Deconvolution Networks

Pattern Recognition & Machine Learning Laboratory
Jinha Lim, Aug 8, 2021

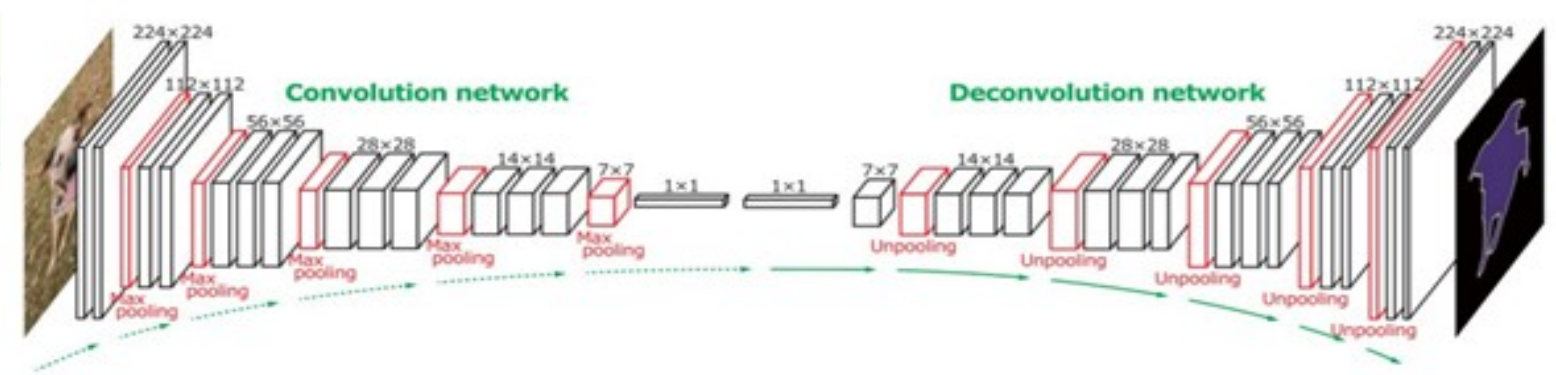
Learning Deconvolution Network for semantic Segmentation

- **Motivation**
  - Network has a predefined fixed-size reception feild
  - Detailed structures of an object are often lost or smoothed
    - Cause by label is too coarse and deconvolution procedure is overly simple
- **Approach**
  - Deconvolution network
    - CNN architecture designed to generate large output
    - Enables dense output score prediction
  - Instance-wise prediction
    - Inference on object proposals, then aggregate
    - Enables recognition of objects with multiple scales
- **Evaluation**
  - PASCAL VOC 2012 segmentation benchmark
    - contains 1456 test images and 20 object categories

Method	bkg	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
Hypercolumn [11]	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
MSRA-CFM [3]	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
FCN8s [19]	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
TTI-Zoomout-16 [20]	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
DeepLab-CRF [1]	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
DeconvNet	92.7	85.9	42.6	78.9	62.5	66.6	87.4	77.8	79.5	26.3	73.4	60.2	70.8	76.5	79.6	77.7	58.2	77.4	52.9	75.2	59.8	69.6
DeconvNet+CRF	92.9	87.8	41.9	80.6	63.9	67.3	**88.1**	78.4	81.3	25.9	73.7	61.2	72.0	77.0	79.9	78.7	59.5	78.3	**55.0**	75.2	61.5	70.5
EDeconvNet	92.9	88.4	39.7	79.0	63.0	67.7	87.1	**81.5**	84.4	27.8	76.1	61.2	78.0	79.3	83.1	79.3	58.0	82.5	52.3	80.1	64.0	71.7
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**
* WSSL [21]	93.2	85.3	36.2	**84.8**	61.2	67.5	84.7	81.4	81.0	**30.8**	73.8	53.8	77.5	76.5	82.3	**81.6**	56.3	78.9	52.3	76.6	63.3	70.4
* BoxSup [2]	**93.6**	86.4	35.5	79.7	**65.2**	65.2	84.3	78.5	83.7	30.5	76.2	**62.6**	**79.3**	76.1	82.1	81.3	57.0	78.2	**55.0**	72.5	**68.1**	71.0

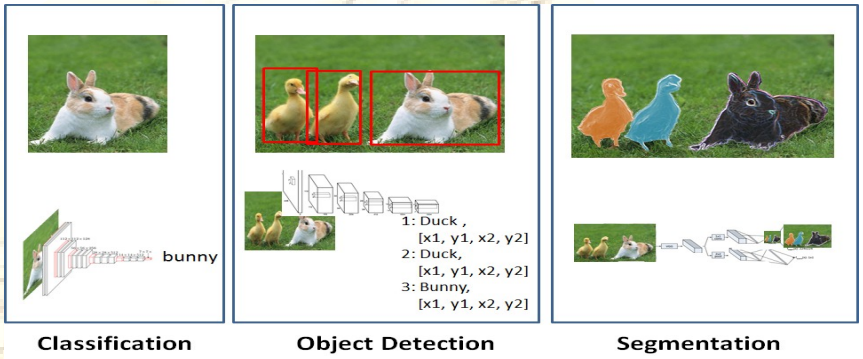
Introduction

- **CNN (Convolutional Neural Network)**
  - A multi-layered feed-forward neural network, made by stacking many hidden layers on top of each other in sequence
  - The hidden layers are typically covolutional layers follwed by activation layers, some of them followed by pooling layers
- **Deconvolution Network**
  - The operation inverse to convolution\
  - Also known as upsampling process
    - synonymous with expansion, on a sequence of smaples of a signal or other continuous function, produces an approximation of the sequence that have been obtained by sampling the signal at a higher rate
  - Designed to remove or reverse the blurring present in microscope images included by the limited aperture of the objective



Motivation

- **Image segmentation**
  - Classification: labeling one image as whole
  - Detection: separating different ob-ject and boxing each objects
  - Segmentation: recognition of objects in the image with pixel level detail



Limitation of Fully Convolutional Network (FCN)

Idea as to design network as stack of convolutional layers to make predictions to all at once

To maintain size of input, exponential parameter was required

Skip architectrue: suggests to skip some layer and feeds the output of one layer as the inpuut to the net layer as well as some other layer

Methods (1/3)

- **Padding**
  - Padding works by extending the area of which CNN processes an image
  - The kernel is the neural filter which moves accross the image, scanning each pixel and converting the data into different format size
  - Padding is added to the frame of the image to allow for more space for kernel to cover the image

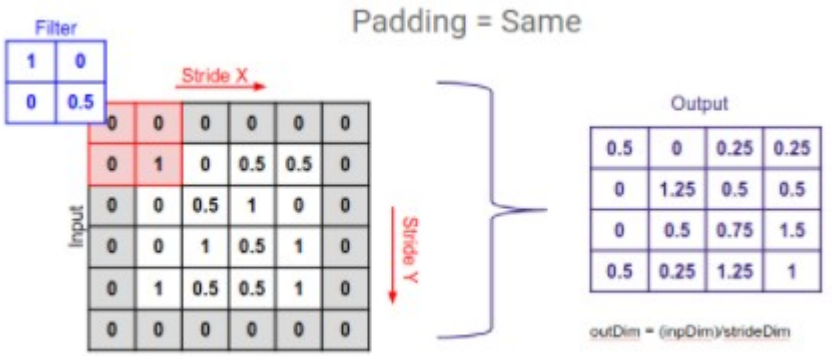


Figure downsampling using padding method

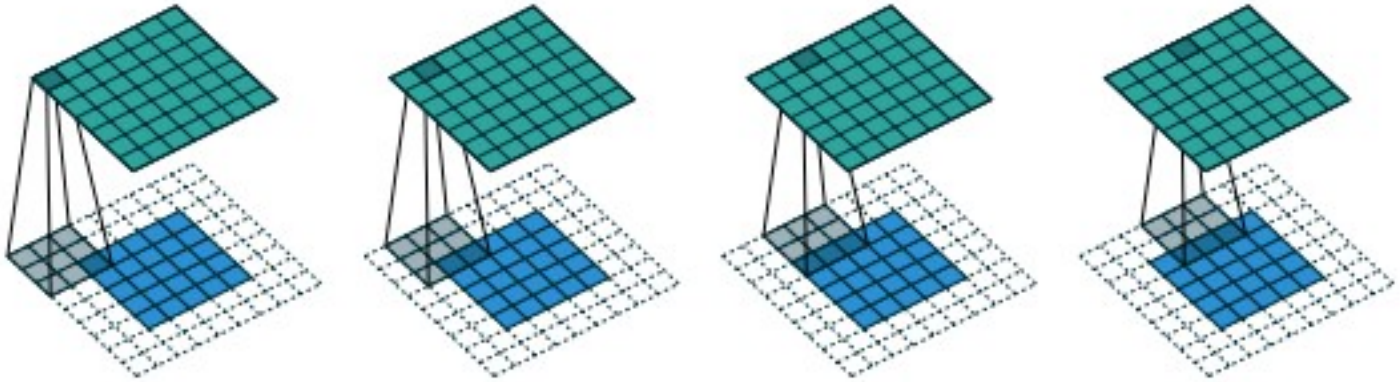


Figure convolving a 3 x 3 kernel over a 5 x 5 input using full padding and unit strides

Methods (2/3)

- **Stride**
  - Parameter of the neural network's filter that modifies the amount of movement over the image or video
  - If a neural network's stride is set to 1, the filter move one pixel, or unit, at a time
  - The size of filter affects the encoded output volume, stride is often set to a whole integer

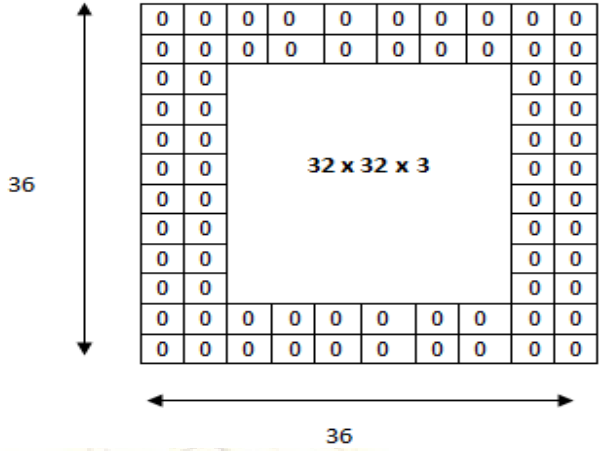


Figure applying 5 x 5 x 3 filters to a 32 x 32 x 3 input volume

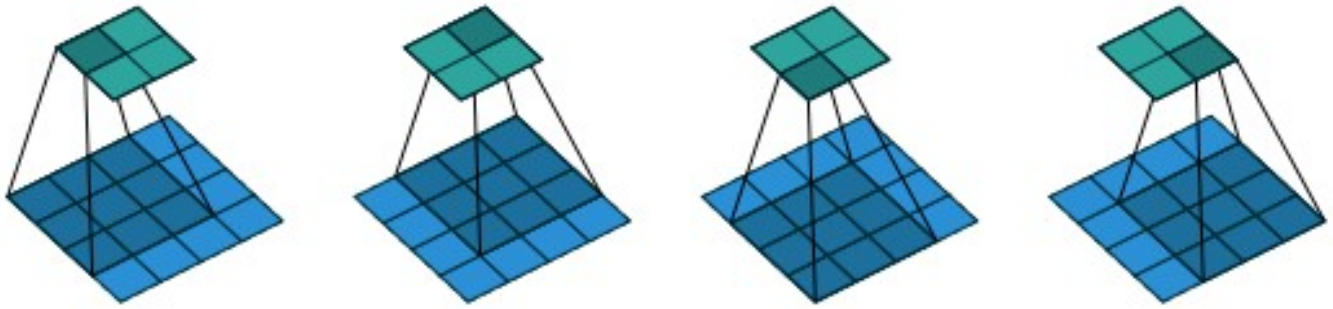
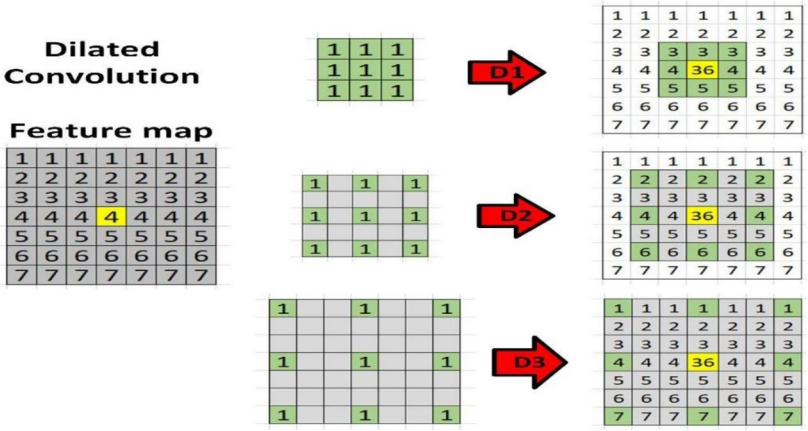


Figure Convolving a 3 × 3 kernel over a 4 × 4 input using unit strides

Methods (3/3)

- **Dilation**
  - Defines a spacing between the values in a kernel
  - A 3x3 kernel with a dilation rate of 2 will have the same field of view as a 5x5 kernel, while only using 9 parameters
  - Delivers a wider field of view at the same computational cost



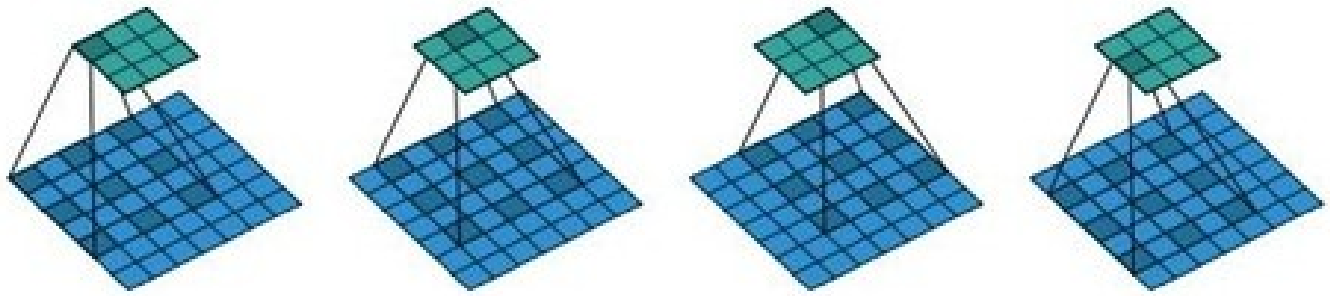
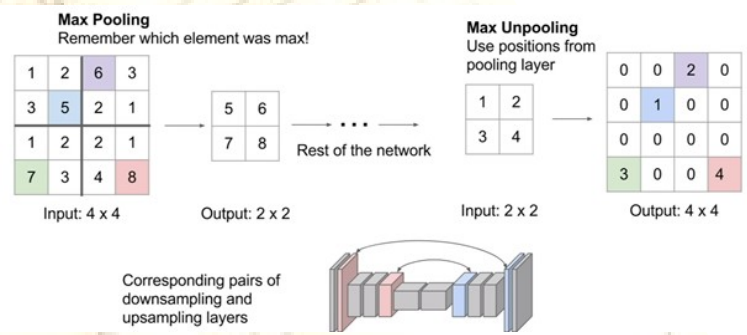


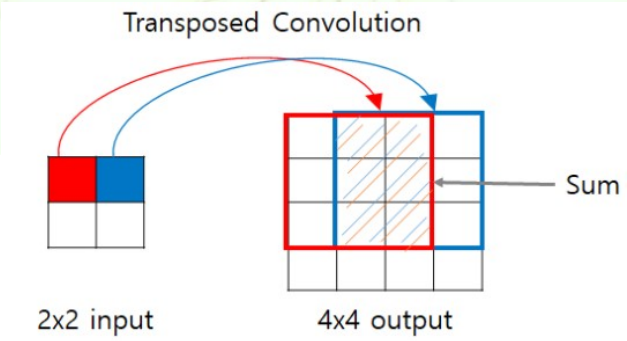
Figure Convolving a 3 × 3 kernel over a 7 x 7 input with a dilation factor of 2

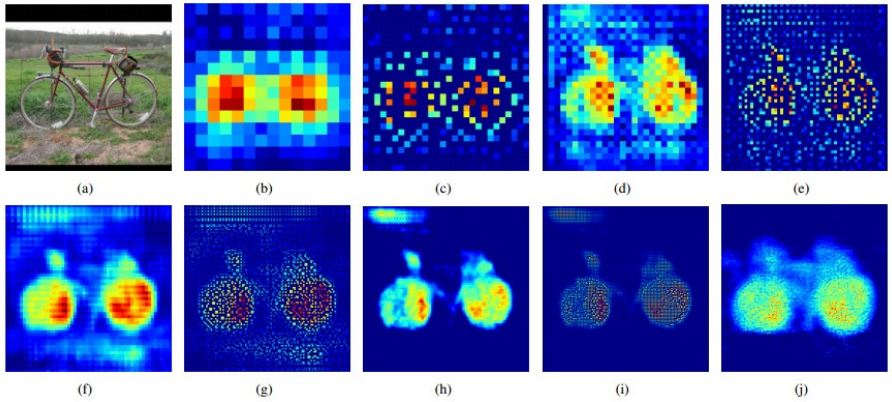
DeepConvNet

- **Unpooling**
  - **Max Pooling**
    - remembers which element was max
    - restores the element using positions from pooling layer
    - reduces information loss compared to Bed of Nails unpooling



- **Deconvolution**
  - Output of unpooling layer is enlarged, yet sparse activation map
  - Deconvolution layers densify the sparse activations obtained by unpooling through convolution-like operations with multiple learned filters





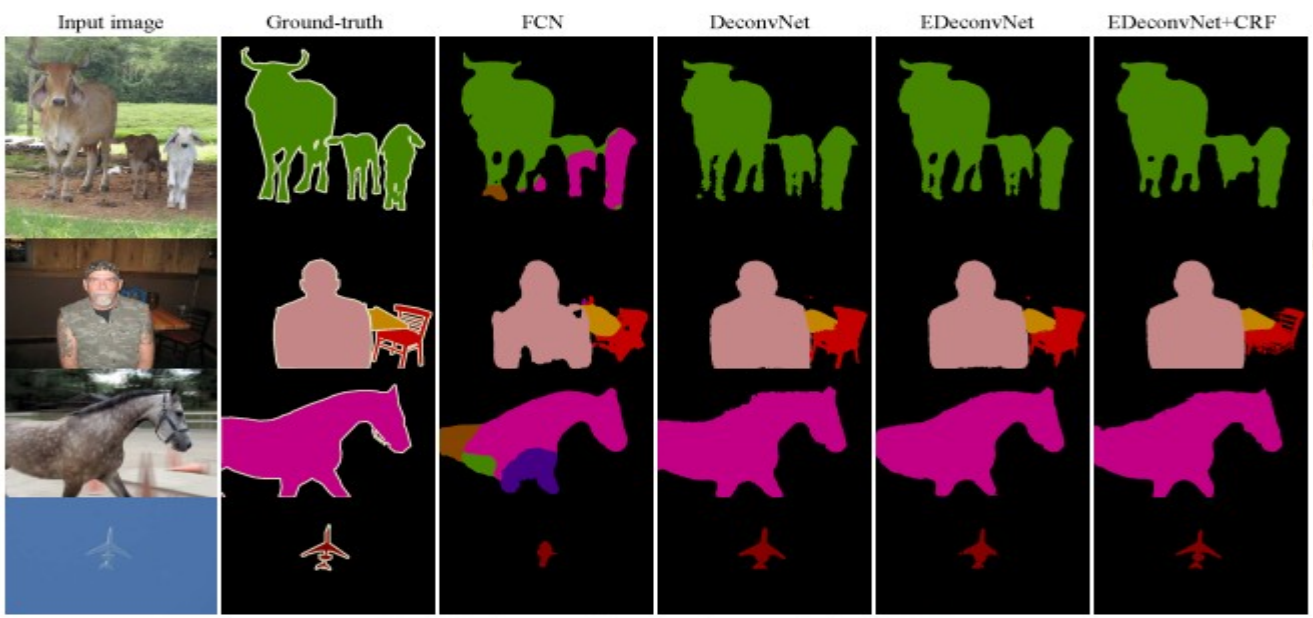
Implementations

- **Experiment**
  - Dataset
    - Employed PASCAL VOC 2012 segmentation dataset for training
  - Training data construction
    - Two-stage training strategy and separate training dataset in each stage
    - Annotated object in training images, and extended the box 1.2 time larger to in-clude local context around the object
      - CRF (Conditional random field) as post processing
      - post processing enhances accuracy by approx 1%
    - On second stage, each training example was extracted from object proposal, where all relevant calss labels are used for annotation
- **Evaluation**
  - Adopt com6 evaluation protocol that measures scores based on Intersection over Union (IoU) between ground truth and predicted segments
  - Improved further performance through an ensemble with FCN-8s
    - imporves IoU about 10.3% and 3.1% point with respect to FCN-8s and DeconvNet

Results

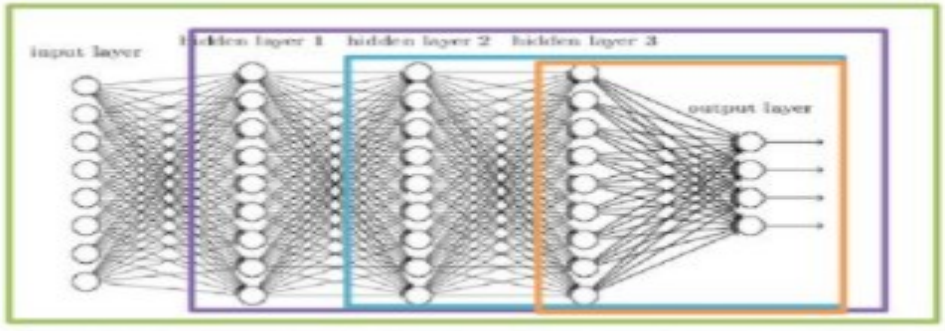
Table 1. Evaluation results on PASCAL VOC 2012 test set. (Asterisk (*) denotes the algorithms that also use Microsoft COCO for training.)

Method	bkg	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
Hypercolumn [11]	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
MSRA-CFM [3]	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
FCN8s [19]	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
TTI-Zoomout-16 [20]	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
DeepLab-CRF [1]	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
DeconvNet	92.7	85.9	42.6	78.9	62.5	66.6	87.4	77.8	79.5	26.3	73.4	60.2	70.8	76.5	79.6	77.7	58.2	77.4	52.9	75.2	59.8	69.6
DeconvNet+CRF	92.9	87.8	41.9	80.6	63.9	67.3	**88.1**	78.4	81.3	25.9	73.7	61.2	72.0	77.0	79.9	78.7	59.5	78.3	**55.0**	75.2	61.5	70.5
EDeconvNet	92.9	88.4	39.7	79.0	63.0	67.7	87.1	**81.5**	84.4	27.8	76.1	61.2	78.0	79.3	83.1	79.3	58.0	82.5	52.3	80.1	64.0	71.7
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**
* WSSL [21]	93.2	85.3	36.2	**84.8**	61.2	67.5	84.7	81.4	81.0	**30.8**	73.8	53.8	77.5	76.5	82.3	**81.6**	56.3	78.9	52.3	76.6	63.3	70.4
* BoxSup [2]	**93.6**	86.4	35.5	79.7	**65.2**	65.2	84.3	78.5	83.7	30.5	76.2	**62.6**	**79.3**	76.1	82.1	81.3	57.0	78.2	**55.0**	72.5	**68.1**	71.0



Stress points

- **Two stage training**
  - Two stage training is not critical
    - DeconvNet can be trained well with single stage training
  - Two stage's performance is better
    - According to the research, validation accuracy of two stage training is higher than single stage training
    - Based on previous experiments, this margin could make big difference in mean IOU score
- **Batch Normalization**
  - Normalizing output of every layer to standard Gaussian distribution
  - Without Batch Normalization, DeconvNet stuch in local mnima
  - Maximum segmentation Accuaracy:
    - With Batch Normalization: 92.61 (0.18 loss)
    - W/O Batch Normalization: 62.41 (0.59 loss)



Conclusion

- **Effects of the research**
  - Practical to sharpen images that suffer from fast motion or jiggles during caturing
    - Effective method of enhancing the resolution and contrast of the optical micro-scope
    - Improves the resoulution and contrast of the fused image compared to content-based multiview fusion
    - For instance, fluorescent beads that are ideally collapsed to single intense points can be restored by deconvolution method
- **Future Direction**
  - Data augmentation: MSCOCO
  - Enhance performance on training set
    - Make network more flexible
  - Critical point
    - Dropout does not work in DevconNet