Mixed-dish Recognition with Contextual Relation Networks

Lixi Deng[1,2] Jingjing Chen[3*] Qianru Sun[4] Xiangnan He[5] Sheng Tang[1] Zhaoyan Ming[6]
Yongdong Zhang[1] Tat-Seng Chua[6]
Institute of Computing Technology, Chinese Academy of Sciences[1] University of the Chinese Academy of Sciences[2]
Shanghai Key Lab of Intelligent Information Processing, School of Computer Science, Fudan University[3] Singapore
Management University[4] University of Science and Technology of China[5] National University of Singapore[6]

ABSTRACT

Mixed dish is a food category that contains different dishes mixed in one plate, and is popular in Eastern and Southeast Asia. Recognizing individual dishes in a mixed dish image is important for health related applications, e.g. calculating the nutrition values. However, most existing methods that focus on single dish classification are not applicable to mixed-dish recognition. The new challenge in recognizing mixed-dish images are the complex ingredient combination and severe overlap among different dishes. In order to tackle these problems, we propose a novel approach called contextual relation networks (CR-Nets) that encodes the implicit and explicit contextual relations among multiple dishes using region-level features and label-level co-occurrence, respectively. This is inspired by the intuition that people are likely to choose dishes with common eating habits, e.g., with multiple nutrition but without repeating ingredients. In addition, we collect a large-scale dataset of mixed-dish images that contain 9, 254 mixed-dish images from 6 school canteens in Singapore. Extensive experiments on both our dataset and a smaller-scale public dataset validate that our CR-Nets can achieve top performance for localizing the dishes and recognizing their food categories.

CCS CONCEPTS

• **Computing methodologies** → **Object detection**; **Object recognition**.

KEYWORDS

Food recognition; Context modeling; Multiple dish detection

ACM Reference Format:

Lixi Deng, Jingjing Chen, Qianru Sun, Xiangnan He, Sheng Tang, Zhaoyan Ming, Yongdong Zhang, and Tat-Seng Chua. 2019. Mixed-dish Recognition with Contextual Relation Networks. In *Proceedings of the 27th ACM International Conference on Multimedia (MM' 19), Oct. 21–25, 2019, Nice, France.* ACM, NY, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3351147

* Jingjing Chen is the corresponding author. chenjingjing.tju@gmail.com.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351147

Figure 1: Food images. (a) is a single dish image, (b) is an example of multiple dish on the School Lunch dataset [12], and (c) shows an example of our mixed dish dataset. In (c), we additionally show the annotation of food category and bounding box.

1 INTRODUCTION

"Keep healthy and fit" is one of the main themes of human life. People care more and more about the calorie and nutrition in their daily food. Among the diverse types of food, mixed dish is one of the most popular food in Eastern and Southeast Asian countries. Some examples are given in Figure 1(c). Recognizing such mixed dish means identifying each of the dish category presented in the mixed dish, which is crucial for calorie estimation as well as nutrition estimation. This is a challenging task, due to the high diversity and the severe overlap among the food components.

For example in Figure 1(c), there are four kinds of dishes and there is no clear boundary between any two of them. In the literature of food recognition, there are several works focusing on the recognition on single dish images [5, 9, 13, 16, 17], and some other works on recognizing the multiple dishes separated in different plates [1, 14, 30, 34]. The image examples are given in Figure 1(a) and (b), respectively. In contrast, we consider a more realistic and challenging scenario in which we need to both localize and classify the individual dishes that are presented on one plate and have severe overlaps, see the example in Figure 1(c).

In order to initiate the study on these problems, we build a dataset by collecting 9, 245 mixed dish images from 6 school canteens in Singapore. As shown in Figure 1(c), we annotate the categories and locations of individual dishes appearing in the image. In the experiments, we leverage the location information to boost the dish identification performance. Additionally, as the data is collected from different canteens, we set a more practical and challenging "cross-domain" evaluation setting that each time the test data comes from a new canteen which is not used during training.

As we have mentioned, the mixed dish data indeed contain high diversities and unclear boundaries. In order to tackle these problems, we propose the novel contextual relation networks (CR-Nets) that aim to encode 1) texture patterns learned from a large-scale

single dish dataset, 2) the implicit feature-level contextual relations among the dishes on the same plate (image), and 3) the explicit co-occurrence contextual relation among their labels. Specifically, we propose to transfer the knowledge learned from single dish classification for multiple dish detection. This is basically an idea of transfer learning that has been widely employed [40, 41]. Besides, to improve the dish recognition performance, we propose to make use of both implicit context and explicit context. Implicit context models the contextual relations between region features in the process of implicite attribute detection [41], aiming to incorporate low-level context information, such as color assortment; while explicit context models the co-occurrence relation among dish labels in the process of recognition refinement, aiming to incorporate the high-level context information such as nutritional mutual replenishment. Overall, our approach encodes implicit and explicit contextual relation information in the perspectives of model, data and label, in order to achieve robust and efficient mixed dish recognition.

Our main contributions are three folds.

- A dataset of mixed dish images annotated with bounding boxes and categories. This aims to provide a realistic testing bed and encourage further research on this topic.
- The novel contextual relation networks (CR-Nets) that aims to encode rich information from the perspectives of model, image and label. It offers a superior representation for the recognition and localization of mixed dishes.
- Extensive experiments on the proposed dataset as well as the existing multi-dish dataset [12]. Our CR-Nets achieves consistent improvements over the state-of-the-art models, especially on the most challenging cross-domain setting.

2 RELATED WORK

Overview on food recognition task. Research literature on food image recognition exhibits a high diversity. The most popular direction is to use deep object recognition models specifically on recognizing food images [4, 6, 9, 13, 17, 26, 27, 29, 31]. Some interesting works proposed to leverage the GPS and restaurant menus [2, 3, 20, 44], some focused on personalized food recognition by history data [19, 22, 45], and some emphasized on multi-modal fusion [21] and real-time recognition [24, 33, 46]. Those most related to ours are on multiple dish recognition [1, 14, 30, 34].

Multiple dish recognition models. The initial work [30] on multiple dish recognition has a two-step pipeline that step-1 detects the plate using either a circle detector or a deformable part model (DPM) [15], and step-2 fuses the image features extracted on the detected regions. The following works use more effective detection models, such as YOLO [35] and Faster R-CNN [36] for dish detection and recognition [12] [14]. In [12], Ege *el al.* used Faster R-CNN to first get the bounding boxes of dish region proposals and then conduct a multi-task learning to predict both dish categories and calories. Later, Ege *el al.* [14] proposed a framework that leverage the better-performed YOLO detectors and obtained a higher detection efficiency. Aguilar *el al.* [1] proposed to combine semantic segmentation models FCN [28] with YOLO for dish detection and recognition. In their method, food and non-food regions are first segmented by FCN and then refined by YOLO. This kind of work requires expensive image annotations, especially for training segmentation models, therefore, it is not applicable to large-scale training. Another work [38] tries to generate foodness proposals with a fully convolutional neural network for multiple dish recognition. Compared to Faster R-CNN and YOLO based methods, this work does not require any bounding box labels. Overall, these mentioned approaches were proposed under the assumption that different dishes are in different containers (plates) and each plate only contains a single type of dish, which is different from the case of our mixed dish recognition.

Multiple dish datasets and recognition models. There are several public food datasets, including VIREO Food 172 [6], UEC Food-100 [30], UEC Food-256 [25], Food-101 [5], ChinFood1000 [16], UNIMIB2016 [1], PFID [8] and School Lunch image dataset [12]. However, most of these food datasets are collected for single dish recognition. The exceptions include UEC Food-100, School Lunch image dataset and UNIMIB2016. Nevertheless, these three datasets are relatively small datasets, ranging from 1,027 to 4,877 food images that cover less than 70 dish categories. Besides, in these three datasets, different dishes are presented in different plates which is the simplest situation for multiple dish recognition. Different to these datasets, we collected a mixed dish dataset which contains 9,254 multiple dish images and covers 164 dish categories. More importantly, this dataset considers the most challenging situation for multiple dish recognition, where different dishes are presented in one plate, and some of them may have overlaps with each other. To the best of our knowledge, we are the first to build a mixed dish dataset annotated with bounding boxes and categories.

There are some related works focusing on recognizing mixed food presented in one plate [10, 32]. For example, Myers *et al.* [32] proposed to use deep convolutional neural network (CNN) and conditional random field (CRF) to predict the pixel level labels which is also called multiple dish segmentation. Dehais *et al.* [10] proposed a CNN-based food border map to guide the region growing for food segmentation. These two methods have shown promising segmentation results on western food images on which each food item only contains a single and simple ingredient. In contrast, we aim to handle the more complicated Asian food images on which each term has higher diversity as well as more complex mixed ingredients.

3 CONTEXTUAL RELATION NETWORKS (CR-NETS)

Figure 2 shows the training pipeline of our proposed CR-Nets. It mainly contains three steps in which the network respectively extracts three kinds of contextual relation information from mixed dish data. (1) Pre-training on the large-scale dataset of single dish images. The trained model contains the content (ingredient) information that is tightly related and helpful to the recognition of mixed dishes. (2) Training the dish localization and recognition model on our proposed mixed dish dataset. In this step, the contexts among co-occurring dishes are implicitly encoded in the representation features. (3) Incorporating the label co-occurrence semantics. This step modifies the classification scores (obtained in the last step) based on modeling the dish co-occurrence in the label space.

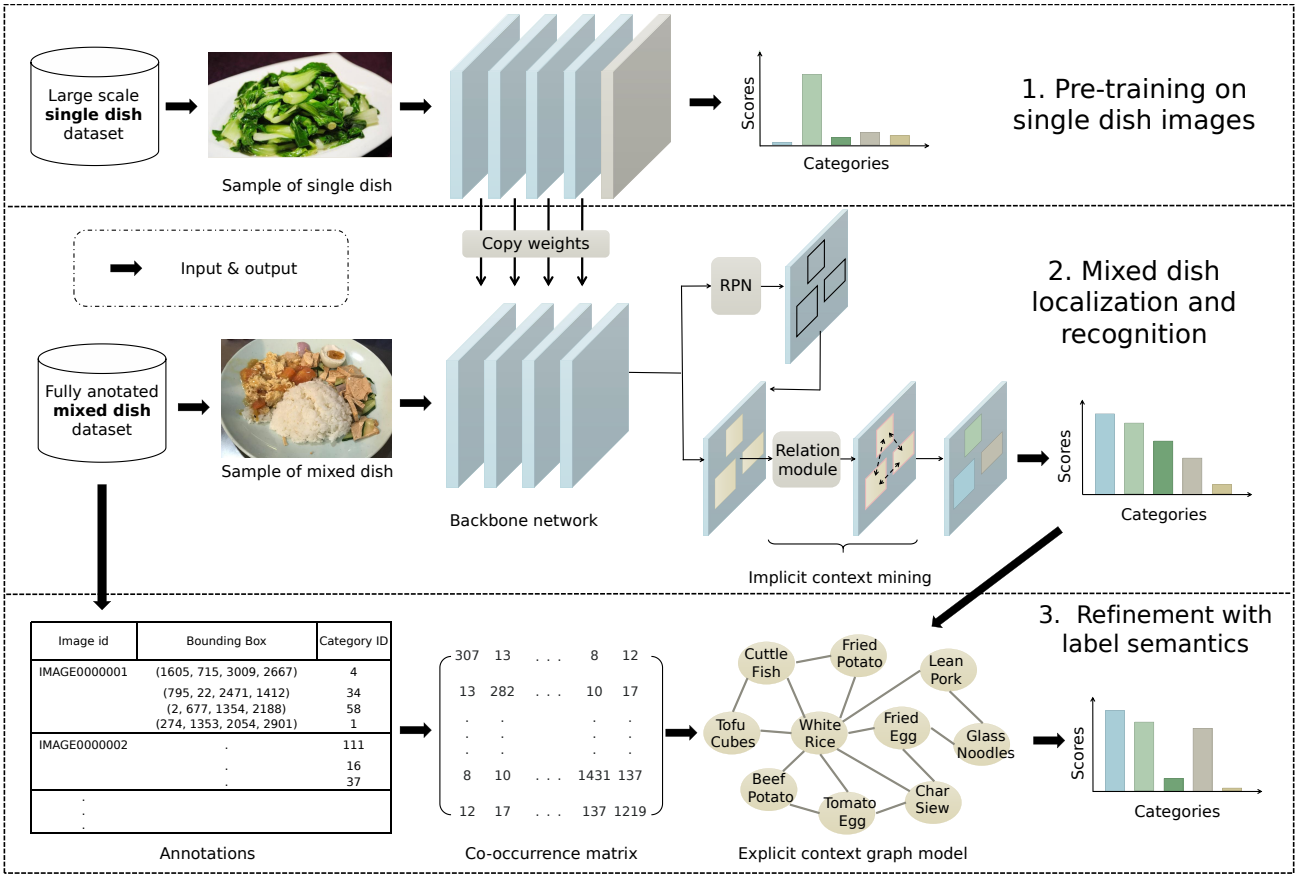

Figure 2: The overview of the proposed CR-Nets for mixed dish recognition. The top block illustrates the pre-training of the feature extractor on a single dish dataset. The middle block shows the region localization and relation encoding on the mixed dish image. The bottom block presents the final step that incorporates the label co-occurrence semantics for refining the recognition results.

3.1 Pre-training on single dish images

As shown in the first block of Figure 2, we train the backbone network of our CR-Nets, e.g. ResNet-50, on the large-scale single dish dataset. Then we copy the pre-trained weights to CR-Nets to enable a fast model adaptation to the mixed dish data. Specifically, the single dish image dataset used in this step contains 264,048 images that cover 751 food categories in southeast Asia. Compared to mixed or multiple dishes, single dish image collection and annotation are cheaper and easier. In fact, a number of existing datasets such as VIREO Food-172 [6] or Food 101 [5] can also be directly leveraged.

It is a fact that mixed dish is the composition of multiple single dishes. When CR-Nets copies the weights and bias from the pre-trained model, it obtains the shared image patterns from the single dish which enables a warm start for the following fine-tuning on mixed dish data. In experiments, we have a careful ablation study on this transfer learning. We conduct pre-training on either the widely used ImageNet [11] or the single dish datasets [5, 6]. The superiority of the second one is clear and consistent in both self-domain and cross-domain settings.

3.2 Mixed dish localization and recognition

As shown in the second block of Figure 2, given a mixed dish image, it goes through the network to locate and classify the dishes it has. This procedure is performed by combining a Faster R-CNN model with a relation module which encodes the feature co-occurrence. A similar application of relation module has been demonstrated to exhibit high efficiency for object detection [23]. In details, the combined network includes a backbone network for feature extraction, a region proposal network (RPN) for localization, and a region recognition network that contains the relation module for classification.

Given a mixed dish image I, we use ResNet-50 [18] as the backbone network to generate $conv4$ feature map, denoted as f. Then, we extract the regional proposals on f to generate the candidate regional locations $\mathbb{P} = \{p_1, p_2 ... p_m\}$. With the feature map f and candidate region locations $\mathbb{P}$ as inputs, we use RoI-pooling to obtain the $7 \times 7 \times 2048$ dimension features. These features go through two fully connected layers to output the region features $\mathbb{F} = \{f_1, f_2 ... f_m\}$. The output is a 1024-dimension feature.

Each feature presents a single image region. We explicitly encode the contextual relations using relation modules. Intuitively, each region may have several implicit relations with the others, such as dish co-occurrence, boundary mixture and overlaps, which are helpful for constructing a richer dish representation. The computing flow of the relation module is given in Figure 3.

Given the n-th region, f_n denotes its feature and p_n denotes the position. Notes that there are different types of implicit relations. Let $f_R^i(n)$ as the relation feature for n-th region under i-th relation type, $f_R^i(n)$ is obtained by

$$f_R^i(n) = \sum_{m=1}^{M} w^{mn} \cdot (\mathbf{W_V} \cdot f_m), \tag{1}$$

where M is the number of regions, f_m is the region feature for m^{th} region, and $\mathbf{W_V}$ is the transformation matrix. Besides, w^{mn} is the weight that indicates the impact of m-th region on n-th region, which considers both the locations and features of m-th and n-th region. Specifically, w^{mn} is computed as follows,

$$w^{mn} = \frac{w_P^{mn} \cdot exp(w_A^{mn})}{\sum_k w_P^{kn} \cdot exp(w_A^{kn})}, \tag{2}$$

where w_A^{mn} and w_P^{mn} represent the weights measuring the importance of m-th region to the n-th region according to appearance feature and location feature, respectively. The w_A^{mn} is obtained as follows,

$$w_A^{mn} = \frac{dot(\mathbf{W_K} f^m, \mathbf{W_Q} f^n)}{\sqrt{d_k}}, \tag{3}$$

where $\mathbf{W_K}$, $\mathbf{W_Q}$ are the transformation matrices, mapping the region features to a lower-dimensional space. d_k is the dimension of transformed features. w_P^{mn} takes the position features of two regions as inputs and embeds them into a vector denoted as ε_p using the method presented in [42]. Besides, w_P^{mn} is obtained by

$$w_P^{mn} = max\{0, \mathbf{W_P} \cdot \varepsilon_P(p_n, p_m)\}, \tag{4}$$

where $\mathbf{W_P}$ is the transformation matrix.

After obtaining N_r relation features, the relation module concatenates them together and sums them with the original region feature f_n to get the new region feature $f_n^{'}$

$$f_n^{'} = f_n + Concat\left[f_R^1(n), ..., f_R^{N_r}(n)\right]. \tag{5}$$

Finally, we obtain the image feature that is assumed to combine the transferred image patterns (from the pre-trained network) and contextural information (from co-occurring image regions). This is followed by the bounding box regression to localize the dish and $softmax$ operation to recognize the dish category.

3.3 Dish recognition refinement with label semantics

As demonstrated in the last block in Figure 2, we refine the recognition scores obtained in the second step, using the holistic co-occurrence statistics of the dish labels. Such statistics depict the co-occurrence of dish categories. They also reveal the geographic information that certain dishes are from certain canteens. Basically, such contextual information can be implicitly extracted by the relation modules during the training in the second step. To enhance its

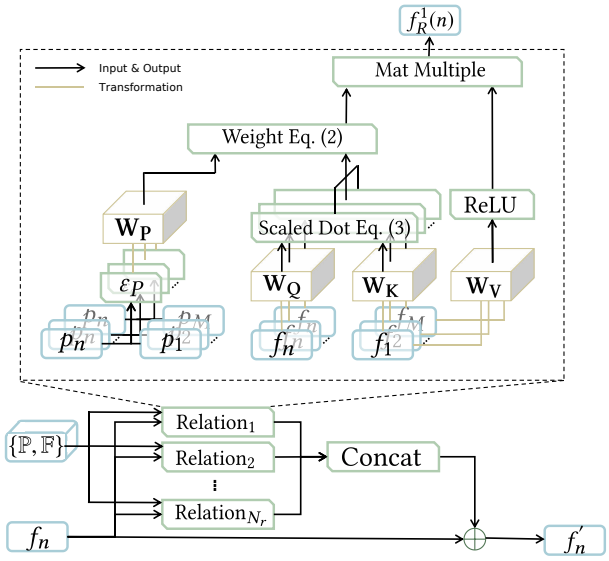

Figure 3: The computation flow of our relation module.

effect on the challenging mixed dish recognition task, its semantic representation is explicitly mined and modeled in our method. In experiments, we conduct ablation study to show its efficiency.

We first compute the recognition scores of all proposals detected (by Faster R-CNN) in the second step. We then filter out those with confidence scores of below 0.001. For the rest, we refine their scores by the following steps.

Let $C = \{c_1, c_2, ..., c_n\}$ be the set of food categories and denote n as the number of the set. The graph G is composed of n vertices V representing the food categories and edges E as the explicit co-occurrence relationships between dishes, with the operation denoted as $\phi()$. Given the graph G, let $P(C)$ be the joint probability of food categories. It is calculated as follows,

$$P(c_1, c_2, ..., c_n) = \frac{1}{Z(\phi)} exp\Big(\sum_{i,j \in C} s_i s_j \phi(i,j)\Big), \tag{6}$$

where s_i indicates the probability of presence of category i, and $Z(\phi)$ is the normalization factor as follows,

$$Z(\phi) = \sum_{C} exp\Big(\sum_{i,j \in C} s_i s_j \phi(i,j)\Big). \tag{7}$$

To learn the graph, we approximate $Z(\phi)$ by Monte Carlo integration and optimize $\phi(\cdot)$ with gradient descent method as in [37]. The energy function consists of unary and binary potentials is then computed by

$$E(y) = \sum_{c \in C} \Phi_u(y_c) + \sum_{(c,v) \in E} \Phi_p(c, v), \tag{7}$$

where E is the set of pairwise cliques. The unary term $\Phi_u(y_c)$ contains the set of dish categories predicted by the second step of dish detection. It is computed by $\Phi_u(y_c) = -log(x_c)$, where x_c is the recognition scores of Faster R-CNN. The pairwise potential $\Phi_p(y_c, y_v)$ demonstrates the joint distribution of dish c and dish

v. For the pairwise co-occurrence, a binary potential is defined as follows,

$$\Phi_p(y_c, y_v) = \sum_{m=1}^{M} w^m(F_{co}(c, v)), \tag{8}$$

where $F_{co}(c, v)$ is the co-occurrence function defined as $F_{co}(c, v) = \frac{co(c,v)}{count(c)+count(v)}$. $co(c, v)$ is the count of co-occurrence between c and v, recorded in the co-occurrence matrix. $count(c)$ and $count(v)$ represent the count of categories c and v, respectively. We employ loopy belief propagation [43] for the minimization. At the inference stage, CRF re-weights the scores of recognition of dishes with context co-occurrence relationship captured in graph G.

4 DATASET CONSTRUCTION

In this section, we introduce the proposed mixed dish dataset. Details are given for image collection, label annotation and image-label statistics.

4.1 Image collection

Images are collected from 6 different school canteens. The lighting conditions and the plate colors vary among different canteens. In 5 canteens, we use cellphone cameras to capture dish photos. For each mixed dish sample, we take 2 pictures from 2 different shooting angles. In the 6-th canteen, we use the camera mounted on the canteen wall to collect videos and then crop the sequence of images to the same size. In total, the dataset contains 9,254 images.

In Figure 4, we show some examples of our dataset. Apart from the unclear boundaries between different dishes, the challenge of this dataset also comes from the fact that the visual appearances of same dish in different canteen can exhibit huge visual variances. For example, the "okra" in Figure 4(a) looks quite different from the "okra" in Figure 4(b), because of different cooking and cutting methods methods.

4.2 Dish annotation

To instruct the annotation process, we collect the list of dish names for each canteen. By merging the overlapped dish categories from different canteens, we totally get 164 categories.

On each image, we annotate the dish categories with each dish in a single bounding box. Annotating bounding boxes on hundreds of dishes for nearly ten thousands of images is extremely tedious. When considering the confusing dish boundaries, it is more tough. First, some foods are mixed in economic rice since all food are placed in one plate, leading to the confusing boundary. Second, some dishes are accidentally separated, or the portion of dishes are obscured leading to some dishes appearing in different positions. Third, certain foods such as "seaweed chicken" consist of several separate integral parts whose number is uncertain, and each part can be considered as a dish. Hence, we stipulate some principles of annotation for eliminating the divergence. If the ingredients of the dish concentrate in the continuous area, regardless of the mixture, all ingredients should be surrounded by the same bounding box. Otherwise, even if the separated parts are logically the ingredients of the same dish, these separated parts are labelled with multiple bounding boxes to minimize the area of the mixture of dishes.

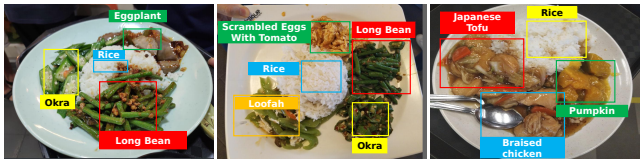

Figure 4: Three mixed dish examples on our new dataset. They are collected from three canteens with different lighting conditions and backgrounds e.g. the shapes and colors of plates.

We recruited 7 trained operators to label the mixed dish dataset. The operator was instructed to label the dishes following the aforementioned principles. In order to guarantee the accuracy of the labels and bounding boxes, we cropped patches from all images according to the annotated bounding boxes to review the annotations. By putting the cropped patches which belong to the same label together, we carefully checked the annotations and corrected the wrong annotations. The entire annotation process took a month.

4.3 Statistics

After annotation, our dataset contains a total of 39,668 bounding boxes and each images contains 4.28 bounding boxes on average. Figure 5(a) shows the distribution of positive samples in dish categories. On average, there are 241 bounding boxes per category. Figure 5(b) further shows the distribution of the annotated bounding box sizes. The sizes of the bounding boxes vary from 0.14% to 71.62% of the image size, and the average sizes of the bounding box is 15.12% of the image size.

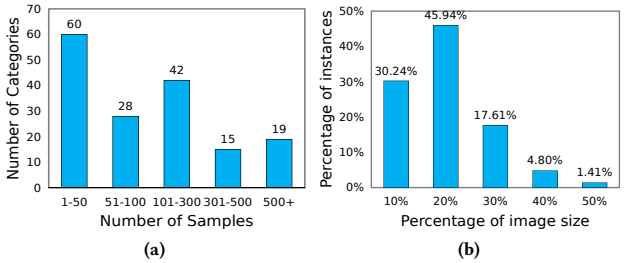

Figure 5: (a) The distribution of dish categories; (b) The distribution of bounding box size.

5 EXPERIMENTS

5.1 Data Splitting

We perform experiments on two datasets, one is the collected mixed dish dataset, and the other one is the School Lunch dataset [12]. As the images in the mixed dish dataset are collected from 6 canteens, we therefore have two manners in terms of data splitting: self-domain splitting and cross-domain splitting. Self-domain splitting ignores the canteen information and randomly split the images into three parts. Among them, 80% of images are selected as training data, and 10% of images are validation data. The remaining 10

Table 1: Contributions of knowledge transfer (Trans), implicit context (IC) and explicit context (EC) on mixed dish dataset. Note that EC is introduced to refine the dish recognition rather than the dish detection results, hence we only report F1 for the models with EC.

	Trans?	IC?	EC?	Cross-domain		Self-domain	
				mAP (%)	F1 (%)	mAP (%)	F1 (%)
(a)				40.63	49.4	73.49	84.59
(b)	✓			44.96	51.05	74.17	84.83
(c)		✓		44.39	51.47	76.32	86.24
(d)			✓	-	50.86	-	85.56
(e)	✓	✓		**47.57**	52.58	**76.89**	86.89
(f)		✓	✓	-	52.37	-	87.63
(g)	✓		✓	-	52.22	-	85.59
(h)	✓	✓	✓	-	**53.55**	-	**88.42**

% images are used for testing. Cross-domain splitting retains the canteen information and splits the data according to the canteen. Under cross-domain setting, images from 5 canteens are chosen as the training data, and the images from the remaining 1 canteen are equally split into validation and testing set. In fact, this is a more realistic setting, as collecting training samples that cover all the canteens is not possible. The evaluation is repeated 6 times such that one testing is performed on each canteen. For evaluation, we only consider the categories of dishes that appear in both training and testing sets. We report the average performances.

The School Lunch dataset is a multi-dish dataset that contains 4,877 images and covers 21 dish categories. Each dish in the images is annotated with a bounding box. Different from our mixed-dish dataset, in the School Lunch dataset, different dishes are presented in different containers and there are clear boundaries between different dishes. Besides, all the images are collected from one canteen. Therefore, this dataset is less challenging. We also conduct experiments on this dataset to verify the effectiveness of the proposed framework on general multi-dish scenarios. Similar to mixed dish dataset, 80% of the images are randomly selected for training, 10% is used for validation set and the remaining 10% is used for testing.

5.2 Implementation Details

The backbone network is initialized with the first four blocks of ResNet-50 network that is pre-trained on large-scale single dish dataset. In this way, the knowledge learned from single dish image recognition is transferred to multi-dish detection. For dish detection, the input images are resized to make their short side equals to 600 pixels. The detection model is trained on 1 NVIDIA TITAN V GPU with the batch size set to 2. The learning rate is set to 0.001, and decays by a factor of 0.1 after the 7^{th} and 12^{th} epoch. In total, the model is trained for 20 epochs. For dish detection, similar to that used in [12], we adopt mean Average Precision (mAP) as the evaluation metric. As our goal is to identify each of the dish type presented in the multiple dish image, we also report the F1 score of the dish recognition results. For dish recognition refinement, the CRF model is trained with the validation set.

Table 2: 10 dish categories that achieve large margin of improvement with knowledge transfer (Trans) on self-domain splitting.

Category	ΔmAP (%)	Number of Samples
Meat Ball	58.50	28
Curry Chicken	50.00	33
Beef Potato	31.15	49
Steamed Bread	23.33	23
Celery	21.05	82
Braised Duck	19.82	98
White Radish	16.51	227
Chicken & Potato	14.41	122
Petai with Prawn	14.29	61
Potato Slice	13.33	44

5.3 Ablation Study

We first investigate the effect of each sub-module in our proposed framework. Table 1 lists the contribution of knowledge transfer, implicit context, explicit context and their combinations towards the performance improvement on mixed dish dataset. From the results, we have the following observations. First, the performance of multiple dish detection and recognition on cross-domain scenario are much lower than that on the self-domain scenario, which demonstrates that cross-domain recognition is more challenging since the same dish provided by different canteens have different visual appearances. Second, with knowledge transfer, it attains higher performance of dish detection on both cross-domain and self-domain settings. In terms of mAP, it improves by around 4% on cross-domain setting and 0.7% on self-domain setting. It's worth noting that the improvement gained from knowledge transfer on cross-domain is larger than that on self-domain. This is maybe because that pre-training on single dish dataset enables the system to learn better features to cope with the visual variance of the same dish in different canteens. Third, with implicit context, the performance of dish detection improves by around 4% on cross-domain setting and 3% on self-domain setting. The results demonstrate that implicit context is quite effective in improving the dish detection performance. Fourth, explicit context improves the dish recognition performances by around 1% on both the cross-domain and self-domain settings, which verifies that label-occurrence is also useful for mixed dish recognition. Fifth, compared with using only implicit context or explicit context, considering both contexts achieves better mixed dish recognition performance. This demonstrates that implicit context and explicit context are complementary to each other. Lastly, by combining knowledge transfer, implicit context and explicit context, the model achieves the best mixed dish recognition performance on both settings.

To obtain deep insights on how the transfer learning influences the detection results, we list 10 dish categories that gain large improvement from knowledge transfer in Table 2. Basically, most of these categories have a limited number of training samples, which is insufficient for learning good discriminative features. By transferring the knowledge learned from the single dish dataset, the model is able to learn good food features and quickly adapt to the task of multiple dish detection with only a few training samples.

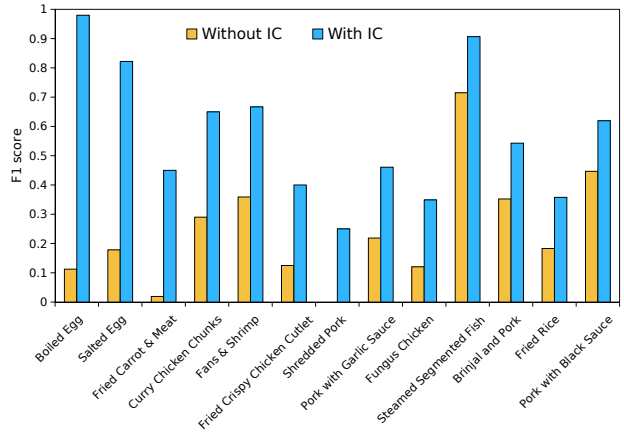

Figure 6: 13 dish categories that achieve large margin of improvement with implicit context on self-domain splitting.

For example, “meat ball" contains only 28 training samples. With knowledge transfer, performance improvement can be as high as 58.5% in terms of mAP.

Figure 6 further lists 13 dish categories that gain large improvement from implicit context. As shown in the Figure 6, dishes that appeared in small size (small portion) in the images, such as “boiled egg" and “salted egg”, gain large improvement from implicit context. The detection performance of “boiled egg” and “salted egg” is 87% and 64% respectively. This is mainly because the implicit context considers the interaction between region features and enhances the features of the dish in small size.

Table 3 shows a few categories that record large improvement with explicit context in terms of F1. For most of the categories, explicit context helps to reduce the false positive predictions, hence lead to higher precision. The examples include “fungus & chicken” and “chicken leg”, in which both achieve 50% improvement in terms of precision. This is mainly due to the fact that people tend not to order dishes with repeating ingredients. There are 16 dishes with chicken as the major ingredient in mixed dish dataset. By modeling the co-occurrence among dishes, the explicit context is effective in reducing false positive predictions that share the same major ingredients. While increasing the precision, explicit context sometimes decreases the recall. For example, by considering label co-occurrence, the recall of both “fried segmented fish” and “fungus & chicken” decrease by more than 25%. This is mainly due to the facts that the training sample of these two categories are relatively small, which results in bias in the learned co-occurrence matrix and leads to decrease in recall on testing set.

Figure 7 further shows three examples, comparing the influence of knowledge transfer and implicit context on mixed dish detection. In general, with knowledge transfer and implicit context, the model can achieve better dish detection results. With knowledge transfer, the model is able to learn better features and reduce the confusion between visually similar dishes. For example, in Figure 7(a), despite “beef stomach” has similar visual appearance with “stir-fried meat”, the model with knowledge transfer successfully remove the false detection of “beef stomach”. Another example is “braised pork trotter” in Figure 7(b). Without knowledge transfer,

Table 3: The changes of recall, precision and F1 score of 10 categories that achieve large margin of improvement with explicit context on self-domain splitting.

	ΔPrecision (%)	ΔRecall (%)	ΔF1 (%)
Chicken Leg	50.00	50.00	33.33
Crab	25.00	25.00	14.29
Glass Noodles	20.00	20.00	11.11
Chinese Cabbage	14.19	37.26	9.97
Potato Slice	13.33	33.33	8.89
Marinated Tofu	15.37	18.07	8.59
Shrimp & Celery	11.19	38.46	8.02
Beans & Eggs	14.12	14.12	7.63
Fried Segmented Fish	33.33	-26.67	7.14
Fungus & Chicken	50.00	-25.00	6.67

the model confuses “chicken & potato” with“braised pork trotter” and predicts “braised pork trotter” with a higher confidence score. Through knowledge transfer, the false detection “braised pork trotter” has been successfully removed. While knowledge transfer helps to remove false defections, implicit context, on the other hand, is effective in improving the recall. As implicit context models the contextual relations between region features, it benefits the predictions of dish in small size (portion). For example, as shown in Figure 7(b) and (c), despite that “salted egg” and “scrambled egg” are small in size, our model is still able to detect them with high confidence score with the help of implicit context.

5.4 Performance comparison

We compare our approach with several baseline methods on both mixed dish dataset and School Lunch dataset. On the proposed mixed dish dataset, we compare with 4 baseline methods which are listed as follows.

- **ResNet-50.** We fine-tune ResNet-50 pretrained on ILSVRC with our mixed dish dataset. As mixed dish recognition is a multi-label recognition problem, we replace the soft-max loss with sigmoid cross-entropy loss.
- **ResNet-50*.** In order to transfer the knowledge learned from single dish recognition, we first pre-train ResNet-50 on large scale single dish dataset, and then fine-tune the model for mixed dish recognition.
- **Region-wise.** The region-wise recognition model divides the feature map of the input image into several grids and performs classification on each grid. The final recognition results are obtained by max pooling the probability distribution across different regions. As illustrated in [7], performing multi-label recognition on region level could lead to significant performance improvement as compared with that at image-level. Hence we also compare the proposed approach against region-wise multi-label recognition method proposed in [7].
- **Faster R-CNN.** Faster R-CNN is the state-of-the-art method on School Lunch dataset [12] for multiple dish detection and recognition. Therefore, we also make Faster R-CNN as one of the baseline. We set a threshold on the confidence score of the detected bounding boxes, and evaluate the recall, precision and F1 score. The threshold is set to 0.5.

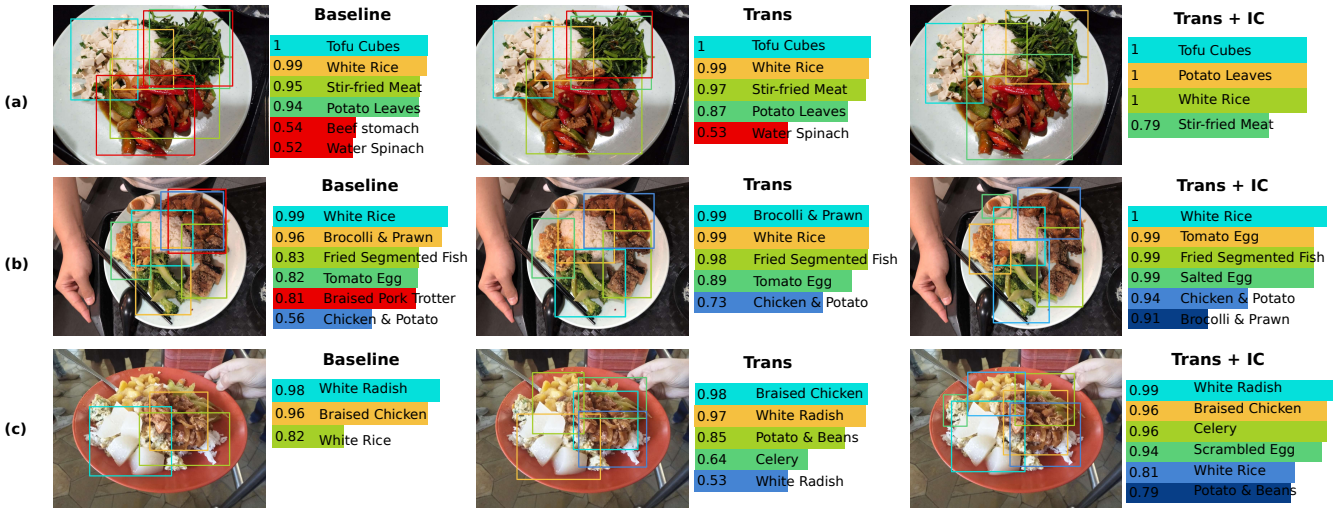

Figure 7: Examples of test images showing effect of transfer learning and implicit context in improving dish detection. False positives are marked in red.

Table 4: Performance comparison of our approach with various existing methods on mixed dish dataset. The evaluations are done on self-domain setting.

	Precision (%)	Recall (%)	F1 (%)
ResNet-50	44.70	44.79	44.74
ResNet-50*	49.79	49.87	49.82
Region-wise	70.92	70.85	70.88
Faster R-CNN	86.53	82.73	84.59
CR-Nets	**87.74**	**89.12**	**88.42**

Table 4 lists the performance comparison. With pre-training and region-wise multi-label recognition, the performance of mixed dish recognition has gain significant improvement. Faster R-CNN, which requires the bounding box annotation for model training, performs much better than multi-label learning methods for mixed dish recognition. The results basically demonstrate the advantages of studying the mixed dish recognition problem from the perspective of object detection. Compared to Faster R-CNN, our method improves the F1 from 70.88% to 88.42%, which demonstrate that our method is more effective in solving the mixed dish recognition problem.

Table 5 lists the performance comparison between our methods and Faster R-CNN on School Lunch dataset. Faster R-CNN, which utilizes VGG [39] as backbone network, has been reported to achieve the best performance on School Lunch dataset [12]. For fair comparison, we re-implement the Faster R-CNN with ResNet-50 network and report the performances. From the results, for dish detection, our method that utilizes knowledge transfer and implicit context performs better than Faster R-CNN, which improves by 0.36% in terms of map. For dish recognition, our method improves by 1.4% as compared with Faster R-CNN. The results basically verify the effectiveness of our method.

Table 5: Performance comparison on School Lunch dataset.

	mAP(%)	F1(%)
Faster R-CNN (VGG) [12]	90.7	-
Faster R-CNN (ResNet-50)	93.55	93.94
CR-Nets (Trans + IC)	93.91	95.26
CR-Nets	-	**95.30**

6 CONCLUSION

In this paper, we introduced a novel approach named CR-Nets for mixed dish recognition. Three kinds of contextual relations from the model, data and label perspectives are explicitly encoded in the model. In addition, we collected a mixed dish dataset containing over 9k images from 6 school canteens. Extensive experiments on both our dataset and a public Japanese School Lunch dataset validate the effectiveness and efficiency of the proposed CR-Nets. There are two interesting directions for future work. First, there remains a significant performance gap between the cross-domain and self-domain settings. How to leverage few-shot learning technique to improve the performance of the cross-domain scenario is one of the direction that is worth investigating. Second, the proposed framework models implicit context and explicit context separately, which makes it difficult to train the model in end-to-end fashion. Therefore, incorporating the explicit context into the detection framework for end-to-end learning is another direction.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (61525206,61572472,61871004) and NExT++ project supported by National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative.

REFERENCES

[1] Eduardo Aguilar, Beatriz Remeseiro, Marc Bolaños, and Petia Radeva. 2018. Grab, Pay, and Eat: Semantic Food Detection for Smart Restaurants. *IEEE Transactions on Multimedia* 20, 12 (2018), 3266–3275.

[2] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. 2015. Menu-match: Restaurant-specific food logging from images. In *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 844–851.

[3] Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory D Abowd, and Irfan Essa. 2015. Leveraging context to support automated food recognition in restaurants. In *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 580–587.

[4] Marc Bolaños, Aina Ferrà, and Petia Radeva. 2017. Food ingredients recognition through multi-label learning. In *International Conference on Image Analysis and Processing*. Springer, 394–402.

[5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision*. Springer, 446–461.

[6] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 32–41.

[7] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1771–1779.

[8] Mei Chen, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang. 2009. PFID: Pittsburgh fast-food image dataset. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 289–292.

[9] Xin Chen, Yu Zhu, Hua Zhou, Liang Diao, and Dongyan Wang. 2017. ChineseFoodNet: A large-scale image dataset for chinese food recognition. *arXiv preprint arXiv:1705.02743* (2017).

[10] Joachim Dehais, Marios Anthimopoulos, and Stavroula Mougiakakou. 2016. Food image segmentation for dietary assessment. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 23–28.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.

[12] Takumi Ege and Keiji Yanai. 2017. Estimating food calories for multiple-dish food photos. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 646–651.

[13] Takumi Ege and Keiji Yanai. 2017. Simultaneous estimation of food categories and calories with multi-task CNN. In *2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 198–201.

[14] Takumi Ege and Keiji Yanai. 2018. Multi-task learning of dish detection and calorie estimation. In *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*. ACM, 53–58.

[15] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. 2010. Cascade object detection with deformable part models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2241–2248.

[16] Zhihui Fu, Dan Chen, and Hongyu Li. 2017. ChinFood1000: a large benchmark dataset for Chinese food recognition. In *International Conference on Intelligent Computing*. Springer, 273–281.

[17] Hamid Hassannejad, Guido Matrella, Paolo Ciampolini, Ilaria De Munari, Monica Mordonini, and Stefano Cagnoni. 2016. Food image recognition using very deep convolutional networks. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 41–49.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[19] Zhao Heng, Sharmili Roy, Kim-Hui Yap, Alex Kot, and Lingyu Duan. 2018. Personalized Knowledge Distillation-based Mobile Food Recognition. In *Proceedings on the International Conference on Artificial Intelligence (ICAI)*. The Steering Committee of The World Congress in Computer Science, 22–28.

[20] Luis Herranz, Shuqiang Jiang, and Ruihan Xu. 2017. Modeling restaurant context for food recognition. *IEEE Transactions on Multimedia* 19, 2 (2017), 430–440.

[21] Hajime Hoashi, Taichi Joutou, and Keiji Yanai. 2010. Image recognition of 85 food categories by feature fusion. In *2010 IEEE International Symposium on Multimedia*. IEEE, 296–301.

[22] Shota Horiguchi, Sosuke Amano, Makoto Ogawa, and Kiyoharu Aizawa. 2018. Personalized classifier for food image recognition. *IEEE Transactions on Multimedia* 20, 10 (2018), 2836–2848.

[23] Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. 2018. Relation networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3588–3597.

[24] Yoshiyuki Kawano and Keiji Yanai. 2013. Real-time mobile food recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1–7.

[25] Yoshiyuki Kawano and Keiji Yanai. 2014. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *European Conference on Computer Vision*. Springer, 3–17.

[26] Yoshiyuki Kawano and Keiji Yanai. 2014. Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 589–593.

[27] Chang Liu, Yu Cao, Yan Luo, Guanling Chen, Vinod Vokkarane, and Yunsheng Ma. 2016. Deepfood: Deep learning-based food image recognition for computer-aided dietary assessment. In *International Conference on Smart Homes and Health Telematics*. Springer, 37–48.

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3431–3440.

[29] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2018. Wide-slice residual networks for food recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 567–576.

[30] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. 2012. Recognition of multiple-food images by detecting candidate regions. In *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 25–30.

[31] Michele Merler, Hui Wu, Rosario Uceda-Sosa, Quoc-Bao Nguyen, and John R Smith. 2016. Snap, Eat, RepEat: a food recognition engine for dietary logging. In *Proceedings of the 2nd international workshop on multimedia assisted dietary management*. ACM, 31–40.

[32] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision*. 1233–1241.

[33] Zhao-Yan Ming, Jingjing Chen, Yu Cao, Ciarán Forde, Chong-Wah Ngo, and Tat Seng Chua. 2018. Food Photo Recognition for Dietary Tracking: System and Experiment. In *International Conference on Multimedia Modeling*. Springer, 129–141.

[34] Parisa Pouladzadeh and Shervin Shirmohammadi. 2017. Mobile multi-food recognition using deep learning. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 13, 3s (2017), 36.

[35] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[36] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[37] Christian Robert and George Casella. 2013. *Monte Carlo statistical methods*. Springer Science & Business Media.

[38] Wataru Shimoda and Keiji Yanai. 2016. Foodness proposal for multiple food detection by training of single food images. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 13–21.

[39] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[40] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-Transfer Learning for Few-Shot Learning. In *CVPR*.

[41] Qianru Sun, Bernt Schiele, and Mario Fritz. 2017. A Domain Based Approach to Social Relation Recognition. In *CVPR*. 435–444.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[43] Martin J Wainwright, Michael I Jordan, et al. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1, 1–2 (2008), 1–305.

[44] Ruihan Xu, Luis Herranz, Shuqiang Jiang, Shuang Wang, Xinhang Song, and Ramesh Jain. 2015. Geolocalized modeling for dish recognition. *IEEE transactions on multimedia* 17, 8 (2015), 1187–1199.

[45] Qing Yu, Masashi Anzawa, Sosuke Amano, Makoto Ogawa, and Kiyoharu Aizawa. 2018. Food Image Recognition by Personalized Classifier. In *2018 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 171–175.

[46] Weiyu Zhang, Qian Yu, Behjat Siddiquie, Ajay Divakaran, and Harpreet Sawhney. 2015. "Snap-n-Eat": Food Recognition and Nutrition Estimation on a Smartphone. *Journal of diabetes science and technology* 9, 3 (2015), 525–533.