# CNN-based Food Image Segmentation without Pixel-Wise Annotation

## MADIMA 2015 at Genova, Italy

Wataru Shimoda and Keiji Yanai

The University of Electro-Communications, Tokyo, Japan

# Introduction: Food Recognition

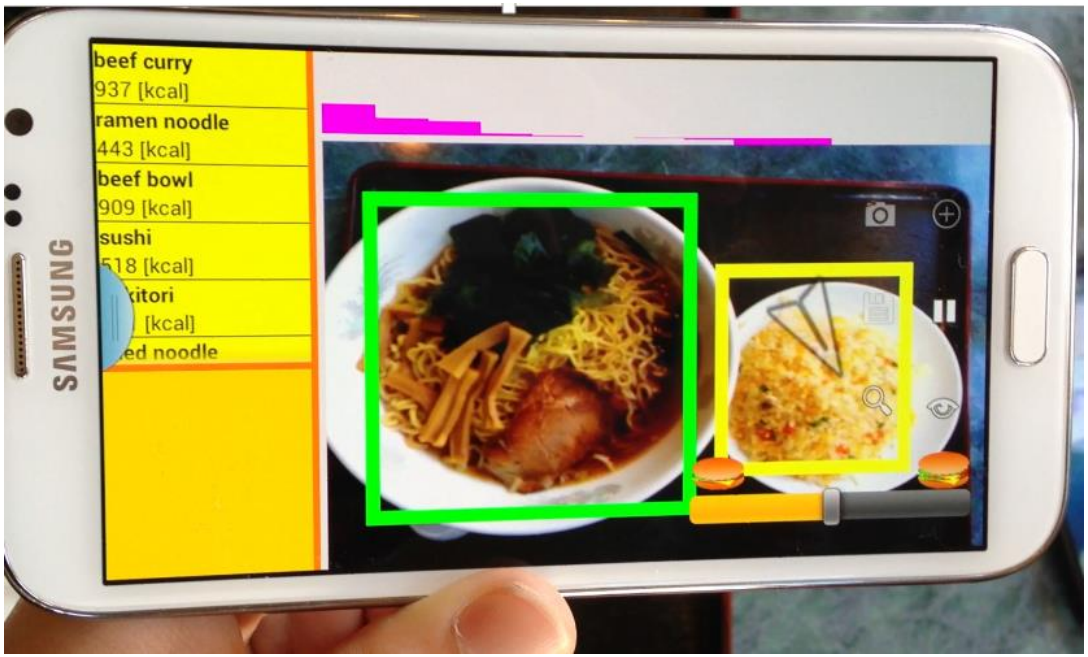
- Food recognition
- Our previous works
  - MKL for food
    [Joutou et al. 2009]
  - UEC FOOD101
    [Matsuda et al. 2012]
  - Food recognition
    on a smartphone
    [Kawano et.al 2013]
    - FOOD CAM


**FOODCAM on Andriod**

# Food segmentation is needed for multiple food items

- Meals sometimes contain multiple food items.
- So far our system needs manual segmentaion.

- In this work, we focus on food segmenta-tion with deep learning method.



# Convolutional Neural Network can be applied for various tasks

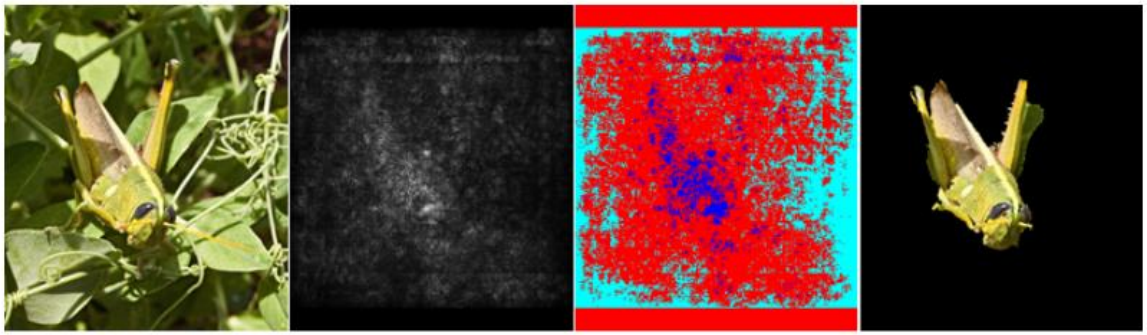
- Convolutional Neural Network (CNN) based method achived the best performance in
  - Object classification


  - Object detection

  - Object Segmentation

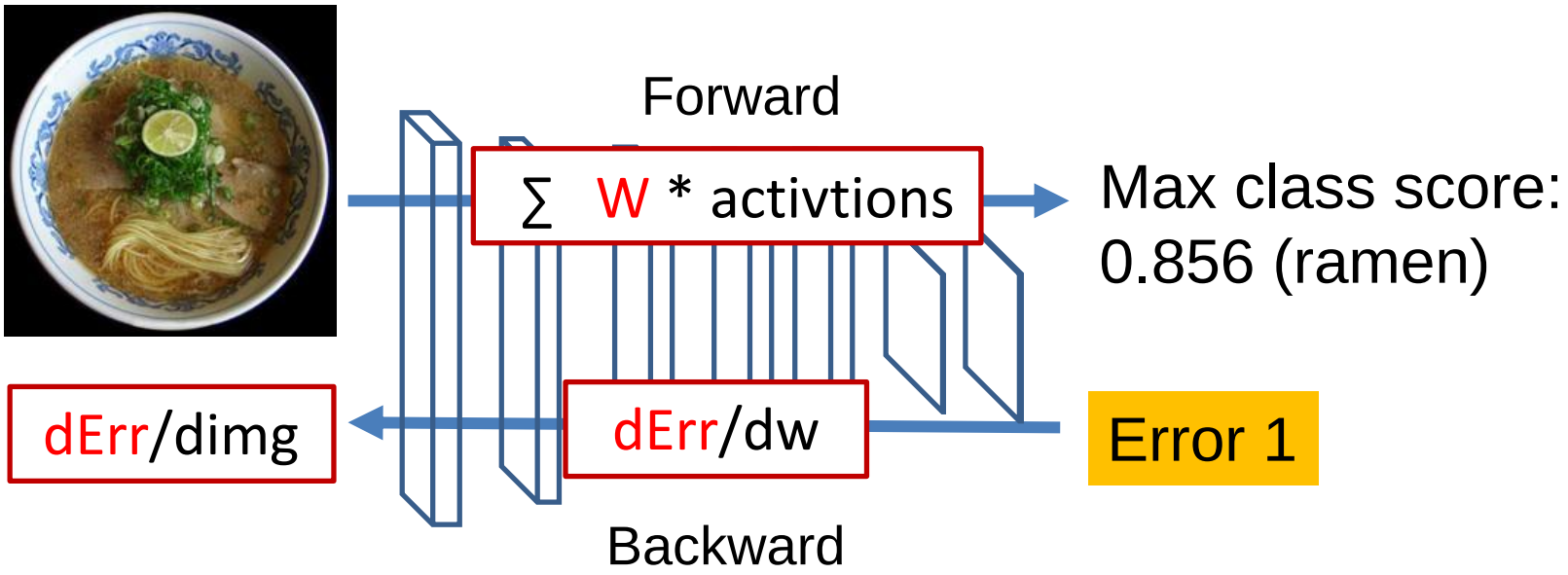
# Segmentation with ConvNet without pixel-level annotation

- Segmentation with object saliency maps computed by Back Propagation(BP) and GrabCut
  - [Simonyan et al. 2014]
  - Prepare only pre-trained CNN for food classification
  - Need no pixel-level annotation (weakly supervised)



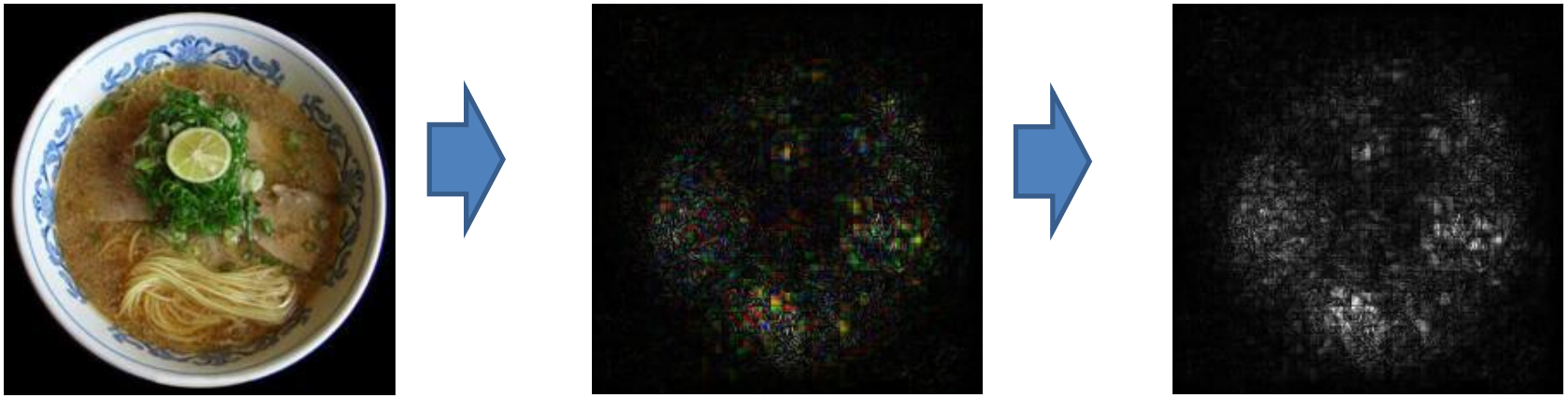
# CNN with back propagation

- SGD with BP is a common training method of CNN
- SGD adjusts weights to minimize error along -dE/dw
- BP chains derivatives(dErr/dw) from top to image



# Object saliency map: Back Propagation to image level

- Magnitude of dE/dI indicates which pixels need to be changed to maximize class score
- High-value pixels are expected to correspond to the object location
- Take a max value among RGB planes of dE/dI

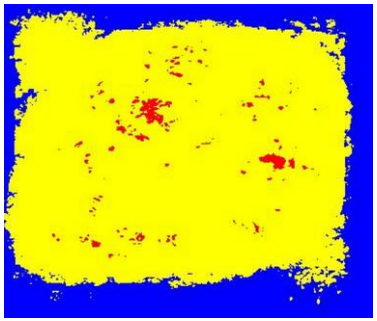


Input image | BP result (dE/dI) | Object saliency map

# Grab cut based on object saliency map

- Graph-cut based segmentation method
- Generate seeds from a saliency map
  - Positive area (upper 5% : red)
  - Negative area (lower 10% : blue)
  - Other (yellow)

Original



Seed

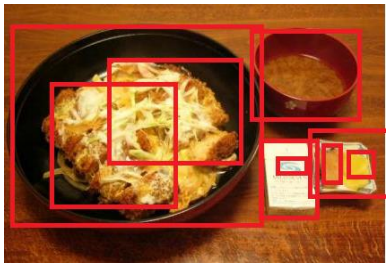
Grab cut result

# Improve BP segmentation

- Weak points
  - separate neighbor object → hard
  - detect small object → hard
- Improvement
  - Rich feature CNN(RCNN) [Girshk et al 2014]
  - Propose many regions and recognize all regions
  - Boost precision at object detection

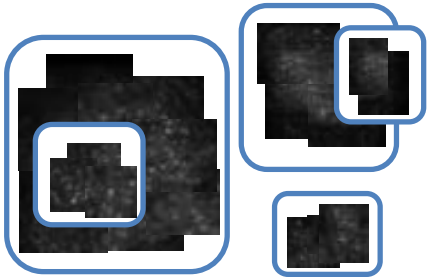
  - Propose many regions and segment all regions

# Proposed method

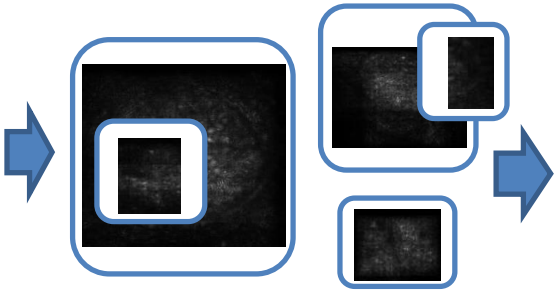


Selective Search(1)

BB grouping(2)



Back propagation(3)



Saliency maps(4)

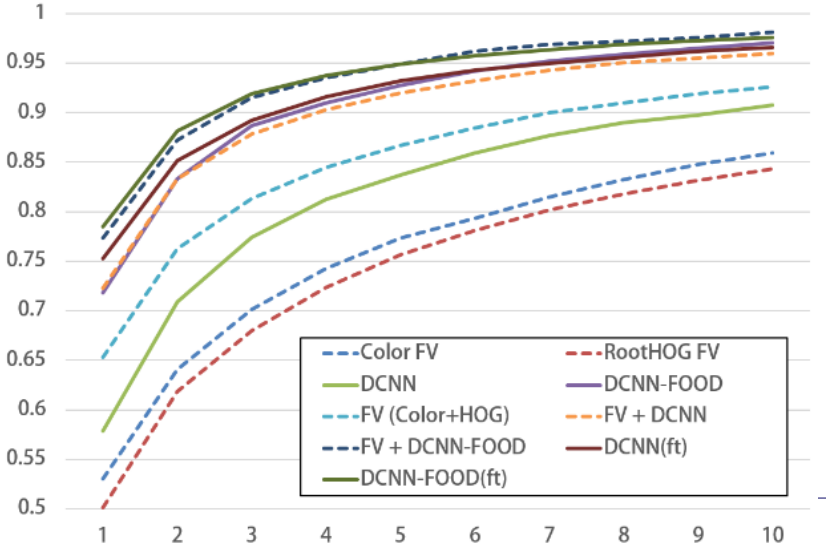
Grab cut (5)

NMS(6)

Result

# Implementation detail

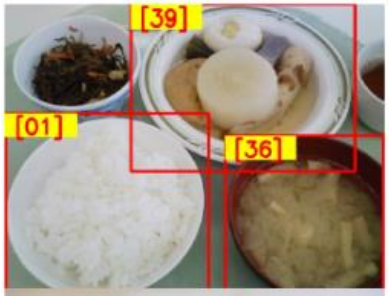
- Trained CNN
  - AlexNet pre-trained ImageNet 2000 classes
    - 1000 class (general) + 1000 class (food-related)
    - About 2 million training images
  - Fine-tuned with UEC-FOOD100
    - 11565 images of 100 classes with bounding box annotation
    - Top-1 78.5 %
    - Top-5 94.9 %

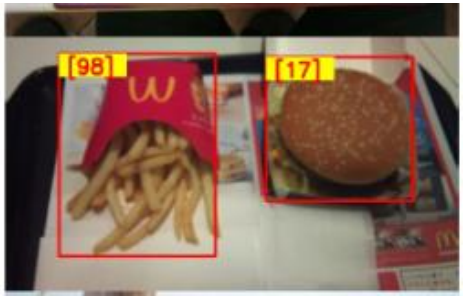


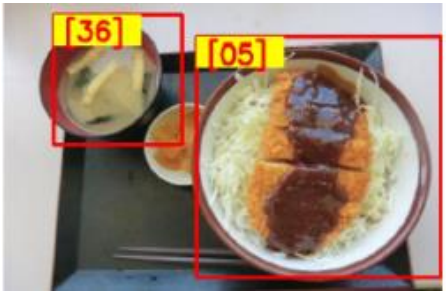
# Experiments

## (1) **Multiple food item images in UEC-Food 100**

- 1175 images
- Have bounding boxes as ground truth (no pixel GT)

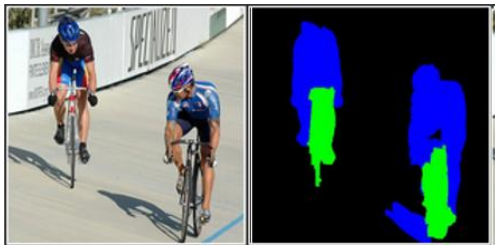


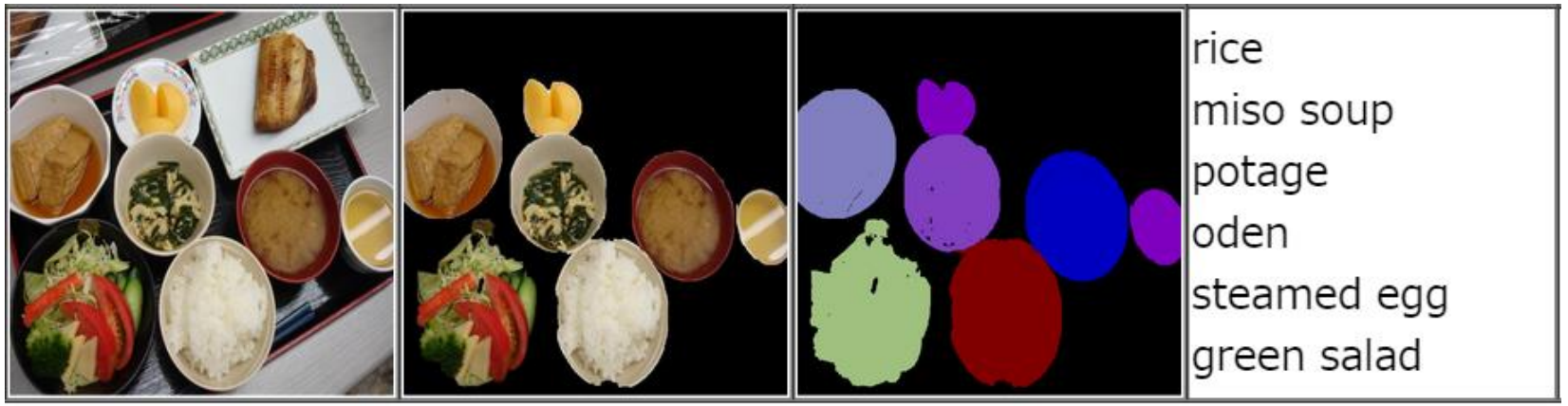

## (2) **Pascal VOC 2012**

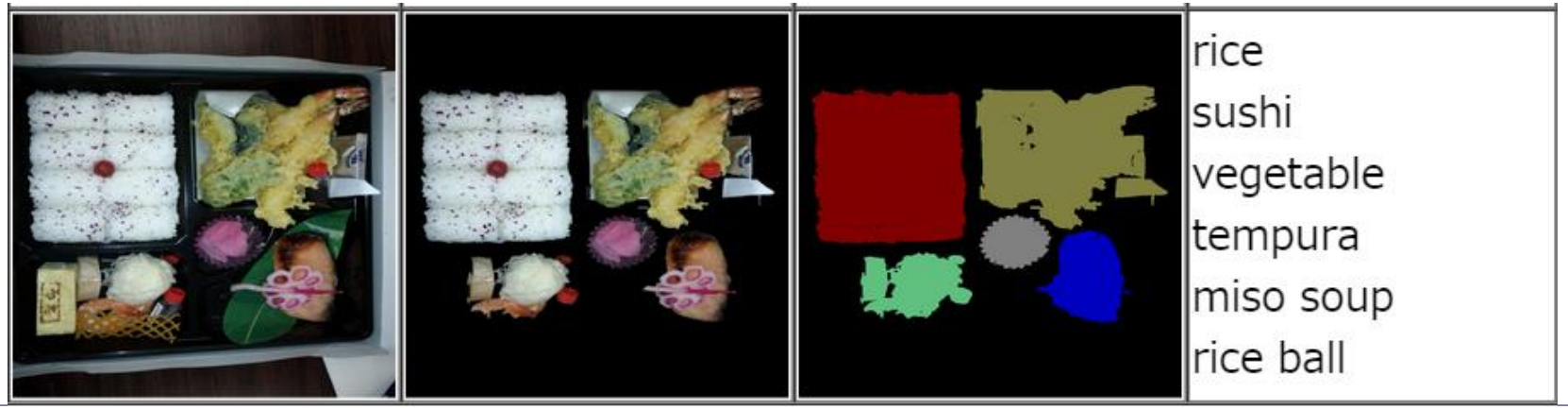
- Have pixel-level GT



# Segmentation results of UEC-Food

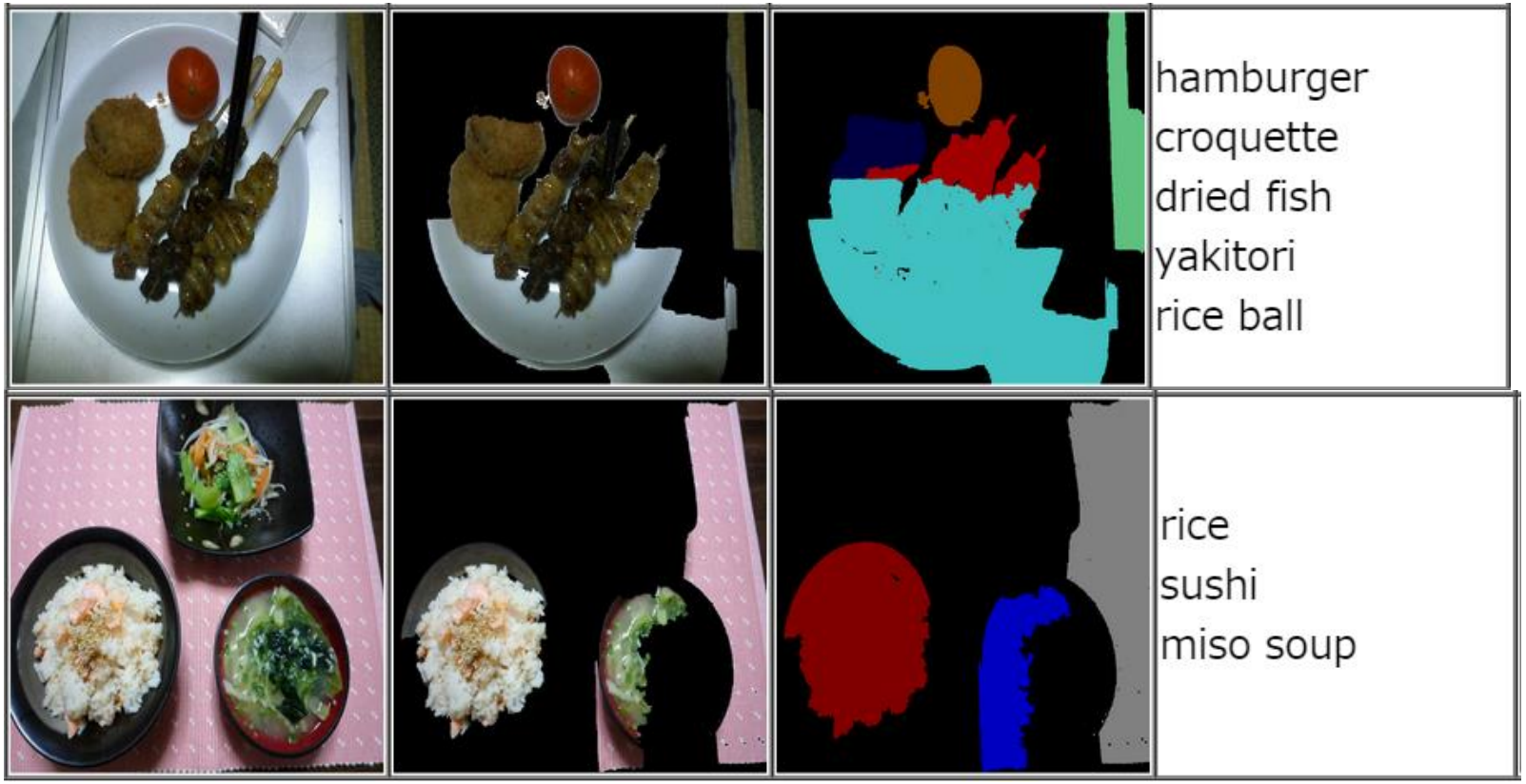
- Success examples



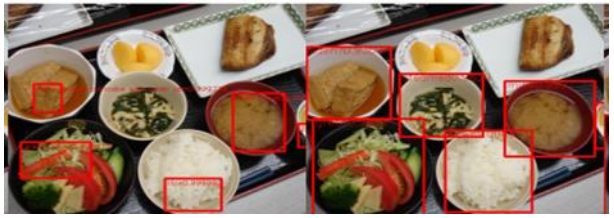
# Segmentation results of UEC-Food

- Failure examples

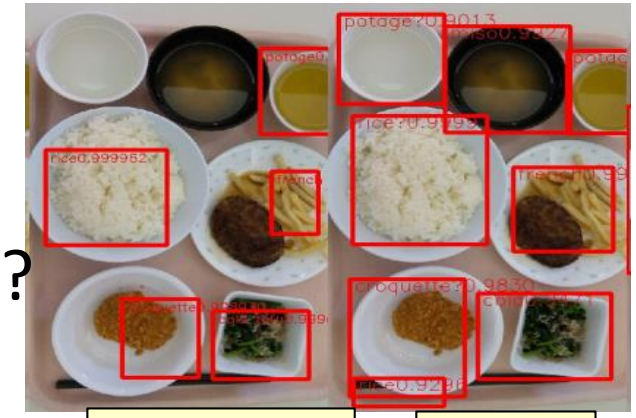


# (1) Evaluation with UEC-Food in the bounding box level

- Compared with R-CNN



- Evaluate BB detection accuracy
  - R-CNN tends to extract smaller BB
  - Food recognition like texture recognition?

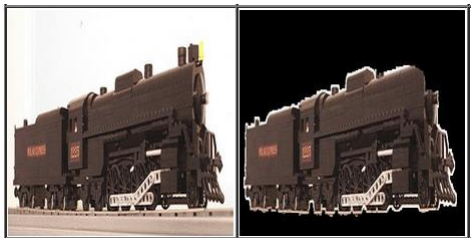



| | 100 class (all) | 53 class (*#item* >= 10) | 11 class (*#item* >= 50) |
|---|---|---|---|
| RCNN | 26.0 | 21.8 | 25.7 |
| Ours | **49.9** | 55.3 | **55.4** |

# (2) Evaluation with Pascal VOC in the pixel level

- PASCAL VOC 2012
  - Generic object images 20 class (bus, dog , etc)
- Evaluate segmentation accuracy in pixes level

| method | mean IU on PASCAL VOC 2012 |
|---|---|
| fully supervised | |
| SDS [4] | 51.6 |
| FCN [9] | 62.2 |
| weakly supervised | |
| ours | 36.4 |
| Pedro-seg [10] | 40.6 |



# Conclusions

- We proposed an improved method of CNN-based segmentation.
- We applied the proposed method to multiple food images in the UEC-FOOD dataset as well as Pascal VOC.

- For future work, we plan to extend CNN-based segmentation with superpixels and CRF.