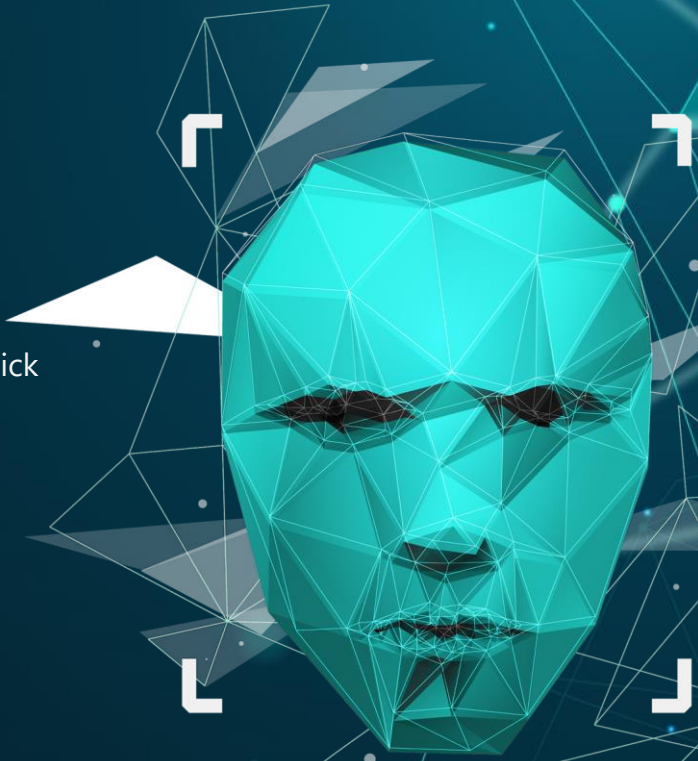
Mask R-CNN

Kaiming He, Georgia Gkioxari, Piotr Dollár, Ross Girshick

Facebook AI Research (FAIR) - 2018



Presenter:
Yoel Bokobza

Motivation

- Mask R-CNN is a conceptually simple, flexible, and general framework for object instance segmentation.
- Mask R-CNN efficiently detects objects in an image while simultaneously generating a high-quality segmentation mask for each instance.
- Mask R-CNN extends Faster R-CNN.
- Faster R-CNN is used for object detection in real-time.

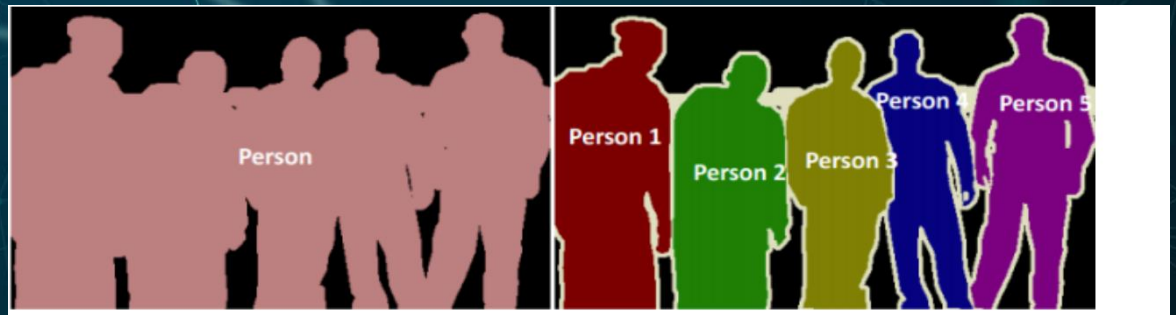
Segmentation

Semantic Segmentation

- The process of assigning a label to every pixel in the image.
- Semantic segmentation treats multiple objects of the same class as a single entity.

Instance Segmentation

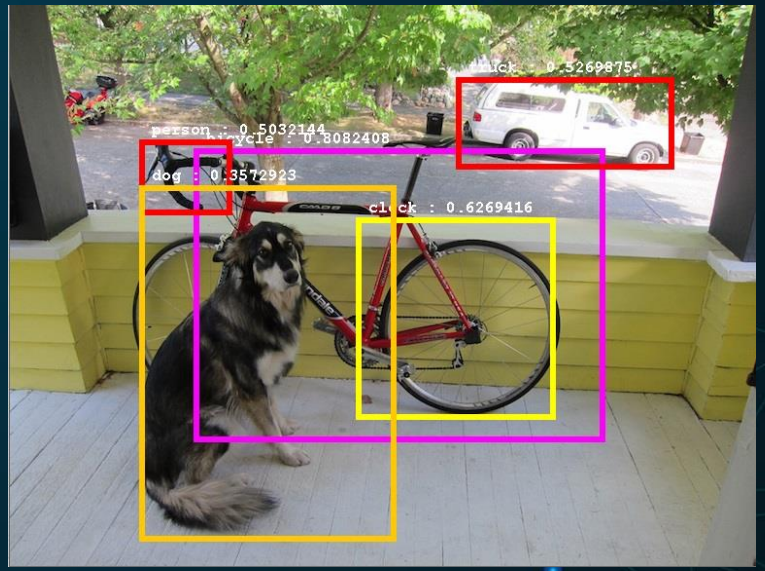
- instance segmentation treats multiple objects of the same class as distinct individual objects (or instances).



Semantic segmentation (left) and Instance segmentation (right)

Object Detection

1. Object detection is an advanced form of image classification where a neural network predicts objects in an image and points them out in the form of bounding boxes.

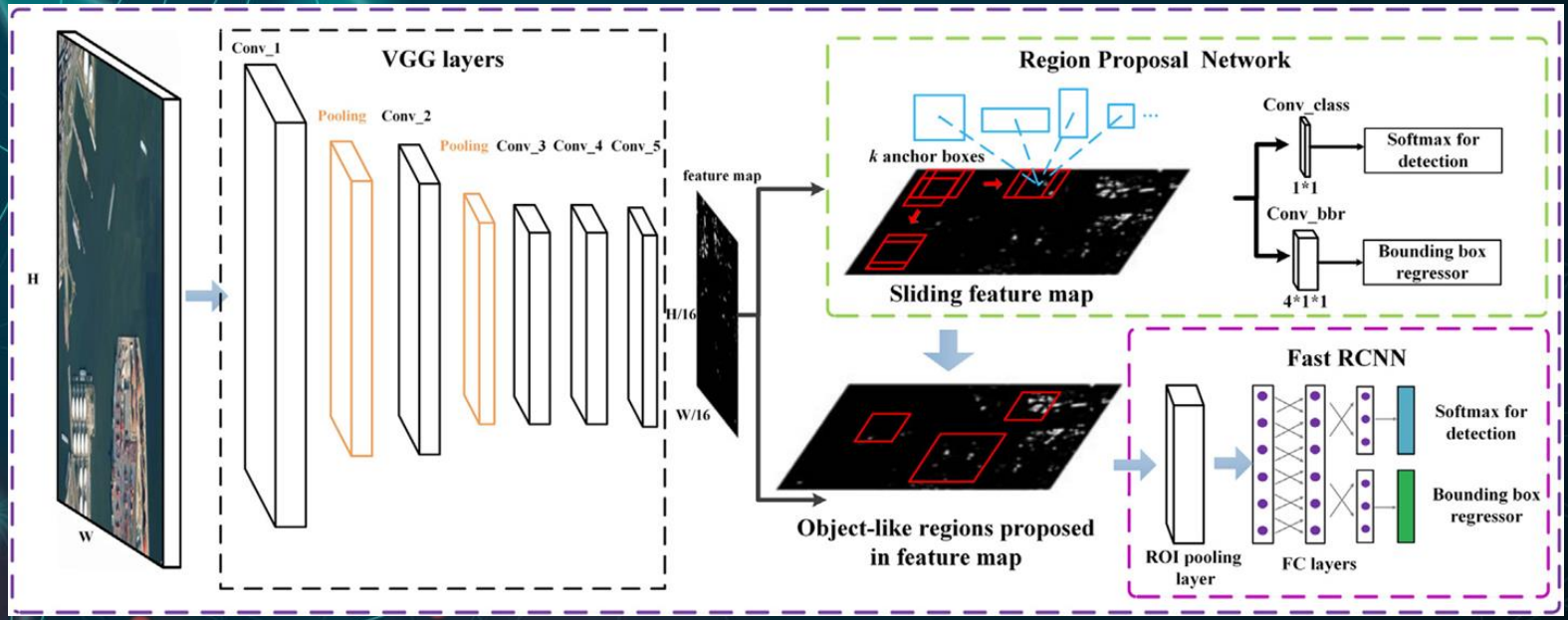


Faster R-CNN

01

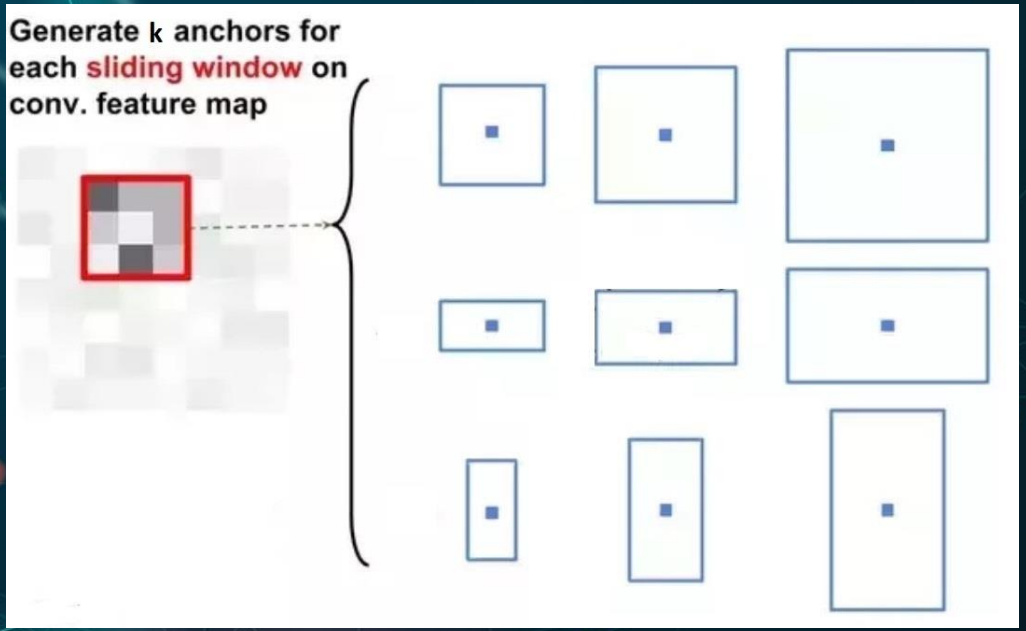
High level Description

- Faster R-CNN is an end-to-end deep convolutional network for object detection.
- Faster R-CNN relays on a deep learning-based approach for region proposals.



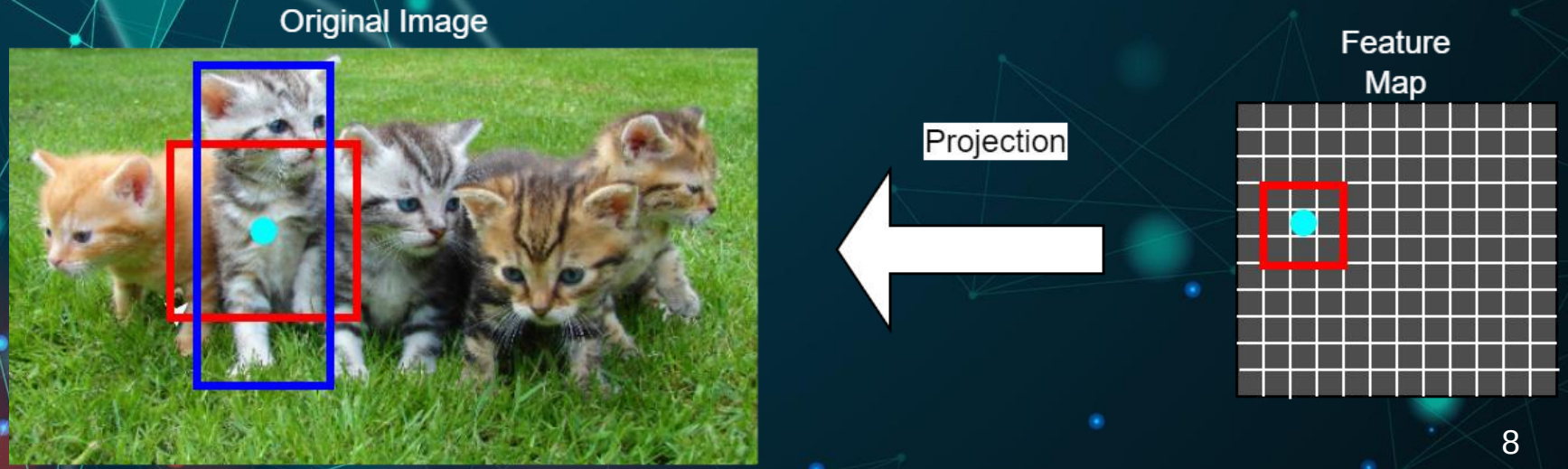
Region-Proposal Network (RPN)

- RPN Input – An image feature map.
- RPN Output - regions of interest (RoIs) in the image.
- RPN slides a small network over the feature map.
- This network takes as input an n × n spatial window of the input feature map.



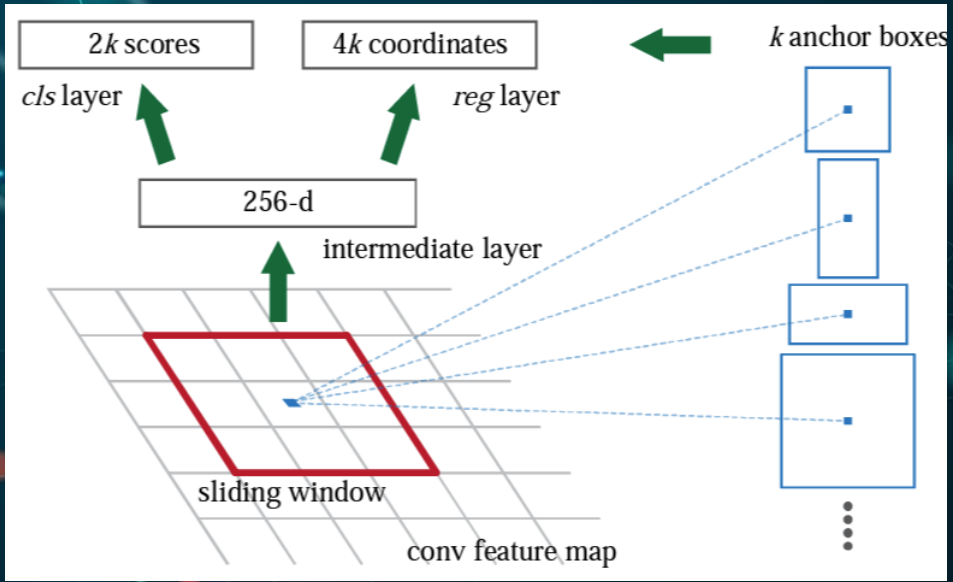
Anchors and Receptive Field

- For any sliding window, multiple region proposals are predicted simultaneously
- These regions referred to as **anchors**
- The spatial window over the feature map is corresponded to an area in the original image according to the receptive field principle.
- For $n = 3$, with VGG-16, the receptive field is 228×228



RPN cont.

- The number of proposals for each location is denoted as k.
- The **reg layer** has 4k outputs encoding the coordinates offset of k boxes.
- The **cls layer** outputs 2k scores that estimate the probability of foreground/background.

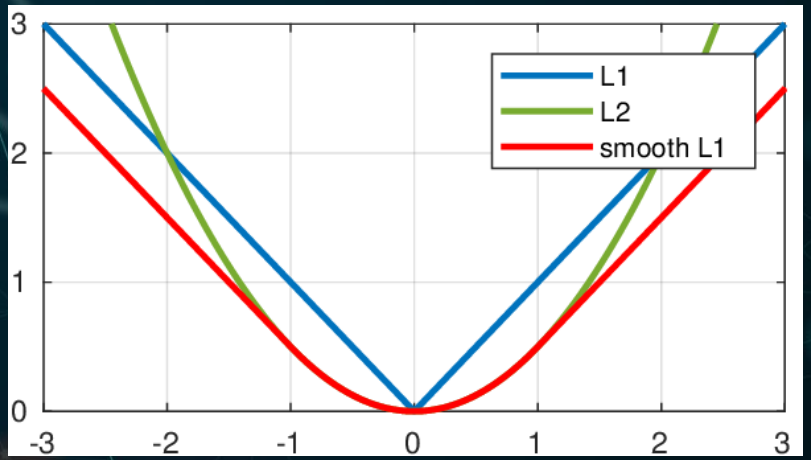


RPN Loss Function

- p - The probability of a RoI to contain an object.
- p^* - The ground truth label.
- $t = (t_x, t_y, t_w, t_h)$ - The predicted bounding-box regression offsets.
- $t^* = (t_x^*, t_y^*, t_w^*, t_h^*)$ – The true bounding-box regression offsets.
- The RPN loss is defined as a weighted summation of to losses $L_{cls}(p, p^*)$, and $L_{box}(t, t^*)$

RPN Loss Function, cont.

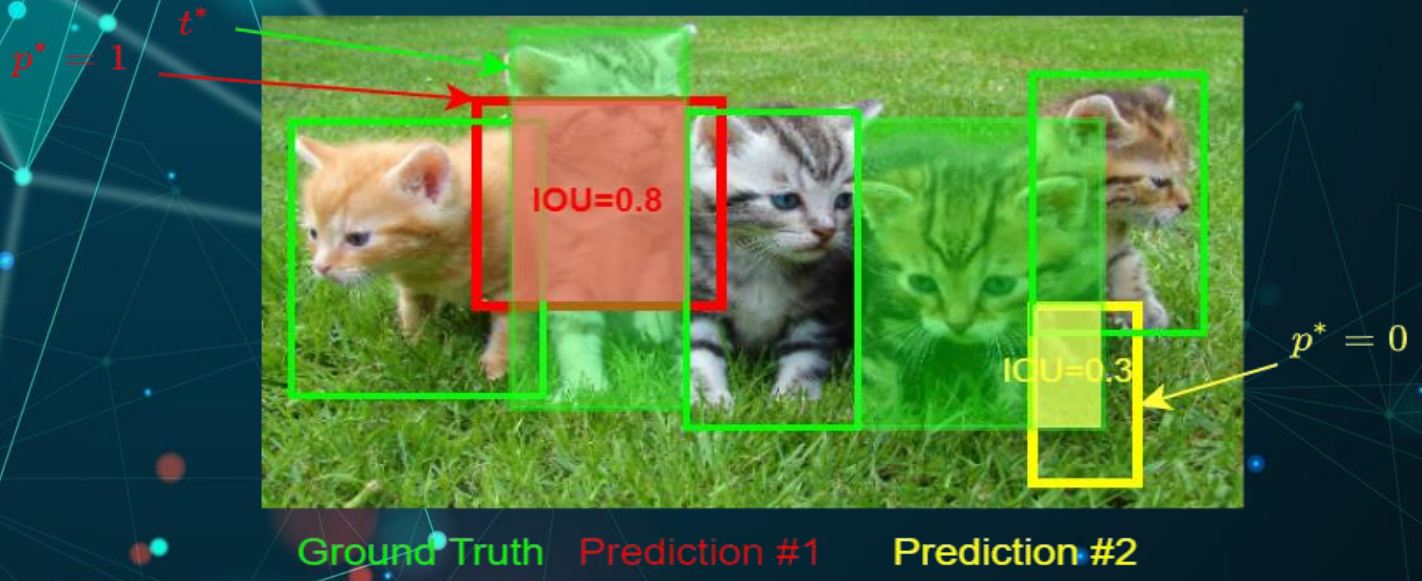
- $L_{cls}(p, p^*)$: $CrossEntropyLoss(p, p^*) = -(p^* \log(p) + (1 - p^*) \log(1 - p))$
- $L_{box}(t, t^*)$: $smooth_{L_1}(t, t^*) = \sum_{i \in \{x,y,w,h\}} \begin{cases} 0.5(t_i - t_i^*)^2, & if\ |t_i - t_i^*| < 1 \\ |t_i - t_i^*| - 0.5, & otherwise \end{cases}$



- The total Loss is $L_{RPN}(p, p^*, t, t^*) = L_{cls}(p, p^*) + \lambda \cdot p^* \cdot L_{box}(t, t^*)$

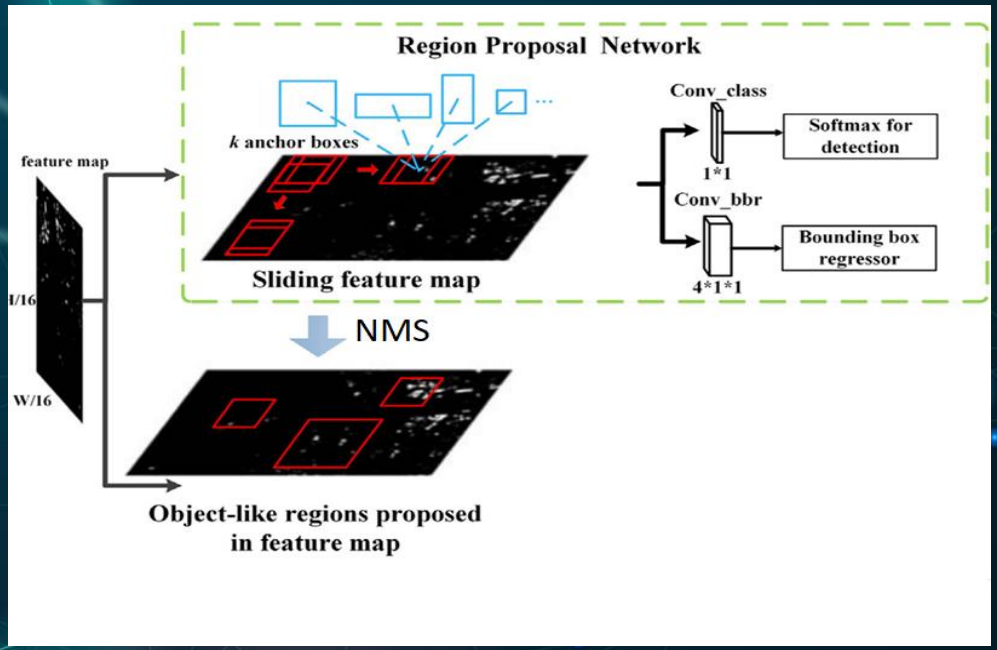
Evaluating Ground Truths

- p^*, and t^* need to be evaluated before computing the loss
- The ground truths are estimated by calculating the IOU for each anchor with each ground truth box
- $p^* = 1$ if the IOU exceeds a threshold T with some ground truth box and 0 otherwise
- t^* is determined according to the ground truth box with the maximal IOU such that IOU$> T$



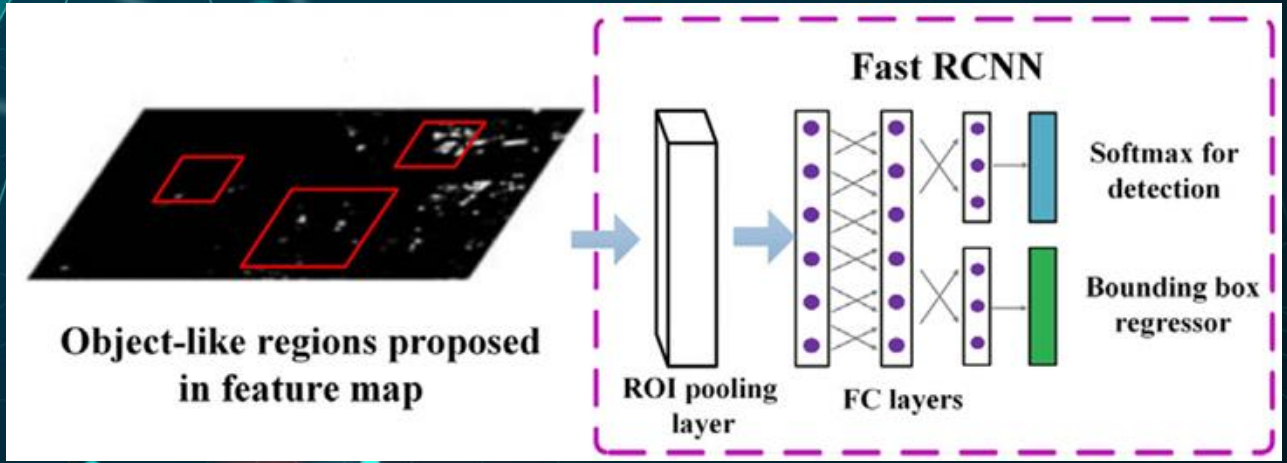
Region of Interest (ROI) Proposals

- Some RPN proposals highly overlap with each other.
- Non-maximum suppression (NMS) algorithm has been applied
- After NMS ,only the top-N ranked proposal regions are used for detection



Fast R-CNN

- **Fast R-CNN input –** The cropped ROIs from the feature map
- **Fast R-CNN output -** The bounding box indices and the class of the object.
- Fast R-CNN is composed of an ROI Pooling mechanism followed by a fully-connected (FC) network.
- ROI Pooling allows the input to the FC to be a fixed size input, independent of the ROIs sizes.



Mask R-CNN

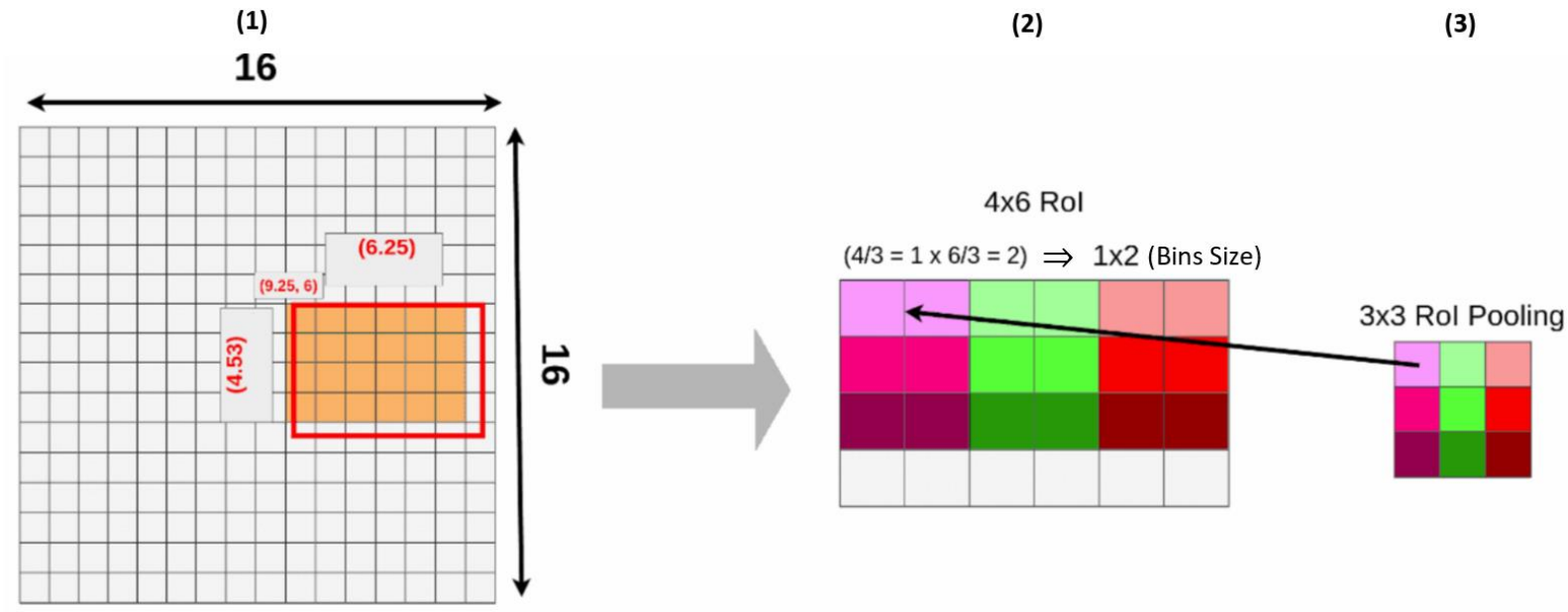
02

Back to Mask R-CNN

- Mask R-CNN extends Faster R-CNN
- Adding a branch for predicting an object mask in parallel with the existing branches for bounding box recognition and classification.
- RoIPool layer has been replaced by RoIAlign.
- This replacement is essential for the segmentation task.

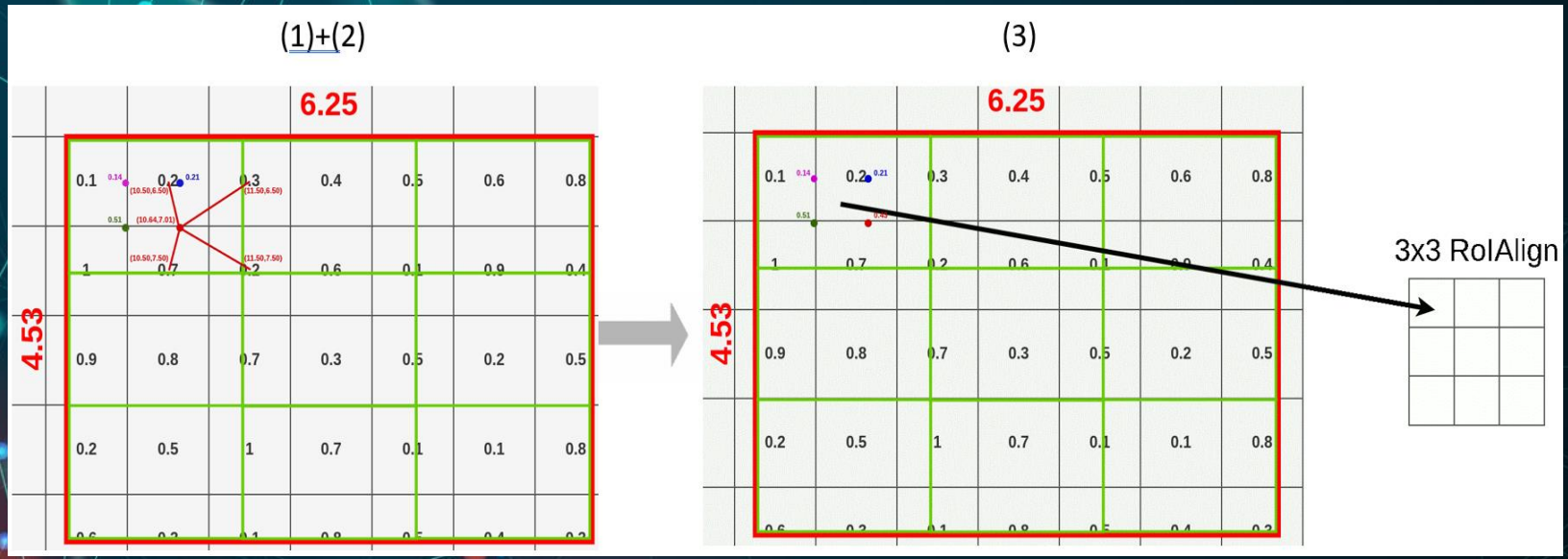
RoIPool

(1) Quantizes a floating-number RoI to the discrete granularity of the feature map.
(2) The quantized RoI is then subdivided into spatial bins which are themselves quantized.
(3) Finally feature values covered by each bin are aggregated (usually by max pooling).



RoIAlign

- This method avoids any quantization of the RoI boundaries
(1) The RoI is divided into bins according to the required output size
(2) A bilinear interpolation is applied at four regularly sampled locations in each RoI bin.
(3) Feature values covered by each bin are aggregated (usually by max pooling).

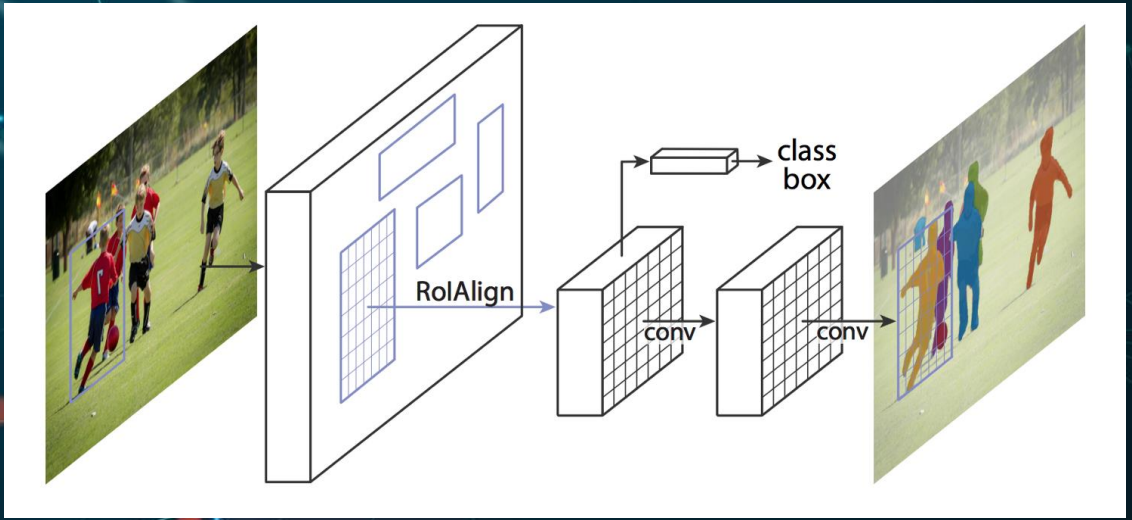


RoIPool VS RoIAlign

- The quantization in RoIPool introduces misalignments between the RoI and the extracted features.
- The quantization leads to a large negative effect on predicting pixel-accurate masks
- RoIAlign layer removes the harsh quantization of RoIPool.
- RoIAlign uses the whole area to pool data from.

Mask Predictions

- The Mask R-CNN workflow:
    1. An image is fed into the RPN which returns a set of RoIs
    2. A RoIAlign is applied for each RoI
    3. The RoIAlign output is passed to a network that parallelly predicts a class, a bounding box offset, and a binary mask.

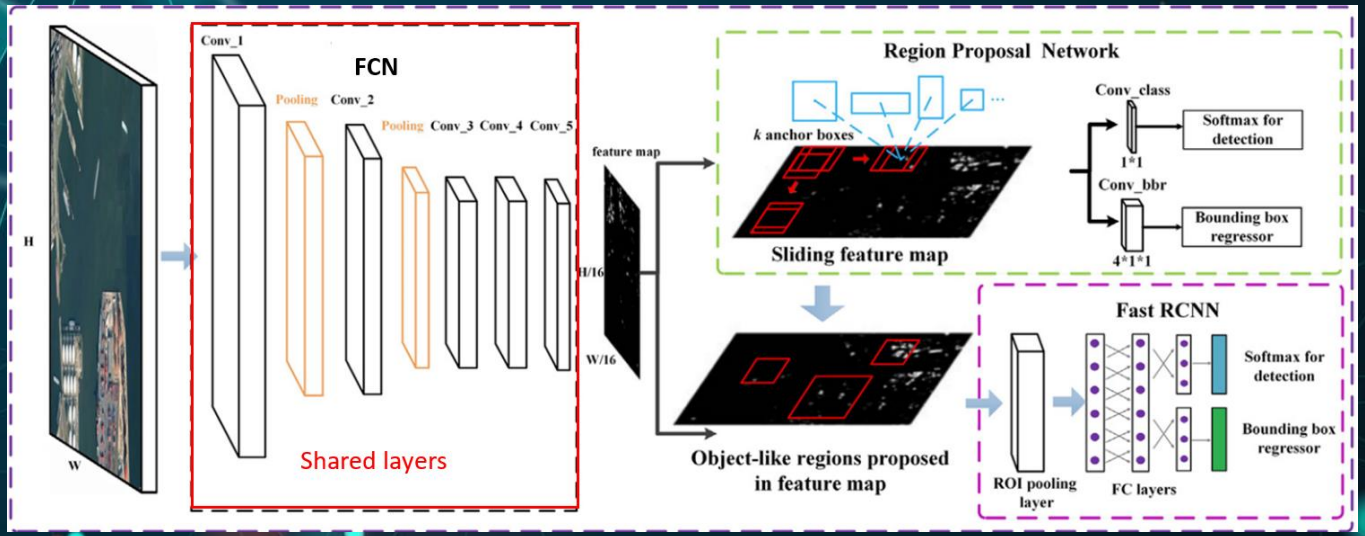


The Loss Function

- The mask branch predicts K binary masks of resolution $m \times m$, one for each of the K classes.
- The loss: $L = L_{cls} + \lambda L_{box} + \gamma L_{mask}$.
- L_{cls} and L_{box} are similar to the losses defined for the RPN loss.
- L_{mask} - The average pixel-wise binary cross-entropy loss.
- L_{mask} is calculated only for the mask corresponding to the ground truth class.
- The L_{mask} decoupled structure allows the network to generate masks without competition among classes.

Training

- The RPN and the Mask R-CNN share FCN features
- These shared layers facilitate fast inference time and reduces the number of weights.
- In the figure below the shared layers are marked by red box.

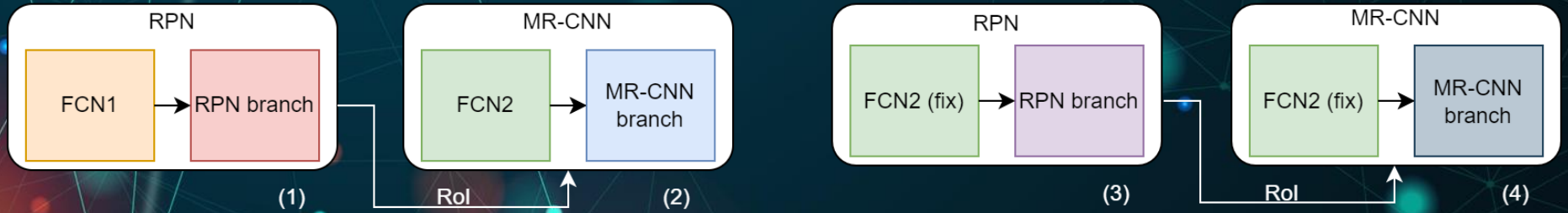


Training – cont.

- The shared layer is achieved by applying a training method referred to as *4-Step Alternating Training:*

1. Initialized the FCN with an ImageNet-pre-trained model and fine-tuned end-to-end for the RPN.
2. Train Mask R-CNN with a separate FCN using the proposals generated by the step-1 RPN.
3. Use the Mask R-CNN FCN to initialize RPN training and fix the FCN weights.
4. Fine-tune the Mask R-CNN, while keeping its FCN weights fixed , using the proposals generated by the step-3 RPN.

- At the end of this process, both RPN and Mask R-CNN share the same FCN



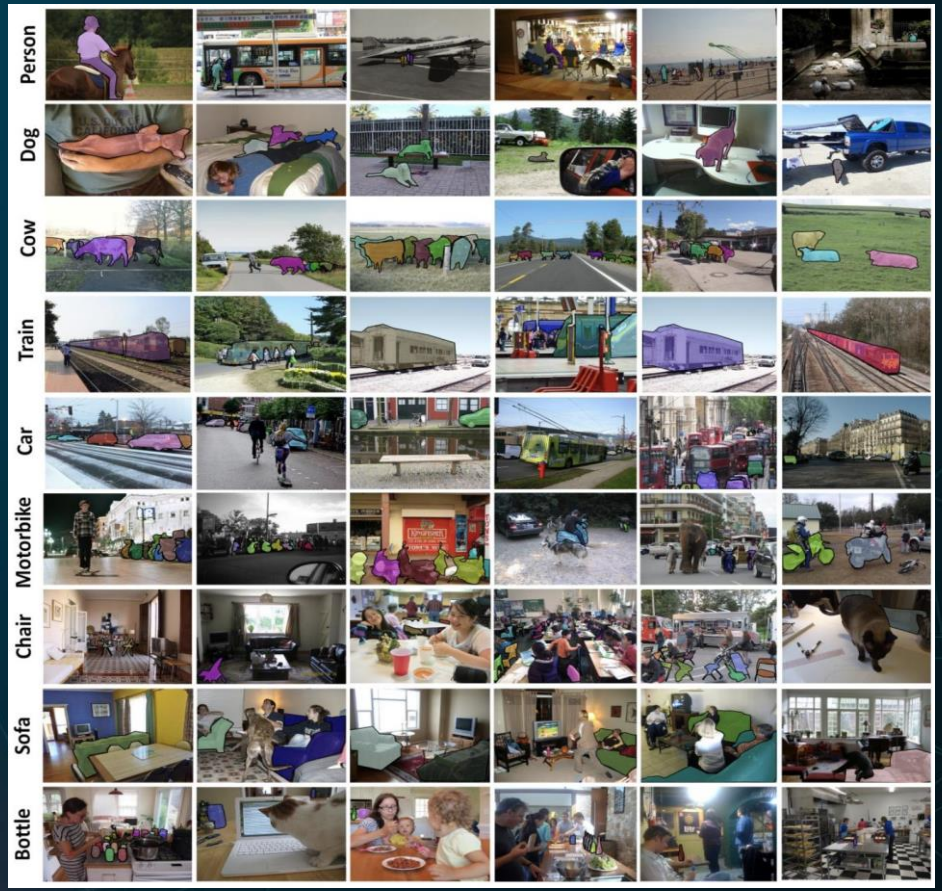
Results

04

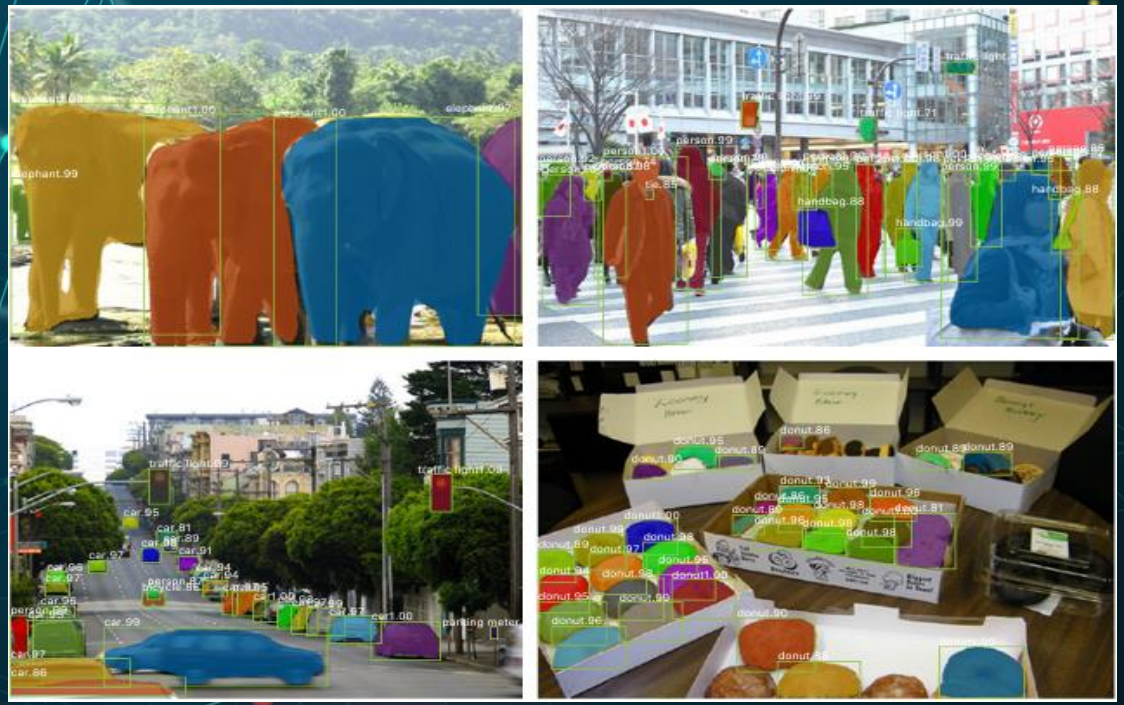
The COCO dataset

- COCO is a large-scale object detection, segmentation, and captioning dataset.

- More than 200K labeled images
- 1.5 million object instances
- 80 object categories
- 5 captions per image
- 250,000 people with keypoints



Mask R-CNN on COCO test images, using ResNet-101-FPN



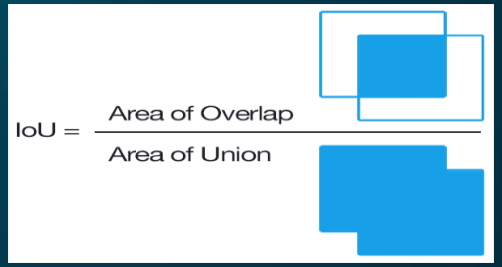
Pose Estimation

- A Human Pose Skeleton represents the orientation of a person in a graphical format.
- Consists of set of coordinates that can be connected to describe the pose of the person

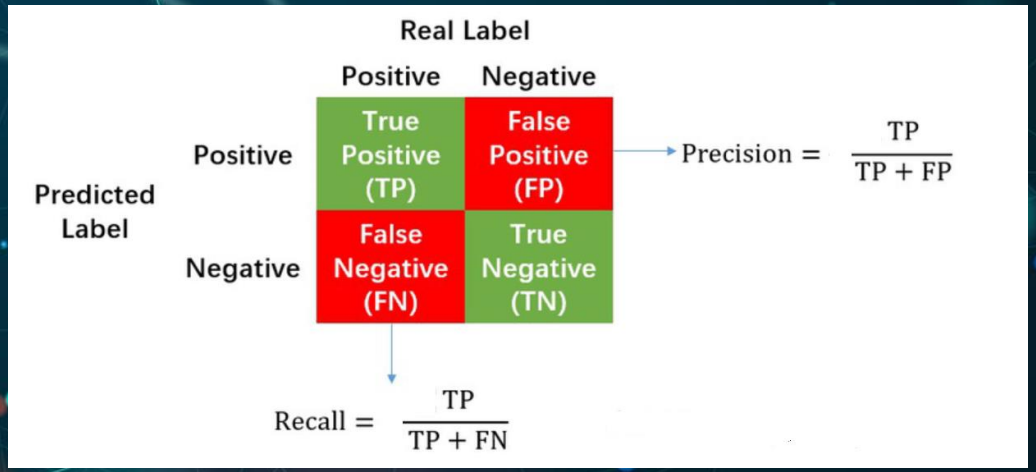
Mask R-CNN on COCO test images, using ResNet-50-FPN

Average Precision (AP)

IOU

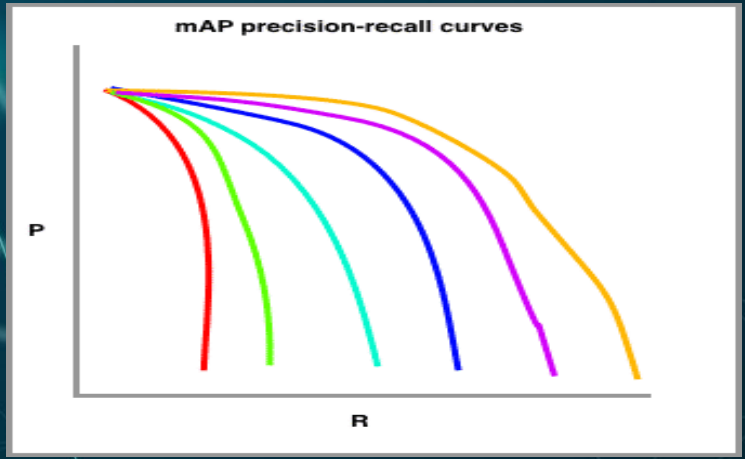


Precision and Recall



Average Precision – cont.

- precision recall curves with the IoU threshold set at varying levels



- AP is calculated by averaging the area under the curves.
- AP_x is calculated by averaging the area under the curve corresponding to $\mathrm{IOU} = x\%$

Results Comparison

Instance segmentation mask AP on COCO test set

	backbone	AP	AP_{50}	AP_{75}
MNC [10]	ResNet-101-C4	24.6	44.3	24.8
FCIS [26] +OHEM	ResNet-101-C5-dilated	29.2	49.5	-
FCIS+++ [26] +OHEM	ResNet-101-C5-dilated	33.6	54.5	-
Mask R-CNN	ResNet-101-C4	33.1	54.9	34.8
Mask R-CNN	ResNet-101-FPN	35.7	58.0	37.8
Mask R-CNN	ResNeXt-101-FPN	**37.1**	**60.0**	**39.4**

Object detection (bounding box) AP on COCO test set

	backbone	AP	AP_{50}	AP_{75}
Faster R-CNN+++ [19]	ResNet-101-C4	34.9	55.7	37.4
Faster R-CNN w FPN [27]	ResNet-101-FPN	36.2	59.1	39.0
Faster R-CNN by G-RMI [21]	Inception-ResNet-v2 [41]	34.7	55.5	36.7
Faster R-CNN w TDM [39]	Inception-ResNet-v2-TDM	36.8	57.7	39.2
Faster R-CNN, RoIAlign	ResNet-101-FPN	37.3	59.6	40.3
Mask R-CNN	ResNet-101-FPN	38.2	60.3	41.7
Mask R-CNN	ResNeXt-101-FPN	**39.8**	**62.3**	**43.4**

Summary

- Mask RCNN simultaneously predicts classes, bounding boxes, and instance segmentation masks.
- Heavily based on Faster R-CNN and extends it.
- Replaces the RoIPool by RoIAlign to reduce segmentation misalignment
- Shows substantial improvements over the SOTA algorithms
- Can also solve pose estimation problems

THANKS!

Any questions?