Object Detection – Part 2

Mu Yadong

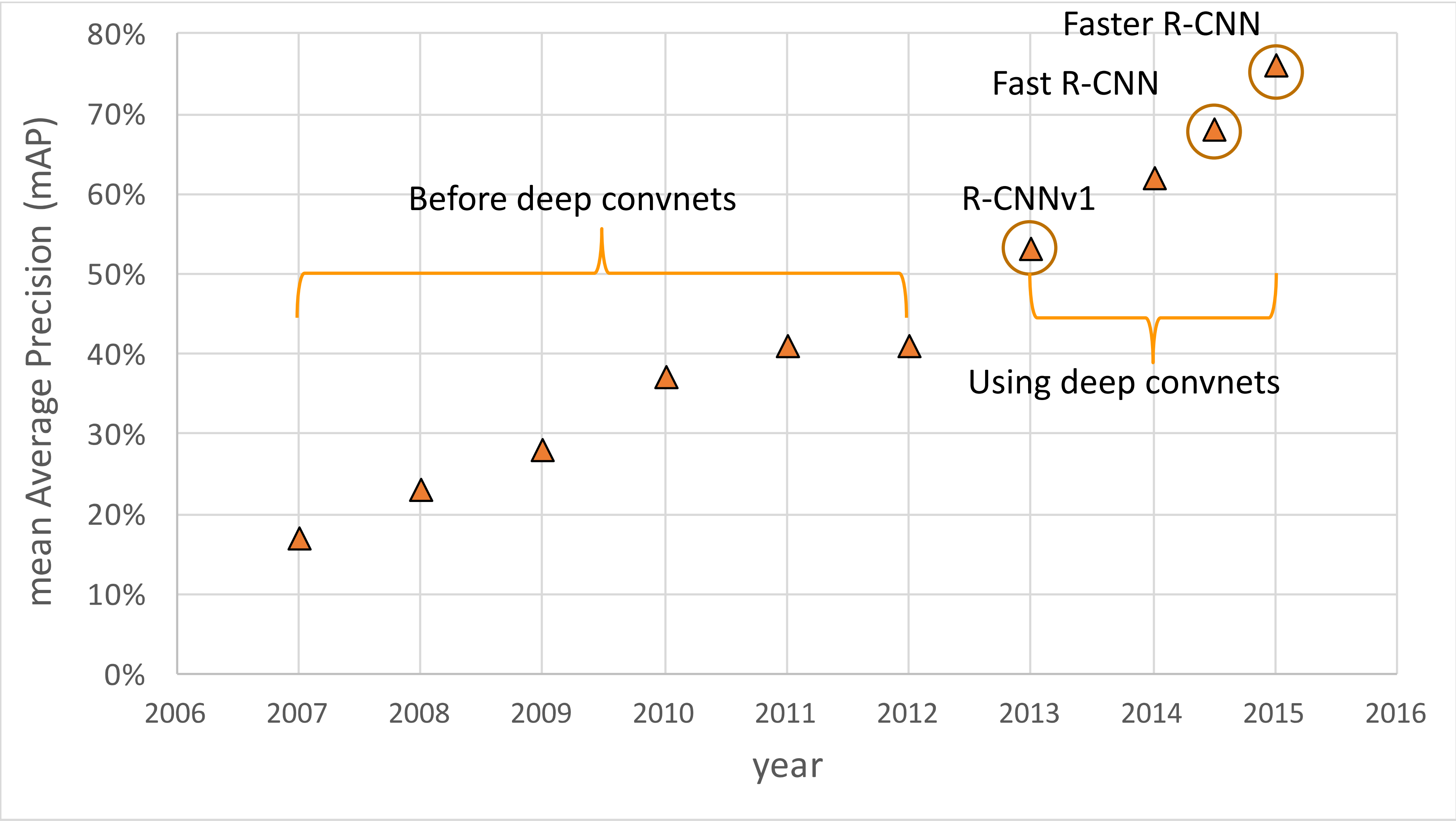
Machine Intelligence Lab
Institute of Computer Science & Technology
Peking University

Most slides are adapted from related courses or tutorials.
Internal use only. Please do not distribute the slides.

Outline

- **R-CNN, Fast R-CNN, Faster R-CNN**
- **YOLO, SSD**
- **Other Extensions**

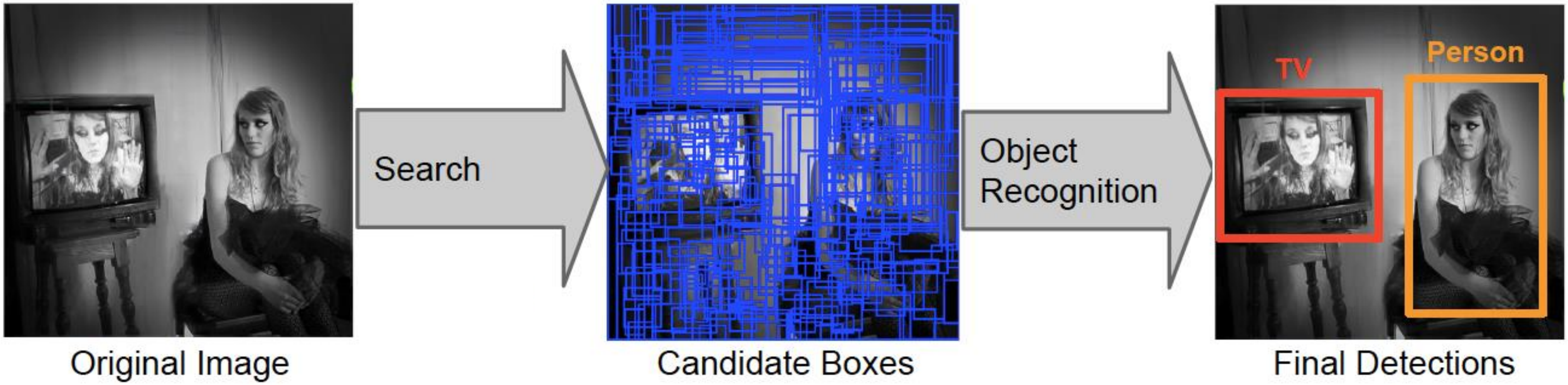
Object detection progress



Convolutional Feature Maps

- See He Kaiming's tutorial slides
- http://mp7.watson.ibm.com/ICCV2015/ObjectDetectionICCV2015.html

Beyond sliding windows: Region proposals



- Advantages:
  - Cuts down on number of regions detector must evaluate
  - Allows detector to use more powerful features and classifiers
  - Uses low-level *perceptual organization* cues
  - Proposal mechanism can be category-independent
  - Proposal mechanism can be trained

Selective search: Basic idea

- Use hierarchical segmentation: start with small *superpixels* and merge based on diverse cues

J. Uijlings, K. van de Sande, T. Gevers, and A. Smeulders, Selective Search for Object Recognition, IJCV 2013

Bounding Box Regression

- Input: A set of N training pairs $\{(P^i, G^i)\}_{i=1,\dots,N}$, where $P^i = (P_x^i, P_y^i, P_w^i, P_h^i)$ and $G^i = (G_x^i, G_y^i, G_w^i, G_h^i)$. (Drop superscript i for simplicity)
- Output: Four functions $d_x(P), d_y(P),\ d_w(P), d_h(P)$

$$\hat{G}_x = P_w d_x(P) + P_x \quad (1)$$

$$\hat{G}_y = P_h d_y(P) + P_y \quad (2)$$

$$\hat{G}_w = P_w \exp(d_w(P)) \quad (3)$$

$$\hat{G}_h = P_h \exp(d_h(P)). \quad (4)$$

- Learn model parameters by optimizing the regularized least squares objective

$$\mathbf{w}_\star = \underset{\hat{\mathbf{w}}_\star}{\operatorname{argmin}} \sum_i^N (t_\star^i - \hat{\mathbf{w}}_\star^{\mathrm{T}} \phi_5(P^i))^2 + \lambda \left\| \hat{\mathbf{w}}_\star \right\|^2 . \quad (5)$$

Bounding Box Regression

- Learn model parameters by optimizing the regularized least squares objective

$$\mathbf{w}_\star = \underset{\hat{\mathbf{w}}_\star}{\operatorname{argmin}} \sum_i^N (t_\star^i - \hat{\mathbf{w}}_\star^{\mathrm{T}} \boldsymbol{\phi}_5(P^i))^2 + \lambda \left\| \hat{\mathbf{w}}_\star \right\|^2 . \tag{5}$$

- The regression targets t_* for the training pair (P, G) are defined as

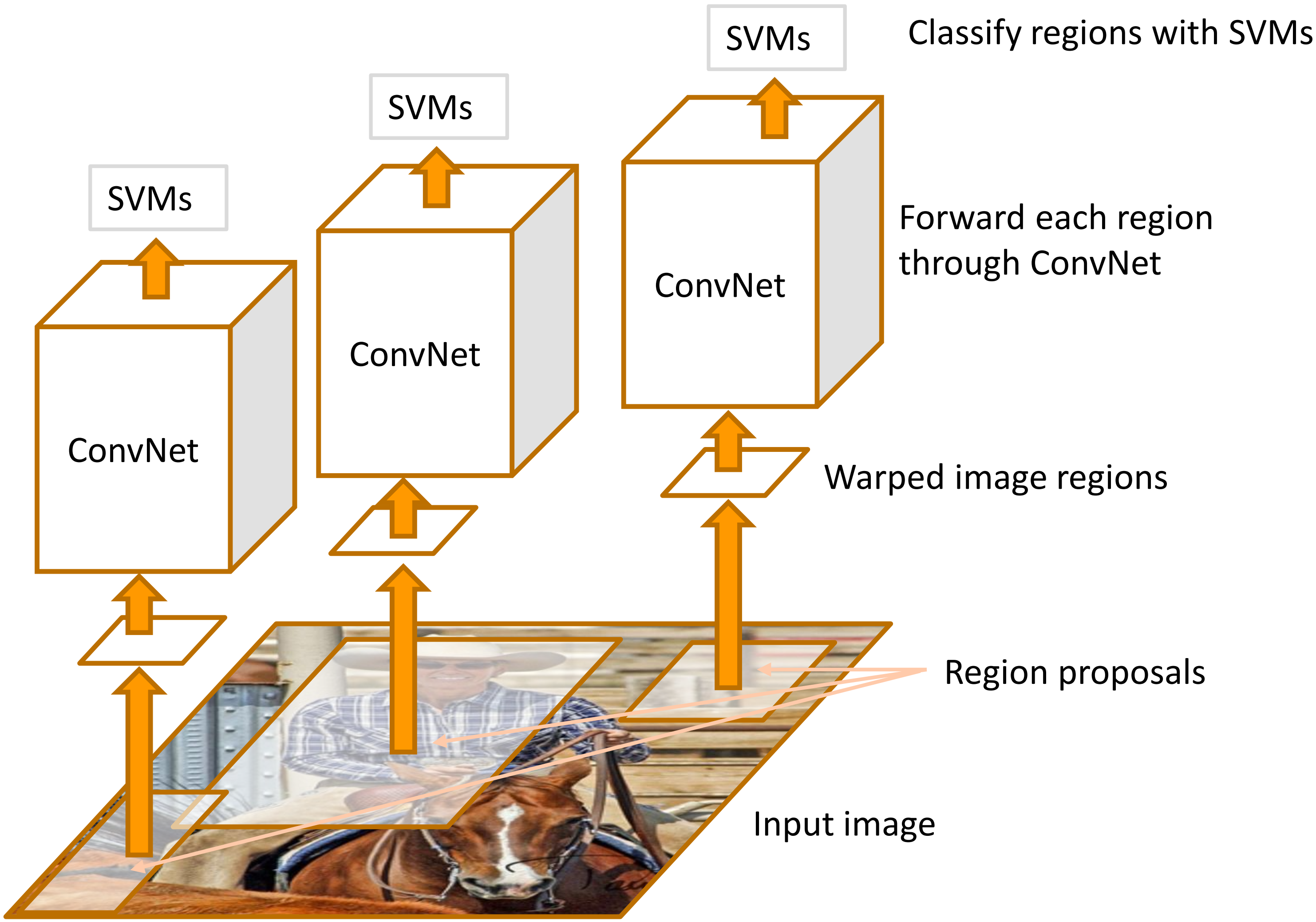
$$t_x = (G_x - P_x)/P_w \tag{6}$$

$$t_y = (G_y - P_y)/P_h \tag{7}$$

$$t_w = \log(G_w/P_w) \tag{8}$$

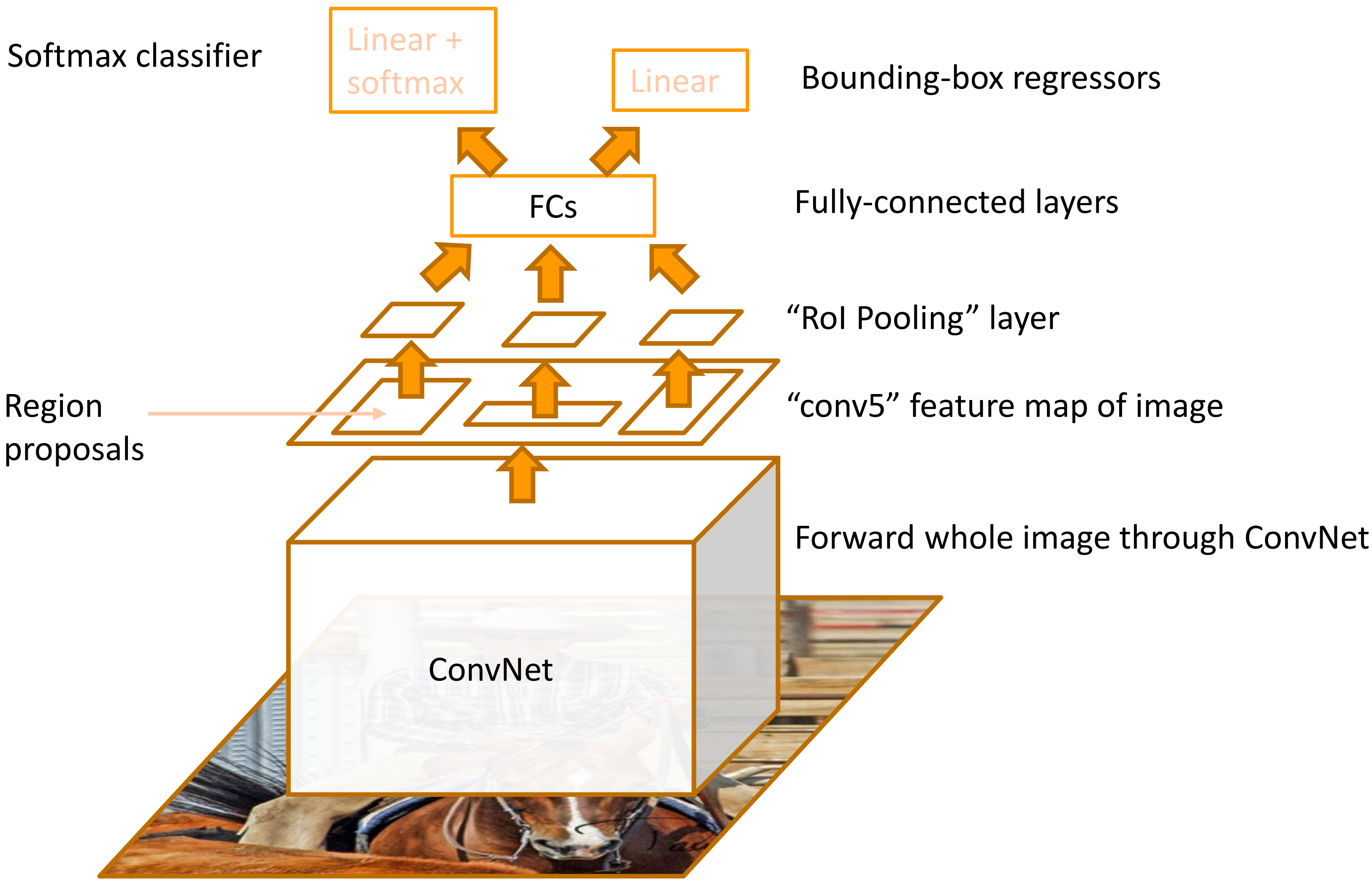
$$t_h = \log(G_h/P_h). \tag{9}$$

Review: R-CNN



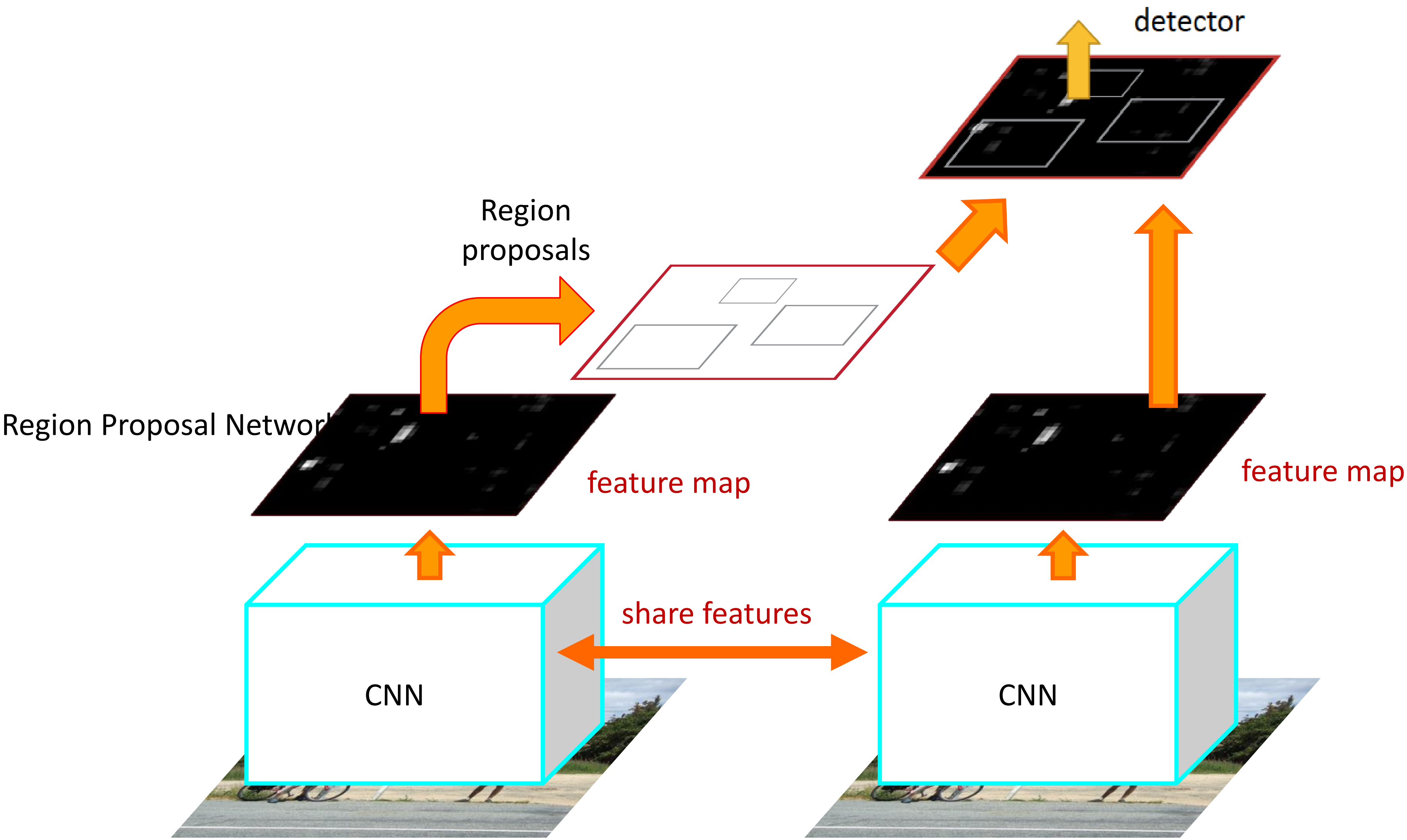
R. Girshick, J. Donahue, T. Darrell, and J. Malik, **Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation**, CVPR 2014.

Review: Fast R-CNN



R. Girshick, Fast R-CNN, ICCV 2015

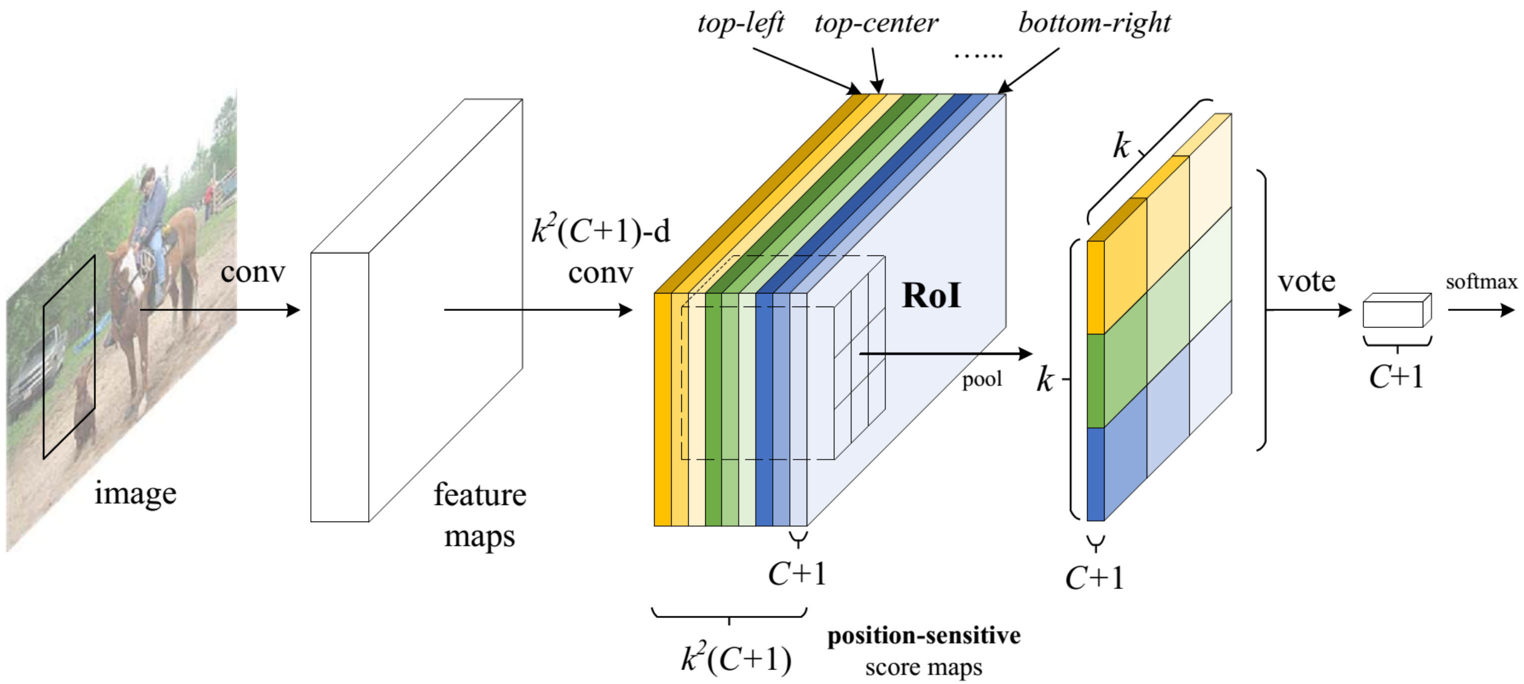
Review: Faster R-CNN



S. Ren, K. He, R. Girshick, and J. Sun, Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks, NIPS 2015

R-FCN

- Abandon FC layers. Use fully convolutional layers.



Jifeng Dai et al., R-FCN: Object Detection via Region-based Fully Convolutional Networks, NIPS 2016

R-FCN

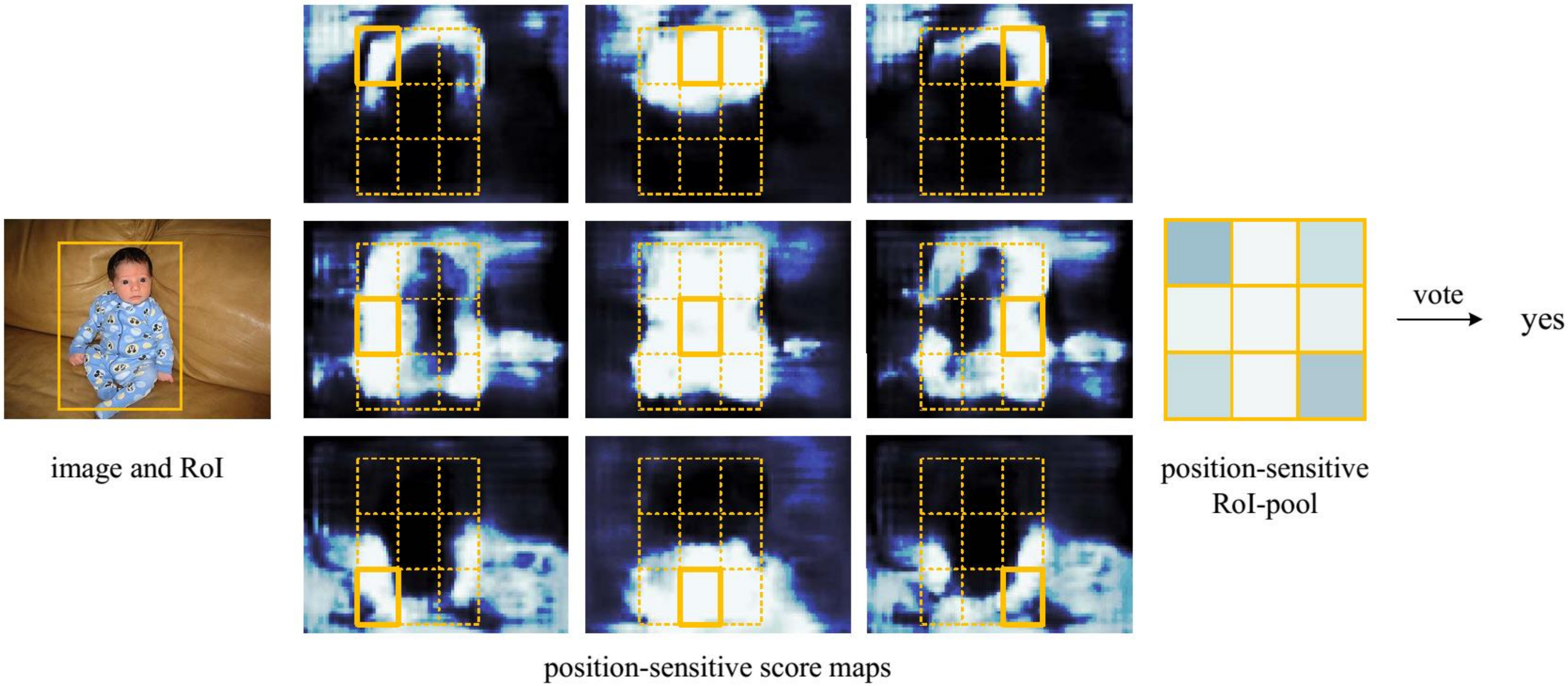
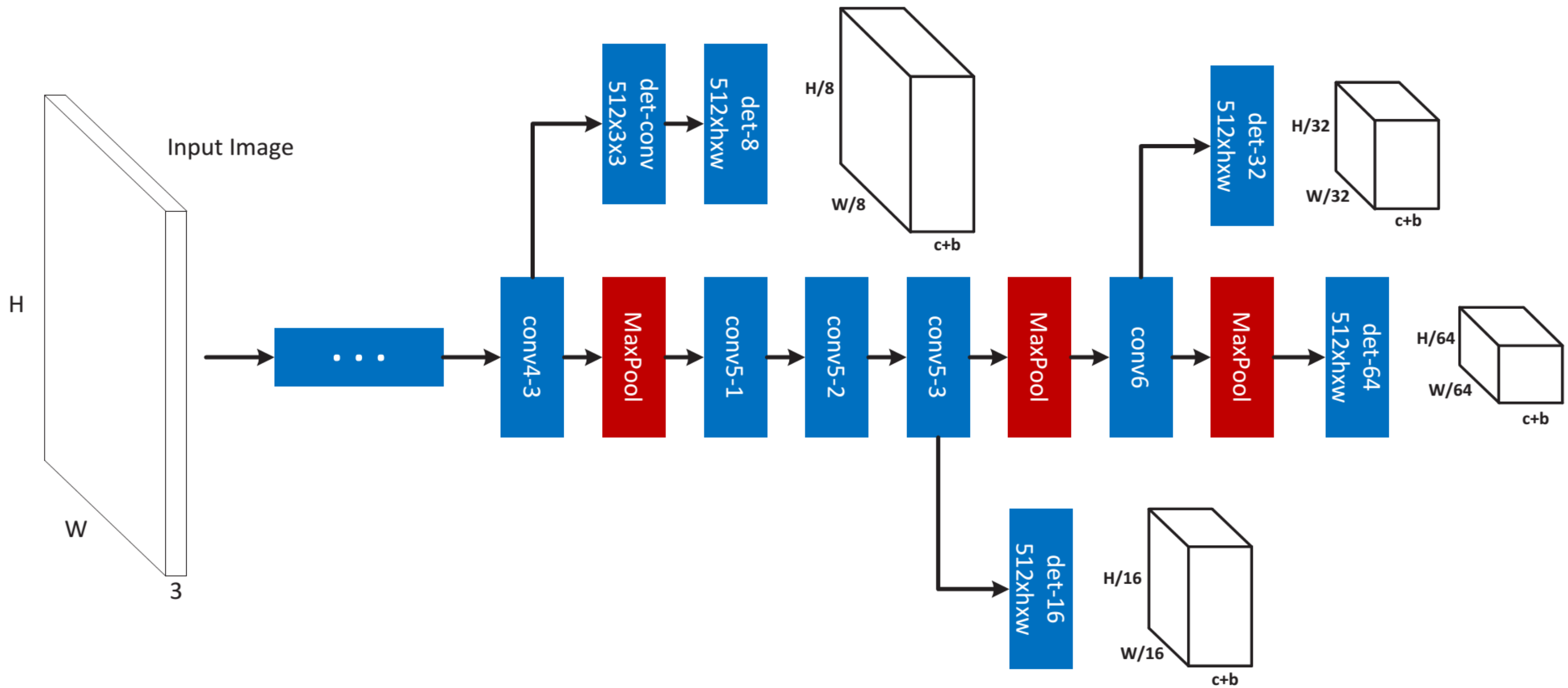


Figure 3: Visualization of R-FCN ($k \times k = 3 \times 3$) for the *person* category.

Multi-Scale R-CNN

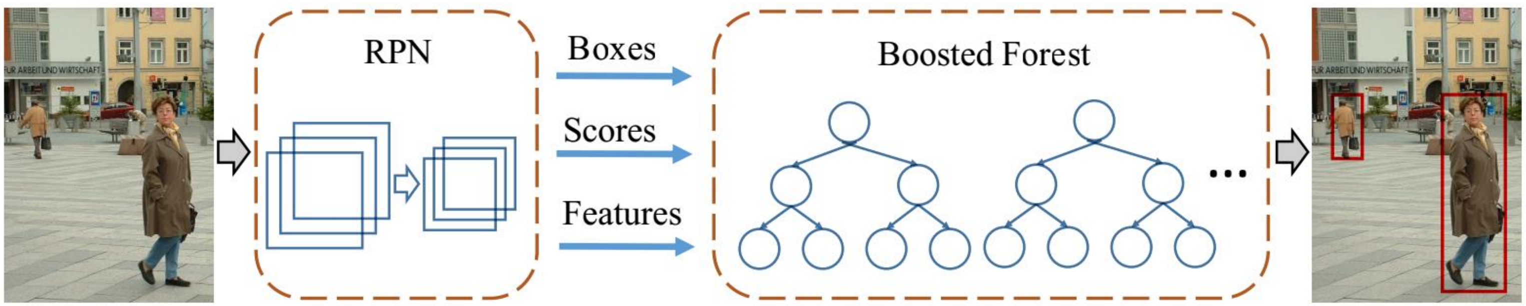
- Objects at different scales



Zhaowei Cai et al, A Unified Multi-scale Deep CNN for Fast Ob ject Detection, ECCV 2016

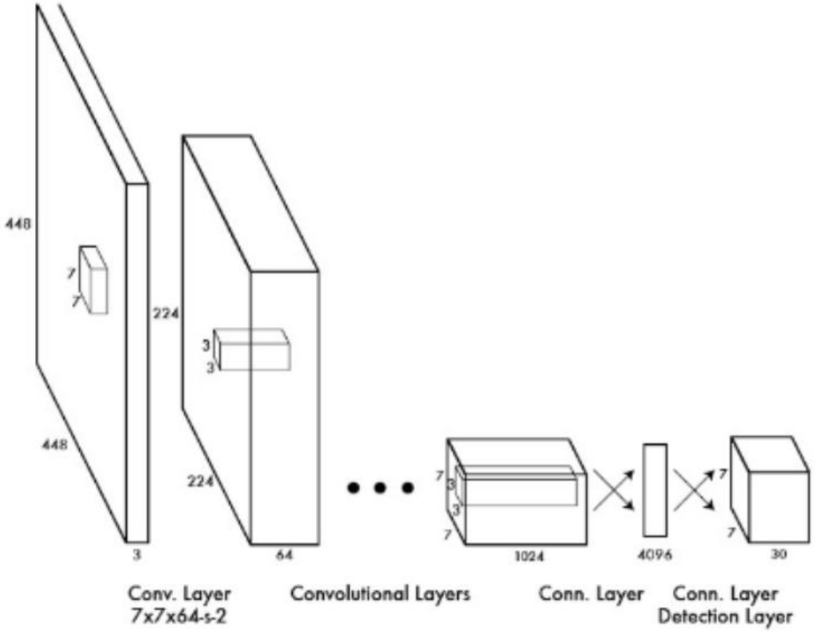
Remove Linear Classifiers

- FC layers are “weak” classifiers
- Replace with non-linear classifiers

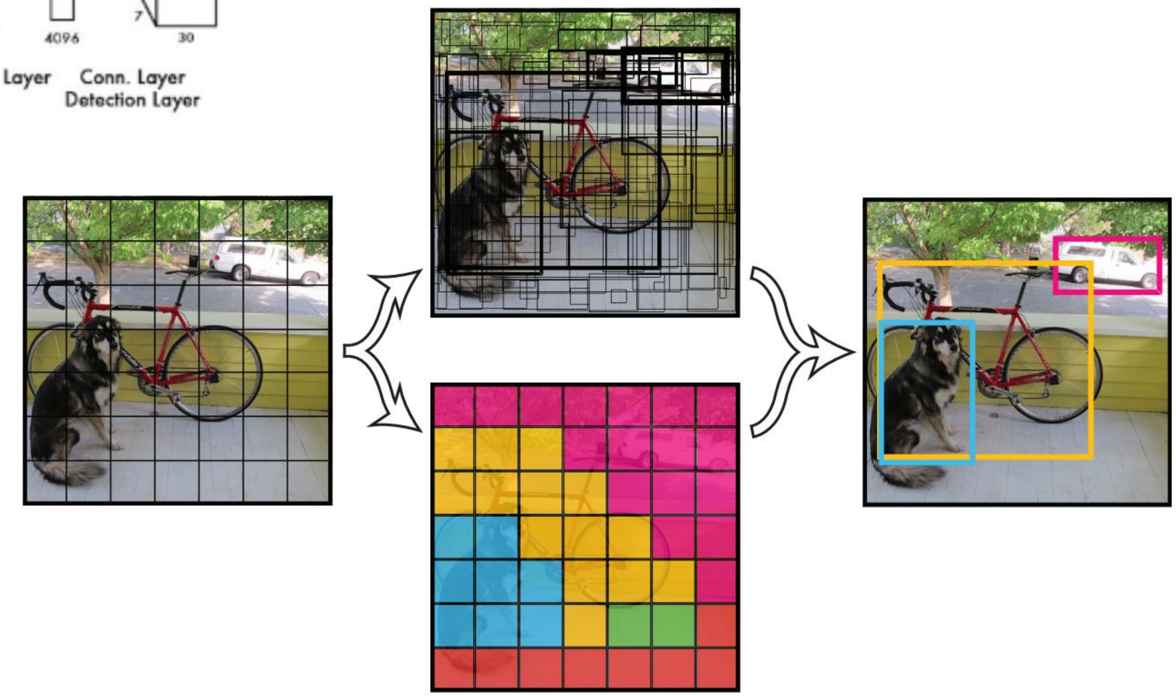


Liliang Zhang et al., Is Faster R-CNN Doing Well for Pedestrian Detection? ECCV 2016

YOLO



Regression instead of classification:
If the center of an object falls into a grid cell, that grid cell is responsible for detecting that object.

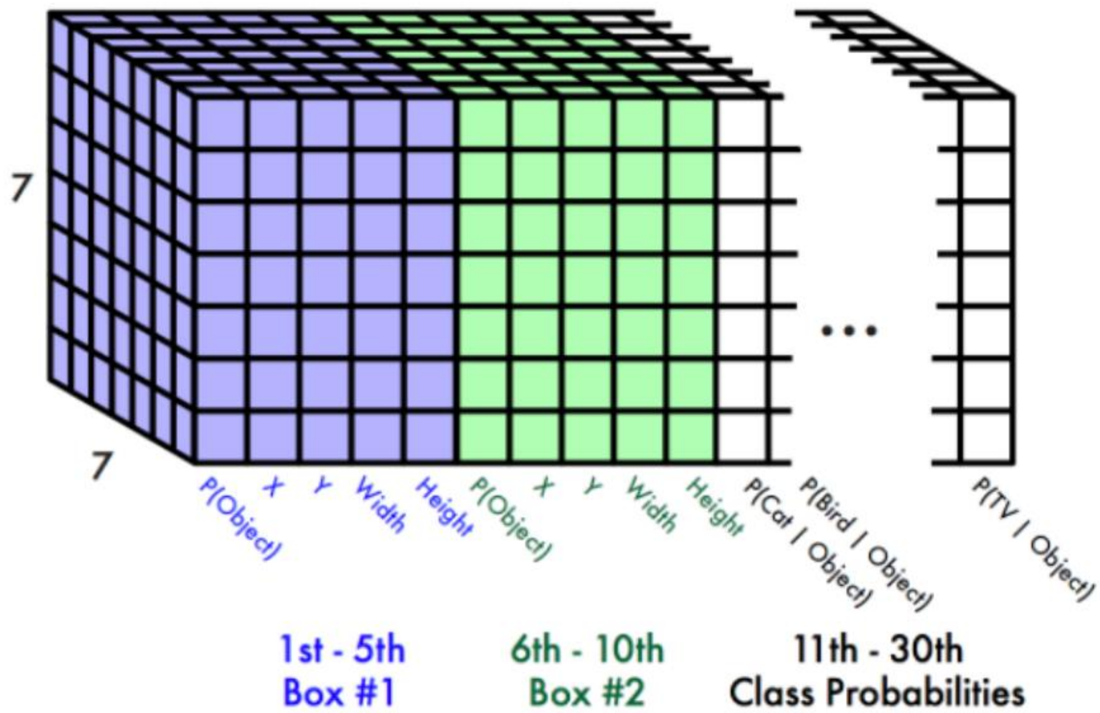


Redmon J, et al. You only look once: Unified, real-time object detection. CVPR2016

YOLO

Each cell predicts:

- For each bounding box:
  - 4 coordinates (x, y, w, h)
  - 1 confidence value
- Some number of class probabilities



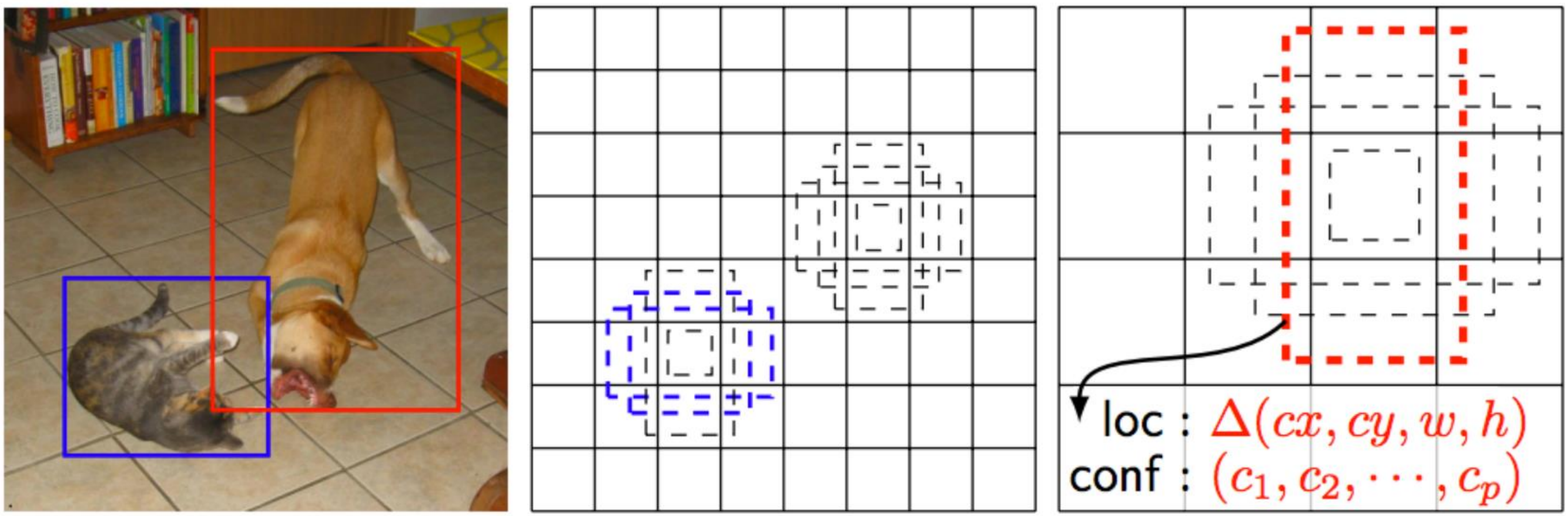
For Pascal VOC:

- 7x7 grid
- 2 bounding boxes / cell
- 20 classes

7 x 7 x (2 x 5 + 20) = 7 x 7 x 30 tensor = **1470 outputs**

SSD

- SSD: YOLO + default box shape + multi-scale

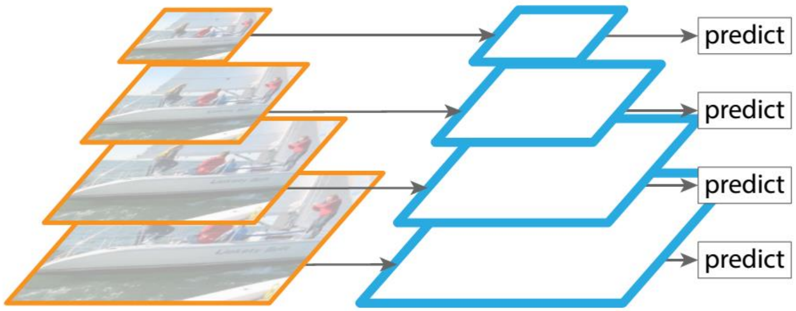


(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

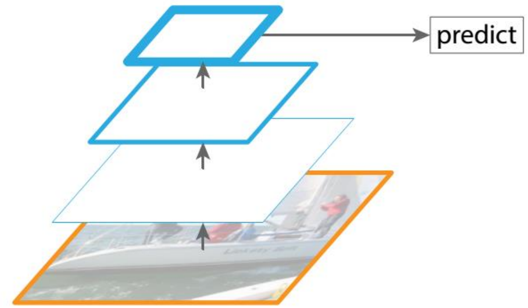
Wei L, et al. SSD: Single Shot MultiBox Detector. ECCV2016

FPN (Feature Pyramid Network)

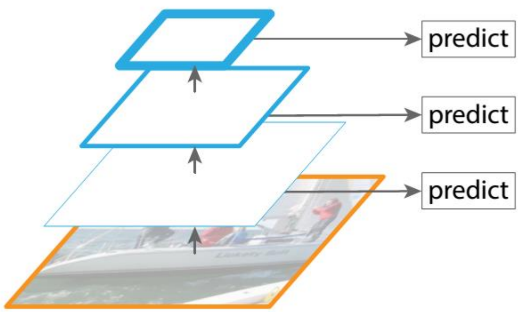
- Explore the power of multiple scales



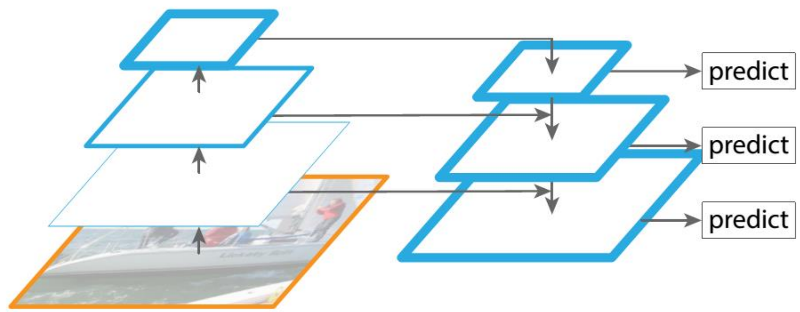
(a) Featurized image pyramid



(b) Single feature map



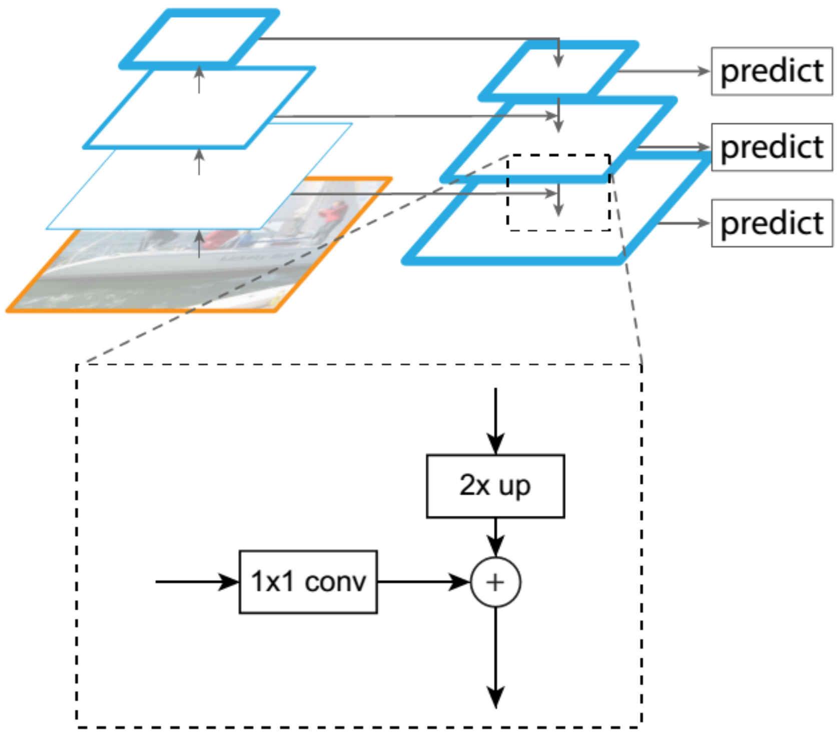
(c) Pyramidal feature hierarchy



(d) Feature Pyramid Network

FPN (Feature Pyramid Network)

- The top-down pathway



Question?