Image-based Food Classification and Volume Estimation for Dietary Assessment: A Review

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

Abstract—A daily dietary assessment method named *24-hour dietary recall* has commonly been used in nutritional epidemiology studies to capture detailed information of the food eaten by the participants to help understand their dietary behaviour. However, in this self-reporting technique, the food types and the portion size reported highly depends on users' subjective judgement which may lead to a biased and inaccurate dietary analysis result. As a result, a variety of visual-based dietary assessment approaches have been proposed recently. While these methods show promises in tackling issues in nutritional epidemiology studies, several challenges and forthcoming opportunities, as detailed in this study, still exist. This study provides an overview of computing algorithms, mathematical models and methodologies used in the field of image-based dietary assessment. It also provides a comprehensive comparison of the state of the art approaches in food recognition and volume/weight estimation in terms of their processing speed, model accuracy, efficiency and constraints. It will be followed by a discussion on deep learning method and its efficacy in dietary assessment. After a comprehensive exploration, we found that integrated dietary assessment systems combining with different approaches could be the potential solution to tackling the challenges in accurate dietary intake assessment.

Index Terms—Dietary assessment, Computer vision, 3D reconstruction, Machine learning, Object recognition, and Mobile technology

I. INTRODUCTION

A Recent National Health Service (NHS) survey [1] in England reported that the proportion of adults who were obese or overweight was 26% and 36% respectively in 2016. Unhealthy food consumption, including nutritional imbalance and excess calorie intake, is one of the reasons which leads to obesity [2]. Commonly used daily dietary assessment methods, such as *24 hour dietary recalls (24HR)*, have proved effective in helping users to understand their dietary behaviour and enable targeted interventions to address the underlying health problems, such as obesity and Type 1 diabetes (T1D) [3]. It is well known that the 24HR is a subjective technique which requires the users to do a face-to-face or telephone interview with dietitians, reporting their food intake with detailed information about the food type and consumed food portion in the previous 24-hour period. Since the procedure of the manual data collection is not carried out directly by experienced dietitians, the consumed portion size can only be estimated by users based on their visual perceptions (e.g., 1 bowl of rice, 1 cup of juice) instead

F. P.-W. Lo, Y. Sun, J. Qiu and B. Lo are with the Hamlyn Centre, Imperial College London, London SW7 2AZ, UK, e-mail:{po.lo15, y.sun16, jianing.qiu17, benny.lo}@imperial.ac.uk

of using weight scales, and thus the portion reported highly depends on their judgement which may lead to a biased and inaccurate dietary analysis result. To address this inaccuracy in dietary assessments, increasing numbers of automatic dietary assessment devices/systems with various sensing modalities, ranging from acoustic sensing approach [4], inertial sensing approach [5] to physiological measurement approach [6], have been studied in the past decade.

The recent advances in computer vision and artificial intelligence have changed every aspect of the way people monitor their health and enabled the introduction of many new applications [7]. A variety of visual-based dietary assessment techniques have been proposed, which can be further divided into image-assisted approach and image-based approach. One of the major difference between these two approaches is that the former is designed to supplement traditional text-based assessment by recalling eating occasions in which manual image analysis will be followed to access the nutrition intake, while the latter allows fully automatic dietary assessment without any human intervention [8]. Despite the great performance in both of these approaches, image-based approach can further reduce workload of dietitians in carrying out dietary assessment. Nevertheless, the implementation of image-based dietary assessment techniques is more complicated since it relies heavily on computing algorithms due to its fully automated characteristics. To facilitate the development and industrialisation of objective dietary assessment technologies, and motivate researchers to improve the accuracy of dietary reporting, an in-depth study on image-based approach is an important step forward.

This study presents an extensive review of algorithms and methodologies used in the field of image-based dietary assessment. After a comprehensive search, a variety of high impact research works have been reviewed. State-of-the-art automatic food recognition and food volume/weight estimation methods are compared with each other to highlight their advantages and obstacles in implementation. In this study, food recognition methods can be divided into two categories which are conventional approaches with manually designed features and end-to-end image recognition with deep learning approaches. Regarding food volume estimation, several approaches have been attempted, ranging from stereo-based approach, model-based approach, perspective transformation approach, depth camera based approach to deep learning approaches. Algorithms and methodologies used in such approaches will be examined and discussed in the paper. Despite there being several comprehensive reviews published on the field of dietary assessment, they
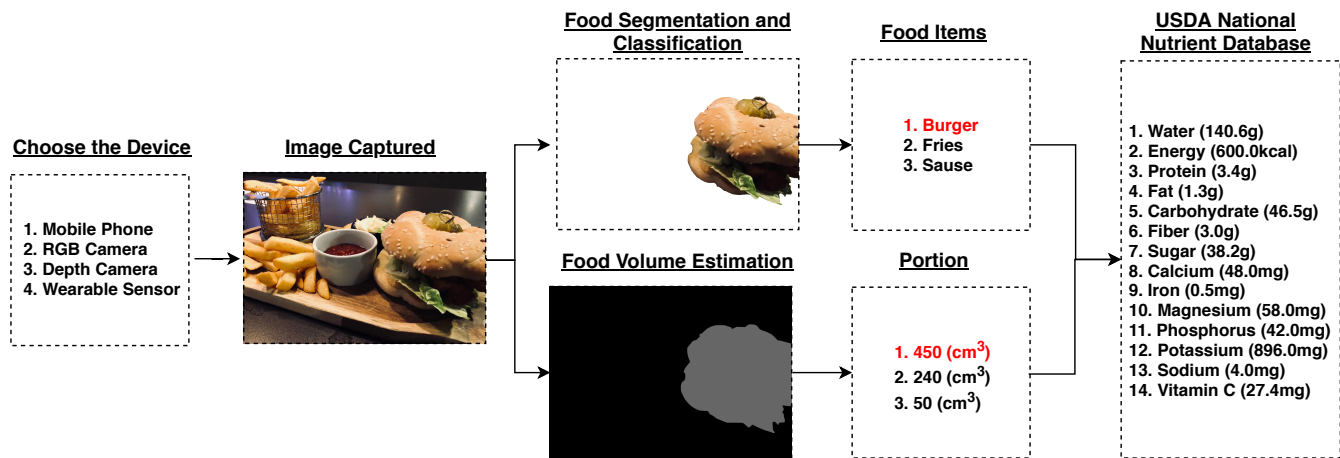
Fig. 1: A system diagram for image-based dietary assessment

have their own particular research focus. Vu et al. published a comprehensive review [9] on sensor-based dietary assessment techniques instead of visual approaches. A more visual-related review paper has been published by [10], but it emphasises mainly on image-assisted approaches. [8] focused on a wide range of methodologies applied on visual approaches, including image-assisted and image-based techniques, for dietary assessment. The procedure, benefits and challenges for visual approaches have been discussed with respect to user comfort, review process and accuracy on nutrition intake. This study, on the other hand, provides an extensive review with the focus on the underlying computing algorithms, mathematical models and methodologies applied in image-based approaches and they are compared and assessed in technical aspects such as processing speed and efficiency, food recognition and volume estimation accuracy and constraints. The main contributions of this study can be summarised as follows: (1) This is, to the best of the authors' knowledge, the first comprehensive review on the state of the art on image-based dietary assessment. (2) A diverse range of methods has been proposed for use in dietary assessment and food quantity control, however, to date no a clear and systematic classification has been done to distinguish them from each other. This study detailed the methods employed and divided them into several major categories to show directions for future research. (3) A critical comparison among different start-of-the-art automatic food recognition and volume estimation approaches is presented in which the advantages and limitations are summarised and concluded. (4) With the advances in artificial intelligence, this study has also explored the feasibility and potential of assessing dietary intake based on deep learning in the future. The rest of the study is organised as follows. Related works on data preparation algorithms used in image-based assessment are reviewed in Section IIA. Section IIB presents the related works on food recognition, which can be divided into two subsections including manually designed feature extraction approaches and deep learning approaches. Food volume estimation using stereo-based, model-based, perspective transformation, depth camera based and deep learning based techniques are detailed in Section IIC. Discussions are provided in Section III.

II. Methodologies and Detailed Information

In this section, the image-based methods proposed to assess dietary intake is detailed as follows: (1) Data preparation methods will firstly be presented to show how they are applied to locate the food items in the images/videos for further reducing memory storage for long time dietary monitoring. (2) Automatic food recognition methods will be explored to show how they are used to assist dietitians in identifying the food items eaten by users. (3) Food volume estimation methods are also shown to explain the underlying theories of using image-based technologies to measure portion size of food items objectively. Furthermore, a system diagram for image-based dietary assessment is presented in Figure 1 .

A. Data preparation

Image-based assessment uses captured images as the main source of input for the analysis. Active and passive methods can be selected to capture images according to the requirements of the applications. Active method normally requires users to take images before and after meals in a deliberate and intentional way, while passive method refers to the techniques which are able to handle image capturing without much human intervention. Despite the convenience of the latter, the implementing strategies are much more complicated. Video surveillance is an important area in machine vision which has long been used for passive monitoring. To monitor food intake, a wearable sensor or mobile device which have a single camera or multiple cameras embedded are required for video capturing. Nevertheless, data storage for continuous long-term video surveillance is a major limitation and problem. The reason is that passive method will continuously generate image frames, which requires a huge amount of memory space of the devices even though the sampling rates or resolution can be reduced. Thus, several algorithms have been proposed to pre-process the images before storage in order to minimise the memory required and prolong the recording time. A previous work by [11] proposed the use of real time image filtering technique which removes redundant images and finds a representative frame based on the similarities of neighbouring

TABLE I: Food recognition methods: Conventional image recognition approach with manually designed features and end-to-end image recognition with deep learning approach

Food Recognition		
Method	Image recognition with manually designed features extraction	End-to-end image recognition with deep learning approach
Procedure	1. Feature points extraction (e.g., SIFT, colour histogram) 2. Coding technique (e.g., BoF, FV) 3. Classification method (e.g., SVM, RF)	1. Deep Convolutional Neural Network (DCNN) applied (e.g., AlexNet, ResNet, GoogLeNet)

*SIFT: Scale-invariant feature transform, BOF: Bag-of-Features, FV: Fisher vector, SVM: Support vector machine, RF: Random forest

frames with respect to the colour, texture and edge profile. Apart from using the representative frame, eliminating frames without food container is one alternative techniques to further reduce memory storage. They presented an algorithm for plate detection which finds circular dining plates based on Canny detector [11], [12]. Edges are firstly converted into curves and the redundant curves (arc) are removed by arc filtering. Convex hull algorithm has been further applied on the remaining arcs to group them. Circular regions can then be detected easily and which can be used to locate food containers/dining plates. Similar idea has been used by [13] to detect the precise location of the dinning plates. Apart from plate detection, food segmentation also plays an important role in data preparation. Increasing works show that segmentation can improve the performance of food recognition. In [11], [14], [15], several algorithms have been proposed to segment images by detecting the boundary of objects with the use of active contours, i.e. edge-based segmentation (Snake model) and region-based segmentation (Chan-Vese model). In [16], the authors extend the idea and develop a modified Chan-Vese algorithm to partition the image. In [13], a novel segmentation technique has been proposed which converts the image into CIELAB colour space. Mean shift filtering is then applied to smooth the fine-grain texture and preserve the dominant colour. This is followed by a region growing algorithm to merge the pixels of similar colours into segments.

B. Food Recognition

With the widespread use of smart phones, many mobile health applications have been launched, e.g., MyDietCoach, Yazio, MyFitnessPal, Foodnotes, MyFoodDiary and FatSecret. Such mobile applications, however, require users to manually enter the food types and consumed weight which are tedious and burdensome [17]. To address the problem, automatic food recognition has been investigated by researchers. Food recognition is a crucial part in the dietary assessment process. Only after recognising the type of food can we further compute the calorie intake and analyse nutritional information. In food recognition, the characteristics of food can be greatly attributed to their surface colours, shapes and texture. Therefore, if a system tries to identify a particular food object, feature descriptors containing those underlying information should be extracted first. In the following section, a summary of previous studies on food recognition is shown. The food recognition methods can mainly be divided into two categories which are conventional image recognition approach with manually designed features and end-to-end image recognition with deep learning approach as shown in Table I.

1. Conventional Image Recognition Approach with Manually Designed Features: The framework for conventional food classification can be divided into two major tasks: Feature extraction and classification. Feature extraction refers to computing a descriptor/ feature vector which can best reveal the underlying visual information. There are several commonly used feature extraction techniques which can extract informative visual data, such as Scale Invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Gabor filter, MR8 filter and Local Binary Patterns (LBP). To enhance the recognition rate of food classification, more sophisticated feature descriptors are developed by fusing different feature vectors. For instance, [18] proposed a technique called multi-view food recognition, i.e. pictures are taken from different viewing angles, which addresses the problem of occlusions and restricted view of using a single image. It starts from generating descriptors based on Difference of Gaussian (DoG) and Scale Invariant Feature Transform (SIFT). With the use of such techniques, the results are invariant to lighting, scaling and affine transformation. Nearest neighbour classifier are then used for food classification. [19] has further extended their idea to improve the efficiency of the SIFT descriptors. They stated that it is computationally expensive to determine the similarity between images using several hundred SIFT features. Thus, they proposed to cluster SIFT features into visual words by using hierarchical k-means clustering algorithms. Visual word, as shown in Figure 2, refers to the most representative descriptor over a particular set of descriptors in the same cluster. Those visual words can be treated as basis to build a visual library. A comprehensive research on Bag-of-Features (BoF) method was conducted by [3]. The study shows that the recognition accuracy can be further improved with more visual words, however, saturation will be reached in using a reasonable large number of features (6000 shown in this study). It also raises an issue that SIFT-based feature extraction technique may fail to produce a sufficient number of feature points and suggests to use random and dense sampling methods, which improves the performance of feature extraction from 69% to 77% and 78% respectively. In addition, their results showed that hsvSIFT and colour moment invariants achieved the best accuracy among other SIFT-based and colour-based techniques. Different linear and non-linear techniques, including linear SVM, non-linear kernel-based SVMs, ANNs and the Random Forests (RF), have been evaluated. Among those techniques, linear SVM outperforms the others marginally by i.e. 2 to 8%. Similar idea has also been proposed by [20], which uses BoF, Segmentation-based Fractal Texture Analysis (SFTA) and color histogram as feature descriptors. Apart from using visual words, other coding techniques are
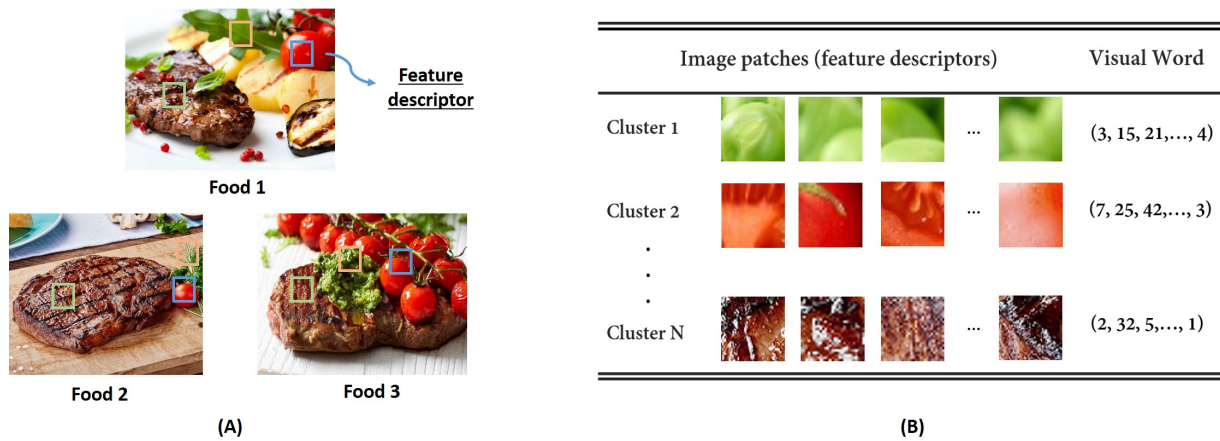
Fig. 2: (A) Feature point extraction (B) Visual word construction

also implemented to improve the efficiency. For example, [21] proposed a real-time food recognition system using RootHoG as features. Without inputting the features directly into classifiers, Fisher Vector (FV) is applied to encode the features, thus the speed of image recognition can be increased. In [22], five types of features are extracted, including color, HOG, SIFT, LBP and MR8 filter. These features are encoded through Locality-constrained Linear Coding (LLC), using a dictionary learned via k-means clustering. Furthermore, a comprehensive research [23] has been done to investigate into the efficiency of using multikernel-based SVM. The authors compare their proposed technique to other methods including SIFT-based nearest neighbour classifier and texture-based SVM and found that their technique outperforms the previous techniques by around 20%. Since the feature descriptors used by their study are different from the previous one, it is really difficult to tell whether linear SVM is better than non-linear kernel based SVM or vice versa in food recognition. An overall comparison among traditional approaches is presented in Table II for further information.

TABLE II: Traditional approaches on food classification

Authors	Methods	Types	Top 1
Kong et al. [18], 2011	Nearest-neighbour classifier; DoG and SIFT	61-food classes	84.0%
Kong et al. [19], 2012	Nearest-neighbour classifier; SIFT and visual words	5-food classes	92.0%
Kawano et al. [24], 2014	one-vs-rest classifier; Fisher vector with RootHoG	256-food classes	50.1%
He et al. [20], 2014	KNN classifier; DCD, MD-SIFT, SCD and SIFT	42-food classes	64.5%
Tamma. et al. [25], 2014	SVM classifier; BoF, SFTA and color histogram	5-food classes	70.0%
Anthim. et al. [3], 2014	SVM classifier; BoF, hsvSIFT and color moment invariant	11-food classes	78.0%
Pouladz. et al. [26], 2015	SVM classifier; color, texture, size, shape features	30-food classes	94.5%
Beijbom et al. [22], 2015	SVM classifier; color, HOG, SIFT, LBP, MR8 filter and LLC	50-food classes	77.4%

*Top 1 accuracy can be computed using (TP + TN)/(TP + TN + FP + FN)

2. End-to-end Image Recognition with Deep Learning Approach: Deep learning has gained much attention due to its outstanding performance in different artificial intelligence applications; however, deep learning approach for food recognition has only been considered in very few works. As there are a huge variety of food types and the food pattern varies significantly under different viewing angles and lighting conditions, standard manually designed features extraction techniques often cannot sufficiently abstract or represents the characteristics of the objects. In 2014, [27] published a paper of food recognition based on deep learning. They applied a convolutional neural network (CNN) to the tasks of dietary monitoring. In their experiments, they compared the efficiency and practicality between CNNs and traditional SVM-based techniques using handcrafted features. The traditional method has the accuracy of around 50-60%, whereas the CNN outperforms them by 10%. Similar idea was proposed in [28] in which image patches are extracted on a grid for each food item, and the patches are then fed to a deep CNN. Google [29] has also proposed a series of deep learning methods which can be used in dietary assessment. They proposed using GoogLeNet pre-trained on ImageNet dataset to perform food recognition. The model has been fine tuned on the publicly available Food101 dataset with 101 food items and examined on it with 79% accuracy. The results outperform the method based on handcrafted features and SVM classifier by 28% which were tested on the same database. Besides, Google also made a contribution on transfer learning which replaced the final 101-way softmax layer from their trained model and plugged in another 41 logistic nodes from MenuMatch. The result is also statistically significant compared to the traditional technique examined on the same database [22]. In a recent study, [17] extended the underlying idea of GoogLeNet and proposed a deep neural network with inception module to carry out food recognition. They have carried out a comprehensive study on the practicality of deep learning technique and examined their network on several large publicly available datasets including UEC-FOOD100, UEC-FOOD256 and Food101. The overall accuracy for Food101 is 77% which gives similar result proposed by Google. More research works based on deep-learning are presented in Table III for further information.

Whereas traditional approaches based on manually extracted features and deep learning approaches were examined using different test datasets, the recognition rate for deep learning methods still outperform traditional ones, even in the case when the datasets have more categories. This further confirmed the efficacy and practicality of using deep learning methods in food recognition.

TABLE III: Deep learning approaches on food classification

Authors	Methods	Dataset	Top 1
Kagaya et al. [27], 2014	CNN	Own dataset	73.7%
Christ. et al. [28], 2015	Patch-wise model+CNN	Own dataset	84.9%
Meyers et al. [29], 2015	CNN+inception	Food-101	79.0%
Liu et al. [30], 2016	CNN+inception	Food-101	77.4%
Pandey et al. [31], 2017	Ensemble Net	Food-101	72.1%
Liu et al. [17], 2018	CNN+inception	Food-101	77.0%
Yu et al. [32], 2018	DLA	Food-101	90.0%
Foresti et al. [33], 2018	WISeR	Food-101	90.3%
Cui et al. [34], 2018	DSTL	Food-101	90.4%
Qiu et al. [35], 2019	PAR-Net	Food-101	90.4%
Min et al. [36], 2019	IG-CMAN	Food-101	90.4%
Tan et al. [37], 2019	EfficientNet	Food-101	93.0%
Kawano et al. [21], 2014	Pre-trained CNN	UEC-FOOD100	72.3%
Yanai et al. [38], 2015	CNN	UEC-FOOD100	78.8%
Hassane. et al. [39], 2016	Inception V3	UEC-FOOD100	81.5%
Liu et al. [30], 2016	CNN+inception	UEC-FOOD100	76.3%
Foresti et al. [33], 2018	WISeR	UEC-FOOD100	89.6%
Yanai et al. [38], 2015	CNN	UEC-FOOD256	67.6%
Hassane. et al. [39], 2016	Inception V3	UEC-FOOD256	76.2%
Liu et al. [30], 2016	CNN+inception	UEC-FOOD256	54.7%
Foresti et al. [33], 2018	WISeR	UEC-FOOD256	83.2%

*Top 1 accuracy can be computed using (TP + TN)/(TP + TN + FP + FN)

3. Case Study: Image-based Approach for Food Recognition: Recent research has explored the use of image-based approach for dietary assessment especially on automatic food recognition. Several teams have already put their research works into practice. In study [18], [19], the authors introduced a mobile phone based food classifier with manually selected features extraction method for health-related research and obesity management. The technique can be used to identify food items with the aim of encouraging a healthy choice. For instance, a test has been carried out in the study to distinguish similar food items such as cheeseburgers, double cheeseburgers and burgers without cheese. With the proposed technique, users are able to select the food items according to their health conditions. A similar idea has been proposed by Ravi et al. [40]. They developed a mobile application for food intake classification which provides real time feedback to users as shown in Figure 3. In addition, [29] examined their food classification system on the real menu from 23 restaurants,

and the top-5 error rate is only around 25%. This illustrates that image-based dietary assessment has great potential in converting research prototypes into real-life applications.
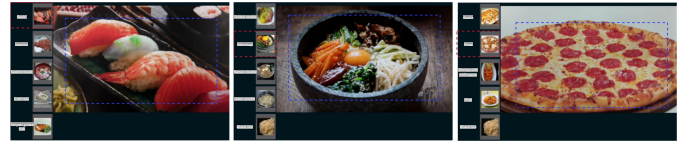

Fig. 3: Screenshots of the mobile application for food classification developed by Imperial College London [40]. The top 5 classes recognised by the app based on the captured frame

C. Food Volume Estimation

To accurately quantify the dietary intake, measuring the portion size or volume of food intake is essential. After a comprehensive exploration, a wide range of research works published between 2009 and 2019 have been reviewed. Several state of the art literature articles about food volume estimation are selected and compared with each other to highlight their advantages and limitations. The differentiation is proposed based on the following five main categories, which is also shown in Table IV.

- **Stereo-based approach:** Stereo-based approach refers to using multiple frames to reconstruct the 3D structure of food objects by finding pixel correspondences between image frames and using the extrinsic parameters to re-project the pixels from image coordinate to world coordinate.
- **Model-based approach:** Model-based approach refers to pre-building shape templates (mathematical models) so that the volume of objects can be determined by model selection followed with model scaling and rotation, which is also known as image registration.
- **Depth camera based approach:** Apart from monocular cameras, other visual sensors are involved in dietary assessment. The most commonly used sensor is the depth camera such as Time Of Flight (TOF) camera. In using depth camera based approach, the actual scale of object items can be obtained without any reference object such as a fiducial marker.
- **Perspective transformation approach:** This refers to the method of estimating object volume based on a single image. By using perspective transformation, a bird's eye view image can be obtained and a rough estimate on the size of the object can be derived. This method does not rely on pre-built shape templates, thus it is typically used to estimate objects with irregular shapes.
- **Deep learning approach:** Deep neural networks have been extensively used in volume estimation. Several research works proposed using a single RGB image to infer the depth map. Voxel representation has been used to present the depth map and the volume can be estimated by counting the number of voxels occupied. In recent, researches explored the use of point cloud completion to achieve volume estimation.

Furthermore, the advantages and challenges for the different approaches are also highlighted in Table V. To evaluate the

TABLE IV: Five types of volume estimation approaches: stereo-based, model-based, perspective transformation, depth camera based and deep learning approach

	Volume Estimation				
Method	Stereo-based	Model-based	Perspective transformation	Depth camera based	Deep learning
Preparation	Camera calibration (Intrinsic matrix)	Model library construction (Pre-build 3D shape models)	Strong constraints required (Camera viewing angle)	Depth camera calibration (Intrinsic matrix)	Model training
Procedure	1. Camera calibration (Extrinsic matrix) 2. 3D model reconstruction 3. 3D meshing and scale determination 4. Volume estimation	1. Food recognition 2. Model selection from the library 3. Model registration by rotating and scaling 4. Volume determination by pre-build models	1. Surface plane equation fitting 2. Perspective transformation 3. Scale determination 4. Volume calculation by geometric relationships	1. Depth map construction 2. 3D model reconstruction based on point cloud or voxel representation 3. Volume estimation	1. Single RGB image captured 2. Depth map estimation/ 3D Shape completion 3. 3D model reconstruction 4. Volume estimation

performance of dietary assessment, various food datasets have been constructed by different research groups, however, most of the publicly known datasets are constructed to examine the performance of food classification only, instead of food volume estimation. To examine the accuracy of volume estimation, authors tend to construct their own databases, which is relatively small in scale compared to the publicly known one. Thus, it seems important to keep in mind which database is used in each case when comparing results. The detailed research methodology and comparison is shown as follows:

1. Stereo-based Approach: Stereo based approach can mainly be divided into multi-view stereo method and binocular stereo method. Multi-view stereo method normally refers to using at least two images taken by a moving camera to reconstruct the 3D structure of the object item. Binocular stereo method uses a binocular camera to capture stereo images and reconstruct the target object structure. Compared to binocular stereo based technique, the multi-view one is more commonly used in food volume estimation due to the low cost and the popularity of monocular cameras. In multi-view stereo method, disparity is an important parameter in estimating depth value in 3D reconstruction. It refers to the distance between two corresponding points in the left and right image of a stereo pair. A single image cannot provide any geometric information about the scene. In the case of using a single image, the same view can be observed in the image even if the target object is placed in different distance from the camera as shown in Figure 4. X_i, where $i = 1, 2, ..., 4$, can all be projected on the left frame shown as X_L. With multi-view images, the actual distance of the object item can be determined. The underlying idea is that X_R will move along the epipolar line when the object item changes its distance from the optical centre, i.e. X_1 to X_4. If the exact location of X_R is determined through point-correspondences, the exact location of the object can be confirmed. This concept is also known as epipolar geometry. The simplified formula of epipolar geometry can be written as
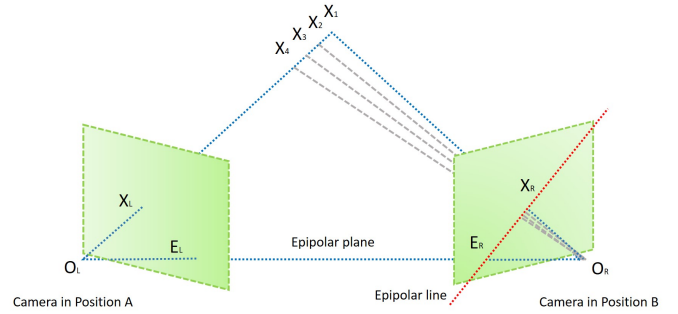

Fig. 4: An illustration diagram of epipolar geometry

follows [41]:

$$x_2^T E x_1 = 0 \qquad (1)$$

where x_2 and x_1 refer to the same point, lying on the normalised camera coordinate (z=1), from two frames matched through point-correspondences and they can be further expressed as $x_1 = [u_1, v_1, 1]^T$ and $x_2 = [u_2, v_2, 1]^T$ respectively. E is the essential matrix which contains the information of camera rotation and transition and it can also be expressed as a vector $\mathbf{e} = [e_1, e_2, ..., e_9]$. The unknown vector $\mathbf{e}$ can be computed by Eight-point algorithm using eight pairs of matched points shown as

$$[u_1 u_2, u_1 v_2, u_1, v_1 u_2, v_1 v_2, v_1, u_2, v_2, 1] \cdot \mathbf{e} = 0 \qquad (2)$$

After E is computed, it can be converted back to rotation and transition using the Singular Value Decomposition(SVD). Once the rotation and translation parameters are obtained, the points can be projected to the world coordinate using Triangulation by equation 3.

$$s_2 x_1^{\wedge} R x_2 + x_1^{\wedge} t = 0 \qquad (3)$$

where s_2 is the depth of x_2, $^{\wedge}$ refers to the outer product here. In 2009, [43] published the first paper on food volume esti-
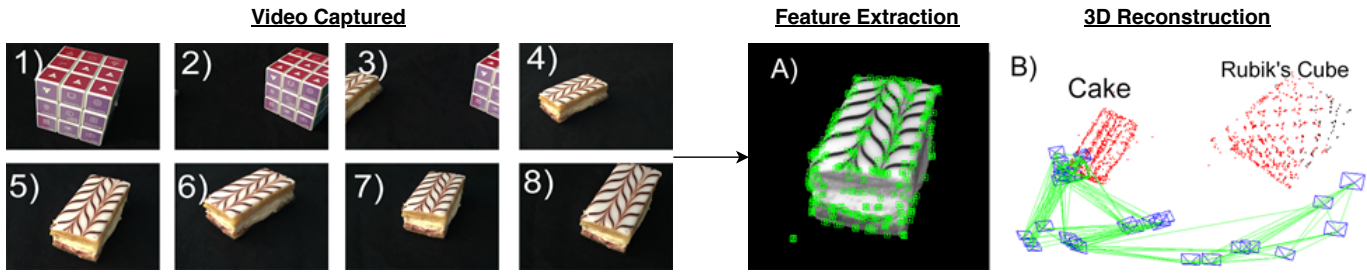
Fig. 5: The implementation of the stereo-based volume estimation using SLAM. The visual odometry can be shown in the figure on the right (the blue objects represent the location of the camera where the frames are captured and the green lines refer to the trajectory of the camera) [42]

mation based on multi-view stereo reconstruction. Afterwards, increasing research works focus on using multiple images taken from different angles to carry out 3D reconstruction for volume estimation. In using such an approach, extrinsic calibration is firstly carried out to determine the geometric relations between captured frames, which is known as relative pose, or between the frames and the reference object, which is known as absolute pose [44]. To perform camera calibration, RANdom SAmple Consensus (RANSAC) scheme is commonly used. RANSAC scheme starts from extracting suitable descriptors, i.e. Harris corners, SIFT and ORB, and n point-correspondences (n is normally set as 4) in different frames. However, features matching in stereo vision is a slow process restricted by epipolar geometry. Each pixel's match can only be found on a slanted line called epipolar line as shown in Figure 4. In order to speed up the process, image rectification has been performed to wrap the images such that the two images appear as if they are captured with just a horizontal displacement without any rotation, which limits the features searching region to a straight horizontal line. This undoubtedly speeds up the processing. Once the camera poses are estimated, the matched feature points of the frames can be projected to world coordinate using triangulation, i.e. Delaunay triangulation [43], in order to construct 3D models. When using a single camera to carry out 3D reconstruction, scale ambiguity is always an issue which should be taken into account. It is for this reason that the scale determination should be carried out to recover the global scale factor in order to find the actual volume of the food object. A fiducial marker with known dimensions, such as a checker-board [41], are normally be placed along with the food items as a reference. The fiducial marker can provide geometric information such as width and height so as to help calibrate the measurement and build the world coordinate. Open source Computer Vision (OpenCV), a library containing a series of algorithms, is widely used along with the fiducial marker to help detect the corners of the checker-board, calculate the intrinsic and extrinsic matrix to determine the focal length and relative camera pose respectively. The mathematical expression for the scale calculation can be shown as equation 4.

$$S = d_{Reference} / d_{Estimated} \quad (4)$$

where $d_{Reference}$ refer to the actual dimension of the object

in the real world scene, and $d_{Estimated}$ refer to the dimension estimated after 3D reconstruction in the world coordinate. In [43], the average error in volume estimation is $5.75(\pm 3.75)\%$ when dense stereo reconstruction has been applied. Despite the efficacy of the dense reconstruction, there are still limitations on the proposed technique. One of the major drawbacks is that it takes around 33 seconds for the dense stereo reconstruction and volume estimation in the paper (if a size of 1600×1200 pixel image is used). In study [44], the authors proposed the use of two-view 3D reconstruction to speed up the processing time to 5.5 seconds per frame in 2017. The method can be divided into three parts including extrinsic calibration, dense reconstruction and volume estimation. In the extrinsic calibration, the authors proposed a modified RANSAC method to carry out relative pose extraction based on SURF descriptors. Local optimisation is carried out by maximising inlier count and minimising inlier distances when a new model is found using RANSAC. Besides, an adaptive threshold estimation has been used to find the inlier threshold instead of a fixed threshold value. With the use of the proposed technique, the mean absolute percentage error ranges from 8.2% to 9.8% examining on Meals-14 dataset. This proposed technique has become the-state-of-art technique and outperformed other stereo-based techniques which have been tested with a similar size database. Similar idea has also been proposed by [45] in which two-view dense stereo reconstruction is carried out to reconstruct 3D food models. Apart from typical stereo-based techniques, a modified stereo-based approach based on Simultaneous Localisation And Mapping (SLAM) has been presented recently [42]. The visual SLAM framework can be divided into four parts including visual odometry, loop closure, back-end optimisation and mapping [41]. In [42], in order to explore the feasibility of SLAM in continuous measurement of food consumption, several experiments have been carried out with their self-collected dataset. The overall volume estimation accuracy can achieve 88.3% if the food item is not consumed and 83.6% when the food item is being consumed. Despite the fact that the accuracy is relatively low compared to the traditional stereo-based approach, the real time property of SLAM, as shown in Figure 5, is worth investigating, as previously proposed techniques mostly relied on post-processing, and SLAM-based approach could better estimate food consumption by continuously capturing and measuring the food while it is being consumed.

2. Model-based Approach: Despite the fact that stereo-based approach can provide more spatial information about the food geometry, it is considered troublesome to capture multi-view images around the food during the meals. Instead of using stereo-based method, model-based techniques have also been proposed. Such approach refers to the use of pre-defined 3D shape models/templates to estimate the target object volume. A variety of 3D models will be constructed and stored in a model library at the beginning. After food recognition, the model corresponding to the label or with similar outlook/characteristics will be selected from the library. The selected models are not always perfectly fit to the object items. Thus, model selection will always have to be rotated, translated and scaled to match the contour of the food item in the image, which is known as model registration. In 2015, [11] proposed a Virtual Reality (VR) approach which makes use of the different mathematical models to superimpose corresponding food items in the real world scene. The volume of the mathematical models have been pre-defined so that the volume of the food items can be estimated by scaling and rotating the model. This idea has been used in Technology Assisted Dietary Assessment (TADA) system as well [46], [16]. A similar model-based technique has also be proposed in [47]. The authors proposed the use of templates from model library to represent the object items and use coordinate representation to determine their location and size. The representation can be written as $E = (x, y, \phi, \Theta, s_x, s_z)$, where x and y refer to the coordinate of x and y-axis respectively, ϕ refers to the elevation angle of the object item, Θ refers to the rotation angle, and s_x and s_z refers to the scale of the object item in x and y-axis respectively. Through pose registration, the predefined models can be projected into the world coordinate followed by volume estimation. In using model-based approach, food volume can be easily obtained even if single viewing angle is used. The limitation of this approach, however, is that the model database should be pre-trained. In such a case, if the object items are unseen and irregular, this will induce a large estimation error.

3. Depth Camera based Approach: In [48], the authors examined the practicality of using a depth camera to estimate food volume. A single depth map has been captured using a 3D sensor system developed by [49] from the top, i.e. bird's eye view. A depth map refers to an image channel that presents the information of the distance between the surfaces of objects and the camera. The depth map is then converted to a voxel representation to estimate the volume. Voxel representation is a commonly used technique to quantify the object size by counting the number of voxels that constructed the 3D models. Similar technique has been used in various research works [29], [48], [50]. While using voxel representation, the reference plane should firstly be determined since the placement of voxels should be perpendicular to the plane surface. In [29], RANSAC has been used to find the plane geometry. In [48], the authors further exploited an alternative method to find the reference surface which is an expectation-maximisation based technique. The technique clusters the image into surface and non-surface regions based on the depth and RGB information using Gaussian Mixture Model (GMM). From the qualitative result, we can see the plane searching is reasonably accurate, however, there is still no quantitative data shown in the paper proving that this technique outperforms RANSAC method. After examining the performance of depth camera based volume estimation, the authors found that 90% of testing samples are overestimated, however, the underlying reason has not been discussed in the paper. Similar problem has been found in [42]. The author stated that the point cloud/voxel representation generated by a top view depth map can only be used to tackle object items with narrow top and wide bottom. It is for the reason that infrared light generated by depth camera cannot reach the bottom if the upper part is wider. Due to the reason of limited viewing angle, this will induce error on volume estimation. To solve this problem, several ways have been proposed by previous research works. For example, [42] proposed to address this problem by carrying out point cloud completion based on symmetry. In [44], the stereo-based approach using multiple images from different angles to reconstruct 3D point cloud, was proposed to address this issue.

4. Perspective Transformation Approach: Despite the efficacy of infra-red light and stereo-based cameras on volume estimation, there has been relatively few depth cameras which have been integrated into mobile or wearable devices for various reasons, such as cost and power constraints. It is for these reasons that much research on the topic of volume estimation from a monocular camera is still gaining interests. Compared with traditional depth camera techniques, estimating volume from single RGB image is much more challenging, since it requires accurate scale calibration of the camera based on a reference object with known dimensions along with various perspective transformations [51]. Besides, single image itself does not have much geometric information, so that strong constraints or assumptions have to be used for 3D reconstruction. For example, [52] presented a technique called the plate method, which makes use of a plate with known dimension. In this approach, coordinate transformations have been carried out to locate the plane and find its corresponding surface equation based on the known radius of the plate. Afterwards, the distance between the optical center of the camera and the plate can be determined. With known distance, the width and height of the objects can be estimated respectively based on simple geometry as shown in Figure 6. In this approach, the line *HG* shown in Figure 6, which refers to the perspective projection of the height of object in real world scene, is required to determine the actual height of the object (point *H* and point *G* correspond to the coordinate of the top and bottom of the object in the image plane respectively). The algorithm has been examined on a self-collected dataset with 10 object items. The overall accuracy is of 88%. Despite the satisfactory performance, there is still limited work on detecting the coordinate of those points without manual operation. In using such approach, users need to locate the top and bottom of the object items by themselves. This also means that the sensors designed based on this approach can only be semi-automatic and may further limit its use in large-scale studies. In addition, to accurately
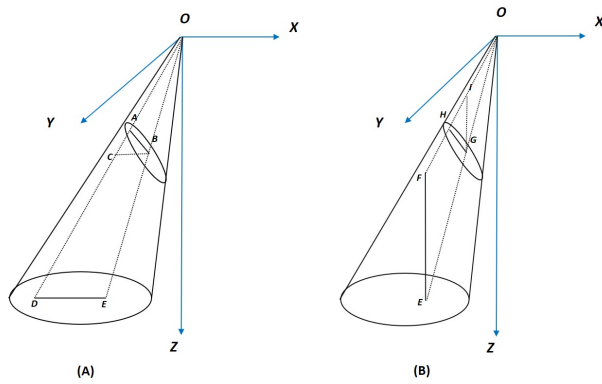
Fig. 6: (A) Width estimation (B) Height estimation based on simple geometry

estimate the food volume based on perspective transformation approach, front (top) view image is always required since the scale can only be determined by the fiducial marker in this way. However, in practice, users will seldom take food images from the top, instead users are used to taking photos from the side or at a convenient angle. Therefore, perspective transformation becomes important in scale estimation since it can convert a geometric distorted image to a front view image. [53] proposed the use of 12 feature points from the checkerboard to determine a $3*3$ projective transformation matrix H. The equation can be shown as equation 5.

$$wp^{'} = Hp \qquad (5)$$

where $p^{'}$ refers to the features point after transformation and p refers to the feature point before transformation. The transformation matrix H is defined up to an arbitrary scale factor. The equation can also be expressed as equation 6.

$$\begin{bmatrix} wx_i^{'} \\ wy_i^{'} \\ w \end{bmatrix} = \begin{bmatrix} h_{00} & h_{01} & h_{02} \\ h_{10} & h_{11} & h_{12} \\ h_{20} & h_{21} & h_{22} \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \qquad (6)$$

Afterwards, Direct Linear Transform (DLT) has been carried out to estimate the transformation matrix/vector. Since H is defined up to an arbitrary scale factor so that we have only 8 unknown parameters (at least 4 points needed). If more points are used, it can be defined as a least squares problem. In order to get an accurate transformation matrix, feature points matching become important. A common way which has been widely used is RANSAC. Furthermore, RANSAC is also a common way to perform plane fitting. This is important to address the problem when the image is captured at an angle but not from the top. For example, if we are going to estimate the object volume based on depth map but the image is not captured from the top view. The depth value cannot be used directly since we do not have the depth value of the horizontal surface, i.e. normally the table or plate, which is occluded by the object. We cannot calculate the height of the object just by subtracting the depth value of object from the depth value of the surface. However, the depth value of the surface can be determined by using plane fitting where the depth value of the non-occluded surface is used to fit a plane. By using this

plane, the depth value of the whole surface can be determined and the height and volume of the object can be calculated. To fit a plane, we need the exact coordinate of minimum 3 points on a plane. The problem is that we do not know which points are on the same surface in the real scene. One solution is to perform RANSAC. To do RANSAC, three points are chosen randomly and we assume those points are on the same plane. After the plane is formed, there may be many other points which will also lie on the plane. We can then compute the distance between those points and the plane and count the number of points which is lying within a predefined threshold, i.e. these points are known as inliers. This process is iterated until the plane formed can fit the largest number of inliers.

5. Deep Learning Approach: In recent years, deep learning has been applied extensively to single images for computer vision applications. The advantage of using deep neural network for food volume estimation is that the scale of the monocular image can be learned from the global cues of the scene without the need of camera calibration, which means reference objects with known dimension are not required. Also, a single RGB image is enough to estimate the volume instead of using multiple view images or a stereo camera approach. With the use of deep learning technique to estimate the food volume, depth is always important information to be used [54]. Though depth cameras have gained increasing popularity in recent years, the majority of mobile or wearable devices are still embedded with a monocular RGB camera. This motivates the use of a model to predict the depthmap. It is well known that estimating depth from single RGB image is a challenging ill-posed problem. However, it is still possible to get a reasonable accuracy by using the models proposed by several research groups. For example, [55] has proposed using multi-scale deep network to predict the depth from a given single RGB image. The underlying idea of the work is to make use of the coarse-scale network to predict the depth of the global features and use a fine-scale network to refine the local features. The work is later extended to a three-scale architecture for further refinement. From the architecture, the feature maps of different scales have been upsampled and fused together to improve the global and local depth estimation [56]. Apart from the use of multiscale architecture, there are still other techniques used in depth estimation. In [57], the author noticed the continuous characteristic of depth value and proposed to address the problem by formulating depth estimation into a continuous conditional random field (CRF) learning problem. With the use of this depth estimation technique, the food volume can be estimated with a single RGB image which makes dietary monitoring more efficient. In recent, Google [29] has proposed using deep neural network to estimate the food volume. Regarding to the coupled nature of depth and volume for every particular object, a CNN architecture has been applied to a single RGB image to estimate the depthmap. The model has been pre-trained based on NYU v2 RGBD dataset and then fine tuned on their newly collected dataset called GFood3d with 60 different meals from various Google cafes i.e.150k frames (This dataset is not disclosed). The data has been collected by RealSenseF200 depth sensor. After the

depthmap is obtained by the trained model, it is converted to a voxel representation. With the voxel representation and the segmented labeling, the volume of the labelled items can be obtained respectively. The performance has been examined by comparing the result with the dataset called NFood-3d which contains food with volume labelled (This dataset is not disclosed). In this work, food segmentation has not been performed since their segmentation model does not work on their dataset due to dissimilar colour and texture property of artificial food and real food. In this case, the volume of segmented food cannot be estimated, instead it is estimated as a whole meal. The error for each meal lies between $50cm^3$ and $400cm^3$. From the experimental results, the performance of this work is difficult to evaluate since the volume of ground truth for each meal is not disclosed as well. Similar research work [58] has also been published. The work aims at regressing the Bread Units (BUs), which is a mathematical function of food volume and bread unit density, by deep neural network in order to assess nutritional information for diabetes patients. The author proposed a 2-stage approach to achieve the BUs estimation. First, a fully convolutional neural network has been trained with the use of the NYU Depth v2 dataset (this is not a food database) to predict the depth map of a given food image. Second, they trained another neural network on top of Resnet-50 proposed in [59] to regress the bread units by using both RGB images and ground-truth depth maps (measured by Microsoft Kinect v2 sensor) as input. The last layer of the network has been replaced with a single neuron with corresponding L_2 cost function rather than using a softmax-layer. Besides, the author states that initialising the weights of neural networks from similar tasks can help promote convergence and reach a higher absolute performance. It is for theses reasons that the model used in stage 2 has been pre-trained on the Food101 database and trained on their manually built food database with 60 western dishes afterwards i.e. 9k images labelled with corresponding BUs. This means that the authors initialise the filter parameters corresponding to the RGB input with the pretrained value and set the parameters corresponding to depth randomly in the initialisation scheme. Last but not least, the estimated depth map generated in the first step will be fused with RGB images to predict the BUs. In this study, the performance of the model has been examined by using BU prediction so that it is difficult to compare this with previously proposed works which estimate the volume. Because of this, we investigated into the depth prediction model in order to evaluate the performance. To estimate the volume from a single RGB image, the accuracy of the inferred depthmap is an important parameter that directly affects the performance and efficacy of volume estimation. However, the depth prediction model proposed by [58], which is the state of the art, still achieve RMSE of 65cm, on the dataset of NYU Depth v2 and achieve 12.9cm, 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 large. Nevertheless, comparing the qualitative results of two proposed works based on deep learning, the latter one infers depth map with more fine local details. Apart from estimating food volume based on predicted

depth image, [60], [61] proposed using shape completion technique to estimate food volume recently. In the study, they stated that the back side of food items cannot be captured due to limited viewing angle, and thus the food volume will be underestimated. To address the problem, shape completion network has been used to complete the occluded region of the food items. Alphashape algorithm is then used to compute the volume of the completed food items. Furthermore, [62] makes use of Generative Adversarial Networks (GAN) to infer food energy (kilo-calories). To conclude, it is still challenging to use deep learning to estimate food volume due to the reason of insufficient information from a single image to accurately reconstructing the 3D objects, and not enough representative training data to train the network. Compared to other volume estimation technique, the error is relatively large, however, volume estimation based on single image is still worth investigating due to the reason of practicality and the ease of use.

6. Case Study: Image-based Approach for Volume Estimation: For volume estimation, several teams have already put their research works into practice. In study [44], the authors have developed a carbohydrate counting system based on stereo-based approach. 77 real meals with known volume have been evaluated using mobile phones and achieved an average error rate of 10%. A clinical trial has also been carried out in their study which explored the feasibility of using the technique in improving self-management of patients with type 1 diabetes. In additional to stereo-based approach, model-based approach is also commonly used in the field of dietary assessment. Ongoing research work [11] aims to develop a multi-purpose, unified wearable sensor using model-based approach to acquire data for the evaluation of dietary intake, which is easier to incorporate into users' daily routines compared to stereo-based approach. A pilot study of seven human subjects was conducted to assess the feasibility of using wearable sensors to achieve dietary food intake measurement. This study indicates that the use of wearable devices for monitoring dietary food intake shows a strong potential to reduce the burdens on users in reporting the food volume. For the other approaches, most of them are developed for research purposes and limited to laboratory settings without any clinical trials and industrial applications at the current stage. More studies are required to carry out in the real world scenario to realise the potential impact of wearable sensors for dietary assessment and personal health study.

III. DISCUSSION

In image/food recognition, several research works have formally validated that deep neural network outperforms traditional approaches, based on manually extracted features, using criterion measures and publicly known databases. Despite existing approaches show promising results in tackling the issue of food recognition, there still exists a wide range of challenges and hurdles in estimating the nutrient intake, as nutritional food information rely mostly on portion size, and the performance of up-to-date volume estimation techniques is not yet satisfactory. Stereo-based techniques rely strongly

TABLE V: High impact research works on volume estimation published between 2009 and 2019

Authors	Brief description	Error	Highlight	Drawbacks
Puri et al. [43] (2009)	Multi-view dense stereo reconstruction (*Stereo-based approach*)	Ranging from 2.0% to 9.5% (Mean volume estimation error); Evaluated on the dataset with 26 types of food objects	Able to handle irregular food shape	Processing time is slow (33 seconds); troublesome to take multiple images; difficult to handle object occlusion
Rahman et al. [45] (2012)	Two-view dense stereo reconstruction (*Stereo-based approach*)	7.7% (Mean volume estimation error); Evaluated on the dataset with 6 types of fruits	Able to handle irregular food shape; Without the need for manual fitting of 3D shape templates to the object items	Less textured object items may lead to wrong 3D volume estimation
Dehais et al. [44] (2017)	Two-view dense stereo reconstruction (*Stereo-based approach*)	Ranging from 8.2% to 9.8% (MAPE in volume estimation); Evaluated on two datasets with 45 and 14 dishes respectively	Fast processing time; a modified RANSAC method	Troublesome to take multiple images; RANSAC easily fails when the food texture is not obvious
Gao et al. [42] (2018)	SLAM-based sparse stereo reconstruction (*Stereo-based approach*)	Ranging from 11.7% to 19.2% in static; Ranging from 16.4% to 27.9% in real time (Mean volume estimation error)	Real time measurement; point cloud completion to tackle limiting viewing angles	Relatively low in accuracy; requires strong assumption
Zhu et al. [16] (2010), Woo et al. [63] (2010)	3D reconstruction based on spherical and prismatic models with a reference card (*Model-based approach*)	5.6% (Mean volume estimation error); Evaluated on 7 types of food items	Detailed information on food segmentation, classification and quantification	Test dataset is relatively small; the error of prismatic models is relatively large
Chae et al. [64] (2011)	Volume estimation using food specific shape templates (*Model-based approach*)	11% for drinks; 8% for bread slices (Mean volume estimation error)	High accuracy in pre-built food shape	Very limited shape templates; difficult to handle unseen shape;
Xu et al. [47] (2013)	3D-model generation and pose estimation have been used (*Model-based approach*)	Ranging from 3.6% to 12.3% (Mean volume estimation error)	High accuracy in pre-trained food shape	Difficult to handle unseen shape; model library required
Jia et al. [65] (2014); Sun et al. [11] (2015)	Virtual reality approach (*Model-based approach*)	20.5% (RMSE in volume estimation); Tested on 100 real-life food objects	Robust volume estimation system (85 out of the 100 object items have errors within 30%)	Difficult to handle unseen shape; model library required
Fang et al. [66] (2015)	Single-view reconstruction based on both shape templates (*Model-based approach*) and prism models (*Perspective transformation approach*)	The proposed method was able to achieve an error of less than 6% in food energy; Evaluated on the dataset with 45 different individual eating occasions	Not relying on manual initialization estimation parameters; automatically estimate volume using the geometric contextual information	The height of the entire horizontal cross-section is assumed to be the same
Shang et al. [67] (2011)	Make use of a mobile structured light system (SLS) to achieve dietary assessment (*Depth camera-based approach*)	The performance of the system is not presented	The system consists of a mobile device and a laser attachment	Troublesome to scan the whole food items to obtain the 3D models
Fang et al. [48] (2016)	Make use of depth camera to estimate volume (*Depth camera-based approach*)	Ranging from 11.0% to 33.9% (Mean volume estimation error)	An alternative method to find plane equation	Depth camera is not always embedded in mobile devices; overestimation problem
Jia et al. [52] (2012); Yue et al. [68] (2012)	Make use of plate method and LED method (*Perspective transformation approach*)	<25% (point clicking method) <10% (wireframe fitting method) (Mean volume estimation error)	Use a plate with known dimension to be the fiducial marker	Semi-automatic; limit the use in large-scale study
He et al. [53] (2013)	Use shape template for foods with regular shapes (*Model-based approach*) and area-based weight estimation for foods with irregular shapes (*Perspective transformation approach*)	The average relative error is 11% for beverage images using shape templates and the error for area-based weight estimation method is 10%.	Able to handle irregular food shape; make a comparison between perspective transformation approach and model-based approach	Area-weight relation may induce large error in the testing procedure
Pouladzadeh et al. [69] (2014)	Make use of area measurement technique to achieve volume estimation (*Perspective transformation approach*)	It achieves a reasonable error of about 10% in the worst case; Evaluated on the dataset with 5 food items	Able to handle irregular food shape; make use of user's thumb for calibration	Strict constraints on image capture (one from above and one from the side)

Yang et al. [70] (2018)	Using a special picture-taking strategy to determine the actual scale of the food objects (*Perspective transformation approach*)	The average absolute error is 16.65% and 47.60% for large and small food objects respectively	Food volume estimation without using a fiducial marker	Troublesome to take the food picture with the bottom of the smartphone sitting on the tabletop
Meyers et al. [29] (2015)	Estimate the depth map through CNN architecture followed by volume calculation (*Deep learning approach*)	50-400ml (Mean volume estimation error); Evaluated on NFood-3d dataset	A comprehensive study on volume estimation using deep neural network	High error in depth map prediction; high error in volume prediction; still limited to the laboratory circumstance
Christ et al. [58] (2017)	2-stage approach: Predict the depth map through FCNN; estimate the bread units through Resnet-50 (*Deep learning approach*)	1.53 bread units (RMSE in bread units); Evaluated on 20 dishes of their own dataset	Confirm the feasibility of using neural network to estimate depth and bread units	Not an end-to-end network; high error in BUs prediction; still limited to the laboratory circumstance
Lo et al. [54] (2018)	Make use of deep learning view synthesis to estimate food volume (*Deep learning approach*)	6.9% for testing on the dataset with 8 synthetic food objects (Mean volume estimation error)	Address the problem of visual occlusion based on deep learning view synthesis	Experiments were developed for research purposes and are still limited to laboratory settings
Fang et al. [71] (2018)	Make use of generative adversarial network to estimate food energy distribution (*Deep learning approach*)	10.89% (Mean energy estimation error); 1875 paired images were used to train the network and 220 paired images were used for testing	Learn the mapping of the food image to the food energy distribution; End-to-end network	Paired RGB images and energy distribution cannot be obtained easily
Lo et al. [61] (2019); Lo et al. [60] (2019)	Make use of 3D point cloud completion method to achieve accurate volume estimation (*Deep learning approach*)	15.3% for testing on real food items; 7.7% for testing on synthetic food items (Mean volume estimation error)	Address the problem of visual occlusion based on deep learning in order to increase the volume estimation accuracy	Paired partial and complete point cloud sets cannot be obtained easily

on feature matching between frames. This special property facilitates the volume estimation of food items in irregular food shape so that a larger variety of food items can be measured automatically without manual intervention and pre-trained model library. However, the drawback is that the 3D models cannot be reconstructed and the volume estimation will fail if the food surface does not have distinctive characteristics or texture. Another concern is that stereo-based approach requires users to capture multiple images from different viewing angles before and after eating, which in turn makes this approach very tedious and not suitable to be applied on wearable sensors for long-term health monitoring and data collection. These findings show that dietary assessment based on a single image seems to be one of the future trends in dietary assessment. High estimation accuracy in model-based approach proves the feasibility of using a single image to assess food intake along with a pre-trained model library. However, the existing research studies on model-based approach have only examined their algorithms on small model libraries consisting of several simple geometric shapes such as cube, sphere, cone, cylinder and etc. Further works are required to develop a more comprehensive model library which is able to deal with food items even in irregular shape. With respect to perspective transformation approach, it usually has strong constraints on the image capturing angle and position. Besides, perspective transformation are required before estimating the scale. Although the processing time has not been discussed in the study using perspective transformation approach, it is reasonable to think that the process is time-consuming due to

the two-stage perspective transformation (to obtain the top and the side view) in order to obtain the scale from the fiducial marker. For deep learning approach, the problem of food volume/weight estimation by means of deep neural network is harder than initially expected and the accuracy rate is relatively low compared to other approaches, which have been validated in the previous section. These findings illustrate that deep learning approach seems to not be able to handle the problem individually without combining with other approaches. Most importantly, most of these approaches are currently developed for research purposes and limited to laboratory settings.

After extensive reviews and comparison, to the best of the authors' knowledge, the depth camera based approach is the most robust and efficient technique compared to the previously mentioned approaches, as it can be used to determine the scale and estimate volume without relying on any fiducial markers. The main challenge that lies ahead for this approach is to tackle the issues of view occlusion. Occlusion refers to information lost when the back of the food items cannot be captured due to the limited viewing angle, and thus the 3D model cannot be perfectly reconstructed. This will inevitably induce errors in volume estimation, i.e. overestimation. It is for this reason that model completion/point cloud completion is required to build on top of the existing methodologies to tackle the issue. While deep learning approach does not show promising results in volume estimation, on the contrary, it has large potential in view synthesis [72], which refers to the technique that synthesises various images captured from different viewing angles based on a single image. Undoubt-

edly, it stands a high chance to improve the performance of model completion/point cloud completion as well as the estimation accuracy. An integrated system based on depth camera based approach along with view synthesis using deep learning methods could be one of the future directions to tackle the issue. Apart from depth camera based approach, SLAM-based technique is also a promising approach. Due to its real time property, it can estimate the volume of the object item as well as dietary intake while the user is consuming the food. Not only can this provide more comprehensive monitoring, but it can also obtain the eating behaviour of the users such as intake sequence and speed, which are not feasible with traditional methods. Further research works to examine these hypothesis and additional validation studies are also required.

IV. CONCLUSION

This is the first review which investigates the underlying algorithms and mathematical models used in the field of dietary assessment especially on food recognition and volume estimation. After a comprehensive review on several state-of-the-art food recognition systems, recent research has found to be focused on exploring the potential of assessing dietary intake based with deep learning approach. Furthermore, the state of the art approaches in food volume estimation are summarised and discussed in this study. Extensive comparison has also been presented to highlight the main advantages and challenges of different approaches. Overall, there is currently a growing potential in integrating different approaches to improve the overall accuracy in food volume estimation. If the challenges can be resolved, image-based dietary assessment will definitely play an important role in nutritional health monitoring in the near future.

REFERENCES

[1] NHS, "Statistics on obesity, physical activity and diet england 2018," *NHS Digital*, 2018. [Online]. Available: https://digital.nhs.uk/.

[2] G. A. Bray and B. M. Popkin, "Dietary fat intake does affect obesity!" *The American Journal of Clinical 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 Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1261–1271, 2014.

[4] O. Amft, "A wearable earpad sensor for chewing monitoring," *IEEE Sensors*, pp. 222–227, Nov. 2010.

[5] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, "Detecting periods of eating during free-living by tracking wrist motion," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1253–1260, Jul. 2014.

[6] M. Farooq, J. M. Fontana, and E. Sazonov, "A novel approach for food intake detection using electroglottography," *Physiological Measurement*, vol. 35, no. 5, p. 739, 2014.

[7] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, no. 10, p. 719, 2018.

[8] C. J. Boushey, M. Spoden, F. M. Zhu, E. J. Delp, and D. A. 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.

[9] T. Vu, F. Lin, N. Alshurafa, and W. Xu, "Wearable food intake monitoring technologies: A comprehensive review," *Computers*, vol. 6, no. 1, 2017.

[10] L. Gemming, J. Utter, C. Ni Mhurchu, and D. A. Kerr, "Image-assisted dietary assessment: A systematic review of the evidence," *Journal of the Academy of Nutrition and Dietetics*, vol. 115, no. 1, pp. 64–77, 2015.

[11] M. Sun, L. E. Burke, T. Baranowski, J. D. Fernstrom, H. Zhang, H.-C. Chen, Y. Bai, Y. Li, C. Li, Y. Yue, *et al.*, "An exploratory study on a chest-worn computer for evaluation of diet, physical activity and lifestyle," *Journal of Healthcare Engineering*, vol. 6, no. 1, pp. 1–22, 2015.

[12] J. Nie, Z. Wei, W. Jia, L. Li, J. D. Fernstrom, R. J. Sclabassi, and M. Sun, "Automatic detection of dining plates for image-based dietary evaluation," *Annual International Conference of the IEEE Engineering in Medicine and Biology (EMBC)*, pp. 4312–4315, 2010.

[13] M. Anthimopoulos, J. Dehais, S. Shevchik, B. H. Ransford, D. Duke, P. Diem, and S. Mougiakakou, "Computer vision-based carbohydrate estimation for type 1 patients with diabetes using smartphones," *Journal of Diabetes Science and Technology*, vol. 9, no. 3, pp. 507–515, 2015.

[14] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International Journal of Computer Vision*, vol. 1, no. 4, pp. 321–331, Jan. 1988.

[15] V. Caselles, R. Kimmel, and G. Sapiro, "Geodesic active contours," *International Journal of Computer Vision*, vol. 22, no. 1, pp. 61–79, Feb. 1997.

[16] F. Zhu, M. Bosch, I. Woo, S. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756–766, Aug. 2010.

[17] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, M. Yunsheng, S. Chen, and P. Hou, "A new deep learning-based food recognition system for dietary assessment on an edge computing service infrastructure," *IEEE Transactions on Services Computing*, vol. 11, pp. 249–261, 2018.

[18] F. Kong and J. Tan, "Dietcam: Regular shape food recognition with a camera phone," *International Conference on Body Sensor Networks*, pp. 127–132, 2011.

[19] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, no. 1, pp. 147–163, 2012.

[20] N. Tammachat and N. Pantuwong, "Calories analysis of food intake using image recognition," in *ICITEE*, IEEE, 2014, pp. 1–4.

[21] Y. Kawano and K. Yanai, "Food image recognition with deep convolutional features," in *ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, ser. UbiComp '14 Adjunct, Seattle, Washington: Association for Computing Machinery, 2014, pp. 589–593.

[22] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar, "Menu-match: Restaurant-specific food logging from images," *IEEE WACV*, pp. 844–851, 2015.

[23] H. He, F. Kong, and J. Tan, "Dietcam: Multiview food recognition using a multikernel svm," *IEEE journal of biomedical and health informatics*, vol. 20, no. 3, pp. 848–855, 2015.

[24] Y. Kawano and K. Yanai, "Foodcam-256: A large-scale real-time mobile food recognitionsystem employing high-dimensional features and compression of classifier weights," in *ACM International Conference on Multimedia*, ser. MM '14, Orlando, Florida, USA: Association for Computing Machinery, 2014, pp. 761–762.

[25] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Analysis of food images: Features and classification," in *IEEE ICIP*, Oct. 2014, pp. 2744–2748.

[26] P. Pouladzadeh, S. Shirmohammadi, A. Bakirov, A. Bulut, and A. Yassine, "Cloud-based svm for food categorization," *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5243–5260, 2015.

[27] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," *Proceedings of the ACM International Conference on Multimedia*, MM '14, pp. 1085–1088, 2014.

[28] S. Christodoulidis, M. Anthimopoulos, and S. Mougiakakou, "Food recognition for dietary assessment using deep convolutional neural networks," in *ICIAP*, Springer, 2015, pp. 458–465.

[29] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, "Im2calories: Towards an automated mobile vision food diary," *ICCV*, pp. 1233–1241, 2015.

[30] C. Liu, Y. Cao, Y. Luo, G. Chen, V. Vokkarane, and Y. Ma, "Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment," in *ICOST*, Springer, 2016, pp. 37–48.

[31] P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan, "Foodnet: Recognizing foods using ensemble of deep networks," *IEEE Signal Processing Letters*, vol. 24, no. 12, pp. 1758–1762, Dec. 2017, ISSN: 1558-2361. DOI: 10.1109/LSP.2017.2758862.

[32] F. Yu, D. Wang, E. Shelhamer, and T. Darrell, "Deep layer aggregation," in *IEEE CVPR*, 2018, pp. 2403–2412.

[33] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *IEEE WACV*, 2018, pp. 567–576.

[34] Y. Cui, Y. Song, C. Sun, A. Howard, and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *IEEE CVPR*, 2018, pp. 4109–4118.

[35] J. Qiu, F. P.-W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," in *BMVC*, 2019.

[36] W. Min, L. Liu, Z. Luo, and S. Jiang, "Ingredient-guided cascaded multi-attention network for food recognition," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 1331–1339.

[37] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2019.

[38] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, IEEE, 2015, pp. 1–6.

[39] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, Amsterdam, The Netherlands: Association for Computing Machinery, 2016, pp. 41–49.

[40] D. Ravì, B. Lo, and G. Yang, "Real-time food intake classification and energy expenditure estimation on a mobile device," in *IEEE BSN*, Jun. 2015, pp. 1–6.

[41] X. Gao, T. Zhang, Y. Liu, and Q. Yan, *14 Lectures on Visual SLAM: From Theory to Practice*. Publishing House of Electronics Industry, 2017.

[42] A. Gao, F. P. W. Lo, and B. Lo, "Food volume estimation for quantifying dietary intake with a wearable camera," *IEEE BSN*, pp. 110–113, 2018.

[43] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," *WACV*, pp. 1–8, 2009.

[44] 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.

[45] M. H. Rahman, Q. Li, M. Pickering, M. Frater, D. Kerr, C. Bouchey, and E. Delp, "Food volume estimation in a mobile phone based dietary assessment system," in *IEEE SITIS*, IEEE, 2012, pp. 988–995.

[46] 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," *IEEE International Symposium on Multimedia*, pp. 290–295, 2010.

[47] C. Xu, Y. He, N. Khanna, C. J. Boushey, and E. J. Delp, "Model-based food volume estimation using 3d pose," *IEEE ICIP*, pp. 2534–2538, 2013.

[48] 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," *IEEE ICIP*, pp. 26–30, 2016.

[49] S. Zhang, "Flexible 3d shape measurement using projector defocusing: Extended measurement range," *Optics Letters*, vol. 35, pp. 934–936, 2010.

[50] F. P.-W. Lo, Y. Sun, and B. Lo, "Depth estimation based on a single close-up image with volumetric annotations in the wild: A pilot study," in *IEEE AIM)*, Jul. 2019, pp. 513–518.

[51] C.-J. Du and D.-W. Sun, "Estimating the surface area and volume of ellipsoidal ham using computer vision," *Journal of Food Engineering*, vol. 73, no. 3, pp. 260–268, 2006.

[52] W. Jia, Y. Yue, J. D. Fernstrom, N. Yao, R. J. Sclabassi, M. H. Fernstrom, and M. Sun, "Imaged based estimation of food volume using circular referents in dietary assessment," *Journal of Food Engineering*, vol. 109, pp. 76–86, 2012.

[53] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Food image analysis: Segmentation, identification and weight estimation," *IEEE ICME*, pp. 1–6, 2013.

[54] F. 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, 2018.

[55] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *CoRR*, vol. abs/1406.2, 2014.

[56] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *IEEE ICCV*, pp. 2650–2658, 2015.

[57] F. Liu, C. Shen, and G. Lin, "Deep Convolutional Neural Fields for Depth Estimation from a Single Image," *IEEE ICCV*, pp. 1–13, 2015.

[58] P. Ferdinand Christ, S. Schlecht, F. Ettlinger, F. Grun, C. Heinle, S. Tatavatry, S.-A. Ahmadi, K. Diepold, and B. H. Menze, "Diabetes60 - Inferring Bread Units From Food Images Using Fully Convolutional Neural Networks," *IEEE ICCV*, pp. 1526–1535, 2017.

[59] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," *IEEE CVPR*, pp. 770–778, 2016.

[60] F. P.-W. Lo, Y. Sun, J. Qiu, and B. P. L. Lo, "Point2volume: A vision-based dietary assessment approach using view synthesis," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 1, pp. 577–586, Jan. 2020.

[61] 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 *IEEE BSN*, May 2019, pp. 1–4.

[62] S. Fang, Z. Shao, R. Mao, C. Fu, D. A. Kerr, C. J. Boushey, E. J. Delp, and F. Zhu, "Single-view food portion estimation: Learning image-to-energy mappings using generative adversarial networks," *CoRR*, vol. abs/1802.09670, 2018.

[63] I. Woo, K. Otsmo, S. Kim, D. S. Ebert, E. J. Delp, and C. J. Boushey, "Automatic portion estimation and visual refinement in mobile dietary assessment," in *Computational Imaging VIII*, International Society for Optics and Photonics, vol. 7533, 2010, 753330.

[64] J. Chae, I. Woo, S. Kim, R. Maciejewski, F. Zhu, E. J. Delp, C. J. Boushey, and D. S. Ebert, "Volume estimation using food specific shape templates in mobile image-based dietary assessment," in *Computational Imaging IX*, International Society for Optics and Photonics, vol. 7873, 2011, 78730K.

[65] W. Jia, H.-C. Chen, Y. Yue, Z. Li, J. Fernstrom, Y. Bai, C. Li, and M. Sun, "Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera," *Public health nutrition*, vol. 17, no. 8, pp. 1671–1681, 2014.

[66] S. Fang, C. Liu, F. Zhu, E. J. Delp, and C. J. Boushey, "Single-view food portion estimation based on geometric models," in *IEEE ISM*, Dec. 2015, pp. 385–390.

[67] J. Shang, M. Duong, E. Pepin, X. Zhang, K. Sandara-Rajan, A. Mamishev, and A. Kristal, "A mobile structured light system for food volume estimation," in *IEEE ICCV Workshops*, IEEE, 2011, pp. 100–101.

[68] Y. Yue, W. Jia, and M. Sun, "Measurement of food volume based on single 2-d image without conventional camera calibration," in *IEEE EMBC*, 2012, pp. 2166–2169.

[69] P. Pouladzadeh, S. Shirmohammadi, and R. Al-Maghrabi, "Measuring calorie and nutrition from food image," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947–1956, 2014.

[70] Y. Yang, W. Jia, T. Bucher, H. Zhang, and M. Sun, "Image-based food portion size estimation using a smartphone without a fiducial marker," *Public health nutrition*, vol. 22, no. 7, pp. 1180–1192, 2019.

[71] S. Fang, Z. Shao, R. Mao, C. Fu, E. J. Delp, F. Zhu, D. A. Kerr, and C. J. Boushey, "Single-view food portion estimation: Learning image-to-energy mappings using generative adversarial networks," in *IEEE ICIP*, IEEE, 2018, pp. 251–255.

[72] E. Park, J. Yang, E. Yumer, D. Ceylan, and A. C. Berg, "Transformation-grounded image generation network for novel 3d view synthesis," *IEEE CVPR*, pp. 702–711, 2017.