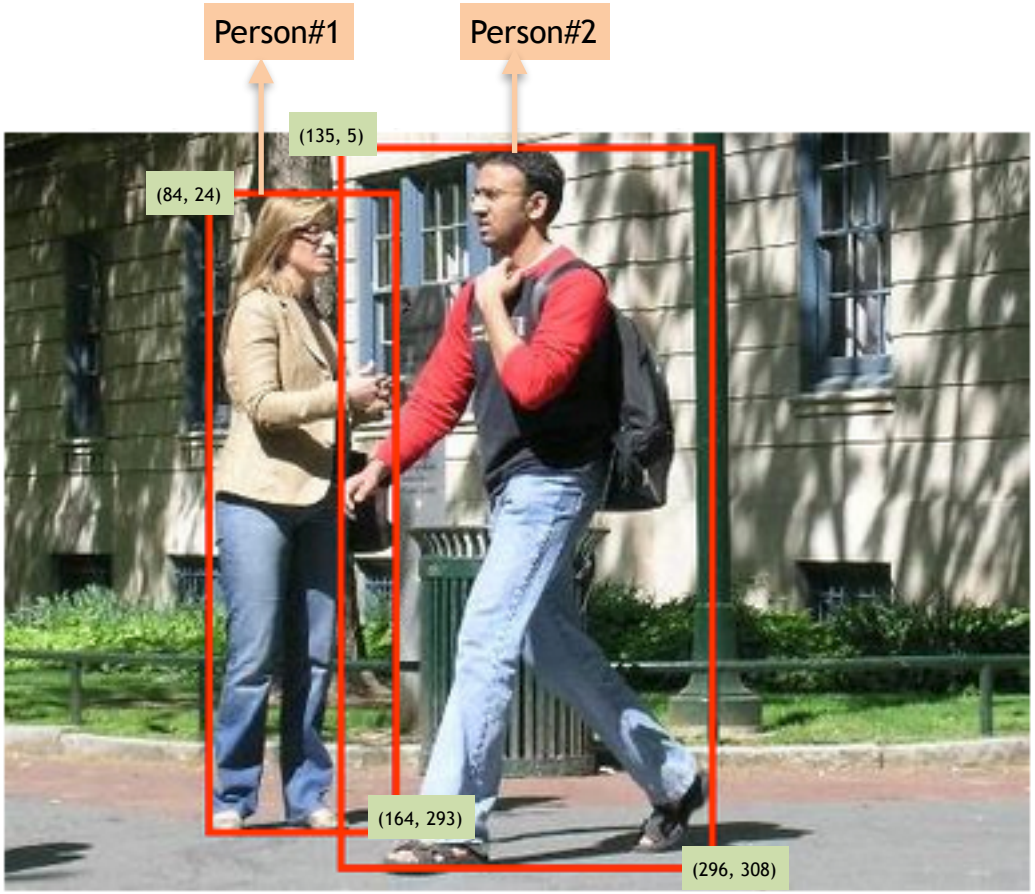
CS195: Computer Vision

Deep Learning-based Object Detection

Wednesday, October 2nd, 2024

Drake UNIVERSITY

Object Detection: what are the locations of objects?



[xmin, ymin, xmax, ymax]

Label	Probs.	Bounding Box Coordinates
Person#1	0.99	[84, 24, 164, 293]
Person#2	0.99	[139, 5, 296, 308]
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…
…	…	…

Input:
1. an image

Detection model

Output:
1. Labels of prediction
2. Probabilities of labels
3. Locations of rectangular bounding boxes associated with each label

Deep learning-based object detections

- **R-CNN**
- Fast RCNN
- Faster RCNN

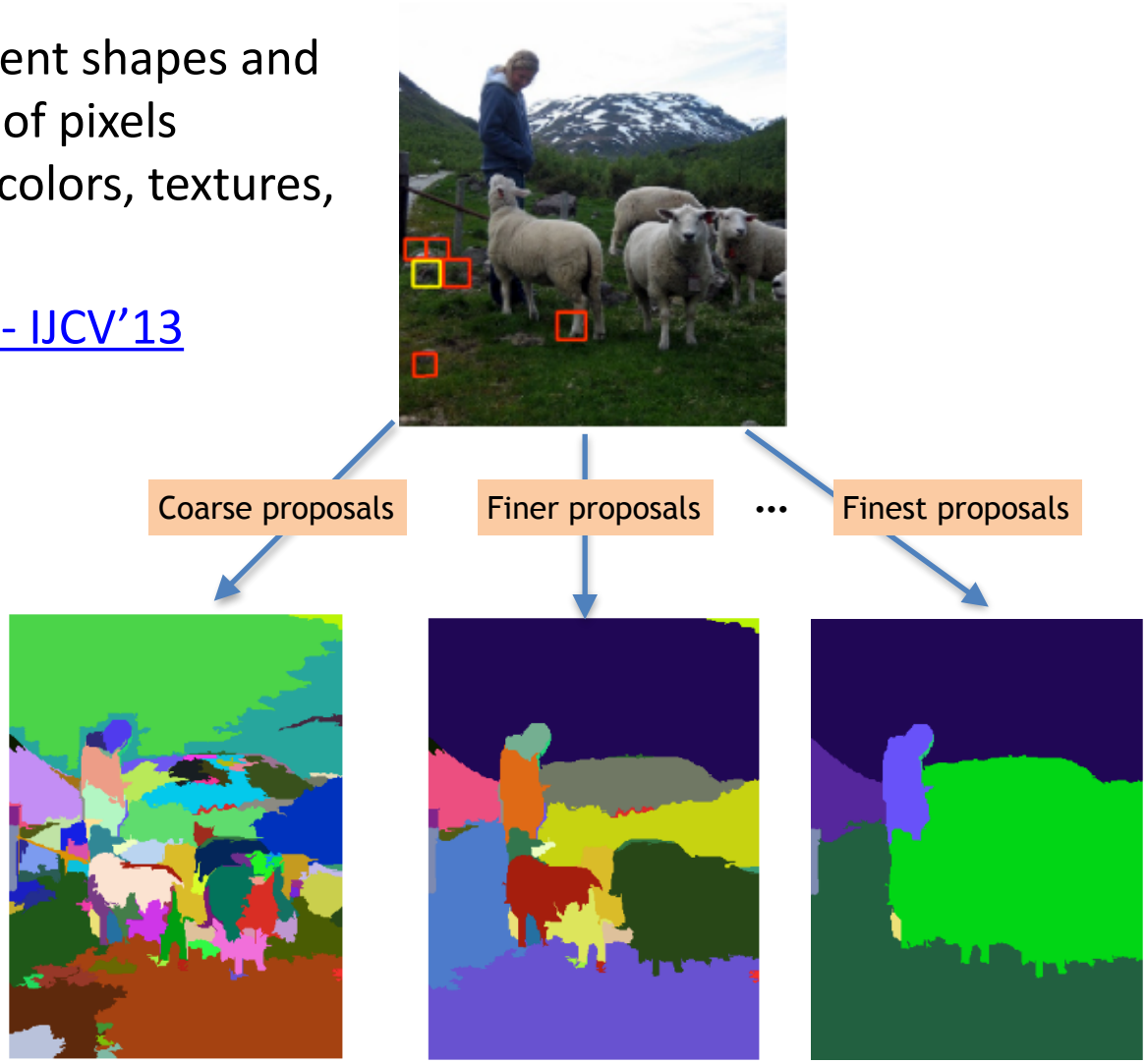
- Mask RCNN

- YOLO
- SSD

R-CNN

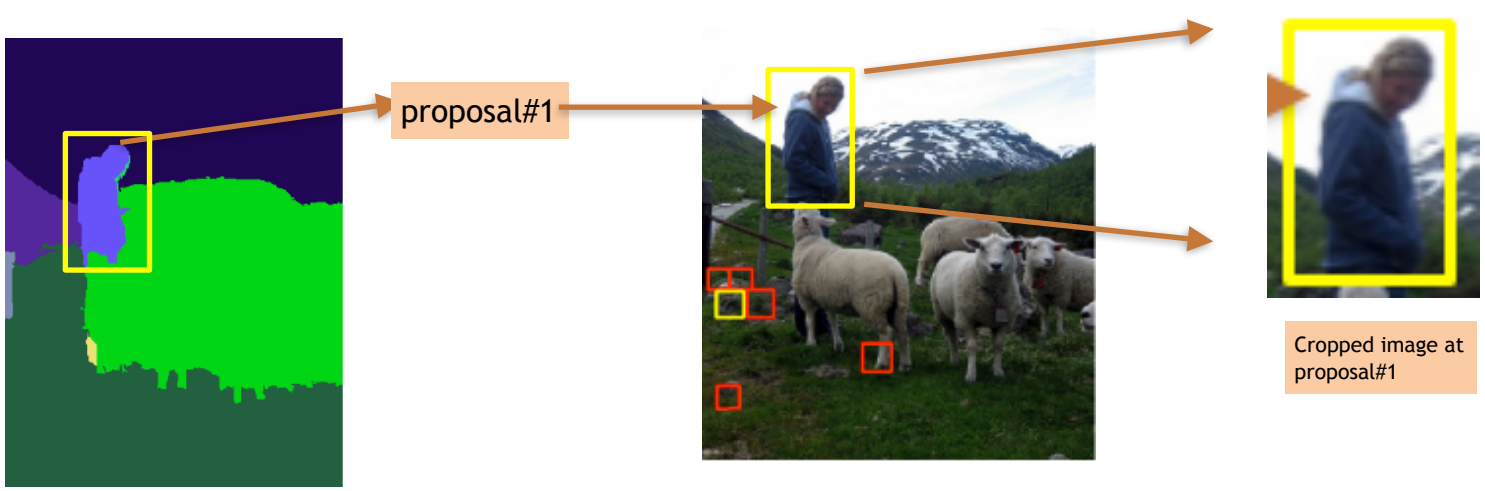
- Generate proposals of different shapes and sizes based on the grouping of pixels together by looking at their colors, textures, and other features

  Reference: Selective Search - IJCV'13



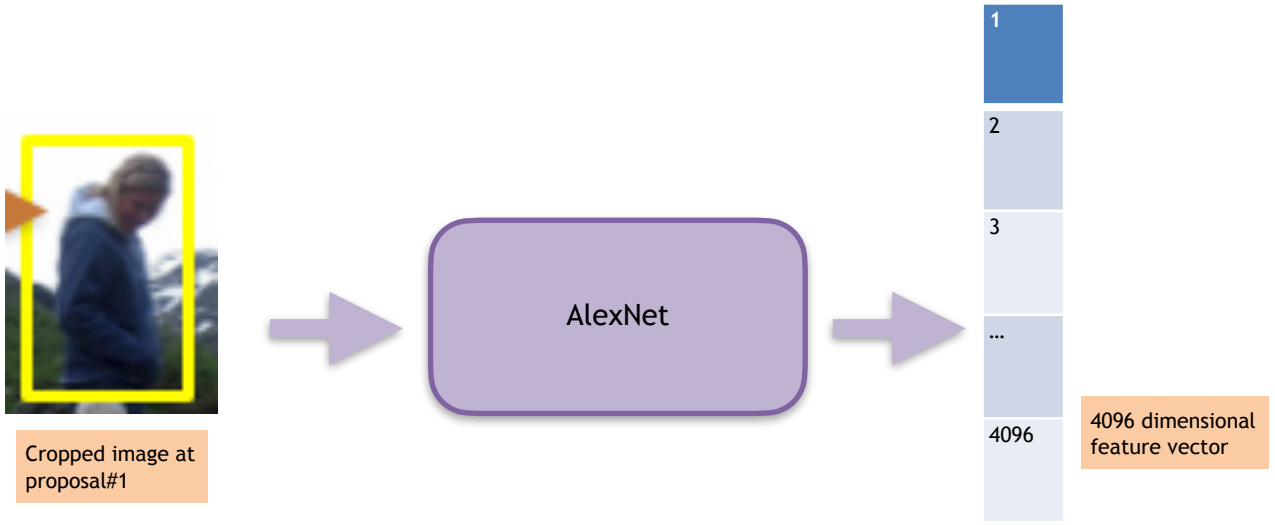
R-CNN

- Extract features for each proposal
  - 2000 proposals of various sizes
  - 4096 dimensional vector from AlexNet's last layer for each proposal



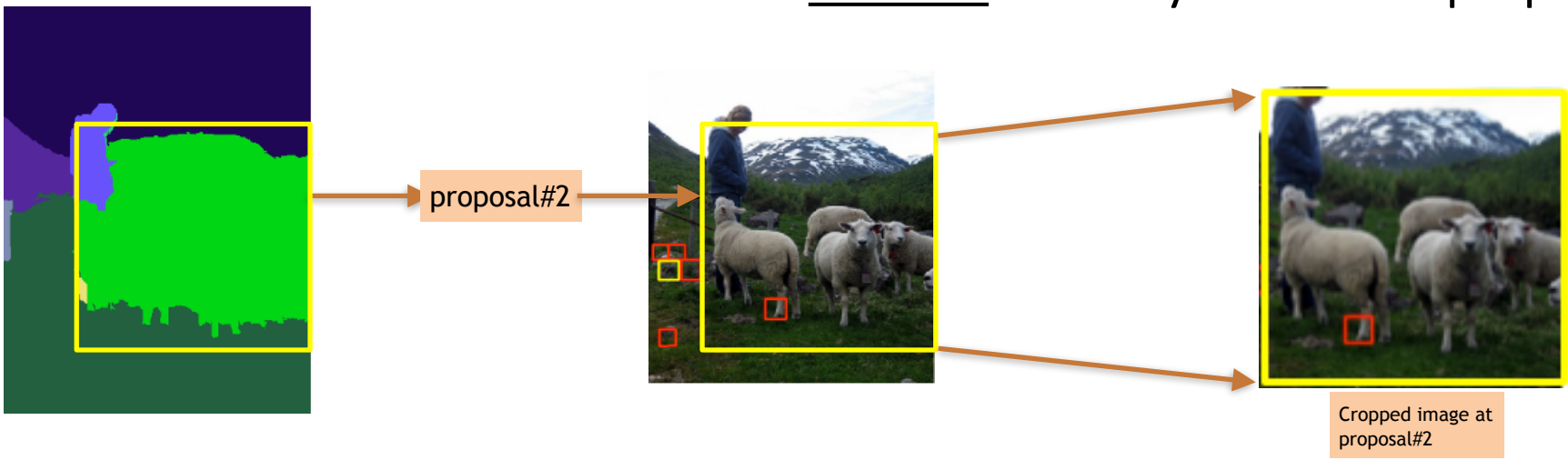
R-CNN

- Extract features for each proposal
  - 2000 proposals of various sizes
  - 4096 dimensional vector from AlexNet's last layer for each proposal



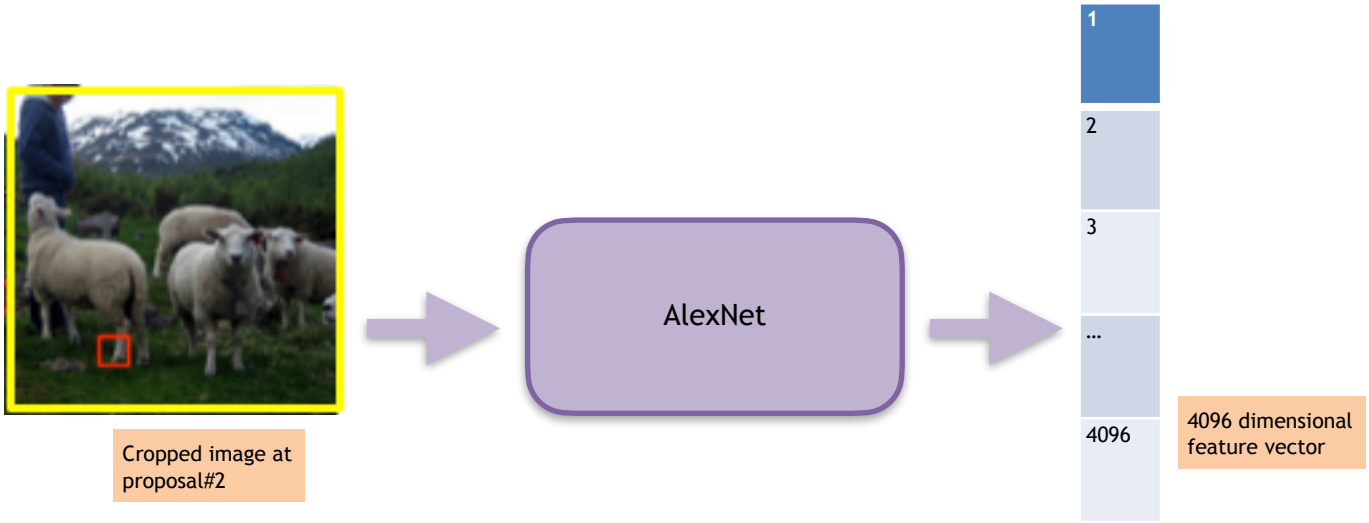
R-CNN

- Extract features for each proposal
  - 2000 proposals of various sizes
  - 4096 dimensional vector from AlexNet's last layer for each proposal



R-CNN

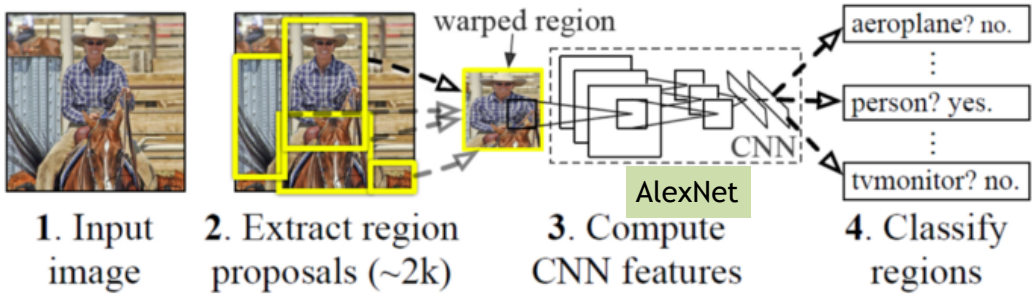
- Extract features for each proposal
  - 2000 proposals of various sizes
  - 4096 dimensional vector from AlexNet's last layer for each proposal



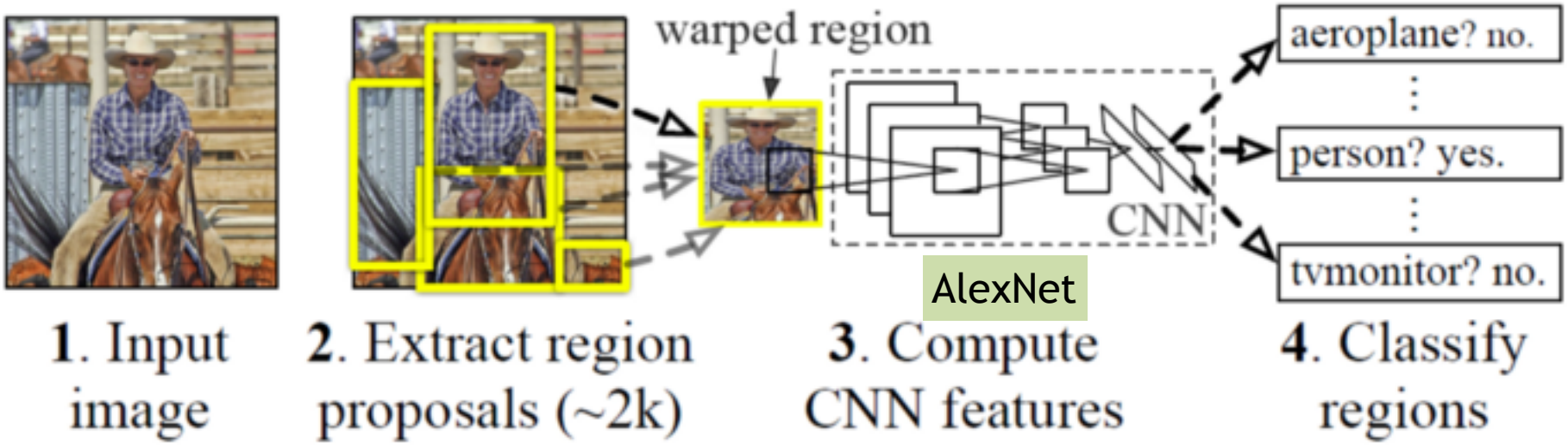
R-CNN

- Generate proposals
- Extract features for each proposal
  - 2000 proposals of various sizes
  - 4096 dimensional vector from AlexNet's last layer for each proposal
- Classify the it into one of the classes using SVM
- Caveat:
  - very slow (low-throughput)
  - Not end-to-end trainable

Rich feature hierarchies for accurate object detection and semantic segmentation - Girshick et al. CVPR'14

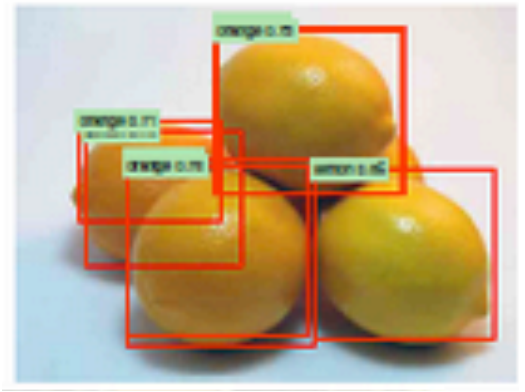


R-CNN



Rich feature hierarchies for accurate object detection and semantic segmentation - Girshick et al. CVPR'14

R-CNN



Rich feature hierarchies for accurate object detection and semantic segmentation - Girshick et al. CVPR'14

Deep learning-based object detections

- R-CNN
- **Fast RCNN**
- Faster RCNN

- Mask RCNN

- YOLO
- SSD

Why it is called “Fast” R-CNN?

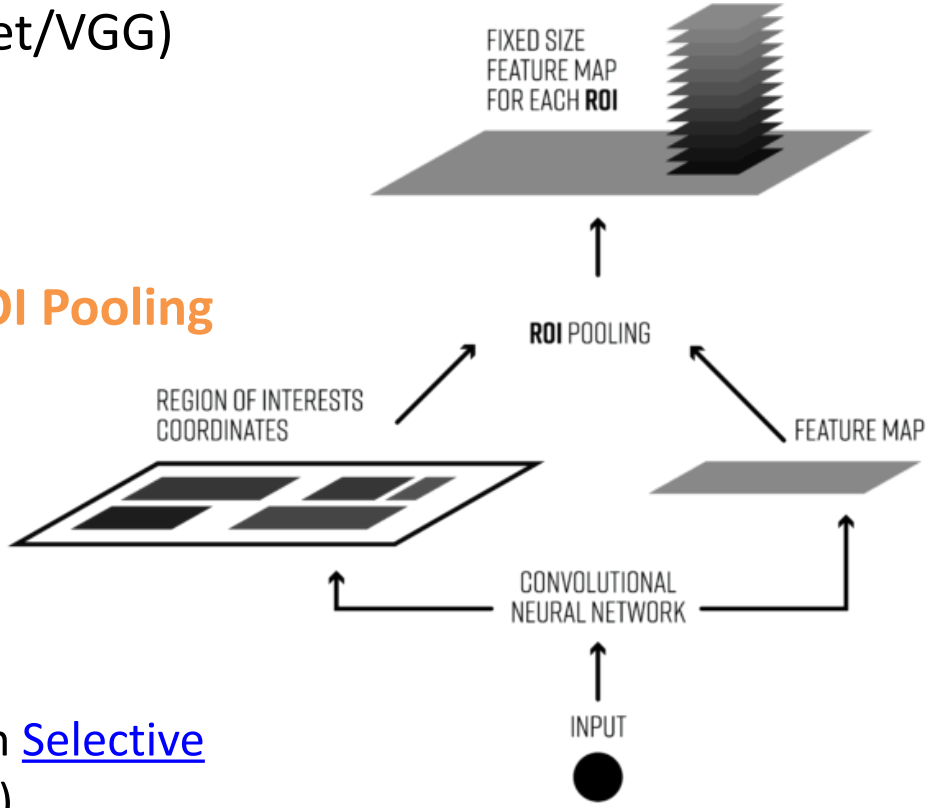
- **Fast R-CNN** is faster than **R-CNN** because we don’t need to feed 2000 region proposals to the convolutional neural network separately

- Instead, for each image, the convolution operation is done only once, and **only one** feature map is generated from it

Fast R-CNN - Girshick et al. ICCV'15

Fast R-CNN

- Feed the image through a CNN (AlexNet/VGG)
- Extract fixed-size feature map using **ROI Pooling**
  - for different Region Of Interests (ROI)
- End-to-end trainable:
  - except proposals generation part from Selective Search - IJCV'13 (R-CNN does that too)



Fast R-CNN - Girshick et al. ICCV'15

Deep learning-based object detections

- R-CNN
- Fast RCNN
- **Faster RCNN**

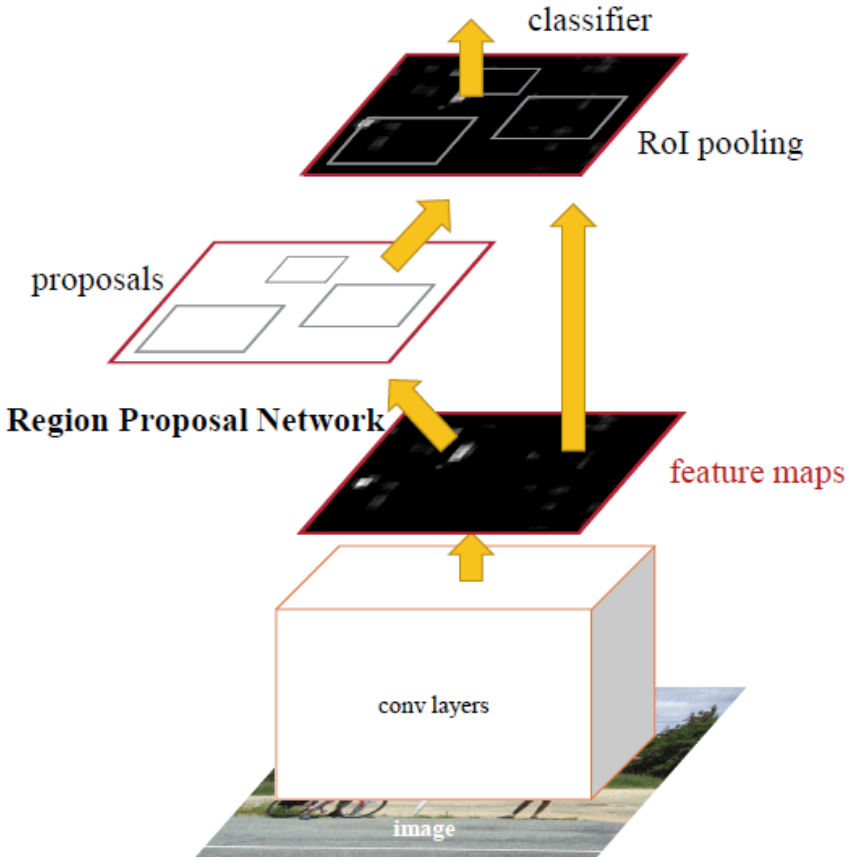
- Mask RCNN

- YOLO
- SSD

Faster R-CNN

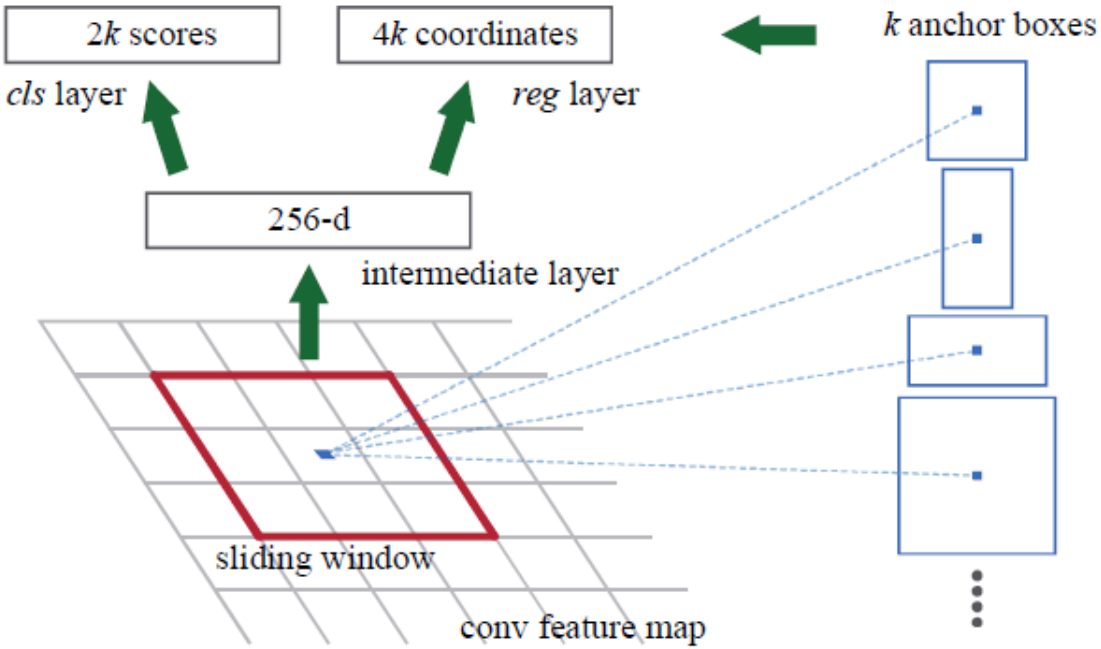
- Introduce a new network which will learn to generate good proposals. It is called **Region Proposal Network (RPN)** which will generate proposals.
- For each proposal, extract fixed-size feature map using **ROI Pooling**
- The network is end-to-end trainable:
  - Unlike R-CNN where different components are trained separately

[Faster R-CNN - Girshick et al. NIPS'15]



Faster R-CNN

- **Region Proposal Network (RPN)**

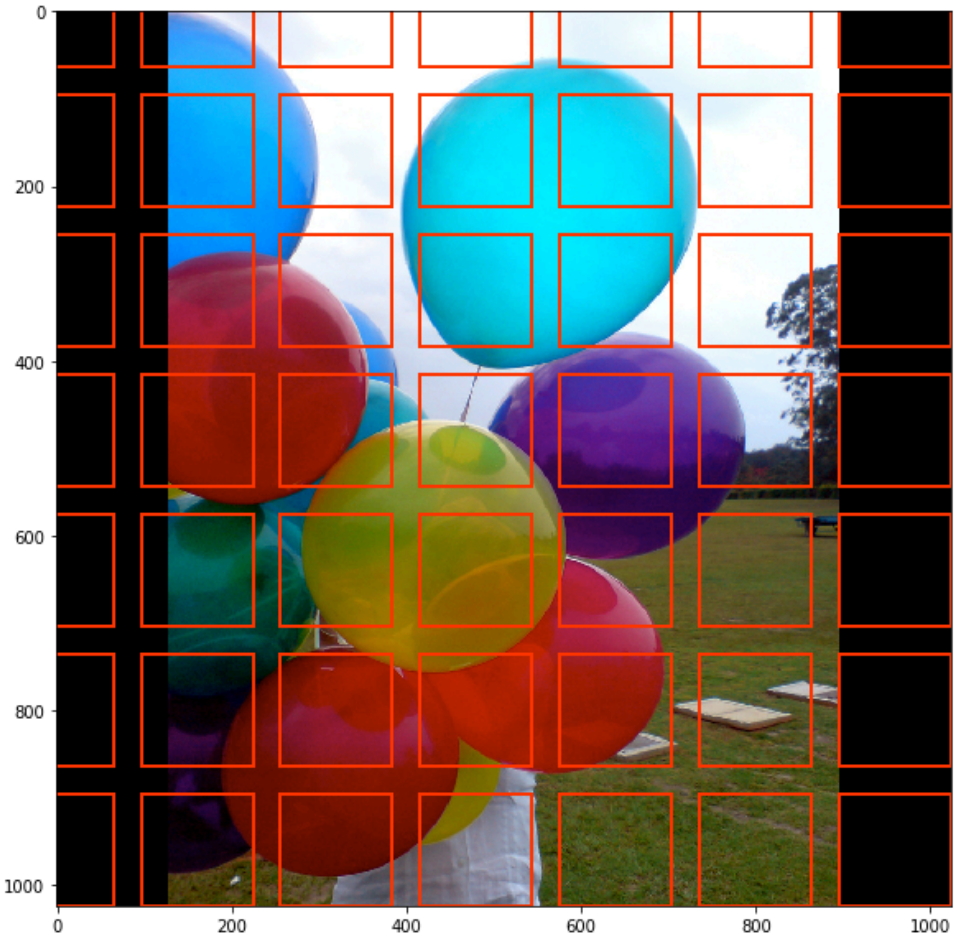


Faster R-CNN - Girshick et al. NIPS'15

Faster R-CNN

- **Region Proposal Network (RPN)**

Faster R-CNN - Girshick et al. NIPS'15



State-of-the-art object detection methods

- R-CNN
- Fast RCNN
- Faster RCNN

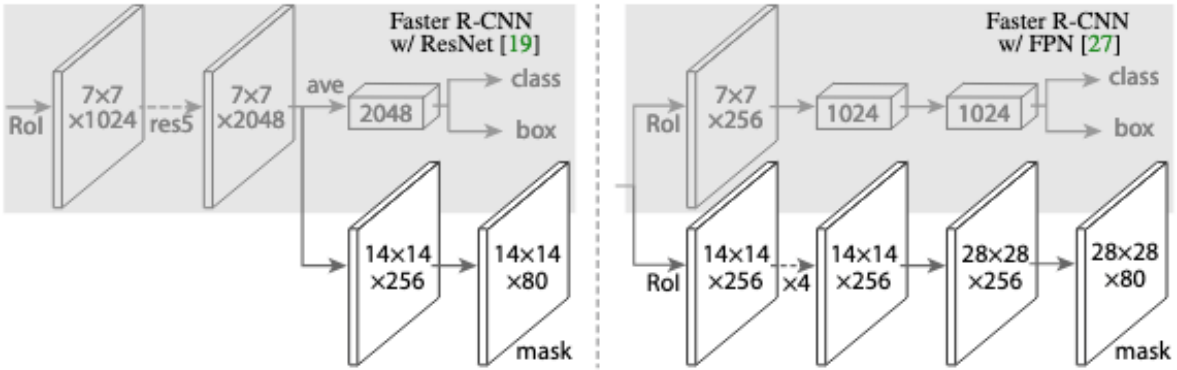
- **Mask RCNN**

- YOLO
- SSD

Mask R-CNN

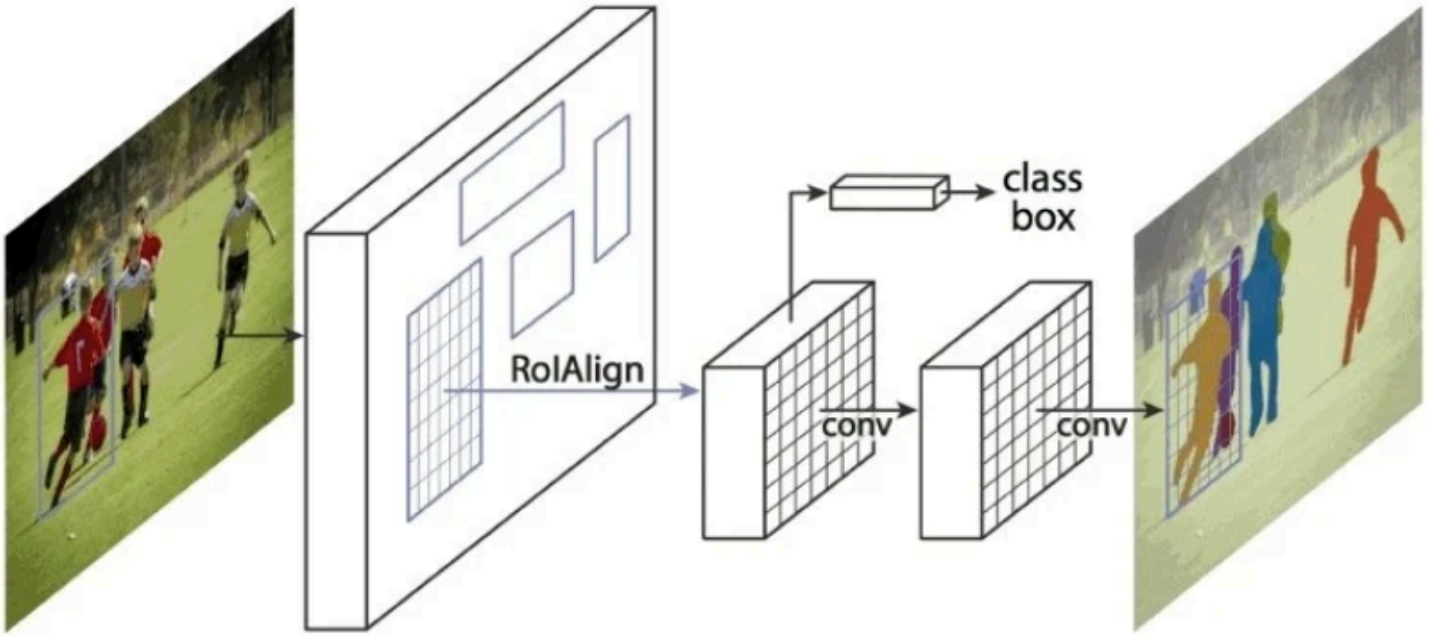
- Built on top of Faster R-CNN
- Faster R-CNN has two outputs for each object
  - a class label and
  - a bounding-box parameters
- Mask R-CNN adds a third branch that outputs the object mask

Mask R-CNN - Girshick et al. CVPR'17



Mask R-CNN

Mask R-CNN - Girshick et al. CVPR'17



Mask R-CNN

Mask R-CNN - Girshick et al. CVPR'17

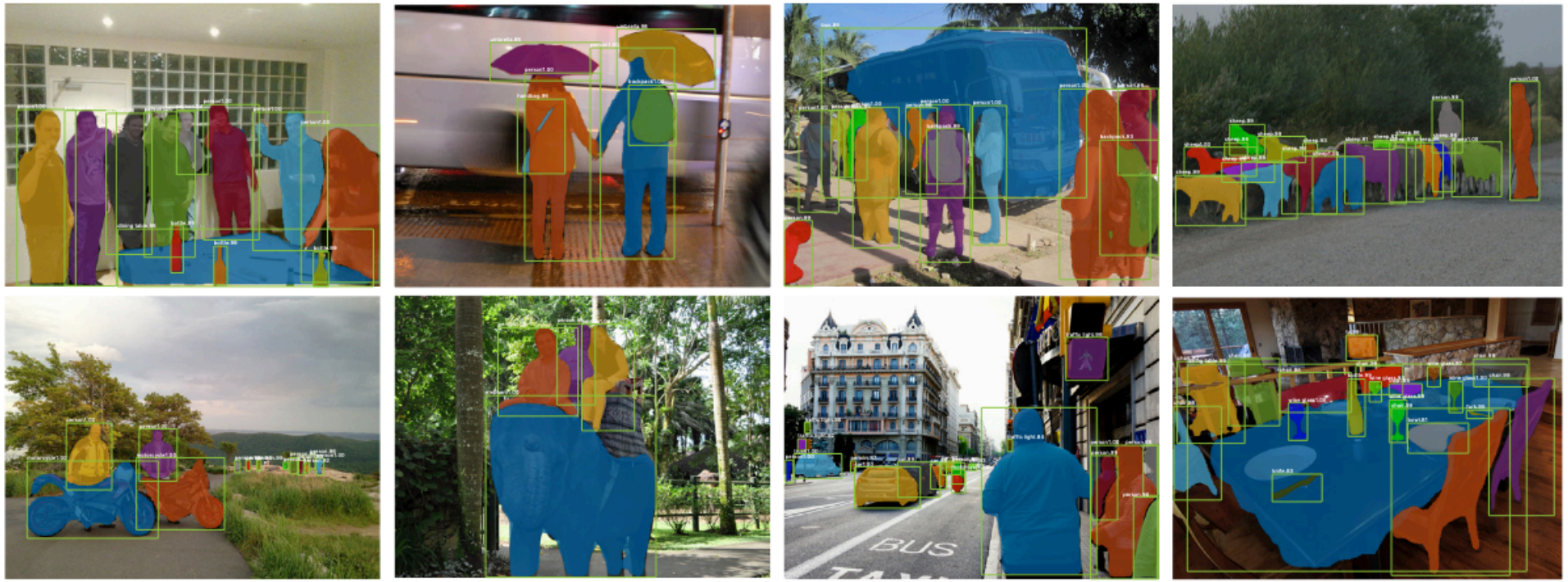


Figure 2. **Mask R-CNN** results on the COCO test set. These results are based on ResNet-101 [19], achieving a *mask* AP of 35.7 and running at 5 fps. Masks are shown in color, and bounding box, category, and confidences are also shown.