Deep Learning for Semantic Segmentation

Fisher Yu i@yf.io
UC Berkeley

Semantic Segmentation

PASCAL data

How Does Human Do it?

- Studied extensively by the Gestalt psychologists
  - Wertheimer, 1938, *Laws of organization in perceptual forms*
- Important factors
  - Similarity, proximity, continuity, symmetry, parallelism, closure and familiarity.

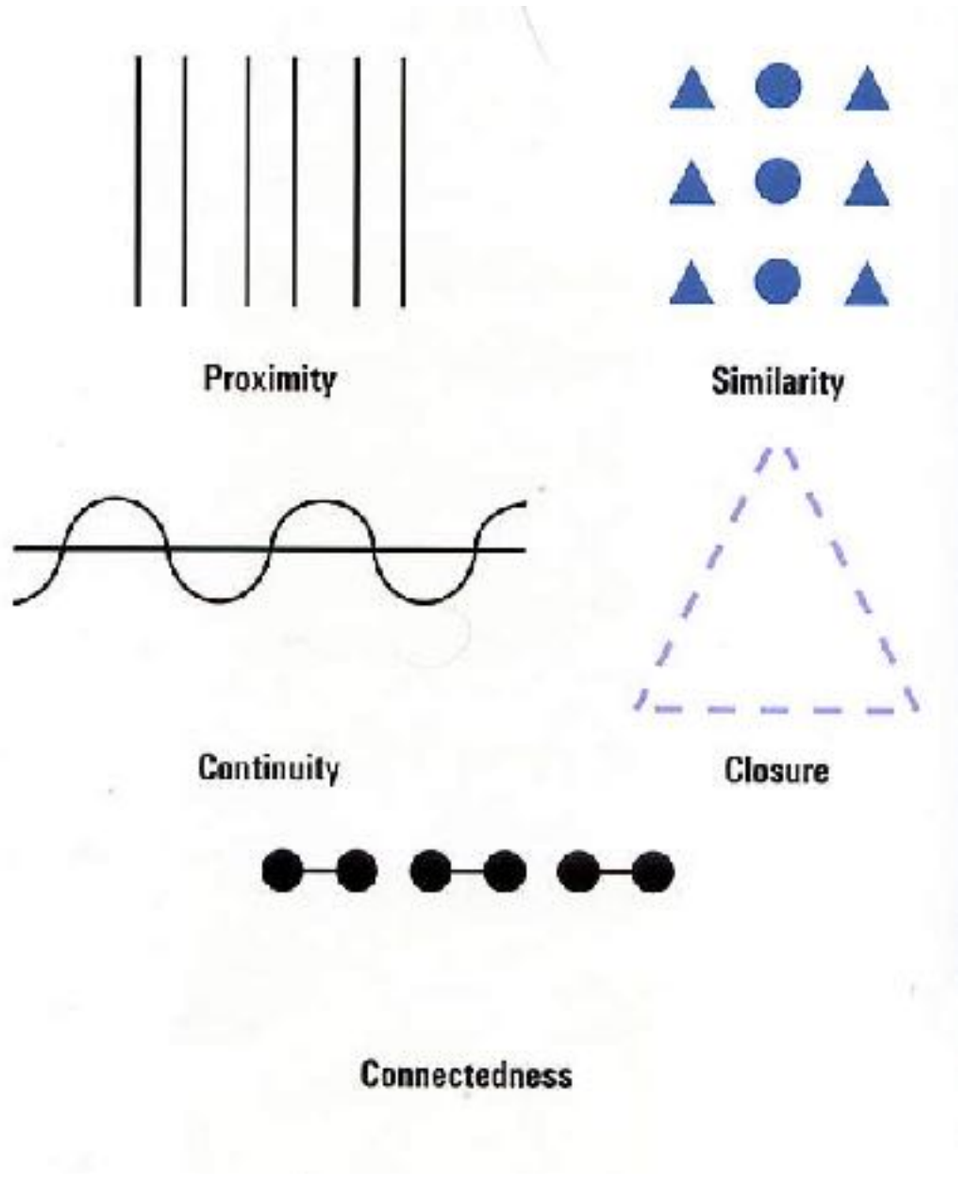
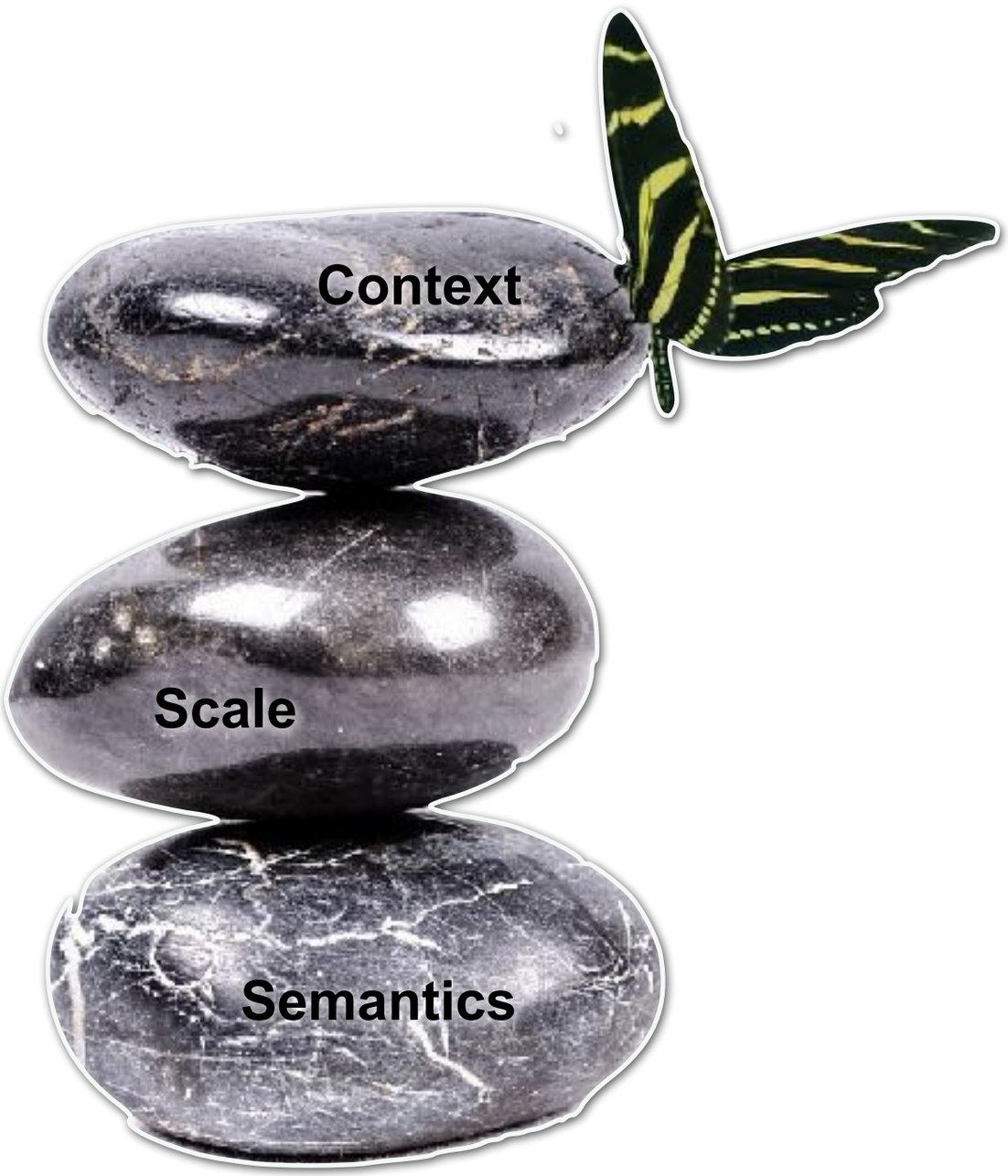
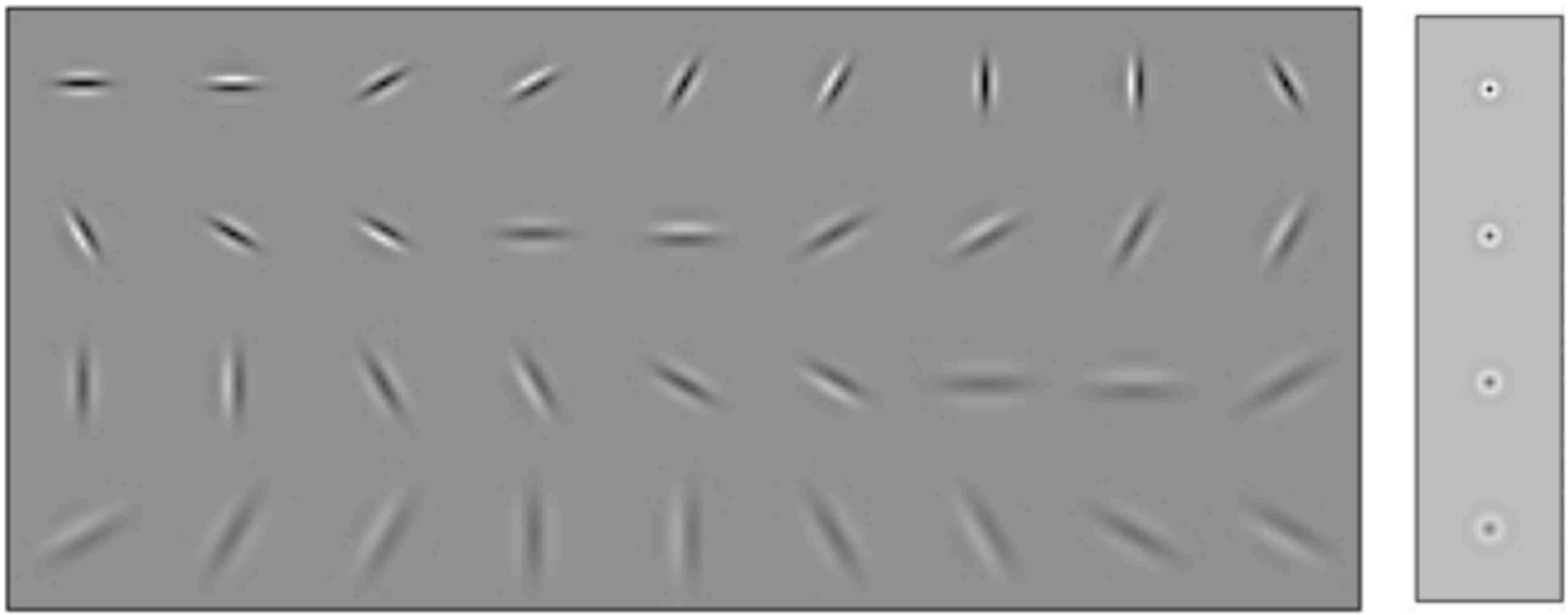


figure @ http://www.rhsmpsychology.com/

How Can Computer Do it?



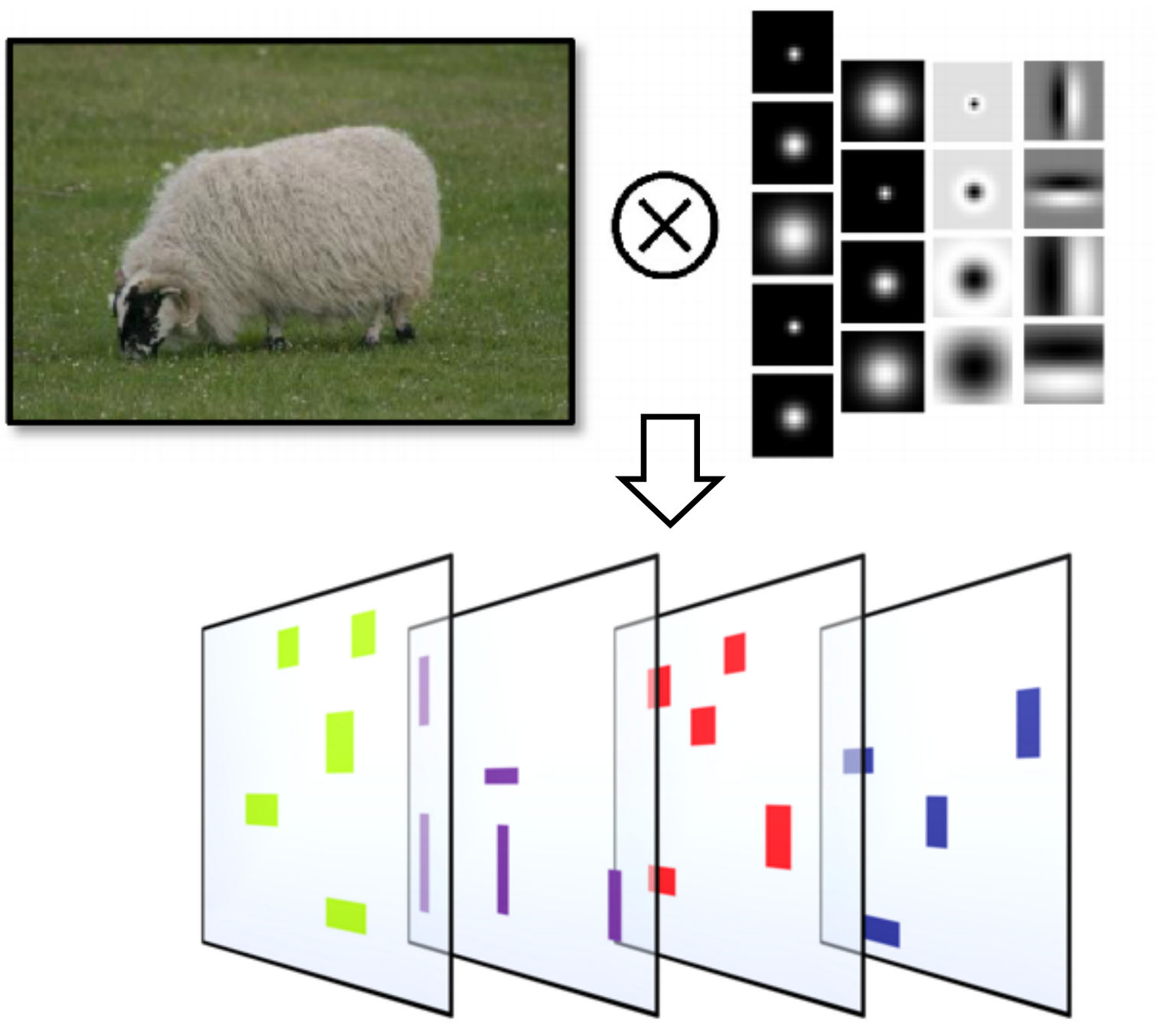
Extract features



Filter bank

Malik, et al. 2001, Contour and Texture Analysis for Image Segmentation

Extract features

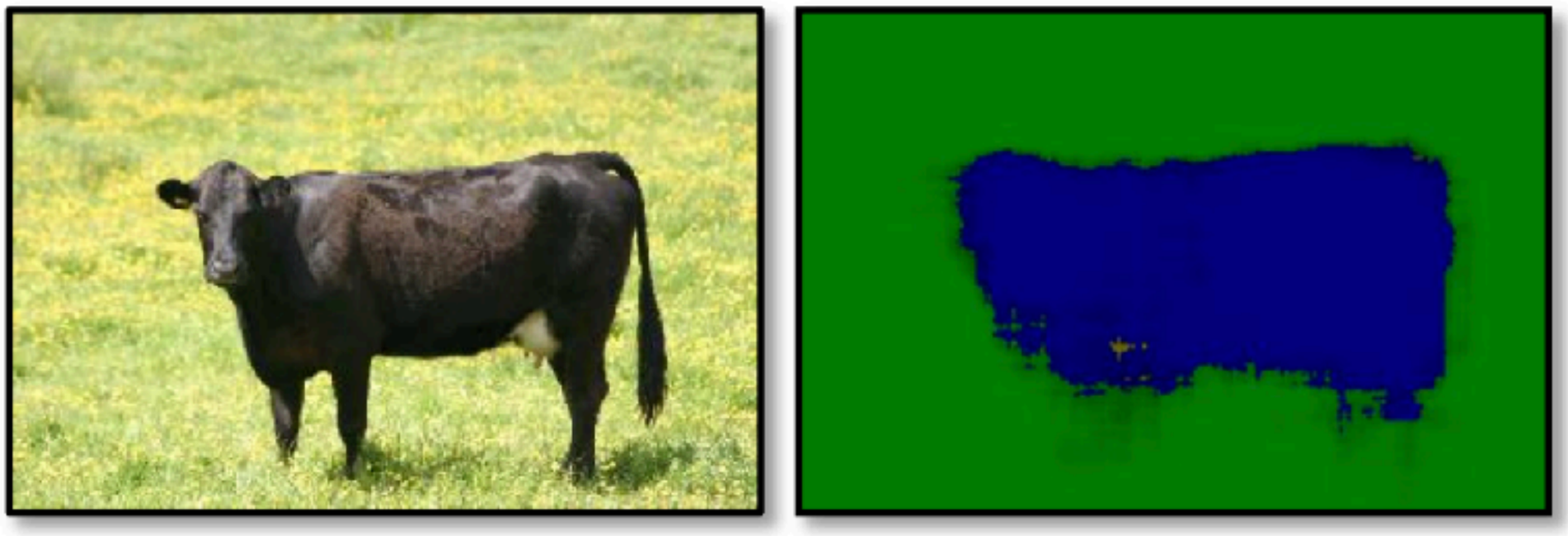


Hypercolumn Transform

Malik, et al. 2001, Contour and Texture Analysis for Image Segmentation

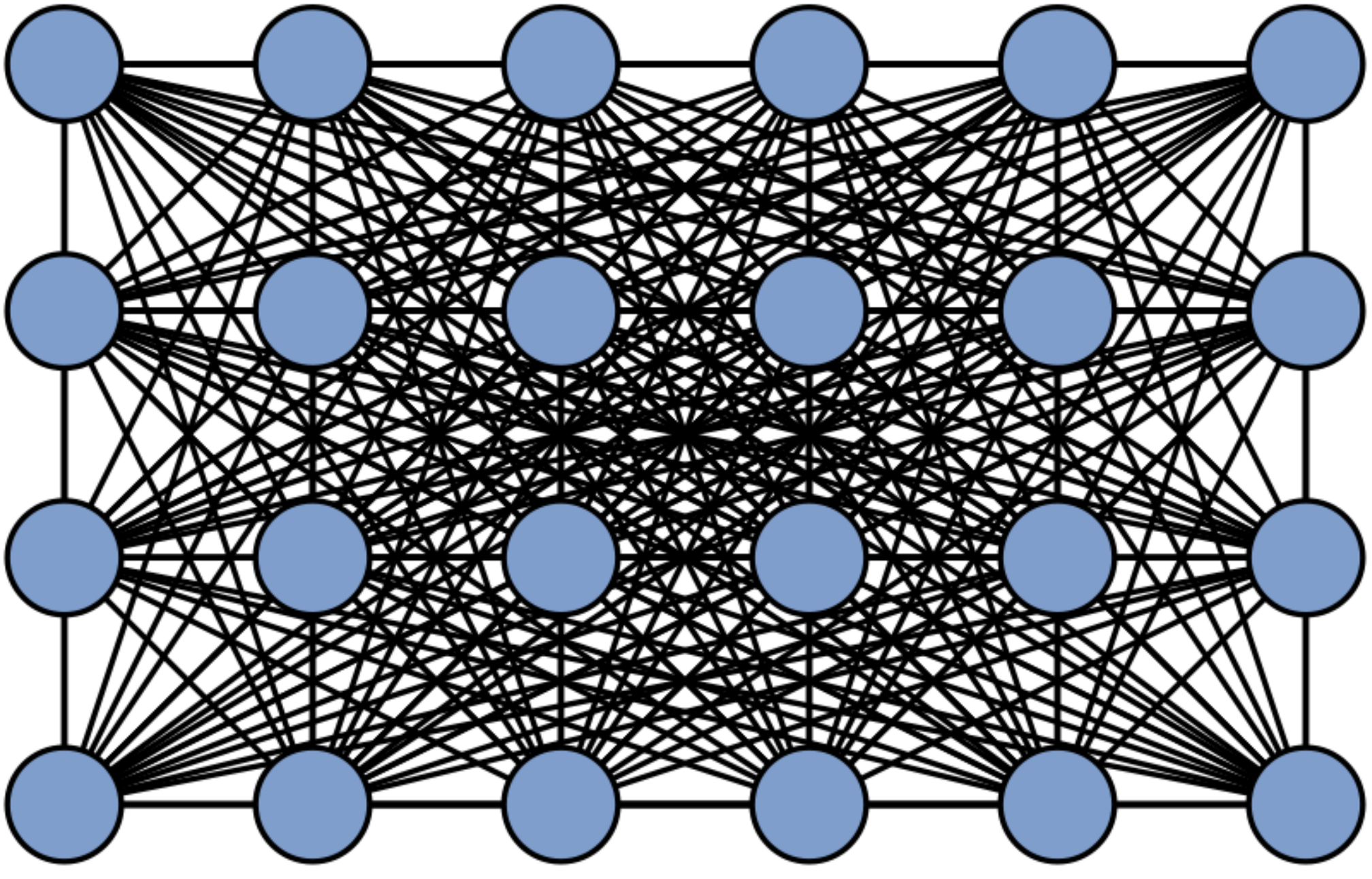
Pixel-wise Classifier

- Joint boost learner



Shotton, et al. 2007, TextonBoost for Image Understanding

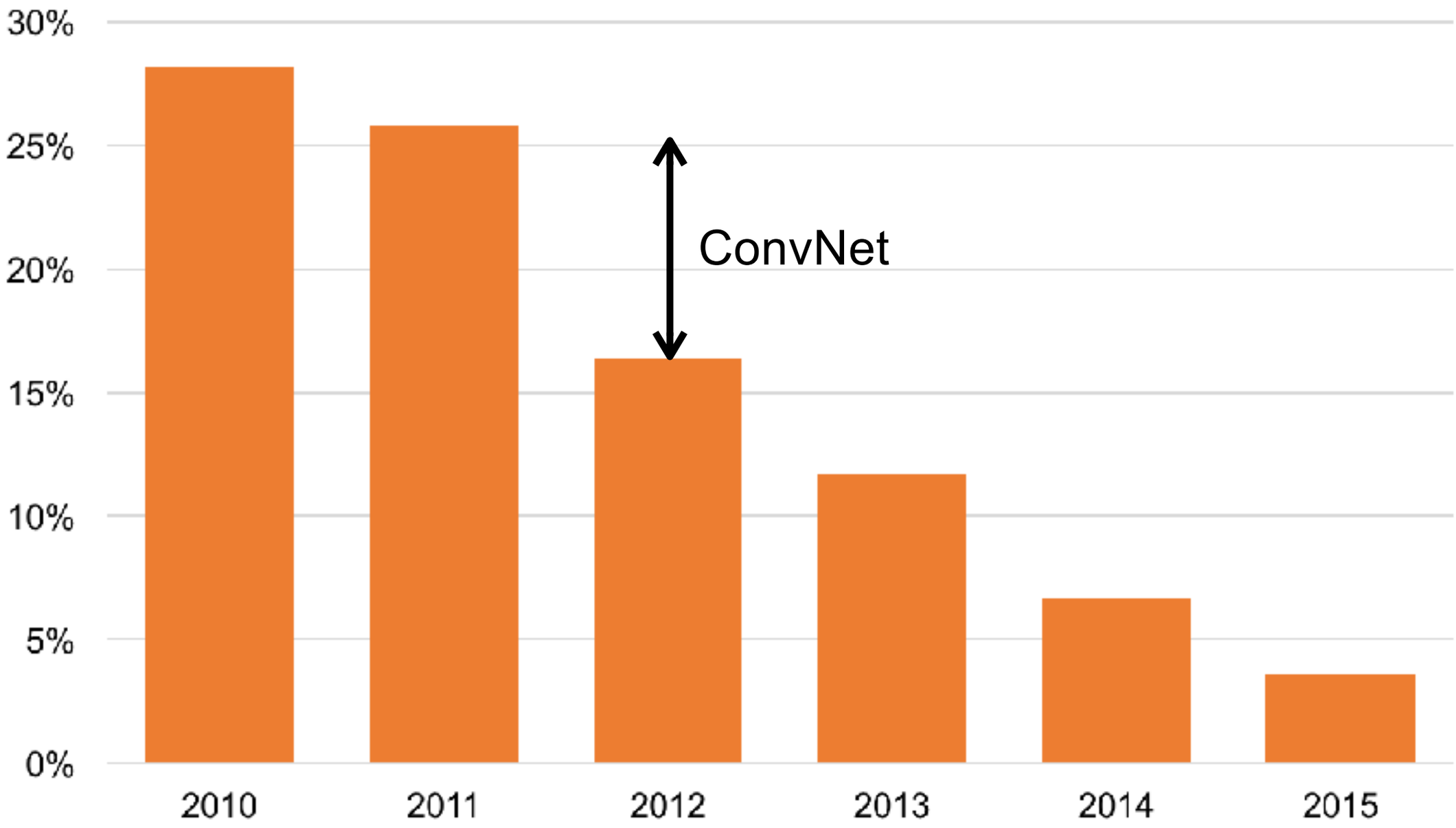
Spatical Consistency



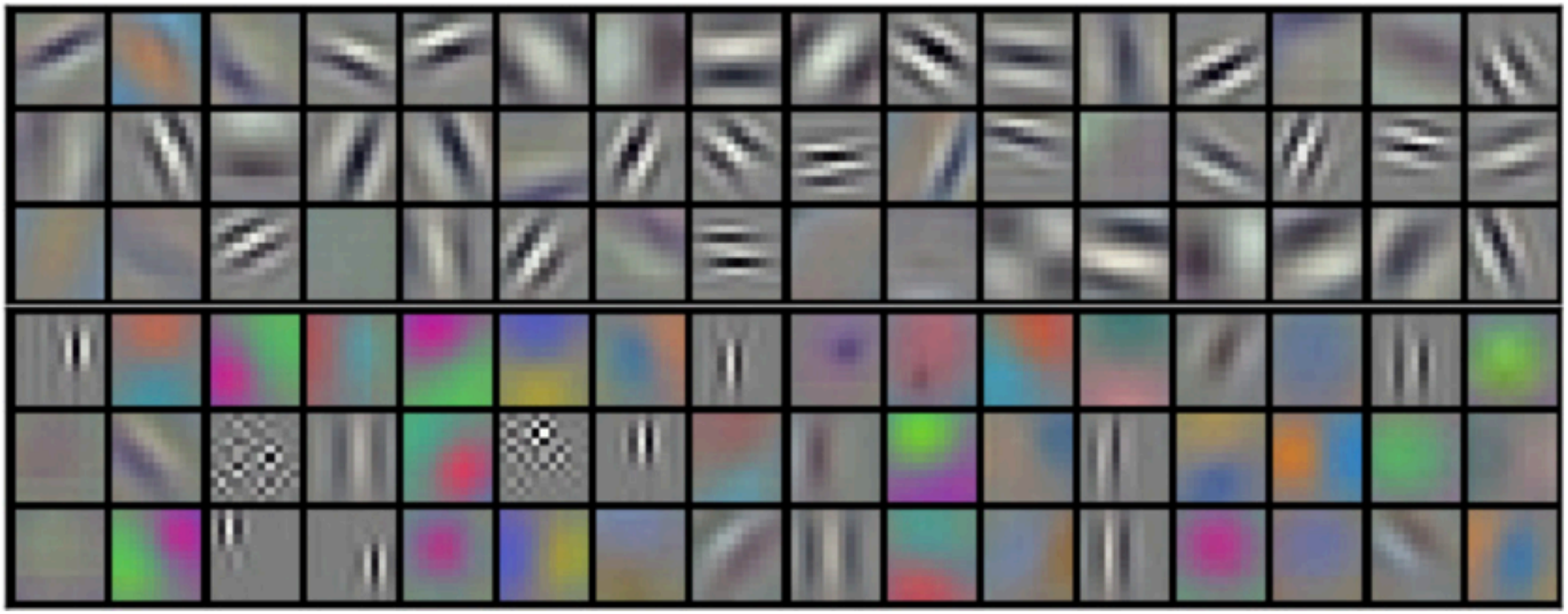
Spatical Consistency

Image Pixel-wise CRF

Deep Learning Solution



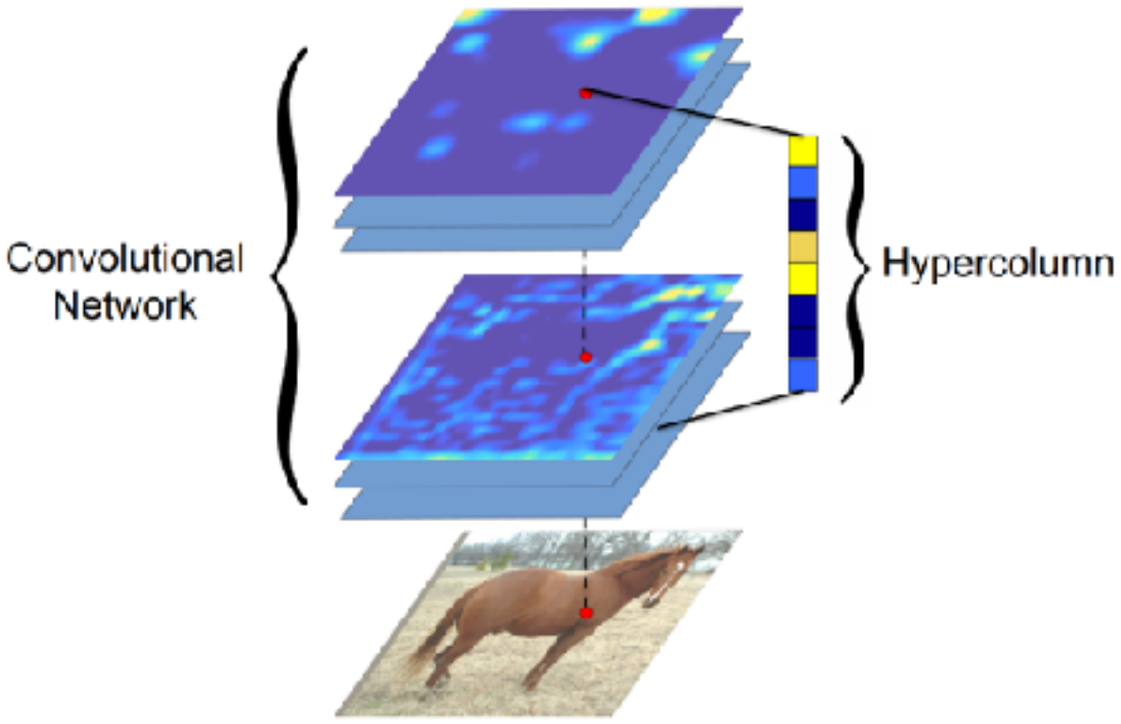
Deep Learning Solution

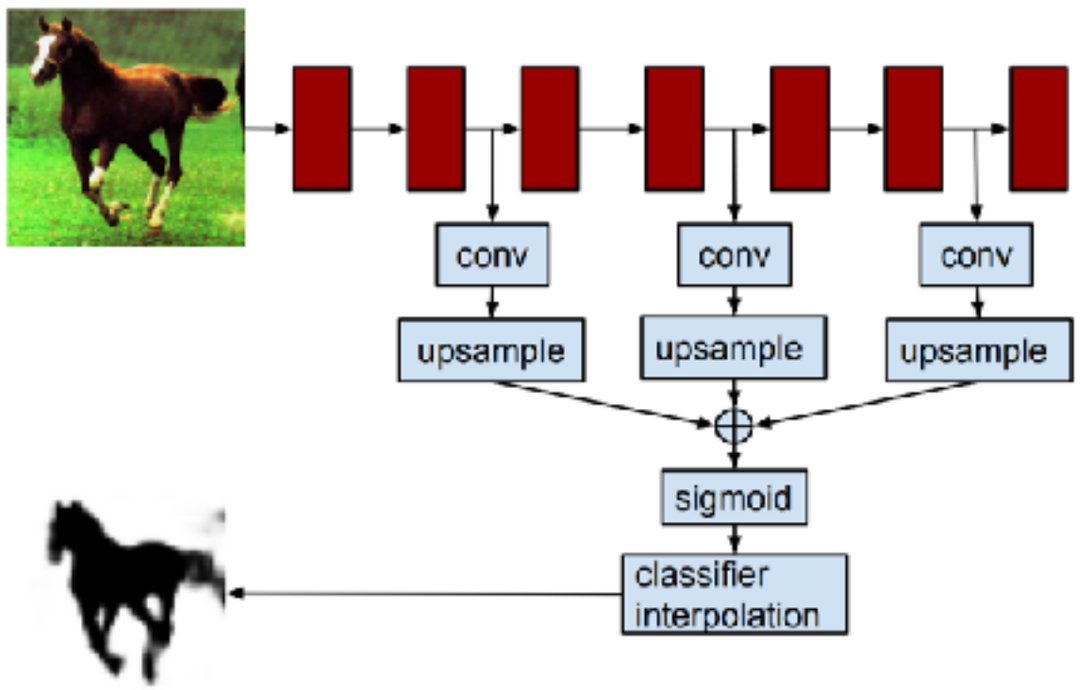


Filters of the first layer in AlexNet

Krizhevsky, et al. 2012, ImageNet Classification with Deep Convolutional Neural Networks

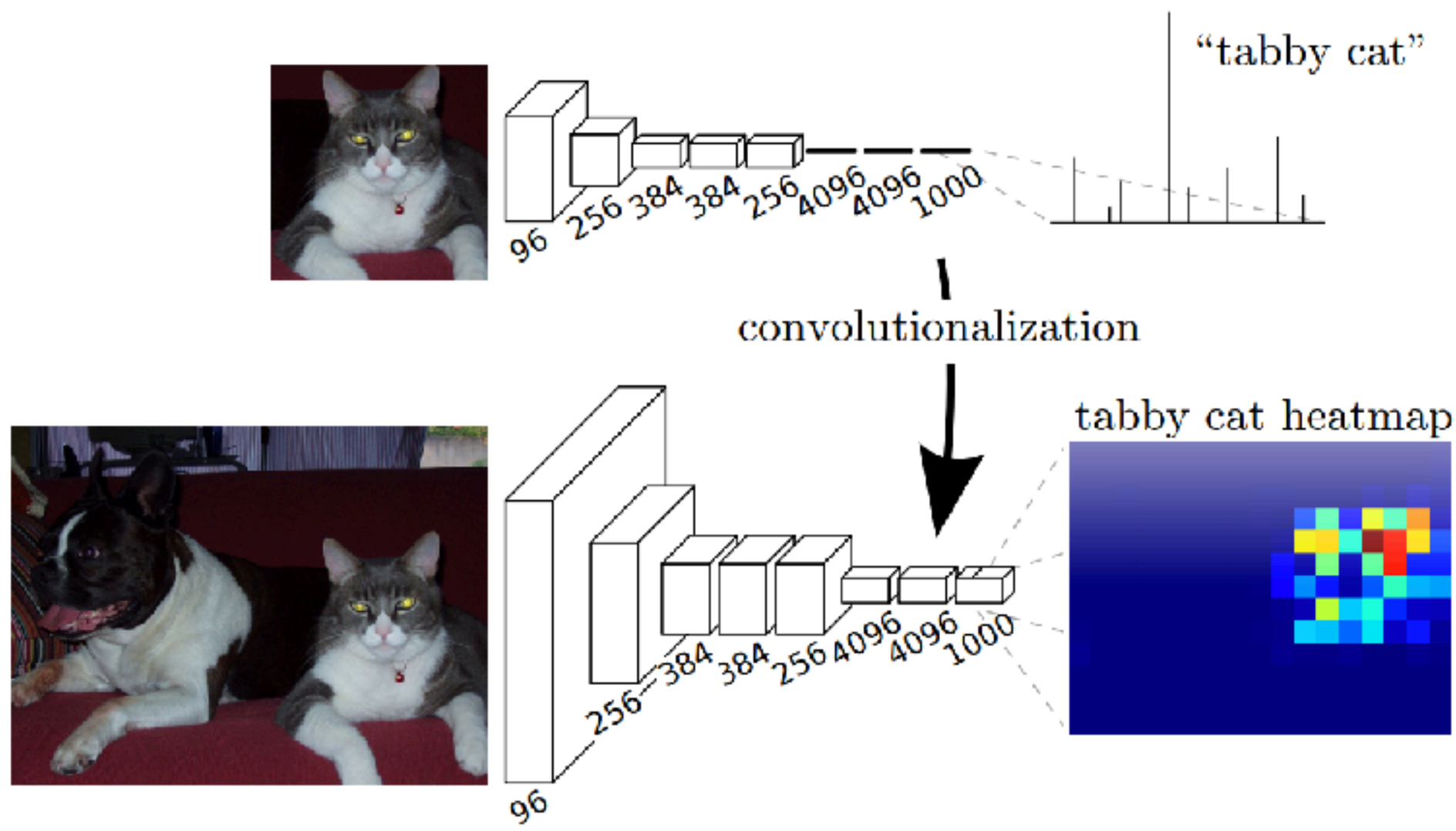
Deep Learning Solution



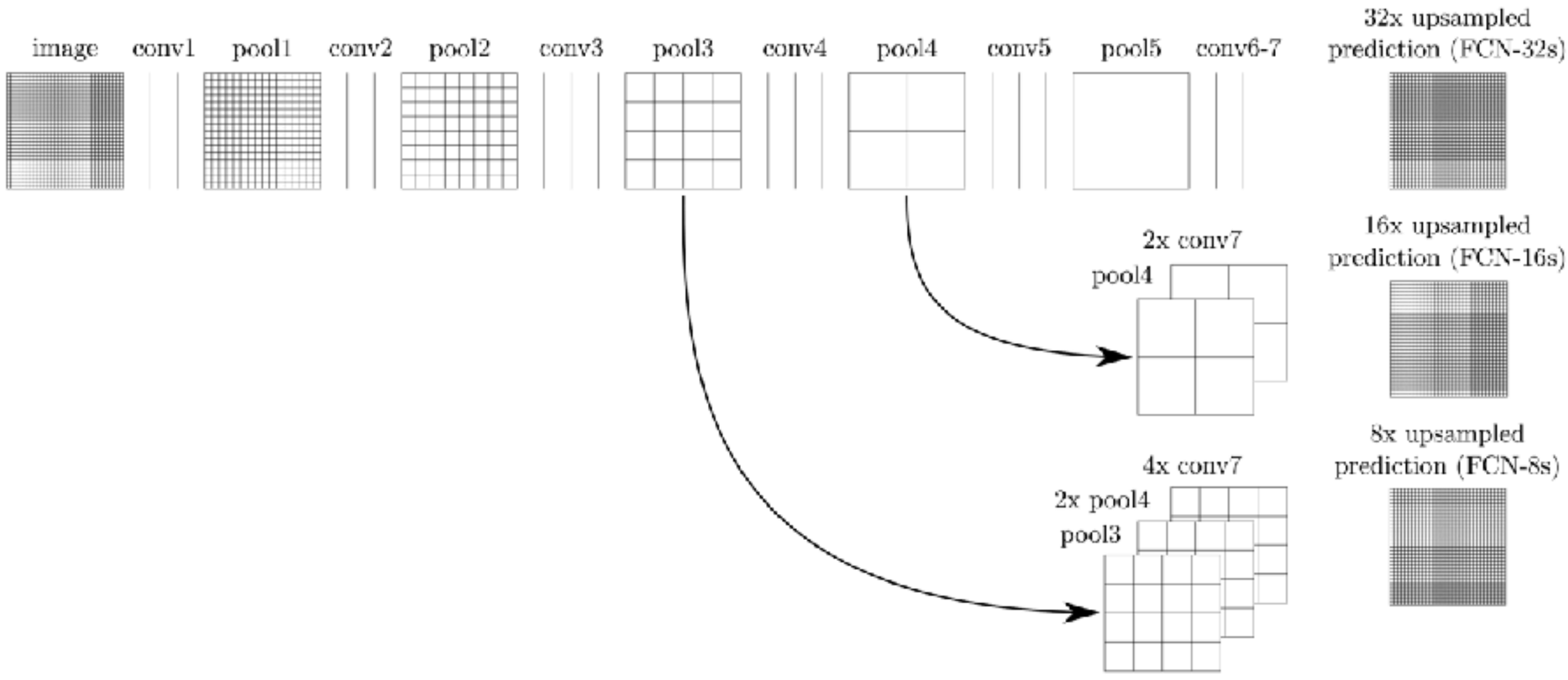


Hariharan, et al. 2014, Hypercolumns

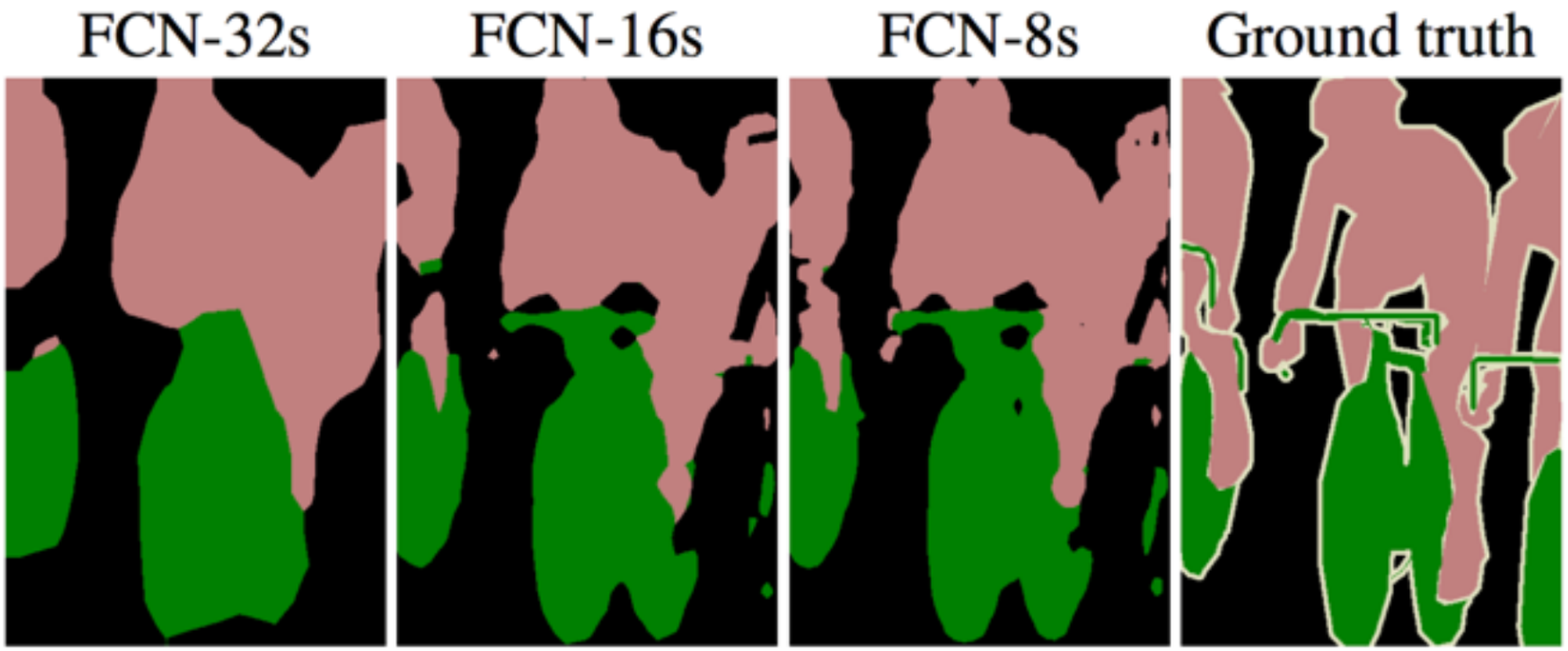
Fully Convolutional Networks



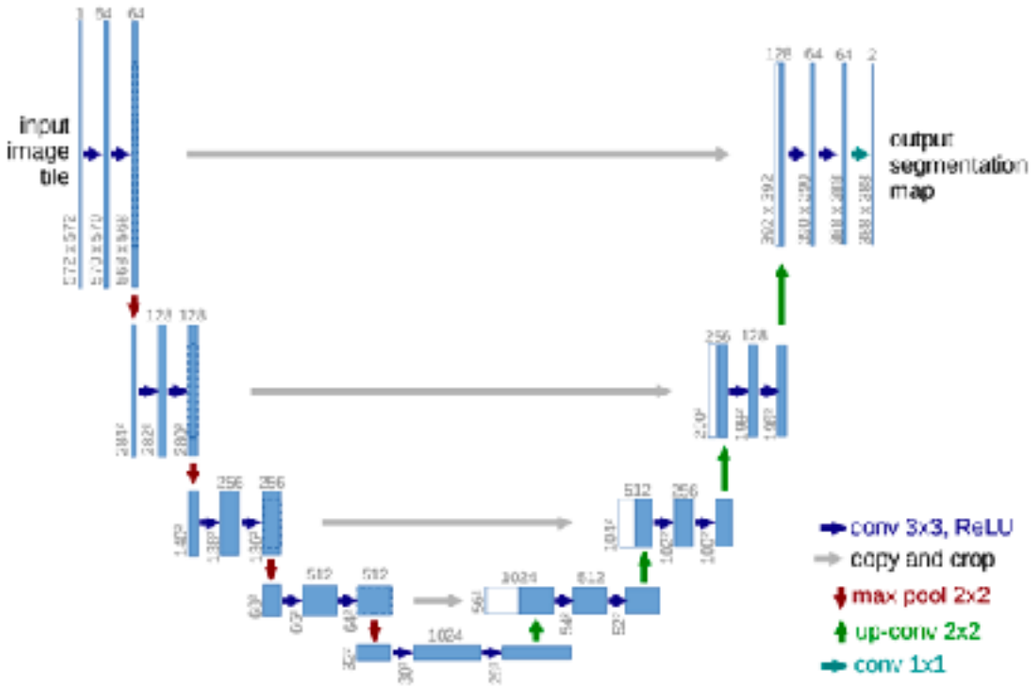
Fully Convolutional Networks



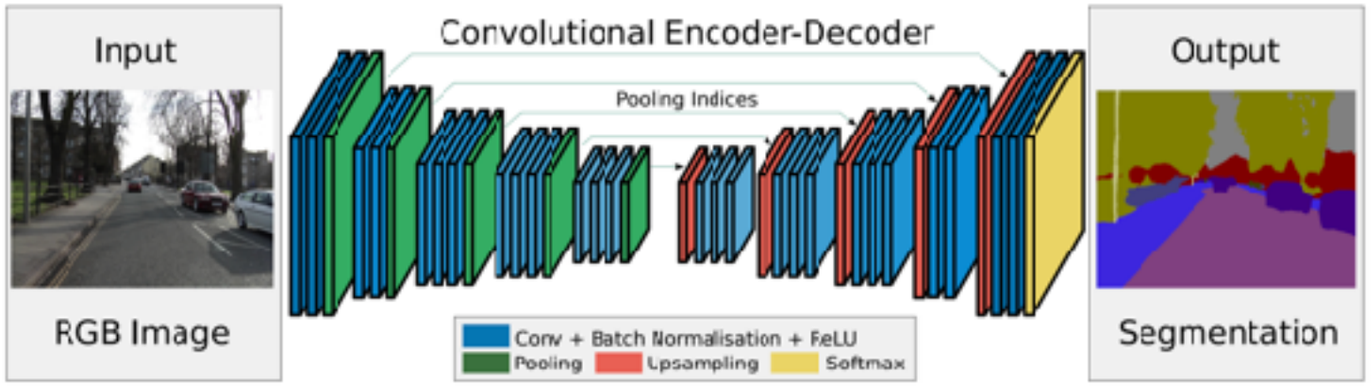
Fully Convolutional Networks



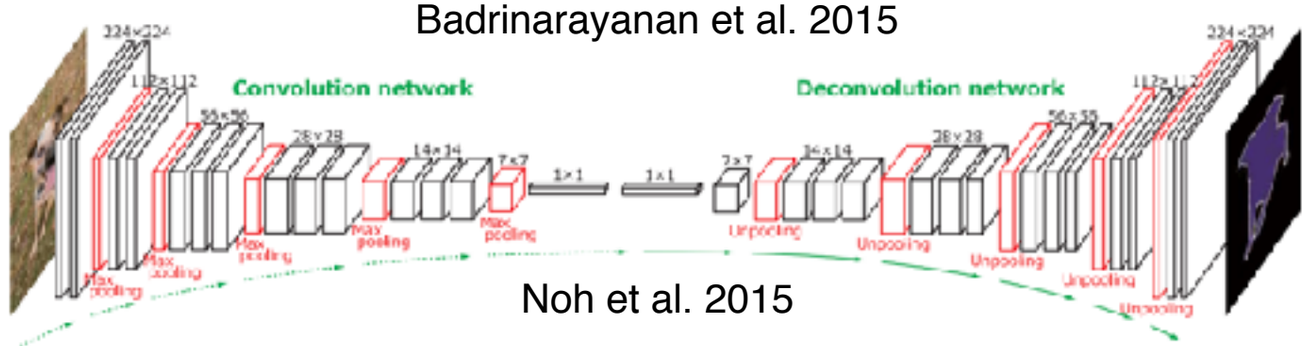
Similar Designs



U-Net



Badrinarayanan et al. 2015

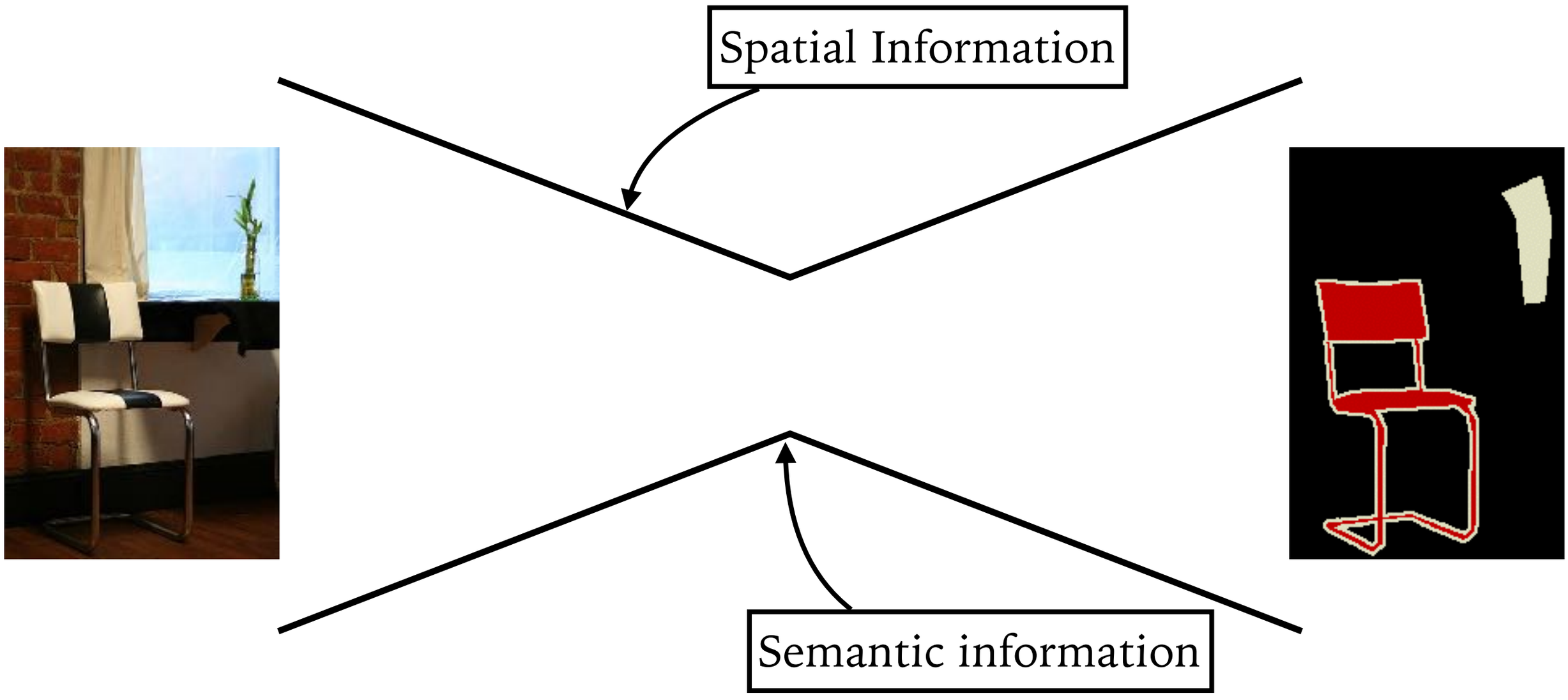


Noh et al. 2015

Fully Convolutional Networks

Chair

Fully Convolutional Networks



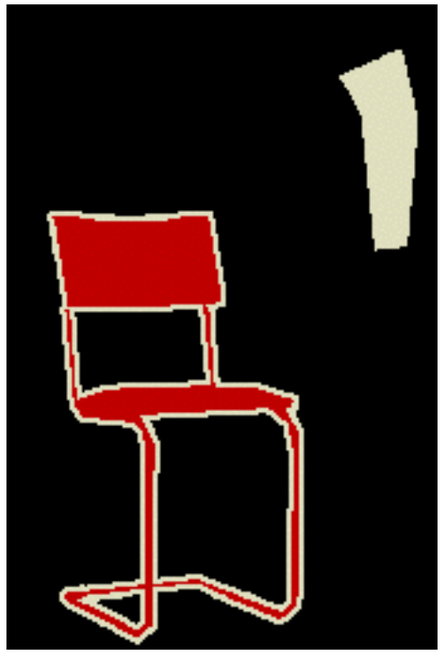
Deconvolution

- Inaccurate boundary
- Miss small objects or thin parts

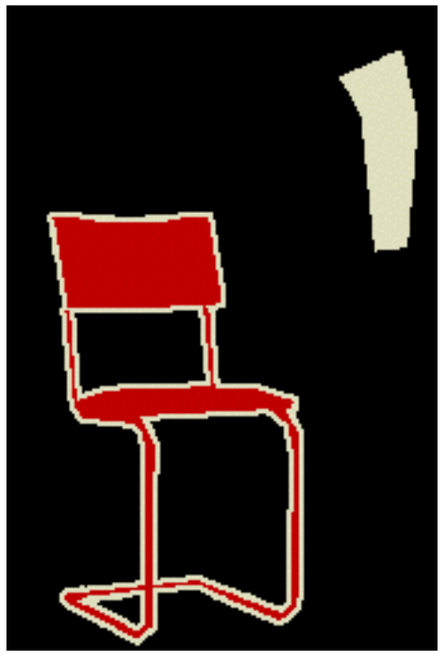
Image

FCN-8s

Ideal Network



Ideal Network



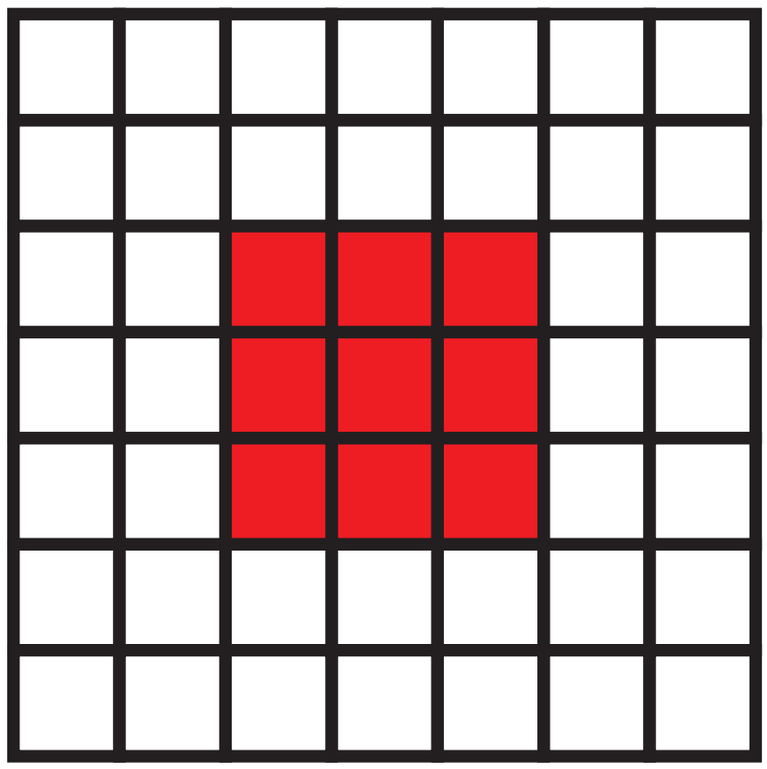
Question

How can we transform classification network without losing spatial information?

Dilated Convolution

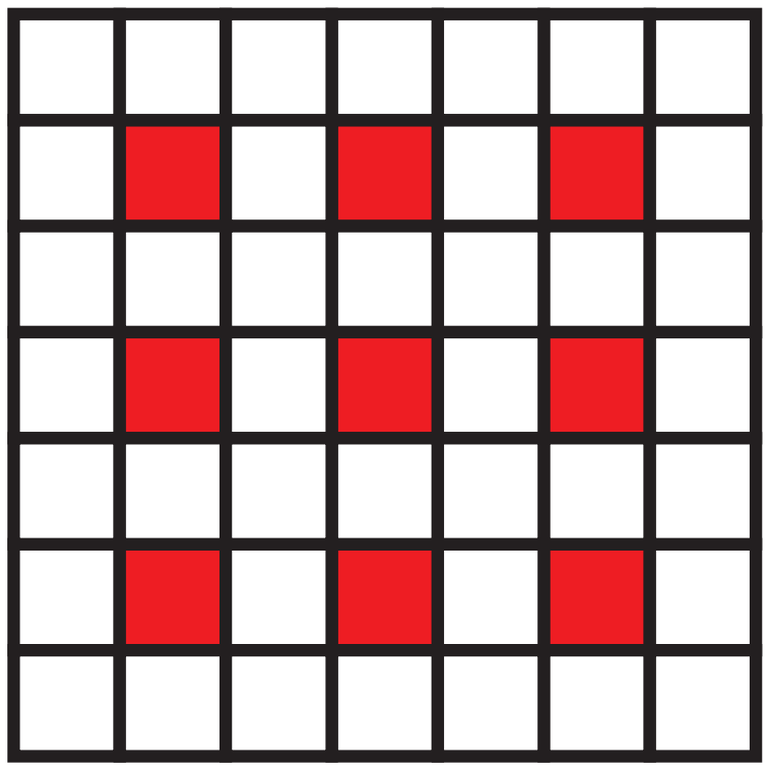
Dilated Convolution

Convolution



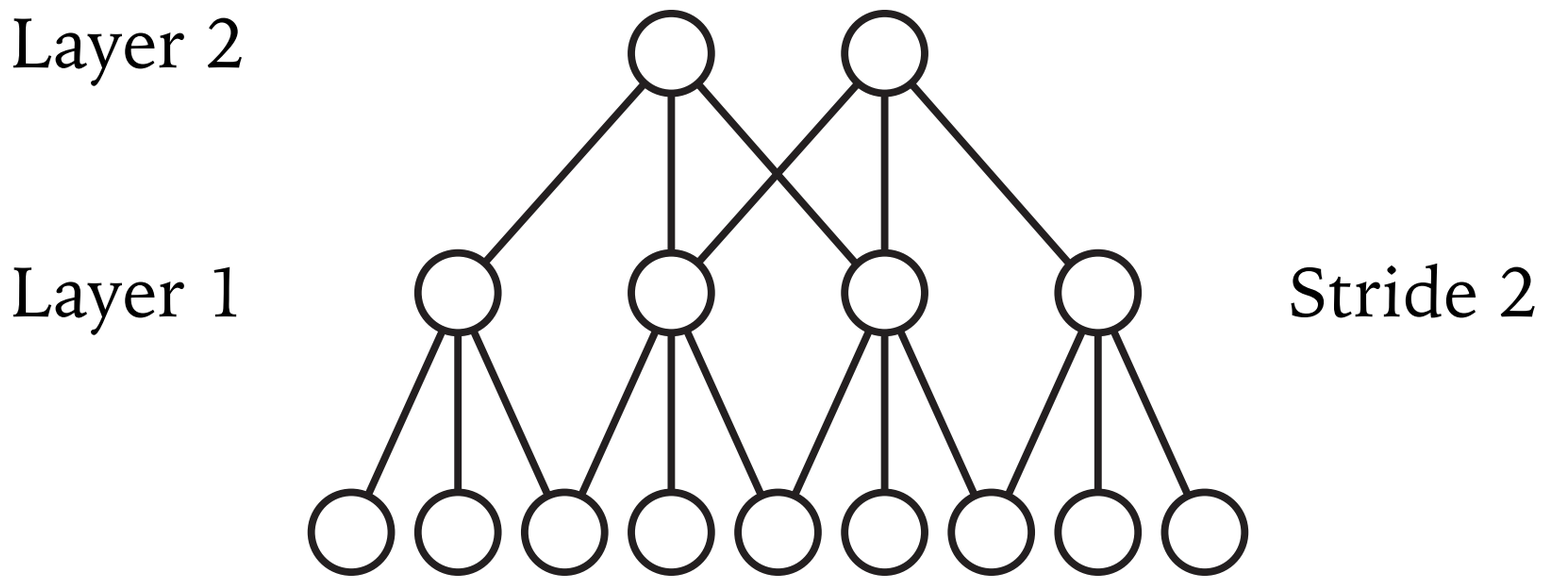
Dilation 1

Dilated Convolution

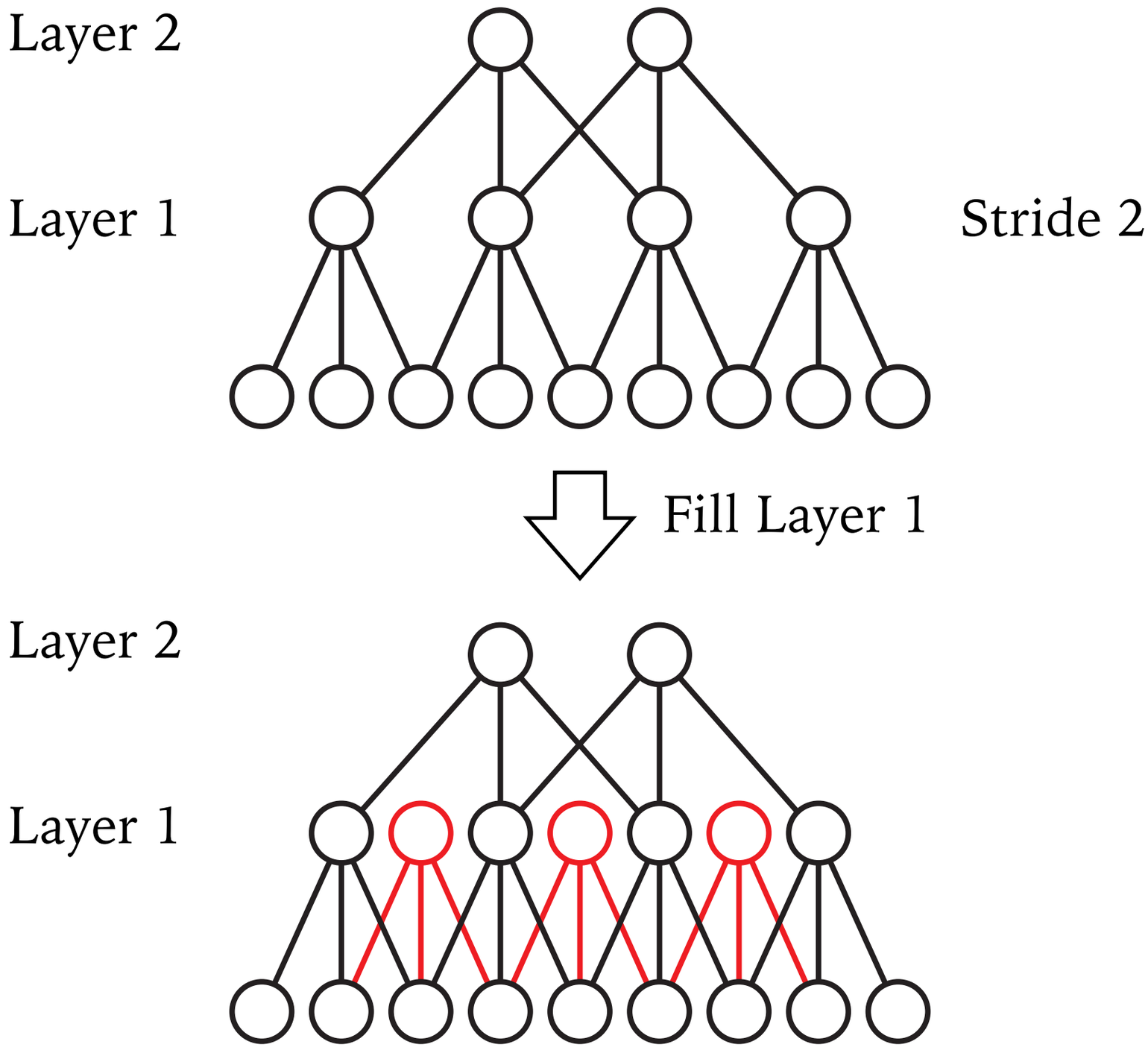


Dilation 2

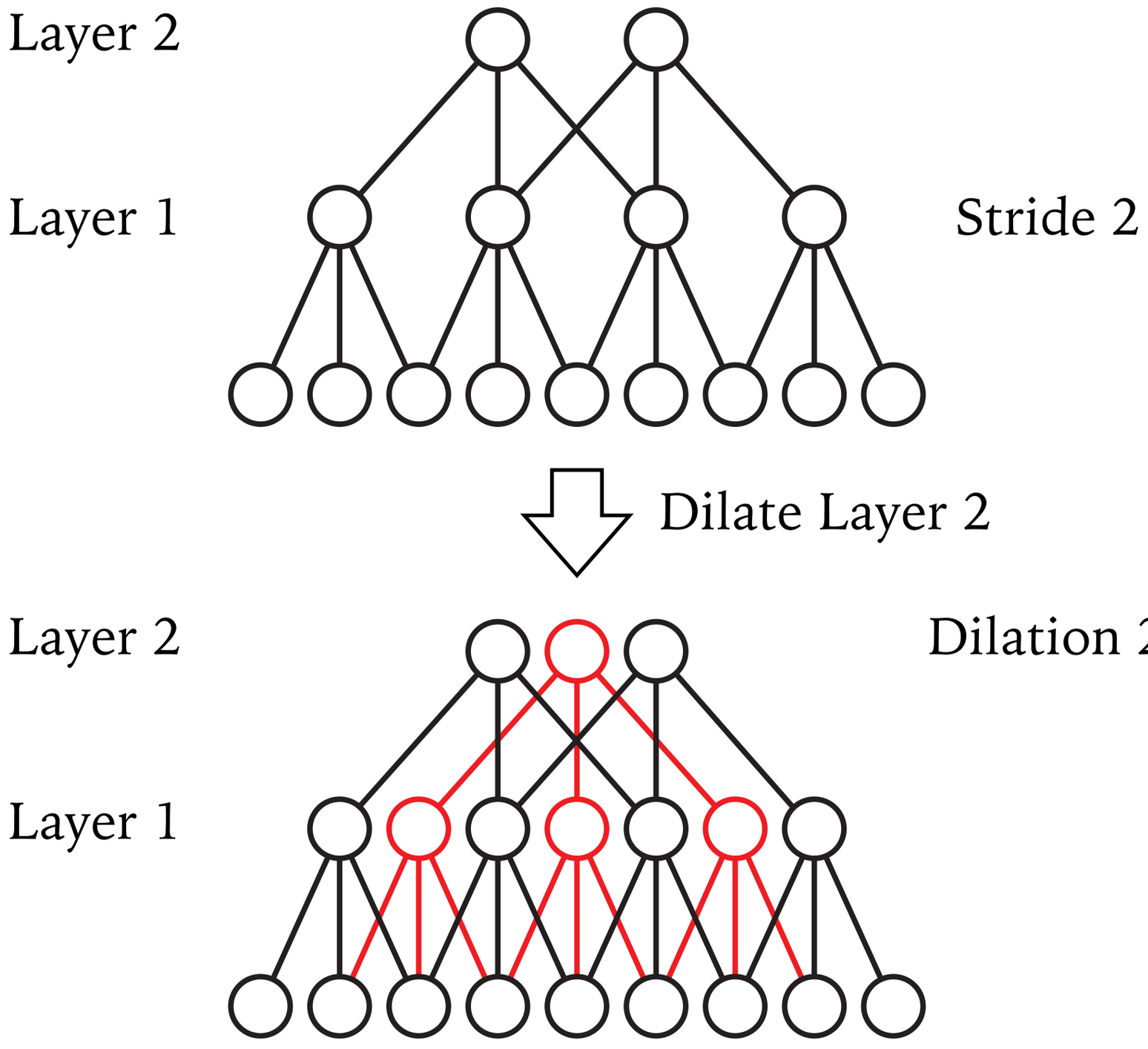
Stride to Dilation



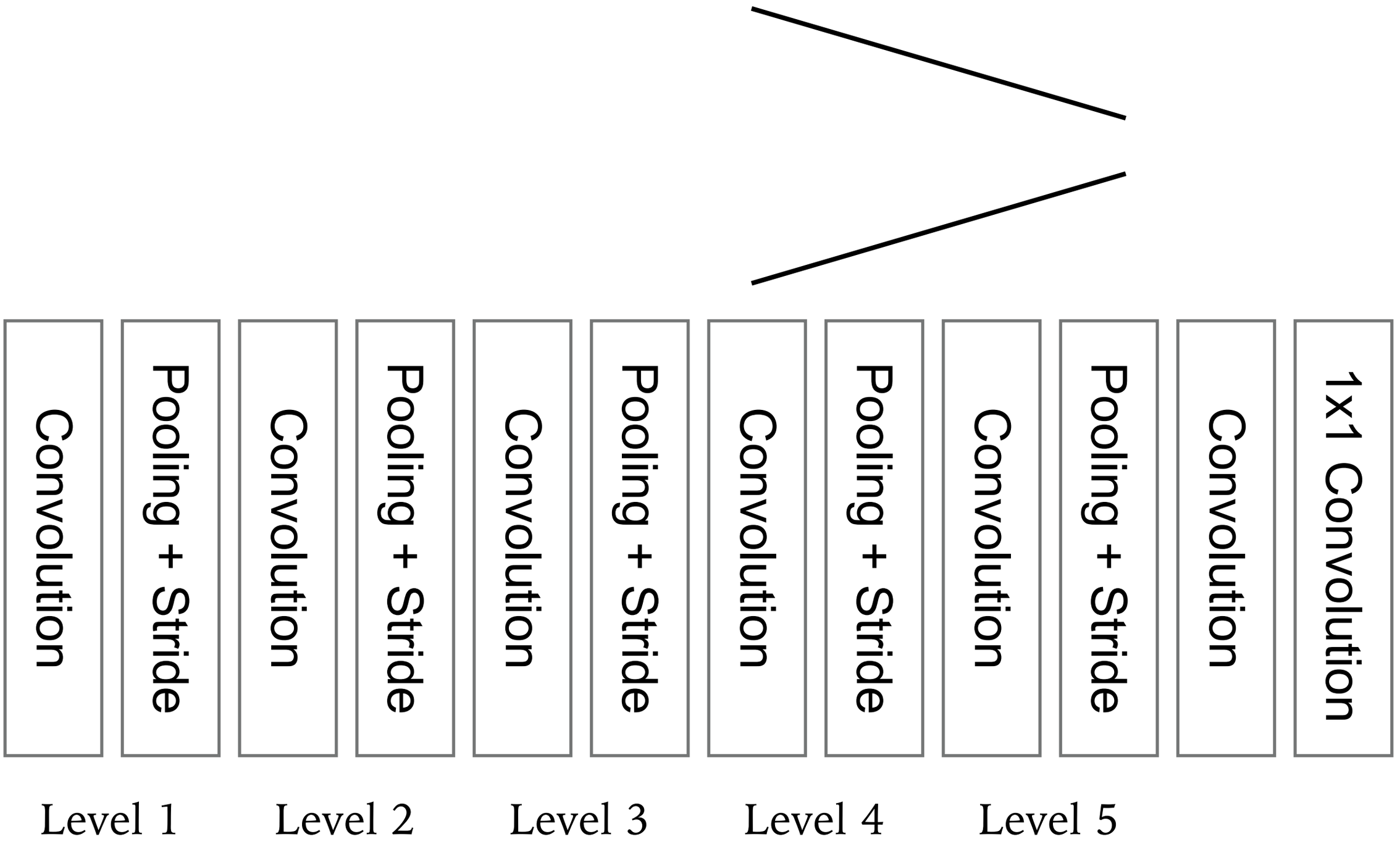
Stride to Dilation



Stride to Dilation

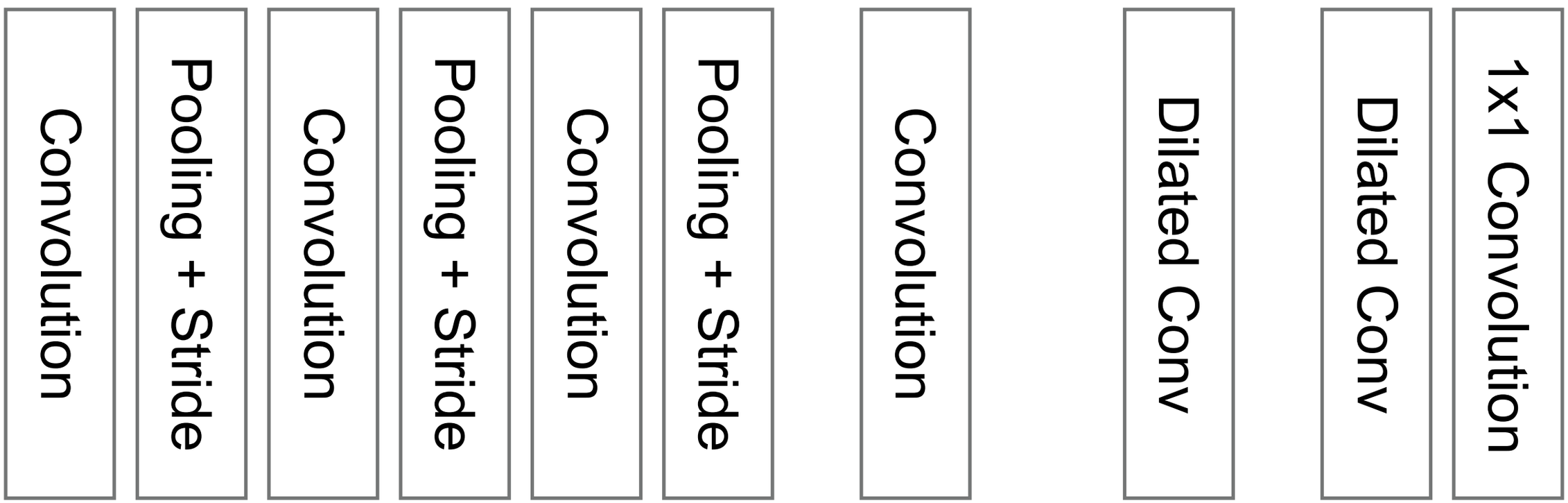


Transform 16-Layer VGG Net



Transform 16-Layer VGG Net

Front End



Front End

Train on PASCAL VOC 2012

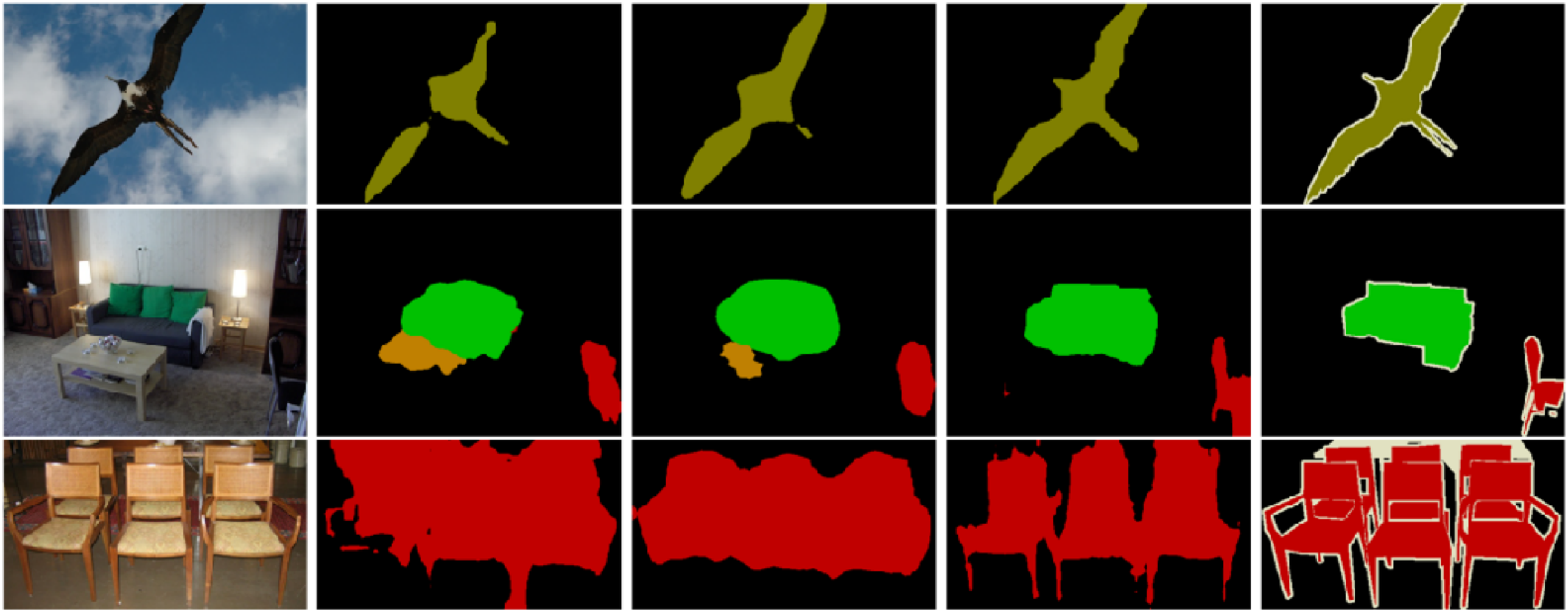


Image FCN-8s DeepLab Front End Ground Truth

Front End

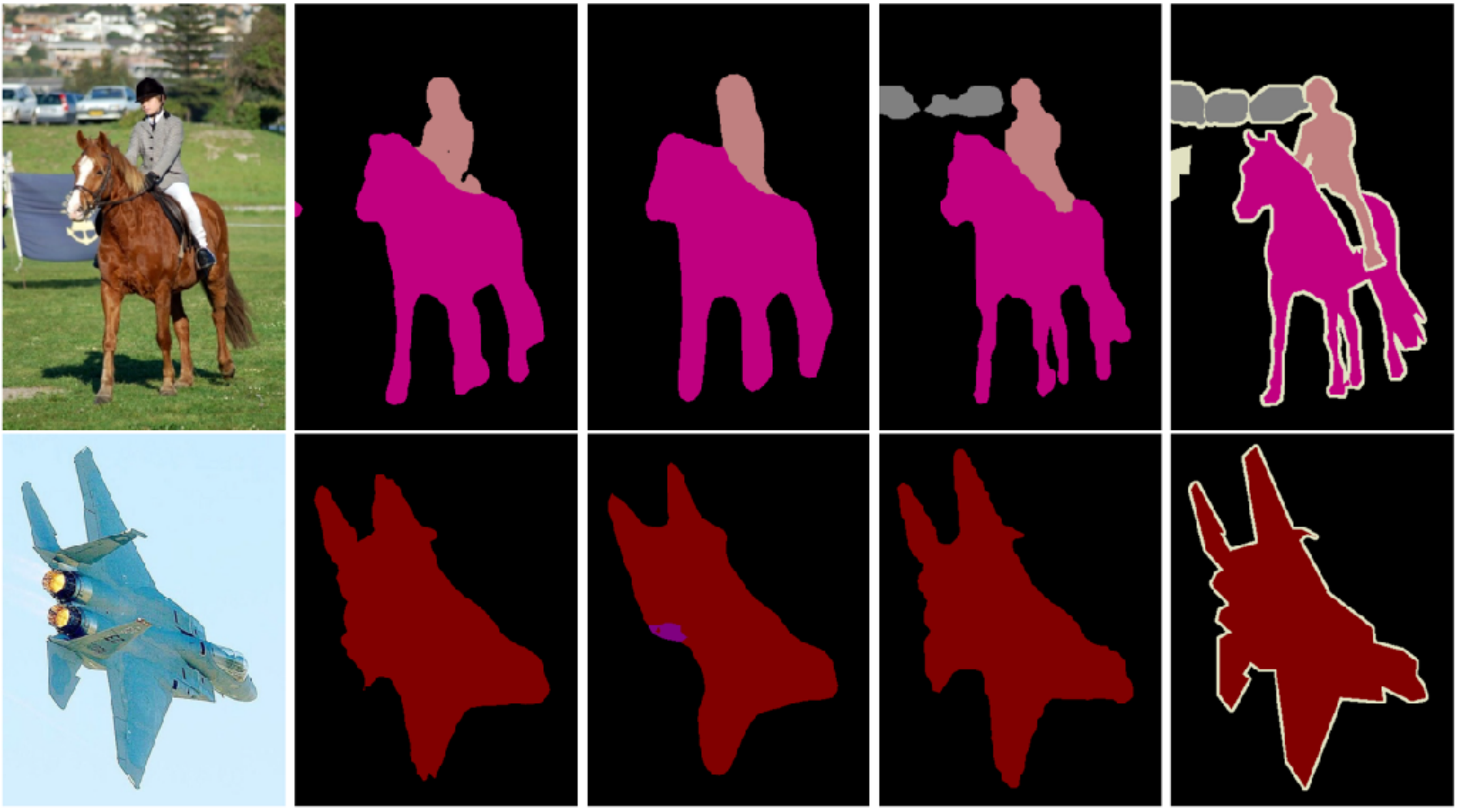


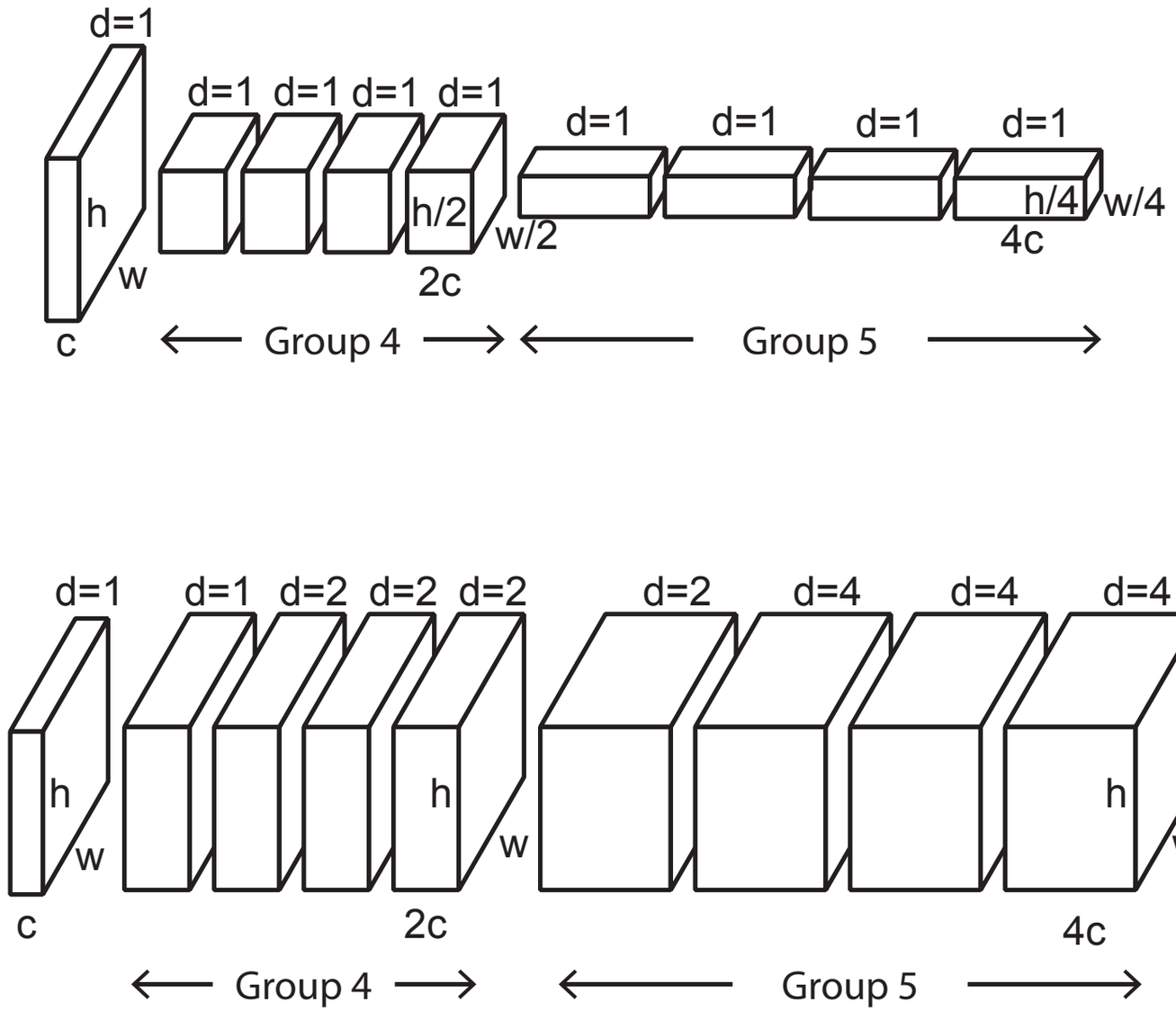
Image FCN-8s DeepLab Front End Ground Truth

Dilated Residual Networks

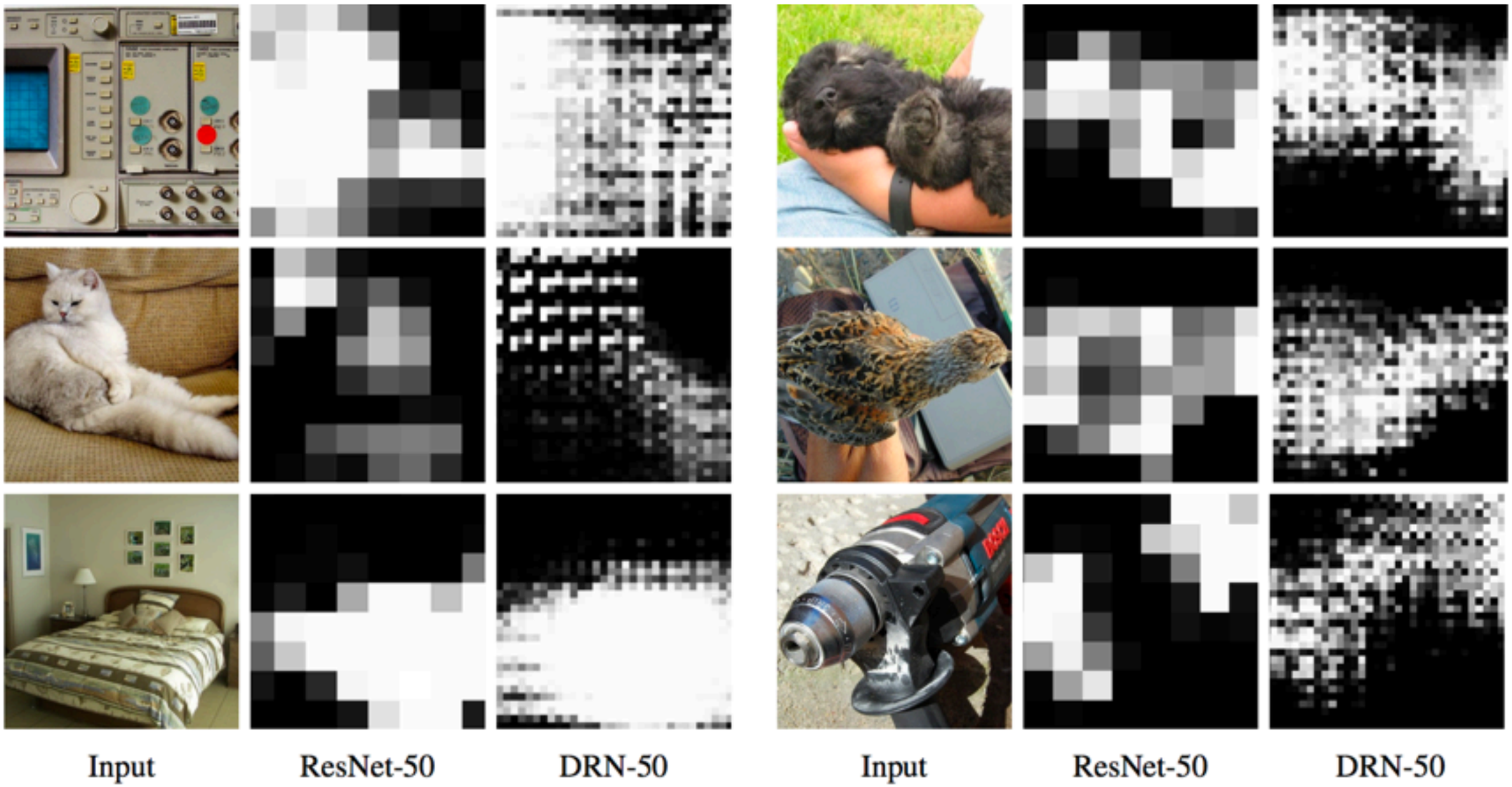
What if we use semantic segmentation network to do image classification?

Does high resolution feature maps benefit the classification task?

Dilated Residual Networks



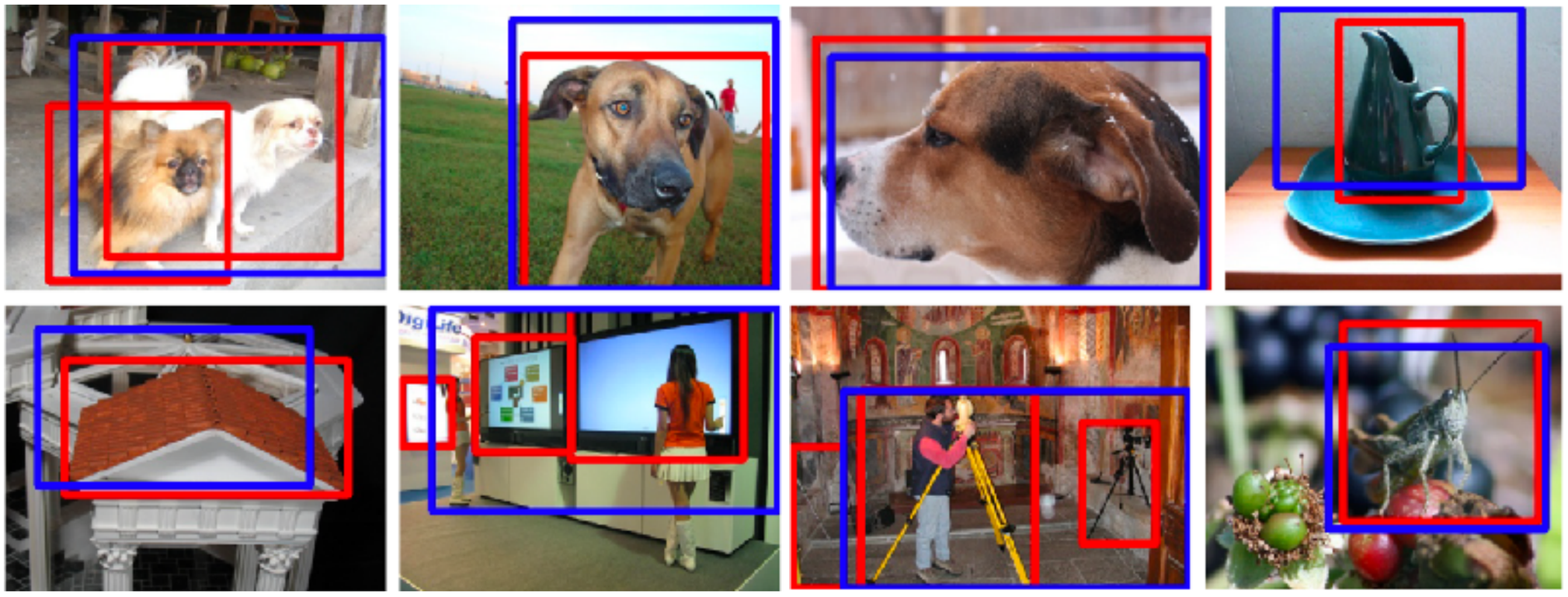
Dilated Residual Networks



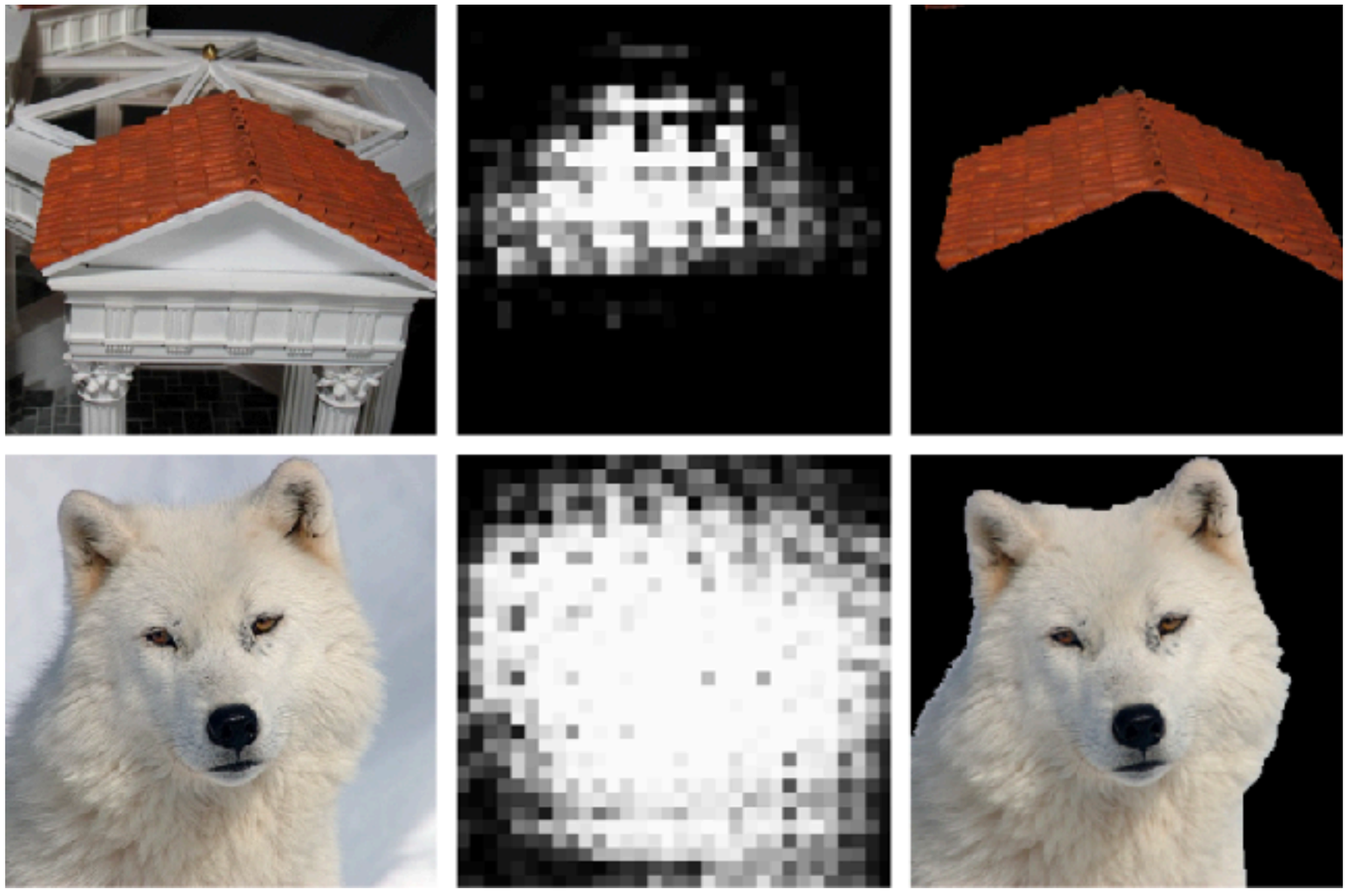
Dilated Residual Networks

	Top-1		Top-5	
# layers	ResNet	DRN	ResNet	DRN
18	30.43	**27.97**	10.76	**9.54**
34	26.73	**24.81**	8.74	**7.54**
50	24.01	**22.94**	7.02	**6.57**

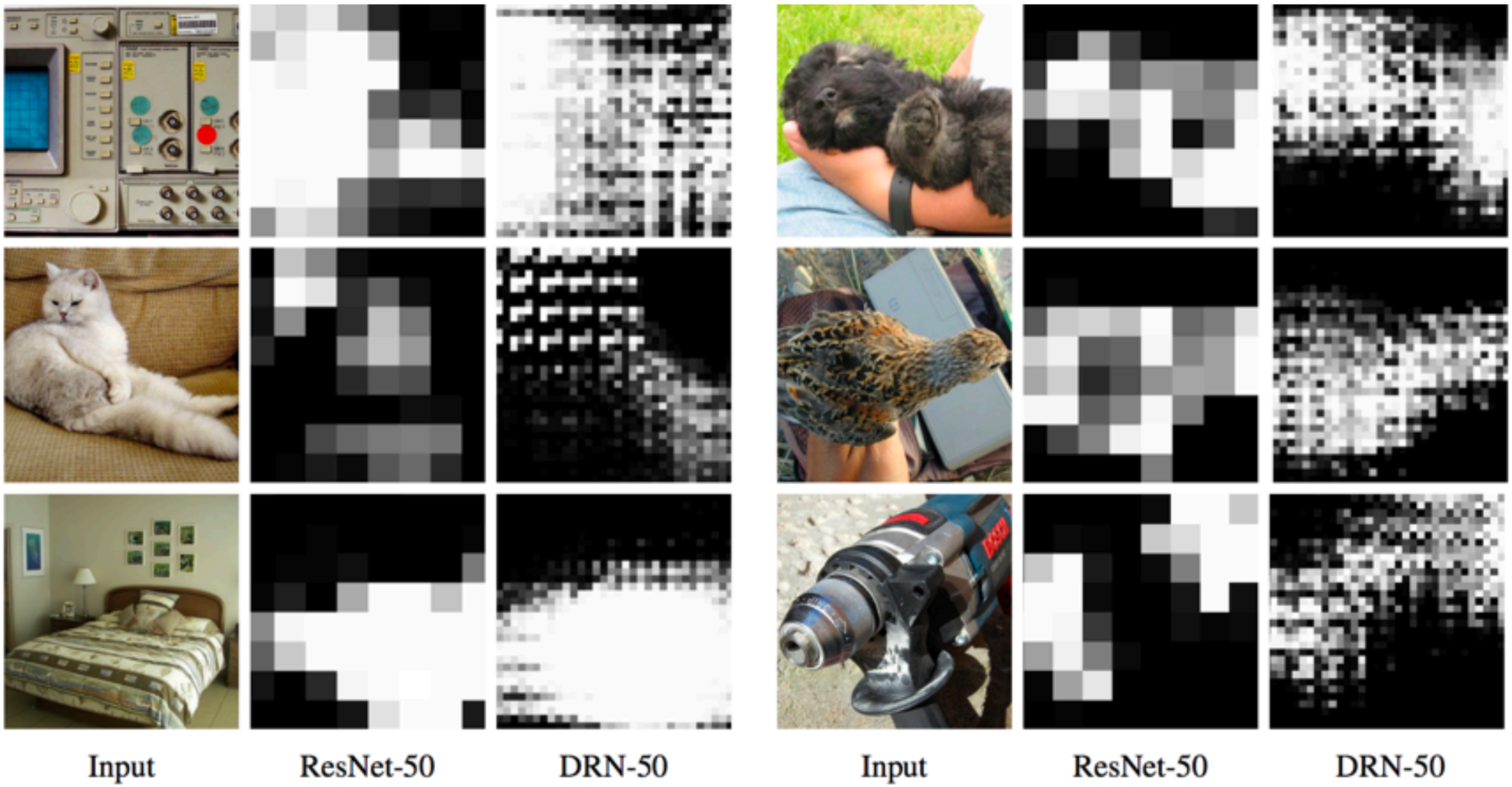
Dilated Residual Networks



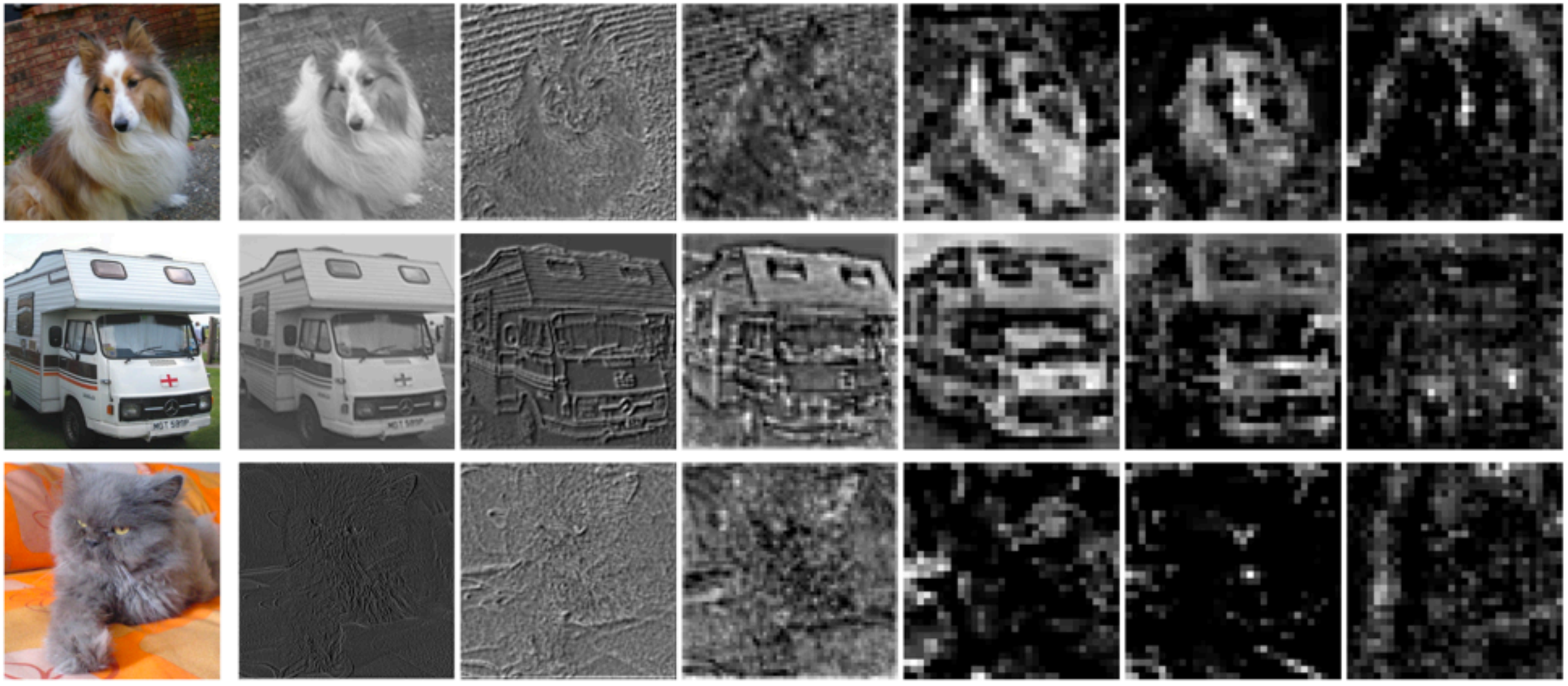
Dilated Residual Networks



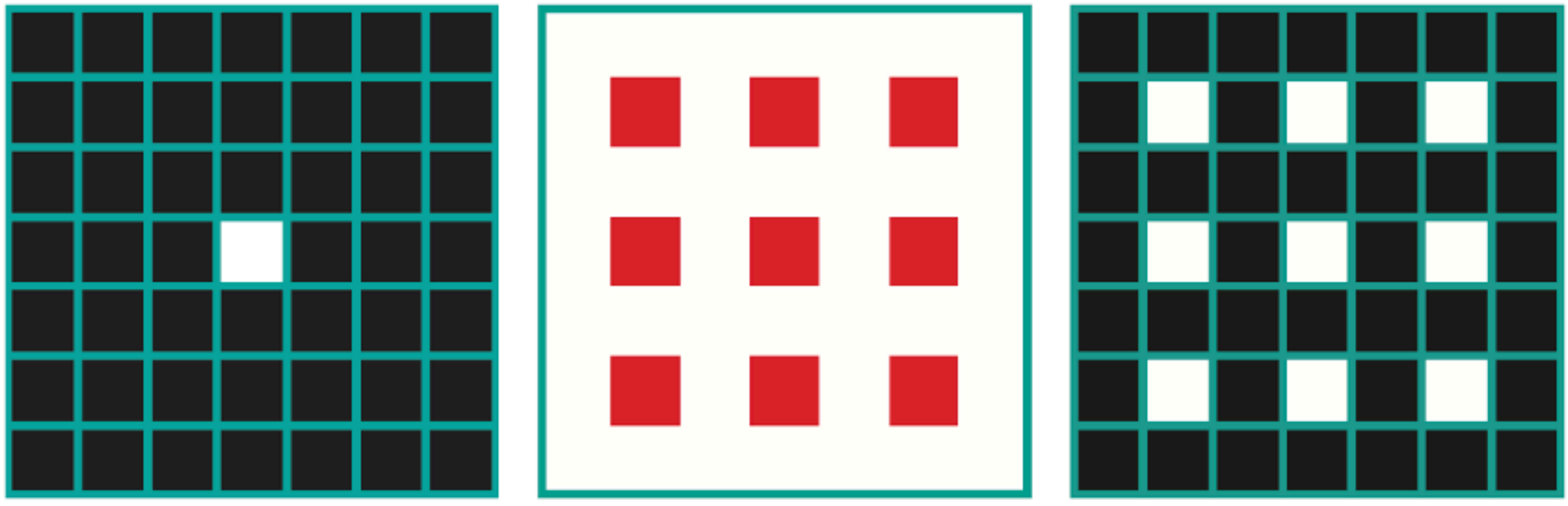
Dilated Residual Networks



Dilated Residual Networks

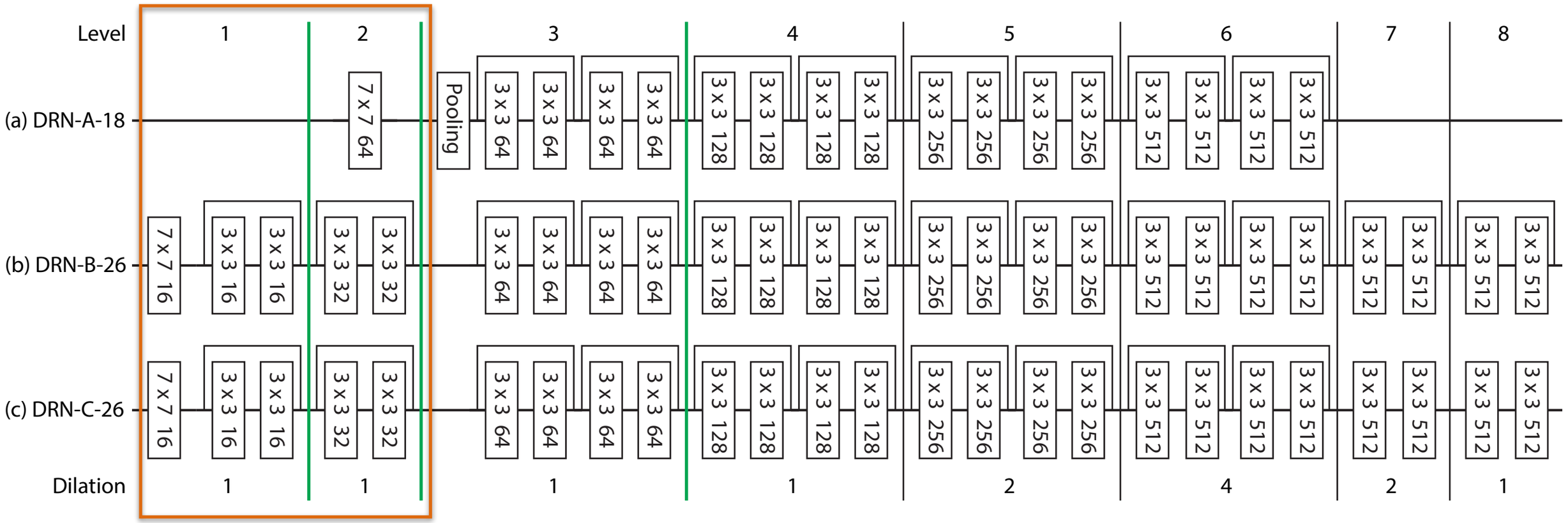


Dilated Residual Networks

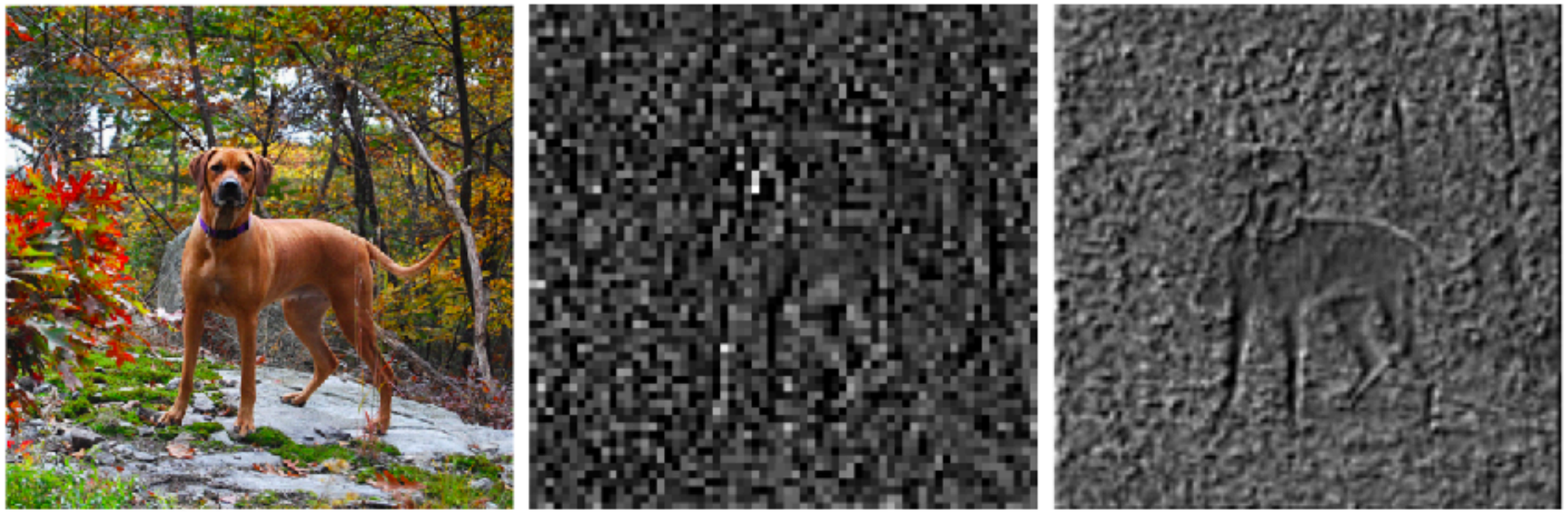


(a) Input (b) Dilation 2 (c) Output

Dilated Residual Networks

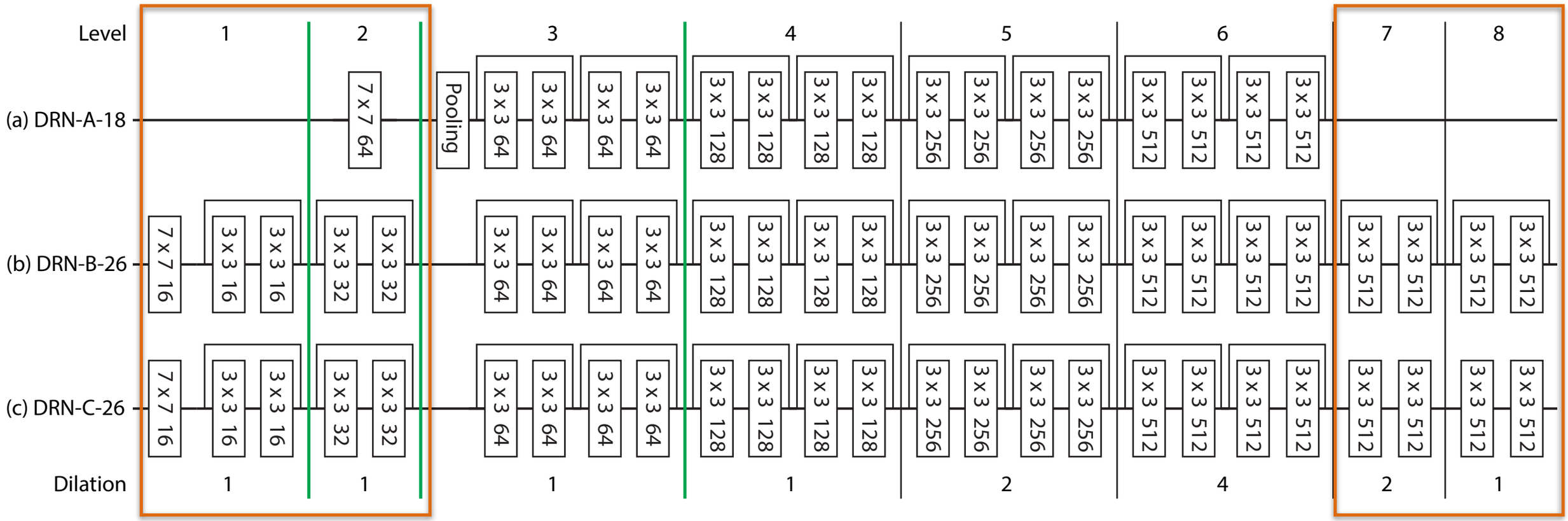


Dilated Residual Networks

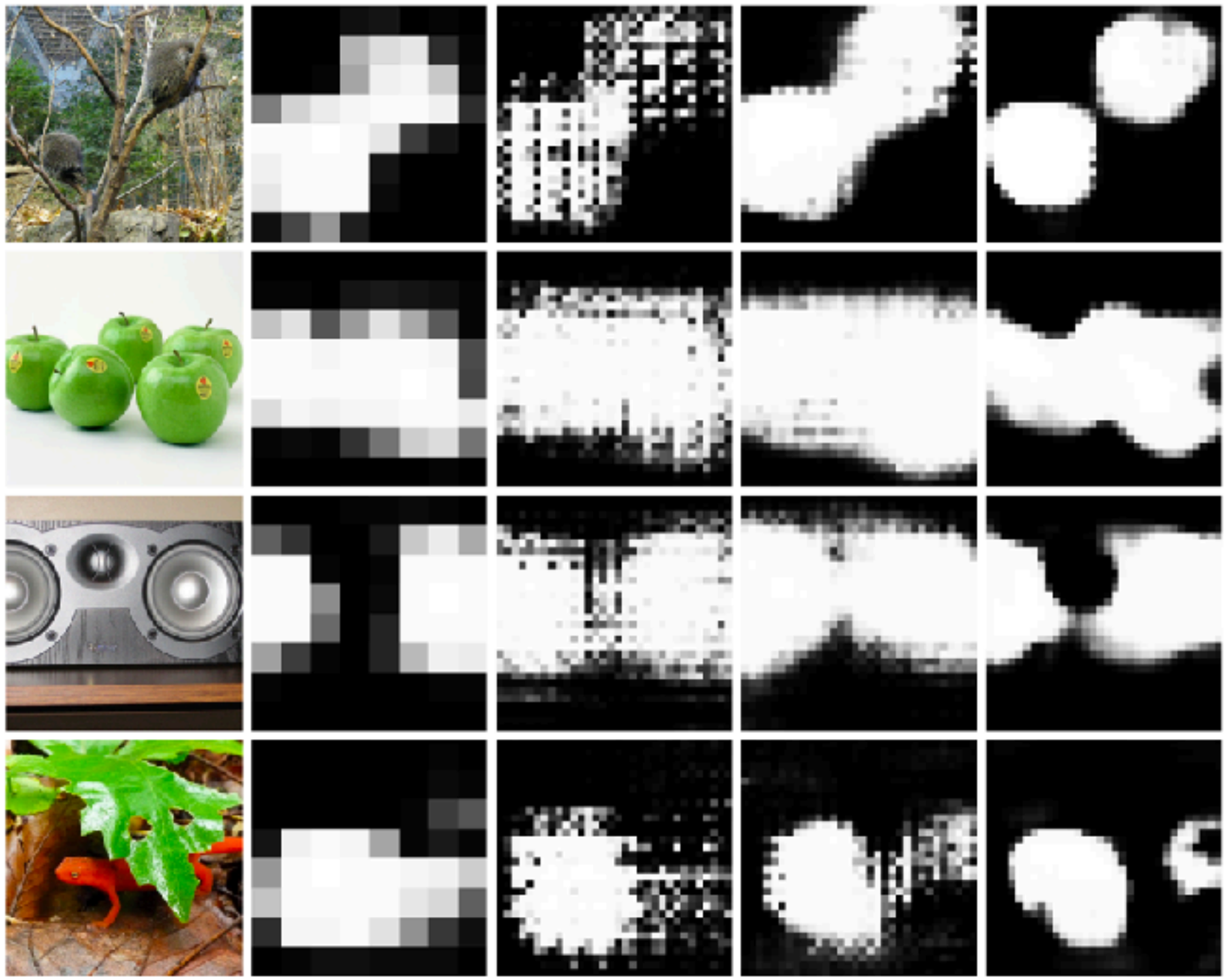


(a) Input (b) DRN-A-18 (c) DRN-B-26

Dilated Residual Networks

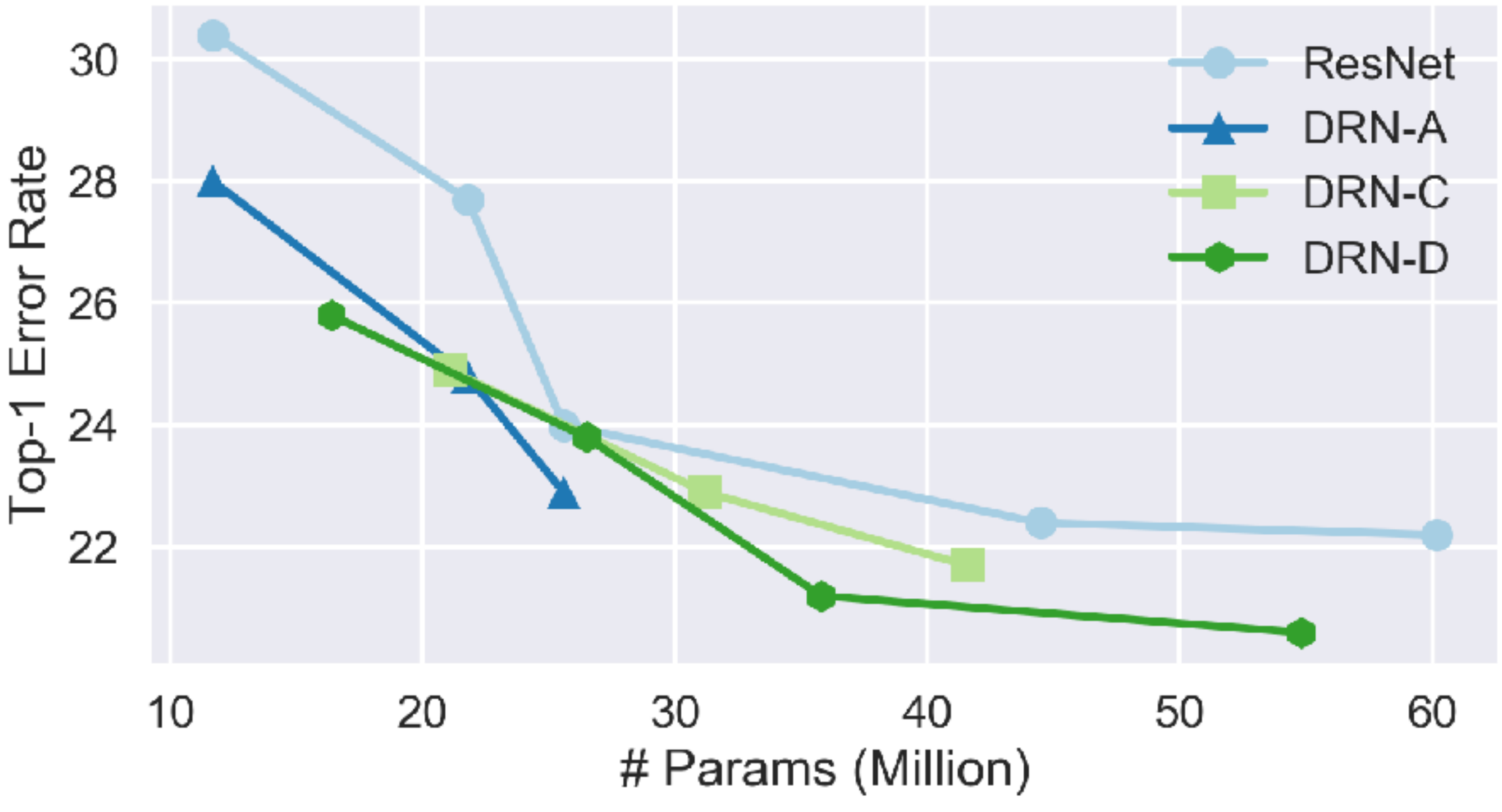


Dilated Residual Networks



(a) Input (b) ResNet-18 (c) DRN-A-18 (d) DRN-B-26 (e) DRN-C-26

Dilated Residual Networks

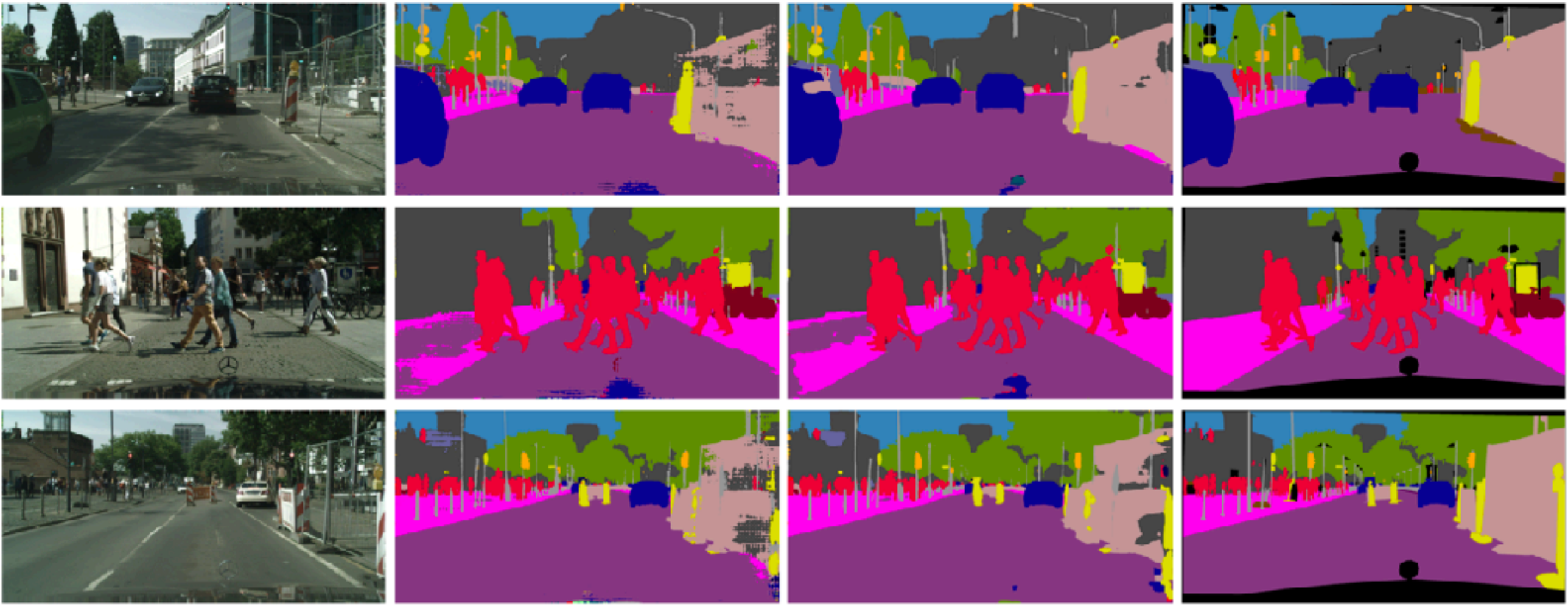


Dilated Residual Networks

	Road	Sidewalk	Building	Wall	Fence	Pole	Light	Sign	Vegetation	Terrain	Sky	Person	Rider	Car	Truck	Bus	Train	Motorcycle	Bicycle	mean IoU
DRN-A-50	96.9	77.4	90.3	35.8	42.8	59.0	**66.8**	74.5	91.6	57.0	93.4	78.7	55.3	92.1	43.2	59.5	36.2	**52.0**	**75.2**	67.3
DRN-C-26	97.4	80.7	90.4	36.1	47.0	56.9	63.8	73.0	91.2	**57.9**	93.4	77.3	53.8	92.7	45.0	70.5	48.4	44.2	72.8	68.0
DRN-C-42	**97.7**	**82.2**	**91.2**	**40.5**	**52.6**	**59.2**	66.7	**74.6**	**91.7**	57.7	**94.1**	**79.1**	**56.0**	**93.6**	**56.0**	**74.3**	**54.7**	50.9	74.1	**70.9**

Cityscapes Semantic Segmentation

Dilated Residual Networks



(a) Input (b) DRN-A-50 (c) DRN-C-26 (d) Ground truth

Questions?