Object Detection

Computer Vision Lab.

Recap: Image classification

- Recognition of visual concepts on an image

Is there a bicycle?	Yes
Is there a person?	Yes
Is there a car?	No

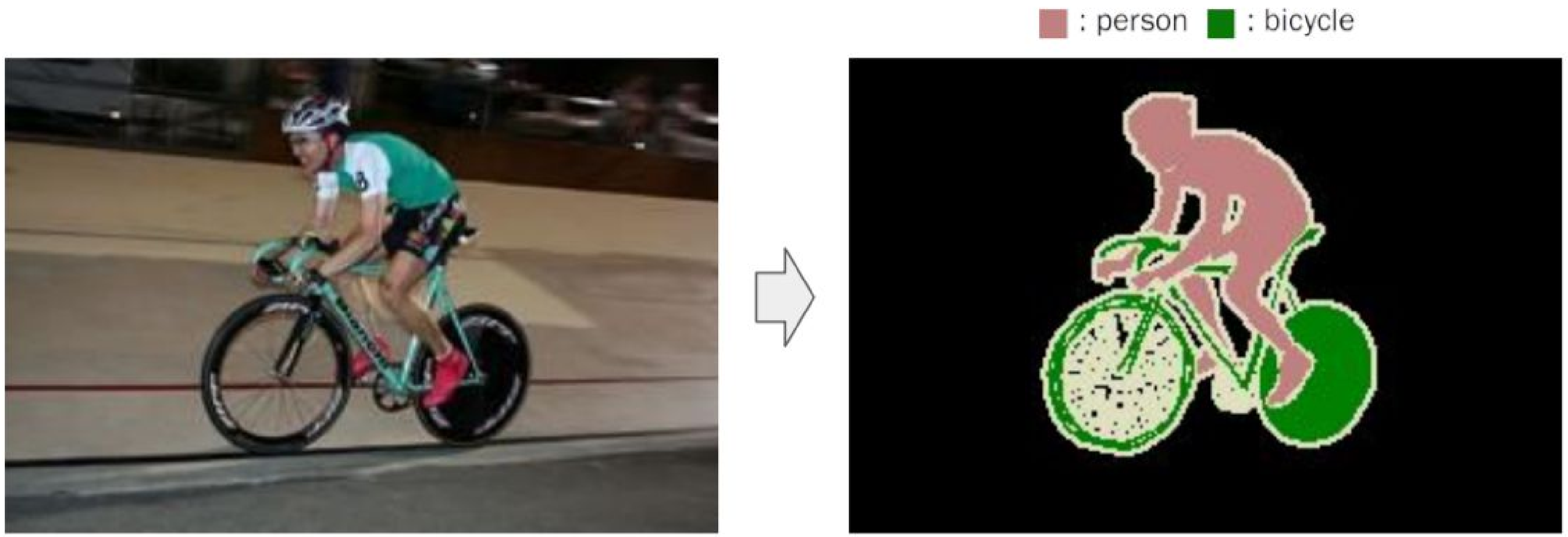
Object detection

- Recognition of visual concepts on an image
- Recognition and box-level localization of visual concepts on an image

Where is a bicycle?
Where is a person?

Recap: Semantic segmentation

- Recognition of visual concepts on an image
- Recognition and box-level localization of visual concepts on an image
- Recognition and pixel-level localization of visual concepts on an image



Challenges in detection and segmentation

- Recognition of visual concepts on an image
- Recognition and box-level localization of visual concepts on an image
- Recognition and pixel-level localization of visual concepts on an image

It requires
- more complicated outputs
  - e.g. box-level or pixel-level class labels
- more fine-grained understanding of the objects
  - e.g. object parts, occlusion, deformation

Challenges in detection and segmentation

- Recognition of visual concepts on an image
- Recognition and box-level localization of visual concepts on an image
- Recognition and pixel-level localization of visual concepts on an image

It requires

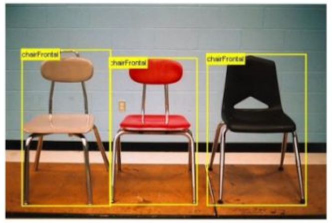
- more complicated outputs
  - e.g. box-level or pixel-level class labels
- more fine-grained understanding of the objects
  - e.g. object parts, occlusion, deformation

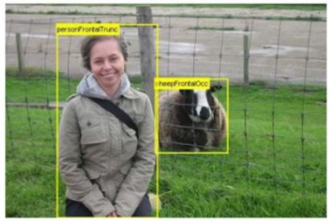
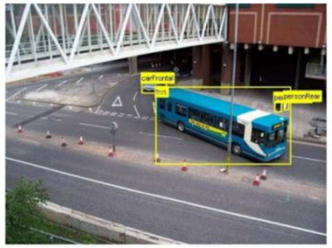
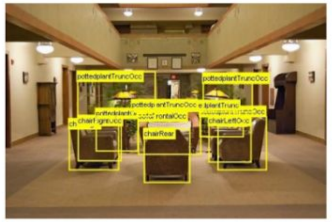
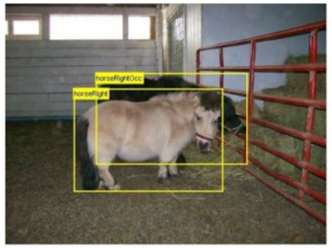
It can be resolved by

- Turing the problem into region-based classification
- Using more supervision during training (and more suitable architecture)

Object detection

- Training data
  - Each image in training set is associated with bounding box annotations
  - How can we learn to generate box label given these training data?

Object detection

- Object detection by regression?

DOG, (x, y, w, h)
CAT, (x, y, w, h)
CAT, (x, y, w, h)
DUCK (x, y, w, h)

= 16 numbers

Object detection

- If there are too many objects to detect?

CAT, (x, y, w, h)
CAT, (x, y, w, h)
....
CAT (x, y, w, h)

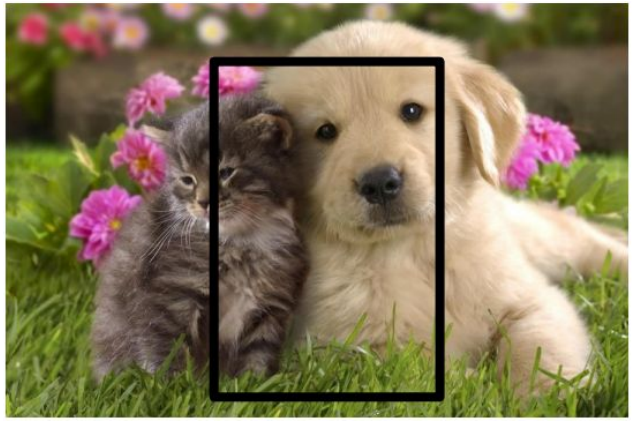
= many numbers

Object detection

- Object detection by region-based classification
  - Extract multiple candidate boxes around potential object locations and examine each box using classification score

CAT? YES!

DOG? NO



CAT? NO

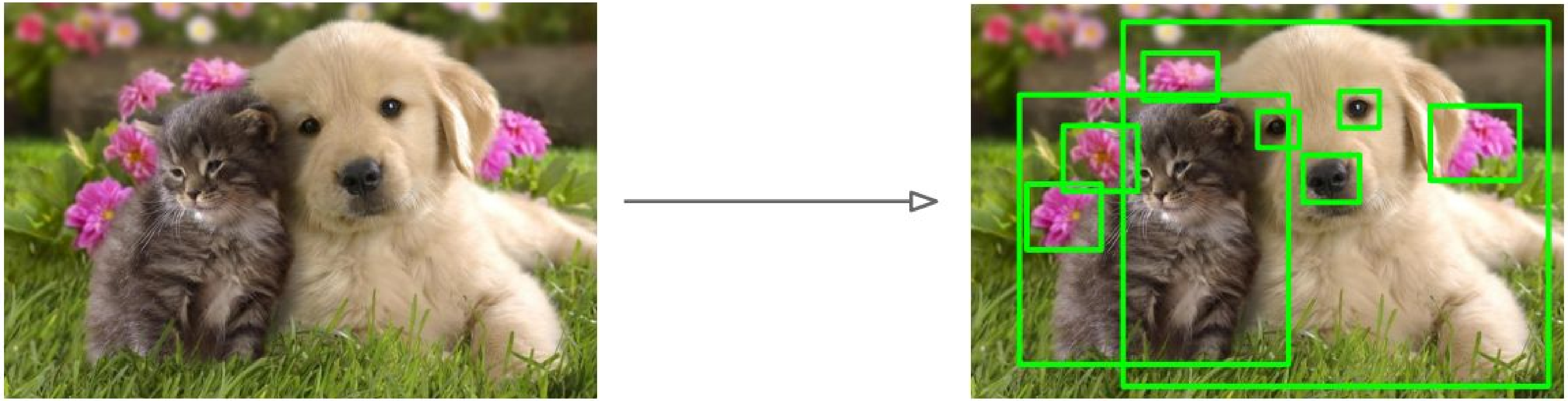
DOG? NO

What are the issues?

- It needs to test many positions and scales and use a computationally demanding such as CNN

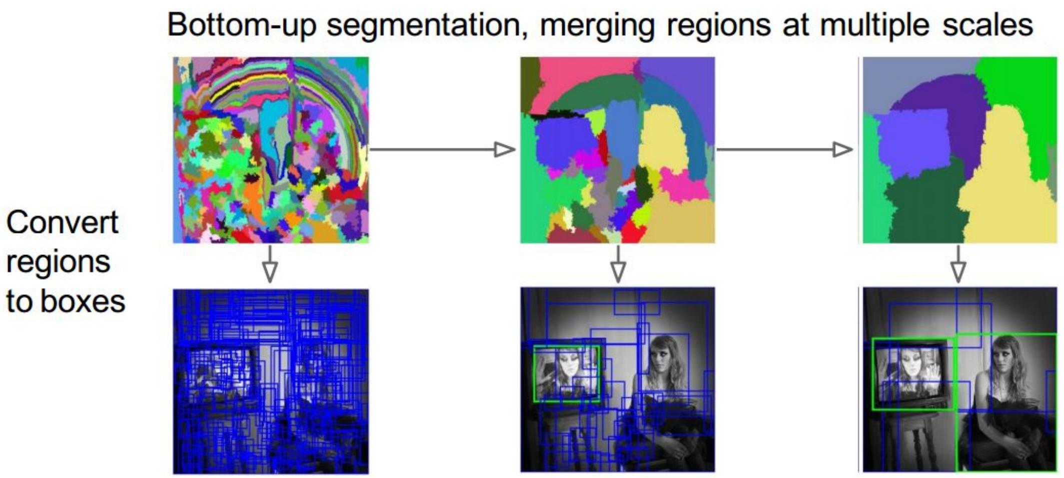
Region proposals

- Find image regions which may contain object.
- “Class-agnostic” object detector



Region Proposals: Selective Search

- Motivation
  - Sliding window approach is not feasible for object detection with CNN
  - We need a more faster to identify object candidates
- Finding object proposals
  - Greedy hierarchical super-pixel segm
  - Diversification of super-pixel construc
    - Using a variety of color spaces
    - Using different similarity measures
    - Varying staring regions



Region Proposals: Many other choices

Method	Approach	Outputs Segments	Outputs Score	Control #proposals	Time (sec.)	Repea-tability	Recall Results	Detection Results
Bing [18]	Window scoring		✓	✓	0.2	★★★	★	·
CPMC [19]	Grouping	✓	✓	✓	250	-	★★	★
EdgeBoxes [20]	Window scoring		✓	✓	0.3	★★	★★★	★★★
Endres [21]	Grouping	✓	✓	✓	100	-	★★★	★★
Geodesic [22]	Grouping	✓		✓	1	★	★★★	★★
MCG [23]	Grouping	✓	✓	✓	30	★	★★★	★★★
Objectness [24]	Window scoring		✓	✓	3	·	★	·
Rahtu [25]	Window scoring		✓	✓	3	·	·	★
RandomizedPrim's [26]	Grouping	✓		✓	1	★	★	★★
Rantalankila [27]	Grouping	✓		✓	10	★★	·	★★
Rigor [28]	Grouping	✓		✓	10	★	★★	★★
SelectiveSearch [29]	Grouping	✓	✓	✓	10	★★	★★★	★★★
Gaussian				✓	0	·	·	★
SlidingWindow				✓	0	★★★	·	·
Superpixels		✓			1	★	·	·
Uniform				✓	0	·	·	·

R-CNN: Regions with CNN features

- State-of-the-art: "Regions with CNN features" (R-CNN)
  Girshick et al, "Region-based Convolutional Networks for Accurate Object Detection and Semantic Segmentation", PAMI 2015 & CVPR 2014.

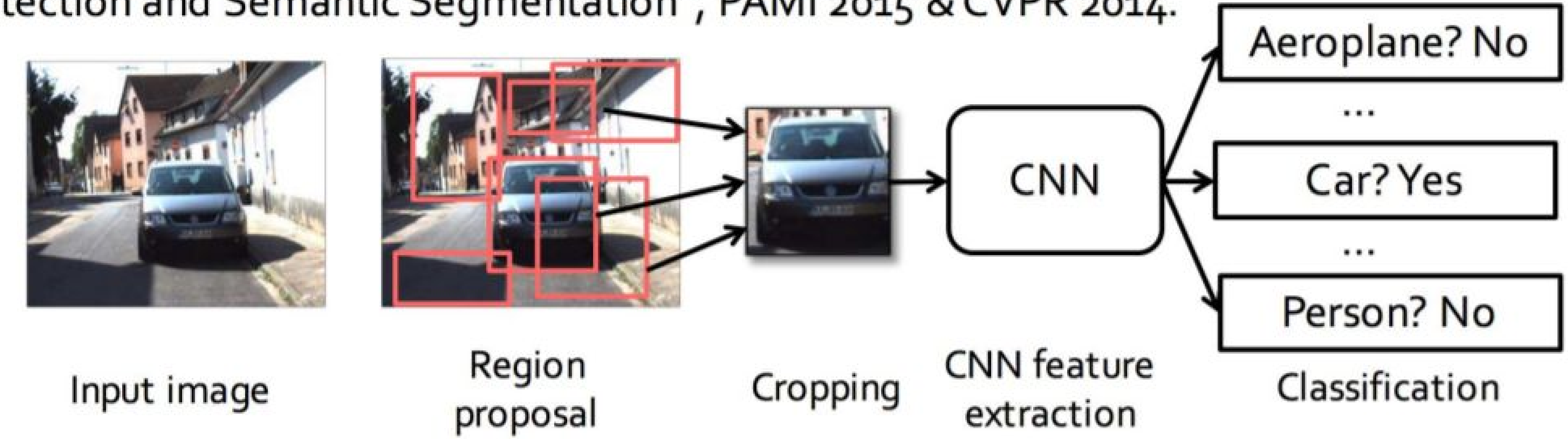
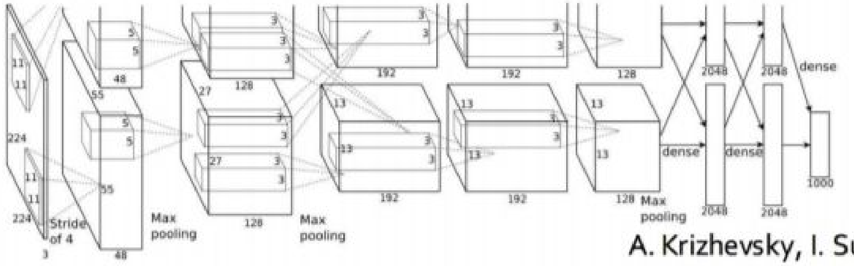


Image adapted from Girshick et al., 2014

R-CNN: Regions with CNN features

1) Convolutional neural network for classification

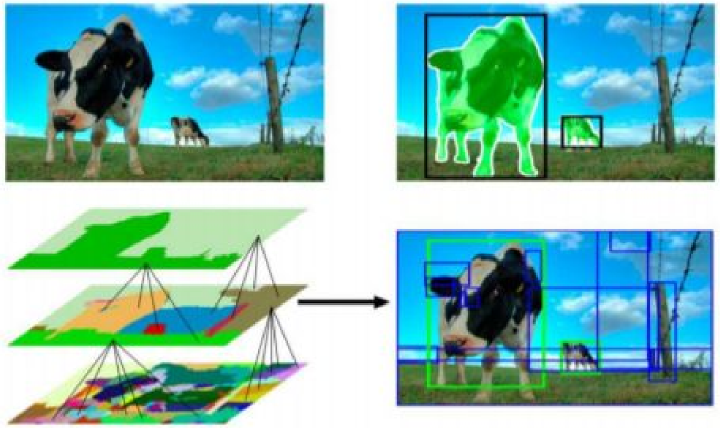


- Pretrained on ImageNet for 1000-category classification
- Finetuned on PASCAL VOC for 20 categories

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *NIPS*, 2012.

2) Selective search for region proposal:

- Hierarchical segmentation → bounding box



K. E. A. Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. *ICCV*, 2011.

Images from Krizhevsky et al. 2012 & Sande et al. 2011

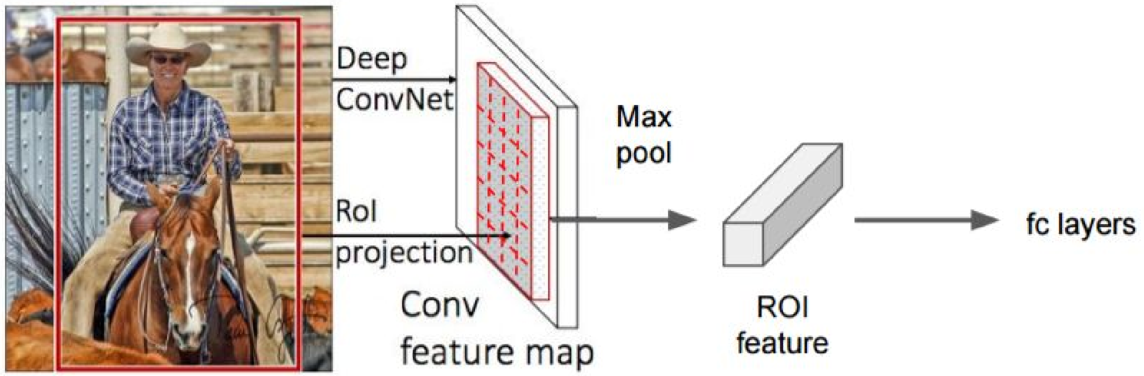
What are the issues?

- Slow processing time
  - It needs to iterate forward propagation of input image patch over all proposals (~ 2000 forward propagations in practice)
- Separate optimization of model components
  - Feature: CNN
  - Classifier: SVM
  - Region proposal: Selective Search Window
  - Post-processing: Bounding box regression

→ It is not desirable to the find optimal combination of all components

Make the operation more efficient

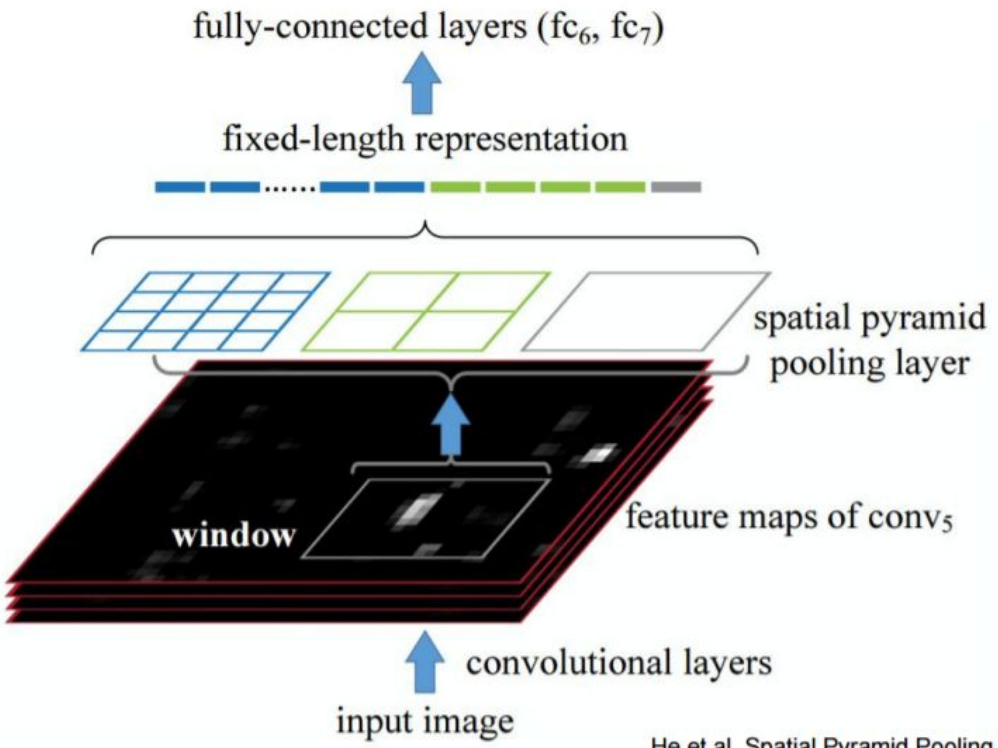
- Reuse feature maps by ROI (Region-Of-Interest) pooling



- Single forward propagation of a whole input image
- Iterative pooling for each bounding box region by ROI pooling
- It reduces processing time significantly

SPPNet: Spatial Pyramid Pooling

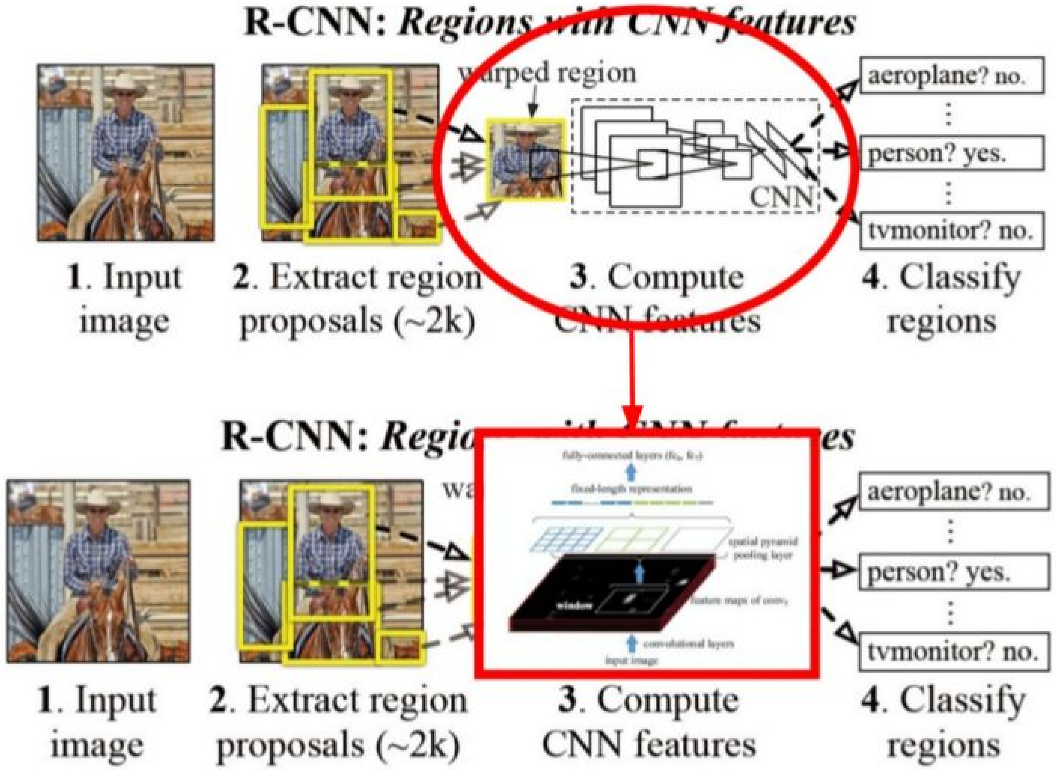
- ROI pooling in multiple spatial pyramid



He et al, Spatial Pyramid Pooling in D

SPPNet: Spatial Pyramid Pooling

- Replace feature extraction step of R-CNN with SPP



Is it done?

Recap: limitations in R-CNN

- Slow processing time
  - It needs to iterate forward propagation of input image patch over all proposals (~ 2000 forward propagations in practice)
- Separate optimization of model components
  - Feature: CNN
  - Classifier: SVM
  - Region proposal: Selective Search Window
  - Post-processing: Bounding box regression

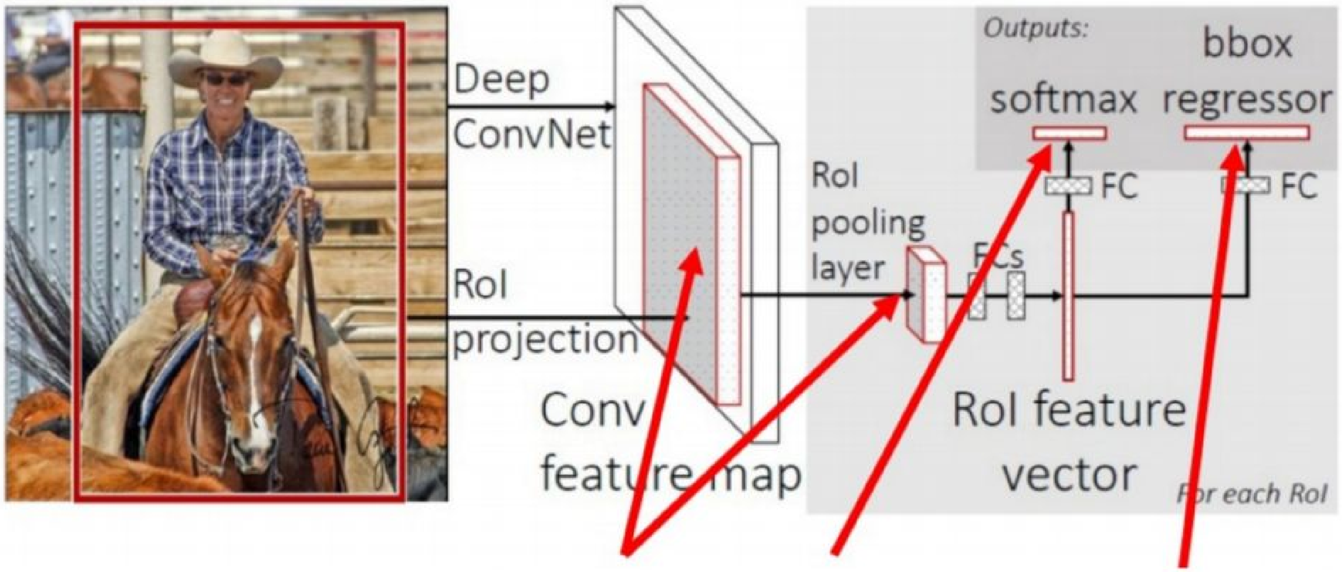
→ It is not desirable to the find optimal combination of all components

R-CNN vs SPPnet

- ~~Slow processing time~~ → **Much faster by ROI-pooling**
  - ~~It needs to iterate forward propagation of input image patch over all proposals (~ 2000 forward propagations in practice)~~
- Separate optimization of model components → **Still optimized in multiple stages**
  - Feature: CNN
  - Classifier: SVM
  - Region proposal: Selective Search Window
  - Post-processing: Bounding box regression

Fast R-CNN

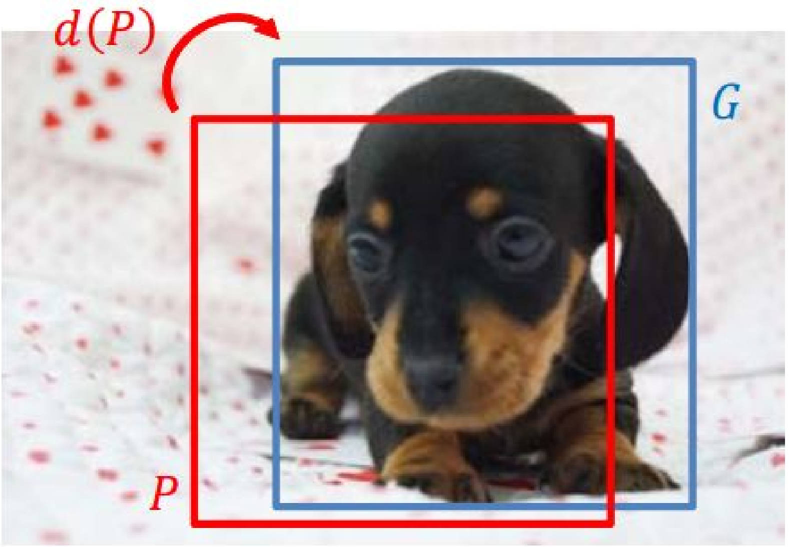
- Optimization of all (post) model components



Joint optimization of feature extractor, classifier, and regressor in a unified framework

Bounding Box Regressor

- Learning a transformation of bounding box
  - Regio
  - Grour
  - Trans



$$\hat{G}_x = P_w d_x(P) + P_x$$
$$\hat{G}_y = P_h d_y(P) + P_y$$
$$\hat{G}_w = P_w \exp\big(d_w(P)\big)$$
$$\hat{G}_h = P_h \exp\big(d_h(P)\big)$$

$$d_i(P) = \mathbf{w}_i^{\mathrm{T}} \boxed{\phi_5(P)}$$

CNN pool5 feature

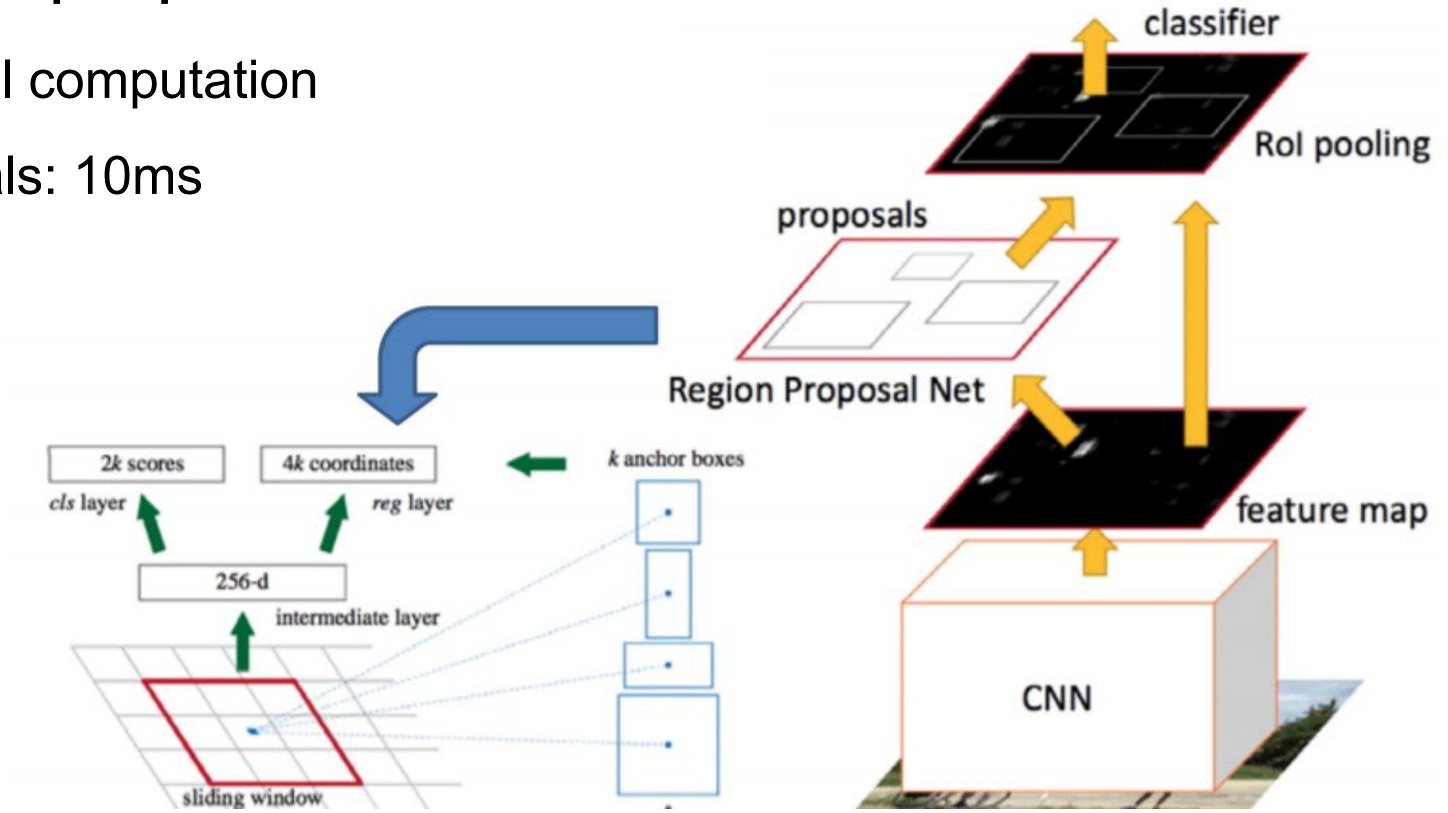
$$\mathbf{w}_i^* = \underset{\mathbf{w}_i}{\operatorname{argmin}} \sum_{k=1}^{N} \left(t_i^k - \mathbf{w}_i^{\mathrm{T}} \phi_5(P^k)\right)^2 + \lambda \|\mathbf{w}_i\|^2$$

SPPNet vs Fast R-CNN

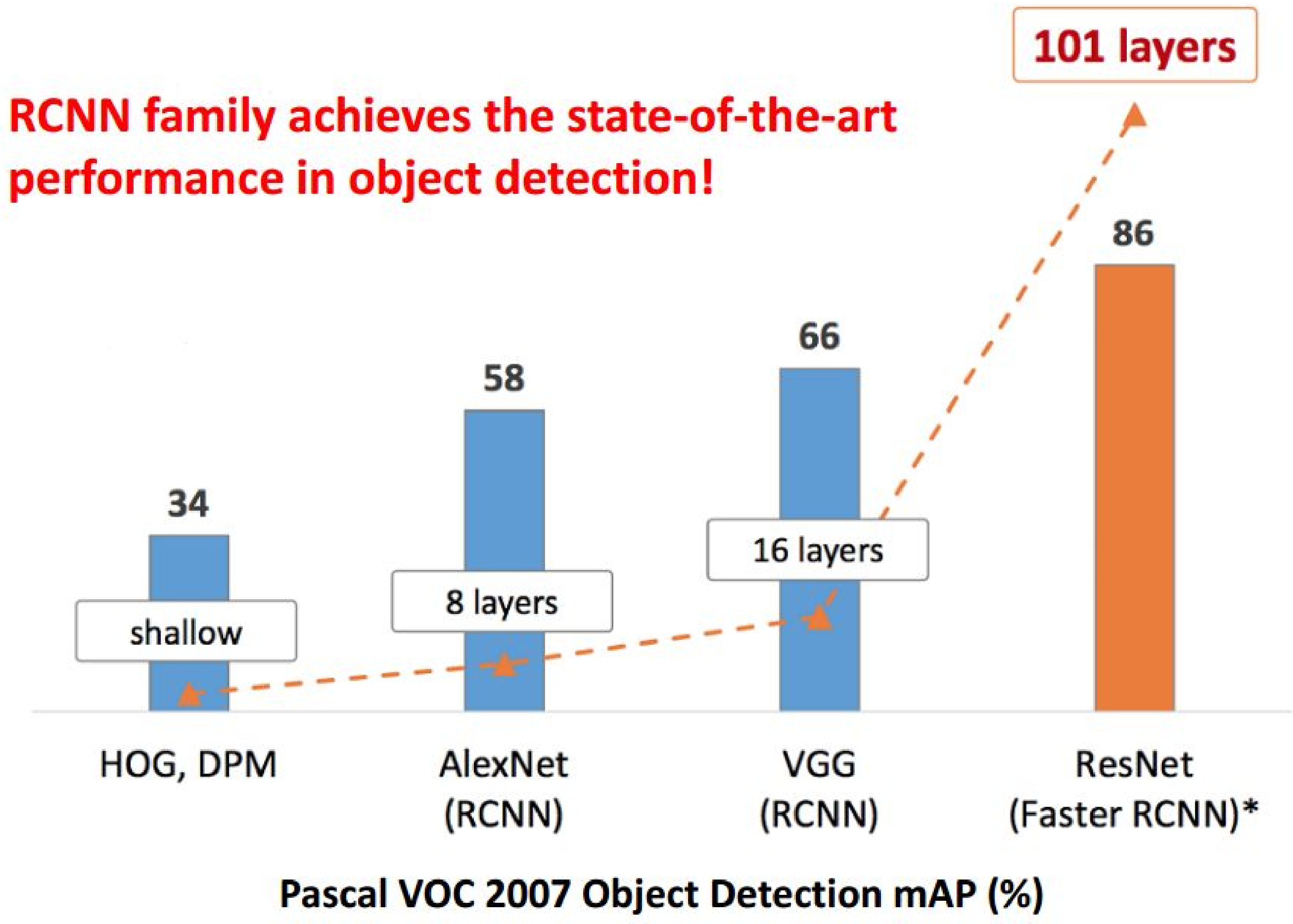
- ~~Slow processing time~~
  - ~~It needs to iterate forward propagation of input image patch over all proposals (~ 2000 forward propagations in practice)~~
- Separate optimization of model components → **Still use SSW**
  - Feature: CNN
  - Classifier: CNN
  - Post-processing: Bounding box regression (CNN)
  - Region proposal: Selective Search Window

Faster R-CNN

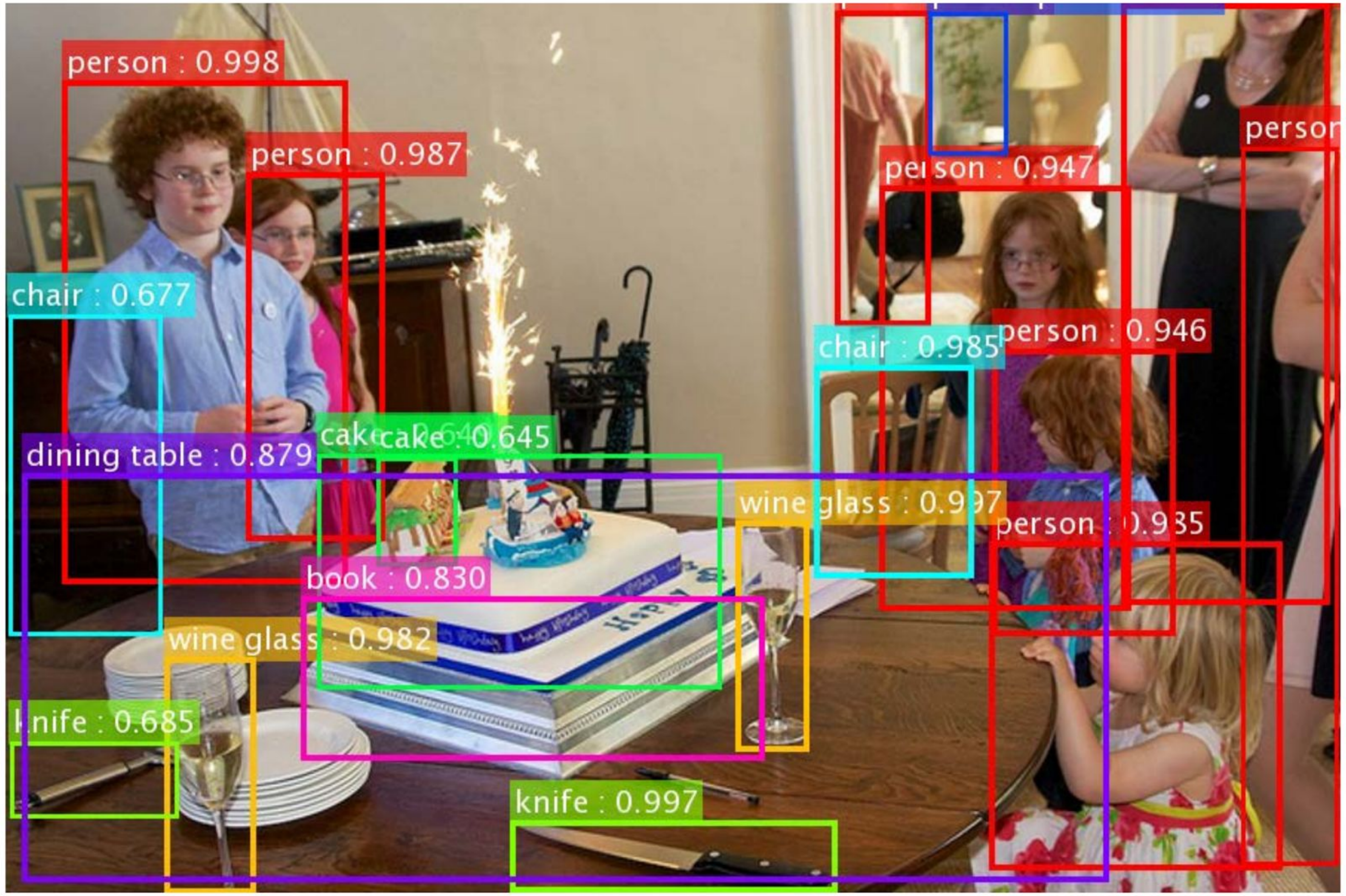
- Fast R-CNN + region-proposal network
  - Integrate region proposal computation
  - Marginal cost of proposals: 10ms



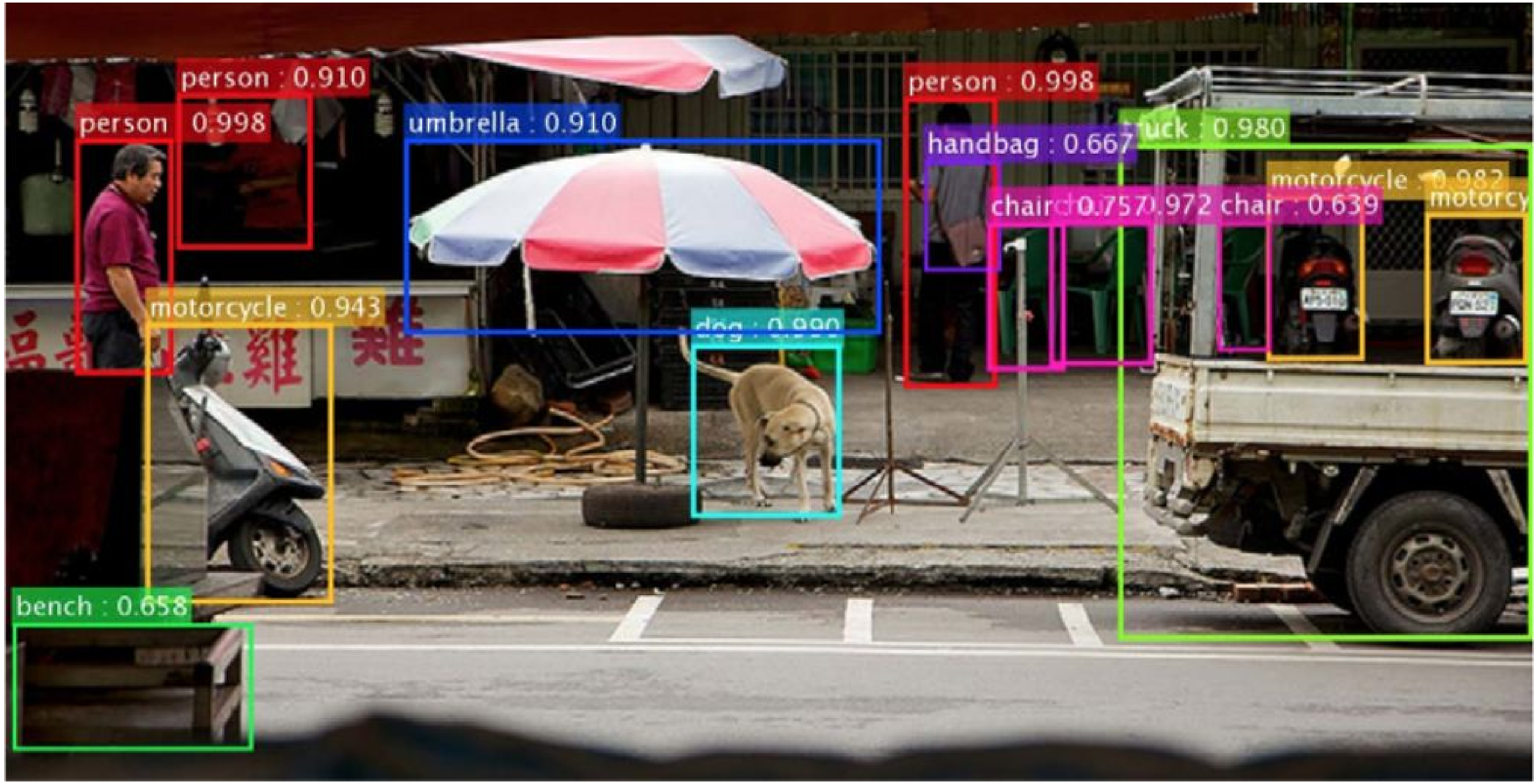
Obeject Detector Performance



Faster RCNN with ResNet



Faster RCNN with ResNet



- Faster R-CNN Original Code (Author’s Code)
  - git clone https://github.com/rbgirshick/py-faster-rcnn.git
- Available Code for TensorFlow
  - git clone https://github.com/endernewton/tf-faster-rcnn.git