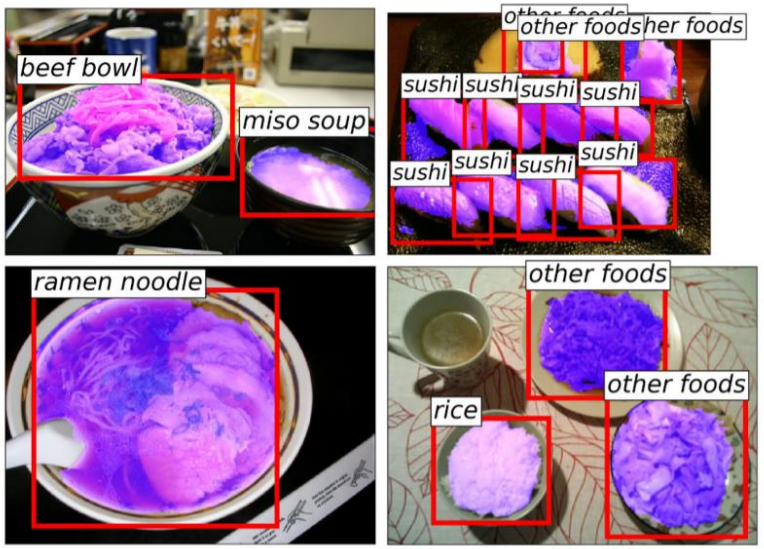
ICME 2020 WEAKLY-SUPERVISED PLATE AND FOOD REGION SEGMENTATION

Wataru Shimoda and Keiji Yanai

The University of Electro Communications, Tokyo, Japan

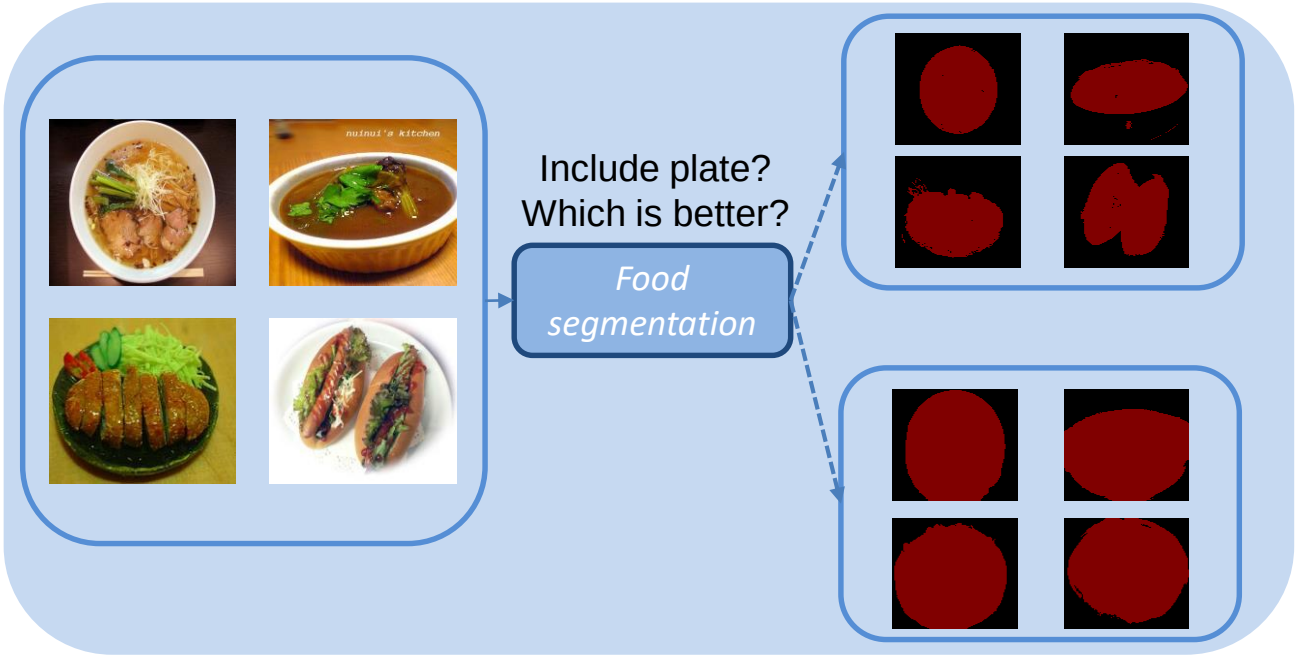
Food segmentation

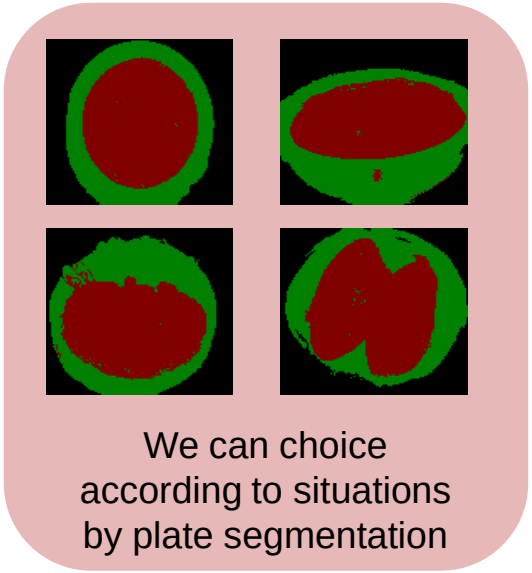
- Various applications
  - Food volume estimation, Food calorie estimation
- Problem
  - Pixel-wise annotation cost
- Approach
  - Weakly-supervised learning
    - Train with image-level labels



Motivation

- A specific problem on food segmentation
  - Should be plate regions in food region?
    - Desirable segmentation is different in case by case





Problem statement

- Food plate segmentation
  - Train plate segmentation model without pixel-wise annotation
  - Use image-level labels for food categories
  - Separate food images into three categories
    - Background
    - Food regions
    - Plate regions

Key idea

S_L^{fg} Food category classifier

Not respond plate regions

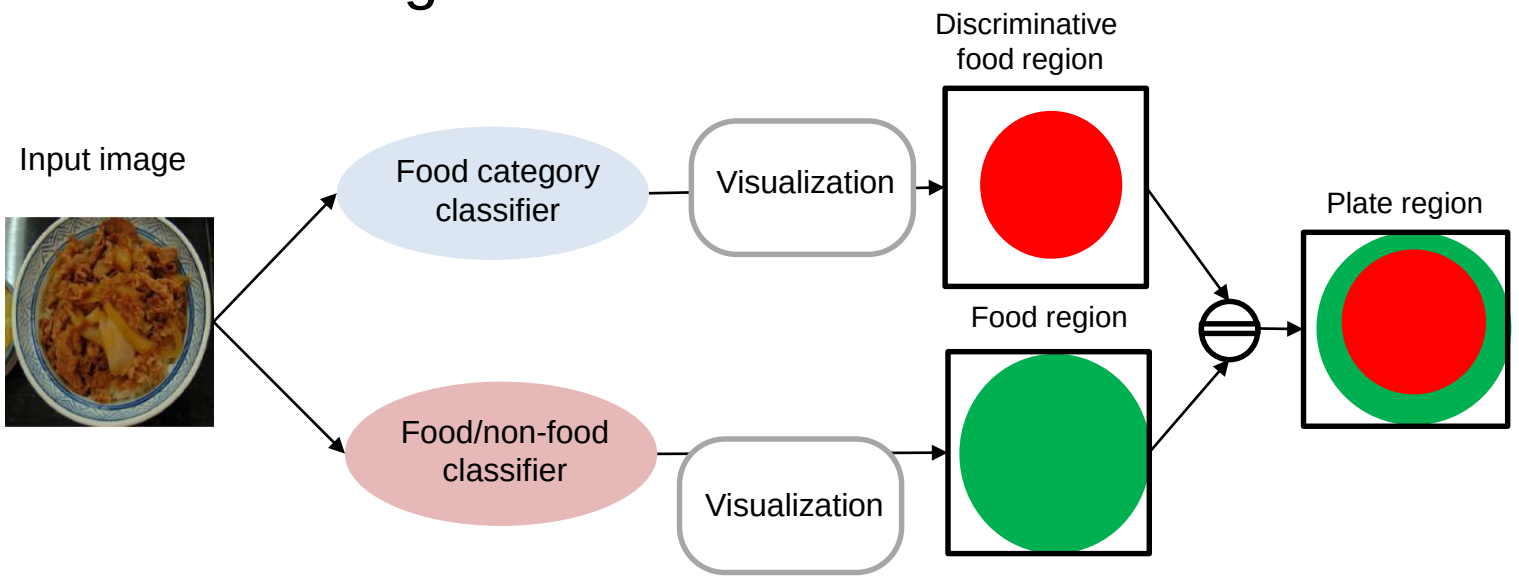
S_F^{fg} Food/Non-Food classifier

Respond plate regions

S_P Difference

Plate regions

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



Approach

- Visualization
  - Class activation map[1]
  - $v_F = CAM(x;\theta_L) \in \mathbb{R}^{2\times H\times W}$, $v_L = CAM(x;\theta_L) \in \mathbb{R}^{C\times H\times W}$
- Mask from visualization
  - $m_{F,cam}$: food/non-food mask, *from* v_F
  - $m^{y}_{L,cam}$: food category mask, *from* v_L *and label* y
    - $m^{r^k}_{L,cam}$: unreliable regions, *upper k class of recognition results*
- Plate mask
  - $m_{P,cam}$: difference of the masks
  - $m_{P,out}$: CRF applied mask

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

S_P:a set of pixels form $m_{P,cam}$

S_F^{fg}:a set of pixels form $m_{F,cam}$

[1] Learning deep features for discriminative localization, Zhou et al., CVPR 2016

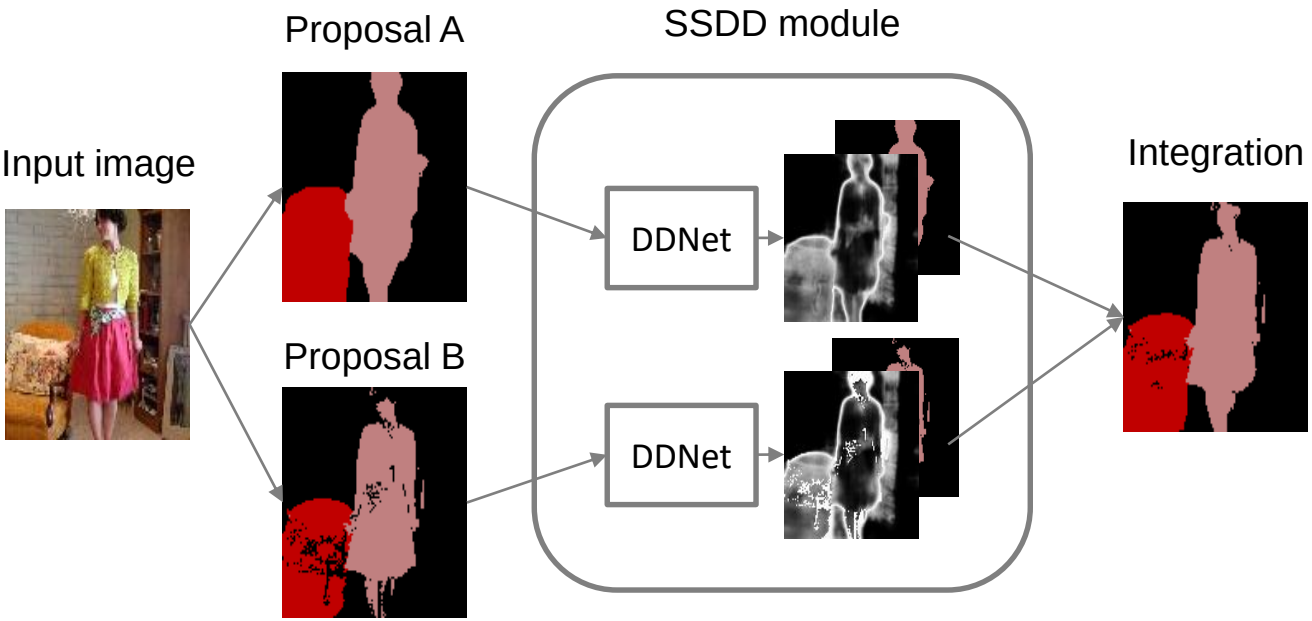
Combination with weakly-supervised food segmentation

- Train plate segmentation model with weakly-supervised segmentation model in end-to-end manner
- Weakly-supervised food segmentation
  - Image-label to segmentation
  - Output setting
    - not include food plate regions
  - Base method
    - [1] Self-Supervised Difference Detection, ICCV 2019

[1] Self-supervised difference detection for weakly-supervised semantic segmentation, Shimoda et al., ICCV 2019

SSDD module

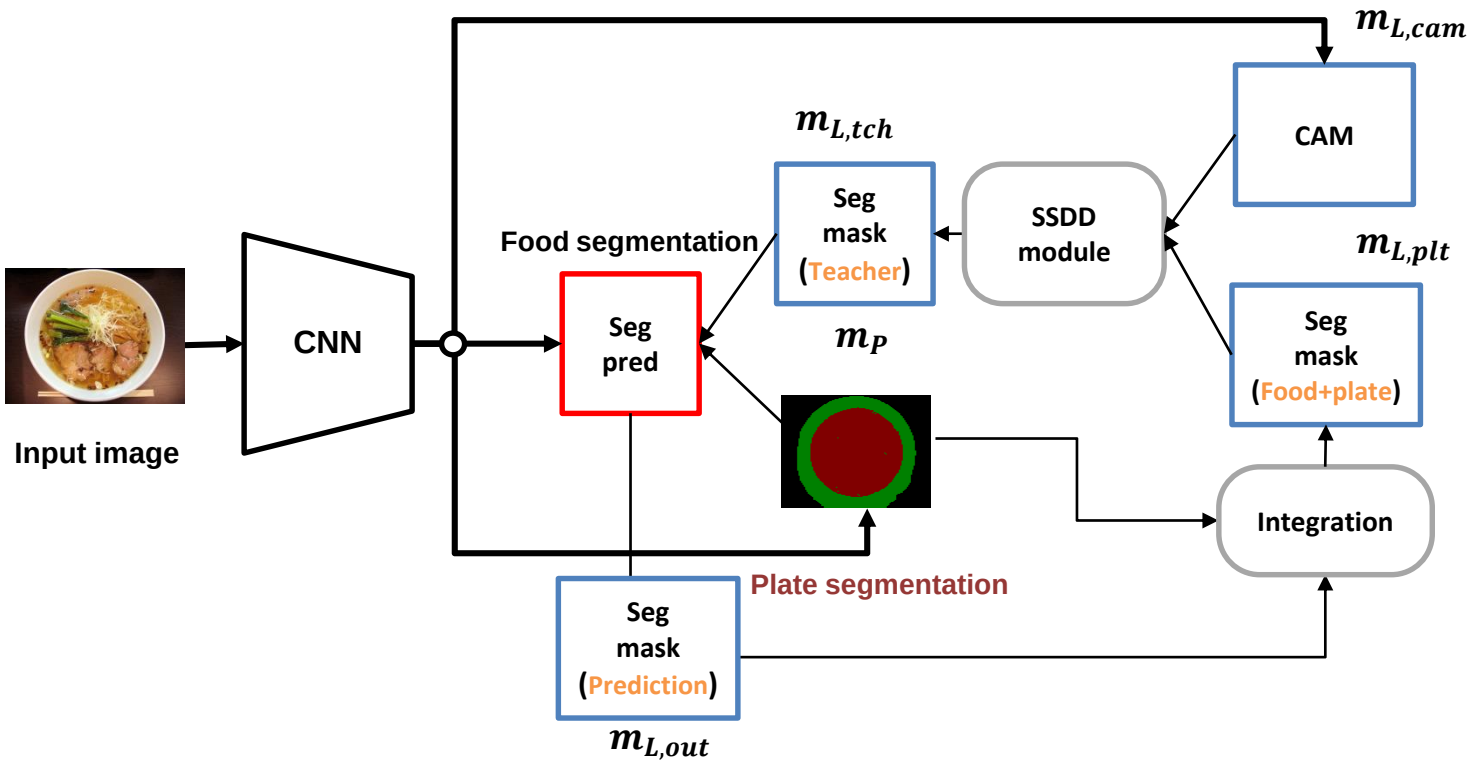
- Integrates two candidate segmentation masks using difference detection for stable refinement
  - [1] Self-Supervised Difference Detection, ICCV 2019



[1] Self-supervised difference detection for weakly-supervised semantic segmentation, Shimoda et al., ICCV 2019

Architecture

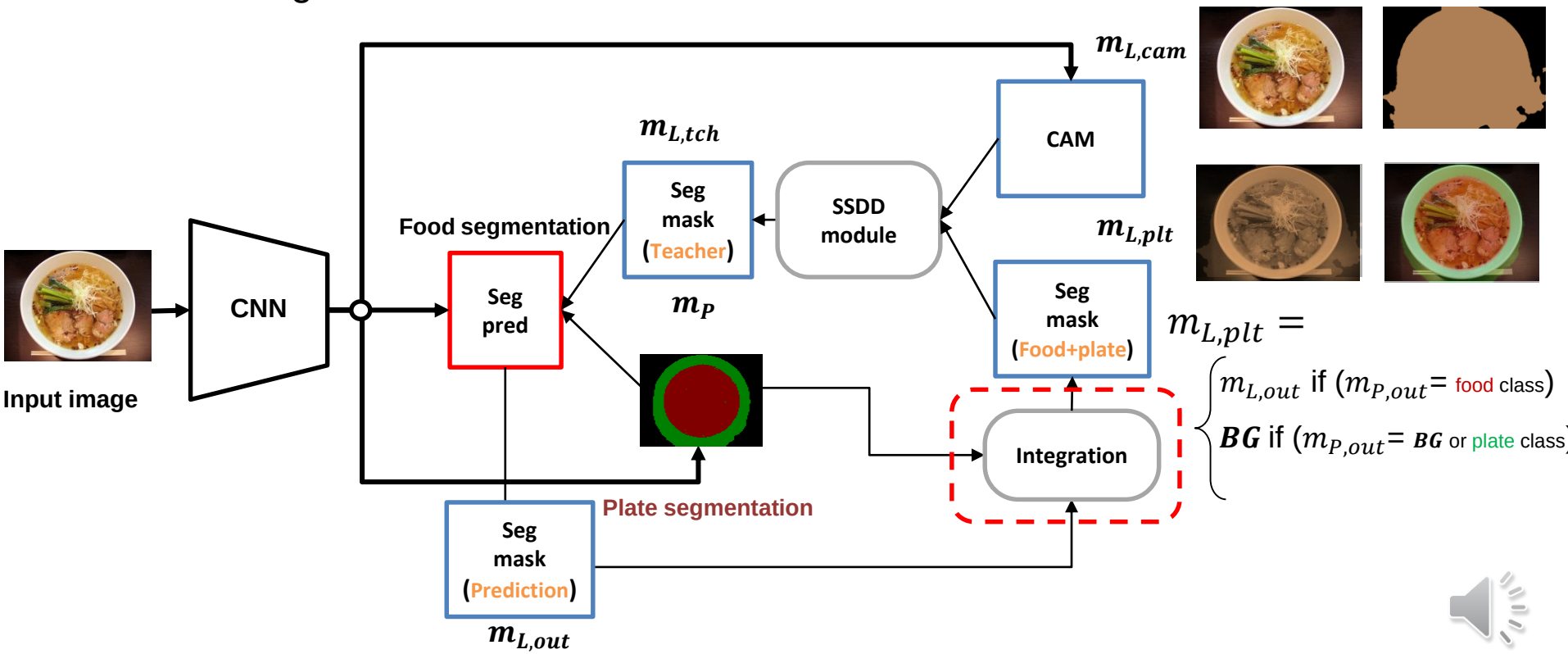
- We use a SSDD module for integration of CAM and food+plate segmentation mask
- We make consistency between the food segmentation model and the plate segmentation model in the food regions with two techniques
  ①Constraining food regions by plate regions
  ②Penalizing background prediction using plate segmentation



Architecture

①Constraining Food Regions by Plate Regions
②Penalizing Background Prediction Using Plate Segmentation

Avoid mixing of the food regions and the plate region by integration based constraining

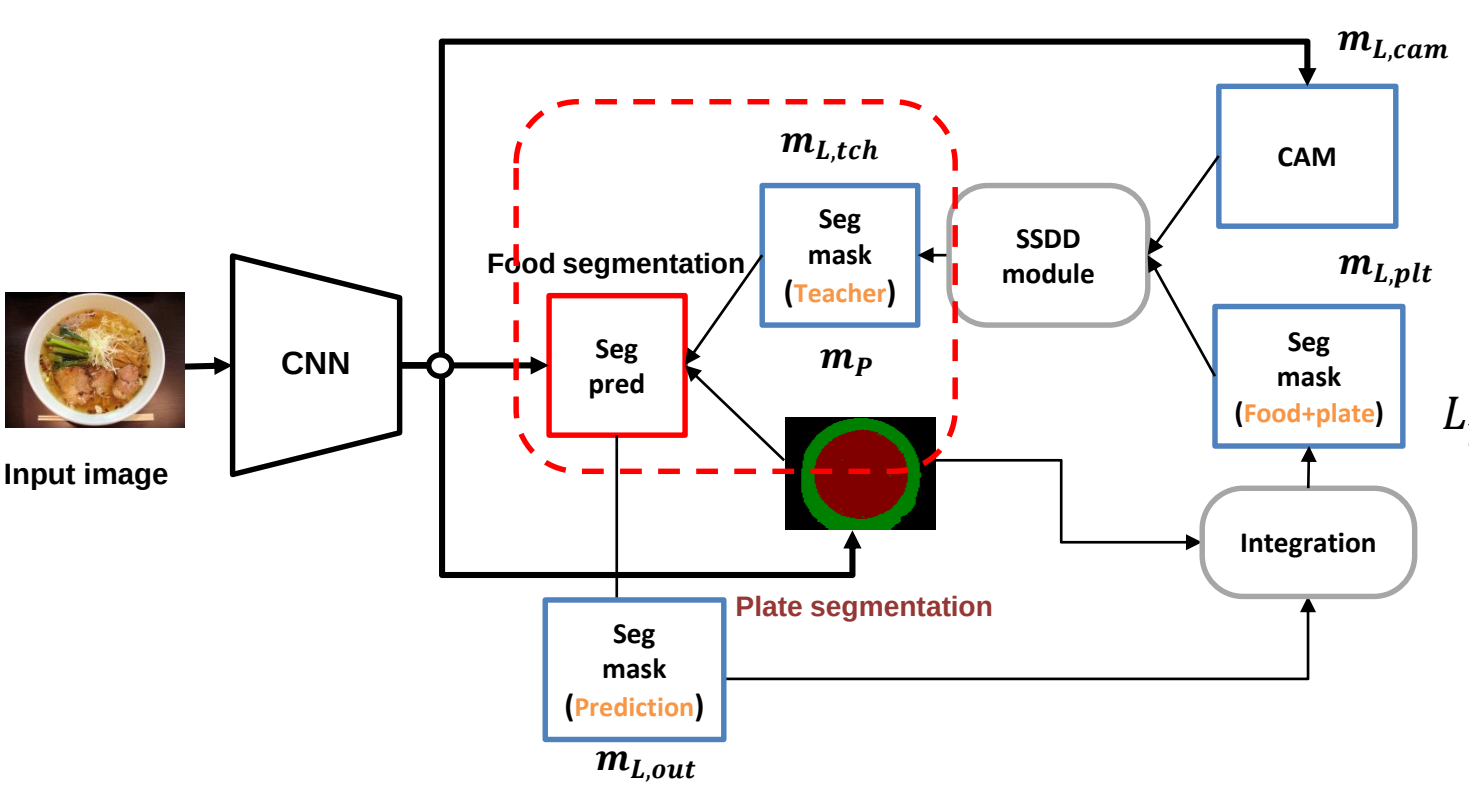


Architecture

①Constraining Food Regions by Plate Regions

②Penalizing Background Prediction Using Plate Segmentation

To limit the outputs of background, we constrain the outputs of the food segmentation model on the background class using a penalty loss



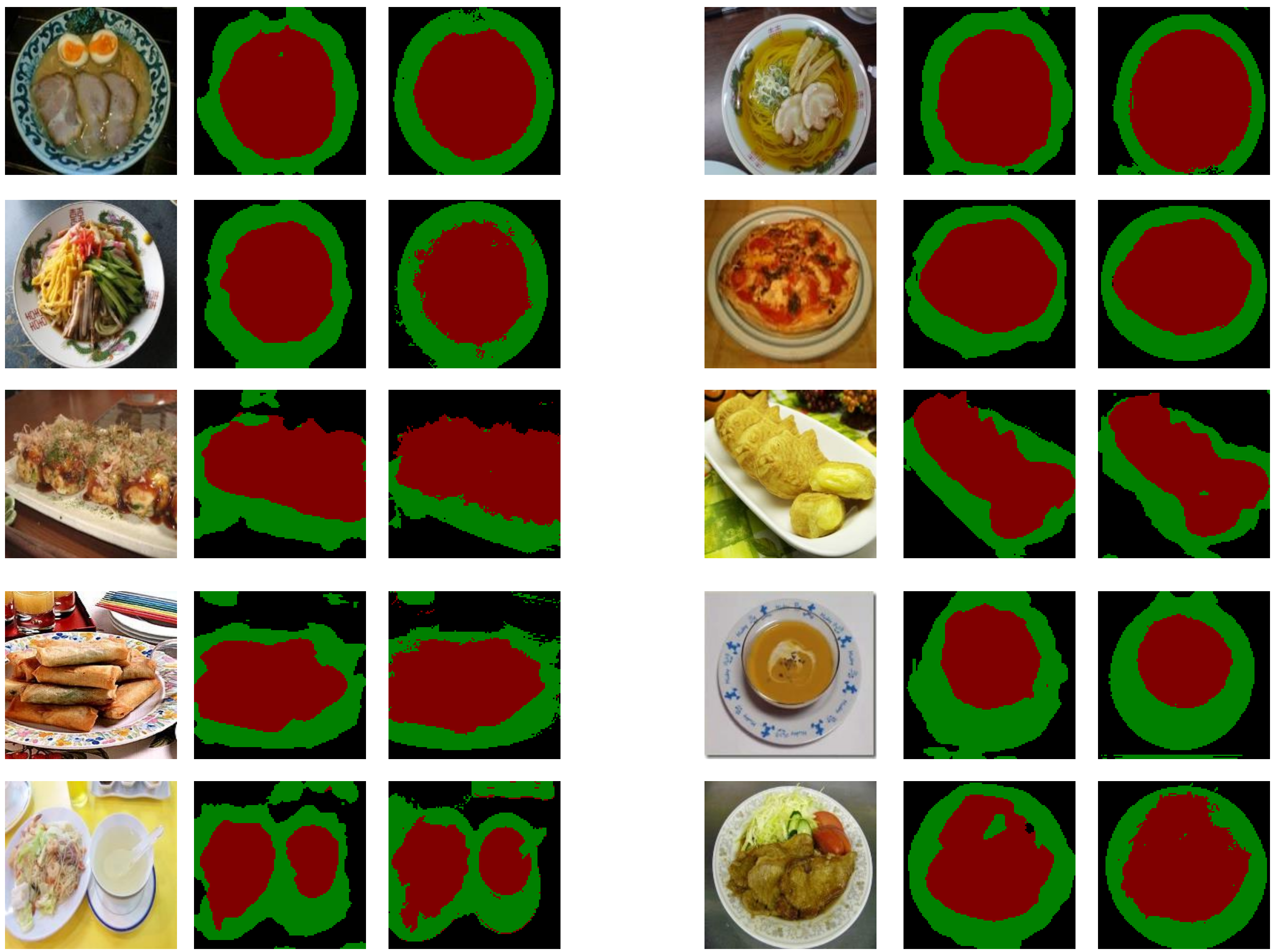
$$L_{penalty} = -\frac{1}{S_{P,out}^{food}} \sum_{u} \log(\tilde{h}_u^{bg})$$

Experiments

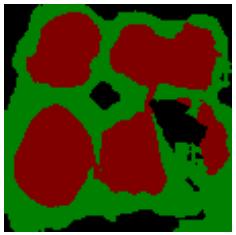
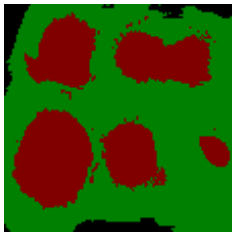
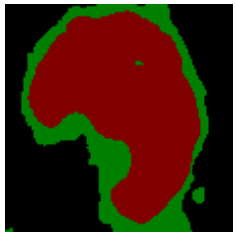
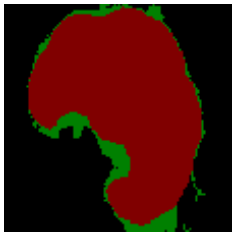
- Dataset
  - UEC FOOD100
    - 100 classes, 10000 images
    - 100 images per class
    - Image-level label and bounding box annotation
    - We annotated pixel-wise annotation to 1000 images for evaluation
  - Train
    - 9000 images with image-level labels
    - 8155 non-food images from Web and Twitter
  - Test
    - 1000 images with pixel-level labels

Plate segmentation results

Input image	Raw plate seg	CRF applied	Input image	Raw plate seg	CRF applied



Failure cases

Input image	Raw plate segmentation	CRF applied segmentation	Input image	Raw plate segmentation	CRF applied segmentation
					
					

Comparison with other weakly-supervised segmentation methods

Quantitative evaluation was performed using weakly supervised food segmentation.
Because we only have pixel-wise ground truth for the food category masks

	mIoU	Pixel Acc
CAM [1]	30.7	65.1
Base method [2]	49.7	78.3
Simple does it [3]†	51.1	81.9
PFSeg(proposed)	55.4	82.6

[1] Learning deep features for discriminative localization, Zhou et al., CVPR 2016

[2] Self-supervised difference detection for weakly-supervised semantic segmentation, Shimoda et al., ICCV 2019

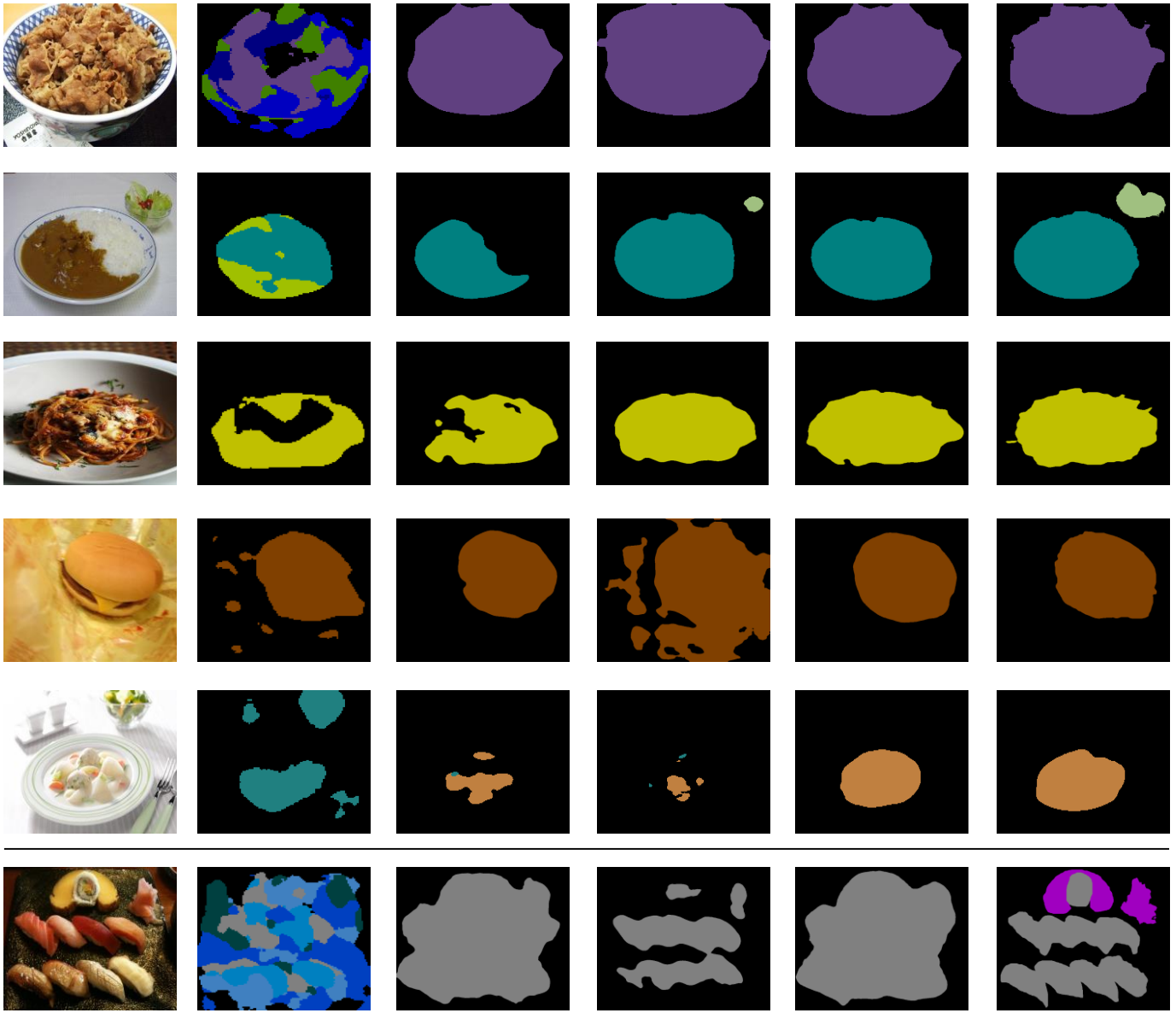
[3] Simple does it: Weakly supervised instance and semantic segmentation, Khoreva et al., CVPR 2017

† use Bounding box annotation and GrabCut

Ablation study

method	Constraining	Penalizing	mIoU	Pixel Acc
(I)	-	-	49.7	78.3
(II)	✔	-	42.9	75.4
(III)	-	✔	52.6	81.0
(IV)	✔	✔	55.4	82.6

Input	CAM	SSDD	Simple Does It	PFseg (Proposed)	Ground truth



Summary

- Predict plate regions without pixel-wise annotations
  - Boost weakly supervised segmentation accuracy using plate segmentation
- Future work
  - Improve inference of plate segmentation on the boundaries in the plate regions and the background
  - Further applications