Estimating Food Calories for Multiple-dish Food Photos

Takumi Ege Keiji Yanai
Department of Informatics, The University of Electro-Communications Tokyo
1-5-1 Chofugaoka, Chofu-shi, Tokyo 182-8585 JAPAN
{ege-t,yanai}@mm.inf.uec.ac.jp

Abstract

A food photo generally includes several kinds of food dishes. In order to recognize food images including multiple dishes, we need to detect each dish in food images. Meanwhile, in recent years, the accuracy of object detection has improved drastically by the appearance of CNN. In this paper, we apply Faster R-CNN [10], a major object detection method, to food photos of multiple dishes. In the experiments we verify by using two kinds of food photo datasets. In addition, this food detector is applied to food calorie estimation for food photos of multiple dishes. We use Faster R-CNN as a food detector to detect each dish in a food image, and the food calorie of each detected dish are estimated by image-based food calorie estimation [2]. In this way, we estimate food calories from a food photo of multiple dishes. In this experiment, we collect food photos of multiple dishes with total food calorie of multiple dishes. Then we estimate food calories from food photos of multiple dishes by combining the food detector and image based food calorie estimation.

1. Introduction

In recent years, due to growing of health consciousness, various food photo recognition applications for recording everyday meals have been proposed. However, most of them needs human assistance especially for the case that multiple dishes are included in food photos. Some applications use CNN for food dish classification, while no applications have used CNN-based dish detection so far. Then, in this paper, by taking advantage of recent CNN-based object detection method, we detect food dishes in a multiple-dish food photo.

Currently, a lot of studies on recognition of food photos have been released so far, most of them cannot perform directly for food images of multiple dishes. However, in reality, we often encounter scenes with multiple food dishes and to recognize these dishes individually is tedious. Previous researches [7] of recognition of multiple food dishes use Grabcut, region growing method, DPM (Deformable part model) [3] and researches using the latest based-CNN method are few. Meanwhile, in recent years, the accuracy of object detection has improved drastically by the appearance of CNN. Therefore, in this paper, we use Faster R-CNN [10] to detect food dishes from a food photo. Faster R-CNN is the basis of the latest research on object detection using CNN, and achieves high-speed and highly accurate detection. Currently, faster and more accurate detection methods [9, 6] have been released so far, however in this research we use Faster R-CNN which is the basis of those studies. Since the object detection by Faster R-CNN outputs rectangular areas of each object, it is possible to aware individual objects of the same category. In the case of a food photo of multiple dishes, this means it is possible to estimate each dish area in the whole image. Therefore, if we combine a system that works for a single-label image and object detection, it is possible to recognize more detailed from a detected rectangular area of each object.

Figure 1. Some example of multiple food dishes images

In this paper, we apply this food detector for food calorie estimation from food photos of multiple dishes. Image-based food calorie estimation [2] is limited to food photos which contains only one dish, so it is effective to combine with this food detector for estimating food calories from food photos of multiple dishes. In this experiment, we use image-based food calorie estimation [2] for food calorie estimation from food photos of multiple dishes. We [2] proposed simultaneous estimation of food categories and calo-

ries for food photos. We collected calorie annotated recipe data from the online cooking recipe sites, and trained multi-task CNN that output food calories and food categories from a food photo that contained only one dish.

Recently, in image recognition of computer vision, the methods using CNN monopolizes the highest accuracy of main tasks. In the recognition of food photos, various CNN-based researches have been released so far, for example food categorization and segmentation of food area. However, there are not many researches on multiple food dish photo using CNN.

One of the CNN-based researches of detection of multiple food dishes is the research by Shimoda et al. [12]. In [12], firstly, region proposals are generated by selective search. Secondly, for each region proposal, the food area is estimated by saliency maps obtained by CNN. Finally, overlapped region proposals are unified by non-maximum suppression (NMS). In practice, their method [12] enables segmentation of food area. Because segmentation is a pixel-by-pixel classification, it can be applied to detection. In addition to the above work, Shimoda et al. [13] also has proposed the method which generates region proposals by CNN. In [13], firstly, region proposals are generated by saliency maps obtained by CNN. Secondly, each region proposal is classified. Finally, overlapped region proposals are unified by non-maximum suppression.

Dehais et al. [1] proposed the other method for food dish segmentation. In [1], firstly, the Border Map which represents rough boundary lines of a food region is obtained by CNN. Then the boundary lines of Border Map are refined by the region growing/merging algorithm. In this paper, we use Faster R-CNN that achieves high-speed and highly accurate detection.

Im2Calories by Myers et al. [8], which estimates food categories, ingredients and these regions of each of the dishes included in a given food photo, and finally outputs food calories by calculation based on the estimated volumes and the calories density corresponding to the estimated food category. In the experiment, there is a problem that the calorie-annotated dataset is insufficient and evaluation is not sufficiently performed. In this paper, we collect food photos of multiple dishes with total food calorie of multiple dishes, and then evaluate food calorie estimation for multiple food dishes photos.

To summarize our contributions in this paper, we use Faster R-CNN [10] for multiple food dishes photos. Then we create a food photo dataset by collecting food photos of multiple dishes from online school lunch sites, in which each of the food photos is annotated with bounding boxes of dishes and their food categories. In this experiments, we use this food detector for food calorie estimation from food photos of multiple dishes. To do that, we collect food photos of multiple dishes which is annotated total food calorie, and then evaluate food calorie estimation for multiple food dishes photos.

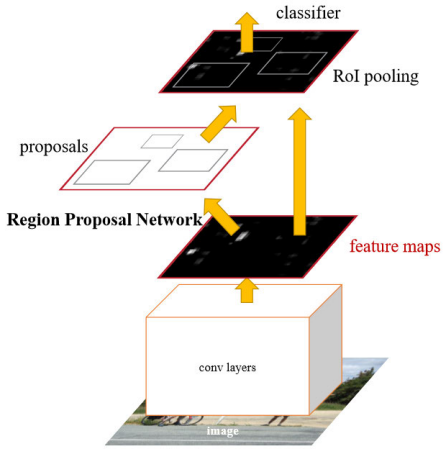

Figure 2. The architecture of network of Faster R-CNN.(This figure is quoted from [11].)

2. Faster R-CNN

In recent years, a lot of methods using CNN have been proposed in the detection task of the visual recognition, and these methods have been greatly successful. For example, an early study is R-CNN [5], which uses CNN for detection. R-CNN generates a large number of region proposals from an input image by selective search, and then uses CNN for each region proposal. Next, there Fast R-CNN [4] is the improved version of R-CNN. In Fast R-CNN, region proposals are generated from a feature map of the entire image, so that CNN is used only once. Then, bounding boxes and class scores for each category are estimated from a region proposal finally. Thus, Fast R-CNN realizes high-speed and highly accurate detection. However, these methods rely on a hand-crafted method of region proposal, and it is bottleneck for the whole system. For these methods, Faster R-CNN [10] generates region proposals by CNN and performs faster and more accurate detection. In this paper, we use the Faster R-CNN for multi-label food detection.

In what follows, Faster R-CNN used in this paper is briefly explained according to [10]. Faster R-CNN is the base of the latest research in the detection task, and consists of two modules. The first module is called Region Proposal Network (RPN) which generates region proposals. This RPN consists of a low-dimensional layer (512 dimensions in VGG16 [14]), followed by a regression layer and a classification layer. The input of RPN is a 3×3 spatial window on the feature map (the output of the last conv layer), and the output is the bounding box positions and object scores. The objectness score is the probability that represents the presence or absence of an object at each bounding-box position. In fact, RPN is implemented by a 3×3 convolutional layer followed by 1×1 convolutional layers. In addition, RPN introduces 9 anchors to deal flexibly with a various shape of an object. This anchor has the shape of 3 scales

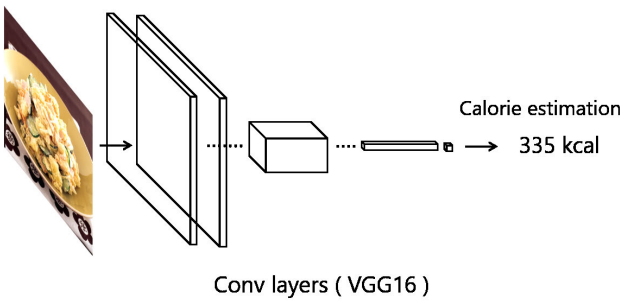

Figure 3. The architecture of food calorie estimation ([2]).

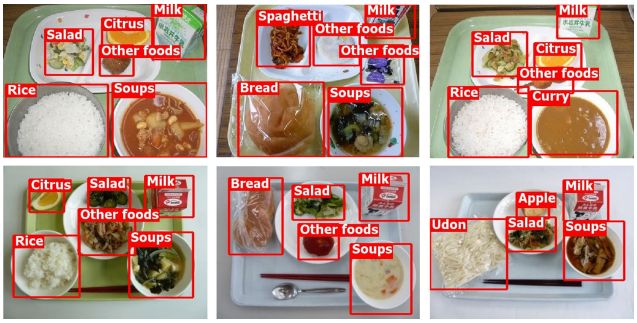

Figure 4. Some examples of school lunch photos in School lunch dataset.

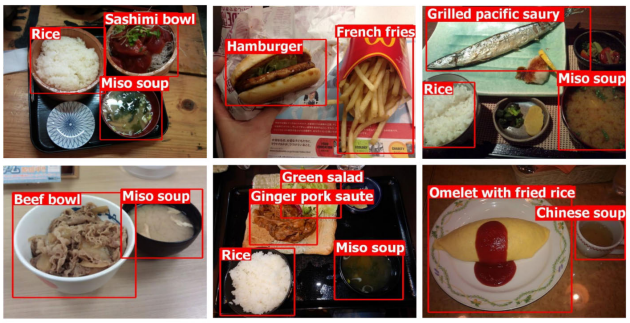

Figure 5. Some examples of multi-label food photos in UEC FOOD-100 [7]

(128^2, 256^2, 512^2 pixels) and 3 aspect ratios (2: 1, 1: 1, 1: 2). Finally, RPN proposes bounding-boxes/objectness scores based on the 9 anchors from a spatial window. The second module is Fast R-CNN detector [4] which refines the region proposals generated by RPN. It is implemented by convolutional layers as well as the RPN. The input of this module is region proposals of RPN, and the outputs are bounding-boxes/class scores for each category. In this way, Faster R-CNN consists of these two modules. Especially, it is important that both RPN and Fast R-CNN networks share a conv layer (Figure 2). An input image forwards conv layers only once, and then region proposals generated by RPN are refined by the Fast R-CNN detector. As a result, Faster R-CNN achieves faster and more accurate detection.

3. Food calorie estimation for food photos of multiple dishes

In this paper, we estimate food calories from food photos according to [2]. In [2], we collected calorie annotated recipe data from the online cooking recipe sites, and trained CNN that output food calories directly from a food photo that contained only one dish. We train the network shown in Figure 3 with the recipe dataset used in [2] for food calorie estimation. The architecture of this network is based on VGG16 [14]. As shown in Figure 3, only the output layer (fc8) is replaced by a single unit which outputs food calorie.

In this experiment, we combine this food calorie estimation network and Faster R-CNN [10] for food photos of multiple dishes. Firstly, we extract bounding boxes of food dishes by Faster R-CNN from a food photo of multiple dishes, and obtain a cropped image corresponding to each bounding box which contains only one dish. Then, we give food calorie estimation network these cropped images one by one. Finally, the total amount of food calories are calculated from food calories estimated from all the cropped images.

4. Datasets

For the evaluation of detection of multiple food dishes, we create a dataset of school lunch images annotated with bounding boxes. In addition, we use UEC FOOD-100 [7] which includes multi-label food images with bounding box annotation. For the evaluation estimating food calorie from multiple food dishes, we also create a dataset of calorie annotated school lunch images. Each image has a total calorie value of all the dishes.

4.1. School lunch dataset with bounding box

This school lunch photo dataset is a food photo dataset prepared in this research which has 21 categories. In Japan, school lunches are usually provided for all school students at lunch time and the menu changes day by day. All 4877 images in this dataset are annotated with bounding boxes. These photos were collected from the school lunch blog[1] [2] on the web being updated by the school lunch centers that is facilities for making school lunches. In these blog, school lunch photos and descriptions about it are posted every day. We create a rough category to cover many dishes, for example fish or meat. Therefore, rough categories contain dishes of various appearances. Figure 4 shows some examples in the built dataset.

[1] http://tate-cook.seesaa.net/
[2] http://blog.canpan.info/takizawa

4.2. UEC FOOD-100

UEC FOOD-100[3] [7] is a Japanese food photo dataset which has multi-label images. This dataset includes more than 100 single-label images for each category, and then 11566 single-label images in total. In addition, this includes 1174 multi-label images in total. All 12740 images in the dataset are annotated with bounding boxes. Figure 5 shows some examples of multi-label images in UEC FOOD-100.

4.3. Calorie annotated school lunch dataset

This school lunch photo dataset is also collected from school lunch blog[4] on the web. This blog posts school lunch photos and food calorie of it everyday. Each school lunch photo of this dataset has a total food calorie. This dataset has 690 images, and is used for food calorie estimation for food photos of multiple dishes.

5. Experiments

5.1. Detection of multiple dishes

In this experiment, to train Faster R-CNN [10], we used an implementation[5] the author released.

We followed the details of this implementation, and we briefly describe it below. We used VGG16 [14], which is pre-trained with the ImageNet dataset. We trained this model by SGD and used a momentum of 0.9 and a weight decay of 0.0005. We used 10^{-3} of learning rate for 50k iterations, and then used 10^{-4} for 20k iterations. To train the two modules, RPN and Fast R-CNN detector, we trained the whole system by Approximate joint training [11]. In the Approximate joint training, the RPN loss and the Fast R-CNN detector loss are back propagated independently, and in the shared conv layer, both losses are combined and back propagated. In this experiment, we trained this model for two datasets.

To evaluate Faster R-CNN, we followed mean Average Precision (mAP) which is the standard evaluation measure of the PASCAL VOC detection task.

5.1.1 School lunch dataset

In this experiment, we used 80% of the food photos in the built dataset for training, and the rest 20% for performance evaluation.

Table 1 shows the results of Average Precision (AP) in each category. As a result of the experiment, mAP was 90.7%, and it is highly accurate detection in most categories. In the school lunch, an arrangement on a dish and its size/volume are roughly unified, so that we think it makes food detection easier. Figure 6 shows some examples of detection results.

[3]http://foodcam.mobi/dataset100.html
[4]http://inzai.ed.jp/kyusyoku/?page_id=32
[5]https://github.com/rbgirshick/py-faster-rcnn

Table 1. The results on detection on School lunch dataset

Milk	99.6
Drinkable yogurt	90.6
Rice	99.7
Mixed rice	82.7
Bread	95.5
White bread	83.7
Udon	98.0
Fish	78.3
Meat	70.8
Salad	94.0
Cherry tomatoes	100.0
Soups	92.2
Curry	95.1
Spicy chili-flavored tofu	99.8
Bibimbap	72.9
Fried noodles	79.9
Spaghetti	90.7
Citrus	99.6
Apple	98.5
Cup desserts	93.1
Other foods	90.4
mean Average Precision	90.7

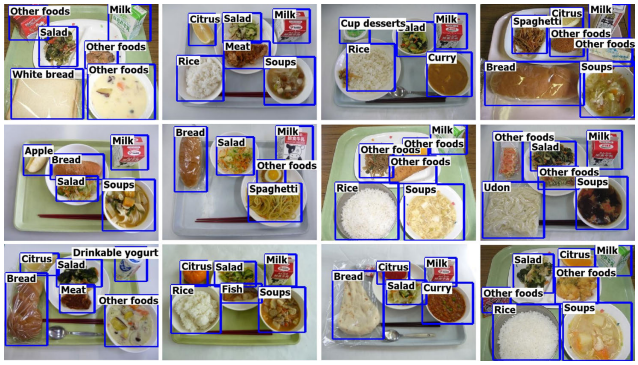

Figure 6. Examples on detection on School lunch dataset.

5.1.2 UEC FOOD-100

In this experiment, to compare with [12], we used all single-label photos in UEC FOOD-100 [7] for training, and all multi-label photos in it for performance evaluation.

For comparison to the exist work on food dish detection based on semantic segmentation using BP by Shimoda et al. [12], according to [12], we show mAP over all the 100 categories, 53 categories (include more than 10 items in the evaluation data), and 11 categories (include more than 50 items in the evaluation data) in Table 2.

Table 2 shows the results of detection with UEC FOOD-100 (Average over food categories). Our result outperformed the result of [12] on 11 class that include a lot of items in evaluation data.

Table 2. The results of detection on UEC FOOD-100 (Average over food categories). R-CNN and [12]'s method are quoted from [12]. mAP over all the 100 categories, 53 categories (include more than 10 items in evaluation data), 11 categories (include more than 50 items in evaluation data).

UEC FOOD-100 mAP(%)	100class (all)	53class (#item≥10)	11class (#item≥50)
R-CNN	26.0	21.8	25.7
[12]'s method (BP)	49.9	55.3	55.4
Faster R-CNN	42.0	46.3	57.9

Table 3. The results of detection on UEC FOOD-100 (Average over all the test images without taking account of categories). mAP over all the 100 categories, 53 categories (include more than 10 items in evaluation data), 11 categories (include more than 50 items in evaluation data).

UEC FOOD-100 mAP(%)	100class (all)	53class (#item≥10)	11class (#item≥50)
[12]'s method (BP)	57.3	58.0	58.8
Faster R-CNN	57.7	59.2	67.0

In particular, the accuracy was improved for food items such as "rice" and "miso soup" which frequently appear in Japanese food photos. Compared with [12], "rice" and "miso soup" improved from 60.0% to 90.2% and from 68.3% to 80.2%. Because these categories often appear in multiple food photos in Japan, it is effective for detection of multiple food photos.

In order to present the result ignoring category, we calculated mAP with weighted average on Table 3 shows the results of detection with UEC FOOD-100 (Average over all the test images without taking account of categories). As shown in Table 3, our result outperformed the result of [12].

The detection speed of Faster R-CNN is about 200 ms on a TITAN X. On the other hand, since [12] uses selective search, it can be thought that detection of [12] takes at least several seconds.

5.2. Food calorie estimation of multiple dishes

In this experiment, for training the network shown in Figure 3, we used Chainer[6] [15] that the framework of Deep Learning, and followed [2] on the details of the parameters. We train the network with the recipe dataset used in [2] for food calorie estimation. This network outputs food calories directly from a food photo that contains only one dish. In [2], they used only 15 food categories in the recipe dataset for multi-task learning of food calories and food categories.

[6]http://chainer.org/

Figure 7. Examples on detection on UEC FOOD-100 [7]

On the other hand, in this work, in order to correspond to every category, we used all recipe data. Then, we excluded photos of multiple dishes, and photos lower than 256×256. For excluding photos with more than one dishes, we used Faster R-CNN trained by UEC FOOD-100 [7]. In the end, we trained the food calorie estimation network by 55020 food images. The size of mini-batch was 8. We used 0.001 of the learning rate for 150k iterations, and then used 0.0001 for 50k iterations. In training term, we performed data augmentation to blur images to address low resolution food images cropped by Faster R-CNN.

As mentioned in Section 3, we combined this food calorie estimation network and Faster R-CNN [10]. Firstly, we extract bounding boxes of food dishes by Faster R-CNN from a food photo of multiple dishes, and obtain crop image of each bounding box which contain only one dish. Then, we give food calorie estimation network these crop images one by one. Finally, the total food calorie are calculated from food calories obtained from crop images. In this experiment, we use calorie annotated school lunch dataset in Section 4.3. Therefore, we use Faster R-CNN in Section 5.1.1 that trained by school lunch dataset. Then in this experiment, since the milk in school lunch is always identical, we fixed the calorie of milk to 134 kcal. The milk is detected by Faster R-CNN, and it is fixed to 134 kcal.

Table 4 shows the results of food calorie estimation for food photos of multiple dishes. We show the average of the relative error representing the ratio between the estimated values and the ground-truth, and the absolute error representing the differences between both. In addition, according to [2], we show the ratio of the estimated value within the relative error of 20% and 40%. As far as we know there is no similar study, so it is difficult to evaluate this results. However, because of the total food calorie, it seems that the relaxation of individual errors are obtained. Figure 6 shows some examples of this results. As Figure 8, food calories are estimated for each dish, finally total food calorie is calculated from the estimated calories of food dishes. The estimated total food calories are strongly influenced by

Table 4. The result of food calorie estimation for multiple dishes. The relative error(rel.err.) representing the ratio between the estimated values and the ground-truth, and the absolute error(abs.err.) representing the differences between the estimated values and the ground-truth. The ($\leq$ 20%err.) and ($\leq$ 40%err.) are the ratio of the estimated value within the relative error of 20% and 40%.

rel.err. (%)	abs.err. (kcal)	$\leq$ 20%err. (%)	$\leq$ 40%err. (%)
21.4	136.8	53.0	85.1

Figure 8. Examples on food calorie estimation on School lunch dataset. The numbers in bounding boxes are estimated food calories of foods in each bounding box (kcal). ES: the estimated total food calorie (kcal). GT: the ground-truth of total food calorie (kcal). Since the numbers in bounding boxes are converted to integer, it does not match the total food calorie(ES).

some detection errors and a part of estimated food calorie. In many cases, the total food calorie are increased by excessive detection that detect food-free background or the same food multiple times. It is thought that it is possible to address with this by processing to integrate these proposal regions or to adjust the non-maximum suppression used in [10].

6. Conclusion

In this paper, we detected food dishes from a food photo of multiple dishes by Faster R-CNN [10]. In this experiment, we created the 21-category school lunch photo dataset by Web image mining. Then, we detected foods in school lunch with high accuracy. In addition, we applied this food detector to food calorie estimation from food photos of multiple dishes. For the evaluation estimating food calorie from food photos of multiple dishes, we also create a dataset of calorie annotated school lunch images. As a results, we estimated total food calories from food photos of multiple dishes.

As future work, we plan to implement multi-task CNN of the food calorie estimation and the food detection. We expect that the accuracy of each task improve by multi-task learning. We also plan to crop higher-resolution images corresponding to the bounding boxes estimated by Faster R-CNN, and use them for more accurate calorie estimation.

References

[1] J. Dehais, M. Anthimopoulos, and S. Mougiakakou. Food image segmentation for dietary assessment. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management*, 2016. 2

[2] T. Ege and K. Yanai. Simultaneous estimation of food categories and calories with multi-task cnn. In *Proc. of IAPR International Conference on Machine Vision Applications(MVA)*, 2017. 1, 3, 5

[3] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 1

[4] R. Girshick. Fast R-CNN. In *Proc. of IEEE International Conference on Computer Vision*, 2015. 2, 3

[5] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014. 2

[6] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *Proc. of European Conference on Computer Vision*, 2016. 1

[7] Y. Matsuda, H. Hajime, and K. Yanai. Recognition of multiple-food images by detecting candidate regions. In *Proc. of IEEE International Conference on Multimedia and Expo*, 2012. 1, 3, 4, 5

[8] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and P. K. Murphy. Im2calories: towards an automated mobile vision food diary. In *Proc. of IEEE International Conference on Computer Vision*, 2015. 2

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, real-time object detection. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2016. 1

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, 2015. 1, 2, 3, 4, 5, 6

[11] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 2, 4

[12] W. Shimoda and K. Yanai. CNN-based food image segmentation without pixel-wise annotation. In *Proc. of IAPR International Conference on Image Analysis and Processing*, 2015. 2, 4, 5

[13] W. Shimoda and K. Yanai. Foodness proposal for multiple food detection by training of single food images. In *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management*, 2016. 2

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *arXiv preprint arXiv:1409.1556*, 2014. 2, 3, 4

[15] S. Tokui, K. Oono, S. Hido, and J. Clayton. Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*, 2015. 5