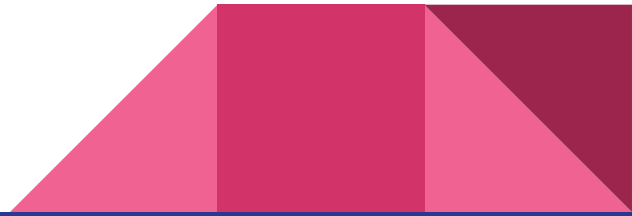
2D Object Detection and Segmentation

Jiayuan Gu
2019.10.29

Outline

- What is object detection and segmentation?
- Object detection: R-CNN
- Segmentation: U-Net



Why do we focus on R-CNN?

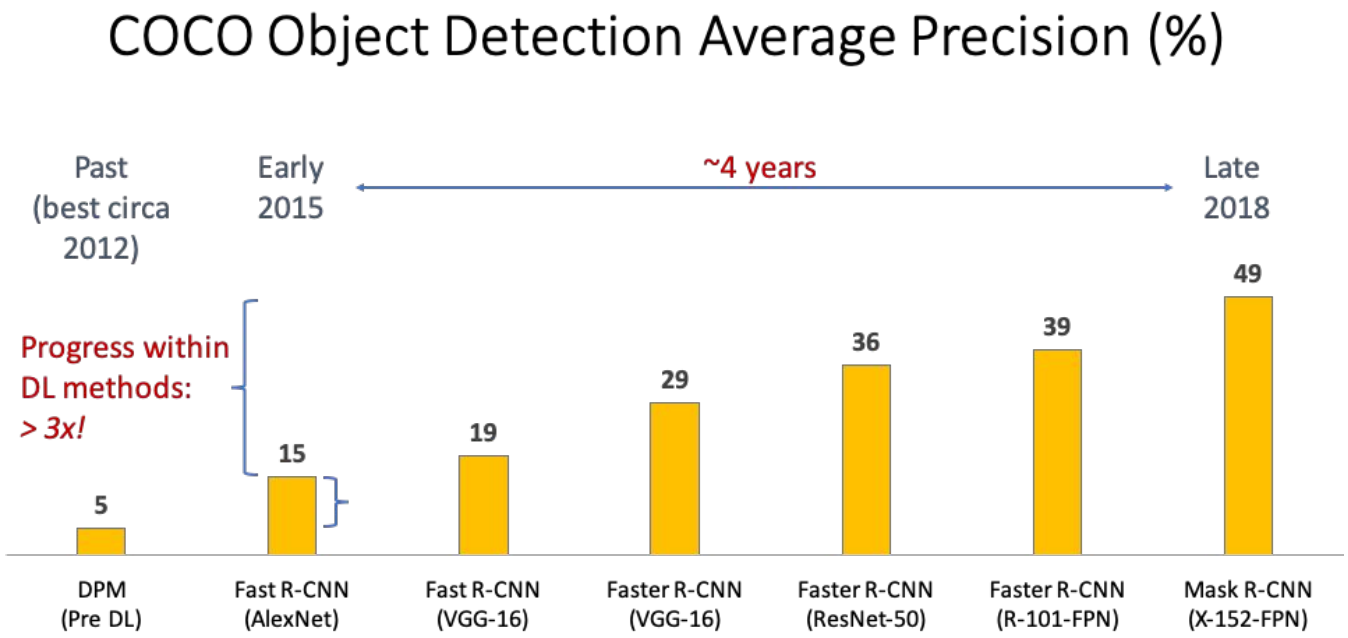


Figure from Ross Girshick tutorial at CVPR 2019

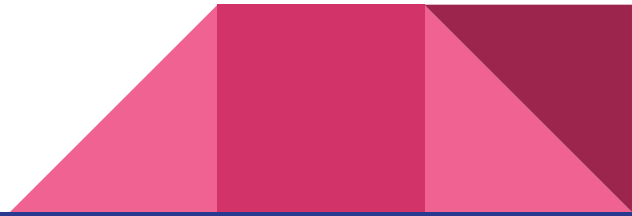
Resources

Tutorials of CVPR

- http://deeplearning.csail.mit.edu
- https://sites.google.com/view/cvpr2018-recognition-tutorial
- http://feichtenhofer.github.io/cvpr2019-recognition-tutorial

Github Repo

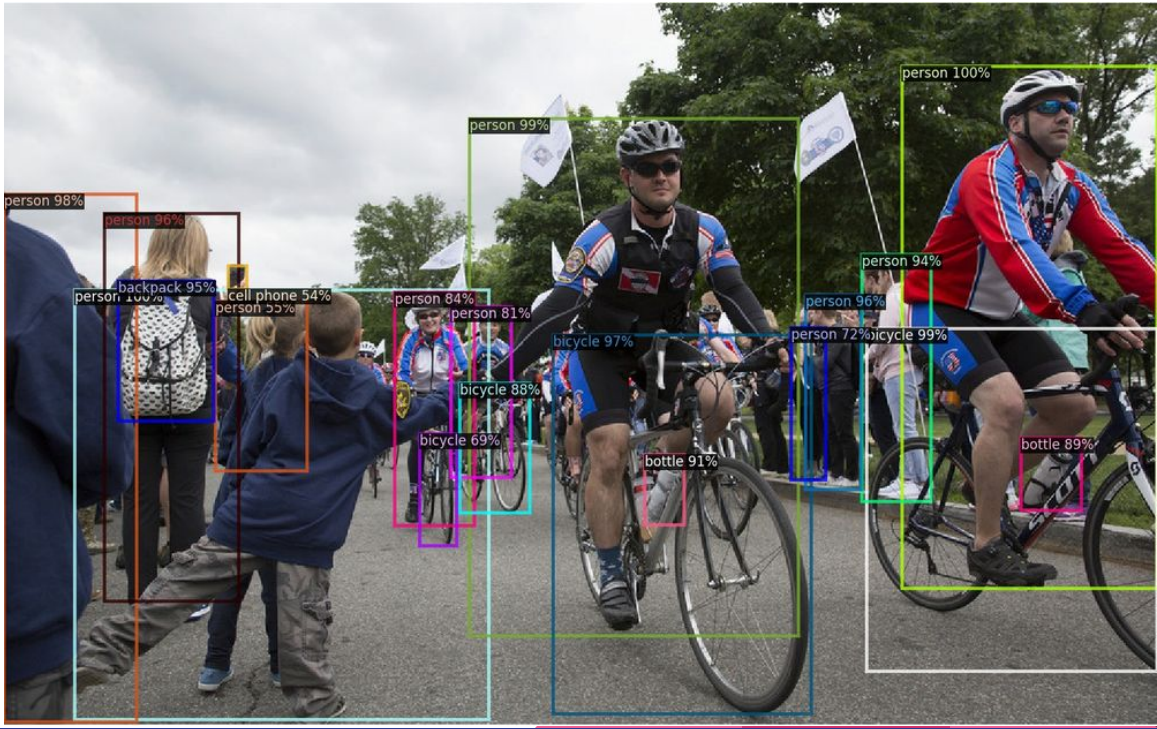
https://github.com/facebookresearch/detectron2



Background

How do we represent objects

- Bounding box



Figures from

How do we represent objects

- Bounding box
- Instance mask

Figures from

How do we represent objects

- Bounding box
- Instance mask
- Keypoint

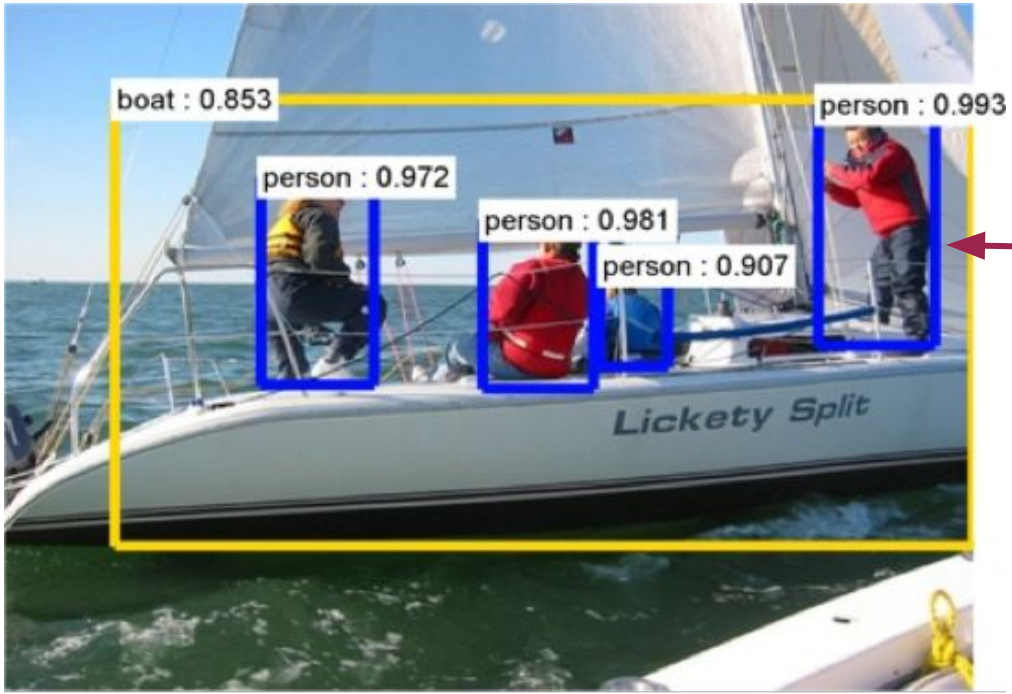
Figures from

How do we represent objects

- **Bounding box**
- Instance mask
- Keypoint

Figures from

Object Detection with Bounding Boxes



What? - Recognition/Classification

Where? - Localization/Regression

"Object detection"

Slides modified from Ross Girshick tutorial at CVPR 2019

Object Detection with Segmentation Masks

What? - Recognition

Where? - Segmentation

"Instance segmentation"

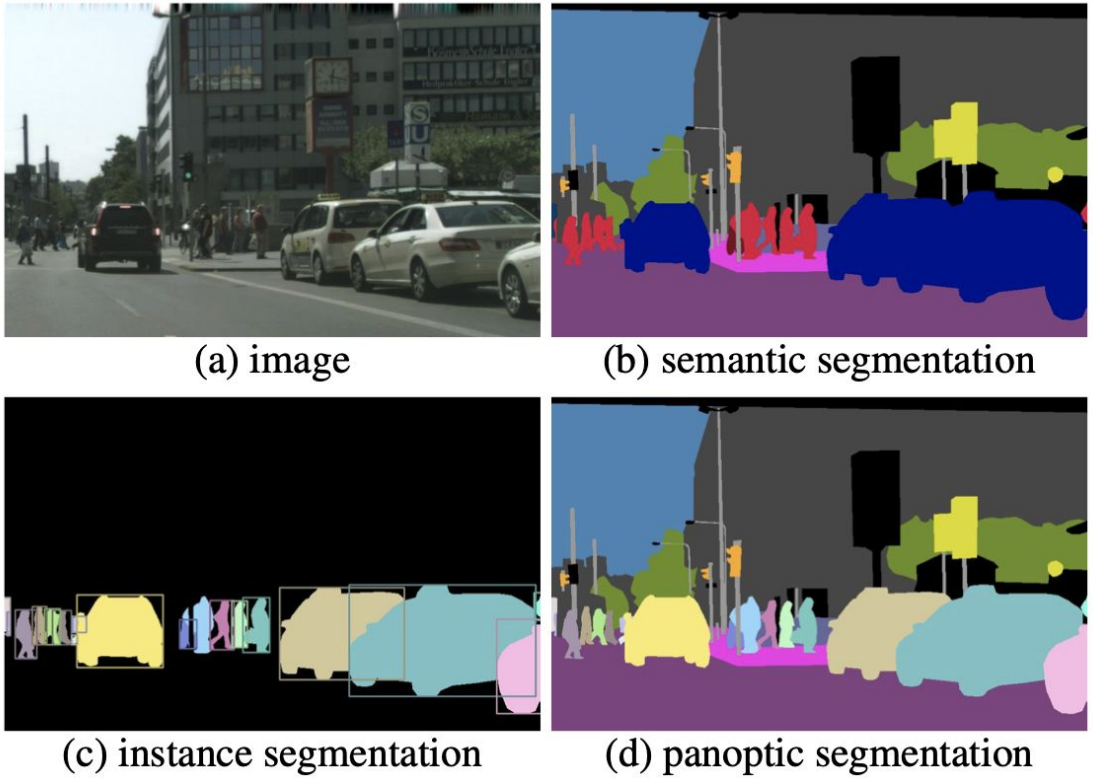
Slides modified from Ross Girshick tutorial at CVPR 2019

Semantic Segmentation

Predict a pixel-wise class label

Stuff: walls, buildings, sky, road

Things: human, cars, bikes



Figures from *Panoptic Segmentation*, CVPR 2019

Datasets

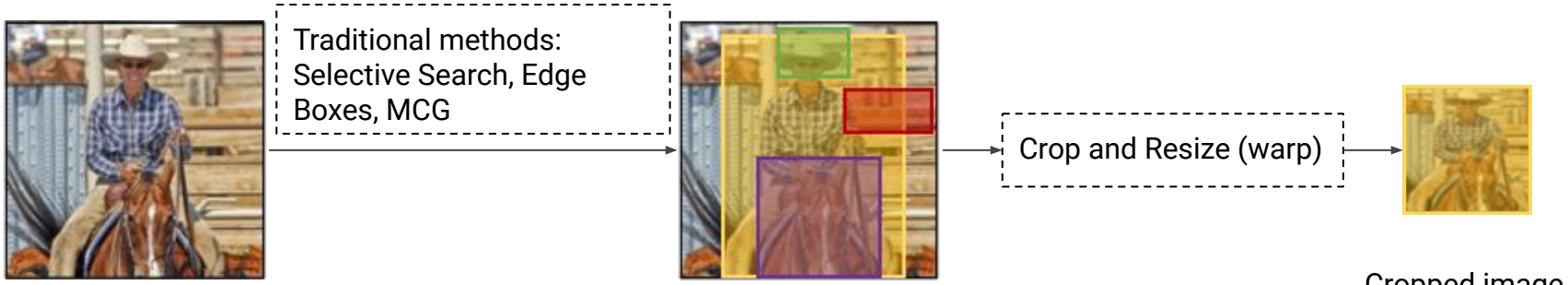


Microsoft COCO

Visual Object Classes Challenge 2012 (VOC2012)

R-CNN: Region-based CNN

Object Detection → Object Classification



We've already reduced object detection to object classification!

R-CNN

Computationally expensive

Input: an image

Traditional methods:
Selective Search, Edge Boxes, MCG

Proposals/Candidates

Cropped image
Region of Interests (RoI)

Crop and Resize (warp)

ConvNet

Class Probability

How probable is it a human?

BBox Regression

How can we modify this bounding box?

Slides modified from Ross Girshick tutorial at CVPR 2019

R-CNN: Crop→ CNN v.s. Fast R-CNN: CNN → Crop

Fast R-CNN

Input: an image

Traditional methods:
Selective Search, Edge Boxes, MCG

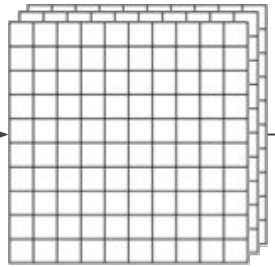
Proposals/Candidates
Region of Interests (RoI)

Class Probability

BBox Regression

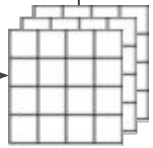
ConvNet
Multilayer Perceptron (MLP)

ConvNet



RoI-Pool

Similar to Crop & Resize

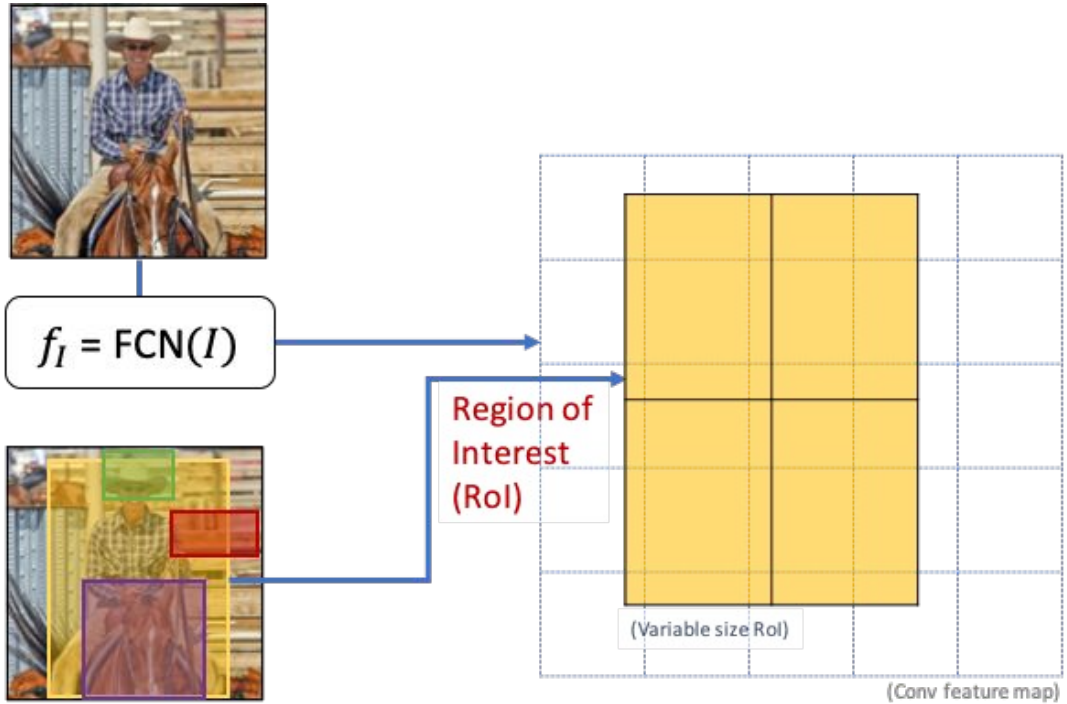


Feature map for an image

Feature map for a RoI

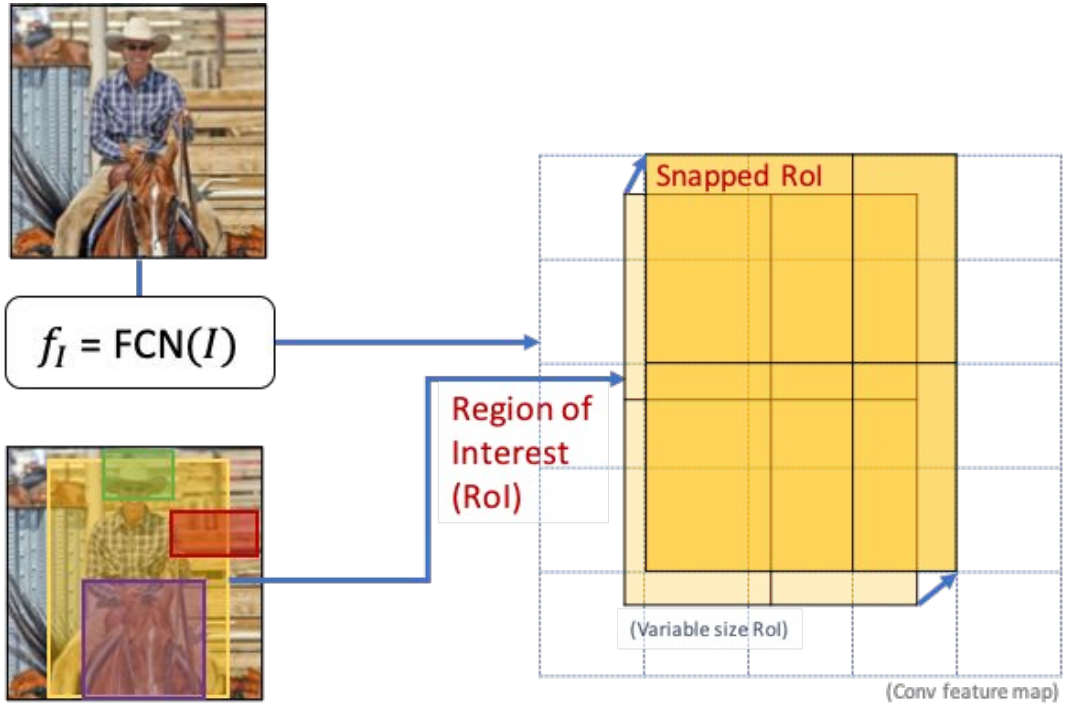
Slides modified from Ross Girshick tutorial at CVPR 2019

RoI Pooling (for each proposal)



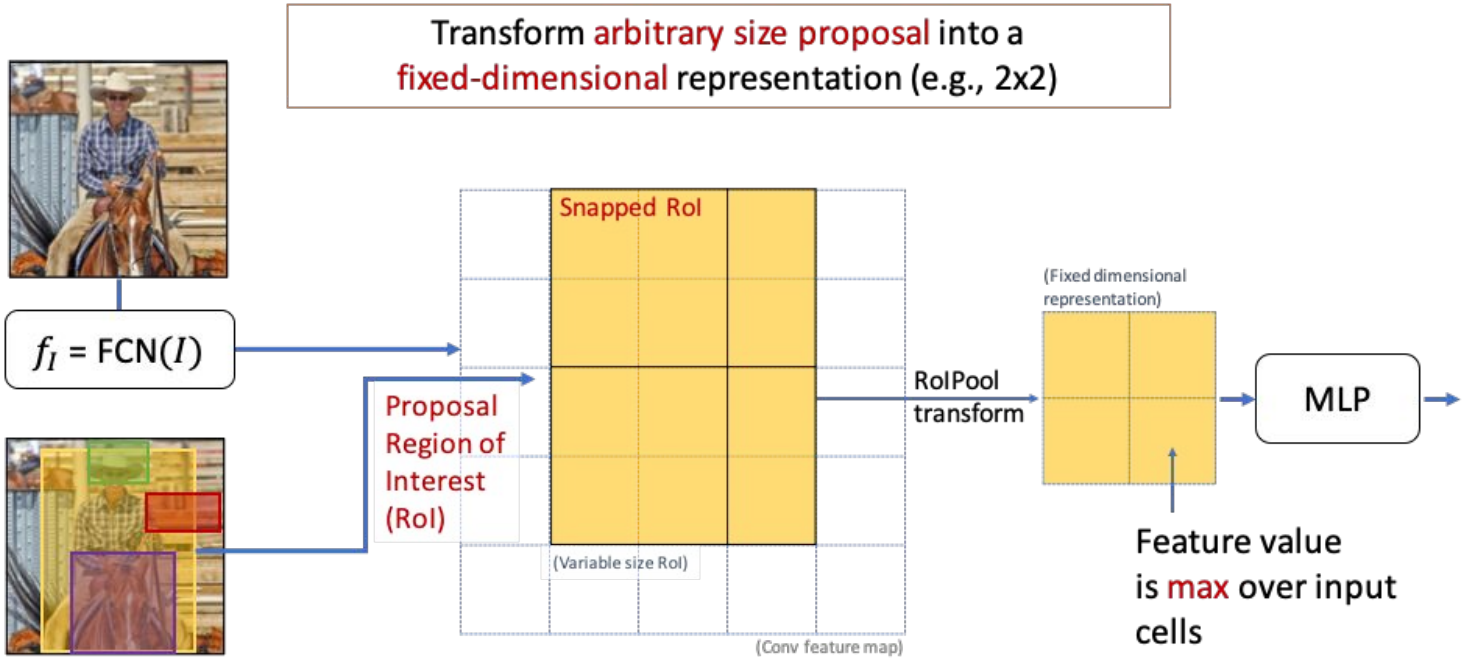
Slides modified from Ross Girshick tutorial at CVPR 2019

RoI Pooling (for each proposal)



Slides modified from Ross Girshick tutorial at CVPR 2019

RoI Pooling (for each proposal)



Slides modified from Ross Girshick tutorial at CVPR 2019

Fast R-CNN

Not good enough

Input: an image

Traditional methods:
Selective Search, Edge Boxes, MCG

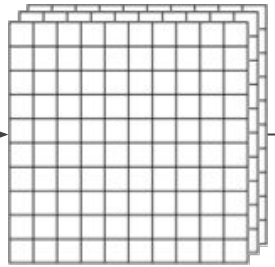
Proposals/Candidates
Region of Interests (RoI)

Class Probability

BBox Regression

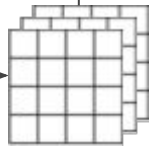
ConvNet
Multilayer Perceptron (MLP)

ConvNet



RoI-Pool

Similar to Crop & Resize

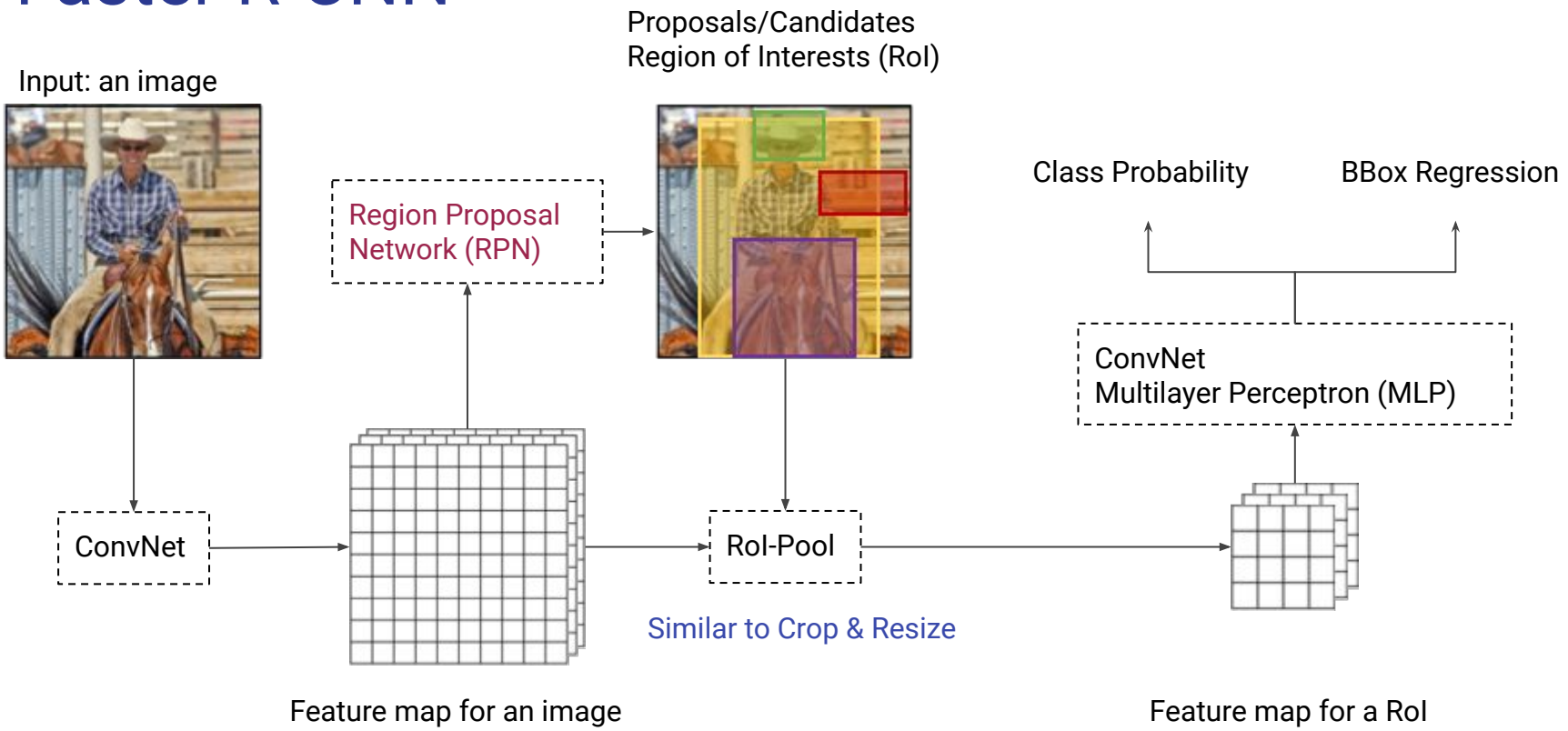


Feature map for an image

Feature map for a RoI

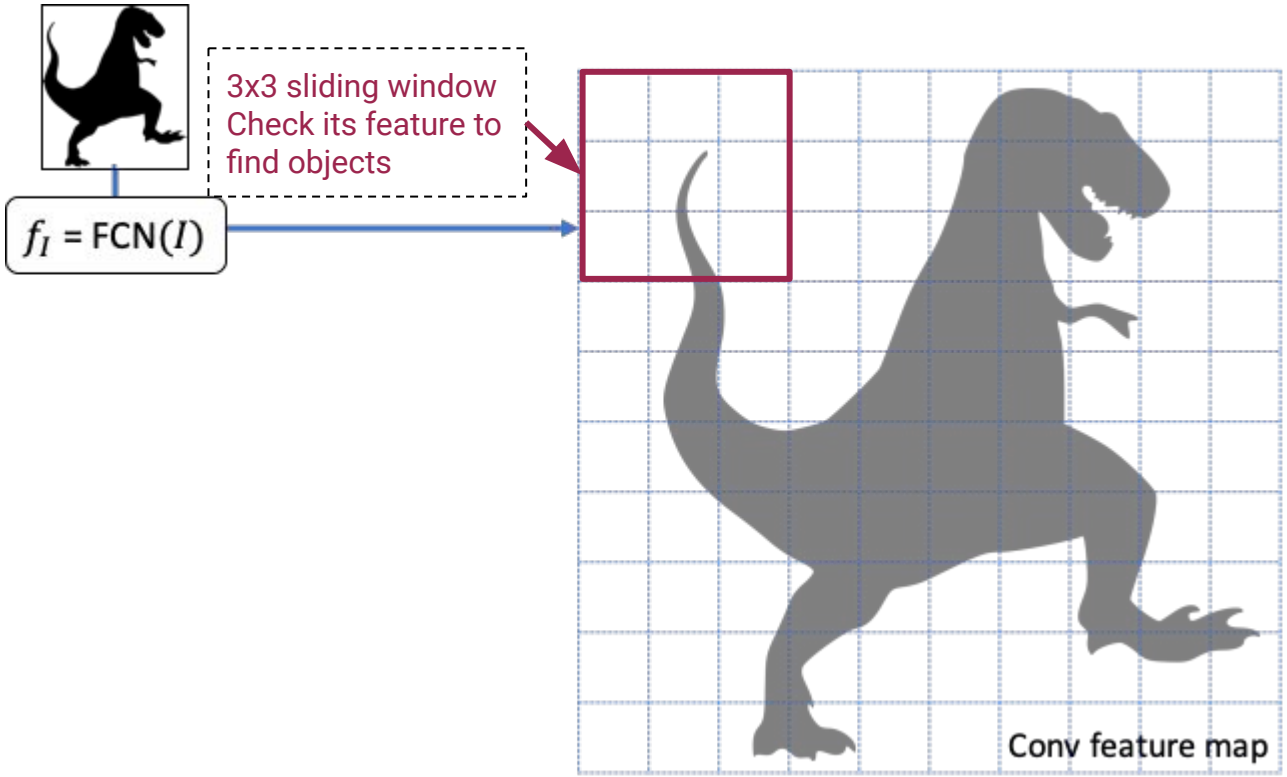
Slides modified from Ross Girshick tutorial at CVPR 2019

Faster R-CNN



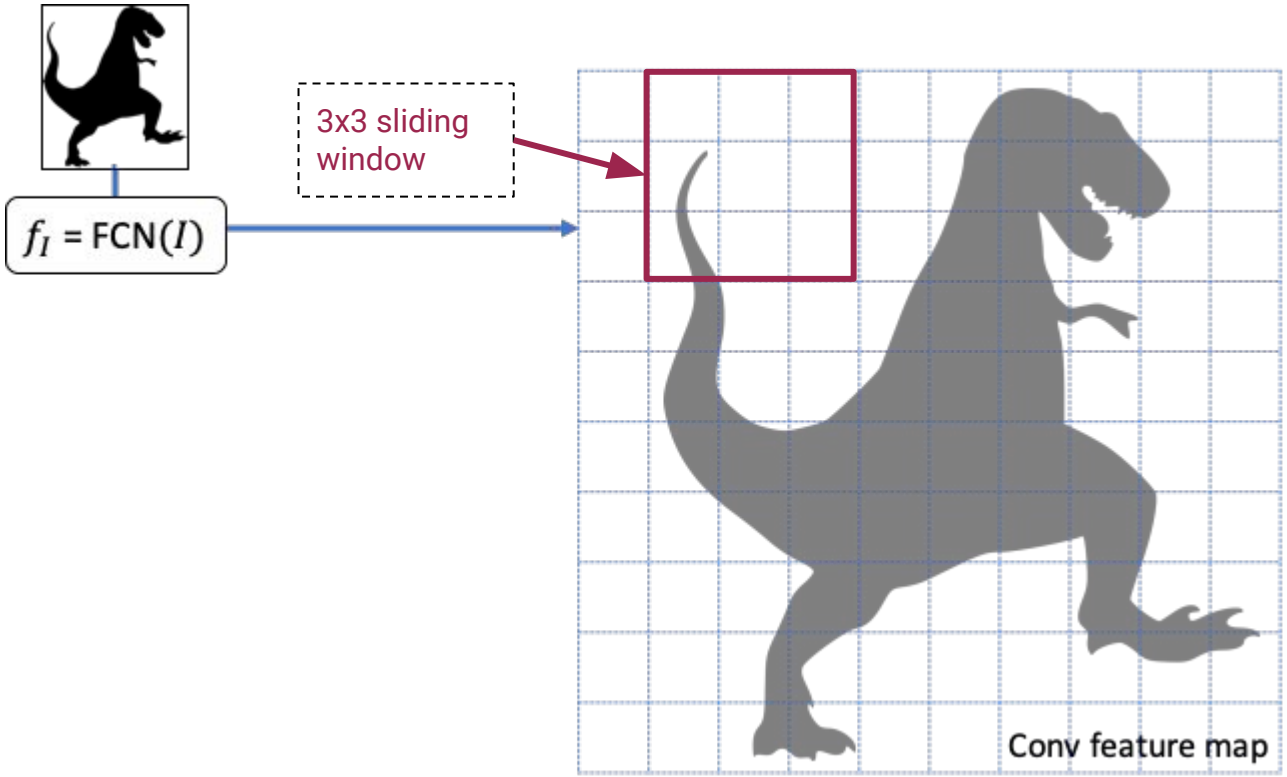
Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Region Proposal Network



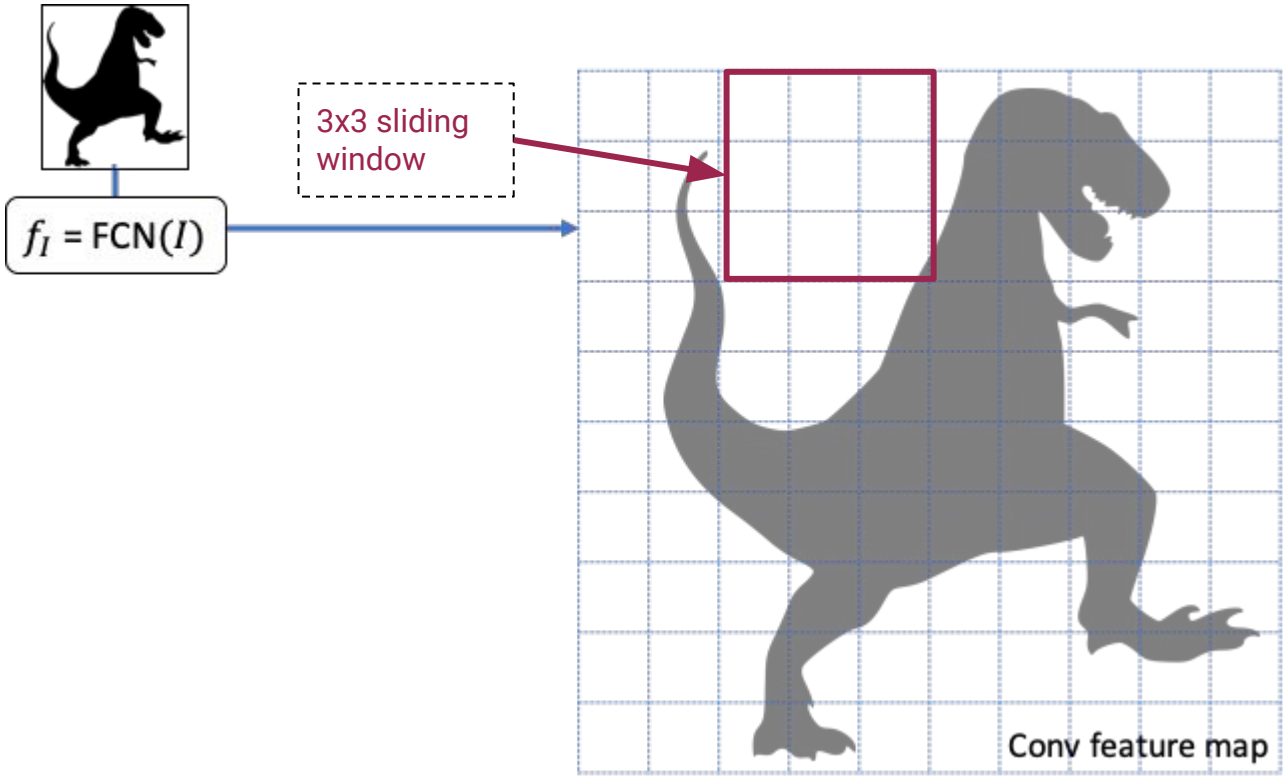
Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Region Proposal Network



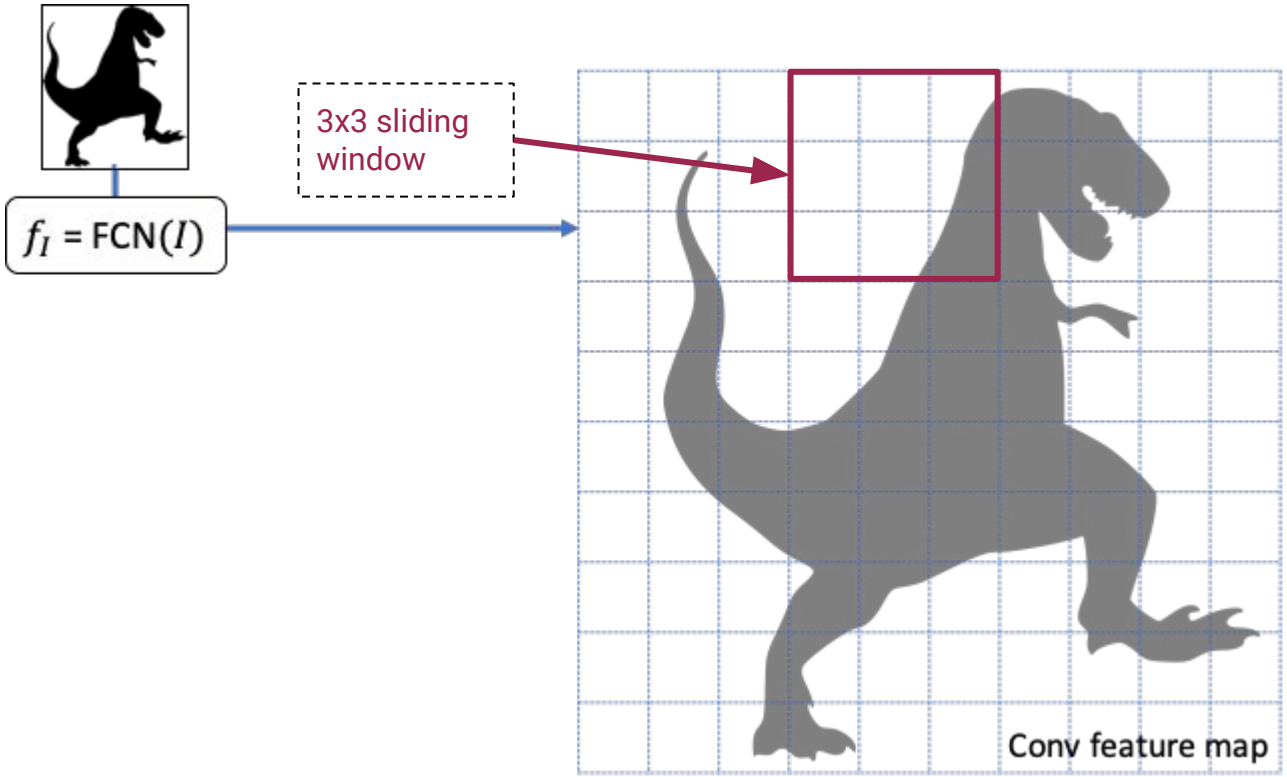
Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Region Proposal Network



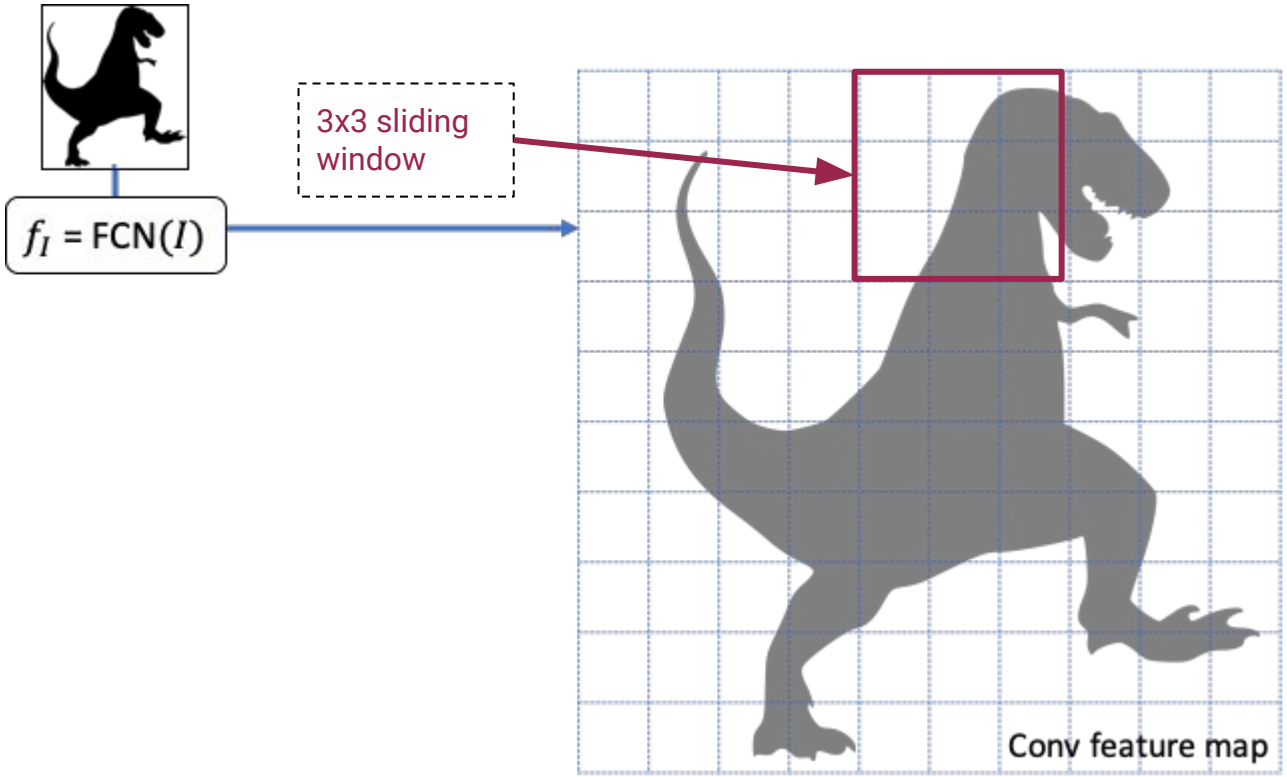
Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Region Proposal Network



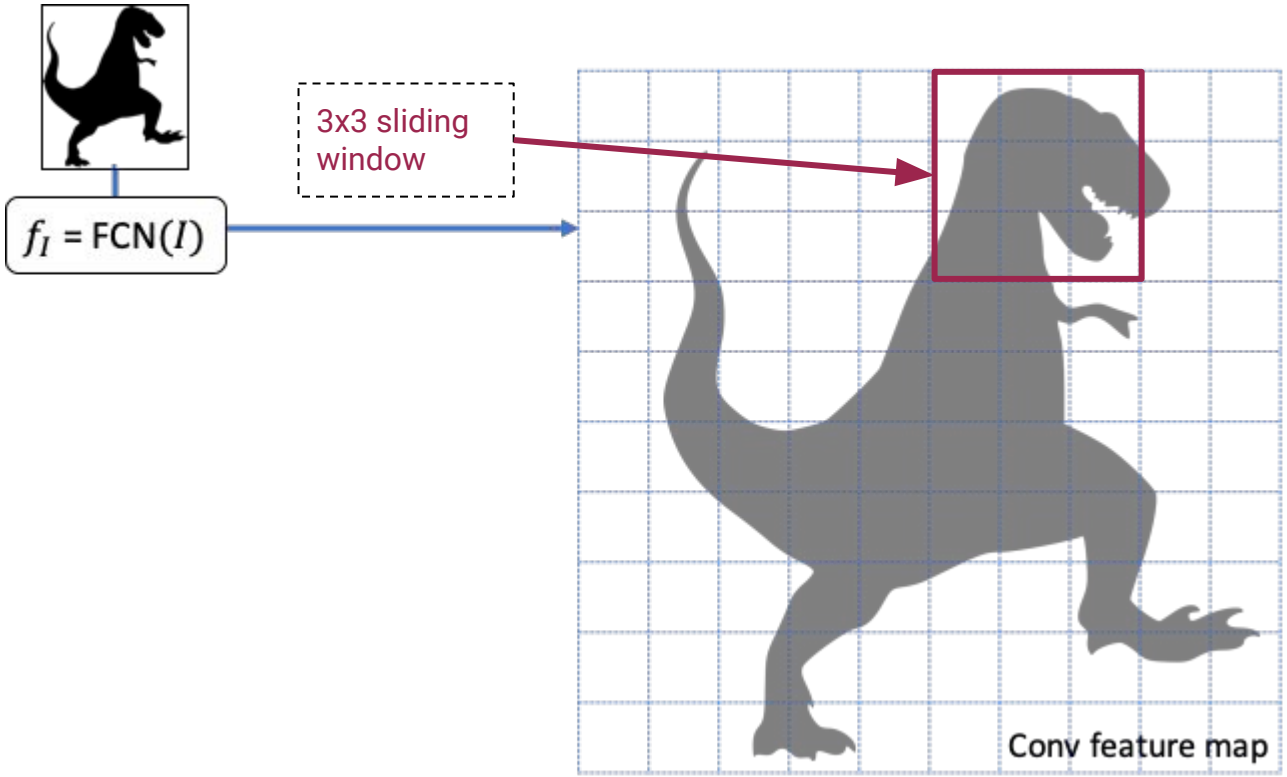
Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Region Proposal Network



Slides modified from Ross Girshick tutorial at CVPR 2019

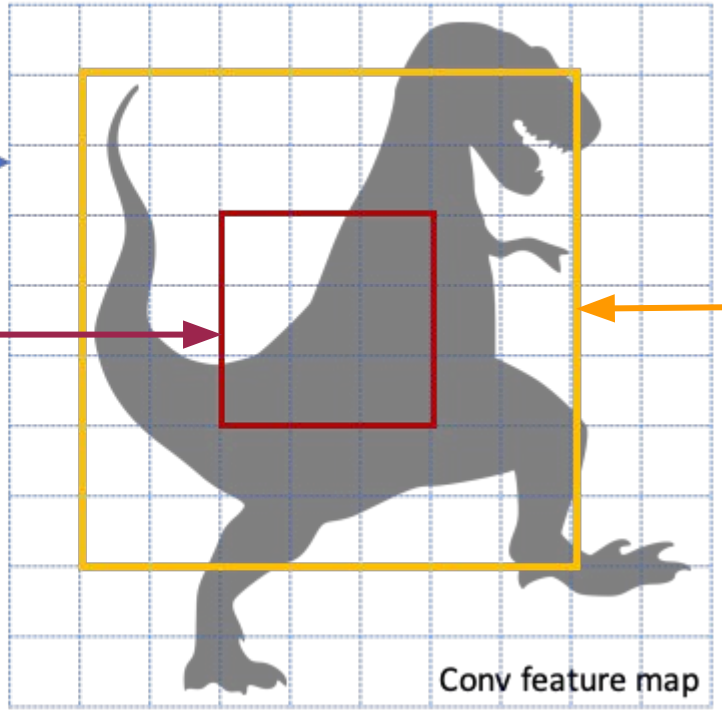
RPN: Region Proposal Network



Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Anchor Box

$f_I = \text{FCN}(I)$



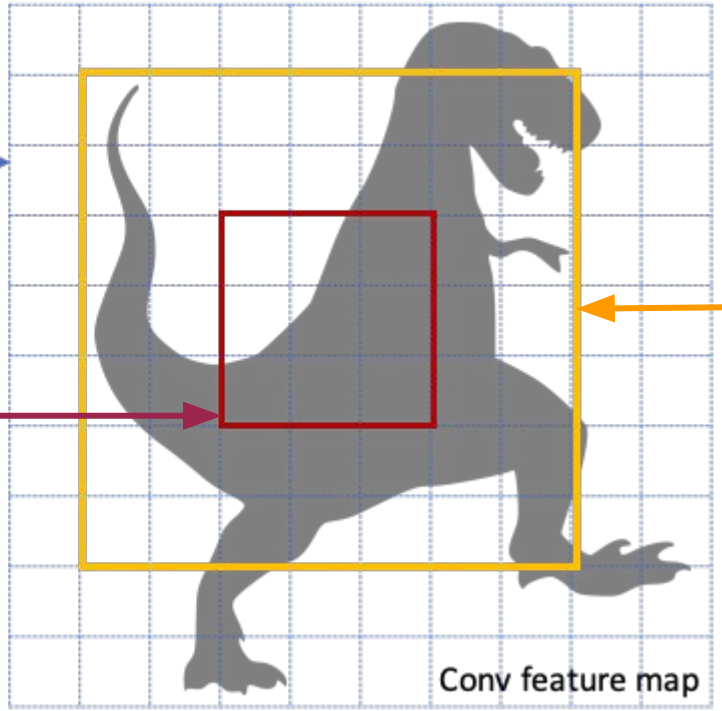
3x3 sliding window
But what's the meaning of this feature?

For this 3x3 window, we predefine a 7x7 "anchor box".

Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Anchor Box

$f_I = \text{FCN}(I)$



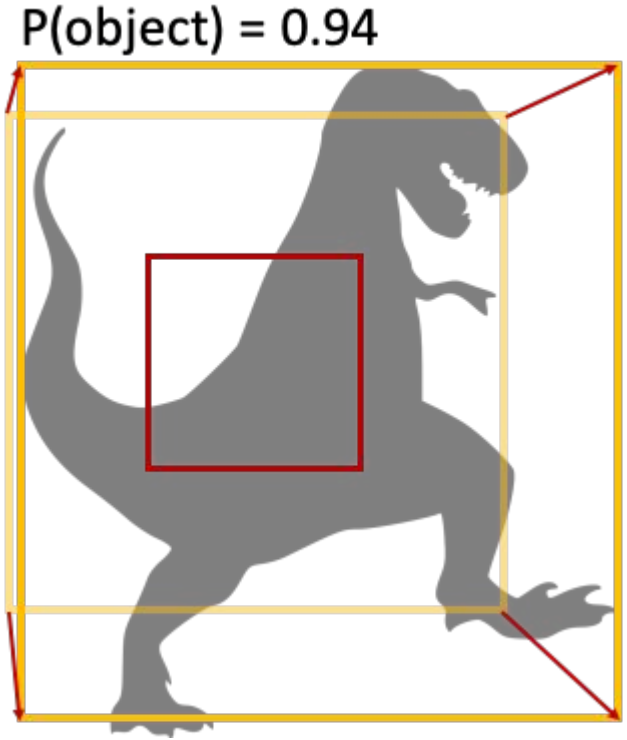
3x3 sliding window
But what's the meaning of this feature?

- Is the anchor box an object?
- Where is the accurate box?

For this 3x3 window, we predefine a 7x7 "anchor box".

Slides modified from Ross Girshick tutorial at CVPR 2019

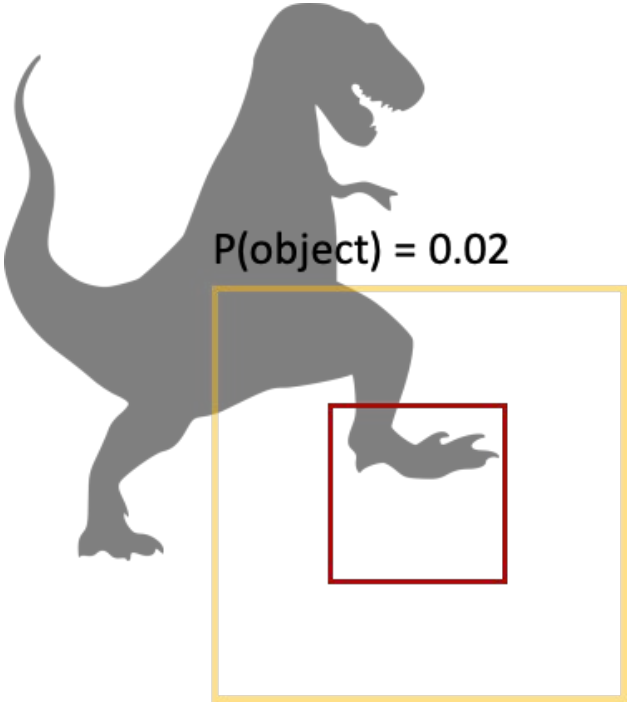
RPN: Prediction (on object)



Direction to the accurate box

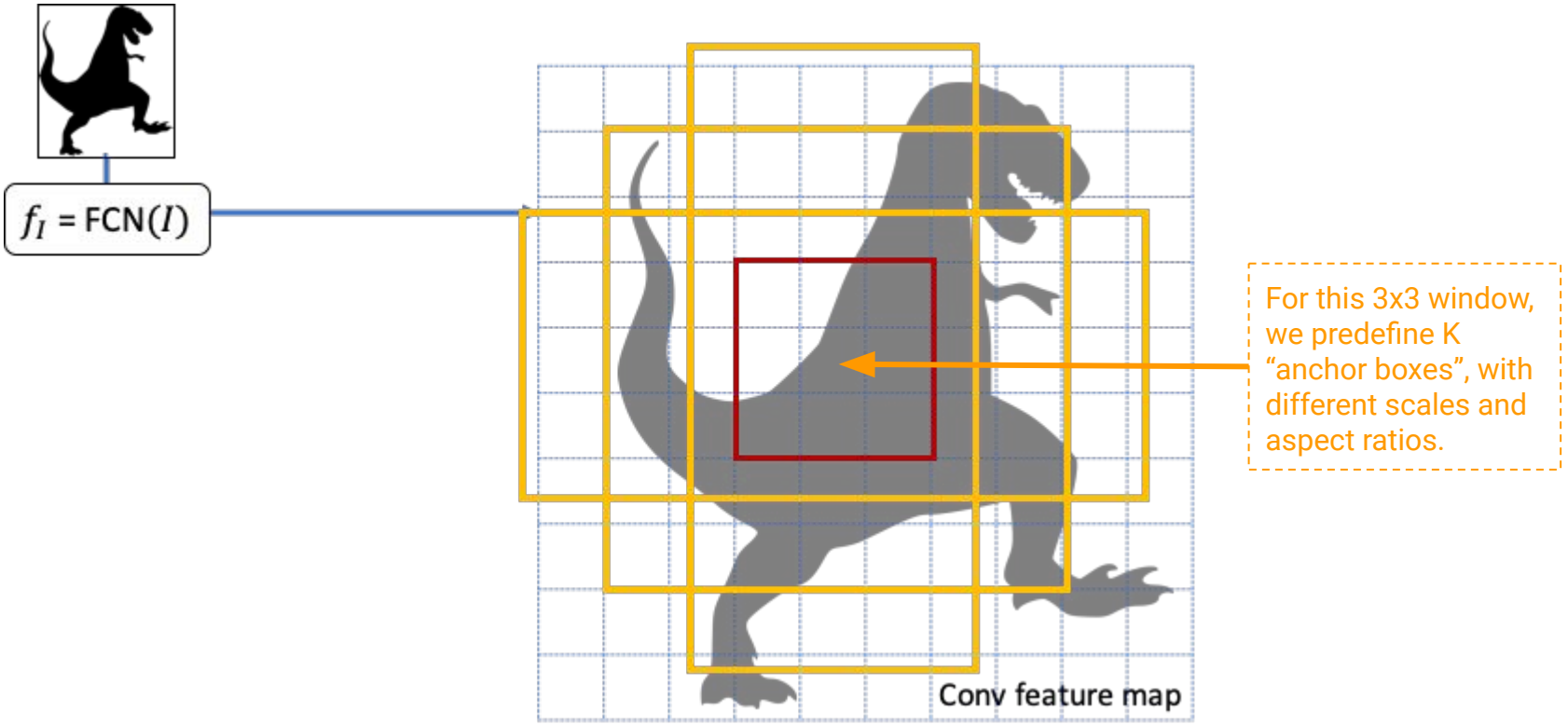
Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Prediction (off object)



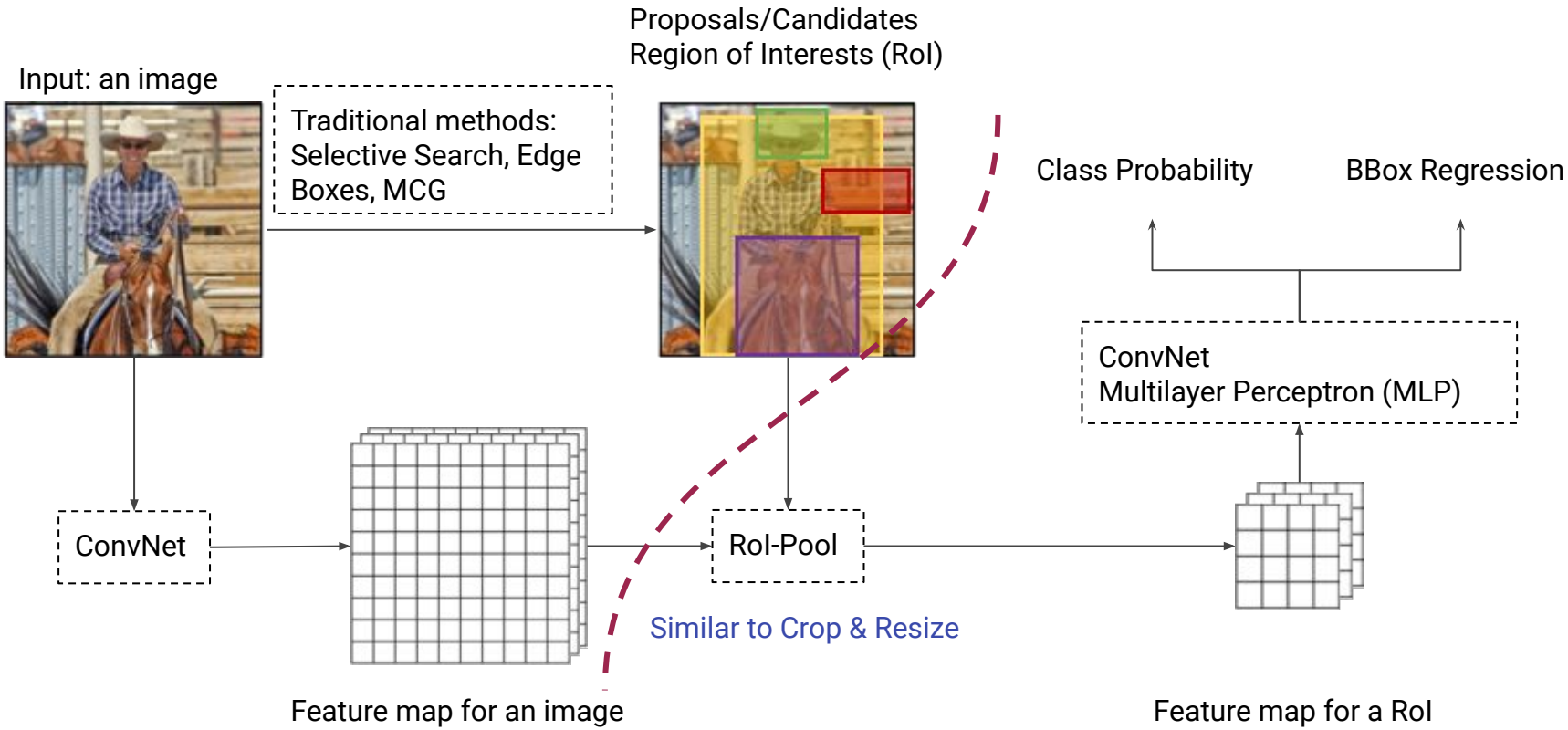
Slides modified from Ross Girshick tutorial at CVPR 2019

RPN: Multiple Anchors

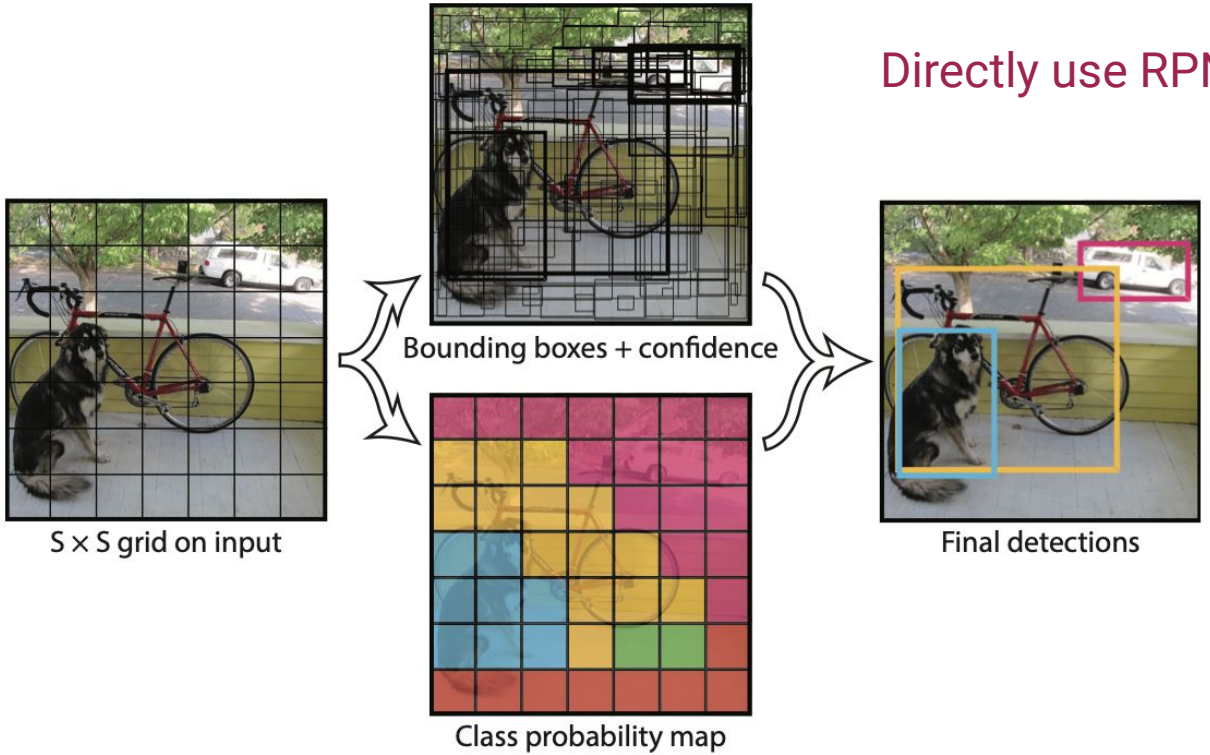


Slides modified from Ross Girshick tutorial at CVPR 2019

Two stages or one stage?



You Look Only Once (YOLO)



Directly use RPN!

Figure from You Only Look Once: Unified, Real-Time Object Detection

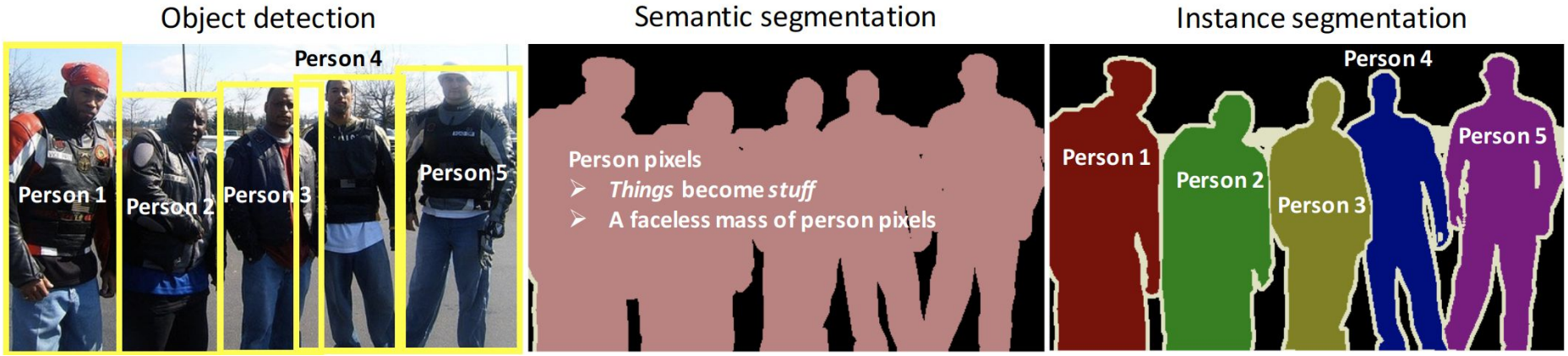
Other Methods

http://web.stanford.edu/class/cs231a/lectures/lecture12_2D_detection.pdf

- VJ Face
- Deformable Part Model
- Implicit Shape Model

U-Net

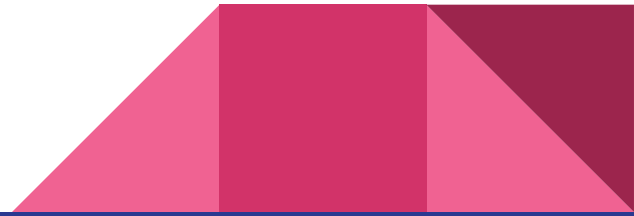
Semantic vs. Instance Segmentation



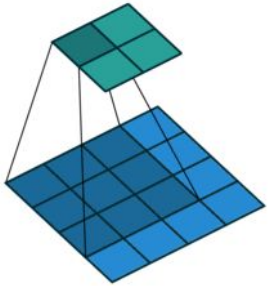
Slides modified from Ross Girshick tutorial at CVPR 2017

How shall we modify CNN?

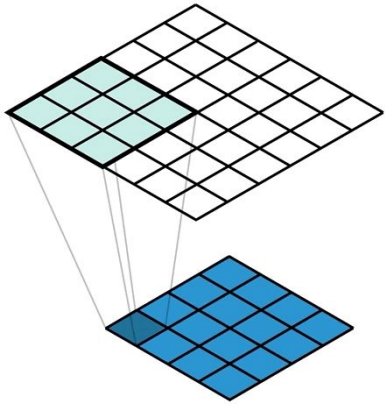
- We need to predict a label for each pixel, but CNN has downsampled our input to a very small scale (e.g. from 224x224 to 7x7).
- Thus, we need a layer to upsample feature maps to the original size.



Transposed Convolution



3x3 Convolution



3x3 Transposed Convolution

http://deeplearning.net/software/theano/tutorial/conv_arithmetic.html
https://medium.com/apache-mxnet/transposed-convolutions-explained-with-ms-excel-52d13030c7e8

Encoder-Decoder

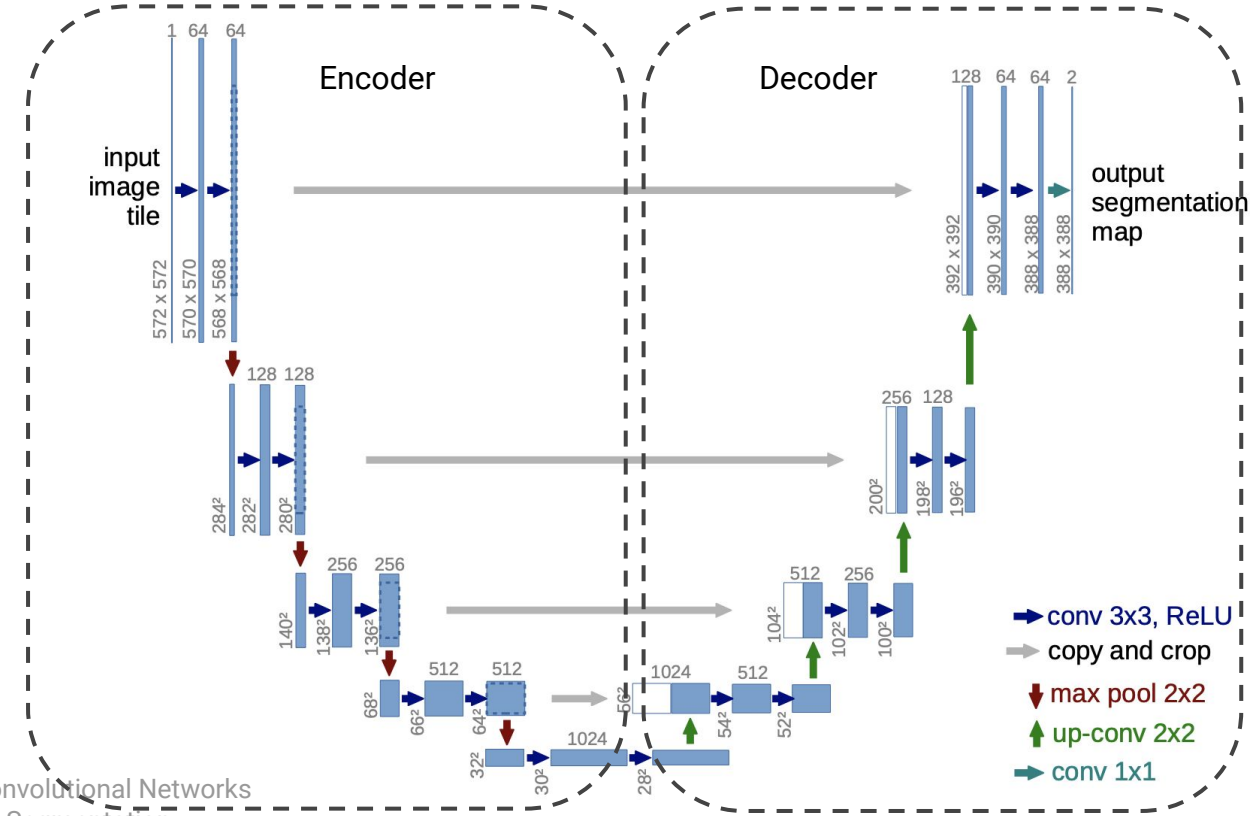


Figure from U-Net: Convolutional Networks for Biomedical Image Segmentation

Transposed Convolution

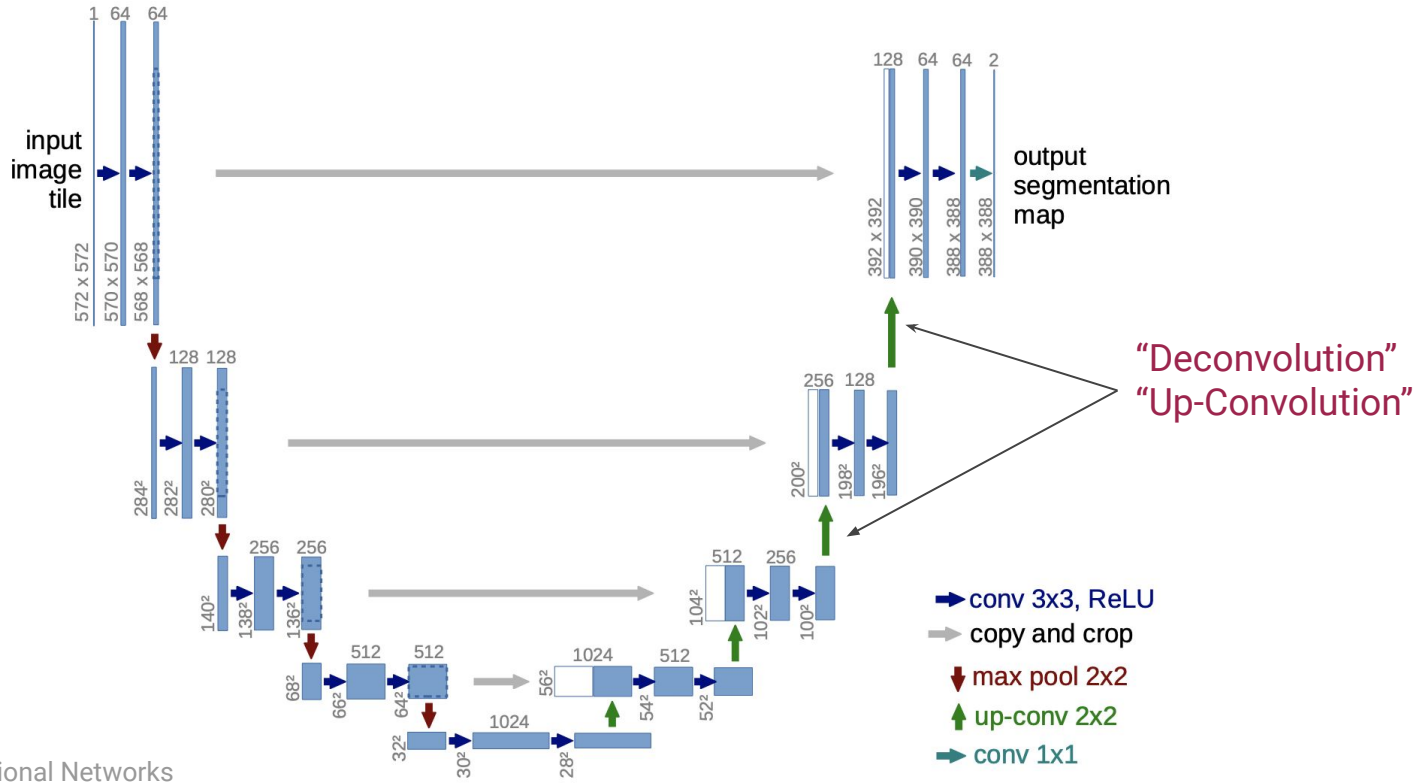


Figure from U-Net: Convolutional Networks for Biomedical Image Segmentation

Skip Layer

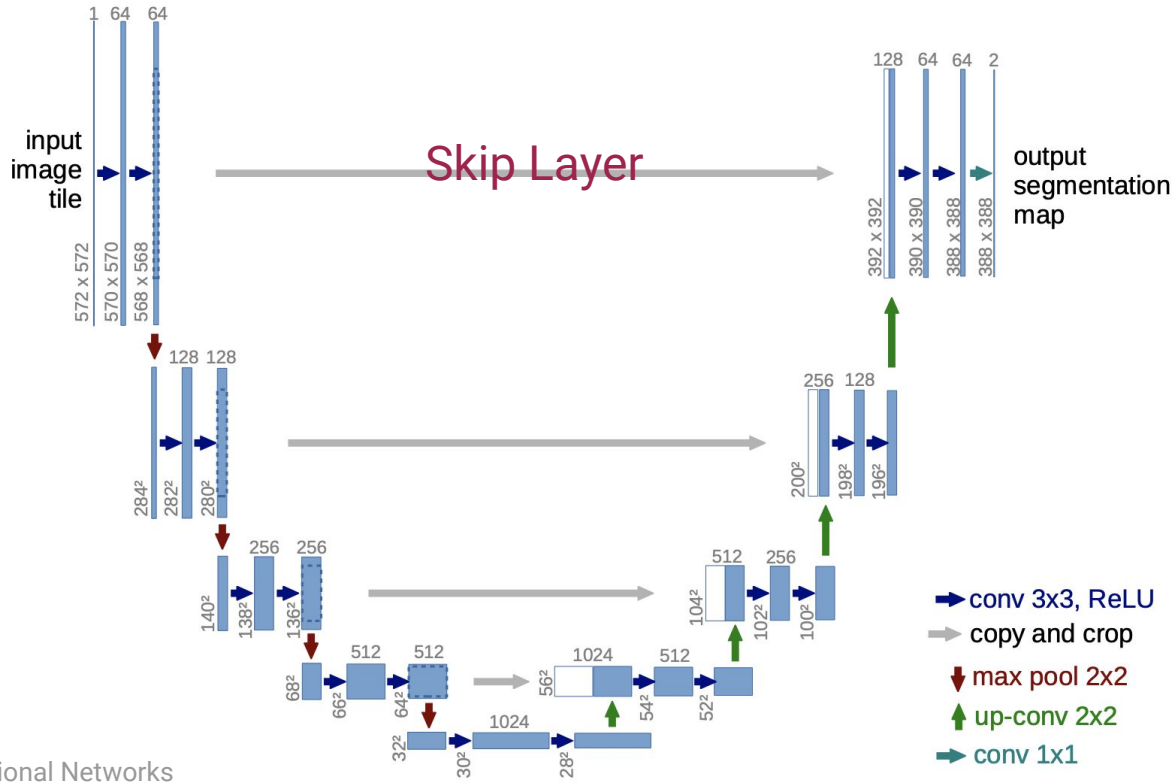


Figure from U-Net: Convolutional Networks for Biomedical Image Segmentation

References

[1] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” pp. 1–9, 2015.

[2] R. Girshick, “Fast R-CNN,” *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 11-18-Dece, pp. 1440–1448, 2016.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pp. 580–587, 2014.

[4] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.

Thanks