Detection & Segmentation – COMP4423 Computer Vision

Xiaoyong Wei (魏驍勇)

x1wei@polyu.edu.hk

Department of Computing
電子計算學系

The Hong Kong Polytechnic University
香港理工大學

Opening Minds • Shaping the Future
啟迪思維 • 成就未來

Outline

>Object detection, Image Segmentation

>Yolo

>UNet

>R-CNN to Mask-RCNN

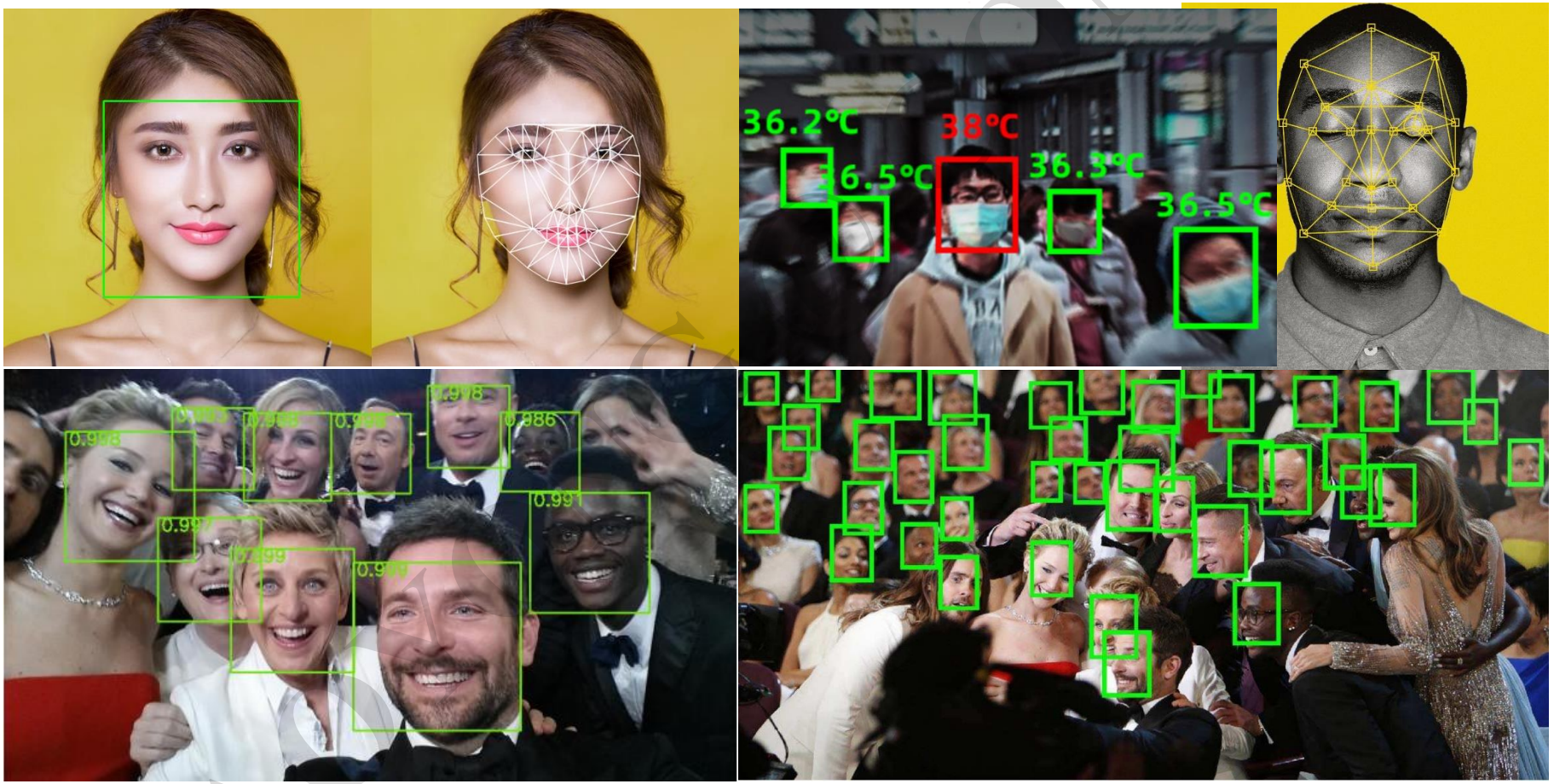
We have introduced deep learning mainly under classification/retrieval setting, because classification is where it started and retrieval is a good example to demonstrate how a classification network can be used for other tasks.

However, there are in fact a wide range of tasks in which we can apply the deep learning.

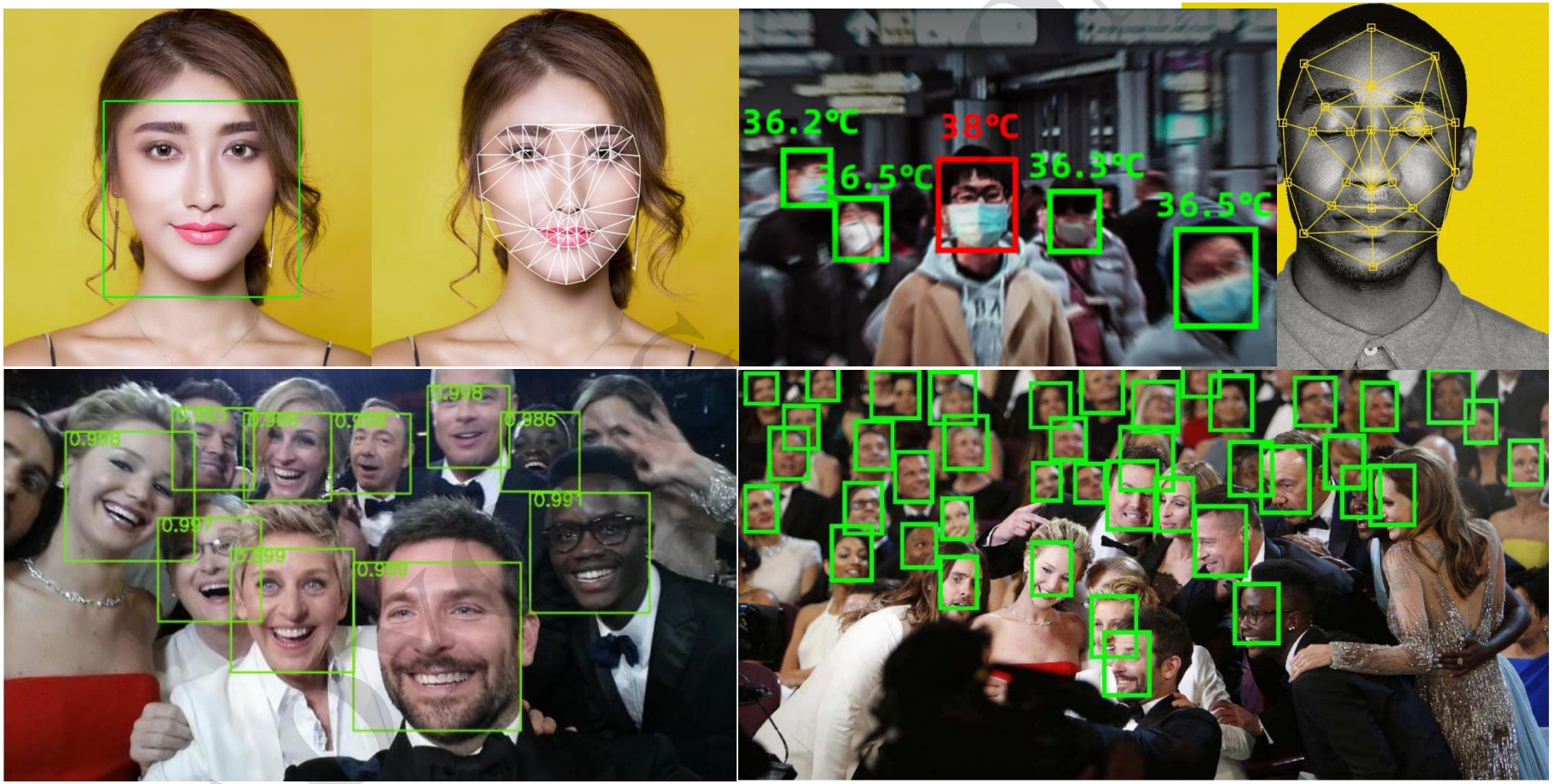
Lets' take **Detection** and **Segmentation** as another set of examples.

Recall facial recognition, while recognition can be implemented using classification models, we have to conduct face detection prior to that.

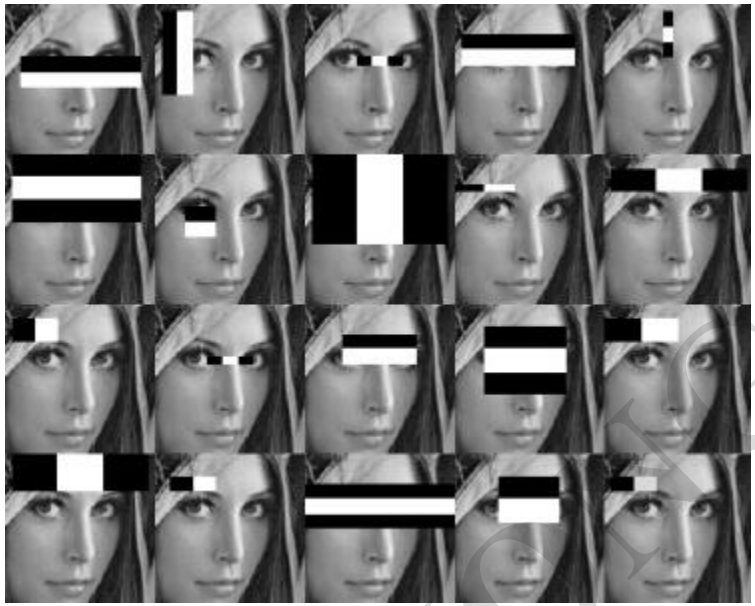
Face Detection

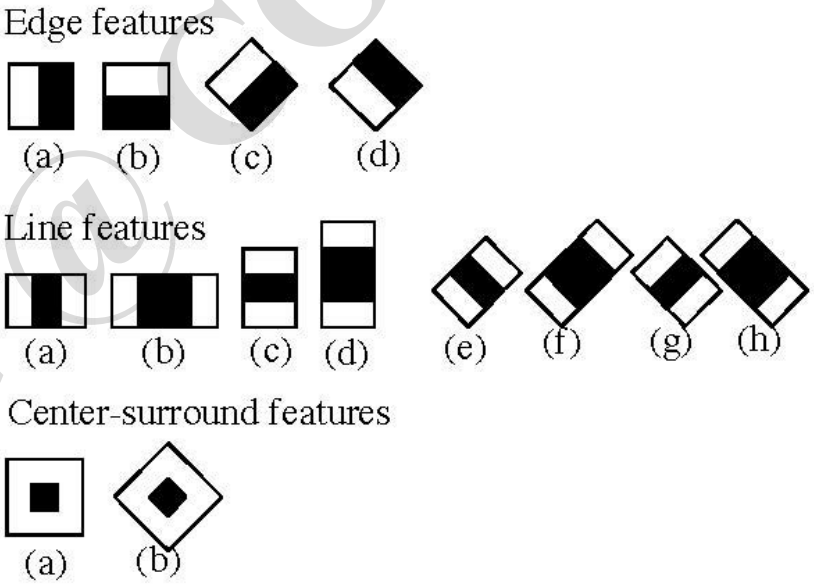


Face Detection



Face Detection





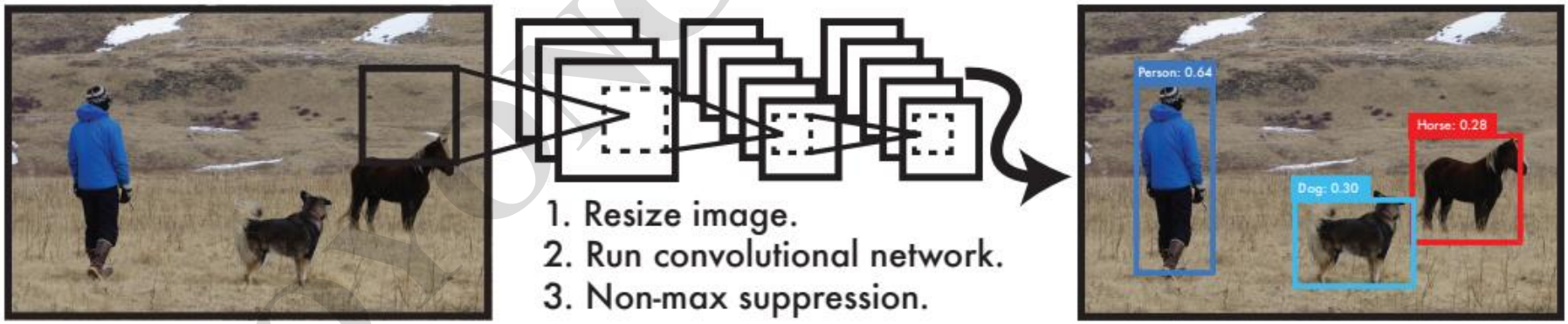
By Voila and Jones

Early methods were trying to model what they observed (e.g.,, the appearance of faces). It is a tedious work and the models are not easy to be comprehensive and generalizable.

Deep learning is a technology that we can feed the data and let the learning figure out the characteristics of the objects by itself. Why don't we use it for detection?

Yolo (You Only Look Once)

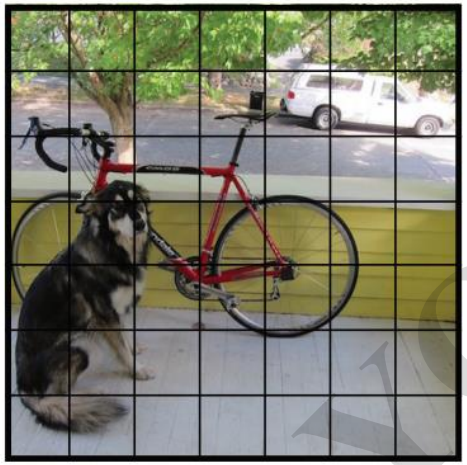
- YOLO is used to predict **what** objects are present and **where** they are
- YOLO applies a single convolutional neural network (CNN) to the entire image, divides the image into grids, and predicts the class probabilities and bounding boxes of each grid.



Redmon J , Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection[J]. IEEE, 2016.

Yolo (You Only Look Once)

- YOLO divides the image into an S × S grid (Figure 1)
- For each grid cell, it predicts B bounding boxes (Figure 2). Each box is with a confidence
- For each grid cell, there are C class probabilities.(Figure 3)



S × S grid on input

Figure 1

Bounding boxes + confidence

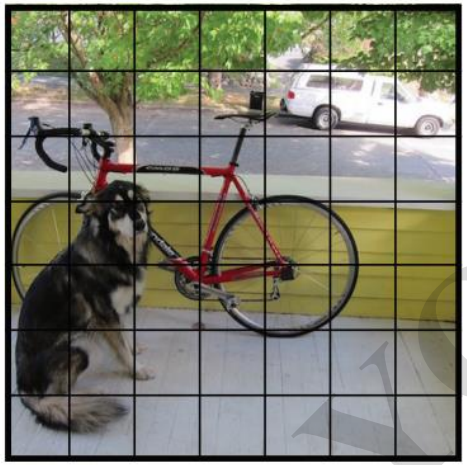
Figure 2

Class probability map

Figure 3

Yolo (You Only Look Once)

- YOLO divides the image into an S × S grid (Figure 1)
- For each grid cell, it predicts B bounding boxes (Figure 2). Each box is with a confidence
- For each grid cell, there are C class probabilities.(Figure 3)



S × S grid on input

Figure 1

Bounding boxes + confidence

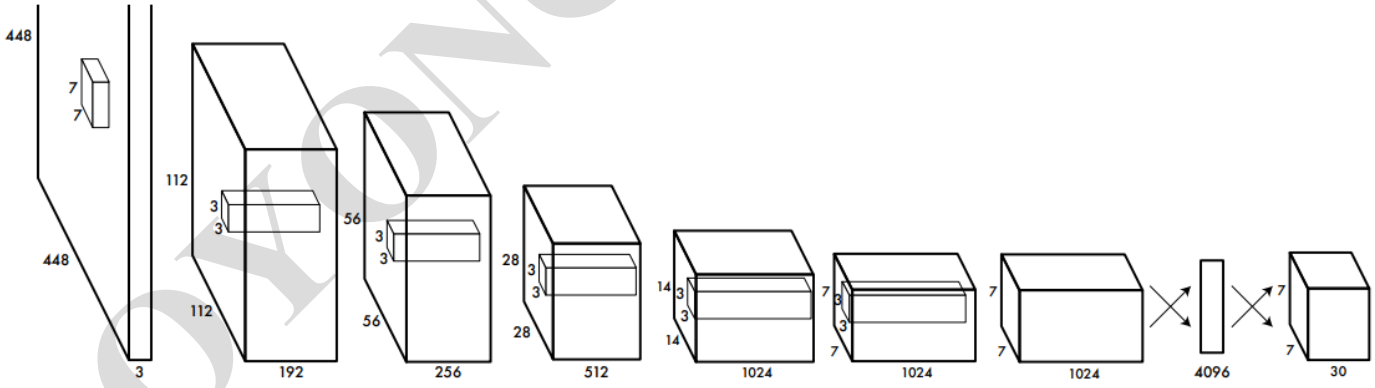
Figure 2

Class probability map

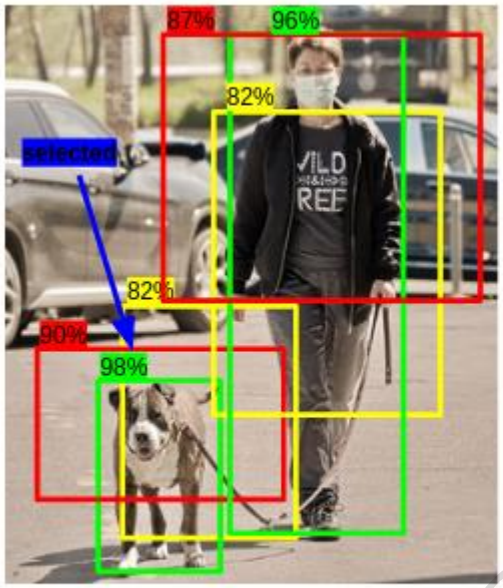
Figure 3

Yolo (You Only Look Once)

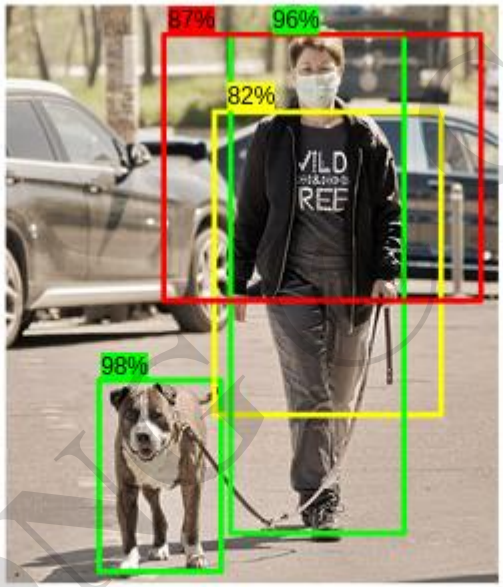
- These predictions are encoded as an S × S × (B ∗5 + C) tensor.
- For example, in YOLO, we use S = 7, B = 2, C = 20. So our prediction is a 7 × 7 × 30 tensor. Then we have a 7×7×2 = 98 Bounding boxes. For each box, it is with a 30-dimensional vector, which consists of 20 classification possibilities, the confidence of 2 boxes, and the positions of 2 boxes (each box requires 4 values)



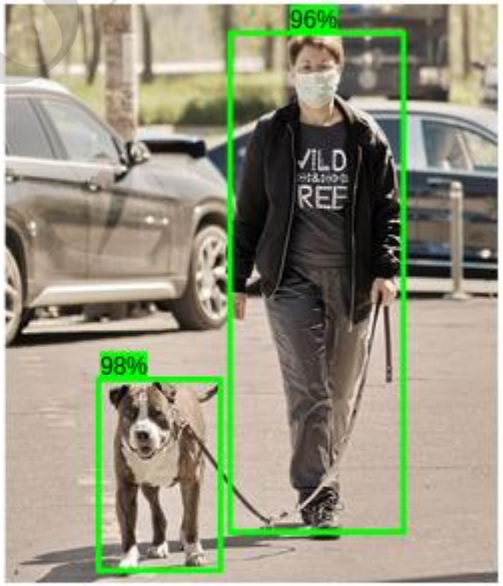
Yolo – Non-Max Suppression



Step 1: Selecting Bounding box with highest score



Step 3: Delete Bounding box with high overlap

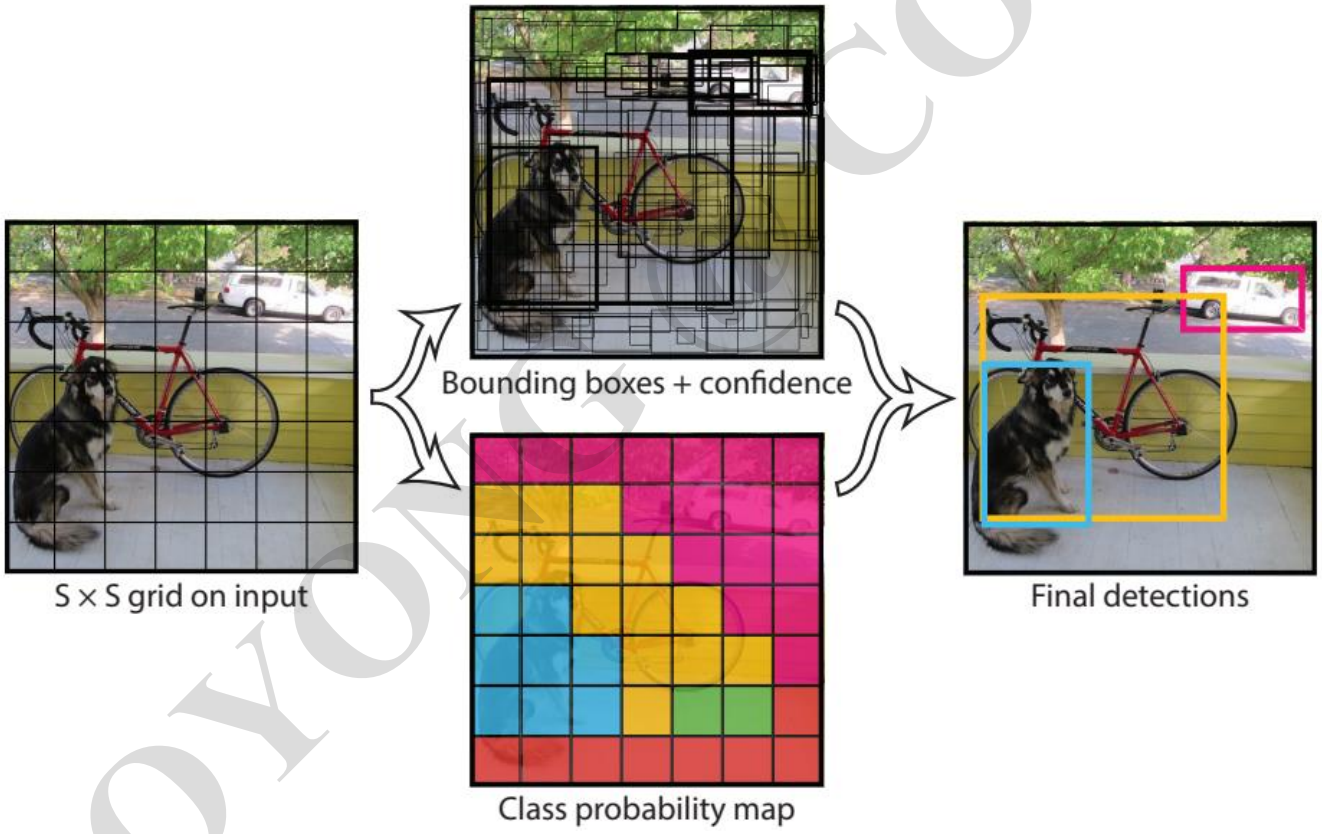


Step 5: Final Output

Objectiveness and Overlap of boxes

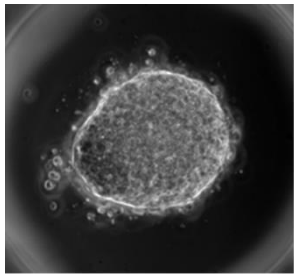
https://www.analyticsvidhya.com/blog/2020/08/selecting-the-right-bounding-box-using-non-max-suppression-with-implementation/

Yolo (You Only Look Once)

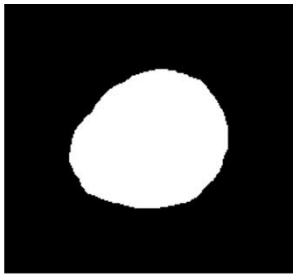


In a lot of applications, we need to find the exact locations of the objects/areas (e.g., lesions in medical images). The predictions are masks rather than bounding boxes.

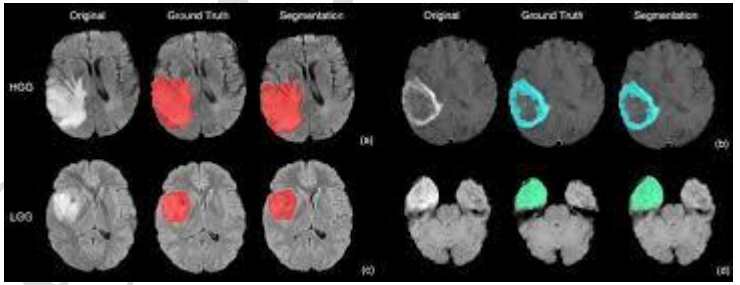
Image Segmentation

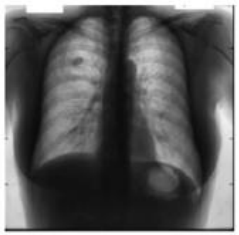


Input Image

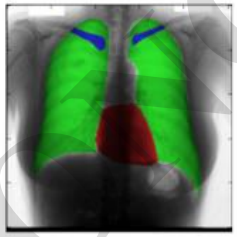


Output Mask

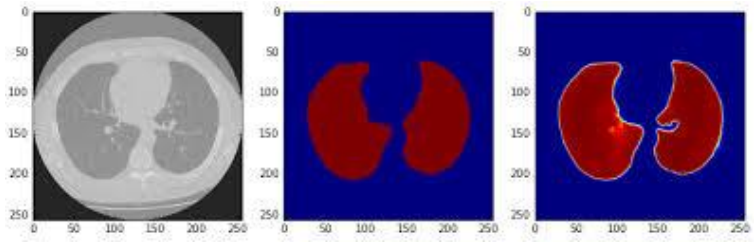




Input Image

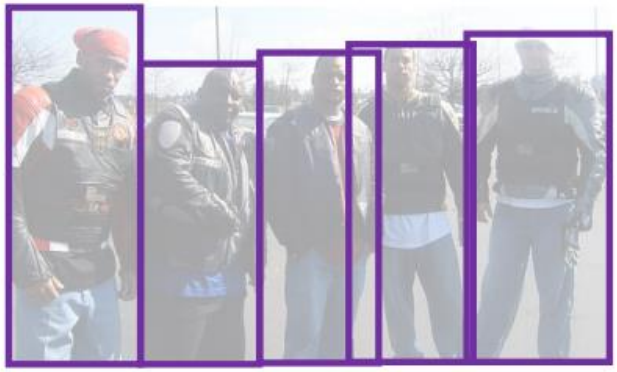


Segmented Image

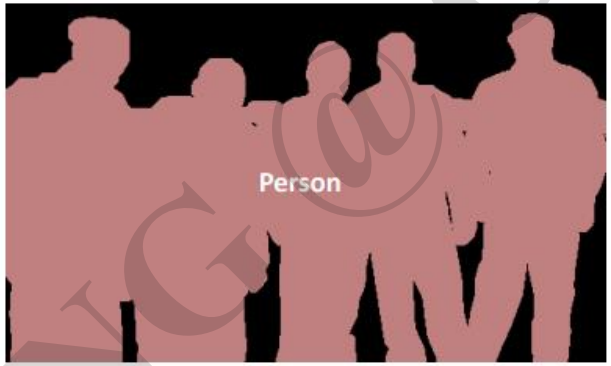


https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47

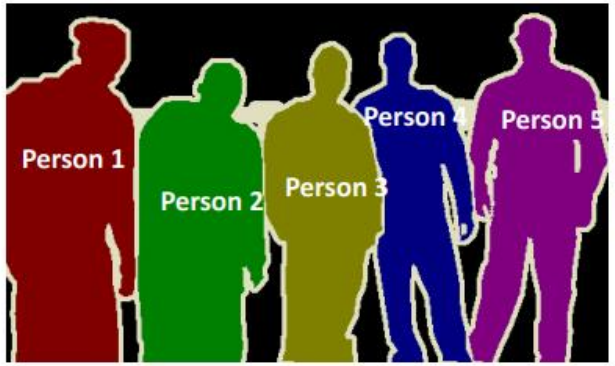
Image Segmentation



Object Detection



Semantic Segmentation

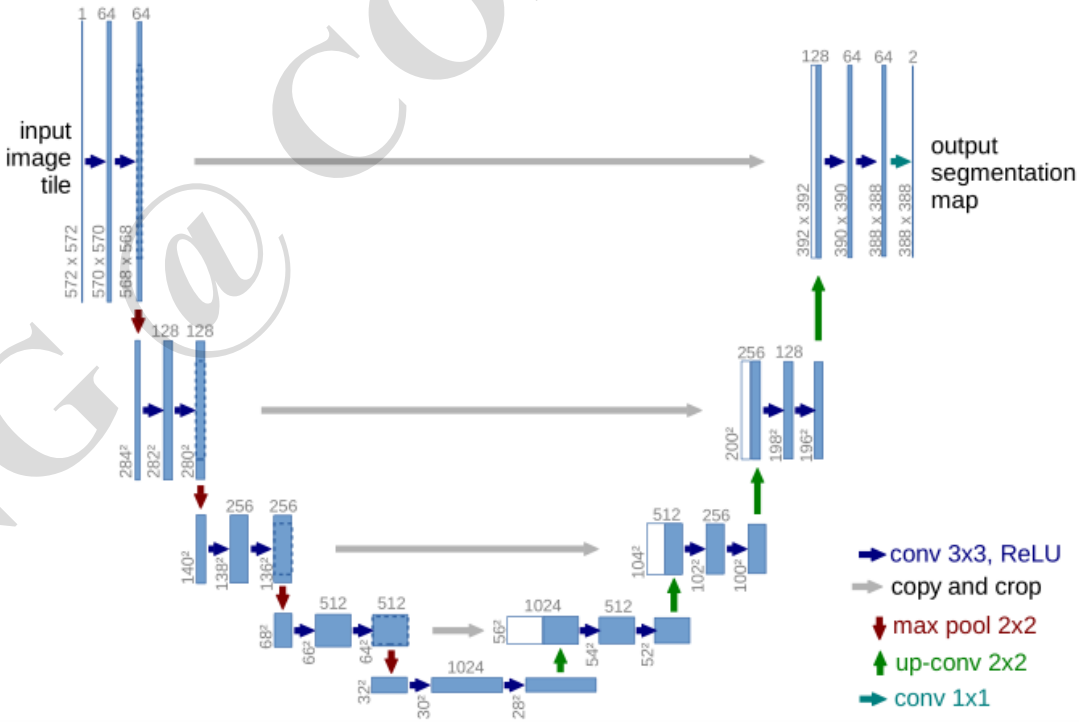


Instance Segmentation

https://towardsdatascience.com/understanding-semantic-segmentation-with-unet-6be4f42d4b47

U-Net

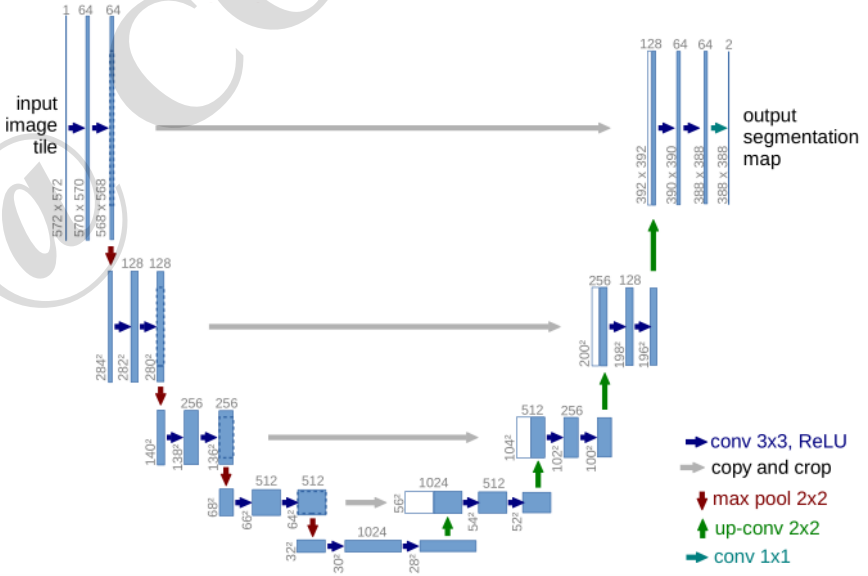
- The U-Net architecture stems from the so-called “fully convolutional network”, which is designed to solve the problem of medical image segmentation.
- The U-Net means that the left part of U is a down-sapling, and the right part of U is an upsampling, which gives it the u-shaped architecture.



Source: Ronneberger O , Fischer P , Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. Springer, Cham, 2015.

U-Net

- The down-sampling is a typical convolutional network, which consists of repeated application of convolutions (the red arrow), each followed by a rectified linear unit (ReLU) and a max pooling operation. During the down-sampling , the spatial information is reduced while the feature information is increased.
- The upsampling combines the feature and spatial information through a sequence of up-convolutions (the green arrow) and concatenations with high-resolution features from the left (the grey arrow).

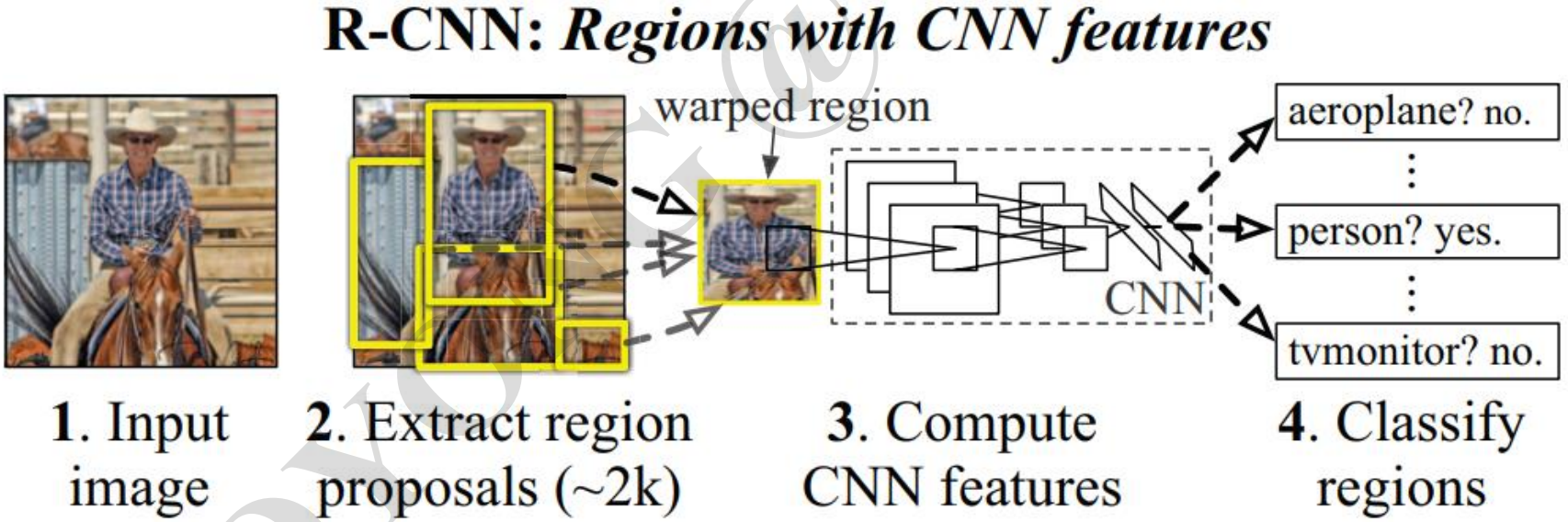


Source: Ronneberger O , Fischer P , Brox T . U-Net: Convolutional Networks for Biomedical Image Segmentation[J]. Springer, Cham, 2015.

To detect the bounding boxes or masks?

RCNN

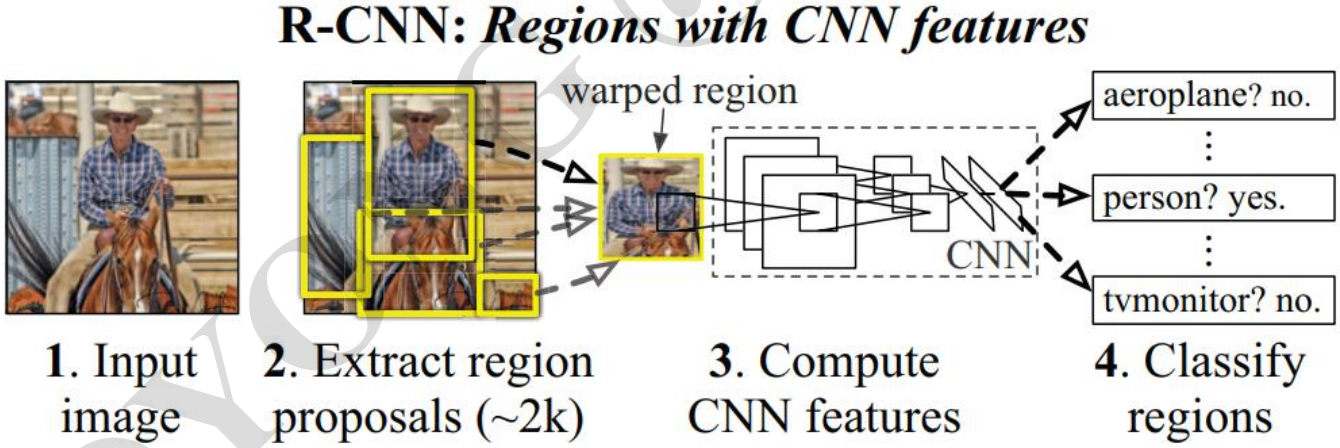
- RCNN is a milestone in the field of target detection. It is the first attempt of target detection using CNNs. It is with superior performance over those of traditional feature extraction methods.



Girshick R , Donahue J , Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. IEEE Computer Society, 2013.

RCNN

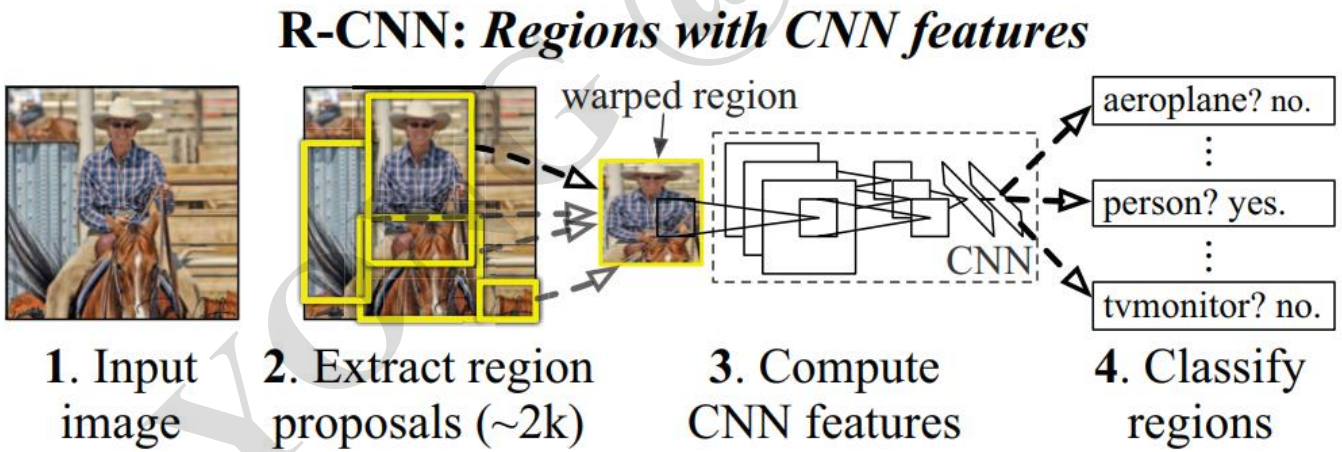
- 1. Input images
- 2. Region proposals: 1K~2K candidate regions were generated for each image (Using Selective Search); Different from the traditional Sliding Window approaches, Selective Search generates regional recommendations based on objectiveness



Girshick R , Donahue J , Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. IEEE Computer Society, 2013.

RCNN

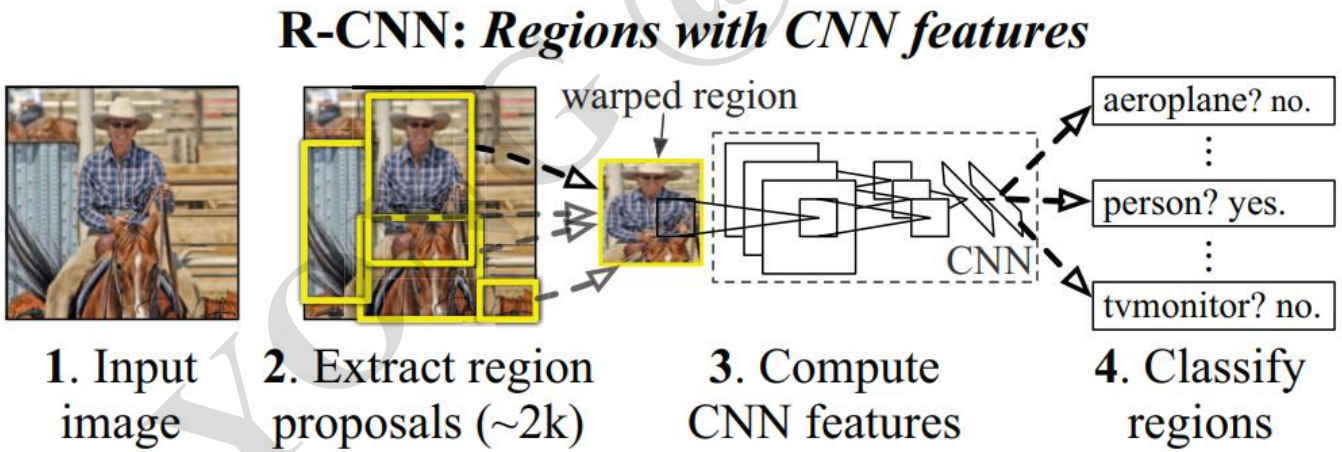
- 3. Feature extraction: for each candidate region, a deep network is used to extract features (e.g., AlextNet, VGG and other CNNs);
- 4. SVM classification : Send the features into a SVM classifier



Girshick R , Donahue J , Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. IEEE Computer Society, 2013.

RCNN

- 5. Bounding box regression based on the region proposals of the SVM



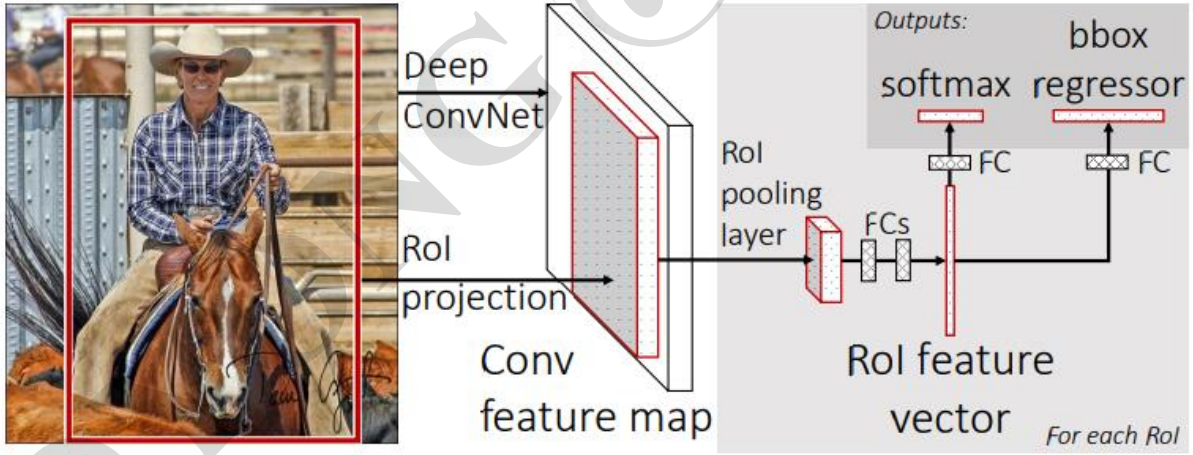
Girshick R , Donahue J , Darrell T , et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. IEEE Computer Society, 2013.

RCNN -> Fast RCNN

- RCNN:
  - Takes too long and consumes too much space

RCNN -> Fast RCNN

- Feed the whole image into the CNN
- Extract the feature for each region from the feature map directly
- Border regression is embedded into the TRAINING of the network



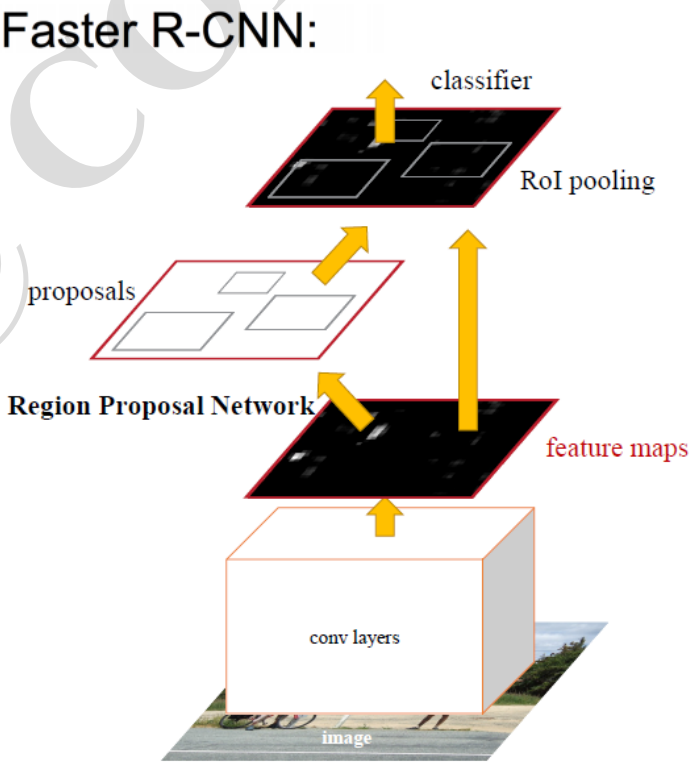
Girshick R . Fast R-CNN[J]. arXiv e-prints, 2015.

Fast RCNN -> Faster RCNN

- Fast-RCNN
  - Selective Search method takes too long

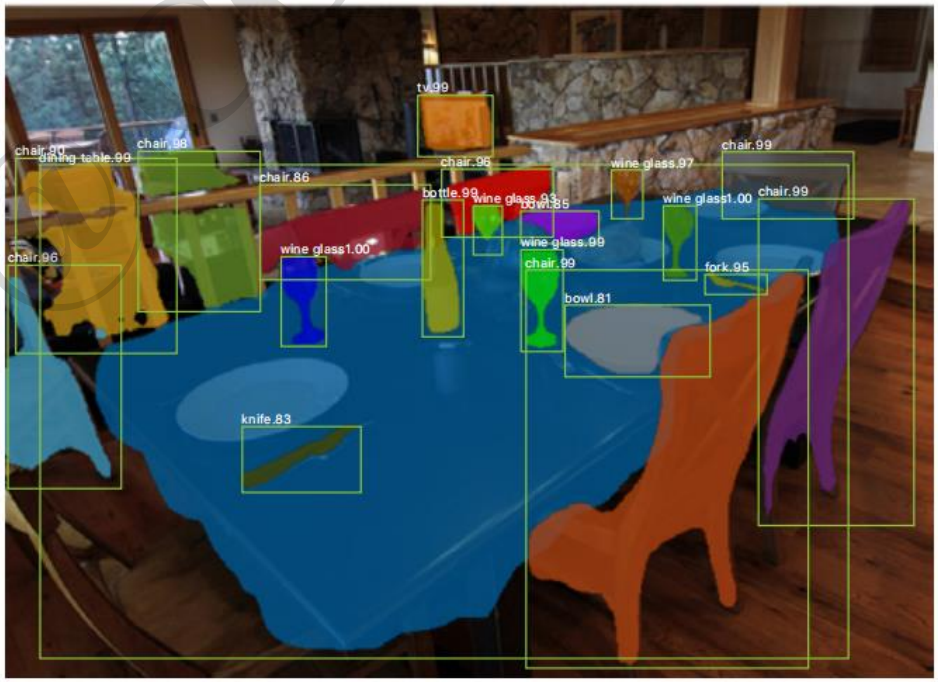
Fast RCNN -> Faster RCNN

- Region Proposal Network (RPN) is used to generate the regions
- The process of finding candidate boxes is integrated into the neural network



Ren S , He K , Girshick R , et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.

Mask RCNN: The latest RCNN



He K , Gkioxari G , Piotr Dollár, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017.

Mask RCNN: The latest RCNN

- Faster-RCNN is a object detection algorithm
- Mask RCNN is not a object detection algorithm, but a semantic segmentation algorithm

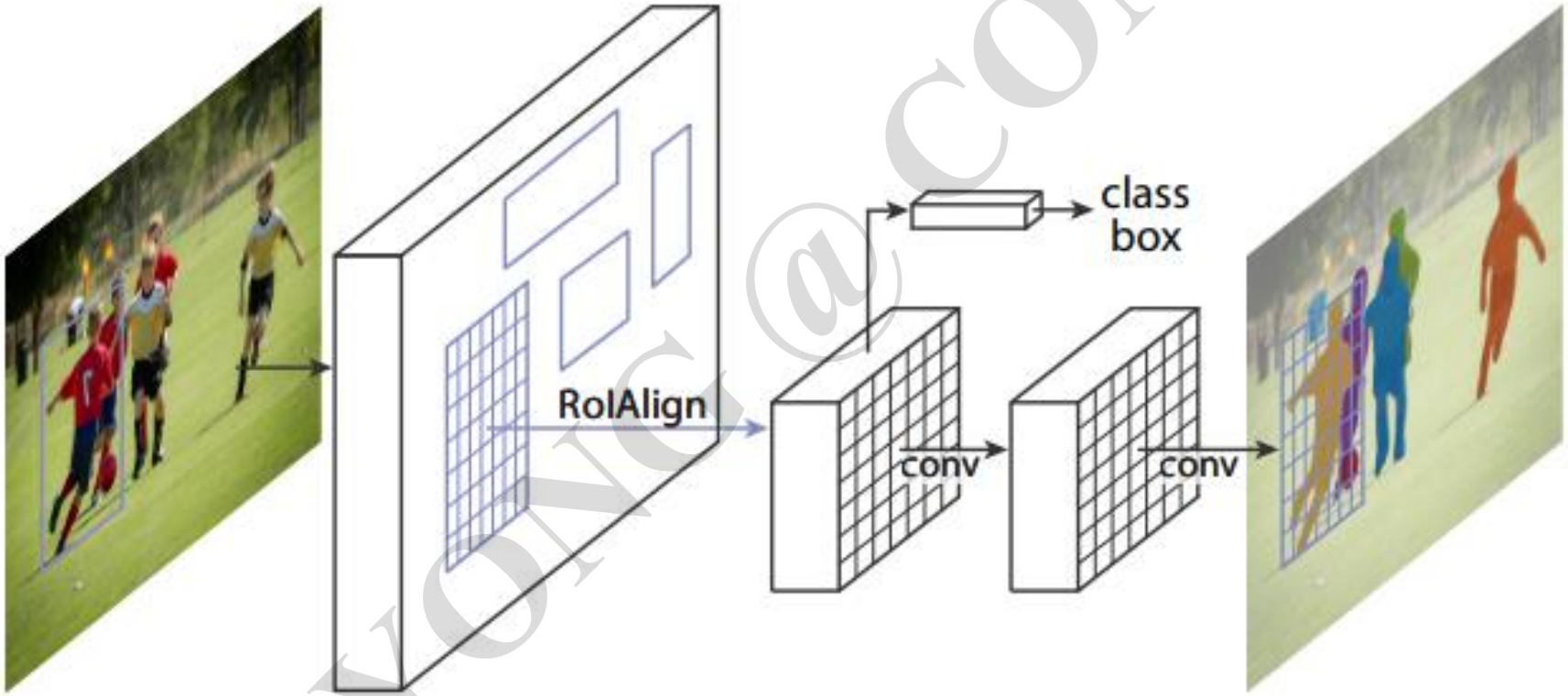
He K , Gkioxari G , Piotr Dollár, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017.

Mask RCNN: The latest RCNN

- ROI Align is used to replace ROI pooling to improve the accuracy of instance segmentation
- The VGG network used in Faster RCNN was replaced with ResNet+FPN. FPN is used to mine multi-scale information, and ResNet combines low-level features with high-level features to facilitate detailed detection.
- FCN layer (Mask layer) is added for semantic segmentation.

He K , Gkioxari G , Piotr Dollár, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017.

Mask RCNN: The latest RCNN



He K , Gkioxari G , Piotr Dollár, et al. Mask R-CNN[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017.

Thank you!