Computer Vision

12. Selected Topics

I-Chen Lin

College of Computer Science,
National Yang Ming Chiao Tung University

Outline

- Learning-based Vision and convolutional neural networks (CNN)

- Modern image segmentation

Many slides about CNN are modified from:

- F.-F. Li, A. Karpathy, J. Johnson, S. Yeung, Convolutional Neural Network for Visual Recognition, course lecture notes.

Oriented cells in the visual cortex

[Hubel and Wiesel 59]

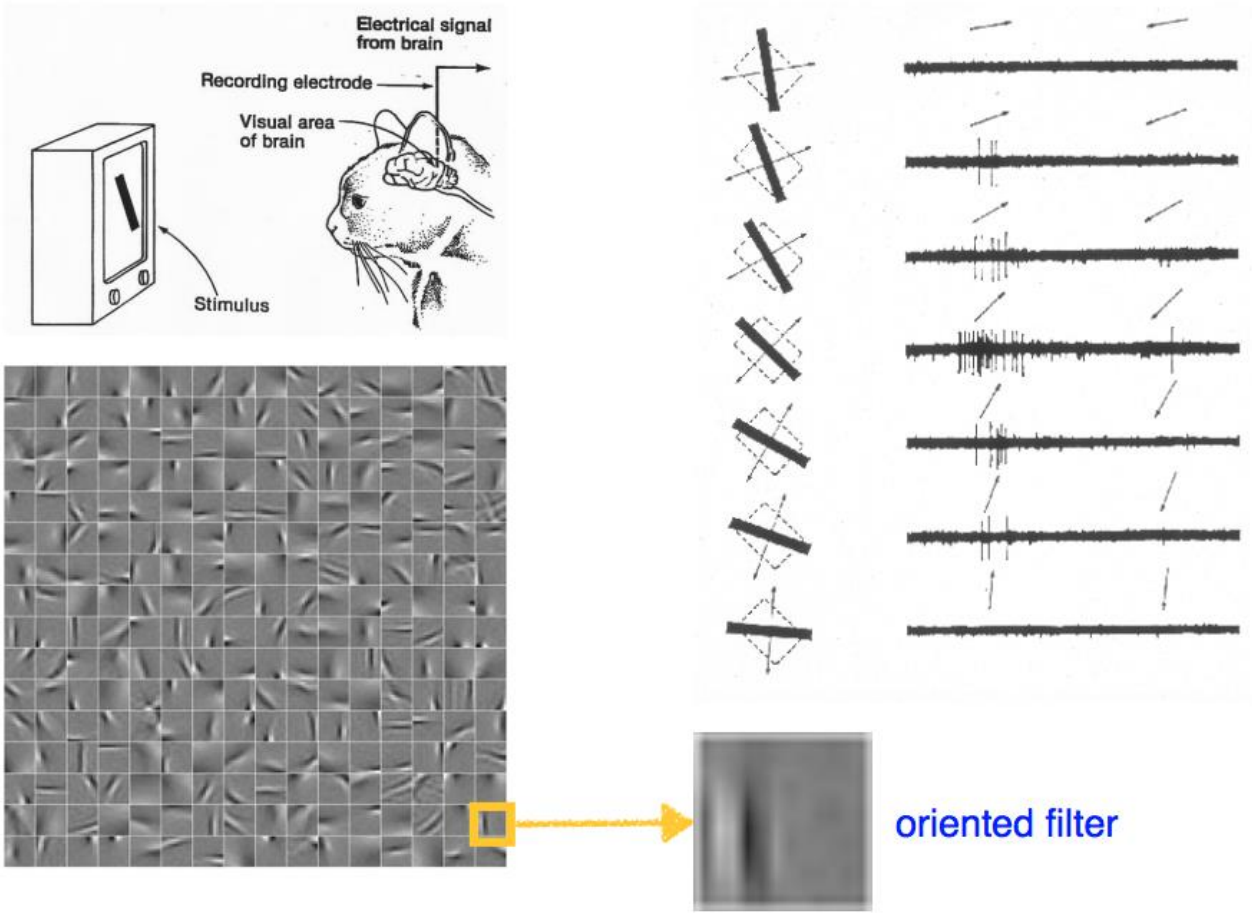


Fig. from: A. Vedaldi, Advanced Convolutional Neural Networks lecture notes.

Hierarchy models for recognition

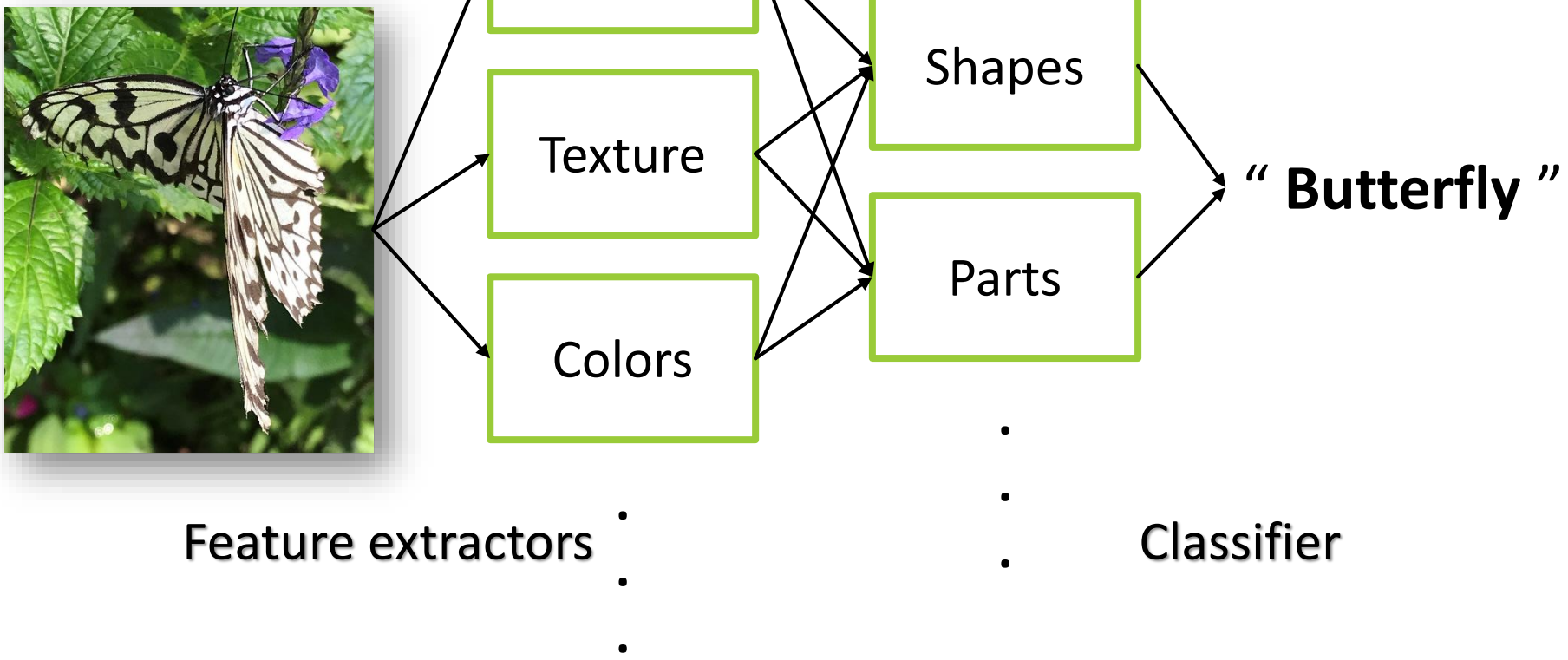
Fig. from : neurdiness.files.wordpress.com/2018/05/fncom-08-00135-g001.jpg

Take object recognition as an example

Brain /
Machine

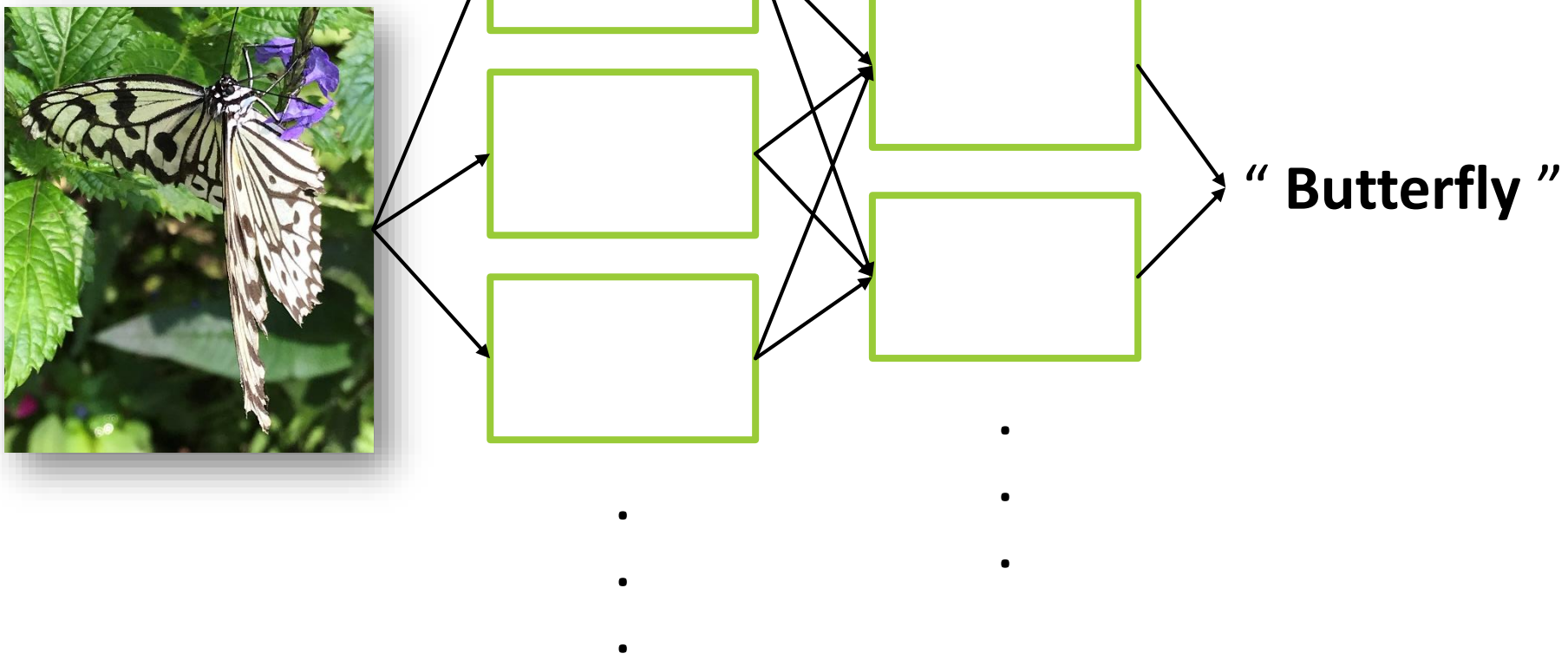
“ **Butterfly** ”

Object recognition



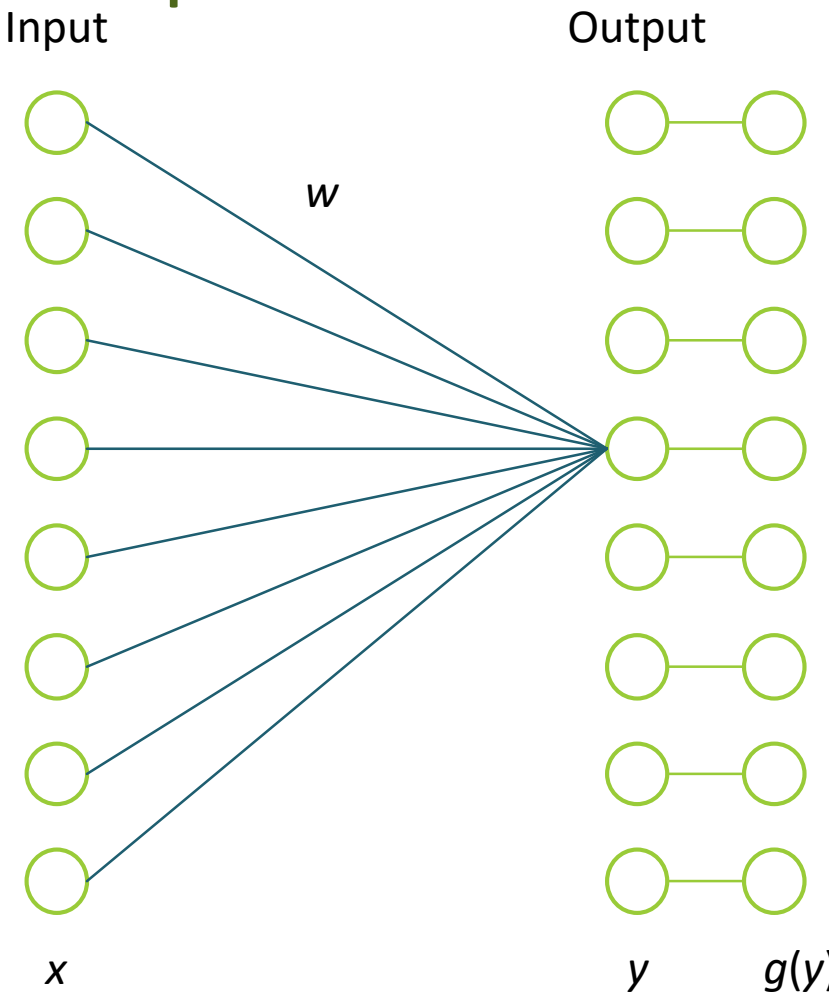
In conventional computer vision work, designers have to assign the roles and operations of a large portion of these blocks.

Artificial neural network

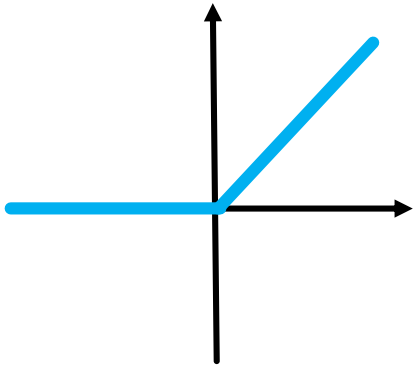


Motivated by Hubel and Wiesel, the roles and operations of these blocks are learnt during training.

Computation in a neural net (layer)



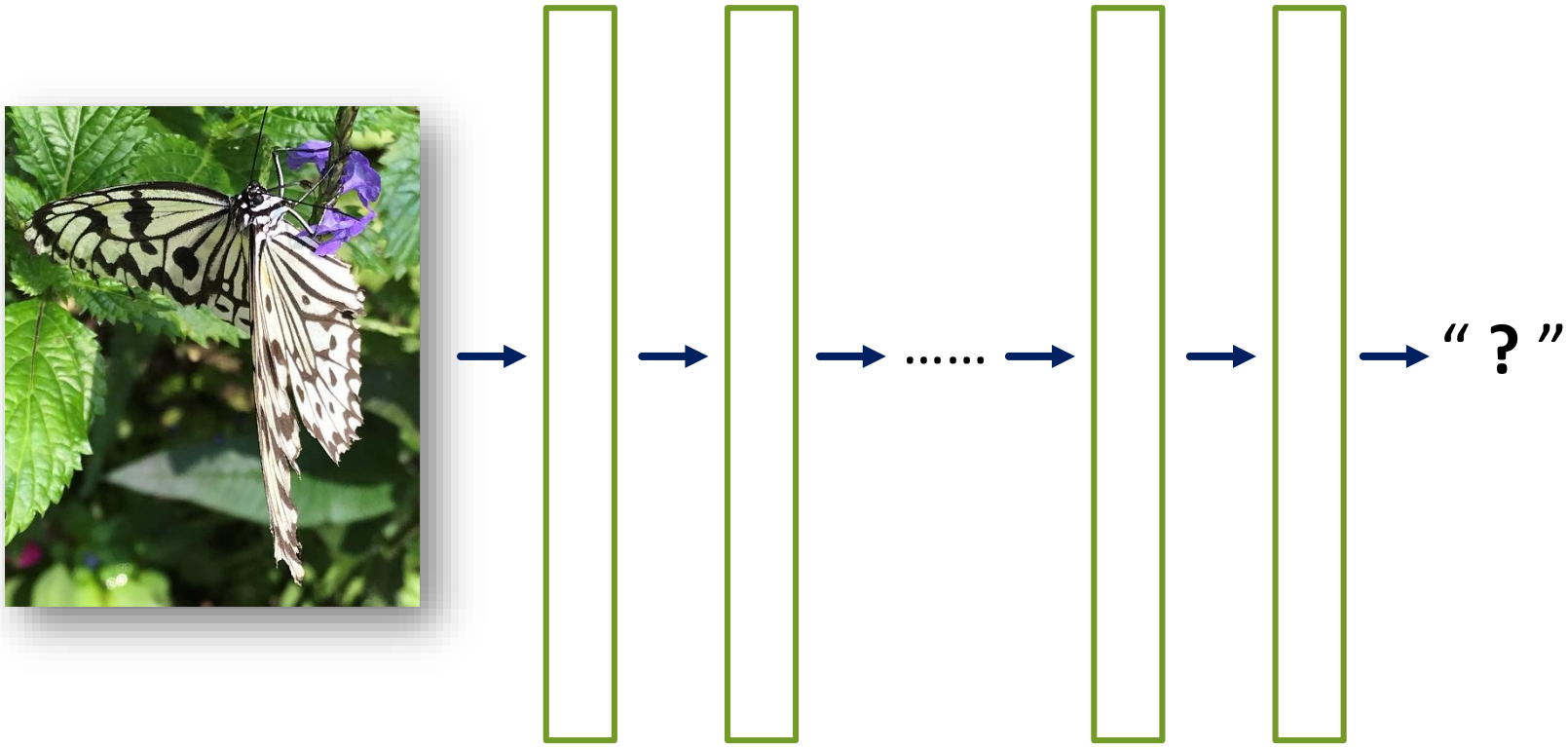
E.g. ReLU (Rectified Linear Unit)



$$g(y) = \max(0, y)$$

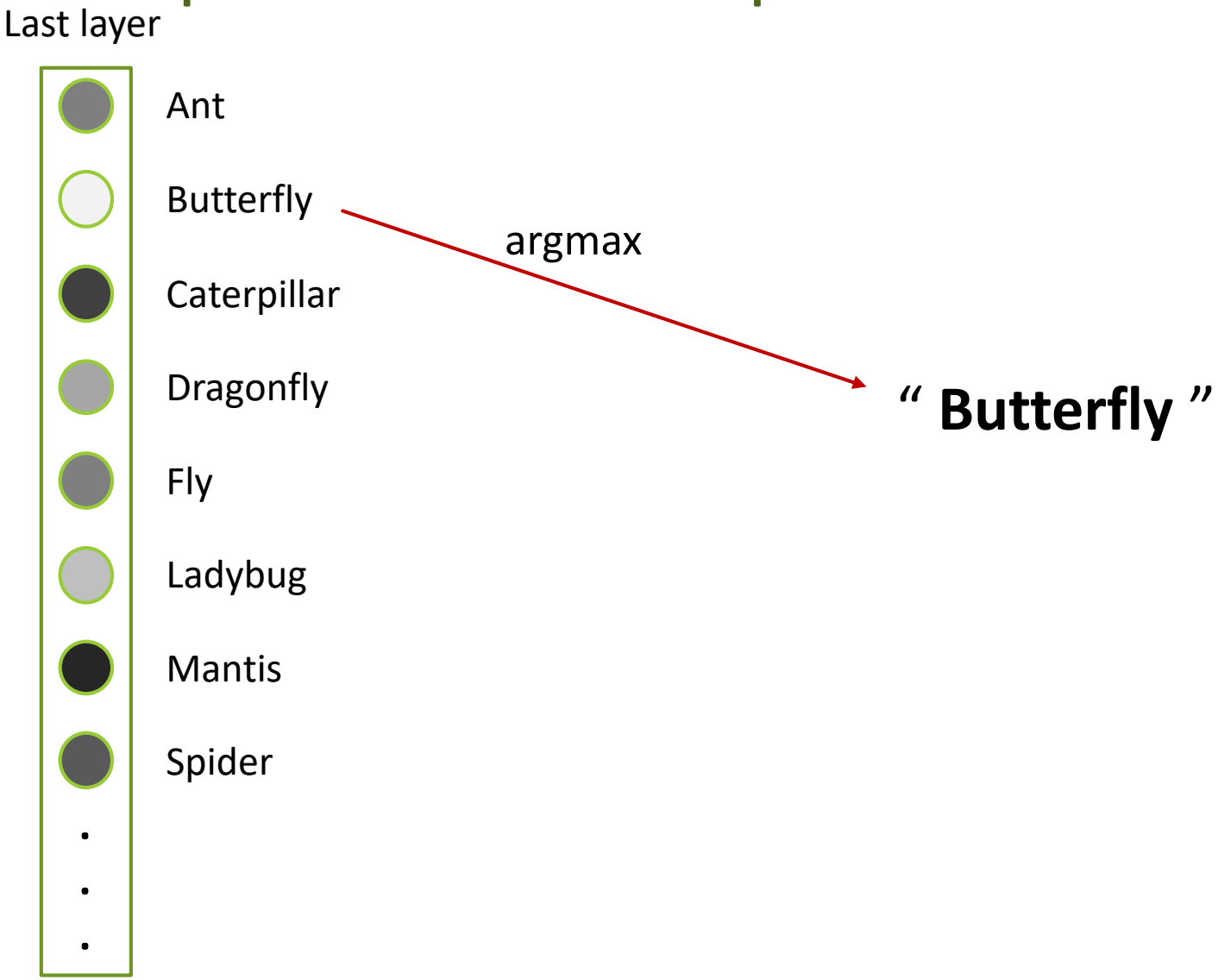
$$\hat{y} = g(y) = g(Wx + b)$$

Computation in a deep neural net



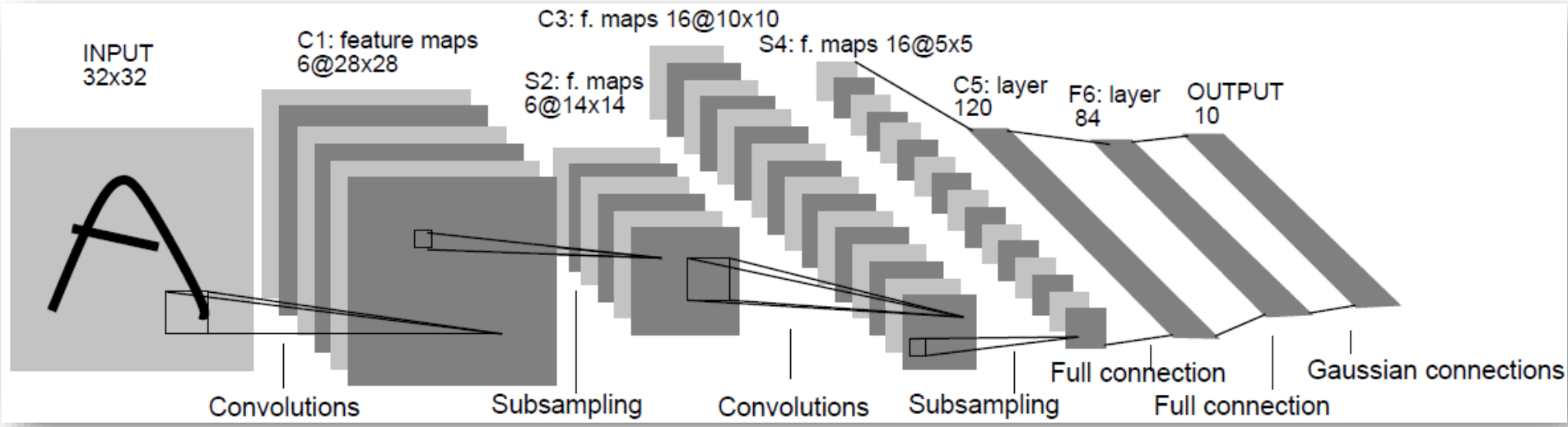
$$\ldots\ldots\ g(W_3 \cdot g(W_2 \cdot g(W_1 \cdot X_1 + b_1) + b_2) + b_3)$$

Computation in a deep neural net



LeNet-5

- Local filtering (local receptive fields)
- Share weights for local filters
- Subsampling with pooling

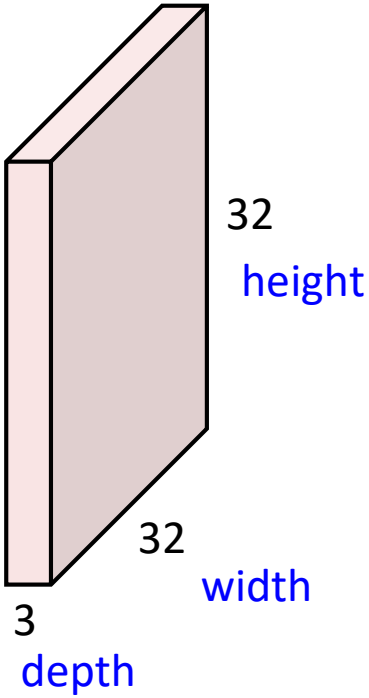


Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," Proc. IEEE, 1998.

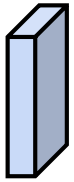
Convolution layer

Filters always extend the full depth of the input volume

32x32x3 image

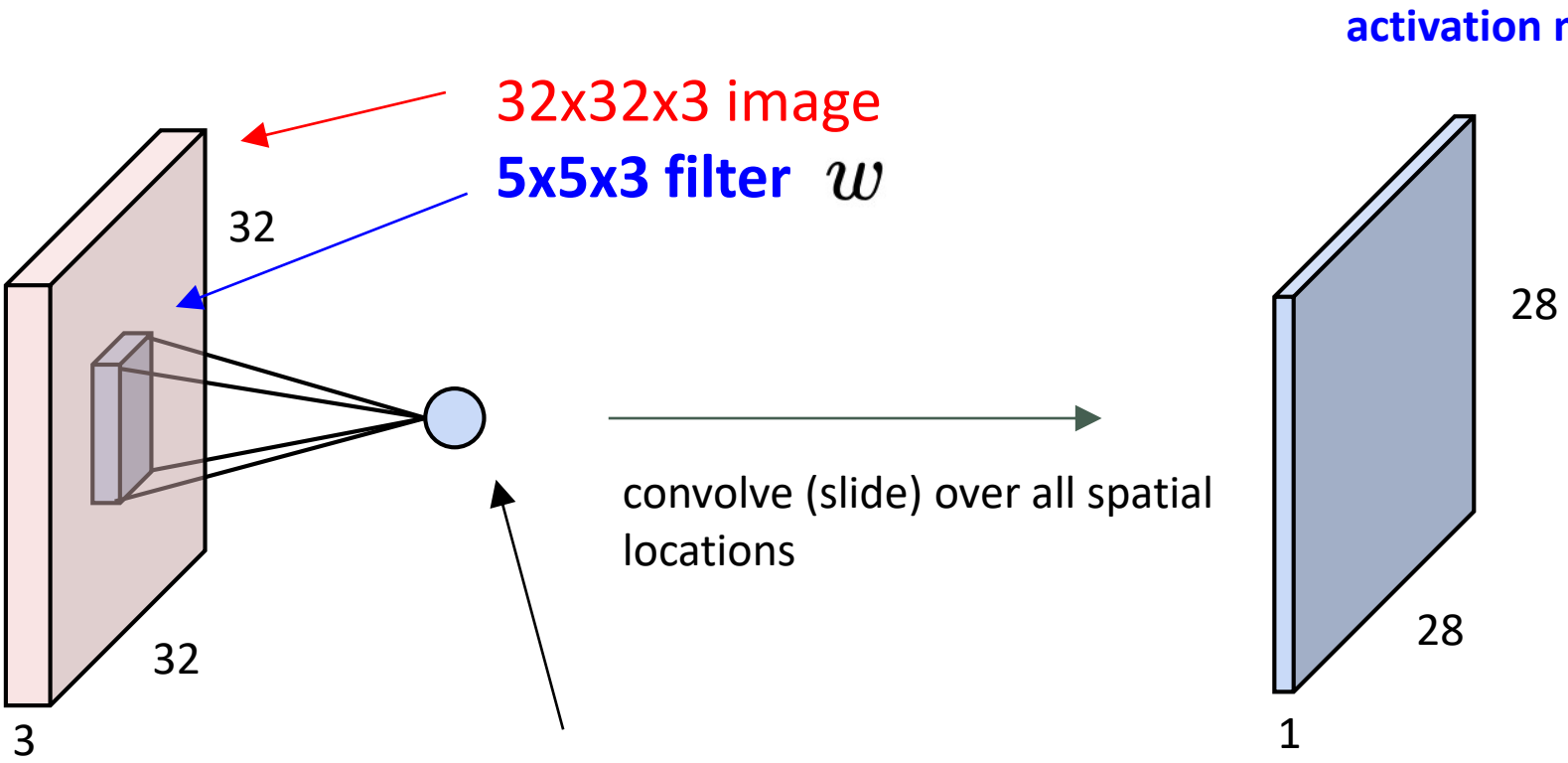


5x5x3 filter



Slides are modified from F.-F. Li, A. Karpathy, J. Johnson, S. Yeung, Convolutional Neural Network for Visual Recognition, course lecture notes.

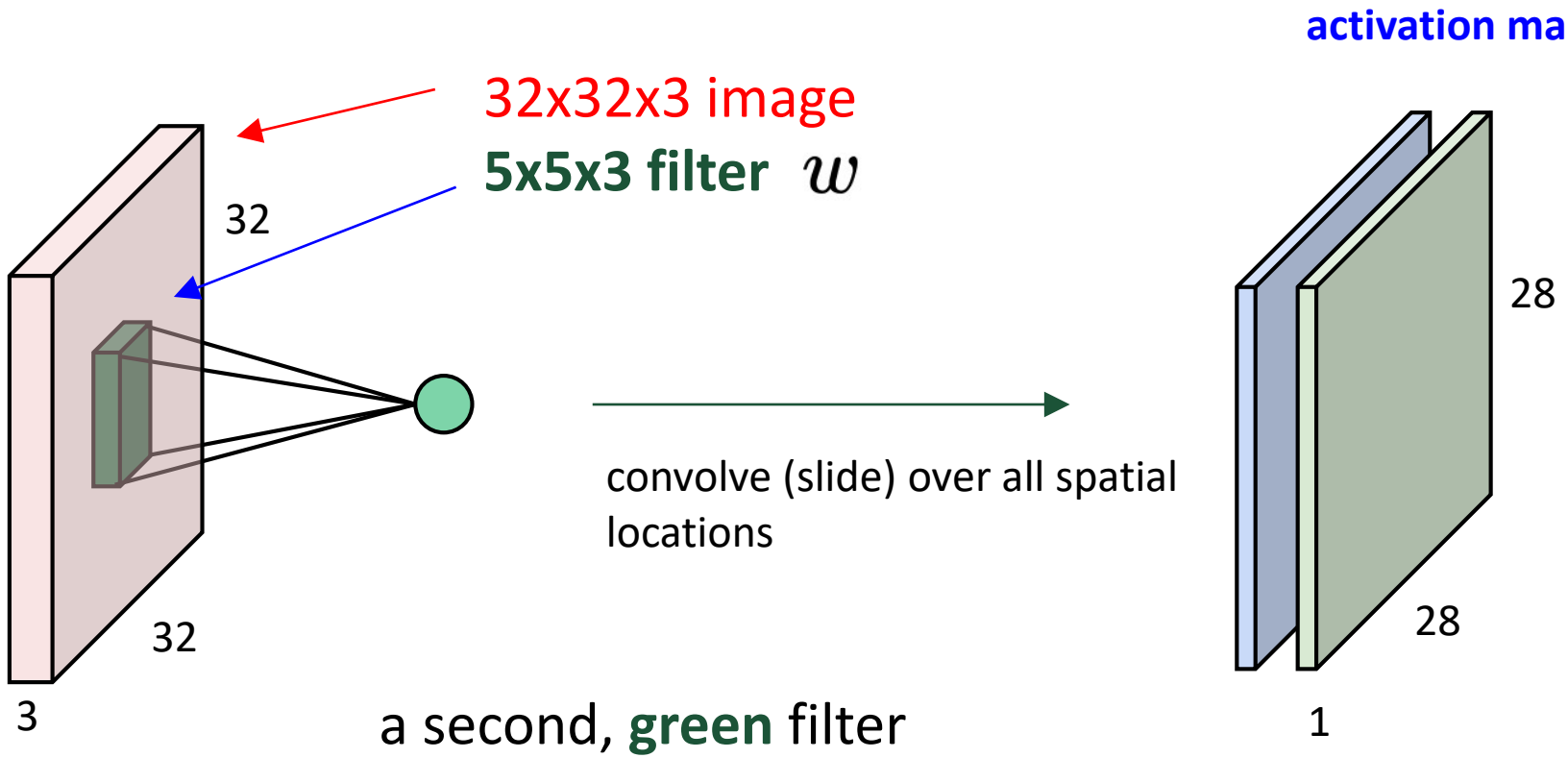
Convolution layer



$$w^T x + b$$

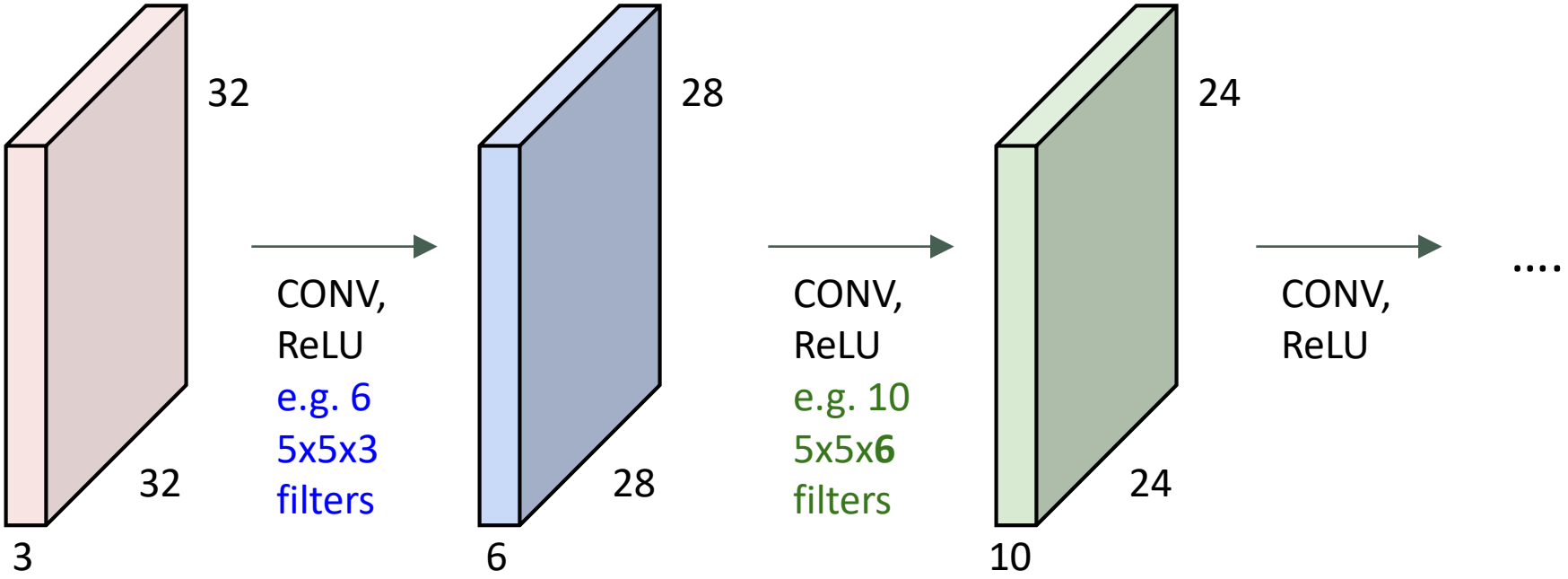
1 number:
the result of taking a dot product between the filter and a small 5x5x3 chunk of the image
(i.e. 5*5*3 = 75-dimensional dot product + bias)

Convolution layer



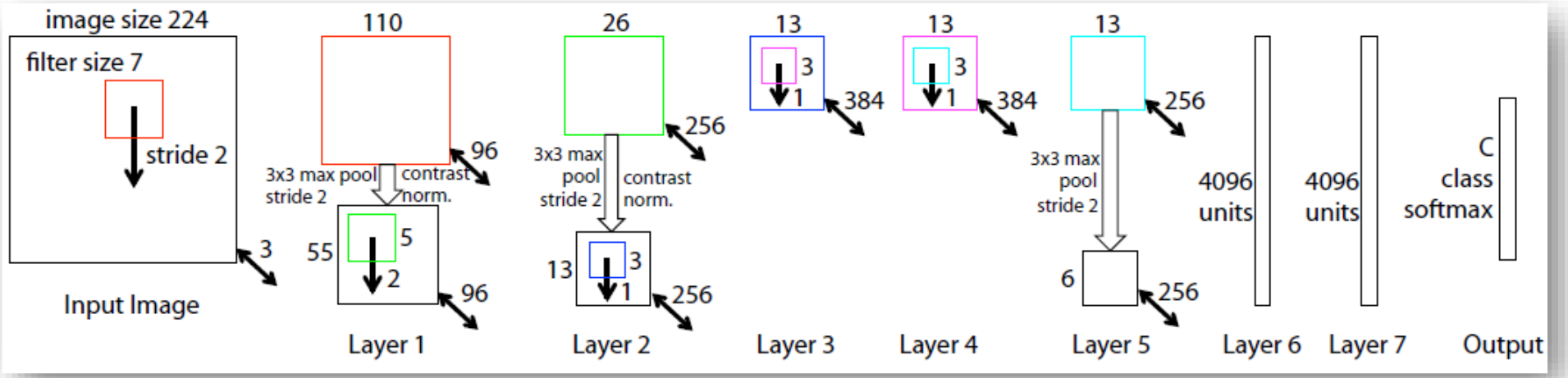
Convolutional neural network

- ConvNet is a sequence of convolutional layers, interspersed with activation functions.



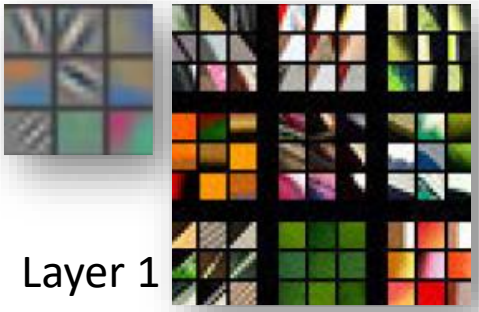
Visualization of ConvNet

- M.D. Zeiler and R. Fergus, “Visualizing and Understanding Convolutional Networks”, ECCV’14.
  - Maps feature activity in intermediate layers back to the input pixel space by deconvolution layers.

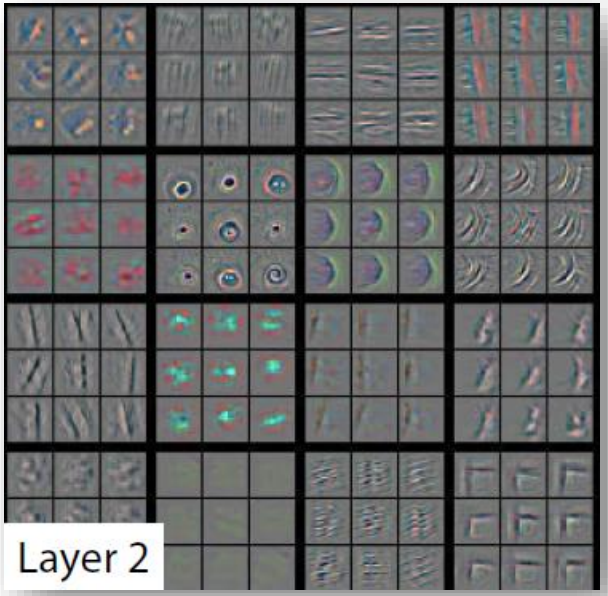


A 8-layer architecture for object recognition: 1~5 conv. Layer; 6,7 fc layers

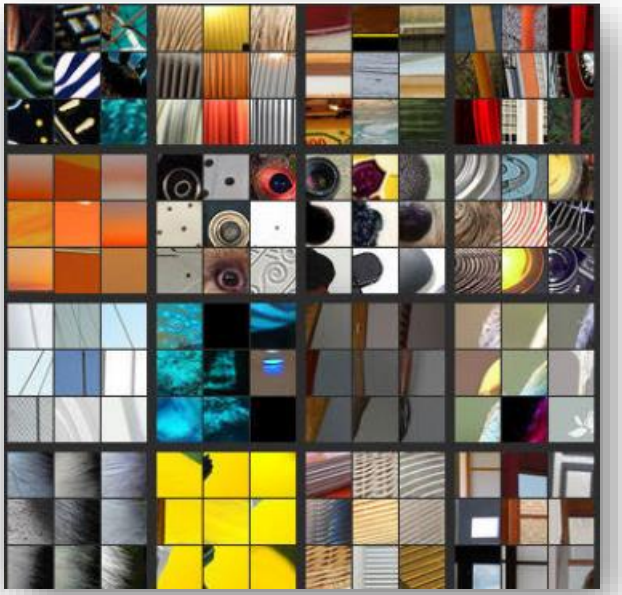
Top activations and their projections back to pixel space

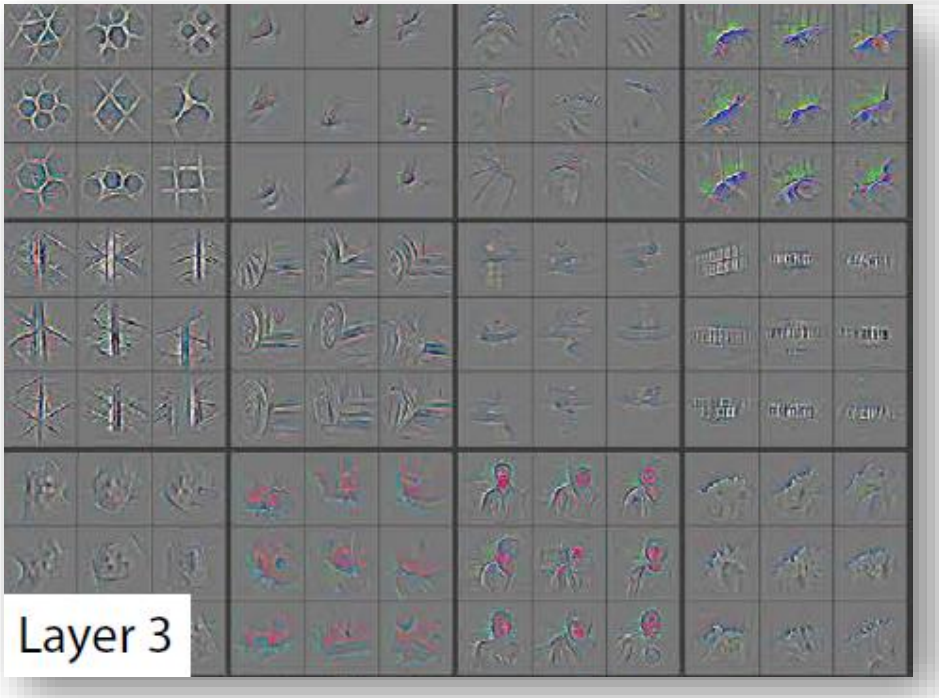


Layer 1



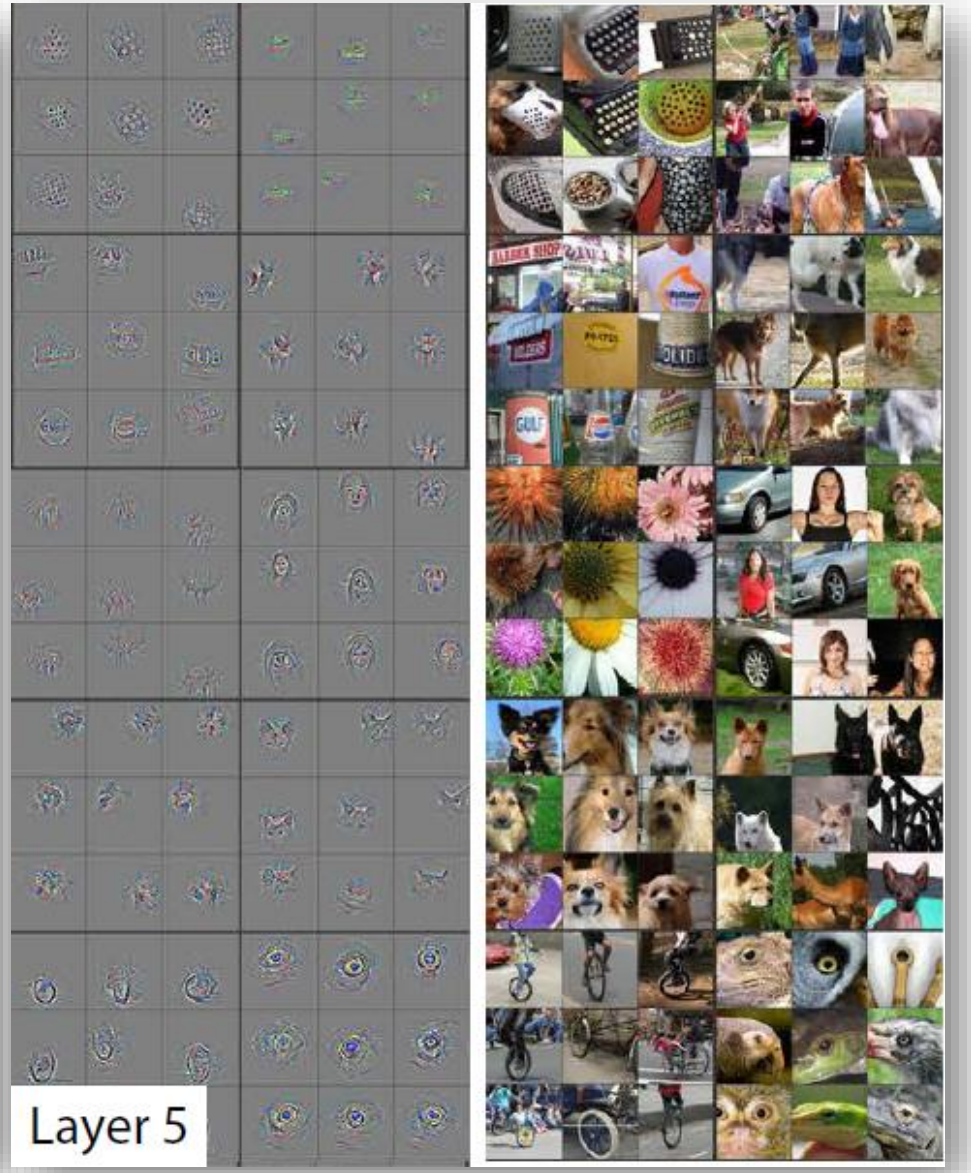
Layer 2



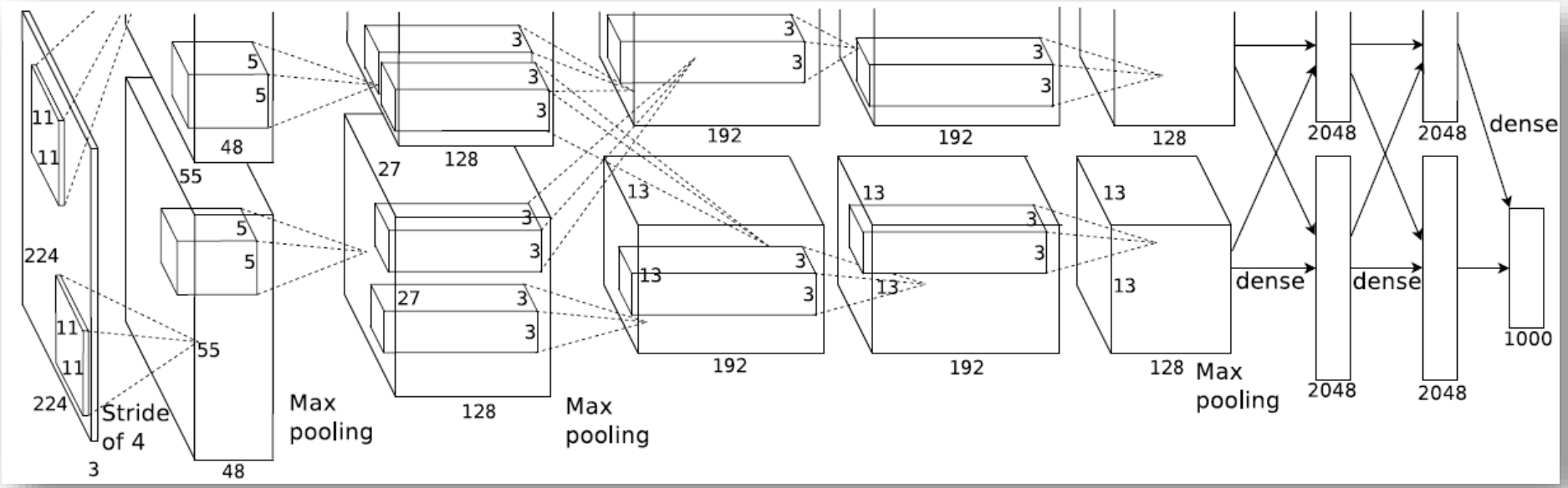


Layer 3

Top activations and their projections back to pixel space



AlexNet



- ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2012 winner
- Input : 224x224x3 (padded to 227)
- 1st layer: (227-11)/4 + 1 = 55
- Due to video-RAM issue, the network spread across 2 GPUs.
  - 1st layer: [55x55x48] x 2

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks, NIPS'12.

VGGNet

- K. Simonyan and A. Zisserman, Very Deep Convolutional Networks for Large-scale Image Recognition, Proc. ICLR'15.
- ILSVRC 2014, 1st and 2nd in localization and classification, respectively.
- No Local Response Normalization (LRN) is applied.

Note: All hidden layers are equipped with the ReLU

VGG16
input
conv3-64
conv3-64
maxpool
conv3-128
conv3-128
maxpool
conv3-256
conv3-256
conv3-256
maxpool
conv3-512
conv3-512
conv3-512
maxpool
conv3-512
conv3-512
conv3-512
maxpool
FC-4096
FC-4096
FC-1000
soft-max

VGG19
input
conv3-64
conv3-64
maxpool
conv3-128
conv3-128
maxpool
conv3-256
conv3-256
conv3-256
conv3-256
maxpool
conv3-512
conv3-512
conv3-512
conv3-512
maxpool
conv3-512
conv3-512
conv3-512
conv3-512
maxpool
FC-4096
FC-4096
FC-1000
soft-max

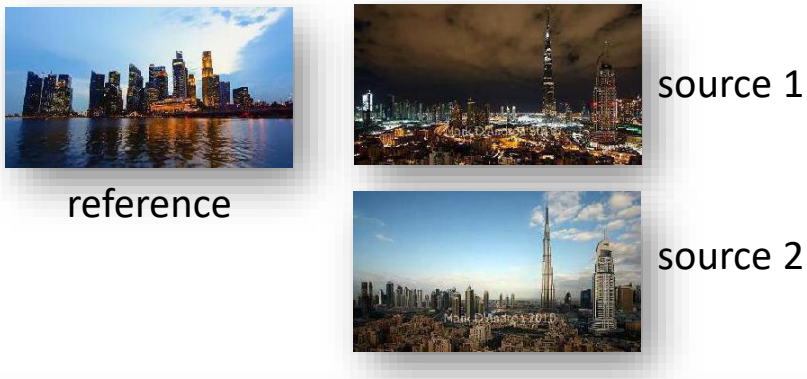
VGGNet (cont.)

- Using multiple small conv to replace a larger one.
  - Effective receptive field: three 3x3 conv (stride 1) layers has same as one 7x7 conv layer.
  - Fewer parameters 3*[(3x3xc)xc] vs. 1*[(7x7xc)xc]
- VGG16/19 are popularly used as an additional loss network for other work.
  - E.g. conv/relu 1-1, 1-2, 3-3… for perceptual loss.
  - FC7 features generalize well to other tasks.

VGG16	alt layer names
input	
conv3-64	conv1-1
conv3-64	conv1-2
maxpool	
conv3-128	conv2-1
conv3-128	conv2-2
maxpool	
conv3-256	conv3-1
conv3-256	conv3-2
conv3-256	conv3-3
maxpool	
conv3-512	conv4-1
conv3-512	conv4-2
conv3-512	conv4-3
maxpool	
conv3-512	conv5-1
conv3-512	conv5-2
conv3-512	conv5-3
maxpool	
FC-4096	fc6
FC-4096	fc7
FC-1000	fc8
soft-max	

Semantic matching using CNN features

- *ReluL*_1, *L* = 5, . . . , 1 in VGG19.



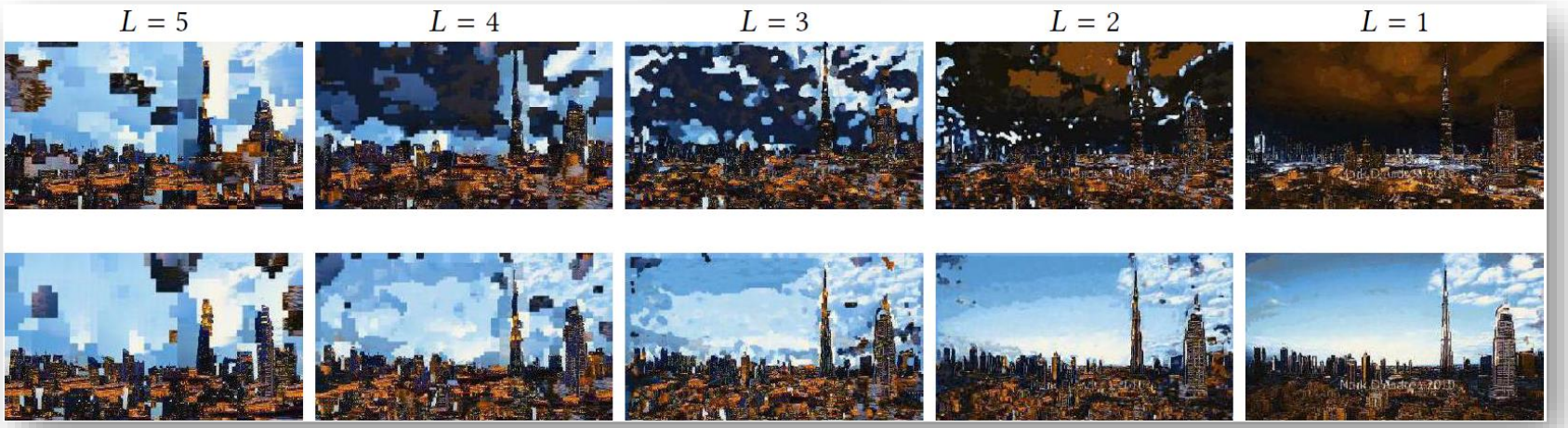


Fig. from He et al., “Progressive Color Transfer With Dense Semantic Correspondences”, ACM Trans. Graph. 2019.

Low-level vs. high level features

- With only low-level features, it is difficult to group regions of various appearance into a meaningful segment.

A castle segmented by mean shift segmentation

- How to apply the learned high-level features for image segmentation?

Computer Vision Tasks

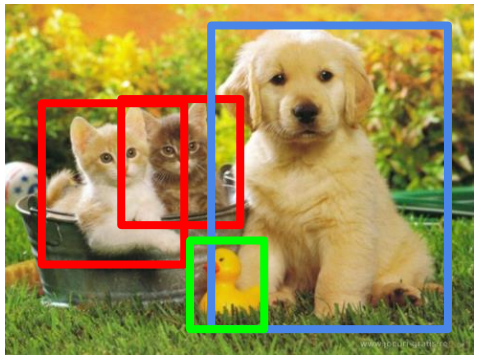
Classification

CAT

Classification + Localization

CAT

Object Detection



CAT, DOG, DUCK

Segmentation

CAT, DOG, DUCK

Single object

Multiple objects

Slides are modified from F.-F. Li, A. Karpathy, J. Johnson, Convolutional Neural Network for Visual Recognition, course lecture notes.

Semantic segmentation

- Label every pixel !
- Don’t differentiate instances (cows)

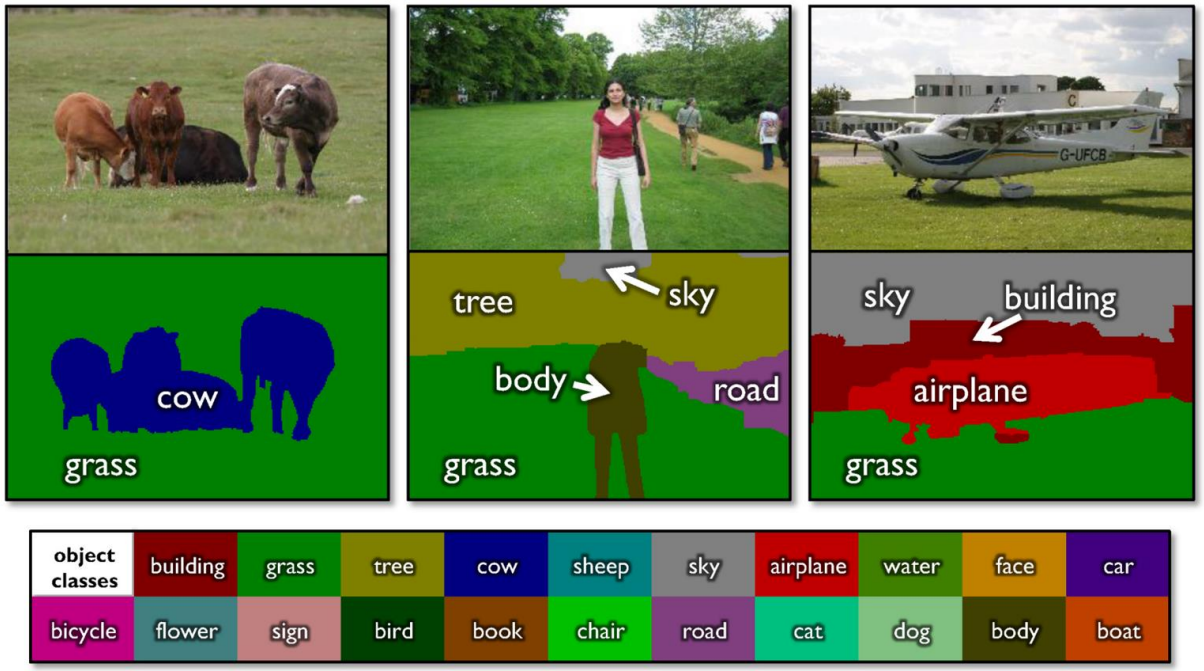


Fig. from Shotton et al, “TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context”, IJCV 2007.

Instance segmentation

- Detect instances, estimate the categories, label pixels.
- “Simultaneous detection and segmentation” !

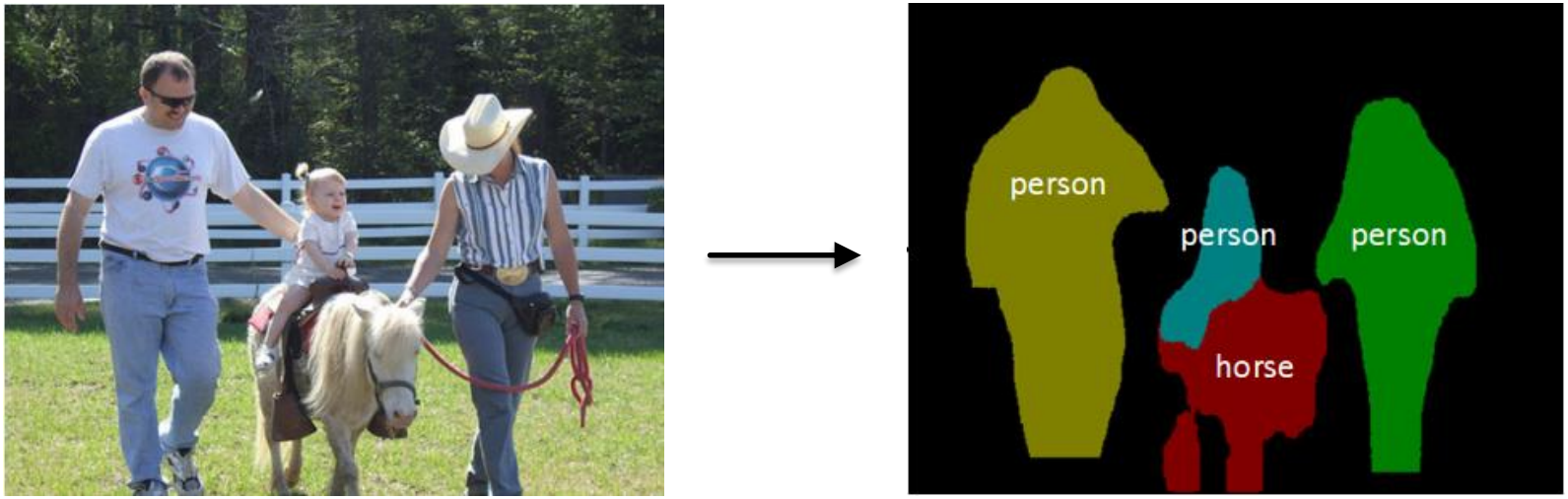
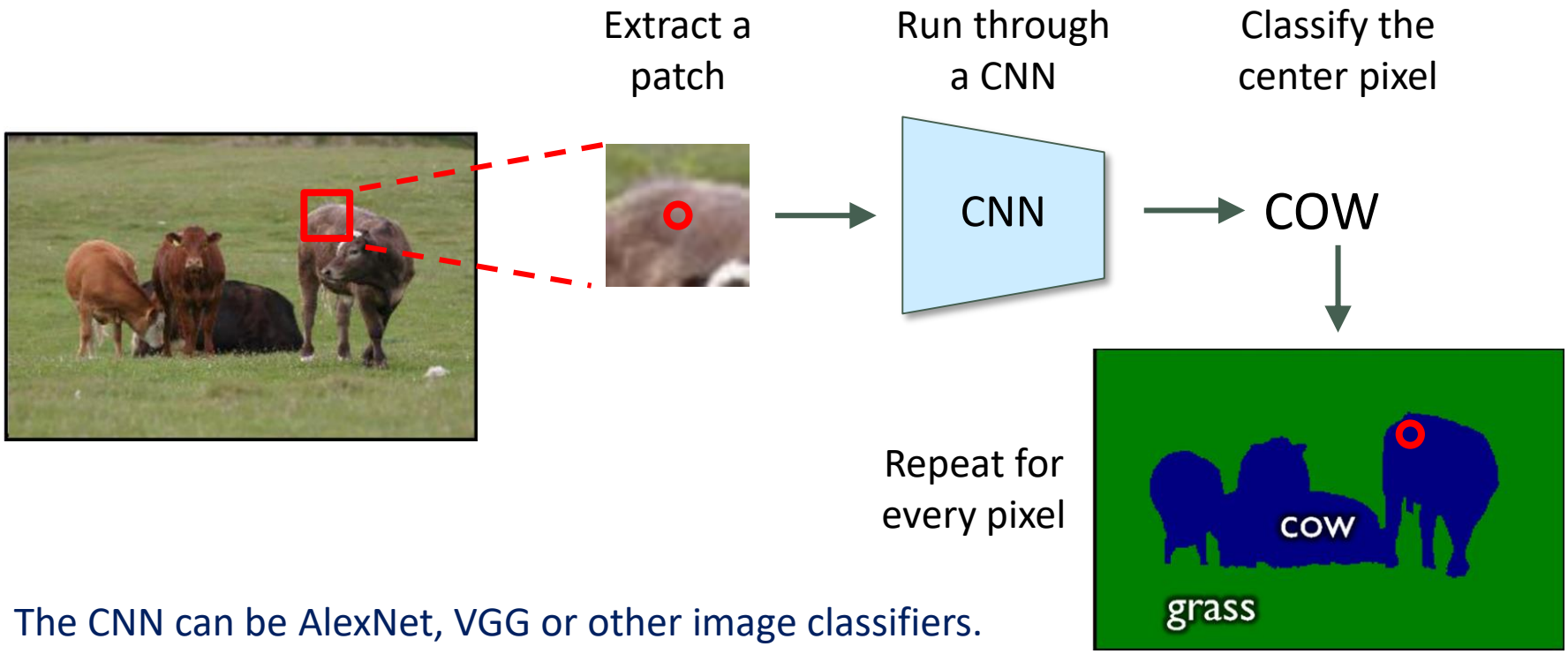


Fig. from Dai et al, “Instance-aware Semantic Segmentation via Multi-task Network Cascades”, Proc. CVPR’16.

Semantic segmentation: sliding windows

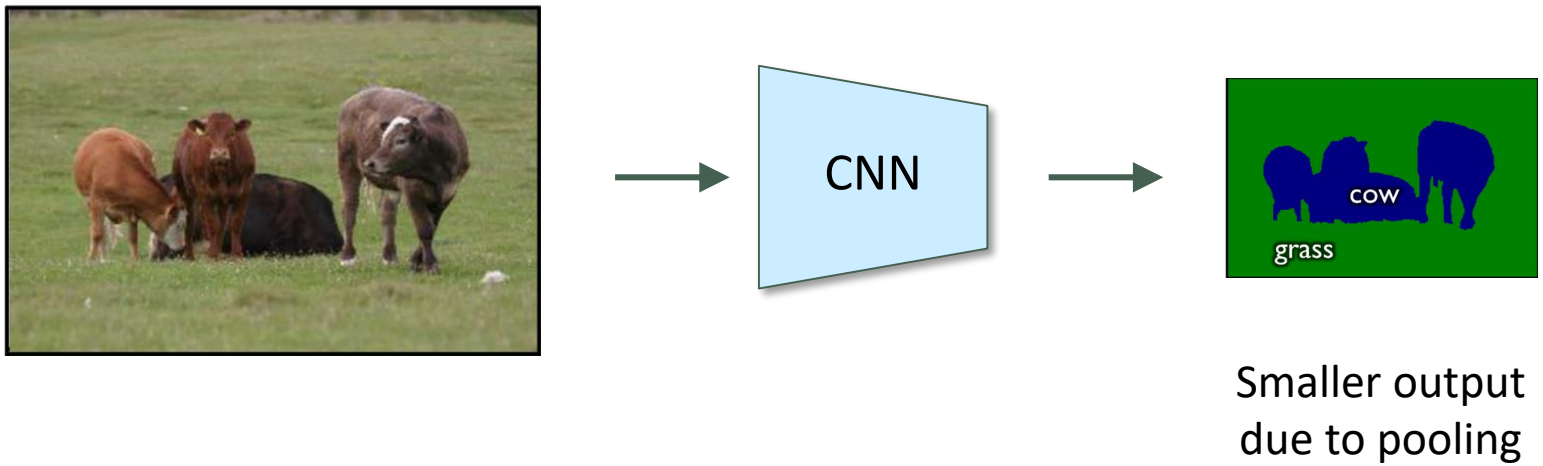


The CNN can be AlexNet, VGG or other image classifiers.

Problem: *Inefficient*!
Not reusing shared features between overlapping patches.

Semantic segmentation: fully convolutional

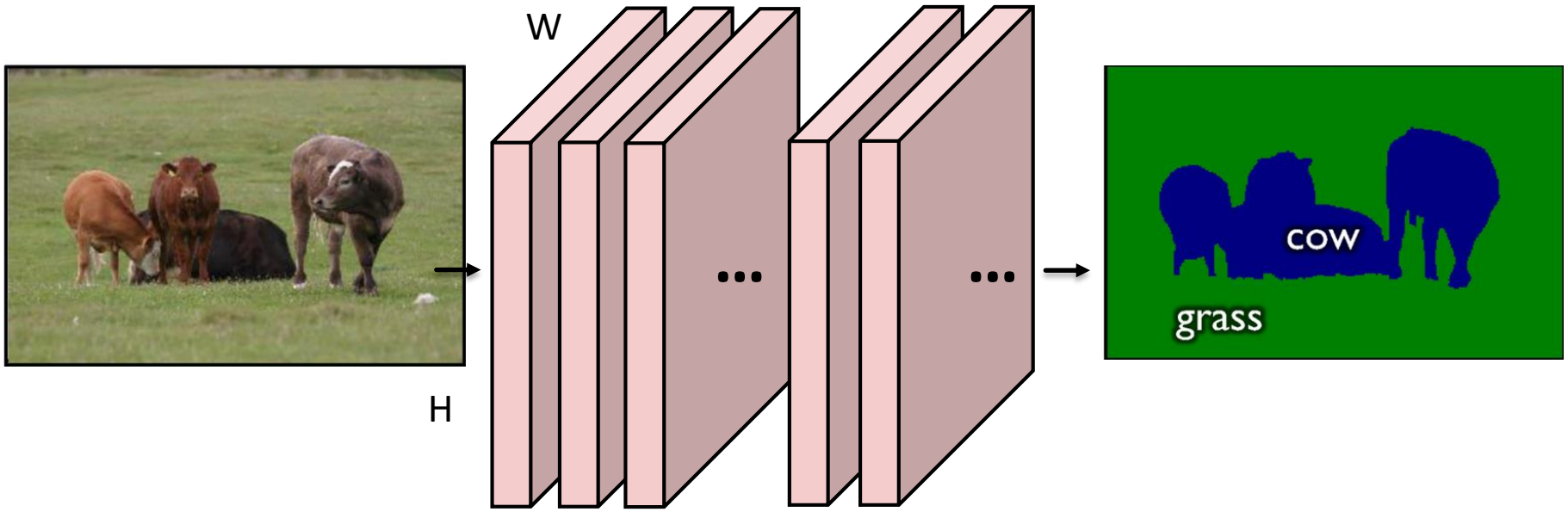
- Apply a fully convolutional network to get labels of all pixels at once.



Smaller output due to pooling

Semantic segmentation: fully convolutional

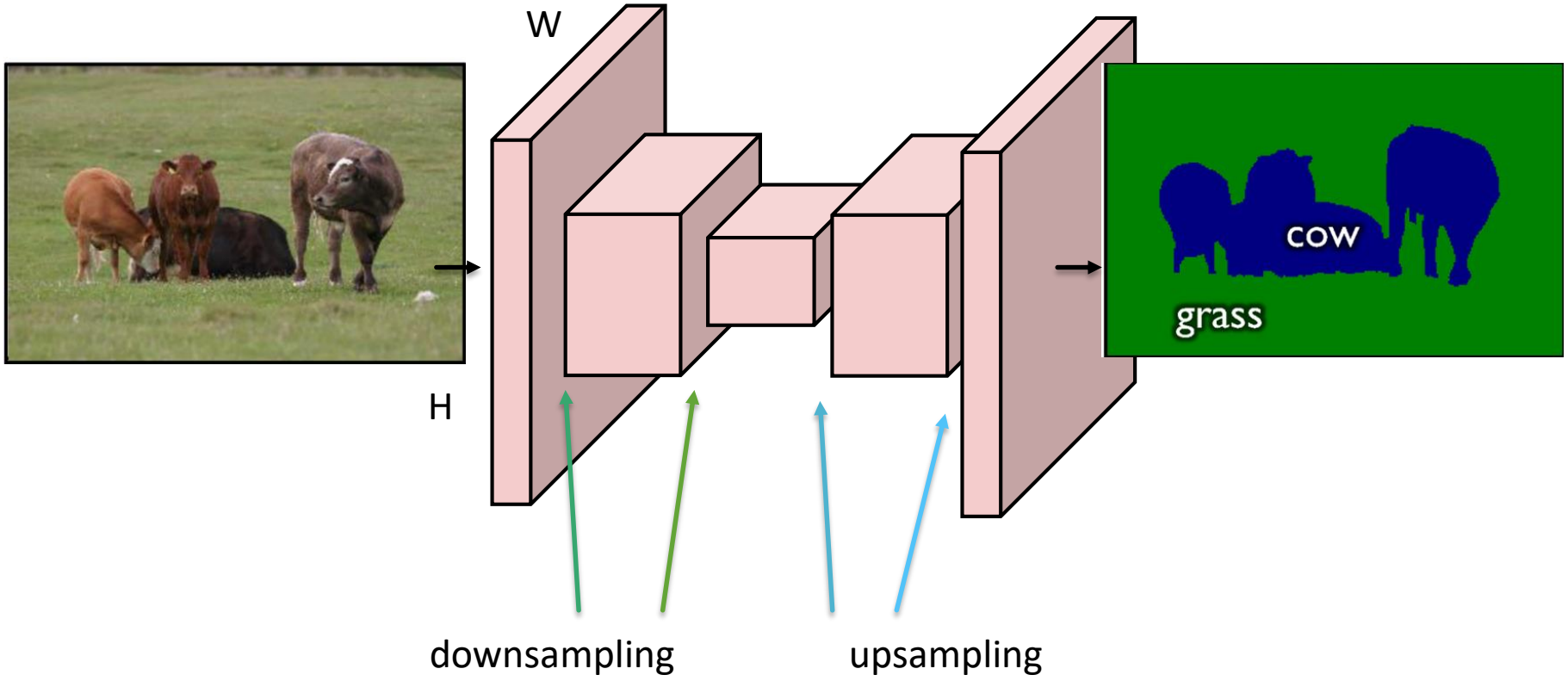
- If we keep the convolutions at original image resolution …..



Expensive and …..

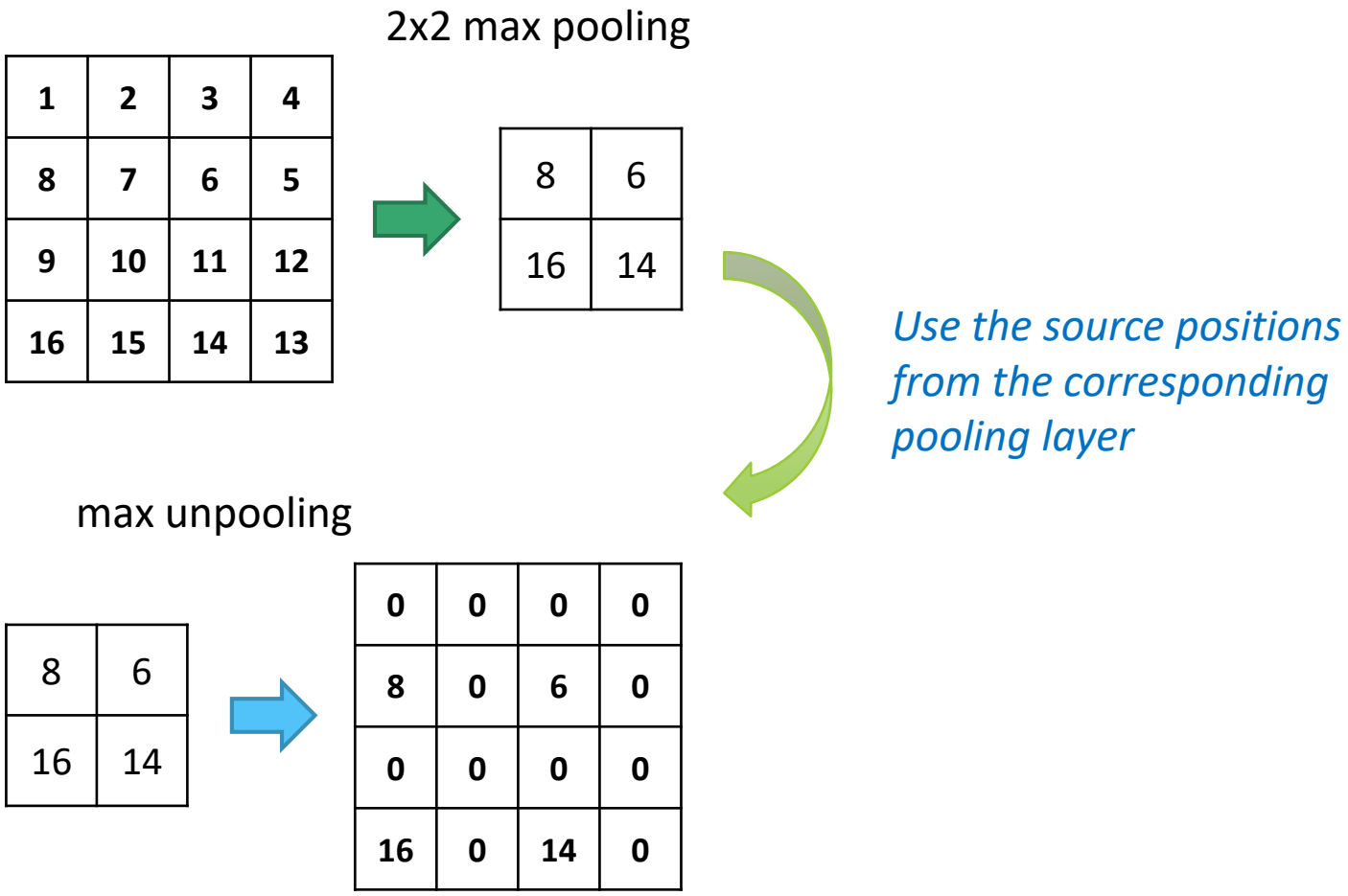
Semantic segmentation: fully convolutional

- The important features can be processed in a low resolution by applying downsampling and upsampling.

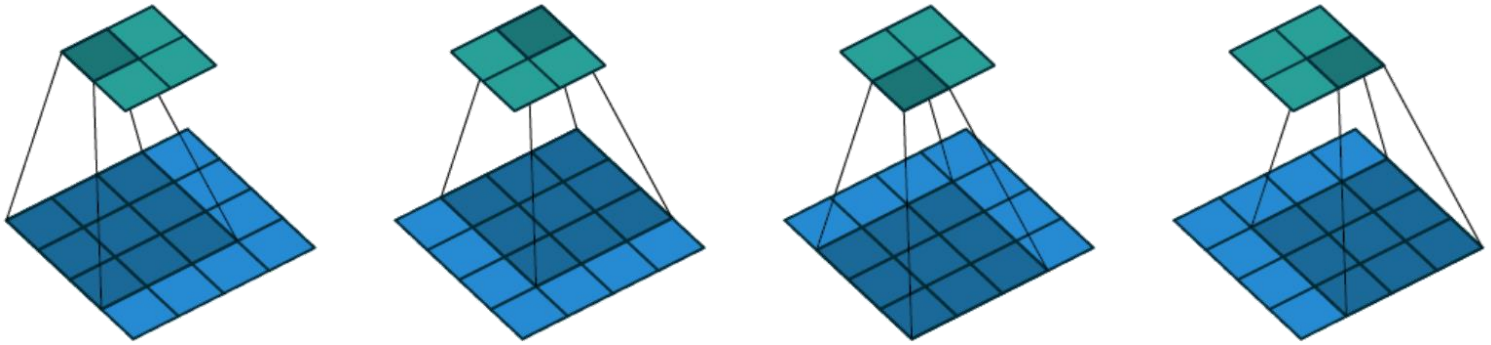


Down- and up-sampling a feature map

- Unpooling: e.g. Max unpooling

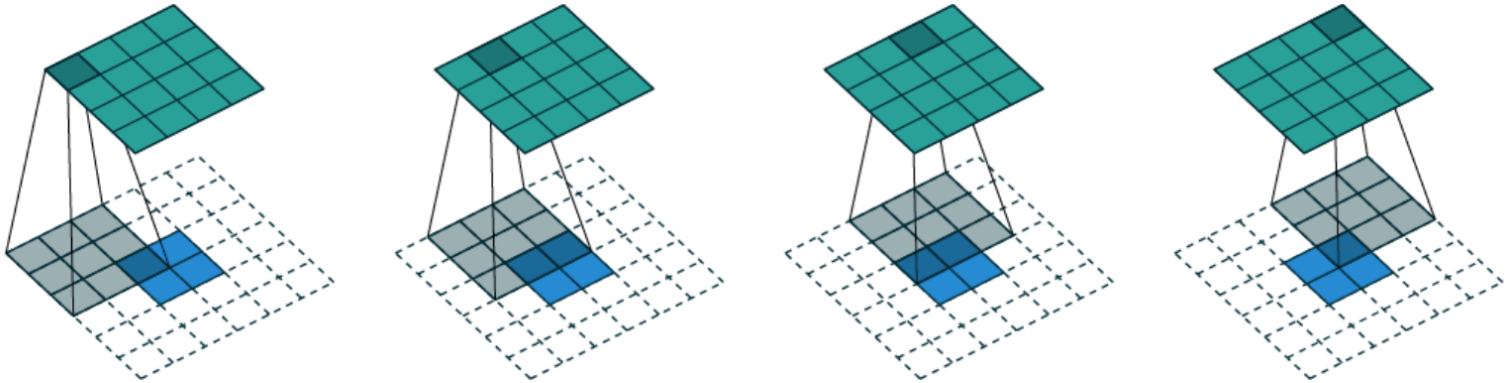


Convolution and transposed conv.



Convolving a 3 x 3 kernel over a 4 x 4 input using unit strides.

Learnable kernels

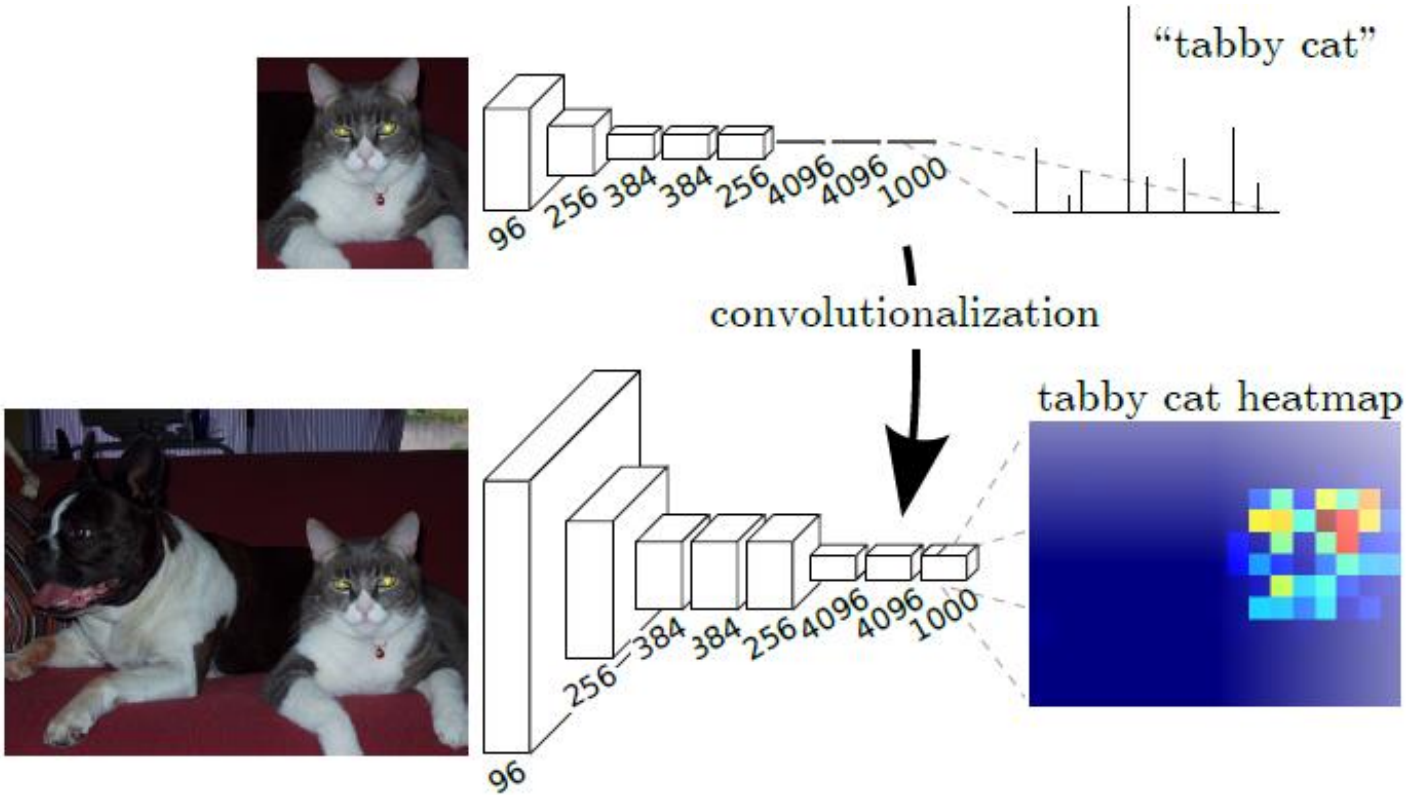


Its **transposed** operation is equivalent to convolving a 3x3 kernel over a 2x2 input padded with a 2x2 border of zeros using unit strides

Figures from V. Dumoulin and F. Visin, “A guide to convolution arithmetic for deep Learning”

Semantic segmentation: FCN

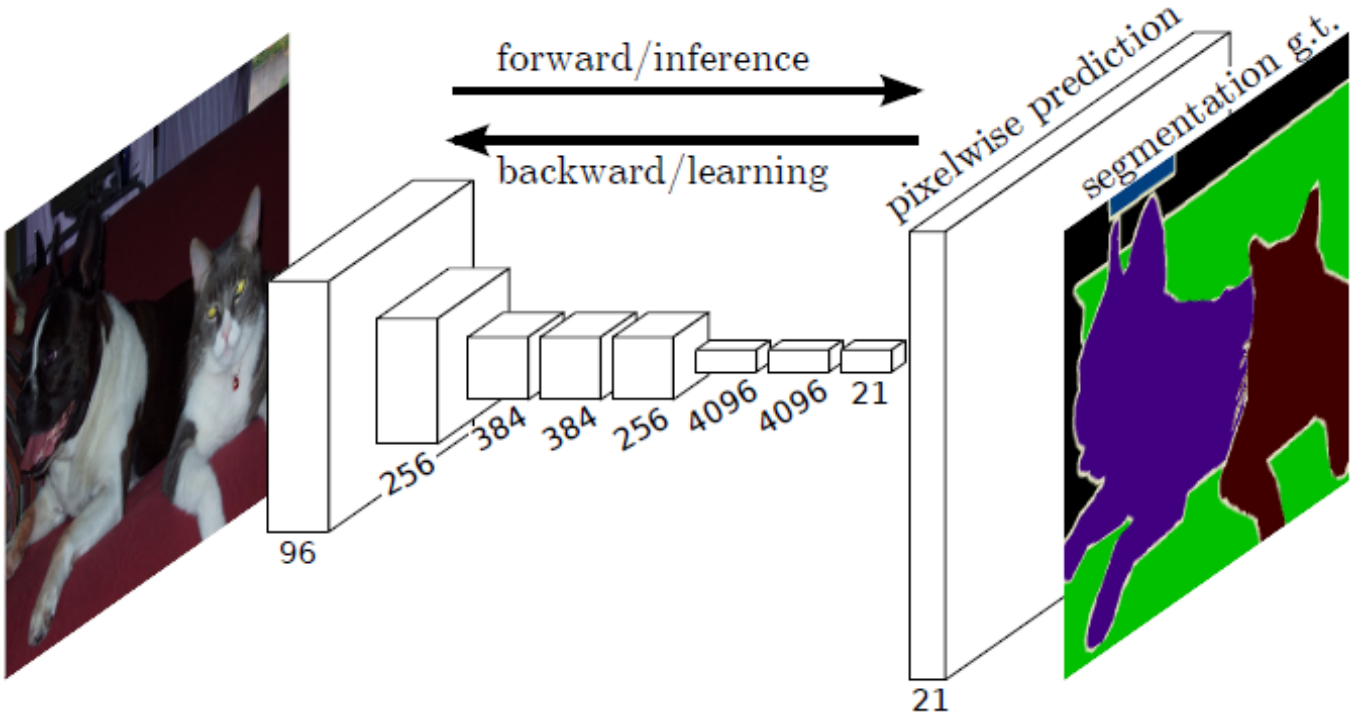
- Adapting classification networks into fully convolutional networks.



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

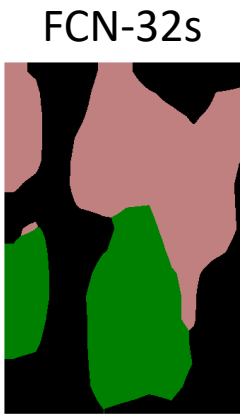
Semantic segmentation: FCN

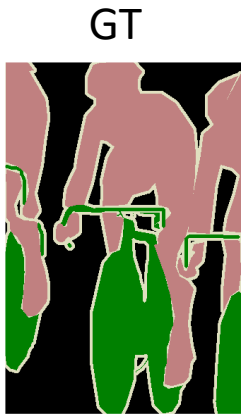
- Dense predictions for per-pixel tasks like semantic segmentation.

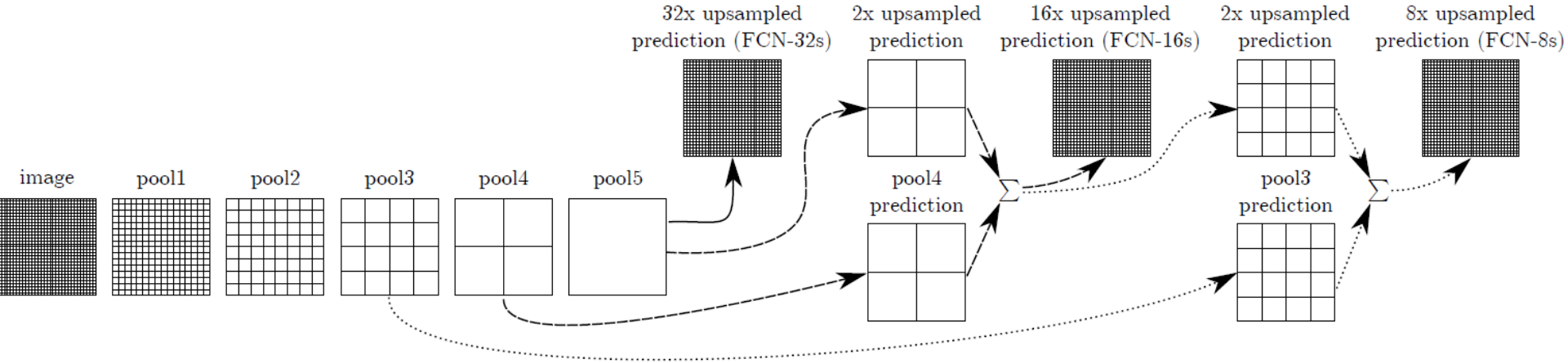


Semantic segmentation: FCN

- Incorporating features of higher-res. can improve the accuracy.

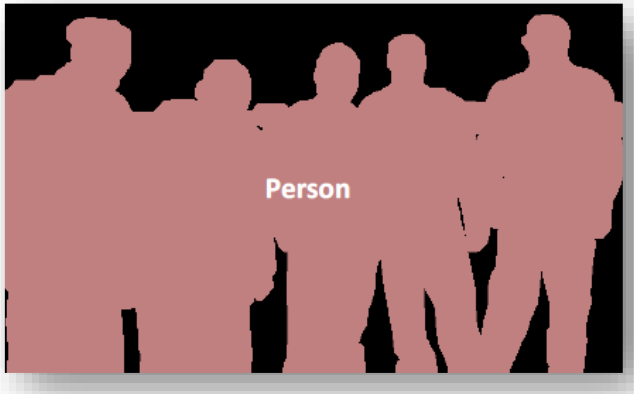




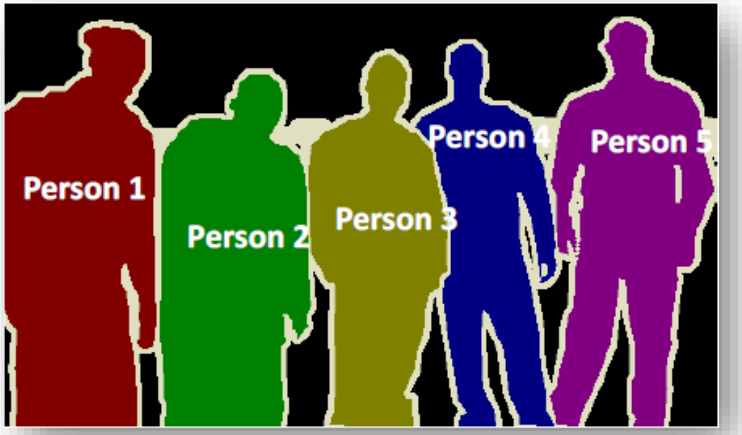


Extension to instance segmentation?

Object detection



Semantic segmentation

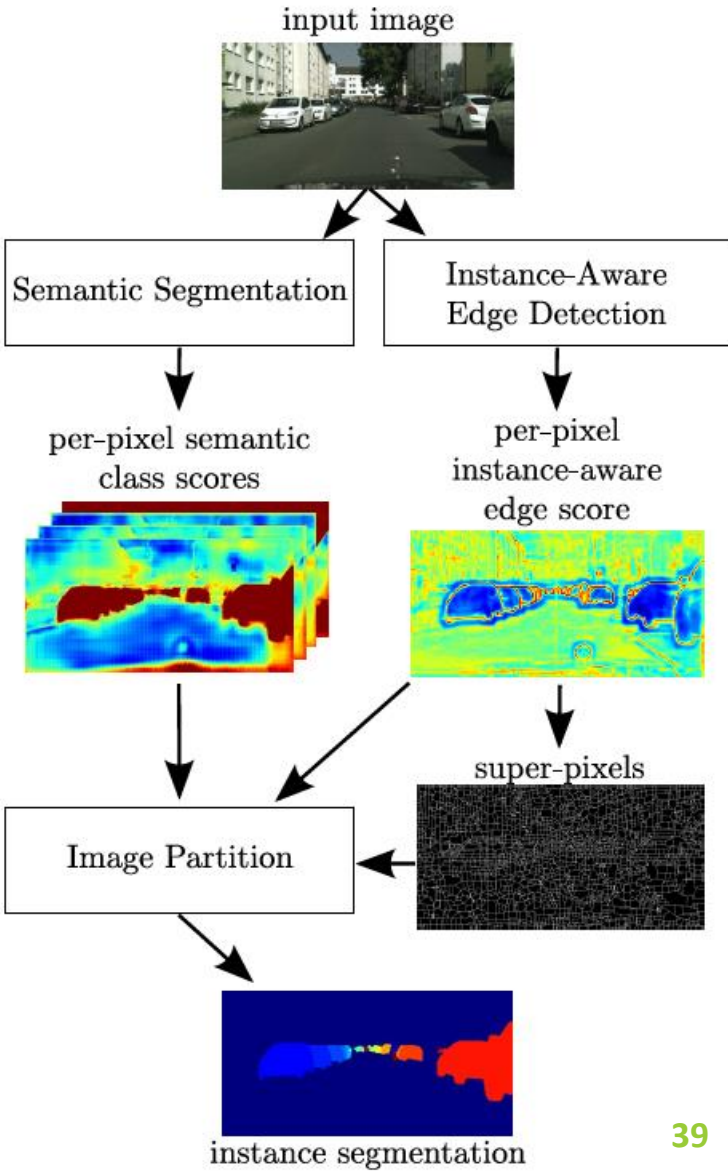


Instance segmentation

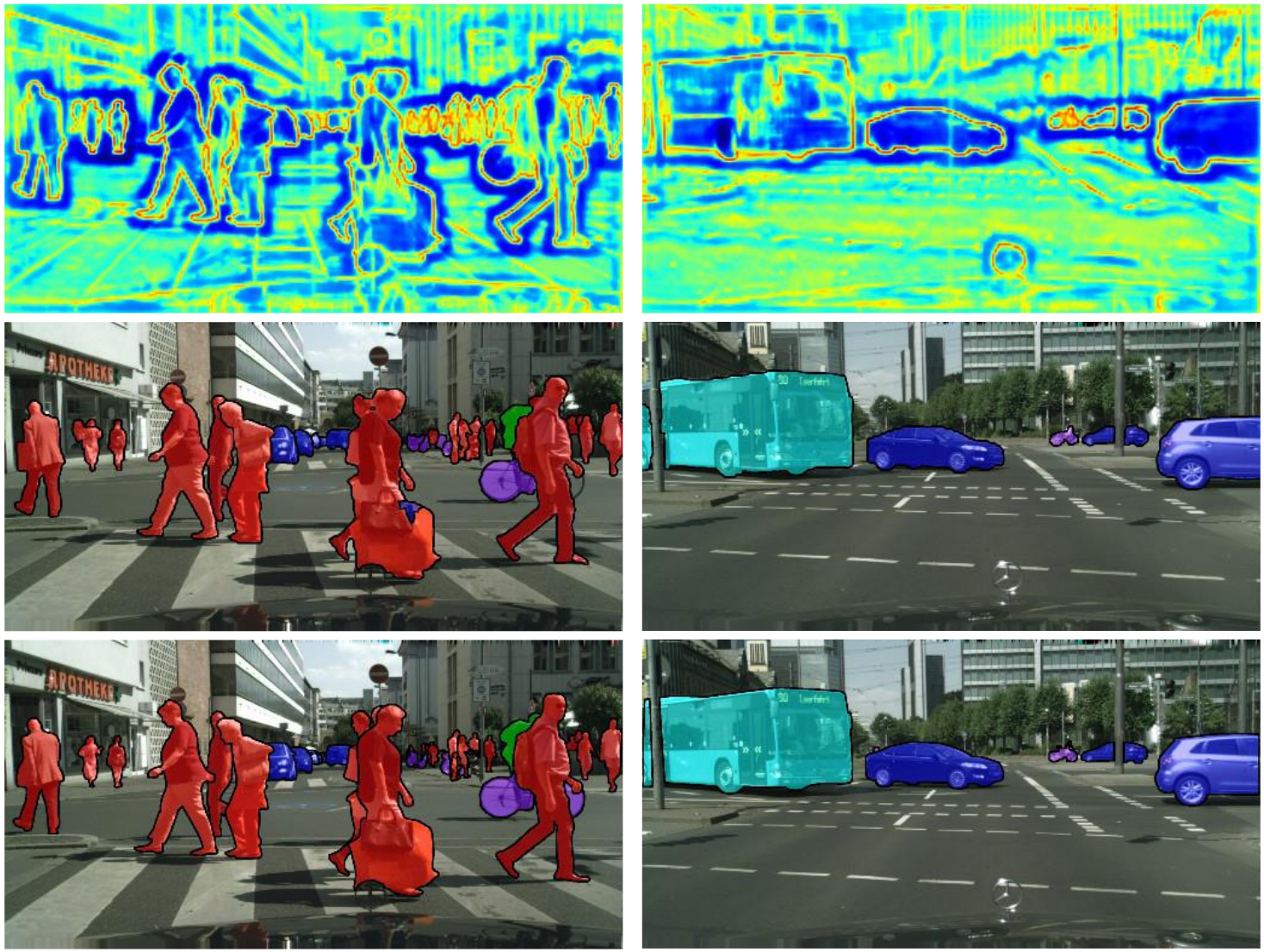
Fig. from He et al., Mask R-CNN: A Perspective on Equivariance

From semantic seg. to instance seg.

- Kirillov et al., InstanceCut: from Edges to Instances with MultiCut, Proc. CVPR'17.
- Applying FCN for instance-aware edge detection.
- Instances are extracted by optimization with Conditional Random Field (CRF).



Instance-aware edge and cut

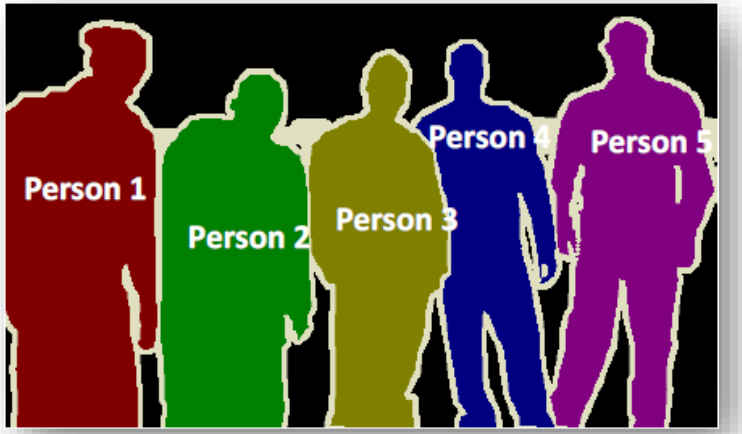


Kirillov et al., InstanceCut: from Edges to Instances with MultiCut, Proc. CVPR'17.

From instance detection to segmentation

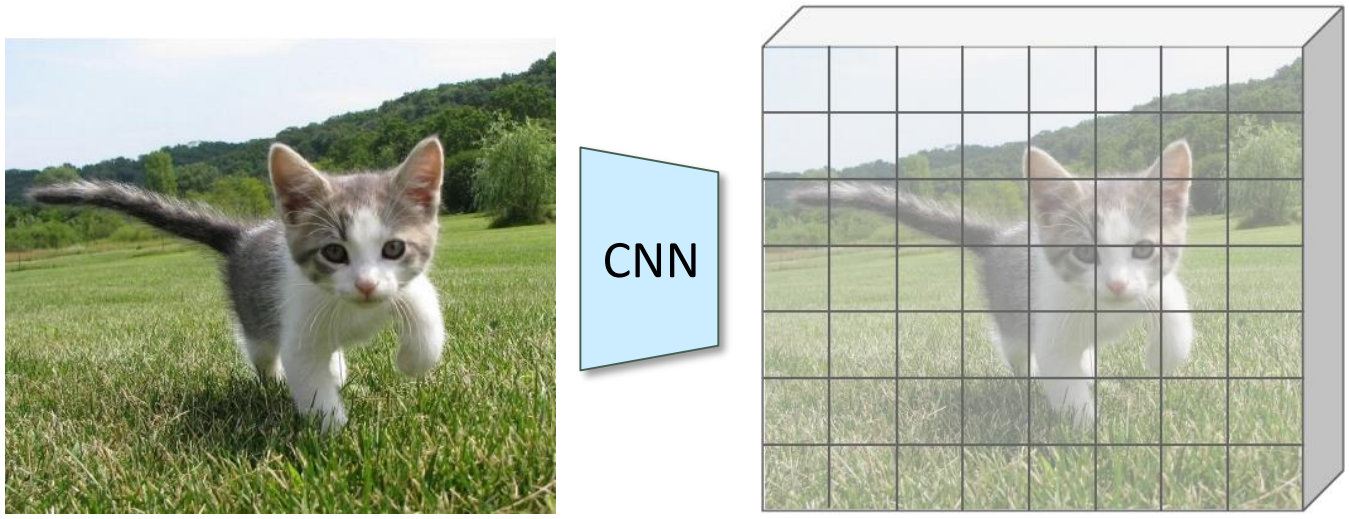
- Could we make use of object/instance detection for instance segmentation ?

- A region of high probability to contain an interesting subject for further analysis is called region proposal or RoI.



Region proposal network

- Proposed by Ren et al, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”, NIPS 2015.



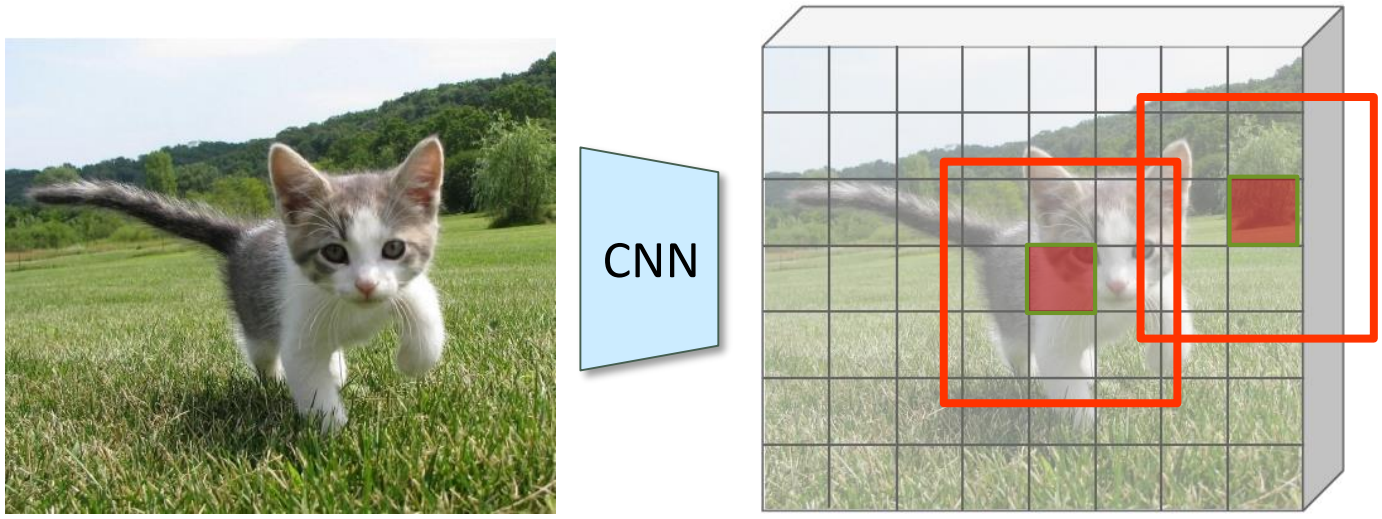
Input Image
(e.g. 3 x 640 x 480)

Image features
(e.g. 512 x 20 x 15)

Slides are modified from F.-F. Li, A. Karpathy, J. Johnson, Convolutional Neural Network for Visual Recognition, course lecture notes.

Region proposal network (cont.)

- Imagine an **anchor** box of fixed size at each point in the feature map.

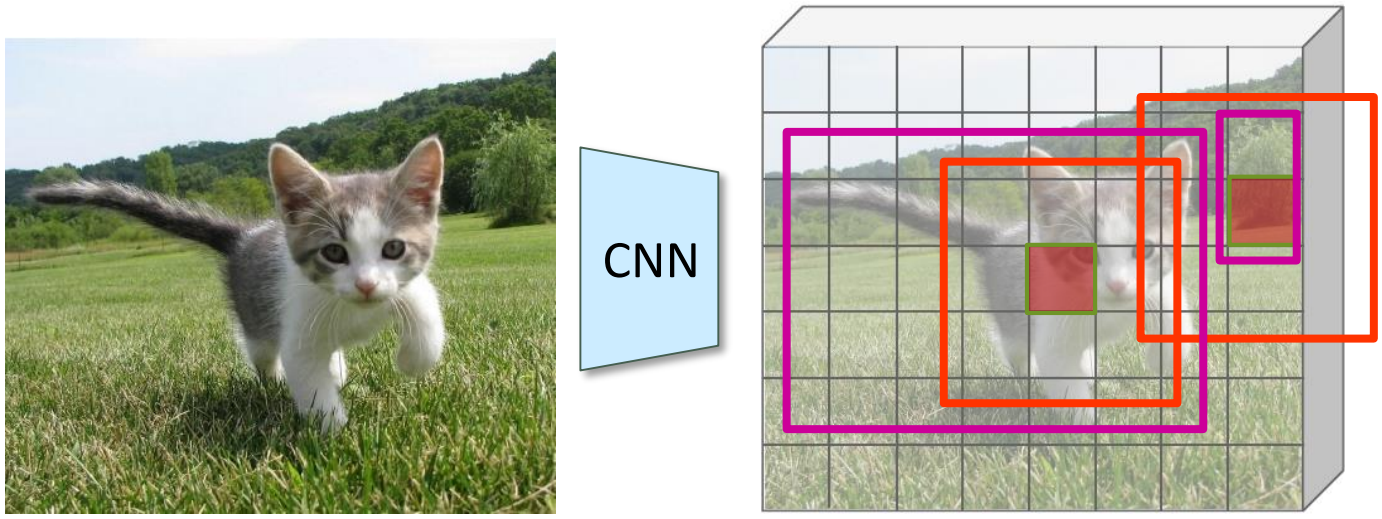


Input Image
(e.g. 3 x 640 x 480)

Image features
(e.g. 512 x 20 x 15)

Region proposal network (cont.)

- For each positive anchor box, also predict a transformation (hw scale, xy shift) to fit for the ideal box.

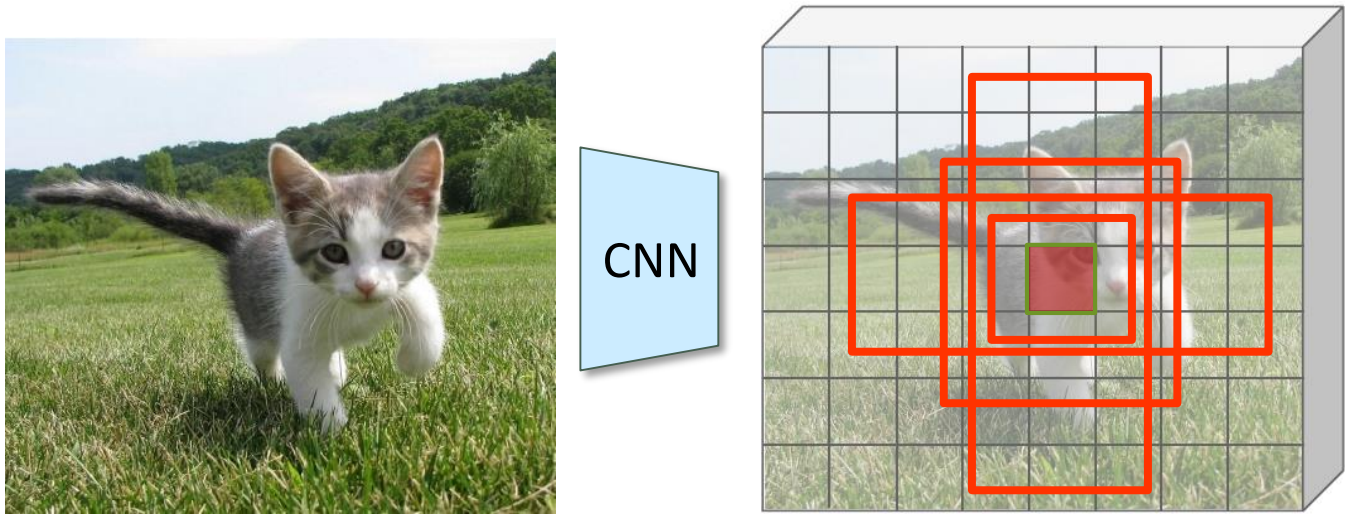


Input Image
(e.g. 3 x 640 x 480)

Image features
(e.g. 512 x 20 x 15)

Region proposal network (cont.)

- In practice use *K* different anchor boxes of different size / scale at each point. (e.g. *K* = 5, 7, 9..)



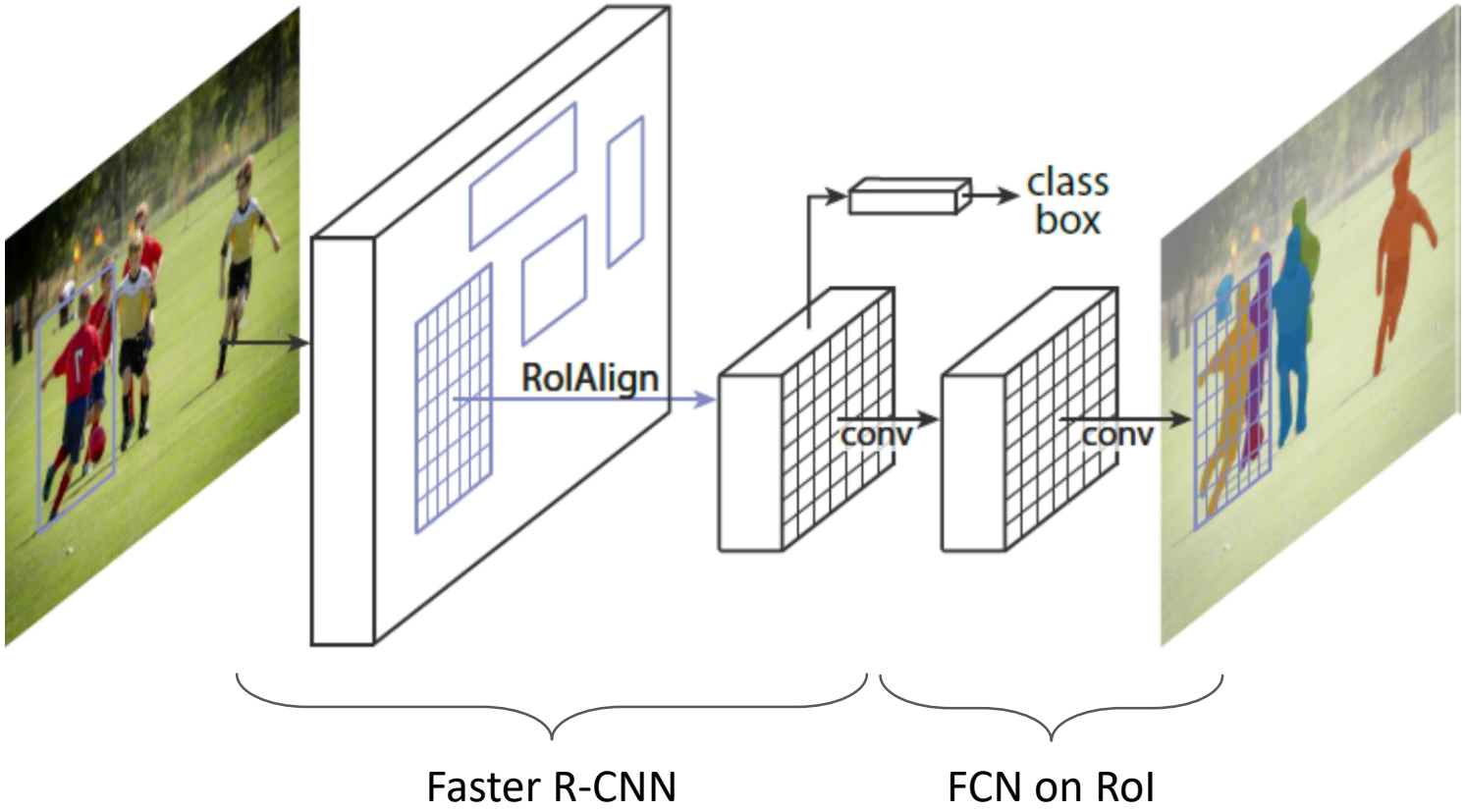
Input Image
(e.g. 3 x 640 x 480)

Image features
(e.g. 512 x 20 x 15)

Sort the *K**20*15 boxes by their scores, and take only the top (e.g.~300) as the proposals

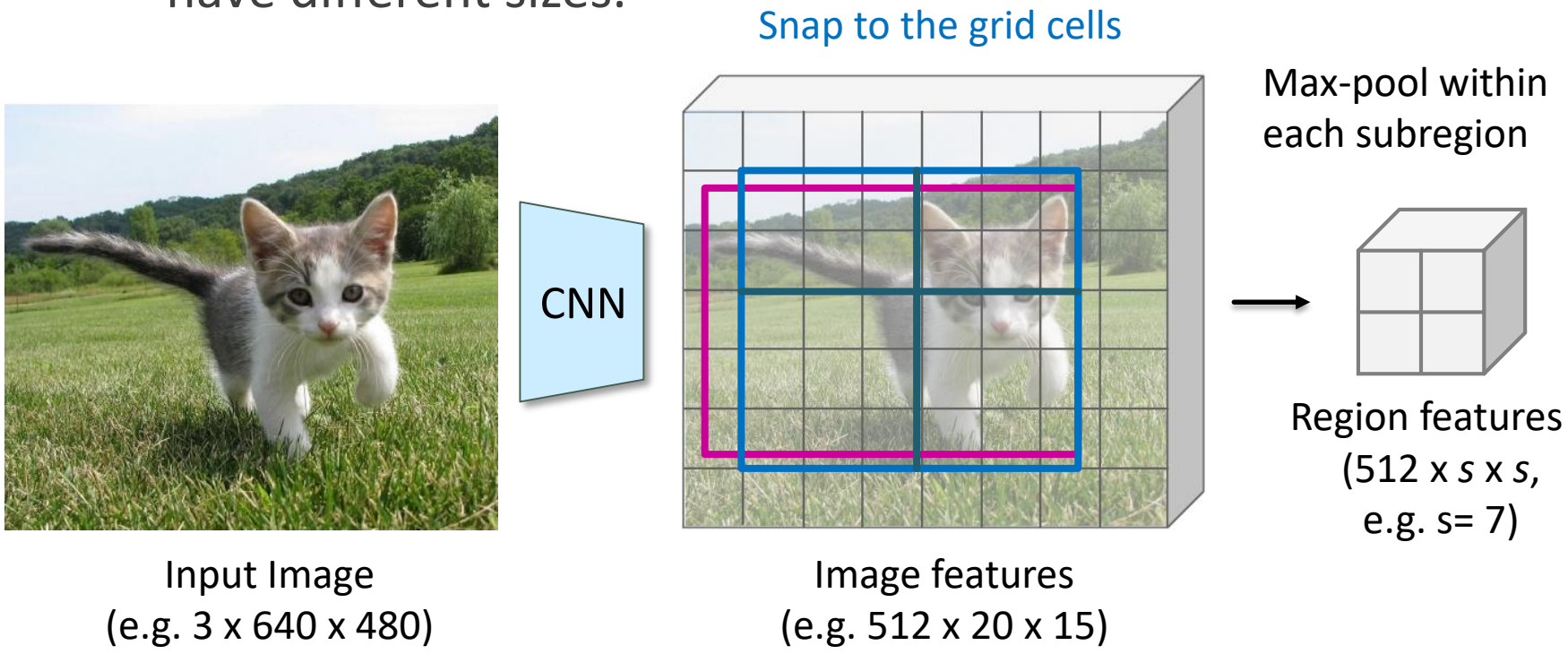
Segmentation from the proposal boxes

- He et al., Mask RCNN, ICCV 2017.



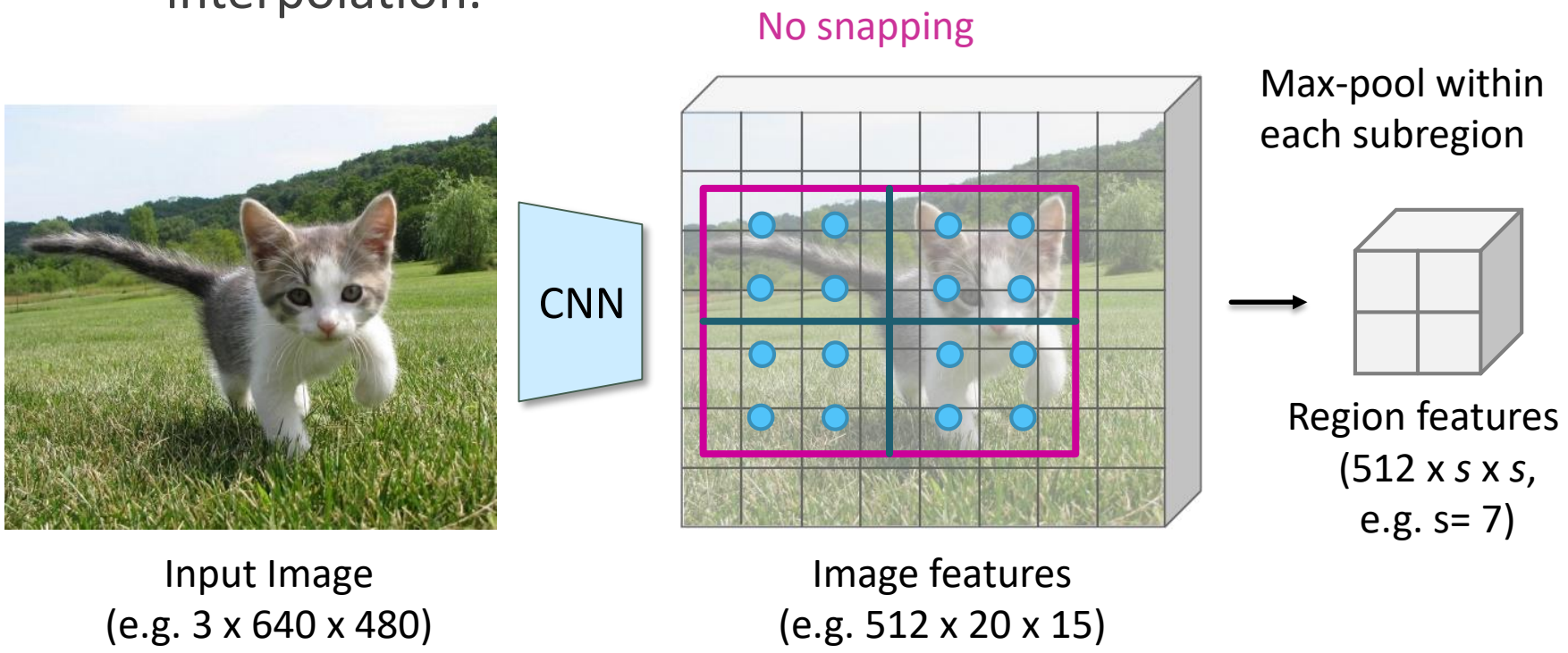
Cropping features: RoI pool

- Region features always the same size even if input regions have different sizes.



Cropping features: RoI align

- Sampling at regular points in each subregion using bilinear interpolation.

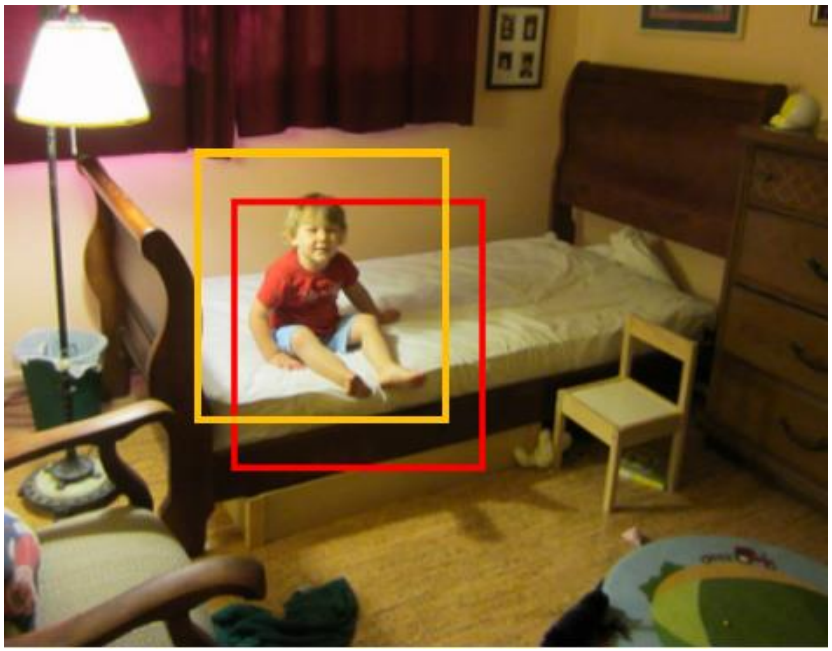


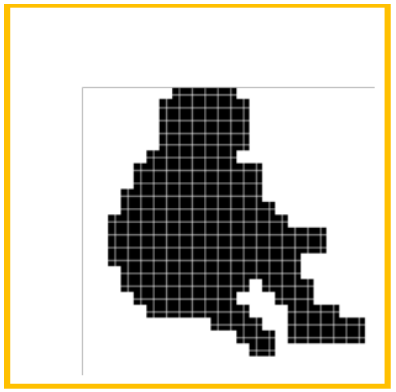
Dividing the region into *s* x *s* subregions

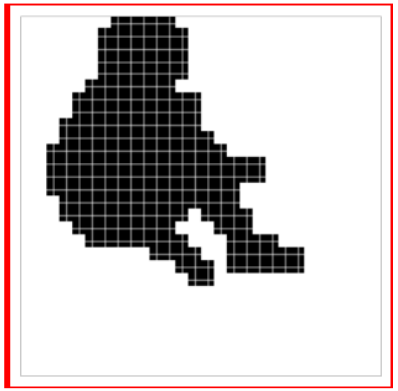
Sampling at regular points

Fully-Conv on RoI

- Translation of object in RoI => Same translation of mask in RoI
- Robust to localization imperfection of RoIs.



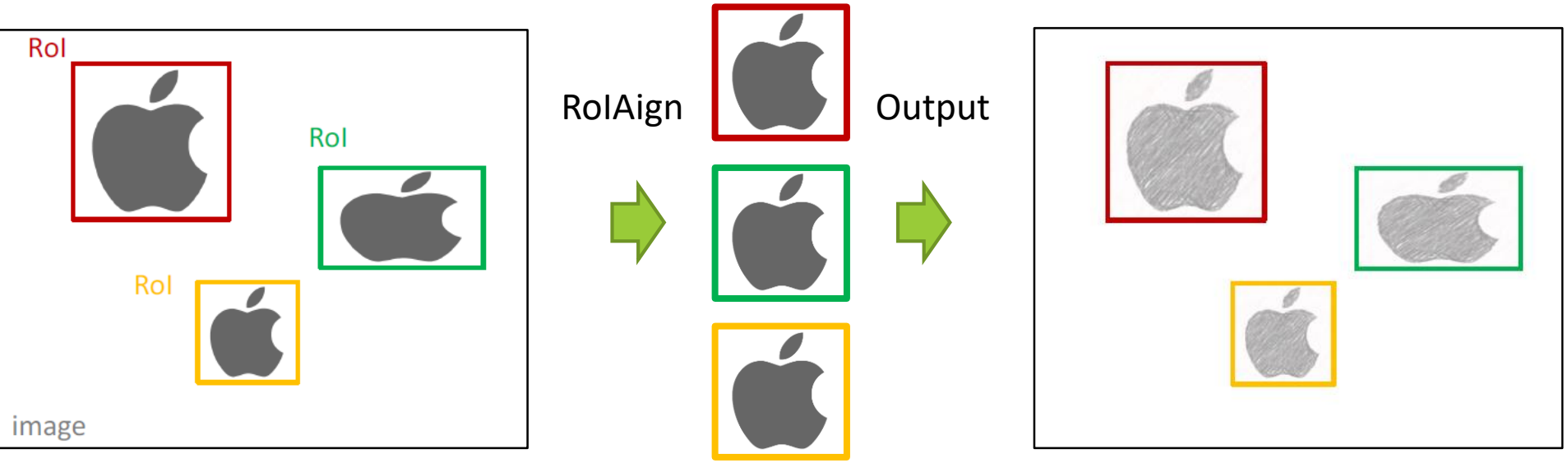




Slides are modified from He et al., Mask R-CNN: A Perspective on Equivariance

Scale-Equivariance of RoIAlign

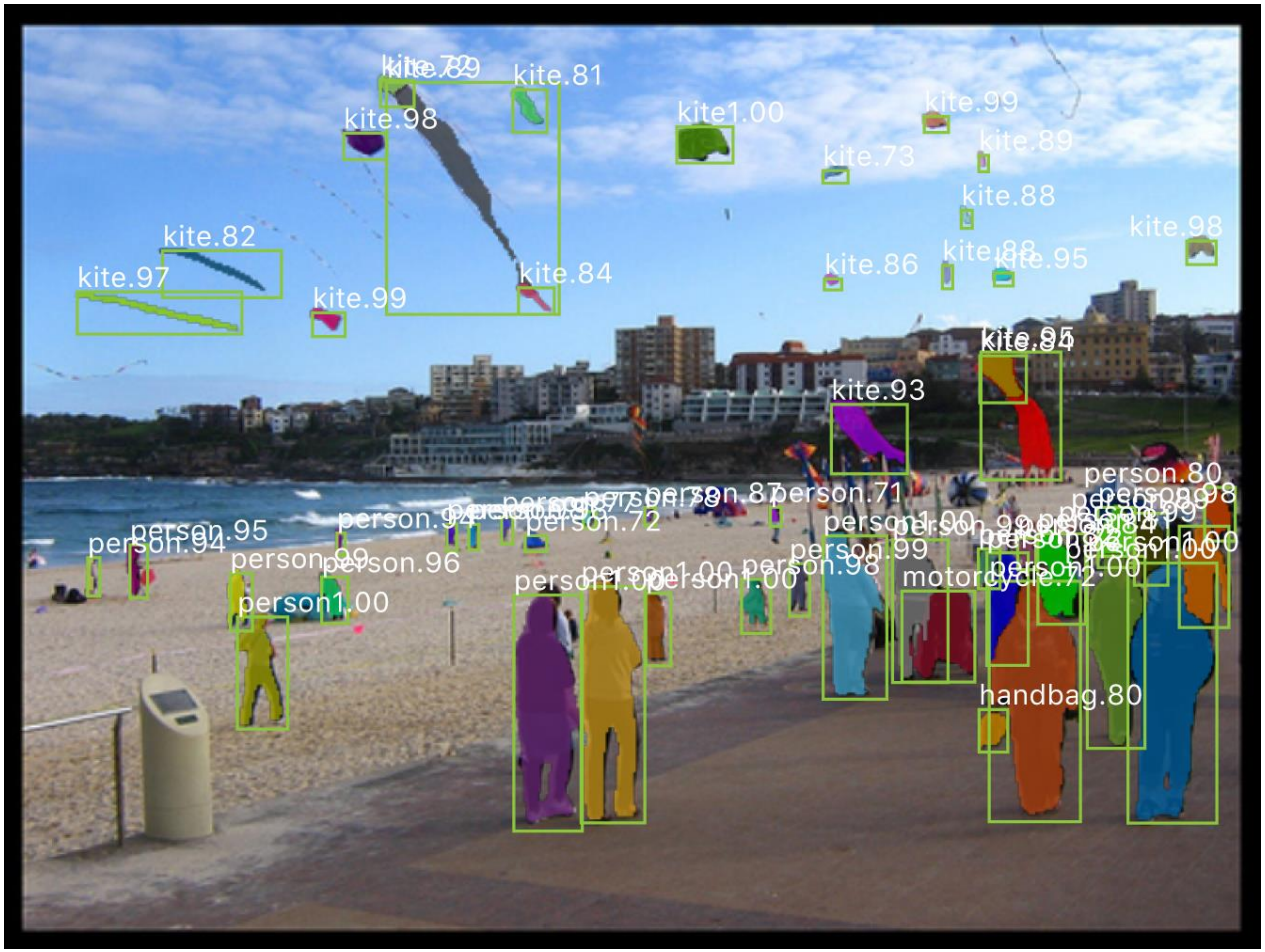
- RoIAlign creates scale-invariant representations
- RoIAlign + “output pasted back” provides scale-equivariance.



Mask R-CNN results

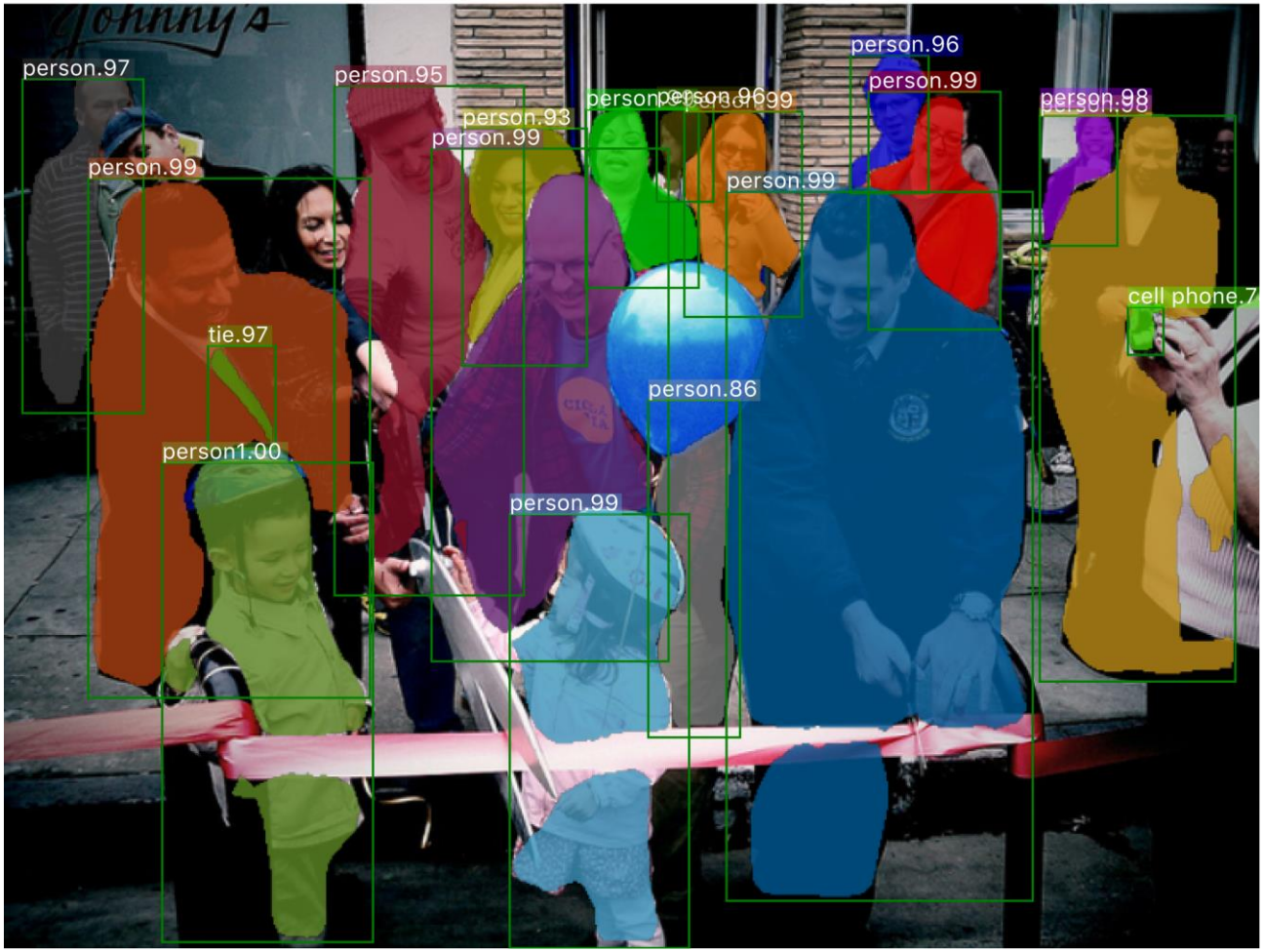
Object surrounded by same-category objects

Mask R-CNN results



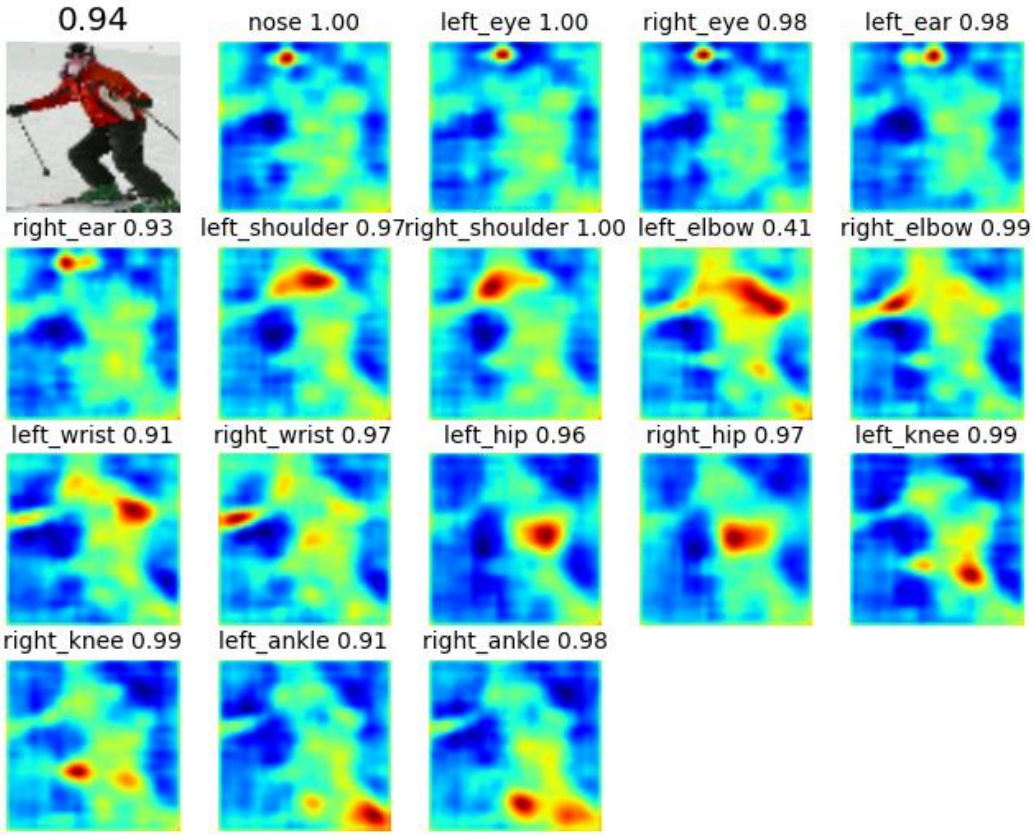
Scenes with small objects

Mask R-CNN failure cases



Mask R-CNN for human pose estimation

- Model a keypoint's location as a one-hot mask
- Adopt Mask R-CNN to predict K masks. (one for each of K keypoint types)



Mask R-CNN for pose

The oversmoothed boundaries

The results of Mask R-CNN

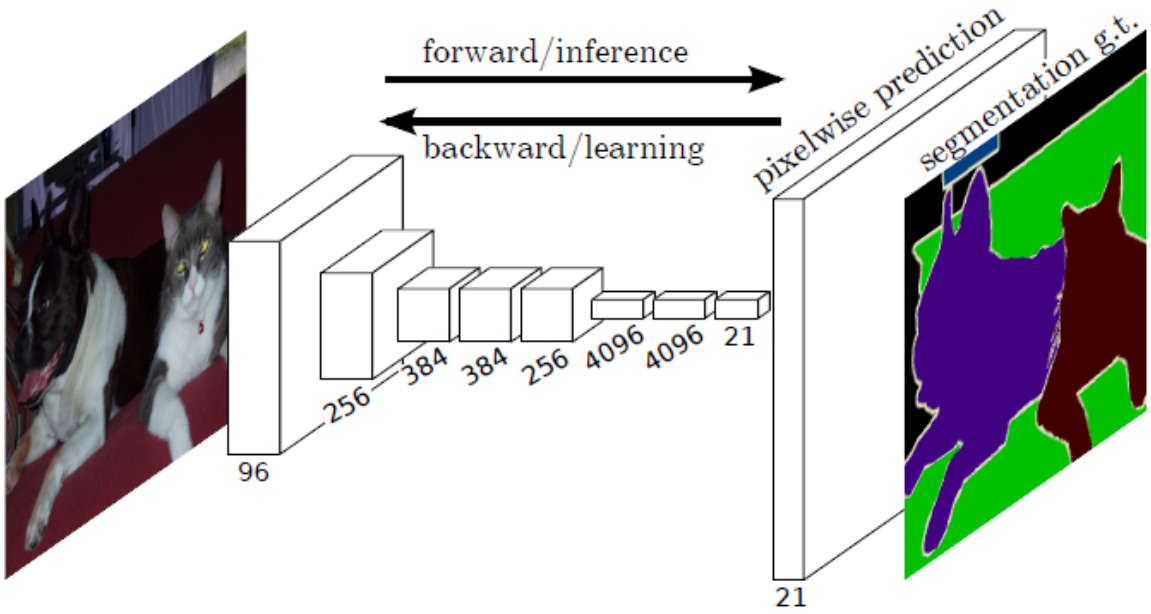
OK ~ Good !

If the boundaries become “crisp”

Better !!

The problem

- Current methods usually a regular grid for upsampling.
  - unnecessarily oversample the smooth areas.
  - simultaneously undersample object boundaries.

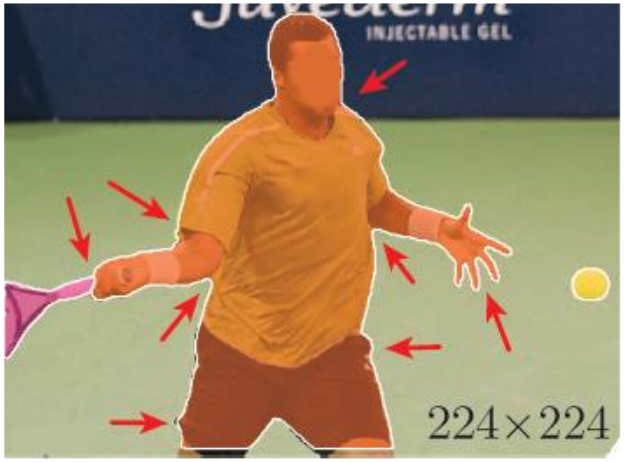


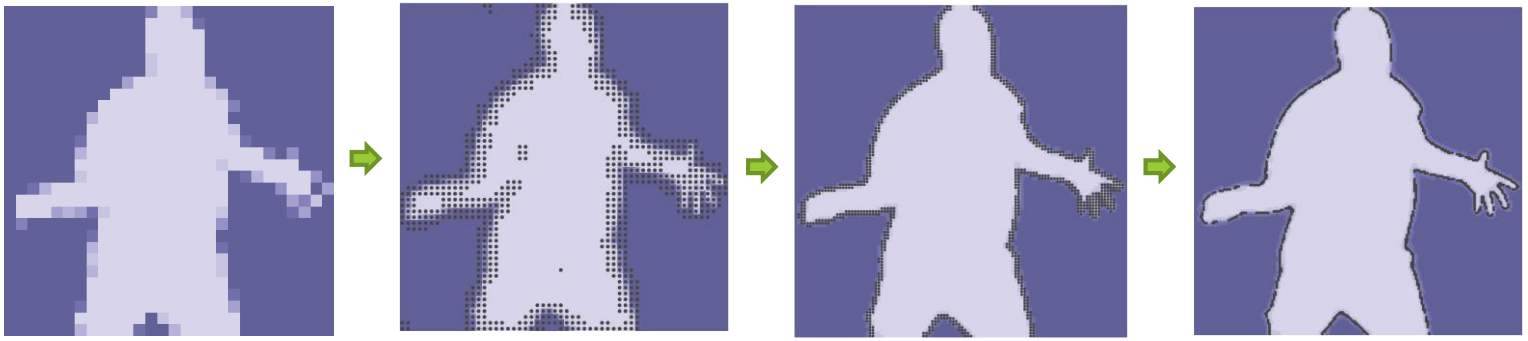
Refining the boundaries

- Kirillov et al., PointRend: Image Segmentation as Rendering, CVPR 2020.

Mask R-CNN

+ PointRend

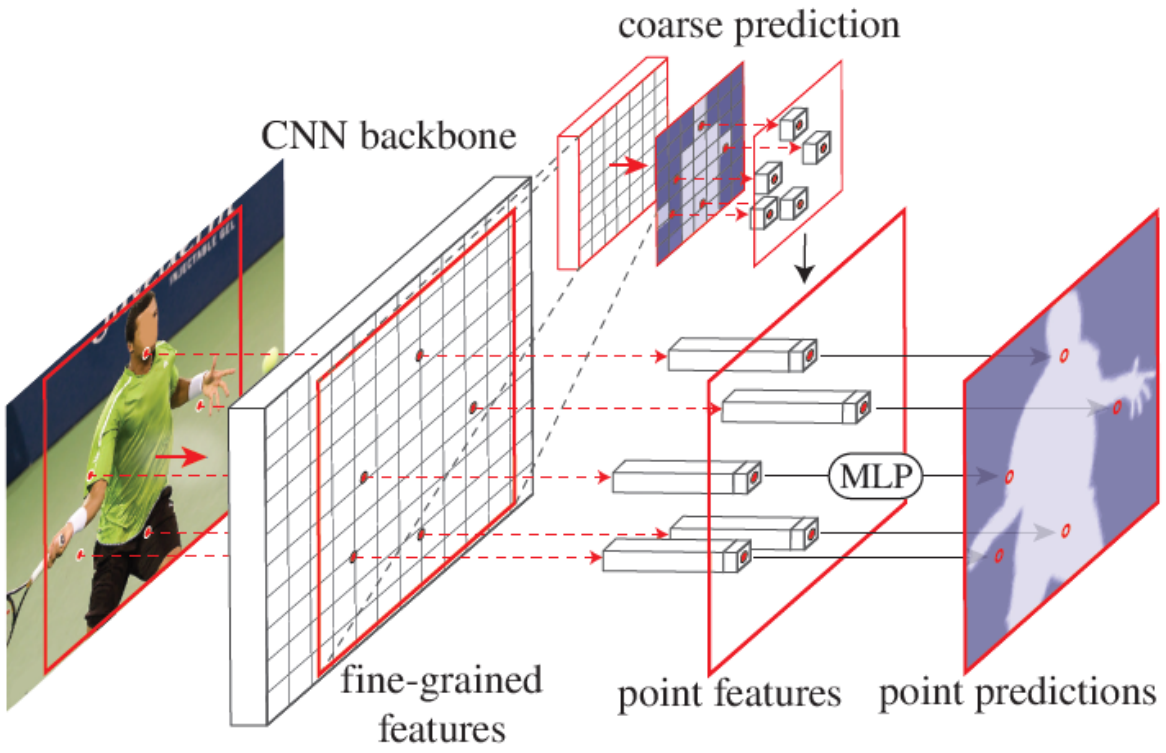




28x28 56x56 112x112 224x224

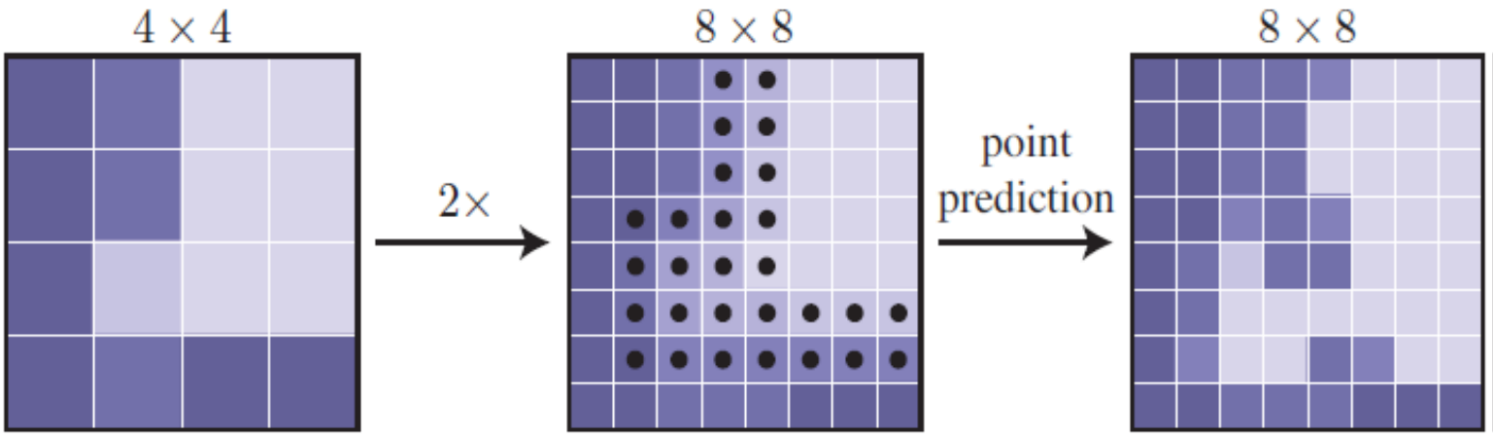
PointRend

- Selects a set of points (red dots) and makes prediction for each point independently with a small MLP.
  - Iteratively refine uncertain regions of the predicted mask.



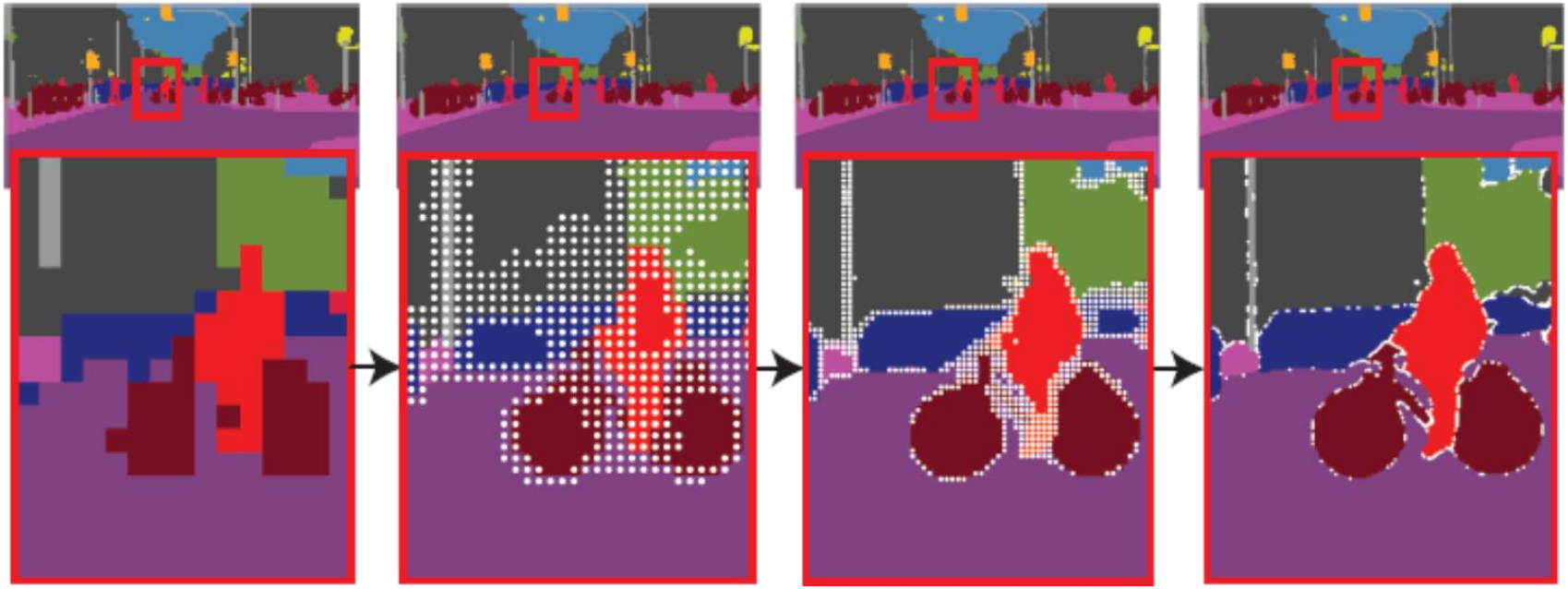
PointRend: in each iteration

- Upsamples previously predicted segmentation by bilinear interpolation.
- Selects the *N* most uncertain points.
  E.g. probabilities closest to 0.5.
- Extracts point-wise features by bilinear interpolation.
- Predicts a label from this features by a small MLP.



PointRend: point selection

- Reaches 1024x2048 resolution (i.e. 2M points) by making predictions for only 32k points



PointRend: results