mAP for Object Detection

Karel Horak
Brno University of Technology / Czech Technical University in Prague
horak@feec.vutbr.cz

Slides modified, courtesy of Jonathan Hui

mAP for Object Detection

- mAP stands for mean Average Precision.
- AP is a popular metric in measuring the accuracy of object detectors like Faster R-CNN, SSD, etc.
- Average precision computes the average precision value for recall value over 0 to 1. It sounds complicated but actually pretty simple as we illustrate it with an example.
- But before that, we will do a quick recap on precision, recall, and IoU first.

Precision & recall

- **Precision** measures how accurate is your predictions. i.e. the percentage of your 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.

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

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

$$TP = \text{True positive}$$

$$TN = \text{True negative}$$

$$FP = \text{False positive}$$

$$FN = \text{False negative}$$

- **F1 score** is a measure combining Precision and Recall.

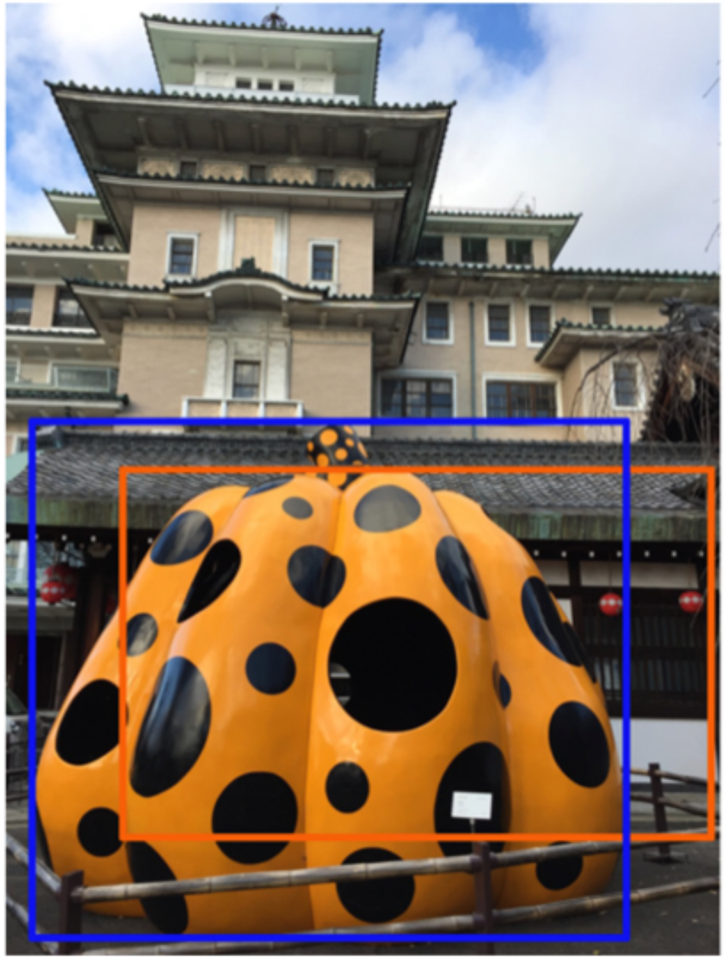
$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

IoU (Intersection over union)

- **IoU** measures the overlap between two boundaries.

- We use that to measure how much our predicted boundary overlaps with the ground truth (the real object boundary).

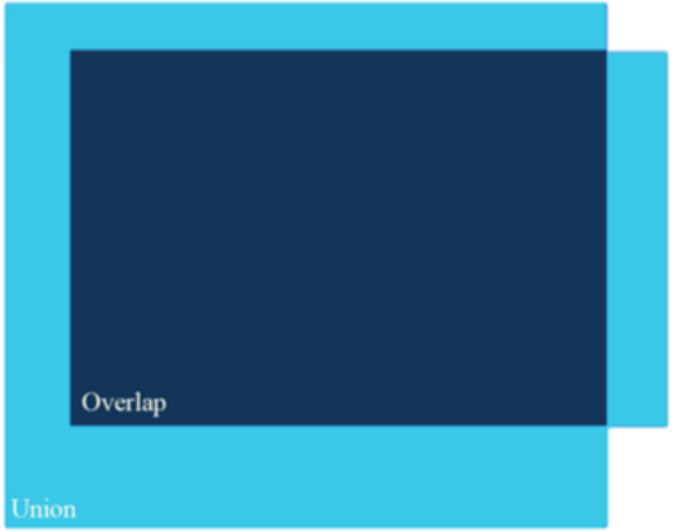
- In some datasets, we predefine an IoU threshold (say 0.5) in classifying whether the prediction is a true positive or a false positive.



□ Ground truth

□ Prediction

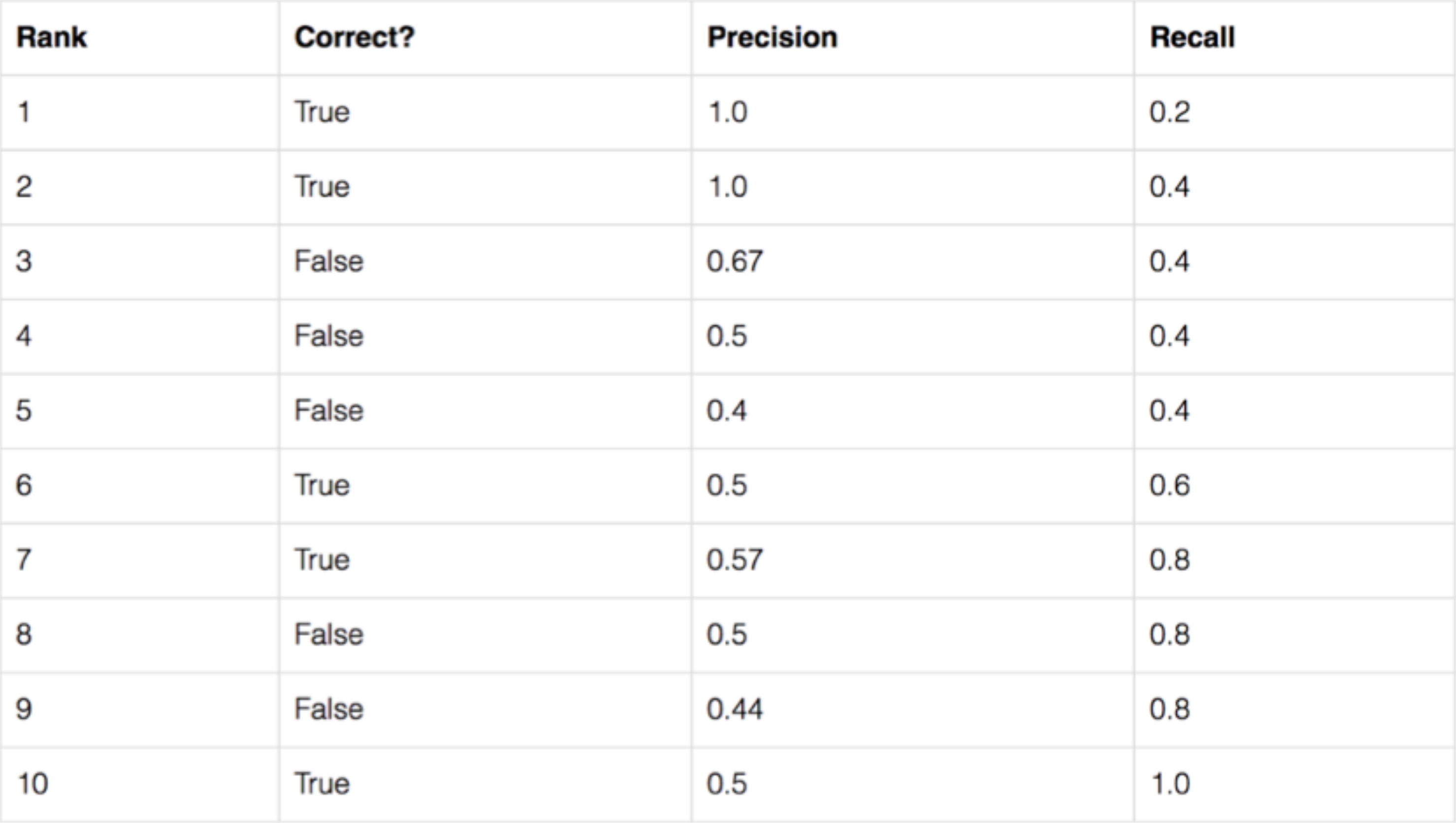
$$IoU = \frac{\text{area of overlap}}{\text{area of union}}$$



Average Precision

- Let us create an over-simplified example in demonstrating the calculation of the average precision.

- In this example, the whole dataset contains five apples only. We collect all the predictions made for apples in all the images and rank it in descending order according to the predicted confidence level. The second column indicates whether the prediction is correct or not. In this example, the prediction is correct if IoU ≥ 0.5.



Rank	Correct?	Precision	Recall
1	True	1.0	0.2
2	True	1.0	0.4
3	False	0.67	0.4
4	False	0.5	0.4
5	False	0.4	0.4
6	True	0.5	0.6
7	True	0.57	0.8
8	False	0.5	0.8
9	False	0.44	0.8
10	True	0.5	1.0

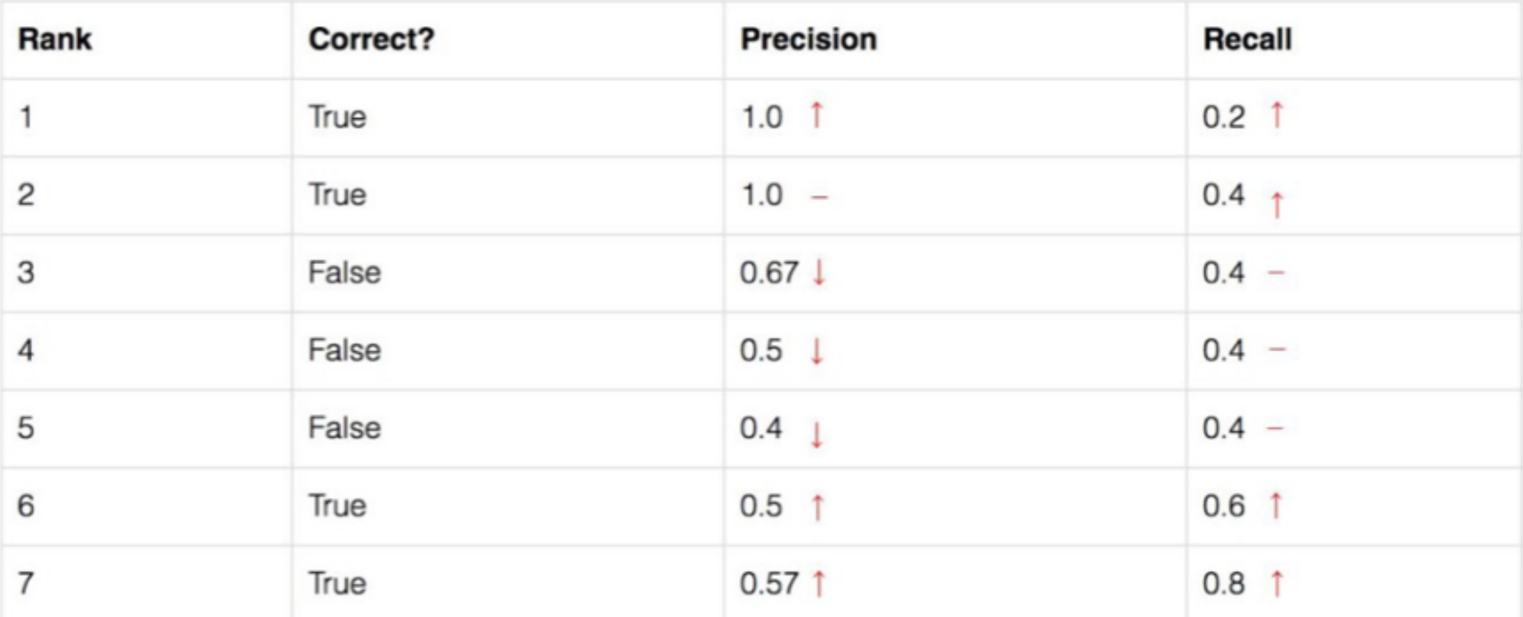
Average Precision

- Let us take the row with rank #3 and demonstrate how precision and recall are calculated first.

- **Precision** is the proportion of TP = 2/3 = 0.67.

- **Recall** is the proportion of TP out of the possible positives = 2/5 = 0.4.

- Recall values increase as we go down the prediction ranking. However, precision has a zigzag pattern — it goes down with false positives and goes up again with true positives.



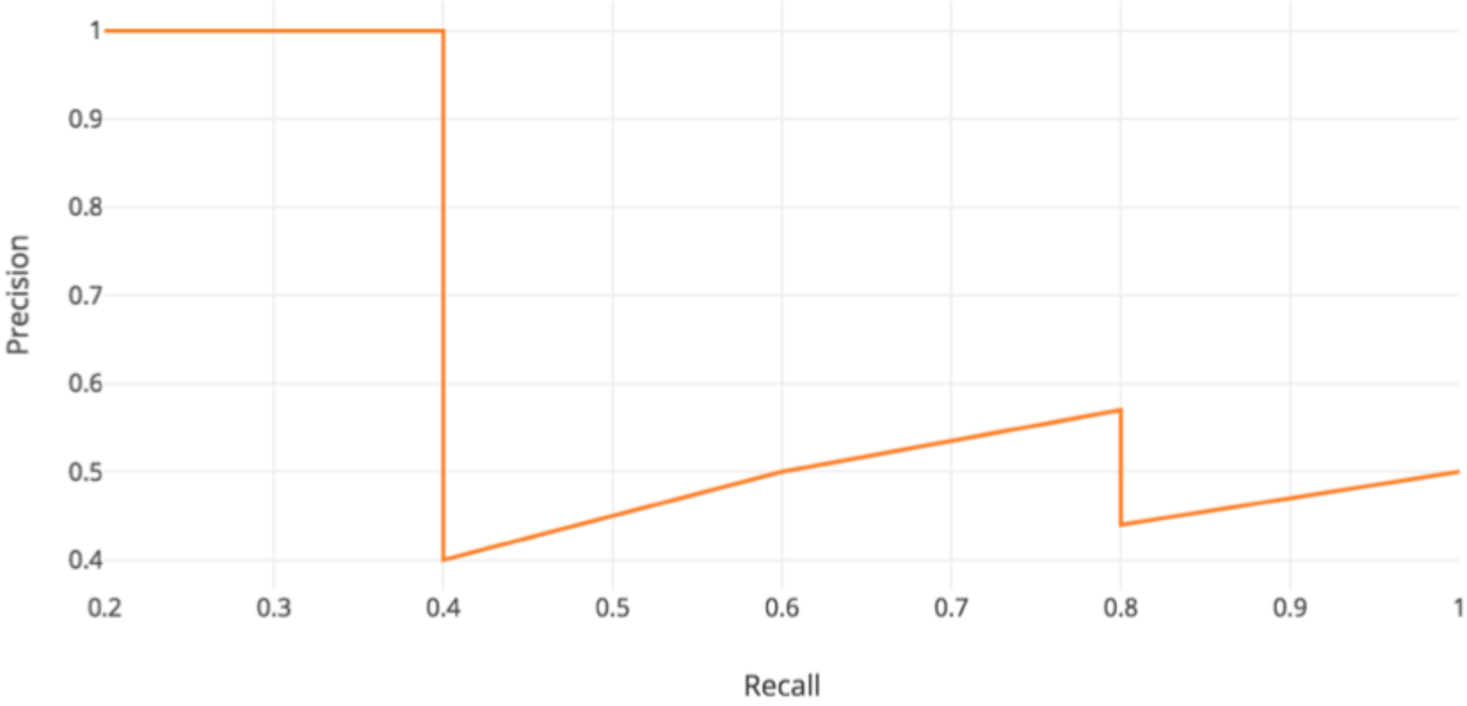
Rank	Correct?	Precision	Recall
1	True	1.0 ↑	0.2 ↑
2	True	1.0 –	0.4 ↑
3	False	0.67 ↓	0.4 –
4	False	0.5 ↓	0.4 –
5	False	0.4 ↓	0.4 –
6	True	0.5 ↑	0.6 ↑
7	True	0.57 ↑	0.8 ↑

Average Precision

- Let us plot the precision against the recall value to see this zig-zag pattern.

- The general definition for the Average Precision (AP) is finding the area under the precision-recall curve.

$$\mathrm{AP} = \int_0^1 p(r)dr$$

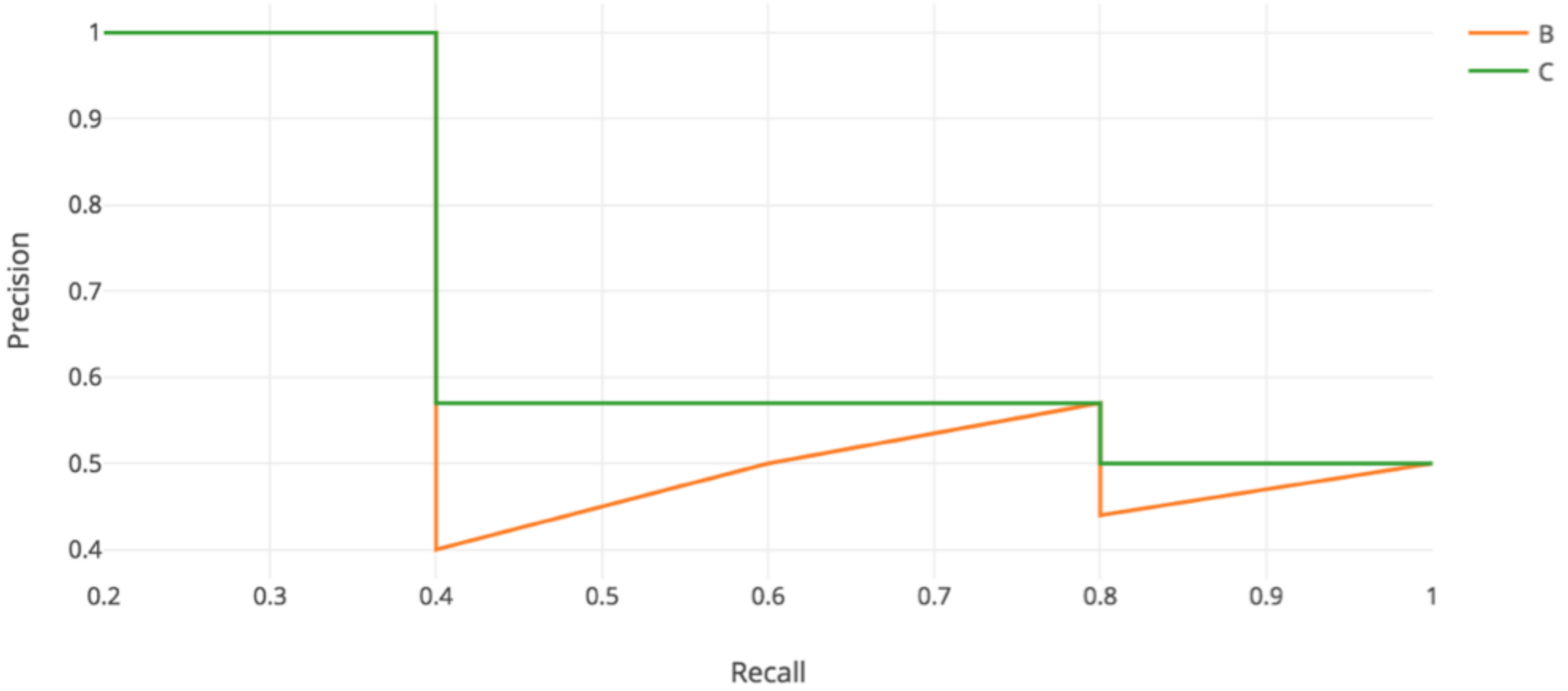


Average Precision

- Precision and recall are always between 0 and 1. Therefore, AP falls within 0 and 1 also.

- Before calculating AP for the object detection, we often smooth out the zigzag pattern first.

- Graphically, at each recall level, we replace each precision value with the maximum precision value to the right of that recall level.



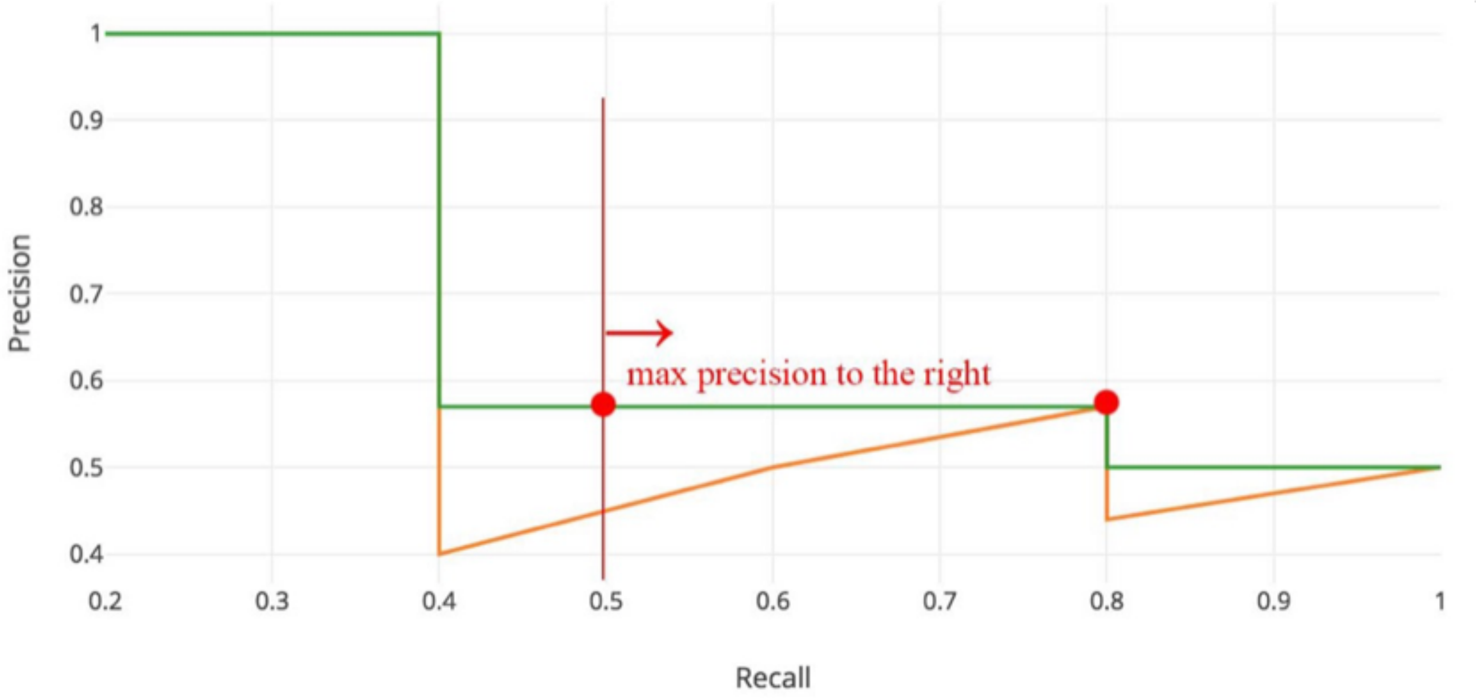
Average Precision

- So the orange line is transformed into the green line and the curve will decrease monotonically instead of the zigzag pattern.

- The calculated AP value will be less suspectable to small variations in the ranking.

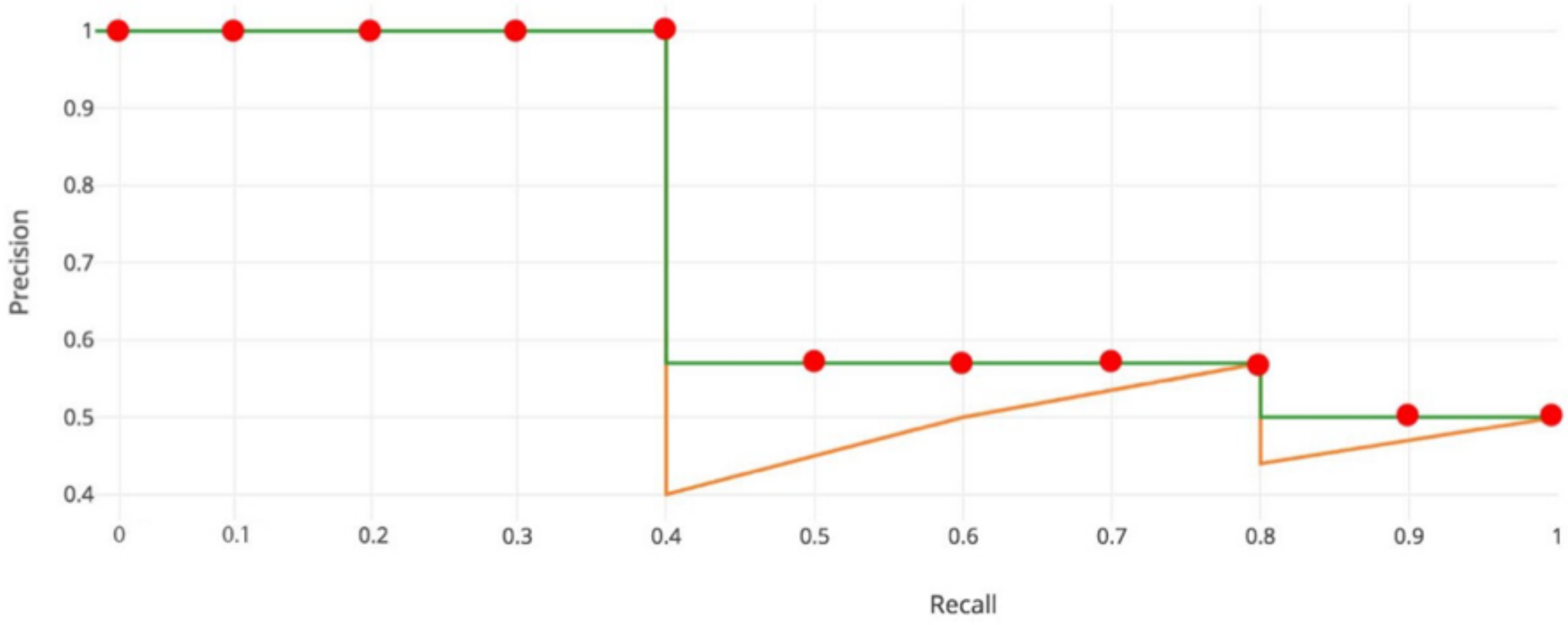
- Mathematically, we replace the precision value for recall r̂ with the maximum precision for any recall ≥ r̂.

$$p_{interp}(r) = \max_{\widetilde{r} \geq r} p(\widetilde{r})$$



Interpolated AP

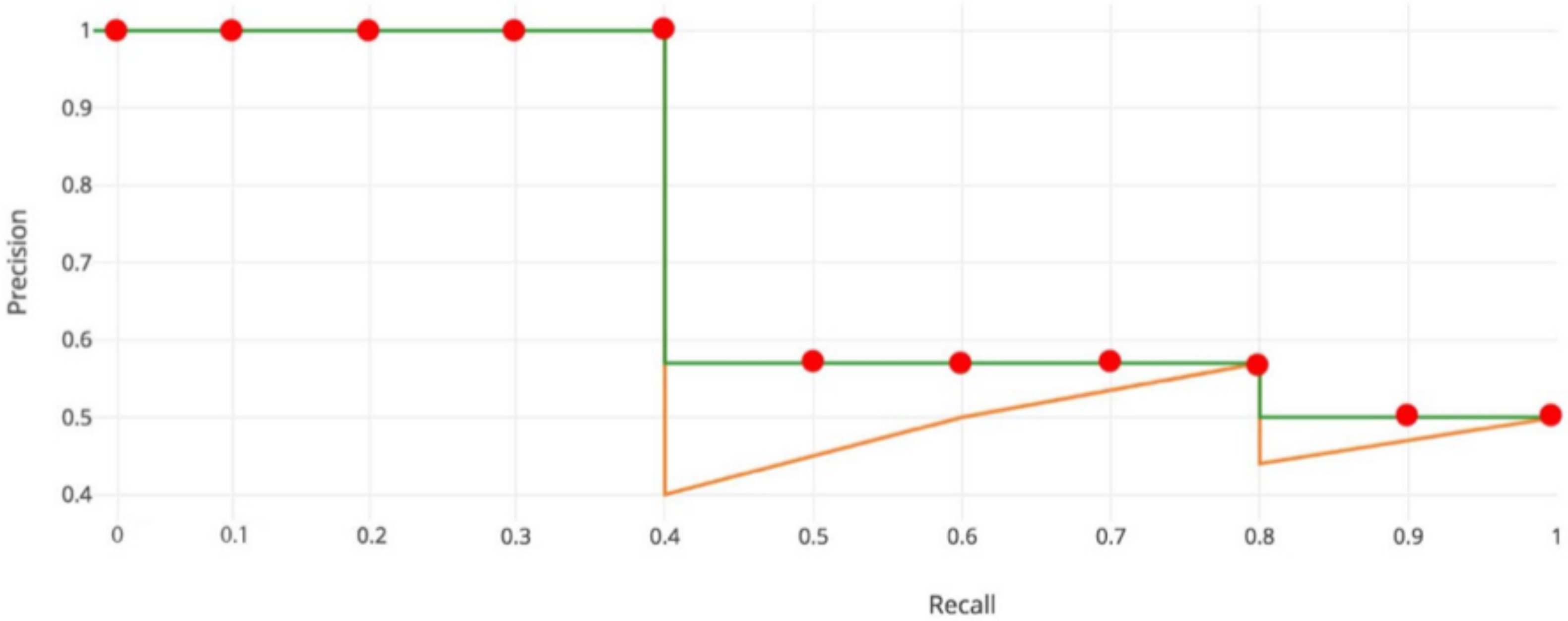
- **PASCAL VOC** is a popular dataset for object detection.
- For the PASCAL VOC challenge, a prediction is positive if IoU ≥ 0.5. Also, if multiple detections of the same object are detected, it counts the first one as a positive while the rest as negatives.
- In Pascal VOC2008, an average for the 11-point interpolated AP is calculated.



Interpolated AP

- First, we divide the recall value from 0 to 1.0 into 11 points — 0, 0.1, 0.2, …, 0.9 and 1.0.
- Next, we compute the average of maximum precision value for these 11 recall values.

$$AP = \frac{1}{11} \times \left(AP_r(0) + AP_r(0.1) + \ldots + AP_r(1.0)\right)$$



- In our example, AP = (5 × 1.0 + 4 × 0.57 + 2 × 0.5)/11

Interpolated AP

- When AP_r turns extremely small, we can assume the remaining terms to be zero. i.e. we do not necessarily make predictions until the recall reaches 100%.

- If the possible maximum precision levels drop to a negligible level, we can stop.

- For 20 different classes in PASCAL VOC, we compute an AP for every class and also provide an average for those 20 AP results.

- However, this interpolated method is an approximation which suffers two issues:

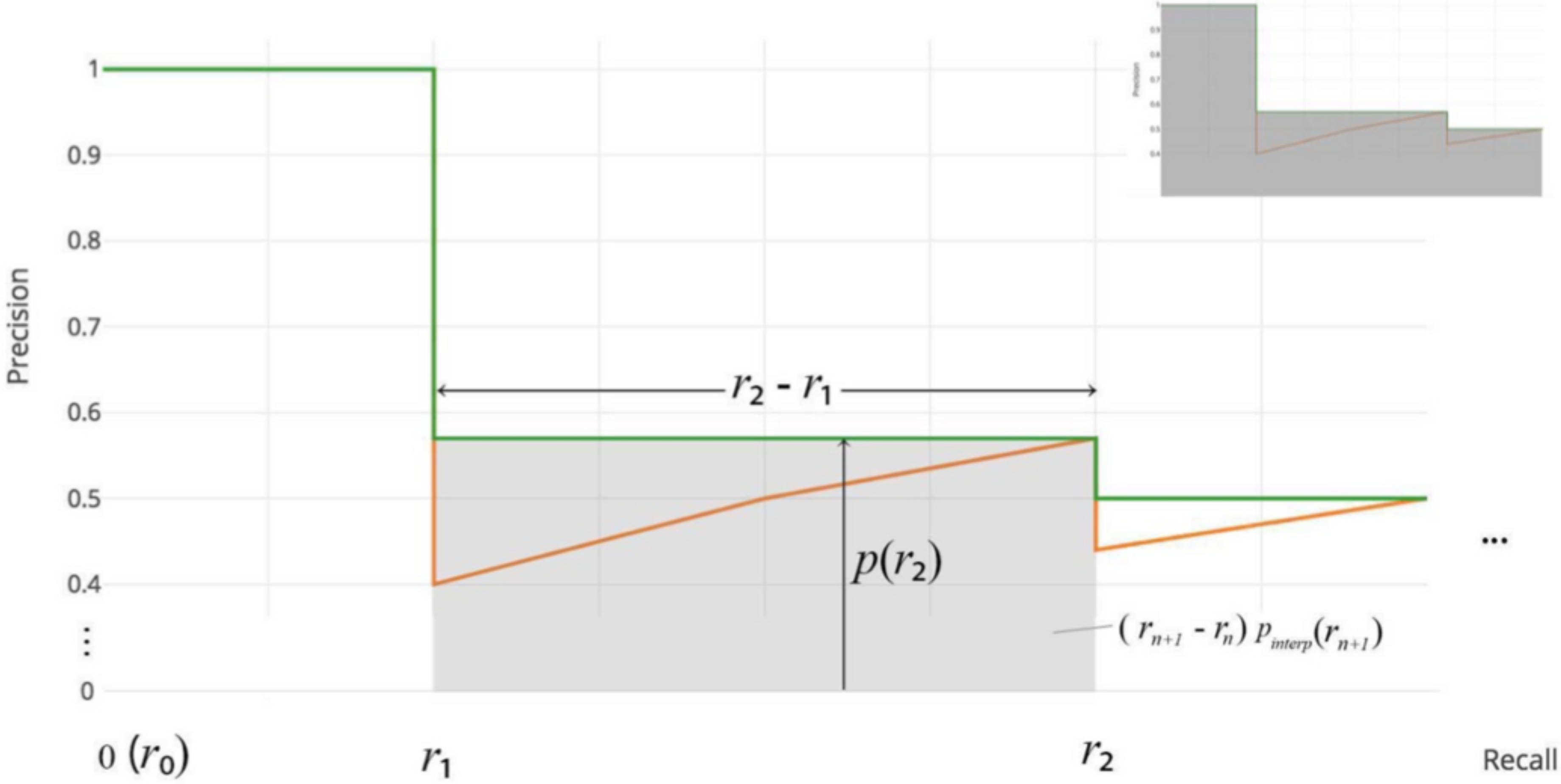
1. It is less precise.
2. Second, it lost the capability in measuring the difference for methods with low AP.

- Therefore, a different AP calculation is adopted after 2008 for PASCAL VOC.

AP (Area under curve AUC)

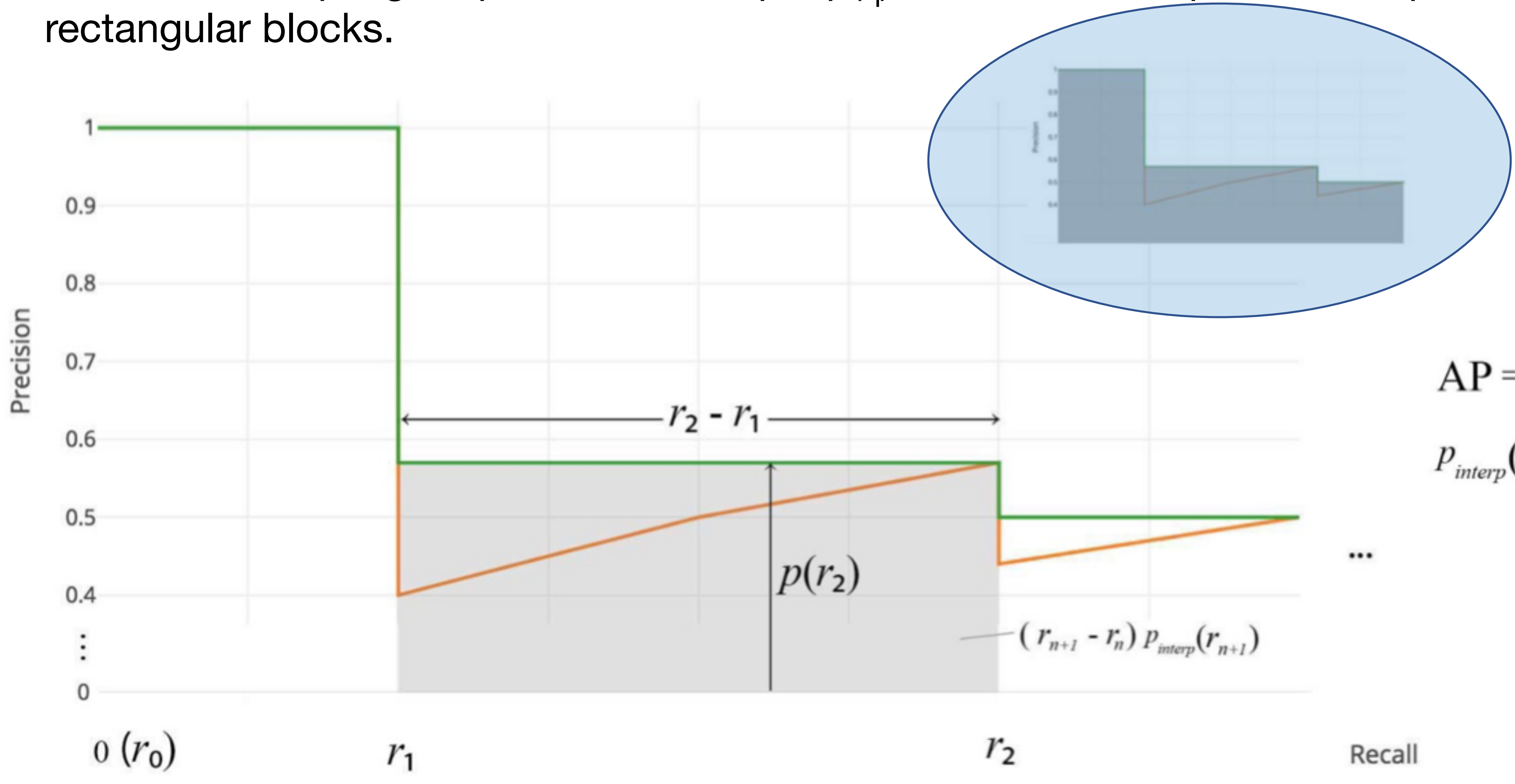
- For later Pascal VOC competitions, VOC2010–2012 samples the curve at all unique recall values (r_1, r_2, …), whenever the maximum precision value drops.

- With this change, we are measuring the exact area under the precision-recall curve after the zigzags are removed.



AP (Area under curve AUC)

- No approximation or interpolation is needed.
- Instead of sampling 11 points, we sample $p(r_i)$ whenever it drops and computes AP as the sum of the rectangular blocks.



$$AP = \Sigma\,(r_{n+1} - r_n)\, p_{interp}(r_{n+1})$$

$$p_{interp}(r_{n+1}) = \max_{\tilde{r} \geq r_{n+1}} p(\tilde{r})$$

COCO mAP

- Latest research papers tend to give results for the COCO dataset only. In COCO mAP, a 101-point interpolated AP definition is used in the calculation.

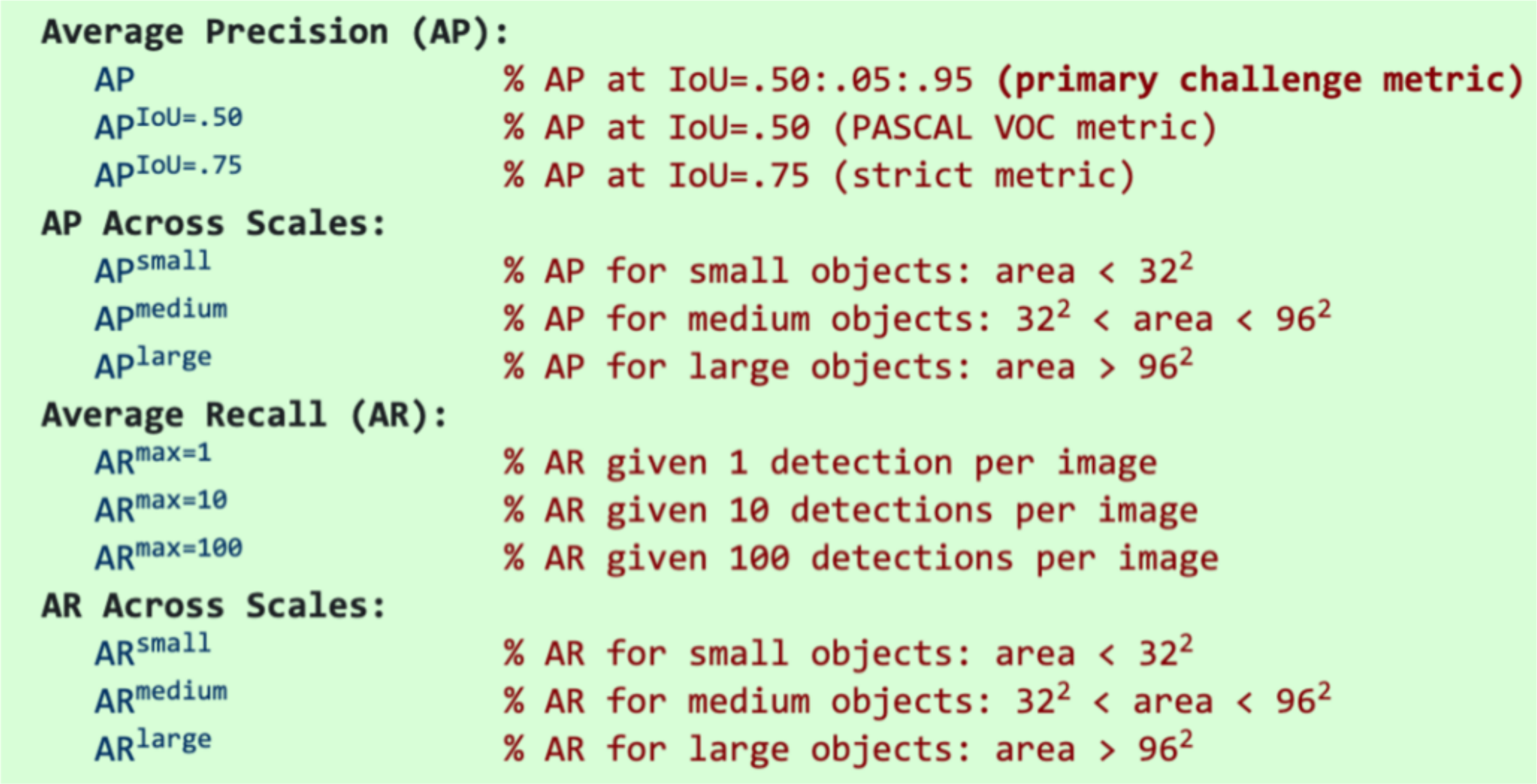
- For COCO, AP is the average over multiple IoU (the minimum IoU to consider a positive match).

- **AP@[.5:.95]** corresponds to the average AP for IoU from 0.5 to 0.95 with a step size of 0.05.

- For the COCO competition, AP is the average over 10 IoU levels on 80 categories (AP@[.50:.05:.95]: start from 0.5 to 0.95 with a step size of 0.05).

COCO mAP

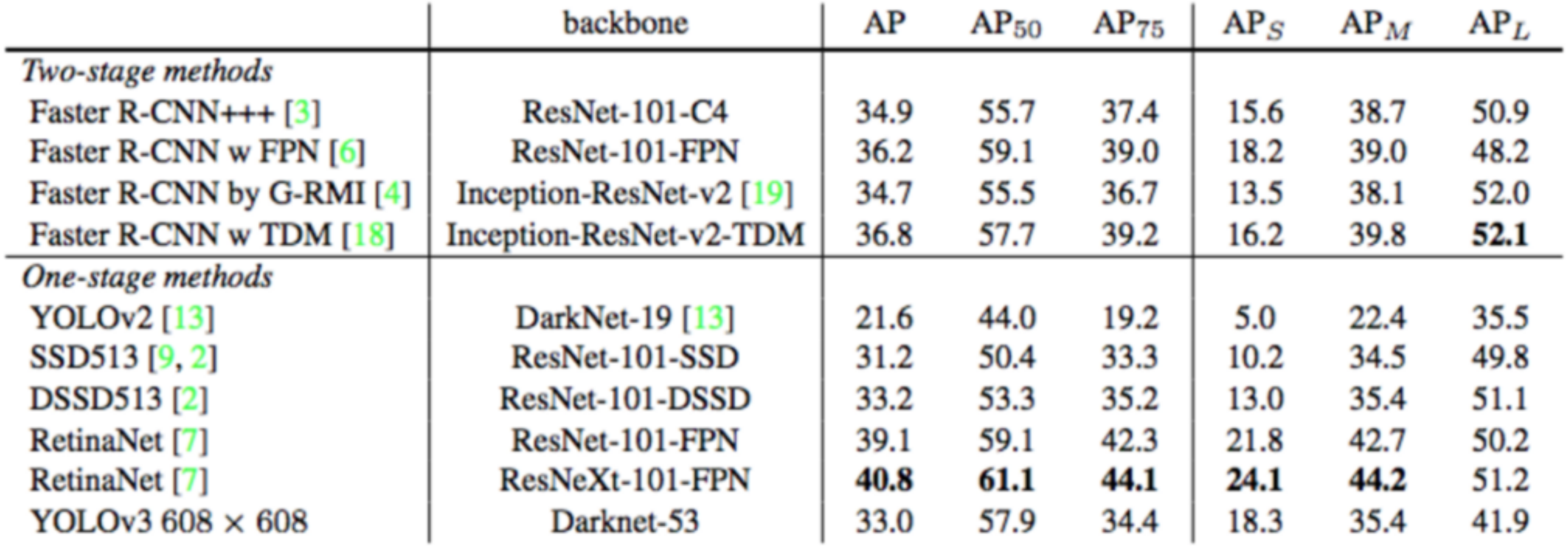
- The following are some other metrics collected for the COCO dataset:



Average Precision (AP):	
AP	% AP at IoU=.50:.05:.95 **(primary challenge metric)**
$AP^{IoU=.50}$	% AP at IoU=.50 (PASCAL VOC metric)
$AP^{IoU=.75}$	% AP at IoU=.75 (strict metric)
AP Across Scales:	
AP^{small}	% AP for small objects: area < 32^2
AP^{medium}	% AP for medium objects: 32^2 < area < 96^2
AP^{large}	% AP for large objects: area > 96^2
Average Recall (AR):	
$AR^{max=1}$	% AR given 1 detection per image
$AR^{max=10}$	% AR given 10 detections per image
$AR^{max=100}$	% AR given 100 detections per image
AR Across Scales:	
AR^{small}	% AR for small objects: area < 32^2
AR^{medium}	% AR for medium objects: 32^2 < area < 96^2
AR^{large}	% AR for large objects: area > 96^2

COCO mAP

- Example: the AP result for the YOLOv3 detector:



	backbone	AP	AP_{50}	AP_{75}	AP_S	AP_M	AP_L
Two-stage methods							
Faster R-CNN+++ [3]	ResNet-101-C4	34.9	55.7	37.4	15.6	38.7	50.9
Faster R-CNN w FPN [6]	ResNet-101-FPN	36.2	59.1	39.0	18.2	39.0	48.2
Faster R-CNN by G-RMI [4]	Inception-ResNet-v2 [19]	34.7	55.5	36.7	13.5	38.1	52.0
Faster R-CNN w TDM [18]	Inception-ResNet-v2-TDM	36.8	57.7	39.2	16.2	39.8	**52.1**
One-stage methods							
YOLOv2 [13]	DarkNet-19 [13]	21.6	44.0	19.2	5.0	22.4	35.5
SSD513 [9, 2]	ResNet-101-SSD	31.2	50.4	33.3	10.2	34.5	49.8
DSSD513 [2]	ResNet-101-DSSD	33.2	53.3	35.2	13.0	35.4	51.1
RetinaNet [7]	ResNet-101-FPN	39.1	59.1	42.3	21.8	42.7	50.2
RetinaNet [7]	ResNeXt-101-FPN	**40.8**	**61.1**	**44.1**	**24.1**	**44.2**	51.2
YOLOv3 608 × 608	Darknet-53	33.0	57.9	34.4	18.3	35.4	41.9

COCO for YOLOv3

- For clarity, AP@.75 means the AP with IoU=0.75.

COCO mAP

- **mAP** (mean average precision) is the average of AP.

- In some context, we compute the AP for each class and average them. But in some context, they mean the same thing. For example, under the COCO context, there is no difference between AP and mAP. Here is the direct quote from COCO:

> *„AP is averaged over all categories. Traditionally, this is called "mean average precision" (mAP). We make no distinction between AP and mAP (and likewise AR and mAR) and assume the difference is clear from context."*

- In ImageNet, the AUC method is used.

- To summarize: even all of them follow the same principle in measurement AP, the exact calculation may vary according to the datasets. Fortunately, development kits are available in calculating this metric.