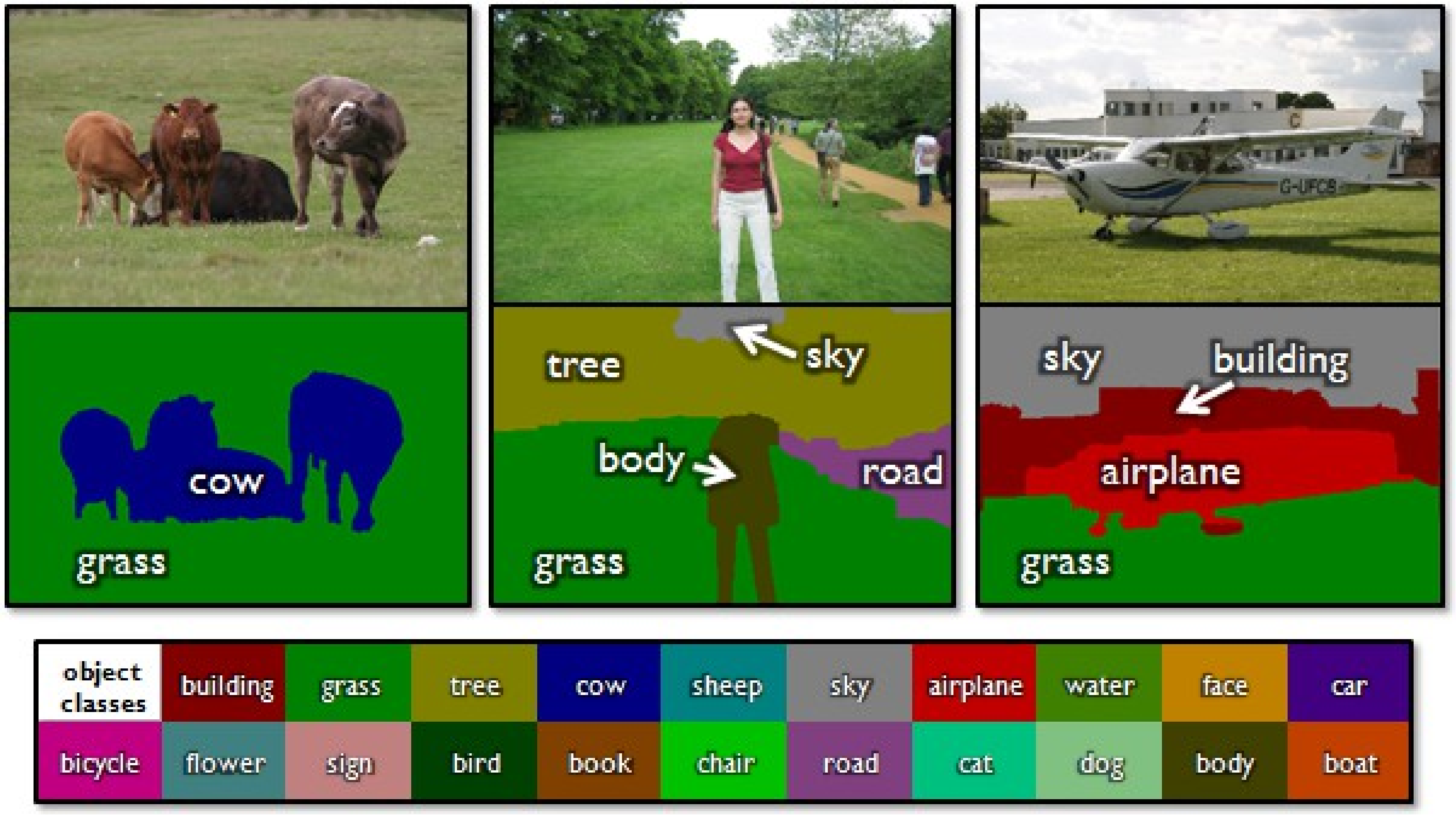
An Overview of Semantic Image Segmentation with Deep Learning

Simone Bonechi

Outline

- Semantic Image Segmentation
- Deep Network for Semantic Segmentation
  - FCN (Fully Convolutional Neural Network)
  - DeconvNet
  - PSPNet (Pyramid Scene Parsing Network)
- Work in progress…

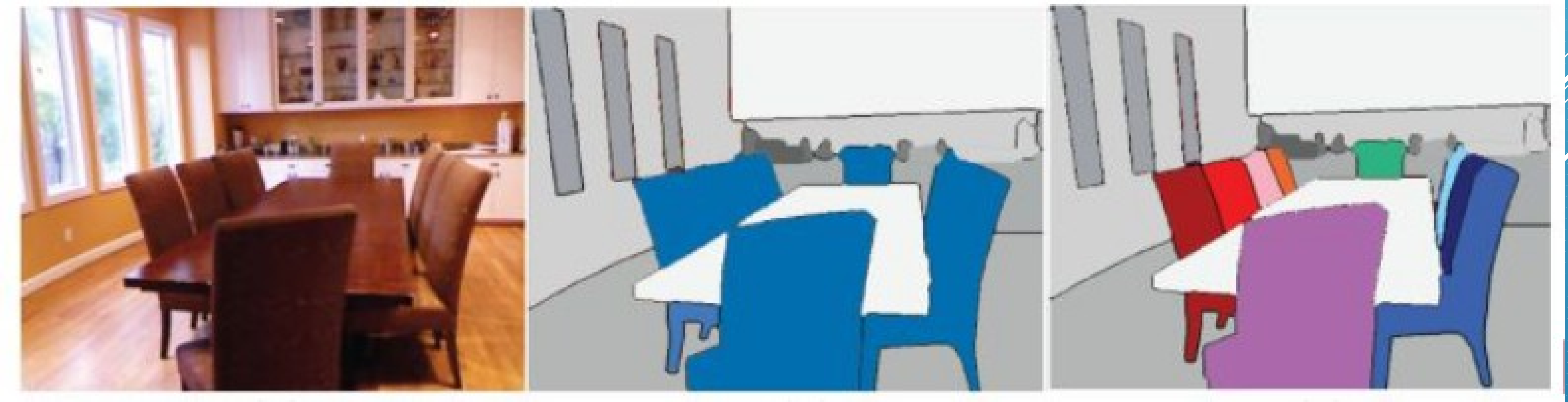
Semantic Image Segmentation



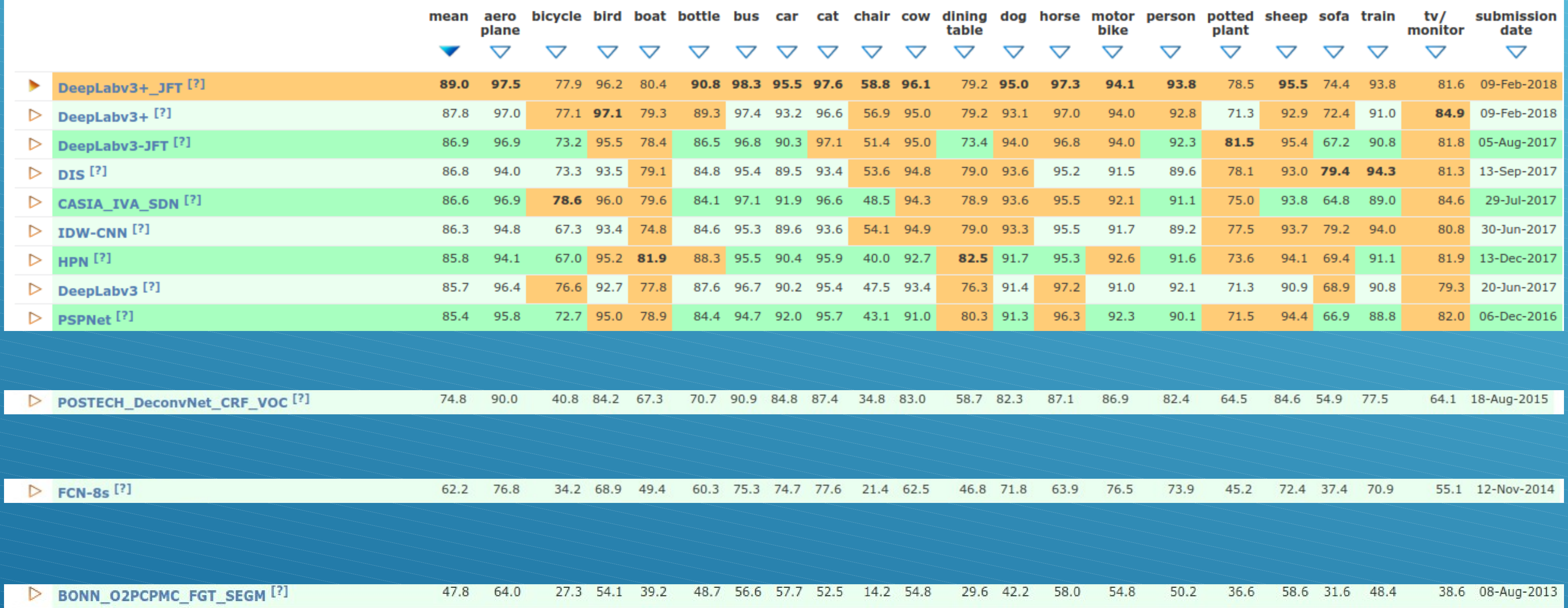
object classes	building	grass	tree	cow	sheep	sky	airplane	water	face	car
bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat

Instance-Level Segmentation

- Its main purpose is to identify objects of the same class and split them into different instances

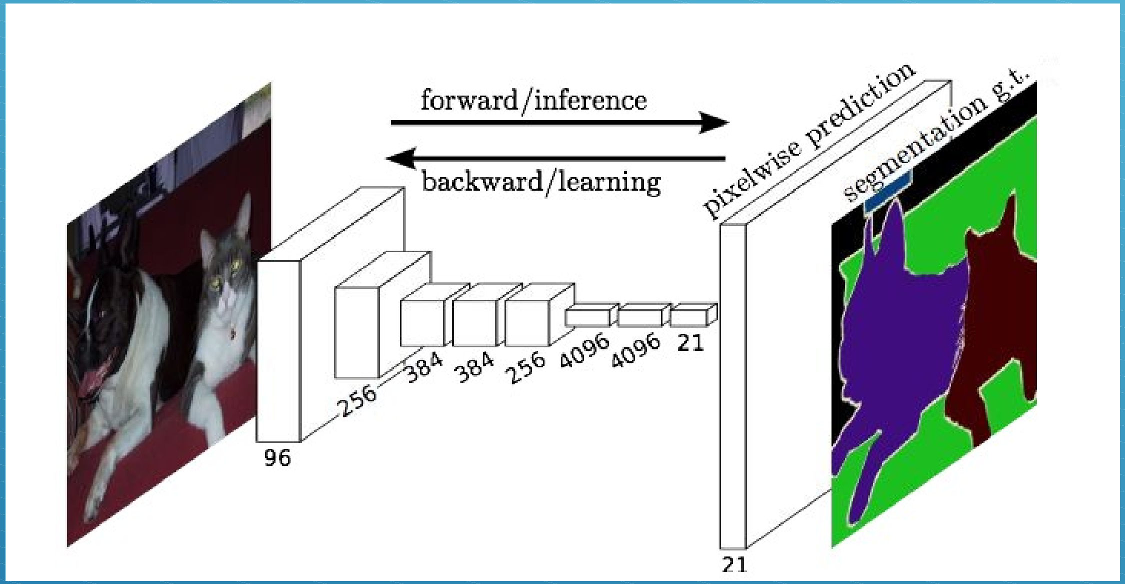


Results on PascalVoc 2012



		mean	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor	submission date
▶	DeepLabv3+_JFT [?]	**89.0**	**97.5**	77.9	96.2	80.4	**90.8**	**98.3**	**95.5**	**97.6**	**58.8**	**96.1**	79.2	**95.0**	**97.3**	**94.1**	**93.8**	78.5	**95.5**	74.4	93.8	81.6	09-Feb-2018
▷	DeepLabv3+ [?]	87.8	97.0	77.1	**97.1**	79.3	89.3	97.4	93.2	96.6	56.9	95.0	79.2	93.1	97.0	94.0	92.8	71.3	92.9	72.4	91.0	**84.9**	09-Feb-2018
▷	DeepLabv3-JFT [?]	86.9	96.9	73.2	95.5	78.4	86.5	96.8	90.3	97.1	51.4	95.0	73.4	94.0	96.8	94.0	92.3	**81.5**	95.4	67.2	90.8	81.8	05-Aug-2017
▷	DIS [?]	86.8	94.0	73.3	93.5	79.1	84.8	95.4	89.5	93.4	53.6	94.8	79.0	93.6	95.2	91.5	89.6	78.1	93.0	**79.4**	**94.3**	81.3	13-Sep-2017
▷	CASIA_IVA_SDN [?]	86.6	96.9	**78.6**	96.0	79.6	84.1	97.1	91.9	96.6	48.5	94.3	78.9	93.6	95.5	92.1	91.1	75.0	93.8	64.8	89.0	84.6	29-Jul-2017
▷	IDW-CNN [?]	86.3	94.8	67.3	93.4	74.8	84.6	95.3	89.6	93.6	54.1	94.9	79.0	93.3	95.5	91.7	89.2	77.5	93.7	79.2	94.0	80.8	30-Jun-2017
▷	HPN [?]	85.8	94.1	67.0	95.2	**81.9**	88.3	95.5	90.4	95.9	40.0	92.7	**82.5**	91.7	95.3	92.6	91.6	73.6	94.1	69.4	91.1	81.9	13-Dec-2017
▷	DeepLabv3 [?]	85.7	96.4	76.6	92.7	77.8	87.6	96.7	90.2	95.4	47.5	93.4	76.3	91.4	97.2	91.0	92.1	71.3	90.9	68.9	90.8	79.3	20-Jun-2017
▷	PSPNet [?]	85.4	95.8	72.7	95.0	78.9	84.4	94.7	92.0	95.7	43.1	91.0	80.3	91.3	96.3	92.3	90.1	71.5	94.4	66.9	88.8	82.0	06-Dec-2016
▷	POSTECH_DeconvNet_CRF_VOC [?]	74.8	90.0	40.8	84.2	67.3	70.7	90.9	84.8	87.4	34.8	83.0	58.7	82.3	87.1	86.9	82.4	64.5	84.6	54.9	77.5	64.1	18-Aug-2015
▷	FCN-8s [?]	62.2	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	12-Nov-2014
▷	BONN_O2PCPMC_FGT_SEGM [?]	47.8	64.0	27.3	54.1	39.2	48.7	56.6	57.7	52.5	14.2	54.8	29.6	42.2	58.0	54.8	50.2	36.6	58.6	31.6	48.4	38.6	08-Aug-2013

Fully Convolutional Neural Network (FCN)



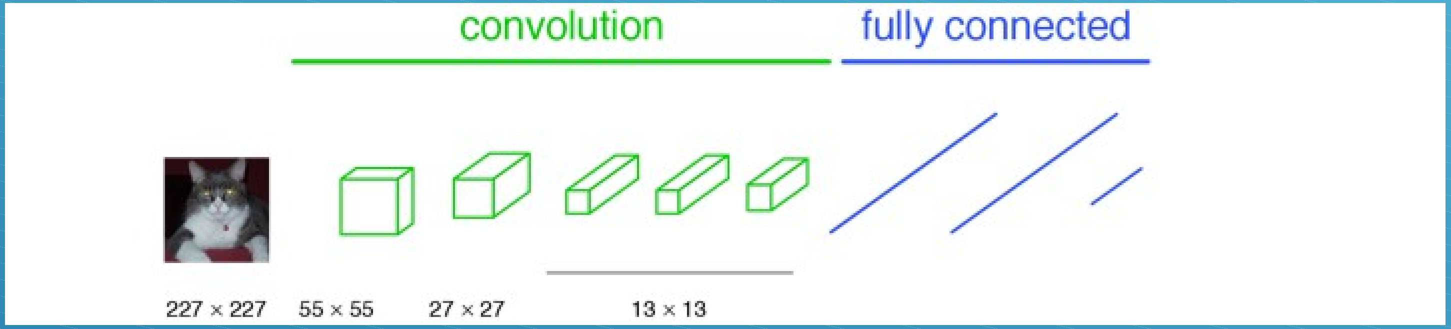
Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3431-3440).

FCN Overview

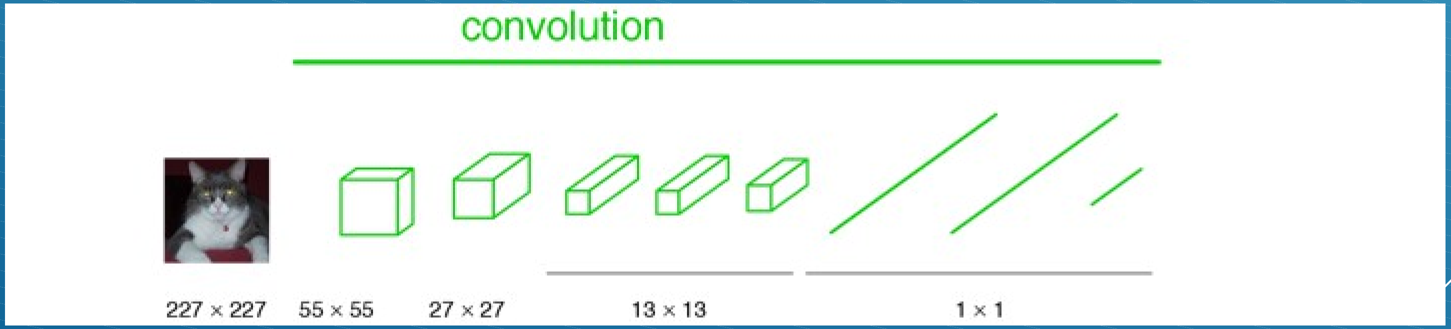
- Tested with AlexNet, VGG and GoogLeNet
- Reinterpret standard classification convnets as “Fully convolutional” networks (FCN) for semantic segmentation
- Combine information from different layers for segmentation

Replace FC with Convolutions

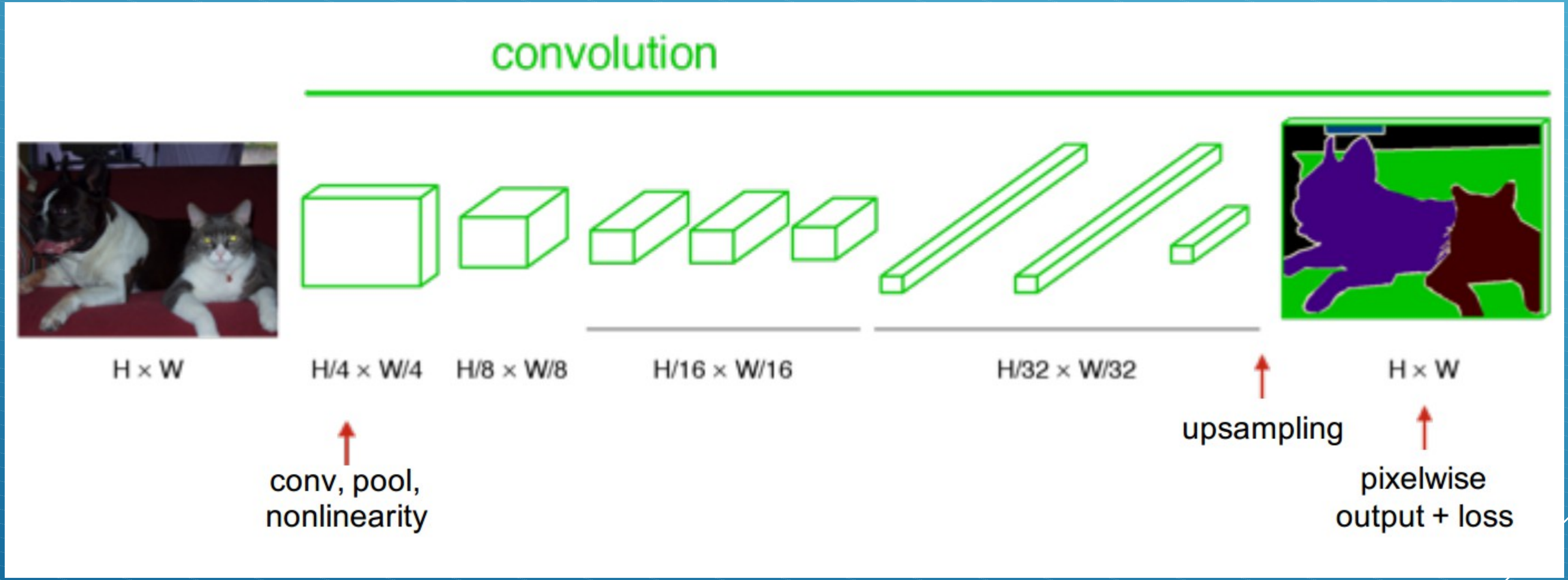
A classification network



Becoming fully convolutional

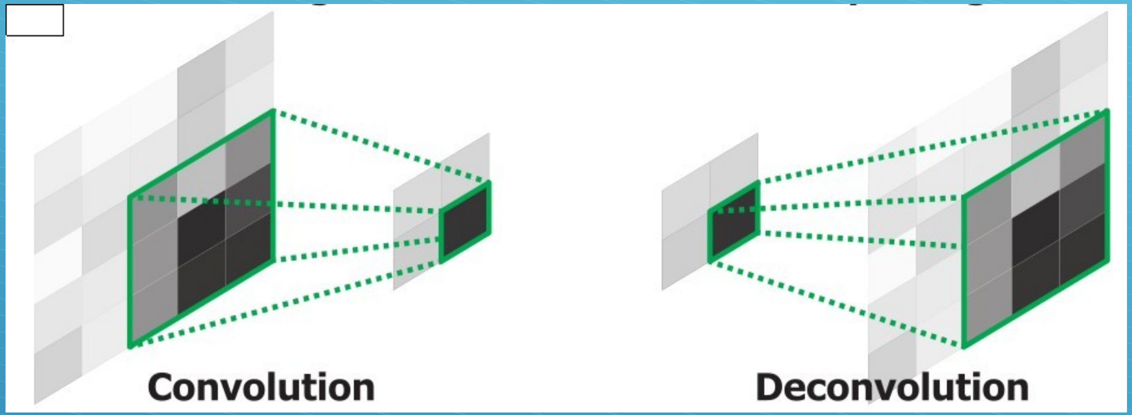


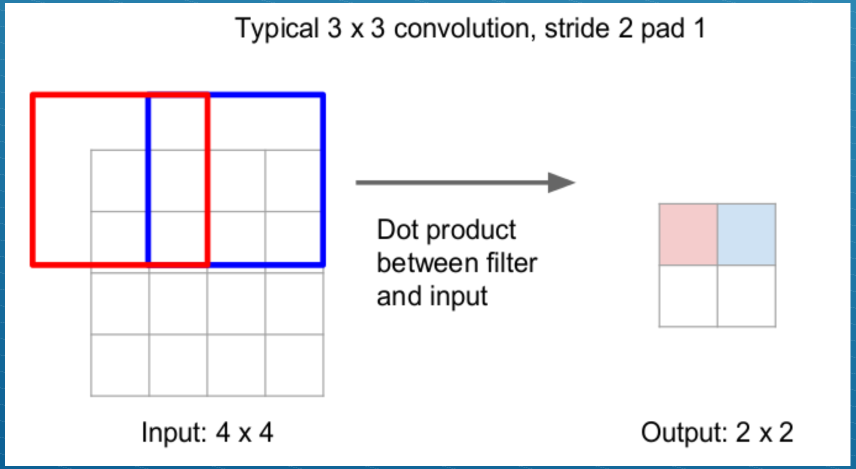
Upsampling the output

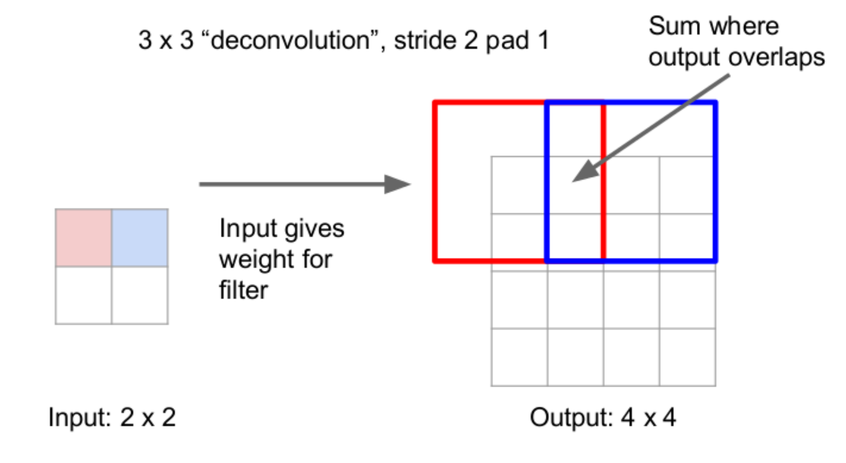


Convolution & Deconvolution

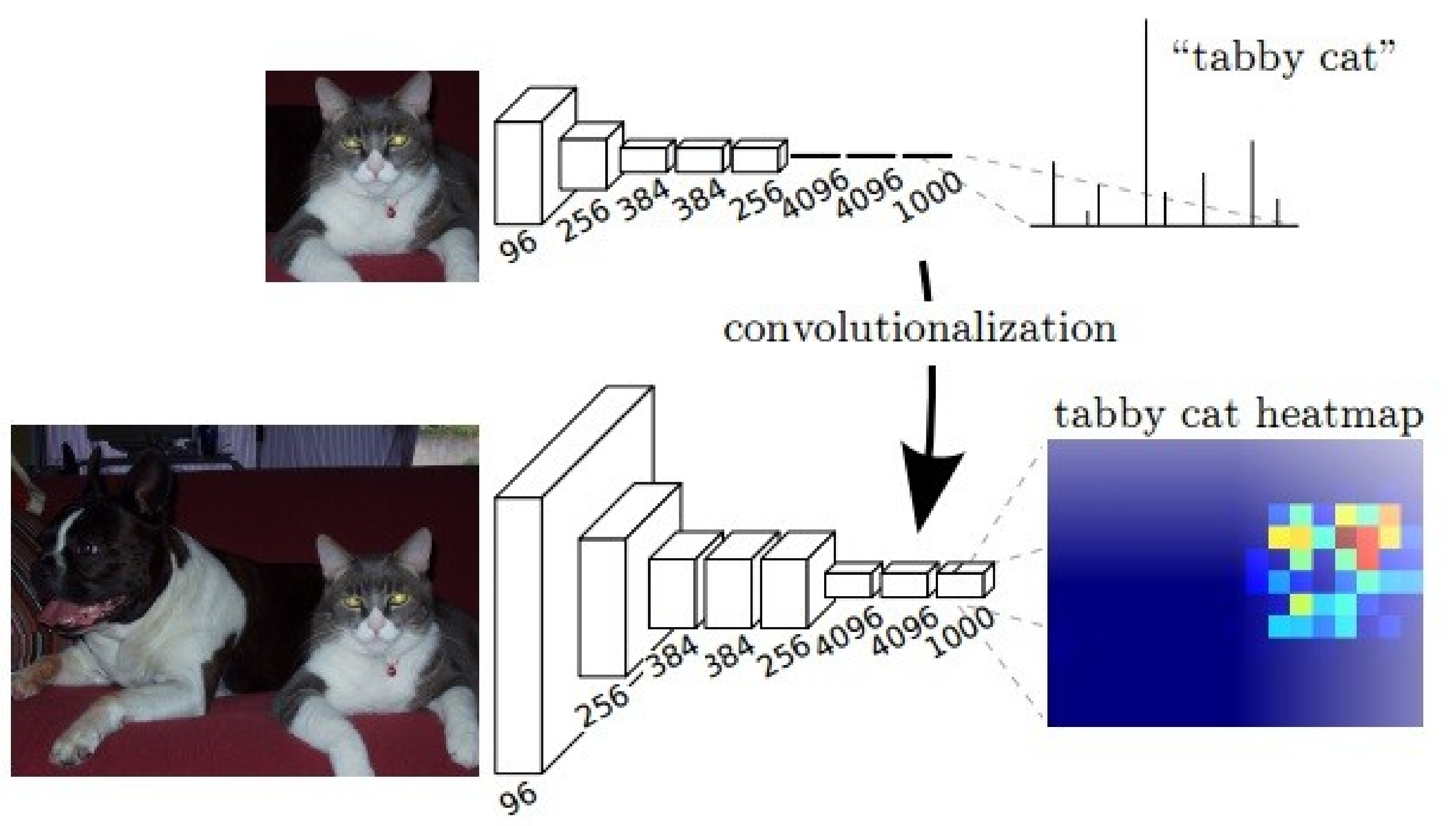
- Deconvolution
- Transposed convolution
- Fractionally strided convolution
- Backward strided convolution
- Upconvolution
- …..




Typical 3 x 3 convolution, stride 2 pad 1
Dot product between filter and input
Input: 4 x 4
Output: 2 x 2

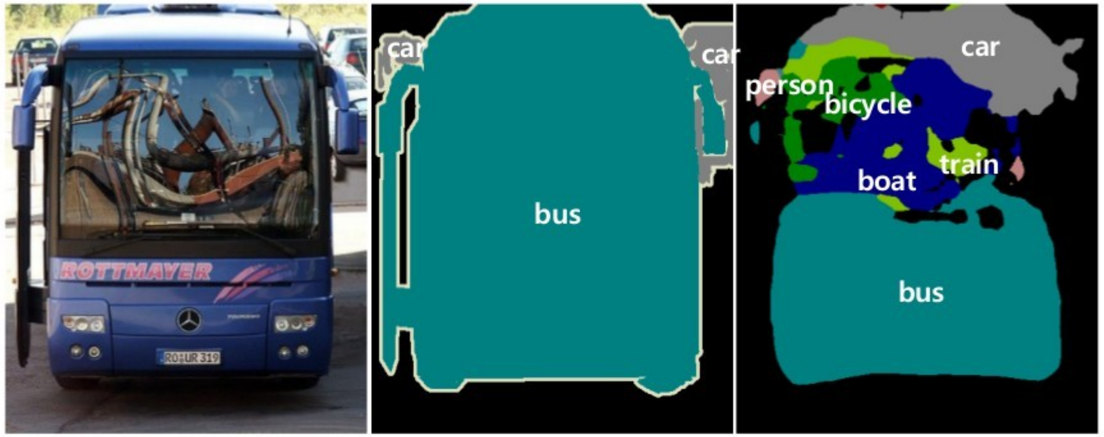

3 x 3 “deconvolution”, stride 2 pad 1
Sum where output overlaps
Input gives weight for filter
Input: 2 x 2
Output: 4 x 4

Upsampling the output

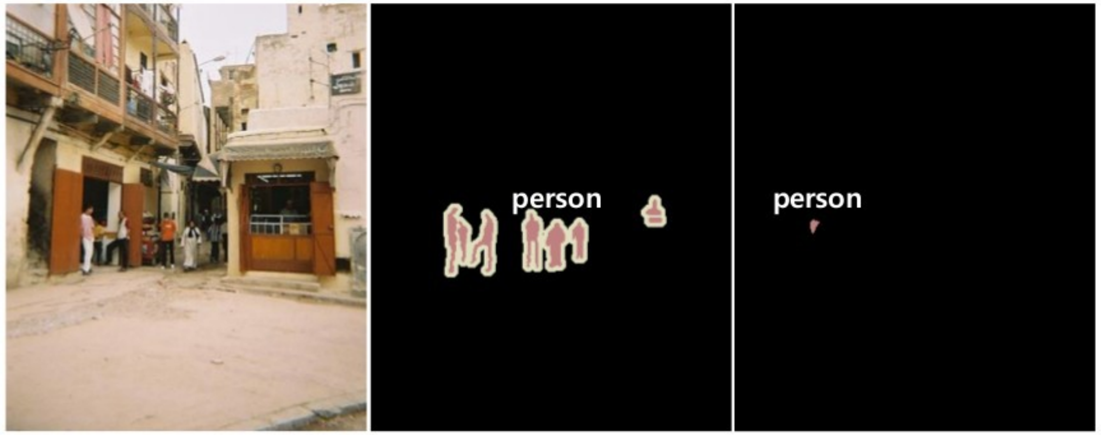


FCN Limitations

- Fixed-size receptive field
  - FCN has fixed-size receptive field; objects substantially larger or smaller than the receptive field may be fragmented or mislabeled
  - Label map is so small, tend to forget detail structures of object

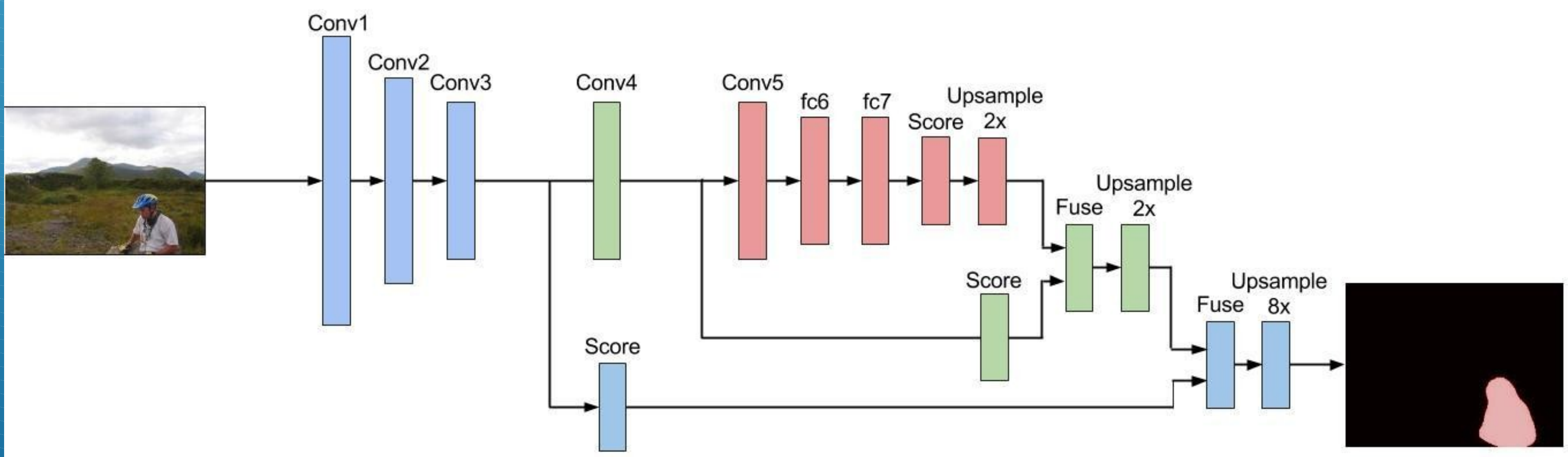


(a) Inconsistent labels due to large object size



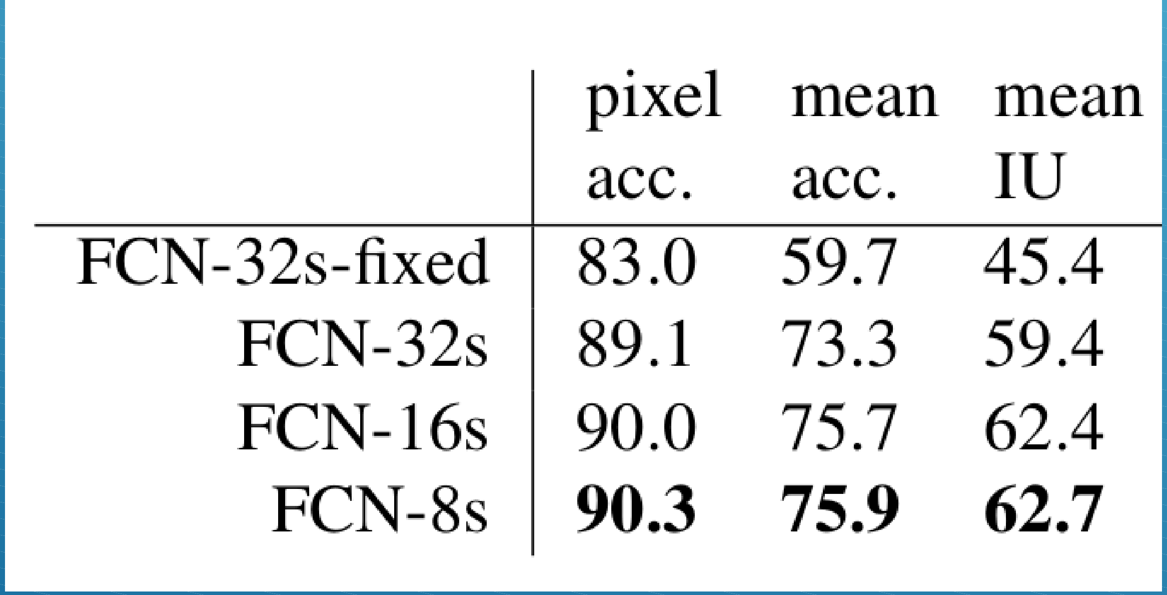
(b) Missing labels due to small object size

FCN skip architecture



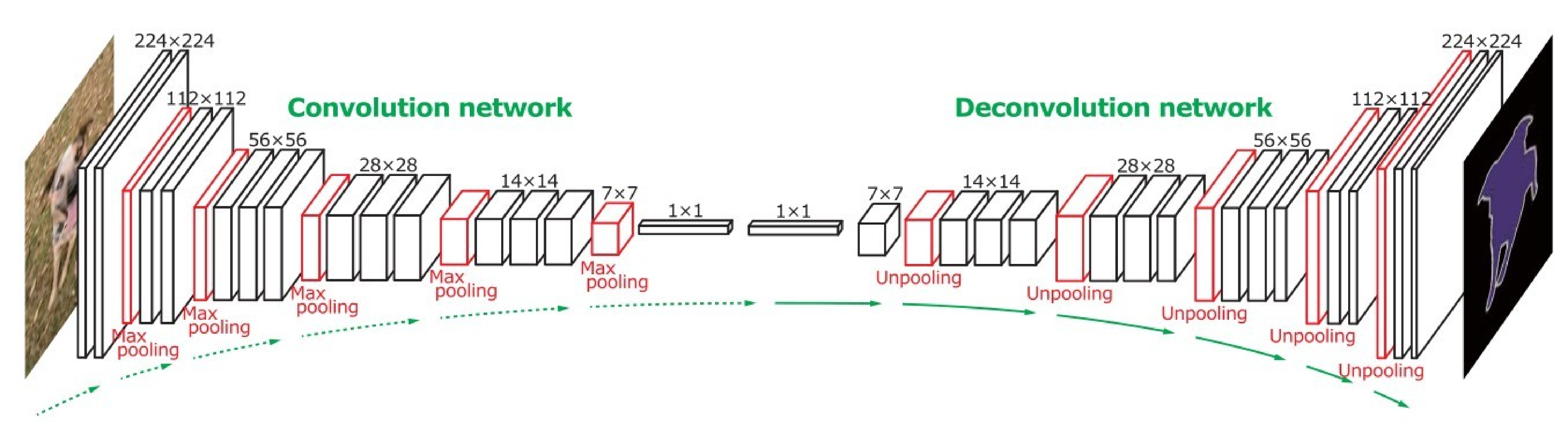
FCN Results

- Results on PascalVOC 2012



	pixel acc.	mean acc.	mean IU
FCN-32s-fixed	83.0	59.7	45.4
FCN-32s	89.1	73.3	59.4
FCN-16s	90.0	75.7	62.4
FCN-8s	**90.3**	**75.9**	**62.7**

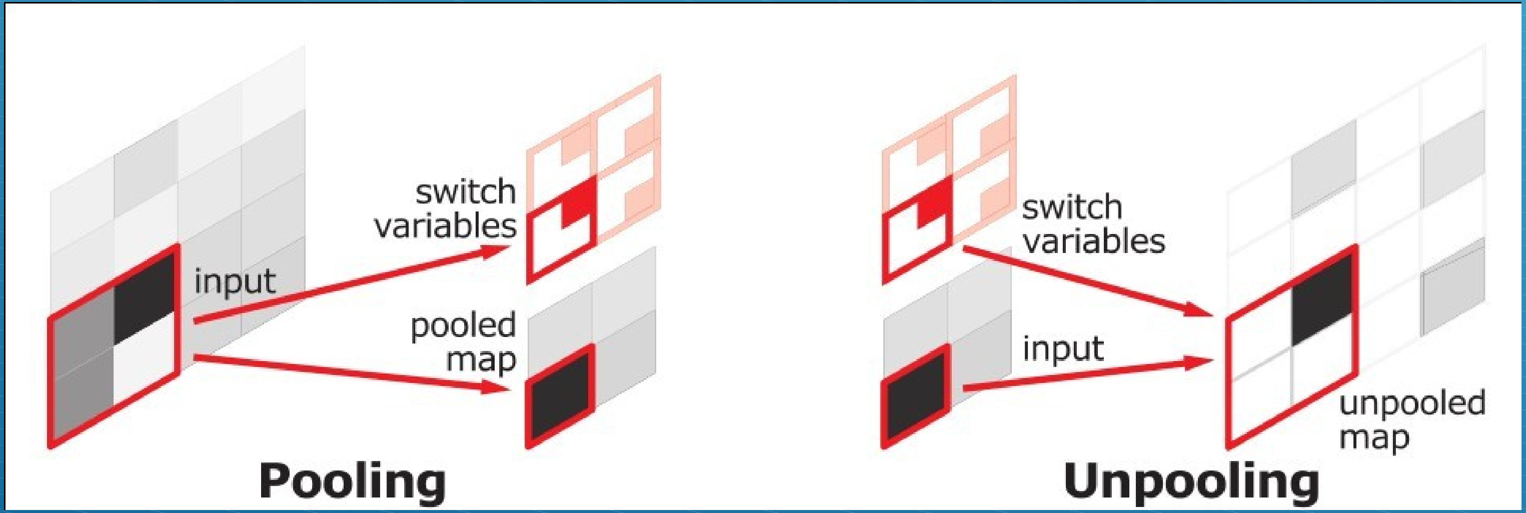
DeconvNet



Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1520-1528).

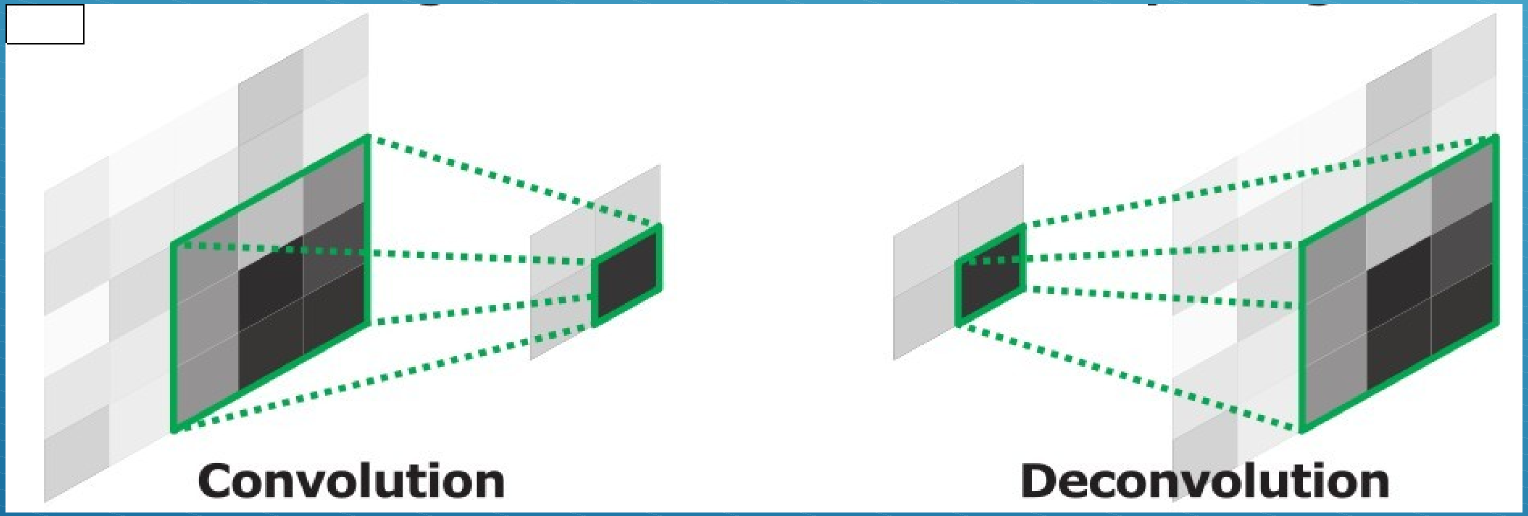
Pooling & Unpooling

- Unpooling
  - Retrieve structure of original activation map
  - Activation size is preserved, but still sparse

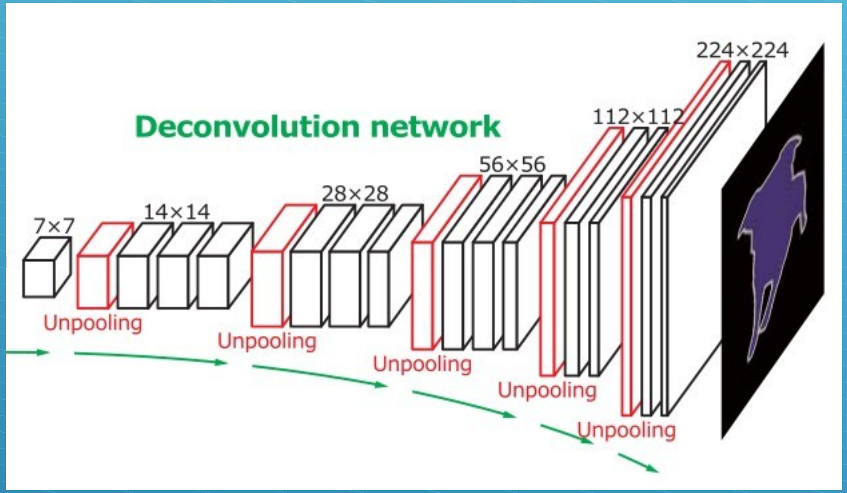


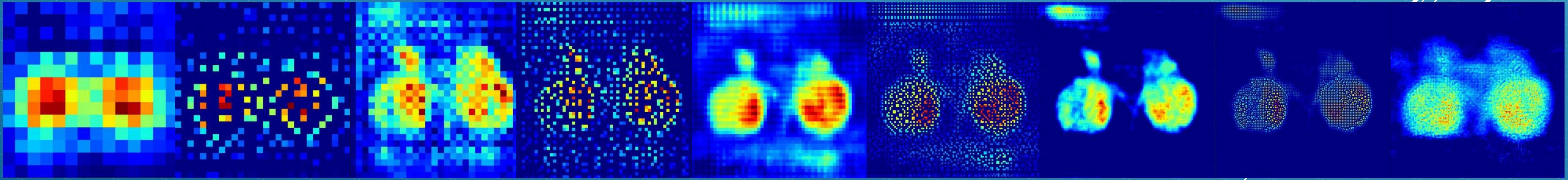
Convolution & Deconvolution

- Deconvolution
  - Densify sparse activation map

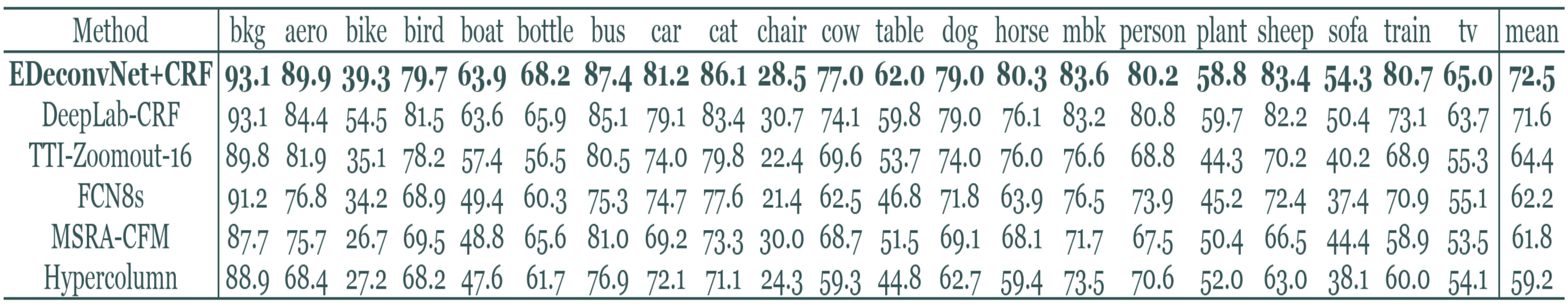


Visualization of activations



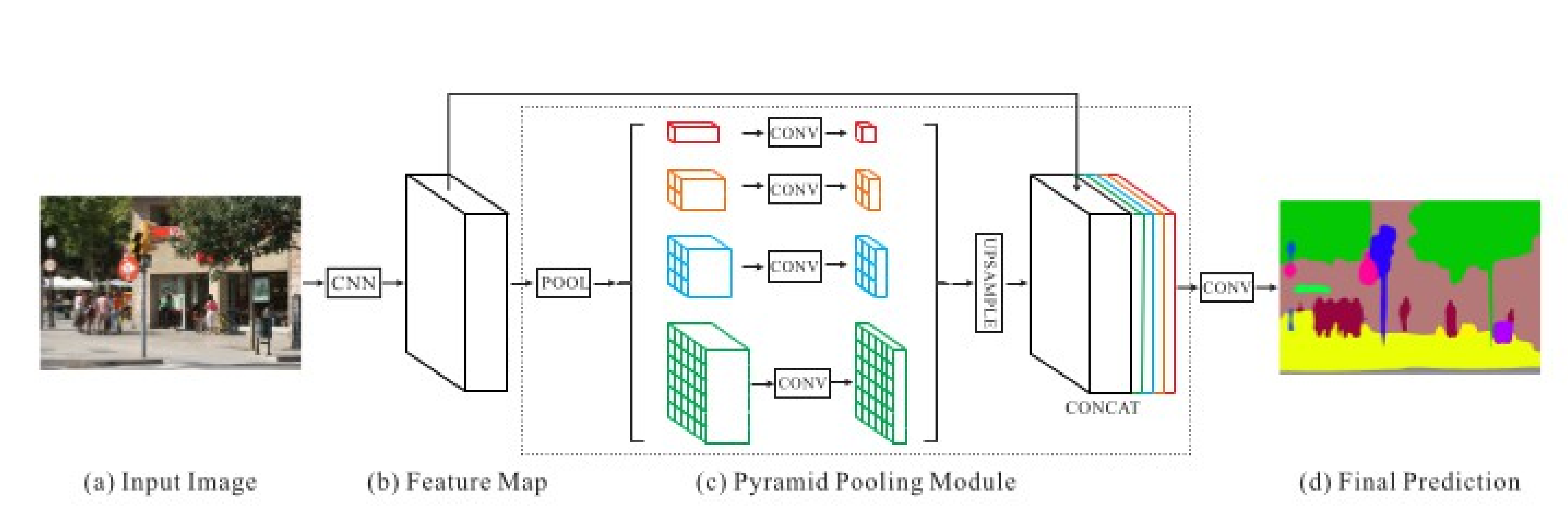


Results - Comparisons



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

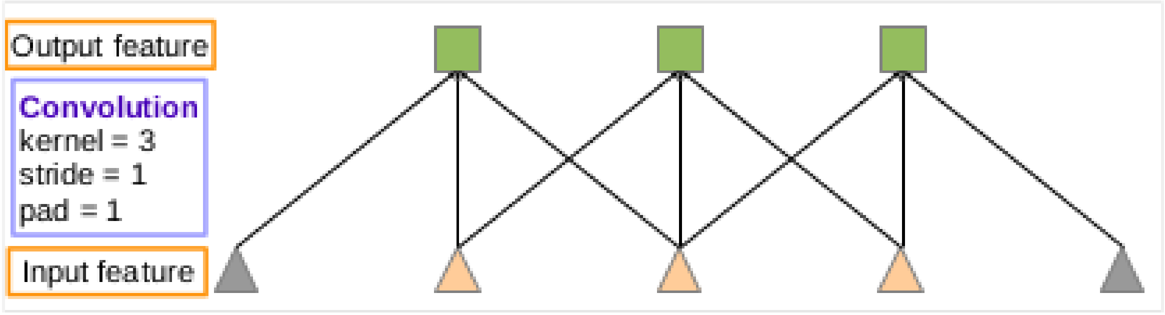
PSP-net



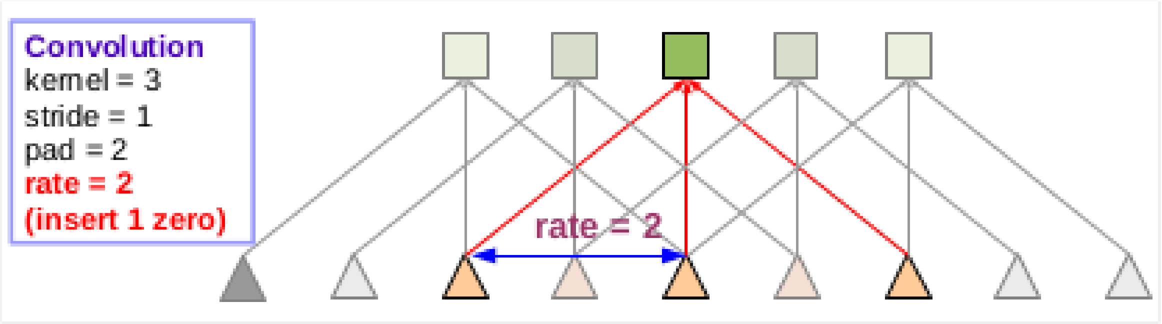
Zhao, Hengshuang, et al. "Pyramid scene parsing network." IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). 2017.

Atrous Convolution

- Upsample with atrous convolution to compute feature densely



(a) Sparse feature extraction



(b) Dense feature extraction

PSPNet Results