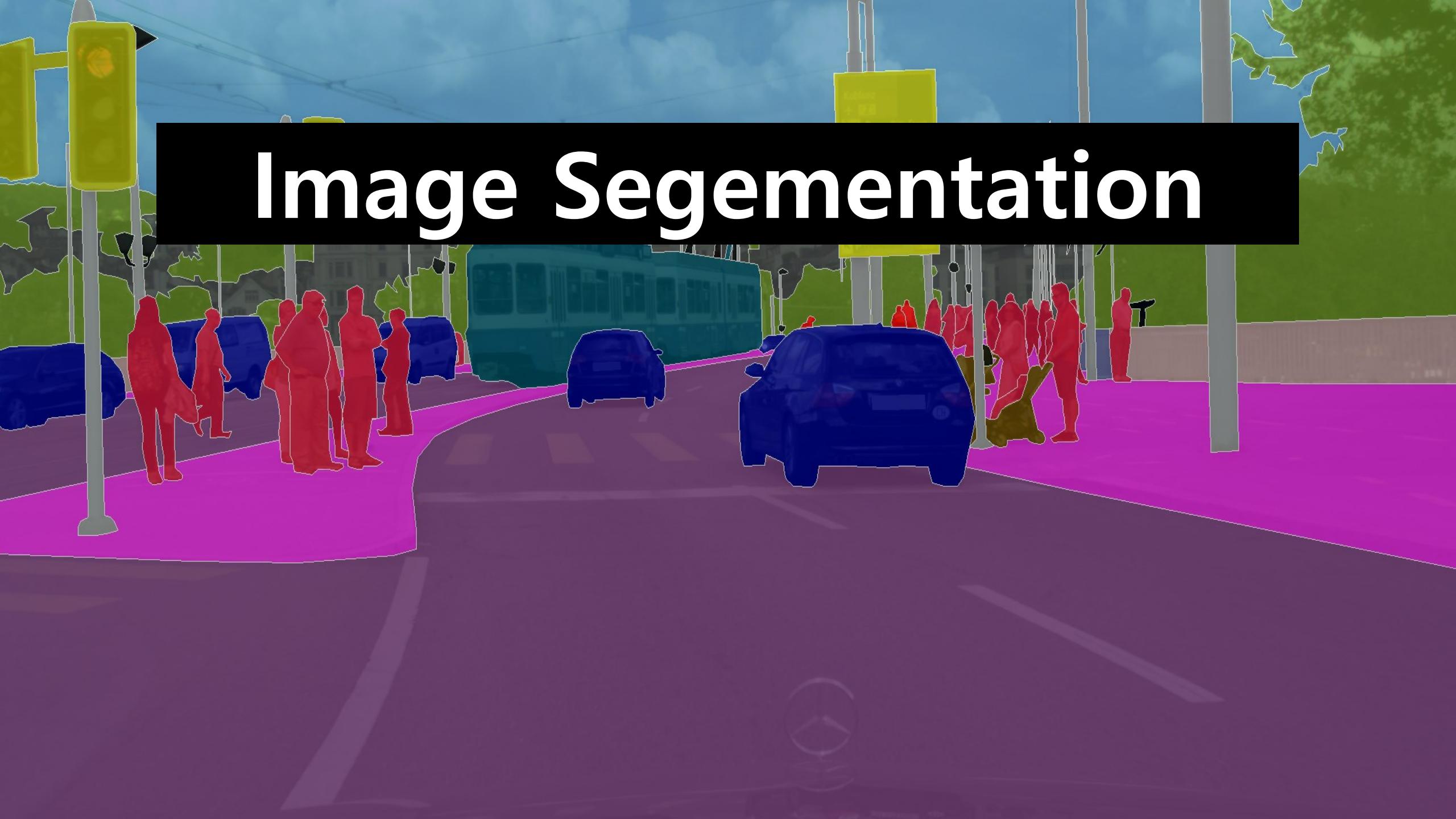


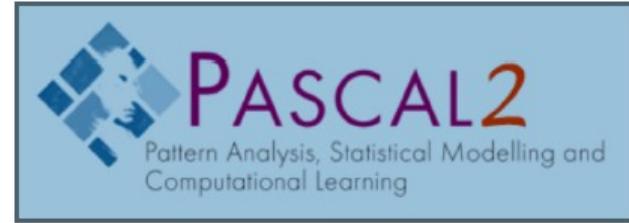
# Image Segmentation



# Semantic Segmentation



- PASCAL VOC segmentation
  - 10K images, 20 classes + bgnd
- MS COCO
  - 100K images, 80 classes + bgnd



# Fully Convolutional Networks for Semantic Segmentation

## Fully Convolutional Networks for Semantic Segmentation

Jonathan Long\*

Evan Shelhamer\*

Trevor Darrell

UC Berkeley

{jonlong, shelhamer, trevor}@cs.berkeley.edu

### Abstract

Convolutional networks are powerful visual models that yield hierarchies of features. We show that convolutional networks by themselves, trained end-to-end, pixels-to-pixels, exceed the state-of-the-art in semantic segmentation. Our key insight is to build “fully convolutional” networks that take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. We define and detail the space of fully convolutional networks, explain their application to spatially dense prediction tasks, and draw connections to prior models. We adapt contemporary classification networks (AlexNet [22],

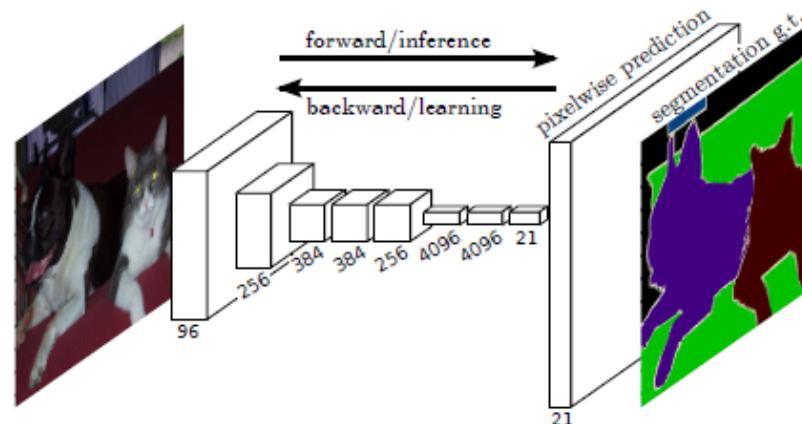
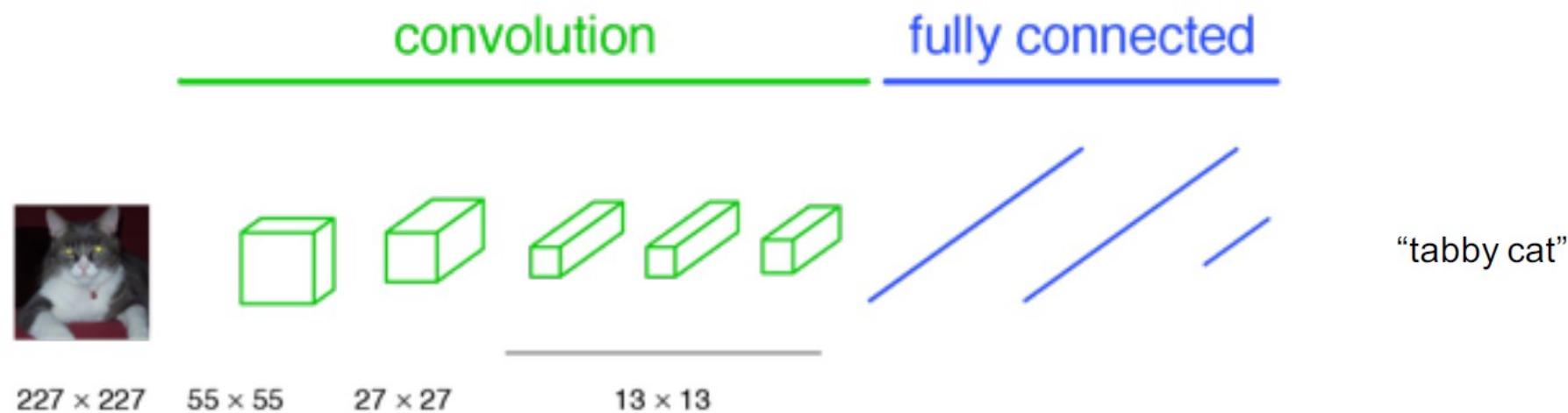


Figure 1. Fully convolutional networks can efficiently learn to make dense predictions for per-pixel tasks like semantic segmentation.

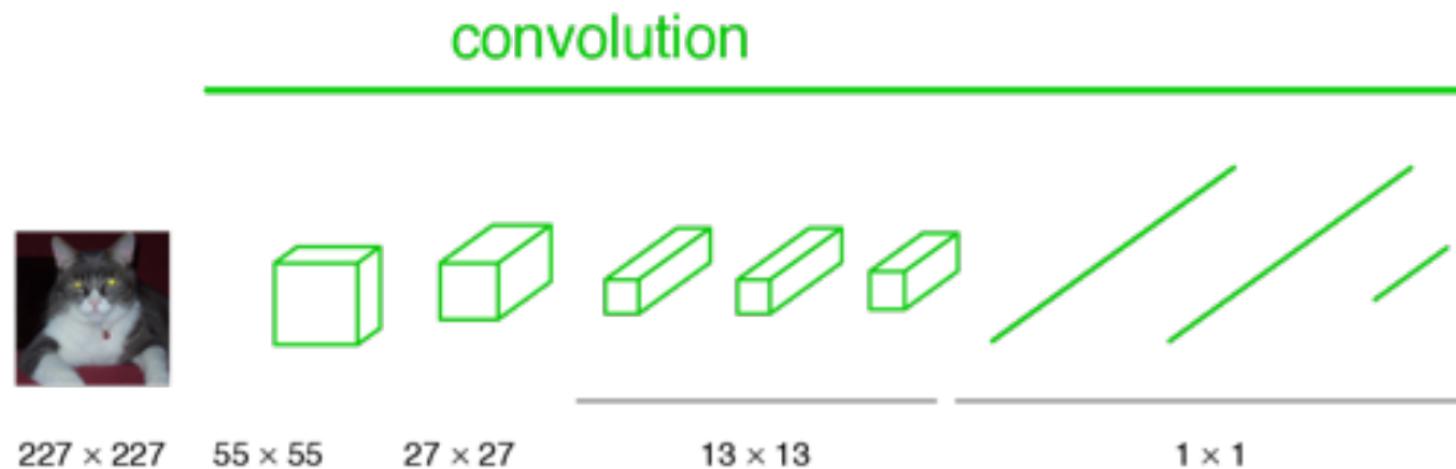
# Fully Convolutional Networks for Semantic Segmentation

- Fully convolutional network(FCN)
  - Pixel-wise prediction with end to end learning
- Fully Convolutional versions of existing networks predict dense outputs from arbitrary-sized inputs

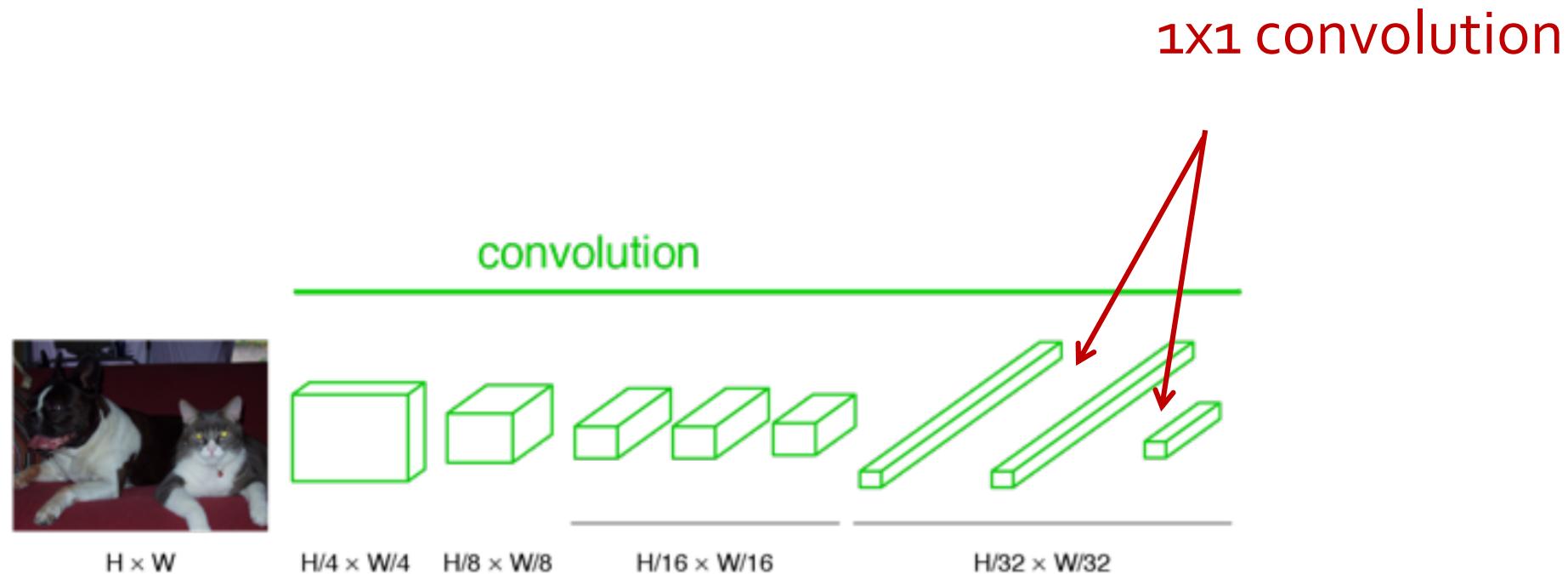
# Classification Network



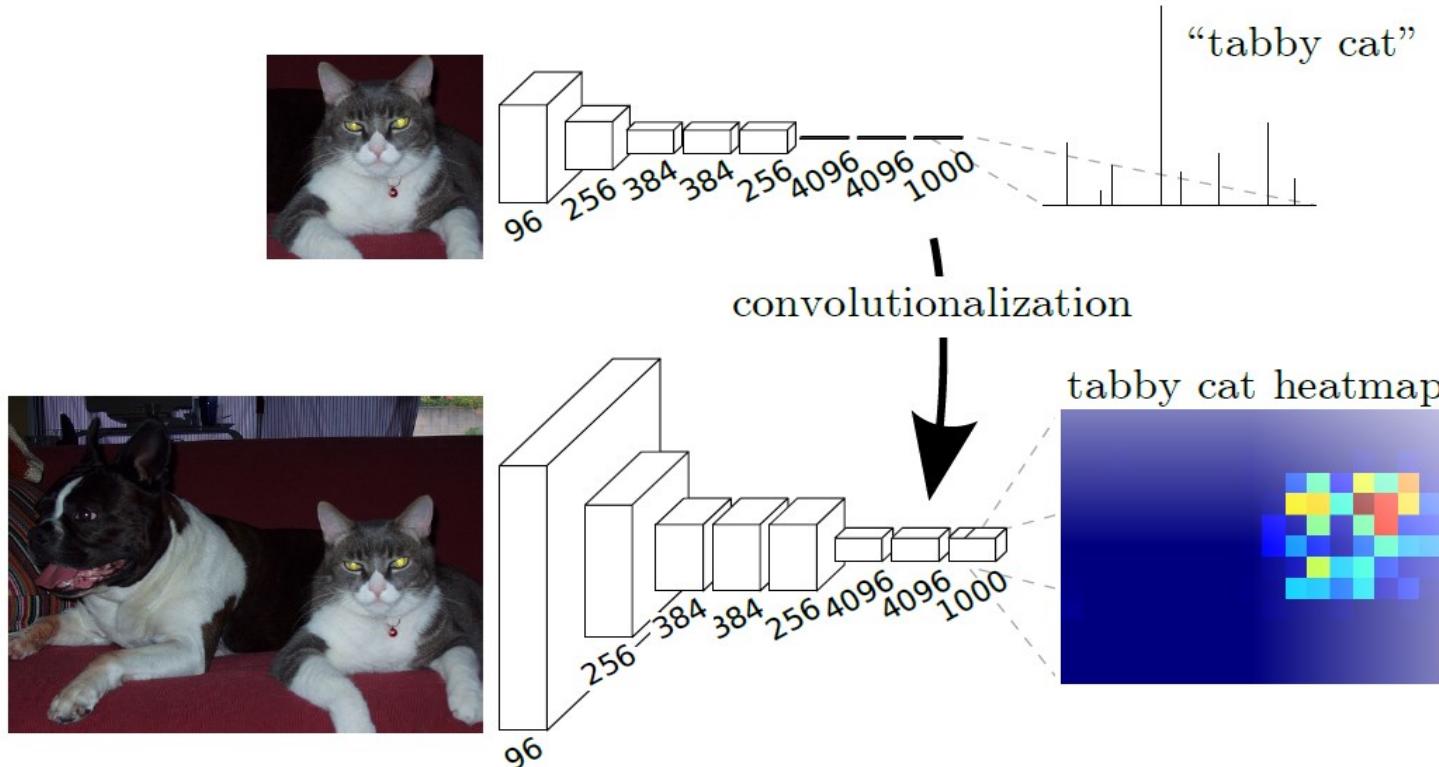
# Becoming Fully Convolutional Network



# Becoming Fully Convolutional Network

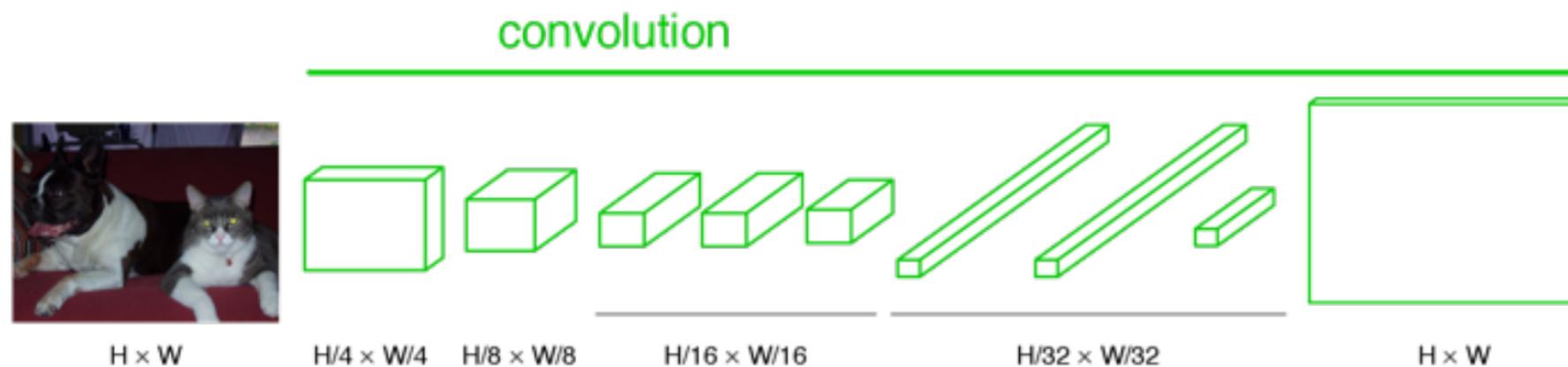


# Heatmap from FCN

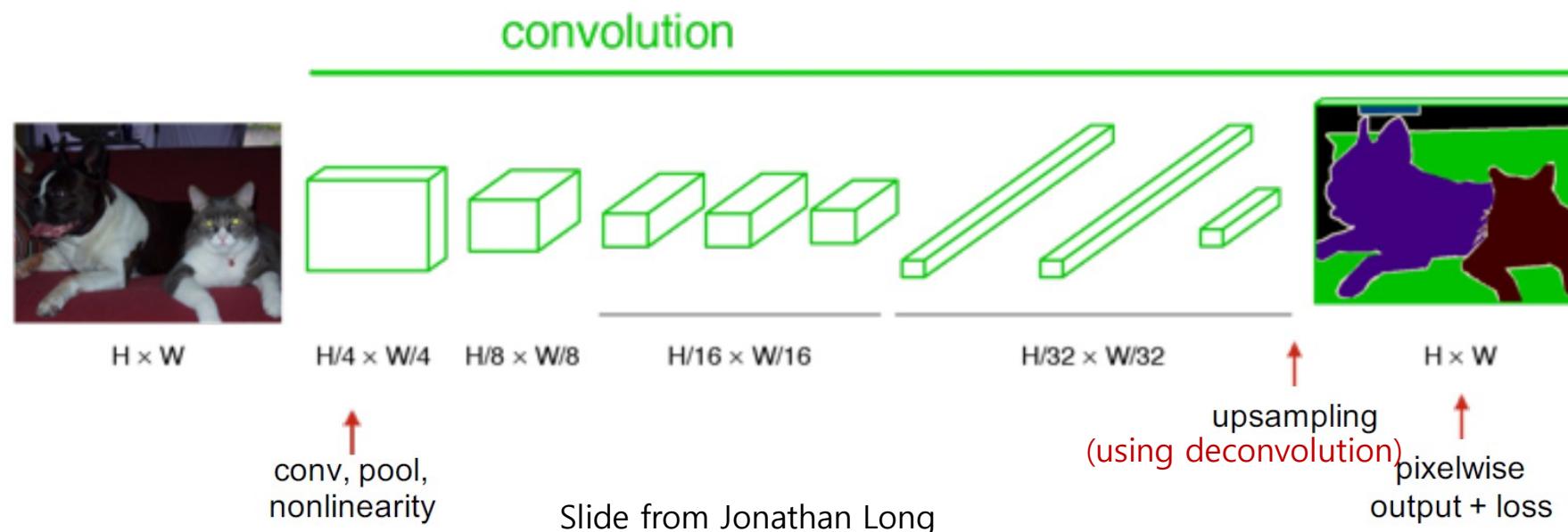


# Upsampling Output

- They use deconvolution for upsampling



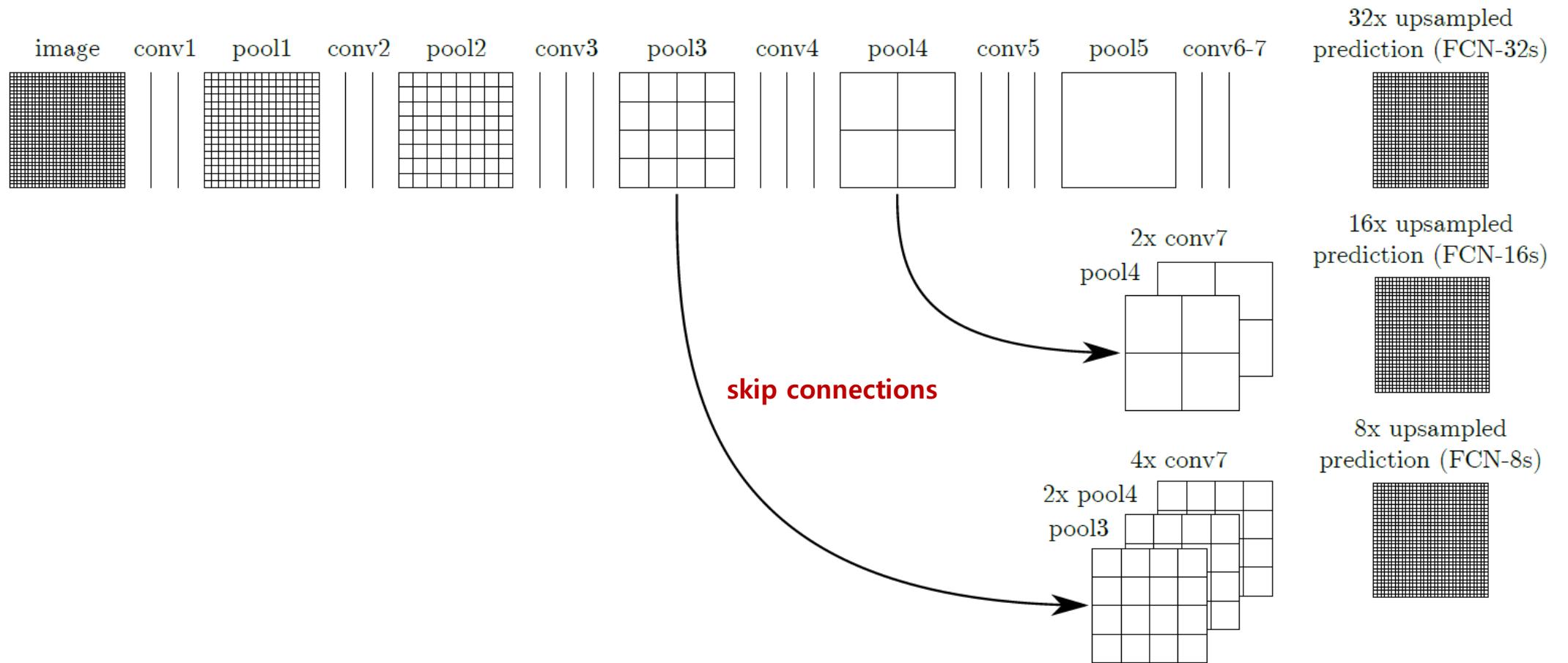
# End to End, Pixels to Pixels Network



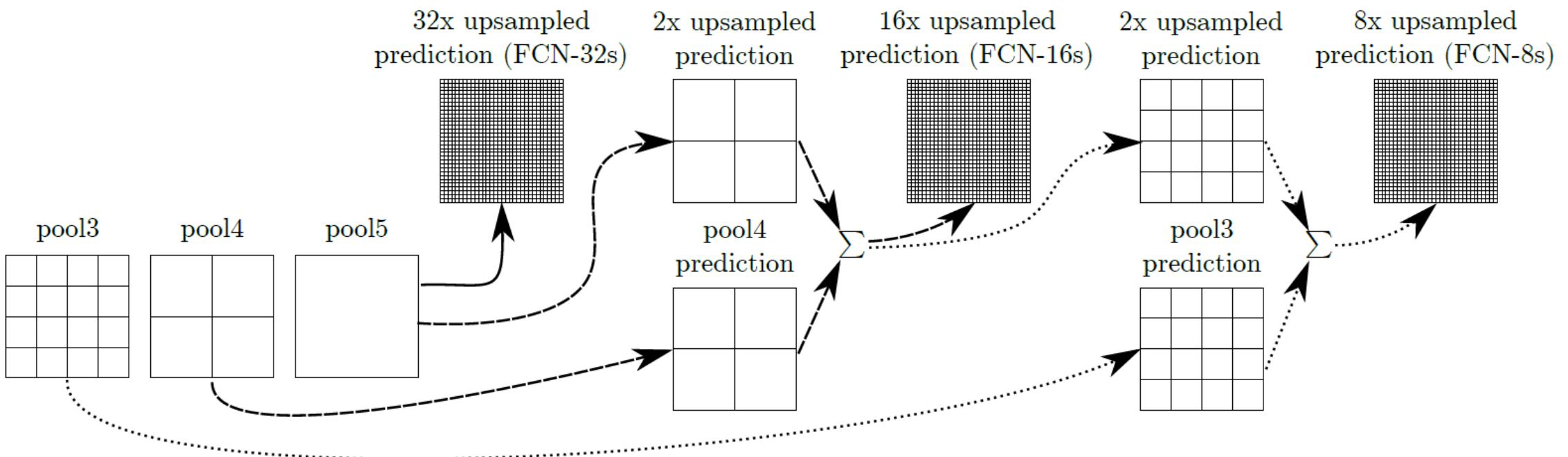
# Classifier to Dense FCN

- Convolutionalize proven classification architectures:
  - AlexNet, VGG, and GoogLeNet (reimplementation)
- Remove classification layer and convert all fully connected layers to convolutions
- Append  $1 \times 1$  convolution with channel dimensions and predict scores at each of the coarse output locations (21 categories for PASCAL)

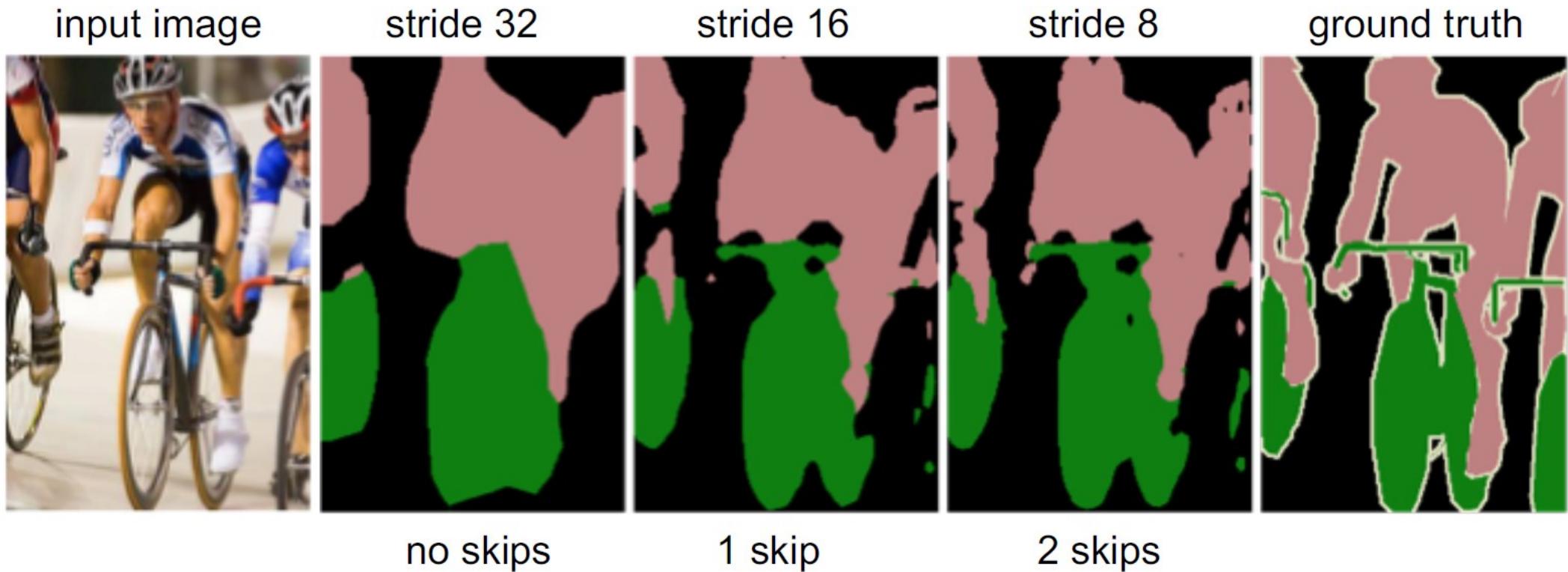
# Skip Layers



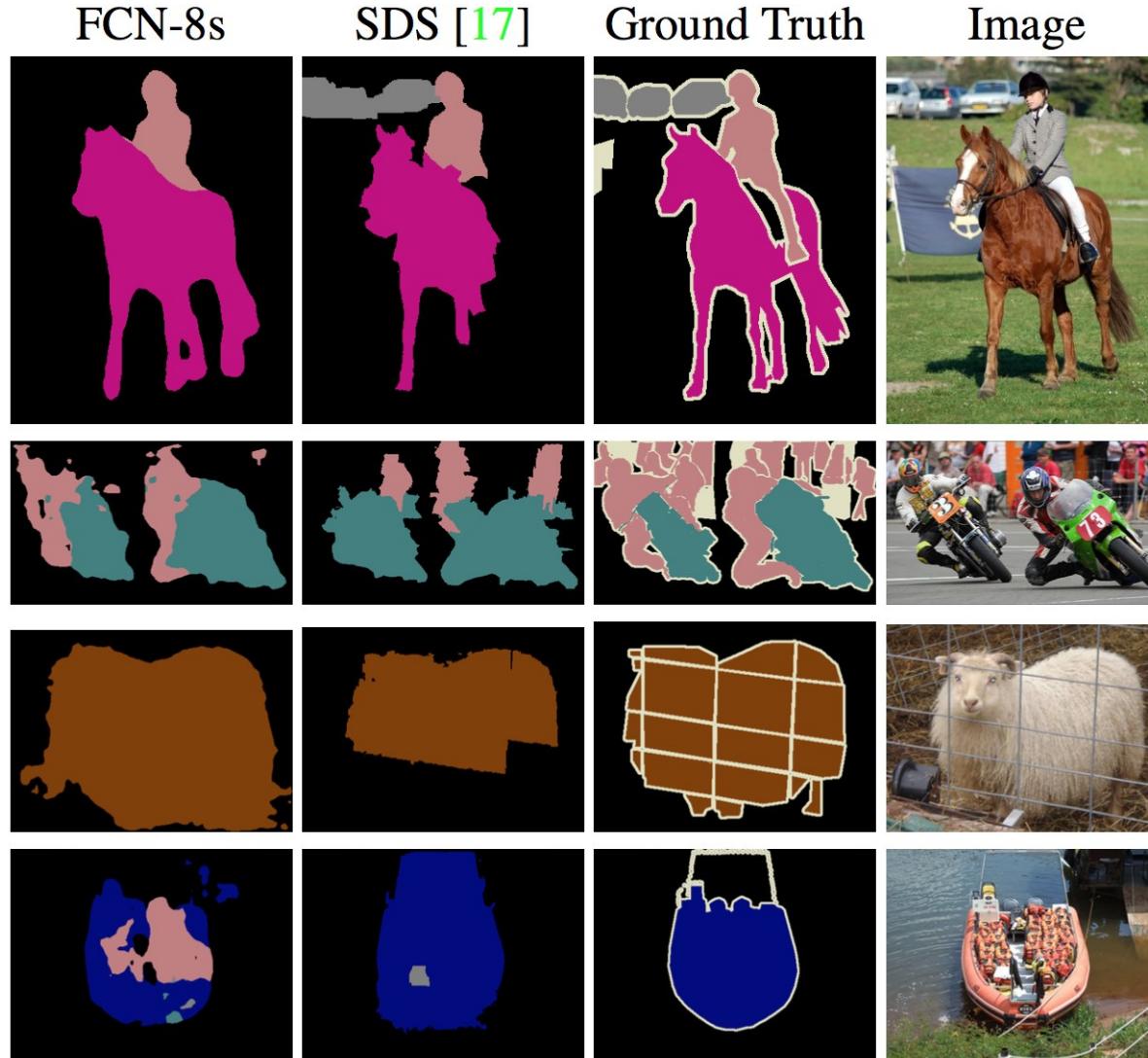
# Skip Layers



# Results



# Results



# Learning Deconvolution Network for Semantic Segmentation

## Learning Deconvolution Network for Semantic Segmentation

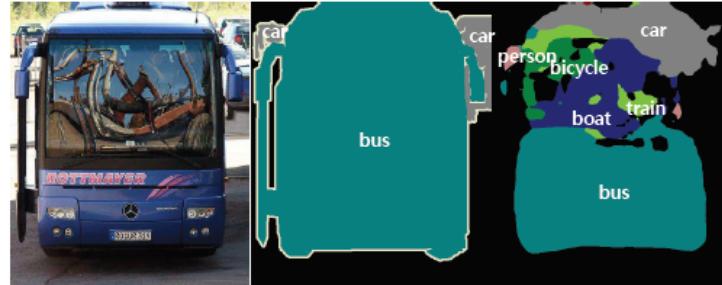
Hyeonwoo Noh

Department of Computer Science and Engineering, POSTECH, Korea

{hyeonwoonoh\_, maga33, bhhan}@postech.ac.kr

### Abstract

We propose a novel semantic segmentation algorithm by learning a deep deconvolution network. We learn the network on top of the convolutional layers adopted from VGG 16-layer net. The deconvolution network is composed of deconvolution and unpooling layers, which identify pixel-wise class labels and predict segmentation masks. We apply the trained network to each proposal in an input image, and construct the final semantic segmentation map by combining the results from all proposals in a simple manner. The proposed algorithm mitigates the limitations of the existing methods based on fully convolutional networks by integrating deep deconvolution network and proposal-wise prediction; our segmentation method typically identifies detailed structures and handles objects in multiple scales naturally. Our network demonstrates outstanding performance



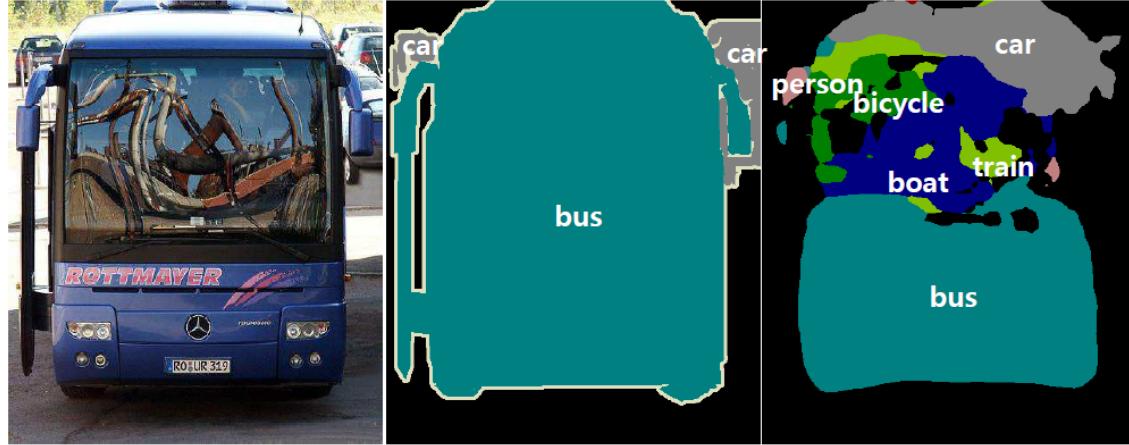
(a) Inconsistent labels due to large object size



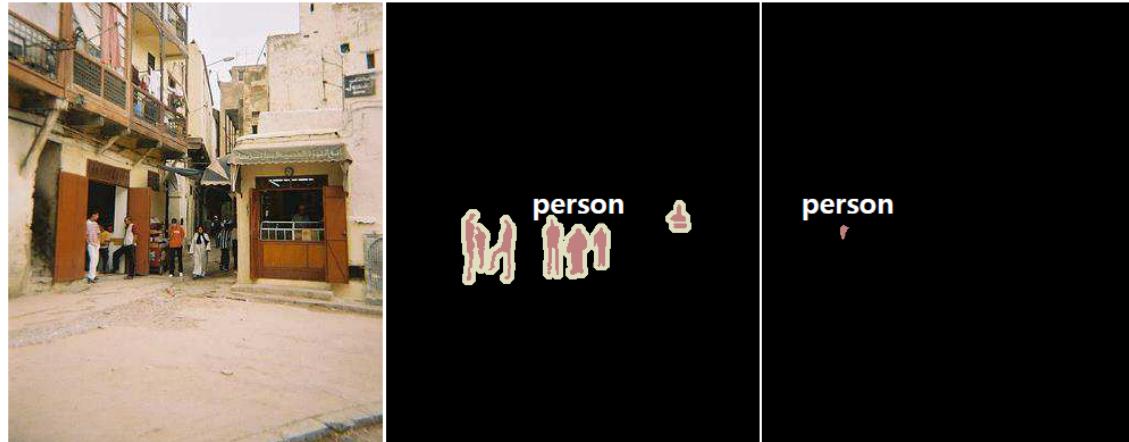
(b) Missing labels due to small object size

# FCN's Limitation

- FCN has a predefined fixed-size receptive field. Therefore, the object that is substantially larger or smaller than the receptive field may be fragmented or mislabeled
- The detailed structures of an object are often lost or smoothed because the label map is too coarse and deconvolution procedure is overly simple



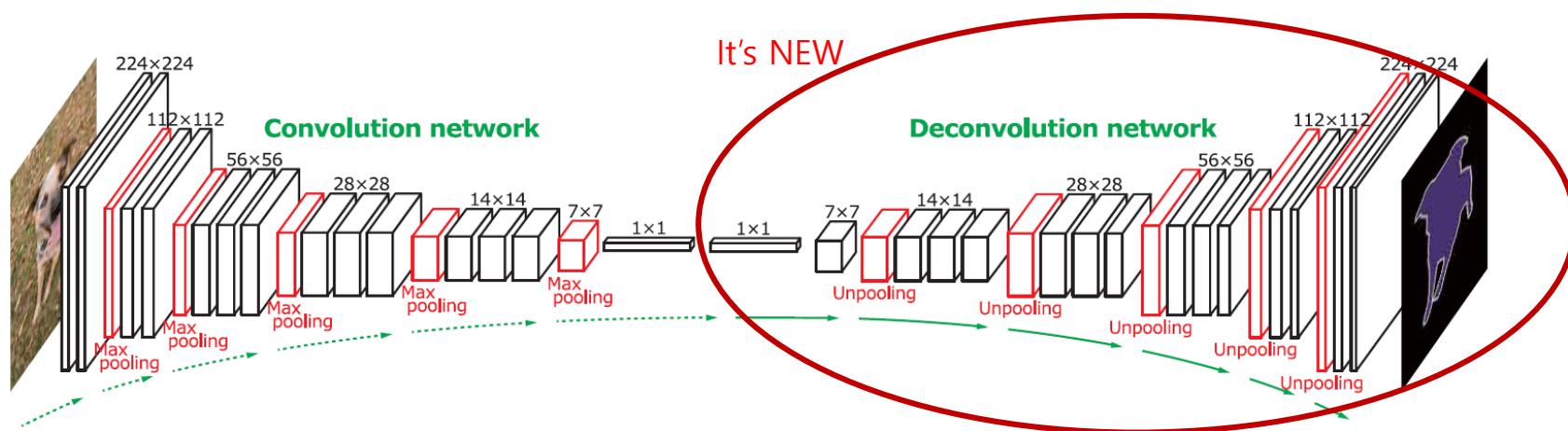
(a) Inconsistent labels due to large object size



(b) Missing labels due to small object size

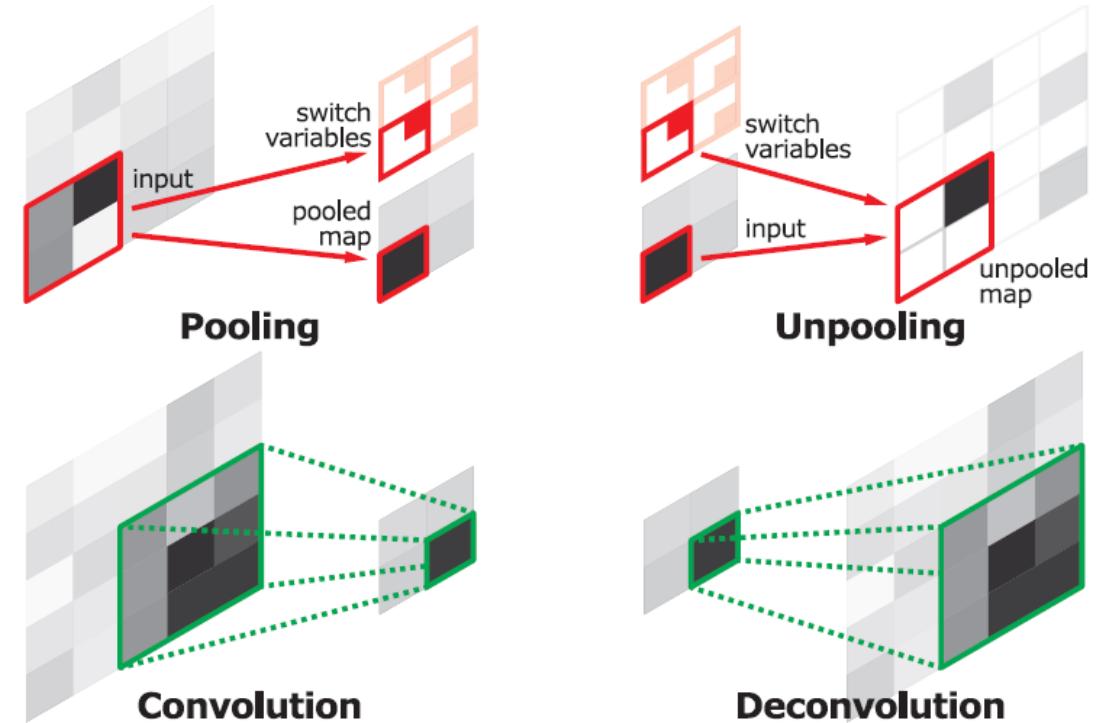
# Deconvolution Network

- To address such issues, Use “**Deconvolution**”!
- Convolution network extract features(VGG-16)
- Deconvolution network generate probability map(same size to input image)
- Probability map indicates probability of each pixel belongs to one of class



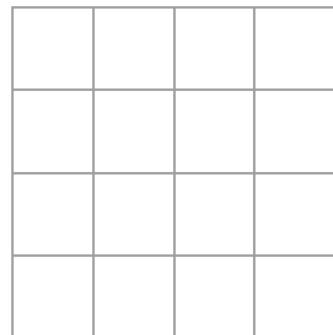
# Unpooling and Deconvolution

- Unpooling
  - Reconstruct structure of original activation map
  - Activation size is preserved but still sparse
- Deconvolution
  - Densify sparse activation map

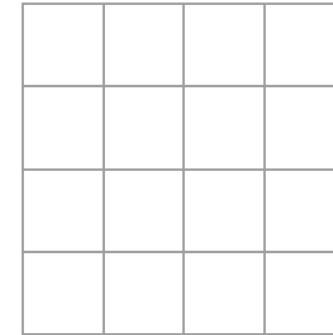


# Learnable Upsampling: “Deconvolution”

Typical  $3 \times 3$  convolution, stride 1 pad 1



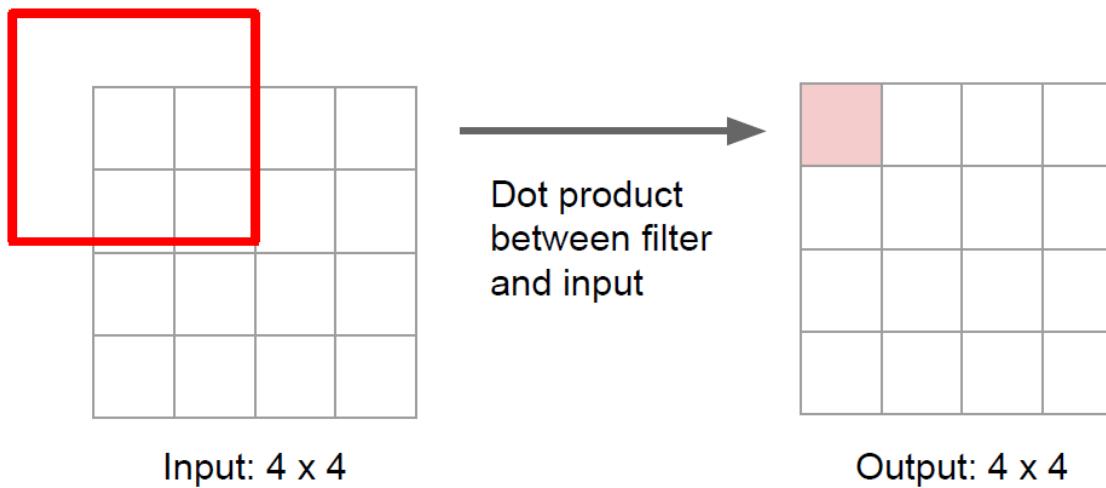
Input:  $4 \times 4$



Output:  $4 \times 4$

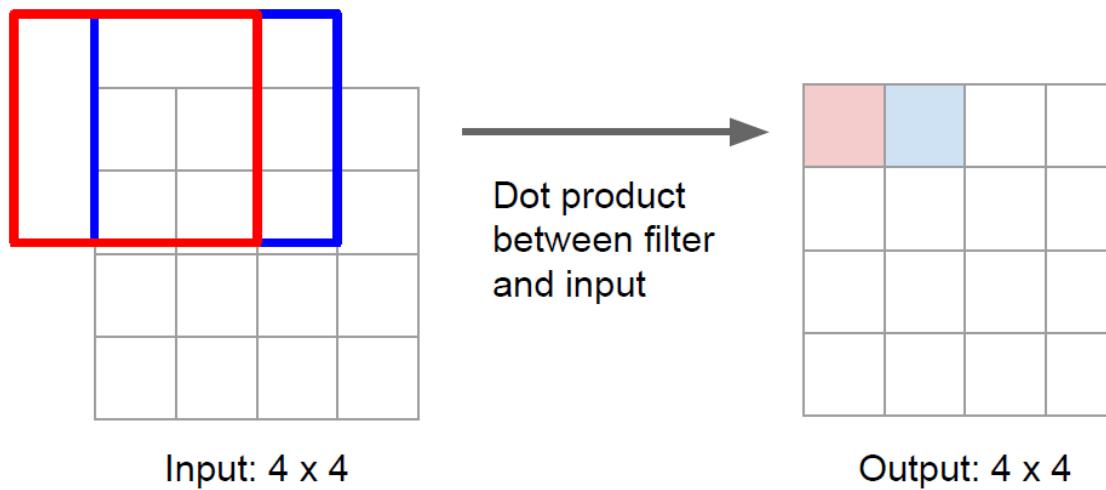
# Learnable Upsampling: “Deconvolution”

Typical  $3 \times 3$  convolution, stride 1 pad 1



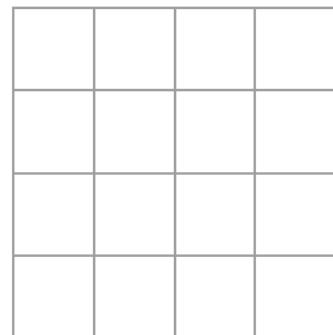
# Learnable Upsampling: “Deconvolution”

Typical  $3 \times 3$  convolution, stride 1 pad 1

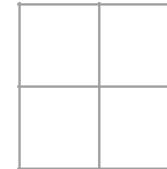


# Learnable Upsampling: “Deconvolution”

Typical  $3 \times 3$  convolution, **stride 2** pad 1



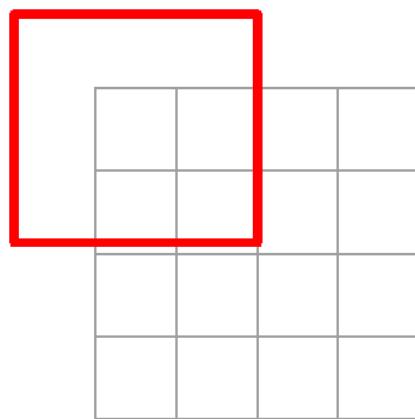
Input:  $4 \times 4$



Output:  $2 \times 2$

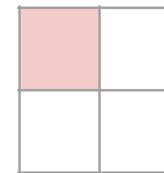
# Learnable Upsampling: “Deconvolution”

Typical  $3 \times 3$  convolution, stride 2 pad 1



Input:  $4 \times 4$

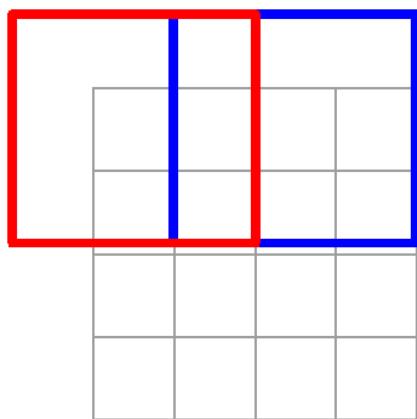
Dot product  
between filter  
and input



Output:  $2 \times 2$

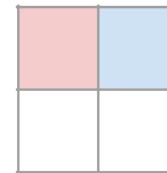
# Learnable Upsampling: “Deconvolution”

Typical  $3 \times 3$  convolution, stride 2 pad 1



Input:  $4 \times 4$

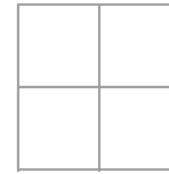
Dot product  
between filter  
and input



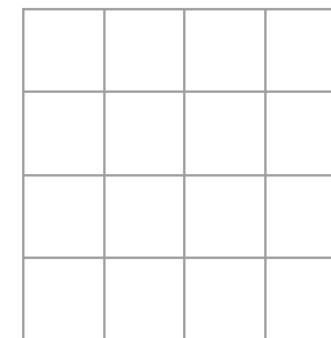
Output:  $2 \times 2$

# Learnable Upsampling: “Deconvolution”

3 x 3 “deconvolution”, stride 2 pad 1

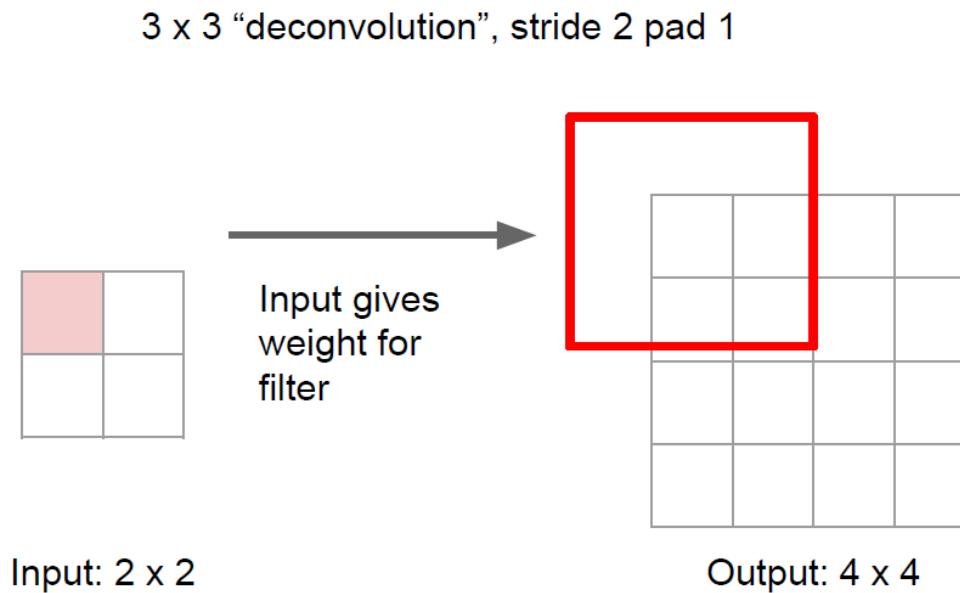


Input: 2 x 2

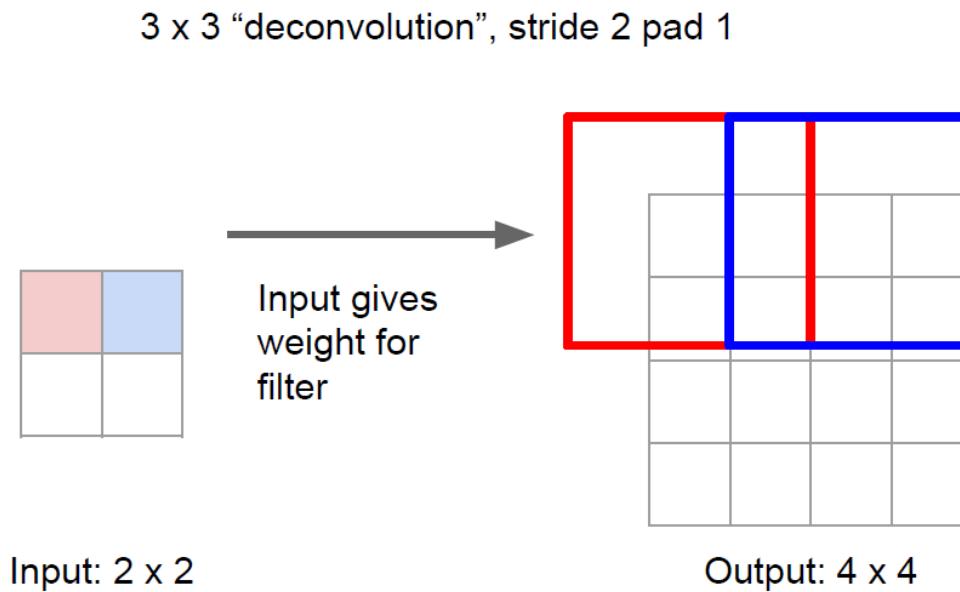


Output: 4 x 4

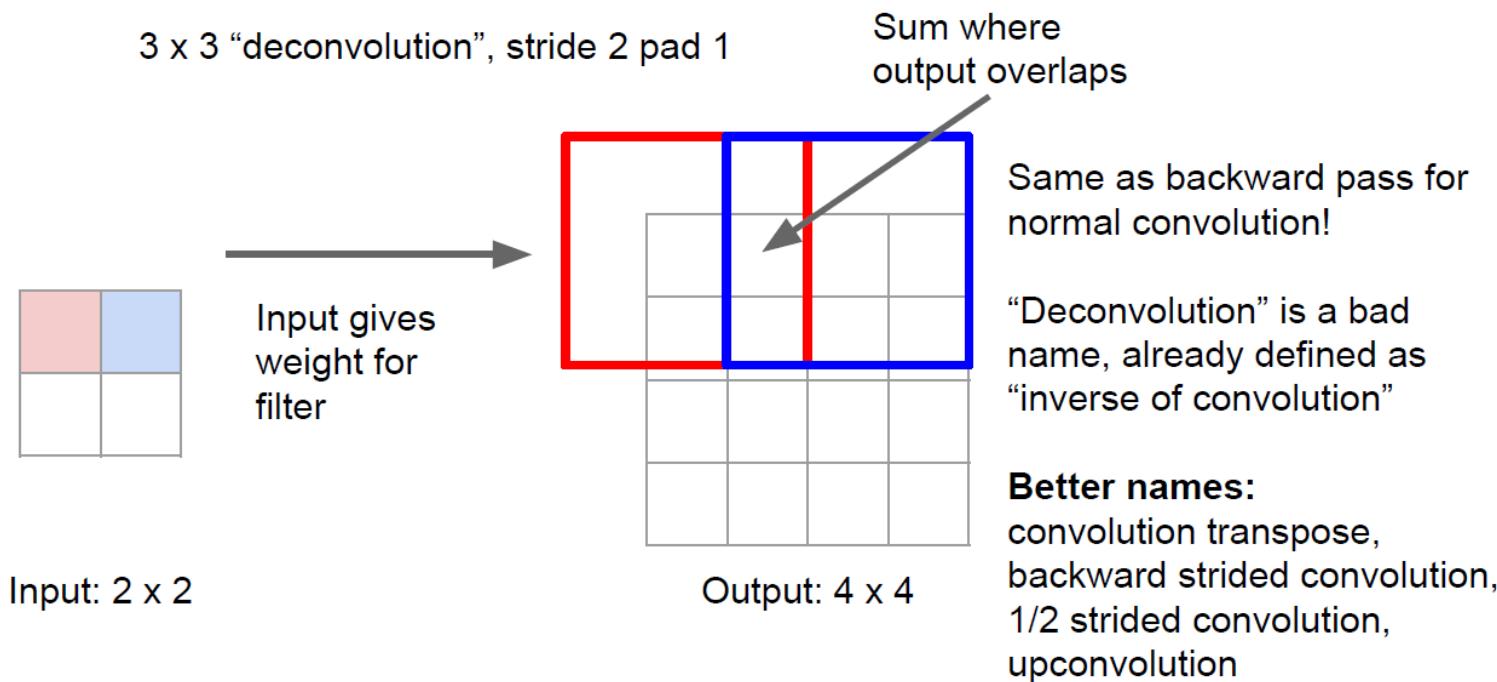
# Learnable Upsampling: “Deconvolution”



# Learnable Upsampling: “Deconvolution”



# Learnable Upsampling: “Deconvolution”



# Analysis of DeconvNet

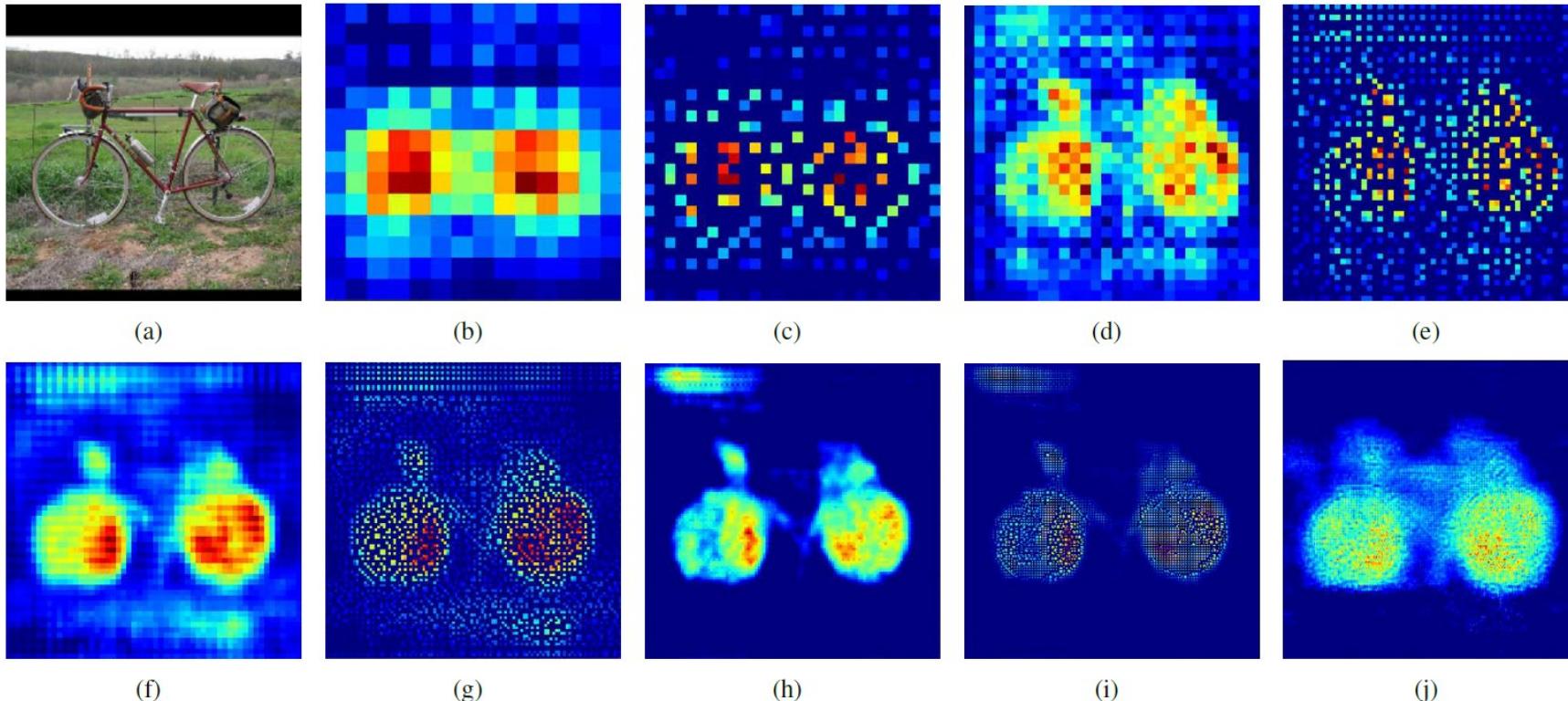
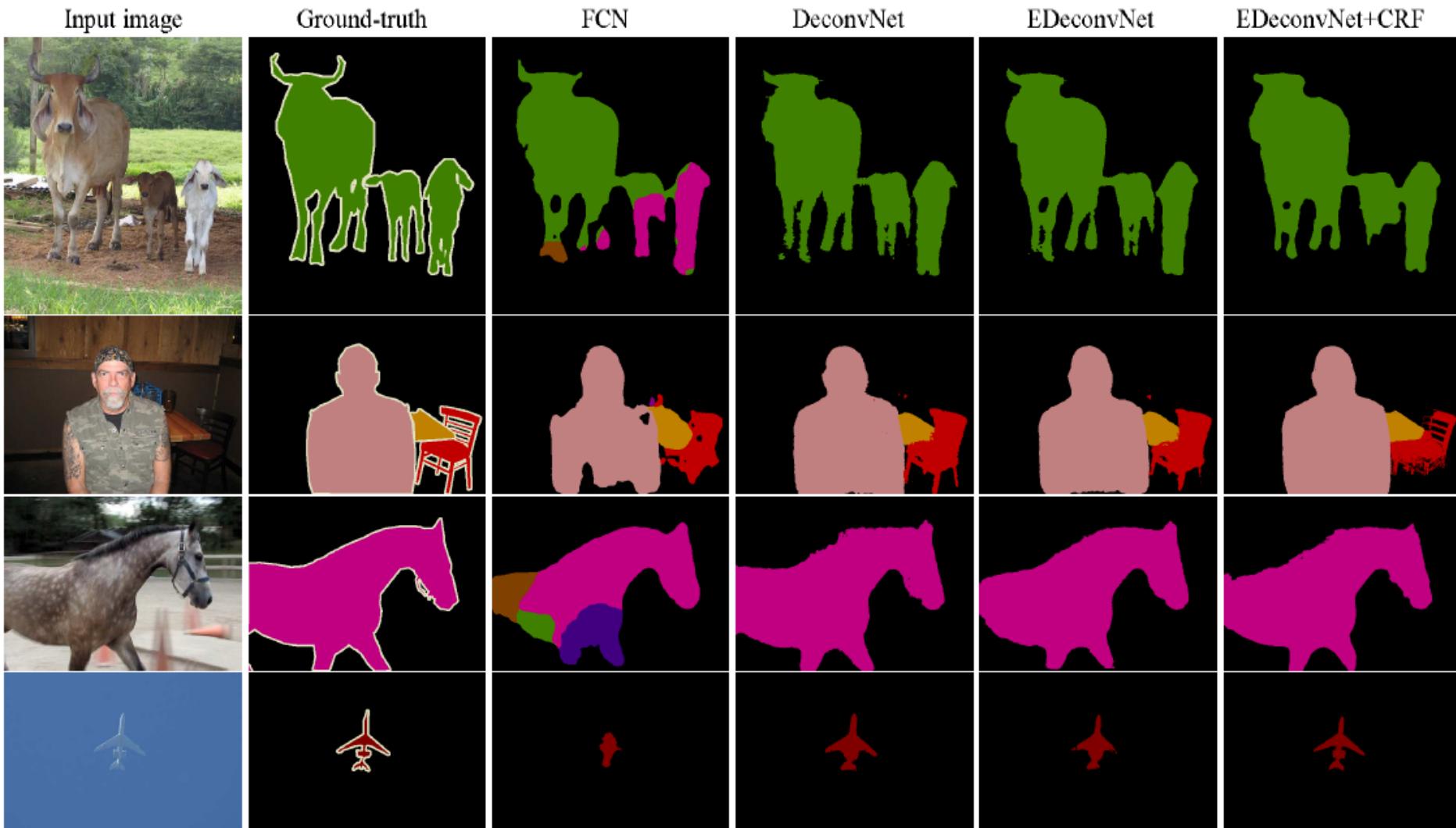


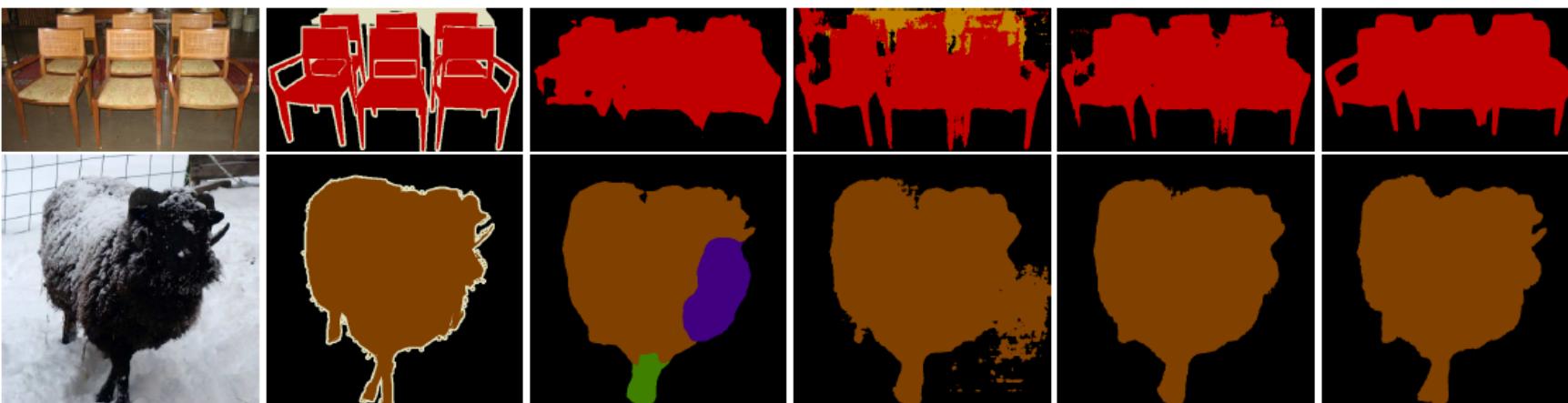
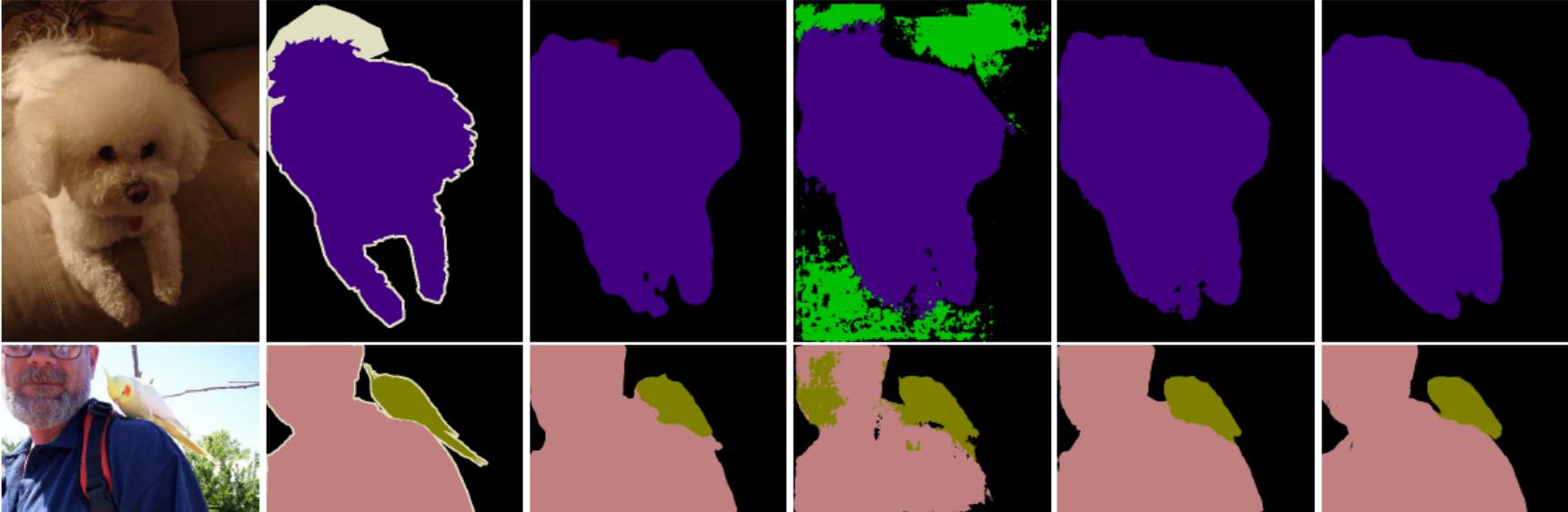
Figure 4. Visualization of activations in our deconvolution network. The activation maps from (b) to (j) correspond to the output maps from lower to higher layers in the deconvolution network. We select the most representative activation in each layer for effective visualization. The image in (a) is an input, and the rest are the outputs from (b) the last  $14 \times 14$  deconvolutional layer, (c) the  $28 \times 28$  unpooling layer, (d) the last  $28 \times 28$  deconvolutional layer, (e) the  $56 \times 56$  unpooling layer, (f) the last  $56 \times 56$  deconvolutional layer, (g) the  $112 \times 112$  unpooling layer, (h) the last  $112 \times 112$  deconvolutional layer, (i) the  $224 \times 224$  unpooling layer and (j) the last  $224 \times 224$  deconvolutional layer. The finer details of the object are revealed, as the features are forward-propagated through the layers in the deconvolution network. Note that noisy activations from background are suppressed through propagation while the activations closely related to the target classes are amplified. It shows that the learned filters in higher deconvolutional layers tend to capture class-specific shape information.

# Results



(a) Examples that our method produces better results than FCN [19].

# Results



# Multi-Scale Context-Aggregation by Dilated Convolutions

## MULTI-SCALE CONTEXT AGGREGATION BY DILATED CONVOLUTIONS

**Fisher Yu**

Princeton University

**Vladlen Koltun**

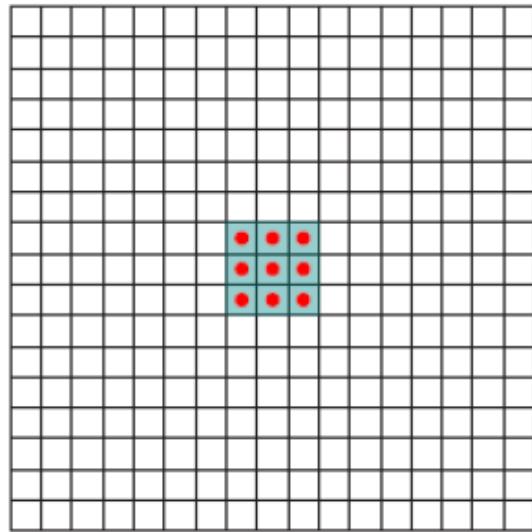
Intel Labs

[cs.CV] 30 Apr 2016

### ABSTRACT

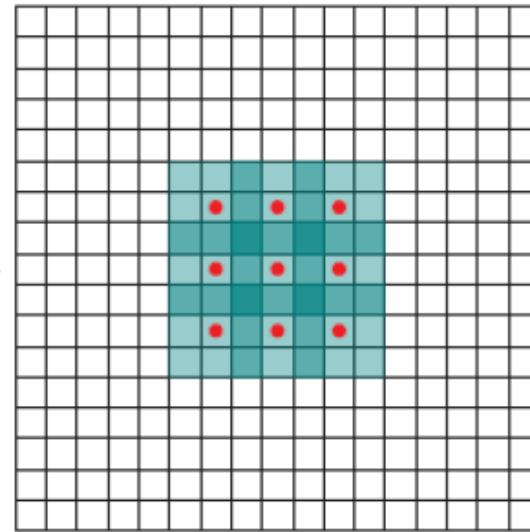
State-of-the-art models for semantic segmentation are based on adaptations of convolutional networks that had originally been designed for image classification. However, dense prediction problems such as semantic segmentation are structurally different from image classification. In this work, we develop a new convolutional network module that is specifically designed for dense prediction. The presented module uses dilated convolutions to systematically aggregate multi-scale contextual information without losing resolution. The architecture is based on the fact that dilated convolutions support exponential expansion of the receptive field without loss of resolution or coverage. We show that the presented context module increases the accuracy of state-of-the-art semantic segmentation systems. In addition, we examine the adaptation of image classification networks to dense prediction and show that simplifying the adapted network can increase accuracy.

# Dilated Convolution



3x3 Conv r=1

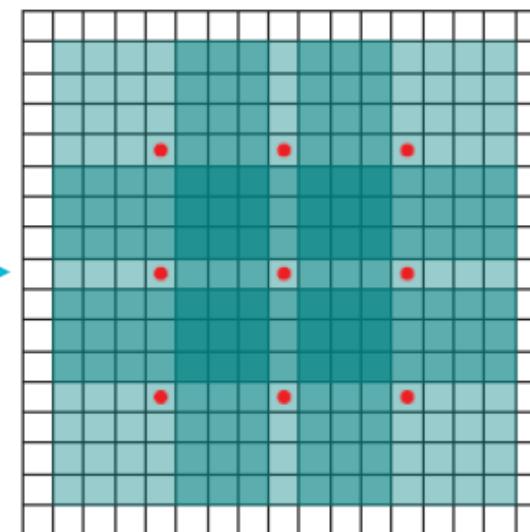
3x3 Range



3x3 Conv r=1

3x3 Conv r=2

7x7 Range



3x3 Conv r=1

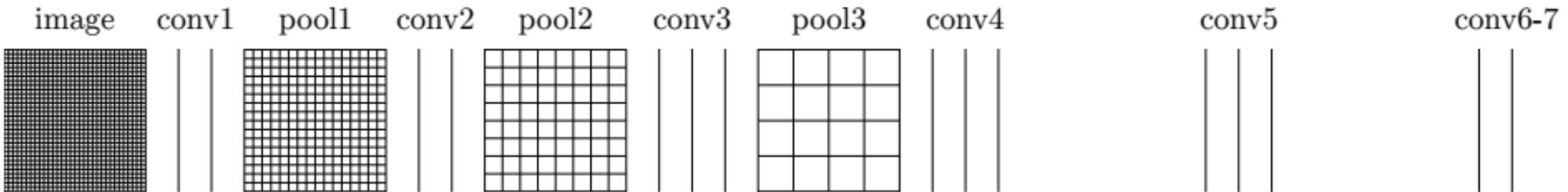
3x3 Conv r=2

3x3 Conv r=4

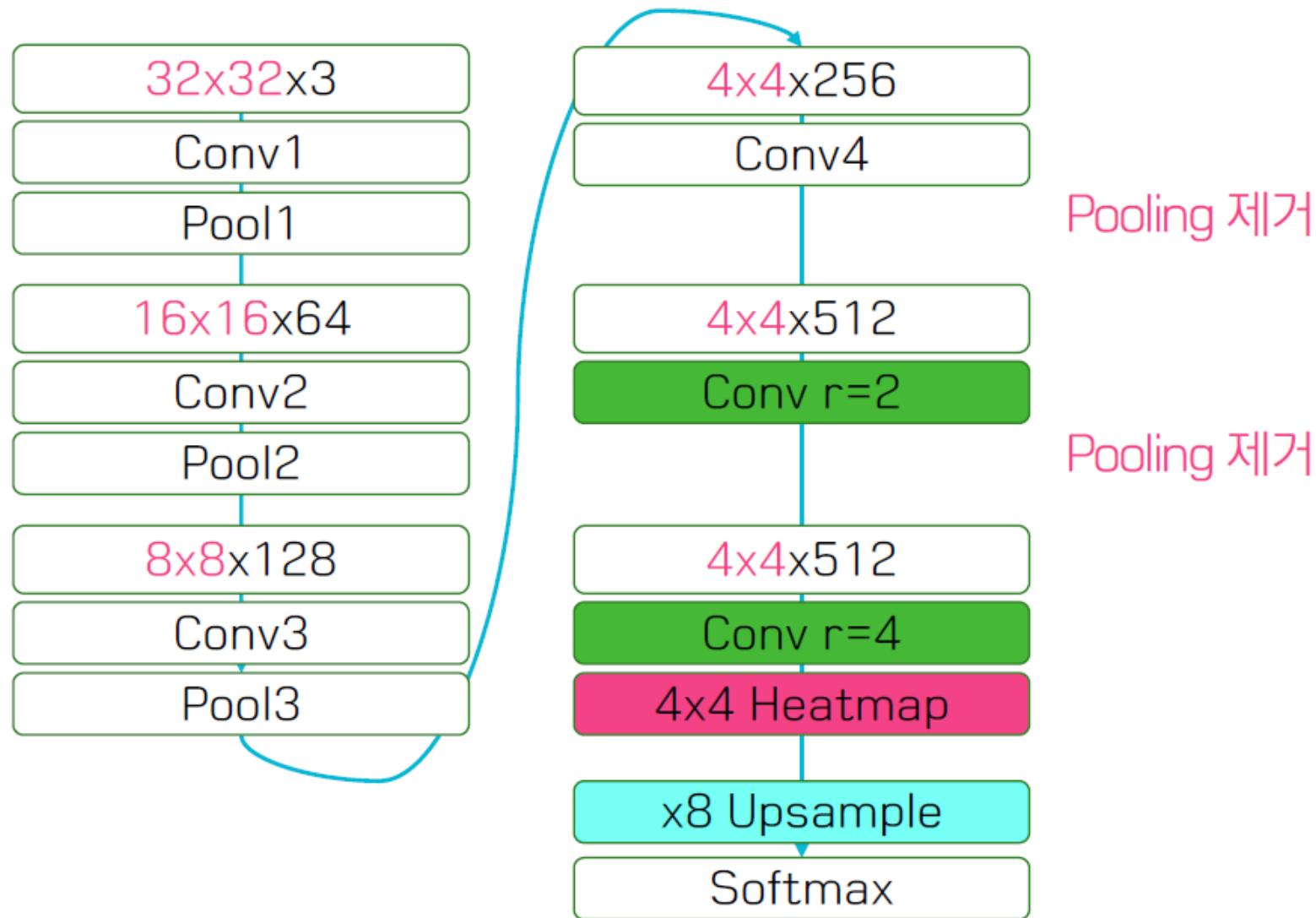
15x15 Range

# Front-End Module

- FCN에서 pool4, pool5 제거
  - conv5 → 2-dilated convolution
  - conv6 → 4-dilated convolution



# Front-End Module

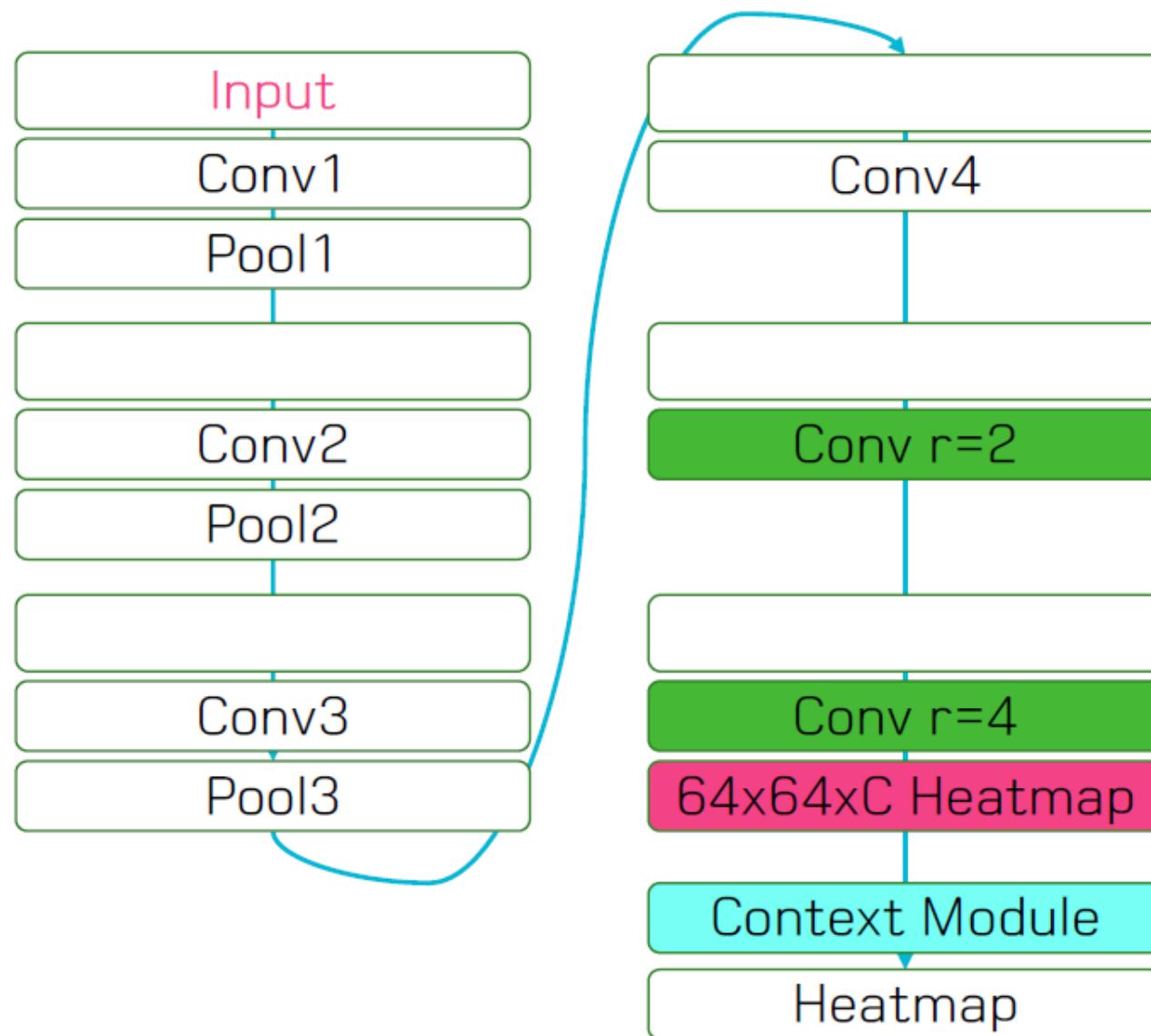


# Context Network

Layer	1	2	3	4	5	6	7	8
Convolution	$3 \times 3$	$3 \times 3$	$3 \times 3$	$3 \times 3$	$3 \times 3$	$3 \times 3$	$3 \times 3$	$1 \times 1$
Dilation	1	1	2	4	8	16	1	1
Truncation	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No
Receptive field	$3 \times 3$	$5 \times 5$	$9 \times 9$	$17 \times 17$	$33 \times 33$	$65 \times 65$	$67 \times 67$	$67 \times 67$
Output channels								
Basic	$C$	$C$	$C$	$C$	$C$	$C$	$C$	$C$
Large	$2C$	$2C$	$4C$	$8C$	$16C$	$32C$	$32C$	$C$

Table 1: Context network architecture. The network processes  $C$  feature maps by aggregating contextual information at progressively increasing scales without losing resolution.

# Front-End Module + Context Module



Size에 맞게 Input Padding  
C = 21 (Class 개수)

# U-Net: Convolutional Networks for Biomedical Image Segmentation

## U-Net: Convolutional Networks for Biomedical Image Segmentation

Olaf Ronneberger, Philipp Fischer, and Thomas Brox

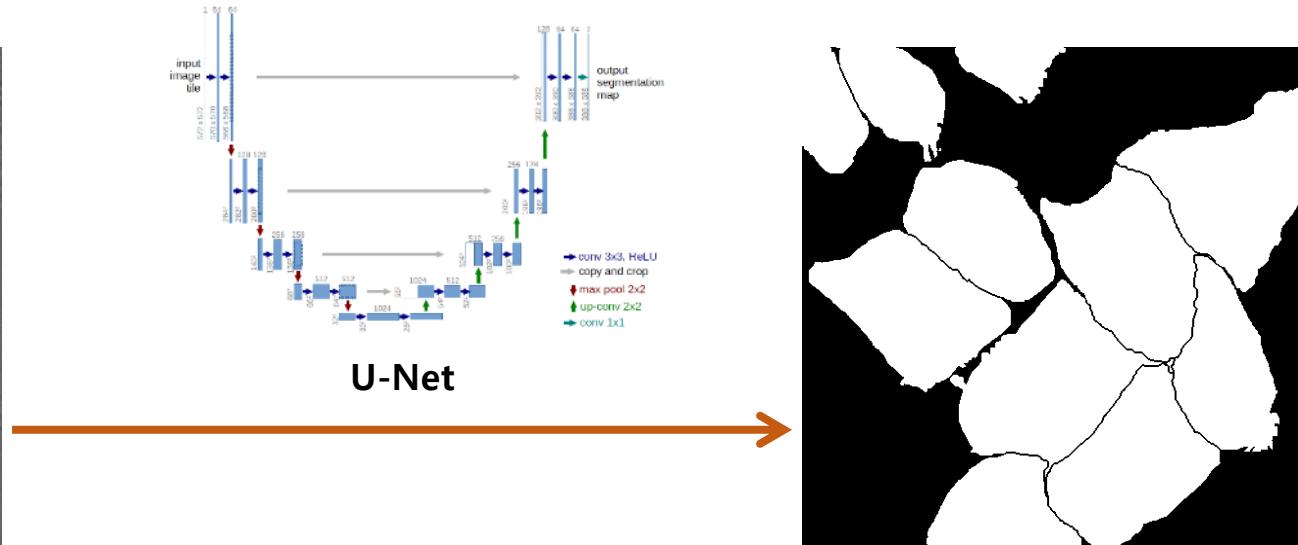
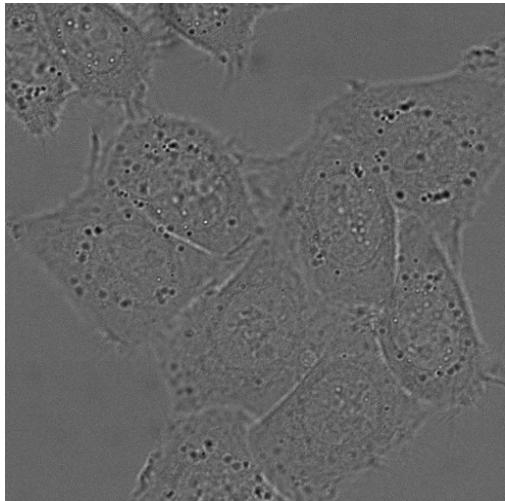
Computer Science Department and BIOSS Centre for Biological Signalling Studies,  
University of Freiburg, Germany

[ronneber@informatik.uni-freiburg.de](mailto:ronneber@informatik.uni-freiburg.de),

WWW home page: <http://lmb.informatik.uni-freiburg.de/>

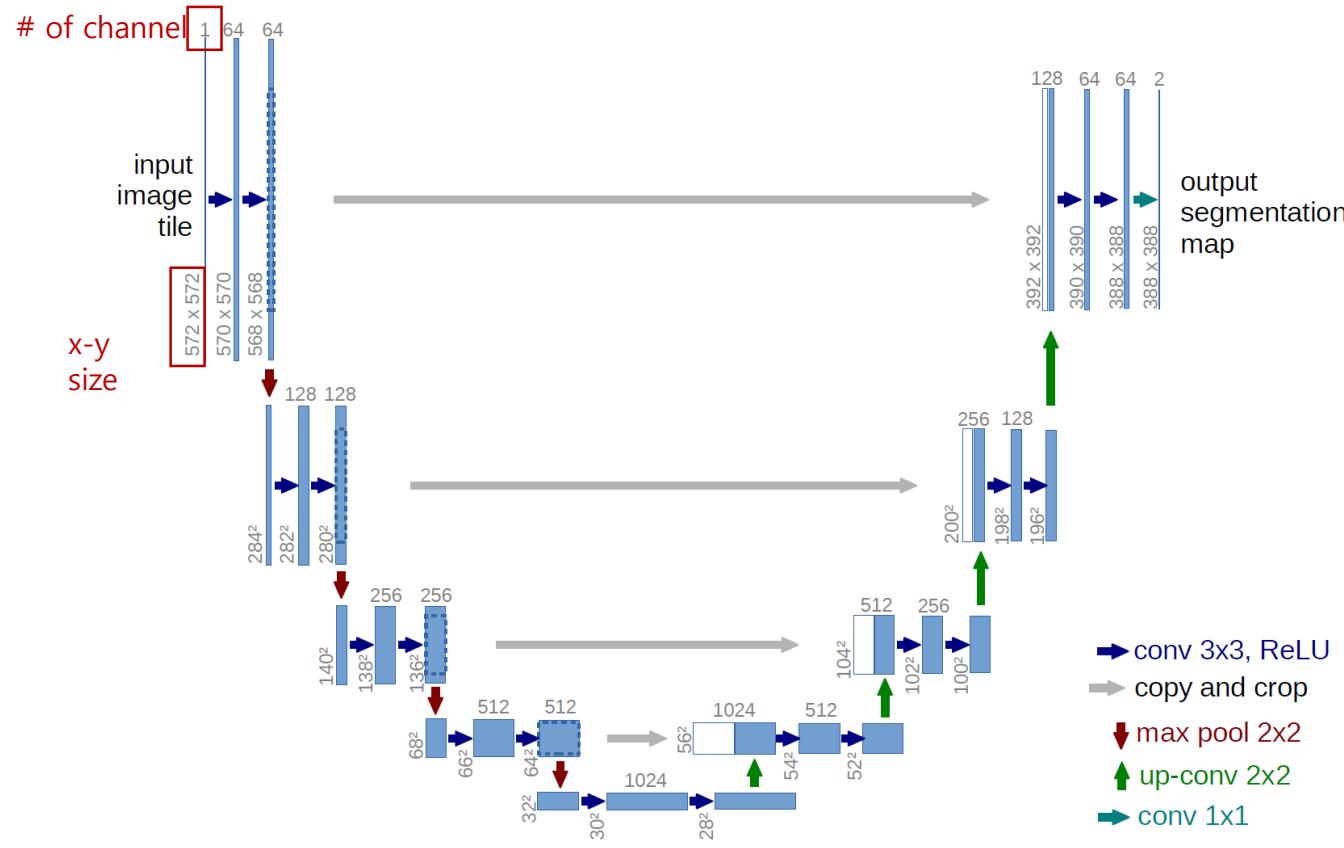
**Abstract.** There is large consent that successful training of deep networks requires many thousand annotated training samples. In this paper, we present a network and training strategy that relies on the strong use of data augmentation to use the available annotated samples more efficiently. The architecture consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. We show that such a network can be trained end-to-end from very few images and outperforms the prior best method (a sliding-window

# Biomedical Image Segmentation with U-net



- U-Net learns segmentation in an end-to-end setting
- Very few annotated images (approx. 30 per application)

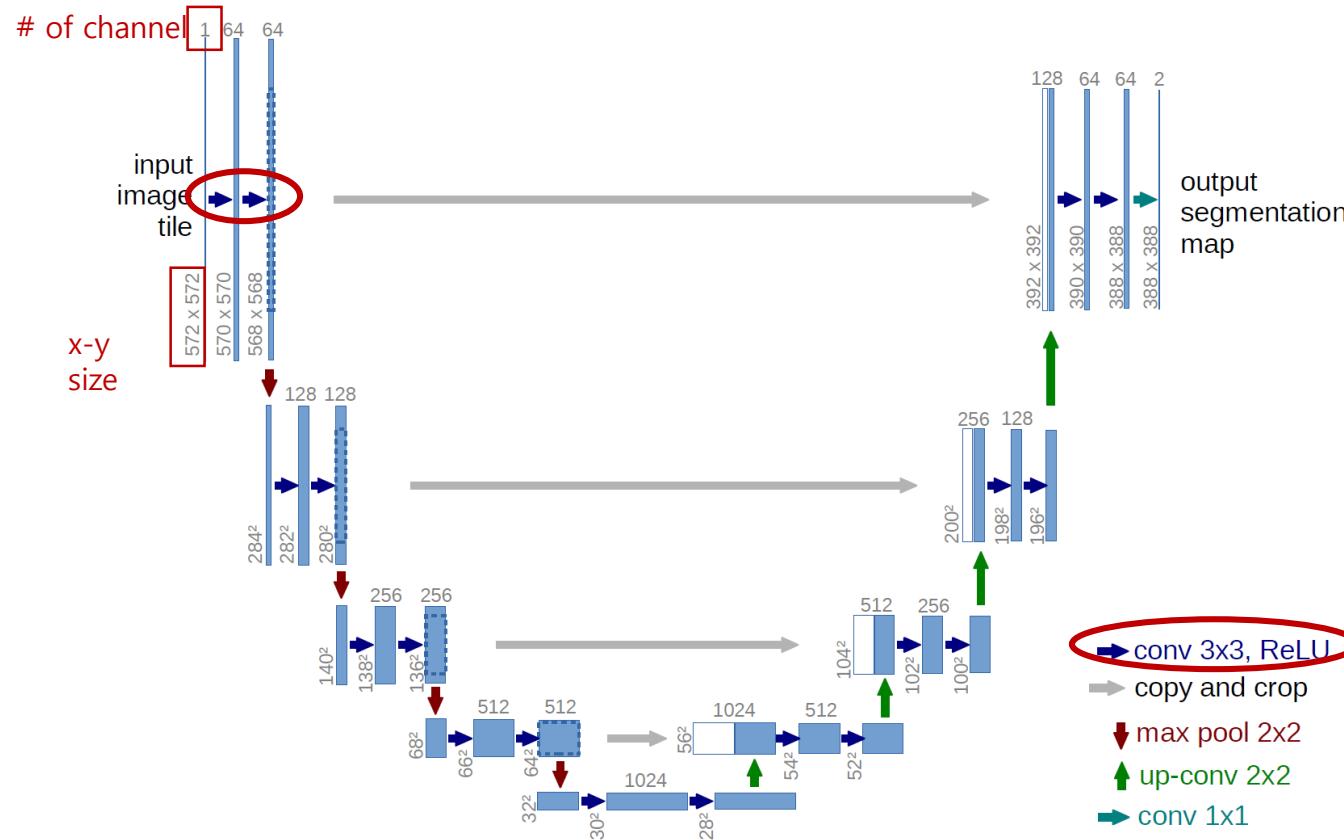
# U-Net Architecture



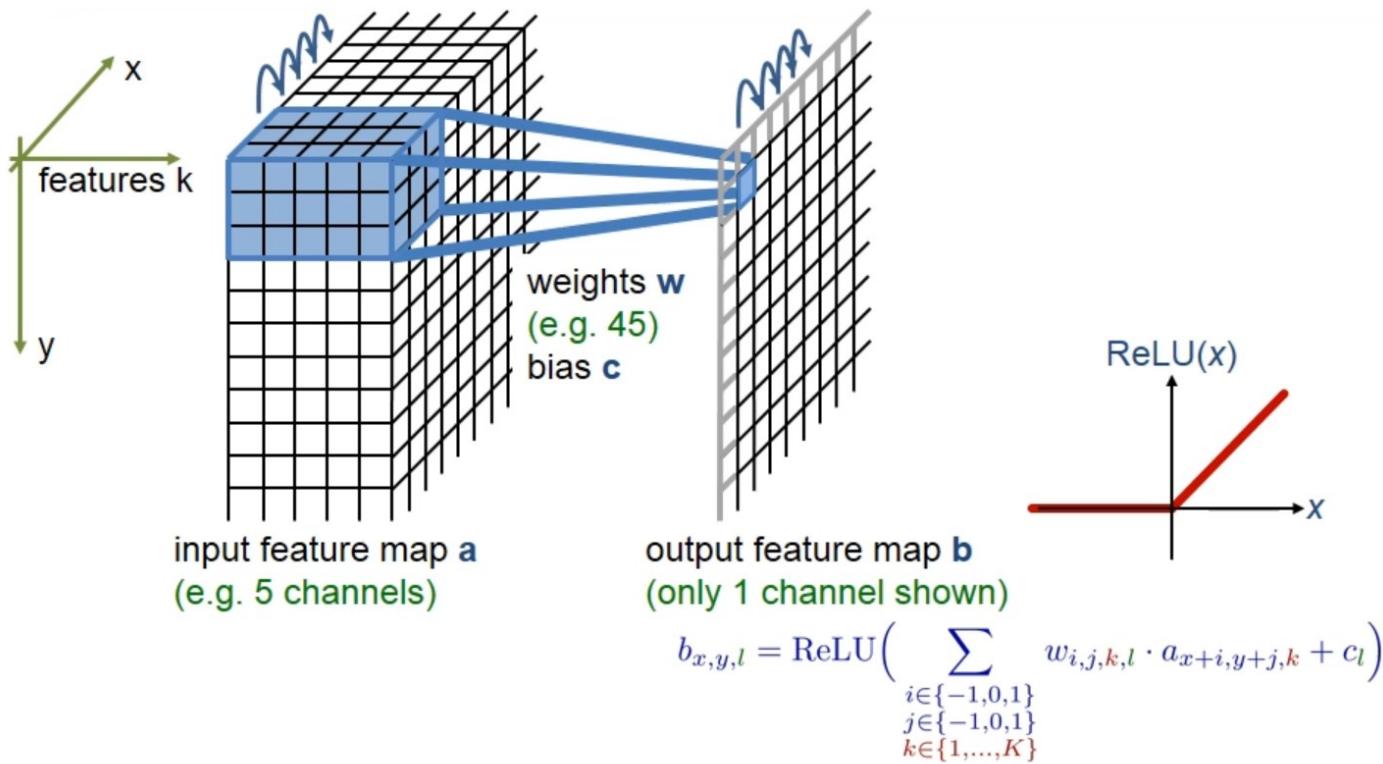
# U-Net Architecture



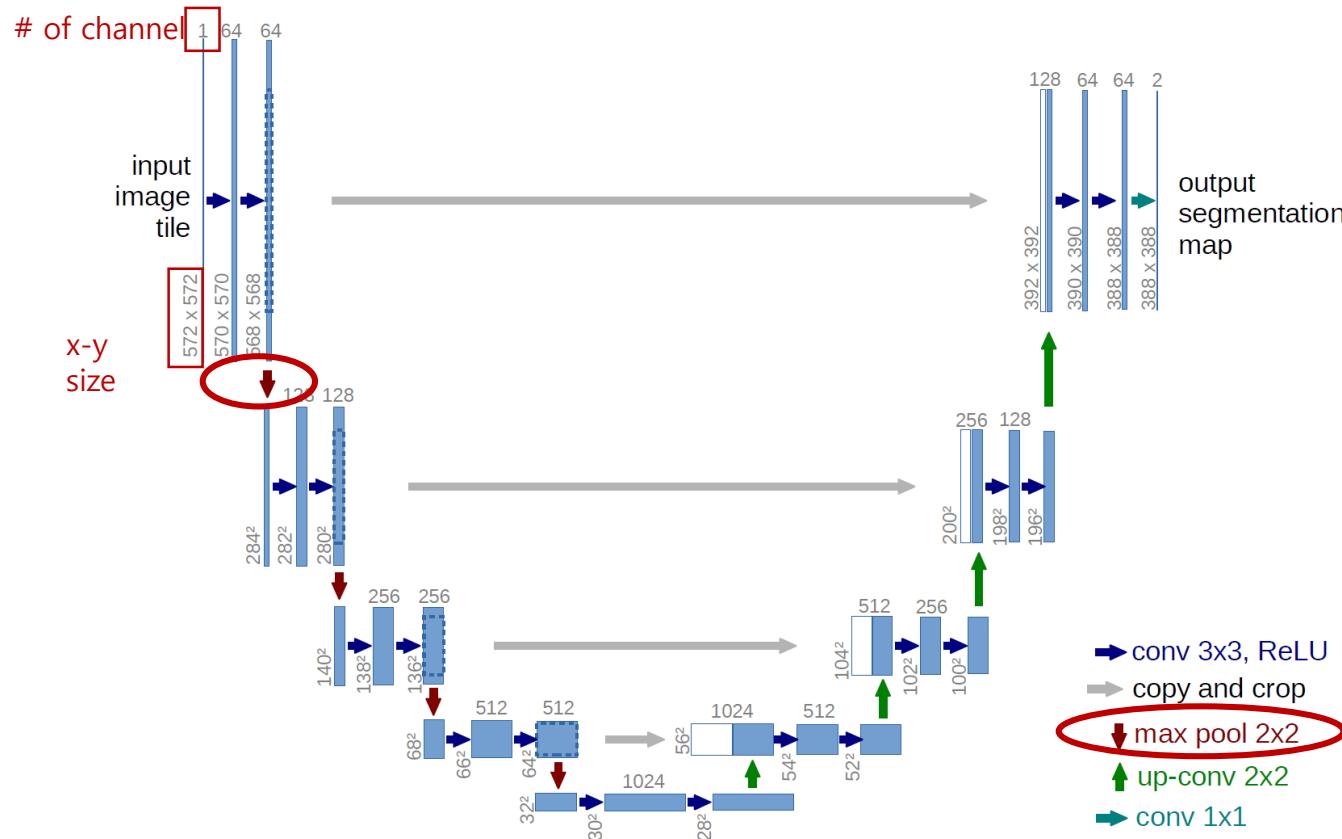
# U-Net Architecture



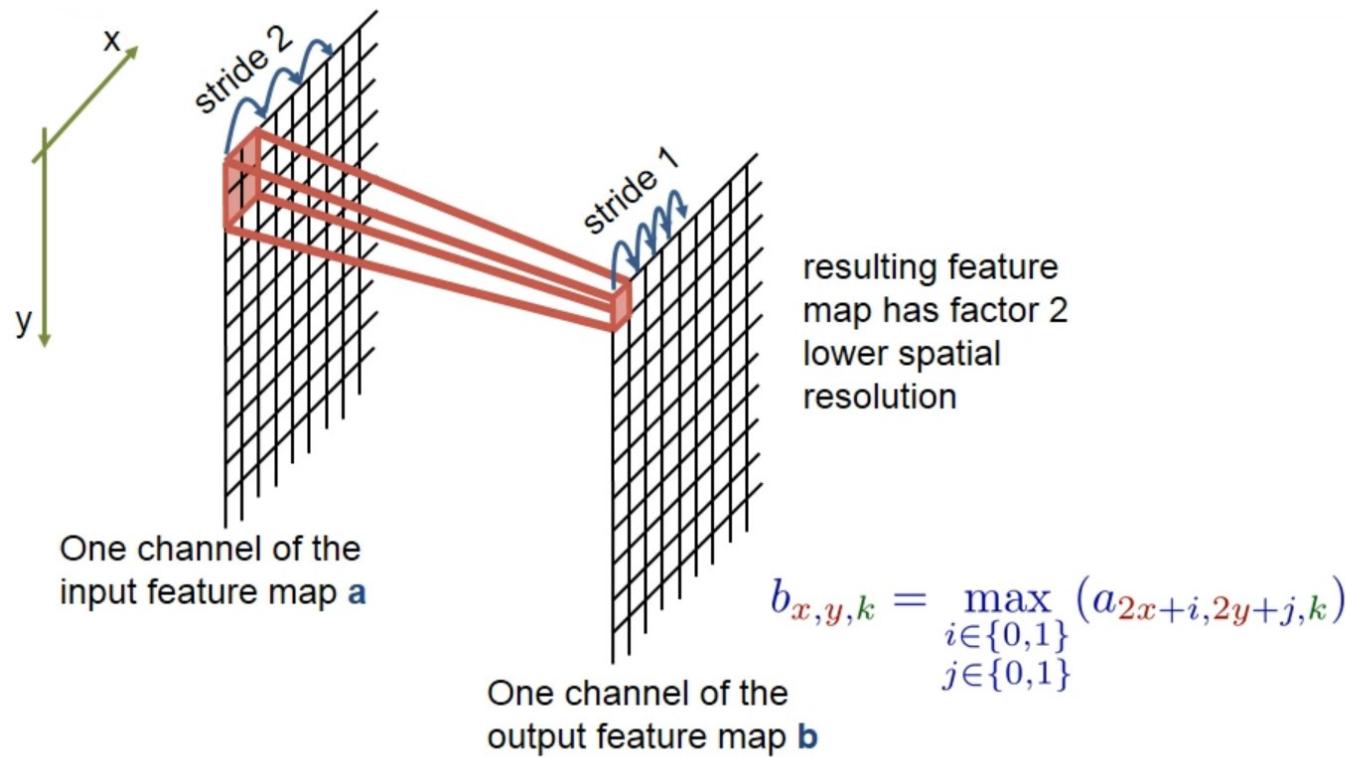
# 3x3 Convolution + ReLU



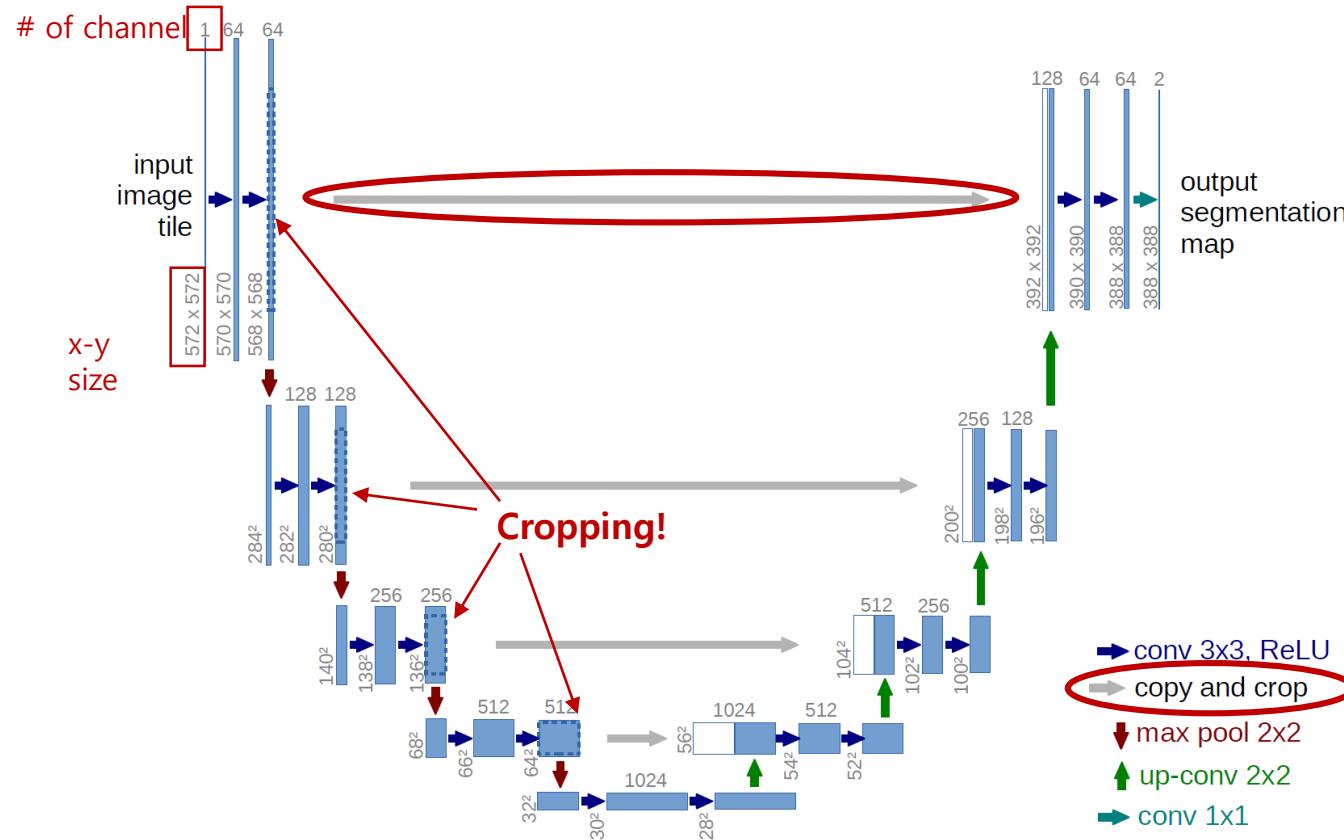
# U-Net Architecture



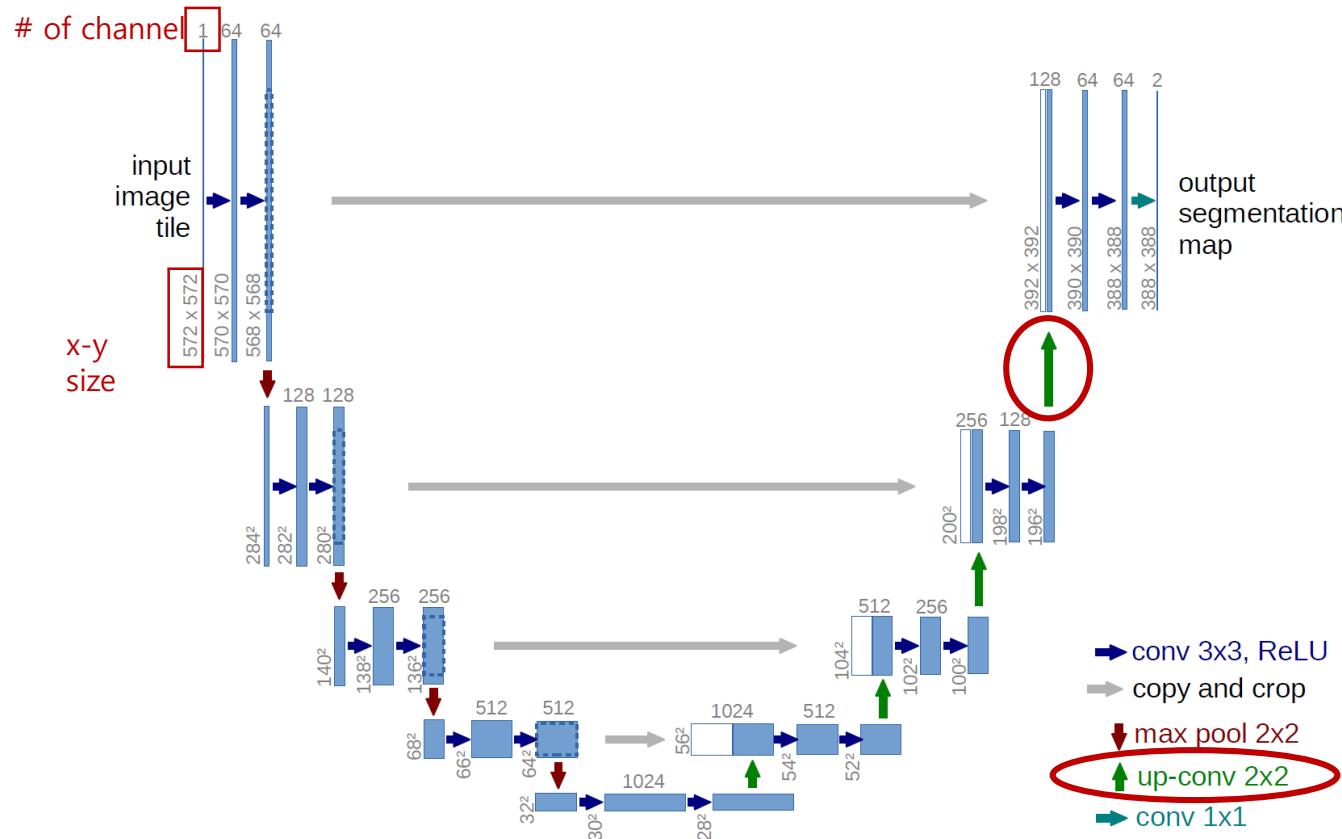
# 2x2 Max-pooling



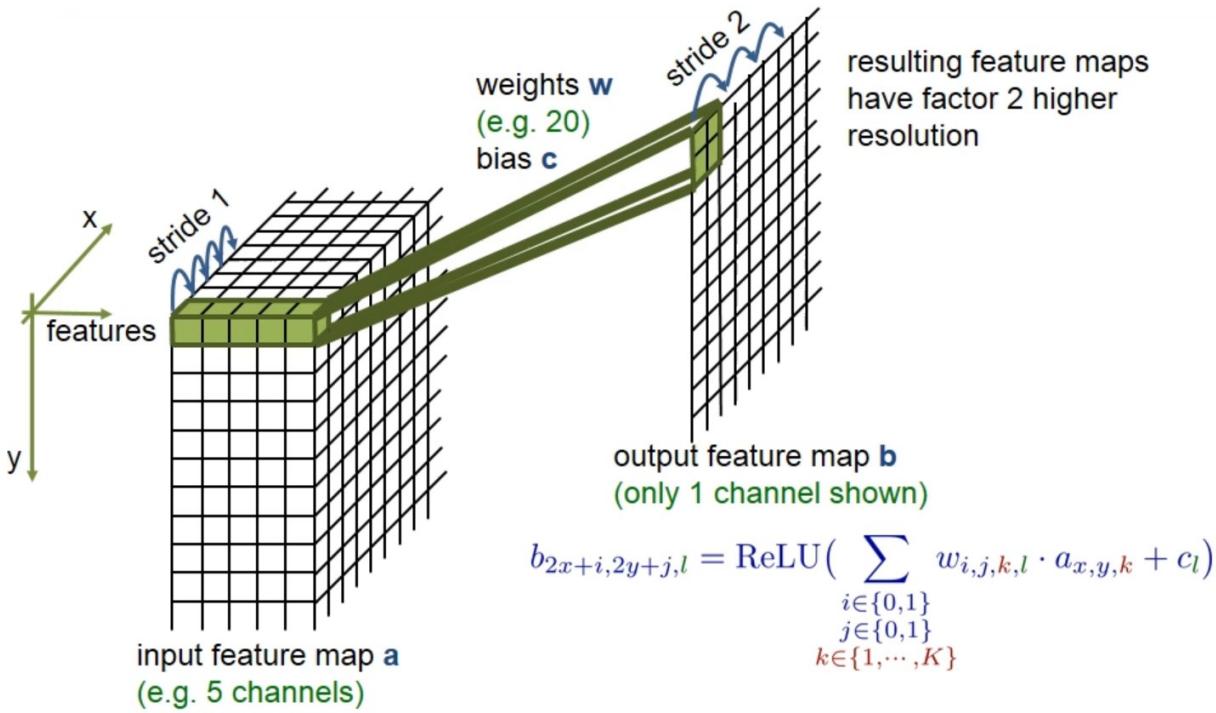
# U-Net Architecture



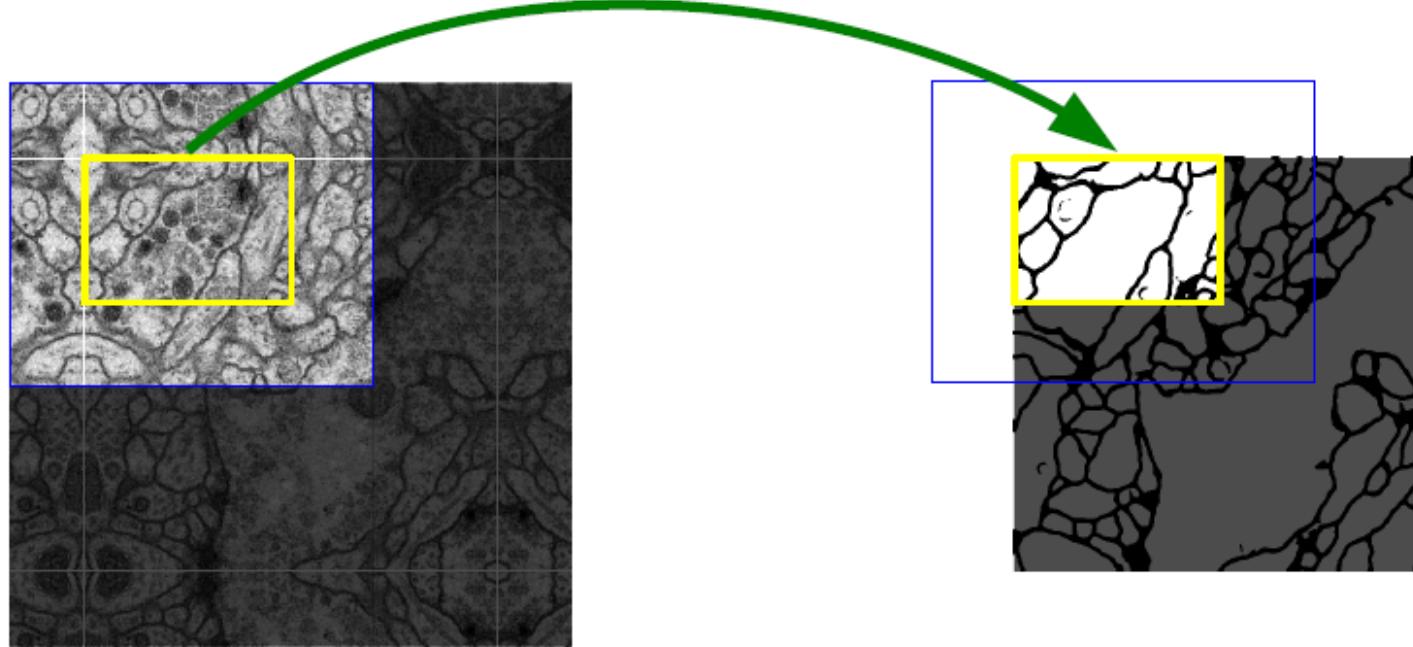
# U-Net Architecture



# 2x2 Up-convolution

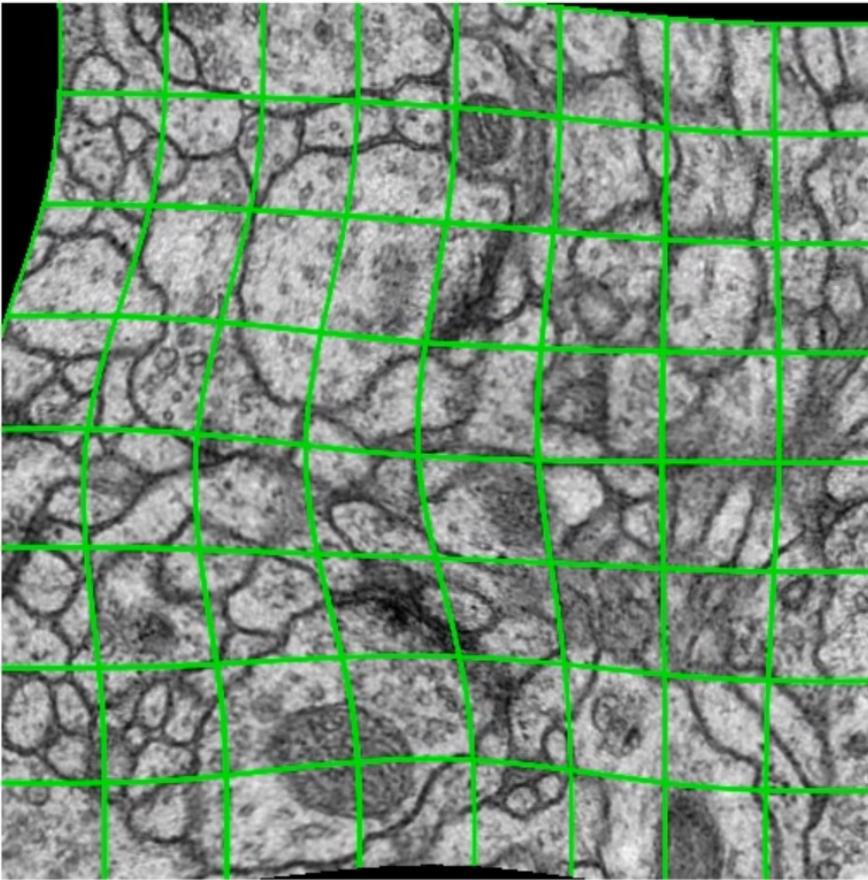


# Overlap-tile strategy for arbitrary large images

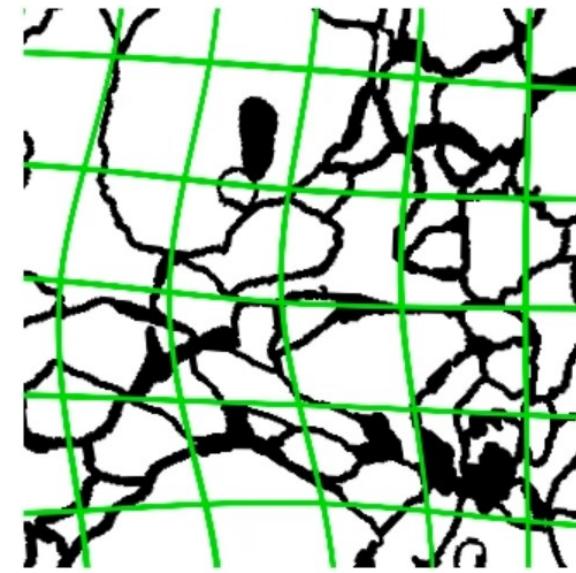


- Segmentation of the yellow area uses input data of the blue area
- Raw data extrapolation by mirroring

# Augment Training Data using Deformations

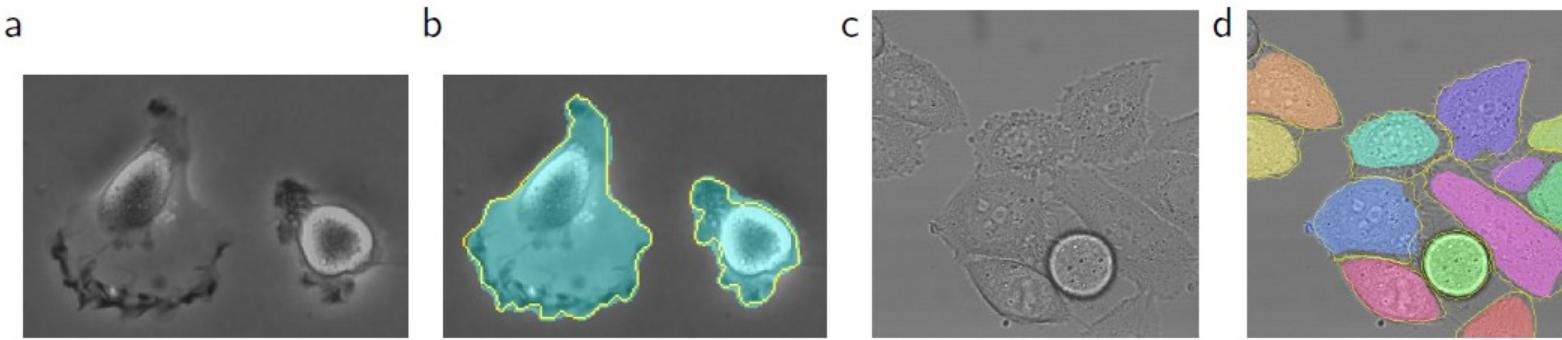


resulting deformed image  
(for visualization: no rotation, no shift, no extrapolation)



correspondingly deformed  
manual labels

# ISBI 2015 Result



Name	PhC-U373	DIC-HeLa
IMCB-SG (2014)	0.2669	0.2935
KTH-SE (2014)	0.7953	0.4607
HOUS-US (2014)	0.5323	-
second-best 2015	0.83	0.46
u-net (2015)	<b>0.9203</b>	<b>0.7756</b>

# Pyramid Scene Parsing Network(PSPNet)

## Pyramid Scene Parsing Network

Hengshuang Zhao<sup>1</sup> Jianping Shi<sup>2</sup> Xiaojuan Qi<sup>1</sup> Xiaogang Wang<sup>1</sup> Jiaya Jia<sup>1</sup>

<sup>1</sup>The Chinese University of Hong Kong <sup>2</sup>SenseTime Group Limited

{hszhao, xjqi, leojia}@cse.cuhk.edu.hk, xgwang@ee.cuhk.edu.hk, shijianping@sensetime.com

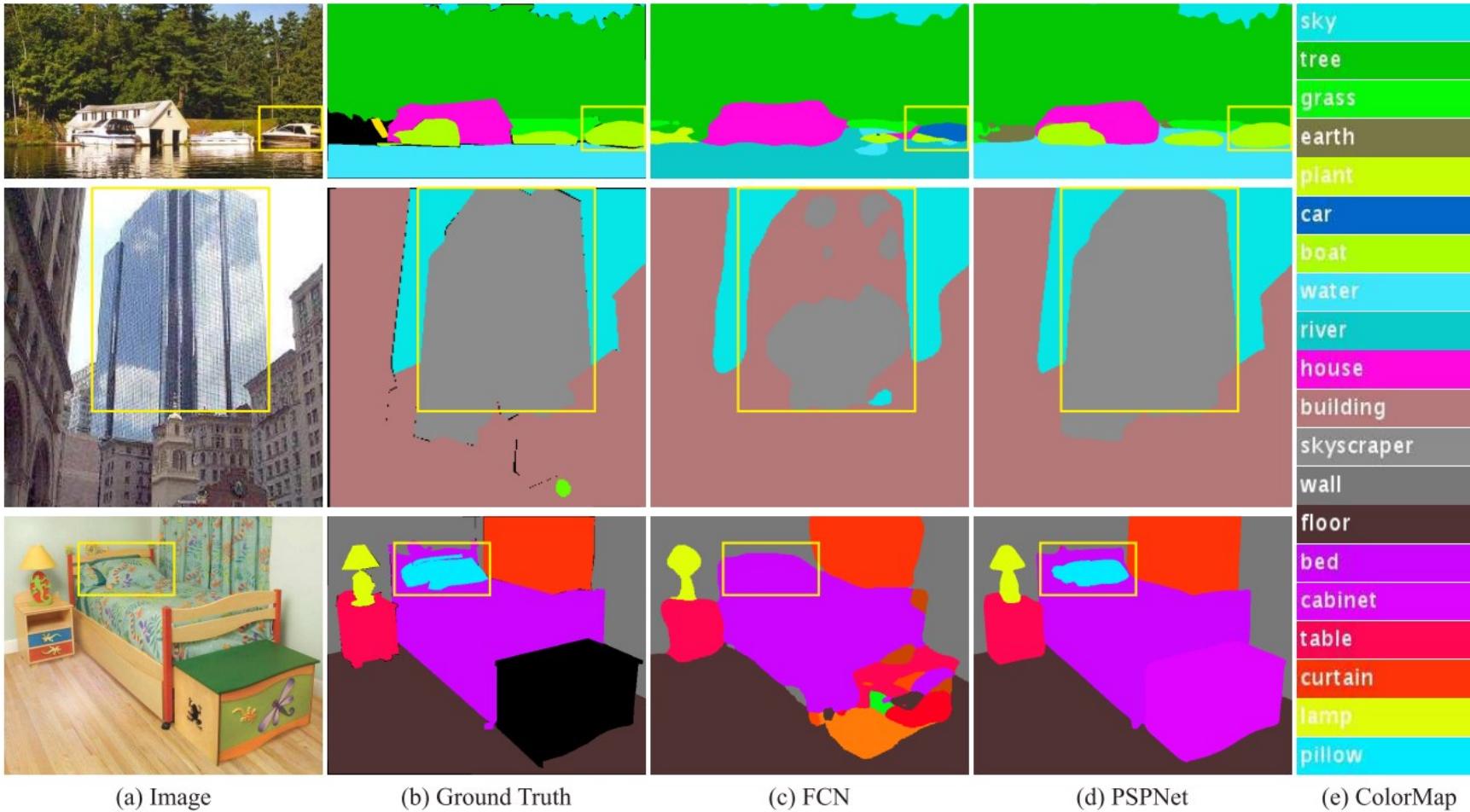
### Abstract

*Scene parsing is challenging for unrestricted open vocabulary and diverse scenes. In this paper, we exploit the capability of global context information by different-region-based context aggregation through our pyramid pooling module together with the proposed pyramid scene parsing network (PSPNet). Our global prior representation is effective to produce good quality results on the scene parsing task, while PSPNet provides a superior framework for pixel-level prediction. The proposed approach achieves state-of-the-art performance on various datasets. It came first in ImageNet scene parsing challenge 2016, PASCAL VOC 2012 benchmark and Cityscapes benchmark. A single PSPNet yields the new record of mIoU accuracy 85.4% on PASCAL VOC 2012, and 88.2% on Cityscapes.*



# Pyramid Scene Parsing Network(PSPNet)

- Deep Network with a suitable global-scene-level prior can much improve the performance of scene parsing



# PSPNet Architecture

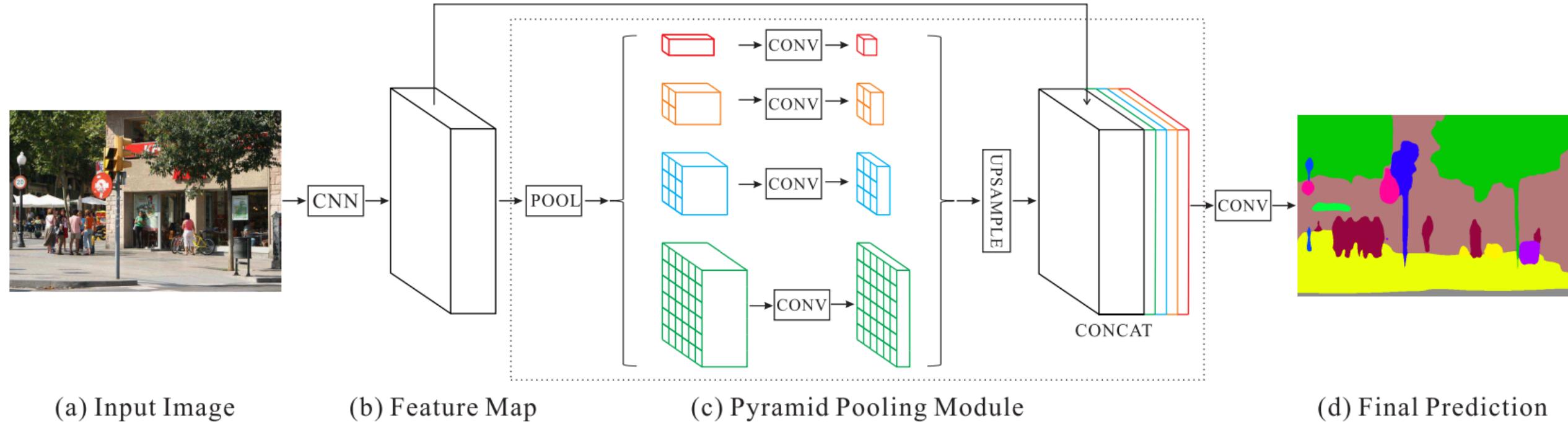
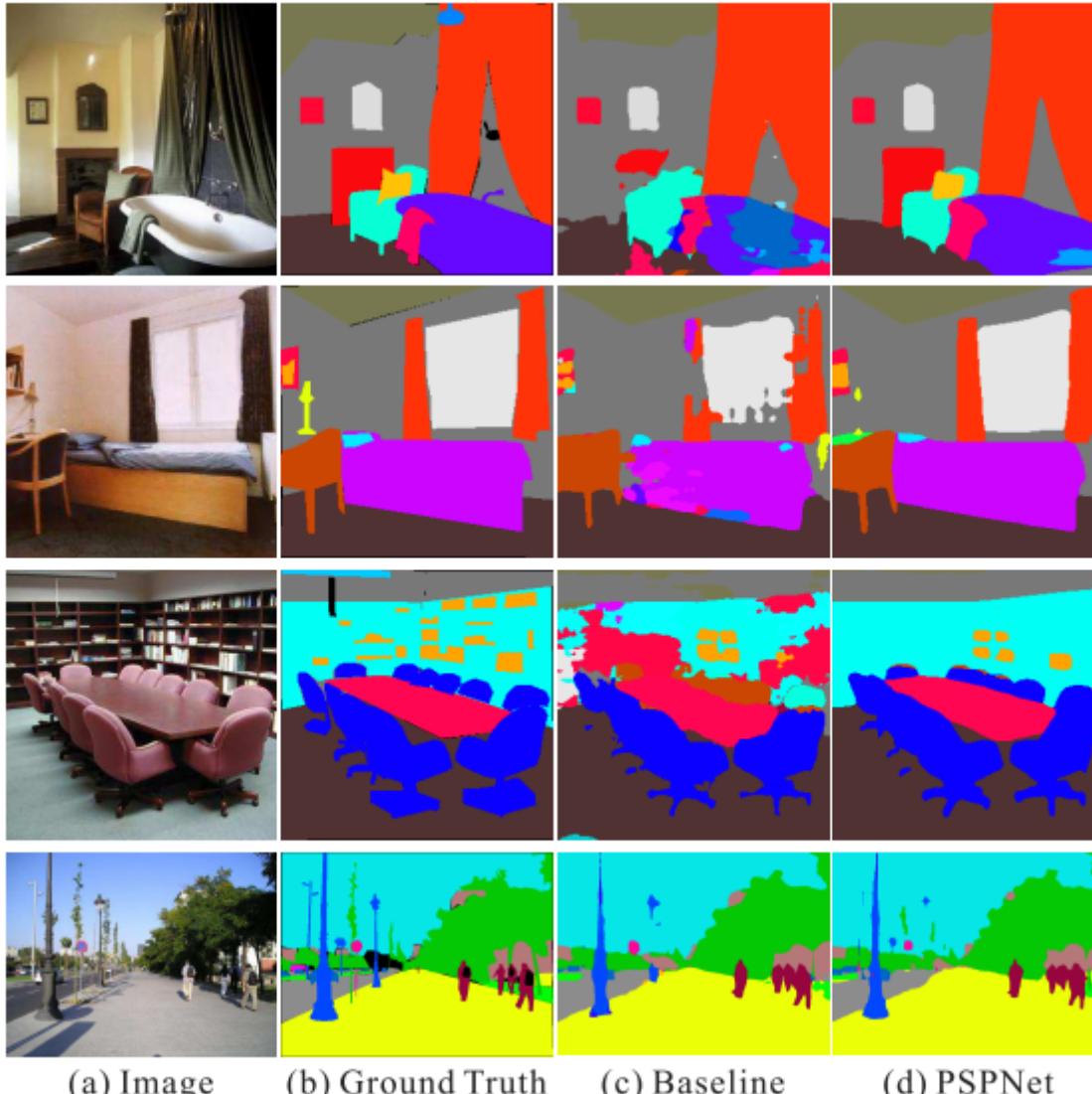


Figure 3. Overview of our proposed PSPNet. Given an input image (a), we first use CNN to get the feature map of the last convolutional layer (b), then a pyramid parsing module is applied to harvest different sub-region representations, followed by upsampling and concatenation layers to form the final feature representation, which carries both local and global context information in (c). Finally, the representation is fed into a convolution layer to get the final per-pixel prediction (d).

# Results



# Results

