food volume can be estimated through meshing. Compared to previous methods, our method has addressed several major challenges in vision-based dietary assessment, such as view occlusion and scale ambiguity, and it outperforms previous approaches in accurate portion size estimation.**

***Index Terms*—Deep learning, dietary assessment, point cloud completion, three-dimensional (3-D) reconstruction, volume estimation.**

I. Introduction

A RECENT National Health Service (NHS) survey [1] disclosed that the proportion of adults in England who were obese or overweight was 26% and 36%, respectively. Unhealthy eating habits, which include nutritional imbalance and excess calorie intake, are the main factors that lead to obesity [2]. Due to the raising awareness of chronic diseases, increasing population pay more attention to their daily food intake. Previous studies indicated that commonly used dietary assessment techniques, such as 24-h dietary recall (24HR), can effectively help people investigate into their dietary behavior and enable targeted interventions to address the underlying health problems, including Type 1 diabetes (T1D) [3]. In 24HR, participants are required to recall the complete profile of the food items eaten in the last 24 h and estimate the respective portion size by naked eyes. To estimate the portion size, however, it relies heavily on individuals' subjective perception, which could be highly biased and inaccurate. It is for these reasons that various objective vision-based methods, ranging from model-based [4], [5], stereo-based [6], depth camera based [7], and deep learning approaches [8], have been proposed. While these approaches present reasonable accuracy in portion size estimation, there still exist several key challenges such as view occlusion and scale ambiguity. Also, another concern is that current approaches require participants to take images from different viewing angles (in 360°) before eating, which in turn complicates the process and is not able to be embedded on wearable devices for long-term health monitoring. With the technological advancements in depth sensing, various existing mobile devices are already equipped with three-dimensional (3-D) cameras. Inspired by [9], a novel vision-based dietary assessment approach based on deep learning view synthesis and depth sensing technique is proposed in this article. This approach aims to address the key problems, such as view occlusion and scale ambiguity, in volume estimation by combining the merits of artificial intelligence and depth sensing capabilities. In using such approach, the food volume can be estimated precisely with a single depth image or a video captured from any convenient position, which in turn facilitates the implementation of pervasive dietary monitoring on wearable devices. The main contributions of this article can be summarized as follows.

1) Two network architectures UNet and VNet are proposed to complete the partial point clouds due to view occlusion and estimate the actual volume (cm^3) of 3-D models, respectively. Both of these networks take raw point cloud as input.
2) A novel data augmentation method is developed to enlarge the dataset of 3-D models using linear latent interpolation to ease the network convergence.
3) Point cloud preprocessing techniques are developed to facilitate volume estimation.
4) A new 3-D food dataset consisting of 4 k models with actual volume labeled is constructed to train and evaluate the proposed volume estimation approach in the wild.

Manuscript received April 10, 2019; revised August 12, 2019; accepted September 13, 2019. Date of publication September 30, 2019; date of current version January 4, 2020. This work was supported by the Innovative Passive Dietary Monitoring Project through the Bill & Melinda Gates Foundation (Opportunity ID: OPP1171395). The work of F. P.-W. Lo was supported by the Lee Family Scholarship. Paper no. TII-19-1351. *(Corresponding author: Frank P.-W. Lo.)*

F. P.-W. Lo and B. P. L. Lo are with the Department of Surgery and Cancer, Imperial College London, SW7 2AZ London, U.K. (e-mail: po.lo15@imperial.ac.uk; benny.lo@imperial.ac.uk).

Y. Sun and J. Qiu are with the Department of Computing, Imperial College London, SW7 2AZ London, U.K. (e-mail: y.sun16@imperial.ac.uk; jianing.qiu17@imperial.ac.uk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TII.2019.2942831

5) The generalization abilities of the point completion networks in handling food items with previously unseen viewing angles, portion size, and shape geometries are compared with previous network architectures.
6) A new vision-based dietary assessment approach is developed by combining real-time 3-D reconstruction and deep learning view synthesis.

II. Related Work

A. Volume Estimation Approaches

With the advances in computer vision, several vision-based approaches have been developed to address the problem of portion estimation. Specifically, they can mainly be categorized into model-based and stereo-based approaches. Model-based approach estimates the portion size by matching the input of the food items with the prebuilt 3-D food templates. For example, a previous work by [4] developed a virtual reality method by using preconstructed 3-D food models with known portion size to superimpose onto the image. This technique requires users to translate, rotate, and scale the models until the contour of the templates matches with the food items. The accuracy of their proposed method in volume estimation can achieve 79.50% on average (across the food items examined by their research team in the wild). Similar works have also been proposed in [10]. However, model-based method does not possess the generalization ability, which hinders the approach in handling objects with irregular shapes and previously unseen objects. Besides, it always requires a certain level of user inputs such as rotating, shifting, and scaling of the prebuilt 3-D models manually to match with the food images. To address this problem, some research studies proposed to use a structure from motion (SfM) technique to reconstruct the 3-D models. SfM-based approach relies heavily on feature points matching between multiple images, estimates the camera positions, and makes use of the extrinsic camera parameters to reproject the feature points from image-to-camera coordinate. This important property facilitates the volume estimation of food items, which are of irregular shape so that a wider range of food items can be estimated automatically without any manual intervention and a large-scale model library. Another 3-D reconstruction technique, which has been extensively used is simultaneous localization and mapping (SLAM). The difference between SfM-based and SLAM-based 3-D reconstruction techniques is that the SLAM-based approach estimates camera motion and reconstructs 3-D models in real-time. In [6], the authors developed a real-time 3-D reconstruction method to estimate the food portion size. This proposed technique can achieve around 83% accuracy examined with similar food types captured in the wild. Despite the promising estimation results, as mentioned in SfM-based and SLAM-based approaches, there are still several challenging problems unresolved. For instance, they require users to capture multiple images from different viewing angles (normally in 360°) during eating, which could be considered as tedious and impractical. Furthermore, it requires feature points extraction and matching. For those food items with smooth surface or less significant texture, feature points cannot be extracted effectively, which leads to failure in loop closure and 3-D reconstruction. Most importantly, reference objects such as fiducial markers are often required to be placed next to the food items for accurate estimation, which makes the whole dietary assessment process inconvenient.

B. Deep Learning in Volume Estimation

In recent years, several research works tried to use deep learning to assess dietary intake. One of the reasons for using such approach is that the scale of the monocular RGB image can be learned implicitly from global cues of the environment without using any intrinsic and extrinsic parameters, which indicates that reference objects or feature matching between frames can be removed. In [11], convolutional neural network (CNN) has been applied to a single RGB image to infer the depth image and estimate the food volume through 3-D voxelization. With the voxel representation, the portion size of each labeled item can be estimated, respectively. Their model is pretrained based on the NYU v2 RGBD dataset and fine-tuned using a self-collected dataset named as GFood3d (captured by RealsenseF200 depth camera) with different kind of meals from Google cafes. To examine the efficacy of their method, they construct another NFood-3 d dataset using 42 food replicas with known portion size. However, for those artificial or real food items with dissimilar color and texture property, the food segmentation fails and the volume of individual food item cannot be inferred, but estimated as a whole meal with error ranging from 50 to 400 cm^3. Considering the number of food types in their meal used for evaluation, the average error for each food item ranges from 16 to 133 cm^3. Similar idea has been proposed in [12], which aims at predicting bread units (BUs), a representation of food portion, for dietary assessment. In their work, CNN is also applied to infer the depth image using a single RGB image. Afterwards, the authors trained another network, which follows the principal of Resnet-50 proposed in [13] by using both RGB images and ground-truth depth images (captured by Microsoft Kinect v2 sensor) as input. Instead of using a softmax layer, the last layer is replaced by a single neuron with L_2 cost function to predict the BUs. For this article, the performance of their proposed approach is evaluated using BUs so that it is not straightforward to compare it directly with other works. It is for this reason that we investigated into their depth prediction model to analyze the performance. As we know, the accuracy of volume estimation relies heavily on depth prediction. However, the depth prediction model proposed by [12] still achieves root mean square error (RMSE) of 65 cm in depth estimation, on the dataset of NYU Depth v2 and achieve 12.9 cm, on their own food dataset. This error is considered to be reasonably small if it is used in mapping or robotic navigation but for the case of food volume estimation, this error is still unsatisfactory.

C. Deep Learning View Synthesis on 3-D Models

From these previous findings, they showed that volume estimation by using depth prediction is inefficient due to the reason of inadequate information given in a single image to precisely reconstruct the 3-D models and insufficient representative

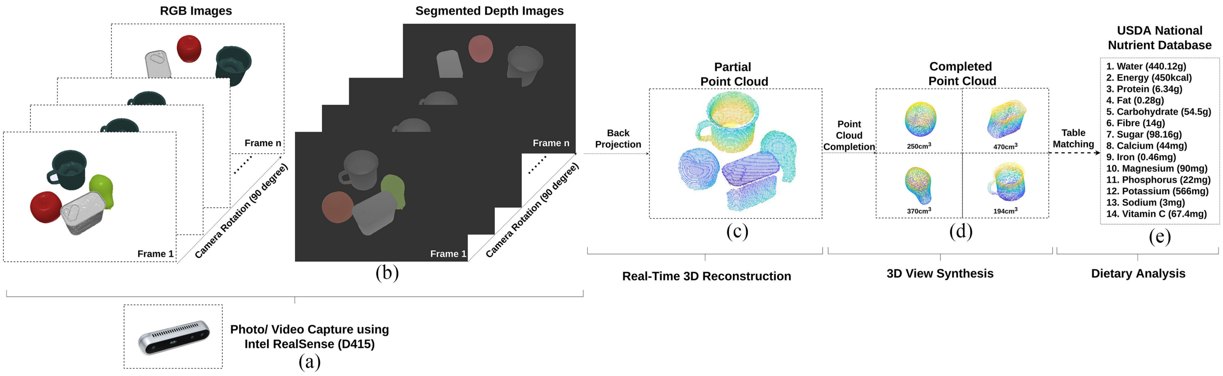


Fig. 1. (a) Photo/Video is captured using Intel RealSense (D415). (b) Food items are segmented out through a fine-tuned Mask R-CNN. (c) The depth image is converted from image coordinate to camera coordinate in order to obtain the partial point cloud for each food item. (d) The partial point cloud is directed to the point completion network to perform 3-D reconstruction. (e) The portion information is linked to nutrient datasets.

training data to train the model; however, volume estimation based on deep learning is still worth investigating due to the reason of practicality and the ease of use. After a comprehensive exploration, we found that an integrated approach based on deep learning and 3-D reconstruction could be one of the potential solutions in aiding dietary assessment. Specifically, deep learning view synthesis [14] can be used along with the SLAM-based approach [6] to estimate the volume of food items without the need for the users to shift and rotate the camera to obtain the complete 360° view of the food. Although a number of research works, such as PointNet [15] and PointNet++ [16], have explored the efficacy of using raw point cloud for classification, there are relatively fewer works using point cloud to perform view synthesis. In [9], the authors proposed an AutoEncoder (AE) architecture to tackle the problem of point cloud completion. Instead of directing the complete point cloud into the AE, partial point cloud is used as the input. Similar idea has been explored by [17]. A point completion network, an encoder–decoder network, was proposed to complete the point cloud with partial input. However, all these works are trained based on ShapeNet in which the models are scaled to a unit cube, losing the information about the portion size of the 3-D models. Also, the partial and complete models are always aligned in canonical coordinates (eight directions), which means the network trained using this dataset is relatively difficult to complete partial point cloud captured in the wild (in any convenient viewing angle).

III. Problem Statement

The unsolved problems can be divided into different parts. First, it is necessary to explore an image capturing technique for dietary assessment without requiring users to take images from inconvenient viewing angles, such as from the back of the food items, and tackle the problem of scale ambiguity. Instead of using common 3-D reconstruction approaches to scan the whole object items, it is preferable to scan the food items only from the front side. To complete the partial point cloud caused by the limited scanning angles, a point completion network is developed to complete the occluded part of the food items. Furthermore, most of the public datasets designed for view synthesis are normalized and centered to ease the image analysis, which hinder portion size estimation. To address this problem, 3-D models of ten commonly seen food categories are constructed, tailored to examine the proposed point completion network in the wild.

IV. Detailed Information and Methods

The pipeline of exploiting 3-D view synthesis in dietary assessment can be divided into different steps. First, a mobile phone with depth sensors or a time-of-flight (ToF) camera, such as Realsense or Kinect, is required to capture a single depth image or a video from any convenient viewing angle, as shown in Fig. 1(a). Second, food items for each frame are segmented out through a fine-tuned Mask R-CNN, as shown in Fig. 1(b). Third, the depth image is converted from image-to-camera coordinate in order to obtain the partial point cloud for each food item, as shown in Fig. 1(c). If a video is captured, a real-time 3-D reconstruction technique, which has been proposed in [6], is used as an alternative choice to reconstruct the partial point cloud with more 3-D information compared to that using a single depth image. Fourth, the partial point cloud is then directed to the point completion network to perform 3-D reconstruction and estimate the portion size of the food items, as shown in Fig. 1(d). Fifth, once the food volume is estimated, the portion information can be linked to nutrient datasets, such as USDA, for detailed dietary analysis [18], as shown as Fig. 1(e). Note that this article mainly focuses on food volume estimation using deep learning view synthesis so that the procedure of dietary analysis will not be covered.

A. Data Augmentation and Mesh Rendering

To examine the efficacy of the proposed volume estimation method using point cloud view synthesis, a large-scale 3-D database is required. Instead of using the benchmark shape repositories such as ShapeNet, which do not involve many food items and are normalized to fit within a unit cube, we used AutoCAD to build a new food dataset, which consists of ten commonly seen food categories including burger, fried rice, pizza, etc. Each category has 20 food models with different shape geometries and portion size. The scale of this dataset, however, is not big enough to train a completion network, which can be applied in the wild. Leveraging the learning representations for 3-D point clouds, a new data augmentation technique is applied to further enlarge the dataset. In [9], their findings showed that

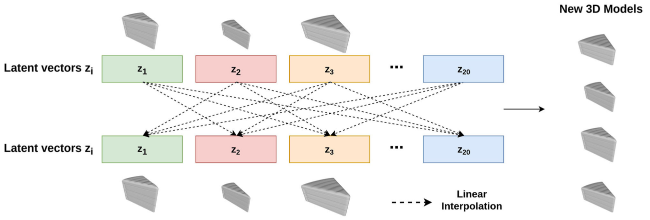


Fig. 2. Linear interpolation on latent space for generating new 3-D models.

TABLE I
RANGE OF EXTRINSIC PARAMETERS (AZIMUTH, ELEVATION, HEIGHT, AND SHIFTING) USED IN MESH RENDERING

Extrinsic Camera Parameters			
Azimuth	Elevation	Height	Shifting
0-360 degree	325-345 degree	0.70-0.90 m	x:-0.1-0.1 m ; y:-0.1-0.1 m

latent vectors, trained by a deep AE, enable shape manipulation easily. Linear interpolation has been used in the latent space among the same category to generate 4 k food models (each category consists of 400 food models) with varying characteristics and portion size, as shown in Fig. 2. The equations for linear interpolation are shown in the following equation:

$$z_i = z_{A_i} + \frac{n}{d}\left(z_{A_i} - z_{B_i}\right) \tag{1}$$

where z_i is an element in a new latent vector, meaning $i = 1, 2, \ldots, 128$, and d represents the number of fraction within the range of the initial vectors. By tuning n, new latent vectors, which represent different shape geometries, can be generated through linear interpolation. The mathematical expression of the new generated latent vector can be written as

$$\mathbf{z} = \begin{bmatrix} z_1 & z_2 & \ldots & z_{128} \end{bmatrix}. \tag{2}$$

After reconstructing new 3-D models using the latent vectors, these models are annotated with their actual volume (cm^3) for portion size estimation, as described in Section IV-B. Furthermore, to evaluate the ability of the point completion network in tackling the view occlusion problems in dietary assessment, another 3-D dataset with occluded food items is constructed through mesh rendering based on the models generated from interpolation. In mesh rendering, the depth images of food items captured from various viewing angles are randomly generated, using the extrinsic camera parameter, as listed in Table I, to simulate photo-taking events in the wild.

After that, we transform the depth images from image-to-camera coordinate based on the intrinsic camera parameters (same as Intel RealSense) to obtain the partial point cloud of food items, as shown in

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = D\mathbf{K}^{-1}\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{K} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

where u, v refer to the coordinates of the depth image and x, y, and z refer to the coordinates of the camera coordinate, D is a scalar number that refers to depthimage(u, v), and $\mathbf{K} \in \mathbb{R}^{3x3}$ refers to the intrinsic matrix. Unlike ShapeNet, the point cloud constructed in this way is not centered and scaled to fit within a unit cube, which enables portion size determination. Before training the point completion network, several point cloud preprocessing techniques are carried out to facilitate the training of the network. First, the point cloud is centered to the origin by subtracting the centroid of the point set, as shown in

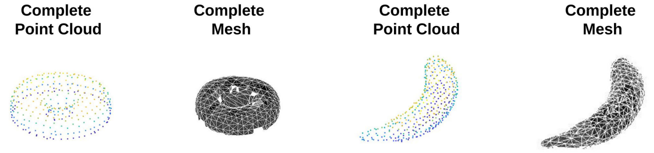


Fig. 3. 3-D meshing of complete point cloud using the proposed method.

$$\text{centroid} = \left(\frac{\sum_i^n x_i}{n}, \frac{\sum_i^n y_i}{n}, \frac{\sum_i^n z_i}{n}\right) \tag{4}$$

where (x_i, y_i, z_i) represents the camera coordinate of data points i and n refers to the total number of points in the point set. This alignment enables the point completion network to tackle food items placed in any position without requiring the food items to be placed in the center of the image. Second, the point cloud is then down-sampled through a voxel grid filter, which takes a spatial average of the data points in every single voxel. Third, a statistical outlier removal filter is applied to remove the outliers from the point set to alleviate the effects of environmental noise. To remove the outliers, we first compute the average point to k nearest neighbors distances ($k$nn) for each point. Then, the point with average distance larger than n standard deviation (S.D.) of the average distance across the point cloud is marked as outlier and removed as shown in

$$f(p_i) = \begin{cases} \text{outlier} & if\ d(p_i) > n(\text{S.D.}) \\ \text{inlier} & if\ d(p_i) \le n(\text{S.D.}) \end{cases} \tag{5}$$

where $d(p_i)$ refers to average distance to k nn for p_i and S.D. represents the standard deviation of the $d(p_i)$ across the point cloud. Note that n, k, and voxel size are determined empirically in this article.

B. Volume Annotation

To annotate the generated models with their corresponding volume, we calculate the bounding polygon of the models with the alpha-shape algorithm [19], [20]. By using alpha-shape algorithm, a sphere with a fixed radius is first defined. Afterwards, the sphere is rotated with its circumference around the models from a chosen starting point until the sphere touches another point lying on the contour. The sphere is transferred to this point and the process continues until reaching loop closure. In Fig. 3, the complete point cloud of a 3-D banana model is converted into a 3-D mesh using the alpha-shape algorithm. Once the 3-D mesh is obtained, the volume of the 3-D models can be deduced easily. Afterwards, all the paired point cloud and its corresponding volume information will be used to construct the training dataset for training the volume estimation network.

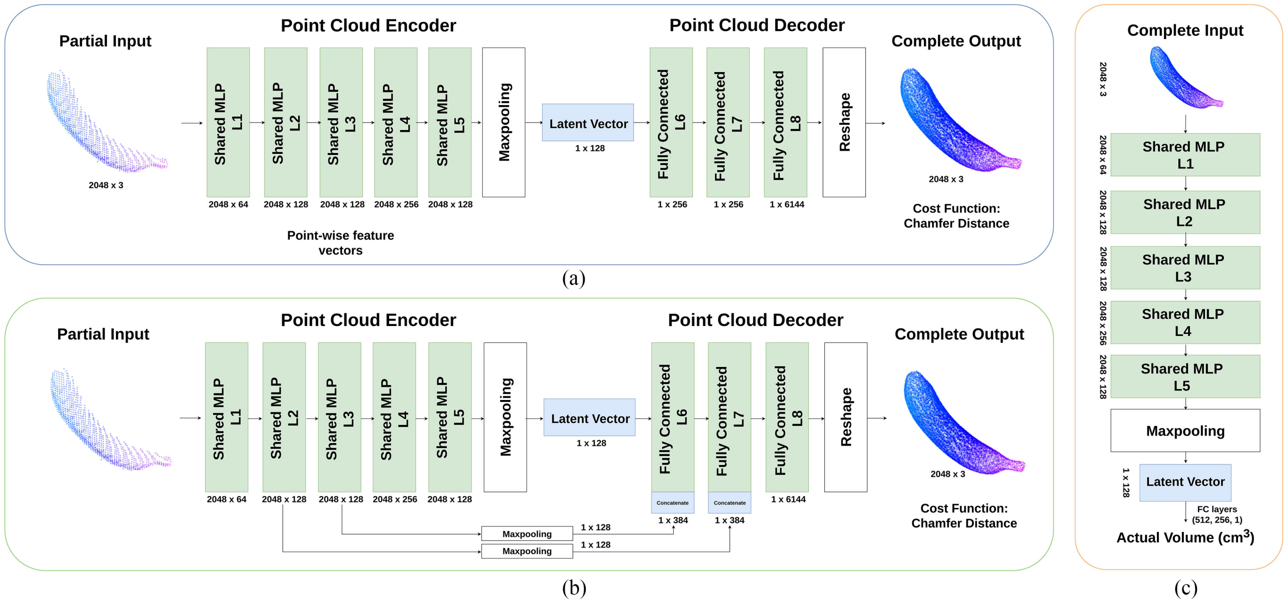


Fig. 4. Network architectures for dietary assessment. (a) Light-weight architecture with FC decoder for view synthesis. (b) UNet architecture for view synthesis. (c) VNet architecture for volume estimation.

One more step is required before using this volume information of 3-D generated models to estimate the actual volume of the original food items. Since the partial point cloud is captured using a depth camera, point-to-point distance should be carefully calibrated, which represents a specific distance in the real world. A real Rubik's cube with known volume (343 cm^3) is used as a scale reference for calibration in this article. A similar calibration method has also been proposed in [6]. When the point cloud is completed using the point completion network to form generated 3-D models, the volume of these generated models can be estimated. By considering the scale difference between the generated 3-D models and the original food items, the volume of the generated 3-D models can be converted to the actual volume of the food items using the calibrated scale/value.

C. Point Completion Network

Due to the problem of view occlusion by taking photos from limited viewing angles, only the 3-D points from one side of the food items can be observed. If using partial point cloud to determine the portion size, the volume of the food items will be largely underestimated. To address this problem, a novel point completion network UNet, shown in Fig. 4(b), is built on top of recent encoder–decoder architectures, tailored to predict the occluded 3-D points using the partial input [9]. In the proposed architecture, the partial point cloud with 2048 points (2048 × 3 matrix) is directed into a feature encoder, which consists of several shared multilayer perception (MLP) layers. Through these MLP layers, each data point p_i is converted into a pointwise feature vector v_i. Since the order of the point set will affect the training of the network, it is necessary to make the point set permutation invariant, which indicates that the order of the point set does not change the geometry they represent. To achieve this, the architecture follows the design of the PointNet [15], which applies a max pooling layer after the MLP to squeeze the feature vectors into a single representation known as a latent vector. For the decoder, several network architectures are compared, including fully connected [9] and UNet architectures, to evaluate the generalization ability in predicting hold-out object items. For FC architecture, it is considered as a light-weight implementation of the point completion network. After passing through the latent vector, it is followed by several FC layers, as shown in Fig. 4(a), to generate the geometrical representation of the complete 3-D models. Regarding to the similarity between partial and complete point clouds, we hypothesize that the pointwise features near the partial input can provide significant guidance in predicting the complete point cloud. In UNet architecture, instead of using pure FC layers, pointwise features are concatenated to these FC layers after passing through max-pooling layers, respectively, as shown in Fig. 4(b). Apart from this, the symmetric version of Chamfer distance (CD) inspired by [14] and [17] is used as the cost function of the point completion networks. In using symmetric CD, penalty will be induced to the cost function if the partial and complete point cloud are not on the same scale. This facilitates scale determination as well as volume estimation, as shown in the following:

$$\text{CD}(G, C) = \frac{1}{|G|} \sum_{g \epsilon G} \min_{c \epsilon C} \|g - c\|_2 + \frac{1}{|C|} \sum_{c \epsilon C} \min_{g \epsilon G} \|c - g\|_2 \tag{6}$$

where G and C refer to the ground truth and complete point cloud, respectively, g and c represent each point in the point cloud.

D. VolumeNet (VNet)

To estimate the portion size, the common approach is to carry out alpha-shape algorithm. While the performance of alpha-shape algorithm shows promising results, there still exist several procedures, which complicate the process such as the radius of the sphere should be determined empirically, as mentioned in Section IV-B, and the estimation error is easily induced when the points are not evenly distributed. Specifically, the real number α refers that the meshed model is constructed by a set of edges and triangles with radii not over $1/\alpha$, which relies heavily on the sampling rate and the geometries of the meshed models. Thus, α is always determined empirically by a user (calibration). To facilitate the dietary assessment and enable automatic quantification, an alternative approach, VNet, as shown in Fig. 4, is also proposed in this article to estimate the portion size without requiring the use of alpha-shape algorithms. To the best of our knowledge, this is the first work on using deep learning to estimate object volume directly by taking raw point cloud as input. The network for volume estimation is similar to the architecture of the point completion network. The feature encoder of VNet follows the design principle of completion network, which also takes raw point cloud as input. By using shared MLP and max-pooling layers, we can also ensure the network to be permutation invariant. However, the complete point cloud is directed into the network instead of the partial one. After the latent vector, three FC layers are followed to infer the actual food volume (cm^3). In this case, a simple L1-norm is used as the cost function, as shown in

$$\text{Cost} = |V_{\text{estimated}} - V_{\text{groundtruth}}| \,. \tag{7}$$

V. Experimental Results and Discussions

A. Performance of Point Completion Network in Handling View Occlusion

To examine the efficacy of the point completion networks, various experiments have been carried out, which aim to evaluate the generalization capability of the point completion networks in handling different scenarios. To construct the training dataset, 2.4 k 3-D models (generated through linear interpolation) from eight food categories, including banana, apple, burger, cake, pizza, orange, rice, and donuts, are used. For each 3-D model, 20 partial inputs from different viewing angles are generated through mesh rendering using the range of extrinsic camera parameters, as shown in Table I. The networks are then trained end-to-end by using 3-D models from these categories with 48 k partial inputs in total. For testing, 3-D food models with unknown geometries, portion size, and viewing angles are directed into the neural network to simulate the real-world photo capturing. Specifically, 800 hold-out models from previously seen categories (16 k partial inputs) and 200 models from another 2 novel categories (4 k partial inputs) are used to evaluate the generalization capability of the networks in tackling food items with previously unseen viewing angles and geometries, respectively. Note that all the networks are trained using Adam optimizer for 750 epochs with the batch size of 100. In Fig. 5, the qualitative results of the point completion networks based on deep learning view synthesis are presented. Partial hold-out food items chosen from eight known categories and two new categories captured from different viewing angles and positions are processed using the FC and UNet architectures, respectively, which provide the evidence that point completion networks are capable of addressing the problem of view occlusion. In previously seen categories, the performance in both of the FC and UNet architectures is comparable; however, we found that the extended UNet has a better generalization ability in predicting new categories (hotdog and muffin). A similar conclusion can also be drawn in the quantitative results, as shown in Table II. Although the training loss measured by CD shows promise in FC architecture, the performance of UNet outperforms FC architecture in testing, which indicates that FC architecture is easier to cause overfitting.

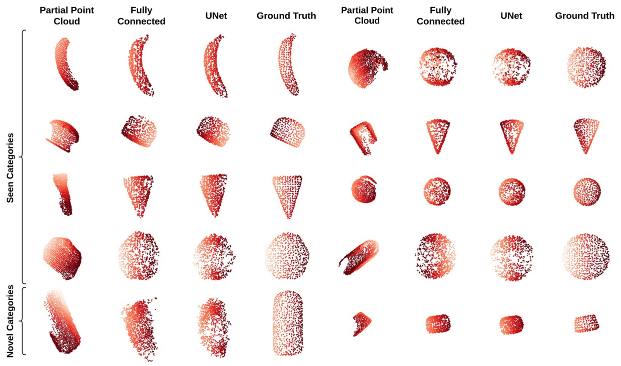


Fig. 5. Qualitative results of point completion networks in handling occluded food items: banana, apple, burger, cake, pizza, orange, rice, donuts, hotdog, and muffin.

B. Performance of Food Volume Estimation Using Deep Learning View Synthesis

The feasibility of using deep learning view synthesis to estimate the actual portion sizes of food items is also evaluated. This experiment is carried out on top of the point cloud completion. Once the partial point clouds are completed, they are converted to 3-D mesh by the alpha-shape algorithm and the food volume is then computed. As shown in Table II, the experimental results of food volume estimation using UNet is promising with average training and testing accuracy up to 95.16% and 92.29%, respectively. The system is also robust with only 5.12% in averaged S.D. for the testing dataset. Furthermore, the accuracy for individual food category is listed out in Table II. It is shown that the categories of cake and muffin have significant improvement in volume estimation by using UNet compared to FC architecture. While the cake belongs to the seen category, the volume accuracy drops sharply when FC architecture is applied. After comprehensive exploration, we found that the main reason for this is due to the large variance between the training and testing datasets for the cake models so that some of them are treated as unseen objects by the FC network. Nevertheless, the UNet architecture is generic enough to tackle the problem without overfitting and predicts the complete models with promising accuracy. Again, these results prove our hypothesis that pointwise feature vectors can provide guidance in completing unseen shape geometries and provide the network with better generalization ability. Most importantly, all these findings conclude that point completion networks can be used to estimate the food volume with unknown

TABLE II
QUANTITATIVE RESULTS OF THE FC AND UNet ARCHITECTURES IN DEEP LEARNING VIEW SYNTHESIS AND VOLUME ESTIMATION

	Fully-Connected (Light-weight)					**UNet (Extended)**				
Food Object Items	Training Loss (CD)	Testing Loss (CD)	Training Accuracy(%) in Volume	Testing Accuracy(%) in Volume	S.D. (%)	Training Loss (CD)	Testing Loss (CD)	Training Accuracy(%) in Volume	Testing Accuracy(%) in Volume	S.D. (%)
Banana	0.15	0.15	-	91.10	16.10	0.19	0.19	-	94.25	5.08
Apple	0.11	0.29	-	93.11	8.64	0.18	0.28	-	95.61	4.10
Burger	0.72	0.76	-	97.91	2.76	0.71	0.85	-	96.49	2.91
Cake	0.15	0.73	-	89.47	11.05	0.21	0.34	-	96.14	2.80
Pizza	0.13	0.26	-	92.21	3.32	0.19	0.27	-	91.18	4.26
Orange	0.09	0.18	-	97.38	2.10	0.18	0.26	-	95.27	3.25
Rice	0.96	1.31	-	94.07	2.36	1.37	2.10	-	94.91	2.85
Donuts	0.11	0.21	-	95.58	5.21	0.18	0.24	-	93.32	4.66
Hotdog	-	2.32	-	79.17	12.67	-	1.75	-	80.35	12.78
Muffin	-	2.46	-	74.55	10.64	-	2.10	-	85.34	8.52
Average	0.30	0.87	97.68	90.45	7.49	0.40	**0.84**	95.16	**92.29**	5.12

*Training loss and training accuracy are calculated based on the training dataset with 8 categories, each category contains 300 models and each model has 20 different viewing angles (48 k partial inputs in total). Testing loss and testing accuracy are calculated based on the testing dataset with 10 categories, each category contains 100 models and each model has 20 different viewing angles (20 k partial inputs in total).

TABLE III
COMPARISON OF THE PERFORMANCE OF VNet TRAINED BY DATASETS WITH AND WITHOUT DATA AUGMENTATION (GENERATED THROUGH LINEAR INTERPOLATION)

	Accuracy and Standard Deviation in Volume Estimation (%)								**Average Accuracy (%)**
	Banana	Apple	Burger	Cake	Pizza	Orange	Rice	Donuts	-
VNet (-Augmentation)	62.46±23.01	60.14±46.81	64.48±24.43	81.91±15.45	49.16±11.62	72.33±22.64	60.19±11.22	78.28±15.01	66.12
VNet (+Augmentation)	90.74±6.93	93.31±28.72	83.59±21.31	87.12±10.11	77.83±9.40	82.59±10.83	72.36±12.73	75.66±16.60	**82.90**

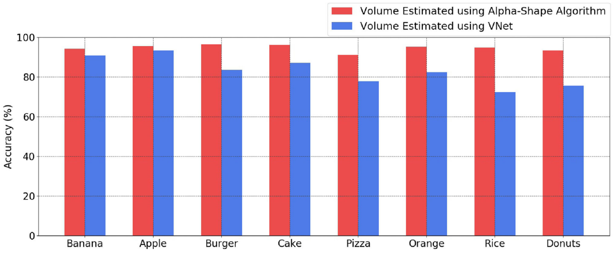


Fig. 6. Comparison of food volume estimation using VNet and alpha-shape algorithm for each category.

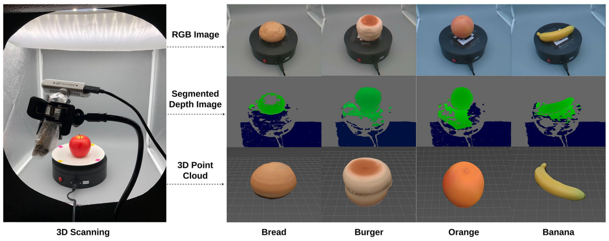


Fig. 7. Experimental platform for obtaining the ground truth volume of food items in the wild.

geometries, portion size, and viewing angles, which in turn makes our proposed method effective in quantifying the portion size consumed by the users.

C. Efficacy of VNet in Volume Estimation

The performance of VNet is evaluated by comparing its results with the results estimated by alpha-shape algorithm. Similar to the previous experiment, hold-out food models are used to examine the trained VNet to ensure the fairness and test its robustness. Thus, the complete point cloud of food items from eight categories with 16 k models completed by UNet are used in this experiment. The average testing accuracy in volume estimation for VNet can achieve up to 82.90%. Considering the light-weight implementation, the VNet got a reasonable drop with error rate in only 12.40% compared to the result estimated by traditional alpha-shape algorithm. Comparison of volume estimation using both of the methods for each category is also shown in Fig. 6. Despite the performance of VNet dropping slightly compared to the alpha-shape approach at the current stage, the VNet is a generic feed forward neural network approach, which is more efficient and easier to use. Furthermore, the performance of data augmentation, which refers to linear interpolation here, is also evaluated in this experiment. Another training dataset is built by eliminating 3-D food models generated by linear interpolation of latent space, and VNet is trained using this newly constructed dataset. The comparison of the performance of VNet trained by datasets with data augmentation and without data augmentation is described in Table III. This illustrates that the proposed data augmentation technique significantly facilitate the training of VNet and help better estimate the volume. Furthermore, it also means that the accuracy of volume estimation relies heavily on the size of the training dataset. From these findings, we hypothesize that the accuracy of VNet can be further improved when more 3-D models are generated using linear interpolation.

D. Point Cloud Completion in the Wild

To further evaluate the robustness of our proposed dietary assessment method, experiments are carried out in the real-world scenarios. Instead of using synthetic dataset, the trained point completion network is evaluated using images of real food items. An experimental platform is set up in a photo studio, as shown in Fig. 7, to obtain the ground truth volume of food items by dense

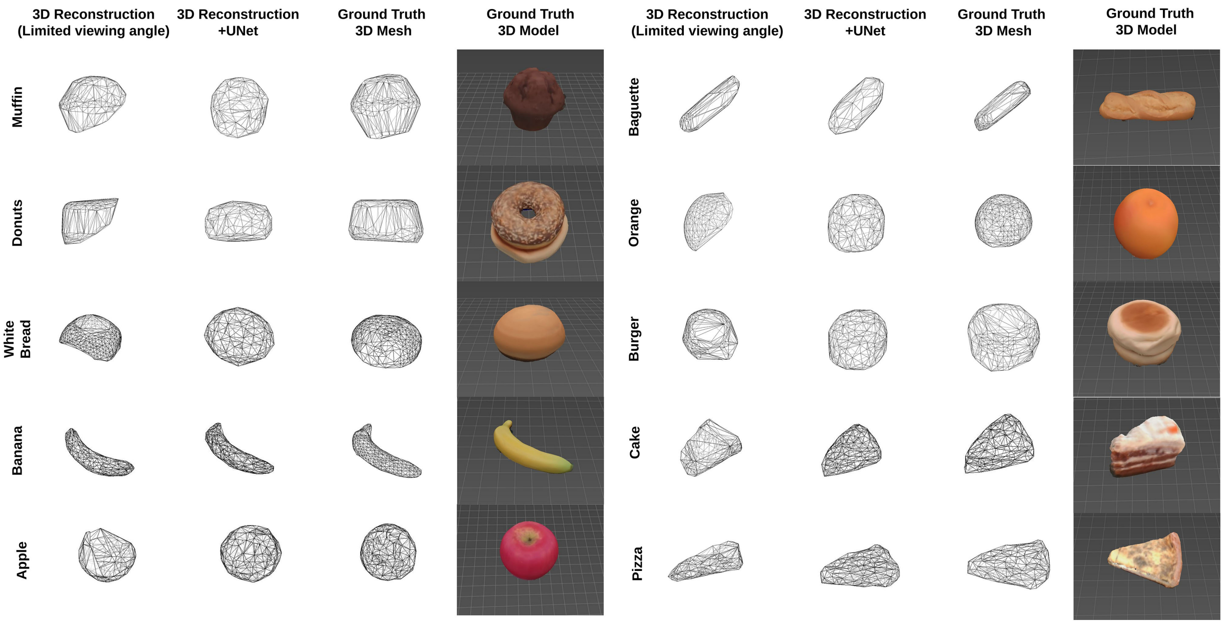


Fig. 8. Examples of using UNet for point cloud completion and 3-D meshing in the wild.

TABLE IV
QUANTITATIVE RESULTS OF FOOD VOLUME ESTIMATION USING 3-D RECONSTRUCTION WITH AND WITHOUT UNET

Food Object Item	Ground Truth (cm^3)	Mean Estimated Volume (cm^3) (3D Reconstruction only)	Accuracy (%)	Mean Estimated Volume (cm^3) (3D Reconstruction +UNet)	Accuracy (%)
Banana	130.34	93.42±7.34	71.67	114.23±12.57	87.64
Apple	248.53	136.76±9.81	55.03	220.21±5.24	88.60
Burger	421.11	391.34±17.30	92.93	436.26±13.92	96.40
Cake	350.57	244.35±13.43	69.70	320.23±14.23	91.34
Pizza	109.24	90.12±8.43	82.50	88.58±10.14	81.10
Orange	298.34	167.73±10.36	56.22	243.95±8.53	81.77
Donuts	351.47	230.53±13.56	65.59	279.10±12.47	79.41
Hotdog	390.86	280.59±21.90	71.78	446.69±17.31	85.72
Muffin	206.17	161.35±17.52	78.26	232.71±11.25	87.13
Baguette	377.31	294.51±22.32	78.10	456.22±15.33	79.08
White Bread	173.81	113.39±23.15	65.24	220.15±13.64	73.33
Average	-	-	71.54	-	**84.68**

3-D reconstruction. The food items are placed on an automatic turning table, which keeps rotating while the depth camera is recording. The 3-D models of real food items are constructed and the ground truth volume can then be obtained using RecFusion, a professional 3-D scanning system, which performs dense 3-D reconstruction and volume estimation. After the ground truth is obtained, the experiment of point cloud completion is carried out. First, videos (six trials for each food item) are captured from convenient viewing angles (only from the front side). A 3-D reconstruction is then applied to reconstruct the partial point cloud as the input of UNet. The qualitative results of using UNet to handle the problem of vision-occlusion are shown in Fig. 8, which present the meshed partial inputs using 3-D reconstruction, meshes of food items completed by UNet, and the ground truth. Further experiments are carried out to examine the performance of the proposed technique in comparison with the method based on 3-D reconstruction only. As listed in Table IV, the experimental results of food volume estimation using both real-time 3-D reconstruction and UNet in the wild are promising with mean accuracy up to 84.68%. In the table, it also indicates that the proposed method can effectively address the problem of view occlusion due to limited viewing angles and ease the implementation of vision-based dietary assessment system.

VI. DISCUSSIONS

A. Comparison With Related Works

The existing research studies on food volume estimation have only examined their algorithms on self-constructed testing datasets with several food items captured in the wild in which there does not exist a benchmark that allows researchers to conduct a fair comparison with previous approaches. Thus, it is reasonable to evaluate our proposed algorithm from the perspective of practicality and implementation. For the model-based approach as presented in [4], [10], and [21], they only examine their algorithms on a few small model libraries consisting of several simple geometric shapes, without considering the case of unseen food shapes. As listed in Tables II and IV, instead, our deep learning view synthesis approach has proven to be effective in tackling unseen food categories, which achieves a level of accuracy comparable to that of seen food categories. With respect to stereo-based approaches [22], [23], they usually have strong requirements on the number of images and their capturing positions. However, our point completion network addresses this problem by completing the partial point cloud of the food items. Compared to [6], our proposed point completion method appears to be more advanced since the previous method can only handle symmetrical food objects. Furthermore, a traditional multiview 3-D reconstruction approach requires a fiducial marker to determine the scale and facilitate the feature matching between frames, which could lead to user compliance issues as the user has to place the markers near the food items and take photos/videos with the markers in the view. Regarding to previous deep learning approaches [11], [12], they rely heavily on depth prediction models. These models can only be trained using paired RGB and depth images captured in the wild, which could be considered as inefficient. Instead, our proposed networks can be trained on synthetic 3-D models, which makes the training process easier. Furthermore, the performance of food volume estimation by means of depth prediction is relatively unsatisfactory as presented in Section II-B. However, in our proposed method, the average error is ranging from 15 to 79 cm^3 as listed in Table IV, which shows promise in the performance. Most importantly, the proposed technique has strong potential in handling view occlusion, which cannot be addressed by existing approaches. As demonstrated in this article, an integrated system based on the depth sensing technique along with deep learning view synthesis should be one future direction in tackling the food volume estimation problem.

VII. Future Works

The proposed technique was initially designed to help dietitians record down the entire portion of food items shown in the meal times. Users are expected to capture the images/videos at the beginning of the meal time for recording the full meal. For the scenarios where the food items are just partially eaten, this article has not yet covered and discussed. In addition, to estimate the exact portion size taken, the algorithm should be able to calculate the remaining food volume after the meal. It is for this reason that a more advanced system could be developed, which consists of newly trained neural networks to recognize the partially eaten food items and estimate remaining food volume in real-time.

VIII. Conclusion

A novel dietary assessment method based on real-time 3-D reconstruction and deep learning view synthesis was presented to estimate food volume in this article. The developed approach showed the feasibility and efficiency in portion size estimation under the circumstances of occluded views. By using the proposed point completion network UNet, the point cloud of the occluded food items can be completed using the prior learned shape and the food volume can be estimated with accuracy up to 92.29%, which not only outperforms other deep-learning-based approach, but also addresses several key challenges in the field of dietary assessment.

References

[1] NHS, "Statistics on obesity, physical activity and diet England 2018," *NHS Digital*, 2018. [Online]. Available: https://files.digital.nhs.uk/publication/0/0/obes-phys-acti-diet-eng-2018-rep.pdf

[2] G. A. Bray and B. M. Popkin, "Dietary fat intake does affect obesity," *Amer. J. Clin. Nutrition*, vol. 68, no. 6, pp. 1157–1173, 1998.

[3] M. M. Anthimopoulos, L. Gianola, L. Scarnato, P. Diem, and S. G. Mougiakakou, "A food recognition system for diabetic patients based on an optimized bag-of-features model," *IEEE J. Biomed. Health Inform.*, vol. 18, no. 4, pp. 1261–1271, Jul. 2014.

[4] M. Sun *et al.*, "An exploratory study on a chest-worn computer for evaluation of diet, physical activity and lifestyle," *J. Healthcare Eng.*, vol. 6, no. 1, pp. 1–22, 2015.

[5] F. Zhu *et al.*, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 4, pp. 756–766, Aug. 2010.

[6] A. Gao, F. P.-W. Lo, and B. Lo, "Food volume estimation for quantifying dietary intake with a wearable camera," in *Proc. IEEE 15th Int. Conf. Wearable Implantable Body Sensor Netw.*, Las Vegas, NV, USA, 2018, pp. 110–113.

[7] S. Fang, F. Zhu, C. Jiang S. Zhang, C. J. Boushey, and E. J. Delp, "A comparison of food portion size estimation using geometric models and depth images," in *Proc. IEEE Int. Conf. Image Process.*, Phoenix, AZ, USA, 2016, pp. 26–30.

[8] F. P. W. Lo *et al.*, "Food volume estimation based on deep learning view synthesis from a single depth map," *Nutrients*, vol. 10, no. 12, 2018, Art. no. E2005.

[9] P. Achlioptas *et al.*, "Learning representations and generative models for 3 d point clouds," 2017, *arXiv:1707.02392*.

[10] C. Xu, Y. He, N. Khanna, C. J. Boushey, and E. J. Delp, "Model-based food volume estimation using 3D pose," in *Proc. IEEE Int. Conf. Image Process.*, Melbourne, VIC, Australia, 2013, pp. 2534–2538.

[11] A. Meyers *et al.*, "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1233–1241.

[12] P. Ferdinand Christ *et al.*, "Diabetes60—Inferring bread units from food images using fully convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Venice, Italy, 2017, pp. 1526–1535.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, 2016, pp. 770–778.

[14] F. P.-W. Lo, Y. Sun, J. Qiu, and B. Lo, "A novel vision-based approach for dietary assessment using deep learning view synthesis," in *Proc. IEEE 16th Int. Conf. Wearable Implantable Body Sensor Netw.*, Chicago, IL, USA, 2019, pp. 1–4.

[15] R. Q. Charles, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[16] C. R. Qi *et al.*, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5099–5108.

[17] W. Yuan, C. Mertz, and M. Hebert, "PCN: Point completion network," in *Proc. Int. Conf. 3D Vis.*, Verona, Italy, 2018, pp. 728–737.

[18] S. Bhagwat, D. B. Haytowitz, and J. M. Holden, "USDA database for the flavonoid content of selected foods, release 3.1," US Dept. Agriculture, Beltsville, MD, USA, 2014.

[19] H. Edelsbrunner, "Smooth surfaces for multi-scale shape representation," in *Foundations of Software Technology and Theoretical Computer Science*. Berlin, Germany: Springer, 1995, pp. 391–412.

[20] N. Akkiraju *et al.*, "Alpha shapes: Definition and software," in *Proc. Int. Comput. Geometry Softw. Workshop*, 1995, vol. 63, pp. 66–70.

[21] N. Khanna, C. J. Boushey, D. Kerr, M. Okos, D. S. Ebert, and E. J. Delp, "An overview of the technology assisted dietary assessment project at purdue university," in *Proc. IEEE Int. Symp. Multimedia*, Taichung, Taiwan, 2010, pp. 290–295.

[22] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," in *Proc. Workshop Appl. Comput. Vis.*, Snowbird, UT, USA, 2009, pp. 1–8.

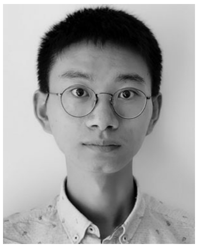
[23] J. Dehais, M. Anthimopoulos, S. Shevchik, and S. Mougiakakou, "Two-view 3D reconstruction for food volume estimation," *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1090–1099, May 2017.

Frank P.-W. Lo received the B.Eng. (First Class Hons.) degree in biomedical engineering and the M.Phil. degree in biomedical engineering from the Department of Electronic Engineering, The Chinese University of Hong Kong, in 2015 and 2017, respectively. He is currently working toward the Ph.D. degree in surgery and cancer with the Department of Surgery and Cancer, Imperial College London, London, U.K.

His current research interests include deep learning, image classification and segmentation, depth estimation, three-dimensional reconstruction, signal processing, blind source separation, and volume estimation.

Yingnan Sun received the B.Eng. (First Class Hons.) degree in electronics with from the University of Liverpool, Liverpool, U.K., in 2015, and the M.Sc. degree in Internet engineering from the Department of Electronic Engineering, University College London, London, U.K., in 2016. He is currently working toward the Ph.D. degree in computing with the Department of Computing, Imperial College London, London.

His current research interests include deep learning, wearable security, Internet of Things, and biometrics.



Jianing Qiu (S'18) received the M.Sc. degree in computing science in 2018, from the Department of Computing, Imperial College London, London, U.K., where he is currently working toward the Ph.D. degree in computing with the Hamlyn Centre.

He was a Visiting Student with Digital Biology Laboratory, University of Missouri, Columbia, MO, USA, in 2016. He is currently a Research Assistant with the Imperial College London. His current research interests include visual recognition and natural language generation.

Mr. Qiu is a member of British Machine Vision Association.

Benny P. L. Lo (SM'16) received the B.A.Sc. degree in electrical engineering from the University of British Columbia, Vancouver, BC, Canada, the M.Sc. (distinction) degree in electronics from King's College London, London, U.K., and the Ph.D. degree in computing from Imperial College London, London, U.K, in 1995, 2000, and 2007 respetively.

He is currently a Senior Lecturer with the Hamlyn Centre and the Department of Surgery and Cancer, Imperial College London. As one of the pioneers in body sensor networks (BSN), he has introduced numerous new platform technologies and novel approaches for wellbeing, and healthcare applications. He has authored or coauthored more than 150 peer-reviewed publications in the field of BSN. His research interests include BSN and wearable technologies for healthcare and wellbeing.

Dr. Lo's work has led to numerous awards, such as the Medical Futures Award. He is an Associate Editor of the IEEE JOURNAL OF BIOMEDICAL HEALTH INFORMATICS, Chair of the IEEE Engineering in Medicine and Biology Society (EMBS) Wearable Biomedical Sensors and Systems Technical Committee, and a member of the IEEE EMBS Standards Committee.