Egocentric Image Captioning for Privacy-Preserved Passive Dietary Intake Monitoring

Jianing Qiu, Frank P.-W. Lo, Xiao Gu, *Graduate Student Member, IEEE*, Modou L. Jobarteh, Wenyan Jia, Tom Baranowski, Matilda Steiner-Asiedu, Alex K. Anderson, Megan A. Mccrory, Edward Sazonov, *Senior Member, IEEE*, Mingui Sun, *Life Senior Member, IEEE*, Gary Frost, and Benny Lo, *Senior Member, IEEE*

Abstract—Camera-based passive dietary intake monitoring is able to continuously capture the eating episodes of a subject, recording rich visual information, such as the type and volume of food being consumed, as well as the eating behaviors of the subject. However, there currently is no method that is able to incorporate these visual clues and provide a comprehensive context of dietary intake from passive recording (e.g., is the subject sharing food with others, what food the subject is eating, and how much food is left in the bowl). On the other hand, privacy is a major concern while egocentric wearable cameras are used for capturing. In this article, we propose a privacy-preserved secure solution (i.e., egocentric image captioning) for dietary assessment with passive monitoring, which unifies food recognition, volume estimation, and scene understanding. By converting images into rich text descriptions, nutritionists can assess individual dietary intake based on the captions instead of the original images, reducing the risk of privacy leakage from images. To this end, an egocentric dietary image captioning dataset has been built, which consists of in-the-wild images captured by head-worn and chest-worn cameras in field studies in Ghana. A novel transformer-based architecture is designed to caption egocentric dietary images. Comprehensive experiments have been conducted to evaluate the effectiveness and to justify the design of the proposed architecture for egocentric dietary image captioning. To the best of our knowledge, this is the first work that applies image captioning for dietary intake assessment in real-life settings.

Index Terms—Egocentric vision, image captioning, passive dietary intake monitoring.

Jianing Qiu and Xiao Gu are with the Department of Computing and the Hamlyn Centre, Imperial College London, SW7 2AZ London, U.K. (e-mail: jianing.qiu17@imperial.ac.uk; xiao.gu17@imperial.ac.uk).

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

Modou L. Jobarteh and Gary Frost are with the Section for Nutrition Research, Department of Metabolism, Digestion and Reproduction, Imperial College London, SW7 2AZ London, U.K. (e-mail: m.jobarteh@imperial.ac.uk; g.frost@imperial.ac.uk).

Wenyan Jia and Mingui Sun are with the Department of Neurological Surgery, University of Pittsburgh, Pittsburgh, PA 15260 USA (e-mail: jiawenyan@gmail.com; drsun@pitt.edu).

Tom Baranowski is with the USDA/ARS Children's Nutrition Research Center, Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030 USA (e-mail: tom.baranowski@bcm.edu).

Matilda Steiner-Asiedu is with the Department of Nutrition and Food Science, University of Ghana, Accra, Ghana (e-mail: tillysteiner@gmail.com).

Alex K. Anderson is with the Department of Foods and Nutrition, University of Georgia, Athens, GA 30602 USA (e-mail: fianko@uga.edu).

Megan A. Mccrory is with the Department of Health Sciences, Boston University, Boston, MA 02215 USA (e-mail: mamccr@bu.edu).

Edward Sazonov is with the Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL 35487 USA (e-mail: esazonov@eng.ua.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCYB.2023.3243999.

Digital Object Identifier 10.1109/TCYB.2023.3243999

I. Introduction

EFFECTIVE dietary intake monitoring allows nutritionists to better understand the diet patterns and nutritional needs of a population, and also allows policy makers to better plan and evaluate nutritional health policy and public health interventions. In nutritional epidemiology, 24-h dietary recall and food frequency questionnaires (FFQs) are the primary dietary assessment tools [1]. Although widely used, they mainly rely on the subjects' memory to recall their past dietary intake, and require nutritionists to collect, analyze, and interpret the dietary data. Thus, these traditional methods are often labor-intensive, inefficient, with the resulting dietary assessment being inaccurate. For these reasons, technological approaches have been developed to automate the dietary intake assessment process and provide objective, more accurate results.

Existing technological approaches for dietary assessment can be categorized as active or passive approaches. Active approaches often require subjects to actively log their dietary intake using tools, such as a smart phone to manually enter the food type and estimated portion size [4], [5], [6]. As it still relies on memory and volitional inputs, the subjectivity of dietary data still exists, and like traditional dietary assessment tools, users often under-report their food intake. In addition, without complete visual recording (even if a few food images are taken at the beginning and end of the meal), active approaches lose the subject's eating details. Such information is crucial for understanding the eating behaviors of the subject, as well as for recognizing the food items and ingredients in a more fine-grained level, as some hidden and occluded ingredients may be revealed during eating [7]. Passive approaches on the other hand use wearable sensors, such as wearable cameras, to monitor, and detect dietary intake. Once the wearable camera is initiated, passive approaches provide pervasive and continuous dietary intake monitoring. Compared

Human annotated captions

1. the subject is having a meal with the family, and the bowl is half empty
2. the subject is eating rice with egg stew, and sharing it with other family members
3. the subject is eating a half bowl of rice with egg stew with other family members
4. the subject is eating rice with egg stew with the family, and the bowl is half empty

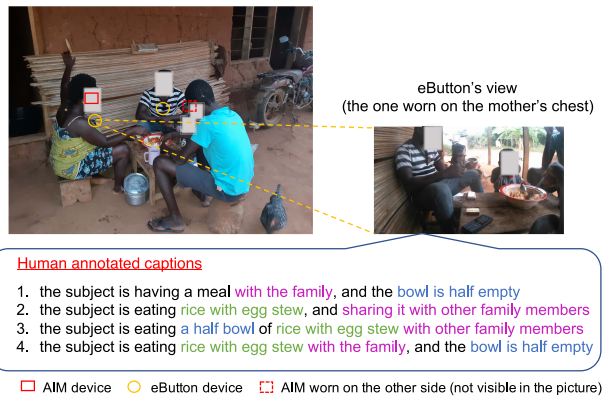☐ AIM device   ○ eButton device   ⬚ AIM worn on the other side (not visible in the picture)

Fig. 1. *Top left:* Food sharing scenario in one household. Wearable cameras were worn by the subjects during the meal. The mother was wearing an AIM [2] and also an eButton [3] device. The father was wearing an eButton. The adolescent child was wearing an AIM. *Top right:* One egocentric image captured by the eButton worn on the mother's chest. *Bottom:* Human-annotated captions for the top-right image. Each caption has dietary-related information in different levels of detail, such as the type of food the subject was eating (*rice with egg stew* in this example), the portion size of the food that can be seen from the image (e.g., *the bowl is half empty*), and whether the subject is sharing food. Faces are masked to protect privacy.

to active approaches, passive ones are designed to capture the entire eating episode and therefore the recorded data are more detailed and comprehensive. As it does not require active participation from the subjects, assessing dietary intake passively is more objective. With the advances in the wearable technology, wearable cameras are becoming cheaper and more reliable. Recent progress in deep learning also enhances the accuracy of computer vision-based food recognition [8], [9], [10], [11], [12], recipe retrieval [13], [14], [15], [16], [17], [18], [19], and food volume estimation [20], [21], [22]. Using wearable cameras to passively monitor dietary intake is therefore a low-cost and effective means for objective and accurate dietary intake assessment.

Nevertheless, in using wearable cameras to record egocentric (first person) images/videos, privacy concerns are the major barrier to the wide acceptability and deployment of wearable cameras in the general public. In the case of dietary intake monitoring, the wearable cameras may capture things that the subject does not want to be captured, for example, the room arrangement, or the faces of other family members who are eating with the subject. We therefore propose to use image captioning to convert each captured dietary image into a rich text description that summarizes the content of the image for dietary assessment. As such, nutritionists can assess the dietary intake of a subject from the text description instead of interpreting from the original image. In addition, the method can potentially be built into the wearable camera, and the device will then only store the text descriptions rather than images, requiring less storage space for each dietary intake recording. This can preserve the subject's privacy, and can also be further extended to generate a nutrient intake report, similar to the medical report generation [23], [24], [25], to give subjects automatic and prompt feedback on their dietary intake.

In dietary assessment, food volume is essential for quantifying the actual dietary intake. However, directly estimating food volume from an RGB image is difficult as food has various and irregular shapes, and food items/ingredients are often occluded. Researchers have tried to estimate food volume from an RGB image by first estimating the size of the food container [26], but this still is not the actual food volume. In this work, we annotate dietary intake images with food portion size measured with respect to its container, for example, half bowl, two bowls, and the bowl is empty. The parsed food portion size information from an image's caption can then be jointly used with the estimated container size to estimate the actual food volume. The details on estimating food volume by combining dietary image captioning and 3-D container reconstruction is shown in Section V-C8. Apart from the portion size information, our annotation also includes the type of the food a subject is eating, for example, okra and plantain, and the action the subject is performing, for example, cut and drink. We present the results of portion size estimation, food recognition, and action recognition by parsing the model-generated image captions in Section V-C5.

To the best of our knowledge, this is the first work that applies image captioning to dietary intake assessment. The contributions of this work are as follows.

1) An egocentric image captioning dataset has been built, which contains in-the-wild dietary images captured by wearable cameras. To the best of our knowledge, this is the first and also the largest egocentric image captioning dataset in both the computer vision and nutritional science communities for dietary assessment. Fig. 1 shows an example of an eating scenario in the field as well as the annotated captions of one egocentric dietary image.

2) A novel transformer-based captioning model has been designed to generate the captions for dietary images. Extensive experiments have been conducted, including captioning egocentric images where domain difference exists due to different viewing angles of the wearable cameras. Our proposed model has also been tested on a public egocentric image captioning dataset. The results show that our model is also effective in captioning egocentric images under other daily life settings.

3) A novel framework of food volume estimation has been proposed, which combines dietary image captioning with 3-D container reconstruction to estimate actual food volume. This novel framework takes the advantage of passive monitoring and image captioning to identify empty containers for better 3-D reconstruction and leverages portion size information parsed from the caption to jointly conduct food volume estimation.

4) To benchmark and quantitatively validate the food volume estimation, we construct another egocentric RGBD dietary intake dataset in a controlled laboratory setting, which contains ground truth food and container volumes, a depth map of each image frame, segmentation mask(s) of the food container(s), and the detailed caption that indicates the eating state of the image frame. Compared to our in-the-wild dataset mentioned above, captions in this laboratory dataset further include more fine-grained food portion size information, for example, a 3/4 bowl of. To the best of our knowledge, our RGBD dataset is the first close-range depth sensing dataset in the research community, that is, precise depth was captured even when the object was as 7-cm close as to the sensor. This novel egocentric RGBD dataset can therefore also

advance future depth-related research beyond food volume estimation. Code and both datasets are available upon request.

The remainder of this article is organized as follows: Section II briefly reviews prior work in passive dietary intake monitoring and image captioning; Section III describes the proposed method in this work; Section IV presents the details of the constructed egocentric dietary image captioning (EgoDIMCAP) dataset; Section V shows the experimental results, including the results of image captioning and those of its derived downstream tasks; and we conclude and discuss some potential future directions in Section VI.

II. RELATED WORK

In this section, we mainly discuss prior work in passive dietary monitoring and image captioning, as they are most relevant to our work.

A. Passive Dietary Monitoring

Based on the type of the wearable sensors used, passive dietary monitoring can be mainly categorized as inertial-based, acoustic-based, or visual-based. Our method is proposed for visual-based passive dietary monitoring.

Inertial-based monitoring systems are able to detect eating episodes and feeding gestures as well as to count overall bites taken in a meal by collecting and analyzing inertial measurement unit (IMU) signals recorded by a wrist-worn device [27], [28], [29], [30], [31], [32], [33], [34], [35]. Acoustic-based monitoring systems, are able to differentiate eating from other daily activities [36], [37], as well as to detect swallows [38], [39] and eating episodes [40] with acoustic signals alone. Studies in [41] and [42] have also examined the effectiveness of combining inertial and acoustic signals in dietary monitoring. In visual-based passive dietary monitoring, wearable cameras are widely used. In [3], a chest-worn camera called eButton was designed for passive dietary monitoring. Once initiated, the eButton continuously captures egocentric images of eating at fixed-time intervals. In [43], an ear-worn dietary intake monitoring device was introduced, in which a miniaturized camera was triggered to record eating episodes so long as the chewing sound was detected by the built-in sensor. In [7], a GoPro camera was mounted on a subject's shoulder to record his/her eating episodes. The number of bites taken and the type of food consumed was then end-to-end deduced from the recorded egocentric dietary intake video. Recently, 360° cameras have been used to monitor and assess dietary intake in food sharing [44], [45] or communal eating [46] scenarios in a passive way. Although visual-based monitoring can lead to more comprehensive dietary assessment, the use of the cameras to record often entails privacy issues.

B. Image Captioning

Image captioning is a cross-modal task, which describes the content of an image with one or a few sentences. Before the deep learning era, early work in image captioning mainly used template- or rule-based approaches [47], [48], [49]. With the advances in deep learning, work in this field starts to switch to end-to-end neural network-based approaches. Typically, a convolutional neural network (CNN) is used to encode the input image, and then a recurrent neural network (RNN) is used to decode its caption [50], [51]. Such encoder–decoder architecture has been widely adopted since then. Attention mechanisms are introduced to image captioning later, and shown to be effective in boosting the performance. To allow a model to attend over different parts of an input image for better caption generation, Xu et al. [52] utilized a grid of convolutional features, whereas Anderson et al. [53] proposed to use regional features extracted from an object detector such as Faster RCNN [54]. In [55], an adaptive attention mechanism was introduced, which utilized a sentinel gate to instruct the model when and where to attend to the image over the course of word generation. Apart from using visual features encoded by a CNN, You et al. [56] also applied attention over semantic attributes predicted from the input image. Li et al. [57] proposed to use text-guided and semantic-guided attentions to correlate vision with language for better caption generation. Yang et al. [58] proposed to minimize the dependency on the previous predicted words, and rather shift more attention on the visual features at a prediction time step in image captioning. Capturing human-object interactions in images and generating captions hierarchically have been studied in [59]. To better model scene semantics, graph convolutional networks (GCNs) have been adopted to caption images [60], [61], [62]. Reinforcement learning has also been attempted to enhance image captioning [63]. Captioning images with other languages, such as Chinese [64] and Japanese [65] has also been studied. To reduce the burden on data annotation, a few studies have proposed to use semisupervised learning [66] or to exploit semantic relevance between unpaired image-caption data [67] for effective image captioning.

More recently, as transformer [68] has shown better performance than RNN across many natural language processing tasks, transformer-based captioning models have also emerged, in which transformers replace RNNs to model geometric relations between detected objects [69], or to model semantic attributes [70] for better image captioning. AoANet was introduced in [71], which integrated with a novel attention on attention (AoA) module to better model the relevance between the queries and attention results in attention-based image captioning. Pan et al. [72] proposed X-Linear Attention Networks (X-LAN), which utilized bilinear pooling to capture the 2nd order feature interactions for captioning images. In [73], memory-augmented attention and meshed cross attention were introduced into the transformer to enhance image captioning. In [74], an auto-parsing network (APN) was proposed, which contains the probabilistic graphical model (PGM) constrained self-attention to boost the transformer-based captioning task. A partially nonautoregressive model was introduced in [75], which was able to retain the accuracy of autoregressive models and enjoy the speedup of nonautoregressive models in image captioning. RSTNet was proposed in [76] recently, which leveraged grid-augmented features and used the adaptive attention mechanism to model visual and nonvisual words for captioning images. With a linguistic transformer and curriculum learning, Dong et al. [77] proposed to use dual GCNs to enhance image captioning. One GCN was designed to model the relationships between objects in each individual image, and the other GCN was designed to capture the additional contextual information

from other similar images. Vision-language pretraining [78], [79] has attracted much attention recently, and shown to be effective in correlating visual and textual features to boost downstream tasks. In [80], ClipCap was designed to leverage vision-language pretraining, and use a mapping network and a language model for image captioning.

Nevertheless, the above image captioning methods have only been evaluated on nonegocentric image captioning datasets such as MSCOCO [81]. Little research has been carried out on egocentric image captioning [82], [83].

The primary goal of this work was to caption egocentric dietary images. As such, the generated captions can be used for dietary assessment instead of the original images, reducing the risk of privacy leakage from images. With the advances in semiconductors and on-node processing, captioning functions can be implemented in the egocentric cameras, and the wearable devices will be able to store only the captions. This will not only remove the privacy concerns, but it will also significantly reduce the volume of data to be stored and prolong the battery life of the egocentric device. To this end, we built the so far largest EgoDIMCAP dataset (also the largest egocentric image captioning dataset). Our model is based on transformer with a novel design adapted for captioning in-the-wild dietary images.

III. METHOD

In this section, we first explain the rationale behind the design of our transformer-based image captioning model, and then present the mathematical formulation of the key modules in our model.

Fig. 2 shows the architecture of the designed transformer-based egocentric dietary image captioning model. The model has a dual stream encoder to encode visual information and a single stream decoder to decode the caption of an input image. In the dual stream encoder, one stream encodes the entire input image, and the other stream encodes regional features pre-extracted from Faster RCNN. We follow [53] to extract regional features from a Faster RCNN model pretrained on the Visual Genome dataset [84]. As our dietary images are largely different from images in the Visual Genome dataset, the pre-trained Faster RCNN model may not output discriminative regional features for the dietary images. Using pre-extracted regional features alone thus may not achieve the best captioning quality. We therefore add another stream in the encoder to encode the entire input image in order to learn representation at a global level with gradient descent. Note that as shown in Fig. 2, all modules within the dashed round purple box are trained simultaneously (i.e., the Faster RCNN is only used for pre-extracting regional features and is not trained during gradient descent, whereas a ResNet [85] is trained with the rest modules to learn global representations of dietary images). We found that training the ResNet simultaneously with the rest modules is essential and can lead to significant improvements in captioning in-the-wild dietary images. We justify this design of our model with extensive experiments, which will be shown in Section V.

Formally, the dietary image captioning process of our model can be formulated as follows.

Given an input dietary image $\mathcal{I}$, we extract N number of regional features $\mathbb{R}^{N \times 2048}$ from a Faster RCNN model
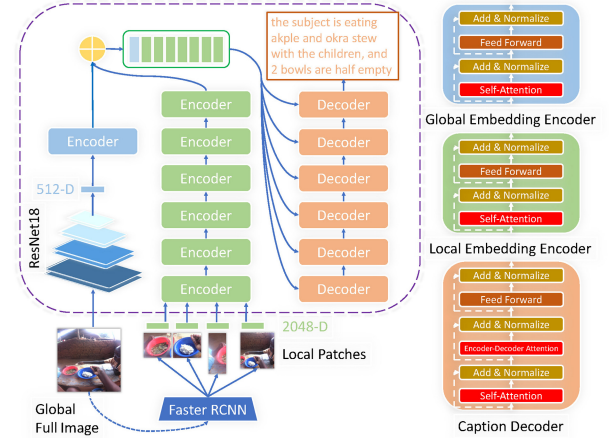

Fig. 2. Architecture of the proposed model for captioning egocentric dietary images. The model has a two-stream encoder and a single-stream decoder. All components within the dashed round purple box are trained together (i.e., apart from the Faster RCNN model which pre-extracts the attentive features, the other parts are trained as an entire for image captioning).

pretrained on Visual Genome where the dimension of each feature vector is 2048. $\mathbb{R}^{N \times 2048}$ is then projected to $\mathbb{P}^{N \times 512}$ using a feed forward layer (not shown in Fig. 2 for simplicity). $\mathbb{P}^{N \times 512}$ is then encoded by a stack of six transformer encoders, each consisting of a self-attention layer and a feed forward layer with residual connection around and layer normalization [86] as shown in the enlarged green box (local embedding encoder) on the right of Fig. 2. After encoding, $\mathbb{P}^{N \times 512}$ is transformed into $\mathbb{L}^{N \times 512}$ which we denote as local embeddings.

In the other stream, the entire image $\mathcal{I}$ is fed into a ResNet, which encodes $\mathcal{I}$ into a 512-D feature vector $\mathbb{F}^{1 \times 512}$ (from ResNet's global average pooling layer). A transformer encoder (see the enlarged blue box) is then used to encode $\mathbb{F}^{1 \times 512}$, transforming it into $\mathbb{G}^{1 \times 512}$ which we denote as the global embedding.

We concatenate $\mathbb{G}^{1 \times 512}$ and $\mathbb{L}^{N \times 512}$, and then feed the resulting visual embeddings $\mathbb{V}^{(N+1) \times 512}$ to the caption decoder, which is a stack of six transformer decoders (see the enlarged orange box). The caption decoder decodes a caption based on the self-attention over the past decoded words and encoder–decoder attention over the visual embeddings $\mathbb{V}^{(N+1) \times 512}$.

The self-attention mechanism in both global and local embedding encoders, and the caption decoder can be mathematically written as follows:

$$\alpha = \text{Attention}(W_q X, W_k X, W_v X) \quad (1)$$

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \quad (2)$$

where W_q, W_k, and W_v are learnable weight matrices. d_k is a scaling factor, which is set to 64 in our experiments. In the global embedding encoder, X is $\mathbb{F}^{1 \times 512}$, whereas in the bottom local embedding encoder, X is $\mathbb{P}^{N \times 512}$, and in the remaining local embedding encoders, X is the output from the encoder directly below. Similarly, in the bottom caption decoder, X is the word embeddings, each of 512-D, whereas in the remaining decoders, X is the output from the decoder directly below.

For encoder–decoder attention in the caption decoder, X multiplied with $\boldsymbol{W_k}$, and $\boldsymbol{W_v}$ is visual embeddings $\mathbb{V}^{(N+1)\times 512}$ from the encoder, and X multiplied with $\boldsymbol{W_q}$ is the output of the self-attention after residual connection and layer normalization in the current caption decoder.

The loss function employed during the training procedure is described in the following. Given a caption with a length of T, we first sum the cross entropy over the entire vocabulary at time step $t(t \leq T)$ as shown in

$$L^{(t)}(\theta) = -\sum_{i=1}^{i=|V|} y_{t,i} \times \log(\hat{y_{t,i}}) \tag{3}$$

where $|V|$ is the vocabulary size, $y_{t,i}$ is the true probability of ith word at time step t, and $\hat{y_{t,i}}$ is the predicted probability of ith word at time step t. θ is a set of learnable parameters in the model.

For a caption of length T, the loss then can be written as follows:

$$\mathcal{L} = \frac{1}{T}\sum_{t=1}^{T} L^{(t)}(\theta) = -\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{i=|V|} y_{t,i} \times \log(\hat{y_{t,i}}). \tag{4}$$

Throughout the training procedure, the model learns to caption egocentric images by minimizing the loss $\mathcal{L}$.

We denote our model as GL-Transformer, as it utilizes the global visual embedding from the entire image and local visual embeddings from the regional features to decode captions.

IV. DATASET

In this section, we introduce our EgoDIMCAP dataset, built using in-the-wild dietary images. In particular, we start by presenting our data collection process, followed by the annotation procedure, dataset details after post-processing, and the final dataset splits used in our experiments.

A. Data Collection

Automatic ingestion monitor (AIM) [2] and eButton [3], were used to capture egocentric images in rural areas in Ghana, Africa. The AIM is attached to the frame of optical glasses, which provides an egocentric view, same as the subject's eyes. The eButton has a 170° range of view and is worn on the subject's chest attached to his/her clothing. The egocentric view from eButton therefore is wider but lower than the AIM. In total, ten households were recruited. Each household had three subjects participating in the study: two adults (mother and father) and one adolescent child. Fathers and adolescent children were given either an AIM or an eButton to wear, whereas mothers were given both to wear. The AIM captured images at an interval of 5–15 s whereas the eButton captured images at an interval of 3–5 s. The captured images of both devices were stored on an internal SD card, and then uploaded to a cloud server after data collection was finished in each household. A detailed study protocol has been published in [87].

B. Data Annotation

Each egocentric image has 1–4 human annotated caption(s). The annotated images are not restricted to dietary intake. Diet-related activities were also annotated, for example, buying

Human annotated captions
1. the subject is having a meal with another subject
2. the subject is eating akple and okra stew with another subject
3. the subject is eating a half bowl of akple and a half bowl of okra stew with another subject
4. the subject is eating akple and okra stew with another subject, and 2 bowls are half empty

Human annotated captions
1. the subject is having a meal with the family, and the bowls are almost empty
2. the subject is eating banku and fish stew with the family, and the bowls are almost empty
3. the subject is eating banku and fish stew with the family, and there isn't much food left
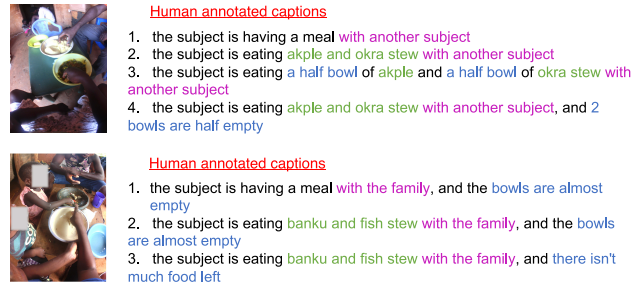
Fig. 3. Samples from our EgoDIMCAP dataset. An annotated sample of eButton-captured images is shown in Fig 1. We show two samples of AIM-captured images in this figure. The captured dietary intake images are annotated in different levels of detail, including the type of food being consumed, the portion size of the food, and whether it is a food-sharing scenario. Faces are masked to protect privacy.

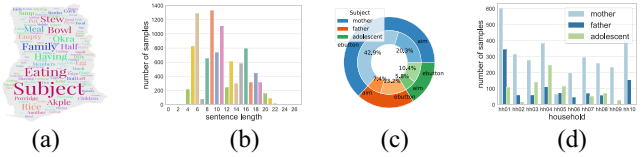|			
(a)	(b)	(c)	(d)

Fig. 4. (a) Tag cloud of our EgoDIMCAP dataset. The size of the word reflects the frequency of that word appearing in the dataset. (b) Sentence statistics. The maximum number of words a sentence has in our dietary image captioning dataset is 27. On average, each sentence has 11.0 words. (c) Data distribution over different subjects (mother, father, and adolescent), and also over different devices (AIM and eButton). (d) Number of data samples collected from each subject in each household.

cooking ingredients in a shop, preparing food, and mother breastfeeding her baby. For dietary intake images, the captions contain portion size information, such as *a half bowl of* and *the bowl is almost empty*. They also contain the information about the type of food the subject is eating or whether the subject is sharing food from the same plate or bowl with other family members. To the best of our knowledge, this is the first dataset that annotates dietary intake images in this level of detail to assist dietary intake assessment. The type of food and the ingredients used in each meal were recorded by field staff in each household. Two annotators then annotated captions based on the record. The portion size information was visually estimated by the annotators and cross-checked by them until a consensus was reached. Fig. 3 shows another two samples from our dataset.

C. Dataset Statistics

After post-processing, 4797 images were used to construct the dataset, among which 1610 images were captured by the AIM and 3187 images were captured by the eButton. Fig. 4(a) shows the tag cloud of the dataset. The frequency of a word in the dataset is reflected by its size in the generated tag cloud. Food types, such as *okra*, *akple*, and *rice*, are very common in the dataset. In terms of portion size information, *half* and *empty* appear frequently in the dataset. The vocabulary size of the dataset is 146. The number of words a caption sentence has ranges from 4 to 27, as shown in Fig. 4(b). On average, each sentence has 11.0 words. Fig. 4(c) shows the data distribution over different subjects and also over different devices. The data from mothers, fathers, and adolescent children account for 63.2%, 20.6%, and 16.2%, respectively. Within data of

TABLE I
COMPARISON BETWEEN EGOCENTRIC IMAGE CAPTIONING DATASETS

Dataset	# Images	Human Annotated Captions
DeepDiary (JVCIR2018) [82]	800	yes
Egoshots (ICLRW2020) [83]	978	no
EgoDIMCAP (Ours)	4797	yes

TABLE II
DIFFERENT DATASET SPLITS FOR EVALUATING THE PERFORMANCE

Dataset Split	Partition	
	# Train Images	# Test Images
I (AIM train and eButton test)	1610	3187
II (eButton train and AIM test)	3187	1610
III (Mixed)	2970	1827

each subject, the eButton data has more samples than the AIM data. Fig. 4(d) shows the number of data samples each subject has in each household. Households No. 6, No. 9, and No. 10 each has one subject's data missing, which is because the wearable camera worn by that subject was not facing in the right angle, and therefore it did not capture useful images for dietary assessment. The rest of the households have data from all subjects recruited in that household.

We name our dataset as egocentric dietary image captioning (EgoDIMCAP) dataset. To the best of our knowledge, our dataset is the largest one in both computer vision and nutritional science communities for egocentric dietary image captioning. Table I compares datasets in the field of egocentric image captioning. Our dataset is five times larger than DeepDiary [82], which is a lifelogging image captioning dataset, and also five times larger than Egoshots [83], which includes real-life egocentric images but with machine-generated captions.

D. Dataset Splits

As AIM and eButton have different viewing angles as mentioned in Section IV-A, the egocentric images captured by these two devices are quite different. Hence, we created three dataset splits to evaluate the performance of egocentric image captioning models. Table II shows the created dataset splits. In split I, images captured by AIM are used for training, and those captured by eButton are used for testing; in split II, it is the opposite; in split III, the training and testing data have images from both AIM and eButton, and images are partitioned into training and testing sets based on their associated captions. To tune the hyperparameters of our model, the training set of split III was partitioned into 80% for training and the rest 20% for validation. After optimal hyperparameters were found, the validation set was merged back, and the original entire training set was used to retrain the model for evaluation on its test set. Experiments on dataset splits I and II used the same set of hyperparameters of split III.

V. EXPERIMENT

In this section, we first describe the baseline methods used for comparison, followed by the implementation details. Comprehensive experiments have been conducted, which includes captioning images on different dataset splits, parsing generated captions to perform portion size estimation, food recognition, and action recognition, as well as justifying the global-local two-stream design of the proposed model with ablation studies. We also show results on a public egocentric image captioning dataset, that is, DeepDiary dataset, which includes daily life egocentric images and human-annotated descriptions, to validate the effectiveness of our model in general egocentric scenarios. Experimental details of combining dietary image captioning and 3-D container reconstruction for estimating actual food volume are elaborated in the end of this section.

A. Baseline Methods

Five state-of-the-art image captioning models and a variant of our proposed model were used as the baselines to compare against our model in captioning egocentric dietary images.
1) *Up-Down (CVPR2018) [53]:* An attention-based model that combines bottom-up features based on Faster RCNN, and the top-down mechanism.
2) *Att2in (CVPR2017) [63]:* An attention-based model in which regional features are only fed to the cell of the internal LSTM.
3) $\mathcal{M}^2$*Transformer (CVPR2020) [73]:* Meshed-Memory Transformer, a transformer-based image captioning model with memory augmented and meshed cross attentions.
4) *X-LAN (CVPR2020) [72]:* X-Linear Attention Network, which captures the 2nd order feature interactions with bilinear attention for image captioning.
5) *ClipCap (CoRR2021) [80]:* An image captioning model based on CLIP [78]. ClipCap uses a mapping network to transform CLIP embeddings to prefix embeddings of a fixed length as the input to its language generation model.
6) *GL-Transformer*:* A variant of the proposed GL-Transformer, in which the ResNet is not trained with the rest of the model (i.e., we use global image features extracted from a ResNet model pretrained on ImageNet [88]).

Standard evaluation metrics BLEU [89], METEOR [90], ROUGEL [91], CIDEr [92], SPICE [93], and WMD [94] in image captioning were adopted to compare the performance of our model against the baselines.

B. Implementation Details

We implemented our model using PyTorch. We used ResNet18 [85] to encode the entire input image to a 512-D feature vector (in the case of GL-Transformer*, a ResNet18 model pretrained on ImageNet was used to pre-extract global image features). The Up-Down, Att2in, $\mathcal{M}^2$Transformer, X-LAN, GL-Transformer*, and GL-Transformer models used the same regional features pre-extracted from Faster RCNN. We pre-extracted CLIP embeddings before training the ClipCap model on our dataset, and we used its default setting during training. The default setting of X-LAN was also adopted during its training on our EgoDIMCAP dataset. For other models during training, we set batch size to 10. Adam [95] was adopted as the optimizer.

TABLE III
WEIGHTED AVERAGE RESULTS WITH ALL THREE DATASET SPLITS

Method	Evaluation Metric								
	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGEL	CIDEr	SPICE	WMD
Up-Down (CVPR2018) [53]	61.3	54.3	48.4	42.0	31.9	61.8	161.3	35.7	57.2
Att2in (CVPR2017) [63]	63.0	56.5	51.0	45.2	33.7	63.3	175.9	36.6	59.2
$\mathcal{M}^2$ Transformer (CVPR2020) [73]	62.5	55.9	50.2	44.1	32.6	63.3	163.4	35.1	59.0
X-LAN (CVPR2020) [72]	60.9	54.8	49.7	44.0	32.5	62.9	167.8	36.5	57.6
ClipCap (CoRR2021) [80]	45.5	41.1	37.9	35.1	43.2	62.7	154.4	43.5	55.6
GL-Transformer* (Ours)	63.4	56.7	51.1	45.1	33.7	63.8	166.4	36.6	58.9
GL-Transformer (Ours)	63.7	57.0	51.6	45.9	34.6	64.3	182.6	38.7	59.5

TABLE IV
RESULTS ON DATASET SPLIT I

Method	Evaluation Metric								
	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGEL	CIDEr	SPICE	WMD
Up-Down (CVPR2018) [53]	57.7	50.0	43.6	36.7	28.6	57.7	115.4	30.7	52.9
Att2in (CVPR2017) [63]	60.4	53.2	47.2	41.0	31.1	60.2	133.2	31.8	55.5
$\mathcal{M}^2$ Transformer (CVPR2020) [73]	61.1	53.9	47.8	41.2	30.3	61.7	126.2	31.1	56.7
X-LAN (CVPR2020) [72]	60.4	54.3	49.0	43.1	30.9	61.0	133.2	34.0	55.4
ClipCap (CoRR2021) [80]	45.3	39.6	35.7	32.1	37.1	58.7	112.0	37.2	52.5
GL-Transformer* (Ours)	62.1	55.1	49.3	43.0	31.6	61.7	123.0	33.1	56.2
GL-Transformer (Ours)	62.4	55.3	49.4	43.2	31.6	61.3	142.0	34.3	56.3

TABLE V
RESULTS ON DATASET SPLIT II

Method	Evaluation Metric								
	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGEL	CIDEr	SPICE	WMD
Up-Down (CVPR2018) [53]	60.9	53.7	47.5	40.5	31.6	61.8	168.4	36.9	56.0
Att2in (CVPR2017) [63]	60.9	54.9	49.8	44.1	33.0	62.9	171.1	36.0	57.4
$\mathcal{M}^2$ Transformer (CVPR2020) [73]	60.2	54.2	49.0	43.1	31.8	61.1	167.0	35.5	56.3
X-LAN (CVPR2020) [72]	58.2	51.5	45.7	39.4	30.6	60.1	168.4	35.8	55.0
ClipCap (CoRR2021) [80]	37.9	35.4	33.3	31.4	46.7	63.0	163.6	48.8	55.8
GL-Transformer* (Ours)	59.9	53.2	47.7	41.9	33.7	63.0	176.7	37.4	57.6
GL-Transformer (Ours)	58.9	52.5	47.4	42.2	35.0	64.0	194.1	38.7	59.0

TABLE VI
RESULTS ON DATASET SPLIT III

Method	Evaluation Metric								
	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGEL	CIDEr	SPICE	WMD
Up-Down (CVPR2018) [53]	67.9	62.3	57.6	52.6	38.1	68.9	235.1	43.4	65.7
Att2in (CVPR2017) [63]	69.3	63.6	58.7	53.6	38.8	69.1	254.7	45.5	67.1
$\mathcal{M}^2$ Transformer (CVPR2020) [73]	67.1	60.8	55.6	50.2	37.4	67.9	225.3	41.8	65.4
X-LAN (CVPR2020) [72]	64.3	58.7	54.3	49.6	37.1	68.6	227.8	41.5	63.9
ClipCap (CoRR2021) [80]	52.6	48.7	45.9	43.5	50.7	69.3	220.4	49.9	60.9
GL-Transformer* (Ours)	68.9	62.4	57.1	51.6	37.4	68.2	232.9	43.4	64.6
GL-Transformer (Ours)	70.2	64.1	59.1	54.0	39.4	69.8	243.4	46.4	65.4

The learning rate was set to 0.0005, and they were trained for a maximum of ten epochs. We used eight attention heads in our proposed GL-Transformer model and the same number of heads in its baseline variant and the $\mathcal{M}^2$ Transformer. We set the number of both encoder and decoder layers in $\mathcal{M}^2$ Transformer to 6 for fair comparisons. N in Section III was set to 36 (i.e., we extracted a maximum of 36 regional feature vectors from the pretrained Faster RCNN model).

C. Experimental Results

We first show the overall results of our proposed method and the baseline methods in Table III, and then present the results tested on each separate dataset split in Table IV (Split I), Table V (Split II), and Table VI (Split III). The results in Table III are the weighted average results using all three dataset splits (i.e., the results in Table III are obtained by weighting the results in Tables IV–VI with the number of test images their corresponding split has).

1) Overall Results: As shown in Table III, in 7 out of 9 evaluation metrics, our proposed GL-Transformer model achieves the best scores, and in the remaining two metrics, our model achieves the second-best scores. In particular, for the CIDEr score, our model achieves the best score and has an absolute increase of 6.7 compared to the second-best one. The recent ClipCap model achieves the highest METEOR and SPICE scores, but in the measured BLEU and CIDEr scores, it lags behind other methods by a large margin. The variant of our GL-Transformer model, GL-Transformer* also shows better performance in the measured BLEU1, BLEU2, BLEU3, and ROUGEL compared to other baseline methods.

2) Results on Dataset Split I: Table IV summarizes the results on dataset split I. In this split, egocentric images from AIM were used for training and images from eButton were used for testing. The proposed GL-Transformer model achieved the best results in 5 out of 9 evaluation metrics ($\mathcal{M}^2$ Transformer and ClipCap each topped two metrics in the rest 4 metrics). Although in this split, GL-Transformer* achieved the closest results to GL-Transformer in most metrics, it had a lower CIDEr score than some baseline models. Nevertheless, GL-Transformer increased the CIDEr score from 133.2 (the best baseline CIDEr score) to 142.0. The first row in Fig. 5 shows some qualitative results on this dataset split. In the left example, GL-Transformer was able to generate a caption close to the ground truth. Note that the food (*banku*) in the caption generated by GL-Transformer is actually visually similar to the food (*akple*) in the ground truth caption. Although ClipCap was also able to generate a caption that matches with the ground truth, its generated caption contains repeated words of akple at the end. X-LAN generated a reasonable caption, but compared to the ground truth and the one generated by GL-Transformer, its caption does not indicate what type of food the subject was making. The rest four models failed to caption the image correctly in this example. In the right example, the caption generated by GL-Transformer perfectly matches with the ground truth. The Up-Down and Att2in models in this example were not able to generate correct or even close captions. $\mathcal{M}^2$ Transformer in this example was able to describe the image with close estimations of the portion size and food type, and ClipCap was able to understand that the subject was having a meal (so were GL-Transformer* and X-LAN) and the bowl was half empty, but it mis-counted the number of food containers.

3) Results on Dataset Split II: Table V compares the results on dataset split II. Note that as the number of eButton images were almost two times larger than that of AIM images (i.e., in this split, there were more training images than split I), all models achieved higher scores in a few evaluation metrics, especially in terms of the CIDEr score. In this split, Att2in achieved the best BLEU scores. GL-Transformer topped ROUGEL, CIDEr, and WMD. Compared to Att2in, it increased CIDEr score from 171.1 to 194.1, which was a 13.4% relative increase. ClipCap topped METEOR and SPICE on this split. We show some qualitative results in the second row of Fig. 5. In the left example, the scene is quite cluttered. ClipCap, GL-Transformer*, and GL-Transformer were able to correctly caption the image as *the subject is cutting okra into dices in the kitchen*, whereas Up-Down, Att2in, $\mathcal{M}^2$ Transformer, and X-LAN all failed in this case. In the right example, all models, except X-LAN, were able to identify that the subject was eating some kind of rice. However, for the Up-Down model, it failed to give the portion size information. For the Att2in model, it mis-recognized the food type as well as mis-estimated the portion size (i.e., *more than half of the bowl is empty* is clearly not correct). Although the $\mathcal{M}^2$ Transformer model correctly recognized the food type, same as the Att2in model, it mis-estimated the portion size. For the ClipCap model, it generated additional but incorrect description about the image, that is, there was no egg stew and the
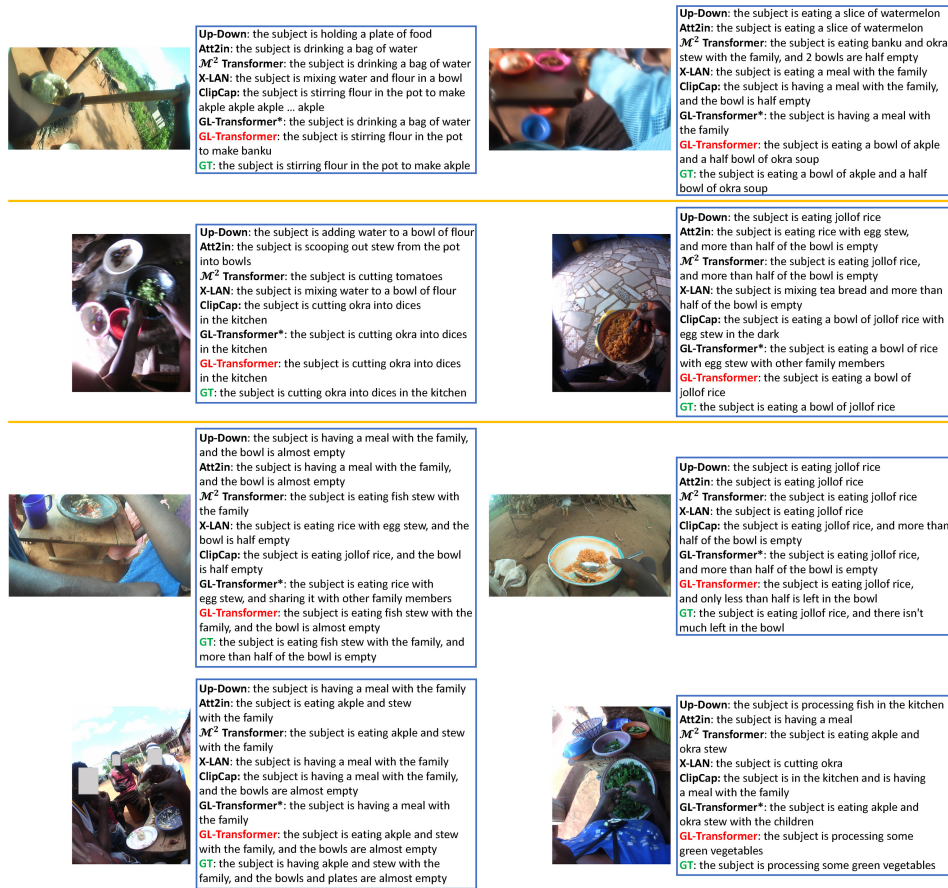
Fig. 5. Qualitative results on dataset splits I (first row), II (second row), and III (bottom two rows). The captions generated by the proposed model GL-Transformer were consistently better than the baselines across all dataset splits, and better matched with the ground truth captions. The faces of the subjects are masked to protect their privacy.

subject was not eating in the dark. For the GL-Transformer*, although it correctly estimated the portion size (i.e., *a bowl of*), it failed to recognize the food type as well as the scene in which the subject was having the meal alone. In this particular example, only GL-Transformer correctly captioned the image.

4) Results on Dataset Split III: Table VI shows the results on dataset split III, in which images from AIM and eButton were mixed and partitioned into training and testing sets based on their captions. Note that the scores achieved by all models across all evaluation metrics were higher than splits I and II. We hypothesize that this is because in splits I and II, the training and testing images were from different devices, and as mentioned earlier, the AIM and eButton had different viewing angles, which caused the domain difference between training and testing. In dataset split III, as the training set contained images from both devices, and so did the testing set, the domain difference was minimized, and therefore all models were able to achieve higher scores across all evaluation metrics. Nevertheless, in this split, our full model GL-Transformer still showed better performance than all six baselines, achieving the highest scores in 5 out of 9 evaluation metrics. The bottom two rows in Fig. 5 show some qualitative results from this split. In the left example of the third row, both Up-Down and Att2in models were able to caption the image appropriately, understanding the subject was having a meal with the

family, but they were not able to caption the image with the type of food the subject was eating. For the estimated portion size, they were actually not too far away from the ground truth. For the $\mathcal{M}^2$ Transformer model, it was able to caption the image with the correct food type and eating scenario, but failed to produce an estimation regarding the food portion size. Both the X-LAN and the ClipCap models failed to recognize the food type and the food portion size in this example. Our GL-Transformer in this example was able to caption the image with correct food type and a close estimated food portion size. In the right example of the third row, all models were able to recognize the subject was eating jollof rice, whereas only the ClipCap, GL-Transformer*, and GL-Transformer were able to caption the image with portion size information close to the ground truth. In the left example of the fourth row, all baseline models were able to recognize that the subject was having a meal with the family. Both Att2in and $\mathcal{M}^2$ Transformer were also able to recognize the food types the subject was eating. However, only GL-Transformer was able to produce a correct portion size estimation for the food the subject was eating as well as the food types in this example. In the right example of the fourth row, all baseline models failed to caption the image correctly. Only GL-Transformer was able to correctly describe the image. The Up-Down model correctly indicated the subject was processing something but failed to recognize its type. We hypothesize that the Att2in model might attend

Up-Down: the subject is breastfeeding her baby
Att2in: the subject is breastfeeding her baby
$\mathcal{M}^2$ Transformer: the subject is breastfeeding her baby
X-LAN: the subject is breastfeeding her baby
ClipCap: the subject is breastfeeding her baby
GL-Transformer*: the subject is breastfeeding her baby
GL-Transformer: the subject is breastfeeding her baby
GT: the subject is breastfeeding her baby

Up-Down: the subject is UNK a bowl of rice into a pot to UNK
Att2in: the subject is mixing water and flour in a bowl
$\mathcal{M}^2$ Transformer: the subject is eating rice
X-LAN: the subject is eating rice with egg stew, and the bowl is half empty
ClipCap: the subject is having a meal with the family, and the bowl is half empty
GL-Transformer*: the subject is eating a half bowl of rice
GL-Transformer: the subject is packaging banku
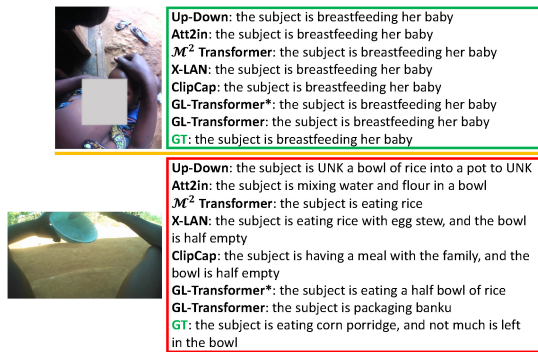**GT**: the subject is eating corn porridge, and not much is left in the bowl

Fig. 6. Qualitative results that all models were able to successfully caption (top row) or failed to caption (bottom row) a given dietary image. Sensitive body parts are masked to protect the subjects' privacy.

TABLE VII
DEFINED PORTION SIZE WORDS/PHRASES, FOOD/INGREDIENT TYPES, AND ACTION TYPES

Portion Size Words/Phrases				
half empty	a bowl	a half bowl	there isn't much	almost empty
not much	small	2 bowls	a plate	a slice
a pot	a cup	less than half	more than half	empty
full	not many	almost full		

Food/Ingredient Types				
fish	okra	meat	soup	flour
water	akple	stew	kenkey	banana
watermelon	green	jollof	roasted corn	corn
plantain	tea	bread	rice	tomato
vegetable	potato	onion	egg	banku
fufu	pineapple	porridge	avocado	chicken

Action Types				
breastfeed	buy	process	cut	cook
take	add	make	stir	have
drink	share	clean	play	eat
select	prepare	put	sit	scoop
hold	mix	package	roast	peel
pour	ground			

TABLE VIII
PORTION SIZE ESTIMATION AND FOOD RECOGNITION AND ACTION-RECOGNITION ACCURACY CALCULATED BASED ON THE GROUND TRUTH AND GENERATED CAPTIONS FROM SPLIT III

Method	Portion Size Estimation (%)	Food Recognition (%)	Action Recognition (%)
Up-Down (CVPR2018) [53]	25.9	38.0	60.4
Att2in (CVPR2017) [63]	28.7	49.9	59.6
$\mathcal{M}^2$ Transformer (CVPR2020) [73]	18.8	**50.0**	58.8
X-LAN (CVPR2020) [72]	22.2	28.1	60.4
ClipCap (CoRR2021) [80]	40.2	39.1	57.5
GL-Transformer* (Ours)	25.9	46.4	62.3
GL-Transformer (Ours)	**42.0**	47.5	**62.7**

results about the portion size estimation, food recognition, and action-recognition accuracy based on the generated captions. As shown in Table VII, we predefined a list of words related to portion size such as *half empty*, a list of food types, such as *okra* and *akple*, and a list of actions. We then parsed both the ground truth caption and the generated caption to obtain their respective portion size words, food types, and actions. We compared how many portion size words in the generated caption match with those in the ground truth to obtain the portion size estimation accuracy. Similarly, we compared food types to obtain food-recognition accuracy, and compared actions to obtain action-recognition accuracy.

Equation (5) illustrates how we calculate these accuracies

$$R = \frac{1}{N} \sum_{i=1}^{N} \frac{C_{gt-gc}^i}{C_{gt}^i} \tag{5}$$

where N is the total number of test captions, C_{gt}^i is the number of predefined word(s)/phrase(s) in the ith ground truth caption, for example, if portion size estimation accuracy is to be calculated, C_{gt}^i is the number of portion size word(s)/phrase(s) in the ith ground truth caption, C_{gt-gc}^i is the number of predefined word(s)/phrase(s) in the ith ground truth caption that also show(s) in the ith generated caption.

Table VIII shows the accuracy calculated from the test set of split III. Our GL-Transformer achieved the best portion size estimation accuracy (18 different portion size descriptions) as well as the best action-recognition accuracy (27 actions). Regarding food recognition, it also achieved decent accuracy of recognizing 30 different types of local food/ingredients. Another straightforward way of conducting portion size estimation, food recognition, or action recognition is to train a classifier on each of these tasks. We therefore conducted further experiments to train three separate classifiers on these three tasks. As a ResNet-18 is integrated to learn global image features in our GL-Transformer, we chose ResNet-18 as the classifiers for fair comparison. Note that as a dietary image may contain multiple food containers, food items, and actions performed by the subject, we trained the classifiers with multilabel classification and calculated Recall on each task. These three classifiers achieved Recall of 39.5%, 47.3%, and 61.2%, respectively, on portion size estimation, food recognition, and action recognition. It is worth noting that a single GL-Transformer can already achieve similar accuracy by parsing its generated captions as shown in Table VIII, whereas to conduct classification directly, it requires three separate classifiers to be trained. Note that as mentioned in Section I, the food portion size parsed from the caption is measured with respect to its container. In order to quantify the food volume, a vision-based food container volume estimation method is

to bowls in the scene and therefore captioned the image as *the subject is having a meal*. For the $\mathcal{M}^2$ Transformer and GL-Transformer* models, we hypothesize that they might mis-recognize the green vegetables as okra stew, and therefore did not caption the image correctly. Although the caption generated by X-LAN seemed appropriate (i.e., okra is a type of vegetable and it is green, and cutting is a means of processing), the exact vegetable was not okra and the subject was not cutting. For the ClipCap model, although it can indicate that the subject was in the kitchen, it mis-recognized the activity the subject was performing.

The above experimental results, including the overall results and results on splits I, II, and III (both quantitatively and qualitatively) demonstrate that the proposed GL-Transformer performs better on captioning egocentric dietary images, and also show that in the GL-Transformer, training ResNet with the rest of the model is essential, so that its encoder can be adapted to learn a better representation for the entire dietary image, leading to better captioning quality.

Fig. 6 shows some examples that all models were able to successfully or failed to caption a given dietary image.

5) Portion Size Estimation and Food Recognition and Action-Recognition Accuracy: As in our EgoDIMCAP dataset, portion size and food types are contained in the captions (so long as the portion size and food types are recognizable for the human annotators), and each caption also has word(s) describing the action(s) the subject is performing, such as *eating* and *cooking*. We further present quantitative

TABLE IX
RESULTS OF THE ABLATION STUDIES ON SPLIT I

Method	Evaluation Metric								
	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGEL	CIDEr	SPICE	WMD
G-Transformer	60.3	52.4	46.4	40.3	31.5	60.0	127.8	31.4	55.8
L-Transformer	59.6	52.8	47.5	42.0	32.5	60.7	124.6	31.0	55.2
GL-Transformer (Ours)	62.4	55.3	49.4	43.2	31.6	61.3	142.0	34.3	56.3

G-Transformer: the subject is eating a plate of banku and a half bowl of okra stew
L-Transformer: the subject is holding a plate of food
GL-Transformer: the subject is eating a plate of jollof rice
GT: the subject is eating a bowl of jollof rice

G-Transformer: the subject is roasting corn with the children
L-Transformer: the subject is eating banku and fish stew with the family
GL-Transformer: the subject is eating a bowl of akple and a half bowl of okra stew with the children
GT: the subject is eating a bowl of akple and a half bowl of okra stew with the children

Fig. 7. Qualitative examples of the ablation study.

needed to estimate the container's volume. We show and discuss the practicality of combining dietary image captioning with 3-D container reconstruction to estimate food volume in Section V-C8.

6) Ablation Study: To justify that in the dual stream encoder, both streams are necessary and able to contribute to better captioning performance. We conducted ablation studies and two variants of the proposed GL-Transformer were created.

1) *G-Transformer:* In which the encoder only contains the stream that encodes the entire input image.
2) *L-Transformer:* In which the encoder only contains the stream that encodes the regional features.

The caption decoder in both G-Transformer and L-Transformer is the same as in the GL-Transformer. The ResNet18 in the G-Transformer was also trained with gradient descent using image-caption pairs. We used dataset split I to conduct the ablation studies. As shown in Table IX, G-Transformer and L-Transformer achieved close results in most evaluation metrics. Although the L-Transformer had the best METEOR score, for the rest of evaluation metrics, the GL-Transformer achieved the highest scores. The quantitative results show that our design of using both streams is effective and able to enhance the captioning performance. Fig. 7 shows two examples for visually comparing the caption results of these three models. In the example shown in the top, the ground truth caption is *the subject is eating a bowl of jollof rice*. G-Transformer was able to caption the image as *the subject is eating*, but with incorrect food types; L-Transformer was able to caption the image as *the subject is holding a plate of food*, which is not entirely incorrect; having both global and local visual embeddings, the GL-Transformer was able to caption the image as *the subject is eating a plate of jollof rice*, which is very close to the ground truth caption. The bottom one shows a similar example. The G-Transformer is able to recognize the subject is with his/her children, but fails to recognize the activity (i.e., eating in this case), as well as the food types. The L-Transformer is able to recognize the subject is eating food, but the food types are not entirely correct. By utilizing both global and local features, the GL-Transformer successfully captions the image, which matches the ground truth caption.

7) Results on DeepDiary Dataset: To better evaluate the egocentric image captioning performance of our model, we

TABLE X
RESULTS ON DEEPDIARY DATASET

Method	Evaluation Metrics		
	BLEU4	CIDEr	ROUGEL
Lifelogging (JVCIR2018) [82]	16.0	32.5	42.5
$\mathcal{M}^2$ Transformer (CVPR2020) [73]	16.3	47.0	41.5
X-LAN (CVPR2020) [72]	15.9	44.3	41.3
GL-Transformer (DeepDiary Only)	19.6	**71.8**	42.4
GL-Transformer (COCO + DeepDiary)	**20.6**	68.7	**44.7**

tested it on the publicly released DeepDiary dataset [82], which is an egocentric lifelogging dataset containing daily life images captured by wearable cameras and human-labeled ground truth captions. Due to privacy concerns, the DeepDiary dataset does not release the original captured images, but rather it provides the extracted VGG feature [96] of each image. As our model utilizes both global and local regional features of an image and we only have global features, that is, VGG features, from the DeepDiary dataset, we generated local features for each image as follows: given a VGG feature (4096-D) of an image, we used a sliding window, whose size is 2048, to extract regional features (2048-D). We extracted 36 regional features, and therefore the moving step of the sliding window was set to 58. We also adapted our model as follows: we replaced the ResNet-18 with an MLP to transform the global feature dimension from 4096 to 512, and we kept the rest of the model unchanged. During training, Adam was adopted as the optimizer. Our model was trained with a batch size of 10 and a learning rate of 0.0005. Training was stopped when the loss plateaued. Table X shows the results on the DeepDiary dataset. Compared to the lifelogging model proposed in the DeepDiary dataset, our GL-Transformer trained on DeepDiary only already surpasses it by 3.6 in terms of BLEU4, and achieves a CIDEr score twice as high as the original Lifelogging model. Our model also shows better performance than the $\mathcal{M}^2$ Transformer and X-LAN on the DeepDiary dataset. We also investigated the performance of our model on the DeepDiary dataset by first pretraining our model on the COCO dataset [97], and then fine-tuning it on the DeepDiary dataset. This transfer learning process leads to increases in both BLEU4 and ROUGEL. The decrease in the CIDEr score is hypothesized to be caused by the vocabulary size discrepancy between the COCO and DeepDiary datasets.

8) Combining Image Captioning With 3-D Container Reconstruction: Although our dataset is annotated with portion size information and the proposed captioning model is able to describe the food portion in a dietary intake image, the portion size is still based on the bowl or plate as a reference. To quantify the exact dietary intake, knowing the volume of the bowl or the size of the plate is necessary. This could be achieved using 3-D model reconstruction to estimate the bowl volume or the plate size, and then from the deduced portion size information in the caption, more accurate food volume estimation could be achieved. It is worth noting that directly estimating food volume from an RGB image is not accurate as food items have irregular and various shapes, and often occlude each other. Reconstructing a 3-D food container with food in it is also difficult, as the food will occlude the container, for example, the bottom of the container may be occluded. On the other hand, reconstructing an empty container is relatively simple,
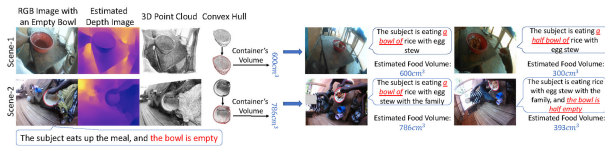
Fig. 8. Proposed framework of combining dietary image captioning with 3-D container reconstruction to estimate food volume with monocular RGB images. We first use the generated image captions to identify the images with empty container(s), and then use a depth estimation network [98] to estimate the depth of the RGB image with empty container(s). The estimated depth image is then projected into the 3-D space to obtain the 3-D point cloud. Afterwards, the 3-D point cloud of the container is extracted, and the 3-D convex hull algorithm is then applied to calculate the actual volume of the food container. The convex hull of a 3-D model is the smallest convex set which contains all the points of a model. Detailed information about volume calculation using the convex hull can be found in our previous works [21], [22]. The obtained volume of the empty food container then can be used in the early images of a dietary intake episode to estimate the actual food volume during eating.

and the reconstruction tends to be more precise as an empty container provides better visibility of its bottom and other inner surface. The advantage of passive dietary monitoring is that it can capture the eating episode of a subject all the way from the start to the end of eating, and at the end of the eating, in most cases, the bowls/plates are close to empty (or at least compared to the start/middle of eating, the bowls/plates are closer to empty). As our image captioning model can indicate whether the container is empty or not in a dietary intake image, we can obtain the image with empty container(s), and reconstruct precise 3-D food container(s) using that image. After 3-D reconstruction, the volume of the container can be obtained, and we can then backtrack the eating episode, and use that volume information to calculate the food volume in the early images with the generated image captions containing food portion size information that is relative to the container, for example, *full* and *half empty*. We show in Fig. 8 a framework of combining image captioning with 3-D container reconstruction to estimate the food volume in the real-world dietary intake scenarios in our EgoDIMCAP dataset.

Note that in order to less affect a subject's dietary intake routine, the real-world data in our EgoDIMCAP dataset were collected using monocular RGB wearable cameras as they are lightweight and can capture data on-board without cables connecting to a computer. The exact food weights were measured in situ but the food volumes were not after food was cooked as it requires measuring cups or spoons, and the transfer of food between its container and measuring cups or spoons may undermine food hygiene.

Quantitative Validation on the EgoDIMCAP-Lab Dataset: In order to quantitatively show the effectiveness of estimating food volume by combining image captioning and 3-D empty food container reconstruction, we therefore collected and constructed a novel RGBD egocentric dietary intake dataset in a laboratory setting. Specifically, Intel RealSense D405, a close-range depth sensor (operating at 7 to 50 cm with submillimeter accuracy), was worn by a subject on the chest to collect the data, which provides the similar viewpoint of the eButton camera used in EgoDIMCAP dataset. Authentic Ghanaian foods were prepared, including *jollof rice*, *okra stew*, and *banku*. Nine different food containers were used as shown in Fig. 9(b).
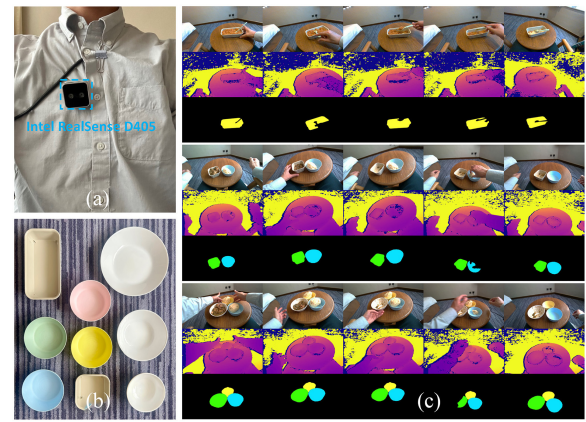

Fig. 9. Laboratory data collection setup. (a) Intel RealSense D405 camera (as highlighted by the blue box) was worn by a subject on the chest to collect egocentric dietary RGBD data. D405 camera can capture valid depth maps in a close-up range from 7 to 50 cm, which is an ideal capture range in dietary intake scenarios as food containers are often close to the human body during eating. (b) Different food containers used in data collection. (c) Data samples from our EgoDIMCAP-Lab dataset. The collected laboratory data contains typical Ghanaian food, different food portion sizes, and different food containers, as well as image captions indicating different eating states. Ground truth food and container volumes, depth maps, and food container segmentation masks are provided in this dataset. Three samples are shown in this figure and each illustrates a complete eating episode of the subject.

In total, 23 eating episodes were collected, which covers eating scenarios, including food in a single/two/three bowl(s), different utensils, and different food combinations. Fig. 9(c) shows three eating episode examples. We downsampled the collected image frames and then removed those unrelated to dietary intake, which resulted in 20 163 paired RGBD image frames. Among them, we selected 1000 frames (distributing over all 23 episodes), and manually annotated the segmentation mask(s) for the food container(s) in them, and detailed caption(s) for each frame. We name this laboratory dataset as EgoDIMCAP-Lab. Note that the food portion size information in EgoDIMCAP-Lab is more fine-grained, e.g., *a 3/4 bowl of* and *a 1/4 bowl of* compared to EgoDIMCAP.

We then conducted four-fold cross validation to evaluate the proposed food volume estimation method, which uses image captioning to indicate images with empty bowl(s) in each eating episode, and reconstructs the 3-D food container using the associated depth map (unlike the pipeline shown in Fig. 8, depth estimation is not required in this laboratory setting as ground truth depth maps were generated synchronously with the RGB images), followed by container volume estimation by the 3-D convex hull algorithm, and finally uses the estimated volume of empty container and food portion size information generated in the caption to deduce the actual food volume.

We compared the food volume estimation error of our proposed method with human estimation. Specifically, dietary intake images before the start of eating of each eating episode were selected. Thirteen volunteers conducted manual estimation of food volumes of these images through a Web interface, and the estimations were collected anonymously.

Equation (6) shows how we obtain the actual food volume using the proposed method for each eating episode

$$V_{\text{food}} = \frac{1}{N} \sum_{i=1}^{N} V_{\text{empty}} \times P_i \tag{6}$$

TABLE XI
FOOD VOLUME ESTIMATION RESULTS ON THE EGODIMCAP-LAB DATASET. G.T. MEANS GROUND TRUTH FOOD VOLUME MEASURED IN CM^3; HUMAN EST. ERR. (MEAN ± STD) IS THE FOOD VOLUME ESTIMATION ERROR OF HUMAN VOLUNTEERS; CAP. + 3-D RECONS. ERR. (MEAN ± STD) IS THE ERROR OF THE PROPOSED METHOD THAT COMBINES IMAGE CAPTIONING AND 3-D EMPTY FOOD CONTAINER RECONSTRUCTION

4-fold Cross Val.	G.T. (cm^3)	Human Est. Err. (cm^3) ↓	Cap. + 3D Recons. (Proposed) Err. (cm^3) ↓
banku-00, single bowl	200	-32 ± 129	11 ± 9
jollof rice-00, single bowl	200	60 ± 193	21 ± 29
jollof rice-01, single bowl	200	36 ± 176	40 ± 20
okra-00, single bowl	150	12 ± 111	-12 ± 18
okra-01, single bowl	200	-72 ± 100	-12 ± 11
okra and banku-00, two bowls	200, 300	-34 ± 146, -97 ± 149	-45 ± 25, -75 ± 25
banku-01, two bowls	200, 200	-46 ± 131, -10 ± 137	-28 ± 23, 30 ± 32
banku-02, single bowl	200	-3 ± 149	40 ± 39
jollof rice-02, single bowl	300	-66 ± 230	34 ± 40
okra-02, single bowl	300	-101 ± 195	-91 ± 14
okra-03, single bowl	400	-176 ± 169	31 ± 28
okra and banku-01, three bowls	200, 200, 300	-9 ± 140, -20 ± 139, -109 ± 155	12 ± 18, -40 ± 19, -84 ± 12
banku-03, single bowl	200	58 ± 221	78 ± 58
banku-04, single bowl	200	40 ± 210	-29 ± 14
jollof rice-03, single bowl	450	-155 ± 238	47 ± 60
jollof rice-04, single bowl	450	116 ± 612	32 ± 47
okra-04, single bowl	400	-146 ± 182	-24 ± 24
okra and banku-02, two bowls	200, 400	0 ± 147, -172 ± 166	53 ± 8, -66 ± 14
banku-05, two bowls	200, 200	31 ± 188, -18 ± 155	6 ± 24, 0 ± 27
jollof rice-05, single bowl	450	-130 ± 310	135 ± 57
okra-05, single bowl	400	-132 ± 263	-129 ± 14
okra-06, single bowl	400	-163 ± 253	-133 ± 7
okra and banku-03, three bowls	200, 200, 400	-16 ± 151, -28 ± 159, -121 ± 208	14 ± 14, -17 ± 22, -90 ± 8
Overall (abs. mean) ↓	274	71	47

where V_{food} is the estimated food volume, V_{empty} is the volume of the empty food container, P_i is the parsed portion size information of the ith image and $P_i = 0$ if no portion size is generated in the caption (P_i in EgoDIMCAP-Lab are numerical values or have numerical counterparts). N is the number of test images (i.e., images before eating starts), and is empirically set to 5 for each eating episode.

Table XI shows the results and comparison of food volume estimation of the food before eating. The results of our proposed method uses GL-Transformer models retrained on EgoDIMCAP-Lab dataset. As shown in Table XI, our proposed method outperforms human volunteers on estimating food volumes.

VI. CONCLUSION

In this work, we proposed the task of captioning egocentric dietary images to assist nutritionists to conduct dietary intake assessment more effectively, and in addition, preserving the subject's privacy. To this end, an in-the-wild EgoDIMCAP dataset has been built, each dietary intake image is annotated with different levels of detail, including the type of food the subject is eating, the food portion size, and whether the subject is sharing food from the same plate or bowl with other individuals. Apart from dietary intake images, relevant images are also annotated in the dataset, such as food preparation. A novel transformer-based captioning model has been proposed and the design of the model has been justified through extensive experiments. A novel framework of estimating food volume by combining image captioning and 3-D container reconstruction has also been proposed along with an egocentric RGBD dietary intake dataset containing multiple data modalities. While extensive experiments have been conducted and we have demonstrated the merits and advantages of our proposed technological concepts for advancing dietary intake assessment, the current work still has some limitations, and potential directions for future research are worth investigating. For example, the proposed food volume estimation framework assumes that the food container is empty or close to empty in the end of the eating episode, and deduces the actual food volume in a two-stage manner. Another potential way of combining captioning and 3-D reconstruction for food

volume estimation is to develop a multimodal framework that fuses caption embedding and 3-D embedding (i.e., semantic and geometric/volumetric information) in a joint space, and estimates the food volume end-to-end. Generating an overall dietary report/diary for a subject based on all images captured by his/her wearable camera can be more intuitive and straightforward for the nutritionists to analyze the nutritional states and needs of the subject. This is challenging but could be one promising future direction. Furthermore, captioning egocentric dietary videos is worth exploring. Utilizing the visual, temporal, and audio features extracted from dietary videos, the more accurate contextual description could be generated for precise dietary assessment.

REFERENCES

[1] J.-S. Shim, K. Oh, and H. C. Kim, "Dietary assessment methods in epidemiologic studies," *Epidemiol. Health*, vol. 36, Jul. 2014, Art. no. e2014009.

[2] A. Doulah, T. Ghosh, D. Hossain, M. H. Imtiaz, and E. Sazonov, "'Automatic ingestion monitor version 2'—A novel wearable device for automatic food intake detection and passive capture of food images," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 2, pp. 568–576, Feb. 2021.

[3] 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.

[4] J.-E. Lee, S. Song, J. S. Ahn, Y. Kim, and J. E. Lee, "Use of a mobile application for self-monitoring dietary intake: Feasibility test and an intervention study," *Nutrients*, vol. 9, no. 7, p. 748, 2017.

[5] L. Wellard-Cole et al., "Relative validity of the eat and track (eat) smartphone app for collection of dietary intake data in 18-to-30-year olds," *Nutrients*, vol. 11, no. 3, p. 621, 2019.

[6] J. L. Lemacks, K. Adams, and A. Lovetere, "Dietary intake reporting accuracy of the Bridge2U mobile application food log compared to control meal and dietary recall methods," *Nutrients*, vol. 11, no. 1, p. 199, 2019.

[7] J. Qiu, F. P.-W. Lo, S. Jiang, Y.-Y. Tsai, Y. Sun, and B. Lo, "Counting bites and recognizing consumed food from videos for passive dietary monitoring," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 5, pp. 1471–1482, May 2021.

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

[9] 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)*, 2015, pp. 1–6.

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

[11] J. Qiu, F. P. W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," in *Proc. BMVC*, 2019, pp. 1–11.

[12] 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.

[13] 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.

[14] 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.

[15] A. Salvador et al., "Learning cross-modal embeddings for cooking recipes and food images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3020–3028.

[16] 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.

[17] J. Marin et al., "Recipe1M+: A dataset for learning cross-modal embeddings for cooking recipes and food images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 187–203, Jan. 2021.

[18] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11572–11581.

[19] B. Zhu, C.-W. Ngo, J. Chen, and Y. Hao, "R2GAN: Cross-modal recipe retrieval with generative adversarial network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 11477–11486.

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

[21] 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.

[22] F. P.-W. Lo, Y. Sun, J. Qiu, and B. P. Lo, "Point2volume: A vision-based dietary assessment approach using view synthesis," *IEEE Trans. Ind. Informat.*, vol. 16, no. 1, pp. 577–586, Jan. 2020.

[23] B. Jing, P. Xie, and E. Xing, "On the automatic generation of medical imaging reports," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguist.*, vol. 1, Jul. 2018, pp. 2577–2586. [Online]. Available: https://www.aclweb.org/anthology/P18-1240

[24] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Hybrid retrieval-generation reinforced agent for medical image report generation," 2018, *arXiv:1805.08298*.

[25] F. Wang, X. Liang, L. Xu, and L. Lin, "Unifying relational sentence generation and retrieval for medical image report composition," *IEEE Trans. Cybern.*, vol. 52, no. 6, pp. 5015–5025, Jun. 2022.

[26] W. Jia et al., "A novel approach to dining bowl reconstruction for image-based food volume estimation," *Sensors*, vol. 22, no. 4, p. 1493, 2022.

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

[28] 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 Inform.*, vol. 18, no. 4, pp. 1253–1260, Jul. 2014.

[29] 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.

[30] 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.

[31] 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 Inform.*, vol. 21, no. 3, pp. 599–606, May 2017.

[32] 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 Ubiquitous Comput. ACM Int. Symp. Wearable Comput.*, 2017, pp. 613–622.

[33] 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, May 2018.

[34] K. Kyritsis, C. Diou, and A. Delopoulos, "Modeling wrist micro-movements to measure in-meal eating behavior from inertial sensor data," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 6, pp. 2325–2334, Nov. 2019.

[35] K. Kyritsis, C. Diou, and A. Delopoulos, "A data driven end-to-end approach for in-the-wild monitoring of eating behavior using smart-watches," *IEEE J. Biomed. Health Inform.*, vol. 25, no. 1, pp. 22–34, Jan. 2021.

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

[37] T. Rahman et al., "BodyBeat: A mobile system for sensing non-speech body sounds." in *Proc. MobiSys*, vol. 14, 2014, pp. 2594368–2594386.

[38] 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. 2010.

[39] T. Olubanjo and M. Ghovanloo, "Real-time swallowing detection based on tracheal acoustics," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)* 2014, pp. 4384–4388.

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

[41] 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 Inform.*, vol. 21, no. 3, pp. 607–618, May 2017.

[42] 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.

[43] J. Liu et al., "An intelligent food-intake monitoring system using wearable sensors," in *Proc. 9th Int. Conf. Wearable Implantable Body Sens. Netw.*, 2012, pp. 154–160.

[44] 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 Sens. Netw. (BSN)*, 2019, pp. 1–4.

[45] J. Lei, J. Qiu, F. P.-W. Lo, and B. Lo, "Assessing individual dietary intake in food sharing scenarios with food and human pose detection," in *Proc. 6th Int. Workshop Multimedia Assist. Dietary Manage. (MADiMa)*, 2021, pp. 549–557.

[46] P. V. Rouast and M. T. Adam, "Learning deep representations for video-based intake gesture detection," *IEEE J. Biomed. Health Inform.*, vol. 24, no. 6, pp. 1727–1737, Jun. 2020.

[47] R. Socher and L. Fei-Fei, "Connecting modalities: Semi-supervised segmentation and annotation of images using unaligned text corpora," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2010, pp. 966–973.

[48] B. Z. Yao, X. Yang, L. Lin, M. W. Lee, and S.-C. Zhu, "I2T: Image parsing to text description," *Proc. IEEE*, vol. 98, no. 8, pp. 1485–1508, Aug. 2010.

[49] M. Mitchell et al., "Midge: Generating image descriptions from computer vision detections," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguist.*, 2012, pp. 747–756.

[50] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3156–3164.

[51] 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.

[52] K. Xu et al., "Show, attend and tell: Neural image caption generation with visual attention," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[53] P. Anderson et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6077–6086.

[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," 2015, *arXiv:1506.01497*.

[55] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 375–383.

[56] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 4651–4659.

[57] X. Li, A. Yuan, and X. Lu, "Vision-to-language tasks based on attributes and attention mechanism," *IEEE Trans. Cybern.*, vol. 51, no. 2, pp. 913–926, Feb. 2021.

[58] L. Yang, H. Wang, P. Tang, and Q. Li, "CaptionNet: A tailor-made recurrent neural network for generating image descriptions," *IEEE Trans. Multimedia*, vol. 23, pp. 835–845, Apr. 2020.

[59] L. Huo, L. Bai, and S.-M. Zhou, "Automatically generating natural language descriptions of images by a deep hierarchical framework," *IEEE Trans. Cybern.*, vol. 52, no. 8, pp. 7441–7452, Aug. 2022.

[60] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 684–699.

[61] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10685–10694.

[62] K. Nguyen, S. Tripathi, B. Du, T. Guha, and T. Q. Nguyen, "In defense of scene graphs for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 1407–1416.

[63] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 7008–7024.

[64] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, "Chinese image caption generation via visual attention and topic modeling," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 1247–1257, Feb. 2022.

[65] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "STAIR captions: Constructing a large-scale Japanese image caption dataset," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguist. (Short Papers)*, vol. 2, 2017, pp. 417–421.

[66] Y. Yang, H. Wei, H. Zhu, D. Yu, H. Xiong, and J. Yang, "Exploiting cross-modal prediction and relation consistency for semisupervised image captioning," *IEEE Trans. Cybern.*, early access, Jul. 27, 2022, doi: 10.1109/TCYB.2022.3156367.

[67] P. Song, D. Guo, J. Zhou, M. Xu, and M. Wang, "Memorial GAN with joint semantic optimization for unpaired image captioning," *IEEE Trans. Cybern.*, early access, May 30, 2022, doi: 10.1109/TCYB.2022.3175012.

[68] A. Vaswani et al., "Attention is all you need," 2017, *arXiv:1706.03762*.

[69] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," 2019, *arXiv:1906.05963*.

[70] G. Li, L. Zhu, P. Liu, and Y. Yang, "Entangled transformer for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 8928–8937.

[71] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4634–4643.

[72] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1–10.

[73] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "Meshed-memory transformer for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10578–10587.

[74] X. Yang, C. Gao, H. Zhang, and J. Cai, "Auto-parsing network for image captioning and visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 2197–2207.

[75] Z. Fei, "Partially non-autoregressive image captioning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2021, pp. 1309–1316.

[76] X. Zhang et al., "RSTNet: Captioning with adaptive attention on visual and non-visual words," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15465–15474.

[77] X. Dong, C. Long, W. Xu, and C. Xiao, "Dual graph convolutional networks with transformer and curriculum learning for image captioning," in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 2615–2624.

[78] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[79] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–17.

[80] R. Mokady, A. Hertz, and A. H. Bermano, "Clipcap: Clip prefix for image captioning," 2021, *arXiv:2111.09734*.

[81] T.-Y. Lin et al., "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[82] C. Fan, Z. Zhang, and D. J. Crandall, "Deepdiary: Lifelogging image captioning and summarization," *J. Vis. Commun. Image Represent.*, vol. 55, pp. 40–55, May 2018.

[83] P. Agarwal, A. Betancourt, V. Panagiotou, and N. Díaz-Rodríguez, "Egoshots, an ego-vision life-logging dataset and semantic fidelity metric to evaluate diversity in image captioning models," in *Proc. ICLR Workshop*, 2020, pp. 1–15.

[84] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, 2017.

[85] 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.

[86] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[87] M. L. Jobarteh et al., "Development and validation of an objective, passive dietary assessment method for estimating food and nutrient intake in households in low-and middle-income countries: A study protocol," *Current Develop. Nutrition*, vol. 4, no. 2, 2020, Art. no. nzaa020.

[88] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[89] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguist.*, 2002, pp. 311–318.

[90] M. Denkowski and A. Lavie, "Meteor universal: Language specific translation evaluation for any target language," in *Proc. 9th Workshop Stat. Mach. Transl.*, 2014, pp. 376–380.

[91] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, 2004, pp. 74–81.

[92] R. Vedantam, C. L. Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 4566–4575.

[93] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 382–398.

[94] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966.

[95] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[96] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[97] X. Chen et al., "Microsoft coco captions: Data collection and evaluation server," 2015, *arXiv:1504.00325*.

[98] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*.

Jianing Qiu is currently pursuing the Ph.D. degree with the Department of Computing, Imperial College London, London, U.K.

His research interests include egocentric vision, surgical vision and robotics, human-object interaction, pervasive sensing, and multimodal learning.

Frank P.-W. Lo received the Ph.D. degree from the Department of Surgery and Cancer, Imperial College London, London, U.K., in 2022.

He is currently a Postdoctoral Research Associate with the Hamlyn Centre for Robotic Surgery, Imperial College London. His current research interests include depth estimation, 3-D reconstruction, signal processing, blind source separation, and healthcare monitoring system.

Xiao Gu, photograph and biography not available at the time of publication.

Modou L. Jobarteh, photograph and biography not available at the time of publication.

Wenyan Jia, photograph and biography not available at the time of publication.

Tom Baranowski, photograph and biography not available at the time of publication.

Matilda Steiner-Asiedu, photograph and biography not available at the time of publication.

Alex K. Anderson, photograph and biography not available at the time of publication.

Megan A. Mccrory, photograph and biography not available at the time of publication.

Edward Sazonov, photograph and biography not available at the time of publication.

Mingui Sun, photograph and biography not available at the time of publication.

Gary Frost, photograph and biography not available at the time of publication.

Benny Lo (Senior Member, IEEE) received the Ph.D. degree from Imperial College London, London, U.K., in 2007.

He is a Reader of the Hamlyn Centre/Dept of Surgery and Cancer, Imperial College London, London, U.K. His research mainly focuses on pervasive sensing technologies, machine learning and medical Internet of things for healthcare, wellbeing, and sports applications.

Dr. Lo's research work has been highly recognized with numerous awards, such as the Best Paper Award IEEE BSN 2021, the Outstanding Interaction Paper Award ICRA 2022. He is an Associate Editor of the IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS and a Topics Editor of the *International Journal of Distributed Sensor Networks*. He leads the B&M Gates Foundation Funded project "An Innovative Passive Dietary Monitoring System" which investigates the use of wearable and AI technologies for dietary assessments.