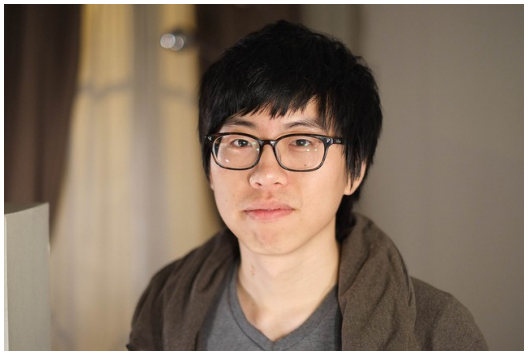
Learning Deconvolution Network for Semantic Segmentation


Hyeonwoo Noh, Seunghoon Hong, Bohyung Han

Mehmet Günel

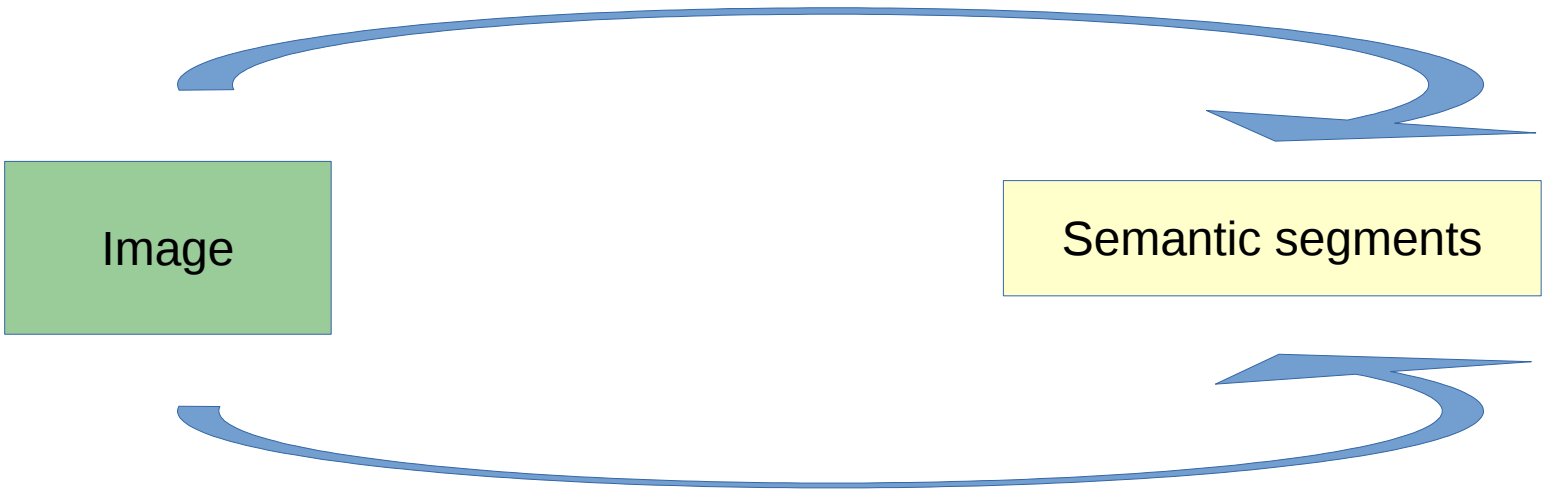
What is this paper about?

- A novel semantic segmentation algorithm
- Convolution & Deconvolution layers
- Fully convolutional network integrated with deep deconvolution network and makes proposal-wise prediction
- Identifies detailed structures and handles objects in multiple scales naturally

Overview - What is and what is not

- Semantic segmentation
  - Scene labeling
  - Pixel-wise classification

Semantically meaningful parts + classify each part into predetermined classes



Classify each pixel!

Problem: Background

- Semantic segmentation algorithms are often formulated to solve structured pixel-wise labeling problems based on CNN
- Conditional random field (CRF) is optionally applied to the output map for fine segmentation
- Network accepts a whole image as an input and performs fast and accurate inference

Problem: Limitations

- Fixed-size receptive field

→ the object that is substantially larger or smaller than the receptive field may be fragmented or mislabeled

→ small objects are often ignored and classified as background

Problem: Limitations

(a) Inconsistent labels due to large object size

Problem: Limitations

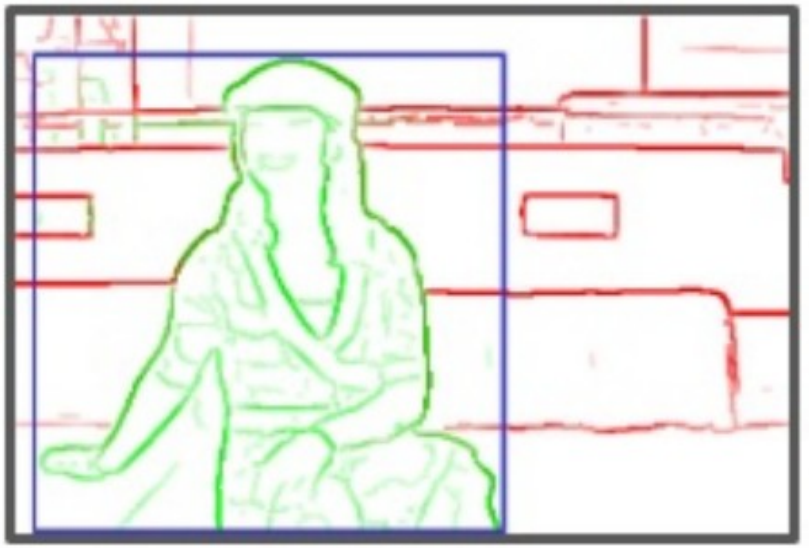
(b) Missing labels due to small object size

Related Work

- J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In CVPR, 2015 (Previous presentation)

- C. L. Zitnick and P. Doll´ar. Edge boxes: Locating object proposals from edges. In ECCV, 2014

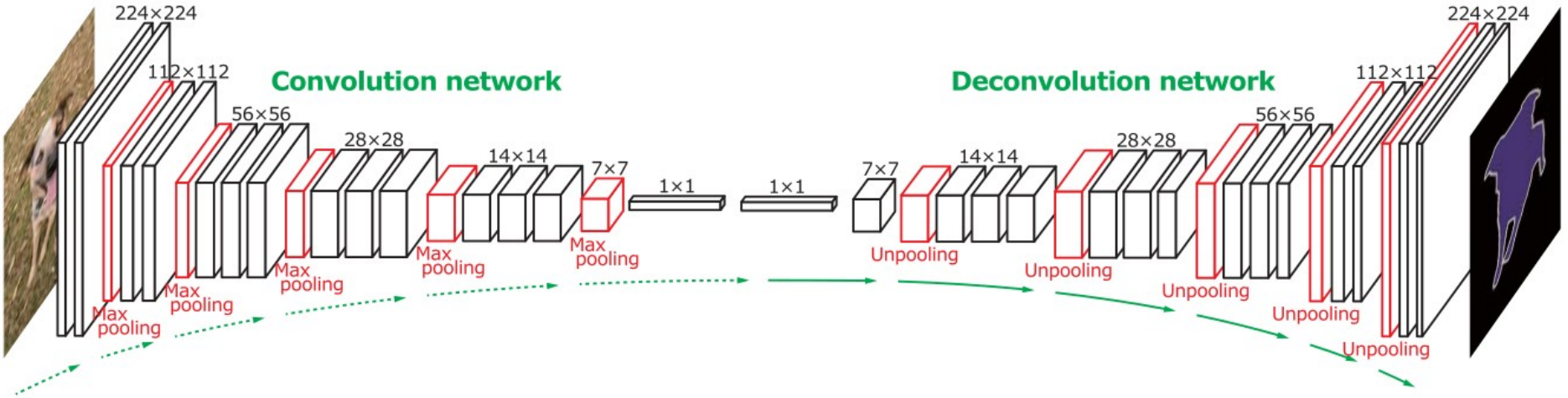
Object proposals



Contributions

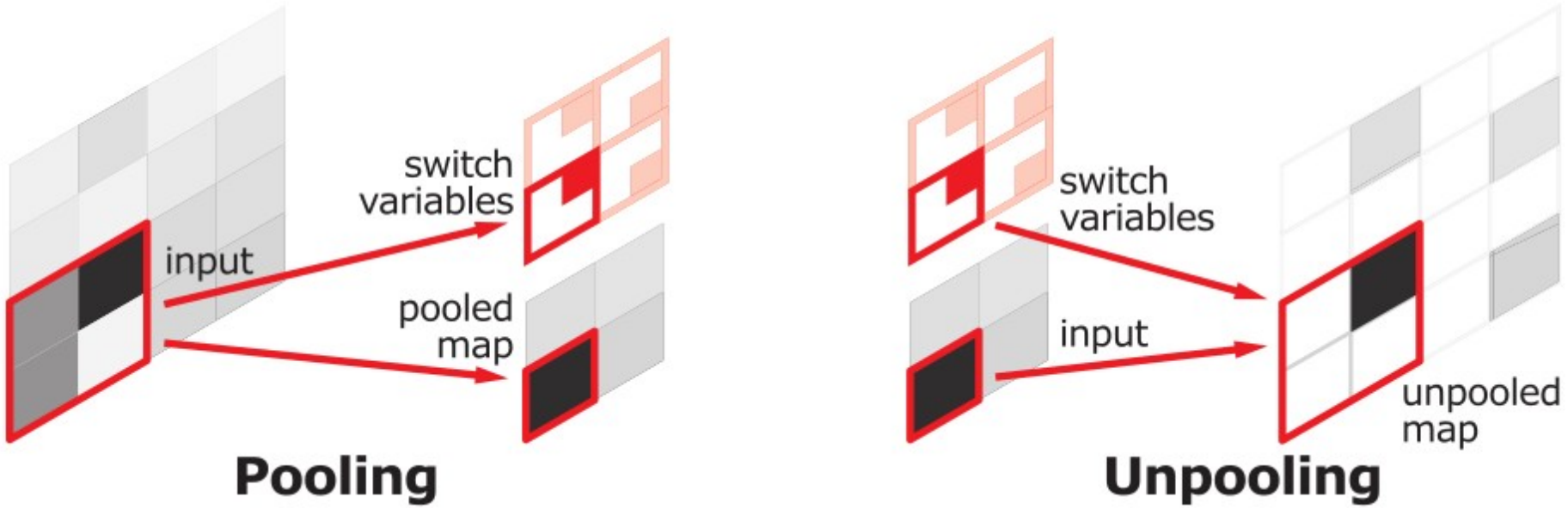
- A multi-layer deconvolution network, which is composed of deconvolution, unpooling, and rectified linear unit (ReLU) layers
- Free from scale issues found in FCN-based methods and identifies finer details of an object
- PASCAL VOC 2012 dataset best accuracy with FCN

Network Model



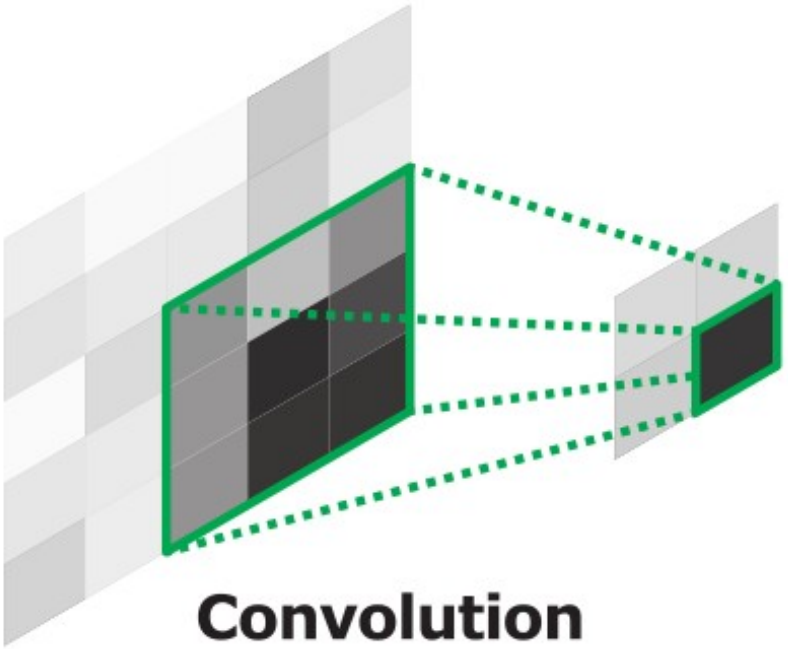
Approximately 252M parameters in total

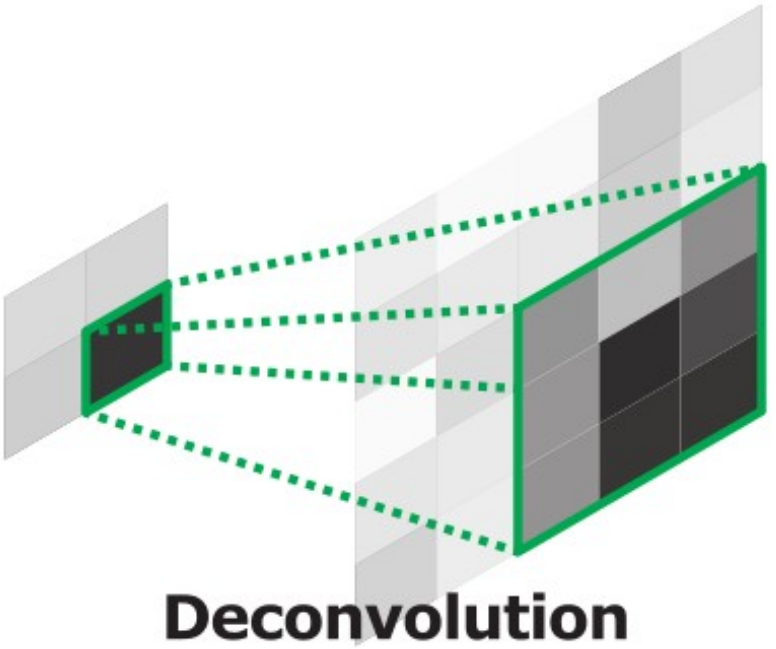
Pooling & Unpooling



Example specific

Convolution & Deconvolution





Class specific

Training Stage

- Batch Normalization
  - Internal covariate shift problem

- Two-stage Training
  - crop object instances using ground-truth annotations
  - utilize object proposals to construct more challenging examples

Segmentation Maps Integration Formula

$$P(x, y, c) = \max_i G_i(x, y, c), \quad \forall i, \tag{1}$$

$$P(x, y, c) = \sum_i G_i(x, y, c), \quad \forall i. \tag{2}$$

Experimental Setup

- PASCAL VOC 2012 segmentation dataset
- All training and validation images are used to train
- They used augmented segmentation annotations
  - Extend the bbox 1.2 times larger to include local context around the object
  - Object & background labeling
  - 250 × 250 input image randomly cropped to 224 × 224 with optional horizontal + flipping
  - The number of training examples is 0.2M and 2.7M in the first and the second stage

Experimental Setup

- Caffe framework
- Stochastic gradient descent with momentum
- Initial learning rate, momentum and weight; 0.01, 0.9 and 0,0005
- VGG 16-layer net pre-trained on ILSVRC
- Network converges after approximately 20K and 40K SGD iterations with mini-batch of 64 samples
- Training takes 6 days (2 days for the first stage and 4 days for the second stage)
- Nvidia GTX Titan X GPU with 12G memory

Inference

- For each testing image, we generate approximately 2000 object proposals, and select top 50 proposals based on their objectness scores

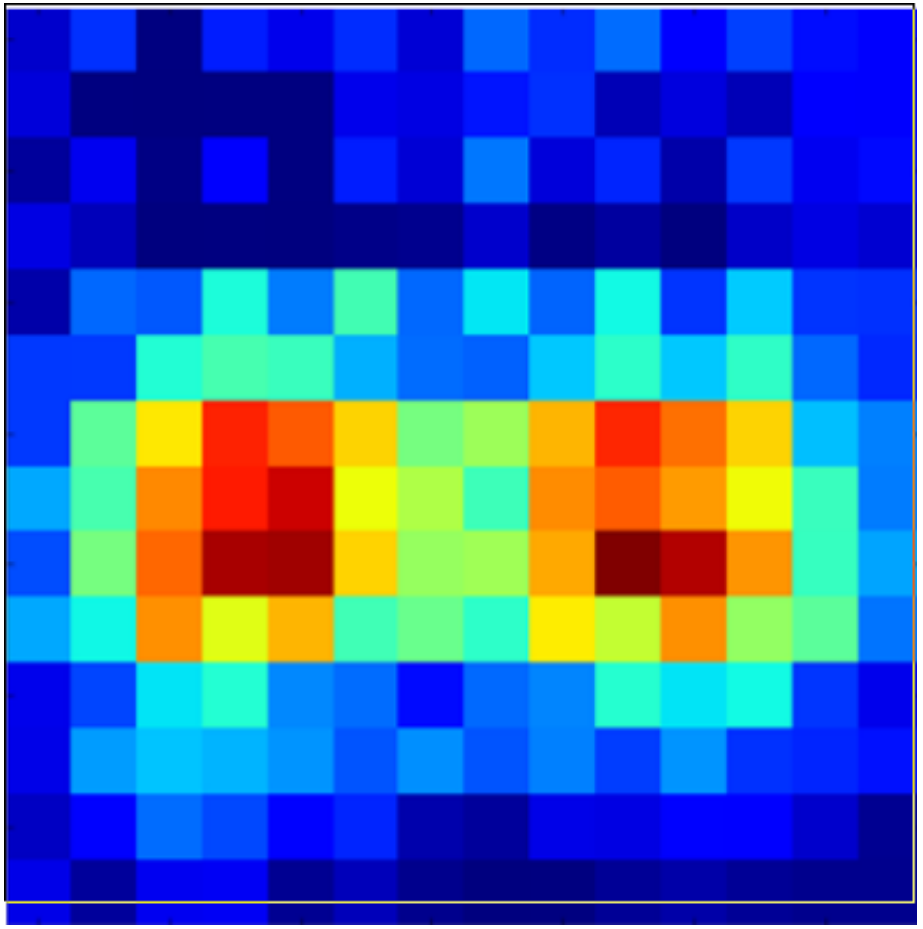
- Compute pixel-wise maximum to aggregate proposal-wise predictions

Evaluation Metrics

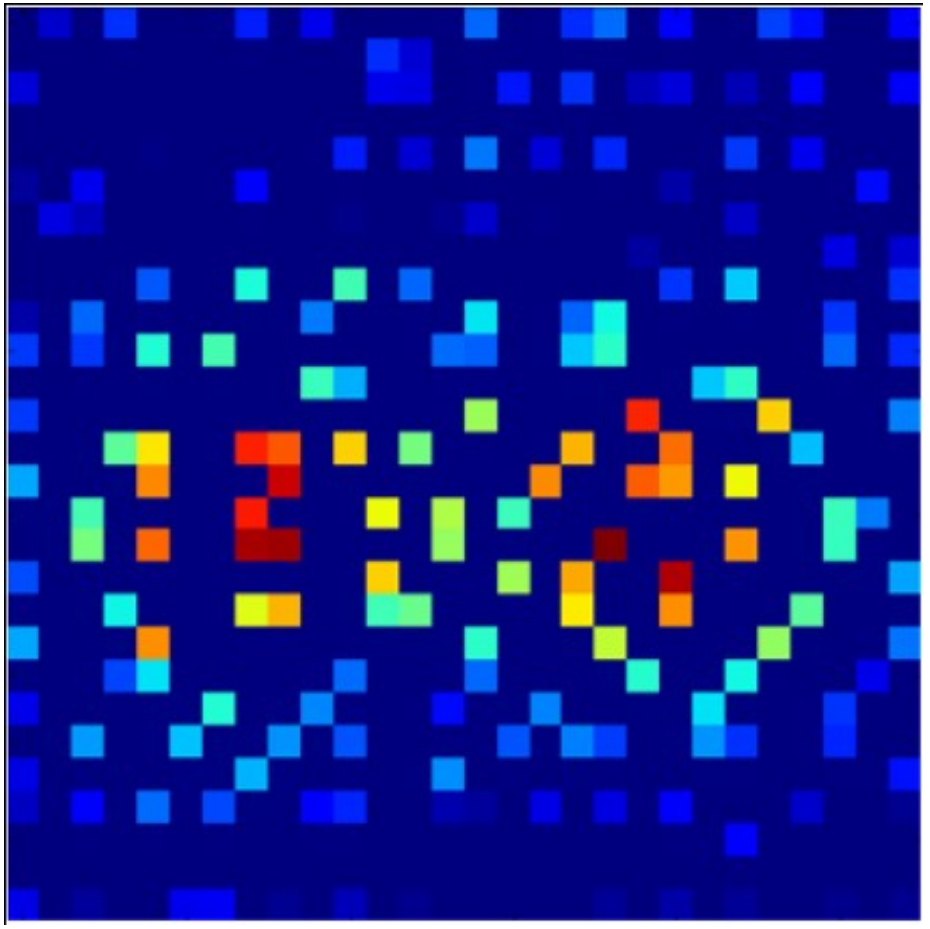
- *comp6* evaluation protocol;
  - intersection over Union (IoU) between ground truth and predicted segmentations

Visualization of activations

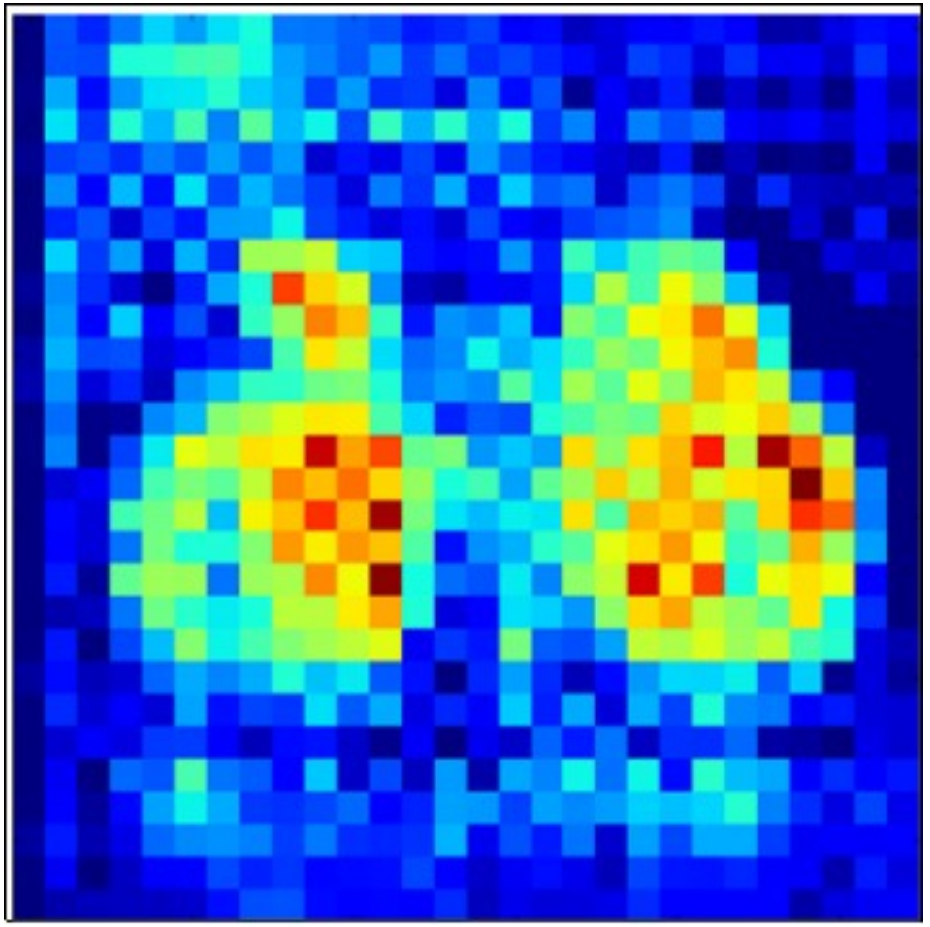
Visualization of activations



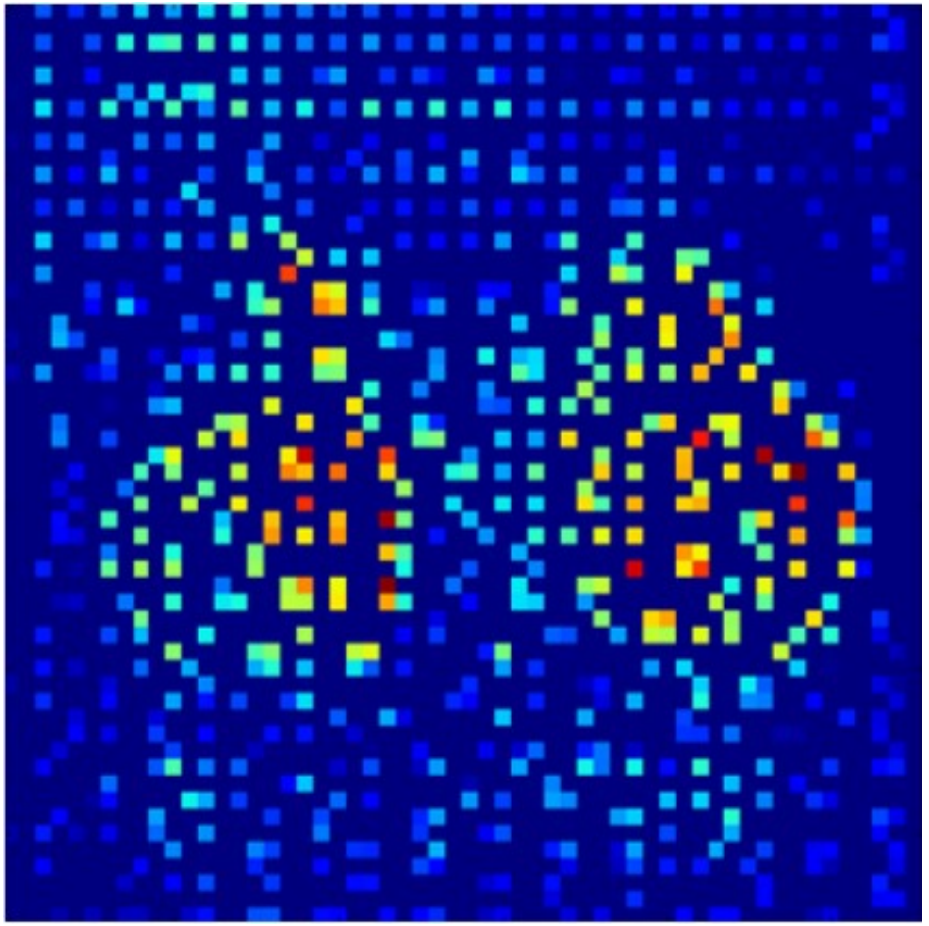
Visualization of activations



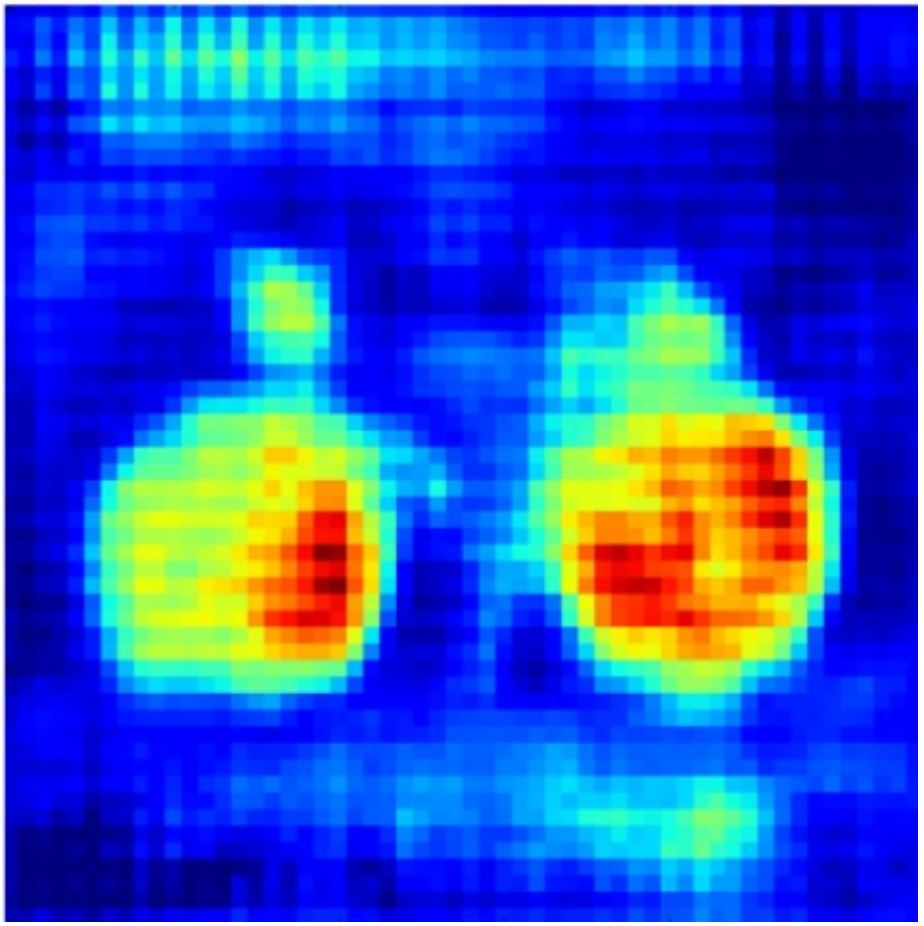
Visualization of activations



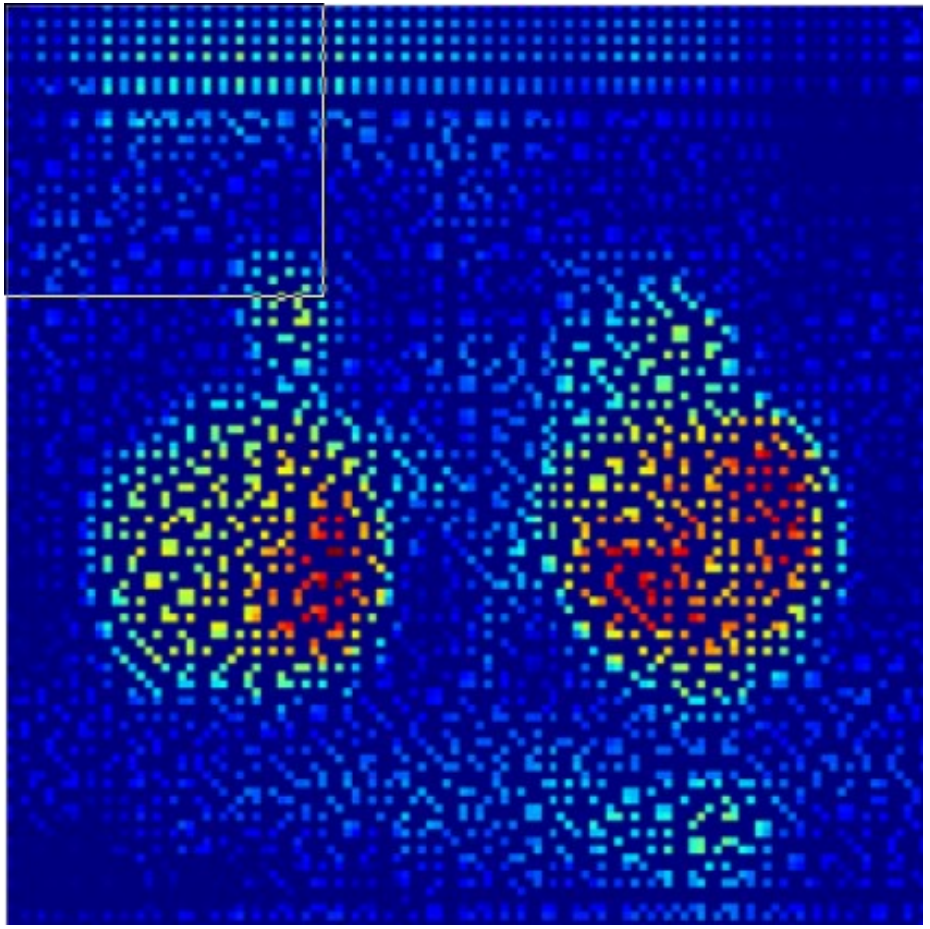
Visualization of activations



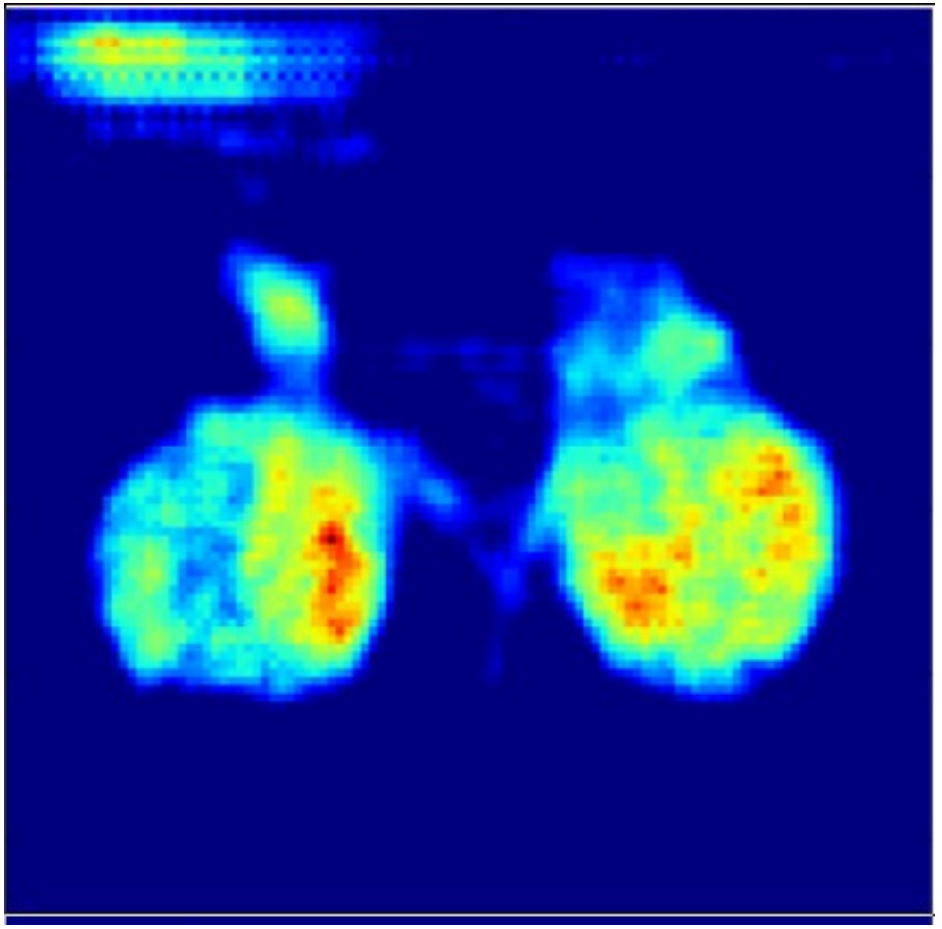
Visualization of activations



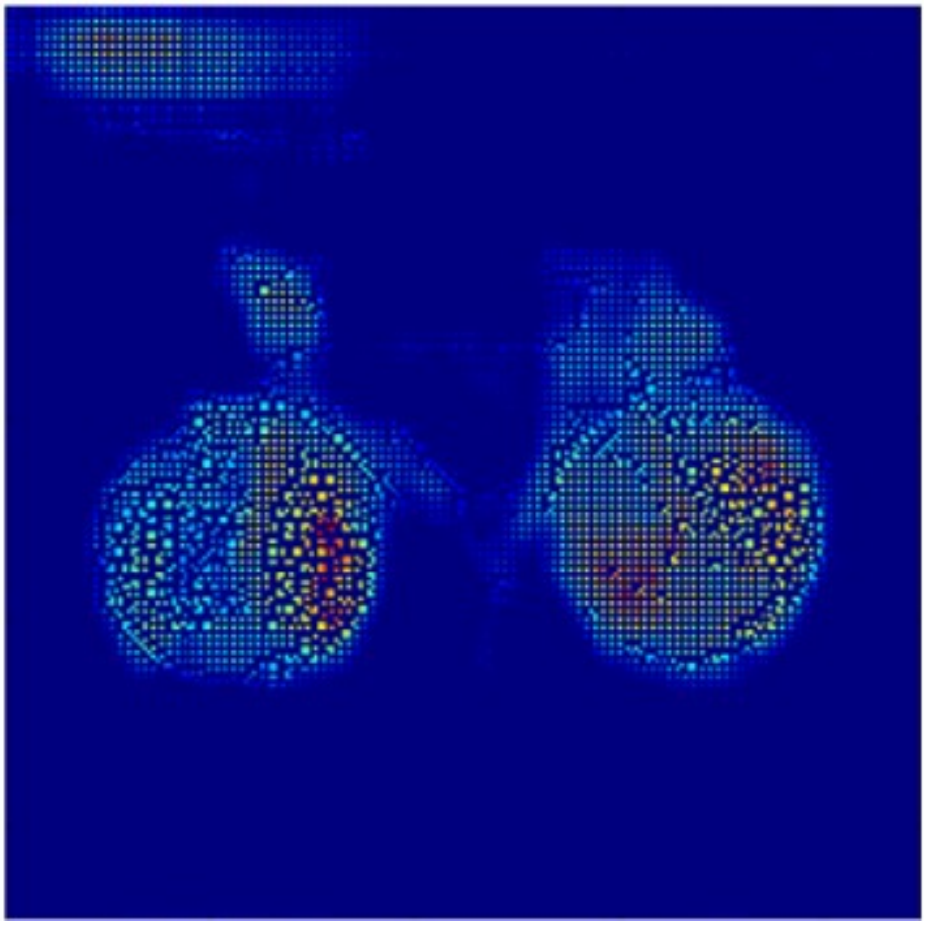
Visualization of activations



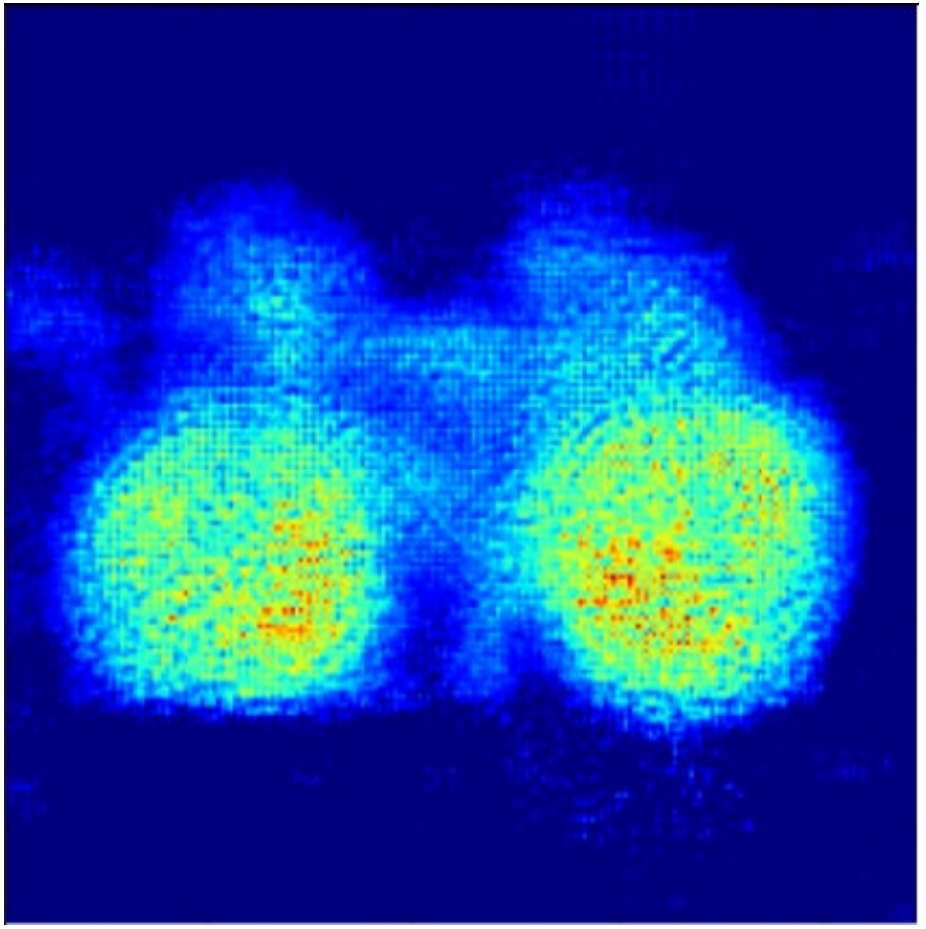
Visualization of activations



Visualization of activations



Visualization of activations



Results

- CRF increase approximately 1% point

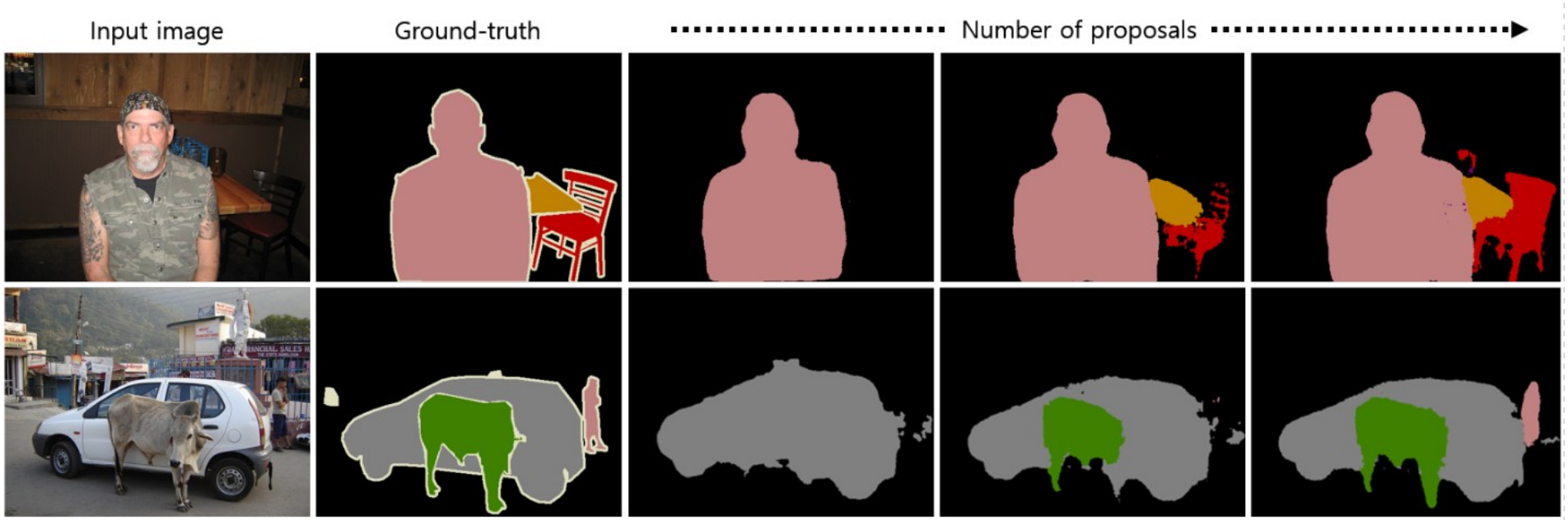
- Ensemble with FCN-8s improves mean IoU about 10.3% and 3.1% point with respect to FCN-8s and DeconvNet

Results - Comparisons

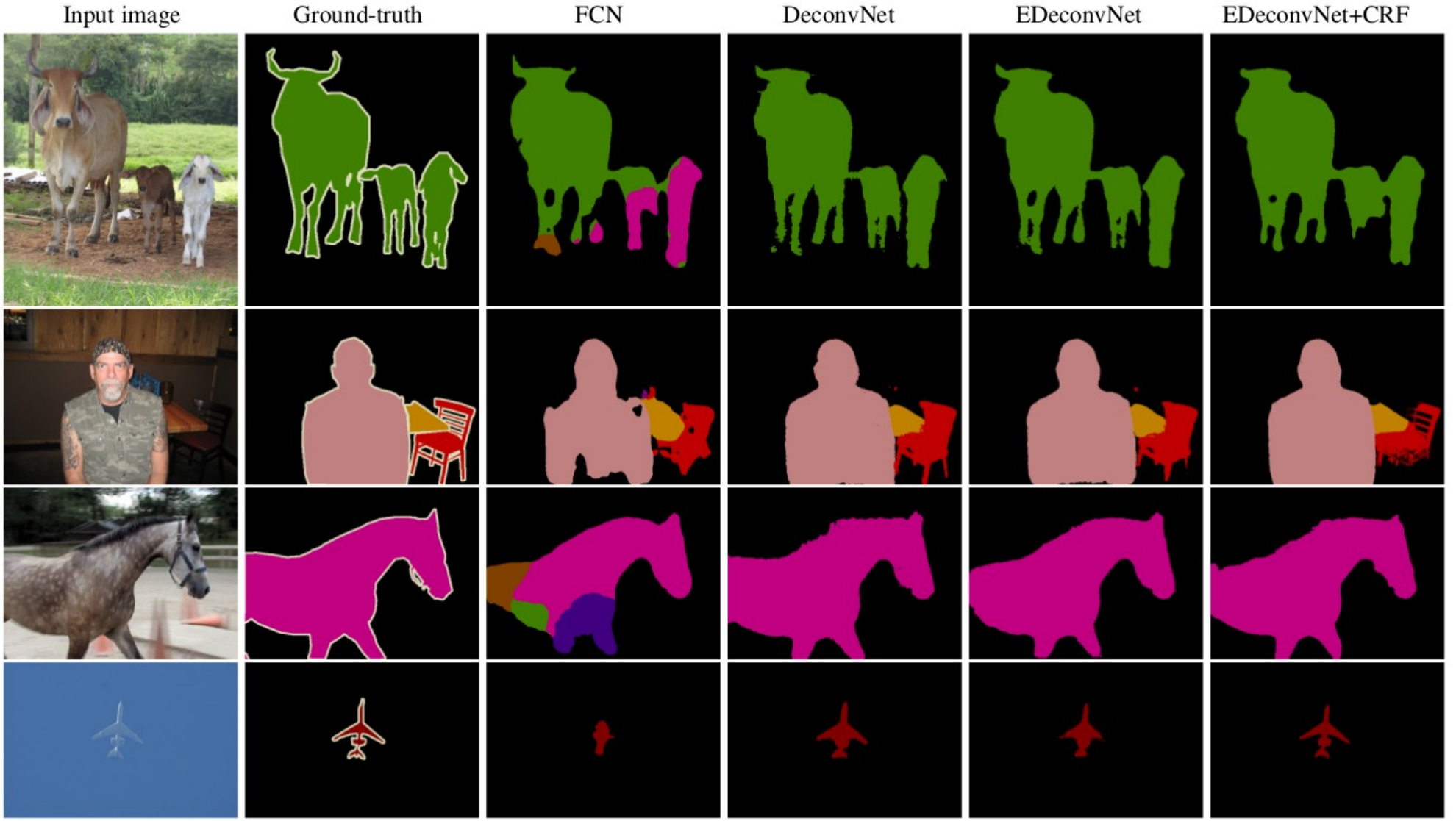
Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
EDeconvNet+CRF	**93.1**	**89.9**	**39.3**	**79.7**	**63.9**	**68.2**	**87.4**	**81.2**	**86.1**	**28.5**	**77.0**	**62.0**	**79.0**	**80.3**	**83.6**	**80.2**	**58.8**	**83.4**	**54.3**	**80.7**	**65.0**	**72.5**
DeepLab-CRF	93.1	84.4	54.5	81.5	63.6	65.9	85.1	79.1	83.4	30.7	74.1	59.8	79.0	76.1	83.2	80.8	59.7	82.2	50.4	73.1	63.7	71.6
TTI-Zoomout-16	89.8	81.9	35.1	78.2	57.4	56.5	80.5	74.0	79.8	22.4	69.6	53.7	74.0	76.0	76.6	68.8	44.3	70.2	40.2	68.9	55.3	64.4
FCN8s	91.2	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	62.2
MSRA-CFM	87.7	75.7	26.7	69.5	48.8	65.6	81.0	69.2	73.3	30.0	68.7	51.5	69.1	68.1	71.7	67.5	50.4	66.5	44.4	58.9	53.5	61.8
Hypercolumn	88.9	68.4	27.2	68.2	47.6	61.7	76.9	72.1	71.1	24.3	59.3	44.8	62.7	59.4	73.5	70.6	52.0	63.0	38.1	60.0	54.1	59.2

Evaluation results on PASCAL VOC 2012 test set. (algorithms trained without additional data)

Results

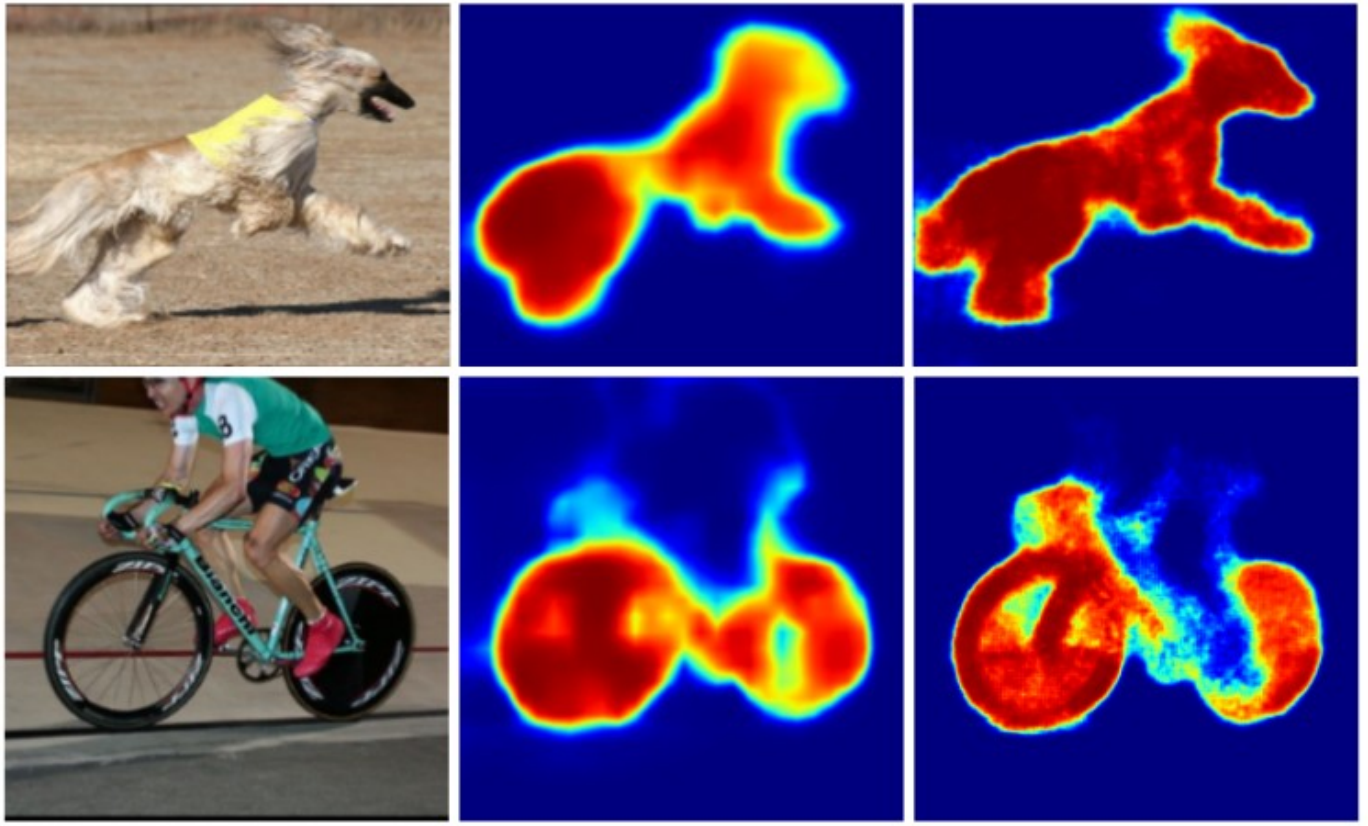


Results - Strengths



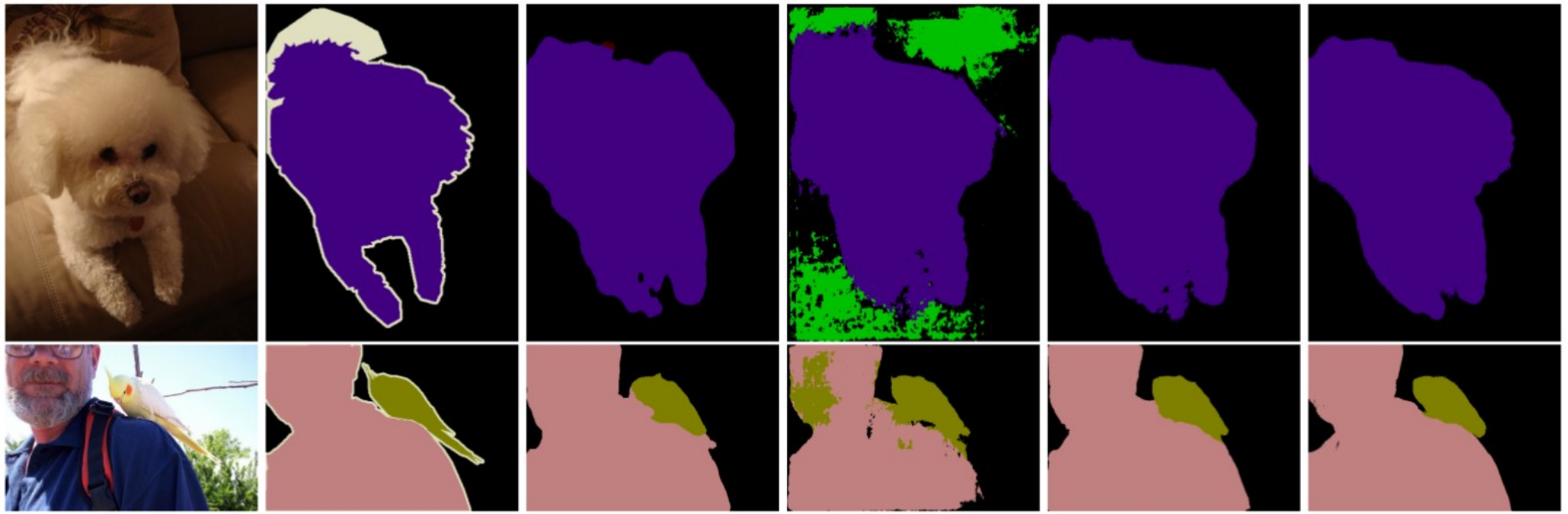
Better results

Results - Strengths



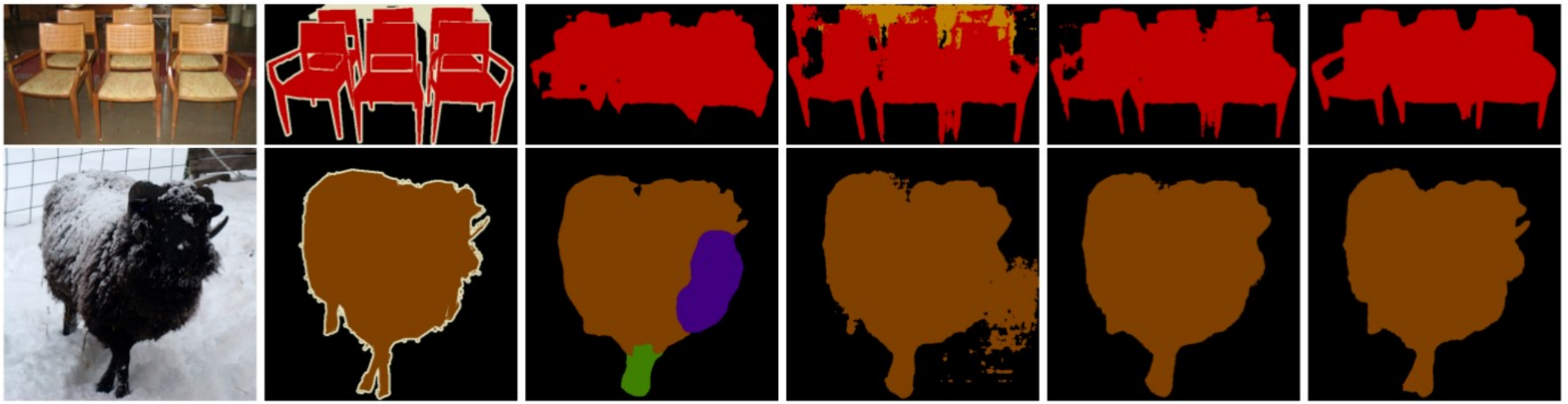
(a) Input image (b) FCN-8s (c) Ours

Results - Weakness



Worse than FCN results

Results



Ensemble results

Conclusions & Future Directions

- A novel semantic segmentation algorithm by learning a deconvolution network
- Elimination of fixed-size receptive field limit in the fully convolutional network
- Ensemble approach of FCN + CRF
- State-of-the-art performance in PASCAL VOC 2012 without external data
- A bigger network with better proposals