YOLO

Real-Time Object Detection

Image Classification

- AlexNet (2012)
- VGG (2014)
- GoogLeNet (2014)
- ResNet (2015)

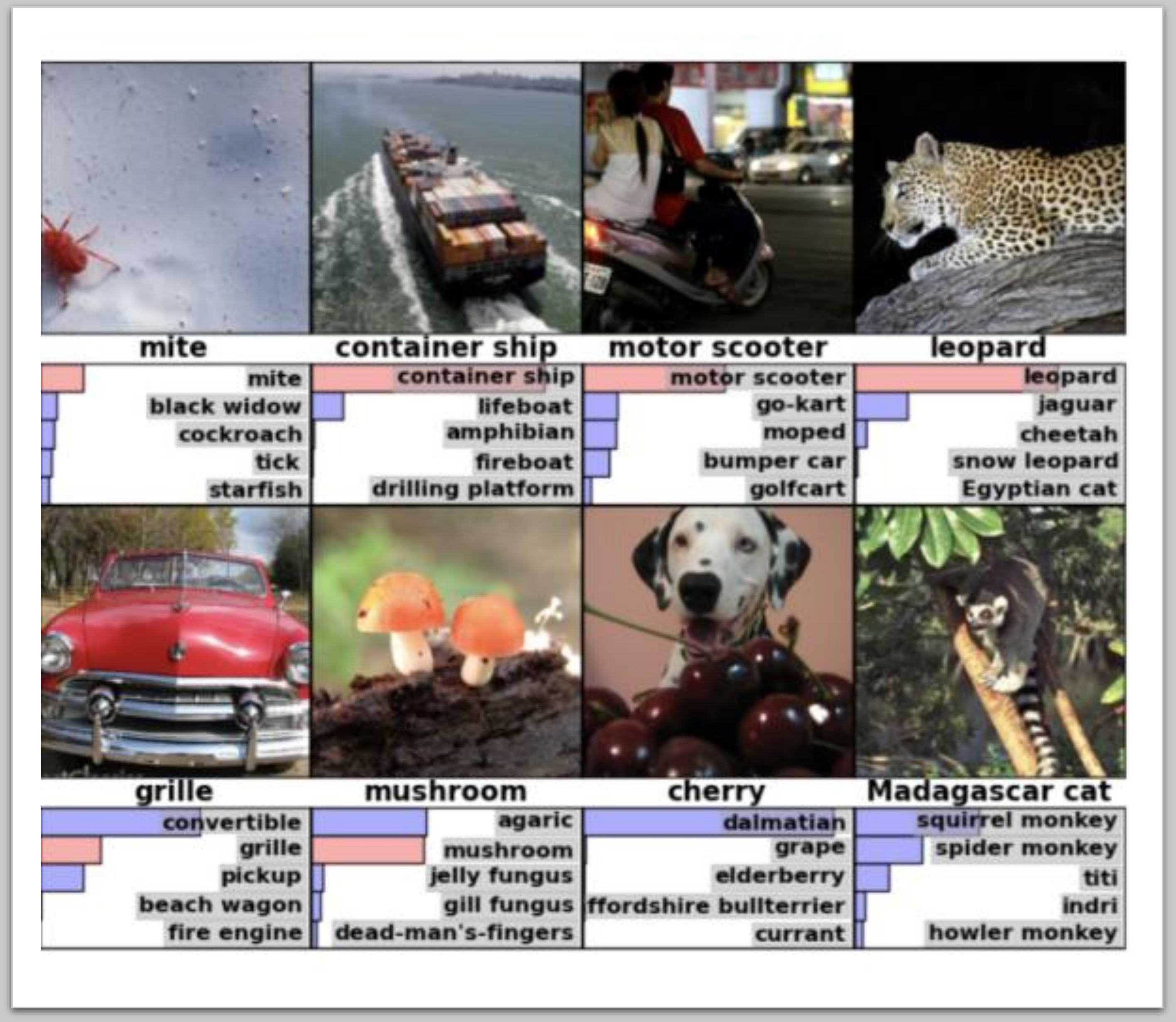
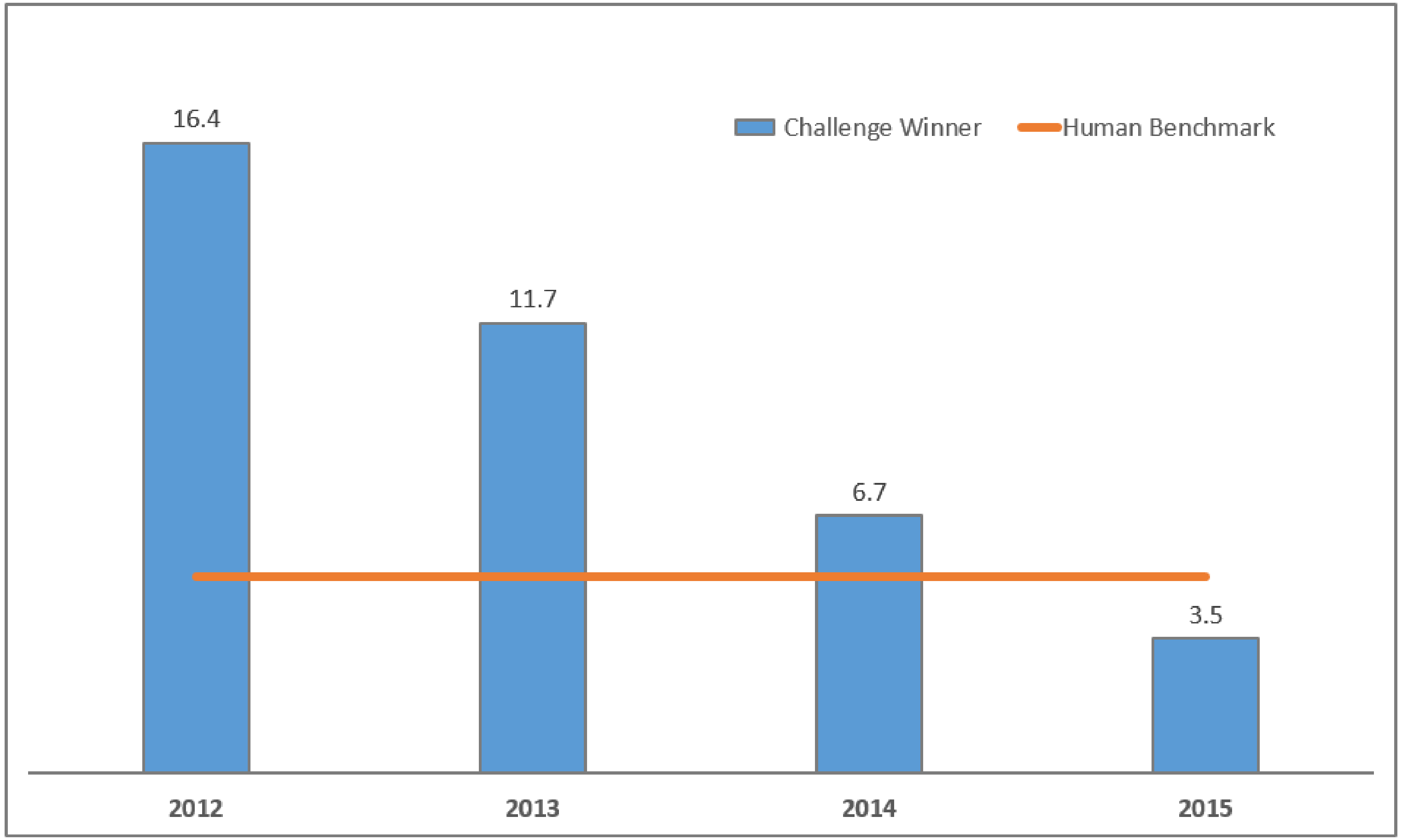
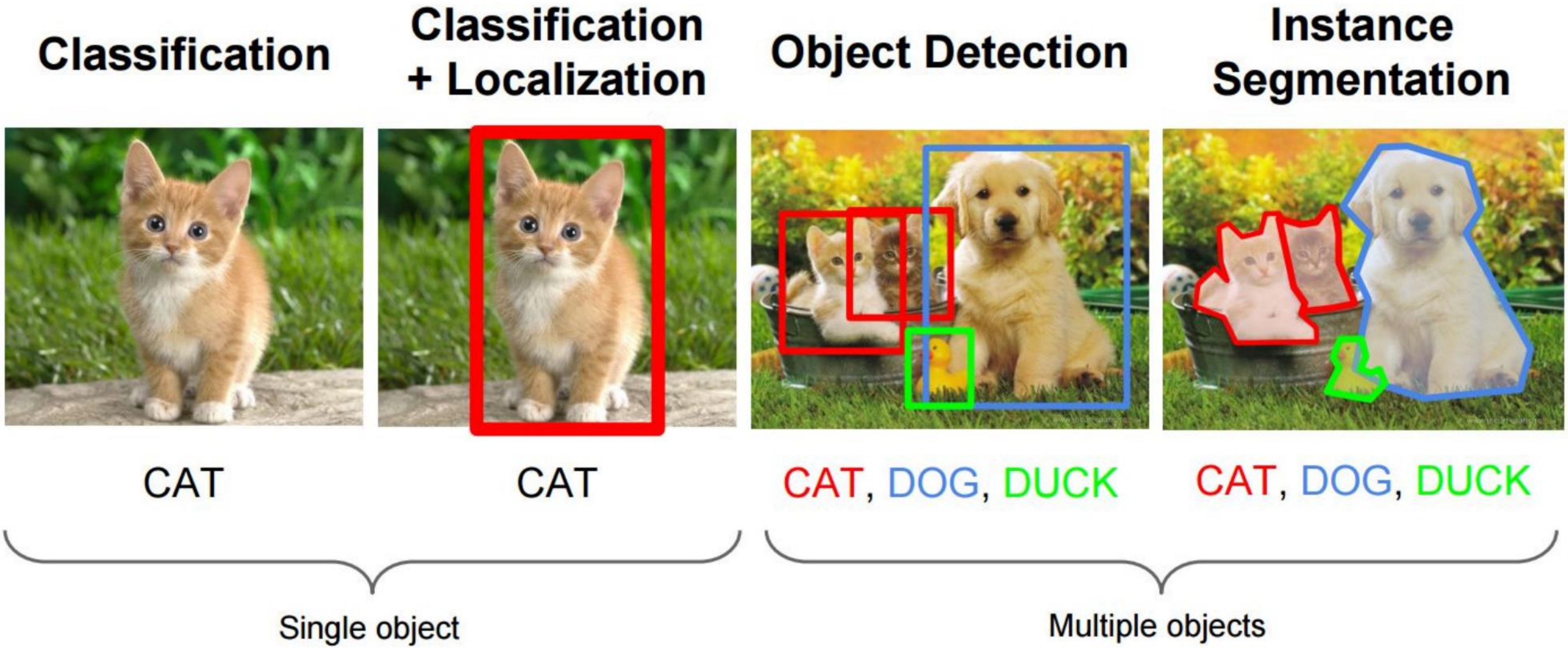


Image Classification



Object Detection



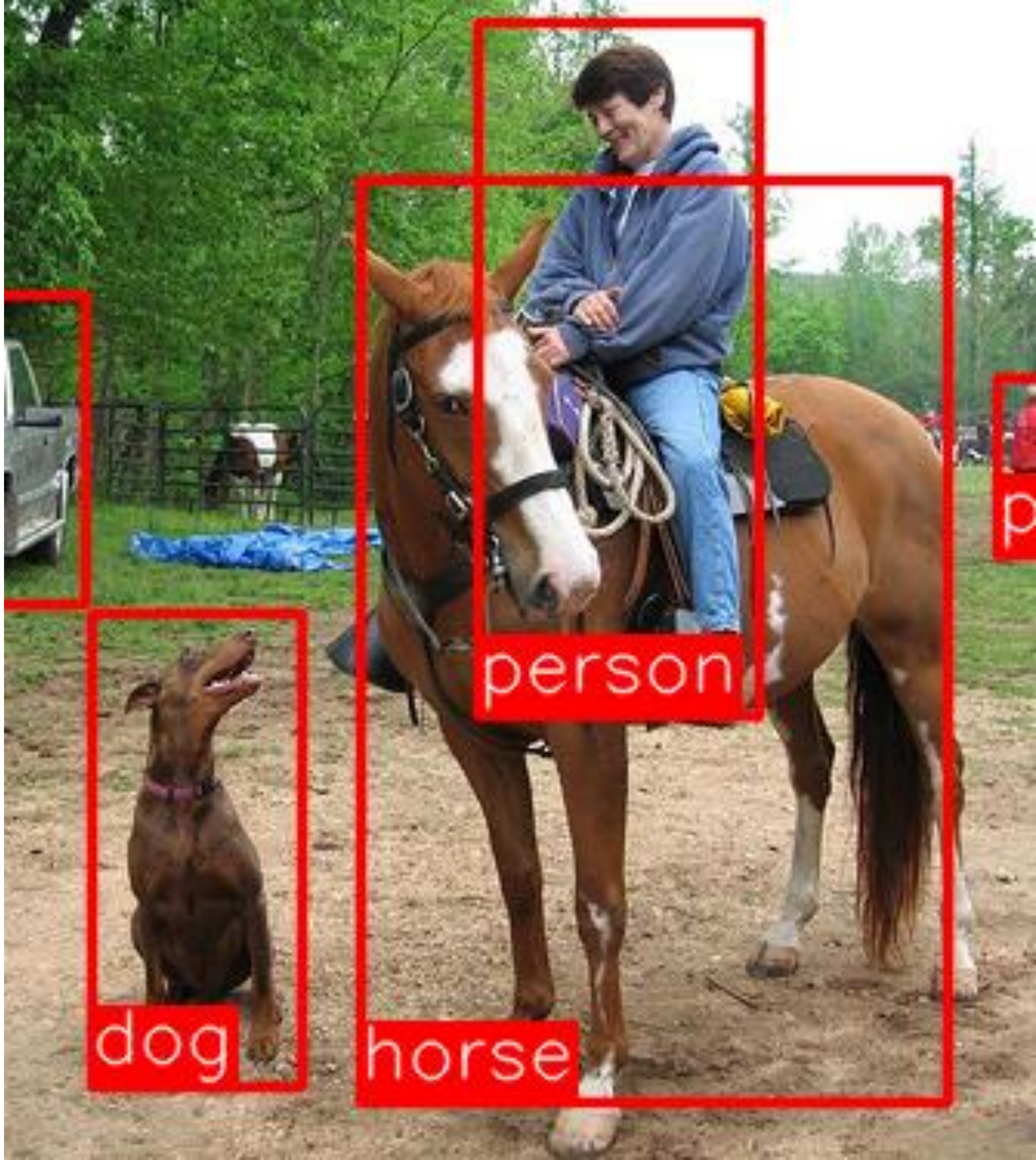
Two-step Detection

- HOG features with sliding windows
- R-CNN: https://arxiv.org/abs/1311.2524
- Fast R-CNN: https://arxiv.org/abs/1504.08083
- Faster R-CNN: https://arxiv.org/abs/1506.01497
- Mask R-CNN: https://arxiv.org/abs/1703.06870

(source: https://blog.athelas.com/a-brief-history-of-cnns-in-image-segmentation-from-r-cnn-to-mask-r-cnn-34ea83205de4)

R-CNN

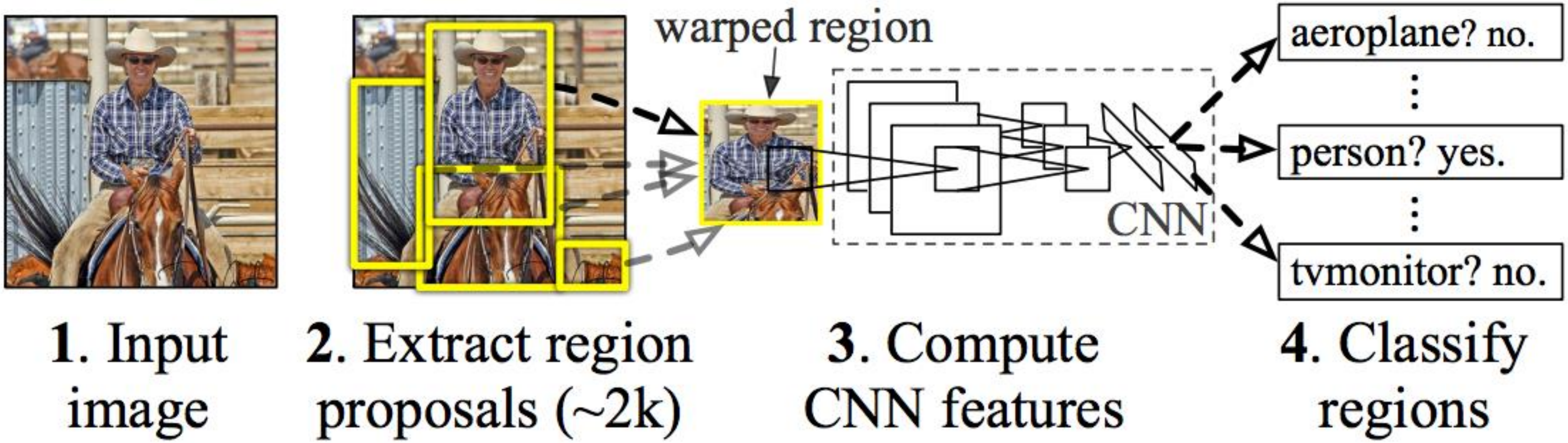
- 2014
- *“This paper is the first to show that a CNN can lead to dramatically higher object detection performance on PASCAL VOC as compared to systems based on simpler HOG-like features.”*



R-CNN

Propose Box	Generate a set of proposals for bounding boxes.
CNN + SVM	Run the images in the bounding boxes through a pre-trained AlexNet and finally an SVM to see what object the image in the box is.
Tighten Box	Run the box through a linear regression model to output tighter coordinates for the box once the object has been classified.

R-CNN: *Regions with CNN features*

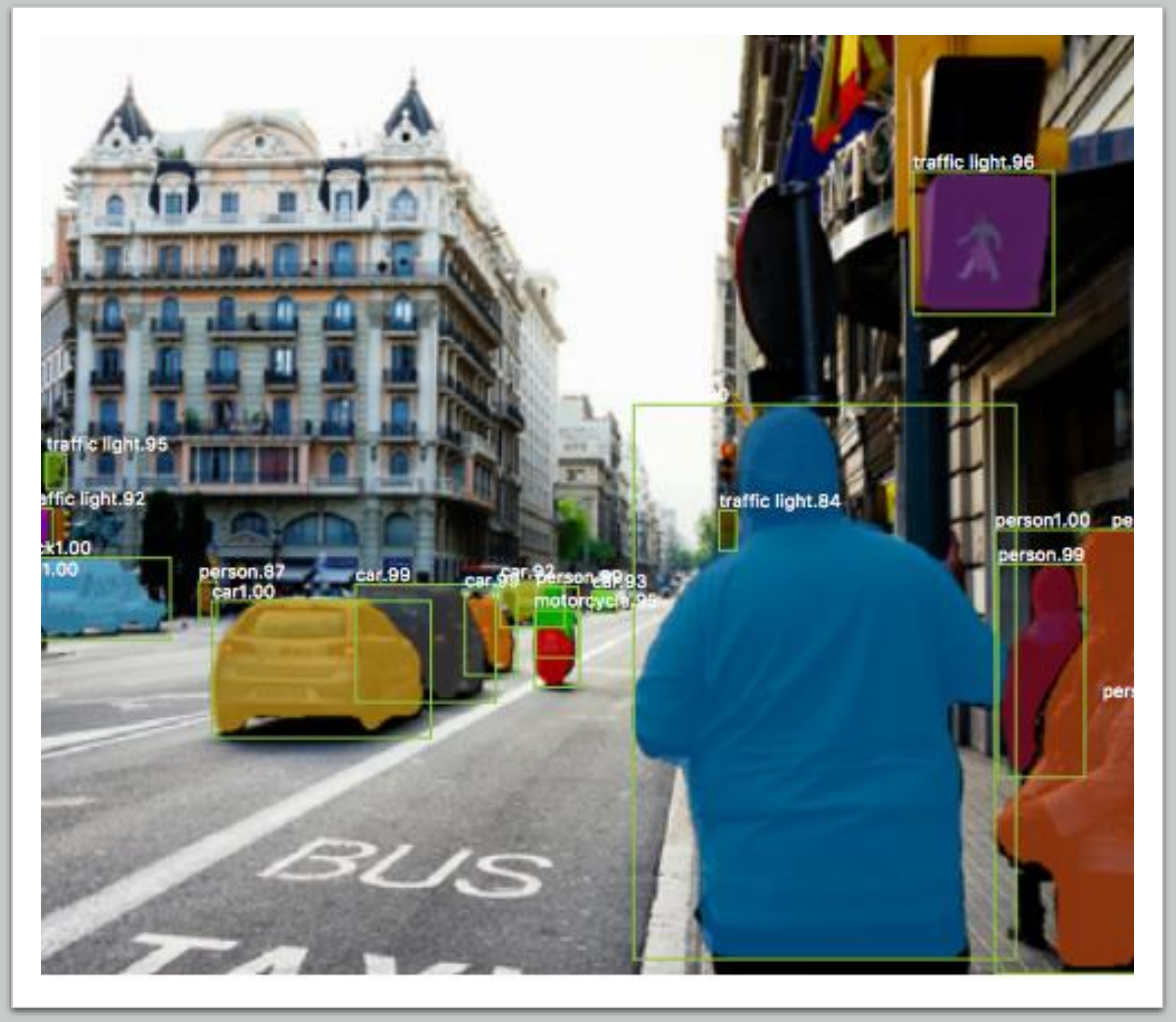


Selective Search

Figure 2: Two examples of our selective search showing the necessity of different scales. On the left we find many objects at different scales. On the right we necessarily find the objects at different scales as the girl is contained by the tv.

Mask R-CNN

- Extending Faster R-CNN for Pixel Level Segmentation



Unified Detection

- Problems with 2 step detection
  - Complex pipeline
  - Slow
  - Hard to optimize each component
- YOLO
  - Use a single ConvNet
  - Run once on entire image

YOLO – 'You Only Look Once'

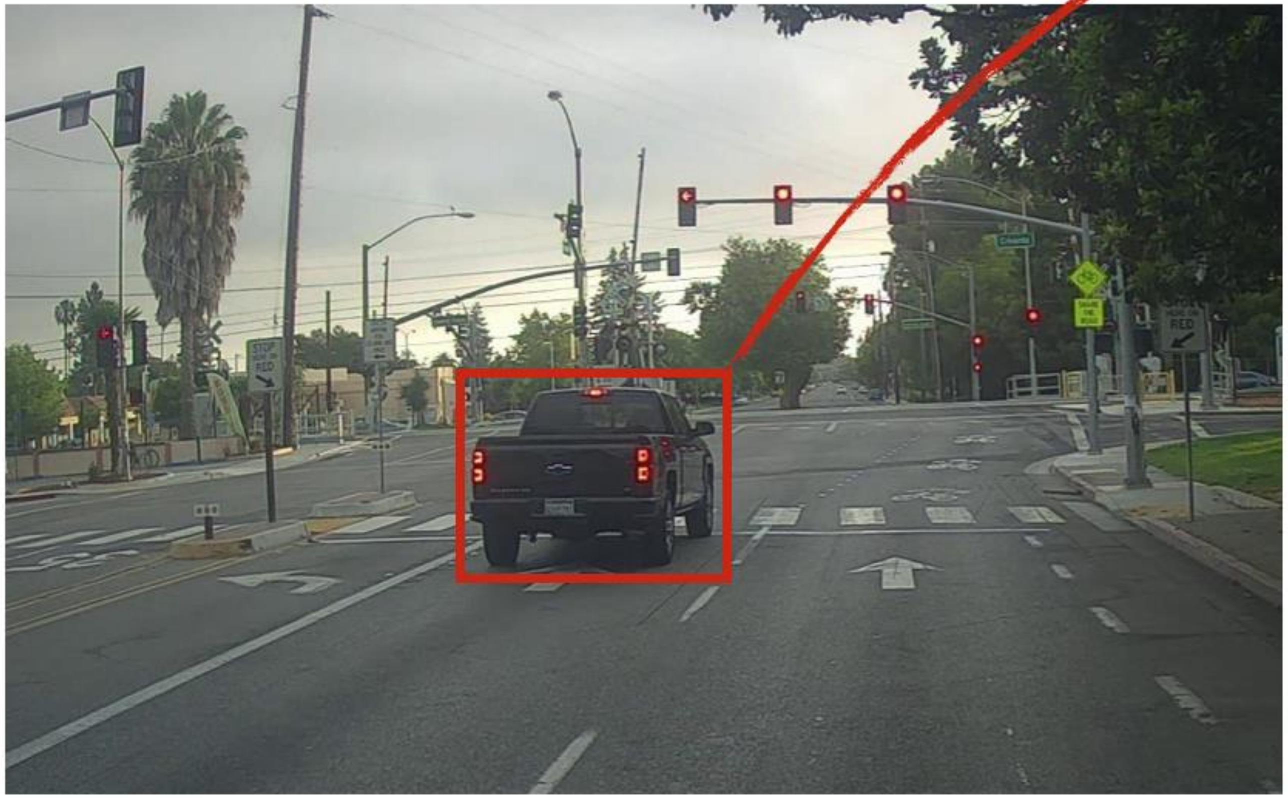
https://www.youtube.com/watch?time_continue=193&v=MPU2HistivI

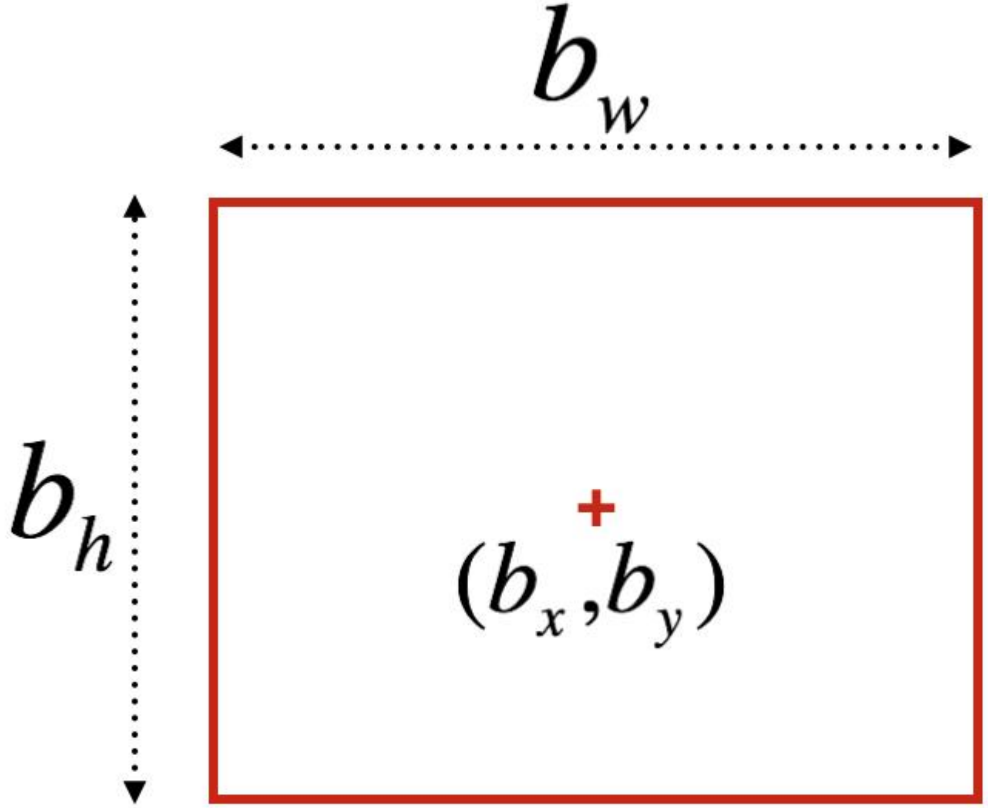
- State-of-the-art object detection system
- COCO test-dev: mAP = 57.9%
- YOLOv3: very fast compared to other models

How it works?

- Other methods: Apply model to image at multiple locations
- YOLO: Apply a single neural network to the full image

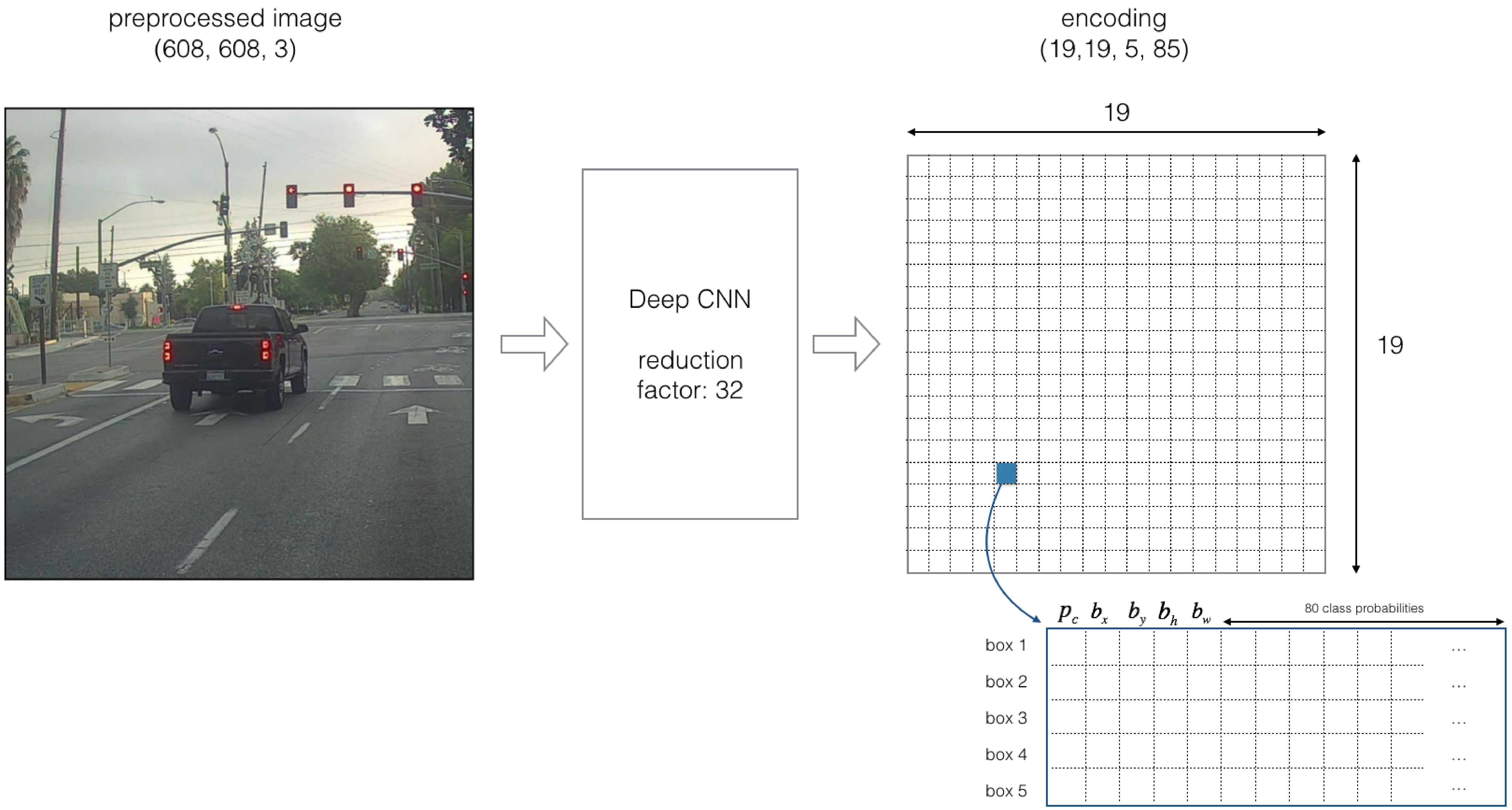
$$y = (p_c, b_x, b_y, b_h, b_w, c)$$



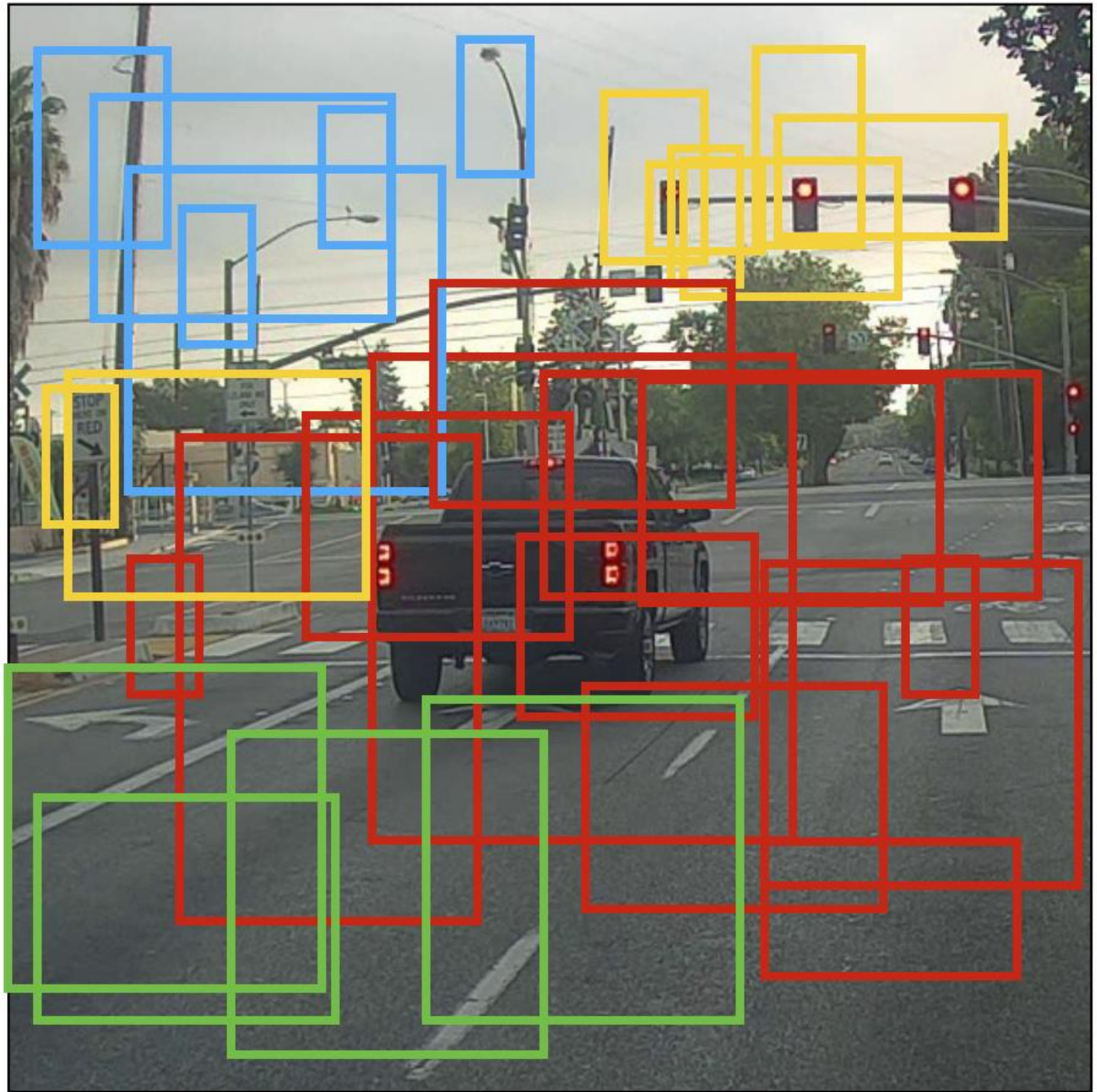


$p_c = 1$: confidence of an object being present in the bounding box

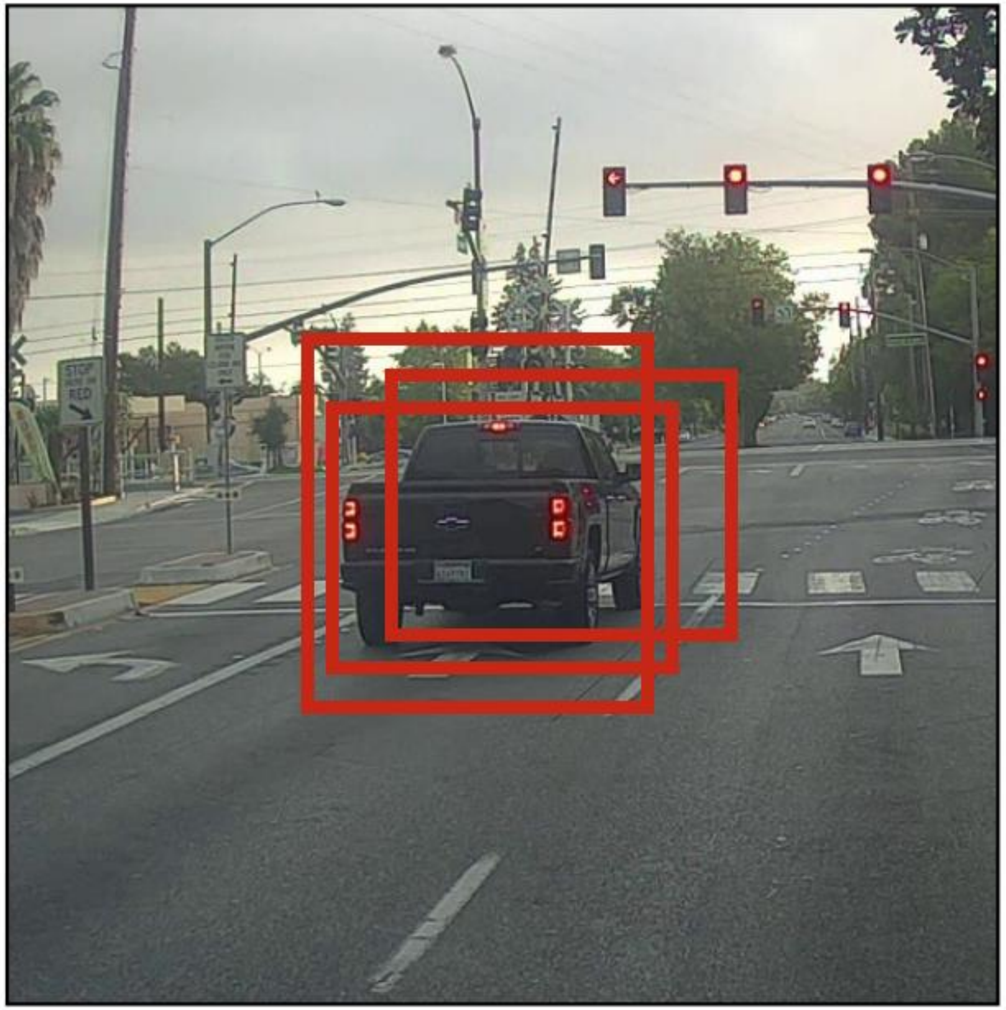
$c = 3$: class of the object being detected (here 3 for “car”)


preprocessed image
(608, 608, 3)
Deep CNN
reduction factor: 32
encoding
(19,19, 5, 85)
19
19
p_c b_x b_y b_h b_w
80 class probabilities
box 1
box 2
box 3
box 4
box 5


car
road sign
tree
traffic light
sky
background



Before non-max suppression



Non-Max Suppression

After non-max suppression

Non-max suppression

1

Select the box that has the highest score.

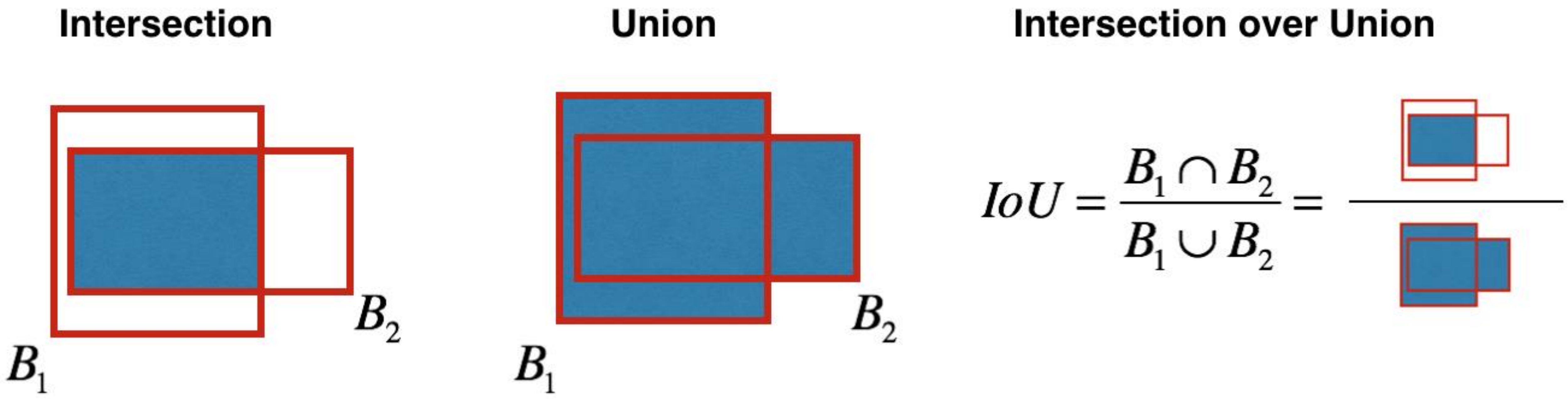
2

Compute its overlap with all other boxes, and remove boxes that overlap it more than iou_threshold.

3

Go back to step 1 and iterate until there's no more boxes with a lower score than the current selected box.

IoU


Intersection
Union
Intersection over Union
B_1
B_2
B_1
B_2
$IoU = \frac{B_1 \cap B_2}{B_1 \cup B_2} =$

Performance

	Yolo	Faster R-CNN (VGG-16)
mAP	63.4	73.2
FPS	45	7

Summary

- Two-step detection: R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN
  - Regional Proposal
  - Classification
- Unified detection: YOLO
  - Single model
  - Non-max suppression
- YOLO can run in real-time