WEAKLY-SUPERVISED PLATE AND FOOD REGION SEGMENTATION

Wataru Shimoda and Keiji Yanai

The University of Electro Communications, Tokyo, Japan

ABSTRACT

In this paper, we propose a novel method to infer plate regions of food images without any pixel-wise annotation. We synthesize plate segmentation masks using difference of visualization in food image classifiers. To be concrete, we use two types of classifiers: a food category classifier and a food/non-food classifier. Using the Class Activation Mapping (CAM) which is one of the basic visualization techniques of CNNs, a food category classifier can highlight food regions containing no plate regions, while a food/non-food category classifier can highlight food regions including plate regions. By taking advantage of the difference between the food regions estimated by visualization of two kinds of the classifiers, in this paper, we demonstrate that we can estimate plate regions without any pixel-wise annotation, and we proposed the approach for boosting the accuracy of weakly-supervised food segmentation using the plate segmentation. In experiments, we show the effectiveness of the proposed approach by evaluating and comparing the accuracy of the weakly-supervised segmentation. The proposed approaches certainly improved an image-level weakly-supervised segmentation method in the food domain and outperformed a well-known bounding box-level weakly-supervised segmentation method.

Index Terms— food image, semantic segmentation, weakly-supervised semantic segmentation, plate region

1. INTRODUCTION

Semantic image segmentation using deep learning has been actively studied especially in the era of deep learning, and its performance has been improved greatly. However, we need pixel-wise annotation to train semantic segmentation model and it requires large annotation cost. Recently, many weakly-supervised segmentation methods have been proposed to resolve the annotation problems. Weakly-supervised segmentation is a task to achieve semantic segmentation by training models with only image-level labels. Weakly-supervised segmentation can reduce large annotation cost because the annotation cost of image-level labels is cheaper than pixel-wise annotation.

In food image recognition, semantic segmentation is one of the important tasks. We can utilize information of semantic segmentation results for food volume estimation and food calorie estimation. However, there are no large scale food semantic segmentation datasets. On the other hand, there are some large scale image recognition datasets, which have image-level labels such as UEC-FOOD100 [1] and FOOD-101 [2]. In addition, food images are closely linked to people's lives, and a large number of meal images are uploaded to Social Networking Sites (SNS). Most of these SNS images are given information such as texts and tags, which can be considered as weakly-supervised labels. It would be of great benefit if we can train semantic segmentation models by classification datasets and vast web images.

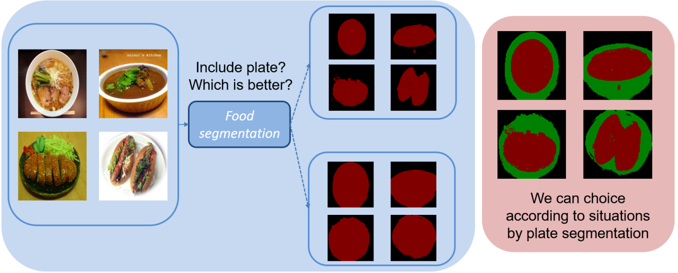

Fig. 1. The motivation and the concept of our proposed approach.

Many weakly-supervised segmentation methods for general objects have been proposed. Though recent weakly-supervised segmentation methods achieve high accuracy on benchmark of weakly-supervised segmentation of general objects, we should consider the difference between general objects and food objects to apply the methods for food images. For example, most of food images include plate regions and it is important that whether or not food segmentation should include plate regions. The solution will vary depending on applications. While, in the case of calorie estimation, it is desirable that the plate regions are excluded from food segmentation, if the aim of food segmentation is inpainting it would be desirable that the plate regions are included in food segmentation. If the plate regions can be inferred, either case can be accommodated. In addition, the information of the plate regions may be beneficial for the refinement of food segmentation. In this paper, we propose a novel method to synthesize plate segmentation masks without any pixel-wise annotation and we utilize the plate segmentation for improvement of the weakly-supervised food segmentation. In Fig.1, we show the motivation and the concept of the proposed approach.

To deduce plate regions without pixel-wise annotation, we train not only a food category classifier but also a food/non-food classifier. In the visualization of the food category classifier, plate regions will not respond because plates are included in most of food images. Therefore, plate regions are not expected to contribute to the recognition of the food category. On the other hand, in the visualization of the food/non-food classifier, plate regions will respond because plates are not included in most of non-food images. Thus, the presence of plate regions is expected to assist recognition by the food/non-food classifier. As we stated, there is a difference in visualization of plate regions between the food category classifier and the food/non-food classifier. We utilize the difference between the visualization of the two classifiers for prediction of plate regions, and synthesize plate segmentation masks. In this paper, we also propose approaches to boost weakly-supervised food segmentation accuracy using the plate segmentation masks. Especially, we make consistency between a food segmentation model and a plate segmentation model in food regions and background. We demonstrate that the proposed approaches can improve a generic weakly-supervised segmentation method in the food domain, and we assess the quality of the plate segmentation by the improvement of the weakly-supervised segmentation method, which utilizes inference of the plate segmentation. To the best of our knowledge and belief, both of the works are the first attempt to extract plate regions from food images without any pixel-wise annotation using visualization techniques, and to boost the accuracy of food segmentation using plate segmentation.

2. RELATED WORK

2.1. Food Recognition

Food image recognition is a promising application of visual object recognition, owing to its potential in estimating food calories and analyzing the eating habits of people for their general well-being. There have been numerous studies on food image recognition that have been published [2, 3, 4, 5, 6, 7, 8, 9].

To estimate the calories associated with the food, the segmentation of food is beneficial. Some studies attempted food region segmentation [5, 10, 4, 11]. Matsuda et al. [5] proposed the use of multiple methods to detect food regions, including Felzenszwalb's deformable part model (DPM) [12], a circle detector, and the JSEG region segmentation method [13]. He et al. [11] employed local variation [14] to segment food regions for estimating the total calories associated with the food in a given food photo. In some studies on mobile food recognition [10, 4], users were asked to point to the rough locations of each food item in an image of food and to perform GrabCut [15] for extracting food item segments.

In addition, there have been several studies on the estimation of calories using computer vision techniques. Kong et al.[16] reconstructed 3D food models using multi-angle pictures and estimated the calories associated with the food using the cubic volume of 3D models. Chen et al.[17] recognized an image and computed the cubic volume using depth information. It must be noted that they obtained depth information using a sensor. 3D base calorie estimation methods tend to be laborious for users. On the other hand, Myers et al.[18] proposed a calorie estimation application called "im2calorie." They obtained each pixel depth information through deep learning prediction and estimated the food calories. However, Myers et al. have not achieved for practical use.

Shimoda et al. [19] proposed weakly-supervised food segmentation and detection methods. They used a back-propagation based visualization method and adapted it to a food domain but they evaluated their methods for only object detection. We also explored efficient weakly-supervised food segmentation method. In this paper, we propose to use a plate segmentation model for boosting weakly-supervised segmentation accuracy. As our best knowledge, this is the first work, which infers plate regions from food images and use it for food image segmentation.

2.2. Weakly-Supervised Segmentation

In the early works of CNN-based weakly-supervised segmentation, visualization based methods have been studied. Since the pixels that contribute to classification have a relationship to regions of target objects, visualization methods can be used as segmentation methods under weakly-supervised setting. Zeiler et al. [20] showed that the derivatives obtained by back-propagation from CNN models trained for classification tasks highlight the region of a target object in an image. Simonyan et al. [21] used the derivatives as the GrabCut seeds and extended a visualization method to a weakly-supervised segmentation method. It is also demonstrated that multi-class objects region also can be captured by difference class specific derivatives [22, 23]. In recent years, Class Activation Mapping (CAM) [24] is widely adopted for generating seed regions for weakly-supervised segmentation methods.

Wei et al. [25] proposed a novel approach to train a fully supervised segmentation model using pixel-level labels obtained by saliency maps under weakly-supervised settings [26]. Ahn et al. [27] proposed a method to learn pixel-level similarity from CRF results, and apply random walk based region refinement, which achieved very high scores on the Pascal VOC 2012 dataset. Shimoda et al. [28] proposed Self-Supervised Difference Detection (SSDD) integrates two segmentation candidates effectively by difference detection. SSDD further improved PSA and achieved the current state-of-the-art on Pascal VOC dataset. In this paper, we use SSDD [28] as a base weakly-supervised segmentation method because of its performance.

3. PLATE SEGMENTATION WITH VISUALIZATION OF FOOD CLASSIFIERS

In this paper, we synthesize plate segmentation masks for learning a plate segmentation model that infers plate regions of food images. To generate plate regions, we use visualization of a food category classifier and a food/non-food classifier. Fig.2 shows the illustration on the idea of the proposed approach.

We assume that $v_L = CAM(x; \theta_L) \in \mathbb{R}^{C \times H \times W}$ is a visualization of the C-class food classifier for input image x generated by Class Activation Mapping (CAM) [24]. In the similar manner, the visualization of the food/non-food classifier is represented by $v_F = CAM(x; \theta_F) \in \mathbb{R}^{2 \times H \times W}$, where θ_L and θ_F are the parameters for the classifiers. Both v_F and v_L should respond to food regions. However, the results of visualization are expected to be different. In particular, while the visualization of the food/non-food classifier returns clear responses in plate regions, the visualization of the food category classifier returns weak responses in plate regions. This is because the plate regions have strong co-occurrence with food images. In this paper, we assume that the difference in v_F and v_L corresponds to plate regions and we synthesize plate segmentation masks by utilizing the difference.

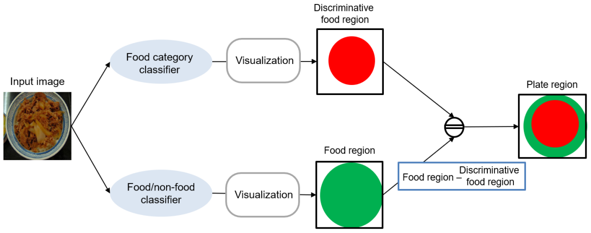

Fig. 2. An illustration of the proposed approach for synthesizing plate segmentation masks using the visualization technique.

Here, we denote the steps in synthesizing plate segmentation masks. First, from v_F, we obtain binary segmentation masks $m_{F,cam}$ whose pixels represent belonging to foods or non-food objects. Secondly, we obtain segmentation masks $m^y_{L,cam}$ for category labels y assigned to images from v_L. If $m_{F,cam}$ and $m^y_{L,cam}$ are able to be extracted correctly, the difference in the masks would be plate regions based on the above assumption. However, the visualization of food category classifier is unreliable because of the difficulty of food classification. Therefore, in this work, in addition to the visualization for the class label, we define unreliable regions obtained from the visualization of the top K classes of the recognition result. In practice, we define unreliable regions $m^{r^K}_{L,cam}$ whose pixels do not overlap with $m^y_{L,cam}$, and just ignore the pixels when training of a plate segmentation model. We set K to 30. We empirically decided this value. We denote the segmentation masks synthesized by the above processing as $m_{P,cam}$. Here, we define a set of pixels for $m_{P,cam}$ as $S_{P,cam}$. This set can be represented by $S_{P,cam} = S^{fg}_{F,cam} - S^{fg}_{L,cam}$, where $S^{fg}_{L,cam}$ is a set of the foreground of the categorical food regions and $S^{fg}_{F,cam}$ is a set of the foreground of the whole food regions. We train the plate segmentation model by the synthesized ternary masks $m_{P,cam}$, which category consists of background, plate regions and food regions. The loss of the plate segmentation model is as follows:

$$\mathcal{L}_{plate} = -\frac{1}{\sum\limits_{k=(0,1,2)} |S^k_{P,cam}|} \sum_{k=(0,1,2)} \sum_{u \in S^k_{P,cam}} \log(h^k_u(x;\theta_P)), \tag{1}$$

where h^k_u is conditional probability of observing any label k at any location u. $S_{P,k}$ is a set of pixels for a class k of the mask $m_{P,cam}$. We apply CRF [29] to the probability map of the plate segmentation model and used the CRF applied results as the final plate segmentation $m_{P,out}$.

4. IMPROVING WEAKLY-SUPERVISED FOOD SEGMENTATION USING PLATE SEGMENTATION

In general, the inside of plate regions are food regions and the outside of plate regions are non-food regions. In this research, we aim to improve the accuracy of weakly-supervised food segmentation by utilizing the relationship between the plate regions and the food regions. To perform weakly-supervised segmentation, we use a method that utilizes Self-Supervised Difference Detection (SSDD) [28]. To improve this further, we propose a new approach which utilizes estimated plate regions. In this section, we describe the details of the approach for making consistency between a food segmentation model and a plate segmentation model. Fig.3 shows an overview of the proposed approach.

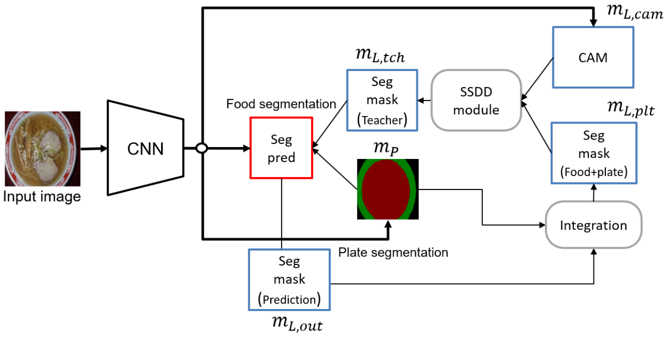

Fig. 3. An overview of the proposed method for refinement of weakly-supervised food segmentation methods.

4.1. Self-Supervised Difference Detection (SSDD) Module

In this paper, we use SSDD [28] as a base weakly-supervised segmentation method that integrates two candidate segmentation masks using difference detection. The proposed method uses a SSDD module, which takes two segmentation masks as inputs and outputs one integrated mask. To be concrete, here, we denote the two segmentation masks as m^K and m^A that has a role of $knowledge$ and $advise$, respectively. The module synthesizes a new segmentation mask m^D by integration of m^K and m^A using inference of difference detection. Difference detection is a task to estimate differences of two segmentation mask. A mask for the difference $M^{K,A} \in \mathbb{R}^{H \times W}$ is defined as following:

$$M^{K,A}_u = \begin{cases} 1 & \text{if } (m^K_u = m^A_u) \\ 0 & \text{if } (m^K_u \neq m^A_u), \end{cases} \tag{2}$$

where $u \in \{1, 2, .., n\}$ indicates a location of pixels, and n is the number of pixels. In the module, we use the Difference Detection network for inference of the difference, $\text{DDnet}(e^h(x;\theta_e), e^l(x;\theta_e), \hat{m}; \theta_d), d \in \mathbb{R}^{H \times W}$, where $\hat{m}$ is a one-hot tensor with the same number of channels to the target class number, θ_d is parameters of DD-Net and $e^h(x;\theta_e)$ is high level features and $e^l(x;\theta_e)$ is low level features extracted from a backbone network such as ResNet. DD-Net takes either of segmentation mask as an input, and outputs the estimation of the difference. We calculate a confidence score $w_u \in \mathbb{R}$ from inferences of the DD-Net d^K and d^A for the masks m^K and m^A:

$$w_u = d^K_u - d^A_u + bias_u, \tag{3}$$

where $bias$ is a hyper parameter for a border of the selection. The refined masks m^D obtained from m^K and m^A are defined by the following expression.

$$m^D_u = \begin{cases} m^A_u & \text{if } (w_u \geq 0) \\ m^K_u & \text{if } (w_u < 0) \end{cases} \tag{4}$$

In this paper, we use mask of CAM $m_{L,cam}$ as $knowledge$ and a synthesized mask using the plate segmentation model

$m_{L,plt}$ as $advice$. From these masks, we generate $m^{L,tch}$ and use it for training of a segmentation model. We describe the detail of $m_{L,plt}$ in the next section.

4.2. Constraining Food Regions by Plate Regions

In standard weakly-supervised food segmentation methods, the food and plate regions may be mixed and it would cause problems in some food-specific applications. In this study, to prevent this we make consistency between the food segmentation model and the plate segmentation model in the food regions. As we stated in Section 4.1, we integrate two segmentation masks $m_{L,cam}$ and $m_{L,plt}$ using the SSDD module. Since the accuracy of the integrated segmentation mask $m_{L,tch}$ depends on the accuracy of the two segmentation masks used for the inputs, the improvement of these inputs would lead better accuracy. Here, we refine the one of the input segmentation mask, which has a role of $advice$. Specifically, we refine the outputs of the food segmentation model $m_{L,out}$ using the outputs of the plate segmentation model $m_{P,out}$, and generate a mask $m_{L,plt}$. To avoid mixing of the food regions and the plate regions we constrain the food regions by below processing:

$$m_{L,plt} = \begin{cases} m_{L,out} & \text{if } (m_{P,out} = food\ class) \\ BG\ class & \text{if } (m_{P,out} = BG\ or\ plate\ class) \end{cases} \tag{5}$$

It is expected that the outputs near by the boundary in food regions and plate regions would be refined by this processing.

4.3. Penalizing Background Prediction Using Plate Segmentation

Since food segmentation is a kind of fine-grained classification, the degree of difficulty is high compared to general object segmentation. Actually, the food segmentation model tends to output background class in regions that are difficult to inference an appropriate category. Therefore, in this section, we limit the outputs of background by making consistency in inference of the food segmentation model and the plate segmentation model. To limit the outputs of background, we constrain the outputs of the food segmentation model on the background class using a penalty loss. The penalty loss minimizes the cross entropy loss for the inverse conditional probability on pixels that belongs inconsistency regions between the food segmentation model and the plate segmentation model. We denote the outputs of the food segmentation model as $h(x;\theta_s)$ and a set of the pixels that are classified as food regions by the plate segmentation model as $S^{food}_{P,out}$. We define penalty loss for the background class as following:

$$\mathcal{L}_{penalty} = -\frac{1}{|S^{food}_{P,out}|} \sum_{u \in S^{food}_{P,out}} \log(\tilde{h}^{bg}_u(x;\theta_{seg})), \tag{6}$$

where $\tilde{h}^{bg}_u(x;\theta_s)$ is conditional probability maps of $background$ class putted negative one on it before softmax function.

4.4. Final Loss for The Food Semantic Segmentation Model

Here, we explain the final loss function for training the food segmentation model. The parameters θ_{seg} of the food segmentation model are trained using the outputs of the SSDD module $m_{L,tch}$ by below equation:

$$\mathcal{L}_{main} = -\frac{1}{\sum_{k\in\hat{y}} |S^k_{L,tch}|} \sum_{k\in\hat{y}} \sum_{u\in S^k_{L,tch}} \log(h^k_u(x;\theta_{seg})). \tag{7}$$

In addition, we also use the loss of $\mathcal{L}_{penalty}$ we stated in Section 4.3 for training the segmentation model. The final loss of the segmentation model is as following:

$$\mathcal{L}_{seg} = \mathcal{L}_{main} + 0.1\mathcal{L}_{penalty} + \mathcal{L}_{plate}. \tag{8}$$

We empirically decided the coefficient of the $\mathcal{L}_{penalty}$.

5. EXPERIMENTS

In the experiments, we used the UEC-FOOD100 dataset [5]. The UEC-FOOD100 dataset [5] consists of 100 class food categories, and each category includes 100 images. Each food item has bounding box annotation, although they have no annotation for segmentation masks. Then, we add new semantic segmentation mask to 10% of UEC-FOOD100 dataset, and used them for evaluation of weakly-supervised segmentation. In addition, we have collected 8155 non-food images from the Web and Twitter, and we use them for training of the food/non-food classifier. We train the proposed model using only image-level labels. The training data does not include bounding information. For training of the classifier models and food segmentation model, we used the 90% of the UEC-FOOD100 dataset.

We evaluate the accuracy of the weakly-supervised segmentation using mean Intersection over Union (mIoU) and Pixel accuracy (Pix acc). mIoU is a standard measurement for semantic segmentation that evaluates the overlap and the union in inference and ground truth. Pix acc is a more simpler measurement that is the accuracy for the all pixels.

5.1. Implementation Details

As semantic segmentation model we used a ResNet-38 model, which is the same architecture used in [28]. The input image size is 448x448 for training and test images and the output feature map size before upsampling is 56x56. These feature map sizes are adjusted to 112 by 112 using simple linear interpolation. Before training of the segmentation model, we trained the food category classifiers with initialization using a pre-trained model of ImageNet. After training of the food category classifiers, we initialized the parameters of backbone models with the food category classifier. The backbone network of the food segmentation model, plate segmentation model and food classifier models are shared, and we trained them in an end-to-end manner. Note that we also continued training of the classifier models. We set an initial learning rate to 1e-3 (1e-2 for initialization without the pre-trained model) and we decreased learning rate with cosine warm up [30]. The batch size for training is 2. For data

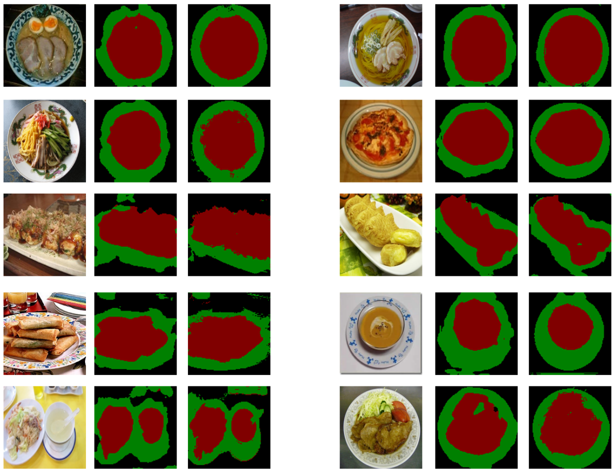

Fig. 4. The examples of the plate segmentation model for the successive results. From left to right, input images, raw plate segmentation masks and CRF applied plate segmentation masks.

augmentation and inference technique, we followed the paper [28]. We implemented the proposed method using PyTorch.

5.2. Qualitative Result of Plate Segmentation and Discussion

In this work, we propose a method to synthesize plate segmentation masks of food images without pixel-wise annotation and we train a plate segmentation model with the synthesized masks. Fig.4 shows some successful examples of plate segmentation. The proposed plate segmentation model excels on inference of plates that have the round shape, but, in several cases, the model can also successfully infer plate regions whose shape is not round such as the case of the middle row in Fig.4. This indicates that the proposed method infers various types of plate regions and the inference does not fall trivial solutions. Fig.5 shows some failure cases. While the proposed plate segmentation model can predict the boundaries between food regions and plate regions, it often fails to capture boundaries between plate regions and background regions. The proposed plate segmentation model also goes wrong on inference for big plates that extend toward outside of the image such as the example of the bottom of the left in Fig.5. We consider that the both of the failure cases are caused by limitations of visualization, that is the whole plate regions do not contribute to the recognition of the food/non-food classifier in these cases. There is also another problem in the plate segmentation model, the plate segmentation model attempts to predict plate regions if there are no plates in images. These problems do not harm the accuracy of weakly-supervised segmentation, however, it would be problems on some other applications. There is still room for improvement in this approach.

5.3. Ablation Study

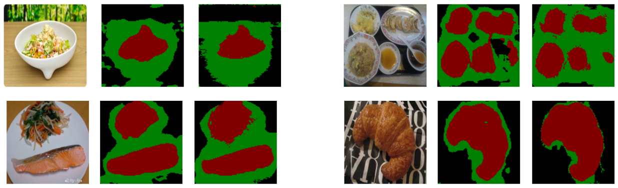

Fig. 5. The examples of the plate segmentation model for the failure cases. From left to right, input images, raw plate segmentation masks and CRF applied plate segmentation masks.

Table 1. Ablation study for the approaches to refine food segmentation by plate segmentation masks.

Method	**Constraining**	**Penalizing**	mIoU	Pix acc
(I)	-	-	49.7	78.3
(II)	✓	-	42.9	75.4
(III)	-	✓	52.6	81.0
(IV)	✓	✓	**55.4**	**82.6**

Here, we study how each of the parts of the proposed approach influences the overall performance. Table 1 shows improvement of the accuracy of weakly-supervised segmentation by the proposed approaches. **Constraining** is the approach proposed in Section 4.2 for reducing overflowed food regions and **Penalizing** is the approach proposed in Section 4.3 for enhancing outputs of food regions on pixels that are often classified as background. The constraint of the food regions using plate segmentation causes large performance dropping because the constraint is too strong and makes the unbalance on inference of the background class though we expected it would be helpful to capture the boundary of the food regions. Penalizing background regions using plate segmentation boosts up the accuracy from 49.7% to 52.6%. This gives evidence that SSDD tend to misclassify on pixels that estimated as background, and plate segmentation can assist to reduce the misclassfication on such pixels. When we incorporate both of the approaches, the constraint of the food regions further leads to the performance boost of 2.8%. These results indicate that the both of approaches help weakly-supervised food segmentation. The balance on the food regions and background regions is important, and plate segmentation is effective on making the balance. We show the qualitative results in Fig.6.

5.4. Comparison with Existing Weakly-Supervised Segmentation Methods

We compare with three existing weakly-supervised segmentation methods. Class Activation Mapping (CAM) [24] is a popular weakly-supervised segmentation method that roughly outputs object location with the ambiguous boundary. SSDD [28] is one of the state-of-the-art method among the current works of weakly-supervised segmentation that greatly improves CAM using CRF and the self-supervised difference detection module. We used SSDD as a base method, and combine the proposed approaches that use plate segmentation. To assess the effectiveness of the proposed approach, we also compare the proposed approach with “Simple Does It” [31]. “Simple Does It” [31] is a well-known bounding box-based weakly-supervised segmentation. While CAM, SSDD and the proposed method are trained with only image-level labels, “Simple Does It” requires bounding box for training, i.e. it uses additional supervision. We compare the proposed method with the most simplest way using GrabCut [15] proposed in the paper [31]. More concretely, the method generates pseudo pixel-level labels from each bounding box by

Table 2. Comparison with existing methods.

Method	mIoU	Pix acc
CAM [24]	30.7	65.1
SSDD (base method) [28]	49.7	78.3
Simple Does It [31]	51.1	81.9
PFSeg (Proposed)	**55.4**	**82.6**

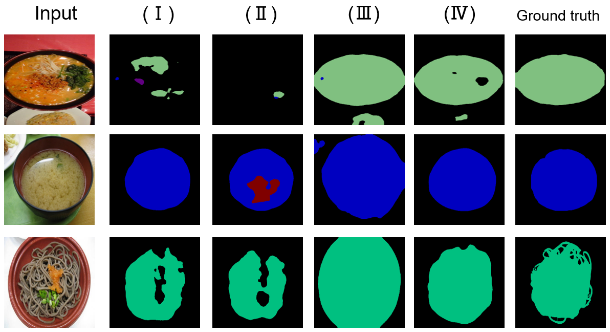

Fig. 6. Examples of the weakly-supervised food segmentation results. (I), (II), (III) and (IV) correspond to Table 1 of the method. From the results, we can observe that the both of the proposed approaches make large effects on the balance of inference for the food regions and background regions, and we can make the good balance by using both of them together.

applying GrabCut [15] and extracting foreground masks. After extracting foreground masks, the method gives the category labels to the foreground masks using the labels of the bounding boxes, then the method trains a segmentation model with the generated segmentation masks. This approach is simple, but a powerful baseline considering the advantage of bounding box information. The performance comparisons are summarized in Table 2. We denoted the proposed method as Plate-based Food Segmentation (PFSeg).

As shown in Table 2, the proposed method achieved 55.4% on mIoU and 82.6% on Pix acc. Compared with the base method, the gain are 5.7 points and 4.3 points on mIoU and Pix acc, respectively. They are also higher than "Simple Does It", which uses bounding boxes as additional training information. These results indicate that the proposed method is efficient and plate segmentation model trained without pixel-wise annotation is beneficial for improving the weakly-supervised food segmentation. Fig.7 (in the supplementary material) shows the examples of the weakly-supervised food segmentation methods.

6. CONCLUSION

In this paper, we proposed a method to synthesize segmentation masks for food plate regions by visualization. We used a food category classifier and a food/non-food classifier for visualization and extracted plate regions from the difference in the visualization of the two types of the classifiers. In addition, we also proposed the approach to make consistency between a food segmentation model and a plate segmentation model, and demonstrated that we boosted the accuracy of weakly-supervised food segmentation using the proposed approach.

Our future works are as follows. First, we would like to improve inference of plate segmentation on the boundaries in the plate regions and the background using other weakly-supervised segmentation techniques, such as affinity model [27]. Second, we also would like to use the plate segmentation model for other applications. We consider that there are many possibilities as the applications, such as food volume estimation, plate separable food image generation, and mask-based food style transfer or domain conversion.

Acknowledgements This work was supported by JSPS KAKENHI Grant Number 17J10261, 15H05915, 17H01745, 17H06100 and 19H04929.

7. REFERENCES

[1] "UEC food 100," http://www.foodcam.mobi/dataset.html.

[2] L. Bossard, M. Guillaumin, and L. V. Gool, "Food-101 - mining discriminative components with random forests," in *ECCV*, 2014.

[3] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," in *ACM MM*, 2014, pp. 1085–1088.

[4] Y. Kawano and K. Yanai, "Foodcam: A real-time food recognition system on a smartphone," *Multimedia Tools and Applications*, pp. 1–25, 2014.

[5] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *ICME*, 2012, pp. 1554–1564.

[6] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, "Combining global and local features for food identification in dietary assessment," in *Proc. of IEEE International Conference on Image Processing*, 2011.

[7] S. Yang, M. Chen, D. Pomerleau, and R. Sukthankar, "Food recognition using statistics of pairwise local features," in *CVPR*, 2010.

[8] S. Jiang, W. Min, L. Liu, and Z. Luo, "Multi-scale multi-view deep feature aggregation for food recognition," in *IEEE Trans. Image Processing*, 2020.

[9] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, "A survey on food computing," in *ACM Computing Surveys*, 2019.

[10] Chamin Morikawa, Haruki Sugiyama, and Kiyoharu Aizawa, "Food region segmentation in meal images using touch points," in *Proc. of ACM MM WS on Multimedia for Cooking and Eating Activities (CEA)*, 2012, pp. 7–12.

[11] Y. He, C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, "Food image analysis: Segmentation, identification and weight estimation," in *ICME*, 2013, pp. 1–6.

[12] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[13] Y. Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. on PAMI*, vol. 23, no. 8, pp. 800–810, 2001.

[14] P. F. Felzenszwalb and D. P. Huttenlocher, "Image segmentation using local variation," in *CVPR*, 1998, pp. 98–104.

[15] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," *ACM Transactions on Graphics (TOG)*, vol. 23, no. 3, pp. 309–314, 2004.

[16] F. Kong and J. Tan, "Dietcam: Automatic dietary assessment with mobile camera phones," in *Proc. of Pervasive and Mobile Computing*, 2012, pp. 147–163.

[17] M. Chen, Y. Yang, C. Ho, S. Wang, S. Liu, E. Chang, C. Yeh, and M. Ouhyoung, "Automatic chinese food identification and quantity estimation," in *SIGGRAPH Asia Technical Briefs*, 2012, p. 29.

[18] 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 *The IEEE International Conference on Computer Vision (ICCV)*, 2015.

[19] W. Shimoda and K. Yanai, "Foodness proposal for webly supervised food detection," in *Proc. of ACM MM Workshop on Multimedia Assisted Dietary Management (MADiMa)*, 2016.

[20] M. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *ECCV*, 2014.

[21] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," in *ICLR WS*, 2014.

[22] Z. Jianming, L. Zhe, B. Jonathan, S. Xiaouhui, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *ECCV*, 2016.

[23] W. Shimoda and K. Yanai, "Distinct class saliency maps for weakly supervised semantic segmentation," in *ECCV*, 2016.

[24] B. Zhou, A. Khosla, A. Lapedriza, and A. Oliva, A. Torralba, "Learning deep features for discriminative localization," in *CVPR*, 2016.

[25] Y. Wei, X. Liang, Y. Chen, X. Shen, M. Cheng, Y. Zhao, and S. Yan, "STC: A simple to complex framework for weakly-supervised semantic segmentation," in *IEEE Trans. on PAMI*, 2017.

[26] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013.

[27] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *CVPR*, 2018.

[28] W. Shimoda and K. Yanai, "Self-supervised difference detection for weakly-supervised semantic segmentation," in *ICCV*, 2019.

[29] P. Krahenbuhl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Advances in Neural Information Processing Systems*, 2011.

[30] L. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," in *ICLR*, 2017.

[31] A. Khoreva, R. Benenson, J. Hosang, M. Hein, and B. Schiele, "Simple does it: Weakly supervised instance and semantic segmentation," in *CVPR*, 2017.