Fully Convolutional Networks for Semantic Segmentation

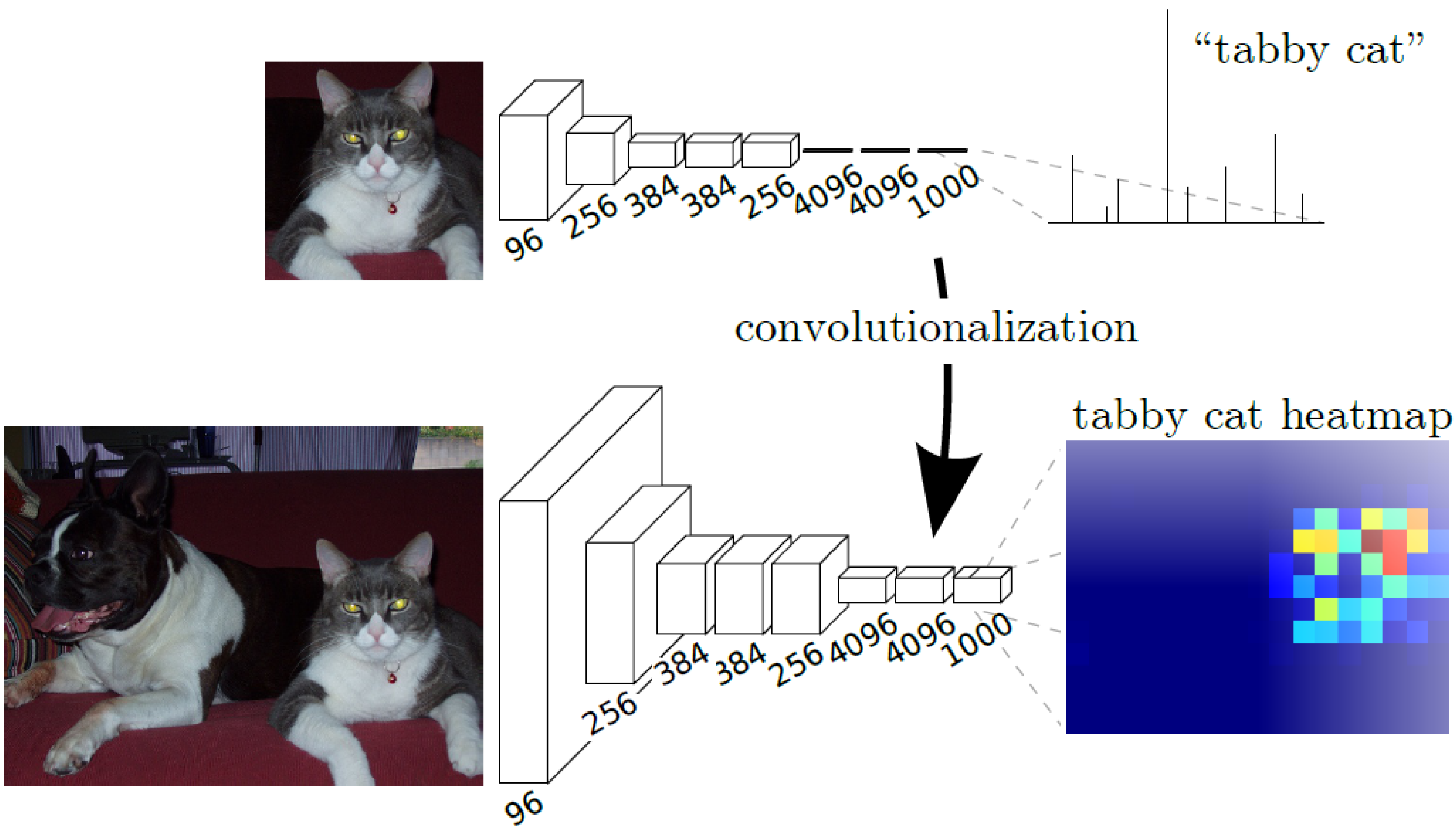
Jonathan Long, Evan Shelhamer & Trevor Darrell (2015)

Proceedings of the IEEE

발표: 문동지

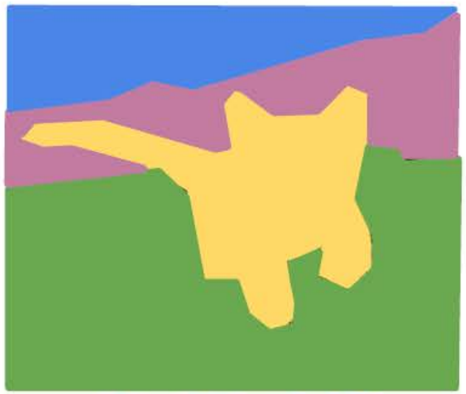
Semantic Segmentation = Pixel Level Classification

Deep Learning(CNN) Image Classification 잘한다
Sematic Segmentation은 결국 Pixel level의 Classification이다
Image Classification으로 Sematic Segmentation을 할 수 있겠다



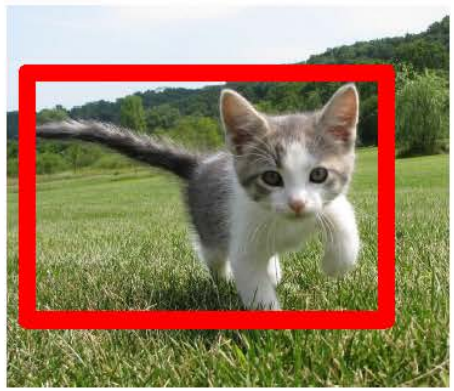
Other Computer Vision Tasks

Semantic Segmentation



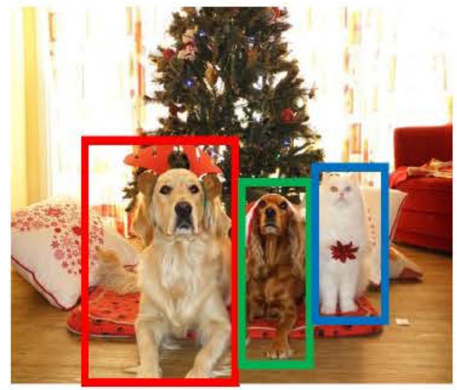
GRASS, CAT, TREE, SKY

Classification + Localization



CAT

Object Detection



DOG, DOG, CAT

Instance Segmentation

DOG, DOG, CAT

No objects, just pixels

Single Object

Multiple Object

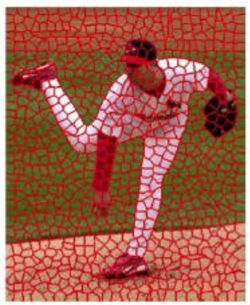
This image is CC0 public domain

Image & Object Segmentation

Image Segmentation

- Group pixels into regions that share some similar properties

Superpixels
(Ren ICCV 2003



Segmenting Images into meaningful objects

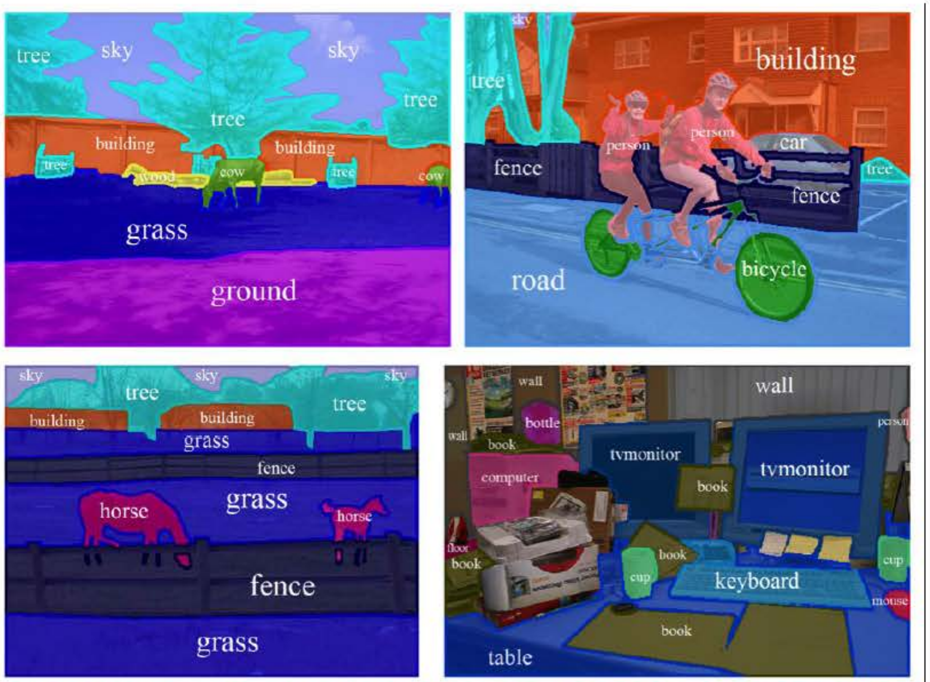
- Object−level segmentation

: accurate localization and recognition

Semantic Segmentation

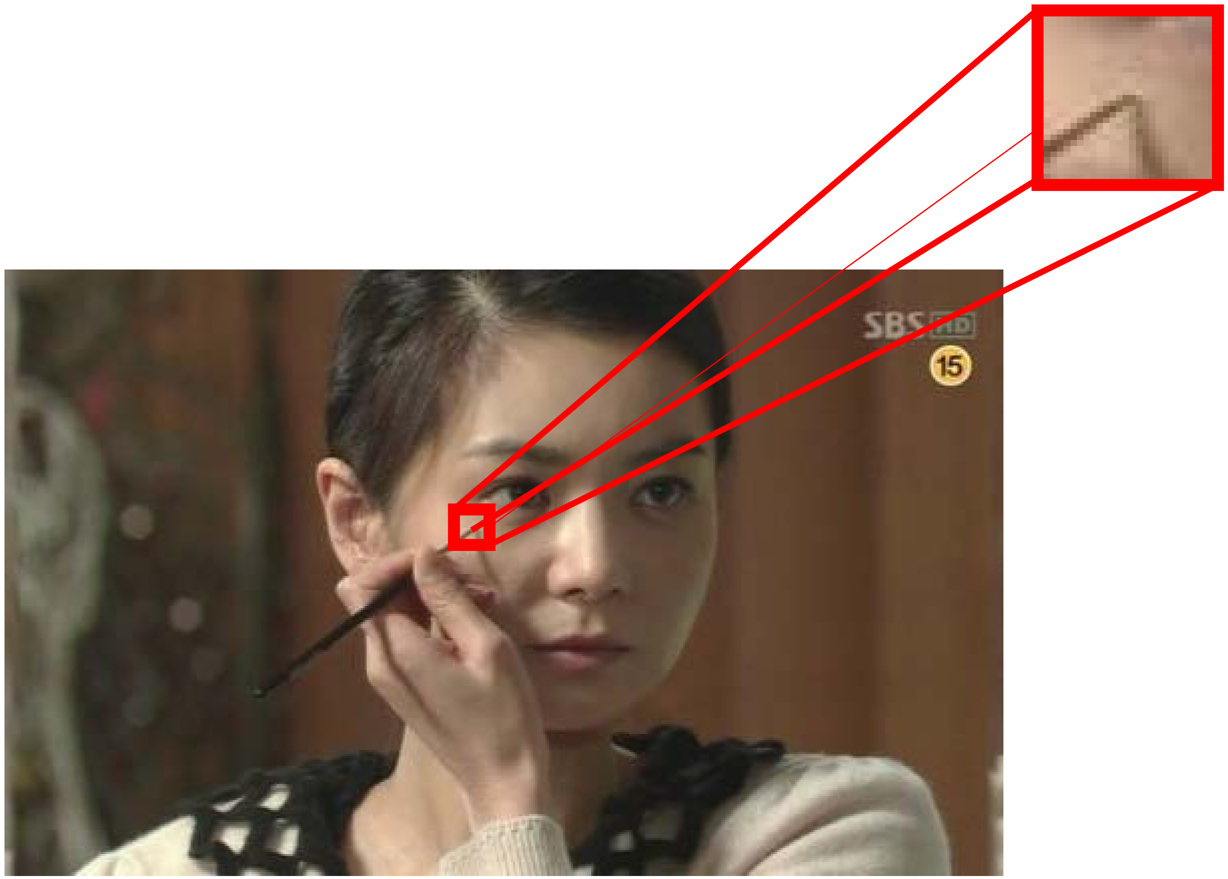
Semantic Segmentation

- Label every pixel: recognize the class of every pixel
- Do not differentiate instances



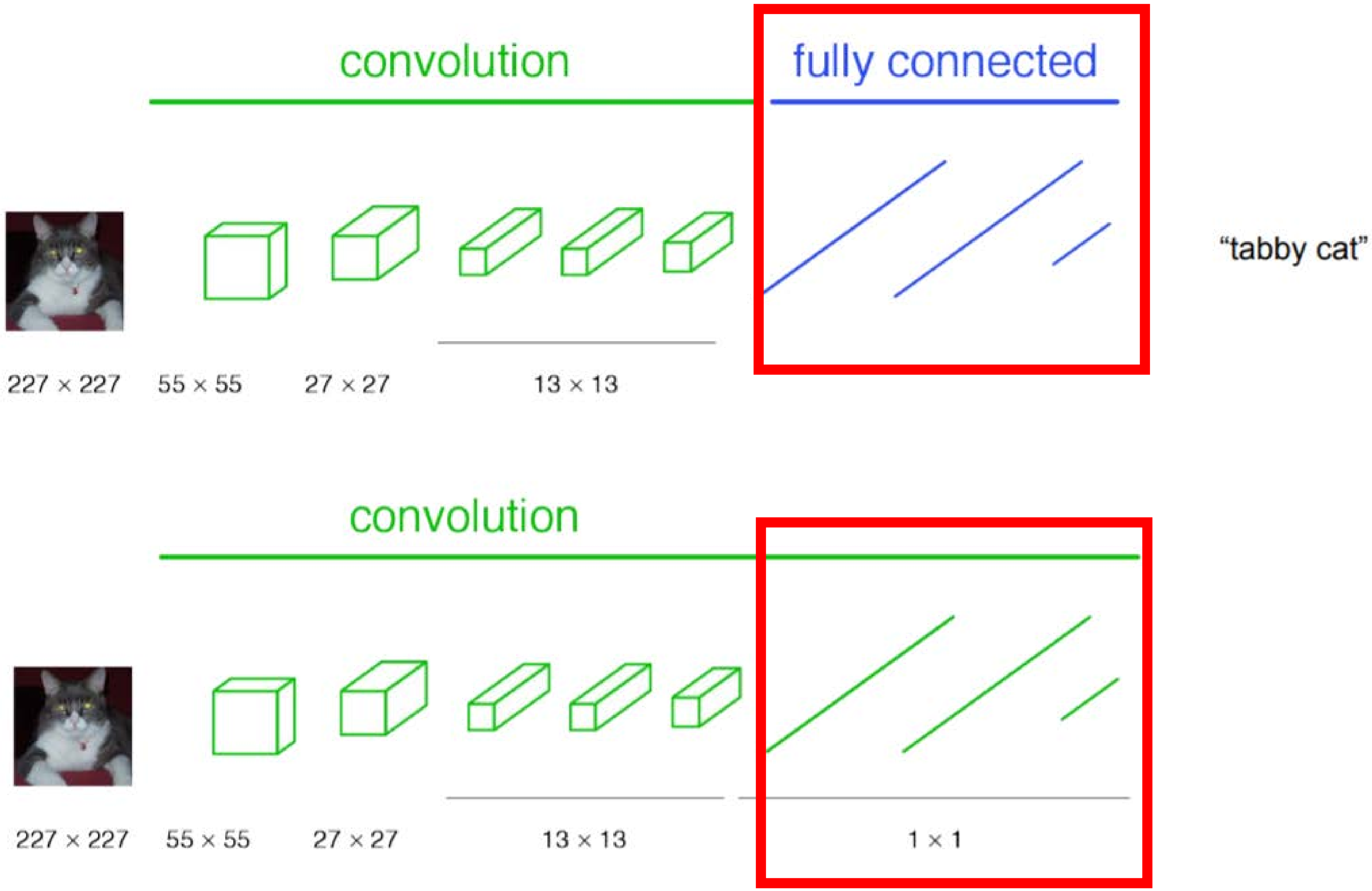
Mottaghi et al, "The role of context for object detection and semantic segmentation in the wild", CVPR 2014

Semantic Segmentation = Pixel Level Classification

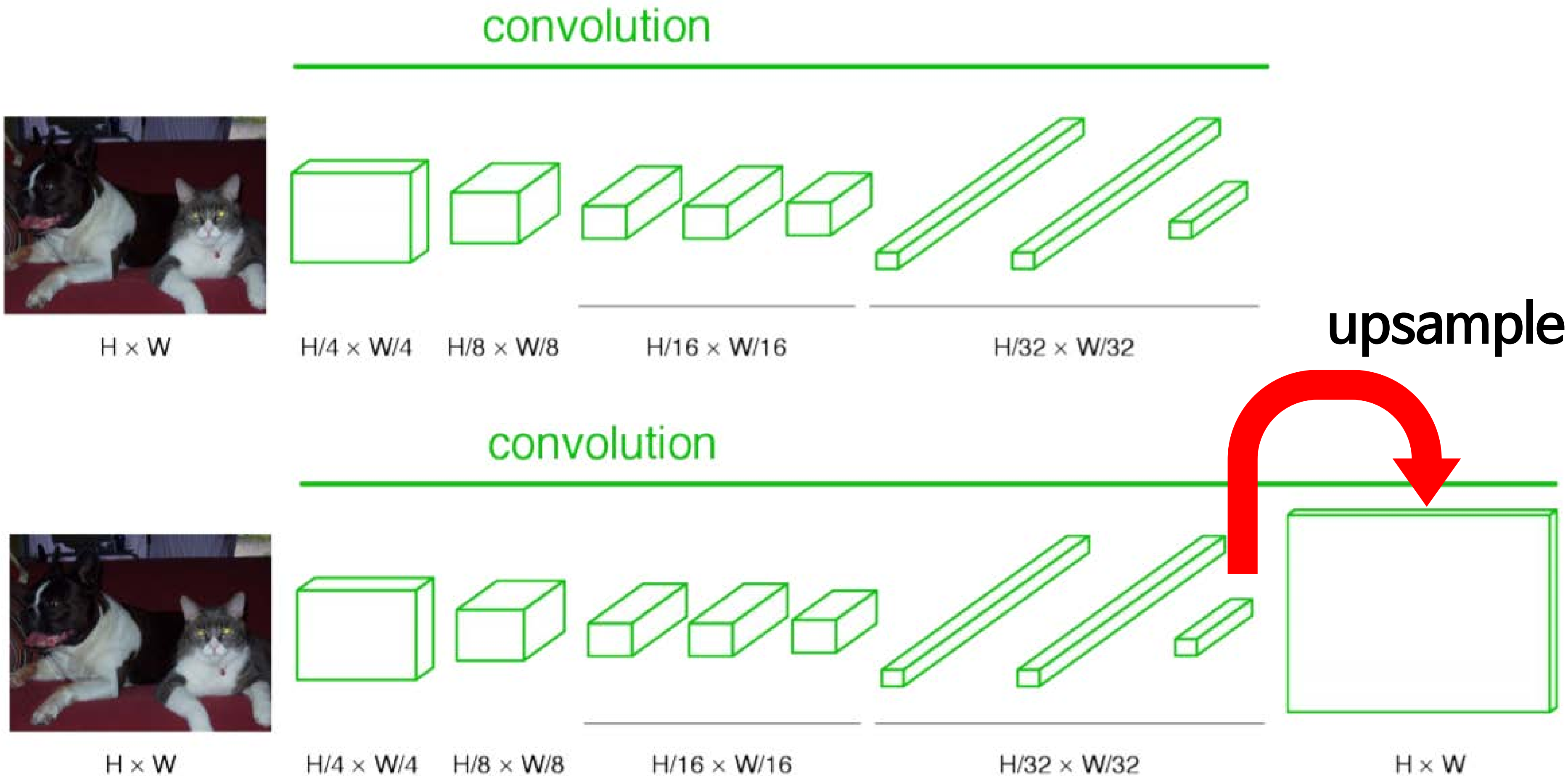


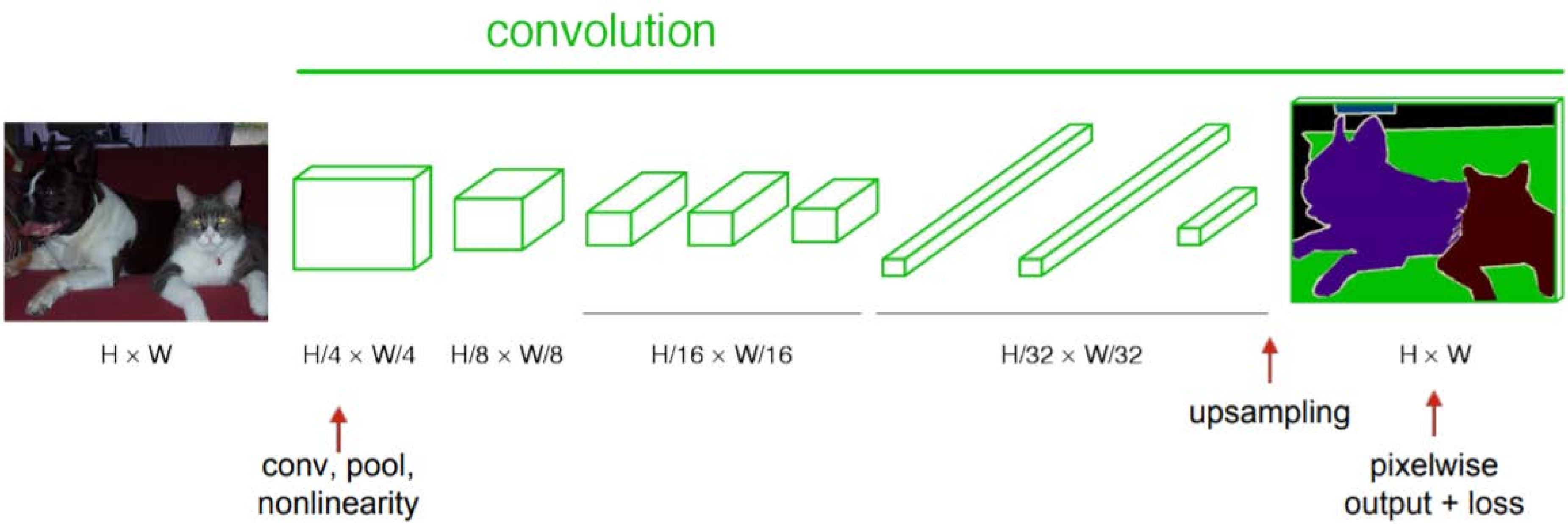
점?

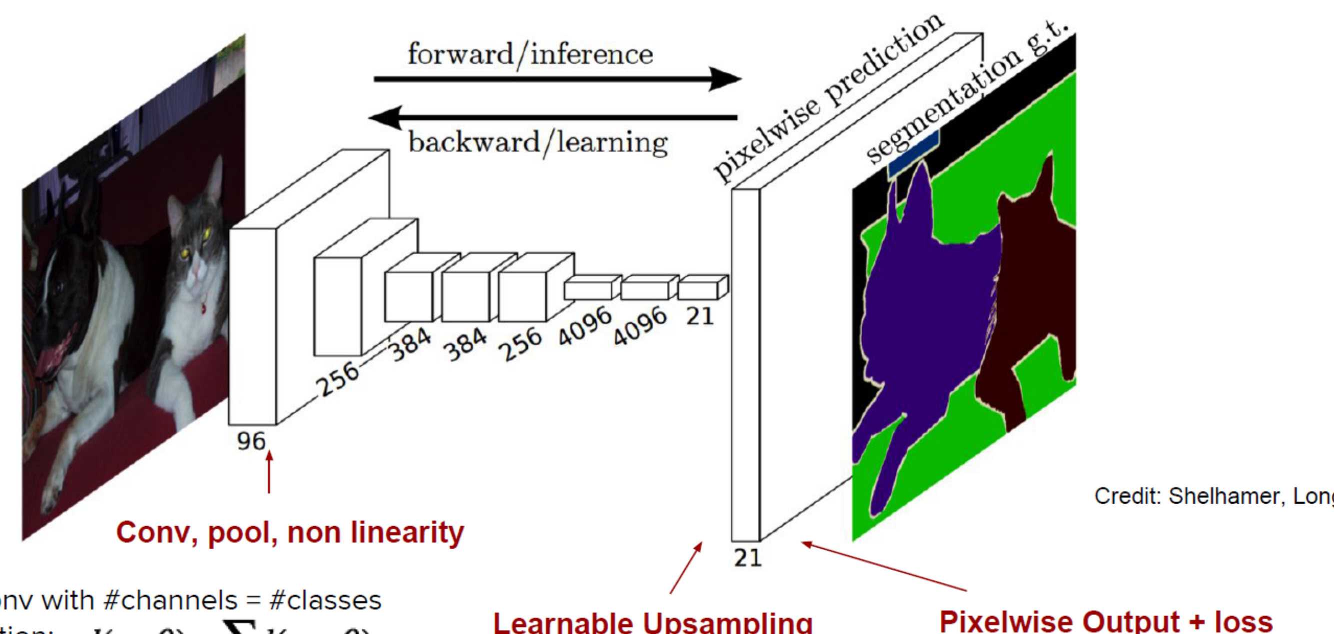
Classification / Semantic Segmentation



Fully convolutional







Final layer is 1X1 conv with #channels = #classes

Pixelwise loss function: $l(x;\ \theta) = \sum_{i,j} l(x_{ij};\ \theta)$

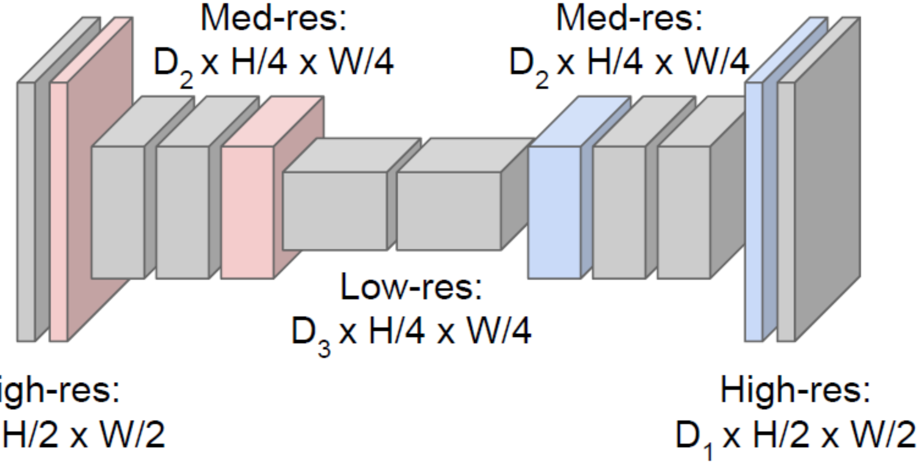
Semantic Segmentation Idea: Fully Convolutional

Downsampling: Pooling, strided convolution

Design network as a bunch of convolutional layers, with **downsampling** and **upsampling** inside the network!

Upsampling: ???

Input:
3 x H x W



Predictions:
H x W

Long, Shelhamer, and Darrell, "Fully Convolutional Networks for Semantic Segmentation", CVPR 2015
Noh et al, "Learning Deconvolution Network for Semantic Segmentation", ICCV 2015

In-Network upsampling: "Unpooling"

Nearest Neighbor

1	2
3	4

Input: 2 x 2

1	1	2	2
1	1	2	2
3	3	4	4
3	3	4	4

Output: 4 x 4

"Bed of Nails"

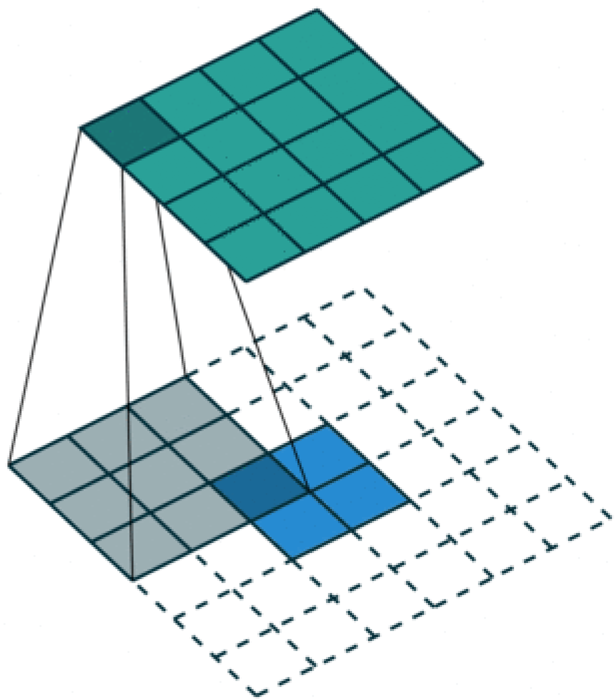
1	2
3	4

Input: 2 x 2

1	0	2	0
0	0	0	0
3	0	4	0
0	0	0	0

Output: 4 x 4

Upsampling Via Decovolution



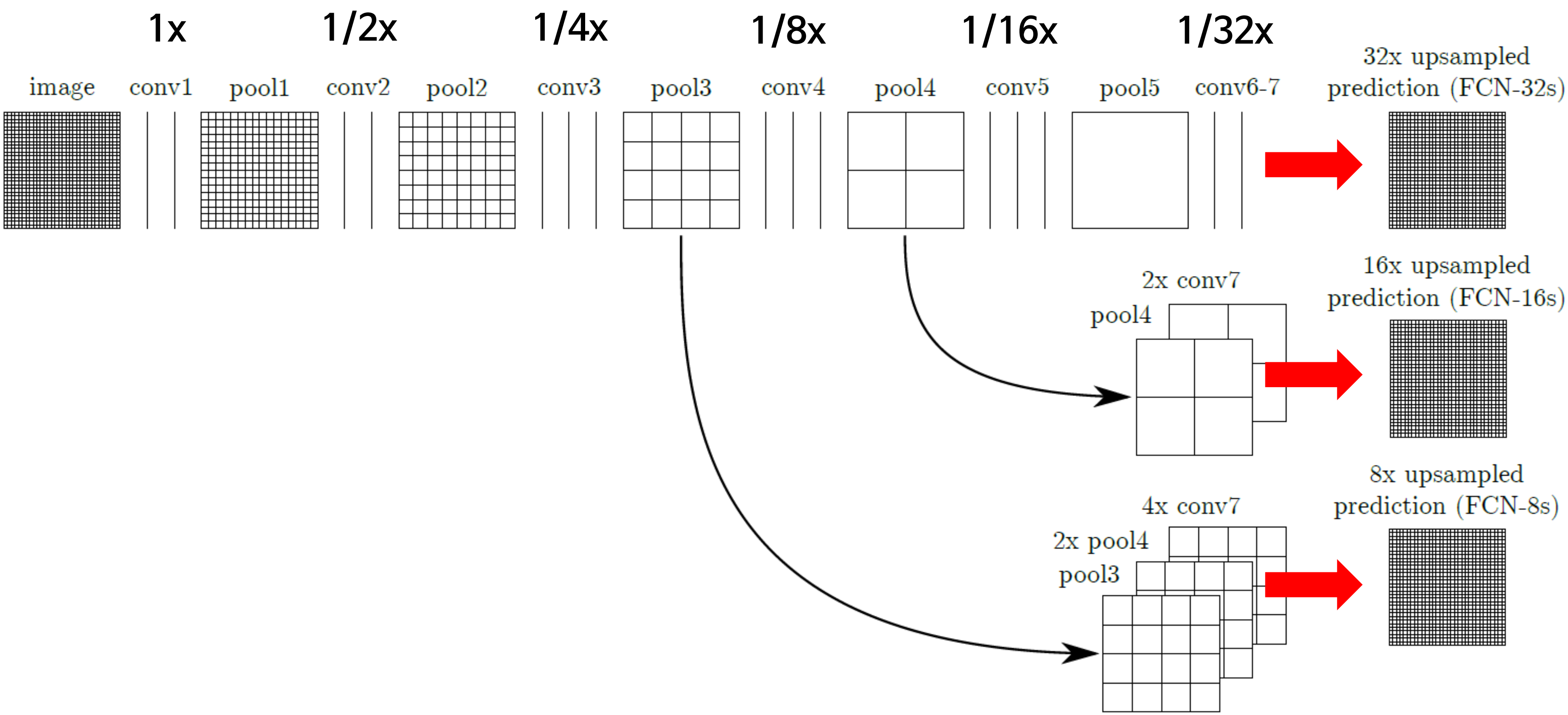
(Blue: Input, Green: Output)

- Convolution is a process getting the output size smaller
- Thus, the name, deconvolution, is coming from when we want to have upsampling to get the output size larger(But the name, deconvolution, is misinterpreted as reverse process of convolution, but it is not)
- And it is also called, **up convolution, and transposed convolution**
- And it is also called **fractional stride convolution** when fractional stride is used

FCN – CNN(AlexNet, VGG, GoogLeNet)

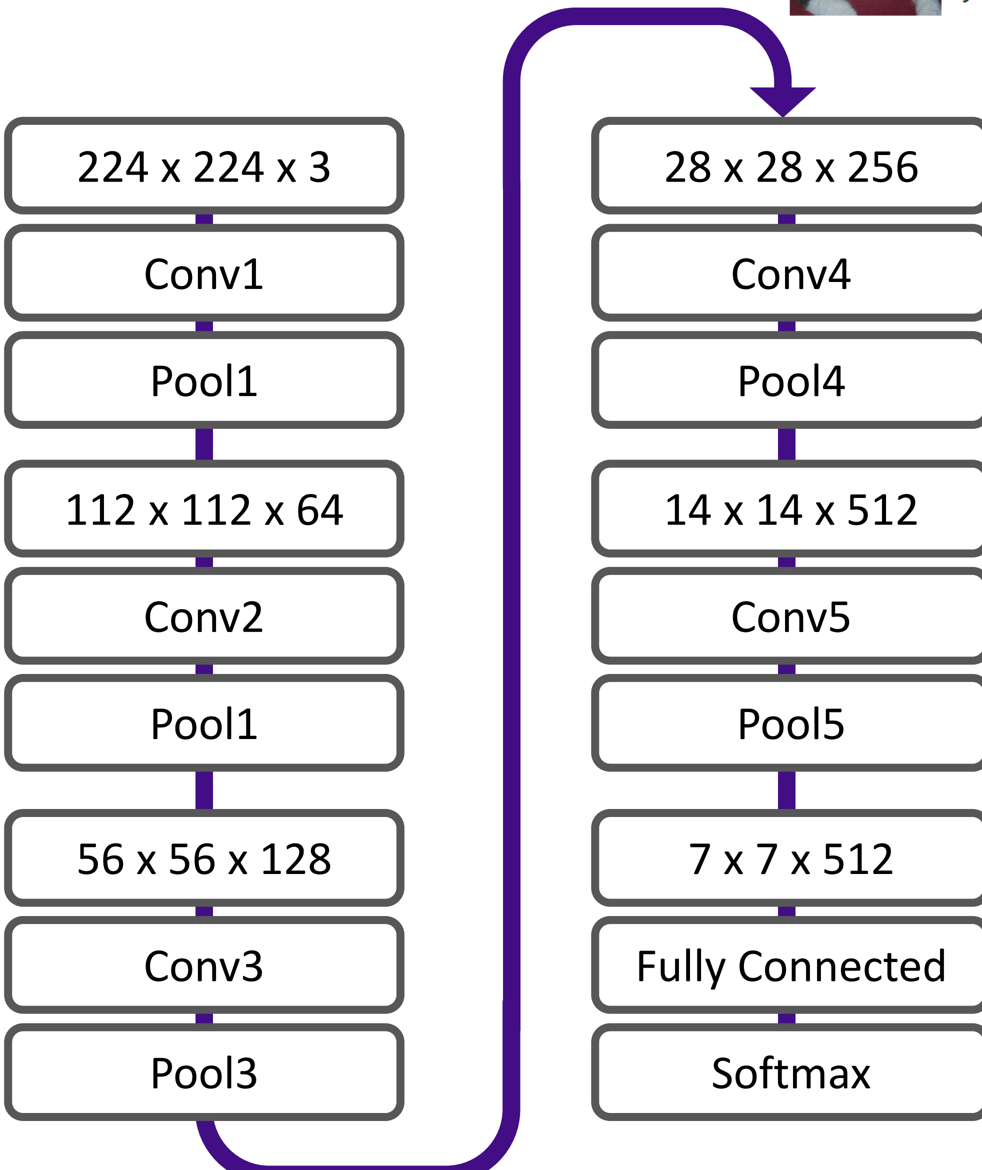
	FCN-AlexNet	FCN-VGG16	FCN-GoogLeNet[4]
mean IU	39.8	**56.0**	42.5
forward time	50 ms	210 ms	59 ms
conv. layers	8	16	22
parameters	57M	134M	6M
rf size	355	404	907
max stride	32	32	32

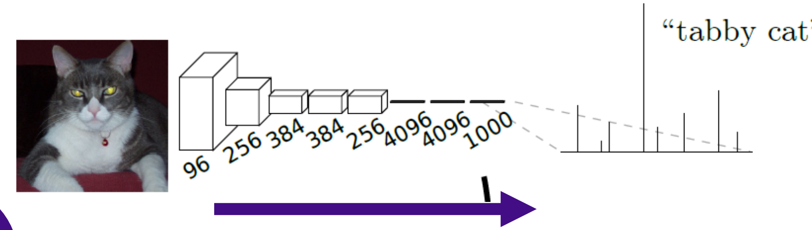
Skip Connection



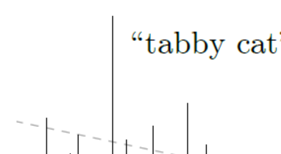
32x up-sample back

Classification

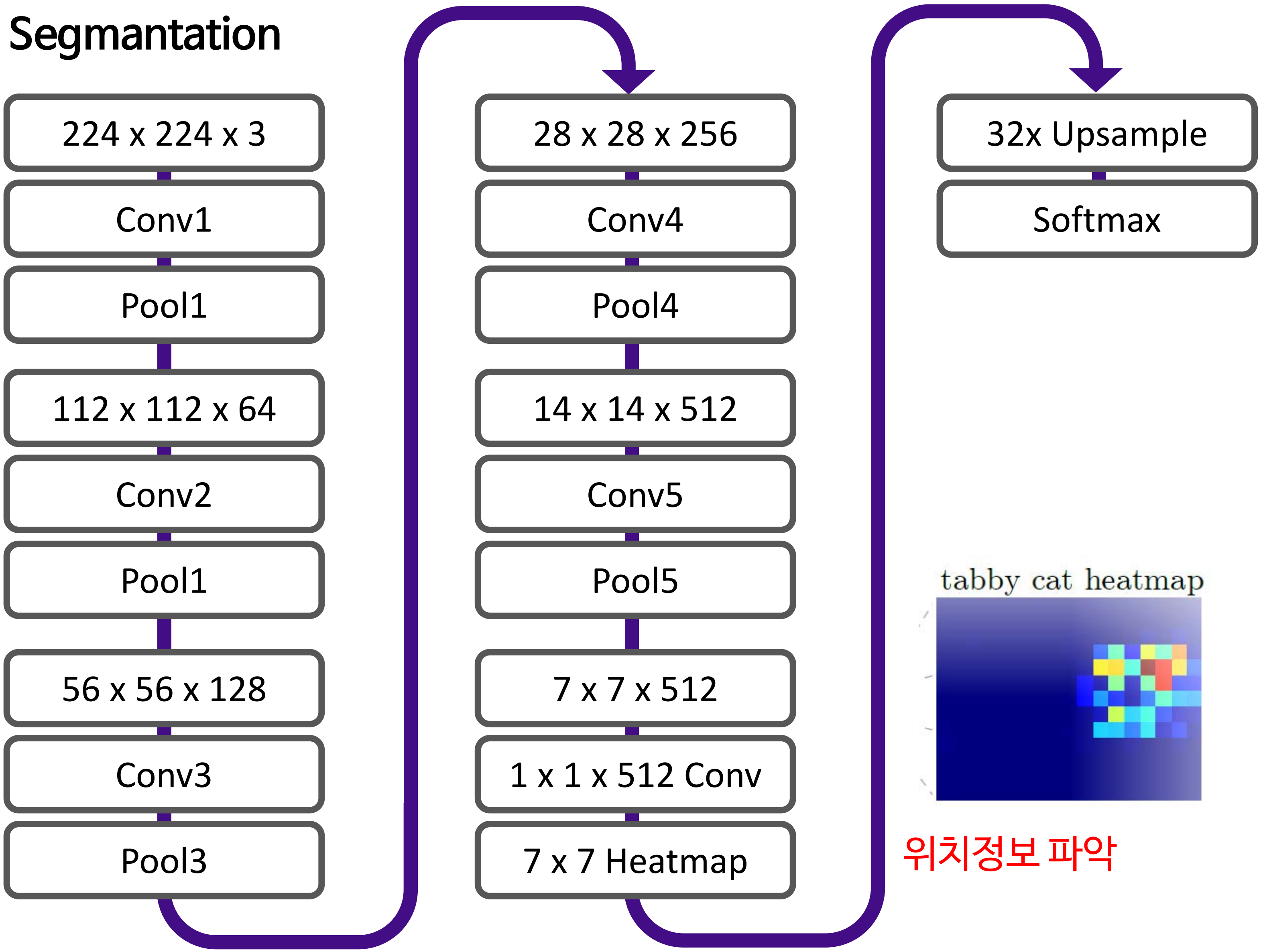




위치정보 소실



Segmentation



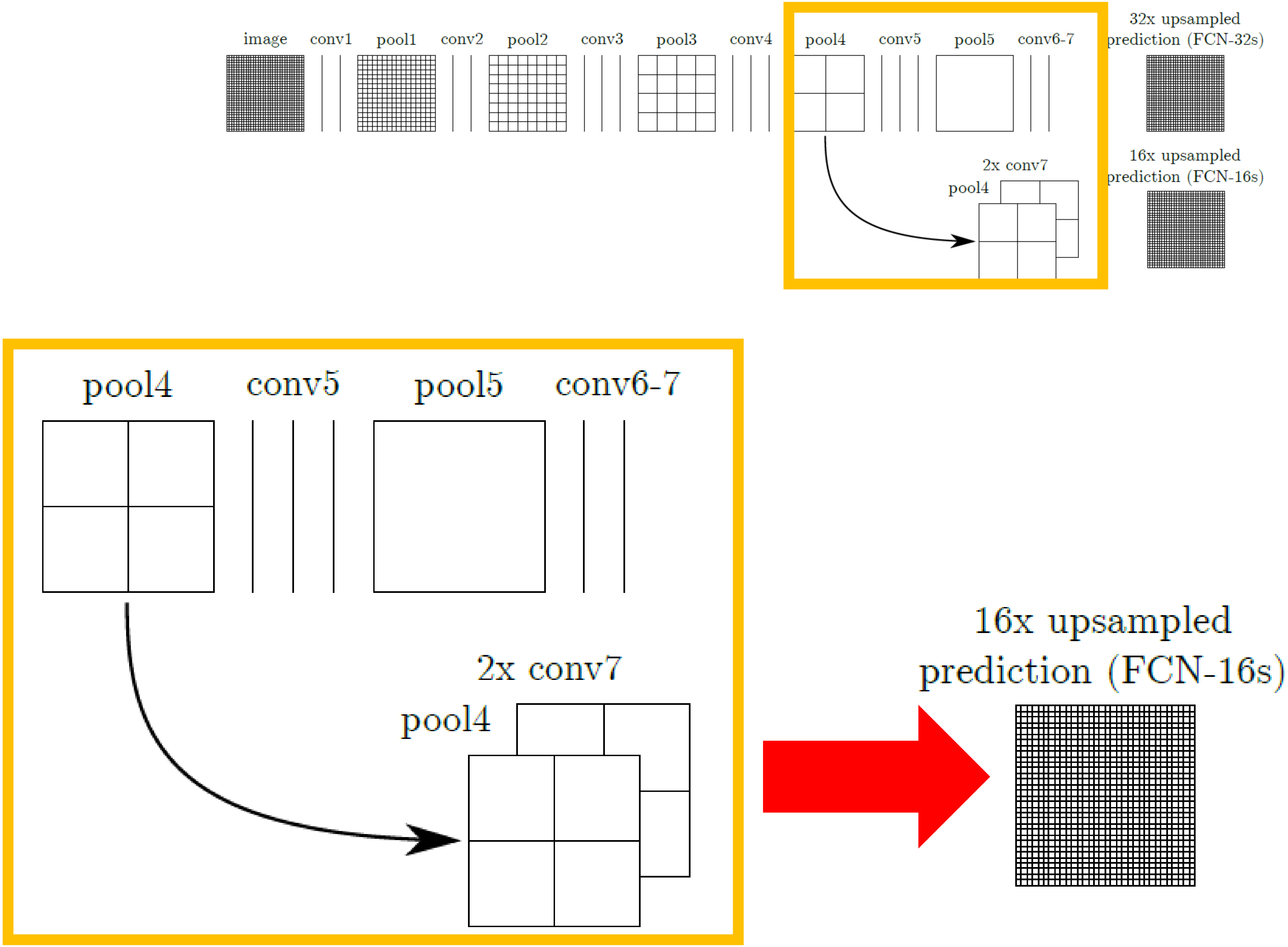


image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
32x upsampled
prediction (FCN-32s)
2x conv7
pool4
16x upsampled
prediction (FCN-16s)
pool4
conv5
pool5
conv6-7
2x conv7
pool4
16x upsampled
prediction (FCN-16s)

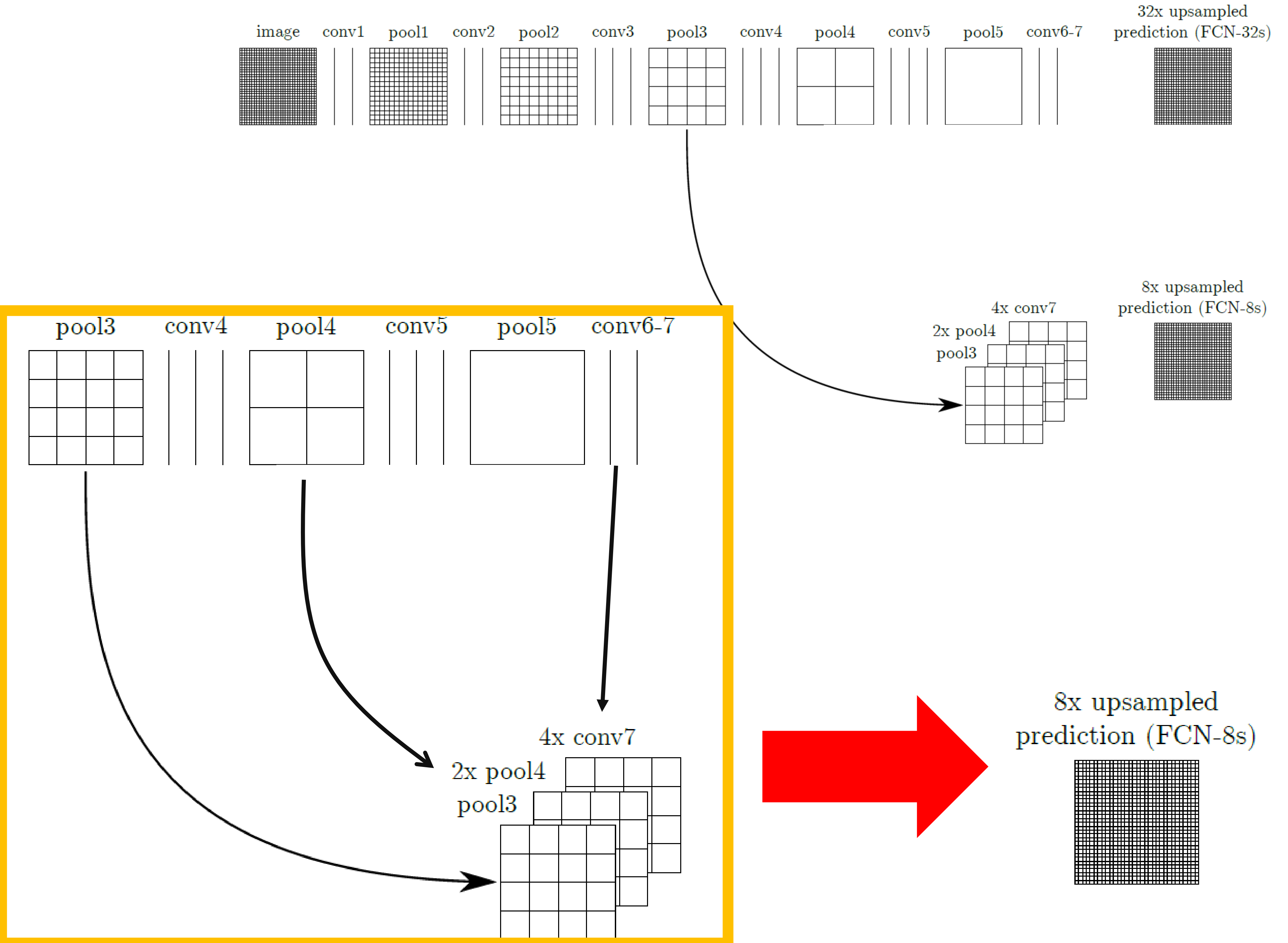
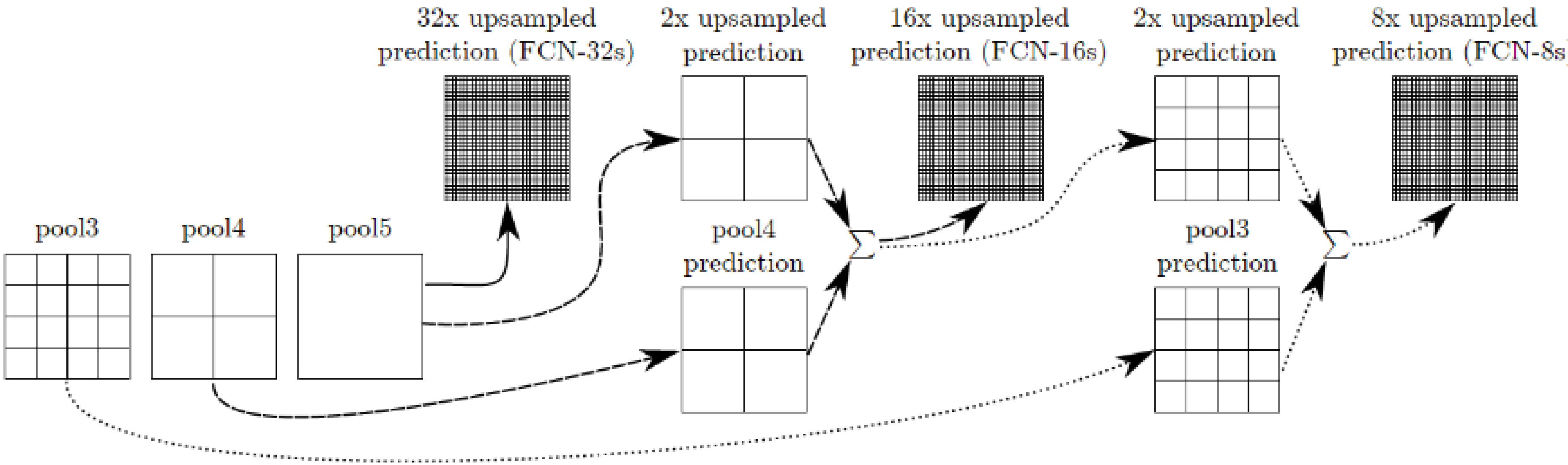


image
conv1
pool1
conv2
pool2
conv3
pool3
conv4
pool4
conv5
pool5
conv6-7
32x upsampled prediction (FCN-32s)
4x conv7
2x pool4
pool3
8x upsampled prediction (FCN-8s)
pool3
conv4
pool4
conv5
pool5
conv6-7
4x conv7
2x pool4
pool3
8x upsampled prediction (FCN-8s)

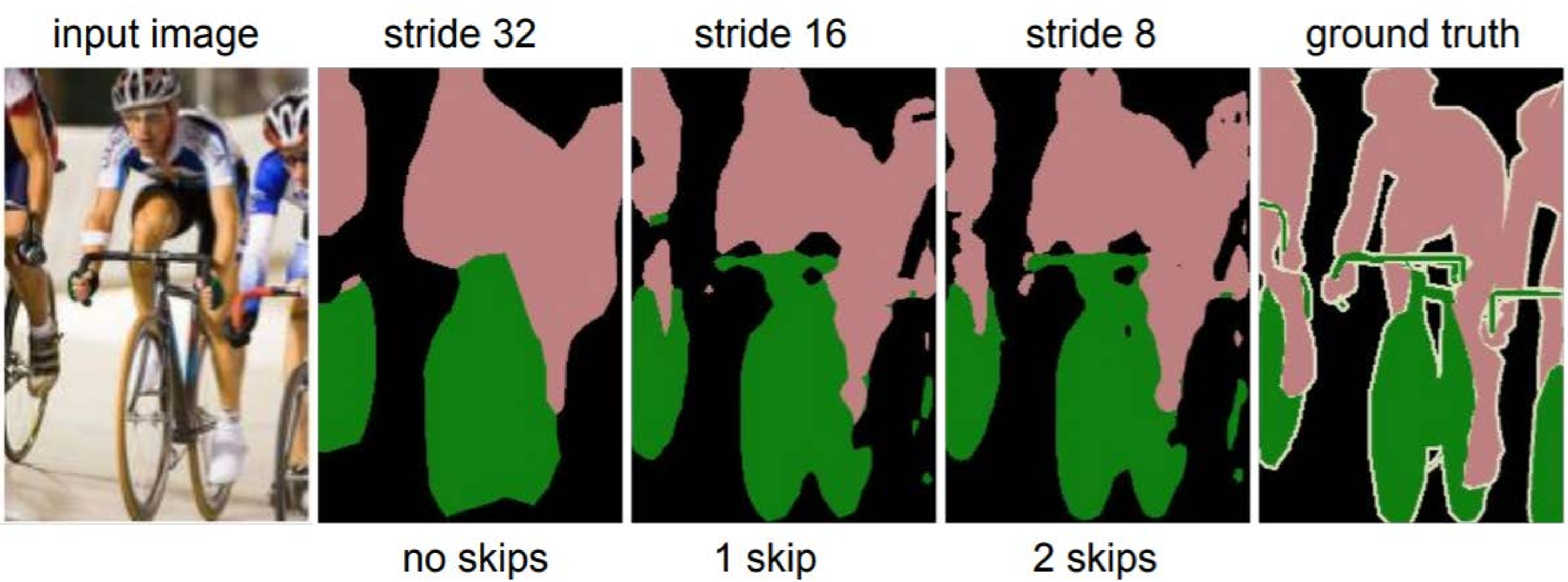


Fusing for FCN-16s and FCN-8s

Comparison of skip FCNs

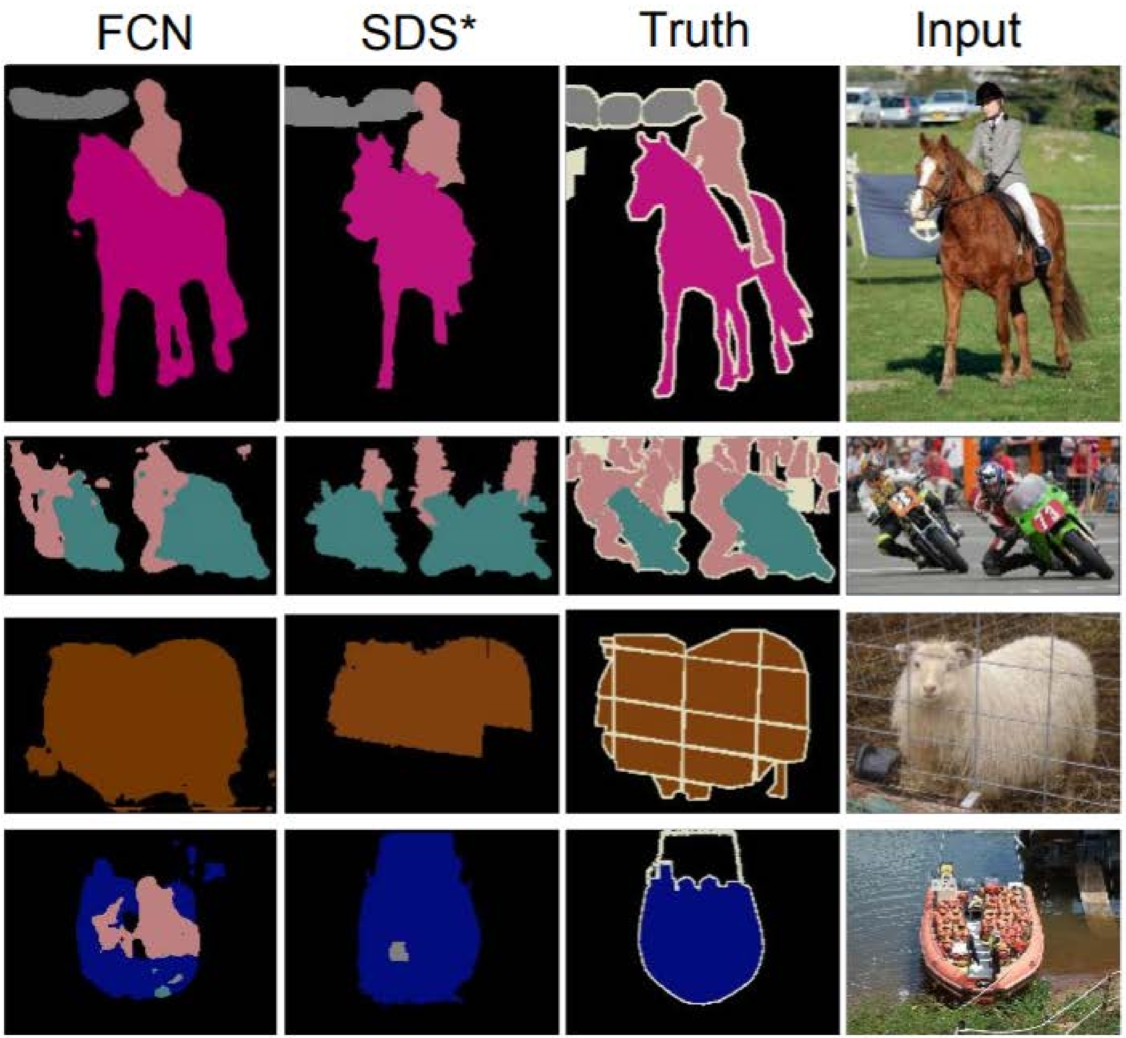
on a subset of PASCAL VOC 2011

	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s-fixed	83.0	59.7	45.4	72.0
FCN-32s	89.1	73.3	59.4	81.4
FCN-16s	90.0	75.7	62.4	83.0
FCN-8s	**90.3**	**75.9**	**62.7**	**83.2**



Fully convolutional segmentation nets

state-of-the-art performace on PASCAL



	pixel acc.	mean acc.	mean IU	f.w. IU	geom. acc.
Liu *et al.* [25]	76.7	-	-	-	-
Tighe *et al.* [36]	-	-	-	-	90.8
Tighe *et al.* [37] 1	75.6	41.1	-	-	-
Tighe *et al.* [37] 2	78.6	39.2	-	-	-
Farabet *et al.* [9] 1	72.3	50.8	-	-	-
Farabet *et al.* [9] 2	78.5	29.6	-	-	-
Pinheiro *et al.* [31]	77.7	29.8	-	-	-
FCN-16s	**85.2**	**51.7**	39.5	76.1	**94.3**

Relative to prior state-of-the-art SDS:

- 30% relative improvement for mean IoU

- 286× faster

*Simultaneous Detection and Segmentation Hariharan et al. ECCV14