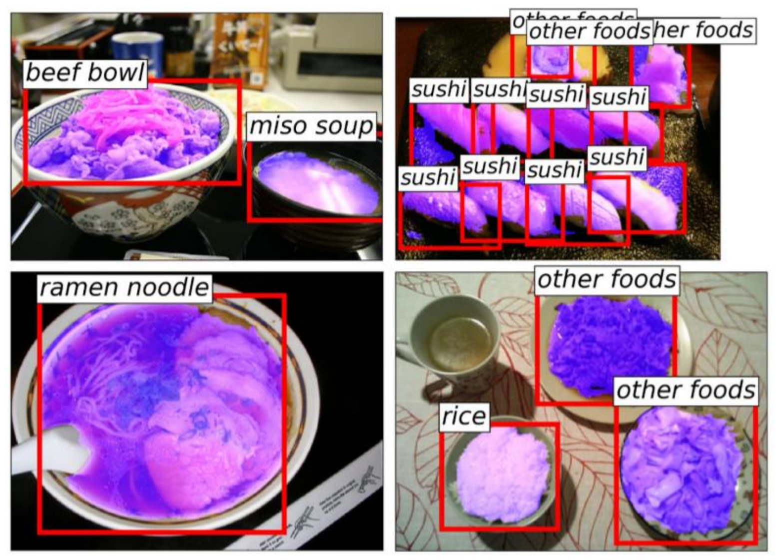
Zero-Annotation Plate Segmentation Using a Food Category Classifier and a Food/Non-Food Classifier

Wataru Shimoda **Keiji Yanai** The University of Electro-Communications, Tokyo, Japan

Background

In food image recognition,
semantic segmentation is one of the important task
There are several applications such as
-food volume estimation
-food calorie estimation



There are no large scale food segmentation datasets.
The foods have many classes and variations
Weakly supervised segmentation is one of the solution for the annotation problem

The plate regions tend to be segmented as food regions
It may cause problems in some applications such as food calorie estimation

Objective

Deduce plate areas
without pixel-wise annotation

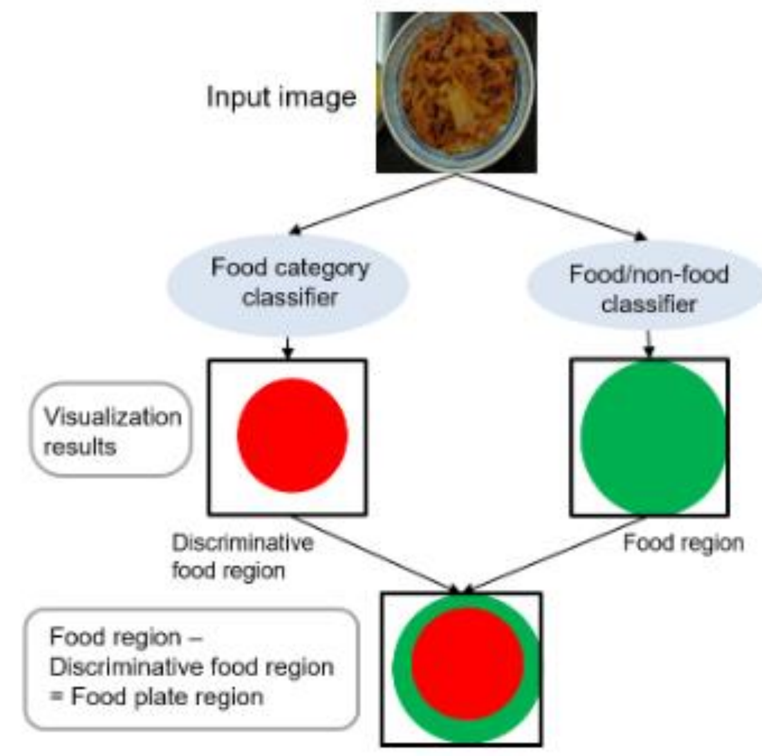
Improve the accuracy of
weakly-supervised food segmentation
using food plate segmentation

Reference

[1] SSDD, Shimoda et al, ICCV 2019
[2] Simple does it, Khoreva et al., CVPR 2017, arxiv:1603.07485

Plate segmentation

Key idea



Highlighted regions by visualization = important regions in classification

Food category classification
The plates does not contribute in food category classification
It is general that food photos include the plates in many food categories.

Food/non-food classification
The plates contribute in food/non-food classification
The photos of the non-food objects usually are not taken with the plates.

Class activation map (CAM)
Food category classification $v_F = CAM(x;\theta_F) \in \mathbb{R}^{2\times H\times W}$
Food/non-food classification $v_L = CAM(x;\theta_L) \in \mathbb{R}^{C\times H\times W}$

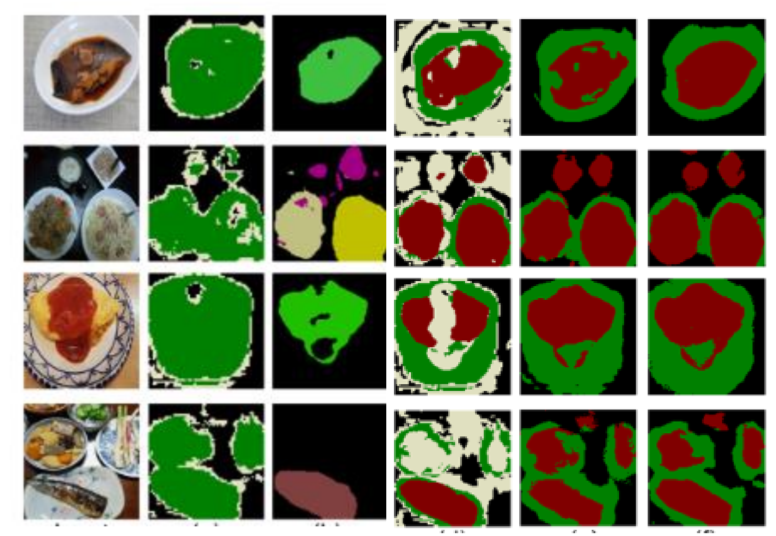
The set of the plate regions: S_P
The set of the whole food regions: S_F^{fg}
The set of the discriminative food regions: S_y^{fg}

$$S_P = S_F^{fg} - S_y^{fg},\ y \in L$$

The loss of the plate segmentation

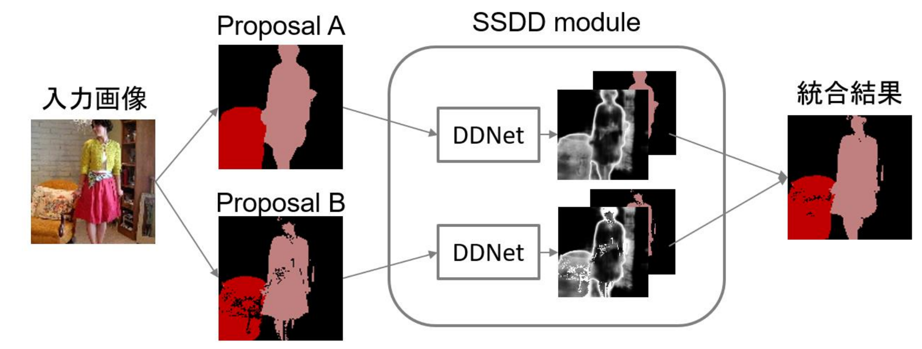
$$\mathcal{L}_{plate} = -\frac{1}{\sum_{k=(0,1,2)} |S_k|} \sum_{k=(0,1,2)} \sum_{u\in S_k} \log(h_u^k(x;\theta_P))$$

$$S_0 = S_F^{bg}, S_1 = \bar{S}_y^{fg} \text{ and } S_2 = S_P$$



Weakly supervised food segmentation

Base method : Self-supervised difference detection[1]

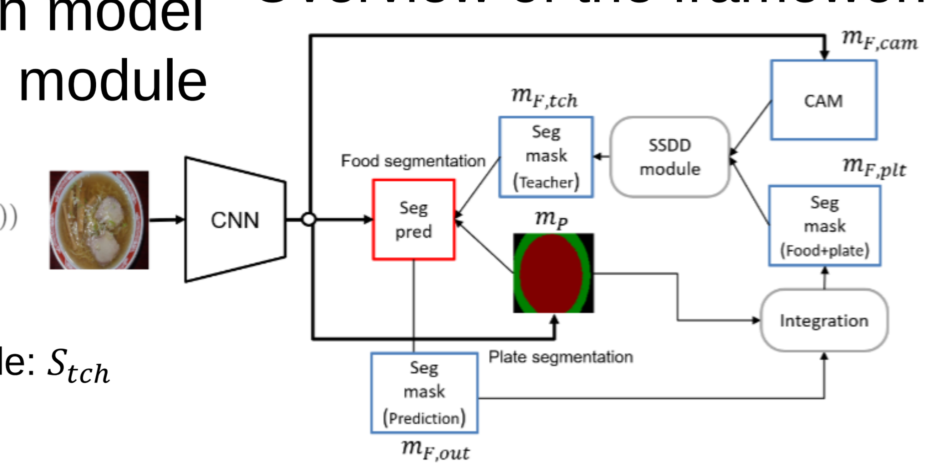


We train the segmentation model with the outputs of SSDD module

$$\mathcal{L}_{main} = -\frac{1}{\sum_{k\in\hat{y}} |S_{F,tch}^k|} \sum_{k\in\hat{y}} \sum_{u\in S_{F,tch}^k} \log(h_u^k(x;\theta_{seg}))$$

The set of the outputs of SSDD module: S_{tch}

Overview of the framework



We improved the base method by:
-(A) Restriction of the foreground by plate regions

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

-(B) Feedback to CAM from the outputs of SSDD module

$$\mathcal{L}_{feedback} = -\frac{1}{|\hat{y}|} \sum_{k\in\hat{y}} \log(p_d^k(x;\theta_{cl})) \quad e_d^k(x;\theta_e) = -\frac{1}{|S_{F,df}^k|} \sum_{u\in S_{F,df}^k} e_h(x;\theta_e)$$

-(C) Penalize the background outputs by the food plate regi

$$\mathcal{L}_{penalty} = -\frac{1}{|S_{P,out}^{food}|} \sum_{u\in S_{P,out}^{food}} \log(-h_u^{bg}(x;\theta_{seg}))$$

The final loss function

$$\mathcal{L}_{final} = \mathcal{L}_{main} + 0.1\mathcal{L}_{feedback} + 0.1\mathcal{L}_{penalty}$$

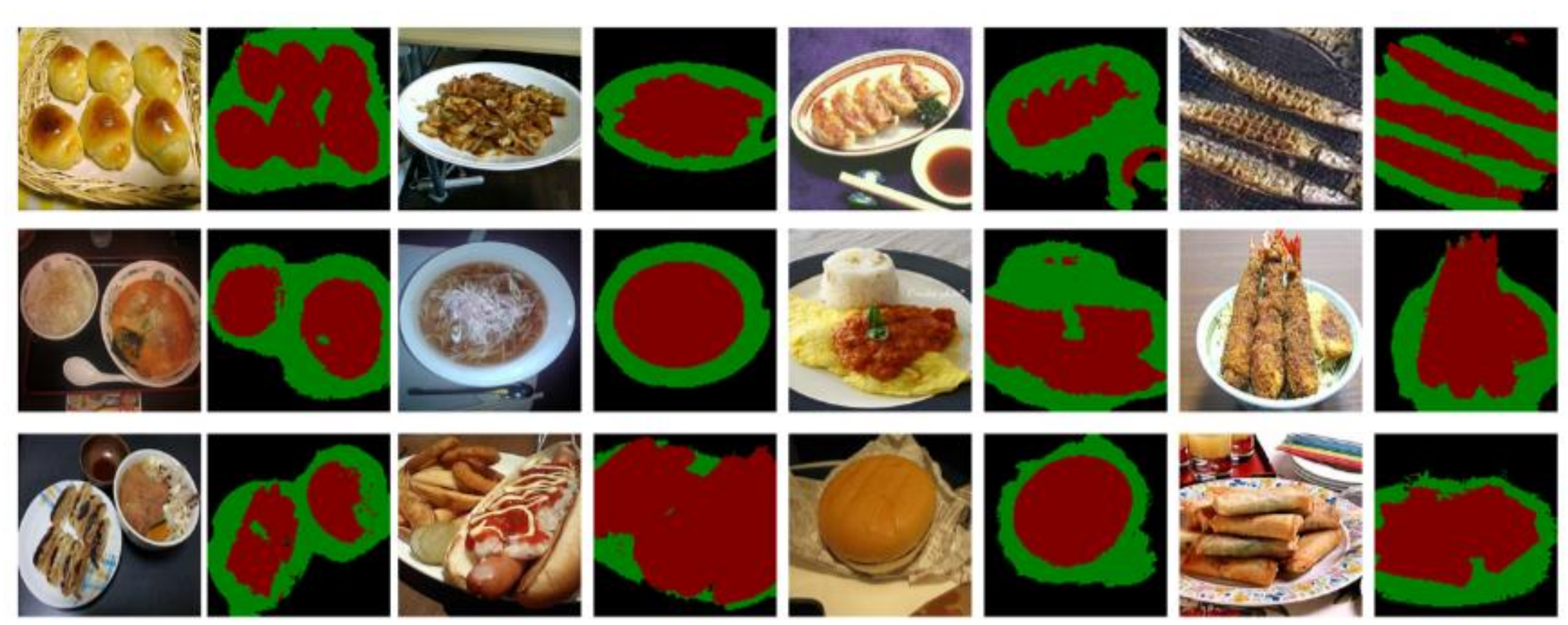
Experiments

Dataset
- UECFOOD101
- 100 classes
- 10000 images

For the evaluations, we annotated pixel-level labels to 1000 images manually
Of course, we used them for only the evaluations

Comparison with other methods
We compared our method with "simple does it" [2]
The compared method use bounding boxes for training
The method has much advantage in the training setting

	mIoU	Pacc
Base method	50.2	77.5
BB annotation + GrabCut [2]	51.1	81.9
proposed	52.3	80.4



Ablation study

	(A)	(B)	(C)	mIoU	Pacc
(I)	-	-	-	50.2	77.5
(II)	-	✔	✔	49.8	78.9
(III)	✔	-	✔	46.0	67.3
(IV)	✔	✔	-	51.2	78.2
(V)	✔	✔	✔	52.3	80.4