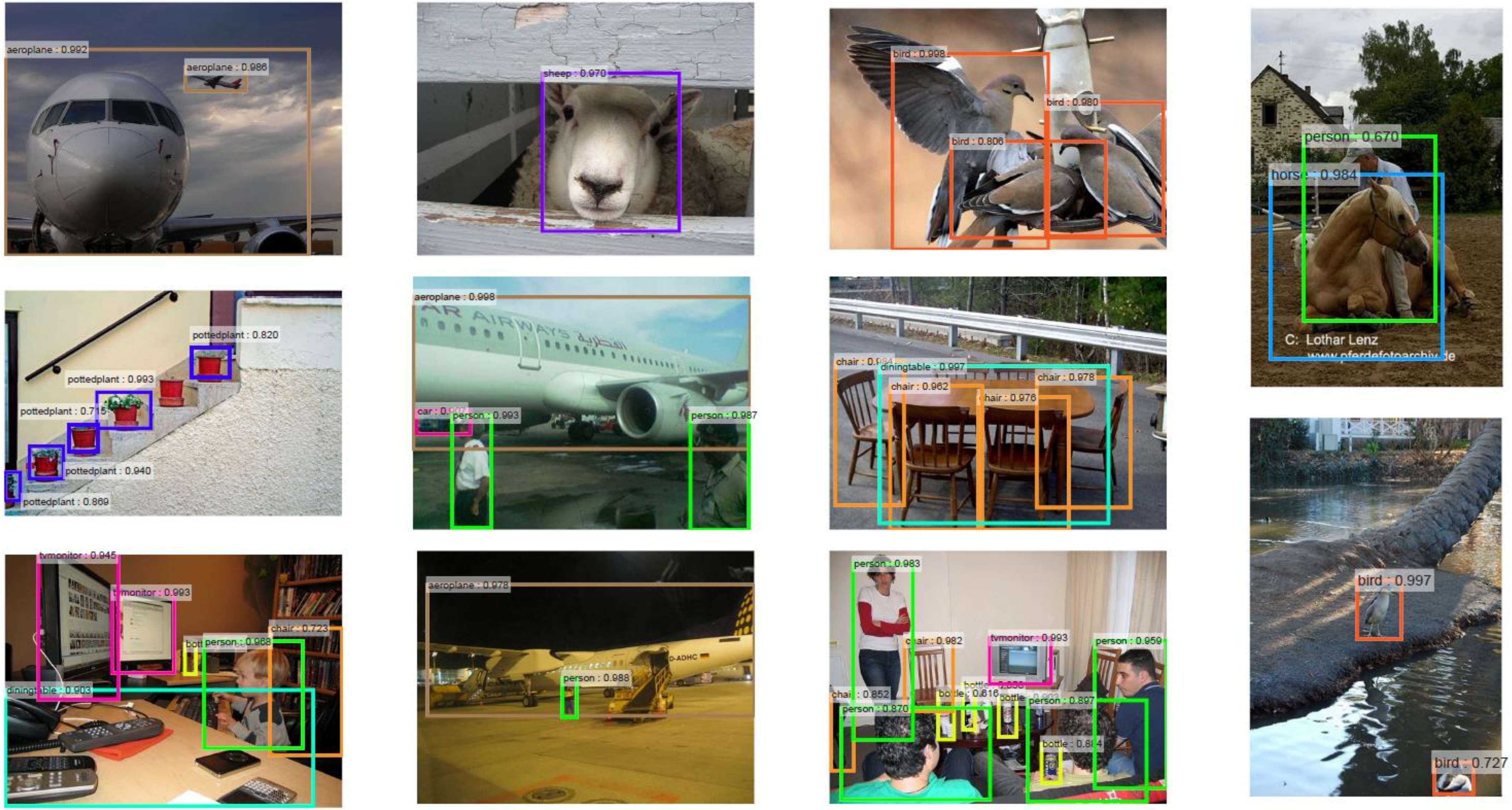
'18.08.22 송석정

DL

Object Detection

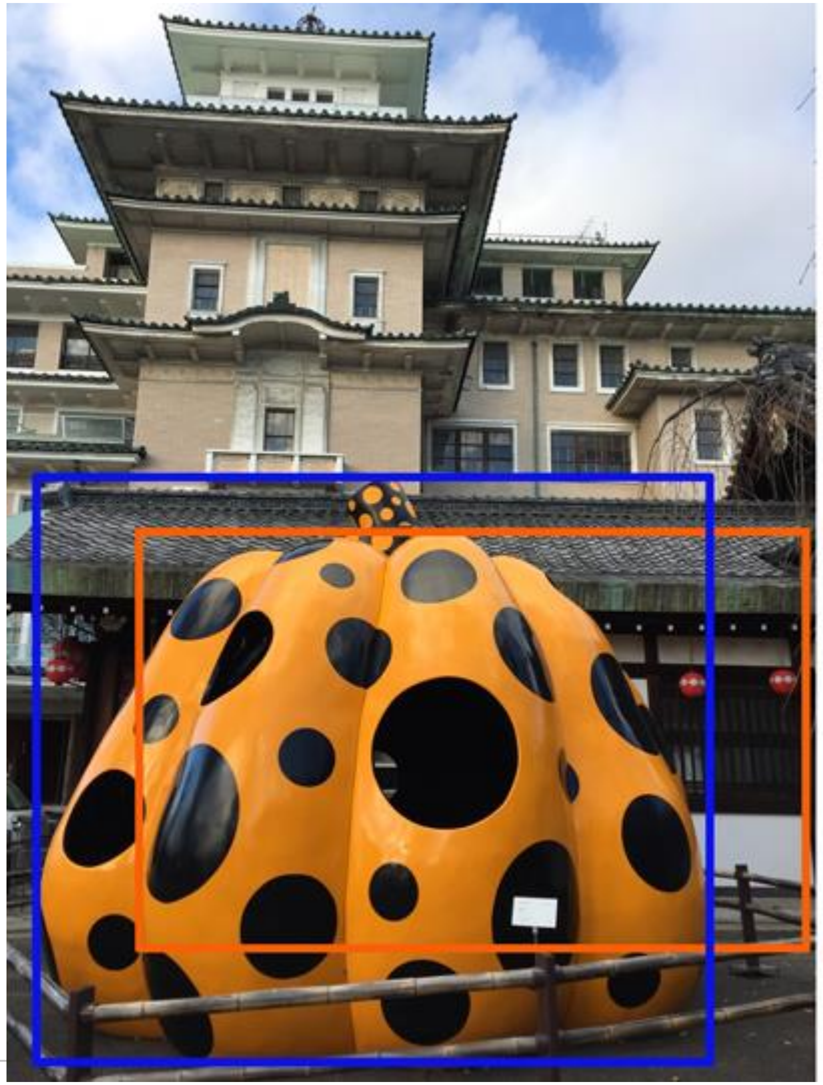
Image Detection?

주어진 image에 존재하는 object를 찾아 label과 bounding box를 출력



Intersection over Union (IoU)

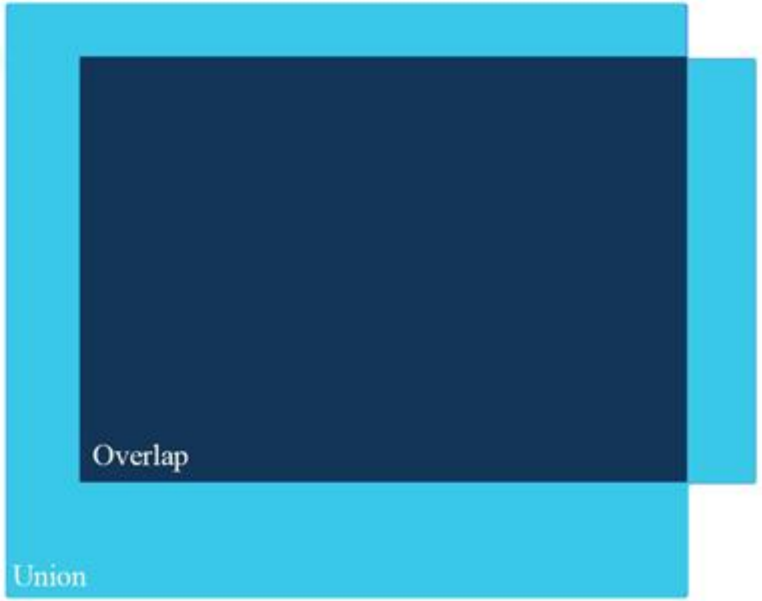
IoU measures how much overlap between 2 regions, This measures how good is our prediction in the object detector with the ground truth (the real object boundary).



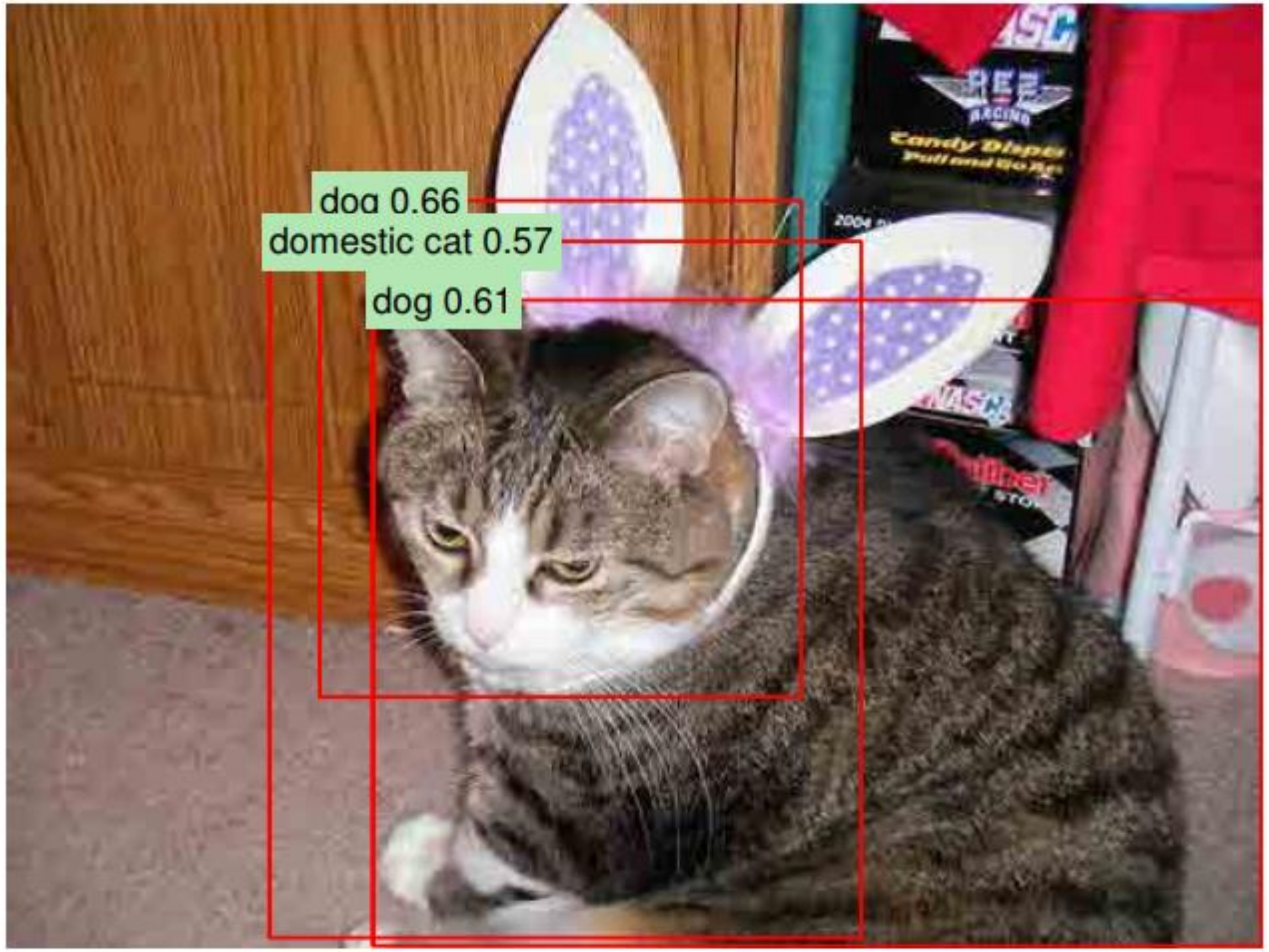
□ Ground truth

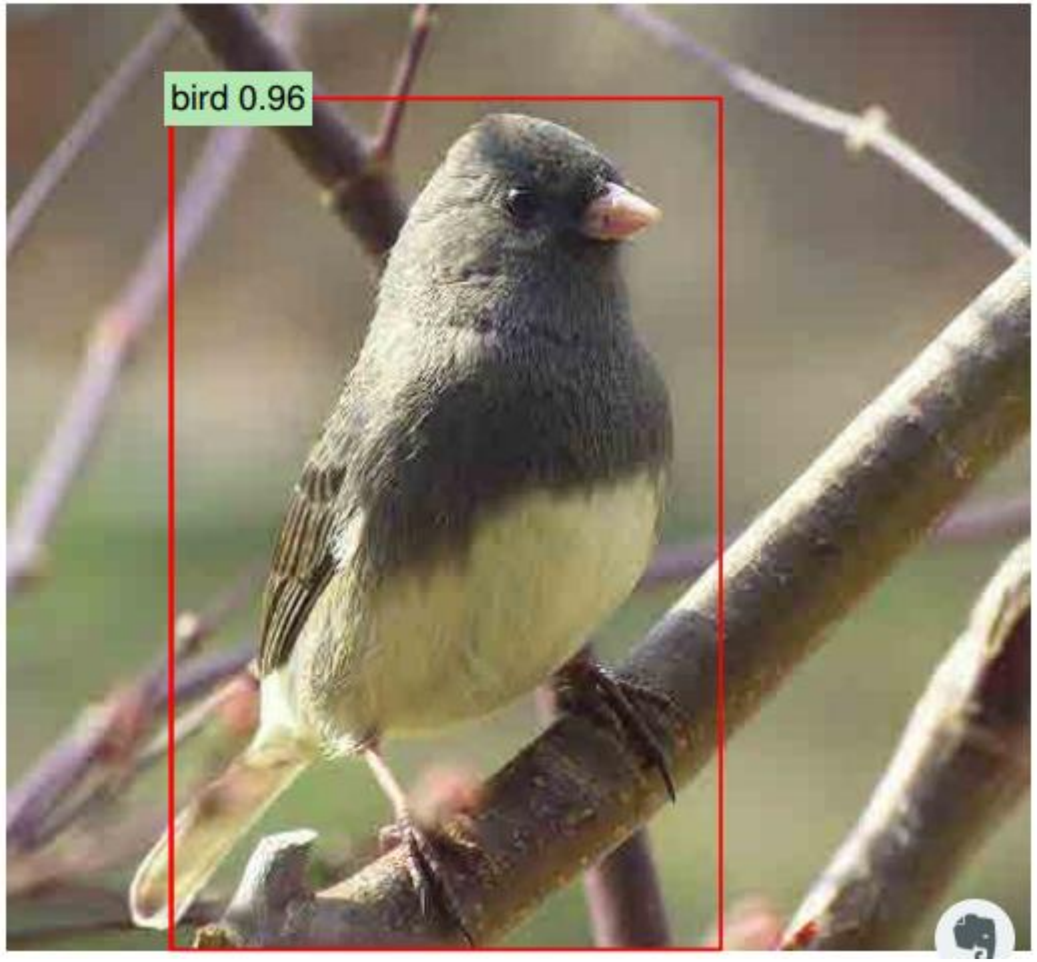
□ Prediction

$$IoU = \frac{\text{area of overlap}}{\text{area of union}}$$



Confidence level





mean Average Precision (mAP)

Object detection 의 평가 기준

It is the average of the maximum precisions at different recall values.

Precision measures how accurate is your predictions.
i.e. the percentage of your positive predictions are correct.

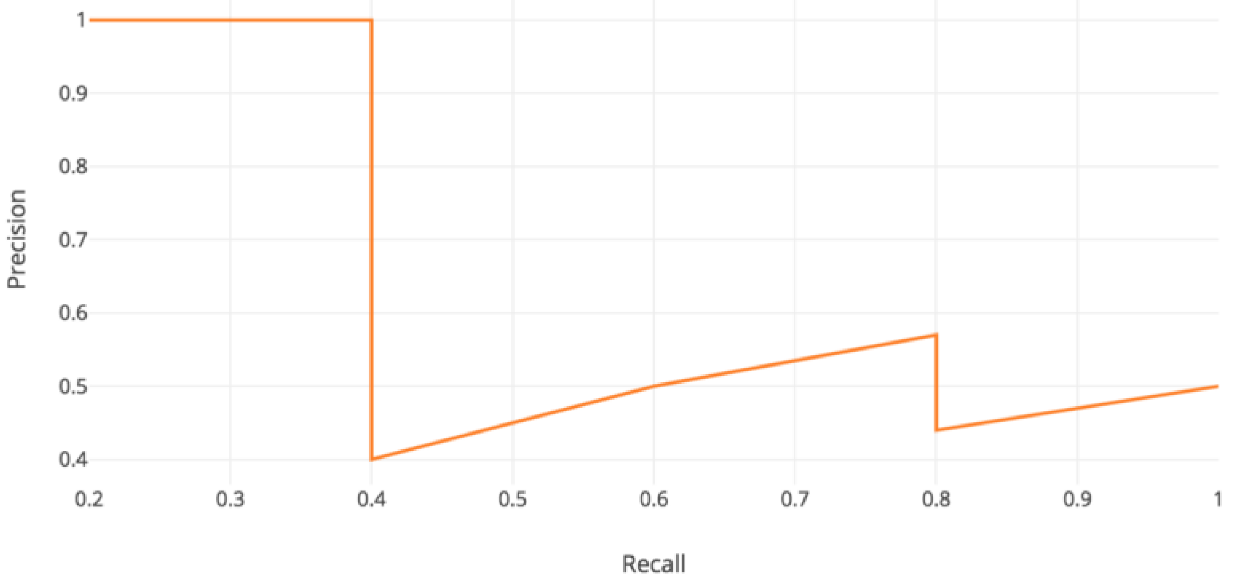
Recall measures how good you find all the positives.
For example, we can find 80% of the possible positive cases in our top K predictions.

TP = True positive
TN = True negative
FP = False positive
FN = False negative

$$Precision = \frac{TP}{TP + FP}$$

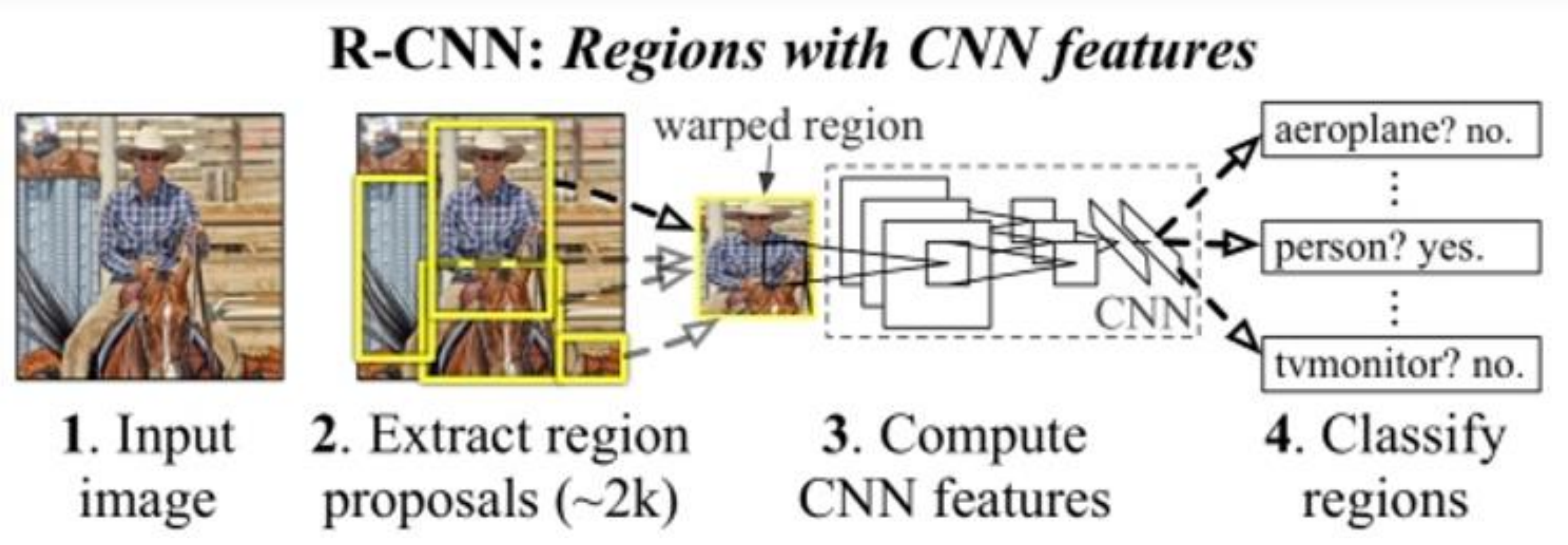
$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$



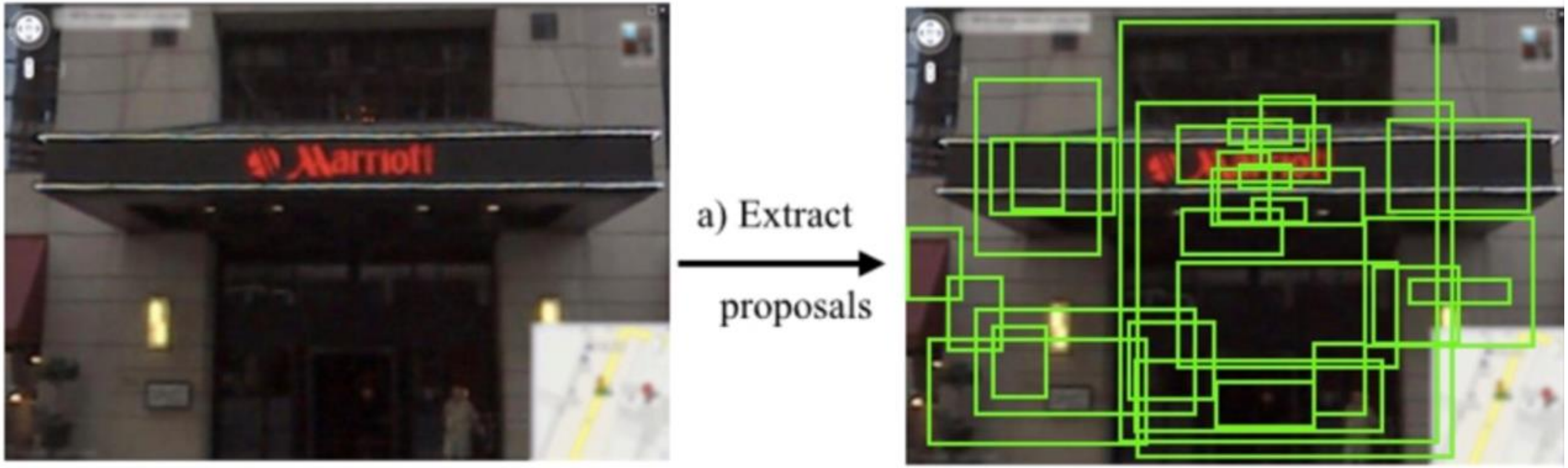
1. Region with CNN (R-CNN)

https://arxiv.org/pdf/1311.2524.pdf



1. Region proposal
2. Region warping
3. SVM

RCNN | Region proposals



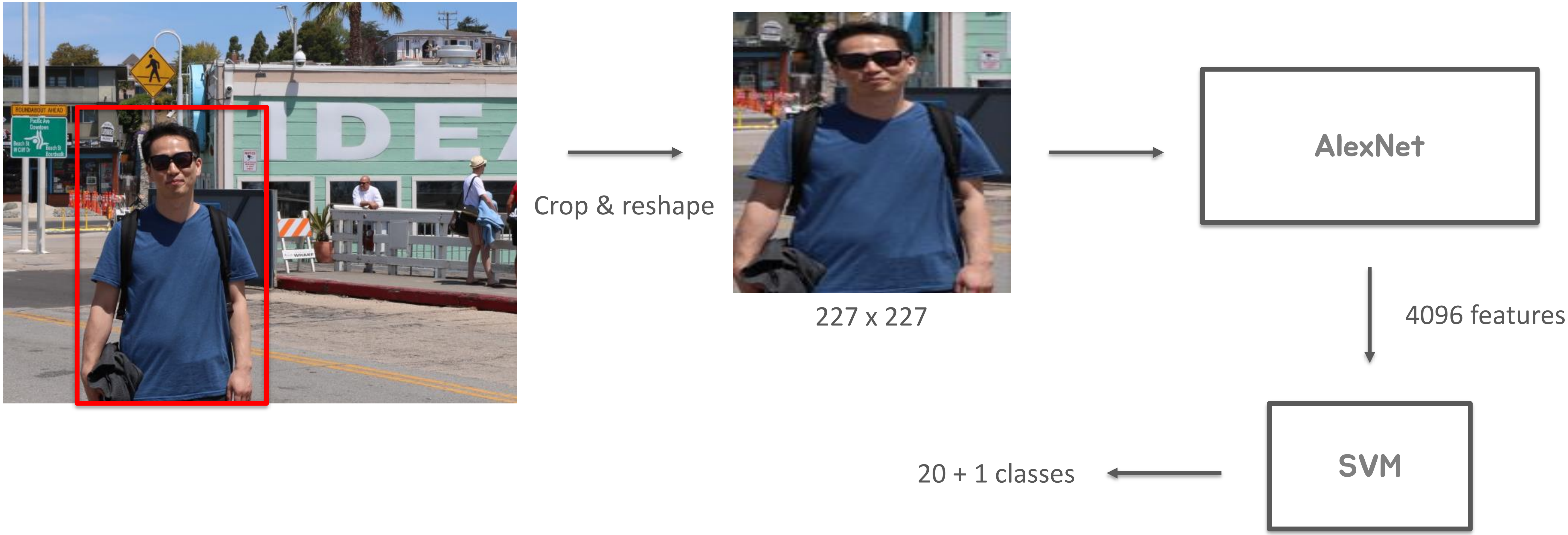
학습하지 않고, Selective Search 라는 알고리즘 사용

색상이나 강도, 패턴 등이 비슷한 인접한 픽셀을 한 region으로 묶는 방식

한 이미지당 약 2000개의 region 추출

RCNN | CNN

AlexNet 사용



Crop & reshape

227 x 227

AlexNet

4096 features

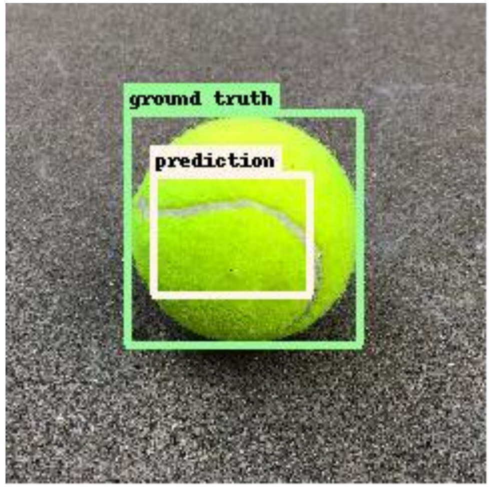
SVM

20 + 1 classes

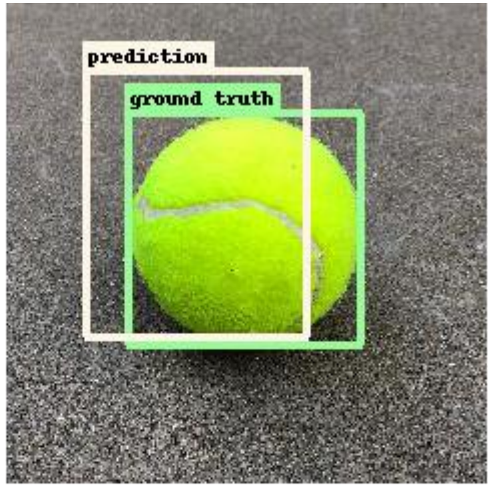
RCNN | Training

Pre-training : ILSVRC2012 data

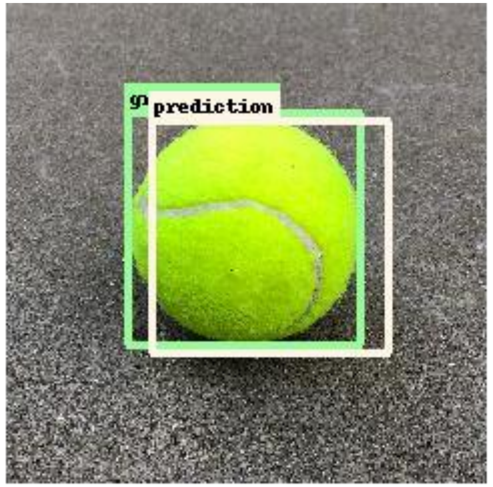
Domain-specific fine-tuning : warped region



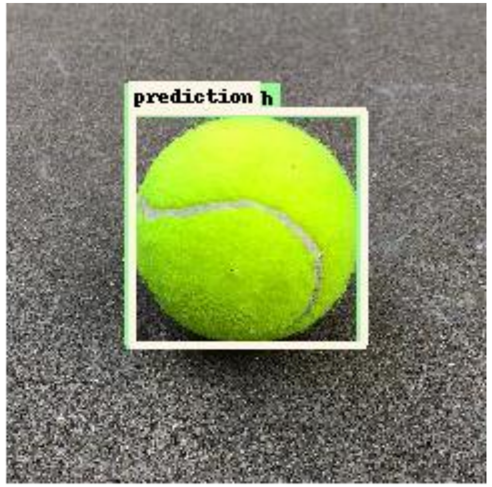
IoU: 35%



IoU: 55%



IoU: 75%



IoU: 95%

If IoU $\leq$ 0.3

→ Background

RCNN | Bounding box regression

Training pairs

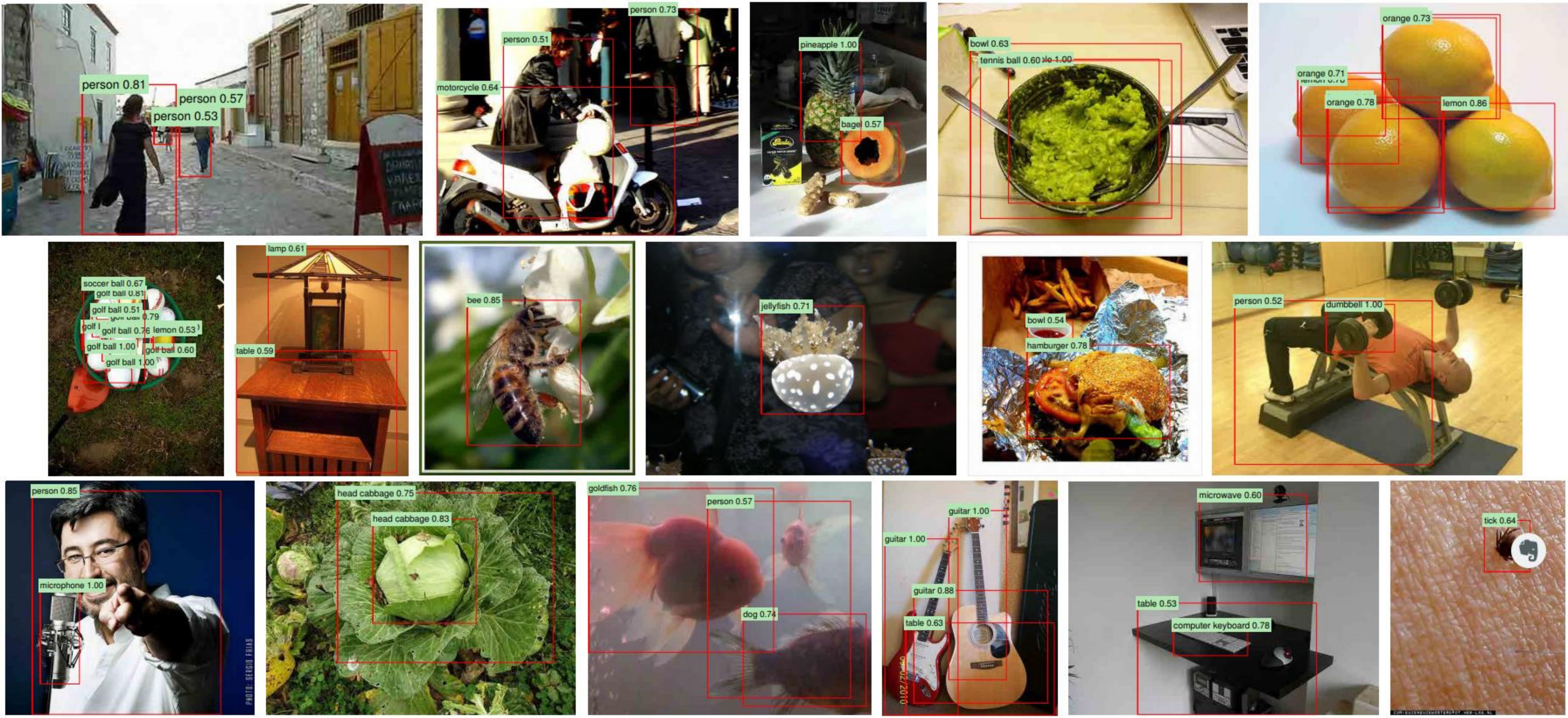
P_x, P_y, P_w, P_h and G_x, G_y, G_w, G_h

Finds

$d_x(P), d_y(P), d_w(P), d_h(P)$

Modeled as linear function of pool_5 features of proposal P

RCNN | Results



RCNN | Drawback

13s / image on a GPU

53s / image on a CPU

2. Spatial Pyramid Pooling (SPPNet)

https://arxiv.org/pdf/1406.4729.pdf

Only one CNN forward running

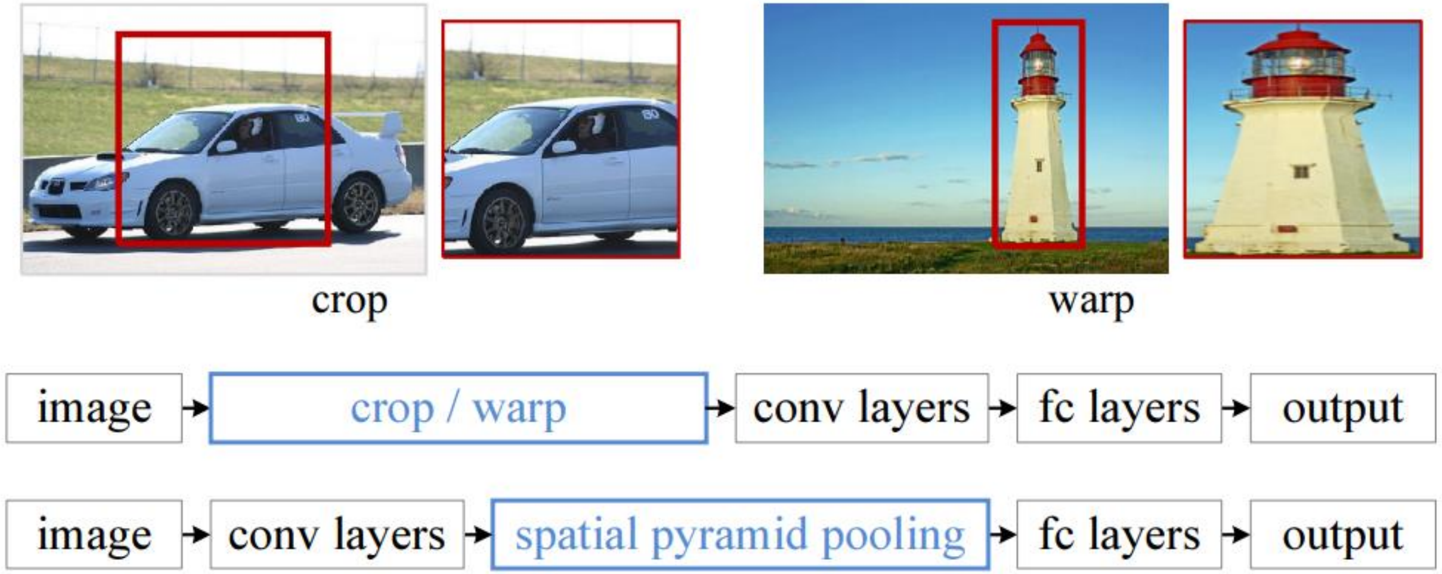


Figure 1: Top: cropping or warping to fit a fixed size. Middle: a conventional CNN. Bottom: our spatial pyramid pooling network structure.

2. Spatial Pyramid Pooling (SPPNet)

Spatial pyramid pooling

Arbitrary size to fixed size

Bounding box: extracted from input image

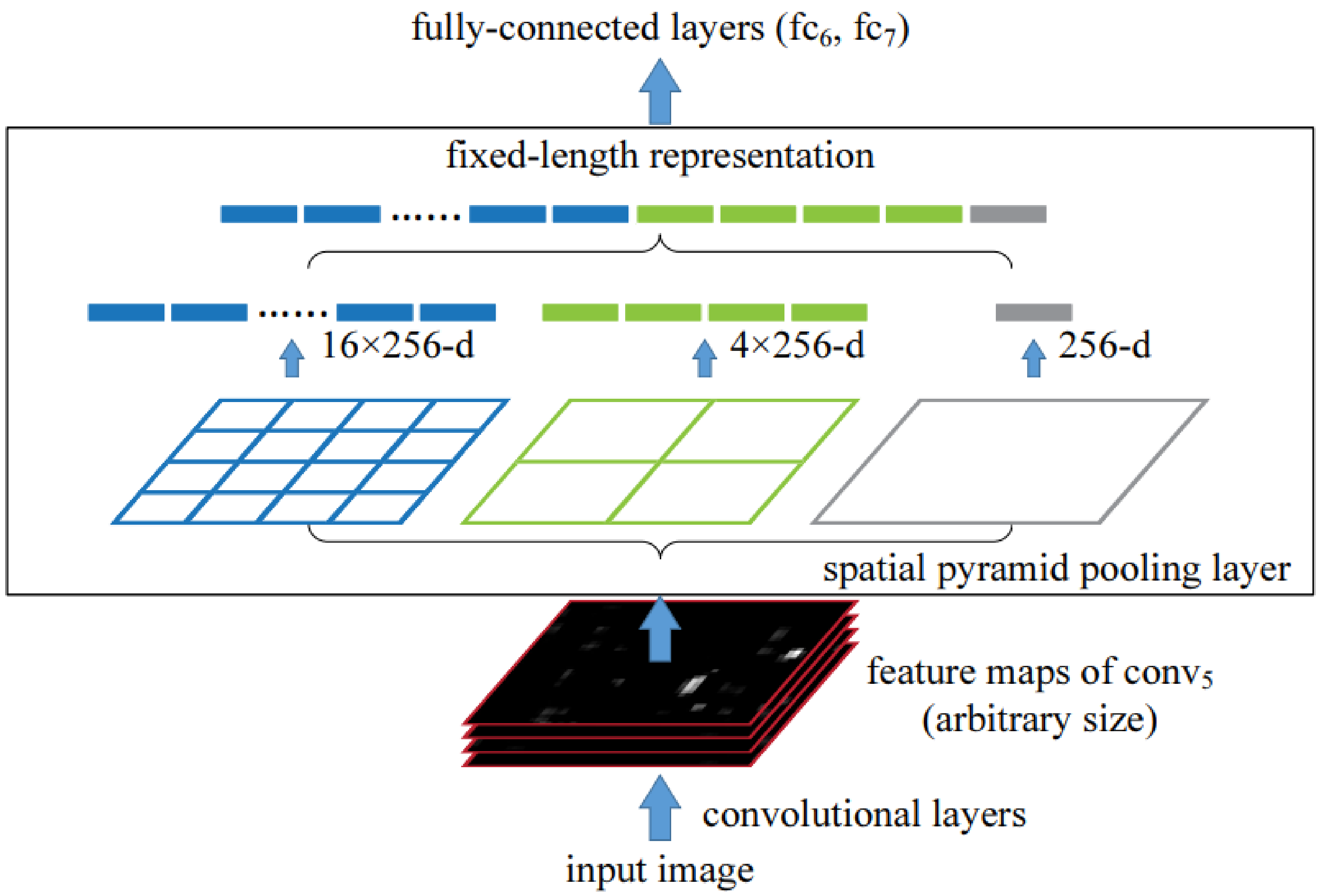

fully-connected layers (fc_6, fc_7)
fixed-length representation
16×256-d
4×256-d
256-d
spatial pyramid pooling layer
feature maps of $conv_5$
(arbitrary size)
convolutional layers
input image

Figure 3: A network structure with a **spatial pyramid pooling layer**. Here 256 is the filter number of the $conv_5$ layer, and $conv_5$ is the last convolutional layer.

https://www.cv-foundation.org/openaccess/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf

3. Fast R-CNN

Single stage training

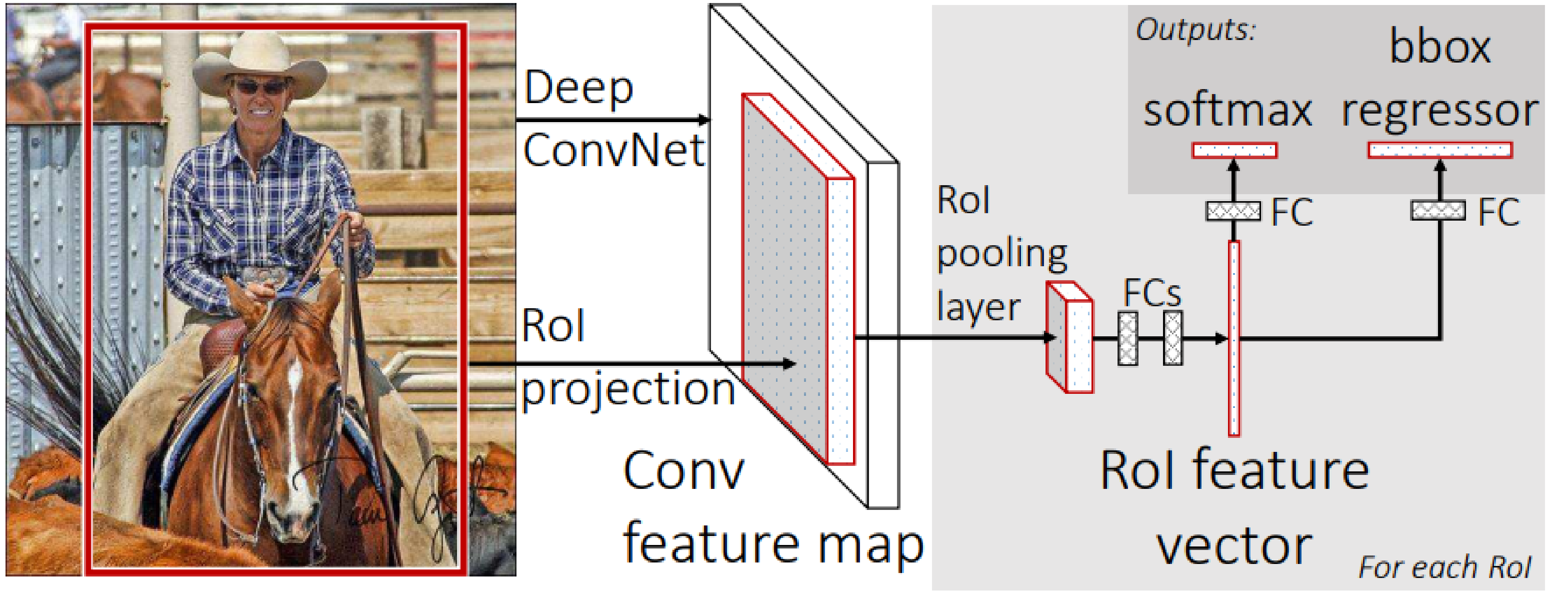


Figure 1. Fast R-CNN architecture. An input image and multiple regions of interest (RoIs) are input into a fully convolutional network. Each RoI is pooled into a fixed-size feature map and then mapped to a feature vector by fully connected layers (FCs). The network has two output vectors per RoI: softmax probabilities and per-class bounding-box regression offsets. The architecture is trained end-to-end with a multi-task loss.

https://arxiv.org/pdf/1506.01497.pdf

4. Faster R-CNN

Region proposed net + Fast R-CNN

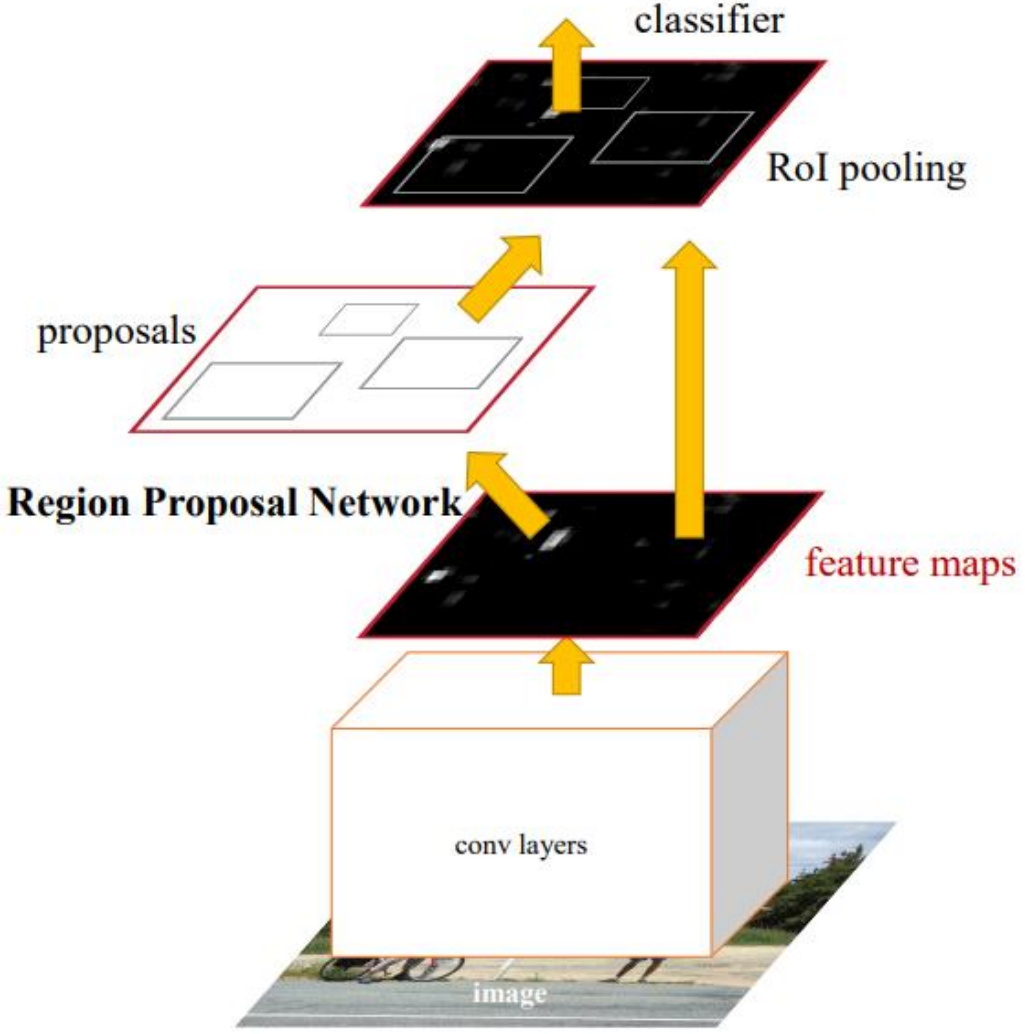


Figure 2: Faster R-CNN is a single, unified network for object detection. The RPN module serves as the 'attention' of this unified network.

https://arxiv.org/pdf/1506.01497.pdf

4. Faster R-CNN

Region proposal net + Fast R-CNN

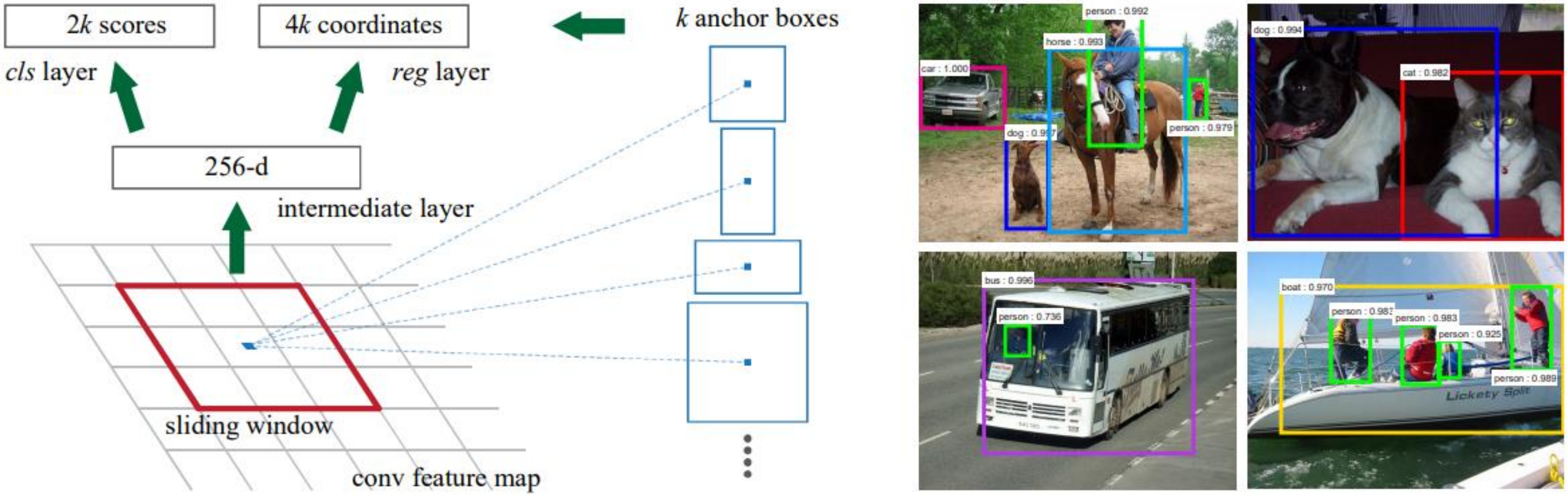


Figure 3: **Left**: Region Proposal Network (RPN). **Right**: Example detections using RPN proposals on PASCAL VOC 2007 test. Our method detects objects in a wide range of scales and aspect ratios.

https://pjreddie.com/media/files/papers/yolo.pdf

5. You Only Look Once (YOLO)

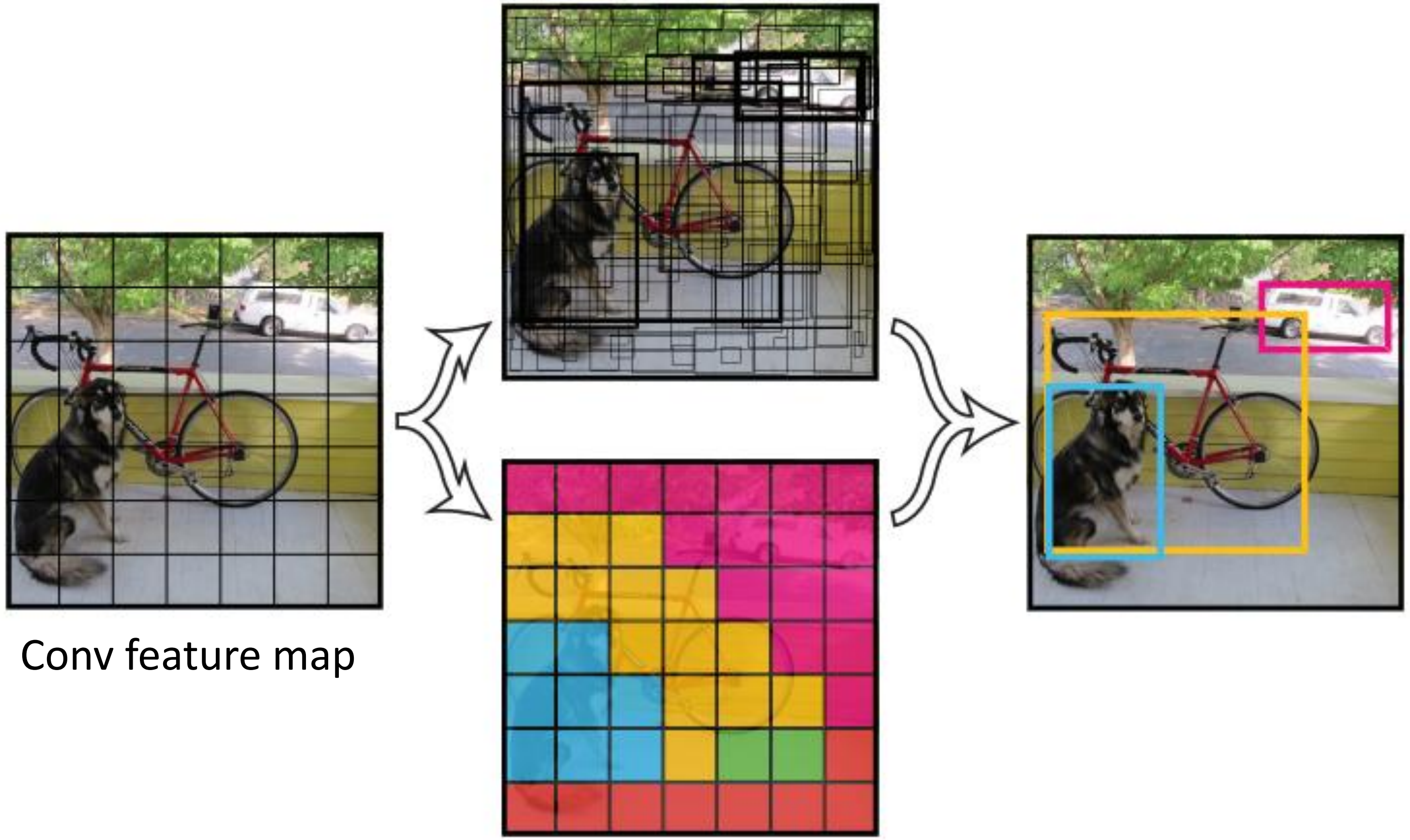
Baseline: 45 fps

Smaller version: 155 fps

5. You Only Look Once (YOLO)

Baseline: 45 fps
Smaller version: 155 fps

Classify only one class per one grid



Conv feature map

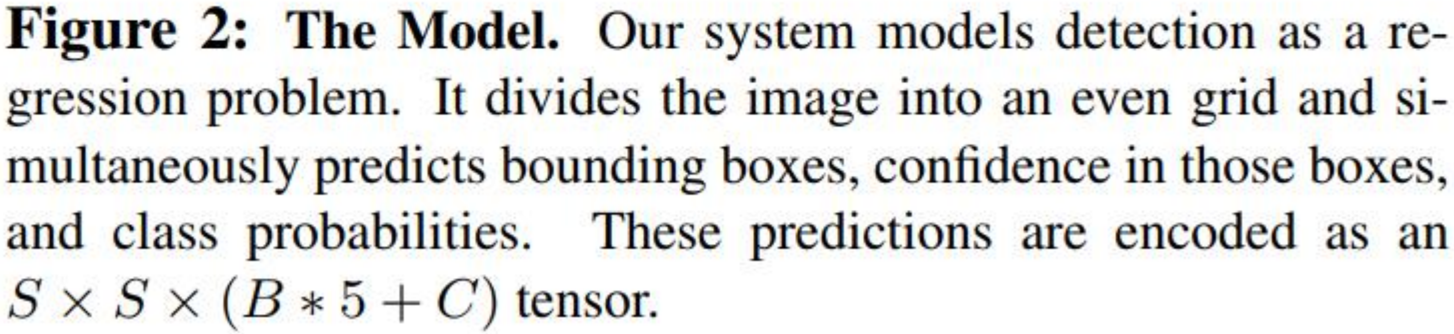


Figure 2: The Model. Our system models detection as a regression problem. It divides the image into an even grid and simultaneously predicts bounding boxes, confidence in those boxes, and class probabilities. These predictions are encoded as an $S \times S \times (B * 5 + C)$ tensor.

Bbox + confidence = 5 C: classes

5. You Only Look Once (YOLO)

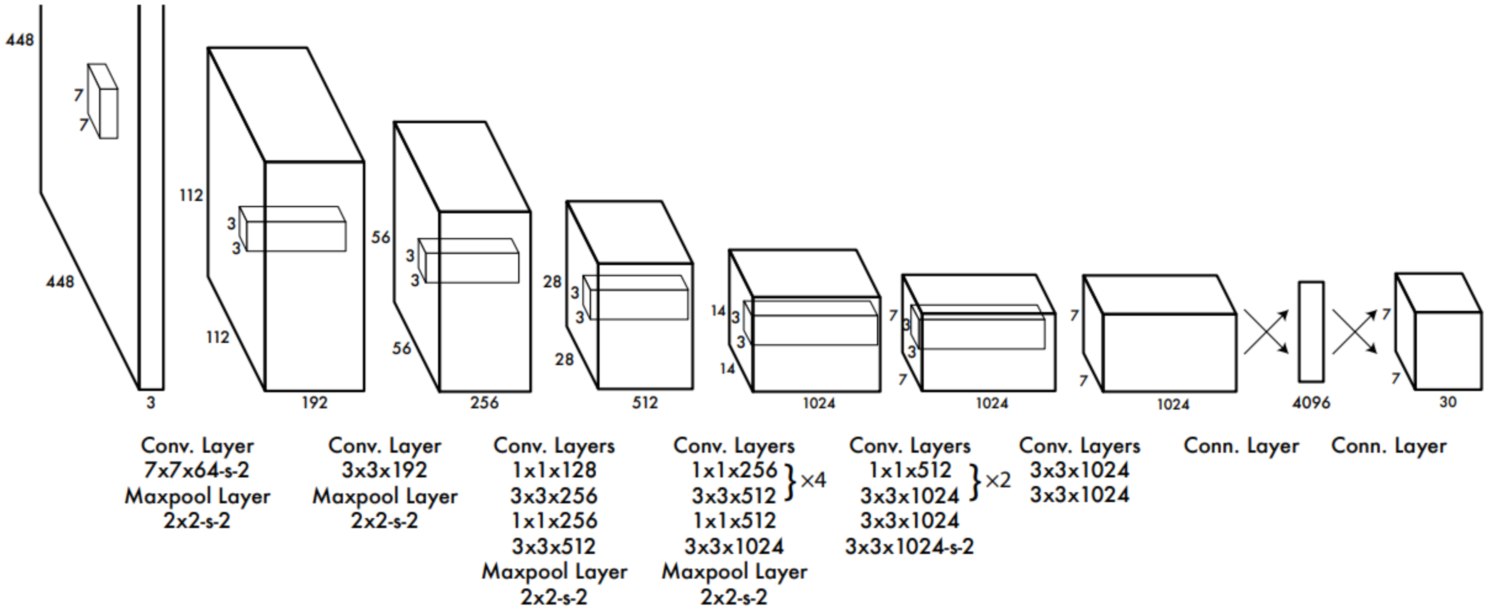


Figure 3: The Architecture. Our detection network has 24 convolutional layers followed by 2 fully connected layers. Alternating 1 × 1 convolutional layers reduce the features space from preceding layers. We pretrain the convolutional layers on the ImageNet classification task at half the resolution (224 × 224 input image) and then double the resolution for detection.

S: 7
B: 2
C: 20