Proceedings Paper:
Cai, Q. and Abhayaratne, C. orcid.org/0000-0002-2799-7395 (2023) SSDB-Net: a single-step dual branch network for weakly supervised semantic segmentation of food images. In: 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP) Proceedings. 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), 27-29 Sep 2023, Poitiers, France. Institute of Electrical and Electronics Engineers (IEEE) . ISBN 9798350338942

https://doi.org/10.1109/mmsp59012.2023.10337656

SSDB-Net: A Single-Step Dual Branch Network for Weakly Supervised Semantic Segmentation of Food Images

Qingdong Cai and Charith Abhayaratne
Department of Electronic and Electrical Engineering, *The University of Sheffield*
Sheffield, S1 3JD, United Kingdom
qcai7@sheffield.ac.uk, c.abhayaratne@sheffield.ac.uk

Abstract—**Food image segmentation, as a critical task in food and nutrition research, promotes the development of various application domains such as calorie and nutrition estimation, food recommender systems, and daily food monitoring systems. Currently, most of the research is focused on food and non-food segmentation, which simply segments the food and background regions. Differently, semantic food segmentation can identify different specific food ingredients in a food image and provide more detailed and accurate information such as object location, shape and class. This is a more challenging but meaningful task, because the same food may appear in completely different colours, shapes and textures in different dishes, and correspondingly less researched. From the implementation perspective, most previous research is based on deep learning methods with pixel-level labelled data. However, annotating pixel-level labels requires extremely high labour costs. In this paper, a novel single-step dual branch network (SSDB-Net) is proposed to achieve weakly supervised semantic food segmentation. To our knowledge, this research is the first time proposing weakly supervised semantic food segmentation with image-level labels based on convolutional neural networks (CNN). It may serve as a benchmark for future food segmentation research. Our proposal method resulted in an mIoU of 14.79%, for 104 categories in the FoodSeg103 dataset compared to 11.49% of the state-of-the-art WSSS used in other domains.**

Index Terms—**Weakly supervised semantic segmentation, semantic food segmentation, food image analysis**

I. Introduction

In recent years, deep learning has been applied in food research, such as, dish classification [1], [2], food image retrieval [3] and food segmentation [4]. Deep learning-based food segmentation has brought considerable benefits to food-related research, such as calorie and nutrient intake computation [5], personalised food recommendation [6] and daily food/health monitoring systems [7], [8]. However, most of the current research focuses on food and non-food segmentation, which only separates food regions from the non-food regions. These segmentation results are insufficient for further application in accurate food classification or food recommendation [5]. In order to obtaining more detailed and accurate semantic information, semantic food segmentation is proposed. It classifies food region pixels into specific food ingredients and provides finer-grained food segmentation results [9], [10]. However, networks of semantic food segmentation are difficult to train. Since obtaining pixel-level labels ground truth for training networks is time-consuming and labour-intensive. In addition, learning to extract food features is also challenging due to the varied features of the same food in different images [9].

Semantic segmentation algorithms based on deep learning have achieved satisfactory results, but only in a fully supervised setting where the training datasets contain pixel-level annotation [1]. These kinds of datasets are expensive and time-consuming for specific fields. For example, labelling an image from the Cityscapes dataset takes an average of 1.5 hours [11]. In addition, the recently released Segment Anything model (SAM) is trained on a dataset of more than 1 billion masks [12]. Such datasets are difficult to afford for ordinary companies and individuals. In order to alleviate the difficulty of labelling, researchers have proposed weakly supervised semantic segmentation (WSSS) methods [13], which only needs image-level annotations or bounding box-level labelled data for training, but outputs object masks (pixel-wise labels) [14]. In this paper, we explore image-level annotation of WSSS in the food domain.

Image-level annotations do not contain any information about the target shape, size, colour, or how many instances exist in an image. This missing information significantly increases the complexity of the segmentation task. In the semantic food segmentation field, there are problems such as different foods showing similar characteristics, and certain food ingredients showing different features under different cooking methods. These issues further increase the difficulty of training networks in the weakly supervised setting. In order to better complete the weakly supervised semantic food segmentation, we first implement the food WSSS using the classic weakly supervised method, then we propose a new network for better implementation of the food WSSS. The main contributions of our research are as follows:

1 This research is the first attempt to explore WSSS in the food domain utilising image-level annotations to achieve pixel-level result.

2 We propose a novel single-step dual branch network (SSDB-Net) to improve performance for food WSSS

tasks based on image-level annotation.

3 We confirm the necessity of the network retrained on the Food101 dataset [15] to achieve better performance on the FoodSeg103 dataset [9].

II. Related Works

A. Fully and weakly Supervised Semantic Segmentation

After fully convolutional networks were proposed, the CNN-based algorithm became the first choice to complete the semantic segmentation task, and achieved breakthrough performance [16]. However, these studies require a large amount of pixel-level annotation datasets and are difficult to be applied in other specific domains due to pixel-level annotation. In order to reduce data requirements, researchers began to study training semantic segmentation networks with weakly supervised information, whihc is called WSSS. WSSS is usually implemented based on Class Activation Maps (CAM) [17], but CAM only activates the most distinguished regions of objects. For solving this problem, various training strategies have been proposed [13], [18], [19], but these methods are mainly based on general object categories dataset, and there is no research on semantic food segmentation.

B. Food full and weakly supervised semantic segmentation

Food segmentation based on CNN has attracted increasing attention in recent years. In earlier years, a method combining bounding boxes and saliency maps to determine the region of food was proposed [20]. The bounding boxes are applied to provide the region proposal information, and the saliency maps are applied to estimate and refine food regions. This framework, which first roughly locates and then refines, is also utilised by other researchers [21]. They obtained a rough boundary of the food region from CNN and refined the boundary via a region merging and growing algorithm. After them, food and non-food segmentation is achieved by segmenting the background, because researchers found that the features of the background are easier to extract by the network than the features of food [22]. However, the above mentioned studies focus on food and non-food segmentation, which only provides limited semantic information for downstream tasks. In order to obtain richer semantic information, an automatic food analysis system based on the DeepLab algorithm was developed [23] [7]. Similarly, food semantic segmentation is also implemented on the Food201 food database based on the DeepLab segmentation algorithm [5]. Last year, Bayesian theory was introduced into deep algorithms to alleviate inaccurate predictions of CNN [24]. Recently, thermal data (RGB-T) was introduced to achieve food image segmentation by combining RGB food images [25]. The thermal data is obtained by the acquisition equipment designed by the researchers themselves. Their study provides a new perspective on how to achieve food segmentation, which is helpful for developing food segmentation. In addition, excellent review work has been completed about fully supervised semantic food segmentation [10]. The performances of different segmentation algorithms are evaluated based on their own food dataset, which contains 5000 images of 50 different food categories.

To the best of our knowledge, there are only a few studies on weakly supervised food segmentation [26], [27]. However, their research is on food and non-food segments and cannot provide detailed information about the food ingredients. In addition, the proposed method by Wang $et\ al.$ [26] was only tested on their own unpublished dataset. This can only provide limited help for future research on weakly supervised semantic food segmentation. Our study further explores weakly supervised semantic food segmentation on a publicly available dataset, FoodSeg103, with 104 categories [9].

III. Methodology

A. Class activation map and pseudo ground-truth

CAM is widely utilised to generate the initial pseudo segmentation ground-truth for WSSS, because it can identify the object regions when a classification network predicts results. Therefore, it provides an effective way to train networks using image-level annotations but generate masks (pixel-wise labels). Specifically, the last convolution layer feature maps, $\mathbf{F} \in \mathbb{R}^{h\times w\times d}$, are weighted by the classification weights, $\mathbf{W} \in \mathbb{R}^{c\times d}$, to get the activation value **M**. CAM is obtained by eliminating negative activation values and scaling to [0,1] through relu function and max normalisation. Therefore, the CAM calculation equation for class c is as follows:

$$\mathrm{CAM_c(F,W)} = \frac{\mathrm{relu(M_c)}}{\mathrm{max(relu(M_c))}},\ \text{where, } \mathrm{M_c} = \sum_{\mathrm{i\in d}} \mathrm{W_{c,i}F_{:,i}}, \tag{1}$$

where h, w and d denote the height, width and channel of the feature maps, respectively and c denotes the number of classes.

The CAM obtained by Eq. (1) is the activation map of the object class. But, it only implies the probability value of belonging to the target classes. In order to obtain semantic segmentation results, a background class score, $\mathbf{BG} \in \mathbb{R}^{h\times w\times 1}$, needs to be set and concatenated to the final result calculation. Therefore, the calculation equation for masks is as follow:

$$y_{j,k} = \sum_{j,k\in h,w} \mathrm{argmax}(P_{j,k,:}), \mathrm{P} = \mathrm{concat(BG, CAM)}, \tag{2}$$

where j, k represents the corresponding spatial point position. The value of **BG** is generally set according to network structure and empirical. The generated masks are generally applied in subsequent training as pseudo ground-truth.

B. Our single-step dual branch network

Most WSSS methods are based on a multi-step framework. Those methods first obtain the pseudo segmentation ground-truth by training a network under image-level labels. Then they train semantic segmentation based on the pseudo ground-truth. Obviously, the quality of the pseudo ground-truth determines the final result. The inaccurate pseudo ground-truth may introduce too much noise during the second network training

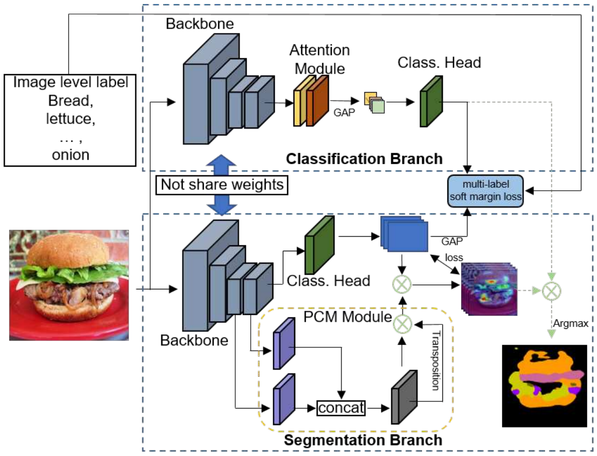

Fig. 1. The SSDB-Net structure with separating-backbone version (SSDB-II)

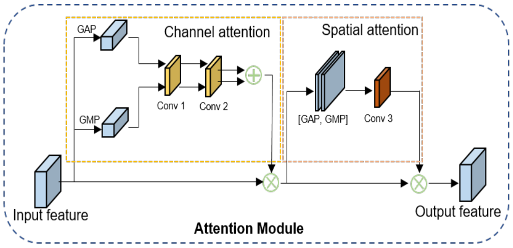

Fig. 2. The channel attention and spatial attention module, the [GAP, GMP] means doing global average pooling and global max pooling Operation to a feature map in channel dimension respectively, and concatenate these result.

process and may lead to decrease of the final performance. To alleviate this problem we propose SSDB-Net which is trained by the image-level label and directly outputs segmentation results. In order to fully consider the performance of the network in the food field and to get more conclusive study the food WSSS, we design two versions of SSDB-Net: one with shared-backbone (we call SSDB-I) and the other with separated-backbone (we call SSDB-II). The biggest difference between the two is whether they share the same backbone network. We only show the separating-backbone version in Fig. 1, because it performs better.

Our SSDB-Net contains two branches, the first is the classification branch, which realises the multi-label classification of food images. The other is for segmentation. In training processing, the classification and segmentation branches are trained simultaneously based on image-level annotations. In the inference stage (the green dotted line part), the classification branch outputs the category results and multiplies them with the segmentation branch results to obtain the CAM of target categories. The final result is obtained according to Eq. 2. Each part of the network will be described in detail in the following sections.

1) backbone network: ResNet-38 is a wider but shallower network than the original ResNet (e.g. ResNet-18, ResNet-50 and ResNet-200), which achieves satisfactory results in both classification and semantic segmentation tasks [28]. The network parameters are modified to adapt the WSSS task. The convolutional layers in the last five blocks are replaced by à trous convolutional layers with stride equals to 1, and padding number and dilation rates are modified to ensure the output stride equals to 1. Moreover, the global average pooling (GAP) layer and the fully connected (FC) layer are removed. Finally, the network output stride equals to 8. This setting ensures the final output feature map is large enough, which is conducive to returning to the original size. It also increases the receptive field of the network to effectively collect the contextual information of the image.

2) Attention module and classification loss: An attention module is added to improve the classification branch performance. This module mainly consists of channel attention and spatial attention modules as shown in Fig. 2. The input features $\mathbf{F} \in \mathbb{R}^{h\times w\times d}$ are fed into GAP and global maximum pooling (GMP) layer to obtain $\mathbf{F}_1 \in \mathbb{R}^{1\times 1\times d}$ and $\mathbf{F}_2 \in \mathbb{R}^{1\times 1\times d}$, respectively. These pooling layers are followed by two trainable convolutional layers. The channel ratios of input and output are 16 and 1/16, respectively. The spatial attention can highlight the value of the object location and suppress the value of the non-target location. The spatial attention mechanism performs a global average operation and a global maximum operation on the input features respectively in the channel dimension. After pooling operation, a convolutional layer with both kernel size and padding equal to 7 is utilised. After the attention module, a global average layer and 1*1 convolutional layer is added to the classification branch to achieve food classification with multi-label soft margin loss. The loss equation is as follows:

$$\begin{aligned} l_{cla}(z,l) = -\frac{1}{C}\sum_{c=1}^{C}[l_c\log(\frac{1}{1+e^{-z_c}}) \\ +(1-l_c)\log\frac{e^{-z_c}}{1+e^{-z_c}}], \end{aligned} \tag{3}$$

where z is the predicted result vector, l is the image label and C is the foreground object category number.

3) Self-supervised equivariant attention mechanism: The Self-supervised equivariant attention mechanism (SEAM) is a classic method in the WSSS domain [19]. It utilises equivariant visual priors, which refers to the same image outputs different CAM at different sizes, to obtain a more complete and accurate CAM. During the training process, the image $\mathbf{I}_o$ is fed into the network, and the classification results $\mathbf{PRE}_o$ is obtained after a 1*1 convolutional layer (the green cube in Fig. 1). $\mathbf{PRE}_o$ indicates the classification result of the image and is applied to obtain the original $\mathbf{CAM}_o^o$. The multi label soft margin loss is calculated based on the global average of $\mathbf{PRE}_o$. On the other hand, the two intermediate layer features $\mathbf{F}_4$ and $\mathbf{F}_5$ are extracted from block 4 and block 5, respectively. They are input into 1*1 convolutional layers (the purple cube) to obtain $\mathbf{F}_4^{'}$ and $\mathbf{F}_5^{'}$ respectively. These feature will concatenated and fed into other 1*1 convolutional layer (the gray cube).

The result after convolution is multiplied by its own transpose to obtain the similarity matrix $\mathbf{M}_o$. This similarity matrix is multiplied with the result of the original $\mathbf{CAM}_o^o$ to obtain the refined result $\mathbf{CAM}_o^r$. In addition, $\mathbf{I}_o$ is scaled to 0.3 times its original size to get $\mathbf{I}_s$. The corresponding result, such as classification result $\mathbf{PRE}_s$, original $\mathbf{CAM}_s^o$, and refine $\mathbf{CAM}_s^r$, is obtained. The multi label classification loss is also calculated by $\mathbf{PRE}_s$. In addition, L_1 loss between $\mathbf{CAM}_o^o$ and $\mathbf{CAM}_s^o$, $\mathbf{CAM}_o^r$ and $\mathbf{CAM}_s^o$, and $\mathbf{CAM}_o^o$ and $\mathbf{CAM}_s^r$ is computed. This module greatly improves the accuracy of CAM.

IV. Performance evaluation

A. Experimental setup

We use two datasets, Food101 and FoodSeg103, to implement semantic food segmentation. Food101 is a food classification dataset, and FoodSeg103 is a semantic food segmentation dataset. Food101 contains 101 food categories, and each category contains about 1000 images. 75% of data is utilised as the training set, and 25% of the data is test set. The FoodSeg103 dataset is a very challenging semantic food segmentation data, which has 104 categories with background. However, there are 7118 images in total, including 4983 in the training set and 2135 in the test set. Each image includes different food categories and instances.

Firstly, the backbone network is retrained on Food101, then the weakly supervised semantic food segmentation is implemented on the Food103 database. Resnet-38 is chosen as the backbone network in this research as mentioned before. During retraining, common data augmentation methods such as flipping, rotation, Gaussian blur, and colour dithering are applied. Food images are cropped and resized to 448*448. The batch size is 16 and the learning rate is 0.001. In the WSSS training process, the same data augmentation method is adopted and the learning rate is 0.005 and weights decay rate is 0.0005. The networks are all trained on a 3090 GPU.

B. Evaluation metrics

The mean intersection over union (mIoU) is utilised as the main evaluation metrics for semantic segmentation. Besides, F1-score, Precision and Recall are performed as the classification evaluation metrics.

$$\mathrm{mIoU} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ji}-p_{ii}}, \quad (4)$$

where, p_{ii} represents the number of true positive pixels of i category; p_{ij} represents the sum of true positive and false negative pixels; p_{ji} represents the sum of true positive and false positive pixels; and k represents the number of categories.

C. Experimental Results

Tab. I shows the performance of the proposed methods in different evaluation metrics. These methods include the SEAM method [19], the proposed SSDB-I and SSDB-II. The first four columns represent the CAM and final results of the three methods under different values for **BG** equalling to 0.1, 0.15, 0.2 and 0.25 respectively. The CAM of the SEAM method represents the pseudo segmentation ground-truth obtained by the first step network. The CAM in SSDB-I and SSDB-II refer to the segmentation results in the case of the classification label of the ground truth. This result implies that the best results can be achieved by these two frameworks when the output of the classification head is perfectly accurate. The final results represent the real result of these methods. The final result of the SSDB-II is 4.56 % more than the SEAM under **BG** = 0.15, and 2.49% more than the SSDB-I. Under other **BG** conditions, the results of SSDB-II are also higher than the other two methods. To implement food WSSS on FoodSeg103, it would be better to choose two branches to output classification and segmentation results respectively. In addition, the classification results of SSDB-I and SSDB-II are validated by F1-score, Precision and Recall. The performance of SSDB-II classification head is 0.5% and 0.3% higher than the SSDB-I in Recall and F1-score. This further implies that independent branches, especially if the backbone networks are all separated, would be better.

TABLE I
THE SEMANTIC FOOD SEGMENTATION RESULT AND FOOD CLASSSIFICATION PERFORMANCE

Model name		Semantic Segmentation Results (mIoU)				Classification results		
		BG=0.1	BG=0.15	BG=0.2	BG=0.25	Precision	Recall	F1-score
SEAM [19]	CAM	27.68	26.98	25.87	23.92	-	-	-
	Fina result	11.49	10.75	10.12	9.06			
SSDB-I	CAM	29.62	26.84	23.74	20.92	72.28	47.85	54.42
	Fina result	13.07	12.04	11.02	10.05			
SSDB-II	CAM	27.8	27.68	26.19	24.29	71.47	48.35	54.72
	Fina result	**14.79**	**14.53**	**13.77**	**12.75**			

The SSDB-II outperforms the other two structures mainly due to two reasons. Firstly, the SSDB-II can effectively isolate the influence of noise in the data on network training. The first step of the SEAM is able to produce better pseudo-segmentation ground-truth compared to the SSDB-I and SSDB-II (according to the first, third and fifth rows in Tab. I). However, the SEAM method has the worst final results. We believe that training on noisy pseudo-segmented ground-truth has a large impact on network performance. Differently, the SSDB-I and SSDB-II do not have this problem, because both networks are trained on manually labeled image-level ground-truth, which accurately expresses the food categories in the sample images. Although the image-level data cannot provide the shape and location information of the objects, it also avoids introducing a large amount of noise in the network training process and guarantees the performance of the network.

Secondly, the SSDB-II alleviates the contradiction between the semantic segmentation task and the classification task under the same training information. In the fully supervised setting, the pixel-wise masks may force the semantic segmentation network to retain more detailed and correct information of objects. However, the network is not provided with explicit supervision information in the weakly supervised setting. This may cause the network to incorrectly activate pixels, and affect both branches. The SSDB-II approach alleviates this problem due to separate semantic segmentation task network and classification task network, which extract features according to

TABLE II
THE IOUS OF TOP 10 CATEGORIES AND MIOU FOR THE OVERALL DATSET. THE BEST VALUES ARE SHOWN IN BOLD FACE FONT

Class name	IoU		
	SEAM [19]	SSDB-I	SSDB-II
Background	**68.32**	64.77	64.5
Corn	50.09	**58.42**	51.81
Green beans	48.61	**52.79**	51.2
Broccoli	48.3	**58.67**	50.94
Seaweed	1.63	0	**44.59**
noodles	30.84	39.39	**44.37**
Strawberry	36.79	**44.55**	43.98
Rice	28.81	**43.71**	43.4
Asparagus	28.27	37.35	**40.9**
French beans	35.37	**38.88**	37.86
mIoU of Top 10 classes	37.7	43.85	**47.36**
mIoU of all classes	11.49	13.07	**14.79**

their own tasks.

Since FoodSeg103 dataset has 104 classes (including background), it is difficult to show intersection over union (IoU) of each class due to lack of space. We show IoUs for top 10 of food categories and the overall averages for all 3 methods with **BG** equal to 0.1 in Tab. II. The IoUs of some categories are satisfactory. However, Tab. II also shows some categories, such as, the Seaweed category, are not recognized during the segmentation process. The IoU of Seaweed category based on SEAM and SSDB-I are 1.63% and 0%, respectively. Differently, its IoU is 44.59% under SSDB-II method. This may be because the features of the Seaweed category are similar to other categories, and Seaweed can be correctly classified during SSDB-II training, but the SSDB-I and SEAM methods classify Seaweed class to another. In addition, some visualisations of semantic food segmentation results are shown in Fig. 3.

We verified the effect of the attention module in food WSSS and the results are shown in Tab. III. Overall, the attention module has a positive impact on the network, whether in sharing-backbone (SSDB-I) or the separating-backbone (SSDB-II). However, attention-based networks have not achieved the best results. Since the attention module focuses on only one category of images and suppresses the simultaneous appearance of multiple categories due to the maximisation operations in the channel attention and spatial attention modules. We believe this problem can be improved by adding multiple attention modules. Compared to the impact of attention modules, sharing the same network for two tasks has a worse impact on achieving food WSSS. This further confirms the previous analysis.

Moreover, We found that pre-trained weights have a large impact on food WSSS. Tab. IV shows the segmentation results of the network after retraining on Food101 dataset and the pre-trained only model. The retrained model performs significantly better than the model without retrained, it is possible that the representations learned by retraining may be more suitable for FoodSeg103 than the general representations learned from ImageNet [29]. In addition, the variability of food features may further increase the difficulty of network training, when labels do not provide enough information. Therefore, we suggest that retraining is a good training strategy when implementing the food WSSS task.

TABLE III
SEGMENTATION RESULTS OF SSDB-I AND SSDB-II NETWORKS WITH AND WITHOUT ATTENTION MODULE, WHEN **BG**=0.1

Model name		mIoU
SSDB-I	w/ attention	13.07
	w/o attention	11.39
SSDB-II	w/ attention	12.7
	w/o attention	14.79

TABLE IV
NETWORK SEGMENTATION RESULTS AFTER NON RETRAINING AND RETRAINING

Framework name		mIoU
Non-Retraining	SEAM	2.24
	SSDB-I	4.14
	SSDB-II	4.66
Retraining	SEAM	10.75
	SSDB-I	12.04
	SSDB-II	**14.53**

V. CONCLUSIONS

In this paper, we have proposed a novel approach for WSSS in food images with image-level annotations to obtain pixel-wise results. Considering the characteristics of the food domain, we have proposed SSDB-Net to better realise food WSSS. We have designed two versions of SSDB-Net: one with shared- backbone (SSDB-I) and the other with separated-backbone (SSDB-II). Our two networks resulted in mIoU of 13.07% and 14.79%, respectively for 104 categories, compared to 11.49% of the state-of-the-art SEAM. Based on the experimental results, we find that weakly supervised semantic food segmentation can be best accomplished by implementing segmentation tasks and classification tasks in branches that do not share weights. Moreover, we have verified the benefit of retraining the network on Food101 to achieve better weakly supervised segmentation on FoodSeg103.

ACKNOWLEDGMENT

We are grateful to the Singapore Management University for providing the FoodSeg103 database.

REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in Proc. of the European Conf. on Computer Vision (ECCV), 2018, pp. 801–818.

[2] L. Deng, J. Chen, Q. Sun, X. He, S. Tang, Z. Ming, Y. Zhang, and T. S. Chua, "Mixed-dish recognition with contextual relation networks," in Proc. of the 27th ACM Int'l Conf. on multimedia, 2019, pp. 112–120.

[3] G. Ciocca, P. Napoletano, and R. Schettini, "Learning CNN-based features for retrieval of food images," in New Trends in Image Analysis and Processing–ICIAP 2017: ICIAP International Workshops, WBICV, SSPandBE, 3AS, RGBD, NIVAR, IWBAAS, and MADiMa 2017, Catania, Italy, September 11-15, 2017, Revised Selected Papers 19. Springer, 2017, pp. 426

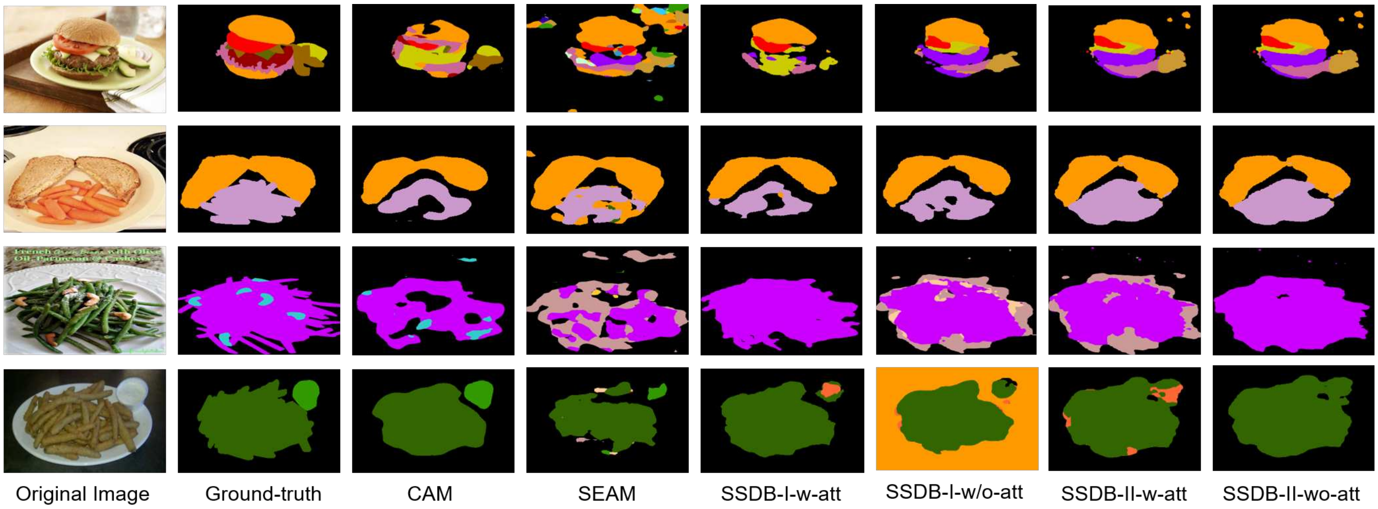

Fig. 3. The visualisation of weakly supervised semantic food segmentation results. All results are obtained under **BG**=0.1. The CAM is obtained by SSDB-II without attention to network structure. SSDB-1-w-att means the results come from SSDB-I network with attention module, and SSDB-II-w/o-att means the results come from SSDB-II without attention module.

[4] V. C. Burkapalli and P. C. Patil, "Food image segmentation using edge adaptive based deepcnns," International Journal of Intelligent Unmanned Systems, vol. 8, no. 4, pp. 243–252, 2020.

[5] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, "Im2calories: towards an automated mobile vision food diary," in Proc. of the IEEE international Conf. on Computer Vision, 2015, pp. 1233–1241.

[6] S. Hamdollahi Oskouei and M. Hashemzadeh, "Foodrecnet: a comprehensively personalized food recommender system using deep neural networks," Knowledge and Information Systems, pp. 1–23, 2023.

[7] S. Aslan, G. Ciocca, and R. Schettini, "Semantic food segmentation for automatic dietary monitoring," in 2018 IEEE 8th International Conf. on Consumer Electronics-Berlin (ICCE-Berlin). IEEE, 2018, pp. 1–6

[8] D. Sahoo, W. Hao, S. Ke, W. Xiongwei, H. Le, P. Achananuparp, E.-P. Lim, and S. C. Hoi, "FoodAI: Food image recognition via deep learning for smart food logging," in Proc. of the 25th ACM SIGKDD International Conf. on Knowledge Discovery & Data Mining, 2019, pp. 2260–2268.

[9] X. Wu, X. Fu, Y. Liu, E.-P. Lim, S. C. Hoi, and Q. Sun, "A large-scale benchmark for food image segmentation," in Proc. of the 29th ACM International Conf. on Multimedia, 2021, pp. 506–515.

[10] S. Aslan, G. Ciocca, D. Mazzini, and R. Schettini, "Benchmarking algorithms for food localization and semantic segmentation," International Journal of Machine Learning and Cybernetics, vol. 11, no. 12, pp. 2827–2847, 2020.

[11] Q. Li, A. Arnab, and P. H. Torr, "Weakly-and semi-supervised panoptic segmentation," in Proc. of the European Conf. on Computer Vision (ECCV), 2018, pp. 102–118.

[12] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," arXiv preprint arXiv:2304.02643, 2023.

[13] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in Proc. of the IEEE conf. on Computer Vision and Pattern Recognition, 2017, pp. 1568–1576.

[14] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2018, pp. 4981–4990.

[15] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101–mining discriminative components with random forests," in Computer Vision–ECCV 2014: 13th European Conf., Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. Springer, 2014, pp. 446–461.

[16] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," arXiv preprint arXiv:1706.05587, 2017.

[17] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, 2016, pp. 2921–2929.

[18] L. Ru, B. Du, Y. Zhan, and C. Wu, "Weakly-supervised semantic segmentation with visual words learning and hybrid pooling," International Journal of Computer Vision, vol. 130, no. 4, pp. 1127–1144, 2022.

[19] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in Proc. of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition, 2020, pp. 12 275–12 284

[20] W. Shimoda and K. Yanai, "CNN-based food image segmentation without pixel-wise annotation," in New Trends in Image Analysis and Processing–ICIAP 2015 Workshops: ICIAP 2015 International Workshops, BioFor, CTMR, RHEUMA, ISCA, MADiMa, SBMI, and QoEM, Genoa, Italy, September 7-8, 2015, Proc. 18. Springer, 2015, pp. 449–457.

[21] J. Dehais, M. Anthimopoulos, and S. Mougiakakou, "Food image segmentation for dietary assessment," in Proc. of the 2nd international workshop on multimedia assisted dietary management, 2016, pp. 23–28.

[22] E. Aguilar, B. Remeseiro, M. Bola˜nos, and P. Radeva, "Grab, pay, and eat: Semantic food detection for smart restaurants," IEEE Transactions on Multimedia, vol. 20, no. 12, pp. 3266–3275, 2018.

[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," arXiv preprint arXiv:1412.7062, 2014.

[24] E. Aguilar, B. Nagarajan, B. Remeseiro, and P. Radeva, "Bayesian deep learning for semantic segmentation of food images," Computers and Electrical Engineering, vol. 103, p. 108380, 2022.

[25] V. B. Raju, M. H. Imtiaz, and E. Sazonov, "Food image segmentation using multi-modal imaging sensors with color and thermal data," Sensors, vol. 23, no. 2, p. 560, 2023.

[26] Y. Wang, F. Zhu, C. J. Boushey, and E. J. Delp, "Weakly supervised food image segmentation using class activation maps," in 2017 IEEE International Conf. on Image Proc. (ICIP). IEEE, 2017, pp. 1277–1281.

[27] W. Shimoda and K. Yanai, "Weakly-supervised plate and food region segmentation," in 2020 IEEE International Conf. on Multimedia and Expo (ICME). IEEE, 2020, pp. 1–6.

[28] Z. Wu, C. Shen, and A. Van Den Hengel, "Wider or deeper: Revisiting the resnet model for visual recognition," Pattern Recognition, vol. 90, pp. 119–133, 2019.

[29] B. Zoph, G. Ghiasi, T.-Y. Lin, Y. Cui, H. Liu, E. D. Cubuk, and Q. Le, "Rethinking pre-training and self-training," Advances in Neural Information proc. Systems, vol. 33, pp. 3833–3845, 2020.