Counting Bites and Recognizing Consumed Food from Videos for Passive Dietary Monitoring

Jianing Qiu, Frank P.-W. Lo, Shuo Jiang, Ya-Yen Tsai, Yingnan Sun, and Benny Lo, *Senior Member, IEEE*

***Abstract*—Assessing dietary intake in epidemiological studies are predominantly based on self-reports, which are subjective, inefficient, and also prone to error. Technological approaches are therefore emerging to provide objective dietary assessments. Using only egocentric dietary intake videos, this work aims to provide accurate estimation on individual dietary intake through recognizing consumed food items and counting the number of bites taken. This is different from previous studies that rely on inertial sensing to count bites, and also previous studies that only recognize visible food items but not consumed ones. As a subject may not consume all food items visible in a meal, recognizing those consumed food items is more valuable. A new dataset that has 1,022 dietary intake video clips was constructed to validate our concept of bite counting and consumed food item recognition from egocentric videos. 12 subjects participated and 52 meals were captured. A total of 66 unique food items, including food ingredients and drinks, were labelled in the dataset along with a total of 2,039 labelled bites. Deep neural networks were used to perform bite counting and food item recognition in an end-to-end manner. Experiments have shown that counting bites directly from video clips can reach 74.15% top-1 accuracy (classifying between 0-4 bites in 20-second clips), and a MSE value of 0.312 (when using regression). Our experiments on video-based food recognition also show that recognizing consumed food items is indeed harder than recognizing visible ones, with a drop of 25% in F1 score.**

***Index Terms*—Bite counting, dietary intake monitoring, food recognition, video understanding.**

I. Introduction

The 2019 Lancet Series on *The Double Burden of Malnutrition* highlights that nearly 2.3 billion children and adults are estimated to be overweight globally and more than 150 million children are affected by stunting [1]. Despite continued efforts to prevent malnutrition in all its forms, recent estimates reveal that we are still far from reaching the World Health Assembly global nutrition targets set for 2025 to improve the nutritional status of young children [2]. It is clear that reducing the risk of malnutrition requires effective diet interventions; however, in current epidemiological studies, dietary assessment methods, such as 24-hour dietary recall and food frequency questionnaires (FFQs) are often inaccurate and inefficient as their assessments are predominantly based on self-reports which depend on respondents' memories and require intensive efforts to collect, process, and interpret [3]. To meet the need for objective and accurate dietary assessments, sensing-based technological approaches are emerging with the aim to transform these conventional subjective assessments. Current technological approaches include image-based sensing, sensing with eating action unit (EAU), and biochemical sensing [4]. EAU-based sensing approaches are primarily designed for detecting eating actions and some are able to produce reasonable estimates with the use of inertial sensors [5]–[13], but one evident limitation of the EAU-based approaches is that they can hardly recognize food items. While image-based approaches can recognize or segment food items more accurately, they still face challenges when processing images that have hidden, occluded, or deformed food items. In addition, many image-based approaches tend to only process few images or a single food image well captured before eating [14]–[16], which is insufficient to determine the exact food consumption. It is for this reason that videos are a more promising and reliable source to estimate food intake, as they can capture the entire eating episode with much information, such as eating actions, chewing, and swallowing sound when the camera is appropriately positioned. As videos contain both spatial and temporal information, apart from basic food recognition, the number of bites taken within an eating episode can also be derived. Pilot study in [5] has shown the potential correlation between the number of bites and dietary intake. For these reasons, this work investigates a video approach for passive dietary assessment. Specifically, we focus on two targets, which have not yet been explored in any prior work on dietary assessments with egocentric videos. The first target of this work is to count the

Manuscript received June 2, 2020; revised July 28, 2020 and September 1, 2020; accepted September 5, 2020. Date of publication September 8, 2020; date of current version May 11, 2021. *(Corresponding author: Jianing Qiu.)*

Jianing Qiu and Ya-Yen Tsai are with the Hamlyn Centre, Imperial College London, London SW7 2AZ, U.K., and also with the Department of Computing, Imperial College London, London SW7 2AZ, U.K. (e-mail: jianing.qiu17@imperial.ac.uk; y.tsai17@imperial.ac.uk).

Frank P.-W. Lo, Yingnan Sun, and Benny Lo are with the Hamlyn Centre, Imperial College London, London SW7 2AZ, U.K., and also with the Department of Surgery and Cancer, Imperial College London, London SW7 2AZ, U.K. (e-mail: po.lo15@imperial.ac.uk; y.sun16@imperial.ac.uk; benny.lo@imperial.ac.uk).

Shuo Jiang is with the Hamlyn Centre, Imperial College London, London SW7 2AZ, U.K., and also with the State Key Laboratory of Mechanical System and Vibration, School of Mechanical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: jiangshuo@sjtu.edu.cn).

Digital Object Identifier 10.1109/JBHI.2020.3022815

number of bites within a time period from recorded dietary intake videos. In the domain of video understanding, there has been growing interest in end-to-end temporal counting [17]–[19], especially with the adoption of deep learning [20]. Thus, in this work, deep neural networks have been used to count the number of bites in an end-to-end fashion. An accurate bite count can be used to measure a subject's eating speed and infer the eating behaviour (e.g., whether the subject always binge-eats). The second target is to perform fine-grained food recognition from videos. Unlike prior works that only focus on recognizing a main category of food (e.g., pasta, or fried rice), this work advances food recognition to a much fine-grained level. In the case of having a meal, some hidden or occluded food ingredients may start to be revealed as eating progresses, and videos can well capture these hidden ingredients during the process. Hence, in this work, three sets of fine-grained food recognition are performed: 1) classifying a meal into a fine-grained class (e.g., *prawn_pasta* or *mixed_seafood_pasta)*; 2) recognizing all visible food items (e.g., food ingredients of a meal, and drinks) within a recorded eating episode; 3) As a subject may not consume all food items in a meal, we further propose to recognize and identify those food items consumed by the subject. To achieve these two targets and evaluate the performance of passive dietary intake monitoring with videos, a new dataset has been constructed which contains egocentric videos capturing eating episodes, and extensive experiments have been conducted with the use of state-of-the-art video understanding networks. Our key findings include: 1) bite counting can be treated as both classification and regression problems. End-to-end bite counting is feasible, and by using only visual clues, the accuracy on unseen subjects can reach 74.15% when classifying between 0-4 bites in 20-second clips, and a low MSE value of 0.312 when using regression; 2) recognizing consumed food items in an end-to-end fashion is more difficult than recognizing all visible food items (a 25% accuracy gap was observed between these two cases), which we conjecture is because recognizing consumed food items needs to take more information into account, such as eating actions and hand-food interactions. To the best of our knowledge, this is the first work that solely uses egocentric videos of eating episodes to count the number of bites and recognize consumed food with the aim of providing automatic and objective dietary assessments.

The rest of this paper is organized as follows: Section II discusses prior technological approaches to dietary intake assessments as well as state-of-the-art video recognition networks. Section III presents details of collecting and constructing the dataset. Methods are described in Section IV, followed by the analysis of experimental results in Section V. We discuss a few limitations of our work in Section VI and conclude in Section VII.

II. Related Work

A. Technological Approaches for Dietary Assessment

With the ubiquitous use of wearable technologies and the advances in associated data analysis algorithms, there has been a rapid increase in developing automatic and objective dietary assessment systems. EAU-based and image-based sensing systems are mostly related to our work, and are the two most common approaches for objective dietary assessments to date. Thus, we mainly discuss prior works on EAU-based and image-based systems in this section.

1) EAU-based Systems: Detecting eating action units (EAUs) [4] such as feeding gestures is a straightforward way to assist dietary intake monitoring and assessment. EAU-based systems are mainly based on inertial or acoustic sensing. In systems with inertial sensors, it is common to use wrist-worn devices to detect eating episodes, feeding gestures, or to count the number of bites from accelerometer and gyroscope signals [5]–[13]. In systems with acoustic sensors, acoustic signals are used to detect eating episodes [21], swallows [22], [23], or to distinguish eating from other activities [24], [25]. The effects of using both inertial and acoustic sensing have also been examined in [26] and [27]. Other sensing modalities have been explored in EAU-based systems include piezoelectric signals [28] and electromyography (EMG) [29] for chewing detection. Despite reasonable performance in these eating-related detection and recognition, EAU-based systems face a number of limitations. One of the limitations is that they can hardly distinguish between food categories, especially for those inertial sensing-based systems. To train machine learning models with collected inertial or acoustic signals, some EAU-based systems require an additional camera to be set up to record the eating episodes in order to obtain ground truth labels [5], [7]–[13], [21], [27]. This additional setup is cumbersome and since a camera is already in use for obtaining the ground truth, directly using visual information to perform eating-related tasks may be more efficient.

2) Image-based Systems: Image-based systems usually consist of a single or sometimes multiple cameras for recording dietary intake in the form of images or videos, and a set of associated computer vision algorithms for food recognition [30]–[33], segmentation [34]–[37], and volume estimation [38], [39]. Sun *et al.* [40] designed a chest-worn camera that takes meal pictures periodically during an eating episode. Liu *et al.* [41] designed an ear-worn device which continuously monitors the sound of chewing, and once it has been detected, a miniaturized camera inside the device will be triggered to record eating episodes. In addition to these devices specifically designed for capturing dietary intake, smartphones are also commonly used to capture and process food images [14]–[16], [42], [43]. Using smartphones also offers an opportunity to utilize restaurant information [42], [43], such as the menu and recipes through GPS localization, which could largely narrow down the number of food categories and potentially provide more accurate nutrient intake estimation. As using recipes facilitates dietary intake assessments, there is growing interest in developing cross-modal food image and recipe retrieval [44]–[49]. Recently, efforts to assess individual dietary intake in communal eating scenarios have also been made with the use of a 360 camera [50], [51]. Fine-grained food ingredient recognition has also been studied to enhance general food recognition [52], or to perform recipe retrieval [53], but so far studies have only been carried out in recognizing ingredients from food images rather than from dietary intake videos. In [54], clustering images sampled from egocentric videos into *food* and *non-food* classes has been attempted, but the types of food were not recognized. For more comprehensive reviews of image-based approaches, we refer readers to [55] and [56].

Although bite counting is one of our targets, it is tackled through vision instead of inertial sensing [5], [9], [12], [13], and our system essentially is a type of image-based systems. Furthermore, we seek an end-to-end approach to counting bites in this work (i.e., given a video clip of eating, the network directly predicts the number of bites taken in that clip, which is realized by solving bite counting as a classification or a regression problem in the context of end-to-end temporal counting [17]–[19]). This is different from prior works on bite detection with inertial sensing [12] and intake gesture detection from videos [51], both relying on sliding windows to localize each bite temporally. The former detects bites on the basis of predefined wrist micro-movements during eating, and the latter detects intake gestures by first estimating the probability of each frame being *intake* or *non-intake* frame and then assigning a local maximum as an intake event. Temporally localizing each bite is one way to achieve bite counting, but as shown in this work, end-to-end prediction of the bite count is also feasible and able to produce decent results. In a recent work, end-to-end bite detection with inertial data has been introduced [13]. Although we both adopt the end-to-end concept, the methodologies we proposed are completely different, and are targeted for different modalities (i.e., visual vs. inertial).

B. Deep Networks for Video Recognition

Both 2D convolutional neural networks (2D CNNs) and 3D CNNs have been used in video recognition. As a 2D CNN itself does not come with the ability to model temporal evolution in a video, deep architectures built with 2D CNNs normally introduce additional temporal modelling scheme to enhance recognition performance. For example, Two-Stream [57] uses two separate streams to process a video, a spatial stream for capturing appearance and a temporal stream for motion; TSN [58] decomposes a video into short time segments and fuses prediction from each segments at the end; TRN [59] proposes a relational module in order to better capture temporal relations in videos; CNN+LSTM [60] builds a LSTM [61] on top of a CNN to integrate temporal information. Despite being less computationally expensive than using 3D CNNs, their ability to model along the temporal dimension is limited. However, although better at learning temporal information, 3D CNNs, such as C3D [62] and I3D [63] are computationally heavy. Balancing between accuracy and efficiency is thus an important factor in the design of deep architectures for video recognition. In this work, we adopted two state-of-the-art network architectures from video recognition: TSM [64] and SlowFast [65]. TSM, short for temporal shift module, is a module that can be inserted into a 2D CNN that gives the 2D CNN comparable performance to 3D CNN in video understanding. Thus, TSM has the advantages of being both efficient and accurate for video recognition. SlowFast uses two pathways operating at different frame rates, one at low rate to learn spatial semantics (Slow pathway) and the other at high rate to capture motion (Fast pathway). Despite the use of 3D CNNs in SlowFast, it is still relatively lightweight as its Fast pathway is designed to have low channel capacity. We notice that in intake gesture detection [51], SlowFast was also adopted. Their use of SlowFast was to perform a per-frame *intake* and *non-intake* binary classification, whereas we adopted SlowFast to directly estimate the bite count of a video clip as well as to recognize food items.

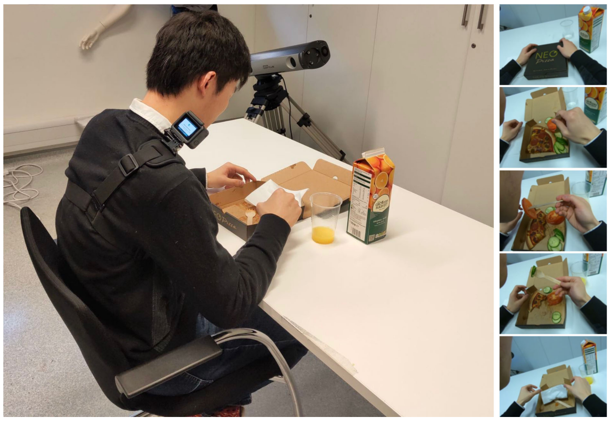

Fig. 1. A subject wears a GoPro Hero 7 Black camera, which records the entire eating episode in a passive way. The camera is mounted on the shoulder at the same side as the subject's dominant hand. This position provides the camera with a good view that can capture both the mouth and dominant hand, which facilitates bite counting and consumed food item recognition from the recorded egocentric video.

III. Dataset

In order to validate our concept of bite counting and consumed food recognition from dietary intake videos, a new egocentric video dataset was constructed. Egocentric videos record the visual field of a subject during eating, which offers a better opportunity to understand the subject's eating behavior (e.g., like or dislike a certain type of ingredients) compared to other sensing modalities such as inertial sensing. In addition, as they can be recorded in a passive way, less interruption will be caused during eating. In practice, people are generally unwilling to wear cameras in their daily life due to privacy concerns. On the other hand, patients could often be more compliant in wearing devices for collecting data for diagnostic purposes. This passive and egocentric way of dietary intake recording is therefore suitable to be applied in care homes and hospitals where the need for dietary monitoring exists and staff can help the user to wear the device and initiate the recording.

A. Data Collection

Data collection was conducted in a laboratory setting. 12 healthy subjects were recruited and asked to come at their normal mealtimes to record dietary intake videos. A GoPro Hero 7 Black camera was mounted on the subject's shoulder, the one at the same side as their dominant hand, before they start eating. This position allows the camera to better capture both the mouth and dominant hand, which facilitates subsequent bite counting and consumed food recognition. Figure 1 illustrates the setup for data collection.

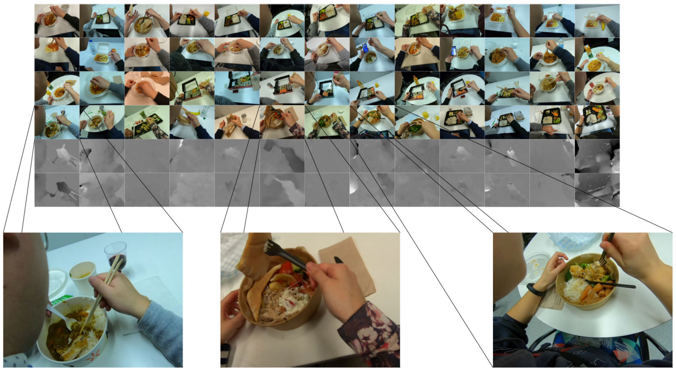

Fig. 2. One RGB frame is shown for each dietary intake video (top four rows). The extracted optical flows of the fourth row are shown in the bottom (x and y directions).

TABLE I
66 Unique Food Items Labelled in the Dataset

potato	chicken	beef	soda	water	orange_juice
red_wine	champagne	green_onion	seaweed	dumpling	broccoli
napa_cabbage	chili_pepper	rice	chicken_katsu	curry	pickled_radish
onion	tofu	miso_soup	takuan	celery	green_bean
pork_ribs	pita_bread	cucumber	tomato	chips	tzatziki
salmon	pepper	sweet_and_sour_chicken	baked_beans	sausage	mushroom
hash_browns	scrambled_eggs	bacon	banana	apple	carrot
melon_milk	prawn	mussel	pasta	squid	tomato_sauce
clam	cheese	sponge_cake	sushi_salmon	sushi_tuna	sushi_prawn
sushi_vegetable	sushi_vegetable_meat	inari	soy_sauce	edamame	wasabi
red_cabbage	sweet_chilli_sauce	sandwich_prawn	teriyaki_sauce	lemon	lettuce

Invisible ingredients, such as oil and salt, are not labelled.

The resolution of the GoPro camera was set to 1920 × 1440 and it recorded videos at 30 fps. Subjects were asked to eat their meal as they normally do (e.g., they were free to read their messages, or browse the web on their phones while eating). There was no restriction on how much the subject should eat, and they were free to leave some food items uneaten if they dislike (e.g., a subject may not like pickles in a meal). The only restriction was no talking. The camera was turned off once the subject finished eating.

After collecting the data, two labellers were involved in data annotations. One annotated all bite counts and food items, and the other double checked the annotation. The dataset will be made available upon request.

B. Data Statistics

A total of 52 dietary intake videos (i.e., 52 meals) is used in this study. Figure 2 shows some snapshots of the videos recorded. Meals consumed by the subjects can be categorized into 8 main classes (please refer to the inner circle of Figure 3), and some of them can be further categorized into fine-grained meal classes as shown in the middle circle. The outer circle of Figure 3 displays visible food items, including ingredients and drinks, in each meal class. In total, there are 66 unique and visible food items labelled in these meals (please refer to Table I for a full list of labelled food items), with the maximum number of visible food items a meal class has being 19 (*sushi_non-vegetarian*) and the minimum being 4 (*sandwich_meal_deal*). Eating with a fork and knife, a spoon, chopsticks, or hands can all be found in the recorded videos. The average time the subjects spent finishing a meal is 9 m 57 s, with the longest time being 27 m 57 s and the shortest time being 3 m 48 s.

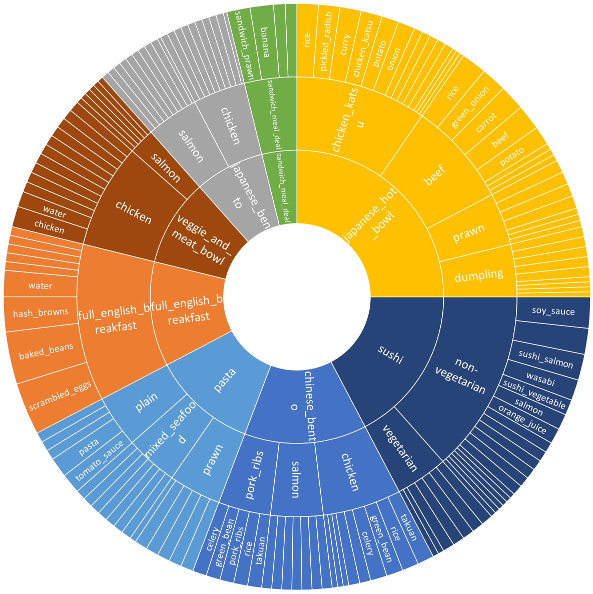

Fig. 3. Meal statistics. The inner circle shows 8 main meal categories. Some meals can be further categorized as shown in the middle circle, which results in 18 fine-grained meal categories. The outer circle shows visible ingredients and drinks identified in each meal. Due to the limitation of font size, not all visible items are displayed in the outer circle.

C. Constructing Dataset for Bite Counting

One of our work's objectives is to count the number of bites from the video data. In order to feed the video data into a deep neural network and count bites in an end-to-end manner, each raw dietary intake video was first split into a set of 20-second video clips using FFmpeg.[1] The reasons for the length of video clips to be 20 seconds are: 1) most video recognition networks are designed for processing short video clips (e.g., 10-second clips from the Kinetics dataset [66]); 2) compared to 10 seconds, cutting a dietary intake video into 20-second clips is able to reduce the number of cutting points at which a bite may happen. Following [5], a bite in this work is defined as food being fully placed into the mouth. Video clips that may cause ambiguities were excluded (e.g., a bite cannot be verified because it happens outside the camera view, but this can be solved by using a wide-angle camera (or the wide FOV mode of GoPro) in future work). In total, 1,022 video clips are valid among those extracted from 52 dietary intake videos.[2] The number of bites the subjects take in a 20-second interval ranges from 0 to 9 as shown in Figure 4, with the average number of bites taken in 20 seconds being 1.995 (the dataset has a total of 2,039 bites recorded). In 987

[1] https://ffmpeg.org/

[2] clips that are non-valid and therefore were excluded are: 1) clips in which a bite happens outside the camera view; 2) an incomplete bite occurs in the beginning or the end of the clip; 3) the last 1 or 2 clips of a video in which the subject has stopped eating or the length of that clip is less than 20 s; 4) we also purposely discarded some clips of long videos (over 20 mins) to balance the dataset.

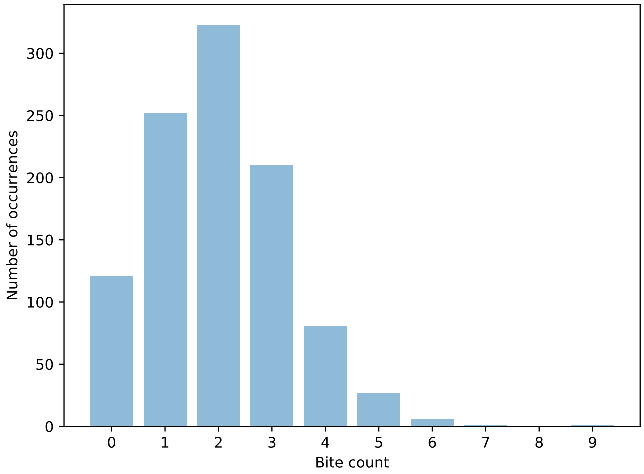

Fig. 4. The number of times each bite count occurs in the dataset. The number of bites the subjects take in a 20 s interval ranges from 0 to 9.

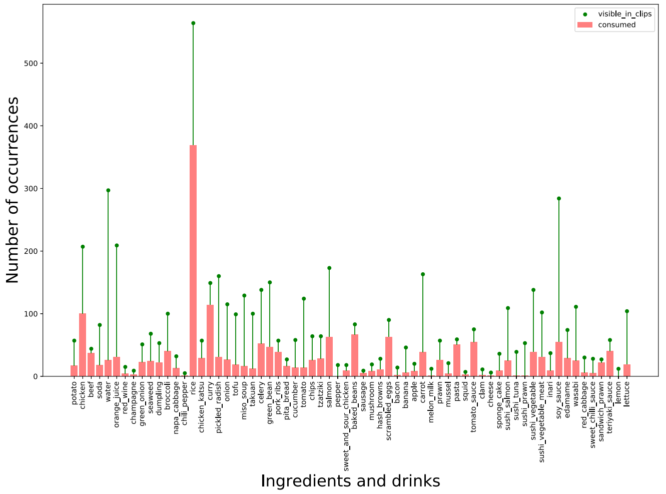

Fig. 5. Food item (ingredient and drink) statistics. Two different sets are shown: 1) visible ingredients and drinks in all video clips; 2) consumed ingredients and drinks in all video clips.

out of 1,022 video clips, subjects take less than 5 bites, with all numbers (0 to 4) occurring in more than 50 clips, which provides sufficient data to validate our bite counting approach. Therefore, we only used these 987 video clips to construct the dataset for bite counting. The dataset was then split into training and test sets, with the training set having 631 clips (from 32 videos) and the test set 356 clips (from the rest 20 videos). It is worth noting that the dataset splits strictly avoid overlapping each other, i.e., there is no such case that the same subject having the same meal occurs in the both training and test sets. As bite counting relies on capturing motion of taking a bite, we also calculated optical flows using the TV-L1 algorithm [67] as an additional modality for bite counting. The number of times of drinking was labelled separately. However, as we observed that subjects rarely drink in video clips, which results in most of the label being 0 (i.e., drinking 0 times), counting drinking times was not studied in this work. Therefore, this dataset is exclusively a bite counting dataset.

D. Constructing Dataset for General Food Recognition

Existing food recognition systems are mostly focused on recognizing food types from images [30]–[33]. Recognizing food from videos may sound inefficient but has its own merits, especially in pervasive dietary intake monitoring scenarios. In certain scenarios, the type of food may not be recognized on the basis of a single food image, especially when the image does not contain discriminative parts of the food. Videos also have a better chance to capture previously hidden and occluded ingredients and other food items during eating, as they may be revealed as eating continues. Therefore, we used all 1,022 video clips to construct a dataset for three different general food recognition tasks: 1) recognizing 8 main meal classes; 2) recognizing 18 fine-grained meal classes; 3) recognizing all visible food items in a video clip, which include food ingredients and also drinks. The dataset was split into 655 clips (from 32 videos) for training and 367 (from the rest 20 videos) for testing. The green lines in Figure 5 show the number of occurrences of visible ingredients and drinks in these 1,022 video clips, with *rice* occurring the most, visible in 564 clips and *chili_pepper* the least, only visible in 5 clips.

E. Constructing Dataset for Consumed Food Recognition

To recognize consumed food items, as one of our main objectives, all consumed food items in a video clip were manually labelled. The red bars in Figure 5 show the statistics of consumed food items. Although *chicken* is visible in 207 video clips, it is really consumed by the subjects in only 100 clips. The same dataset splits as for bite counting were adopted for consumed food recognition.

IV. Method

In this work, we aim to achieve bite counting and food recognition in an end-to-end manner from egocentric videos. To this end, deep neural networks were used. Two state-of-the-art networks from the domain of video recognition were adopted: TSM [64] and SlowFast [65]. A batch of frames are sampled from a video clip as the input to both TSM and SlowFast, and the networks directly predict the number of bites taken in that clip. Visible and consumed food items are also recognized but with separately trained networks.

A. Bite Counting: Classification or Regression

Bite counting can be formulated as either a classification or a regression problem as bite count is both categorical and numerical. Previous works in video question answering tend to consider temporal counting as a regression problem [17]–[19]. In this work, we investigate both classification and regression as the solutions to bite counting. For classification, deep neural networks were trained with standard cross-entropy loss. For regression, networks were trained with ℓ_2 loss, which measures

TABLE II
OVERALL RESULTS OF BITE COUNTING. CLS STANDS FOR CLASSIFICATION AND REG STANDS FOR REGRESSION

Model	Backbone	Pretrain	Modality	#Frame	#Crop	CLS (Top-1)	REG (Acc)	REG (MSE)
TSM	2D ResNet-50	ImageNet	RGB	8	1	42.42	37.64	1.118
TSM	2D ResNet-50	ImageNet	RGB	16	1	48.88	43.82	0.933
TSM	2D ResNet-50	Kinetics	RGB	8	1	50.28	48.32	0.673
TSM	2D ResNet-50	Kinetics	RGB	16	1	60.67	60.67	0.427
TSM	2D ResNet-101	Something-V2	RGB	8	1	54.78	48.03	0.661
TSM	2D ResNet-50	Something-V2	RGB	16	1	62.64	62.64	0.436
TSM	2D ResNet-50	Kinetics	Flow	8	1	51.12	45.51	0.599
TSM	2D ResNet-50	Kinetics	Flow	16	1	55.62	55.34	0.438
$\text{TSM}_{\text{Ensemble}}$	2D ResNet-50	Kinetics + Kinetics	RGB + Flow	16 + 16	1	57.58	60.96	0.364
$\text{TSM}_{\text{Ensemble}}$	2D ResNet-50	Kinetics + Something-V2	RGB + RGB	16 + 16	1	63.20	62.64	0.383
$\text{TSM}_{\text{Ensemble}}$	2D ResNet-50	Kinetics + Something-V2 + Kinetics	RGB + RGB + Flow	16 + 16 + 16	1	59.83	**63.20**	**0.352**
Slow-only	3D ResNet-50	Kinetics	RGB	8	1	46.91	41.29	0.747
Slow-only	3D ResNet-50	Kinetics	RGB	8	3	48.88	41.01	0.738
SlowFast	3D ResNet-50	Kinetics	RGB	4 + 32	1	60.11	49.72	0.568
SlowFast	3D ResNet-50	Kinetics	RGB	4 + 32	3	61.24	49.16	0.567
SlowFast	3D ResNet-50	Kinetics	RGB	8 + 32	1	63.76	52.25	0.550
SlowFast	3D ResNet-50	Kinetics	RGB	8 + 32	3	**64.89**	51.12	0.557

the mean squared error between the predicted value and the true bite count.

B. Food Recognition: Multi-Class and Multi-Label

Food recognition in this work includes classifying a meal into a single category and recognizing all visible or consumed food items in a video clip. Classifying a meal (i.e., either into 8 main meal classes or 18 fine-grained classes) is a standard multi-class classification problem. Therefore, the networks were trained with cross-entropy loss. Recognizing food items, including ingredients and drinks, is a multi-label classification problem. Therefore, the networks were trained with binary cross-entropy loss.

C. Evaluation Metrics

For bite counting as a classification problem and the meal classification, we calculated the top-1 accuracy. For bite counting as a regression problem, two evaluation metrics were adopted. The first is mean squared error (MSE) measured between the predicted values and true labels (i.e., bite count). The second is to calculate accuracy as in the classification. The predicted value is first rounded to the closest integer and accuracy is then calculated by dividing the sum of correctly predicted values after rounding with the total number. For food item recognition, both recognizing visible ones and consumed ones, F1 score was used after binarizing multi-label predictions with a threshold of 0.5.

V. EXPERIMENTS

A. Implementation Details

All network models were fine-tuned on our dietary video dataset using SGD. For TSM, different pretrained models were used to initialize, which include pretrained models from ImageNet [68], Kinetics [66], and Something-Something-V2 [69] datasets. For SlowFast, we used models pretrained on Kinetics to initialize. TSM models were fine-tuned for 50 epochs with a learning rate starting at 0.001 and decayed by 10 at epoch 10, 20, and 40. SlowFast models were fine-tuned for 64 epochs with a base learning rate of 0.1, and 16 warmup epochs with a start learning rate of 0.01. The inputs to both TSM and SlowFast in training are 224 × 224 crops from sampled frames. TSM models were tested with 1-crop (center 224 × 224 crops of input frames). SlowFast were tested with both 1-crop and 3-crop (left, center, and right 256 × 256 crops).

B. Results of Bite Counting

The overall results of bite counting are summarized in Table II, which include the accuracy of both classification and regression. It can be observed that using more frames of a video clip generally leads to better results. This is especially true when TSM was used. As shown in the first 8 rows of Table II, no matter what modality was used as the input, RGB frames or optical flows, using 16 frames always produces higher accuracy than just using 8 frames, even if a deeper network (i.e., ResNet-101 [70]) was used to process 8 frames as the input. In addition, the accuracy of using optical flows as the single input modality generally is no better than that of using RGB frames (a 5% decline can be observed when comparing the 4th and 8th rows). However, in solving bite counting as a regression problem, the addition of optical flows is able to improve the accuracy. An ensemble of a TSM model trained with RGB frames and another one trained with optical flows produces slight improvements in accuracy, from 60.67% to 60.96%, and also decreases the MSE value, from 0.427 to 0.364 as shown in the 4th and 9th rows. An ensemble of three TSM models with one of them trained with optical flows is also able to improve the accuracy compared to an ensemble without the optical flow-trained model. Accuracy is increased from 62.64% to 63.20% and the MSE value is decreased from 0.383 to 0.352 as shown in the 10th and 11th rows. Nevertheless, in solving bite counting as a classification problem, combining optical flow with RGB modality results in a decrease in accuracy. Similar to TSM models, SlowFast models with 8 frames as the input to the slow pathway produce higher accuracy than 4 frames, when the input to the fast pathway is fixed to 32 frames. We also used a Slow-only architecture [65], i.e., without the fast pathway in SlowFast, to count the number of bites. The results of

TABLE III
RESULTS OF BITE COUNTING ON 4-FOLD INTER-SUBJECT CROSS VALIDATION

Metric	Model	S01&S02&S03	S04&S05&S06	S07&S08&S09	S10&S11&S12	Avg.
CLS (Top-1)	TSM	58.89	55.70	**60.95**	**74.15**	62.42
	SlowFast	**60.08**	**59.15**	58.57	72.11	**62.48**
REG (Acc)	TSM	67.59	48.81	**62.86**	**64.63**	**60.97**
	SlowFast	**70.36**	**49.34**	59.05	59.86	59.65
REG (MSE)	TSM	**0.393**	**0.593**	**0.411**	**0.312**	**0.427**
	SlowFast	0.414	0.635	0.479	0.412	0.485

Each subject appears only in either the training or the test set. We used 9 subjects for training and the rest 3 subjects for testing in each fold.
The configuration of backbone, pretrain, modality, #frame, and #crop for the TSM and SlowFast models reported in the table is (2D ResNet-50, Something-V2, RGB, 16, 1) and (3D ResNet-50, Kinetics, RGB, 8+32, 3), respectively.
The number of testing clips of each fold (from left to right): 253, 377, 210, and 147.

Slow-only architecture are not comparable to those of SlowFast, which indicates that capturing the motion of taking a bite is important for accurate bite counting. SlowFast and Slow-only networks were also tested with 3-crop. The use of 3 crops is able to boost the accuracy in classification, but slightly decreases the accuracy in regression and has no significant impact on the measured MSE values. Overall, the highest accuracy of bite counting we obtained is 64.89% when it is solved as a classification problem. Solving bite counting as a regression problem can also produce reasonable accuracy of 63.20% and achieve a low MSE of 0.352 (each clip in the test set has an average bite count of 1.885).

We further conducted a 4-fold inter-subject cross validation to evaluate the performance of the proposed bite counting method on unseen subjects. The dataset introduced in Section III-C was re-split, with 9 subjects used for training and the rest 3 subjects for testing in each fold. The results are summarized in Table III. Although after averaging across all 4 folds, the results as shown in the last column are close to the reported results in Table II, it is impressive to find that on the testing split of subjects #10, #11, and #12, TSM yields the top-1 accuracy of 74.15% as a classification model, and the MSE of 0.312 as a regression model. For SlowFast, it achieves 70.36% regression accuracy on the split of subjects #1, #2, and #3.

C. Results of General Food Recognition

Table IV shows the overall results of general food recognition, which include meal classification, and visible food item recognition. Note that models were trained only with RGB modality in these tasks. In terms of meal classification, a ResNet-50 with TSM embedded achieves the best accuracy when the input is set to 8 frames, 97.55% accuracy of classifying a meal into 8 main classes and 54.77% accuracy of classifying it into 18 fine-grained classes. We hypothesize the accuracy of classifying a meal into 18 fine-grained classes being not satisfactory is because 1) the meals under the same main category only have a single or few ingredients that can distinguish between them; 2) and in some clips, these discriminative ingredients have already been consumed, which confuses the models and results in the meal being misrecognized as another similar meal under the same main category. The visualized confusion matrix also verifies that most misclassification happens between meals that belong to a same main category (e.g., meals under the category of *japanese_hot_bowl* as shown in both Figures 6 and 7). Note that due to the limited number of *veggie_and_meat_bowl_salmon* samples, this class was only included in the training set, and that is why the last row of the confusion matrix is all zeros. In terms of recognizing all visible food items from a clip, SlowFast produces the best overall F1 score of 65.0% and TSM is also able to produce a reasonable overall score of 59.5%. In order to better understand which visible food items are well recognized, we calculate a F1 score for each food item (i.e., given a food item, its F1 score is calculated using a set of its true labels and a set of its predicted labels from all testing clips). The results are shown in Figure 8. In general, the F1 scores of most food items produced by SlowFast are also higher than TSM. In all food items that have a non-zero F1 score, 48 items in total for the SlowFast model, 14 of them have the F1 score over 80%, with *scrambled_eggs*, *tomato_sauce*, and *sweet_chilli_sauce* having the F1 score of 100%, indicating that the model correctly recognizes these three items in all clips that contain them, and in clips without them, the model indeed estimates that they are not present. For the TSM model, its recognition produces 44 food items that have a non-zero F1 score, and 9 of them have the F1 score over 80%.

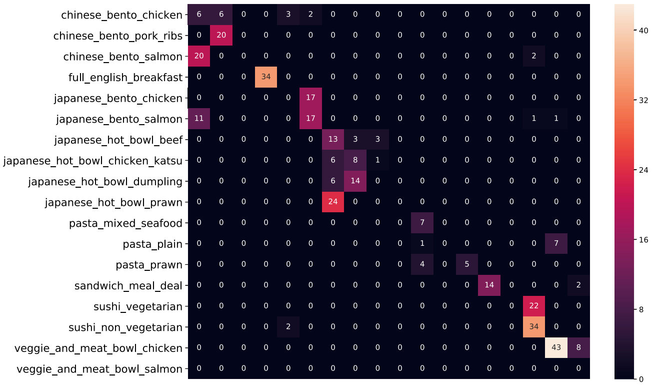

Fig. 6. Confusion matrix of a TSM model (top-1 accuracy of 54.77% on the test set) classifying meals into 18 classes.

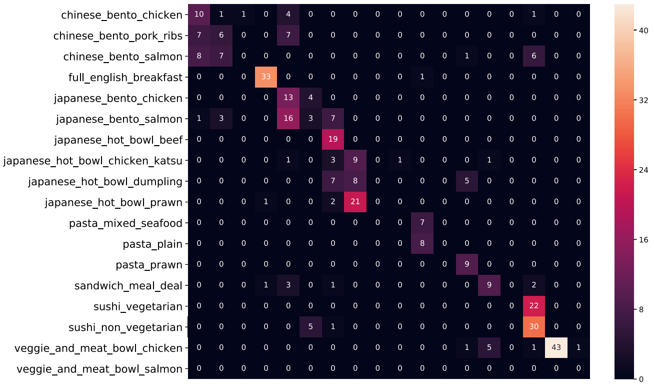

Fig. 7. Confusion matrix of a SlowFast model (top-1 accuracy of 52.04% on the test set) classifying meals into 18 classes.

TABLE IV
OVERALL RESULTS OF GENERAL FOOD RECOGNITION. MAIN REPRESENTS CLASSIFYING A MEAL INTO A MAIN CATEGORY. FG REPRESENTS FINE-GRAINED MEAL CLASSIFICATION. V-ITEM REPRESENTS VISIBLE FOOD ITEM RECOGNITION

Model	Backbone	Pretrain	#Frame	#Crop	Main (Top-1)	FG (Top-1)	V-Item (F1)
TSM	2D ResNet-50	Kinetics	8	1	**97.55**	**54.77**	41.5
TSM	2D ResNet-50	Kinetics	16	1	95.10	51.77	**59.5**
TSM	2D ResNet-101	Something-V2	8	1	84.47	40.05	-
TSM	2D ResNet-50	Something-V2	16	1	85.01	36.51	-
Slow-only	3D ResNet-50	Kinetics	4	1	92.92	46.87	55.6
Slow-only	3D ResNet-50	Kinetics	4	3	93.19	47.14	58.1
Slow-only	3D ResNet-50	Kinetics	8	1	89.37	46.05	56.8
Slow-only	3D ResNet-50	Kinetics	8	3	89.92	45.50	58.5
SlowFast	3D ResNet-50	Kinetics	4 + 32	1	**93.73**	51.23	57.8
SlowFast	3D ResNet-50	Kinetics	4 + 32	3	93.46	**52.04**	58.9
SlowFast	3D ResNet-50	Kinetics	8 + 32	1	88.56	44.14	64.6
SlowFast	3D ResNet-50	Kinetics	8 + 32	3	88.28	45.23	**65.0**

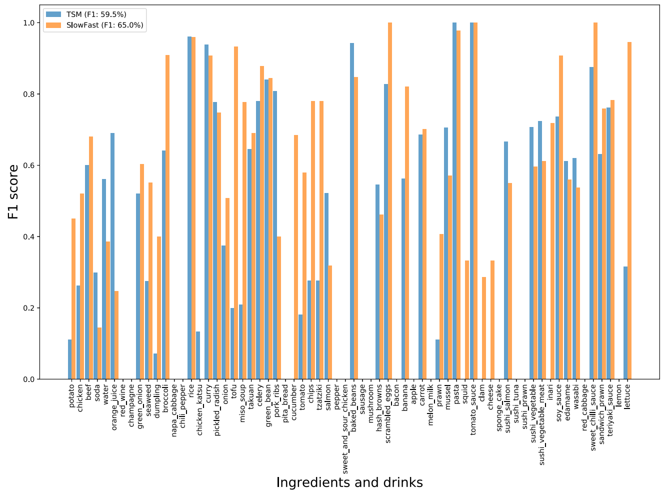

Fig. 8. F1 scores calculated separately for each food item in order to better understand which items are well recognized.

D. Results of Recognizing Consumed Food Items

Table V summarizes the results of recognizing consumed food items. All models were trained with only RGB modality. Same to the outcomes of bite counting, using more frames generally leads to higher accuracy, as it carries more spatial and temporal information. Using 16 frames as the input to a 2D ResNet-50 network with TSM embedded has an 18% increase in F1 score compared to using 8 frames. We also implemented and trained a two-head 2D ResNet-50 which also has TSM embedded, one head is for recognizing all visible food items in a clip and the other for recognizing the consumed ones. The losses from these two heads were summed and backpropagated to update network's parameters. This was motivated by the fact that the consumed food items should only be the ones visible in a clip, so by being aware of visible food items in the clip from one head, the other head ideally could better recognize consumed food items. The result shown in the 3 rd row indicates this design is effective (a 0.8% increase compared to a single head).

TABLE V
OVERALL RESULTS OF RECOGNIZING CONSUMED FOOD ITEMS

Model	Backbone	Pretrain	#Frame	#Crop	F1
TSM	2D ResNet-50	Kinetics	8	1	17.5
TSM	2D ResNet-50	Kinetics	16	1	35.5
$\text{TSM}_{\text{Two-Head}}$	2D ResNet-50	Kinetics	16	1	36.3
Slow-only	3D ResNet-50	Kinetics	8	1	32.7
Slow-only	3D ResNet-50	Kinetics	8	3	34.7
SlowFast	3D ResNet-50	Kinetics	4 + 32	1	31.0
SlowFast	3D ResNet-50	Kinetics	4 + 32	3	33.3
SlowFast	3D ResNet-50	Kinetics	8 + 32	1	36.8
SlowFast	3D ResNet-50	Kinetics	8 + 32	3	38.6
TSM	2D ResNet-50	Kinetics	16	1	35.7
$\text{TSM}_{\text{Two-Head}}$	2D ResNet-50	Kinetics	16	1	**37.8**
SlowFast	3D ResNet-50	Kinetics	8 + 32	1	38.6
SlowFast	3D ResNet-50	Kinetics	8 + 32	3	**40.5**

The results from the middle 6 rows (Slow-only and SlowFast) verify that recognizing consumed food items requires a network capable of capturing both the action of taking a bite and the appearance of that bite. Therefore, sufficient visual (comparing the results of SlowFast 8 + 32 and SlowFast 4 + 32) and motion (comparing the results of SlowFast 8 + 32 and Slow-only) clues are important, but visual clues seem to play a more important role in recognizing consumed food items as the Slow-only model trained with 8 frames shows higher F1 scores than the SlowFast model trained with 4 (slow) + 32 (fast) frames. We also used the dataset for general food recognition to train and test TSM, $\text{TSM}_{\text{Two-Head}}$, and SlowFast. As the dataset is slightly larger, the resulting testing accuracy is also higher as shown in the bottom 4 rows. The SlowFast (8 + 32) model, based on 3D ResNet-50, produces the highest F1 score of 40.5% when tested with 3 crops, and using a 2D ResNet-50 with TSM, the highest F1 score is 37.8%. The F1 score of each food item is shown in Figure 9. Although the overall F1 score of the SlowFast model is only 2.7% higher than that of the TSM model, the visualized F1 scores for each individual food item show that the SlowFast model was able to recognize a wider range of food items that

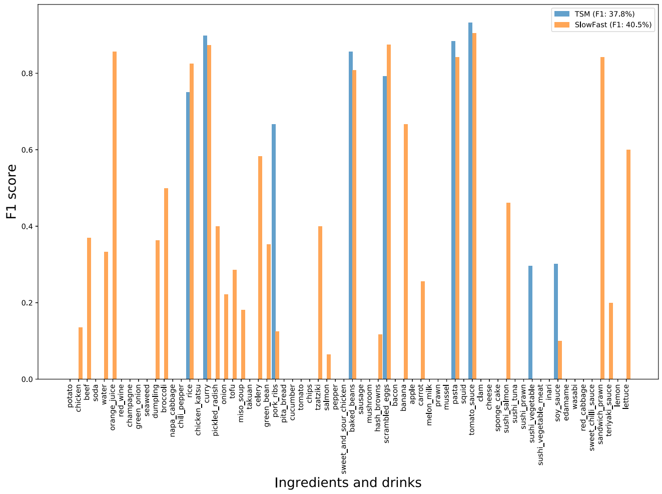

Fig. 9. F1 scores calculated separately for each food item in order to better understand which **consumed** items are well recognized.

were consumed by the subjects, with 29 food items having a non-zero F1 score, and 8 of them having the F1 score over 80%.

E. Visualization of Multi-Label Recognition

Figure 11 shows some samples of the results of recognizing visible and consumed food items on the test set. The top 4 rows are the samples of recognizing visible food items. The SlowFast model generally can recognize more true positives than the TSM model. For example, all ingredients and drinks visible in the clips shown in the 2nd and 3 rd rows are correctly recognized by the SlowFast model. In the 4th row, however, while the SlowFast recognizes 7 out of 8 food items correctly, it misrecognizes the main ingredient *chicken* as *salmon*. This misrecognition of the main ingredient can also be observed in the results of the TSM model (please refer to the 1st and 4th rows), which also explains the results of classifying a meal into 18 fine-grained classes being not satisfactory, as it is difficult for the models to capture the main ingredient that distinguishes between similar meals.

In the case of recognizing consumed food items (bottom 4 rows in Figure 11), the SlowFast model also recognizes more true positives. It is worth noting that in the 5th row, although the SlowFast model fails to recognize that the subject has eaten *pickled_radish*, it still outputs a close ingredient that the subject is estimated to have eaten (i.e., *carrot*). In the 6th row, it is encouraging that the SlowFast model is able to recognize that the subject has eaten *celery* and *green_bean* even though these two ingredients appear to be so small and close to each other in the clip.

VI. Discussion

Although this work offers a new insight into using egocentric videos to count bites and recognize consumed food items for dietary assessments, there are some areas that this work has not investigated. First, in this work, counting bites was implemented as an end-to-end manner (i.e., given a set of sampled frames from a video clip as the input, the network directly outputs the estimated number of bites taken). Another alternative way of counting bites in a video clip is by explicitly tracking and analyzing the movements of hands, which is more complicated but may lead to better results. We also show preliminary results of counting bites for an entire meal in Figure 10. Our current solution was first using regression to estimate bite counts in 20-second intervals, and then aggregating the bite counts from these intervals to produce an overall estimation for the entire meal. Tested on 3 unseen subjects (the other 9 subjects were used for training), it yields satisfactory results, with an average error of 7.76 bites when directly aggregating initial values of regression from 20-second intervals, and an average error of 7.71 bites when aggregating integer bite counts after rounding the initial values (the subjects take an actual average of 42.57 bites). It is worth noting that in practice, it is more appropriate to solve bite counting as a regression problem as the number of bites people take in fixed intervals becomes uncertain, which makes classification unlikely. As we cut an dietary intake video into a set of short intervals, the issue of bites happening at interval cut points could affect the overall performance. In the current dataset, bite counts are labelled as integers, and we excluded clips that have an incomplete bite taking event at the start or end of the clip that misses the moment of fully placing food into the mouth. The effect of this issue can be mitigated by labelling bite counts as decimals (e.g., 3.3 and 2.7 bites for 2 adjacent clips if a bite happens at the cut point), and we conjecture the accuracy of estimating the bite count of an entire meal will further increase. Such labelling will be investigated in our future work. In addition, the estimated bite counts of video clips may possibly complement each other when aggregated for an entire meal. This may have contributed to the high accuracy of entire meal bite estimation in some cases. However, the bite counting accuracy of individual clips is still important and needs to be

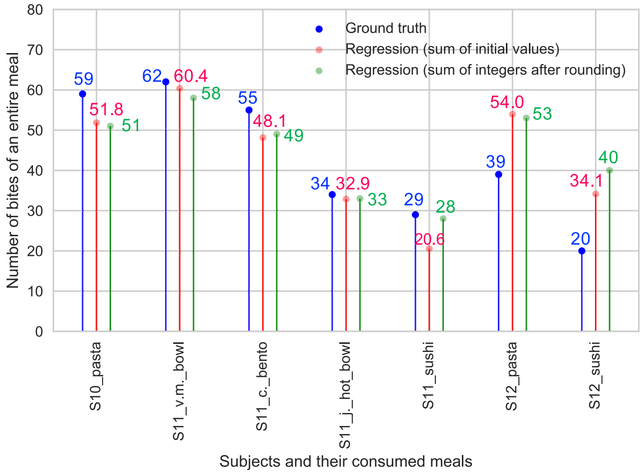

Fig. 10. Estimation of the number of bites taken in an entire meal for subjects #10, #11, and #12 (we chose these 3 subjects because they show the best accuracy in the inter-subject cross validation). Results are reported using a TSM model (2D ResNet-50, Something-V2, RGB, 16, 1) trained on the other 9 subjects (874 clips out of the whole 1,022 clips) with ℓ_2 loss. The entire videos of the reported 3 subjects were used (i.e., none of their split clips were abandoned.)

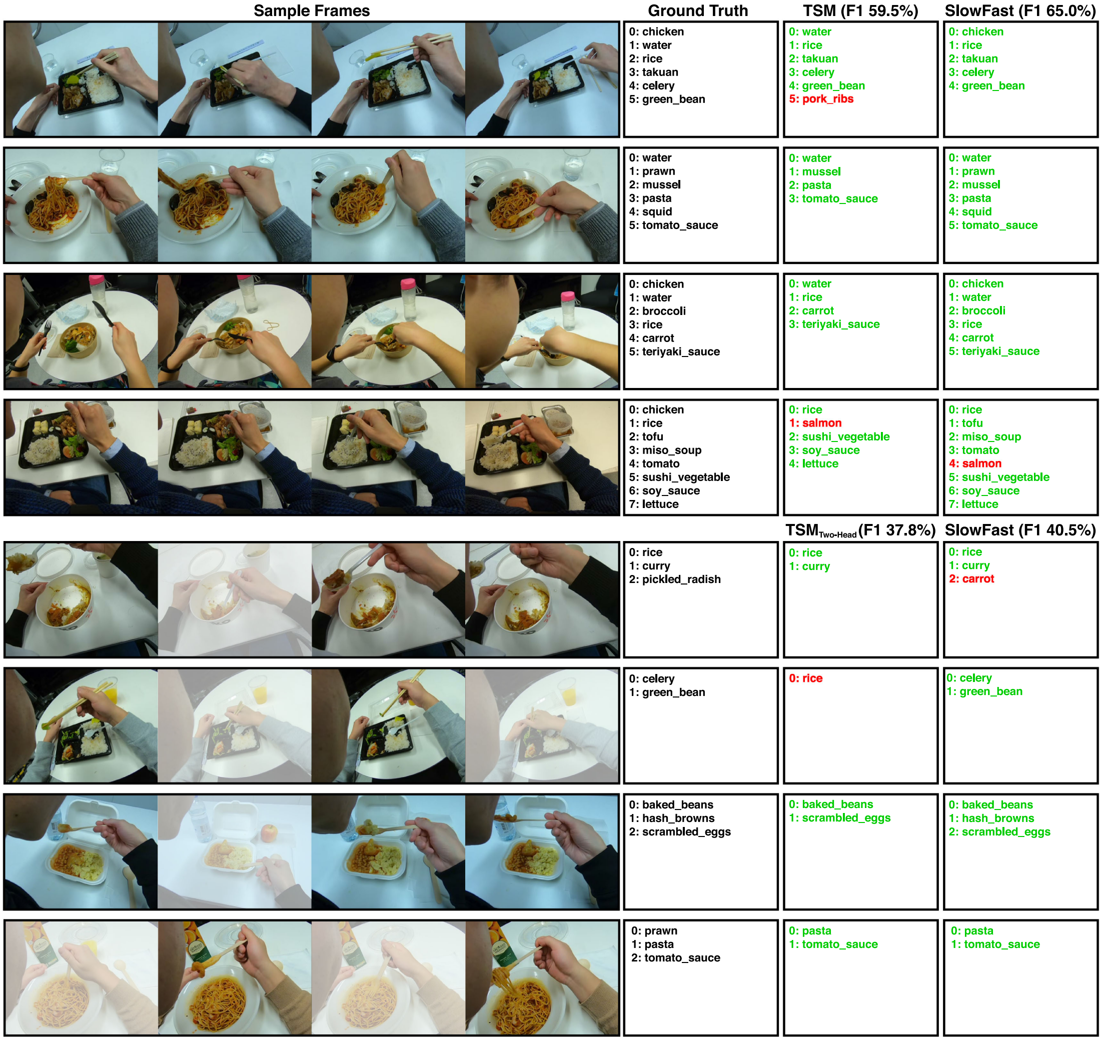

Fig. 11. Recognized food items (ingredients and drinks). Top 4 rows are samples of recognizing visible food items and bottom 4 rows are samples of recognizing **consumed** food items in a clip. True positives are indicated using green color and false positives are in red color.

further improved in the future in order to make estimation for entire meals more robust. In addition, in this work, we explicitly ensured that the number of frames sampled as the input to the network was more than the number of bites taken so that the network has sufficient information to estimate bite counts. An appropriate sampling rate still needs to be investigated, if the dataset expands and includes more subjects with different eating speeds. Second, given the fact that all deep network models were only trained with weak supervision for food recognition (i.e., no bounding boxes or masks provided), although the results so far are reasonable, we conjecture that better results could be achieved by 1) labelling consumed food or all visible food items with bounding boxes or masks, or 2) using categorical labels or visual attention techniques to localize food items [33], [52]. Third, this work does not investigate bite size estimation. Estimating bite size is an essential part of automatic dietary assessments, and this needs to be investigated in future research. Fourth, in this work, recognizing consumed food items is at a whole video clip level. Recognizing what subjects take in each individual bite may provide more fine-grained information, and benefit bite size estimation. To achieve this, temporally localizing each bite and then recognizing what that bite contains is one way. Another alternative way is, similar to recipe generation from images [71], to decode consumed food items in chronological order from video clips, but this requires consumed food items in a video clip also be annotated in chronological order (e.g., an annotation could be like [rice, [SEP], rice, chicken, [SEP], celery, [END]] where [SEP] and [END] are special tokens to separate bites and to indicate the end of prediction, respectively). Although this alternative way can recognize consumed food items and associate them with each individual bite (bite count is also obtained) in an end-to-end manner, its efficacy needs to be validated in future work. Despite not to the fine-grained food item level, a very recent work using CTC loss and deep networks has shown the success in simultaneously detecting intake events and classifying them as eating or drinking in both video and inertial data [72]. Fifth, counting the number of times of drinking has not been investigated in this work, although drinks have been considered as one of food items and included in food recognition. To produce a more comprehensive estimate of overall food consumption, counting drinking times is also important. As drinking occurs far less times than taking bites in the current dataset, a dataset contains sufficient drinking samples is needed, and we leave this to future work. Sixth, as the data used in this work were collected from a laboratory setting, which may be generalized to a hospital setting, assessing dietary intake in-the-wild with videos still needs further investigation.

As the camera is mounted on the shoulder, its captured videos also contain audio signals that are useful for dietary assessment, such as the sound of chewing and swallowing during eating. Thus, it is also worth investigating the fusion of visual and audio signals from egocentric videos, which may yield better accuracy in dietary intake assessments.

VII. Conclusion

In this work, we have proposed to count the number of bites and recognize consumed food items in egocentric videos for passive dietary intake monitoring. Experimental results show that an end-to-end manner of bite counting and consumed food recognition is feasible with the use of deep neural networks. However, consumed food item recognition is still challenging compared to conventional visible food item recognition or meal classification. Using videos as the source for dietary intake assessments is a promising option, but efforts to improve accuracy are still needed. Building on this work, our future plans are to expand current dataset, address limitations mentioned in the discussion section, and also to design new models that produce more accurate assessments.

VII. Acknowledgment

This work is supported by the Innovative Passive Dietary Monitoring Project funded by the Bill & Melinda Gates Foundation (Opportunity ID: OPP1171395).

References

[1] B. M. Popkin, C. Corvalan, and L. M. Grummer-Strawn, “Dynamics of the double burden of malnutrition and the changing nutrition reality,” *The Lancet*, 2019.

[2] WHO *et al.*, “Unicef/who/the world bank group joint child malnutrition estimates: levels and trends in child malnutrition: key findings of the 2019 edition,” 2019.

[3] J.-S. Shim, K. Oh, and H. C. Kim, “Dietary assessment methods in epidemiologic studies,” *Epidemiology Health*, vol. 36, 2014.

[4] N. Alshurafa *et al.*, “Counting bites with bits: Expert workshop addressing calorie and macronutrient intake monitoring,” *J. Med. Int. Res.*, vol. 21, no. 12, p. e14904, 2019.

[5] Y. Dong, A. Hoover, J. Scisco, and E. Muth, “A new method for measuring meal intake in humans via automated wrist motion tracking,” *Appl. Psychophysiology and Biofeedback*, vol. 37, no. 3, pp. 205–215, 2012.

[6] Y. Dong, J. Scisco, M. Wilson, E. Muth, and A. Hoover, “Detecting periods of eating during free-living by tracking wrist motion,” *IEEE J. Biomed. Health Inf.*, vol. 18, no. 4, pp. 1253–1260, 2013.

[7] E. Thomaz, I. Essa, and G. D. Abowd, “A practical approach for recognizing eating moments with wrist-mounted inertial sensing,” in *Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput.*, 2015, pp. 1029–1040.

[8] S. Zhang, R. Alharbi, W. Stogin, M. Pourhomayun, B. Spring, and N. Alshurafa, “Food watch: detecting and characterizing eating episodes through feeding gestures,” in *Proc. 11th EAI Int. Conf. Body Area Netw.*, 2016, pp. 91–96.

[9] Y. Shen, J. Salley, E. Muth, and A. Hoover, “Assessing the accuracy of a wrist motion tracking method for counting bites across demographic and food variables,” *IEEE J. Biomed. Health Inf.*, vol. 21, no. 3, pp. 599–606, 2016.

[10] S. Zhang, R. Alharbi, M. Nicholson, and N. Alshurafa, “When generalized eating detection machine learning models fail in the field,” in *Proc. ACM Int. Joint Conf. Pervasive and Ubiquitous Comput. Proc. ACM Int. Symp. Wearable Comput.*, 2017, pp. 613–622.

[11] S. Zhang, W. Stogin, and N. Alshurafa, “I sense overeating: Motif-based machine learning framework to detect overeating using wrist-worn sensing,” *Inf. Fusion*, vol. 41, pp. 37–47, 2018.

[12] K. Kyritsis, C. Diou, and A. Delopoulos, “Modeling wrist micromovements to measure in-meal eating behavior from inertial sensor data,” *IEEE J. Biomed. Health Inf.*, vol. 23, no. 6, pp. 2325–2334, 2019.

[13] K. Kyritsis, C. Diou, and A. Delopoulos, “A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smartwatches,” *IEEE J. Biomed. Health Inf.*, 2020.

[14] W. Zhang, Q. Yu, B. Siddiquie, A. Divakaran, and H. Sawhney, “asnap-n-eata food recognition and nutrition estimation on a smartphone,” *J. Diabetes Sci. Technol.*, vol. 9, no. 3, pp. 525–533, 2015.

[15] Y. Kawano and K. Yanai, “Foodcam: A real-time food recognition system on a smartphone,” *Multimedia Tools Appl.*, vol. 74, no. 14, pp. 5263–5287, 2015.

[16] D. Ravì, B. Lo, and G.-Z. Yang, “Real-time food intake classification and energy expenditure estimation on a mobile device,” in *Proc. IEEE 12th Int. Conf. Wearable Implantable Body Sensor Netw.* ., 2015, pp. 1–6.

[17] Y. Jang, Y. Song, Y. Yu, Y. Kim, and G. Kim, “Tgif-qa: Toward spatio-temporal reasoning in visual question answering,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2758–2766.

[18] J. Gao, R. Ge, K. Chen, and R. Nevatia, “Motion-appearance co-memory networks for video question answering,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 6576–6585.

[19] C. Fan, X. Zhang, S. Zhang, W. Wang, C. Zhang, and H. Huang, “Heterogeneous memory enhanced multimodal attention model for video question answering,” in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 1999–2007.

[20] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[21] S. Bi *et al.*, “Auracle: Detecting eating episodes with an ear-mounted sensor,” *Proc. ACM Int., Mobile, Wearable Ubiquitous Technol.*, vol. 2, no. 3, pp. 1–27, 2018.

[22] E. S. Sazonov, O. Makeyev, S. Schuckers, P. Lopez-Meyer, E. L. Melanson, and M. R. Neuman, “Automatic detection of swallowing events by acoustical means for applications of monitoring of ingestive behavior,” *IEEE Trans. Biomed. Eng.*, vol. 57, no. 3, pp. 626–633, Mar. 2009.

[23] T. Olubanjo and M. Ghovanloo, “Real-time swallowing detection based on tracheal acoustics,” in *Proc. IEEE Int. Conf. Acoustics, Speech Signal Process.* ., 2014, pp. 4384–4388.

[24] K. Yatani and K. N. Truong, “Bodyscope: a wearable acoustic sensor for activity recognition,” in *Proc. ACM Conf. Ubiquitous Comput.*, 2012, pp. 341–350.

[25] T. Rahman *et al.* “Bodybeat: a mobile system for sensing non-speech body sounds.” in *MobiSys*, vol. 14, no. 10.1145. Citeseer, 2014, pp. 2 594 368–2 594 386.

[26] V. Papapanagiotou, C. Diou, L. Zhou, J. van den Boer, M. Mars, and A. Delopoulos, “A novel chewing detection system based on ppg, audio, and accelerometry,” *IEEE J. Biomed. Health Inf.*, vol. 21, no. 3, pp. 607–618, 2016.

[27] A. Bedri *et al.*, “Earbit: using wearable sensors to detect eating episodes in unconstrained environments,” *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 1, no. 3, pp. 1–20, 2017.

[28] E. S. Sazonov and J. M. Fontana, “A sensor system for automatic detection of food intake through non-invasive monitoring of chewing,” *IEEE Sen,. J.*, vol. 12, no. 5, pp. 1340–1348, May 2011.

[29] R. Zhang and O. Amft, “Monitoring chewing and eating in free-living using smart eyeglasses,” *IEEE J. Biomed. Health Inf.*, vol. 22, no. 1, pp. 23–32, 2017.

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

[31] 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.*, 2015, pp. 1–6.

[32] N. Martinel, G. L. Foresti, and C. Micheloni, “Wide-slice residual networks for food recognition,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 567–576.

[33] J. Qiu, F. P.-W. Lo, Y. Sun, S. Wang, and L. Benny, “Mining discriminative food regions for accurate food recognition,” in *Proc. Brit. Mach. Vision Conf.*, 2019, pp. 588–598.

[34] F. Zhu, M. Bosch, N. Khanna, C. J. Boushey, and E. J. Delp, “Multiple hypotheses image segmentation and classification with application to dietary assessment,” *IEEE J. Biomed. Health Inf.*, vol. 19, no. 1, pp. 377–388, 2014.

[35] Y. Wang, C. Liu, F. Zhu, C. J. Boushey, and E. J. Delp, “Efficient superpixel based segmentation for food image analysis,” in *Proc. IEEE Int. Conf. Image Process.* ., 2016, pp. 2544–2548.

[36] W. Shimoda and K. Yanai, “Cnn-based food image segmentation without pixel-wise annotation,” in *Proc. Int. Conf. Image Anal. Process.*. Springer, 2015, pp. 449–457.

[37] Y. Wang, F. Zhu, C. J. Boushey, and E. J. Delp, “Weakly supervised food image segmentation using class activation maps,” in *Proc. IEEE Int. Conf. Image Process.* ., 2017, pp. 1277–1281.

[38] F. P.-W. Lo, Y. Sun, J. Qiu, and B. Lo, “Food volume estimation based on deep learning view synthesis from a single depth map,” *Nutrients*, vol. 10, no. 12, p. 2005, 2018.

[39] P. W. Lo, Y. Sun, J. Qiu, and B. Lo, “Point2volume: A vision-based dietary assessment approach using view synthesis,” *IEEE Trans. Ind. Inf.*, vol. 16, no. 1, Jan. 2020.

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

[41] J. Liu, E. Johns, L. Atallah, C. Pettitt, B. Lo, G. Frost, and G.-Z. Yang, “An intelligent food-intake monitoring system using wearable sensors,” in *Proc. Ninth Int. Conf. Wearable Implantable Body Sen. Net.*, 2012, pp. 154–160.

[42] A. Meyers *et al.*“Im2calories: towards an automated mobile vision food diary,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1233–1241.

[43] O. Beijbom, N. Joshi, D. Morris, S. Saponas, and S. Khullar, “Menu-match: Restaurant-specific food logging from images,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2015, pp. 844–851.

[44] W. Min, S. Jiang, J. Sang, H. Wang, X. Liu, and L. Herranz, “Being a supercook: Joint food attributes and multimodal content modeling for recipe retrieval and exploration,” *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1100–1113, May 2017.

[45] A. Salvador, N. Hynes, Y. Aytar, J. Marin, F. Ofli, I. Weber, and A. Torralba, “Learning cross-modal embeddings for cooking recipes and food images,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 3020–3028.

[46] M. Carvalho, R. Cadène, D. Picard, L. Soulier, N. Thome, and M. Cord, “Cross-modal retrieval in the cooking context: Learning semantic text-image embeddings,” in *Proc. 41st Int. ACM SIGIR Conf. Res. & Develop. Inf. Retrieval*, 2018, pp. 35–44.

[47] J. Marin *et al.*, “Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, 2019.

[48] H. Wang, D. Sahoo, C. Liu, E.-p. Lim, and S. C. Hoi, “Learning cross-modal embeddings with adversarial networks for cooking recipes and food images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 572–11 581.

[49] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, “R2gan: Cross-modal recipe retrieval with generative adversarial network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11 477–11 486.

[50] J. Qiu, F. P.-W. Lo, and B. Lo, “Assessing individual dietary intake in food sharing scenarios with a 360 camera and deep learning,” in *Proc. IEEE 16th Int. Conf. Wearable Implantable Body Sensor Netw.* ., 2019, pp. 1–4.

[51] P. V. Rouast and M. T. Adam, “Learning deep representations for video-based intake gesture detection,” *IEEE J. Biomed. Health Inf.*, vol. 24, no. 6, Jun. 2020.

[52] W. Min, L. Liu, Z. Luo, and S. Jiang, “Ingredient-guided cascaded multi-attention network for food recognition,” in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 1331–1339.

[53] J. Chen and C.-W. Ngo, “Deep-based ingredient recognition for cooking recipe retrieval,” in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 32–41.

[54] A. Doulah and E. Sazonov, “Clustering of food intake images into food and non-food categories,” in *Int. Conf. Bioinformatics Biomed. Eng.*. Springer, 2017, pp. 454–463.

[55] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, “A survey on food computing,” *ACM Comput. Surv.*, vol. 52, no. 5, pp. 1–36, 2019.

[56] F. P. W. Lo, Y. Sun, J. Qiu, and B. Lo, “Image-based food classification and volume estimation for dietary assessment: A review,” *IEEE J. Biomed. Health Inf.*, vol. 24, no. 7, pp. 1926–1939, 2020.

[57] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Adv. Neural Inf. Process. Syst.*, 2014, pp. 568–576.

[58] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, “Temporal segment networks: Towards good practices for deep action recognition,” in *Proc. Eur. Conf. Comput. Vis*. Springer, 2016, pp. 20–36.

[59] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, “Temporal relational reasoning in videos,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 803–818.

[60] J. Donahue *et al.*“Long-term recurrent convolutional networks for visual recognition and description,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 2625–2634.

[61] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[62] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[63] J. Carreira and A. Zisserman, “Quo vadis, action recognition? a new model and the kinetics dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6299–6308.

[64] J. Lin, C. Gan, and S. Han, “Tsm: Temporal shift module for efficient video understanding,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 7083–7093.

[65] C. Feichtenhofer, H. Fan, J. Malik, and K. He, “Slowfast networks for video recognition,” in *Proc. IEEE Int. Conf. Comput. Vision*, 2019, pp. 6202–6211.

[66] W. Kay *et al.*, “The kinetics human action video dataset,” 2017, *arXiv:1705.06950*.

[67] C. Zach, T. Pock, and H. Bischof, “A duality based approach for realtime tv-l 1 optical flow,” in *Proc. Joint Pattern Recognit. Symp*. Springer, 2007, pp. 214–223.

[68] O. Russakovsky *et al.*, “Imagenet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

[69] R. Goyal *et al.*, “The “something something” video database for learning and evaluating visual common sense,” in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017.

[70] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[71] A. Salvador, M. Drozdzal, X. Giro-i Nieto, and A. Romero, “Inverse cooking: Recipe generation from food images,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10 453–10 462.

[72] P. V. Rouast and M. T. Adam, “Single-stage intake gesture detection using ctc loss and extended prefix beam search,”2020, *arXiv:2008.02999*.