# Semantic Segmentation of Images

## Using Deep Convolutional Neural Networks

**Fahimeh Fooladgar**

Image Processing Lab (IPL)
Department of Computer Engineering
Sharif University of Technology

July 2019

Amirkabir Artificial Intelligence Summer Summit

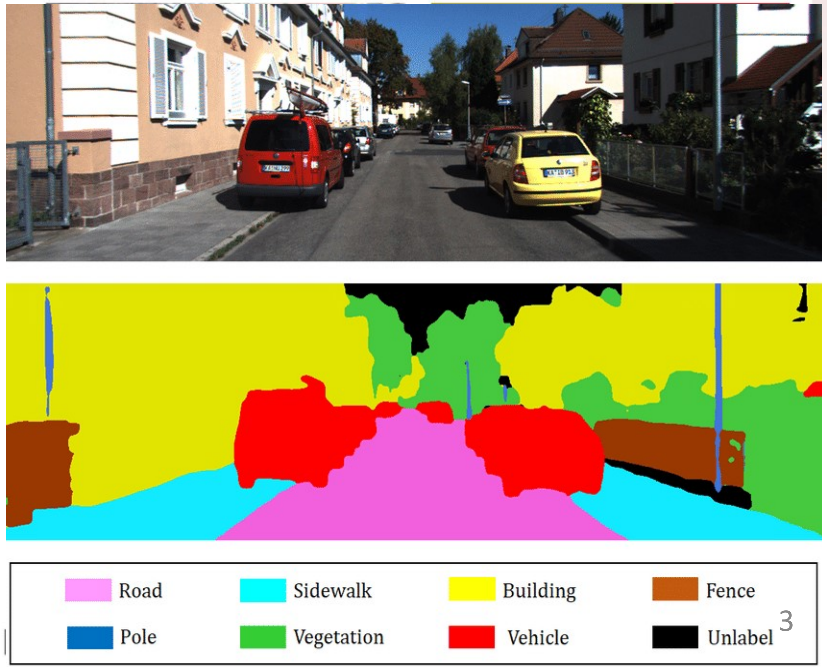
# Outline

- Introduction
  - Definition & Applications
- Approaches
  - Traditional Methods
  - Deep Learning Methods
- Available datasets and other directions

# Introduction

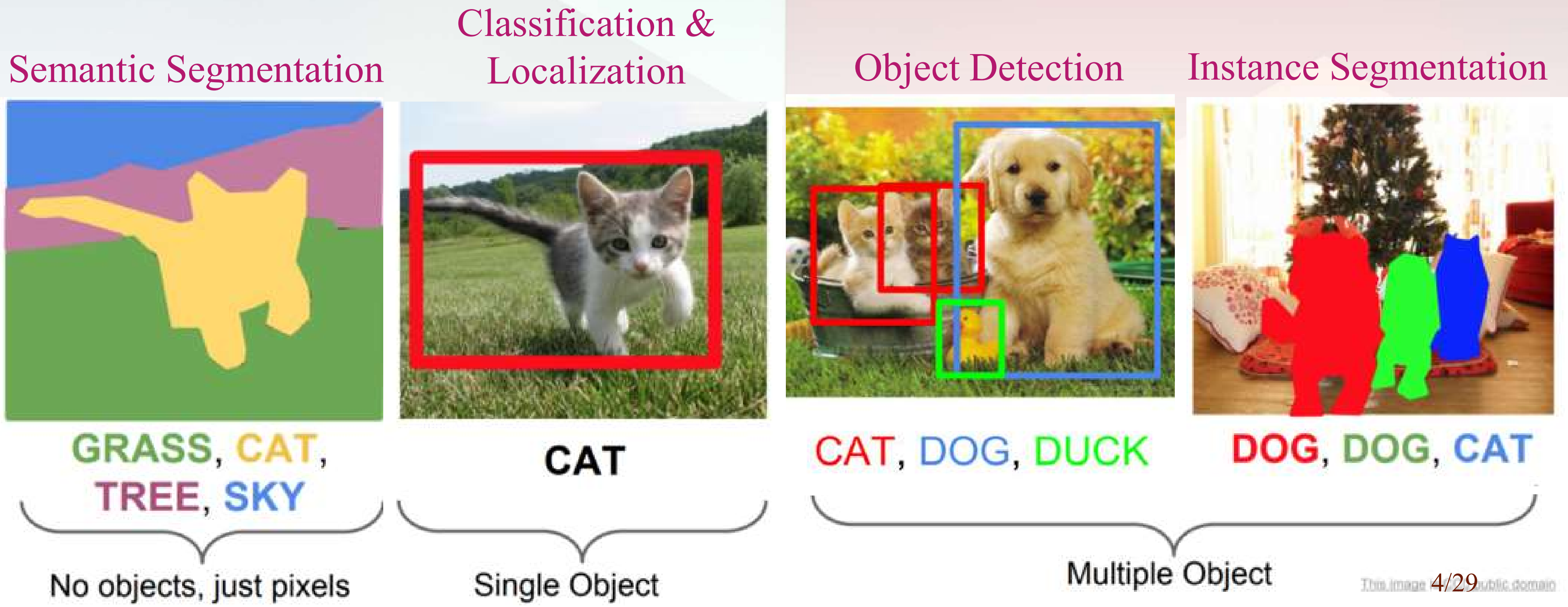
## Image Segmentation

## Semantic Segmentation

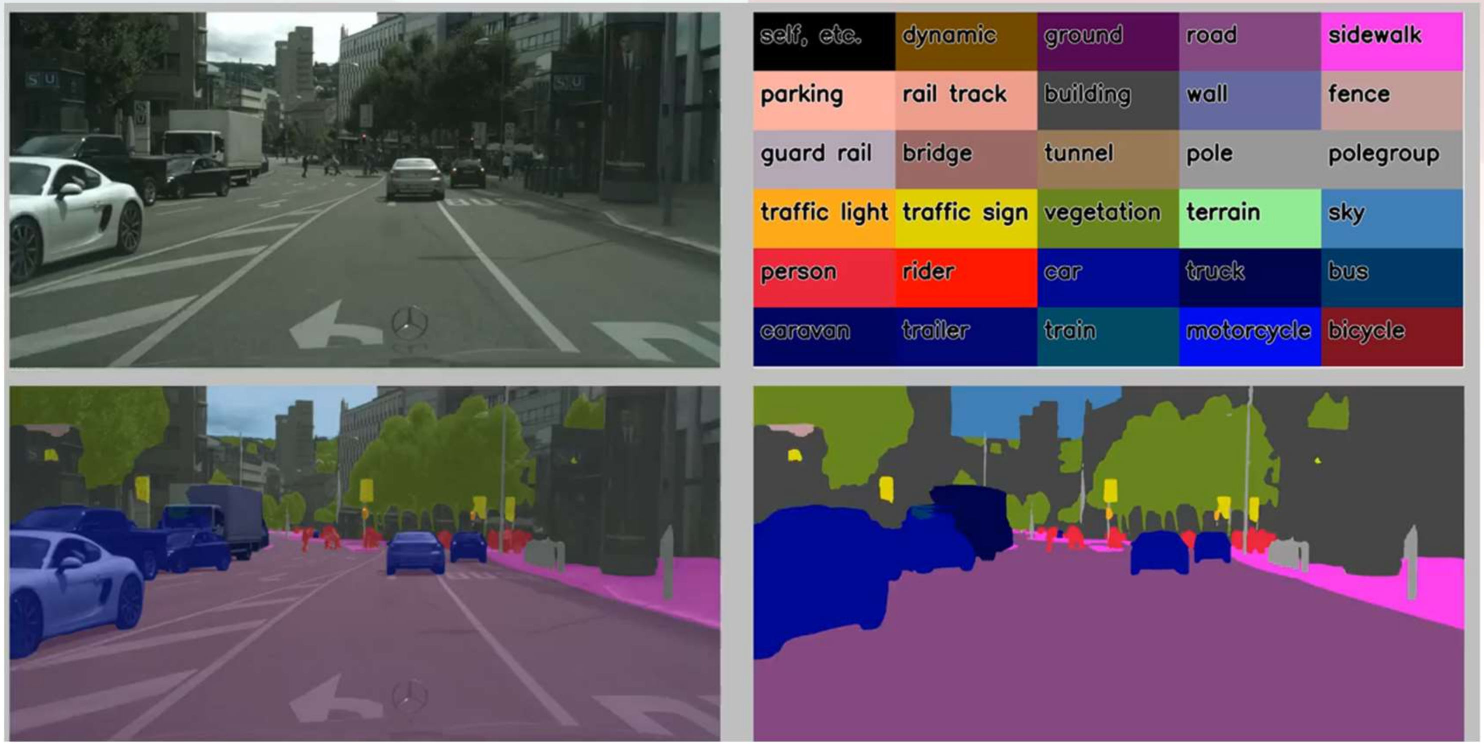


# Introduction (cont.)

- Semantic segmentation classifying each and very pixel

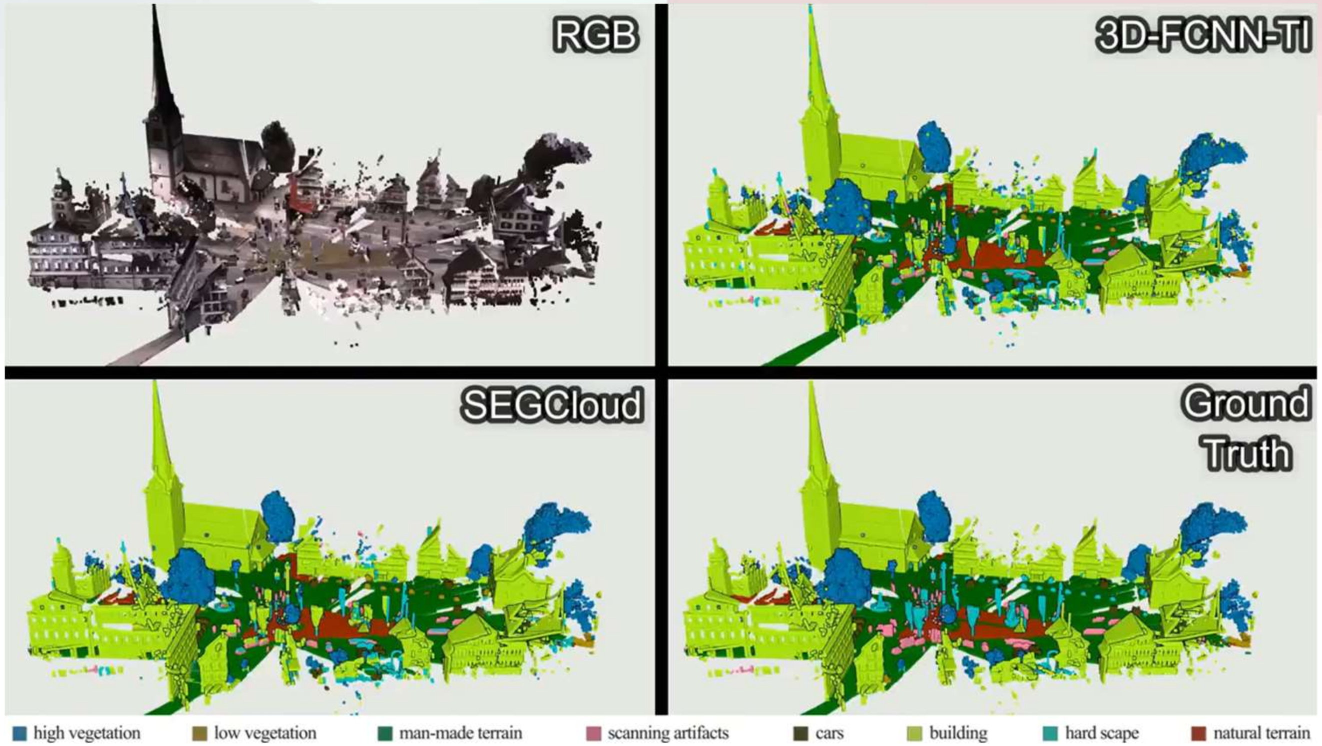


# Applications: Self-Driving Car

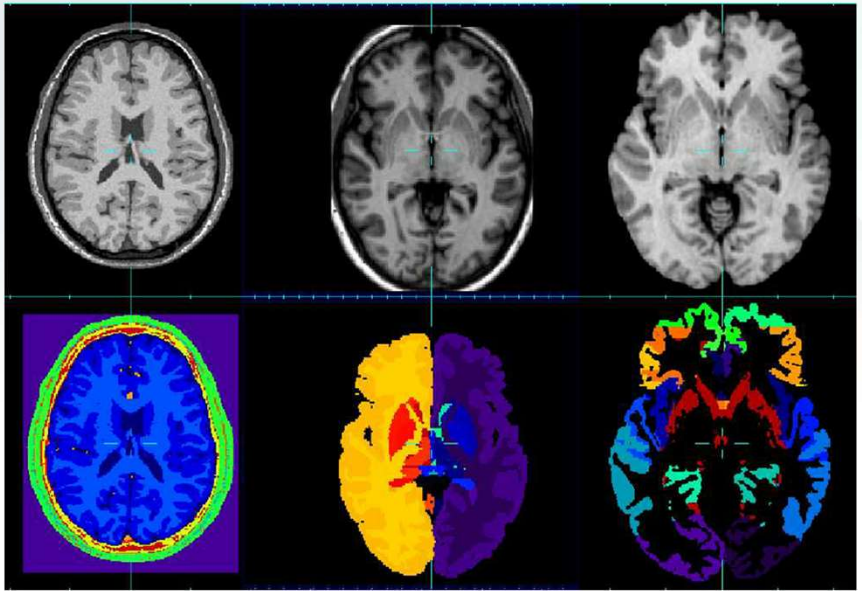


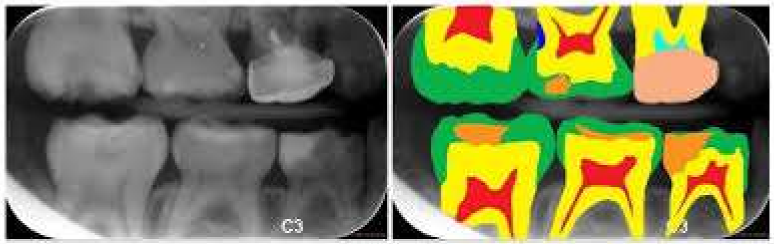
# Applications: Robotic

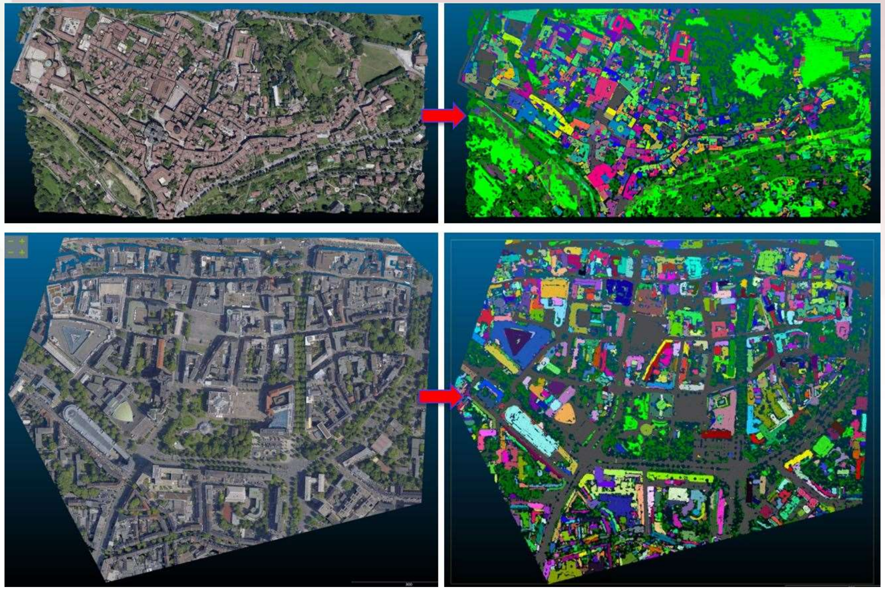
# Applications: 3D Scene Understanding

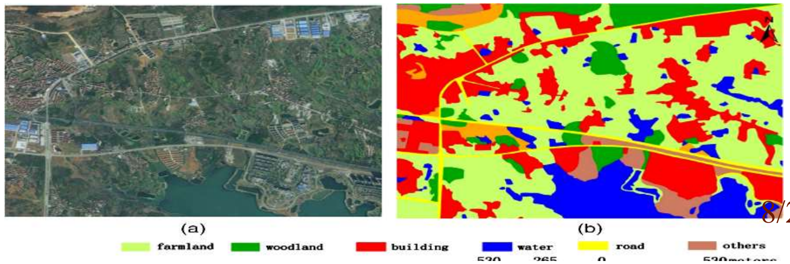


# Others …

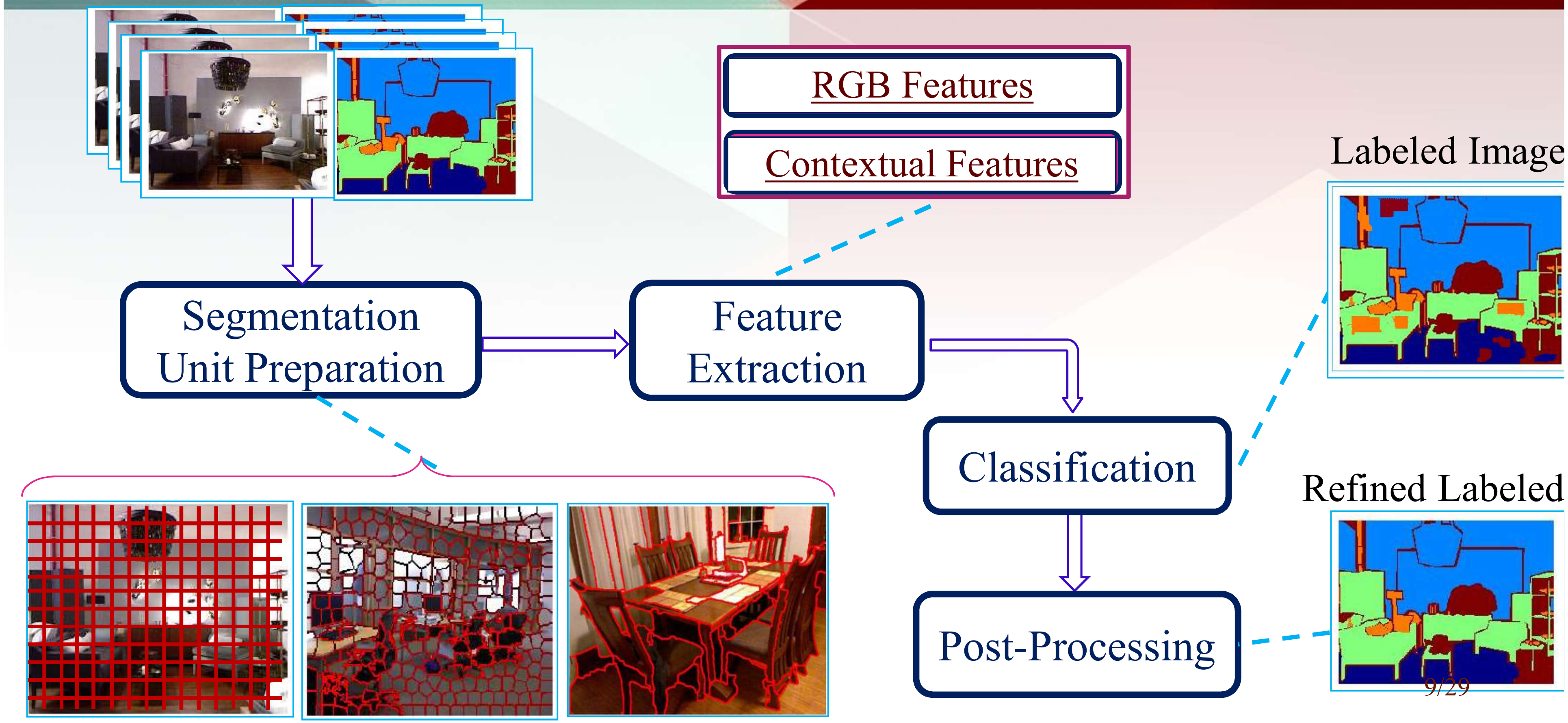


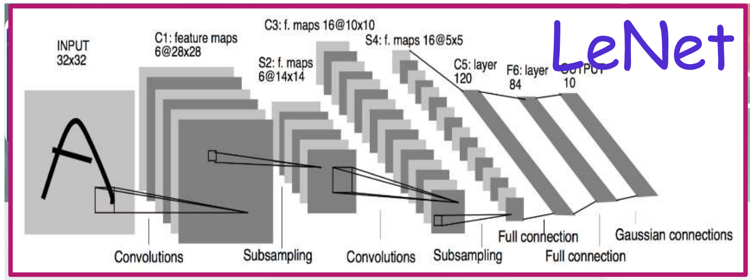


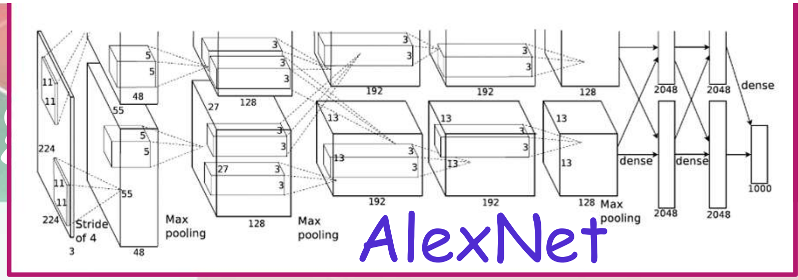


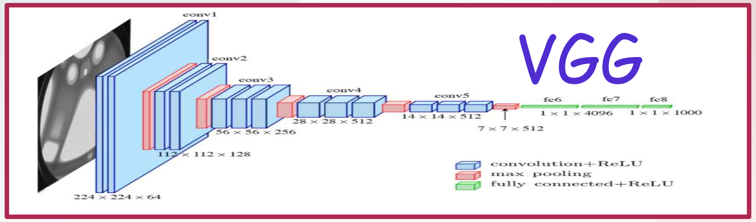


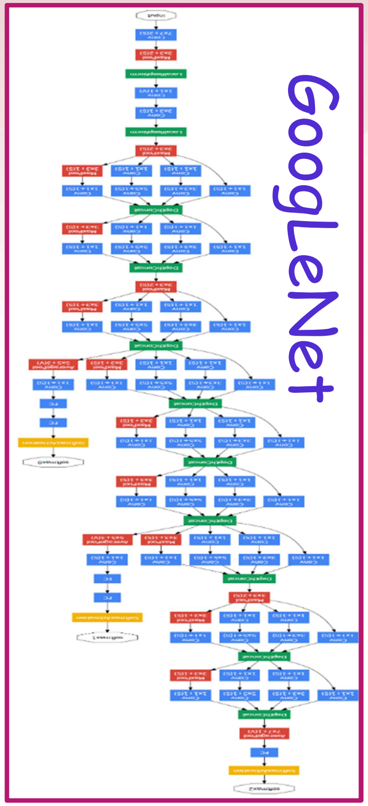
# Traditional Methods

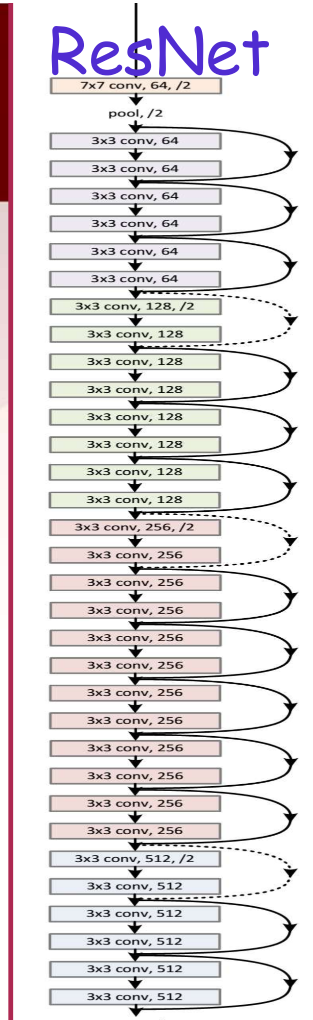


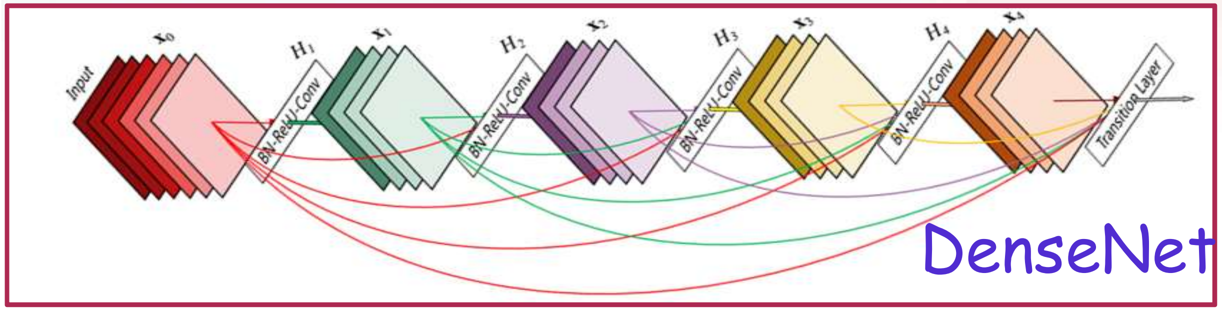

LeNet
INPUT 32x32
C1: feature maps 6@28x28
S2: f. maps 6@14x14
C3: f. maps 16@10x10
S4: f. maps 16@5x5
C5: layer 120
F6: layer 84
OUTPUT 10
Convolutions
Subsampling
Convolutions
Subsampling
Full connection
Full connection
Gaussian connections

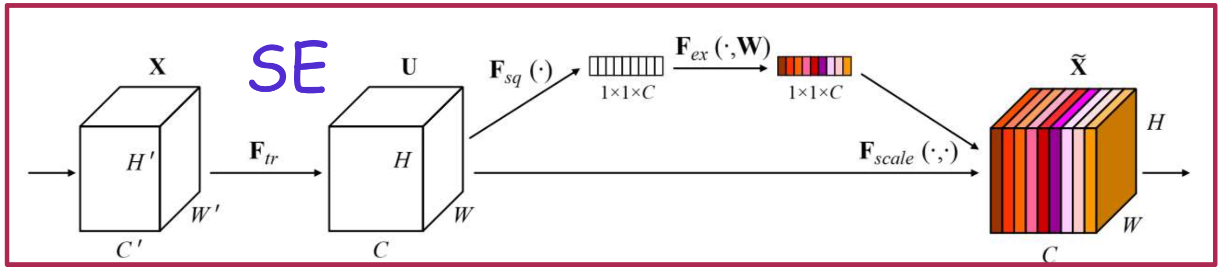

AlexNet
224
224
Stride of 4
Max pooling
Max pooling
Max pooling
dense
dense
dense
2048
2048
1000


VGG
conv1
conv2
conv3
conv4
conv5
fc6
fc7
fc8
224 × 224 × 64
112 × 112 × 128
56 × 56 × 256
28 × 28 × 512
14 × 14 × 512
7 × 7 × 512
1 × 1 × 4096
1 × 1 × 1000
convolution+ReLU
max pooling
fully connected+ReLU

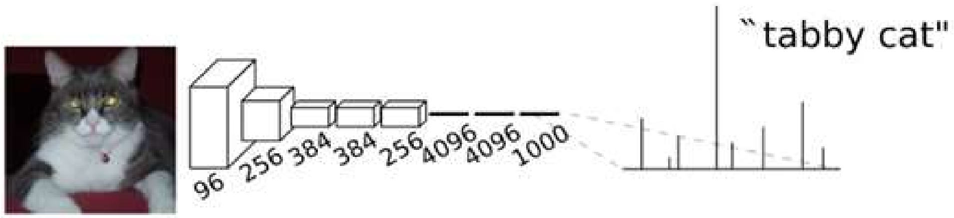

GoogLeNet

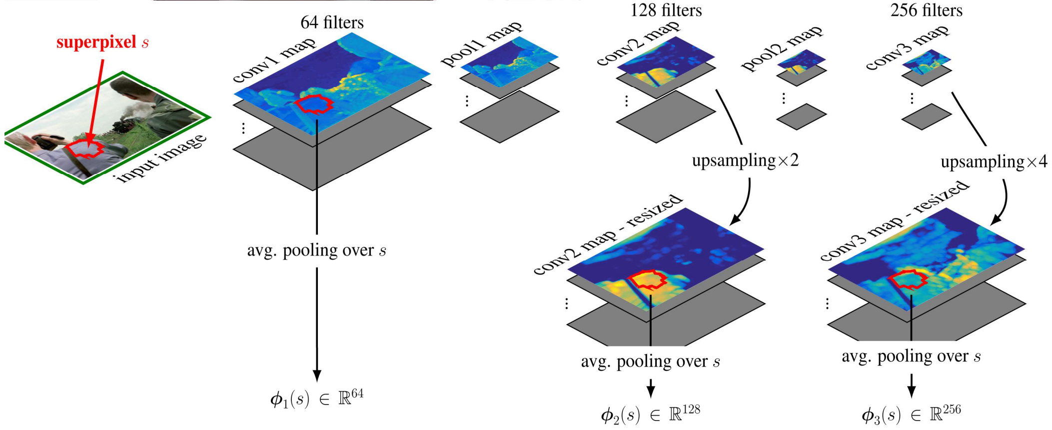

ResNet
7x7 conv, 64, /2
pool, /2
3x3 conv, 64
3x3 conv, 128, /2
3x3 conv, 128
3x3 conv, 256, /2
3x3 conv, 256
3x3 conv, 512, /2
3x3 conv, 512
avg pool
fc 1000

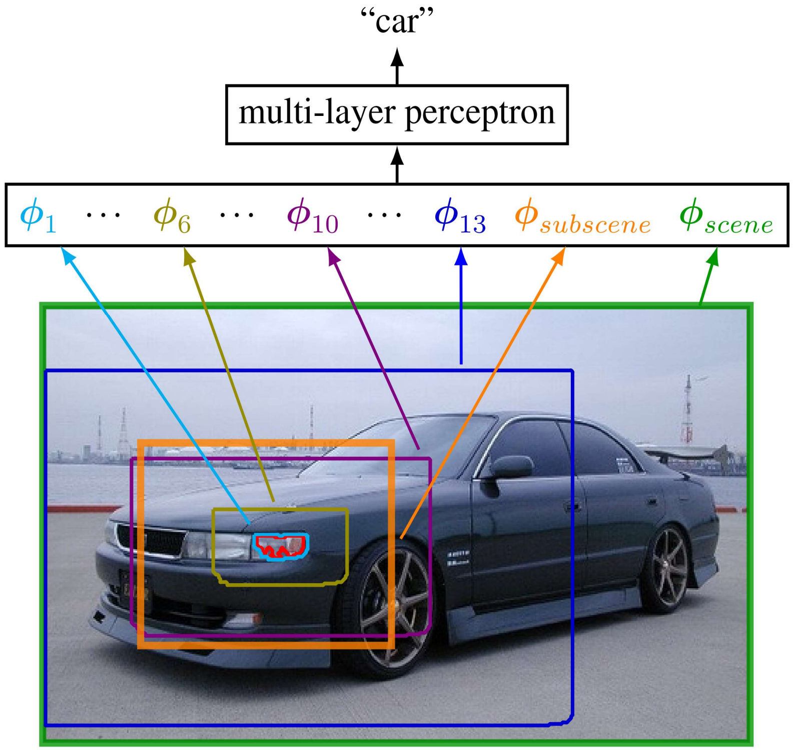

Input
$x_0$
$H_1$
$x_1$
$H_2$
$x_2$
$H_3$
$x_3$
$H_4$
$x_4$
BN-ReLU-Conv
Transition Layer
DenseNet


SE
X
U
$\tilde{X}$
$F_{tr}$
$F_{sq}(\cdot)$
$F_{ex}(\cdot, W)$
$F_{scale}(\cdot,\cdot)$
1×1×C
1×1×C
H'
W'
C'
H
W
C

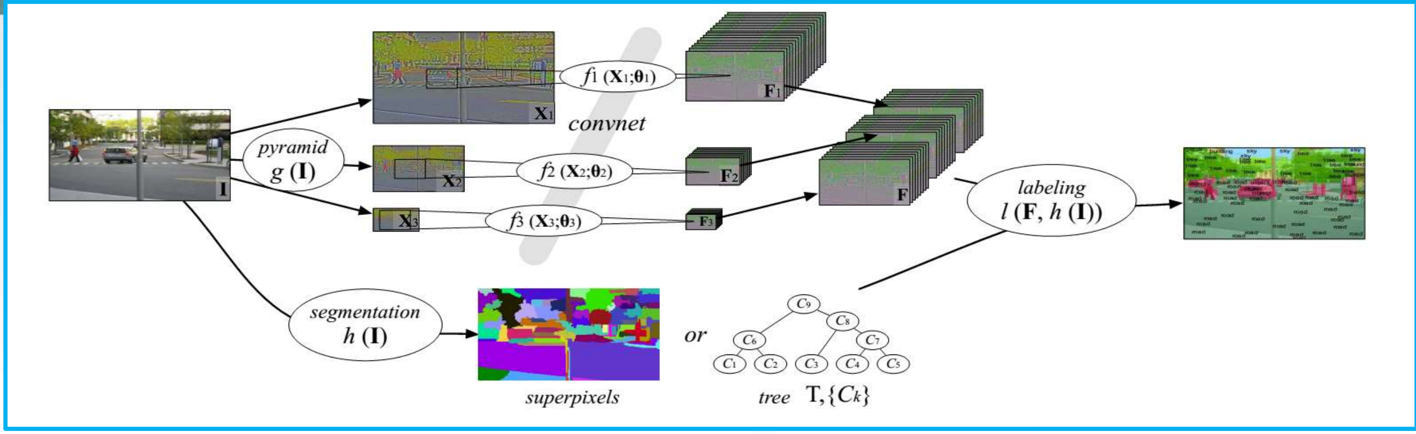
# Early Deep Learning Methods

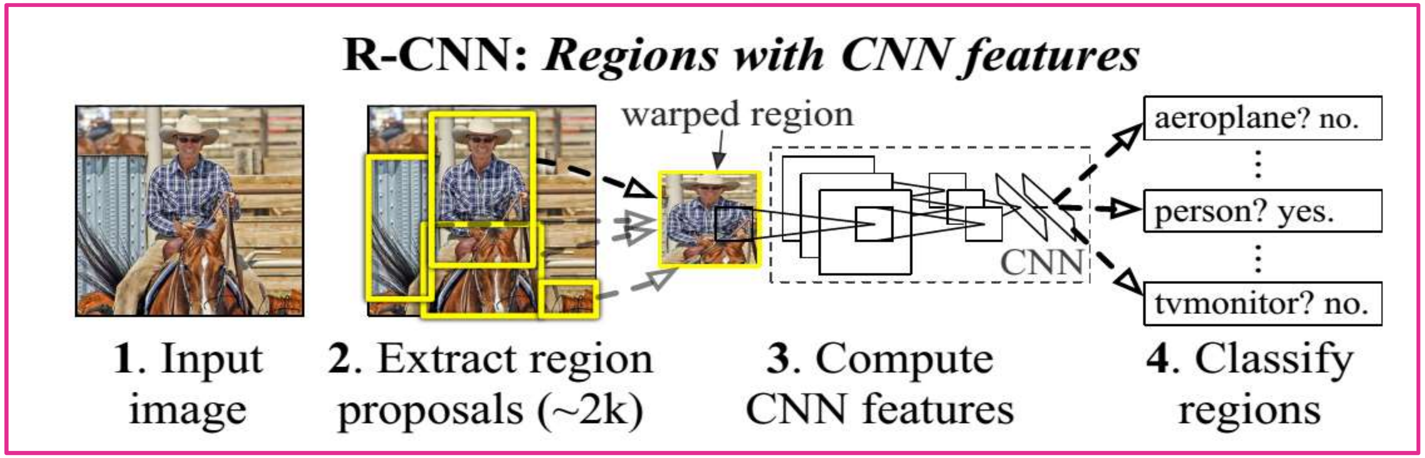


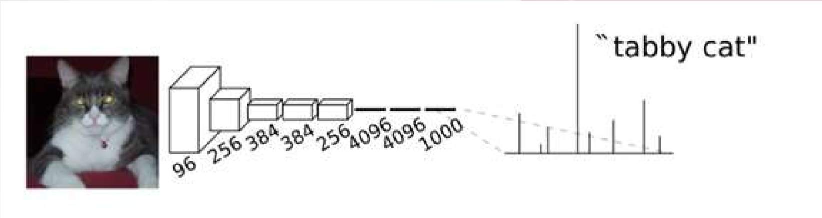



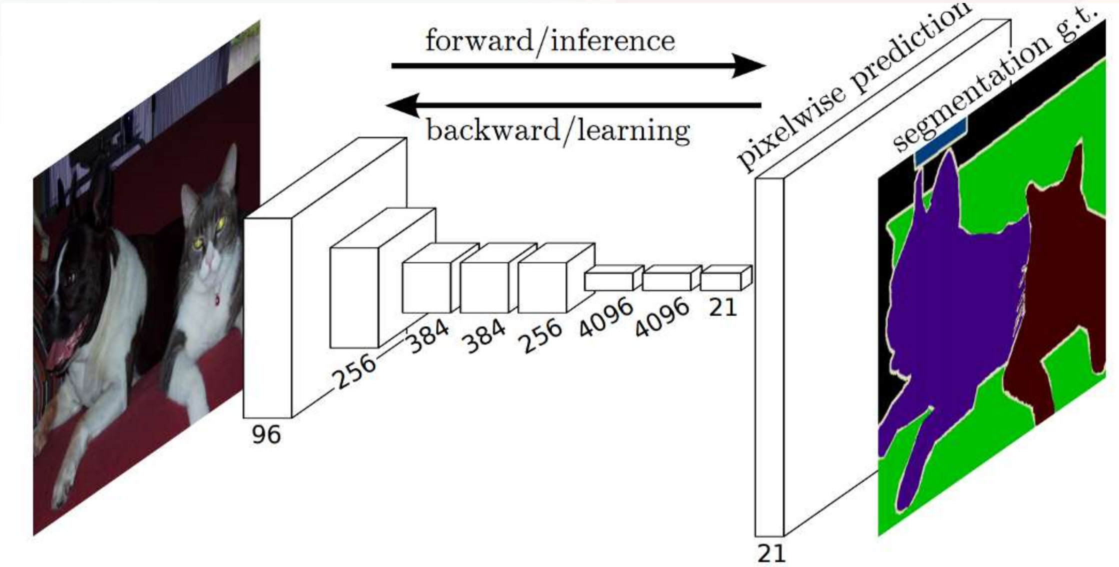
# Early Deep Learning Methods (cont.)



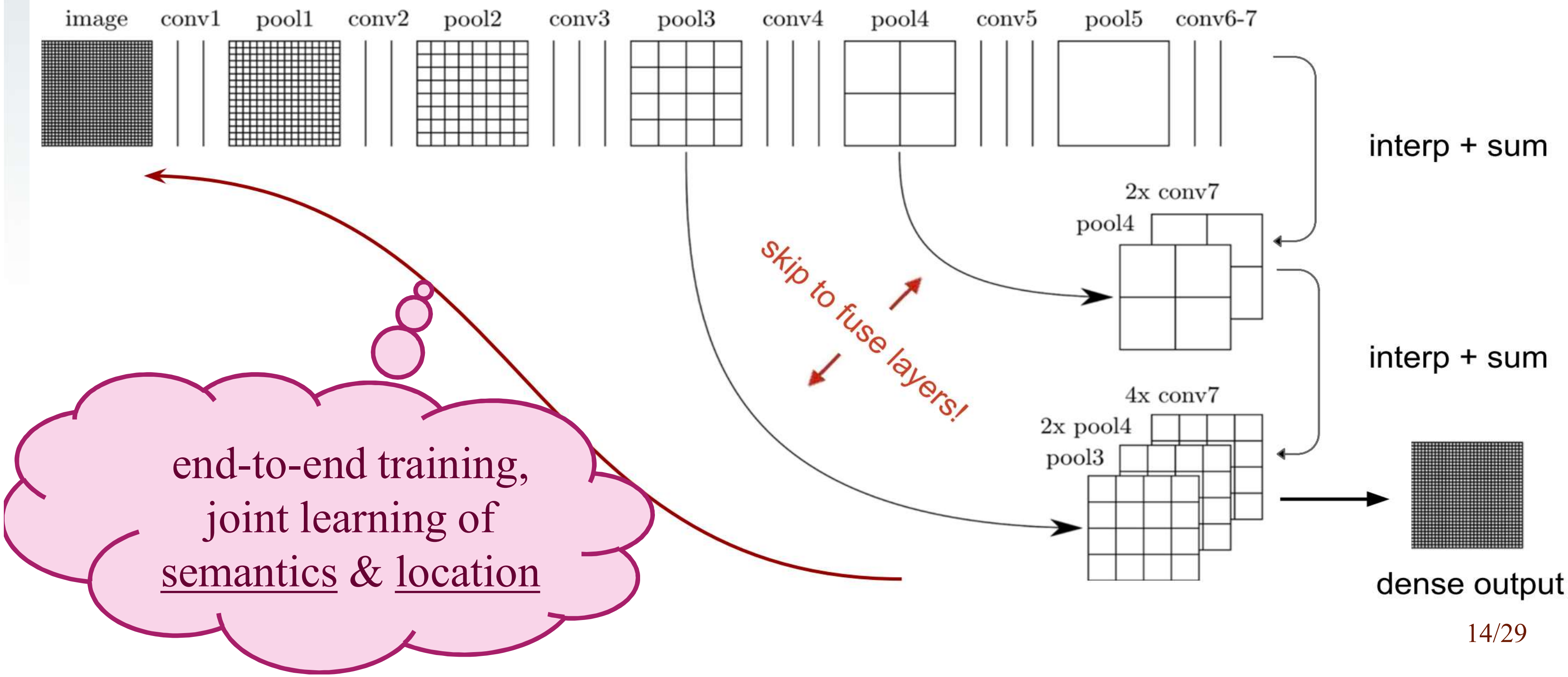



# Fully Convolutional Networks



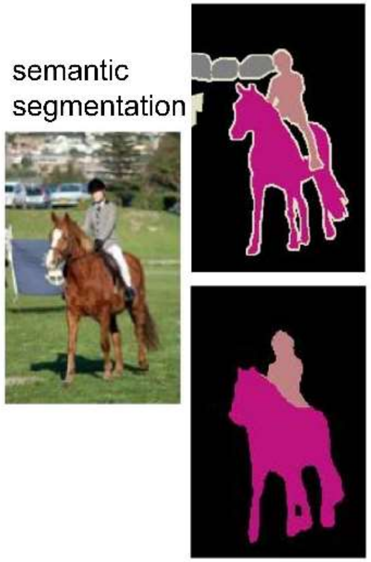



# Fully Convolutional Networks (cont.)

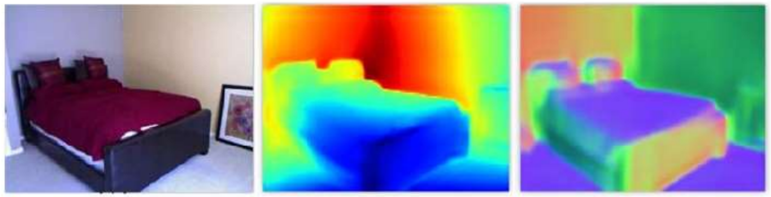


# Fully Convolutional Networks (cont.)

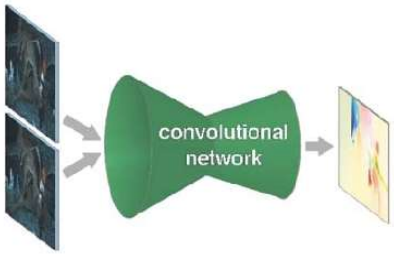
semantic segmentation



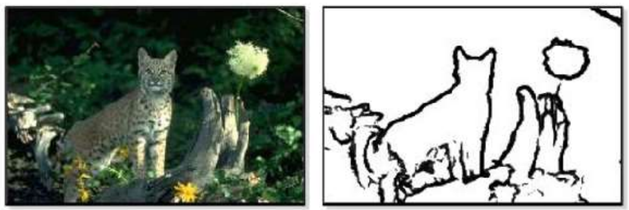
monocular depth + normals Eigen & Fergus 2015



colorization Zhang et al.2016



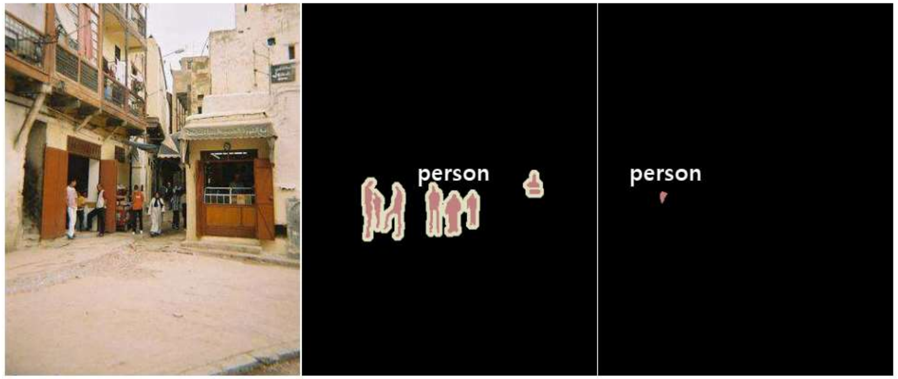

optical flow Fischer et al. 2015



boundary prediction Xie & Tu 2015

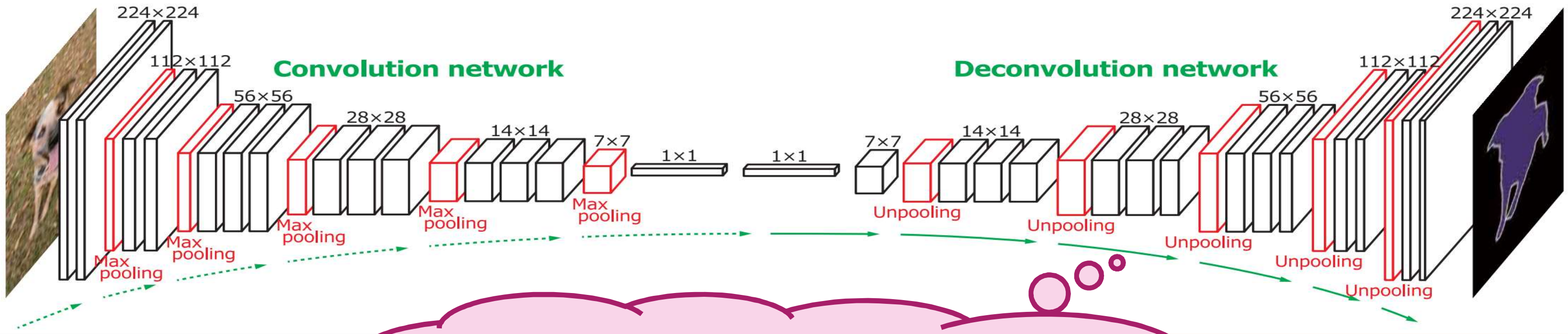
# Limitations of FCN


(a) Inconsistent labels due to large object size

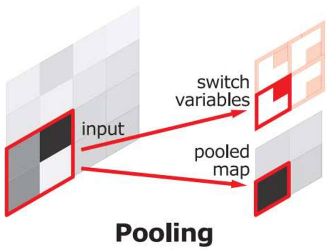


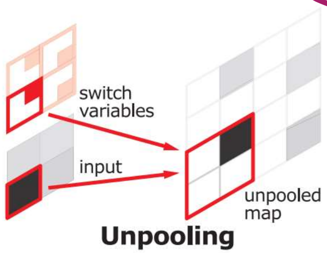

(b) Missing labels due to small object size

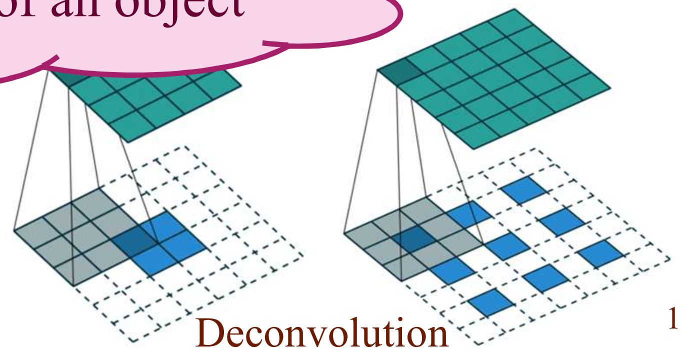
# DeconvNet



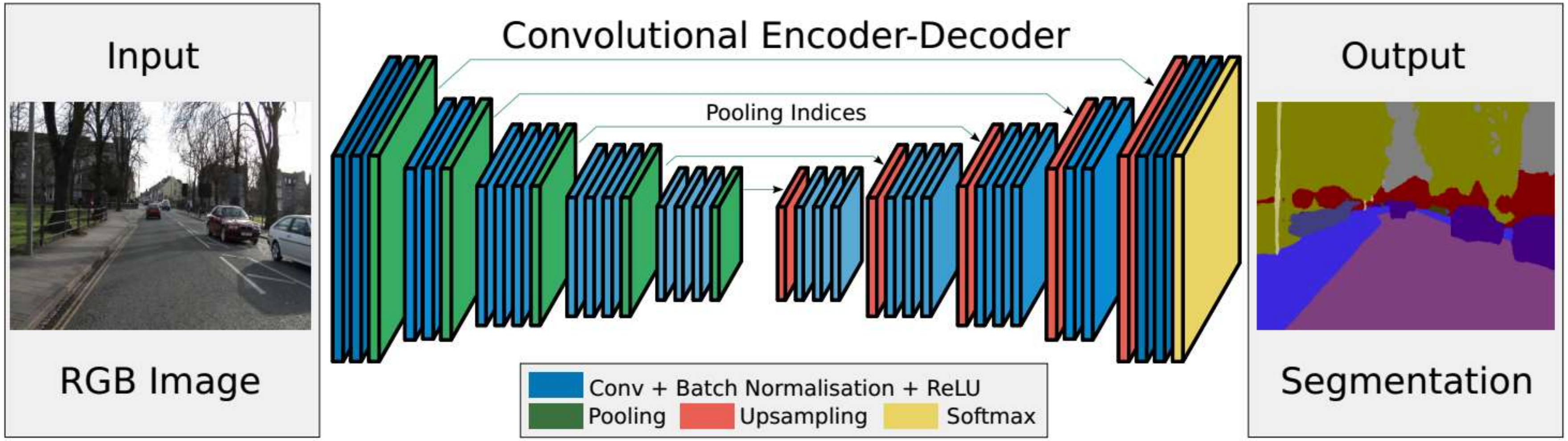

Reconstruction of coarse-to-fine structures of an object





Deconvolution

# SegNet

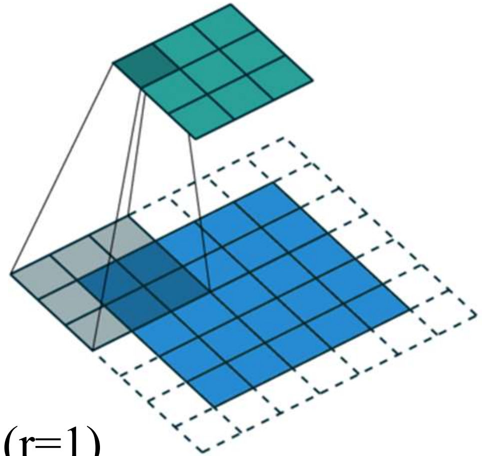



Smaller model size
Efficient memory requirements

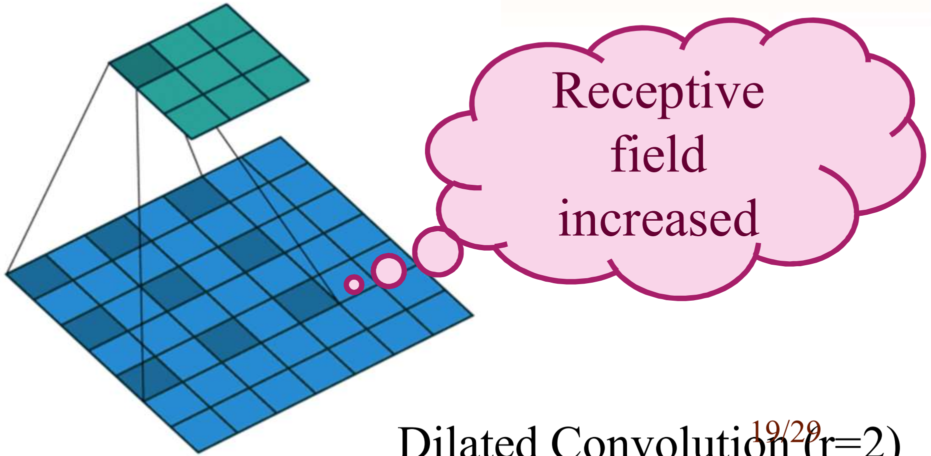
# Atrous convolution

Small output feature maps → Rough segmentation results → Reduce accuracy

Removing all down-sampling → Reduce receptive field → Reduce context → Reduce accuracy

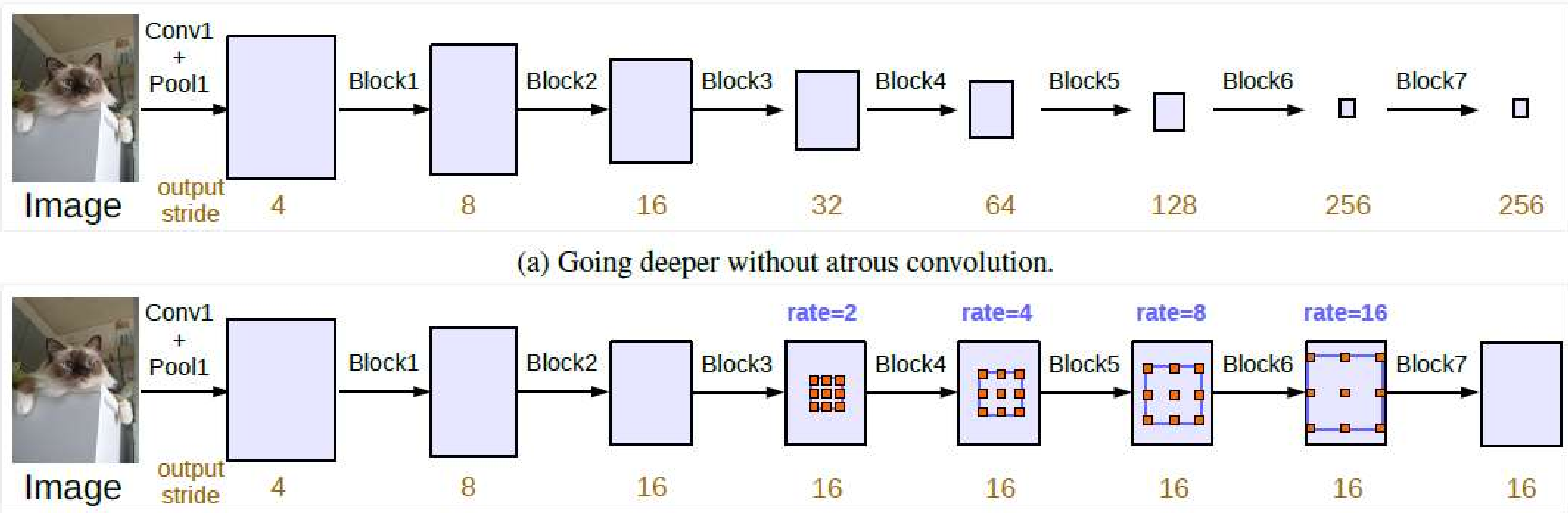


Standard Convolution (r=1)




Dilated Convolution (r=2)

# DeepLab



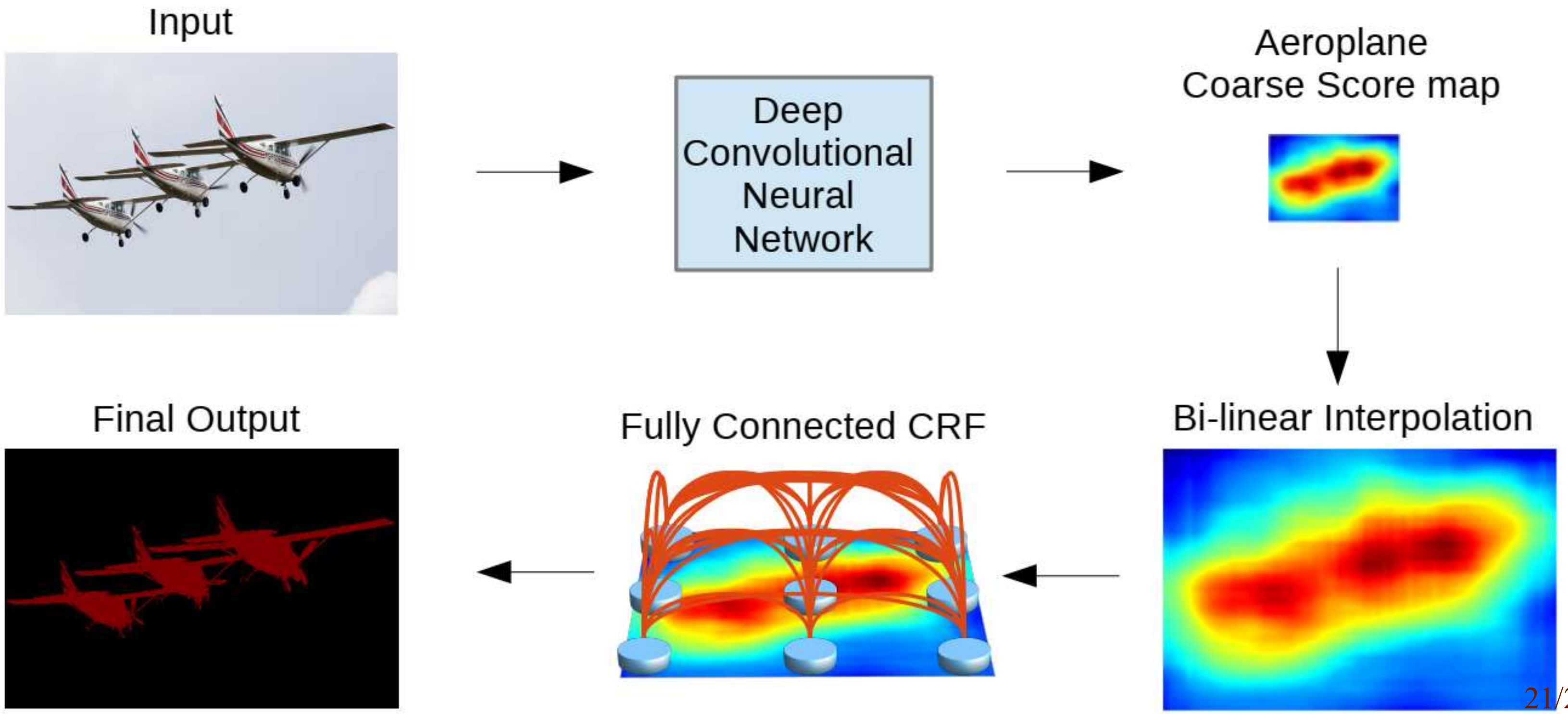

(a) Going deeper without atrous convolution.



(b) Going deeper with atrous convolution. Atrous convolution with $rate > 1$ is applied after block3 when $output\_stride = 16$.

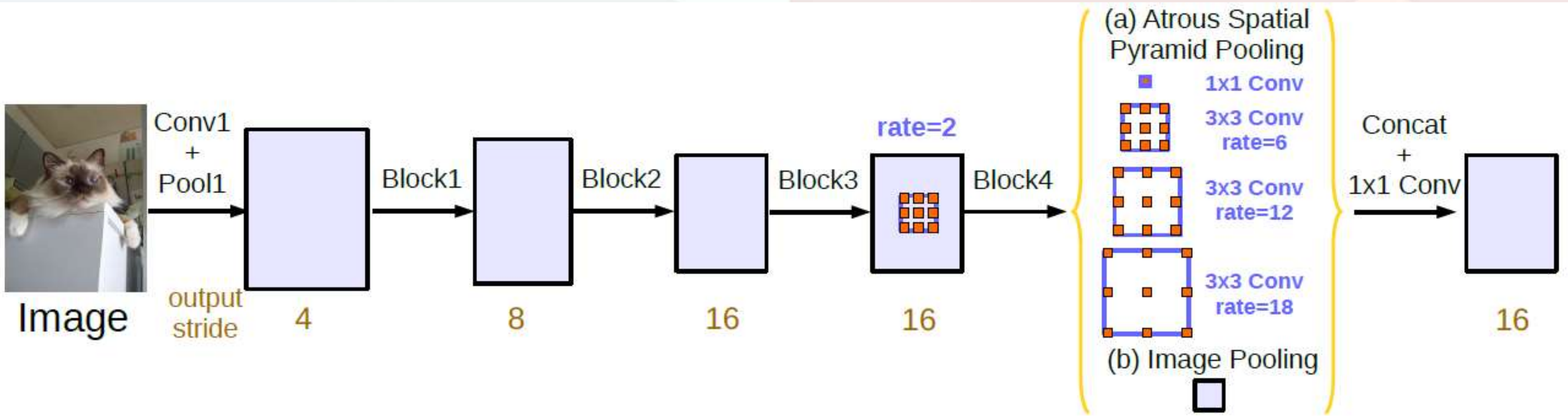
Keep stride constant but larger field-of-view

# DeepLab



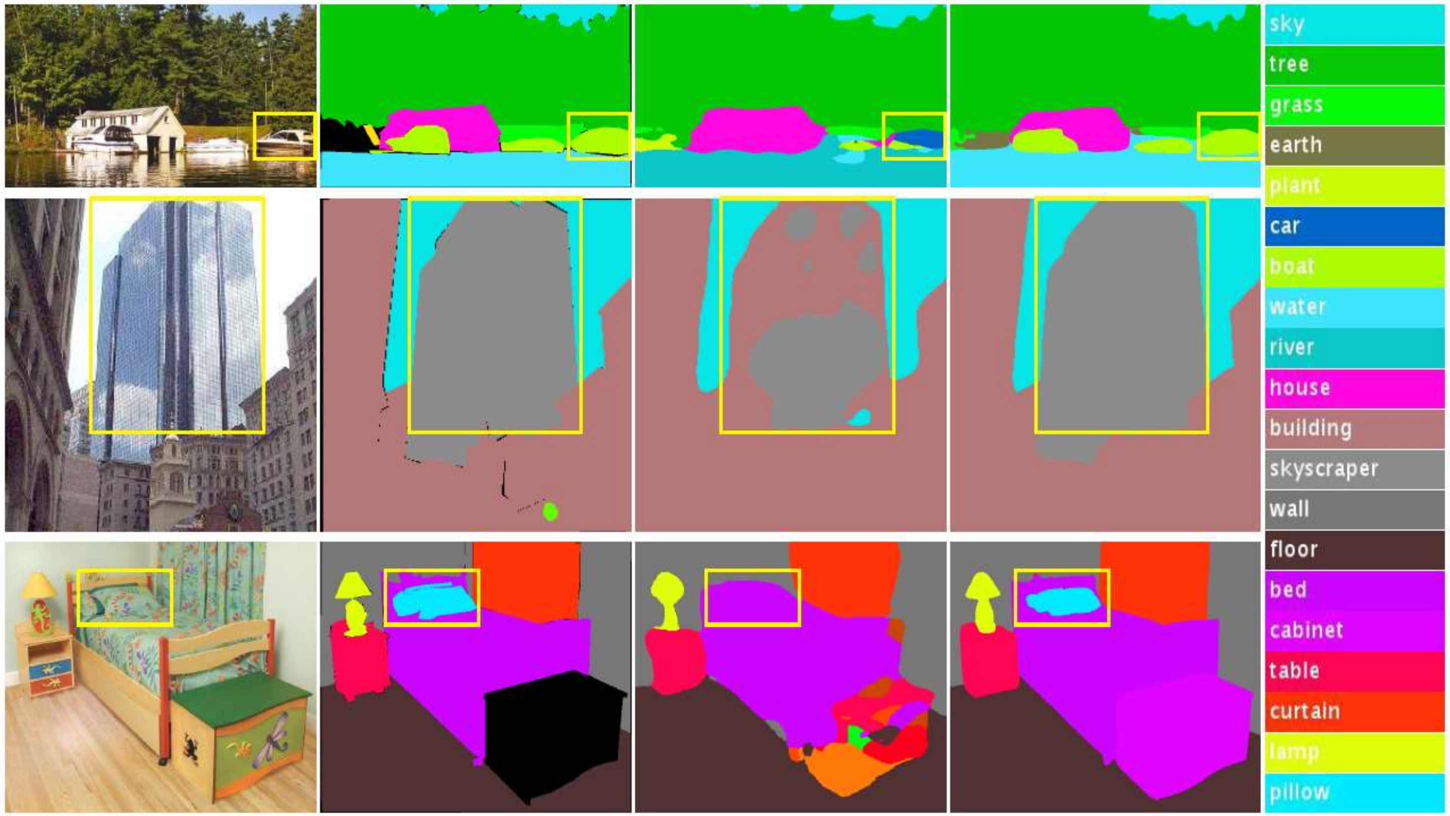
Input
Deep Convolutional Neural Network
Aeroplane Coarse Score map
Bi-linear Interpolation
Fully Connected CRF
Final Output

# DeepLabv2



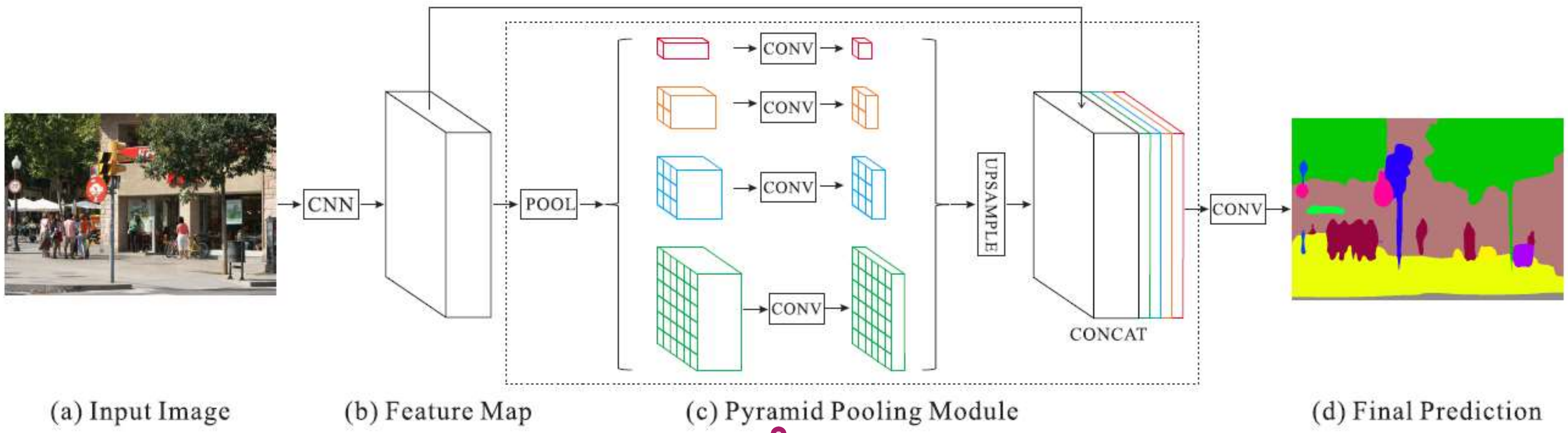

**Atrous Spatial Pyramid Pooling (ASPP)**

# PSPNet: **Pyramid Scene Parsing Network**



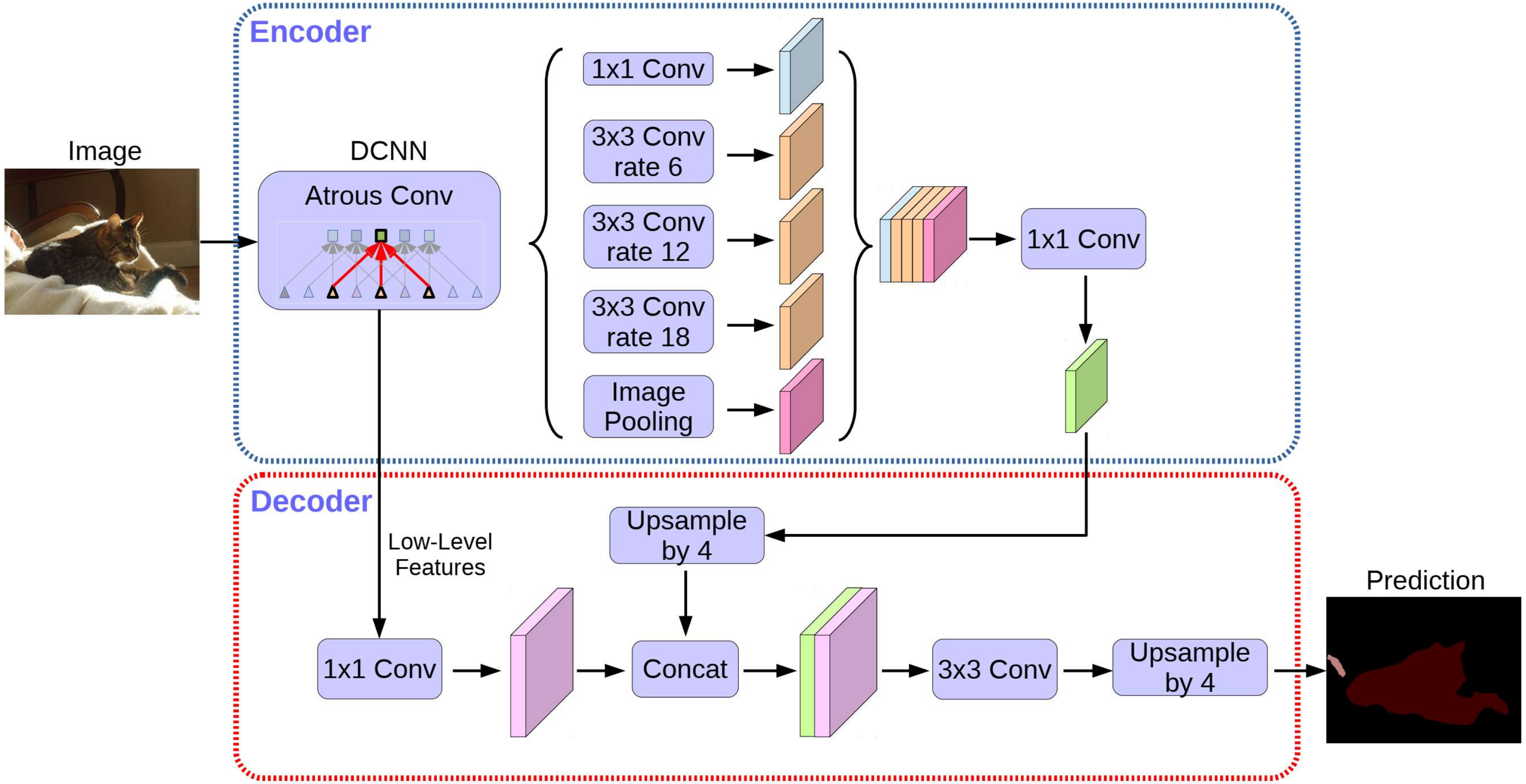

(a) Image (b) Ground Truth (c) FCN (d) PSPNet (e) ColorMap

# PSPNet: Pyramid Scene Parsing Network




obtained global information of the image

# DeepLab v3+



Encoder
Image
DCNN
Atrous Conv
1x1 Conv
3x3 Conv rate 6
3x3 Conv rate 12
3x3 Conv rate 18
Image Pooling
1x1 Conv
Decoder
Low-Level Features
Upsample by 4
1x1 Conv
Concat
3x3 Conv
Upsample by 4
Prediction

# Some Results

<u>Datasets</u>

- ✓ ADE2K
  - 20K/2K/3K + 150 classes
- ✓ PASCAL VOC 2012
  - 10582/1449/1456 + 20 classes
- ✓ Cityscapes
  - 2975/500/1525 +19 classes

- ✓ CamVid, SUN-RGBD, NYU-V2, Stanford-2D-3D-S, SceneNet, Matterport3D

| Method | PascalVoc | Cityscape | ADE2K |
|---|---|---|---|
| Zoom-out | 69.6 | - | - |
| FCN | 62.2 | 65.3 | 44.8 |
| SegNet | - | 57.0 | 40.8 |
| DeconvNet | 72.5 | | - |
| DeepLab | 79.7 | 70.4 | - |
| PSPNet | 85.4 | 78.4 | 57.2 |
| DeepLabv3+ | 89.0 | 82.1 | - |

# Other directions

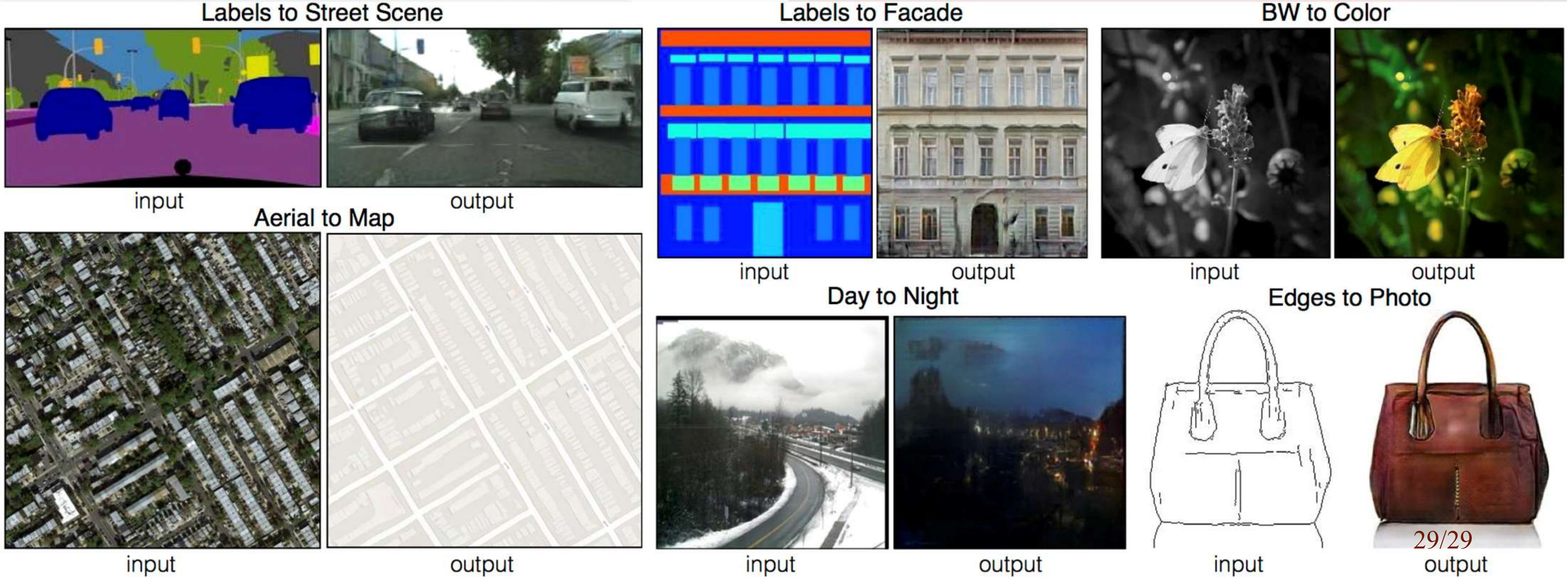
- RGB-D Semantic Segmentation

# Other directions

- 3D Pointclouds Semantic Segmentation

# Other directions

- Semi-supervised semantic segmentation using GAN



# References


- Girshick, Ross, et al. "Rich feature hierarchies for accurate object detection and semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2014.
- Long, Jonathan, Evan Shelhamer, and Trevor Darrell. "Fully convolutional networks for semantic segmentation." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- Noh, Hyeonwoo, Seunghoon Hong, and Bohyung Han. "Learning deconvolution network for semantic segmentation." Proceedings of the IEEE international conference on computer vision. 2015.
- Mostajabi, Mohammadreza, Payman Yadollahpour, and Gregory Shakhnarovich. "Feedforward semantic segmentation with zoom-out features." Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
- Chen, Liang-Chieh, et al. "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs." IEEE transactions on pattern analysis and machine intelligence 40.4 (2017): 834-848.
- Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. "Segnet: A deep convolutional encoder-decoder architecture for image segmentation." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2481-2495.
- Chen, Liang-Chieh, et al. "Encoder-decoder with atrous separable convolution for semantic image segmentation." Proceedings of the European conference on computer vision (ECCV). 2018.
- Zhao, Hengshuang, et al. "Pyramid scene parsing network." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- Isola, Phillip, et al. "Image-to-image translation with conditional adversarial networks." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
- Hazirbas, Caner, et al. "Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture." Asian conference on computer vision. Springer, Cham, 2016.
- Li, Zhen, et al. "Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling." European conference on computer vision. Springer, Cham, 2016.

# Thanks for your Attention!

?