Received January 1, 2020, accepted January 16, 2020, date of publication February 13, 2020, date of current version March 18, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973625

DeepFood: Food Image Analysis and Dietary Assessment via Deep Model

LANDU JIANG[1,2], (Member, IEEE), BOJIA QIU[2], XUE LIU[2], (Fellow, IEEE), CHENXI HUANG[1], AND KUNHUI LIN[1]

[1]School of Informatics, Xiamen University, Xiamen 361005, China
[2]School of Computer Science, McGill University, Montreal, QC H3A 2A7, Canada

Corresponding authors: Xue Liu (xueliu@cs.mcgill.ca), Chenxi Huang (supermonkeyxi@xmu.edu.cn),
and Kunhui Lin (linkunhuixmu@163.com)

ABSTRACT Food is essential for human life and has been the concern of many healthcare conventions. Nowadays new dietary assessment and nutrition analysis tools enable more opportunities to help people understand their daily eating habits, exploring nutrition patterns and maintain a healthy diet. In this paper, we develop a deep model based food recognition and dietary assessment system to study and analyze food items from daily meal images (e.g., captured by smartphone). Specifically, we propose a three-step algorithm to recognize multi-item (food) images by detecting candidate regions and using deep convolutional neural network (CNN) for object classification. The system first generates multiple region of proposals on input images by applying the Region Proposal Network (RPN) derived from Faster R-CNN model. It then indentifies each region of proposals by mapping them into feature maps, and classifies them into different food categories, as well as locating them in the original images. Finally, the system will analyze the nutritional ingredients based on the recognition results and generate a dietary assessment report by calculating the amount of calories, fat, carbohydrate and protein. In the evaluation, we conduct extensive experiments using two popular food image datasets - UEC-FOOD100 and UEC-FOOD256. We also generate a new type of dataset about food items based on FOOD101 with bounding. The model is evaluated through different evaluation metrics. The experimental results show that our system is able to recognize the food items accurately and generate the dietary assessment report efficiently, which will benefit the users with a clear insight of healthy dietary and guide their daily recipe to improve body health and wellness.

INDEX TERMS Food computing, health, dietary assessment, nutrition analysis, image recognition, computer vision.

I. INTRODUCTION

According to the statement from World Health Organization (WHO) [47], obesity and overweight are defined as abnormal or excessive fat accumulation that presents a risk to health. It claims that fundamental cause of such issues is an energy imbalance between calories consumed and expended. Since 2016, there were already over 1.9 billion adults overweight in the world, and the obesity epidemic has been growing steadily but not a single country has been able to reverse it so far. In the United States, in 2019, adult obesity rates now exceed 35% in nine states and 30% in 31 states, the annual medical cost of obesity-related illness healthcare including heart disease, stroke, type 2 diabetes and certain types of cancer are a staggering 190.2 billion US dollars [16], and the medical cost for people who have obesity was $1,429 higher than those of normal weight [22]. Though there are various factors may cause obesity such as certain medications, emotional issues like stress, less exercise and poor sleep quality, eating behavior - what and how people eat is always the major problem that results in weight gain.

Calories as well as other nutrition ingredients like fat, carbohydrate and protein are measures of energy [6], [59]. There are more and more people would like to keep track of what they eat and the amount of nutrition contents they get everyday to see whether they are having a healthy diet. Therefore, an accurate estimation of dietary caloric intake will be very important for well-being. Besides, the rapid development of Internet of Things (IoT) and the explosion of data enhances the social media user experience [8].

The associate editor coordinating the review of this manuscript and approving it for publication was Ying Song.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

People become willing to record, upload and share food images on the websites like Yelp [58], Dianping[1] and Yummly,[2] thus it is now more convenient than ever to find a huge amount of data (images and videos) related to food. It is even more necessary than ever to develop tools for detecting and recognizing food from images in order to promote the analysis of the nutritional ingredients from receipts and track the personal habit of eating and drinking.

Currently, there are three types of most commonly used methods [1] to manually assess dietary intake including diet records [4], 24-hour recall [41] and food frequency questionnaire (FFQ) [42]. For diet records, subjects need to record the food and beverage consumed over three consecutive days (two weekdays and one weekend day). Detailed instructions on how to record intake must be provided by trained staff and the completed records need to be entered into a application such as Nutrition Data System for Research (NDSR) for analysis. By applying 24-hour recall, subjects are asked to report all food/meals consumed in the past 24 hours, which can be done via telephone call or face-to-face interview. The data from subjects are required to be collected and analyzed, an interview for details will be conducted by trained staff [57]. Subjects using FFQ method are asked to report how frequently certain food and beverage items were consumed over a specific period of time (e.g., 1 year). Most FFQs are available in paper or electronic format listing general questions about everyday diet and cooking practice. Software programs are deployed to calculate nutrient intake by multiplying the reported frequency of each food by the amount of nutrient in each food item [62].

Although we already have these gold-standard methods for reporting diet information, at least one drawback exists that we can not ignore - such methods still suffer from bias since the subject is required to estimate their dietary intake by themselves. Dietary assessment finished by participants can result in underreporting and underestimating of food intake [48]. In order to get rid of the bias and improve the accuracy of self-report, many automatic or semi-automatic eating monitoring systems have been proposed. Additionally, recently there are an increasing number of applications built on mobile platforms (i.e., smartphones) for food analysis. For example, Zhu *et al.* [65] proposed a segmentation based food classification method for dietary assessment. They aim to determine the regions where a particular food is located in an image where a particular food is located and correctly identify them by using computer vision techniques. Another cloud-based food calorie measurement system is developed by Pouladzadeh *et al.* [49] using Support Vector Machine (SVM) to recognize food and calculate the calories of each item.

On the other hand, with the rapid development of artificial intelligence and deep learning algorithm, it is more accurate and efficient to analyze food image using deep models. The work from Pouladzadeh *et al.* [50] describe an assistive calorie measurement system for patients and doctors. The system achieve good recognition accuracy on single food portions by leveraging deep convolutional neural networks. It is able to record and measure user daily calorie intake in an convenient and intelligent way. In this paper, we present an automatic food recognition and nutrition analysis system. With just one single photo taken by mobile devices, our system can recognize the food items and analyze the nutrition contents from the meal, as well as providing the user with a thorough dietary assessment report on what they have. With this automatic system, it will help the user to measure their daily intake and enable them in a long-term way to ensure their health. Our contribution is summarized as follows:

- In this paper, we explore a deep model based approach for food recognition and dietary assessment. Specifically, we design and implement a system for food image analysis - output the amount of nutritional ingredients of each food items from daily captured images. A thorough dietary assessment report will be generated based on what you have during the meal.
- We leverage deep convolutional neural network models (e.g. Faster R-CNN) for food detection and identification. Firstly, we use region proposal network to generate thousands of region proposals from the input image. Then a state-of-the-art deep convolutional neural network (i.e., VGG-16) to extract the feature maps from each proposals, and classify them as different food items. In order to reduce the processing time, we also apply a regression module to locate each food item in the image.
- We generate/revise a new food related dataset, named FOOD20-with-bbx, based on the existing dataset FOOD101 [11]. Since the original dataset (FOOD101) only contains image category label. We select 20 most common categories from the dataset, and manually label each food items in the image with bounding box information. By adding this information, our dataset can be used in the evaluation of other detection models.
- We conduct extensive experiments to test our food detection model and nutrition analysis system. We use UEC-FOOD100 [38] and UEC-FOOD256 [32] datasets with bounding box information for detection evaluation. We compare the performance of different models by using the mean Accuracy Precision mAP) and the detection delay/speed. We also trained our model with our new dataset FOOD20 with bounding box to test the robustness of our model in scenarios with a larger range of food categories.

The rest of the paper is organized as follows. In section II, we present a thorough literature review on different approaches towards food detection and nutrition analysis. Several state-of-the-art food recognition and diet analysis systems are introduced. In section III, we discuss the project motivation and design challenges of proposed system. In section IV, we elaborate the whole structure of food

[1] http://www.dianping.com/

[2] https://www.yummly.com/

analysis system, as well as detailed design and implementation of deep learning modules. In section V, we evaluate our proposed system by using different types of image datasets. In section VI, we conclude this paper and discuss possible future works.

II. RELATED WORK

A. FOOD RECOGNITION BASED ON GEOMETRY FEATURES

Food category recognition and analysis has been a popular research area in the field of nutrition study. However, it is relative difficult because food items are deformable objects with significant variations in appearance. The food items may either have a high intra-class variance (similar foods such as beef and steak look very different based on how to cook them), or low inter-class variance (different foods like fish and pork look very similar).

Different approaches have been proposed to recognize food items in image using geometry features such as SIFT [37] descriptor, color histograms or GIST [46], and shape context [9]. Moreover, Felzenszwalb [21] use triangulated polygons to represent a deformable shape for detection and Jiang *et al.* [30] proposes learning a mean shape of the target class based on the thin plate spline parametrization. Besides, Belongie [9] choose n pixels from the contours of a shape and then form $n - 1$ vectors as a description of the shape at the pixel level. Though geometry feature-based approaches work well in object detection for the certain types of items, there are two main problems for food related tasks. The first problem is that geometry feature-based methods need to detect features like edges, counters and key points or landmarks, which may not be available in food images. The other problem is that it is hard to describe the shape of a food item in real world thus calculating shape similarity is very hard.

B. FOOD RECOGNITION BASED ON STATISTICAL FEATURES METHODS

To overcome the problems described above, approaches using statistical features are proposed. Instead of edge or keypoints, the methods focus on local, statistical features like pairs of pixels. Since the statistical distribution of pairwise local features could extract important shape characteristics and spatial relationships between food ingredients, thus facilitating more accurate results in object recognition.

For example, Yang *et al.* [64] explore the spatial relationships between different ingredients (e.g., vegetables and meat in one meal) by employing a multi-steps discriminative classifier. Each pixel in the image is assigned a vector indicating the probability of the pixel belongs to nine food ingredients [54]. A multi-dimensional histogram is generated by using pairwise statistic local features, then the histogram is passed into a multi-class SVM for image classification.

C. FOOD RECOGNITION BASED ON MACHINE LEARNING METHODS

Recently, there has been an increasing number of research conducting experiments and researches toward the fields of food classification, leveraging machine learning/deep learning algorithms.

Aizawa *et al.* [5] proposed a Bayesian framework based approach to facilitates incremental learning for both food detection and food-balance estimation. Bossard *et al.* [11] used Random Forest on the Food-101 test set achieving a classification accuracy with 50.67% by mining discriminative components. The random forest model is used for clustering the superpixels of the training dataset [13]. Other advanced classification techniques were also applied in the work including Improved Fisher Vectors (IFV) [52], Bag-of-Words Histogram (BOW) [36], Randomized Clustering Forests (RCF) [40] and Mid-Level Discriminative Superpixels (MLDS) [56].

As the computational power is getting stronger, convolutional neural network (CNN) and deeper models are also widely used in food recognition and provide better performance. Kagaya *et al.* [31] applied the CNN model in food image classification. They achieved a very high accuracy of 93.8% on food and non-food item detection. The experimental results on food recognition showed that the proposed CNN solution outperformed all other baseline methods - achieved an average accuracy of 73.7% for 10 classes. What is more, a fine-tuned the AlexNet model is used in the work [63]. The method achieved the promising results on public food image datasets so far, with top-1 accuracy of 67.7% for UEC-FOOD-256. Hassannejad *et al.* [26] apply a 54 layers CNN model to evaluate the effectiveness of deep model in food image classification. The model is based on the specifications of Google's image recognition architecture - Inception. In addition, GoogLeNet [60] was used in [39] for food recognition to build a Im2Calories system on Food-101 dataset.

Additionally, researchers start to investigate which features and models are more suitable for the food recognition, and comply them into food analysis system to calculate the calories [7], [23], [33], [34]. In order to automatically estimates the food calories from a food image, multi-task convolutional neural networks is used for simultaneous learning of food calories, categories, ingredients [18]. What's more, a generative adversarial network approach is also proposed for food image analysis [20].

Though food recognition and nutrition contents analysis have been well discussed by above work, two basic challenges remain. Firstly, most of the approaches are dealing with image with single food item. Secondly, it is still time consuming (2 seconds in general) to detect and classify the food in images. In this paper, we aim to address these issues and propose an automatic food recognition system to identify the food from images and generate dietary assessment reports for long-term healthcare plan.

III. DESIGN CONSIDERATION

In this section, we show our design consideration of the food recognition and dietary assessment system. We first introduce our motivation and goals, then explore main technical challenges to build the system.

A. *MOTIVATION*

1) WHY WE NEED AN AUTOMATIC SYSTEM

With the rapid development of smart computing and Internet of Things (IoT), now we have a huge amount of data from social networks and mobile networks everyday. People keep uploading, sharing and recording what they do everyday in case of missing the chance of using them to improve our daily life. Food images, recipes and food diaries become the most popular information to be shared, we can learn the implication to build an automatic nutrition analysi system by taking the advantage of such large-scale datasets. With the help of food recognition and analysis systems, users are able to record their daily meals and assess dietary habits, as well as promote their health.

2) FOOD IDENTIFICATION IS NOT EASY

The computer vision methods make the process of food analysis more reliable and accurate. But there is still a challenging issue needed to be addressed - how to recognize the different types of food in one plate/image correctly. Though there are a great number of image recognition tools available, the methods for food identification are still largely relied on self-reported dietary intakes. It is mainly because that food items are typically deformable compared to other objects in real world. It is hard to define the structure of a food item, and a high intra-class (similar foods look very different) and low inter-class (different foods look very similar) variance also exist. An example of low inter-class variance is shown in Figure 1 (Source: https://www.cupcakeproject.com/cupcake-vs-muffin-update/).

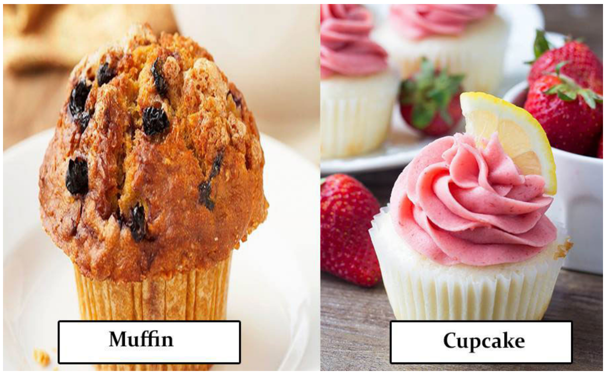


FIGURE 1. Example of low inter-class variance. Muffin and Cupcake are two different types of food, but they look similar on appearance (e.g., shape features).

3) CLASSIFICATION VS DETECTION

How to define the right goal - image classification or object detection is very important, especially for food recognition and nutrition analysis scenarios in our project. We will need object detection if we aim to identify the objects in a food image, for example, count the number of apples in the image. An image can contain different categories of foods, and detection techniques enable us to recognize what kinds of foods we have in the meal. Based on such detection results, we are able to calculate the calories and analyze the food nutirion.

B. *CHALLENGES*

Although the idea of our project sounds simple, many challenges arise in practise. Three main challenges in real food image recognition and analysis are addressed as follows:

1) REGION OF INTEREST

In this project, we need to recognize multiple food items from a single image. As illustrated in the Figure 2 [38], an image for one real meal may contain multiple types of food items. Therefore, a identification scheme that detects and recognizes multiple food items from a single image is desirable. Nowadays, most of food analysis applications are extracting the visual features of different foods from the entire image. Since the background may contain non-food objects with the similar shape or color features as food. Such methods include the background information/features would inevitably mislead the detector that affect the detection accuracy. Region based deep learning based models have been applied and achieved a great success in object detection. The basic idea of these solutions is to use different kinds of region proposal methods to generate the regions of interests (RoIs). These RoIs will dissociate the target objects from the background, making it much easier to extract the features for recognition. However, the many state-of-the-art models that deploied in popular competitions like ImageNet, Large Scale Visual Recognition Challenge (ILSVRC) and MS-COCO have not yet been widely examined for food image datasets.

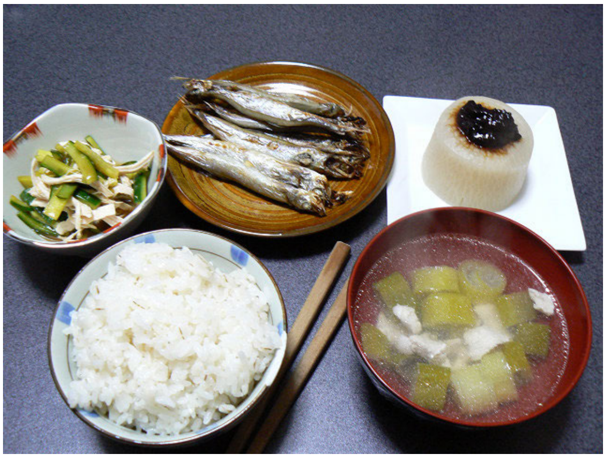


FIGURE 2. Example of the food image to analyze.

2) THE DELAY OF FOOD RECOGNITION

The second challenge in this project is to reduce the processing time for food detection and classification.

Obejct detection algorithms and recognition tasks are always time-consuming. Basically, the more objects in one image, the more time the classification system may take. It is very important to control the processing time for a delay sensitive application. Because users would not like to wait one or two minute to know what is the amount of caloric he/she just ate. Thus it is required to design a system that processes all the images with different numbers of objects (food items) at the acceptable speed, and cost less time.

3) INSUFFICIENT INFORMATION OF NUTRITION CONTENT FOR DIETARY ASSESSMENT

As a dietary assessment system, how to identify a large number of food class and accurately analyze nutrition contents is very challenging. There are a number of smartphone applications for food analysis such as MyFitnessPal [2] and SHealth [3] to help users to record their food intake. Some of the systems enable users to use smartphone cameras to take pictures of the food for recognition. Such work could assist users to achieve dietary goals such as weight loss, allergy management or maintaining a healthy diet. There is still not enough information to have a comprehensive study on what we eat - the number of food classes (collected/recored) is limited. Though crowd sourcing [44] may become a solution, the model itself cost significant power and thus inhibits it from widespread deployment.

C. DIETARY ASSESSMENT AND NUTRITION ANALYSIS

According to a study [19] launched by The International Life Sciences Institute (ILSI), 43 new technology-based dietary assessment tools from 2011 to 2017 are evaluated, including web-based programs, mobile applications as well as wearable devices. The results show that most of the tools (79%) relied on self-reported dietary intakes. While 91% of them used text entry and 33% used digital images to help identify foods. Only 65% of the tools had integrated databases for estimating energy or nutrients, and less than 50% contained features of customization as well as generated automatic reports.

Due to the limitations of the self-report methods in web-based and mobile-based tools, leveraging food image captured by smartphone cameras has been proved as an efficient way to record the daily meals and analyze the components. Image or computer vision based methods for dietary assessment refers to any model that uses images/videos of eating episodes to enhance self-report schemes in previous work or as the primary way to record dietary intake. Such systems are widely tested and validated. Basically, features from the images are extracted for food recognition, and the program check the foods with a searchable image. The food types and portion sizes are then matched to the databse of food and nutrient for dietary studies.

Based on the record, the image-based methods could make the food analysis progress more reliable and accurate. In this paper, we aim to provide an automatic food recognition and dietary assessmentsystem that takes a photo of food items as input. And output the amount of nutri-tional ingredients of each food items from the image. We will address the issues discussed above to improve the accuracy and speed of food recognition.

IV. SYSTEM DESIGN

In this section, we first give an overview of our proposed deep model based system for food recognition and nutrition analysis. Then we discuss the implementation details of each component of the system. For more background information and technical details about our food analysis system, we refer readers to the thesis on this project [10].[3]

A. SYSTEM OVERVIEW

In this paper, we propose a deep learning based system for food item detection and analyze the nutrition components of each meal image. As shown in **Figure 3**, our model consists of three main steps.

- We first extract the regions of interests (ROIs) by applying the Region Proposal Network derived from the Faster R-CNN model. The RoIs would help to separate the food items from the background, and improve the detection model efficiency.

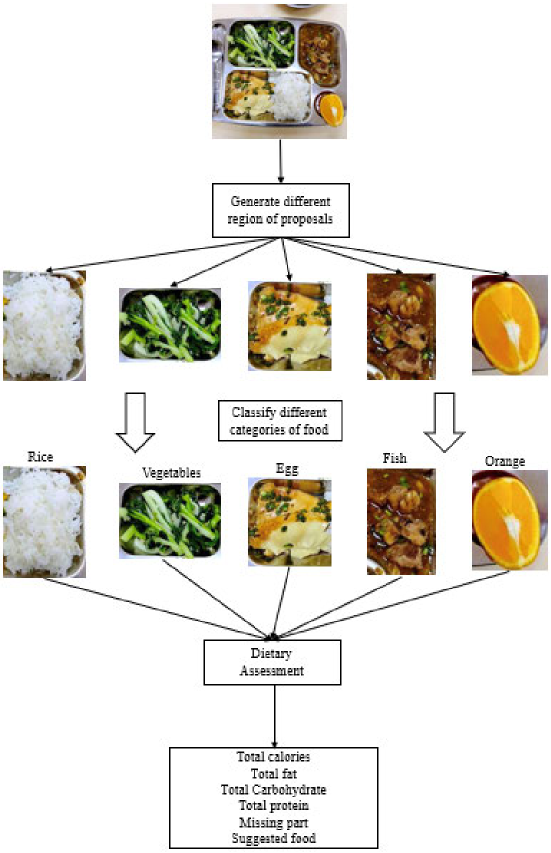


FIGURE 3. The automatic three steps system of food recognition and nutrition analysis system.

[3] Parts of this work have been used in partial fulfillment of the requirements for Bojia Qiu's (co-author of this paper) master thesis at McGill University.

- The second step is to apply a well designed Convolutional Neural Network (CNN) on selected RoIs, and classify them into different food item categories. Meanwhile, an regression module is also used to locate the food coordinates in the image.
- The final step is to use modern technology-based dietary assessment tools for food nutrition analysis, and generate a health report for users based on their meal images.

B. REGION BASED DETECTION

As discussed in section III-B, region based object detection approaches possess leading accuracy on object recognition. As how it is defined, it proposes different regions from the input image, and classify them into different categories. The traditional region-based object objection methods use a sliding window go through the image. It will make the whole proccess extremely slow especially when the deep CNN models are employed.

Early region-based detection CNN (e.g., R-CNN [25] and Fast R-CNN [24]) rely on the input generic region proposals such as selective search [61] and EdgeBox [17], etc. Such hand-crafted process is time consuming due to computational burden of proposal generation. To address this issue, Ren *et al.* [51] found a way to make the region proposal more efficient, called Faster R-CNN. Faster R-CNN has 4 main parts including the feature extraction (a basic convolutional module with convolution layer, relu activation function and pooling layer), region proposal (anchors classified as foreground region or background region), bounding box regression (fix the anchors location) and classification.

In this paper, we apply Faster R-CNN model to detect the food items from the images. This section only briefy introduces the key aspects of the Faster R-CNN, for more technical details we refer readers to the original paper [51].

C. FOOD ITEM DETECTION

As discussed, in order to extract food items from the background in images, we apply Faster R-CNN to detect the food-related regions. Since we only focus on food objects, the feature maps will be less complex and rubost to the noise in the background. However, the model of Faster R-CNN we used was trained with VOC2007 database, which only contains 20 common non-food object types. Therefore we first select a food image database - FOOD100 [38], to get a fine-tune the pre-trained Faster R-CNN model. Then we use the model to extract the RoIs on each food image. The loss function is shown in equation 1.

$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*) \quad (1)$$

where i is the index of the anchor in each batch, p_i is the predicted probability of anchor i. If the anchor is foreground, p_i^* equals to 1, otherwise p_i^* equals to 0. t_i represents the four coordinates of the predicted bounding box. t_i^* is the matching ground-truth box for the foreground anchor.

During the test, we examined several different deep neural network architectures to replace the CNN part in the Faster R-CNN for performance comparisons. For example, we tried 1. the AlexNet [35] architecture with 7 layers to jointly detect and classify objects, 2. VGG-16 [55] network for object recognition developed and trained by Oxford's renowned Visual Geometry Group, as well as the Residual neural network (ResNet) [28], which is builds on constructs known from pyramidal cells in the cerebral cortex.

1) IMPLEMENTATION DETAILS

In the implementation, we first resize the image ($s = 600$ pixels and $l = 1000$ pixels) to make the VGG16 model fit in GPU memory during fine-tuning [24]. We also set 3 different scales for the anchors in the detection - 128^2, 256^2 and 512^2 pixels, and the aspect ratios are $1:1$, $1:2$ and $2:1$. It results in 9 different anchors for each point in the feature map.

In order to control the total number of anchors in one image to make the model more easy to converge, we remove the cross-boundary anchors so that they do not contribute to the loss calculation in the training process. After the filtering, about 6000 anchors left in one image for training. But we still apply the fully convolutional RPN to the entire image during testing period. Though we shrink the nubmer of anchors to 6000, some of them are still highly overlapped with each other. We adopt non-maximum suppression (NMS) [43] on the proposal regions based on their classification score to further reduce the redundancy. NMS algorithm scans the image along with the image gradient direction to eliminate points (set to zero) that do not lie in important edges (not the local maxima). After NMS, one image may have around 2000 proposal regions for detection.

In order to share the convolutional layers, we use 4-step Alternating Training strategy in [51] to train the model instead of learning two networks for RPN and the detection module. Specifically, we first train the RPN module by leveraging a pre-trained model. For each image, a sample of 256 anchors are randomly selected for the loss function of a mini-batch. Secondly, we train the detection model with the proposal anchors generated from RPN. Two models/networks do not share convolutional layers at this stage. Thirdly, we use dection model to initialize RPN training with fixed shared convolutional layers. Only the layers belong to RPN is tuned. At fourth step, we do similar training to detection model using RPN, thus these two models share the same convolutional layers.

2) DATA AUGMENTATION

In addition, we also use data augmentation to enrich the training samples to enhance the system performance. We first flip images horizontally but not vertically. Since most of food item in real world will not be flip vertically. We then apply image rotation on the data with 90 degree, 180 degree and 270 degree.

D. FOOD ITEM CLASSIFICATION

1) OBJECT CLASSIFICATION

The classification module uses the proposal feature maps computed for classification, and calculate the score for each class. At the same time, it apply bounding box regression again to enhance the accuracy for proposal region localization.

In this paper, we use VGGNet as the CNN model, to extract the feature map of the proposed regions and perform classification for food recognition. The maximum number of food item in UECFOOD100 database is 5, so we set highest score of bounding boxes regions as 5. In the classification module, we use VGG-16 model, which contains 16 weight layers in the network with fully connected layers. 4096 dimensional feature vector will be used to accurately classify the object into categories. Since there 100 food categories selected in UEC-FOOD100, the softmax layer contain 100 units. We will also change the structure of the layer corresponding to different datasets with different number of food categories respectively.

Additionally, all deep models used in this work will be pre-trained with the natural image datasets and then trained using food image datasets we proposed. The image samples were randomly split into a training and validation sets.

2) BOUNDING BOX REGRESSION

For bounding box regression, it is because we may not be able to propose a region that perfectly cover the object. Thus we need to tune the initial bounding box to make its coordinates match the ground truth.

As shown in **Figure 4**, the green box G represents the ground truth location, while the original foreground anchors is the blue box P. The bounding box regression is to find a transformation that cast the P to G', which is closer to the ground truth bounding box G. A simple way to achieve this is firstly to use translation method in equation 2 and 3

$$G'_x = A_w \cdot d_x(A) + A_x \tag{2}$$

$$G'_y = A_h \cdot d_y(A) + A_y \tag{3}$$

And then zoom in or out based on equation 4 and 5

$$G'_w = A_w \cdot \exp(d_w(A)) \tag{4}$$

$$G'_h = A_h \cdot \exp(d_h(A)) \tag{5}$$

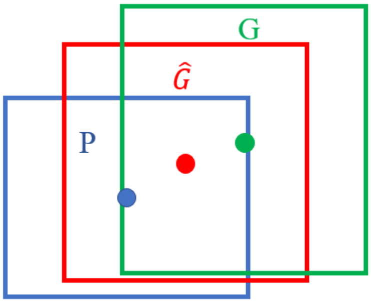


FIGURE 4. Bounding box regression [15].

In this paper, we assume that all above operations are linear. Thus we can learn $d_x(A)$, $d_y(A)$, $d_w(A)$, $d_h(A)$ by using linear regression to fine tune the bounding box coordinates. The linear regression is used to learn weights W for the input feature vector A to match the ground truth t. Where $t = W \cdot A$. The loss is defined as shown in equation 6

$$Loss = \sum_{i}^{N} (t^i - W^T \cdot A^i)^2 \tag{6}$$

E. DIETARY ASSESSMENT

After food item recognition, the system should be able to perform dietary assessment that analyze the nutrition of the meal. In this paper, we mainly focus on calculating calories, fats, carbohydrates and proteins contents from each meal image. Moreover, vegetables, nuts and whole grain food are also healthy food to be recommended.

In real world scenario, the system should ask user basic information about theri body like age, gender, weight, height and activity level, etc. Based on the profile, we can estimate the right amount of food and nutrition they need to keep a healthy diet. Users can simply use their smartphones to take a picture of what they eat in the meal (of course before eat them up). The system will recognize the food in the image and then estimate the nutrition status of the meal. A diet log will be create for easy tracking and remind users to calculate the number they left.

How to determine whether the user are having enough energy and specific kinds of food? We utilize the standard source for dietary assessment - the USDA National Nutrient Database (NNDB) [45], which contains nutritional facts about 8618 basic foods. We can build a reference table of nutrition facts tables based on the data from USDA including all the food items in our datasets. And then calculate the amount of nutritional ingredients for the food items captured by users by mapping the detected food to the reference table.

Table 1 shows an example of our food nutrition reference table. For each food item, we assume the weight is 400 grams,

TABLE 1. Snippet of the reference nutrition facts table. Each row contains the value of calories, fat, carbohydrate and protein.

Food(400g)	Calories	Fat(g)	Carbohydrate(g)	Protein(g)
Steak	1365	63	0	187.3
Ramen	760	29	78	35
Miso Soup	81	3.3	9.8	6.5
Fried Rice	619	12.8	106.8	12.8
Sushi	536	7.7	103.3	13.4
French Fries	428	21.4	57.1	4.8
Takoyaki	1264	92.8	67.2	38.4
Pizza	690	20.4	103.9	26
Hamburger	1086	73.7	0	99.1
		...		

which is a normal size for a single serve. Each row in the reference table presents the amount of calories, fat, carbohydrate and protein of the food. There are 100 food categories in the UEC-FOOD100 dataset, thus the table will have 100 rows. Based on the user profile information, we are able to provide a diet calculator. For example, a 24 year-old man, with weight of 60 kilogram and height of 170 centimeter, with an activity level of moderate, a suitable amount of nutritional ingredients for a healthy diet will be 2399 calories per day, 311 grams of carbohydrate, 109 grams of protein and 80 gram of fat.

V. EVALUATION

In this section, we evaluate our proposed models and methodologies, which consists of three parts: data preprocessing, model training and performance evaluation.

A. DATASETS

In this paper, we use two real-world datasets: UEC-FOOD100 [38], UEC-FOOD256 [32] to evaluate the proposed models. UEC-FOOD100 contains 12740 images in total of 100 categories of different food items while UEC-FOOD256 contains 31395 food images within 256 categories. Most of food images in UEC-FOOD100 are japanese food. There are more than 100 images of each food category in both datasets, with the bounding box information indicating food location in each photo. The basic information of UEC-FOOD100 dataset is summarized in **Table 2**.

TABLE 2. UEC-FOOD100 Dataset statistics.

Total number of categories	100
Total number of images	12740
Number of images with single food item	11566
Number of images with multiple food items	1174
Category with largest amount of images	Miso Soup (729)
Category with least amount of images	Chicken Rice (101)

Though both UEC-FOOD100 and UEC-FOOD256 datasets contain a large number of food images, most of samples are asian food that may not be used to train a more generable model for all different food images. In order to build a robust system that works for a wider range of diet categories, we choose another very commonly used dataset - FOOD101 [11], which contains 101 food categories, with 101000 images. However, the FOOD101 dataset does not come with bounding box information in images. We are not able to train the proposed model using samples in FOOD101. Thus we manually generate bouding box for a portion of images in FOOD101 dataset - at least 300 photos within each selected class. We named subsets from FOOD101 image dataset with bounding box as FOOD20-with-bbx. Several image examples taken from FOOD20-with-bbx are shown in Figure 5.

FIGURE 5. Some examples of food items in the FOOD20-with-bbx Dataset. This dataset includes only single food item.

B. EVALUATION METRICS

In this paper, we use mean Average Precision (mAP) to evaluate our detection model. As in food item detection, evaluation is not as simple as other models, because there are two distinct tasks to measure at the same time.

- **Classification** - Determining whether an object exists in the image.
- **Regression** - Determining the location of the object.

In addition, there are many food classes and the distribution is unbalanced. For example, the number of images in the rice category in UEC-FOOD100 and UEC-FOOD256, is four to five times higher than others. A simple metric measuring accuracy may introduce biases. Thus, a "confidence score" associate with each bounding box could help to assess the model.

1) CLASSIFICATION EVALUATION

Average precision (AP) is a metric that frequently used in measuring the accuracy of object detectors [29]. It computes the average precision value for recall value over 0 to 1.

For each class, AP (7) is basically the area under the precision-recall curve, while the precision-recall curve is computed from the model's detection output. For simplicity, we divide the recall values from 0 to 1.0 into 11 points - 0, 0.1, 0.2, …, 1.0, and calculate the maximum precision measured at each set (8).

$$AP = \int_0^1 Precision(Recall)d(Recall) \tag{7}$$

$$AP = \frac{1}{11}\sum_{Recall_i} Precision(Recall_i) \tag{8}$$

As precision and recall are always between 0 and 1. Therefore AP falls within 0 and 1.

C. REGRESSION EVALUATION

In order to evaluate the how well the regression model estimates the location of the object, we use the intersection over

Union (IoU) metric to measure the overlap between two boundaries [29]. As illustrated in 9, we compute the *area of overlap* between the predicted bounding box and the ground truth (the real object boundary). And the *area of union* is the area encompassed by two boundaries.

$$IoU = \frac{Area\ of\ Overlap}{Area\ of\ Union} \tag{9}$$

A predefined IoU threshold is used in the test to determine whether the object detection is positive or negative. The IoU thresholds may be vary for different tasks, and we set IoU threshold as 0.5 in this project. Finally, the mean Average Precision (mAP) score can be calculated by taking the mean AP over all object classes with the IoU threshold 0.5.

D. EXPERIMENT SETUP

1) EXPERIMENT PLATFORM

We implement our deep model using Keras [14], which is a high-level neural network API written in python. In the backend, we use TensorFlow platform [12] designed by Google.

E. EXPERIMENT RESULTS

1) BASELINES AND EXPERIMENT SETTINGS

We select two baseline models for food detection - R-CNN [25] and a CNN-based food image segmentation model proposed by Shimoda er al. [53]. We follow the environment settings as Shimoda's work to provide a reasonable comparison. We split the dataset UEC-FOOD100 with 80% as training data and the rest 20% as testing set with the momentum of 0.9 and weight decay rate of 0.0005.

2) DETECTION RESULTS

a: UEC-FOOD100 DATASET

Due to the limitations of our computing resource, we can not match the same setting - 40000 iterations as the baseline model [53] did. Thus the training loss for first 239 ephochs in training is shown in Figure 6

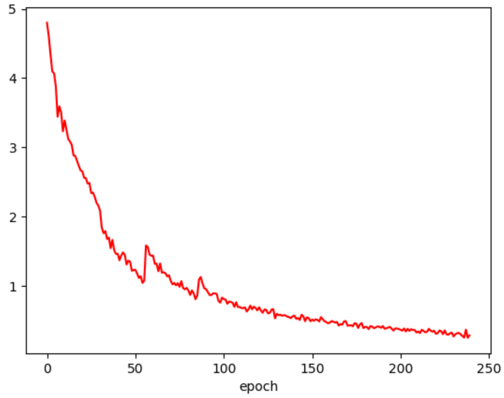


FIGURE 6. Training loss for the 239 epochs during training step for UEC-FOOD100 dataset.

We calculate the mAP of 100 food classes in the dataset. However, in order to compare with the baseline model [53], we select 53 categories of food that have been tested. Each category contain at least 10 images, and over 11 categories contain more than 50 imges. The detection result for UEC-FOOD100 are shown in Table 3. We acquire baseline model results from the original paper/report [53]. We can see from Table 3, even though we have much less iterations in training session, our proposed model still outperform the R-CNN model in certain food categories. The set that get better results (Set 3 compare to Set 1) is because of more sample images in the dataset, thus with a large dataset size with more training data, our model could get much higher accuracy.

TABLE 3. The results in UEC-FOOD100 dataset. Set 1 is the experiment that used all food classes. Set 2 is the experiment that test 53 classes (more than 10 items in test dataset). Set 3 is the experiment that test with 11 classes (more than 50 items in test dataset).

mAP(%)	Set 1	Set 2	Set 3
RCNN-based model	26.0	21.8	25.7
BP-based model	49.9	55.3	55.4
Our model	17.5	23.1	25.5

The main reason that our model results in low mAP in these experiments is the limitation of computing resource, since we made a trade off between training time (number of iterations) and the system performance. Basically, other models may take more than 20000 iterations to train a Faster R-CNN, while we only perform 250 - 400 iterations. Therefore, we are positive on our proposed solution. The model could achieve higher accuracy with a better localization results from RoI method. For example, the classification accuracy could reach to 98% in the training when the region proposal network could generate correct bounding boxes. We believe that with more powerful computational resource, we can achieve a much better result with a much higher interation number. Several examples of successfully detected items are shown in Figure 7. In addition, we also test the speed of our model. Since the processing time depends on both implementation design and hardware used, we only present a general idea that our proposed model outperformed the other two baseline approaches. The R-CNN has the highest delay as it generates the region proposals using selective search [61] algorithm, which consumes a considerable amount of time. While Shimoda's model is much faster in classification, however, they still use the selective search that may have a bottleneck.

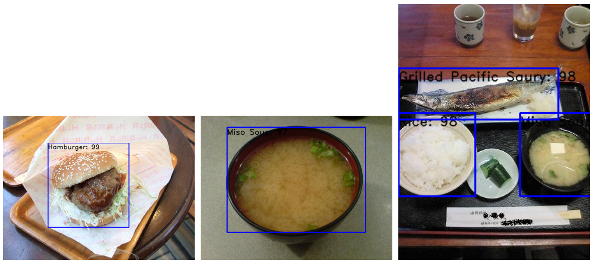


FIGURE 7. Examples of the detection result from UEC-FOOD100.

In this paper, we use region proposal network (RPN) [51] instead of selective search algorithm and simultaneously train the model for classification.

b: UEC-FOOD256 DATASET

We also use UEC-FOOD256 dataset (contains 256 categories of food items) to examine the performance of our proposed model. As shown in Table 4, there are 132 categories with more than 10 items and 21 categories with more than 50 food items in UEC-FOOD256 dataset. We can see from the table that the mAP values for all three sets are worse compared to UEC-FOOD100 dataset. It may caused by the increase of the number of food class. In addition, there are many food categories have similar features (e.g., shape and color) such as *miso soup* and *beef miso soup*, which make it even harder to distinguish.

TABLE 4. **The results in UEC-FOOD256 dataset. Category 1 is the experiment that test with all the classes of food. Category 2 is the experiment that test with 132 classes (more than 10 items in test dataset). Category 3 is the experiment that test with 21 classes (more than 50 items in test dataset).**

mAP(%)	Set 1	Set 2	Set 3
Our model	10.5	13.3	18.3

c: FOOD20-WITH-BBX DATASET

In addition, we also examine our model using our self-modified dataset, FOOD20-with-bbx. The FOOD20 [11] data contains a large number of western food items images which could be used to test the robustness of our proposed model. In FOOD20-with-bbx, we selected 20 categories of food bounded by the bounding boxes with coordinates. To save time, we use the model after pretrained by UEC-FOOD100 dataset. 80% of image samples in FOOD20-with-bbx were used in training session, and the rest 20% of the data were used to examine the mAP values. The experiment results are shown in Table 5. With 500 iterations, our proposed model could achieve the top-1 accuracy in FOOD20-with-bbx dataset as 71.7% and top-5 accuracy comes to 93.1%.

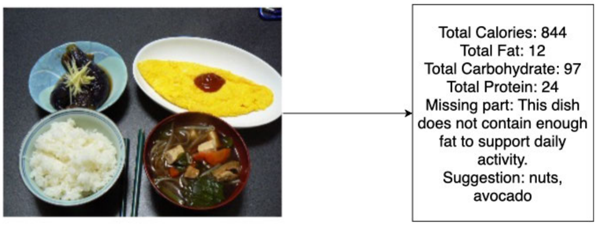
TABLE 5. **The results in FOOD20-with-bbx.**

# of iteration	Top-1 accuracy (%)	Top-5 accuracy (%)
100	28.8	61.1
250	43.6	81.2
500	71.7	93.1

The food recognition results in FOOD20-with-bbx are relatvie better than the other two datasets. It is mainly because that most of images in dataset FOOD20-with-bbx only contains one single food items, and the majority of the food is western style, which make them easier to be distinguished based on shape and texture features.

F. DIETARY ASSESSMENT

In order to promote the healthy diet, the final step of the proposed system is to analyze the nutrient contents of each food image. As discussed in previous section, for each food items detected in the image, we assume basic weight is 400 grams per item, which is a normal size for a single serve. **Figure 8** is an example of dietary assessment in our system. Based on the information collected from the image, we could summarise the dietary assessment report for user everyday.



Total Calories: 844
Total Fat: 12
Total Carbohydrate: 97
Total Protein: 24
Missing part: This dish does not contain enough fat to support daily activity.
Suggestion: nuts, avocado

FIGURE 8. **Example of result from dietary assessment system.**

G. DISCUSSION

Though our proposed model do perform well on different types of datasets, there is still room for the improvement compared with some of the state-of-the-art models.

1) MODEL COMPLEXITY

Model complexity is always major factor that affects the performance of the deep learning model. We need to make a trade off between processing time and the system accuracy due to hardware limitations. In the experiments, we use a VGG-16 neural network for feature extraction and item classification. Due to the architecture (fully-connected nodes) of VGG with 16 layers, the training process could take considerable time - several days for no more than 300 epochs to complete the training session. We may try different models in the future like deep residual networks (ResNet) [28]. "ResNet has lower time complexity than VGG-16/19" is claimed in Kaiming He's presentation [27]. Since the ResNet is constructed by several building blocks. Due to the usage of global average pooling layers rather than fully-connected layers, and the model size becomes substantially smaller. For example, ResNet with 50 layers is 102MB while our proposed model VGG-16 in the test is over 533MB.

2) DATA COLLECTION

The second challenge is to find/generate a good dataset that we can used to capture food images from daily meals. As the problem we met in the evaluation, though we have two popular image datasets UEC-FOOD100 and UEC-FOOD256, most of images in two datasets are Japanese/Asain food items. While FOOD101 dataset has western style food images but no bounding box information. In addtion, a number of food items may have high intra-class variance or low inter-class variance. Items in the same category that have high intra-class variance may look different and the two different types of food with low inter-class variance may have similar appearance. Both high intra-class variance and low inter-class variance issues can significantly affect the accuracy of detection models. To address this issue, we need to search more datasets like FOOD101 [11], to create more FOOD20-with-bbx datasets. In the future, we will continue labeling our FOOD20-with-bbx dataset, to expand this dataset to a larger range of food categories. The combination with other datasets to create a more diverse food dataset is desirable.

VI. CONCLUSION AND FUTURE WORK

In this paper, we explore the food recognition and dietary assessment problem by leveraging deep learning techniques. In particular, to have a better understanding of obejct detection and nutrition analysis, we apply state-of-the-art Faster R-CNN model to generate RoIs and use deep neural network to extract the feature map for food item recognition. We analyze the nutrition of detected food and summarize the report of the meal based on modern technology-based dietary assessment tools. We conduct extensive experiments to evaluate the efficiency and effectiveness of our system. Results show that our proposed solution achieved comparable performance and has great potential to promote healthy dietary and feasible advice.

In the future, we would continue our work on improving our detection system accuracy and reducing processing time. A more comprehensive food analysis scheme such as weight prediction is desirable. In addition, to provide a healthy diet, an automatic diet calculator is in the plan.

REFERENCES

[1] Colorado Clinical and Translational Sciences Institute. (2014). *Most Common Dietary Assessment Methods*. [Online]. Available: https://cctsi.cuanschutz.edu/

[2] *MyFienessPal*, 2018.

[3] *Samsung Health*, 2018.

[4] *Nutrition Data System for Research*, 2019.

[5] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, "Food balance estimation by using personal dietary tendencies in a multimedia food log," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 2176–2185, Dec. 2013.

[6] D. Albanes, "Total calories, body weight, and tumor incidence in mice," *Cancer Res.*, vol. 47, no. 8, pp. 1887–1892, 1987.

[7] S. Ao and C. X. Ling, "Adapting new categories for food recognition with deep representation," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 1196–1203.

[8] K. Ashton, "That 'Internet of Things' thing," *RFID J.*, vol. 22, no. 7, pp. 97–114, 2009.

[9] S. Belongie, J. Malik, and J. Puzicha, "Shape context: A new descriptor for shape matching and object recognition," in *Proc. Adv. Neural Inf. Process. Syst.*, 2001, pp. 831–837.

[10] Q. Bojia, *Food Recognition and Nutrition Analysis Using Deep CNNs*. Montreal, QC, Canada: McGill Univ., 2019.

[11] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 446–461.

[12] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Design Implement. (OSDI)*, 2016, pp. 265–283.

[13] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[14] F. Chollet. (2015). *Keras*. [Online]. Available: https://keras.io

[15] *Bounding Box*, CNblogs, 2019.

[16] C. M. Hales, M. D. Carroll, C. D. Fryar, and C. L. Ogden, "Prevalence of obesity among adults and youth: United States, 2015–2016," Centers Disease Control Prevention, Atlanta, GA, USA, Tech. Rep. 288, 2016.

[17] P. Dollar and C. L. Zitnick, "Fast edge detection using structured forests," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1558–1570, Aug. 2014.

[18] T. Ege and K. Yanai, "Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions," in *Proc. Thematic Workshops ACM Multimedia-Thematic Workshops*, 2017, pp. 367–375.

[19] A. Eldridge, C. Piernas, A.-K. Illner, M. Gibney, M. Gurinović, J. D. Vries, and J. Cade, "Evaluation of new technology-based tools for dietary intake assessment—An ilsi europe dietary intake and exposure task force evaluation," *Nutrients*, vol. 11, no. 1, p. 55, 2019.

[20] 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 *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 251–255.

[21] P. F. Felzenszwalb, "Representation and detection of deformable shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 2, pp. 208–220, Feb. 2005.

[22] E. A. Finkelstein, J. G. Trogdon, J. W. Cohen, and W. Dietz, "Annual medical spending attributable to obesity: Payer-and service-specific estimates," *Health Affairs*, vol. 28, no. 5, pp. w822–w831, Jan. 2009.

[23] Z. Ge, C. McCool, C. Sanderson, and P. Corke, "Modelling local deep convolutional neural network features to improve fine-grained image classification," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 4112–4116.

[24] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1440–1448.

[25] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.

[26] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "Food image recognition using very deep convolutional networks," in *Proc. 2nd Int. Workshop Multimedia Assist. Dietary Manage. (MADiMa)*, 2016, pp. 41–49.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2016, pp. 630–645.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[29] J. Hui. (2018). *mAP (Mean Average Precision) for Object Detection*. [Online]. Available: https://medium.com/@jonathanhui/map-mean-average-precision-for-object-detection45c121a31173

[30] T. Jiang, F. Jurie, and C. Schmid, "Learning shape prior models for object matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 848–855.

[31] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 1085–1088.

[32] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," in *Proc. ECCV Workshop Transferring Adapting Source Knowl. Comput. Vis. (TASK-CV)*, 2014, pp. 3–17.

[33] Y. Kawano and K. Yanai, "Real-time mobile food recognition system," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2013, pp. 1–7.

[34] 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 *Proc. ACM Int. Conf. Multimedia (MM)*, 2014, pp. 761–762.

[35] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[36] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 2, Jun. 2006, pp. 2169–2178.

[37] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[38] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 25–30.

[39] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2Calories: Towards an automated mobile vision food diary," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1233–1241.

[40] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 9, pp. 1632–1646, Sep. 2008.

[41] National Institute of Health (NIH). (2020). *24-hour Dietary Recall (24HR)—Dietary Assessment Primer*. [Online]. Available: https://dietassessmentprimer.cancer.gov/profiles/recall/

[42] National Institute of Health (NIH). (2020). *Food Frequency Questionnaire—Dietary Assessment Primer*. [Online]. Available: https://dietassessmentprimer.cancer.gov/profiles/questionnaire/

[43] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 850–855.

[44] J. Noronha, E. Hysen, H. Zhang, and K. Z. Gajos, "Platemate: Crowdsourcing nutritional analysis from food photographs," in *Proc. 24th Annu. ACM Symp. User Interface Softw. Technol.*, 2011, pp. 1–12.

[45] *Usda National Nutrient Database*, FNDDS, United States Dept. Agriculture, Washington, DC, USA, 2014.

[46] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vis.*, vol. 42, no. 3, pp. 145–175, 2001.

[47] *Overweight and Obesity*, World Health Org., Geneva, Switzerland, 2020.

[48] K. Poslusna, J. Ruprich, J. H. M. de Vries, M. Jakubikova, and P. van't Veer, "Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice," *Brit. J. Nutrition*, vol. 101, no. S2, pp. S73–S85, Jul. 2009.

[49] P. Pouladzadeh, P. Kuhad, S. V. B. Peddi, A. Yassine, and S. Shirmohammadi, "Mobile cloud based food calorie measurement," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jul. 2014, pp. 1–6.

[50] P. Pouladzadeh, P. Kuhad, S. V. B. Peddi, A. Yassine, and S. Shirmohammadi, "Food calorie measurement using deep learning neural network," in *Proc. IEEE Int. Instrum. Meas. Technol. Conf.*, May 2016, pp. 1–6.

[51] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[52] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, no. 3, pp. 222–245, Dec. 2013.

[53] W. Shimoda and K. Yanai, "CNN-based food image segmentation without pixel-wise annotation," in *Proc. Int. Conf. Image Anal. Process.* Cham, Switzerland: Springer, 2015, pp. 449–457.

[54] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[55] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556

[56] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *Computer Vision—ECCV*. Berlin, Germany: Springer, 2012, pp. 73–86.

[57] N. Slimani, P. Ferrari, M. Ocké, A. Welch, H. Boeing, M. van Liere, V. Pala, P. Amiano, A. Lagiou, I. Mattisson, C. Stripp, D. Engeset, R. Charrondière, M. Buzzard, W. van Staveren, and E. Riboli, "Standardization of the 24-hour diet recall calibration method used in the European Prospective Investigation into Cancer and Nutrition (EPIC): General concepts and preliminary results," *Eur. J. Clin. Nutrition*, vol. 54, no. 12, pp. 900–917, Dec. 2000.

[58] J. Stoppelman. *Yelp*. [Online]. Available: https://www.yelp.com/

[59] S. Subramanian and A. Deaton, "The demand for food and calories," *J. Political Economy*, vol. 104, no. 1, pp. 133–162, Feb. 1996.

[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[61] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154–171, Sep. 2013.

[62] W. C. Willett, L. Sampson, M. J. Stampfer, B. Rosner, C. Bain, J. Witschi, C. H. Hennekens, and F. E. Speizer, "Reproducibility and validity of a semiquantitative food frequency questionnaire," *Amer. J. Epidemiol.*, vol. 122, no. 1, pp. 51–65, 1985.

[63] K. Yanai and Y. Kawano, "Food image recognition using deep convolutional network with pre-training and fine-tuning," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops (ICMEW)*, Jun. 2015, pp. 1–6.

[64] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2249–2256.

[65] F. Zhu, M. Bosch, T. Schap, N. Khanna, D. S. Ebert, C. J. Boushey, and E. J. Delp, "Segmentation assisted food classification for dietary assessment," *Proc. SPIE*, vol. 7873, Feb. 2011, Art. no. 78730B.

LANDU JIANG (Member, IEEE) received the B.Eng. degree in information security engineering from Shanghai Jiao Tong University, the M.Sc. degree in construction management, the M.Sc. degree in computer science from the University of Nebraska–Lincoln, and the Ph.D. degree from the School of Computer Science, McGill University. He currently holds the Mitacs Accelerate Industrial postdoctoral position at Aerial Technology and McGill University. His research interests include computer vision, machine learning, smart sensing, wearable and mobile computing, cyber-physical systems, green energy solutions, and online social networks.

BOJIA QIU received the bachelor's degree from Zhejiang University and the master's degree from the School of Computer Science, McGill University. His research interests include applying deep learning techniques for object detection and recognition task. He is also interested in optimizing the structure of neural networks and finding better hyper-parameters to improve the performance of deep model.

XUE LIU (Fellow, IEEE) received the Ph.D. degree in computer science from the University of Illinois at Urbana–Champaign. He worked as the Samuel R. Thompson Chaired Associate Professor with the University of Nebraska–Lincoln and HP Labs, Palo Alto, CA, USA. He is currently a Full Professor with the School of Computer Science and the William Dawson Scholar (Chair Professor) with McGill University. He is also the Vice President R&D, the Chief Scientist, and a Co-Director of the Samsung AI Centre-Montreal. He is also a Professor (courtesy appointment) of mathematics and statistics with McGill University. He is also a Faculty Member and an Associate Member of the Montreal Institute for Learning Algorithms (MILA). He served as the Chief Scientist of Tinder Inc. His research interests include computer and communication networks, real-time and embedded systems, distributed systems, cyber-physical systems, green computing, and smart energies. He has published over 280 research articles in major highly-reputable peer-reviewed International academic journals and conference proceedings, including ACM MobiCom, the IEEE INFOCOM, ICNP, the IEEE Security and Privacy (Oakland), the IEEE RTSS, the IEEE RTAS, ACM/IEEE ICCPS, WWW, ACM UbiComp, ACM KDD, the IEEE ICDE, and so on and various the IEEE/ACM Transactions. He is or had been an Editor/Associate Editor of the IEEE/ACM Transactions on Networking (ToN), *ACM Transactions on Cyber-Physical Systems* (TCPS), the IEEE Transactions on Vehicular Technology (TVT), the IEEE Communications Surveys and Tutorial (COMST), and the IEEE Transactions on Parallel and Distributed Systems (TPDS). He served on various national and international grant review committees or panels.

CHENXI HUANG is currently an Assistant Professor with Xiamen University. His research interests include image processing, image reconstruction, data fusion, three-dimensional visualization, and machine learning, and so on. He has been an Associate Editor of the *Journal of Medical Imaging and Health Informatics*, since 2019.

KUNHUI LIN is currently a Professor with Xiamen University. His research interests include database systems, computer networks, multimedia applications, intelligence information systems, and so on.