Computer Vision

Object detection

Tae-Hyun Oh (오태현)
전자전기공학과
POSTECH

Slide by Dongmin Choi (최동민)
TAs: { Dongmin Choi , Jongha Kim, Juyong Lee, Sungbin Kim } (in alphabetic order)

boostcamp AI Tech

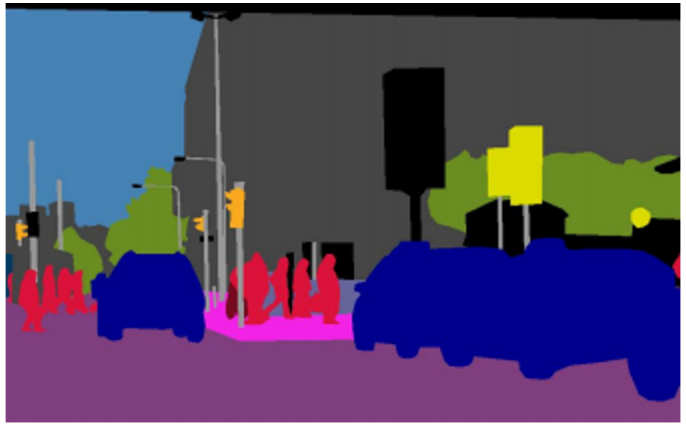
1.
Object detection

1.1 What is object detection?

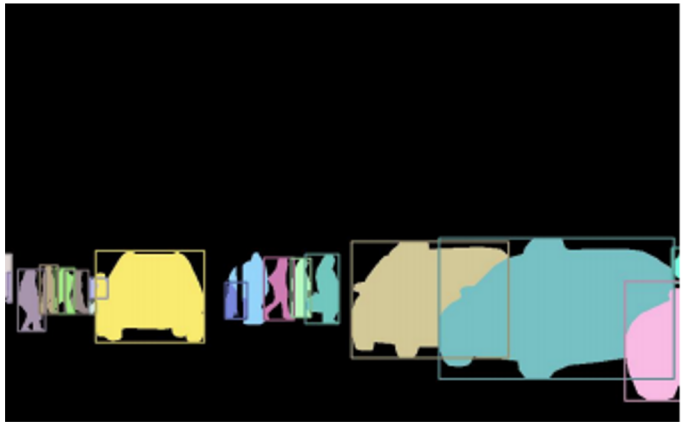
Fundamental image recognition tasks

[Kirillov et al., CVPR 2019]

Image



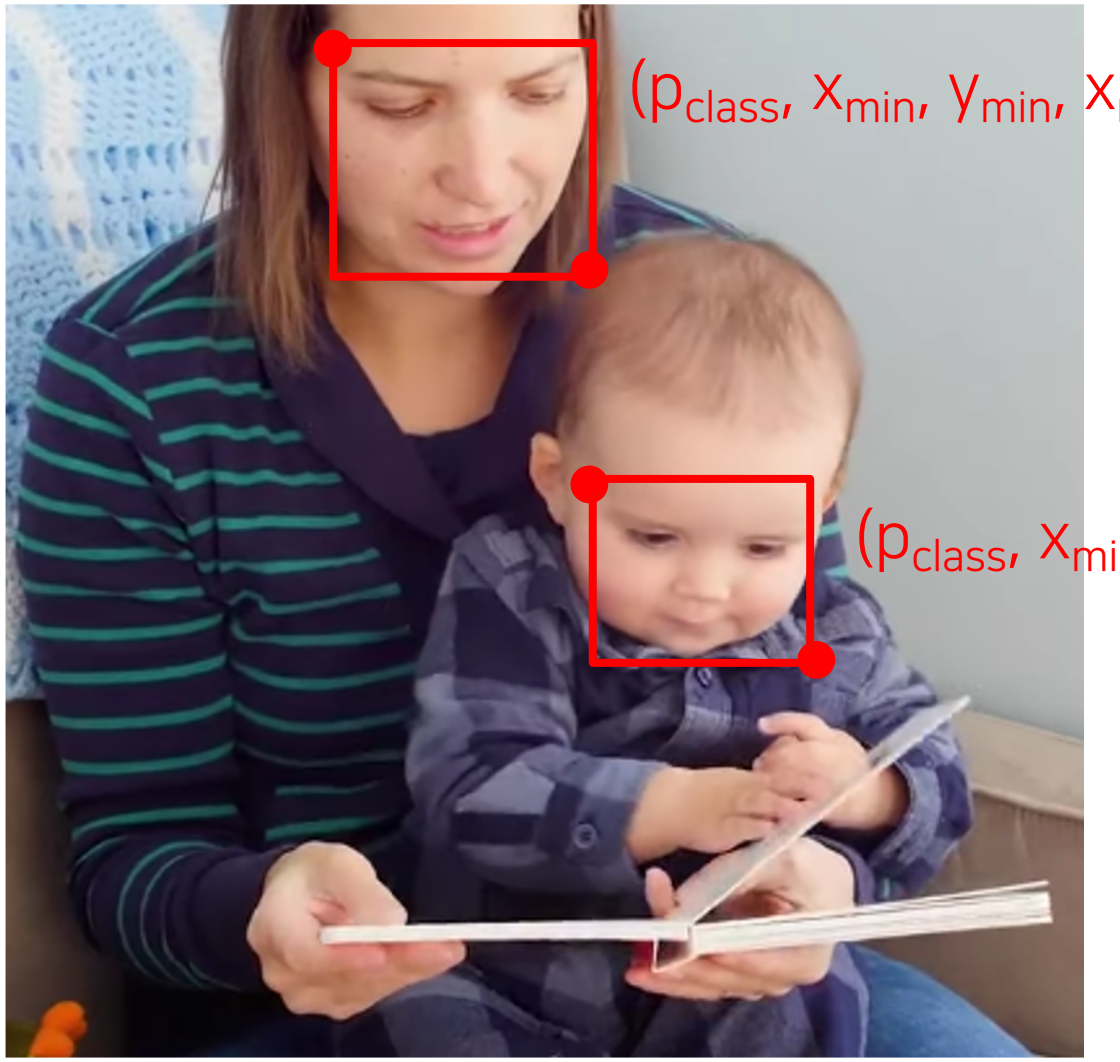
Semantic segmentation



Instance segmentation

Panoptic segmentation

1.1 What is object detection?

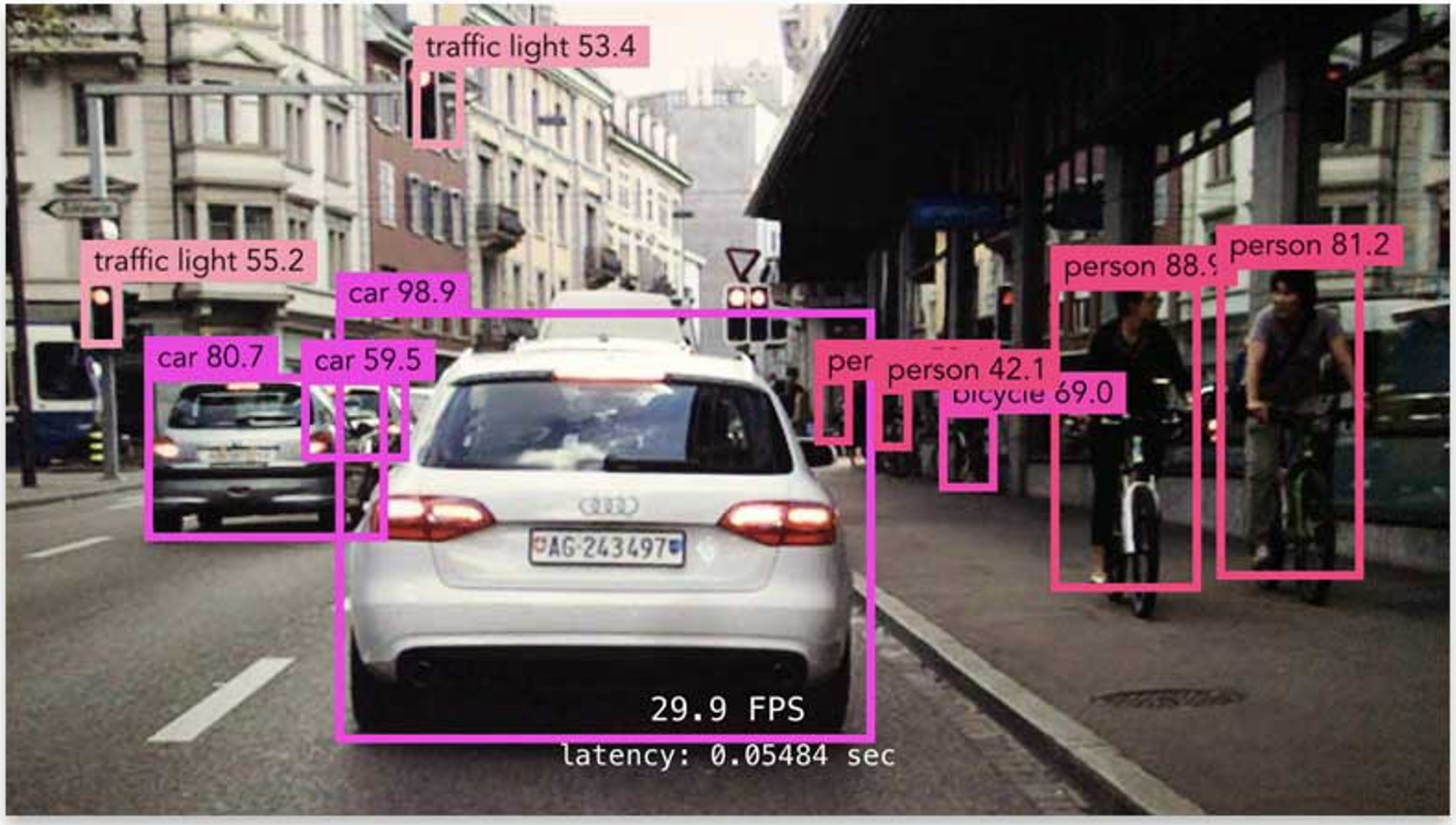


Classification + Box localization

[https://www.youtube.com/watch?v=_uY3O96iv8w]

1.2 What are the applications of object detection?

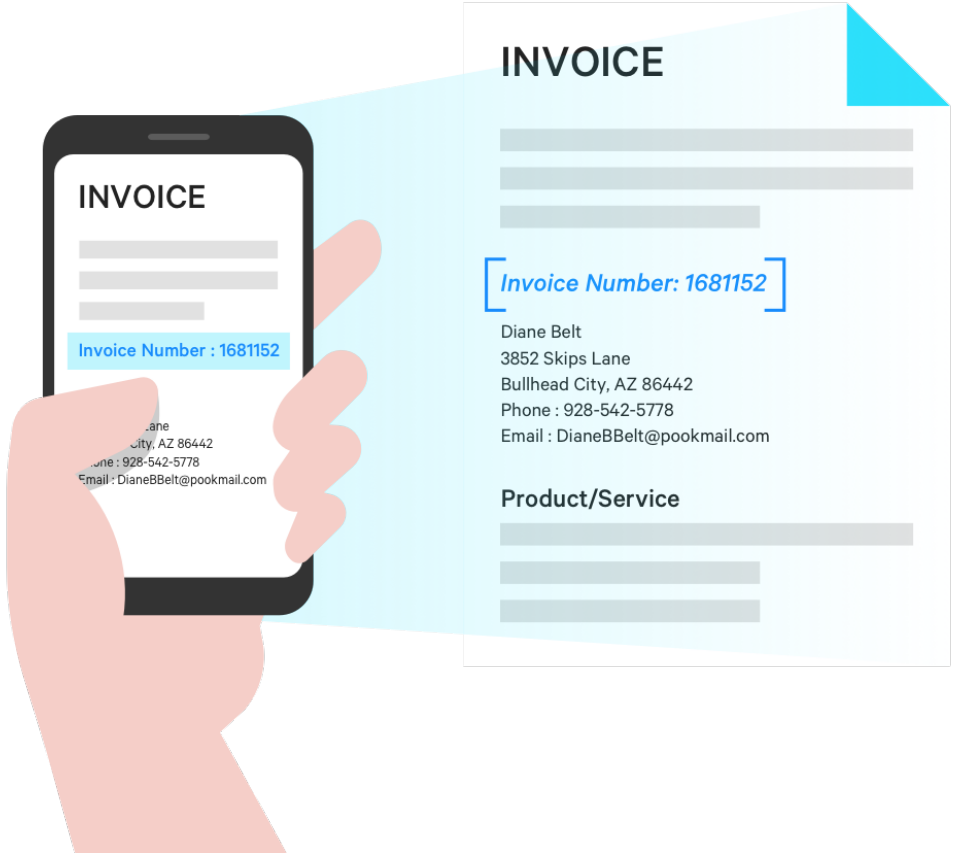
Autonomous driving

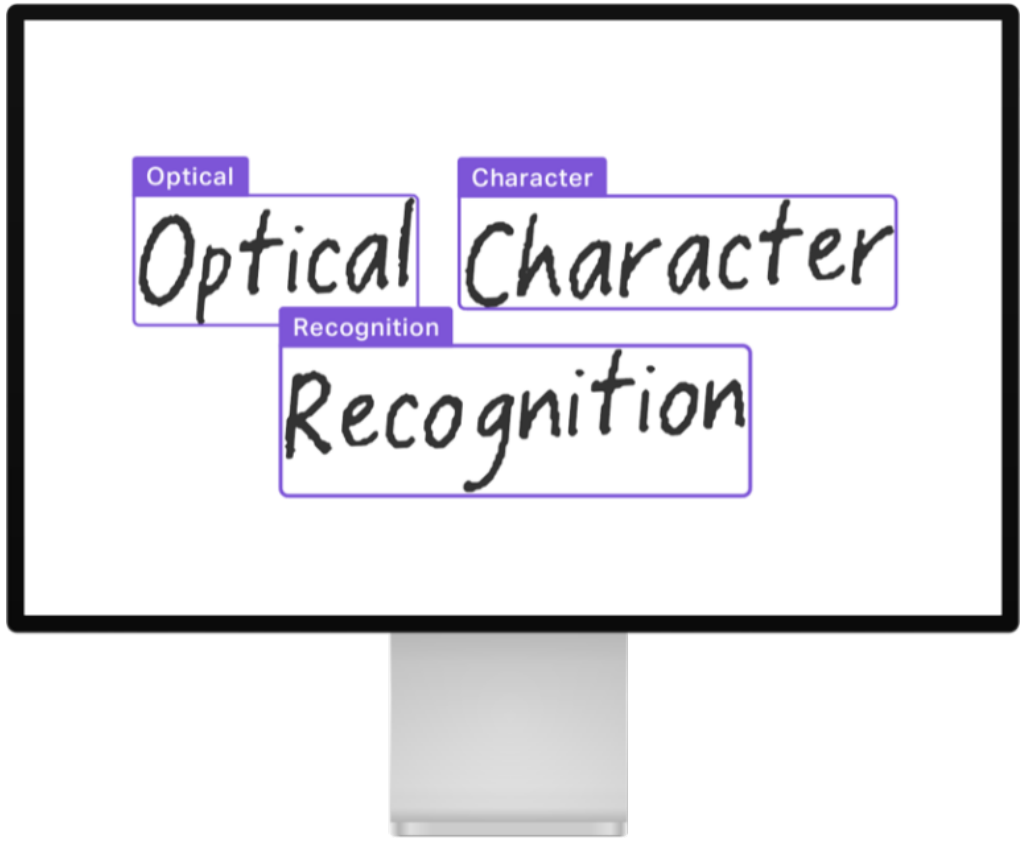


[https://medium.com/@sapiensrobot/real-time-object-detection-705c2d26b6f7]

1.2 What are the applications of object detection?

Optical Character Recognition (OCR)





[https://www.ncloud.com/product/aiService/ocr, https://medium.com/@aicompare/optical-character-recognition-ocr-which-solution-to-choose-cd4f829c4e5]

2.

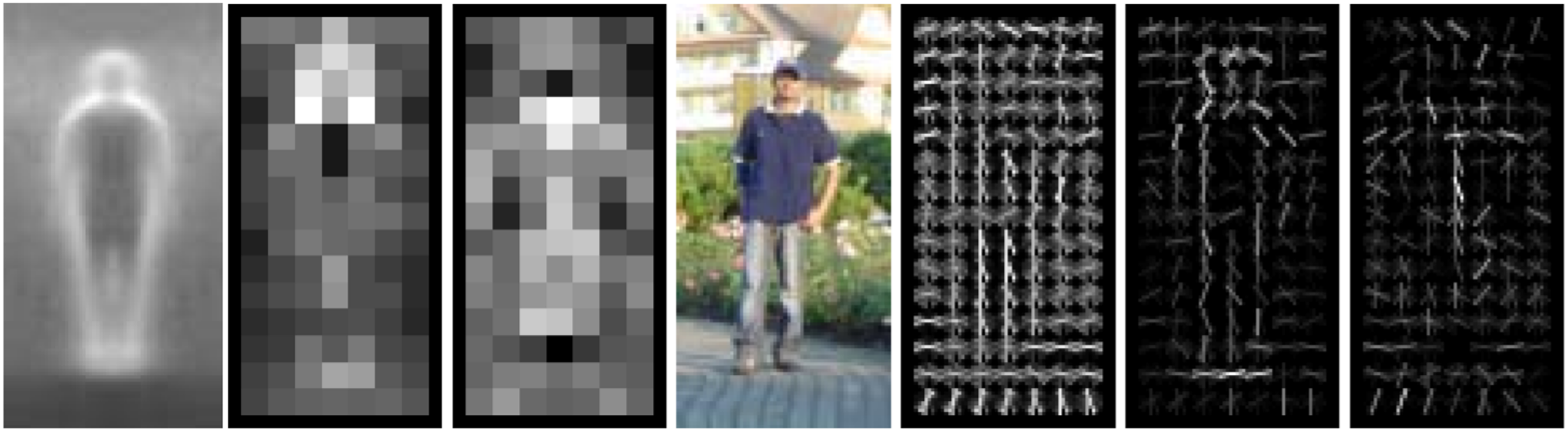
Two-stage detector

2.0 Traditional methods – hand-crafted techniques

Gradient-based detector (e.g., HOG)

[Dalal et al., CVPR 2005]

* HOG = Histogram of Oriented Gradients, SVM = Support Vector Machine



(a) Average Gradient

(b) max (+) SVM weight

(c) max (-) SVM weight

(d) Image

(e) R-HOG descriptor

(f) R-HOG w/ (+) SVM

(g) R-HOG w/ (-) SVM

2.0 Traditional methods – hand-crafted techniques

Selective search

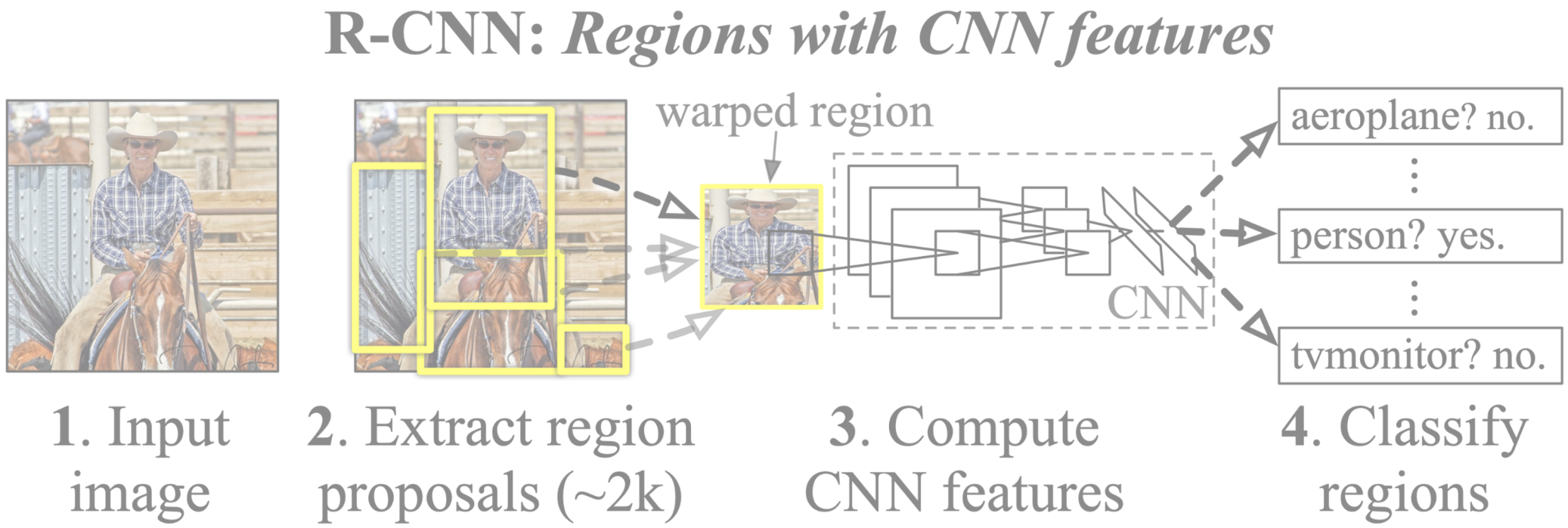
[Uijlings et al., IJCV 2013]

1. Over-segmentation
2. Iteratively merging similar regions
3. Extracting candidate boxes from all remaining segmentations

2.1 R-CNN

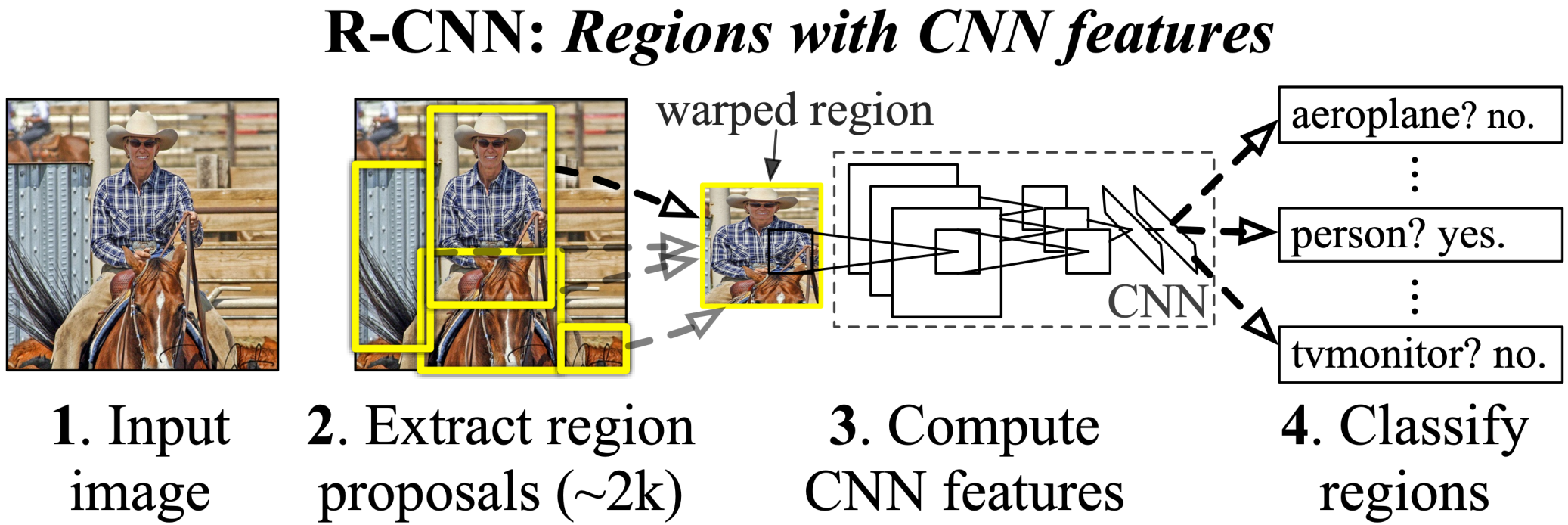
[Girshick et al., CVPR 2014]

Directly leverage image classification networks for object detection



2.1 R-CNN

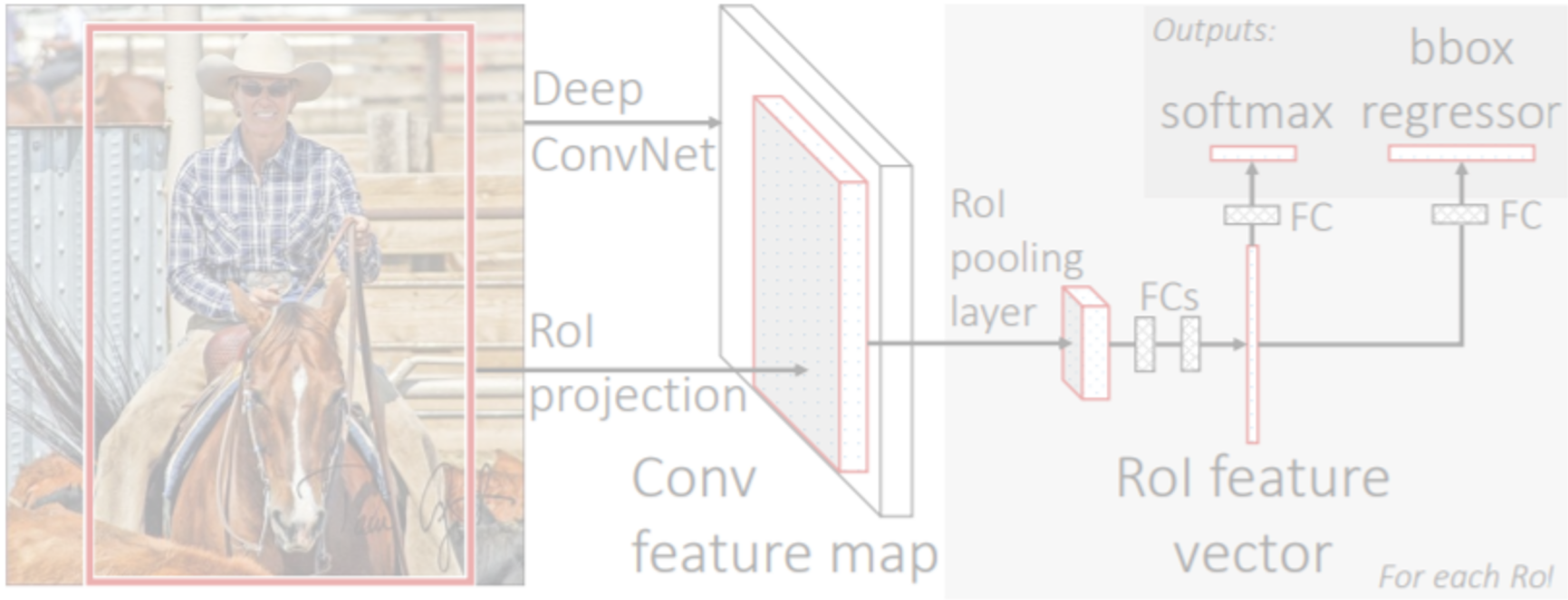
[Girshick et al., CVPR 2014]



2.2 Fast R-CNN

[Girshick et al., ICCV 2015]

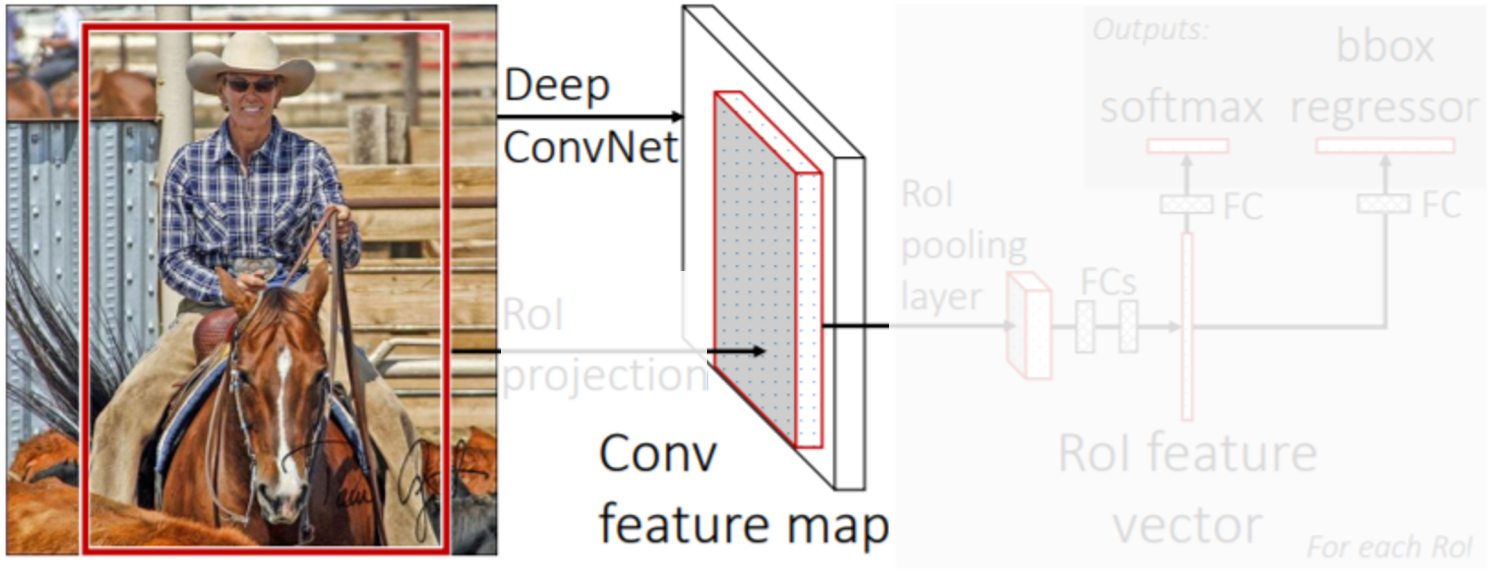
Recycle a pre-computed feature for multiple object detection



2.2 Fast R-CNN

Fast R-CNN

[Girshick et al., ICCV 2015]

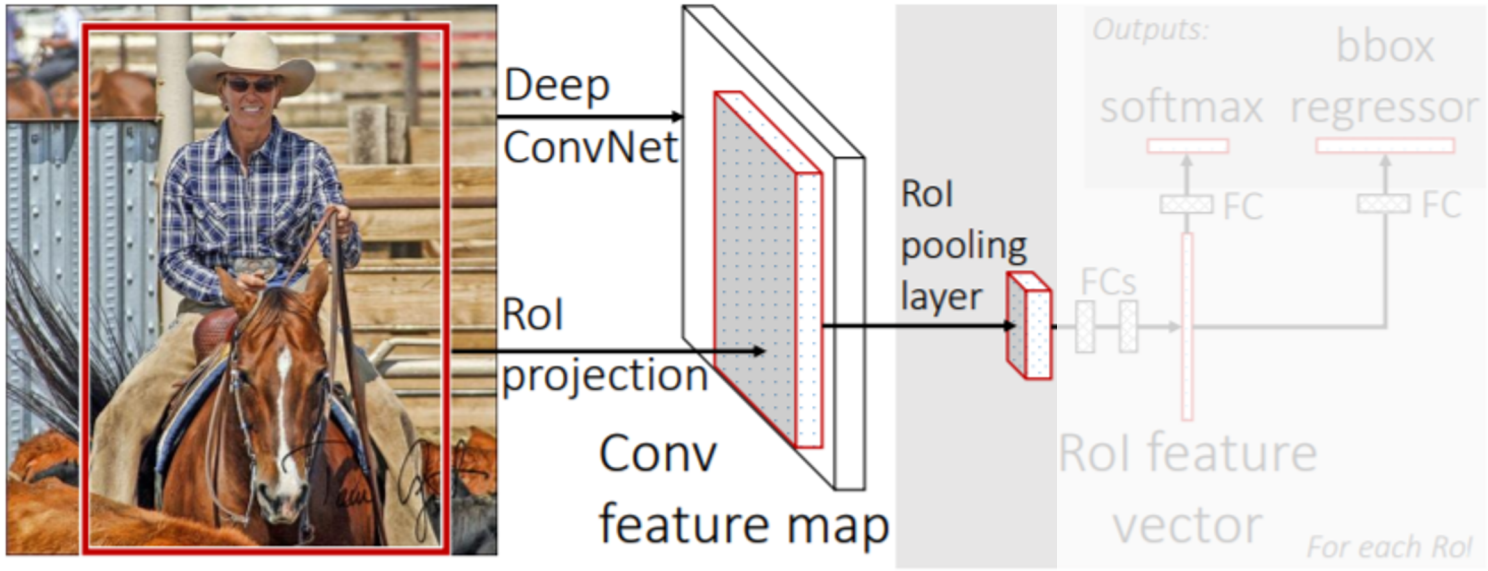


1. Conv. feature map from the original image

2.2 Fast R-CNN

[Girshick et al., ICCV 2015]

* RoI = Region of Interest

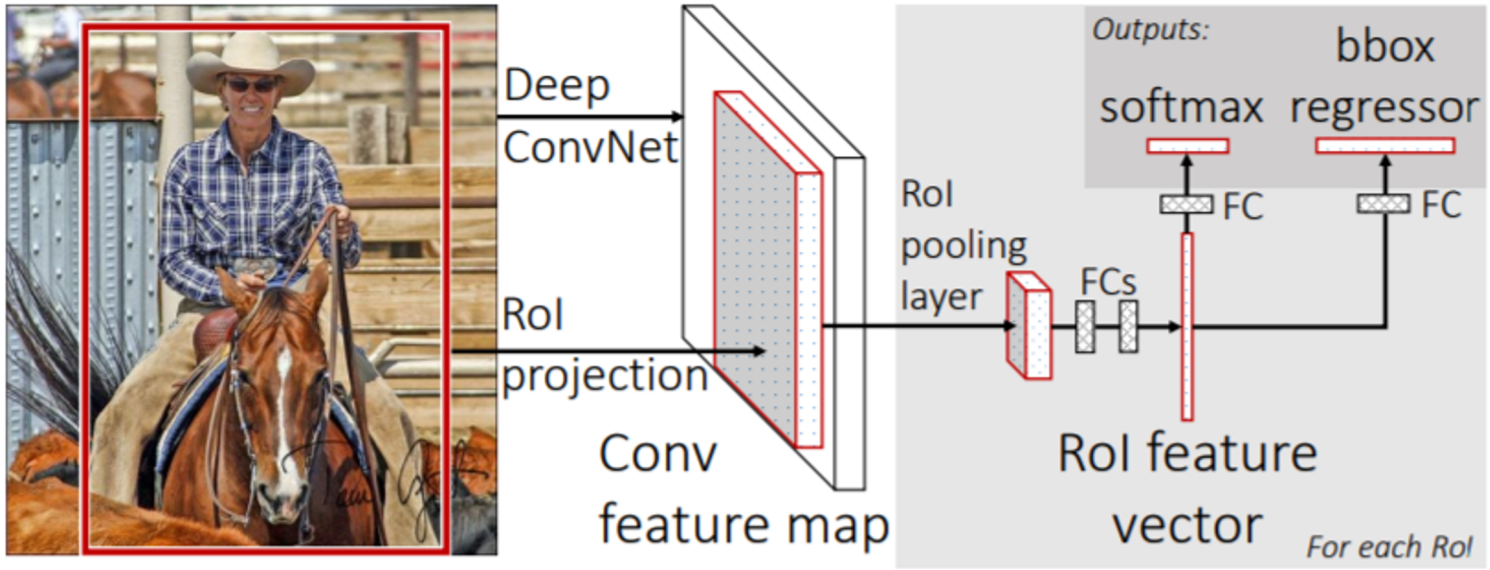


1. Conv. feature map from the original image
2. RoI feature extraction from the feature map through RoI pooling

2.2 Fast R-CNN

[Girshick et al., ICCV 2015]

* RoI = Region of Interest

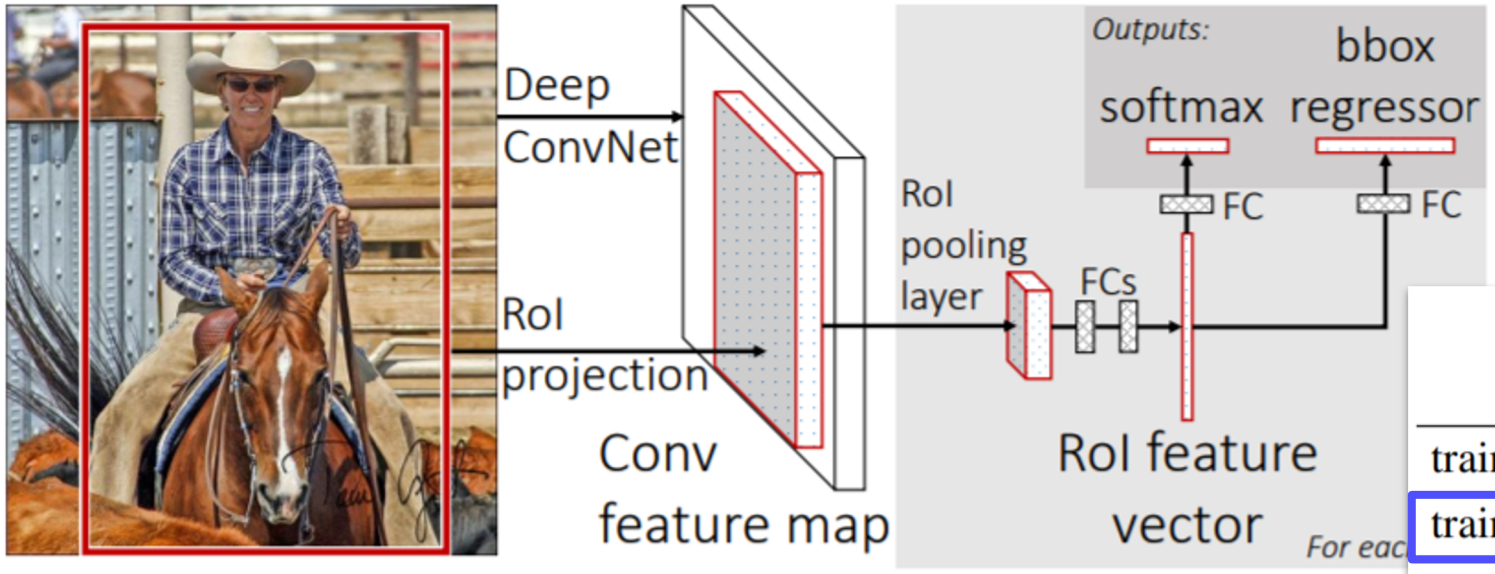


1. Conv. feature map from the original image
2. RoI feature extraction from the feature map through RoI pooling
3. Class and box prediction for each RoI

2.2 Fast R-CNN

[Girshick et al., ICCV 2015]

* RoI = Region of Interest



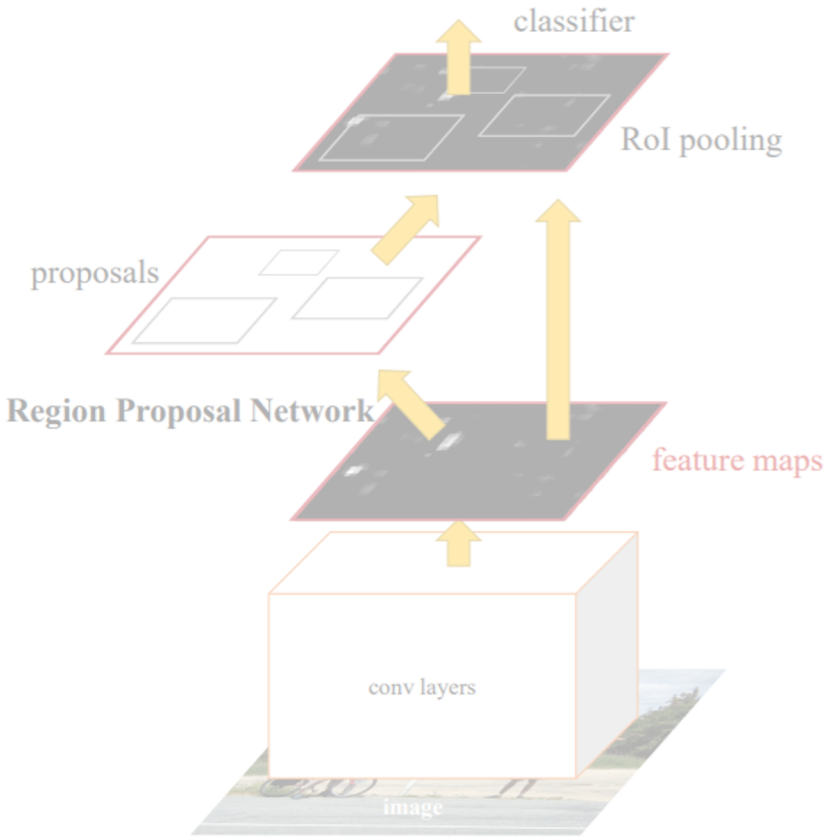
	Fast R-CNN			R-CNN		
	S	**M**	**L**	**S**	**M**	**L**
train time (h)	**1.2**	2.0	9.5	22	28	84
train speedup	**18.3**×	14.0×	8.8×	1×	1×	1×
test rate (s/im)	0.10	0.15	0.32	9.8	12.1	47.0
▷ with SVD	**0.06**	0.08	0.22	-	-	-
test speedup	98×	80×	146×	1×	1×	1×
▷ with SVD	169×	150×	**213**×	-	-	-
VOC07 mAP	57.1	59.2	**66.9**	58.5	60.2	66.0
▷ with SVD	56.5	58.7	66.6	-	-	-

1. Conv. feature map from the original image
2. RoI feature extraction from the feature map through RoI pooling
3. Class and box prediction for each RoI

2.3 Faster R-CNN

[Ren et al., NeurIPS, 2015]

End-to-end object detection by neural region proposal



2.3 Faster R-CNN

$$IoU = \frac{\text{Area of Overlap}}{\text{Area of Union}}$$

IoU (Intersection over Union)

: A metric commonly used in object detection

[https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/]

2.3 Faster R-CNN

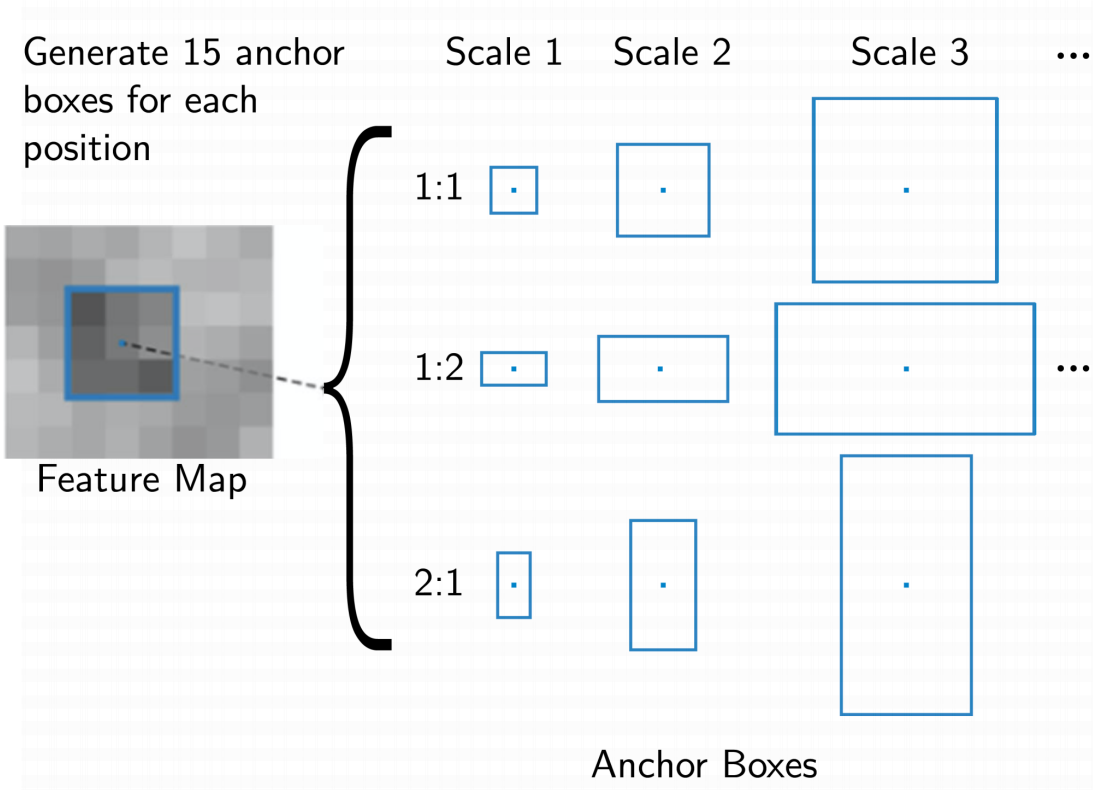


Illustration from [Ferguson et al. 2018]

Anchor boxes

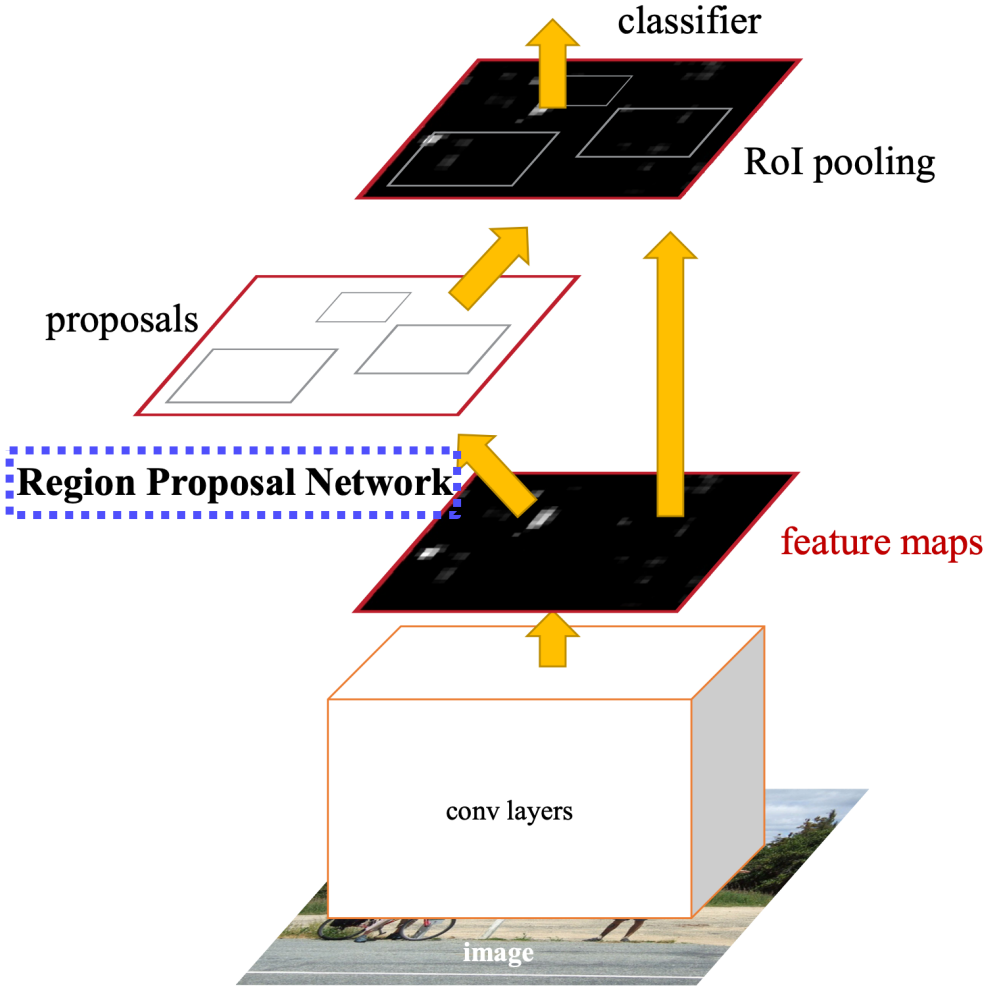
- A set of pre-defined bounding boxes
- IoU with GT > 0.7 ⇒ positive sample
- IoU with GT < 0.3 ⇒ negative sample

2.3 Faster R-CNN

[Ren et al., NeurIPS, 2015]

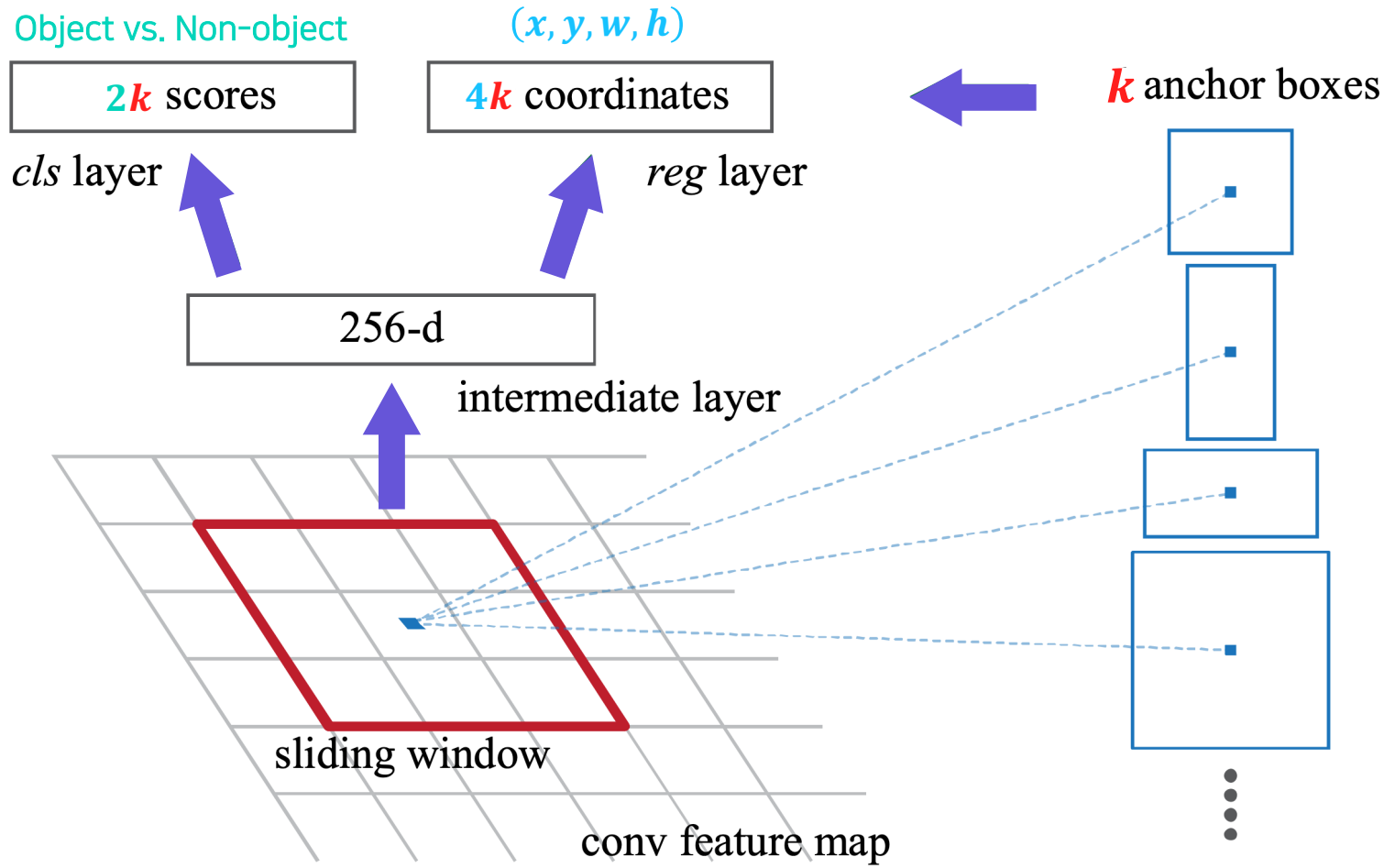
Time-consuming selective search

Region Proposal Network (RPN)



2.3 Faster R-CNN

[Ren et al., NeurIPS, 2015]



2.3 Faster R-CNN

Non-Maximum Suppression (NMS)

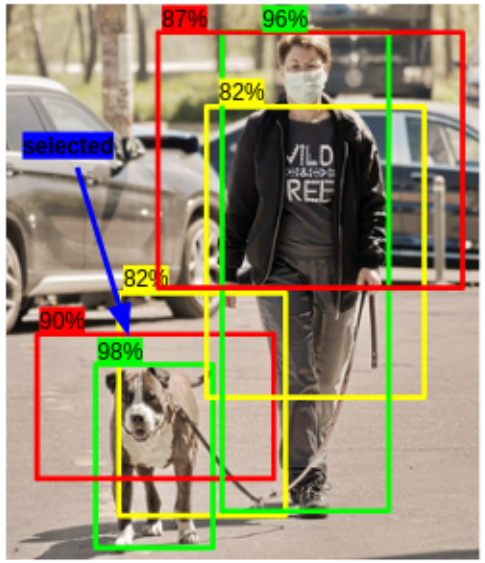
Step 1: Select the box with the highest objectiveness score

Step 2: Compare IoU of this box with other boxes

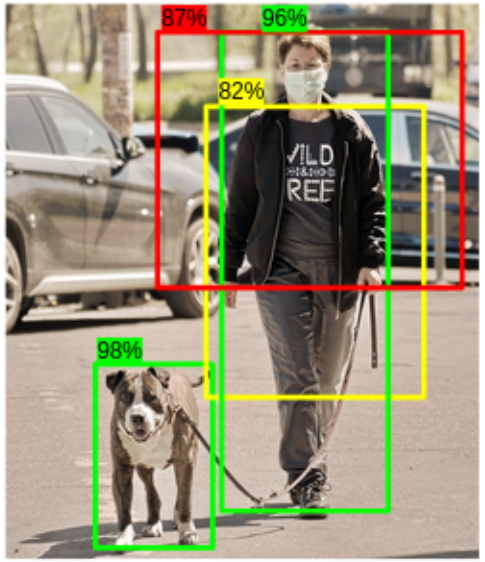
Step 3: Remove the bounding boxes with IoU ≥ 50%

Step 4: Move to the next highest objectiveness score

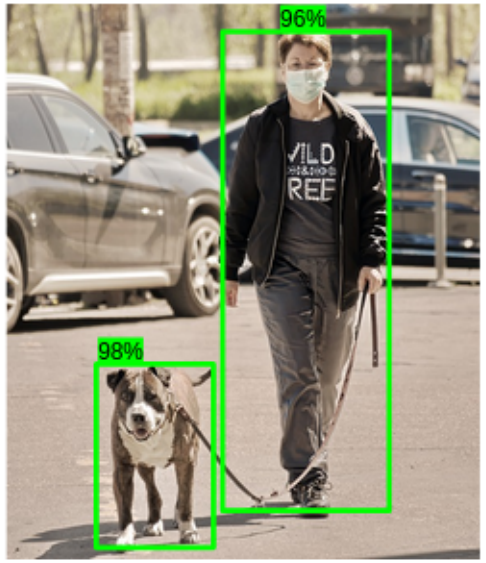
Step 5: Repeat steps 2-4



Step 1: Selecting Bounding box with highest score



Step 3: Delete Bounding box with high overlap

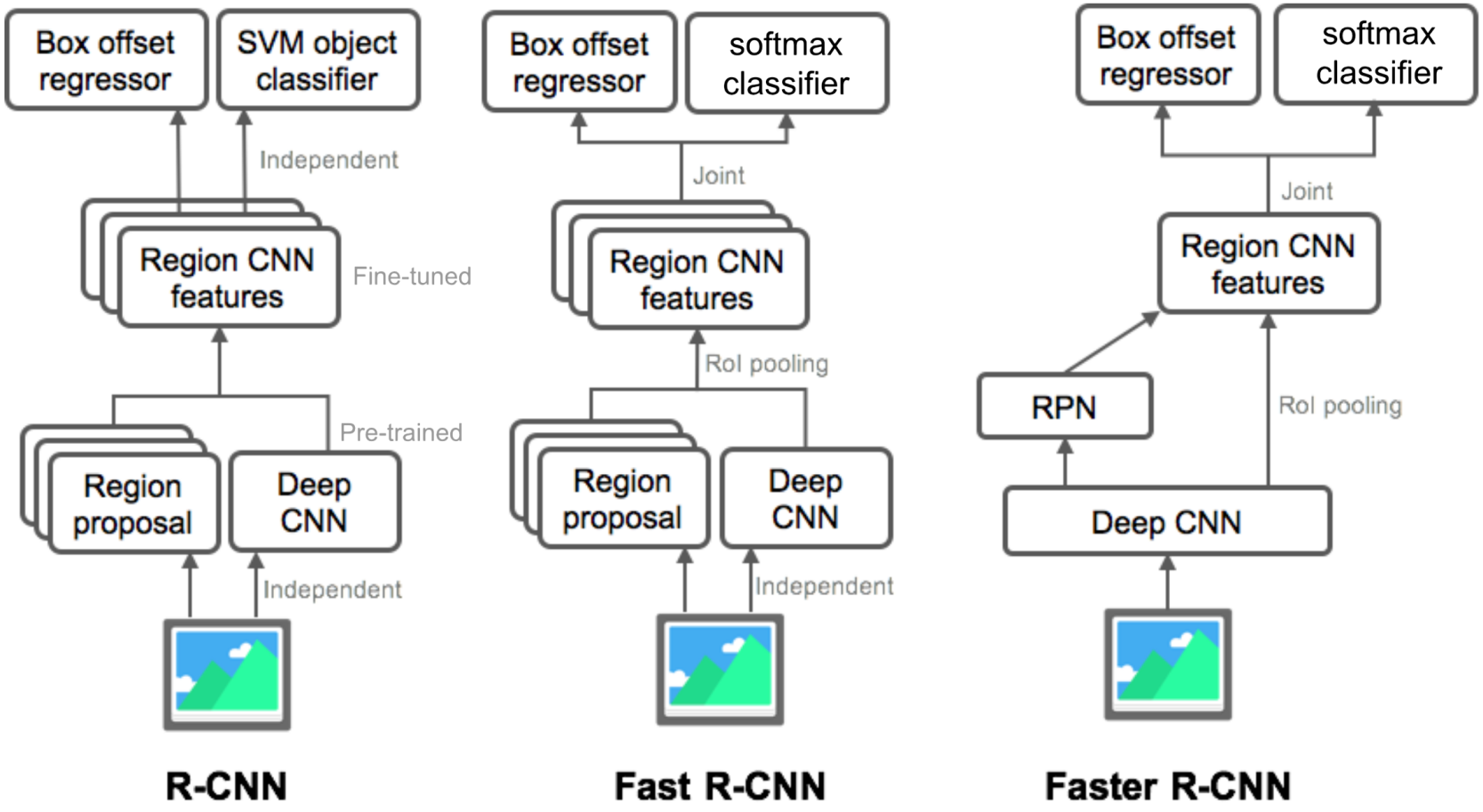


Step 5: Final Output

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

2.3 Faster R-CNN

Summary of the R-CNN family



[https://lilianweng.github.io/lil-log/2017/12/31/object-recognition-for-dummies-part-3.html]

3.

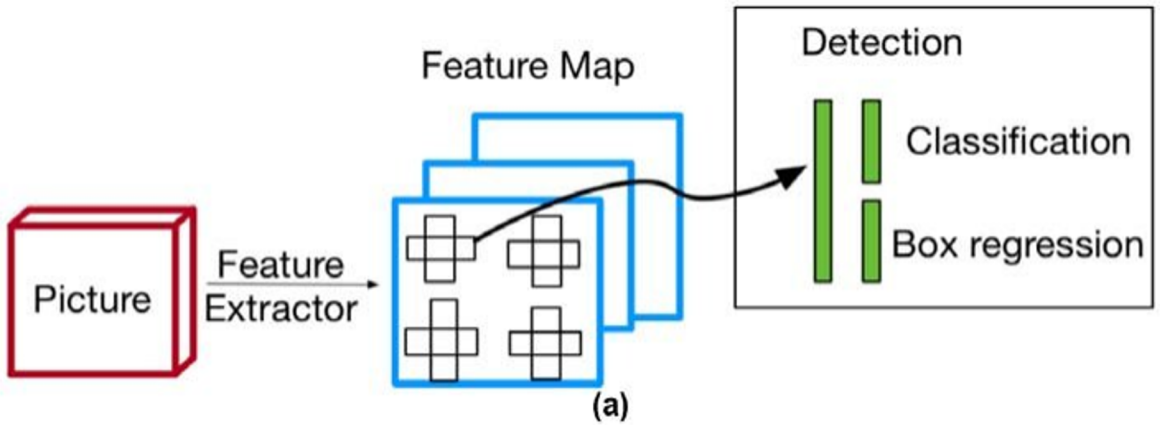
Single-stage detector

3.0 Comparison with two-stage detectors

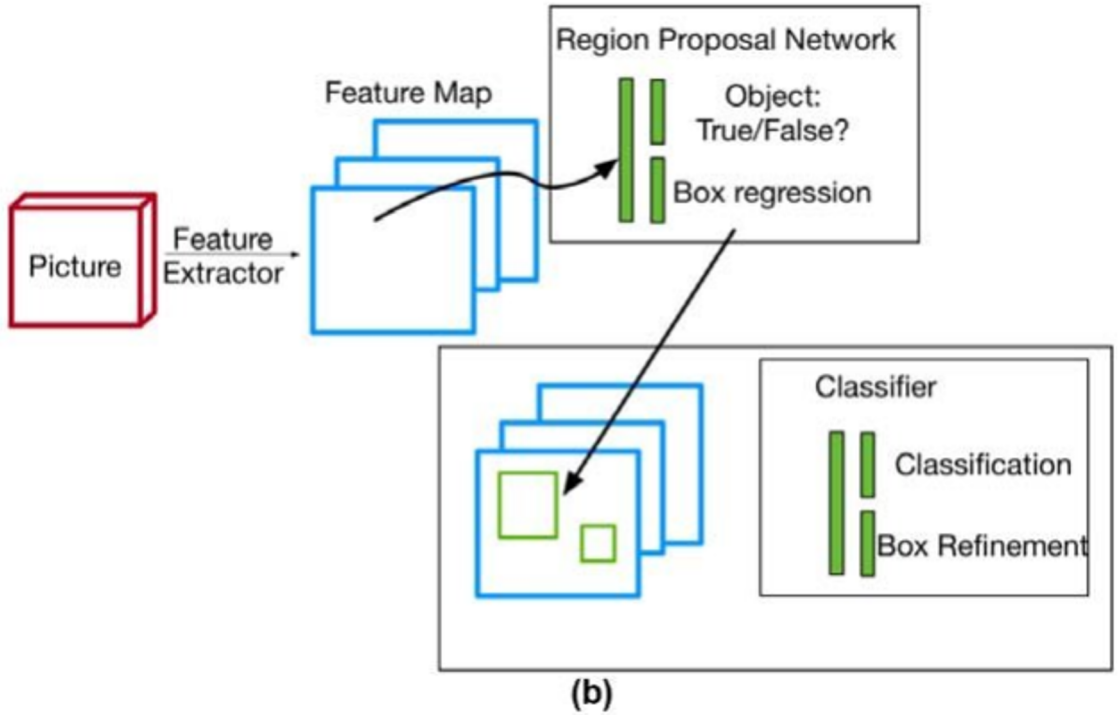
One-stage vs. two-stage

Illustrations from [Ndonhong et al., Offshore Technology Conference 2019]

- No explicit RoI pooling



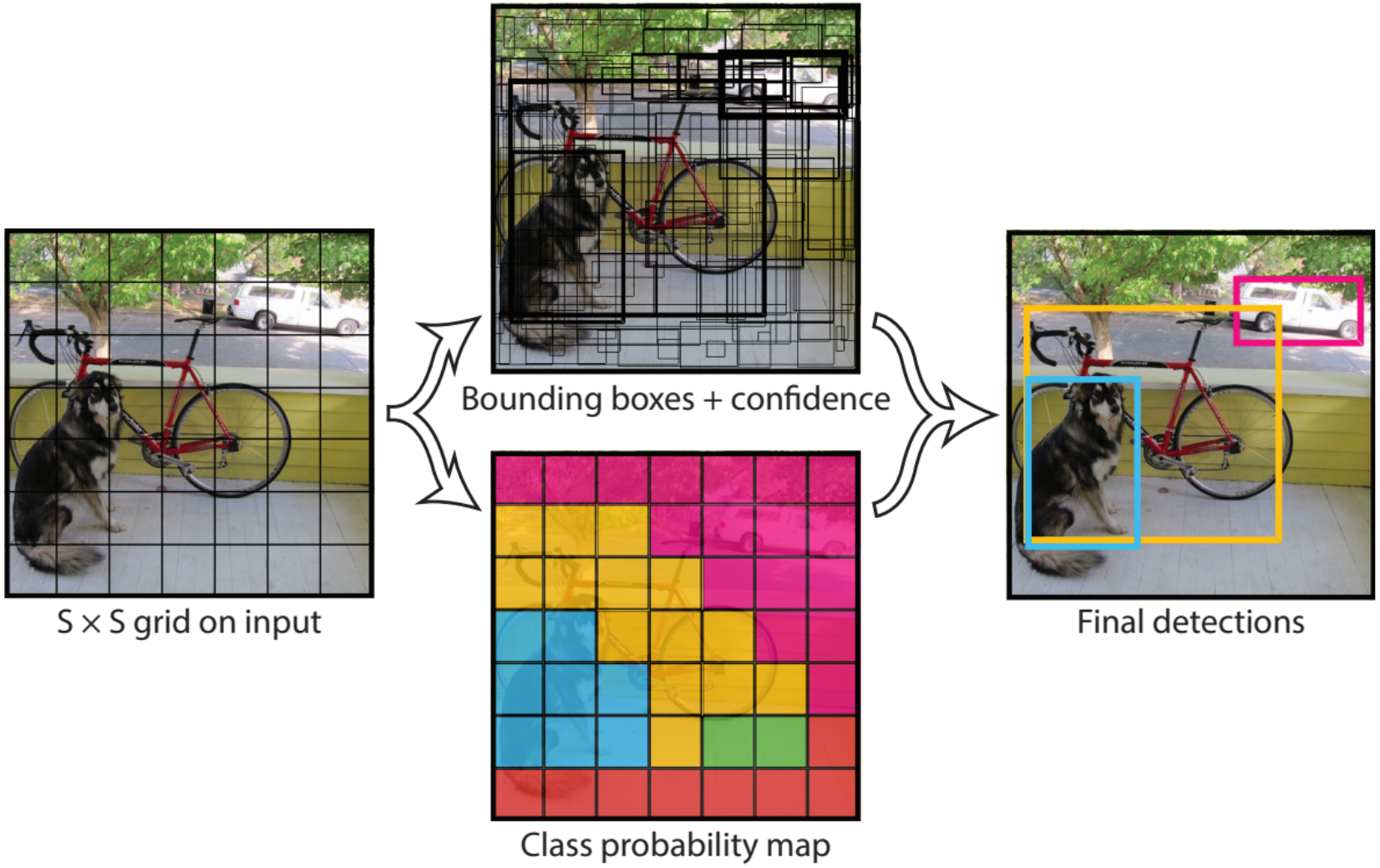
One-stage detector



Two-stage detector

3.1 You only look once (YOLO)

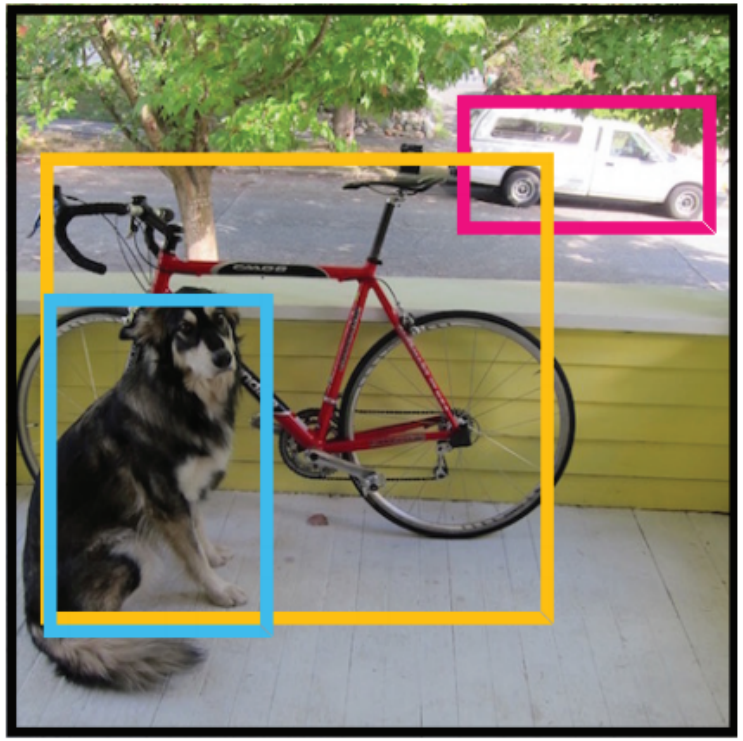
[Redmon et al., CVPR 2016]

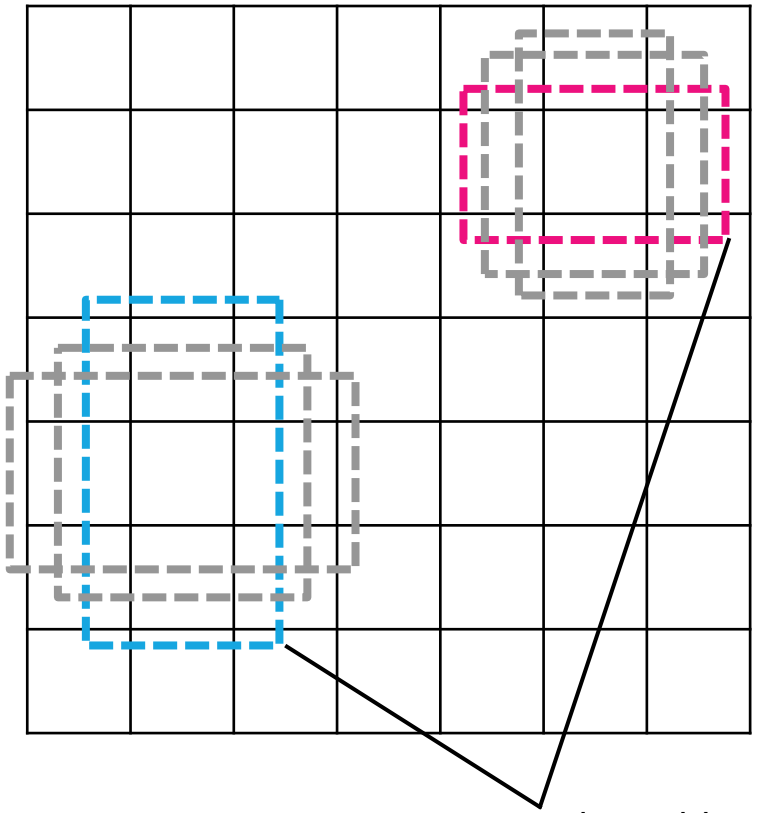


3.1 YOLO

[Redmon et al., CVPR 2016]

* Anchor boxes for yellow bicycle are omitted



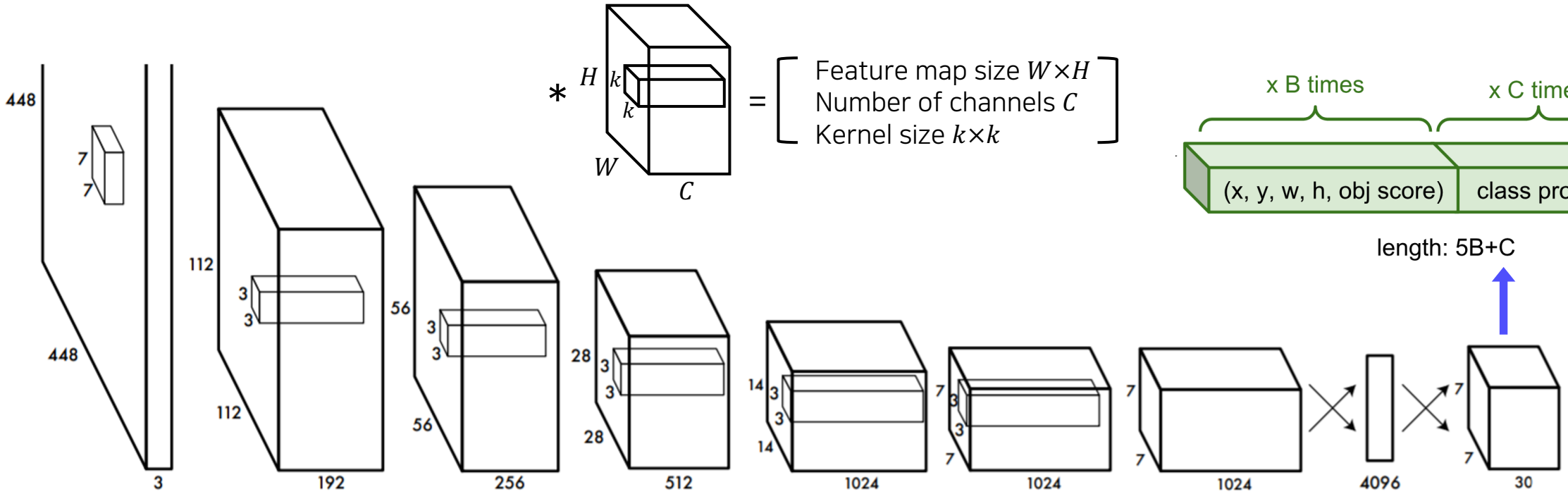


Colored boxes are
matched positive samples

3.1 YOLO

One-stage detector

[Redmon et al., CVPR 2016]



Conv. Layer	Conv. Layer	Conv. Layers	Conv. Layers	Conv. Layers	Conv. Layers	Conn. Layer	Conn. Layer
7x7x64-s-2	3x3x192	1x1x128	1x1x256 } ×4	1x1x512 } ×2	3x3x1024		
Maxpool Layer	Maxpool Layer	3x3x256	3x3x512	3x3x1024	3x3x1024		
2x2-s-2	2x2-s-2	1x1x256	1x1x512	3x3x1024			
		3x3x512	3x3x1024	3x3x1024-s-2			
		Maxpool Layer	Maxpool Layer				
		2x2-s-2	2x2-s-2				

< DarkNet >

[https://lilianweng.github.io/lil-log/2018/12/27/object-detection-part-4.html]

3.1 YOLO

One-stage detector

[Redmon et al., CVPR 2016]

* mAP : mean Average Precision (metrics)

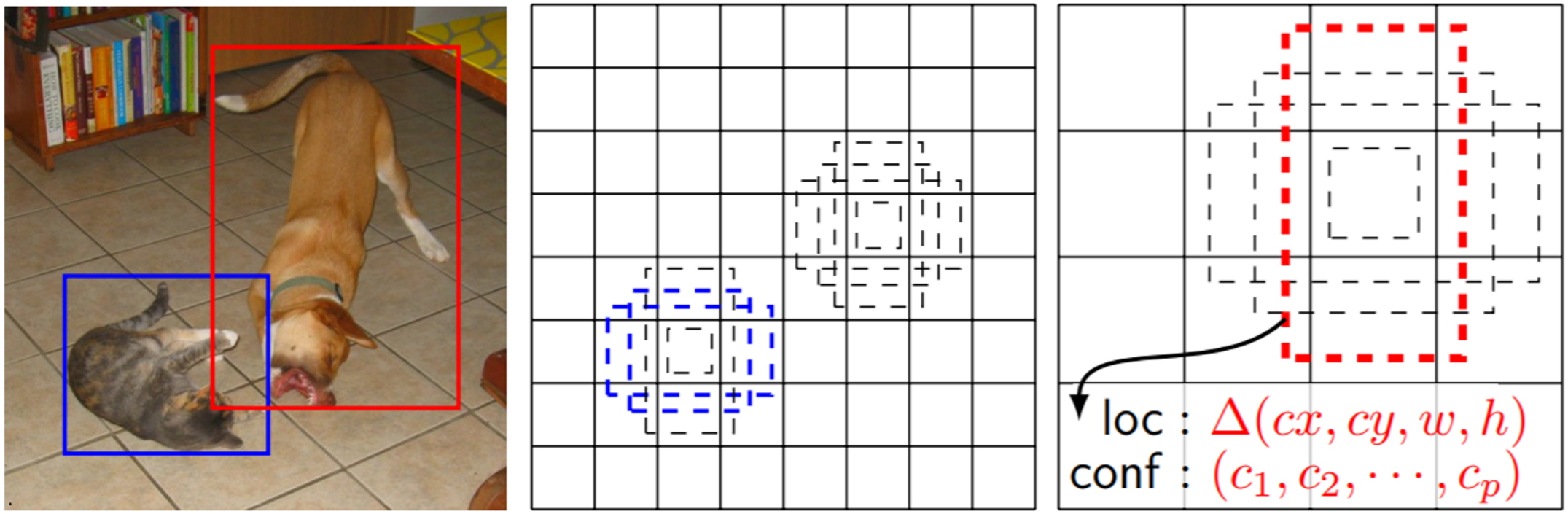
* FPS : Frame Per Second (speed)

Real-Time Detectors	Train	mAP	FPS
100Hz DPM [31]	2007	16.0	100
30Hz DPM [31]	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM [38]	2007	30.4	15
R-CNN Minus R [20]	2007	53.5	6
Fast R-CNN [14]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[28]	2007+2012	73.2	7
Faster R-CNN ZF [28]	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21

3.2 Single Shot MultiBox Detector (SSD)

One-stage detector

[Liu et al., ECCV 2016]



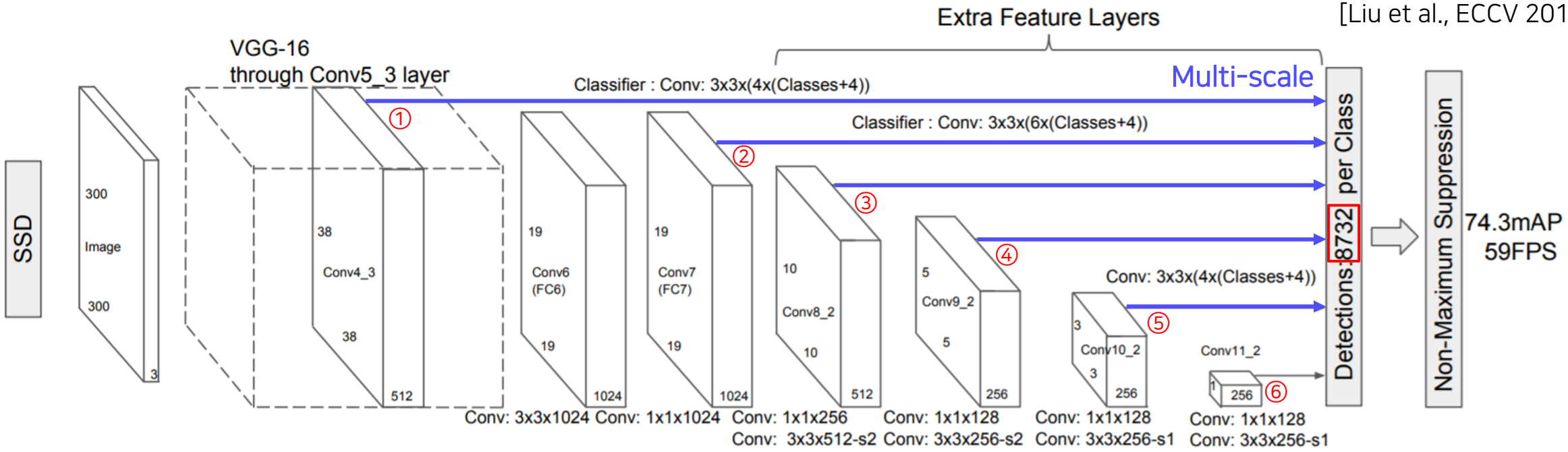
(a) Image with GT boxes (b) 8×8 feature map (c) 4×4 feature map

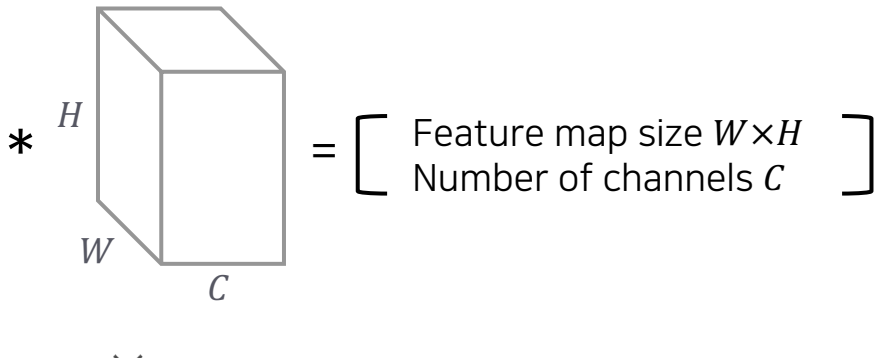
The use of multi-scale outputs attached to multiple feature maps
enable effectively modeling a diverse space of possible box shapes

3.2 SSD

One-stage detector

[Liu et al., ECCV 2016]





$*$ = [Feature map size $W \times H$
Number of channels C]

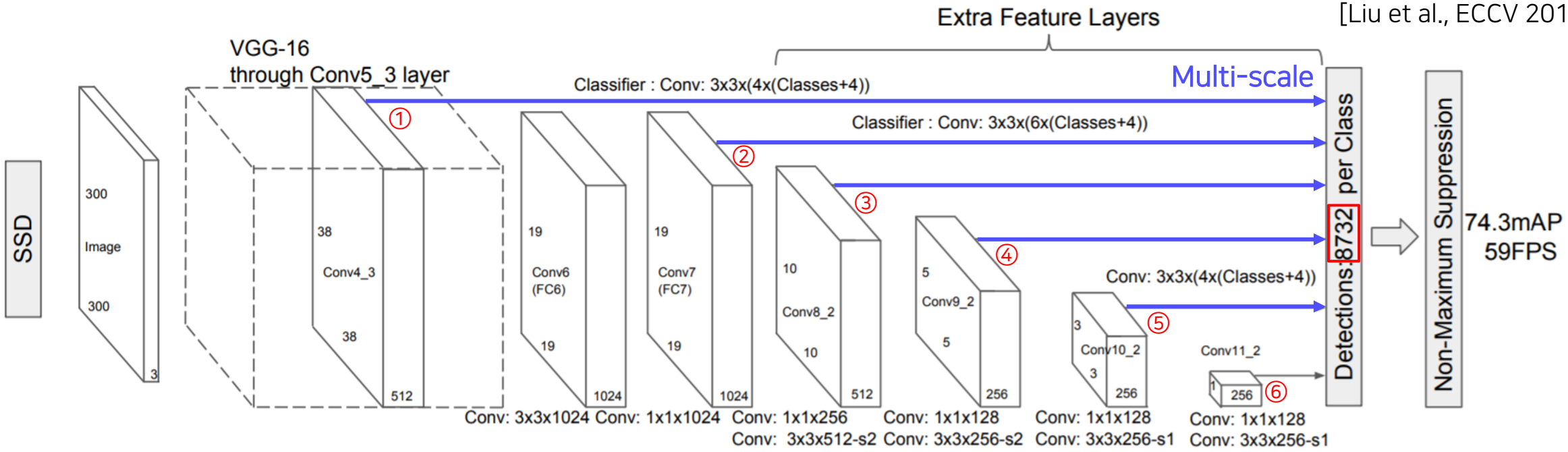
$\times$

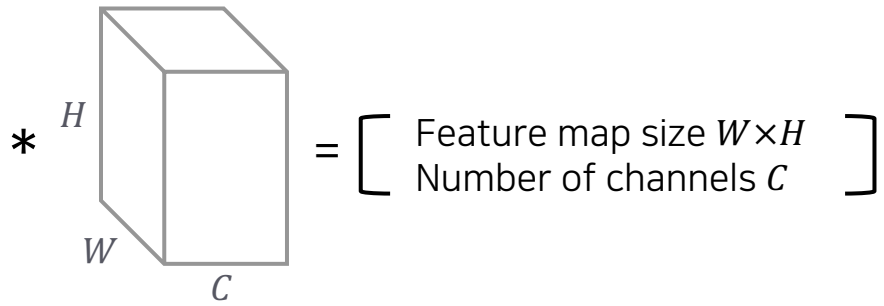
[https://medium.com/@lh.long/what-is-8732-boxes-in-ssd-7df6904a78ae]

3.2 SSD

One-stage detector

[Liu et al., ECCV 2016]





Feature map size

Number of anchor boxes for each pixel in feature map

$* \overline{W \times H} \times N$

① $38 \times 38 \times 4 = 5776$
② $19 \times 19 \times 6 = 2166$
③ $10 \times 10 \times 6 = 600$
④ $5 \times 5 \times 6 = 150$
⑤ $3 \times 3 \times 4 = 36$
⑥ $1 \times 1 \times 4 = 4$

Total number of anchor boxes

$\rightarrow 5776 + 2166 + 600 + 150 + 36 + 4 = 8732$

[https://medium.com/@lh.long/what-is-8732-boxes-in-ssd-7df6904a78ae]

3.2 SSD

[Liu et al., ECCV 2016]

* SSD300 = SSD for 300×300 input

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

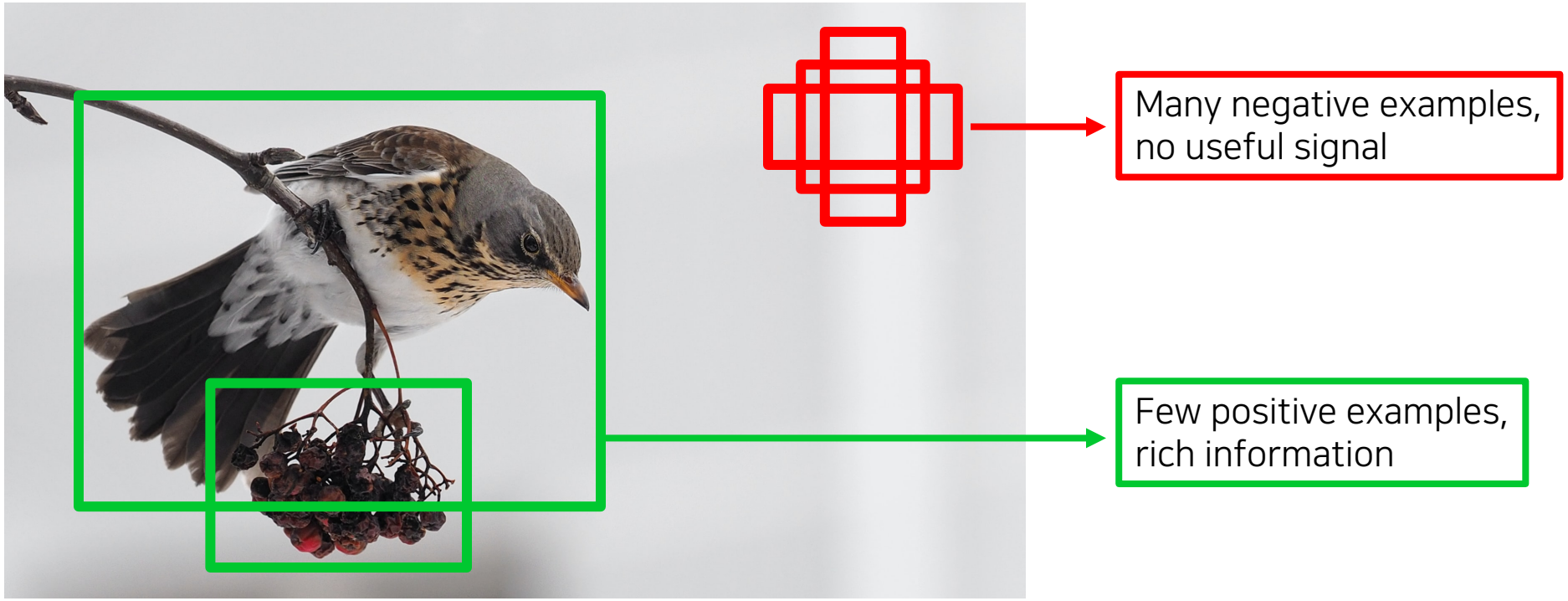
Table 7: **Results on Pascal VOC2007 `test`.** SSD300 is the only real-time detection method that can achieve above 70% mAP. By using a larger input image, SSD512 outperforms all methods on accuracy while maintaining a close to real-time speed.

4.

Two-stage detector vs. one-stage detector

4.1 Focal loss

Class imbalance problem

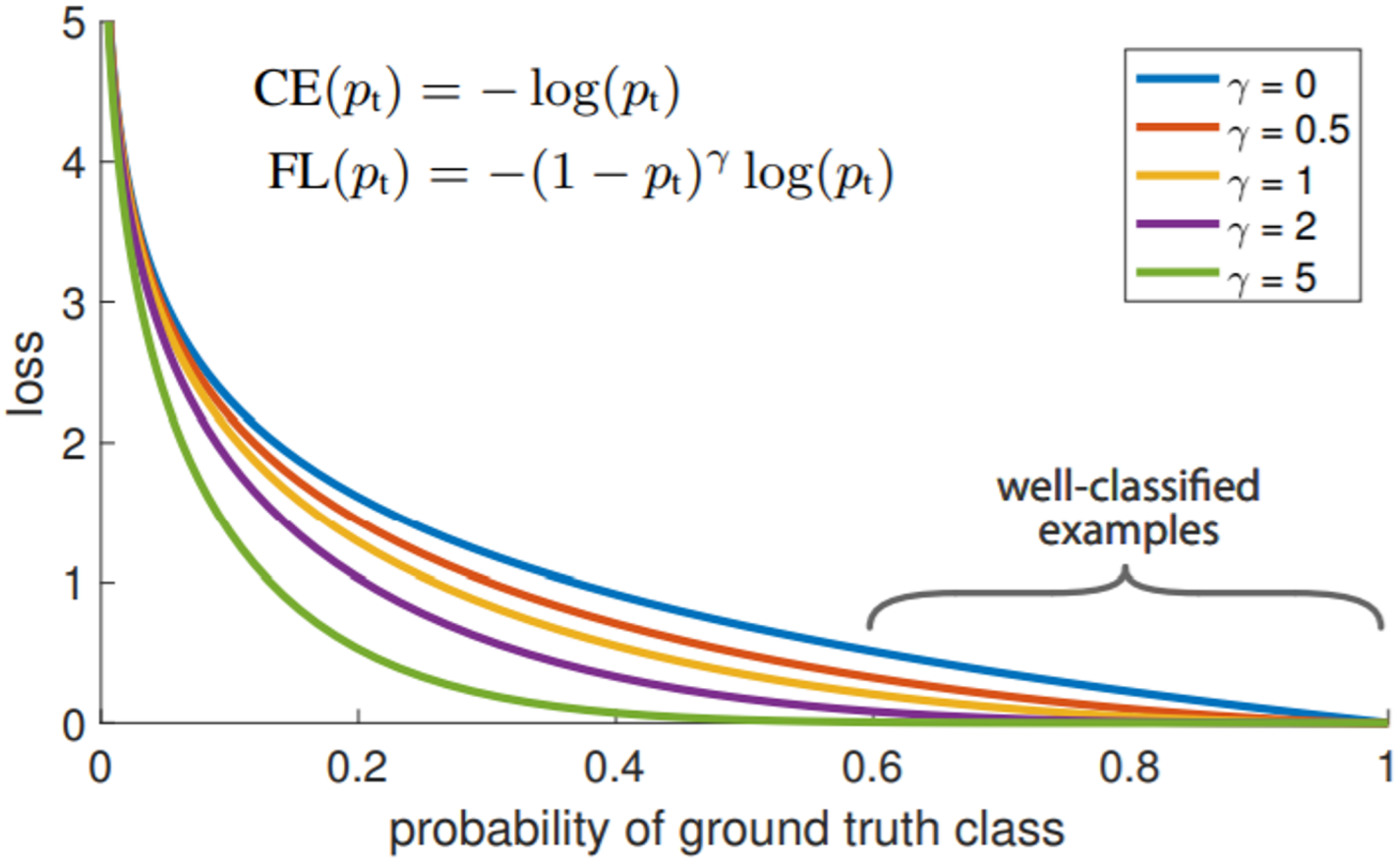


Class imbalance (#neg. anchor boxes >> #pos. anchor boxes)

Single-stage detectors (e.g., YOLO, SSD) are prone to this problem…

[https://unsplash.com/photos/GmVv2VmIxlc]

[Lin et al., ICCV 2017]



Focal loss

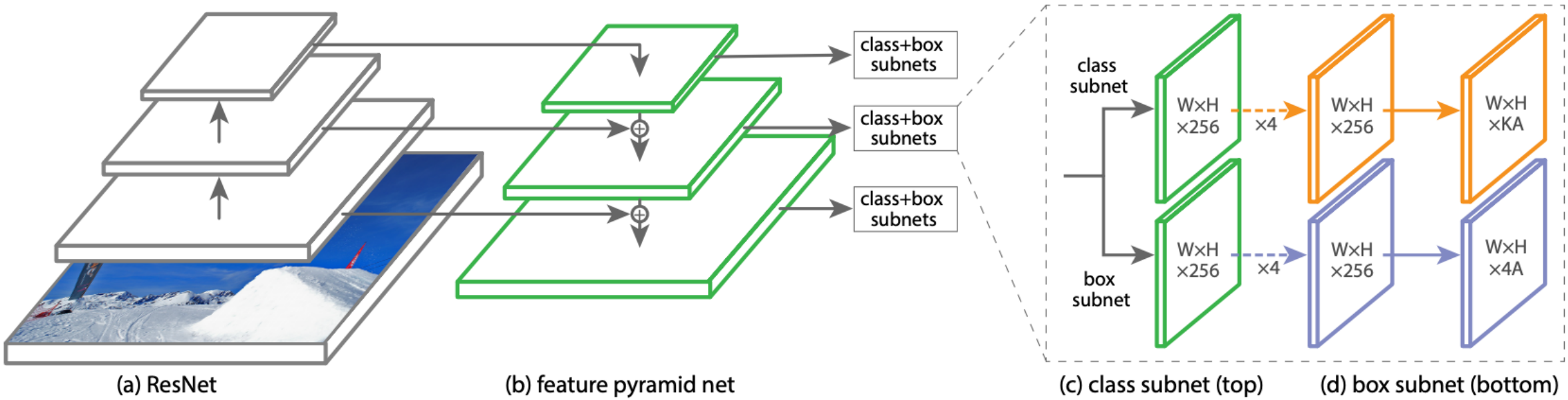
- Improved cross entropy loss
- Deal with class imbalance
  - Over-weights hard or misclassified examples
  - Down-weights easy examples

4.2 RetinaNet

RetinaNet is a one-stage network

[Lin et al., ICCV 2017]

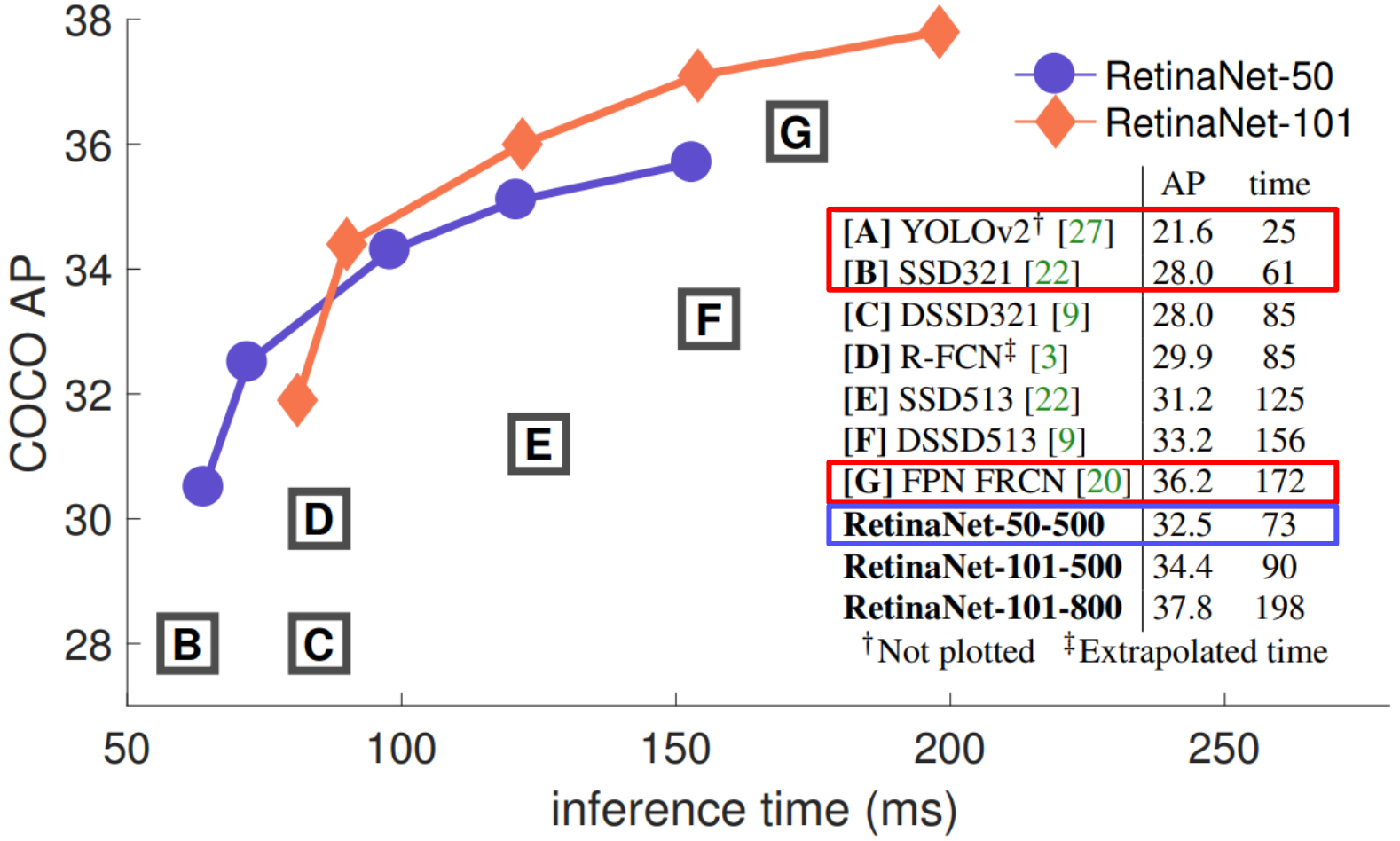
Feature Pyramid Networks (FPN) + class/box prediction branches



4.2 RetinaNet

[Lin et al., ICCV 2017]

* RetinaNet-50-500 = ResNet-50 backbone RetinaNet for 500×500 input



	AP	time
[A] YOLOv2† [27]	21.6	25
[B] SSD321 [22]	28.0	61
[C] DSSD321 [9]	28.0	85
[D] R-FCN‡ [3]	29.9	85
[E] SSD513 [22]	31.2	125
[F] DSSD513 [9]	33.2	156
[G] FPN FRCN [20]	36.2	172
RetinaNet-50-500	32.5	73
RetinaNet-101-500	34.4	90
RetinaNet-101-800	37.8	198

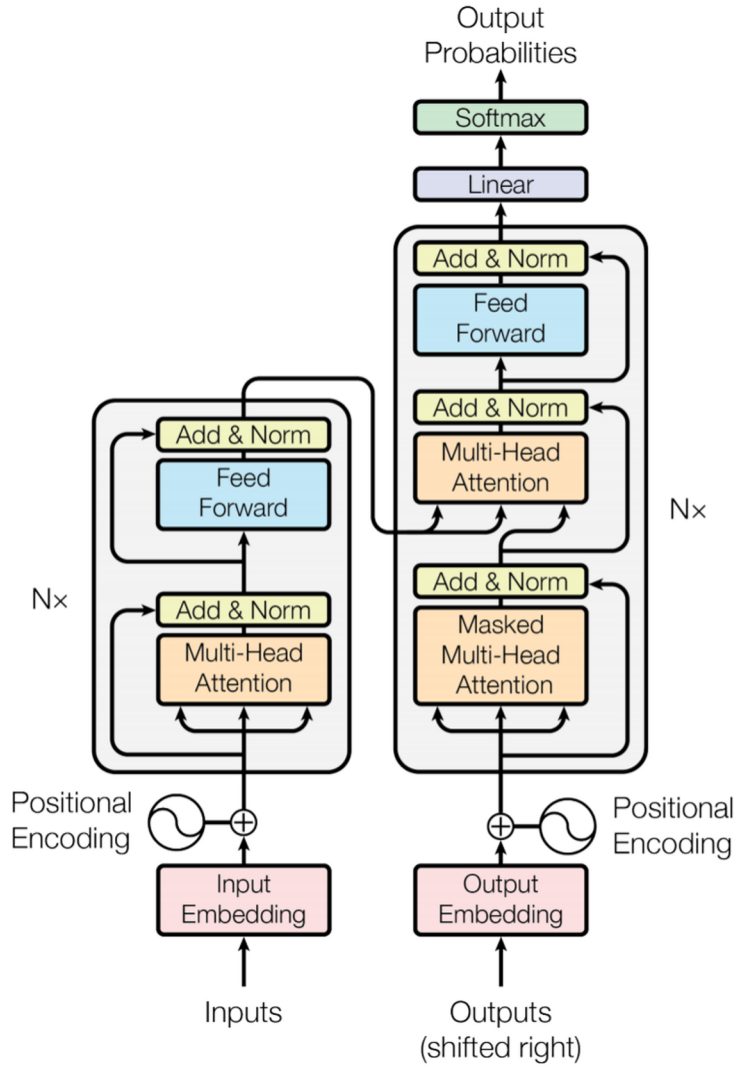
†Not plotted ‡Extrapolated time

5.

Detection with Transformer

Transformer

[Vaswani et al., NeurIPS 2017]



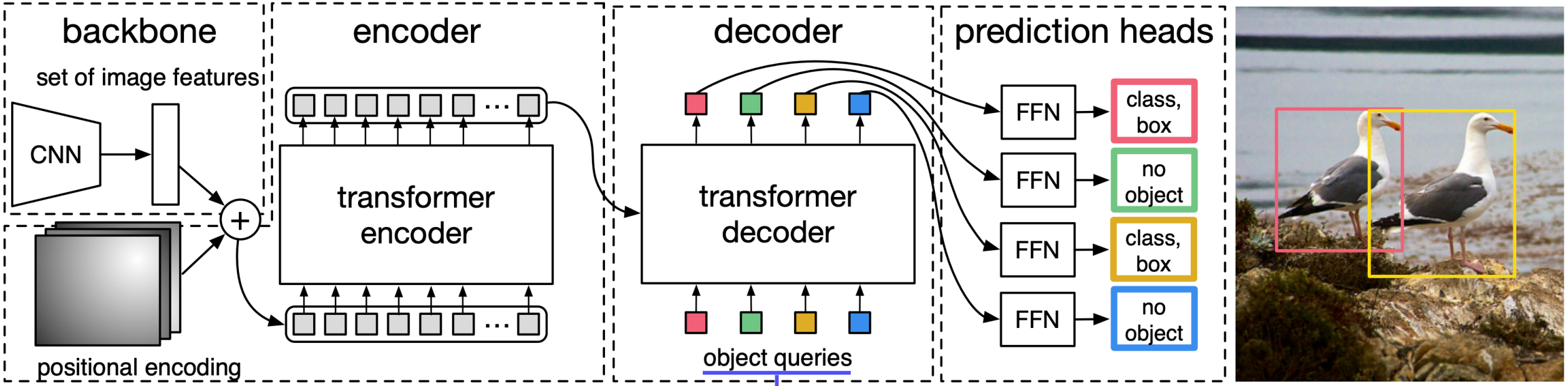
Transformer has shown a great success in **NLP**

Why not **extending** Transformer to computer vision tasks!

- ViT (Vision Transformer) by Google
- DeiT (Data-efficient image Transformer) by Facebook
- DETR (DEtection TRansformer) by Facebook

DETR

[Carion et al., ECCV 2020]



Object query: Learned positional encodings for querying

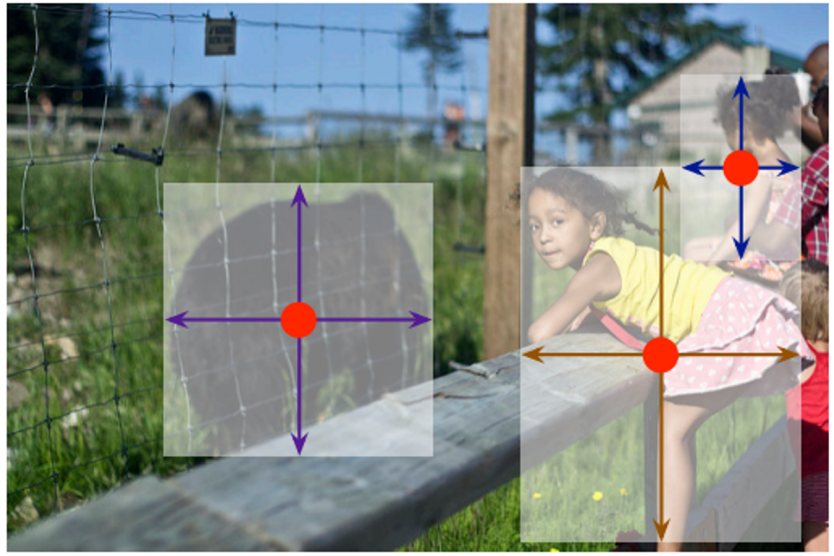
Hyper-parameter:
(N object queries = max. N objects exist in a single image)

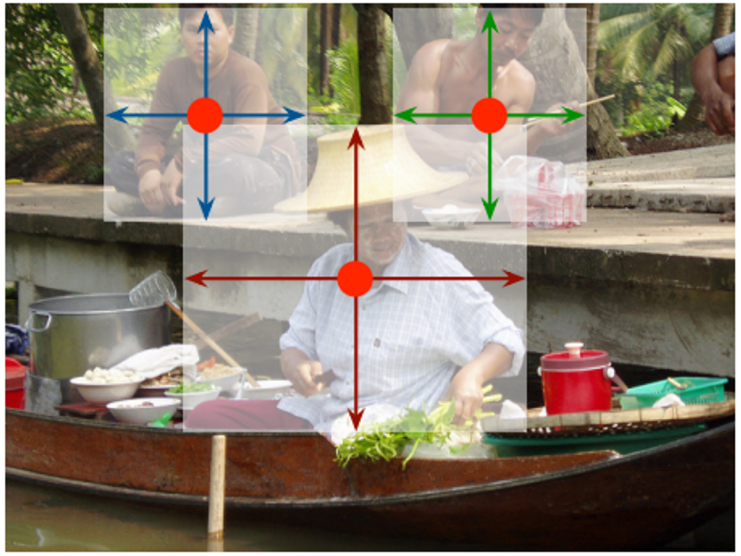
Further reading

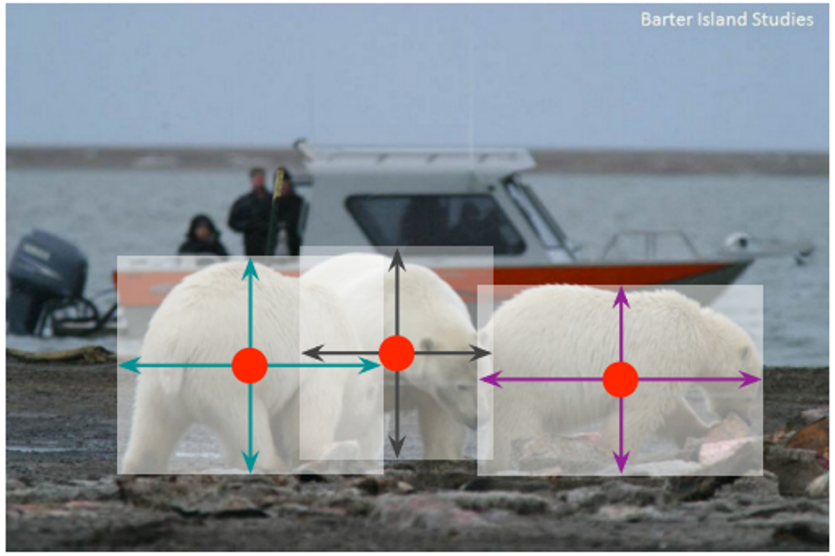
Detecting objects as points

[Zhou et al., arXiv 2019]

- Bounding box can be represented by other ways (left-top, right-bottom, centroid & size)
- Idea: Let's detect objects using corresponding points!
- CornerNet/CenterNet will be covered in Lecture 7







Reference

1. Object detection

- Kirillov et al., Panoptic Segmentation, CVPR 2019

2. Two-stage detector (R-CNN Family)

- Dalal et al., Histograms of Oriented Gradients for Human Detection, CVPR 2005
- Uijlings et al., Selective Search for Object Recognition, IJCV 2013
- Girshick et al., Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation, CVPR 2014
- Girshick et al., Fast R-CNN, ICCV 2015
- Ferguson et al., Detection and Segmentation of Manufacturing Defects with Convolutional Neural Networks and Transfer Learning, Smart and Sustainable Manufacturing Systems 2018
- Ren et al., Faster R-CNN: Towards Real-Time Object detection with Region Proposal Networks, NeurIPS 2015

3. Single-stage detector

- Ndonhong et al., Wellbore Schematics to Structured Data Using Artificial Intelligence Tools, Offshore Technology Conference 2019
- Redmon et al., You Only Look Once: Unified, Real-Time Object detection, CVPR 2016
- Liu et al., SSD: Single Shot MultiBox Detector, ECCV 2016

×

Reference

4. Single-stage detector vs. two-stage detector

- Lin et al., Focal loss for Dense Object detection, ICCV 2017

5. Detection with Transformer

- Vaswani et al., Attention Is All You Need, NeurIPS 2017
- Carion et al., End-to-end Object Detection with Transformers, ECCV 2020
- Zhou et al., Objects as Points, arXiv 2019

End of Document
Thank You.