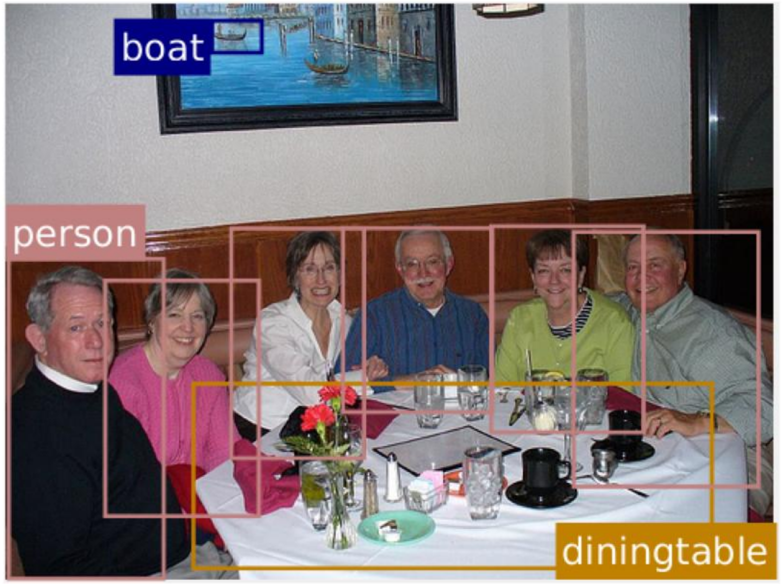
Semantic Segmentation

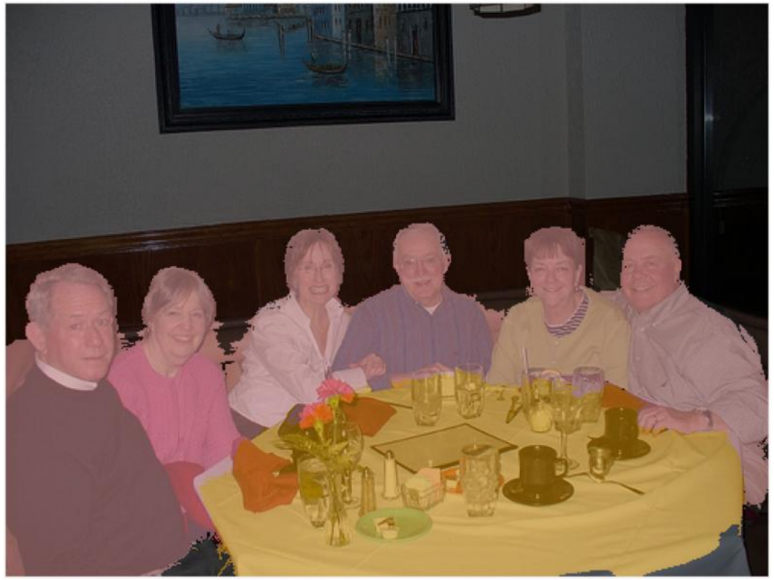
CS 4391 Introduction Computer Vision

Professor Yu Xiang
The University of Texas at Dallas

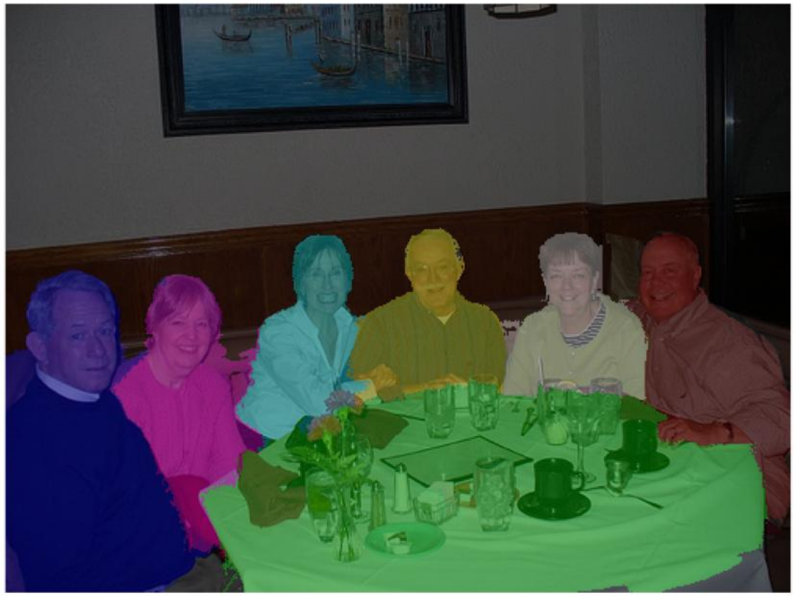
Semantic Understanding



Object Detection



Semantic Segmentation



Instance Segmentation

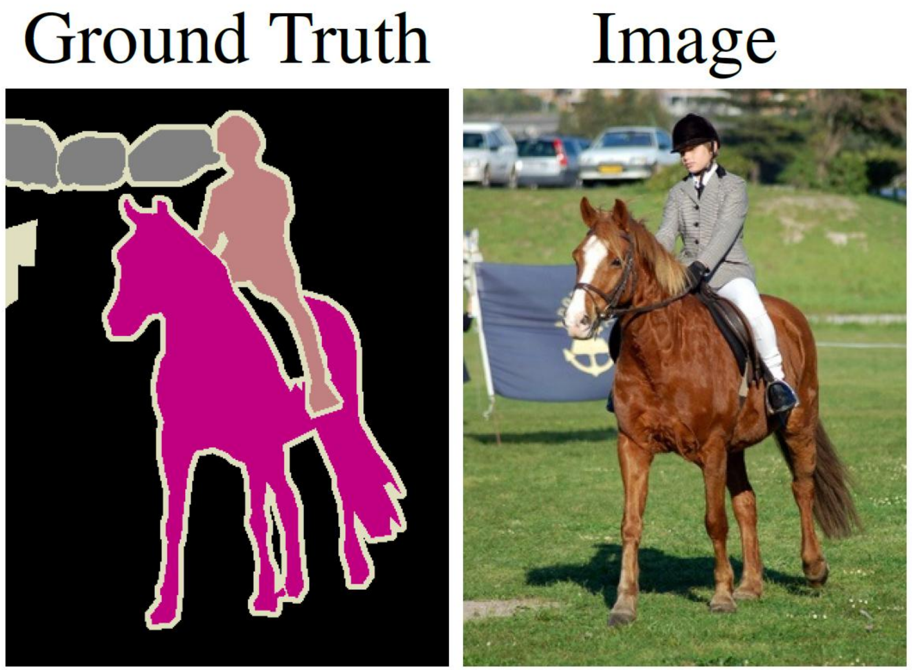
Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation. Arnab et al., IEEE SIGNAL PROCESSING MAGAZINE, 2018

Semantic Segmentation

- Label pixels into semantic classes

- Naïve method
  - Classify each pixel independently

- Better idea
  - Using context of pixels



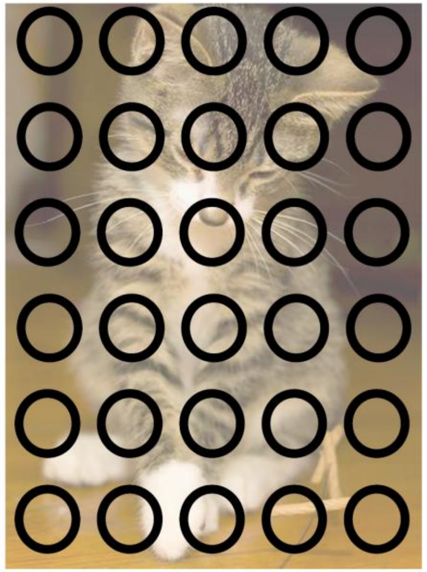
Conditional Random Fields (CRFs)

- Pixel labeling problem

graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

2D grid for images

$X_1 \in \{\text{bg, cat, dog, person}\}$



Conditional Random Fields (CRFs)

- Model the conditional probability distribution

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp\left(-\sum_{c \in \mathcal{C}_{\mathcal{G}}} \phi_c(\mathbf{X}_c|\mathbf{I})\right)$$

label

image

Partition function
(normalization factor)

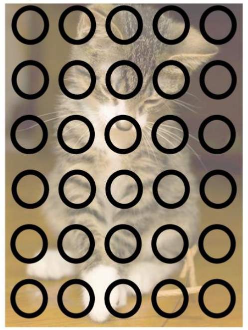
clique

Potential function

graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

2D grid for images

$X_1 \in$ {bg, cat, dog, person}



Conditional Random Fields (CRFs)

$$P(\mathbf{X}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-\sum_{c \in \mathcal{C}_{\mathcal{G}}} \phi_c(\mathbf{X}_c|\mathbf{I}))$$

- Energy function $E(\mathbf{x}|\mathbf{I}) = \sum_{c \in \mathcal{C}_{\mathcal{G}}} \phi_c(\mathbf{x}_c|\mathbf{I})$ $\quad \mathbf{x} \in \mathcal{L}^N$

$$P(\mathbf{x}|\mathbf{I}) = \frac{1}{Z(\mathbf{I})} \exp(-E(\mathbf{x}|\mathbf{I})) \qquad Z(\mathbf{I}) = \sum_{\mathbf{x}} \exp(-E(\mathbf{x}|\mathbf{I}))$$

- Maximum a posteriori (MAP) labeling

$$\mathbf{x}^* = \arg\max_{\mathbf{x} \in \mathcal{L}^N} P(\mathbf{x}|\mathbf{I})$$

Conditional Random Fields (CRFs)

- Unary potential and pairwise potential

$$E(\mathbf{x}, I) := \sum_{u \in V} \psi_u(X_u = x_u | I) + \sum_{\{u,v\} \in \mathcal{E}} \psi_{u,v}(X_u = x_u, X_v = x_v | I)$$

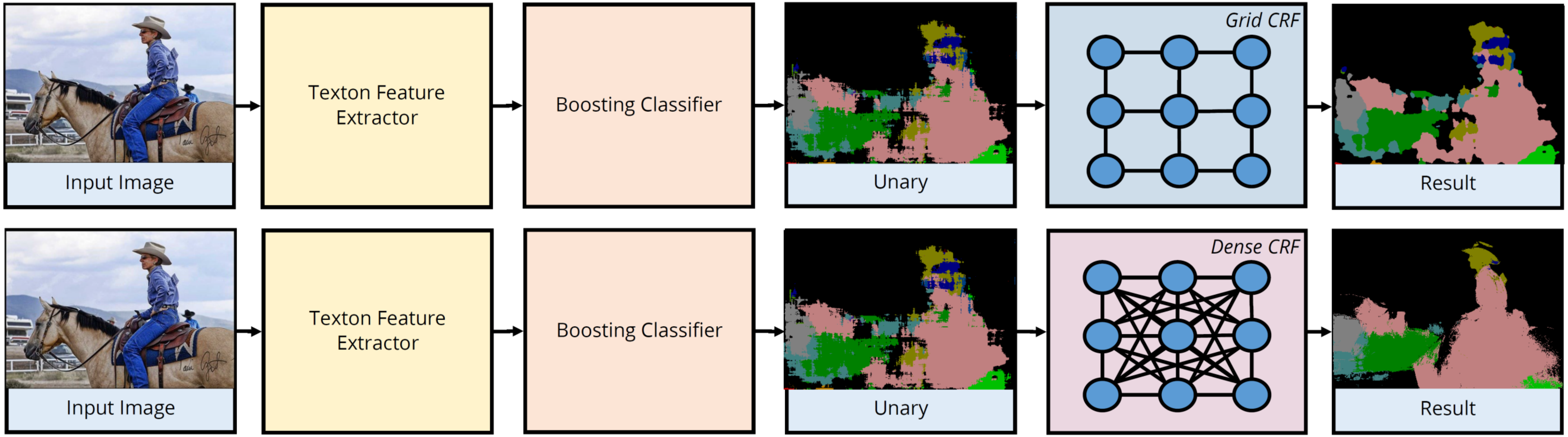
E.g., classifier output

E.g., smoothing pairwise potential $[x_u \neq x_v]$

- Energy minimization problem
  - NP-hard
  - Exact and approximate algorithms exist to obtain acceptable solutions

A Comparative Study of Modern Inference Techniques for Structured Discrete Energy Minimization Problems. Kappes, et al., IJCV, 2015

Conditional Random Fields (CRFs)



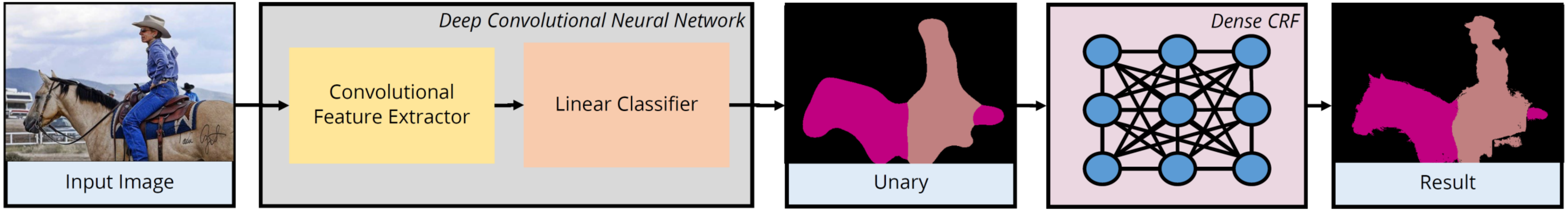
$$E(\mathbf{x}) = \sum_i \psi_u(x_i) + \sum_{i<j} \psi_p(x_i, x_j)$$

Efficient Inference in Fully Connected CRFs with Gaussian Edge Potentials. Krähenbühl & Koltun, NeurIPS, 2011

Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation. Arnab et al., IEEE SIGNAL PROCESSING MAGAZINE, 2018

Combining Neural Networks with CRFs

- Utilize neural networks to compute unary potentials

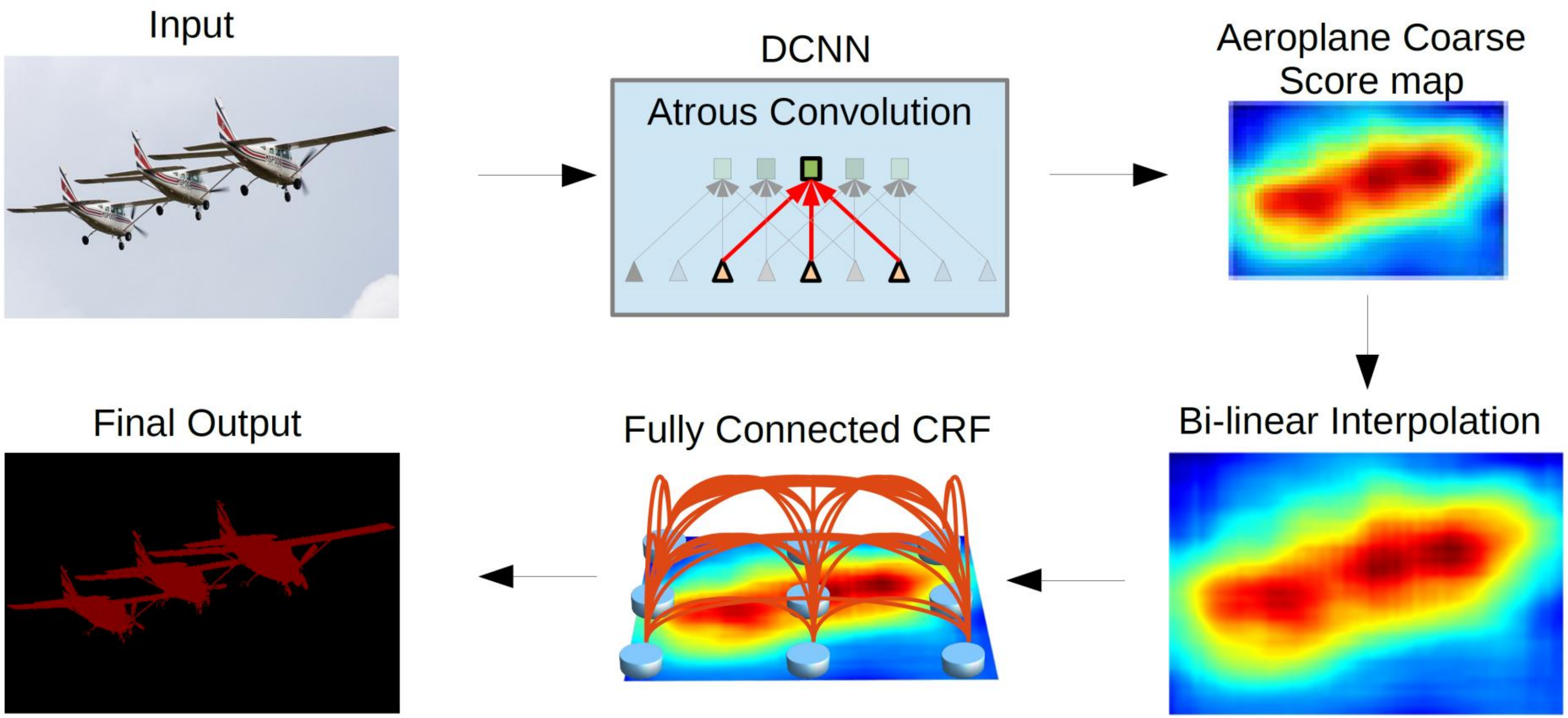


Better classifier

Semantic image segmentation with deep convolutional nets and fully connected CRFs. Chen et al., ICLR, 2015.

Conditional Random Fields Meet Deep Neural Networks for Semantic Segmentation. Arnab et al., IEEE SIGNAL PROCESSING MAGAZINE, 2018

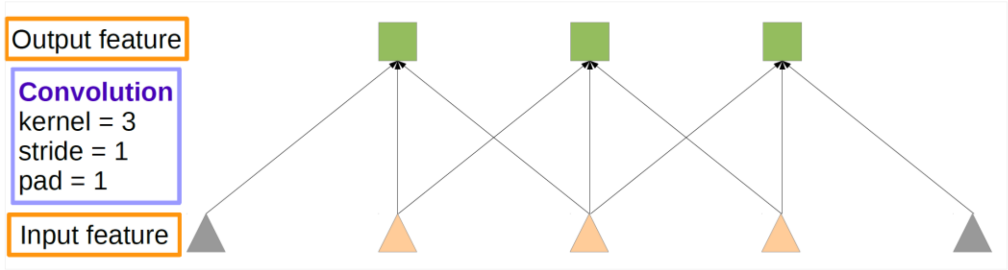
DeepLab



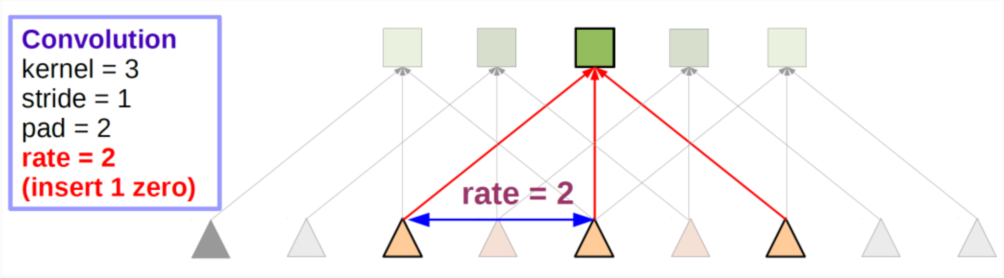
DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. Chen et al., 2016

DeepLab

Atrous convolution

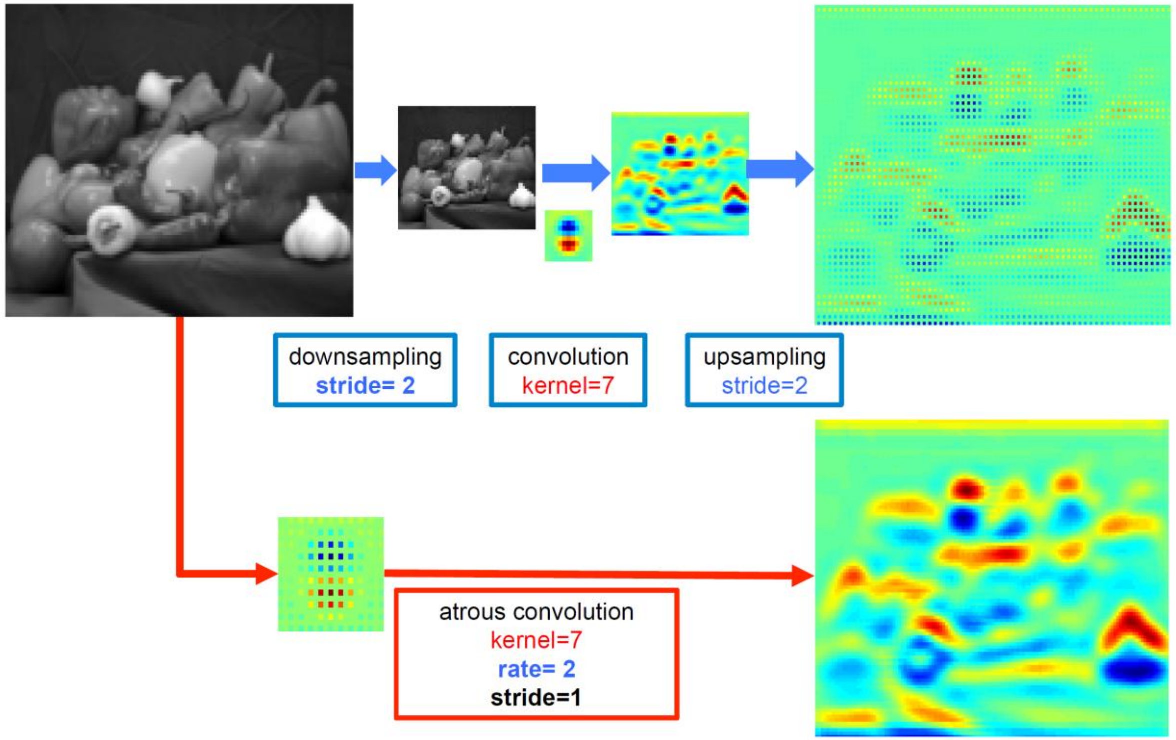


(a) Sparse feature extraction



(b) Dense feature extraction

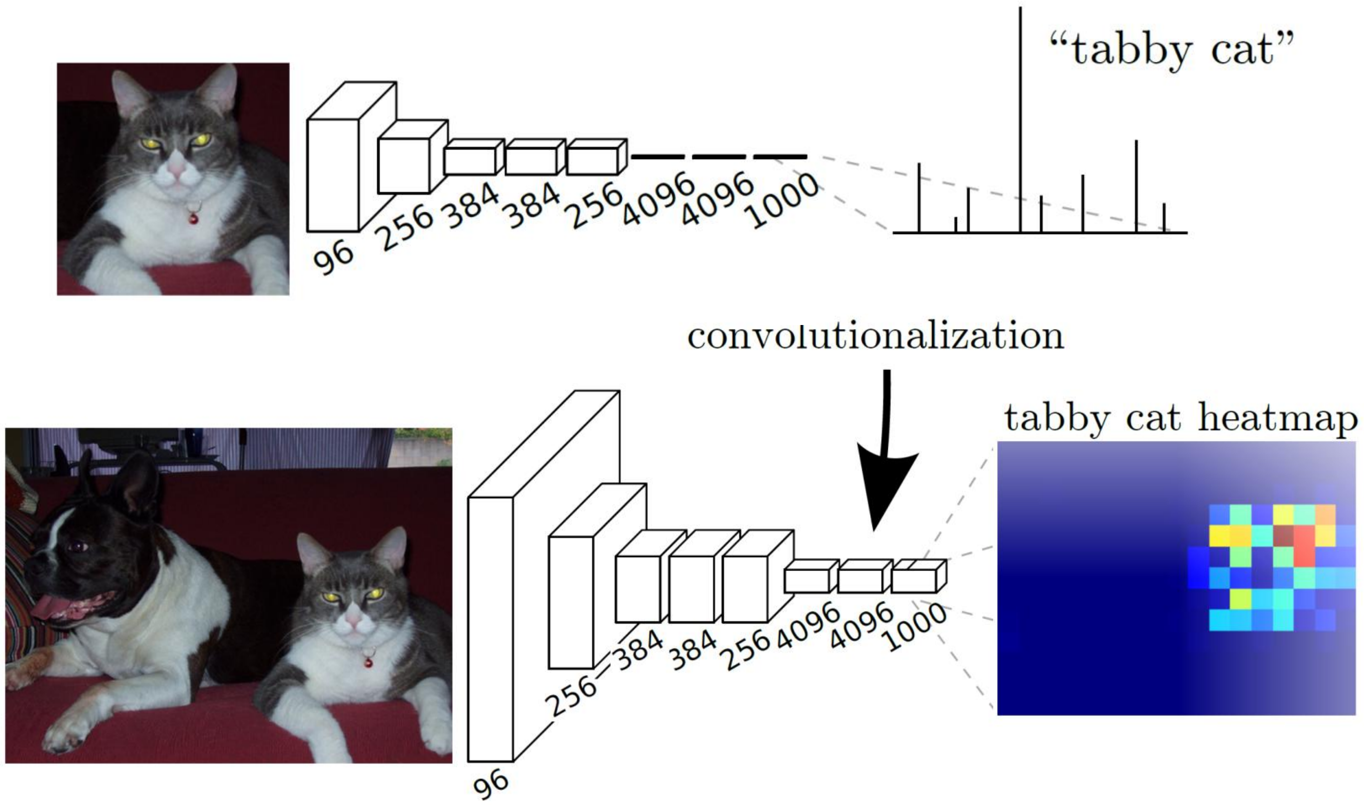
$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k]w[k]$$



DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. Chen et al., 2016

Fully Convolutional Networks

- Adapt classification networks for dense prediction



Treat FC layers as convolutions with kernels that cover the entire input regions

Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015

Fully Convolutional Networks

- Convert AlexNet

[224x224x3] INPUT
[55x55x96] CONV1: 96 11x11 filters at stride 4, pad 0
[27x27x96] MAX POOL1: 3x3 filters at stride 2
[27x27x96] NORM1: Normalization layer
[27x27x256] CONV2: 256 5x5 filters at stride 1, pad 2
[13x13x256] MAX POOL2: 3x3 filters at stride 2
[13x13x256] NORM2: Normalization layer
[13x13x384] CONV3: 384 3x3 filters at stride 1, pad 1
[13x13x384] CONV4: 384 3x3 filters at stride 1, pad 1
[13x13x256] CONV5: 256 3x3 filters at stride 1, pad 1
[6x6x256] MAX POOL3: 3x3 filters at stride 2
[4096] FC6: 4096 neurons
[4096] FC7: 4096 neurons
[1000] FC8: 1000 neurons (class scores)

```
layer {
  name: "fc6"
  type: "Convolution"
  bottom: "pool5"
  top: "fc6"
  convolution_param {
    num_output: 4096
    pad: 0
    kernel_size: 6
    group: 1
    stride: 1
  }
}
```

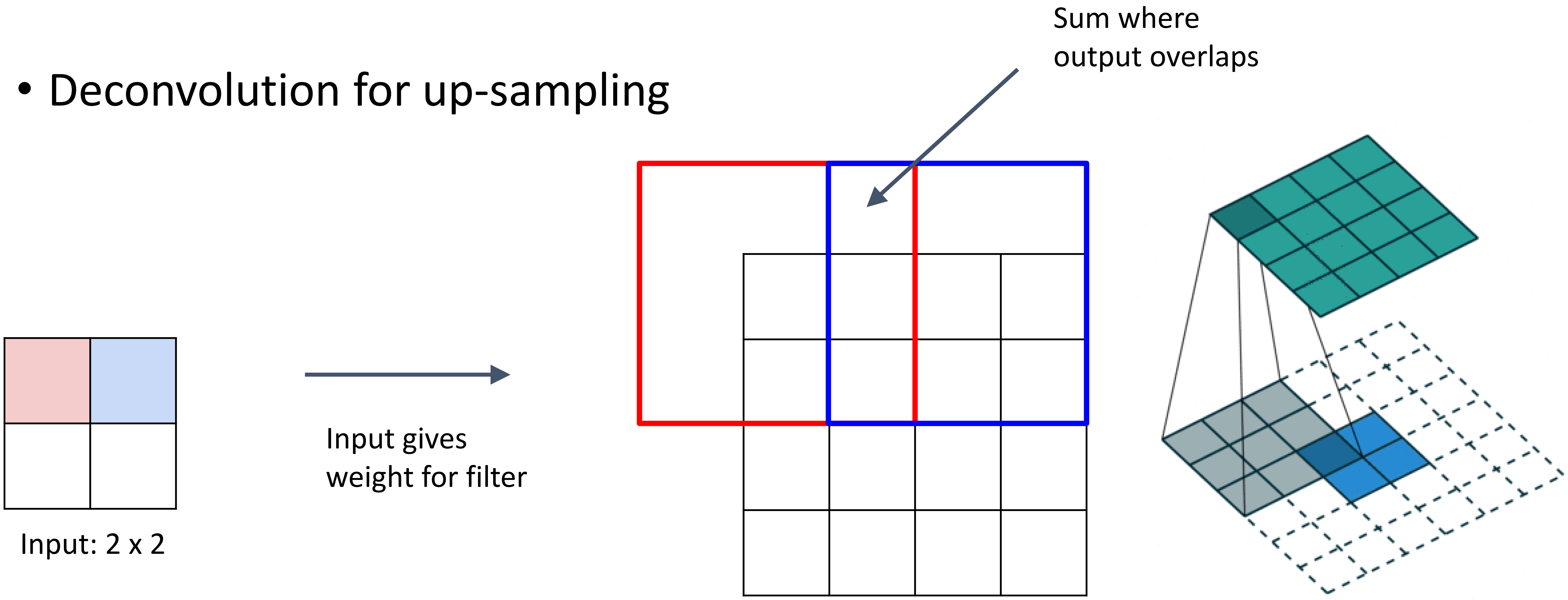
```
layer {
  name: "fc7"
  type: "Convolution"
  bottom: "fc6"
  top: "fc7"
  convolution_param {
    num_output: 4096
    pad: 0
    kernel_size: 1
    group: 1
    stride: 1
  }
}
```

```
layer {
  name: "score_fr"
  type: "Convolution"
  bottom: "fc7"
  top: "score_fr"
  param {
    lr_mult: 1
    decay_mult: 1
  }
  param {
    lr_mult: 2
    decay_mult: 0
  }
  convolution_param {
    num_output: 21
    pad: 0
    kernel_size: 1
  }
}
```

Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015

Fully Convolutional Networks

- Deconvolution for up-sampling



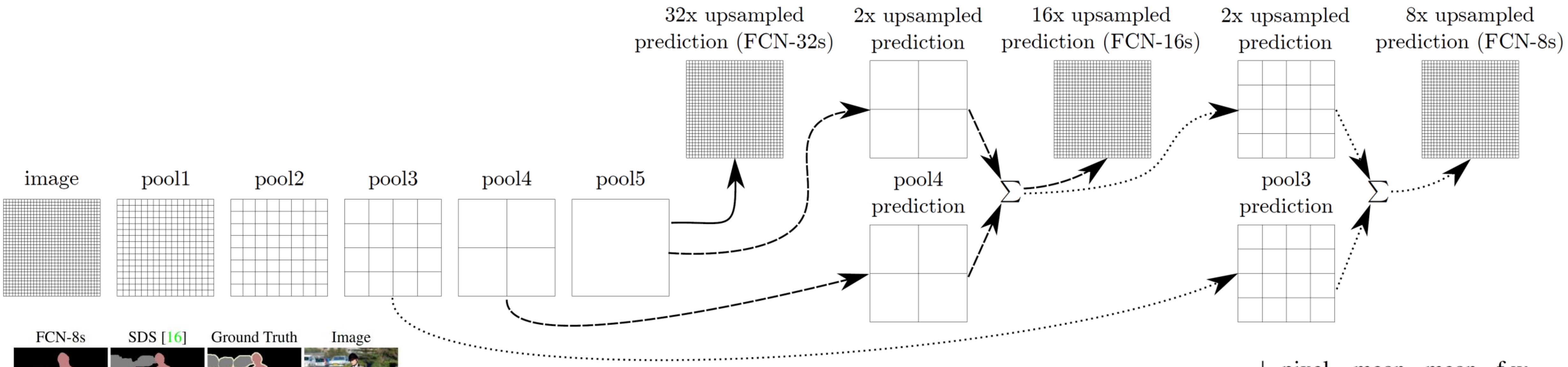
3 x 3 “deconvolution”, stride 2, pad 1

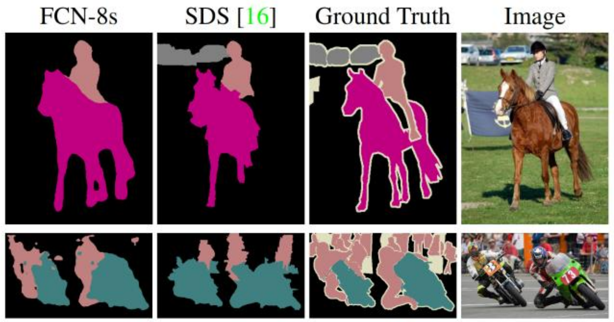
Output: 4 x 4

Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015

Fully Convolutional Networks

- Combine predictions with different resolutions



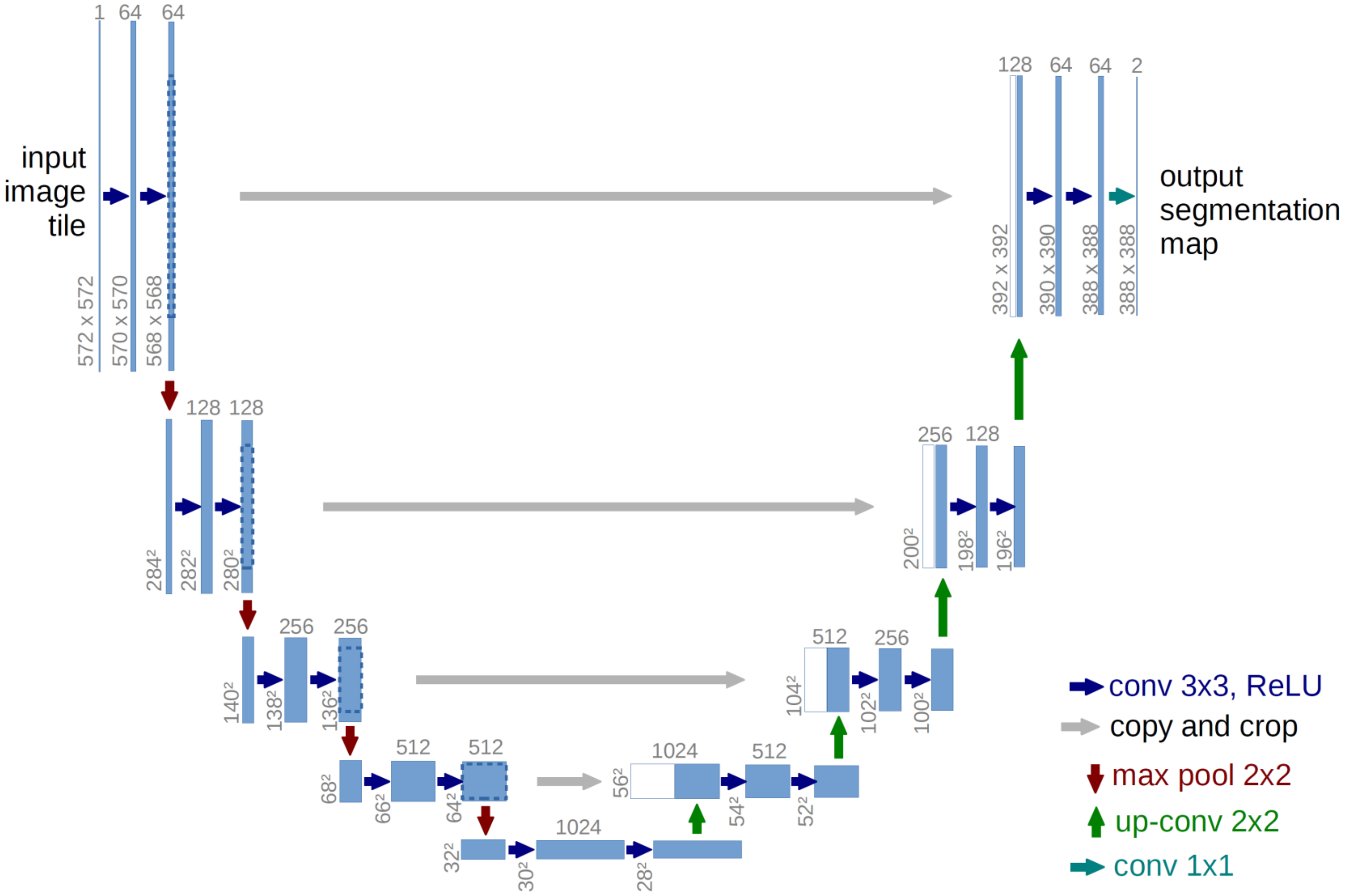


	pixel acc.	mean acc.	mean IU	f.w. IU
FCN-32s-fixed	83.0	59.7	45.4	72.0
FCN-32s	89.1	73.3	59.4	81.4
FCN-16s	90.0	75.7	62.4	83.0
FCN-8s	**90.3**	**75.9**	**62.7**	**83.2**

Fully Convolutional Networks for Semantic Segmentation. Long et al., CVPR, 2015

U-Net

Convolution

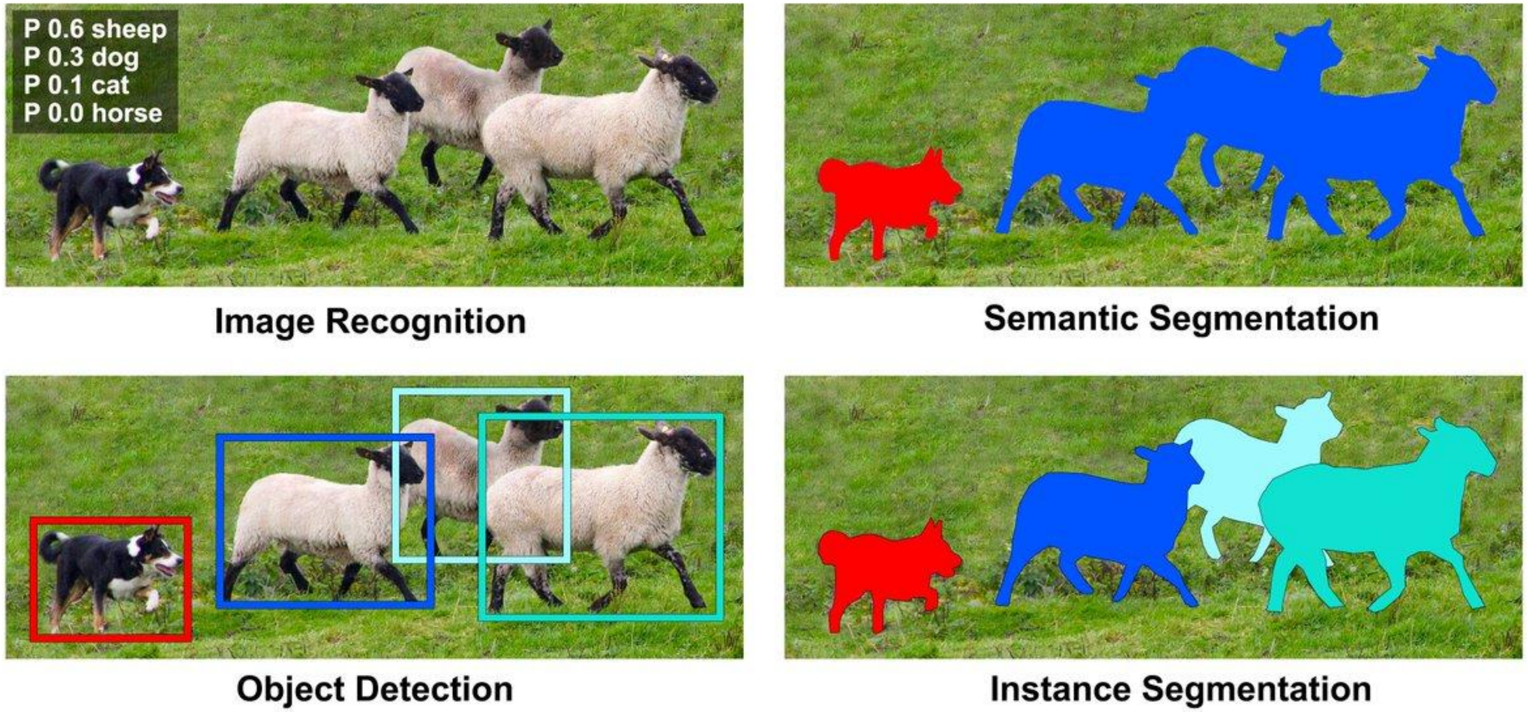


Deconvolution

U-Net: Convolutional Networks for Biomedical Image Segmentation, Ronneberger et al., MICCAI 2015

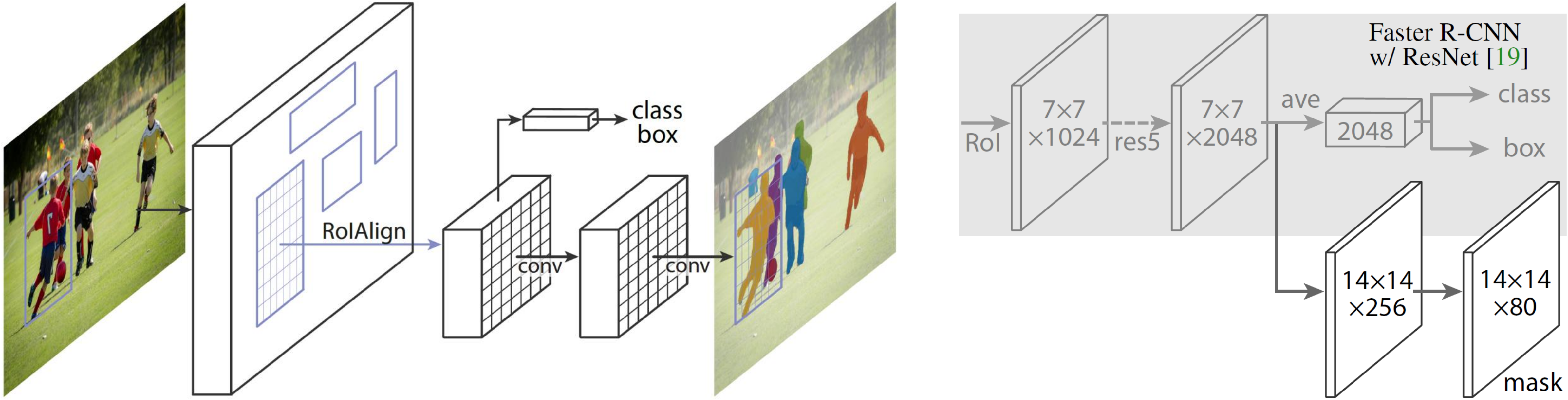
Instance Segmentation

- Separate object instances in the same class
- Detection + segmentation



https://ai-pool.com/d/could-you-explain-me-how-instance-segmentation-works

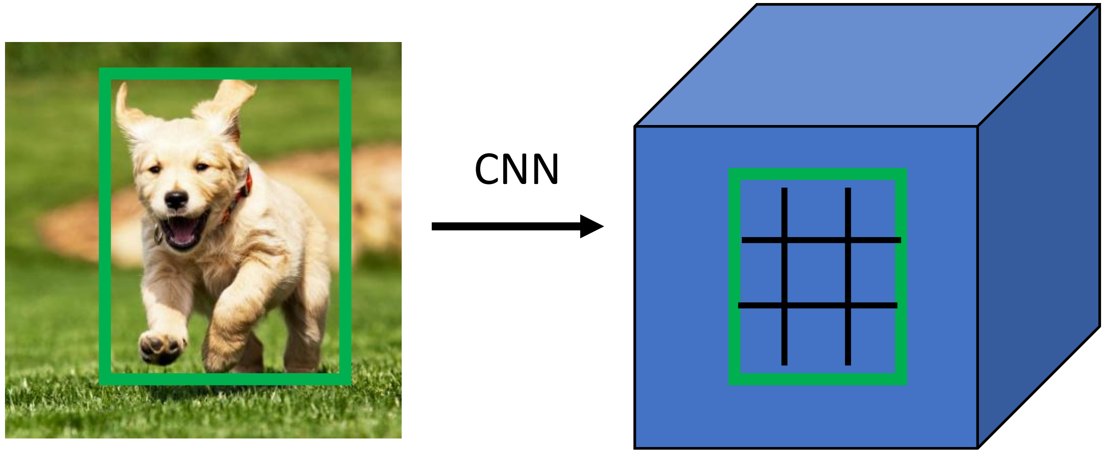
Mask R-CNN



'res5' denotes ResNet's fifth stage

Mask R-CNN. He et al., ICCV, 2017

RoI Pooling vs. RoI Align



RoI

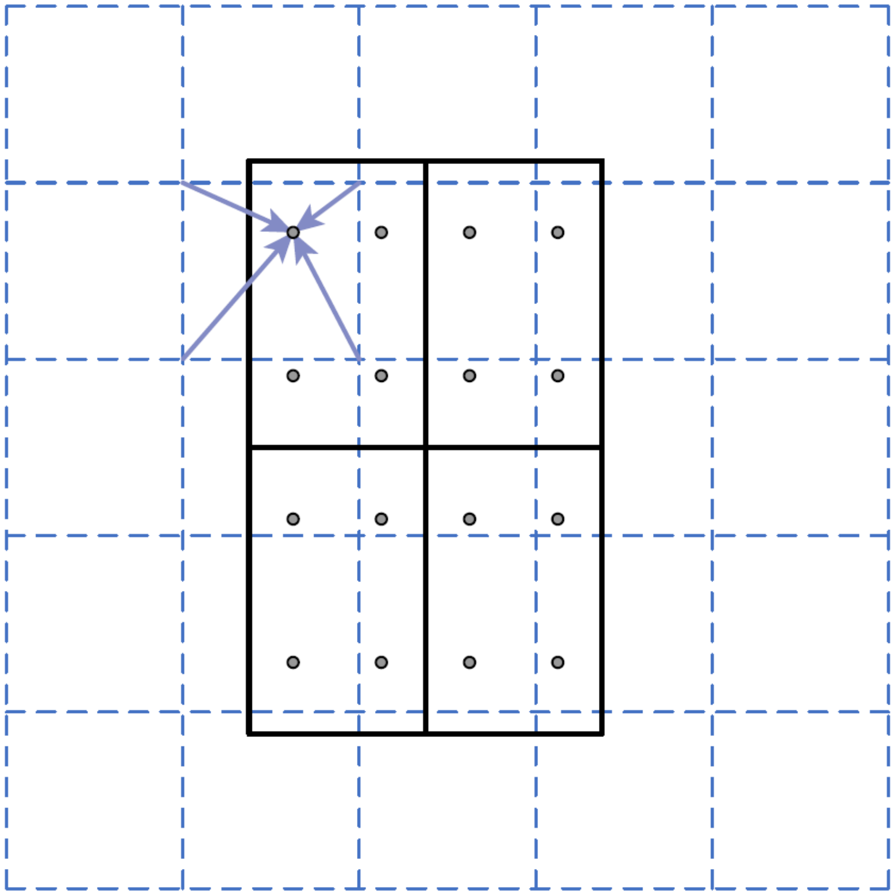
$$(x, y, h, w)$$

RoI mapping to feature map

$$s \times (x, y, h, w)$$

$$s = \frac{1}{16}$$

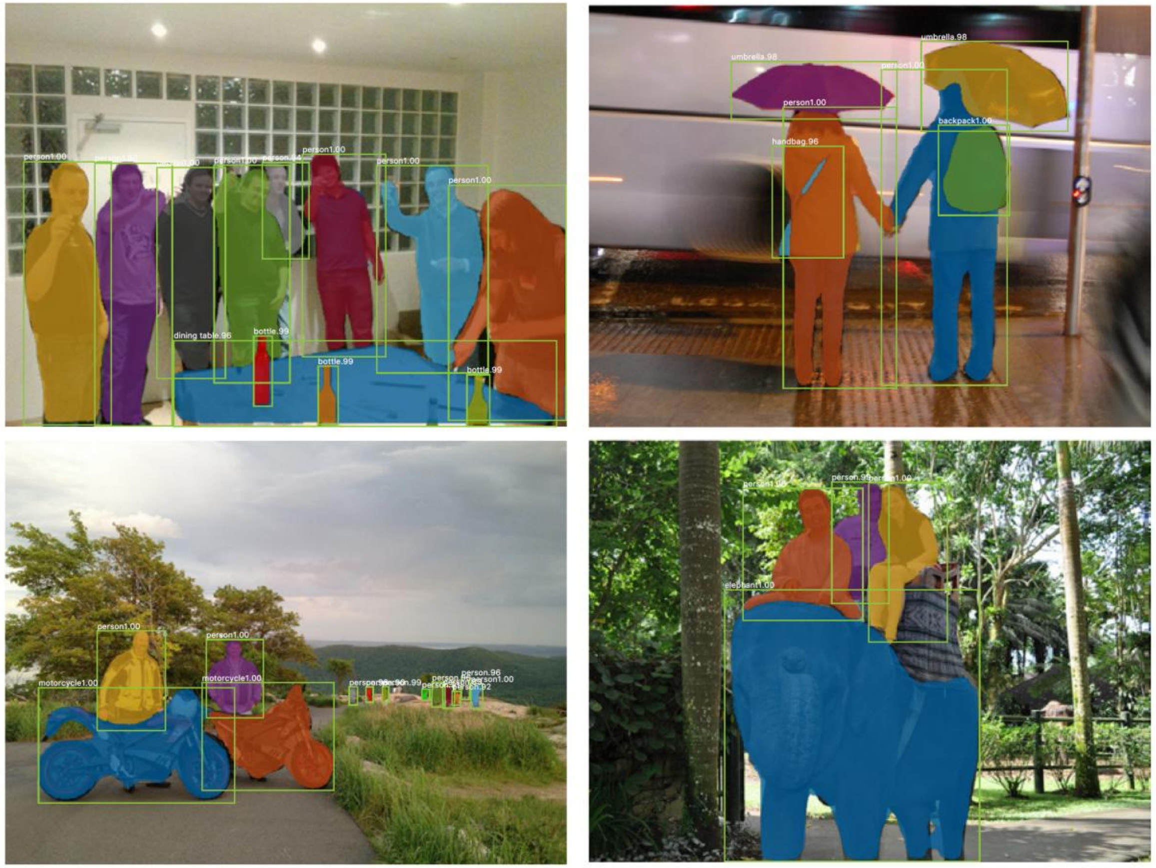
RoI Pooling



RoI Align

Mask R-CNN

	align?	bilinear?	agg.	AP	AP_{50}	AP_{75}
RoIPool [12]			max	26.9	48.8	26.4
RoIWarp [10]		✓	max	27.2	49.2	27.1
		✓	ave	27.1	48.9	27.1
RoIAlign	✓	✓	max	**30.2**	**51.0**	**31.8**
	✓	✓	ave	**30.3**	**51.2**	**31.5**



Mask R-CNN. He et al., ICCV, 2017

Summary

- Semantic segmentation
  - Label pixels into object classes
  - Traditional methods: conditional random fields
  - Deep learning methods: deconvolution, atrous convolution

- Instance segmentation
  - Separate object instances in the same class
  - Detection + segmentation inside each box

Further Reading

- Fully-connect CRFs, 2011 https://arxiv.org/abs/1210.5644
- DeepLab, 2015 https://arxiv.org/abs/1606.00915
- FCN, 2015 https://arxiv.org/abs/1411.4038
- Unet, 2015 https://arxiv.org/abs/1505.04597
- Mask R-CNN, 2017 https://arxiv.org/abs/1703.06870

Final Exam

- Epipolar Geometry
- Convolutional Nueral Networks
- Vision Transformers
- Object Detection