

Object Detection and Segmentation

Lior Yariv

February 28th, 2022



Previously

Classification



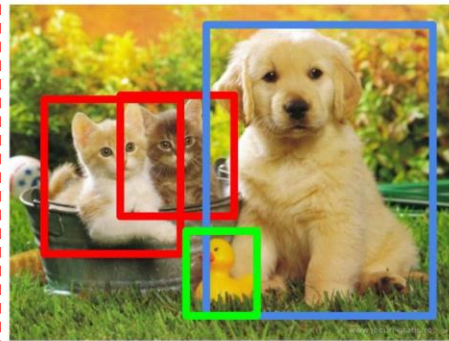
CAT

**Classification
+ Localization**



CAT

Object Detection



CAT, DOG, DUCK

**Semantic
Segmentation**



CAT, DOG, DUCK

**Instance
Segmentation**

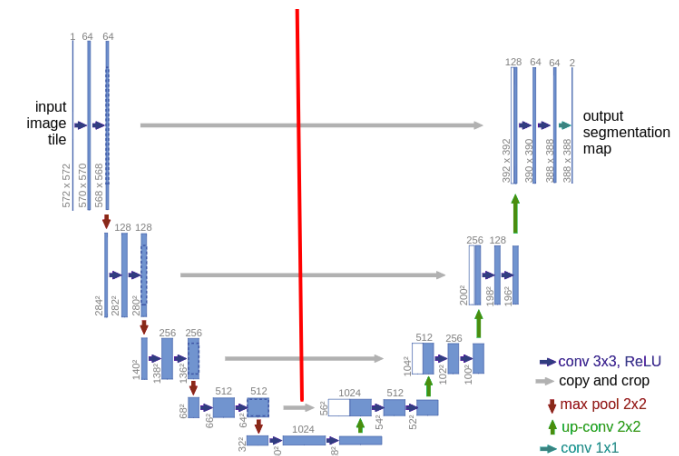
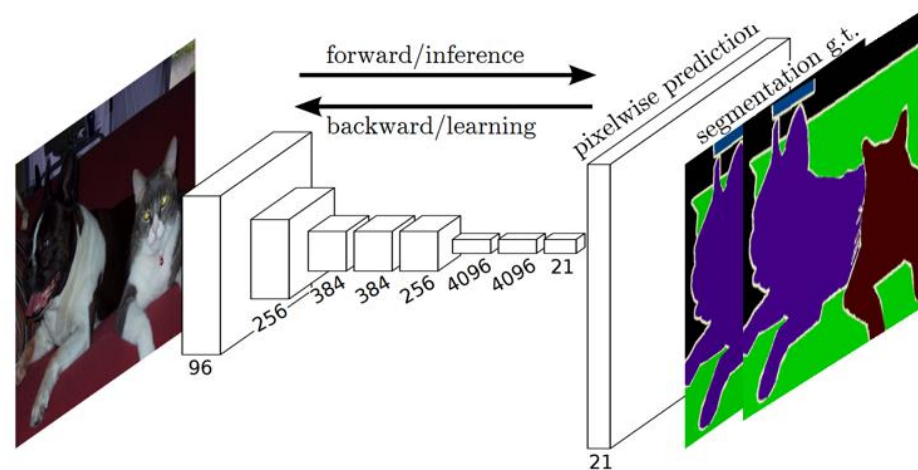
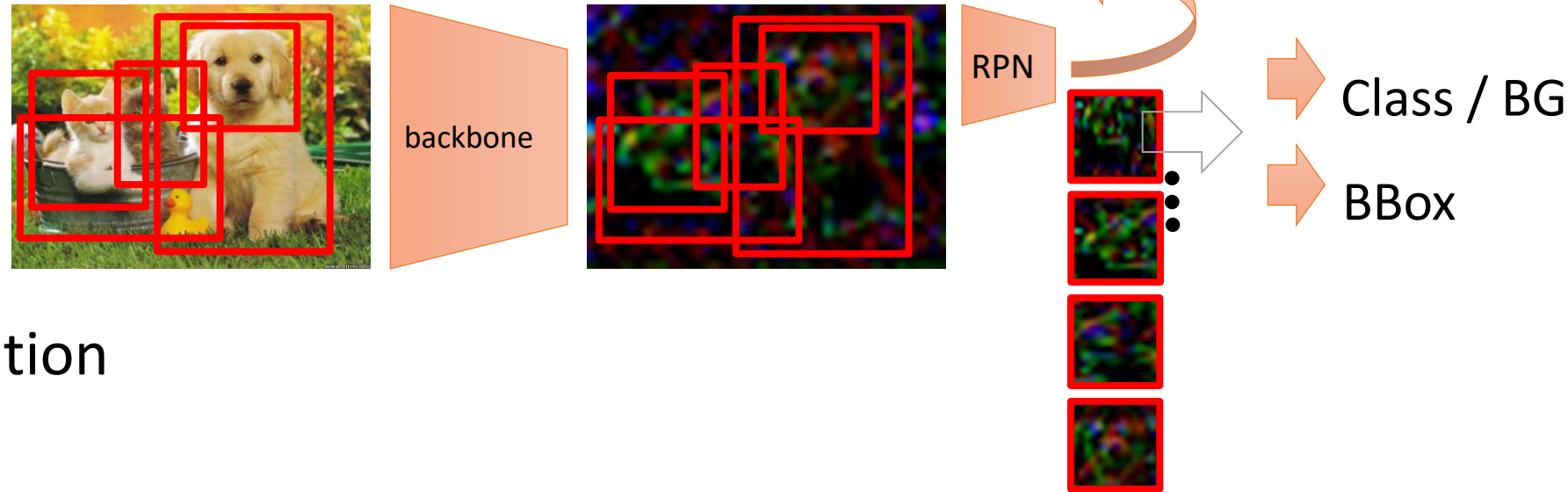


CAT, DOG, DUCK

Previously

- Object detection
 - Faster RCNN

- Semantic segmentation
 - FCN / DeepLab
 - UNet



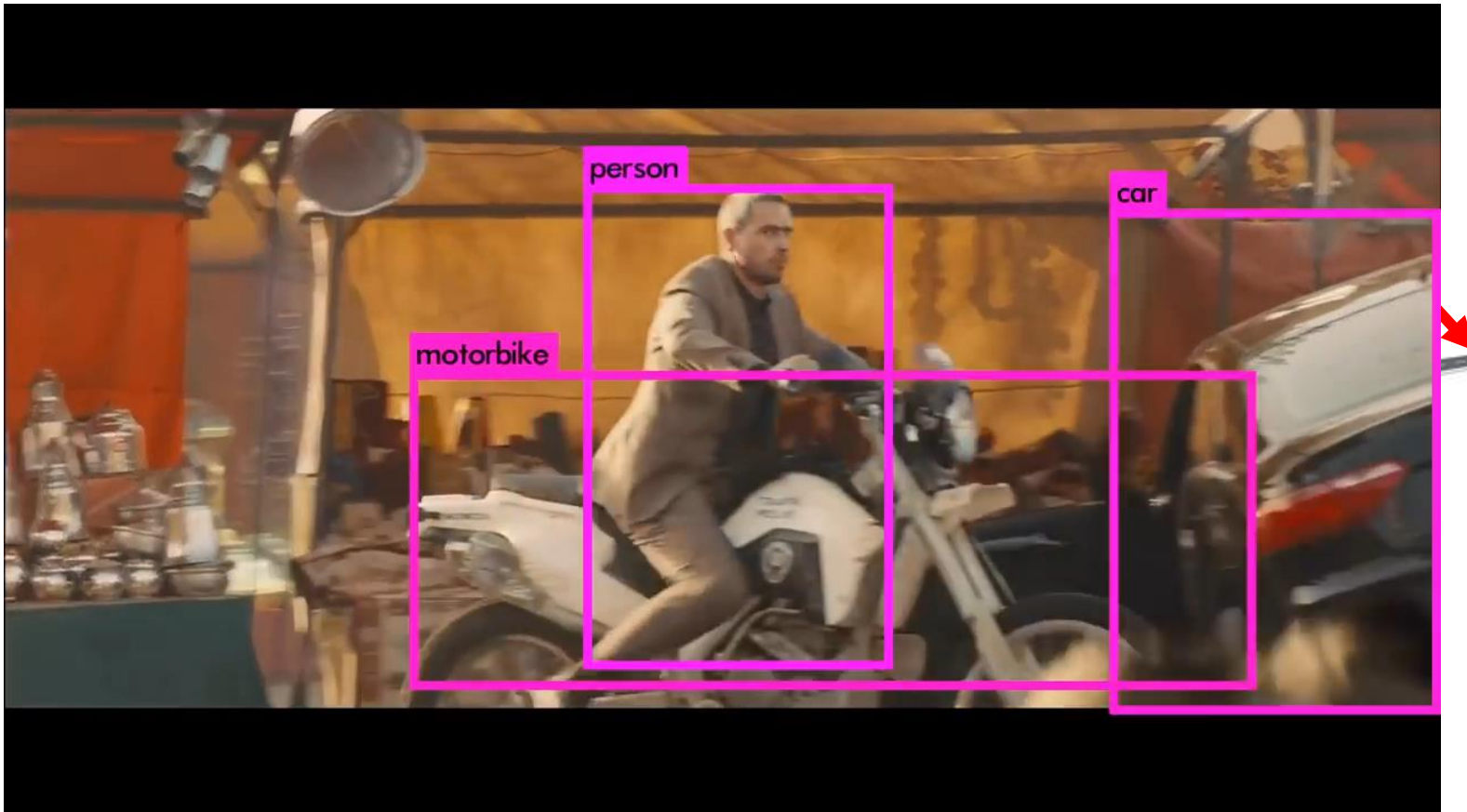
Previously

- Object detection
 - Faster RCNN
- Semantic segmentation
 - FCN / DeepLab
 - UNet

Today

- (Multiple) Object detection
 - YOLO (You Only Look Once)
 - DETR (DEtection TRansformer)
- Semantic Segmentation
 - Segmenter

Object Detection

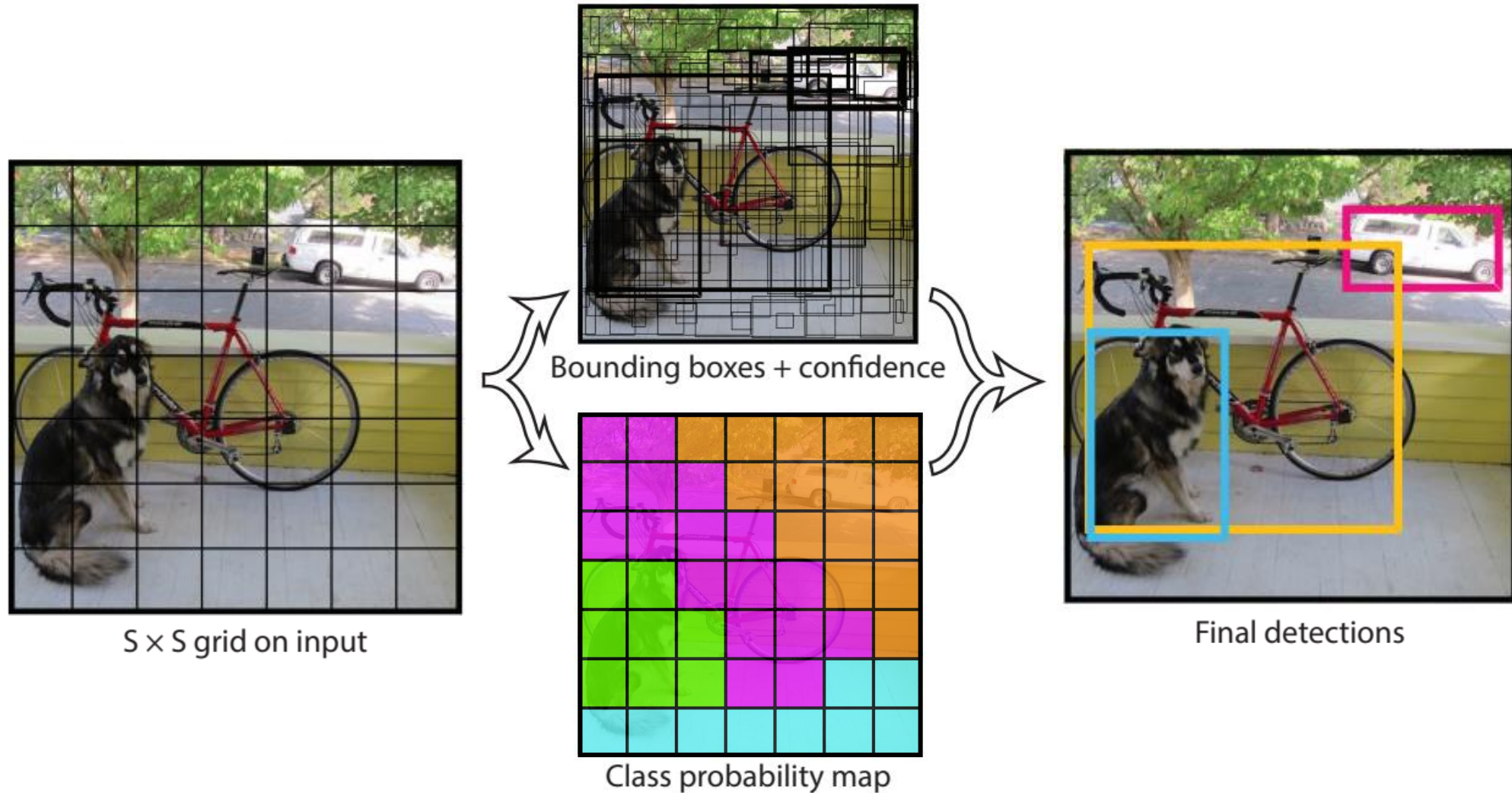


Single Shot:
SSD, **YOLO** ...

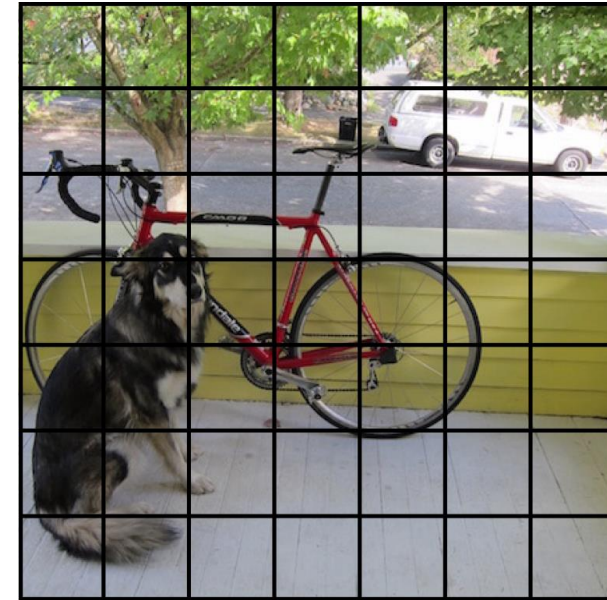
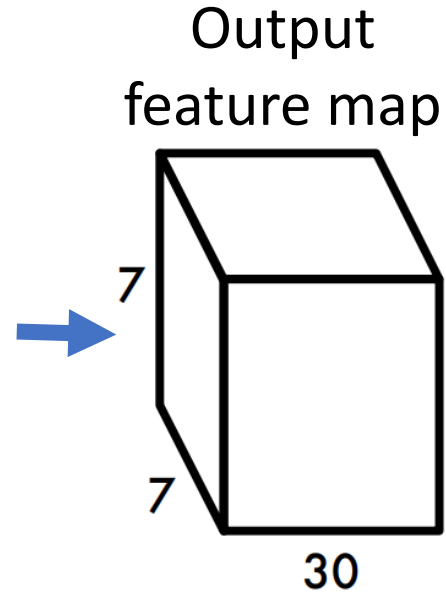
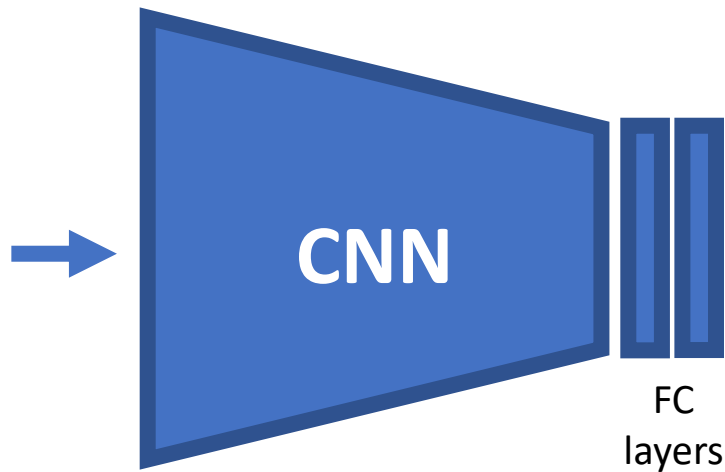
Fast

High false rate

YOLO – Overview



YOLO



YOLO

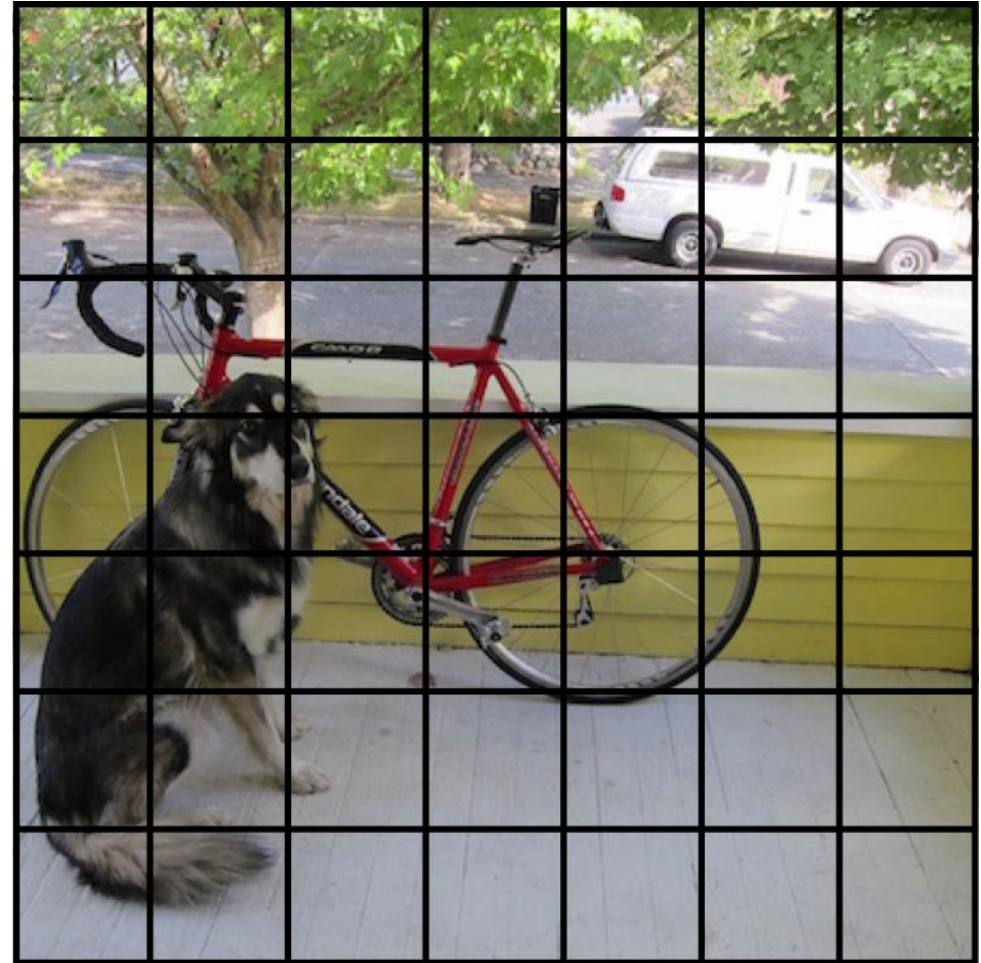


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](#), 2015 (CVPR 2016)

YOLO

Each cell predicts

- $B = 2$ bounding boxes
 (x, y, w, h) + confidence score

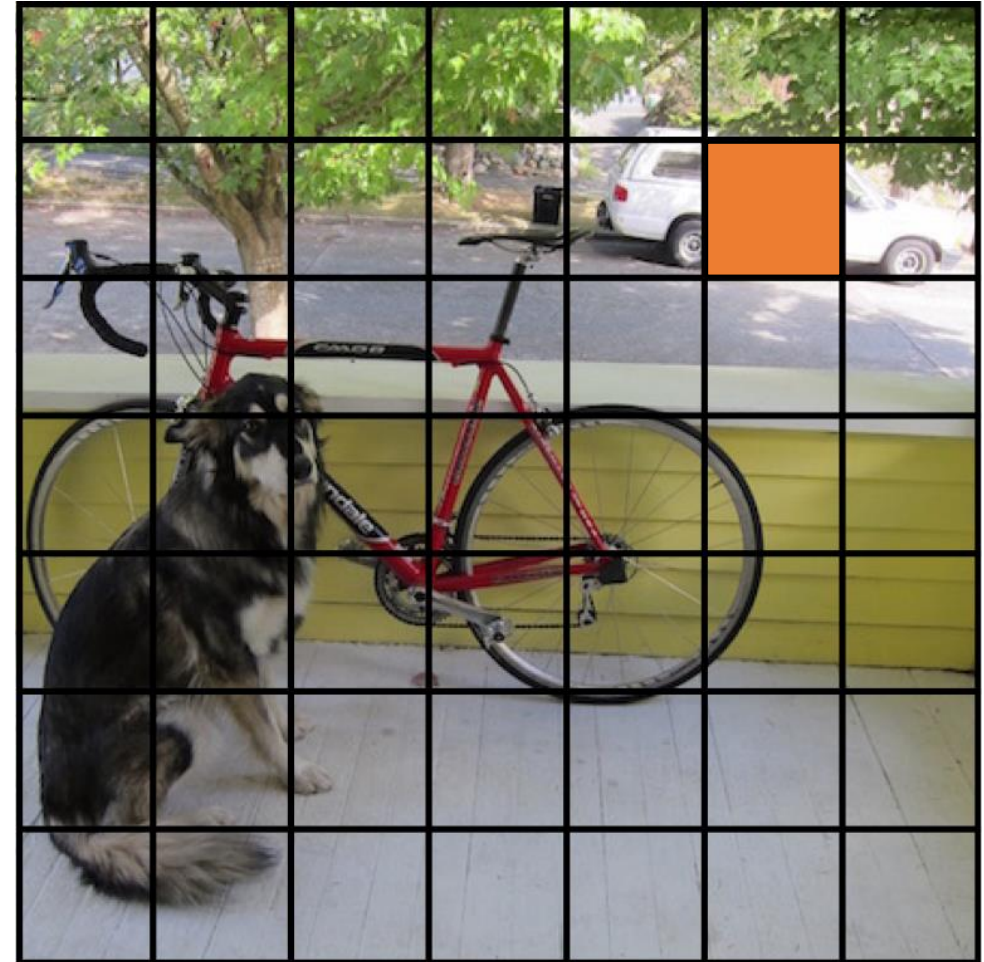


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO

Each cell predicts

- $B = 2$ bounding boxes
 (x, y, w, h) + confidence score

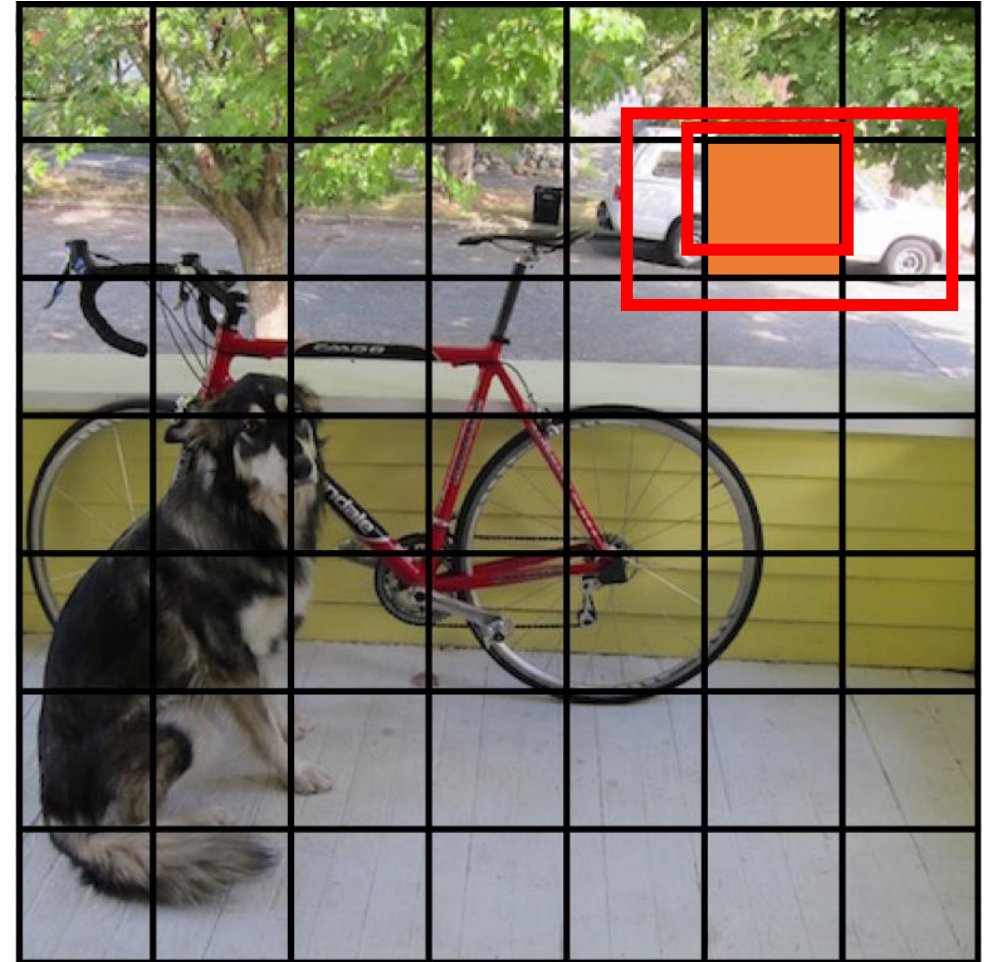


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Inference

Each cell predicts

- $B = 2$ bounding boxes
 (x, y, w, h) + confidence score
- $C = 20$ class probabilities

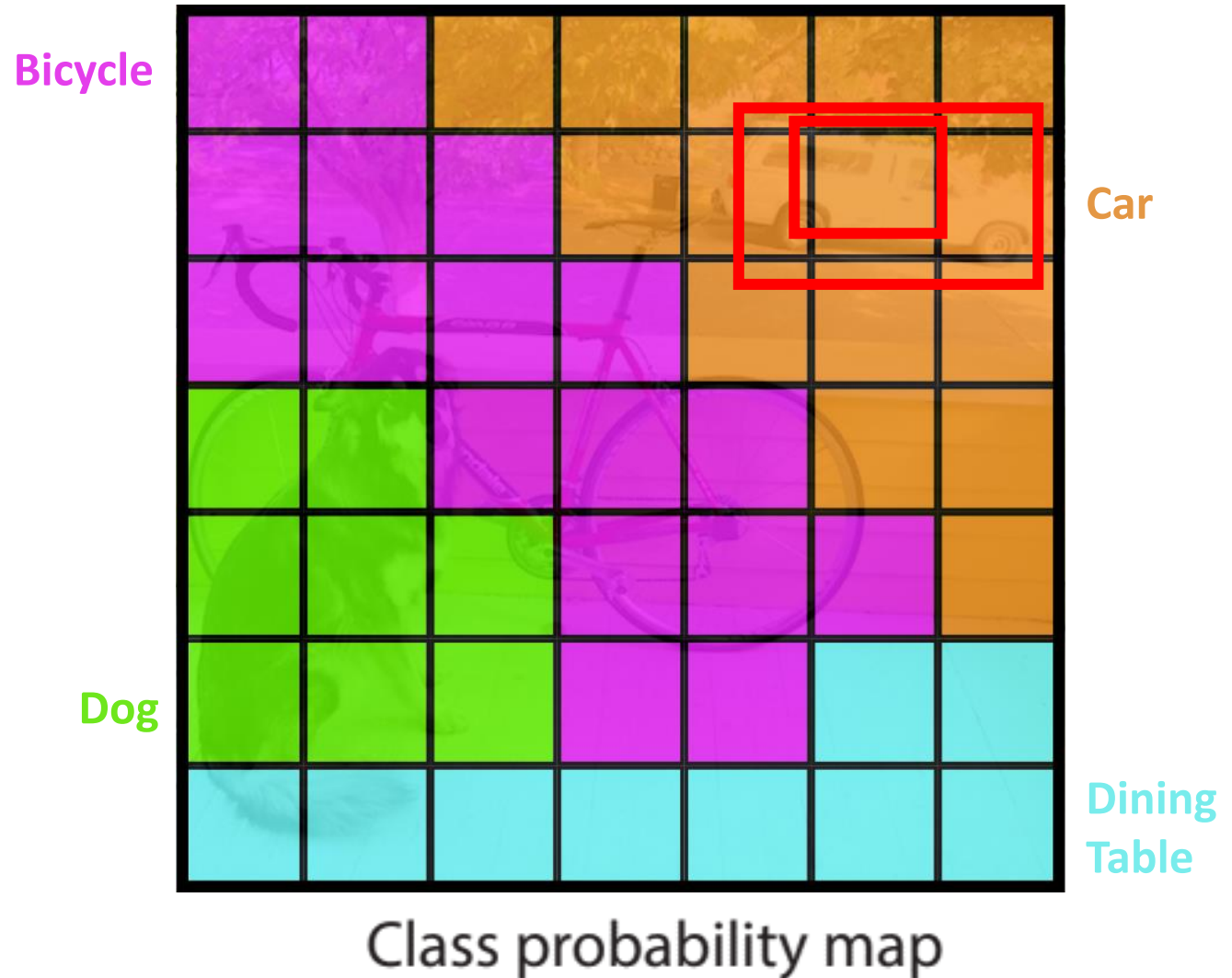
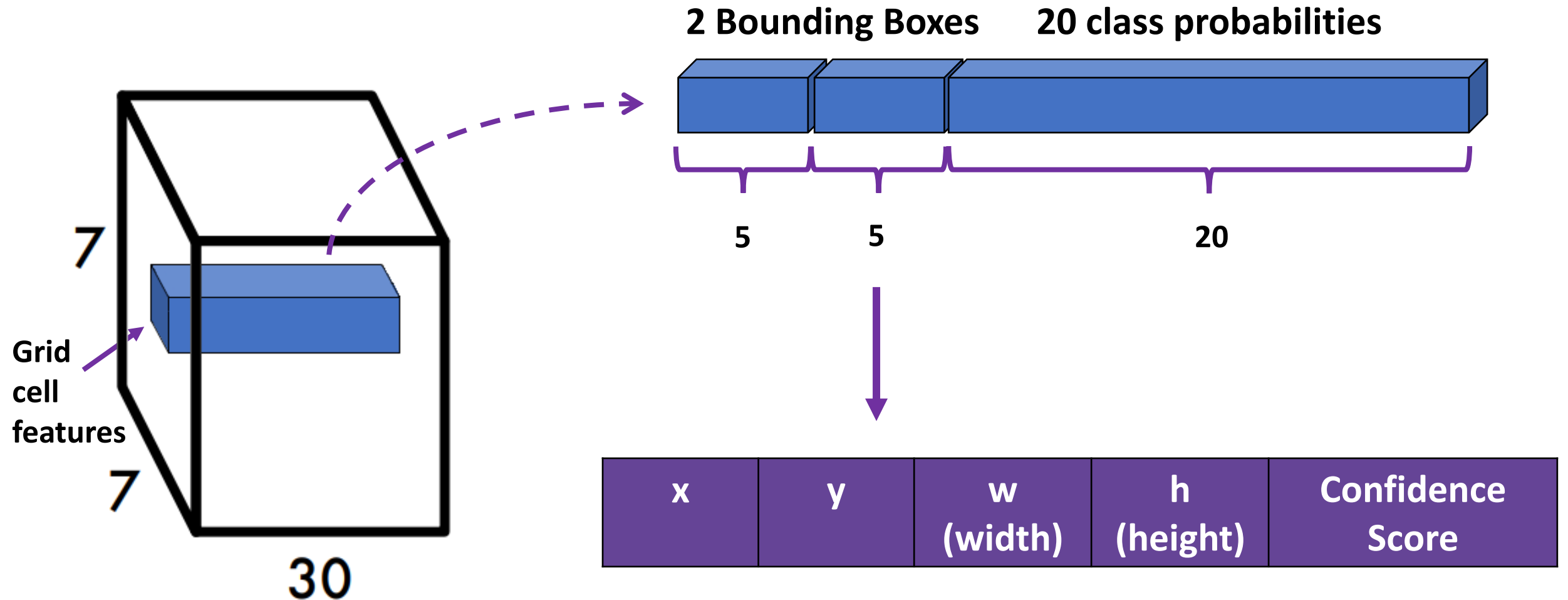


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Output feature map

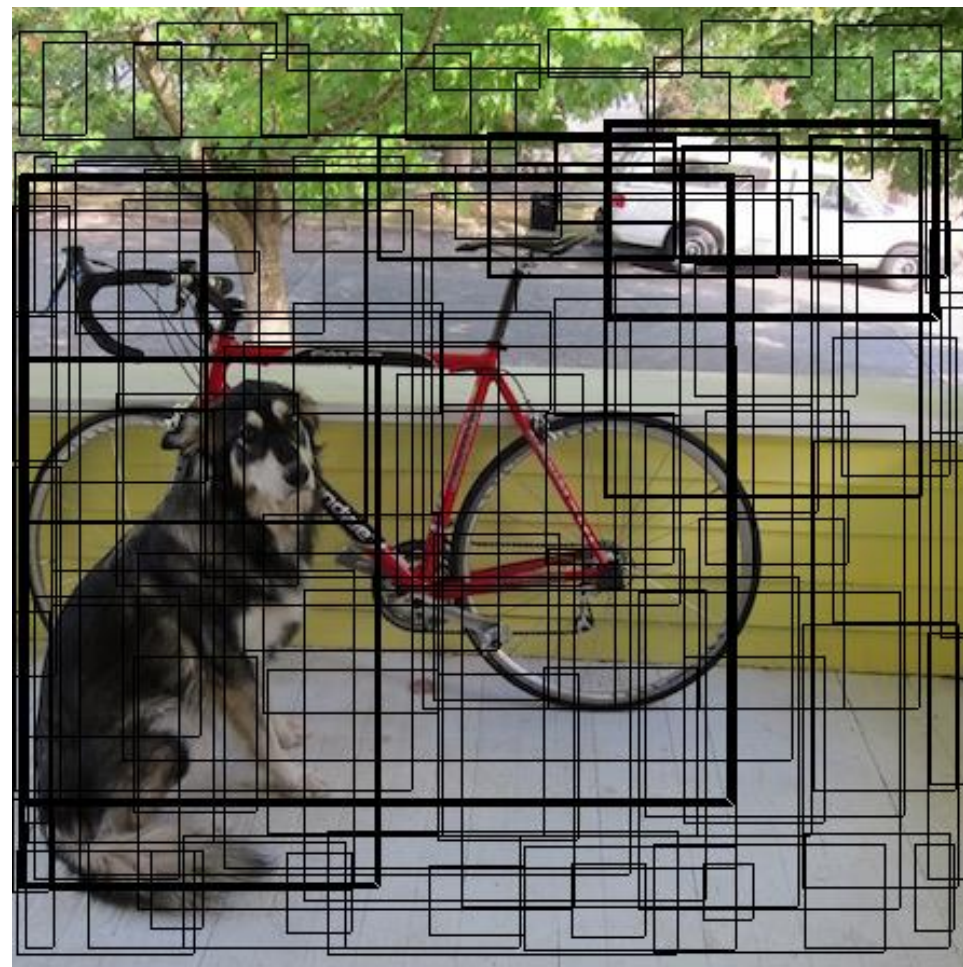


YOLO

Each cell predicts

- $B = 2$ bounding boxes
 $(x, y, w, h) + \text{confidence score}$
- $C = 20$ class probabilities

$S \times S \times B$ Bounding-Boxes ($S=7, B=2 \rightarrow 96$ Bboxes)



$S \times S$ grid on input

Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO

Each cell predicts

- $B = 2$ bounding boxes
 (x, y, w, h) + confidence score
- $C = 20$ class probabilities

$S \times S \times B$ Bounding-Boxes ($S=7, B=2 \rightarrow 96$ Bboxes)



Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO

Each cell predicts

- $B = 2$ bounding boxes
 (x, y, w, h) + confidence score
- $C = 20$ class probabilities
- Apply Non-Maximum Suppression

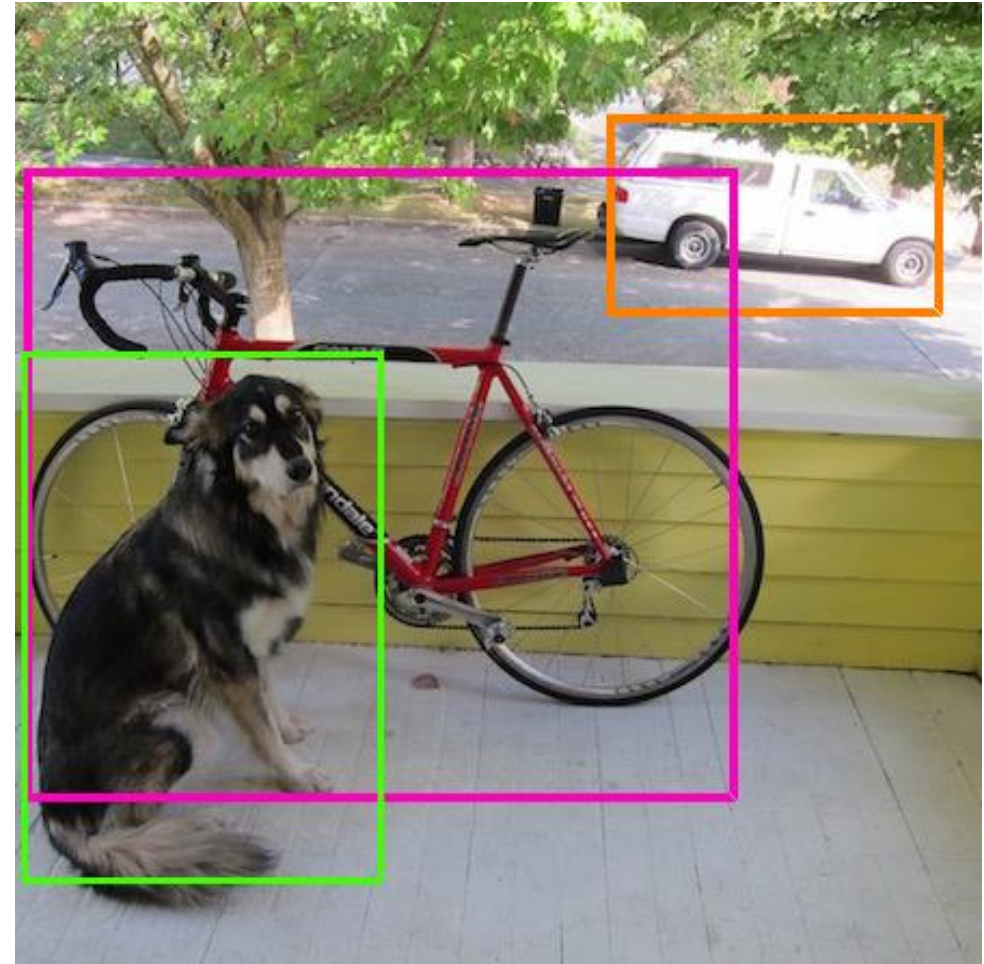
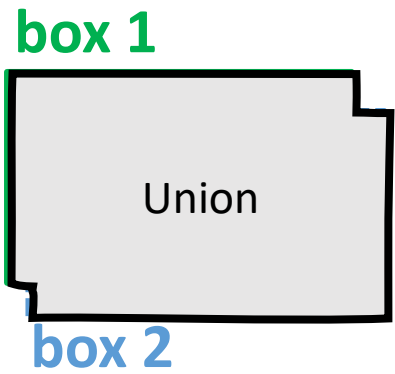


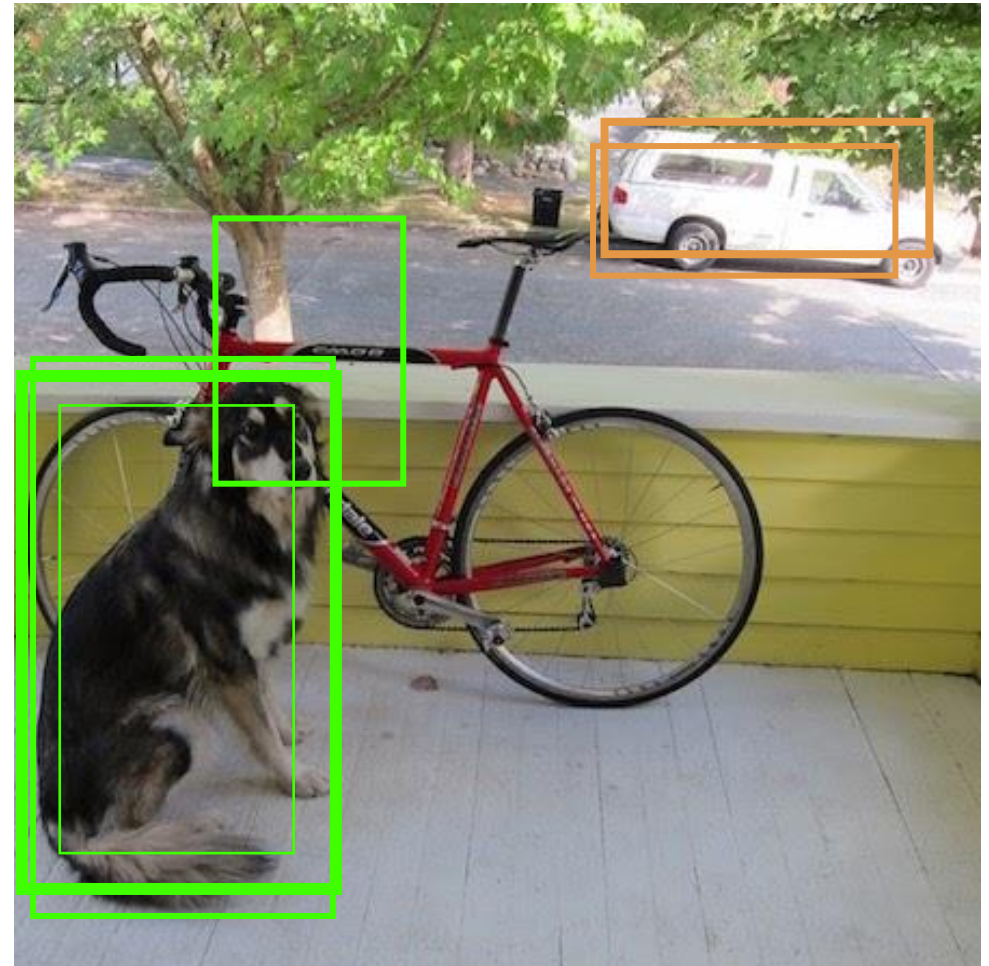
Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO - NMS



IoU = _____

Dog

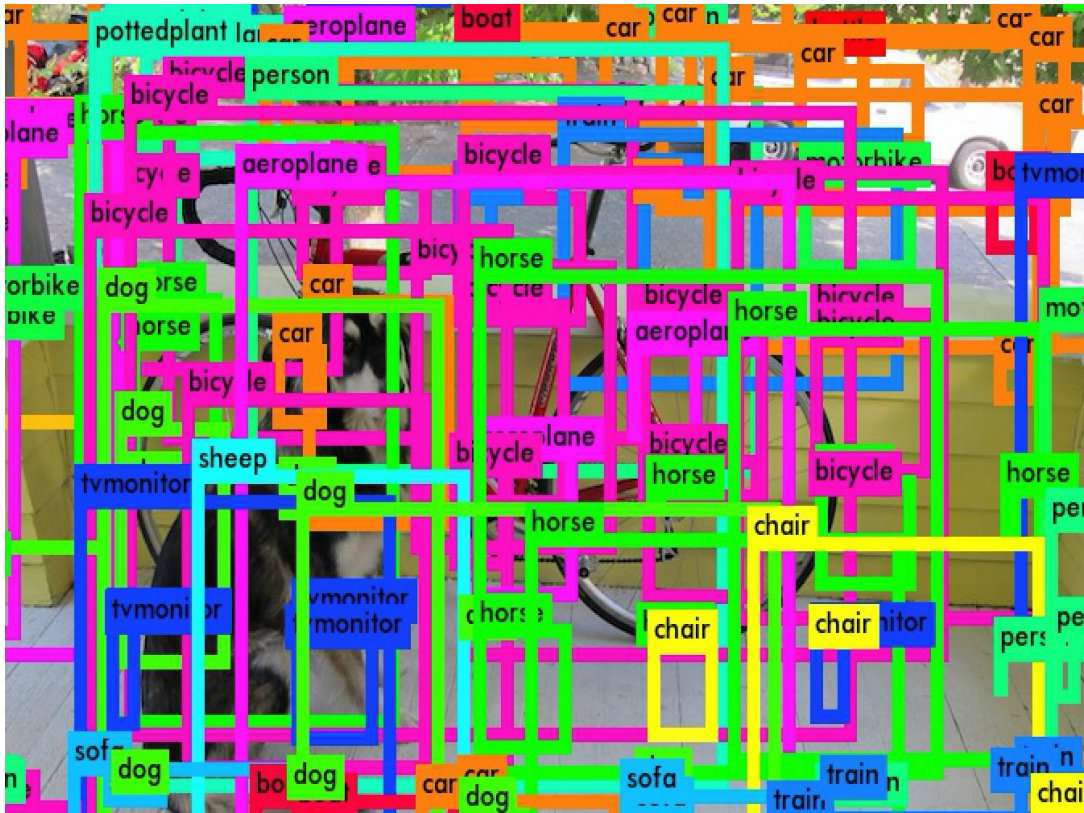


Car

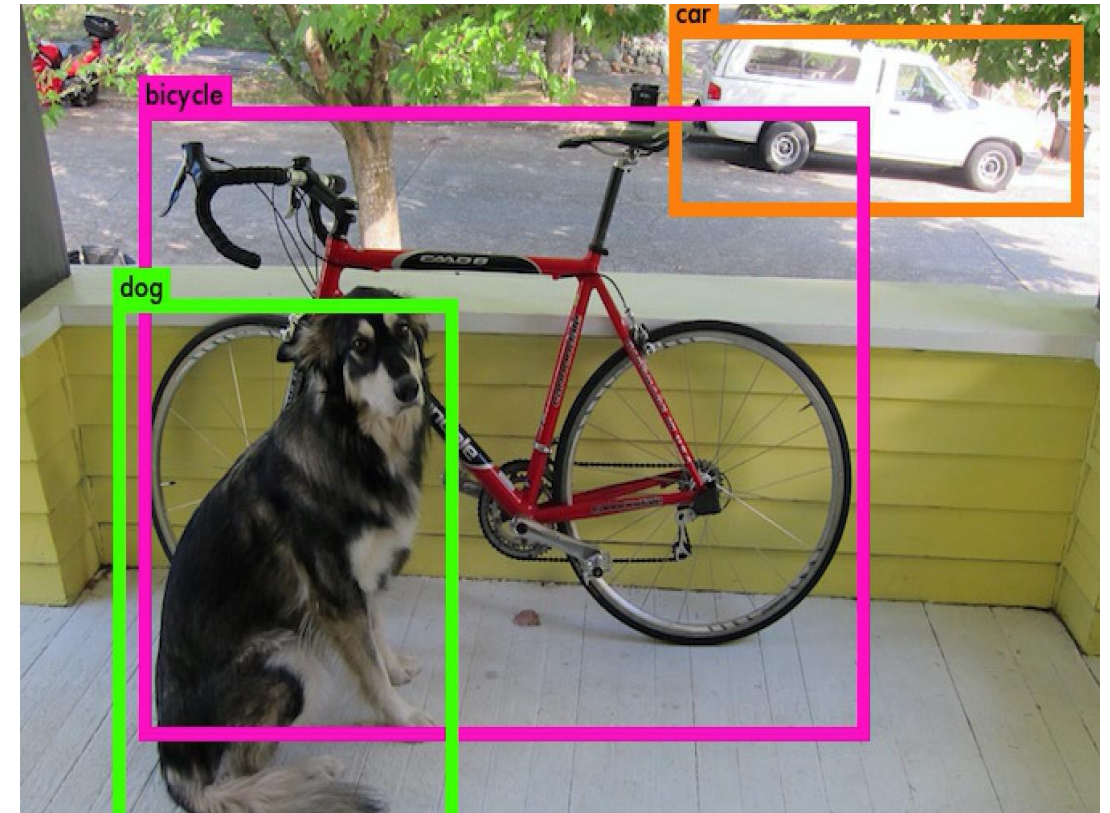
Slide idea: Shai Bagon

YOLO – Role of Threshold

Low Threshold



High Threshold



YOLO – Training (end-to-end)

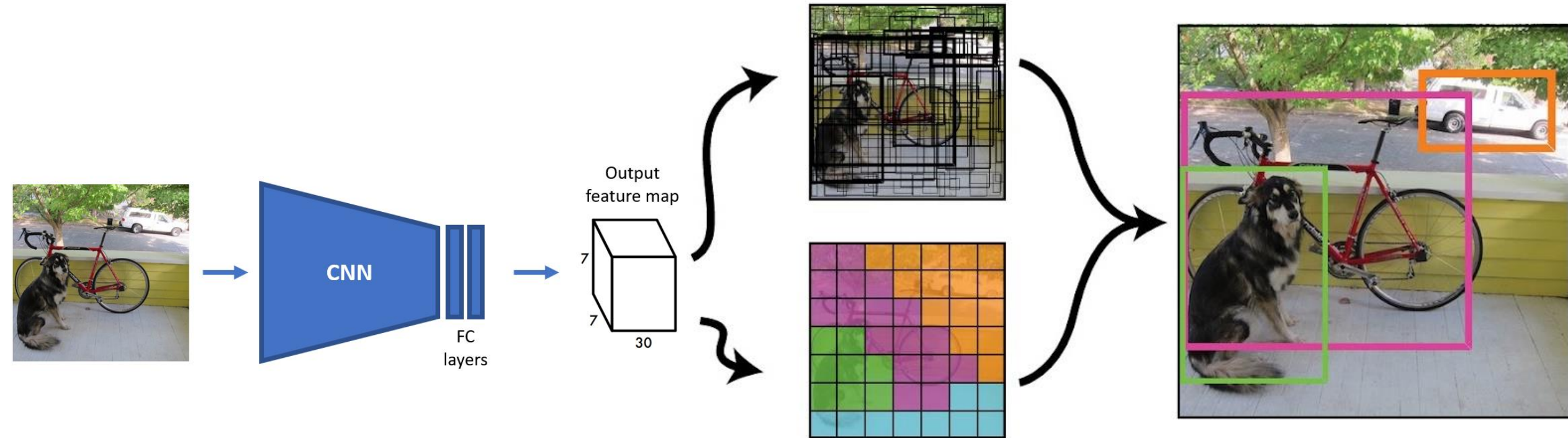
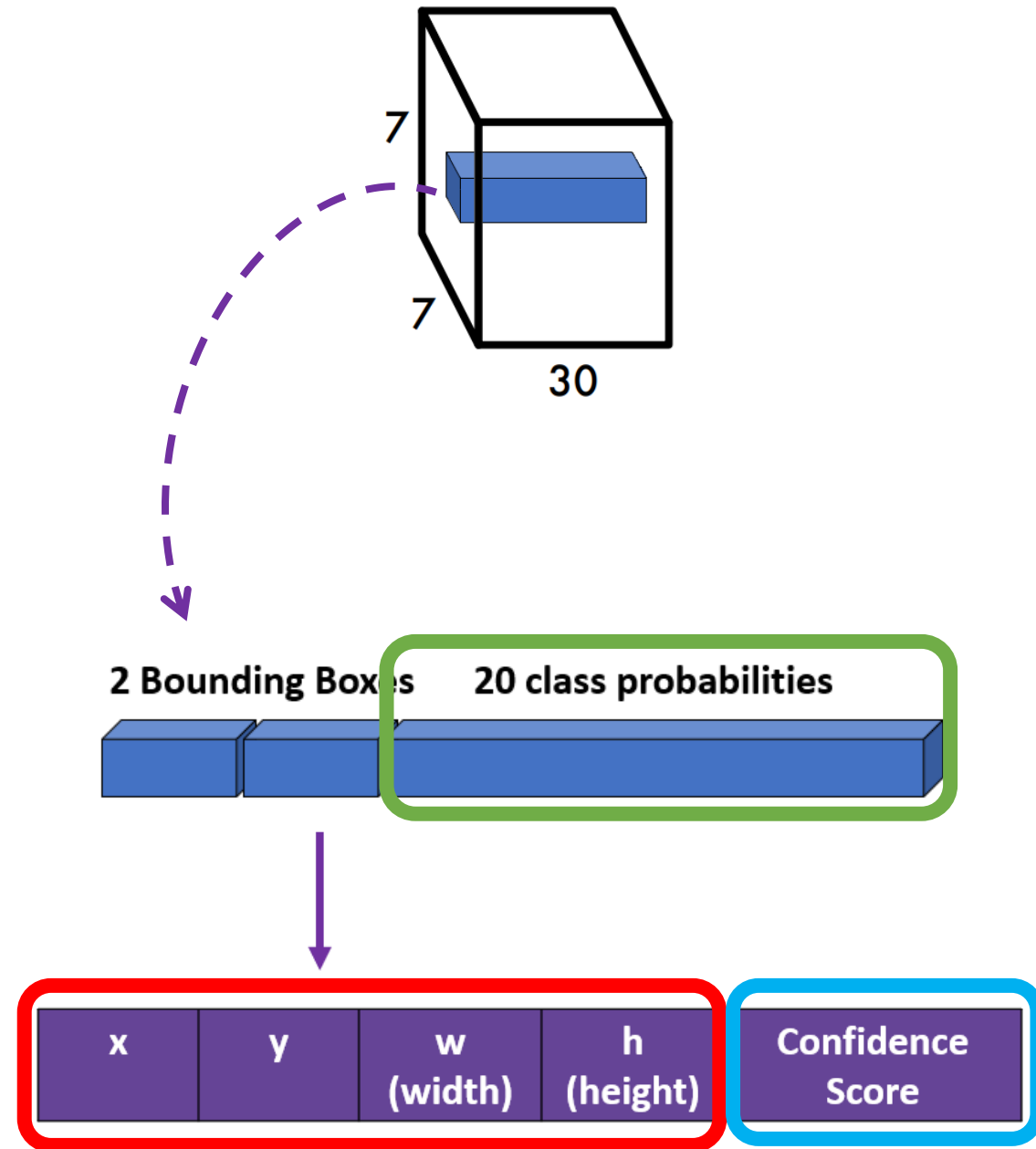


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Loss function

$$\mathcal{L} = \mathcal{L}_{\text{Localization Loss}} + \mathcal{L}_{\text{Confidence Loss}} + \mathcal{L}_{\text{Classification Loss}}$$



YOLO – Training

$$\mathcal{L} = \mathcal{L}_{Localization Loss} + \mathcal{L}_{Confidence Loss} + \mathcal{L}_{Classification Loss}$$

Ground truth bounding box

Center of object

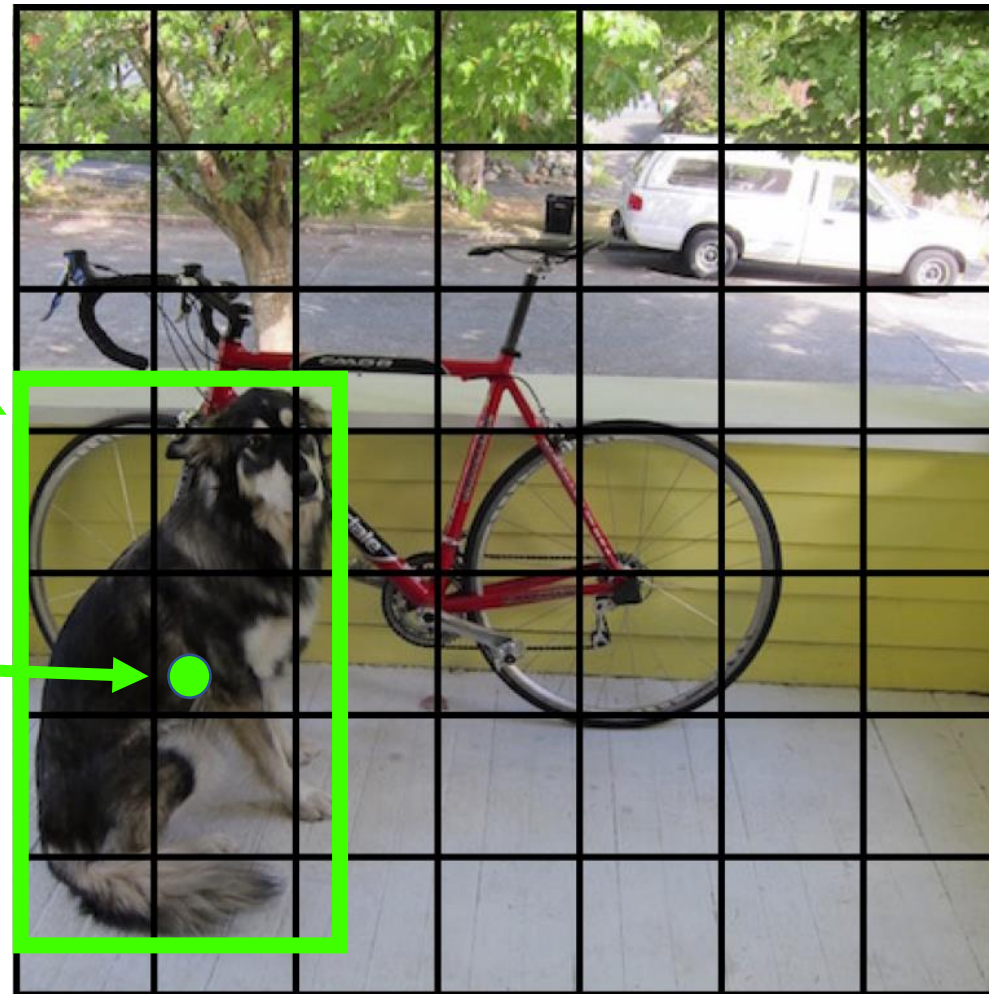


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Training

$$\mathcal{L} = \mathcal{L}_{Localization Loss} + \mathcal{L}_{Confidence Loss} + \mathcal{L}_{Classification Loss}$$

Ground truth bounding box

Assign to specific cell

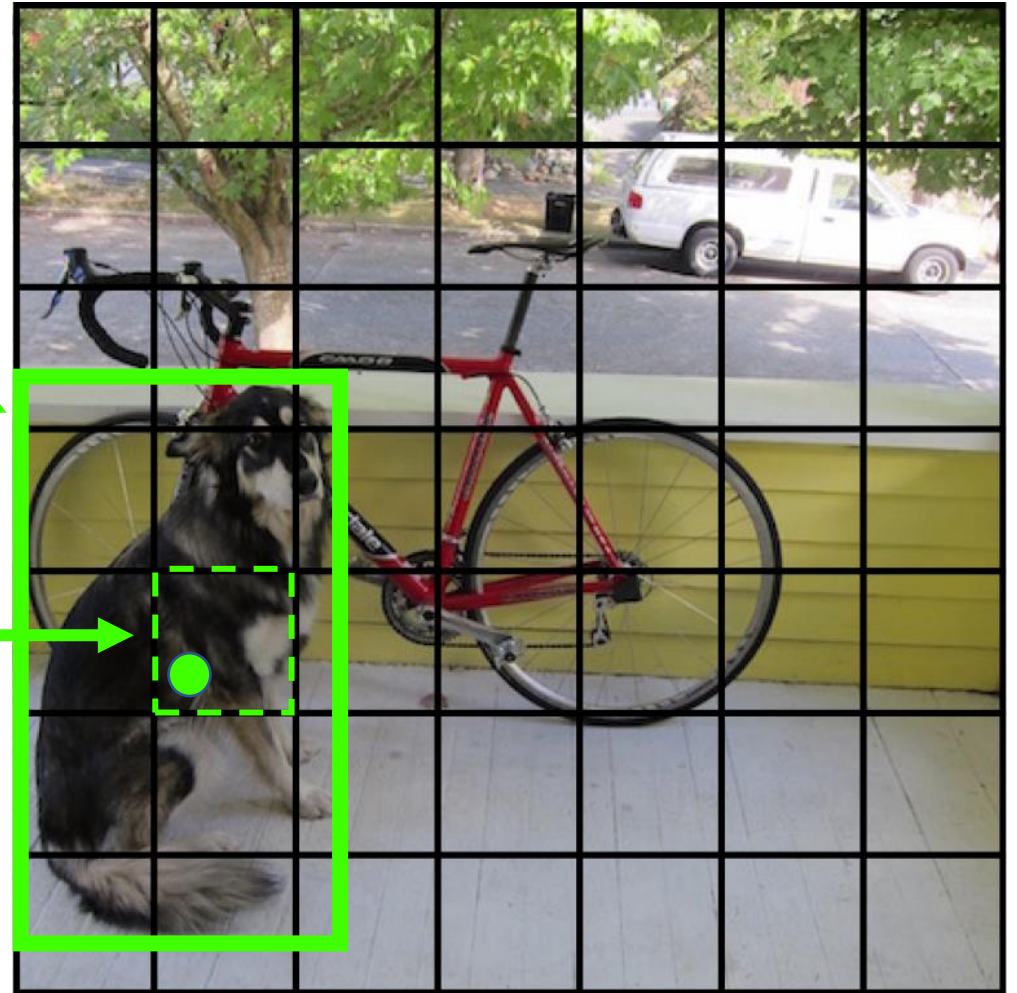


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Training

$$\mathcal{L} = \mathcal{L}_{Localization Loss} + \mathcal{L}_{Confidence Loss} + \mathcal{L}_{Classification Loss}$$

Ground truth bounding box

Supervisory signal:

Dog: 1
Cat: 0
Bike: 0
...

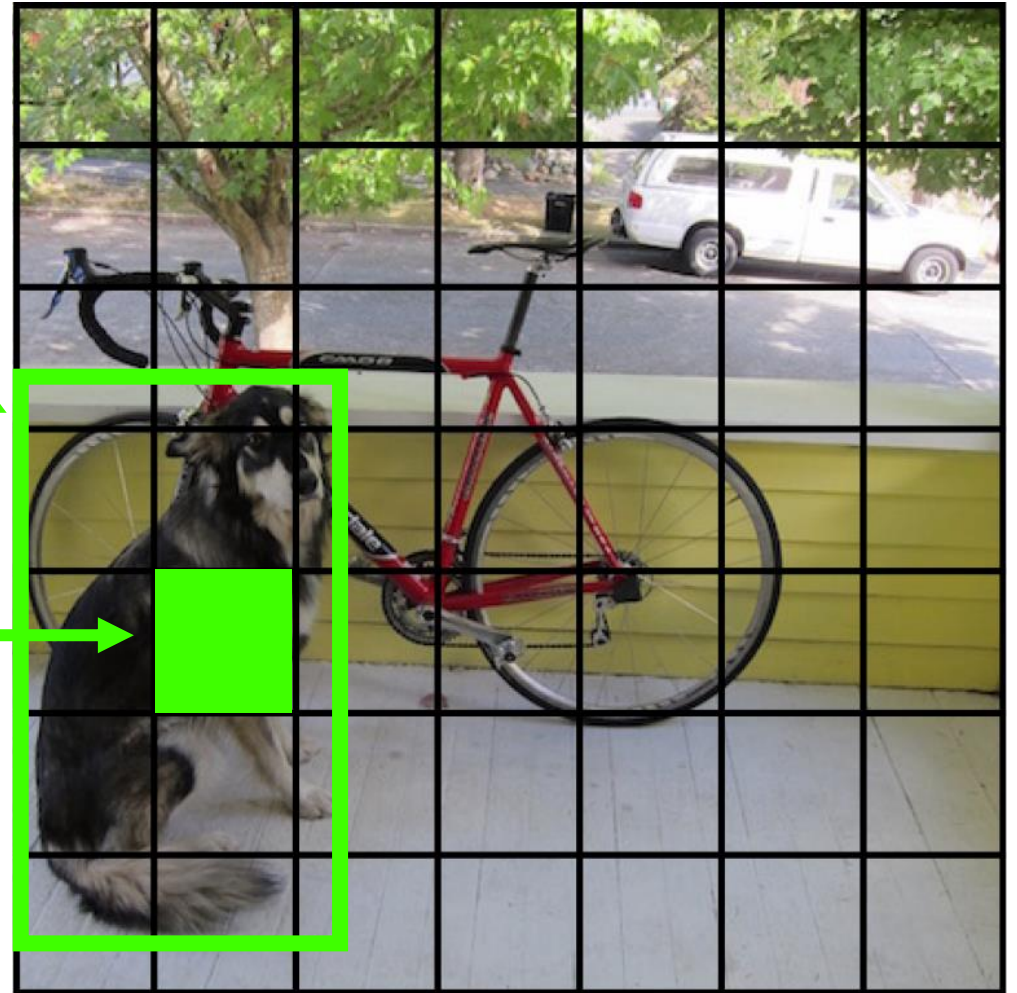


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Training

$$\mathcal{L} = \mathcal{L}_{Localization Loss} + \mathcal{L}_{Confidence Loss} + \mathcal{L}_{Classification Loss}$$

Ground truth bounding box

Look at cell's predicted boxes

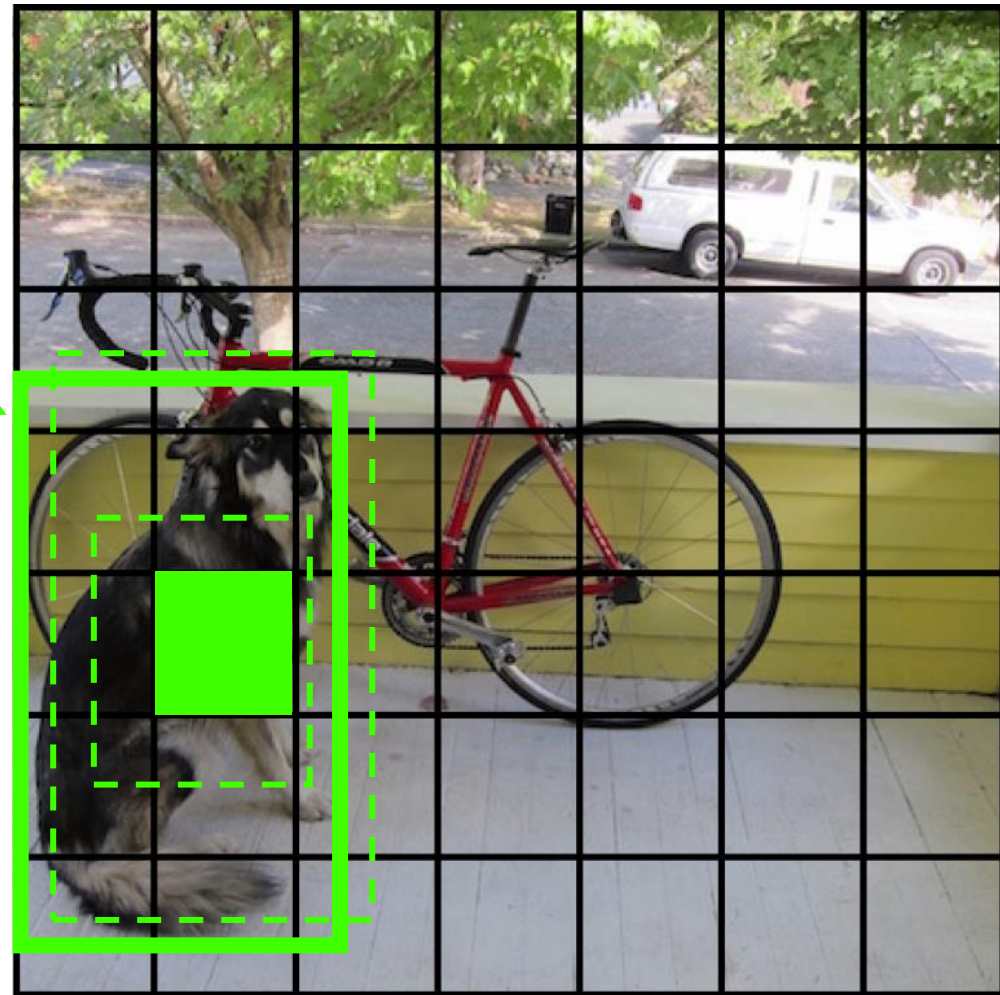


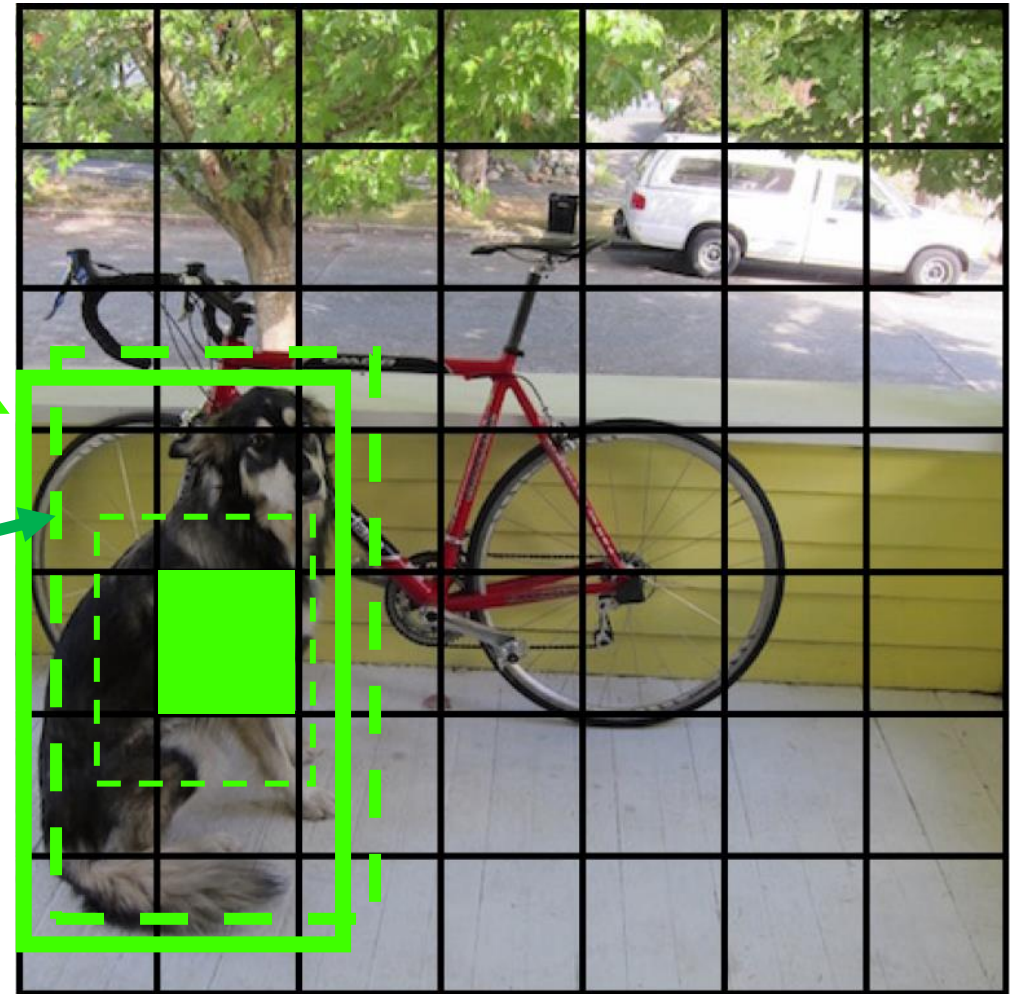
Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Training

$$\mathcal{L} = \mathcal{L}_{\text{Localization Loss}} + \mathcal{L}_{\text{Confidence Loss}} + \mathcal{L}_{\text{Classification Loss}}$$

Ground truth bounding box

Increase confidence score
adjust $\mathcal{L}_{\text{Localization}}$



YOLO – Training

$$\mathcal{L} = \int \cancel{\mathcal{L}_{\text{Localization Loss}}} + \mathcal{L}_{\text{Confidence Loss}} + \mathcal{L}_{\text{Classification Loss}}$$

Ground truth bounding box

Increase confidence score
adjust $L_{\text{Localization}}$

Decrease confidence score
don't adjust $L_{\text{Localization}}$

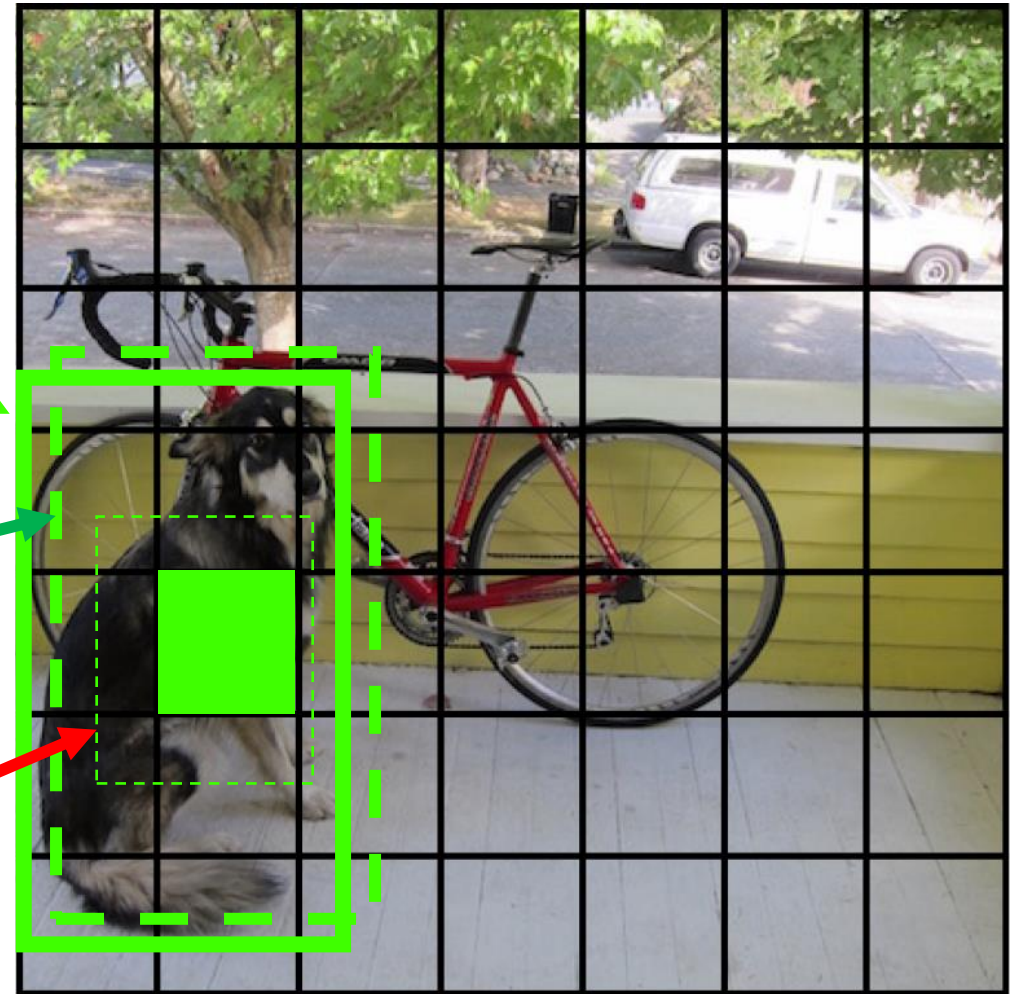


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Training

$$\mathcal{L} = \mathcal{L}_{\text{Localization Loss}} + \mathcal{L}_{\text{Confidence Loss}} + \mathcal{L}_{\text{Classification Loss}}$$

A cell with no ground truth detection

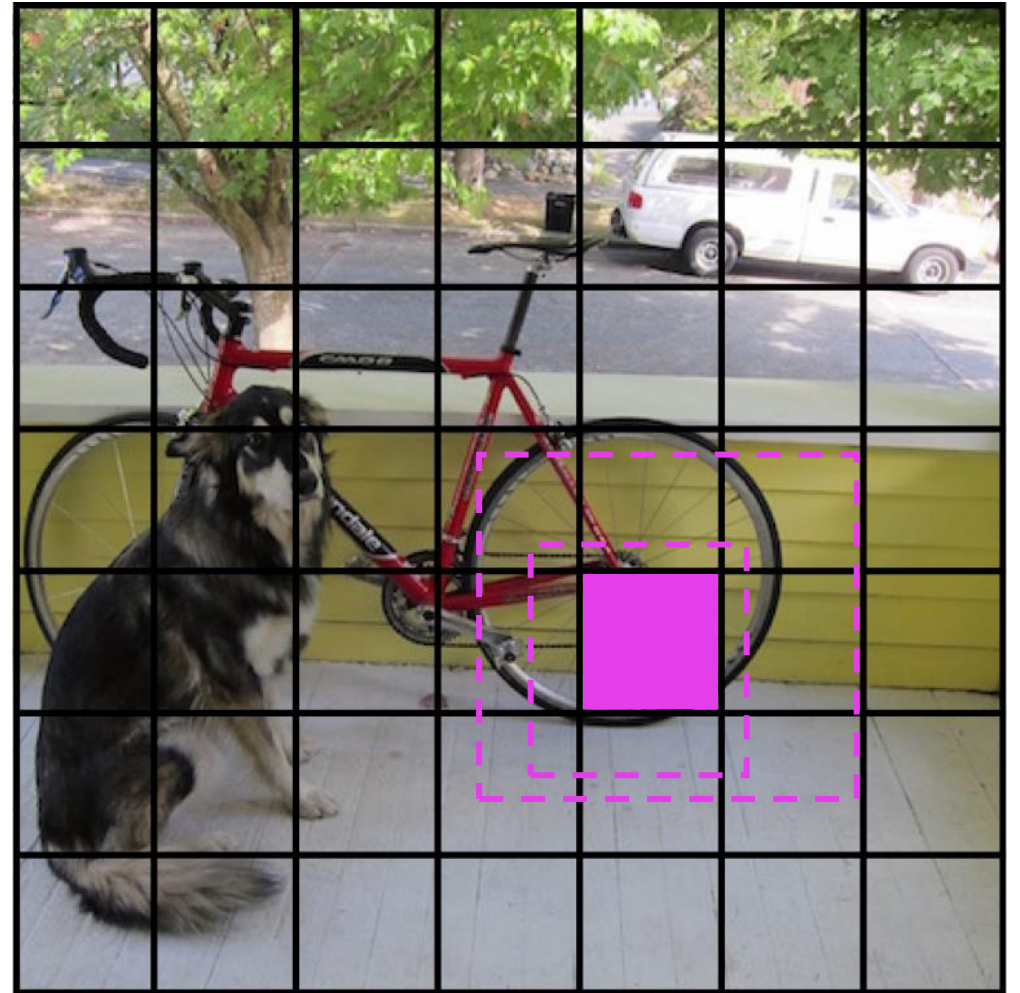


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Training

$$\mathcal{L} = \mathcal{L}_{\text{Localization Loss}} + \mathcal{L}_{\text{Confidence Loss}} + \mathcal{L}_{\text{Classification Loss}}$$

A cell with no ground truth detection

Decrease confidence score



Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Training

$$\mathcal{L} = \int \cancel{\mathcal{L}_{\text{Localization Loss}}} + \mathcal{L}_{\text{Confidence Loss}} + \cancel{\mathcal{L}_{\text{Classification Loss}}}$$

A cell with no ground truth detection

Decrease confidence score

Don't adjust

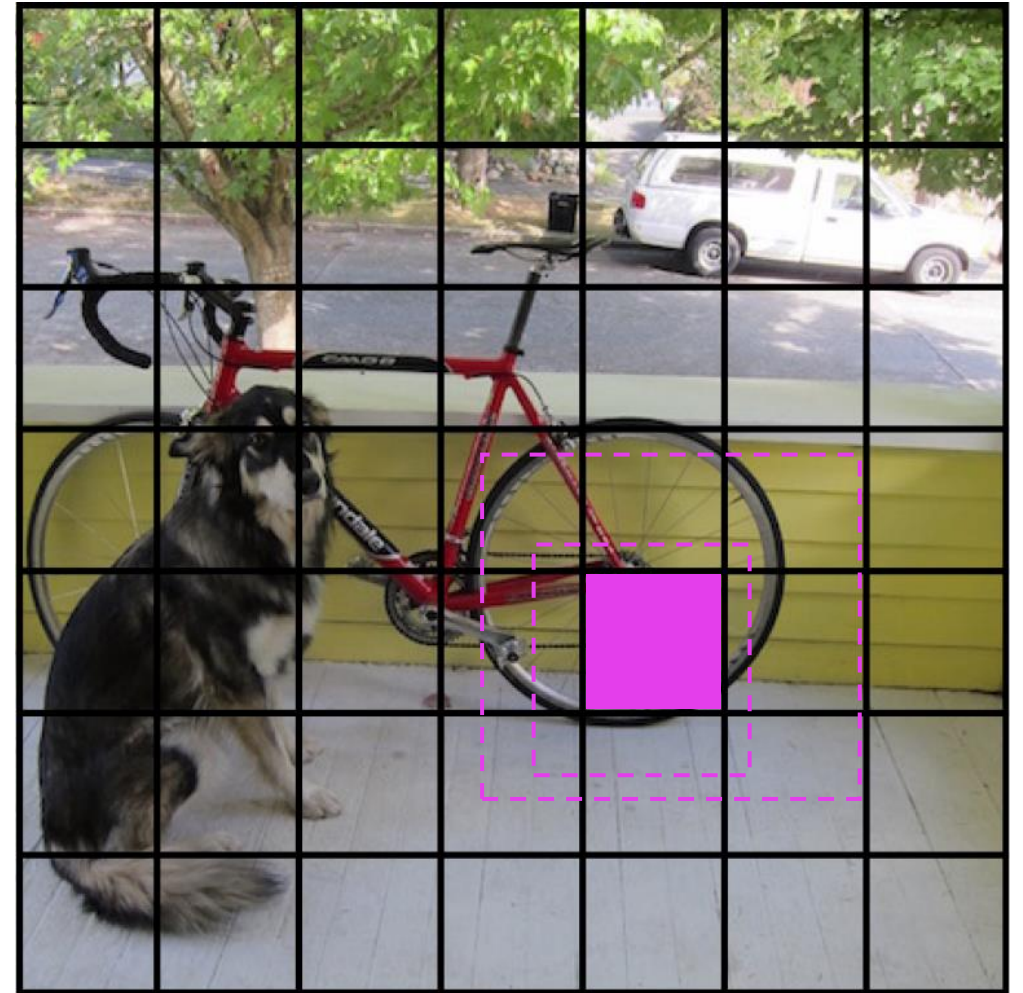
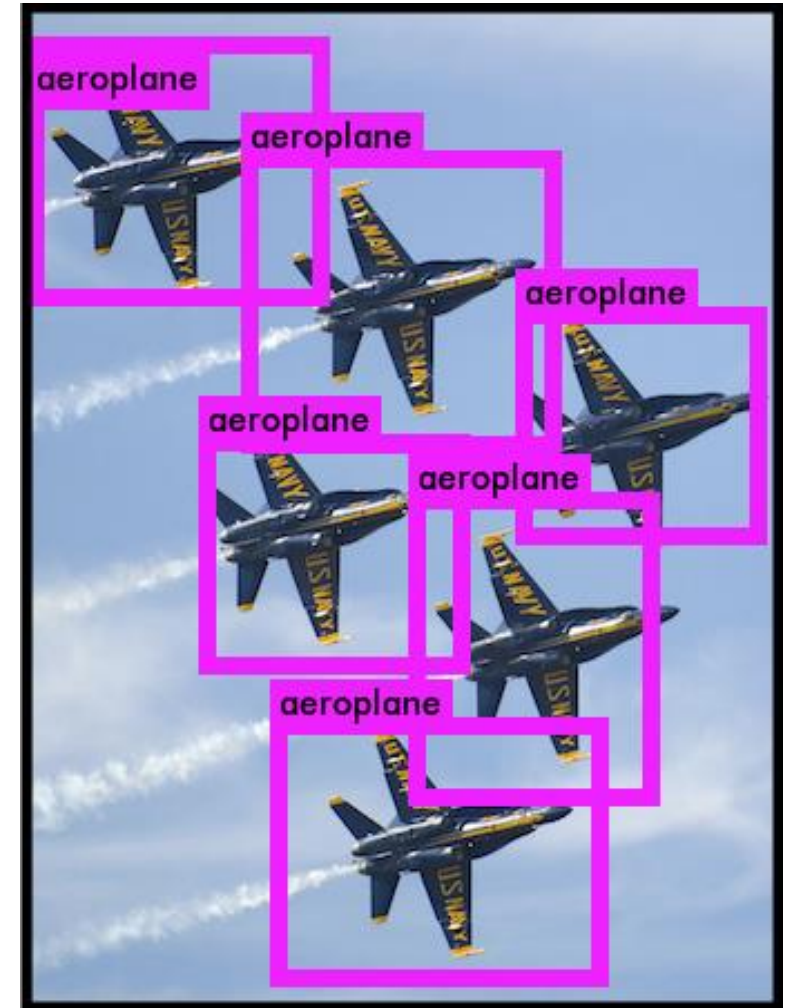
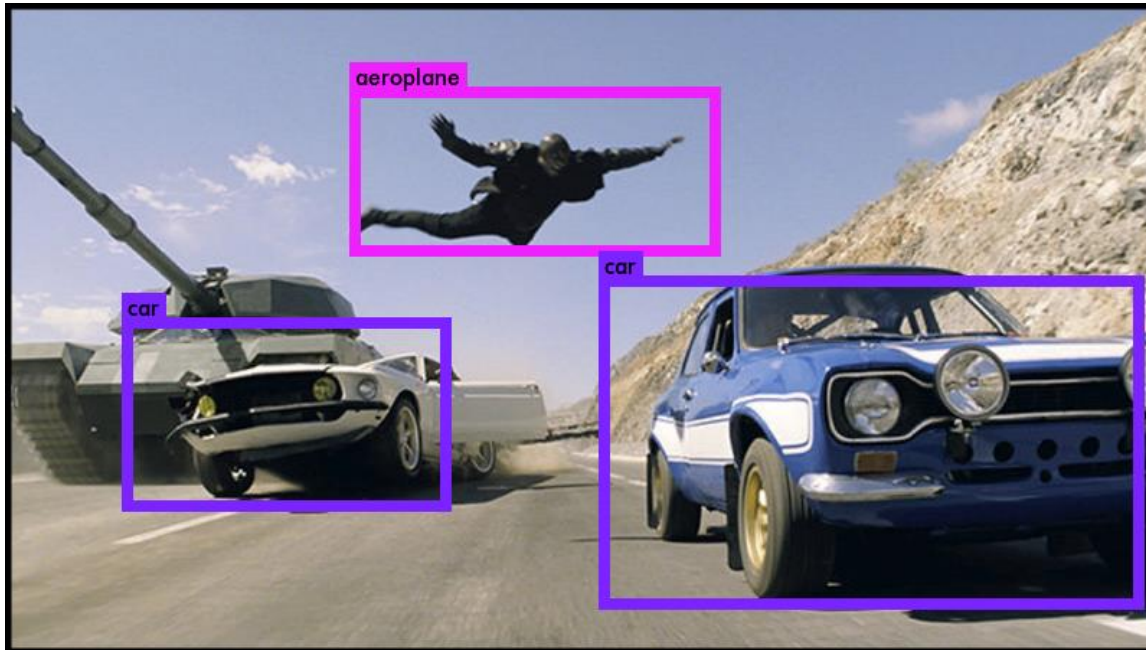


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

YOLO – Benefits

- Fast. Good for real-time processing
- End-to-end training



YOLO – Limitations

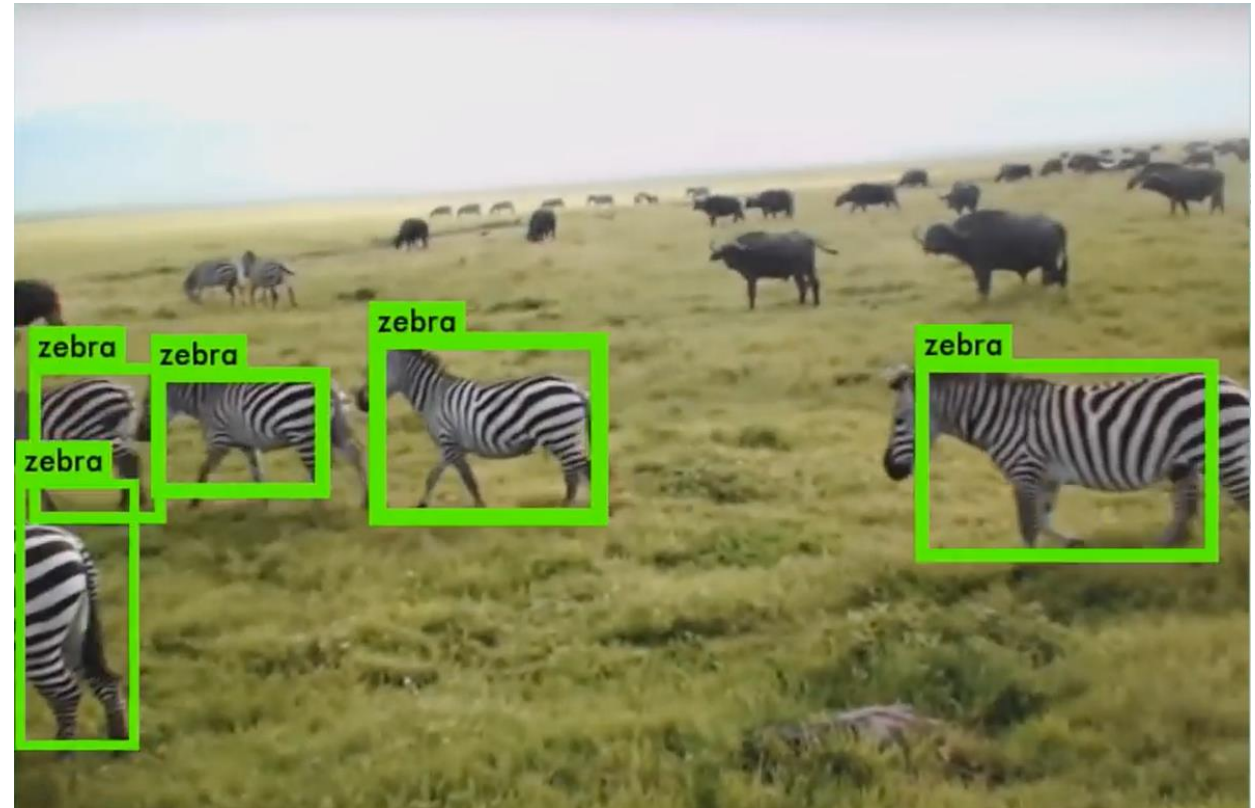


Image credit: <https://pjreddie.com/darknet/yolov1/>



YOLO – Limitations

- Difficult to detect small objects
- Coarse predictions

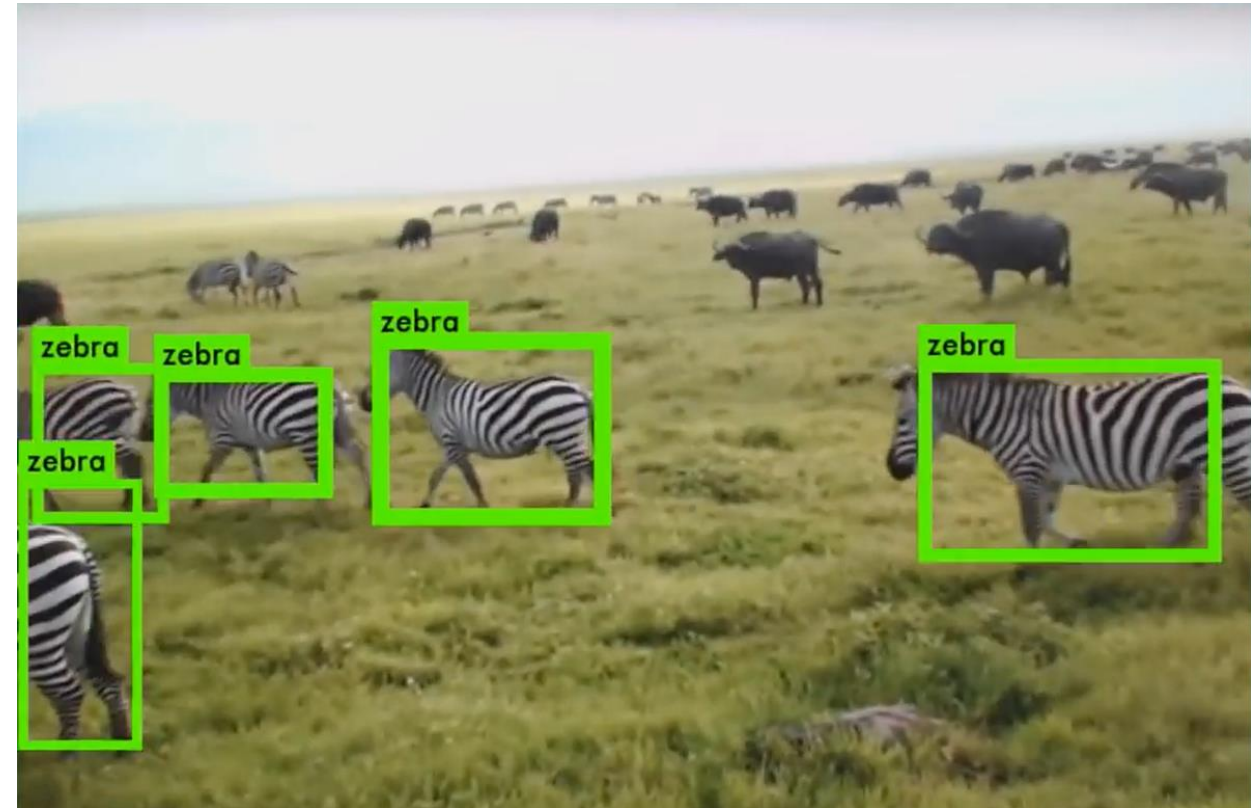
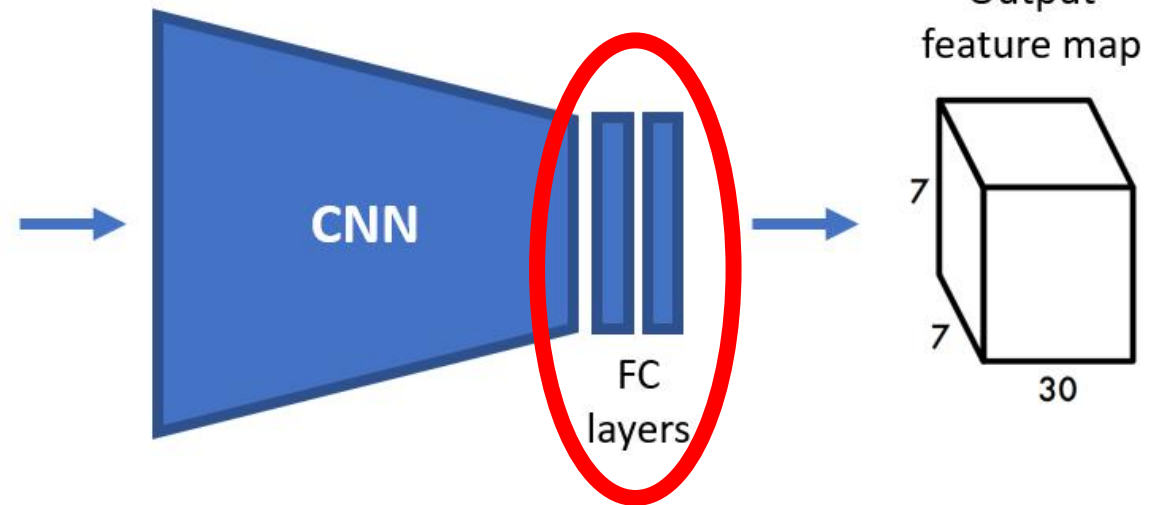


Image credit: <https://pjreddie.com/darknet/yolov1/>

YOLO – Limitations

- Difficult to detect small objects
- Coarse predictions
- Fixed input size



YOLO – Limitations

- Difficult to detect small objects
- Coarse predictions
- Fixed input size
- A grid cell can predict only one class

Solutions:

Remove fc layers!

Predict class per bbox (not per cell)

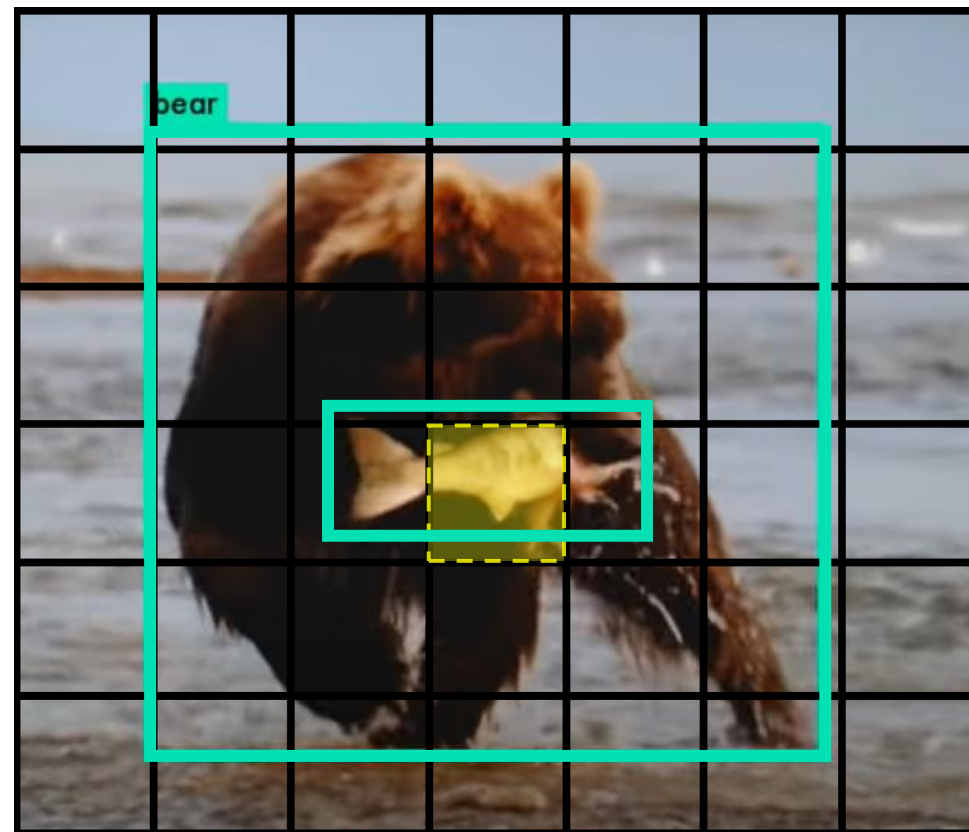


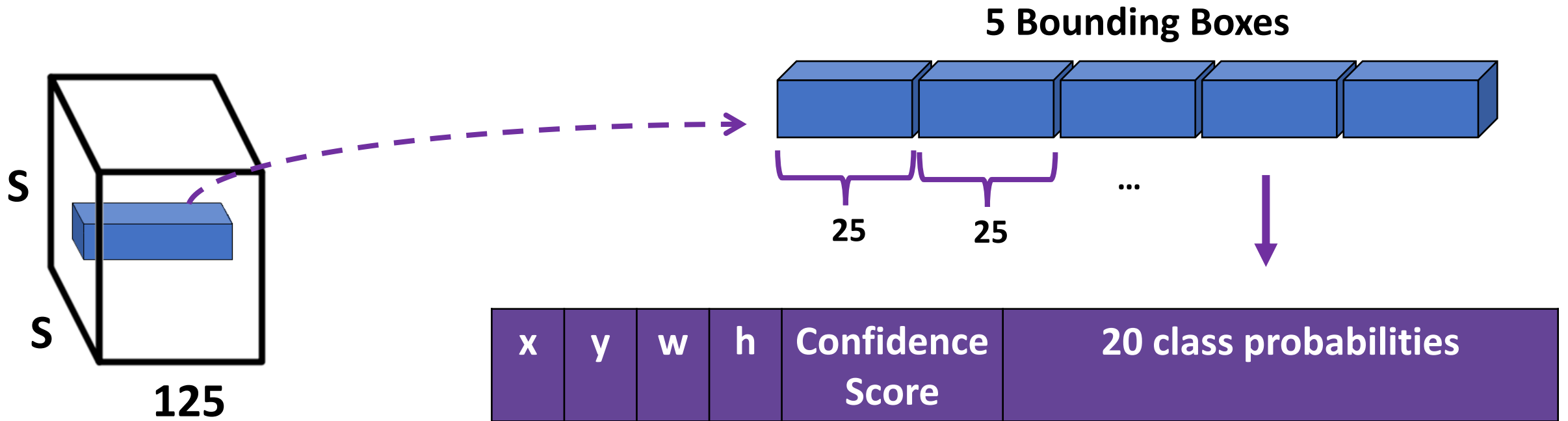
Image credit: <https://pjreddie.com/darknet/yolov1/>

YOLOv2

- Removed fully connected layers

YOLOv2

- Removed fully connected layers
- A grid cell predicts class probabilities for **each** box



YOLOv2

- Removed fully connected layers
- A grid cell predicts class probabilities for **each** box
- Working with anchor boxes (prior bounding boxes)

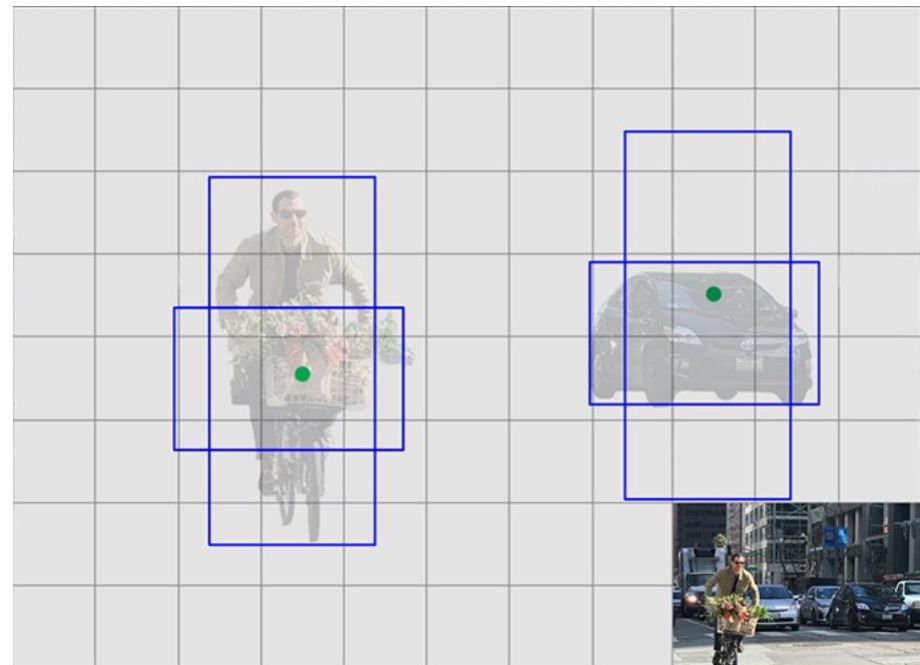
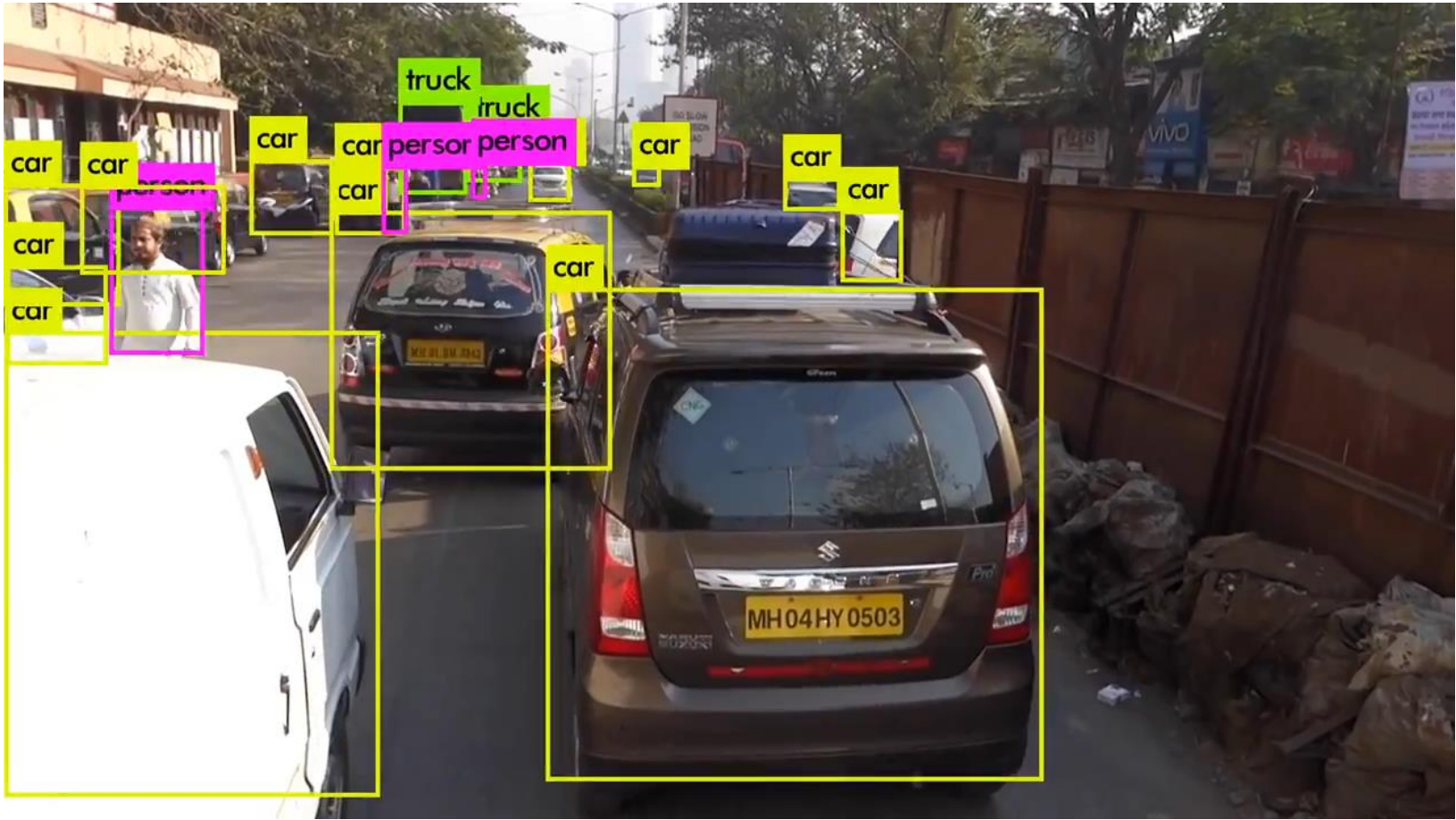


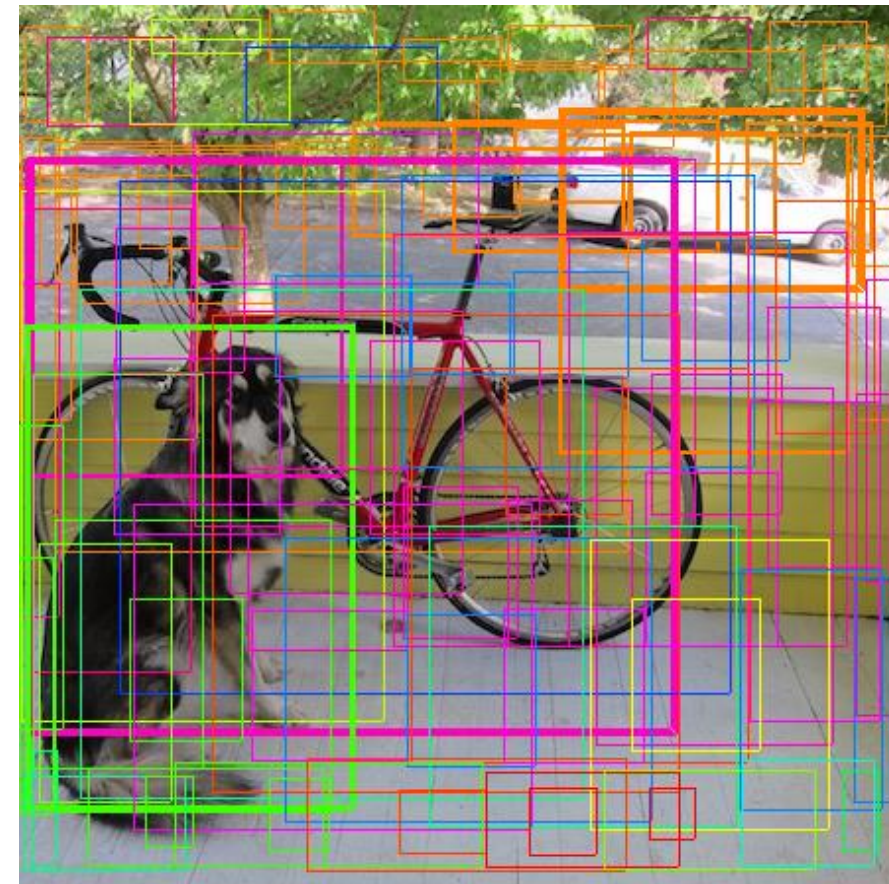
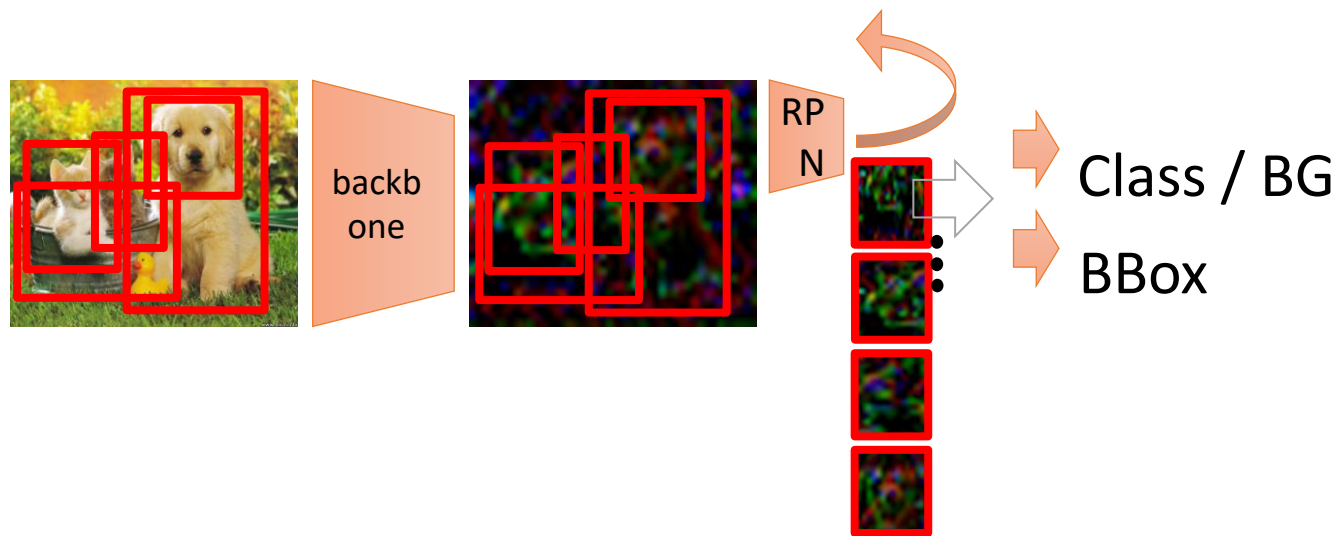
Image credit: [medium](#)

There's always room for improvement!

- YOLOv3
J. Redmon, A. Farhadi. [Yolov3: An incremental improvement](#), 2018
- YOLO v4
A. Bochkovskiy, C. Wang, H. Liao. [Yolov4: Optimal speed and accuracy of object detection](#) (Feb. 2020)
- YOLOv5
[YOLOv5 by ultralytics](#) (June 2020)
- PP-YOLO
X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding, S. Wen. [Pp-yolo: An effective and efficient implementation of object detector](#) (June 2020)
- PP-YOLOv2 (2021)
X. Huang, X. Wang, W. Lv, X. Bai, X. Long, K. Deng, Q. Dang, S. Han, Q. Liu, X. Hu, D. Yu, Y. Ma, O. Yoshie. [PP-YOLOv2: A Practical Object Detector](#) (2021)
- ...



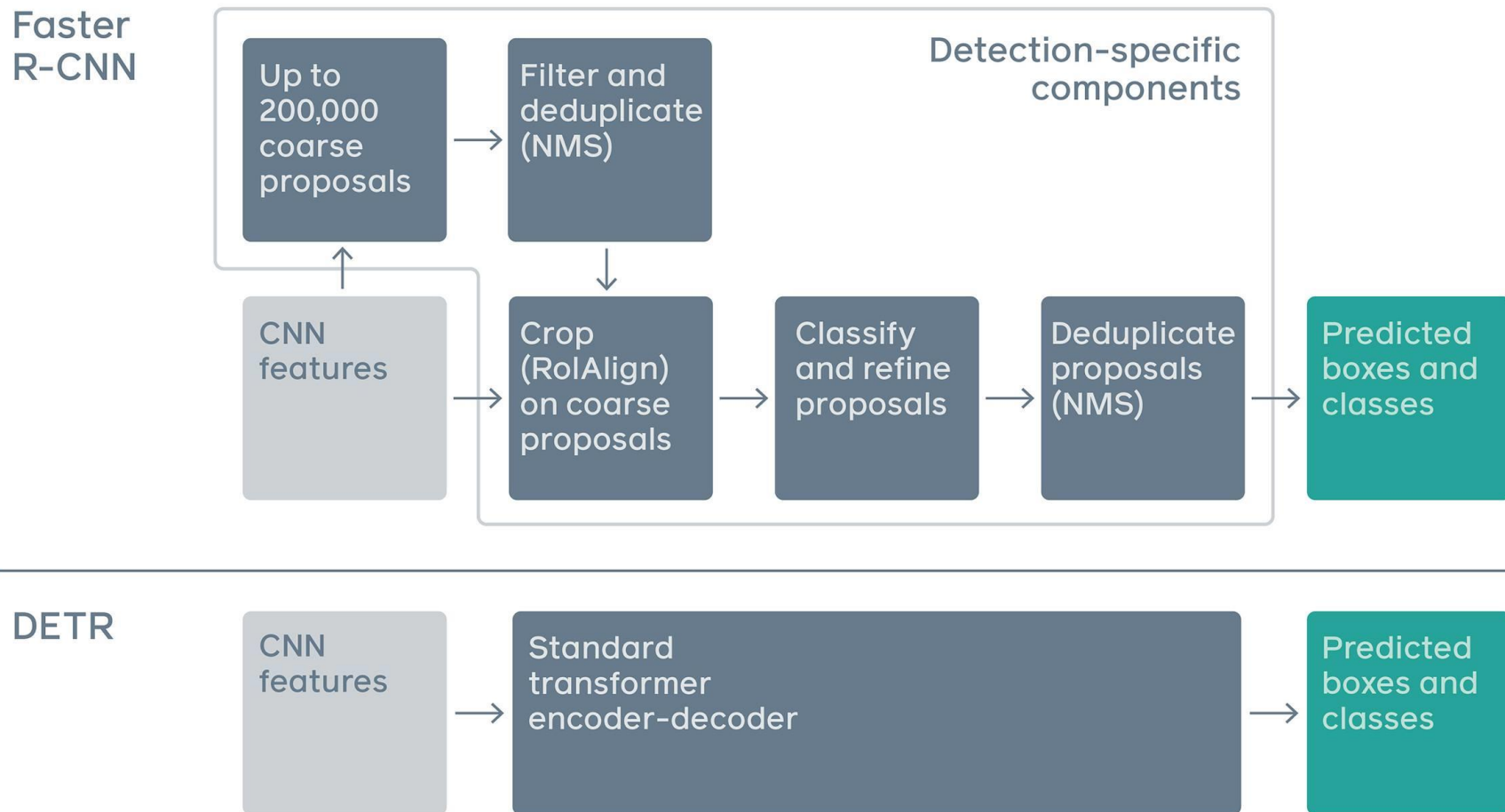
Object Detection So Far..



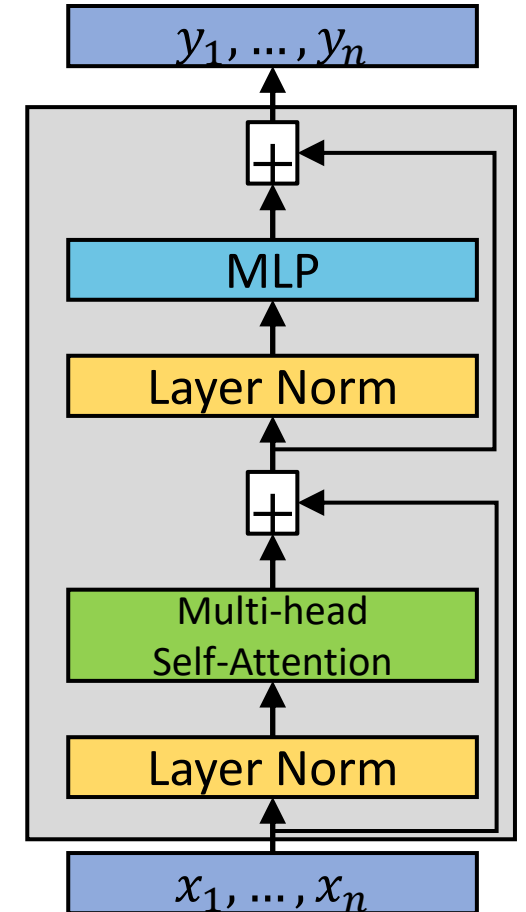
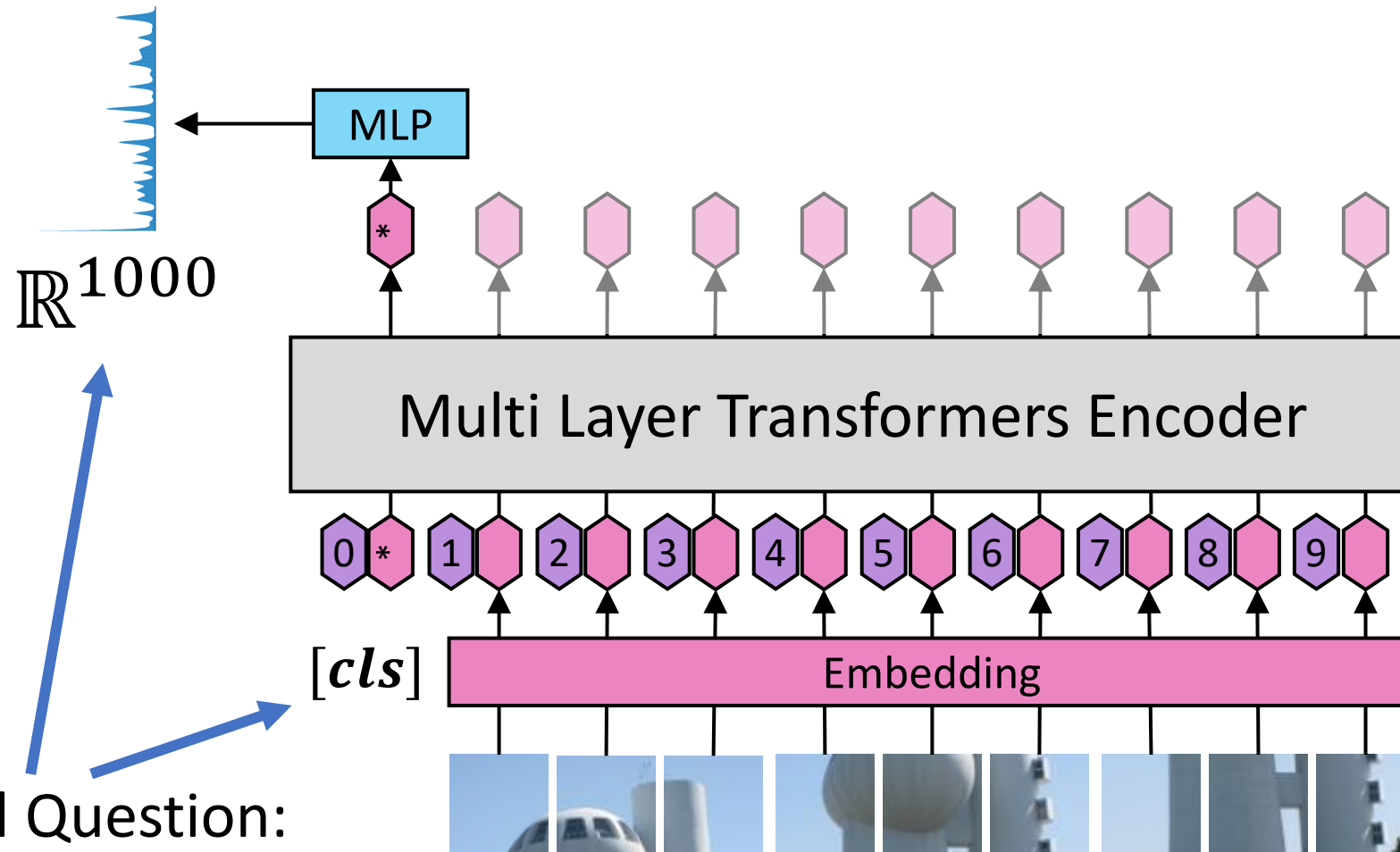
Annoying parts of Detection:
Multiple Bboxes \rightarrow NMS, Threshold

Can we do better? "Set" prediction

DETR: DEtection with TRansformers



Vision Transformers (ViT)

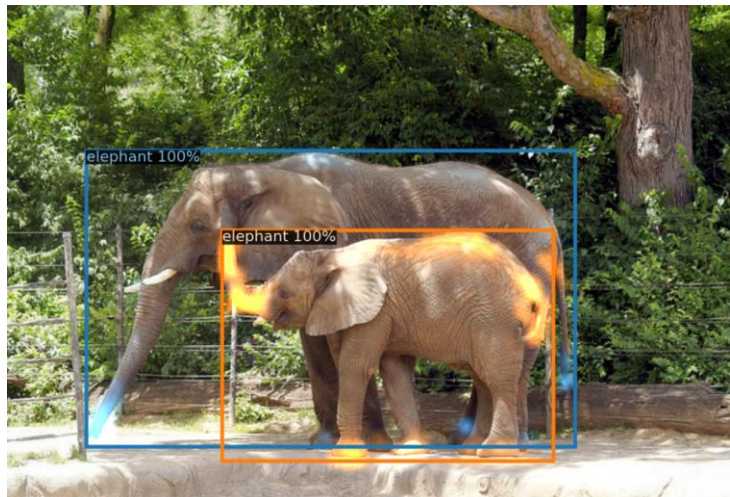
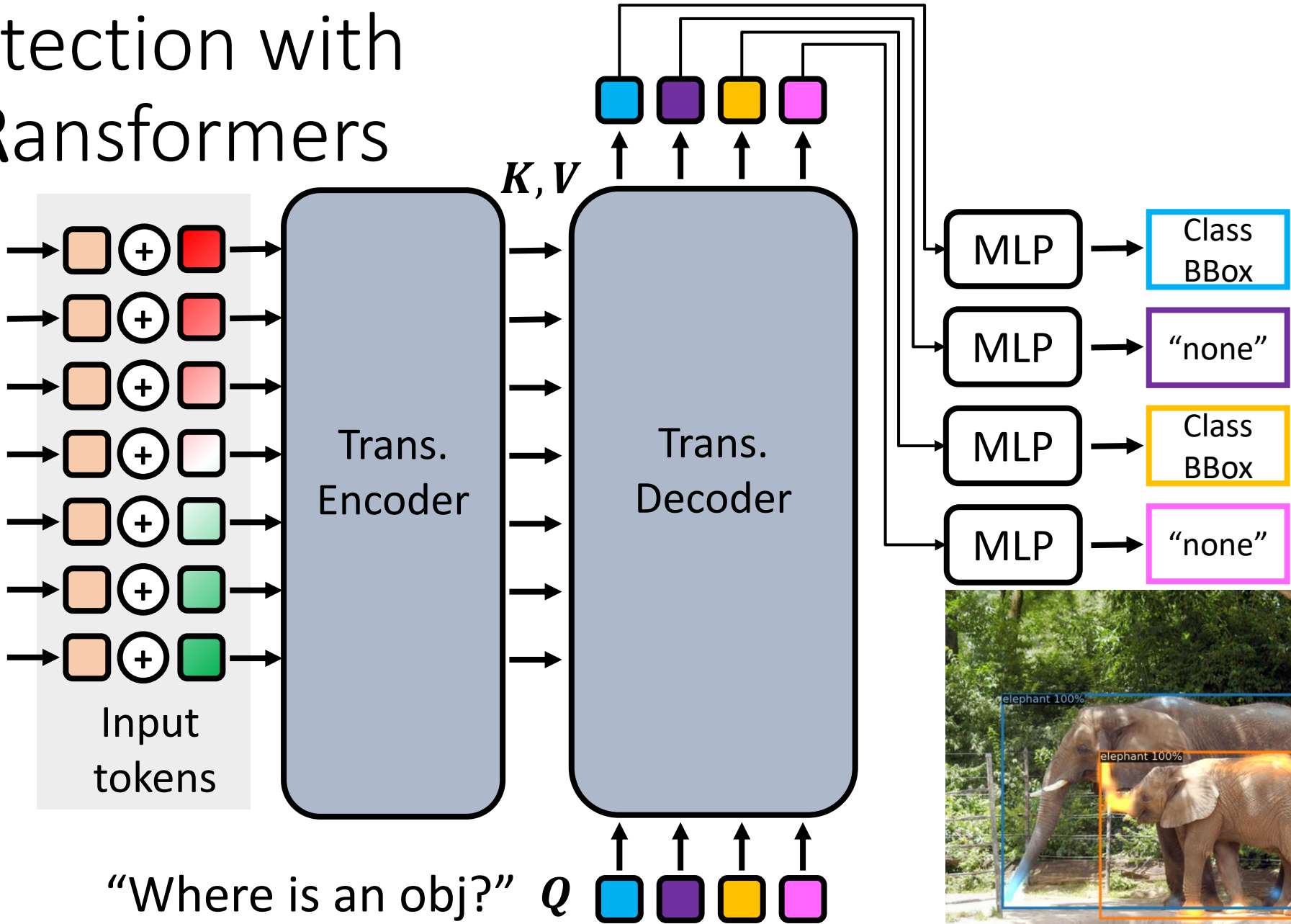


Global Question:
“What is the class?”

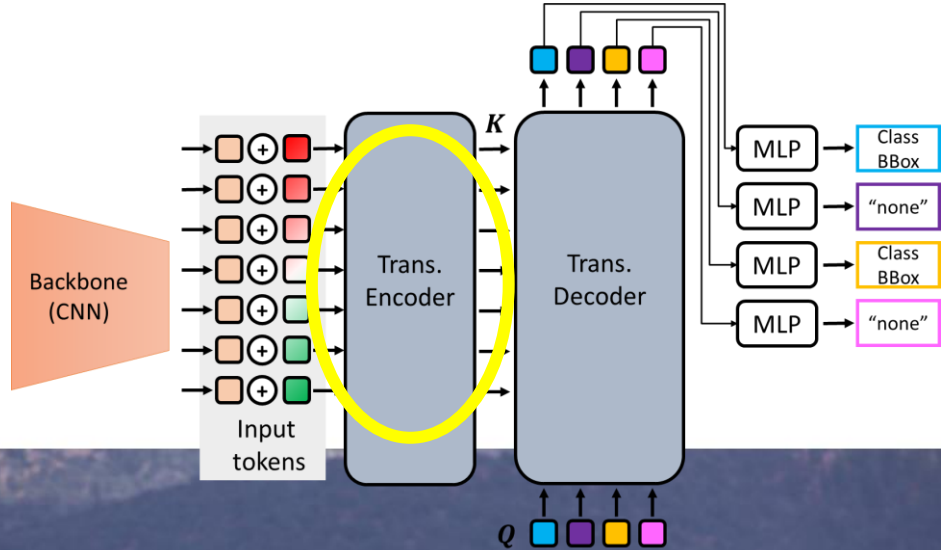
Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. [“An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale”](#) (ICLR 2021)

DETR: DEtection with TRansformers

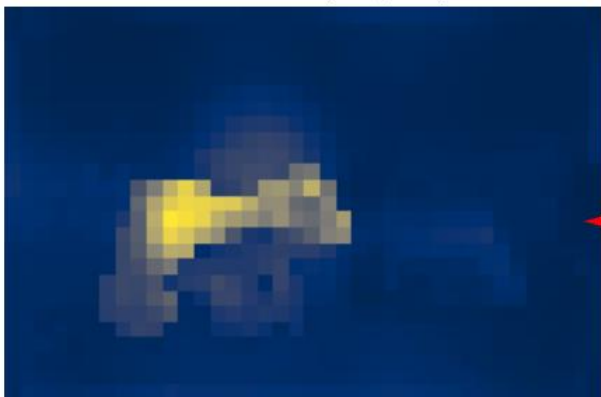
Backbone (CNN)



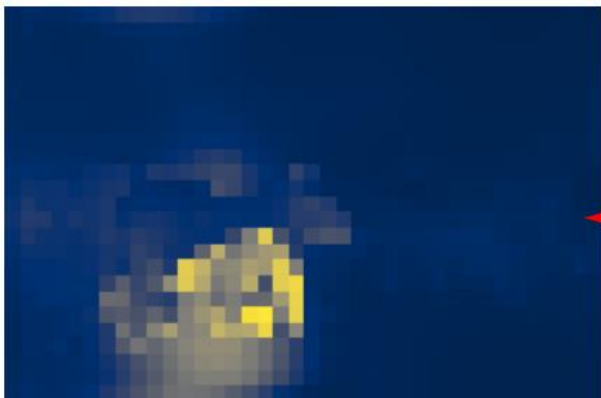
DETR



self-attention(430, 600)



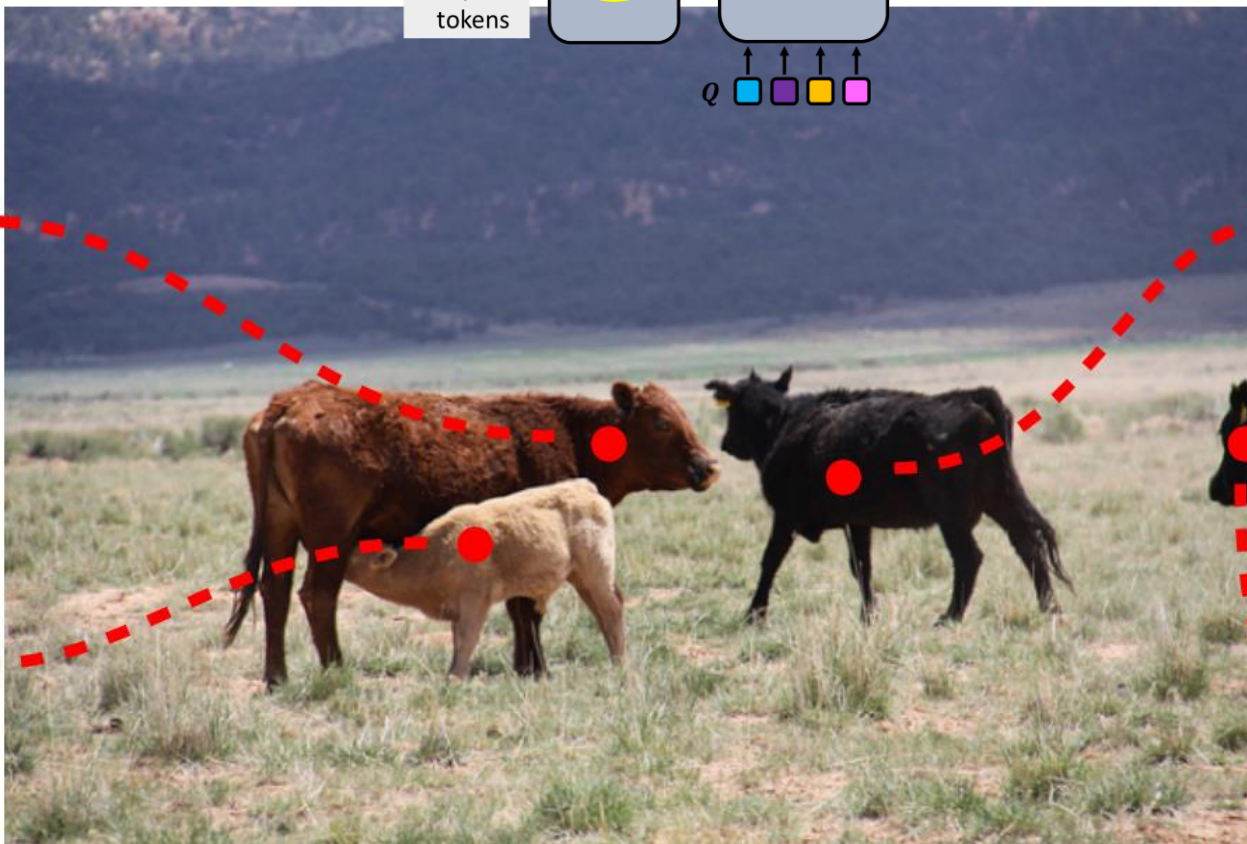
self-attention(520, 450)



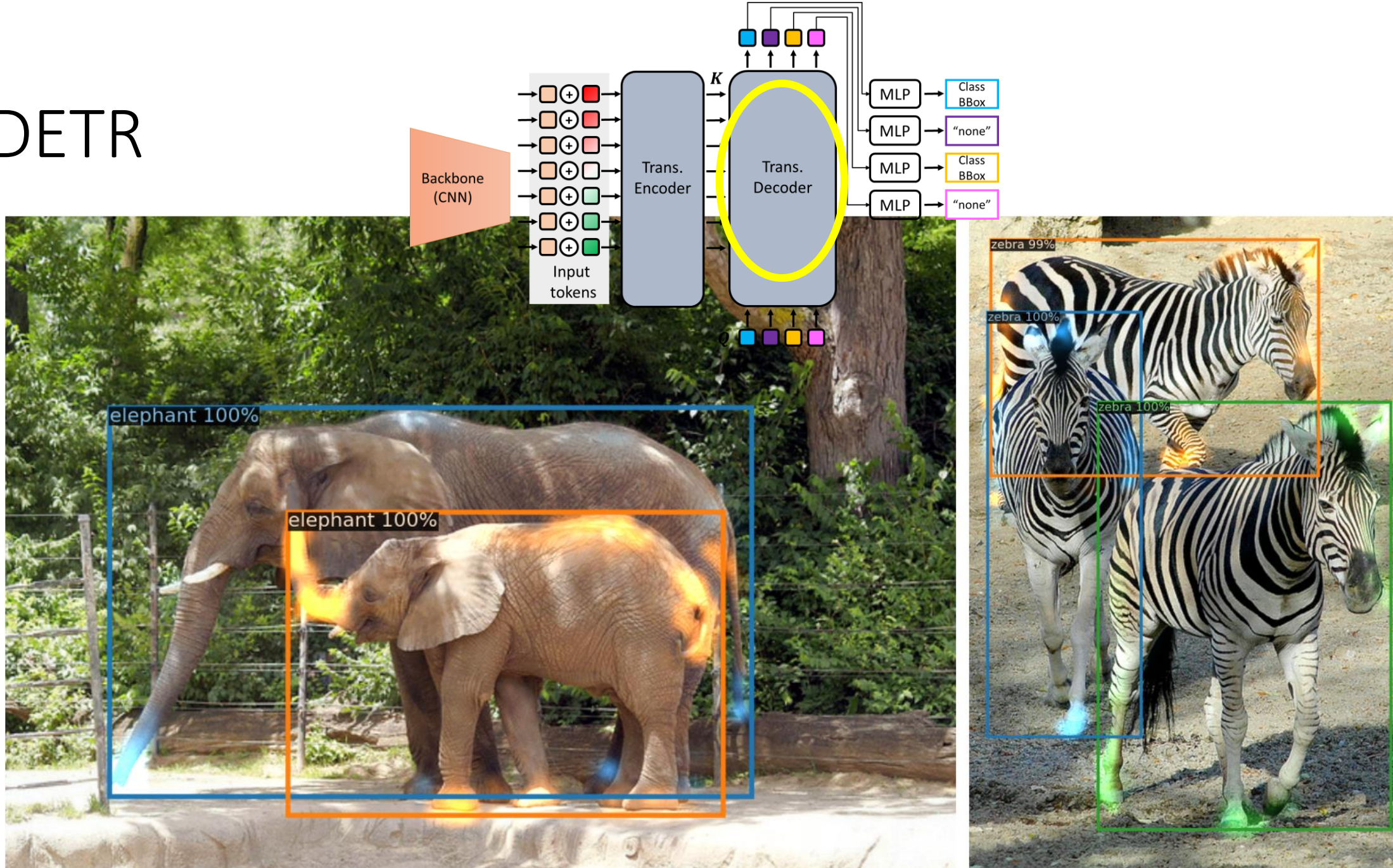
self-attention(450, 830)



self-attention(440, 1200)

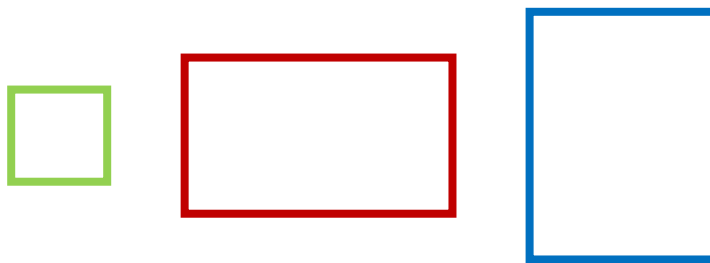
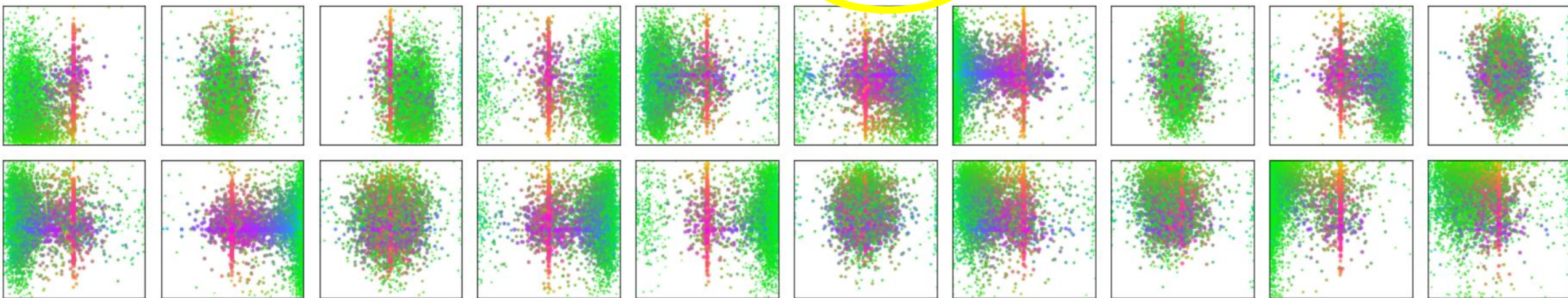
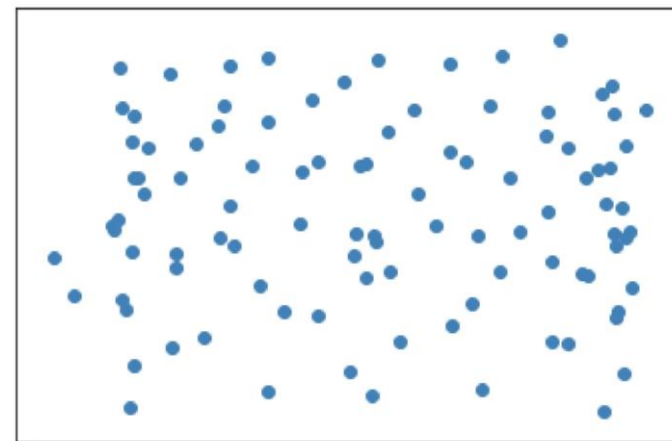
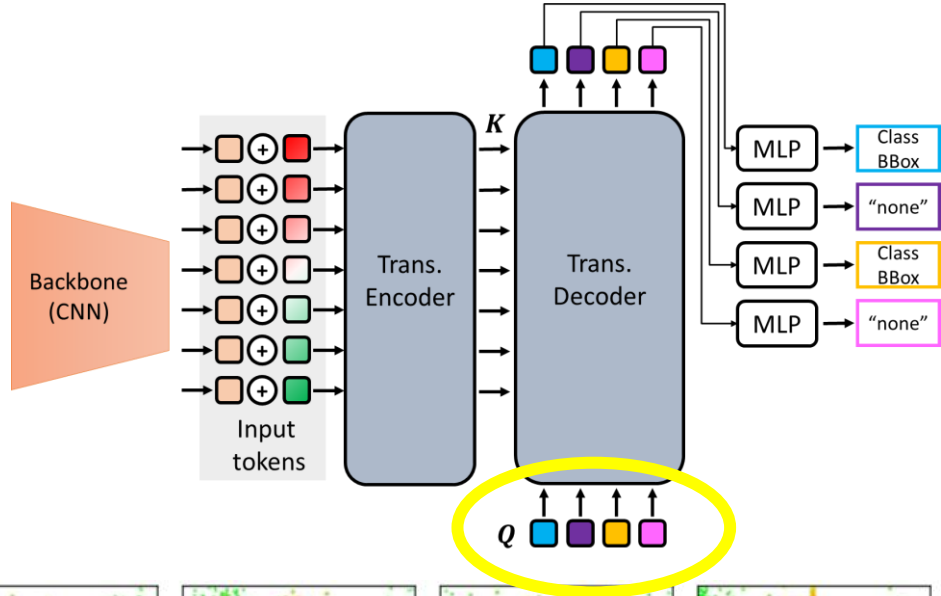


DETR



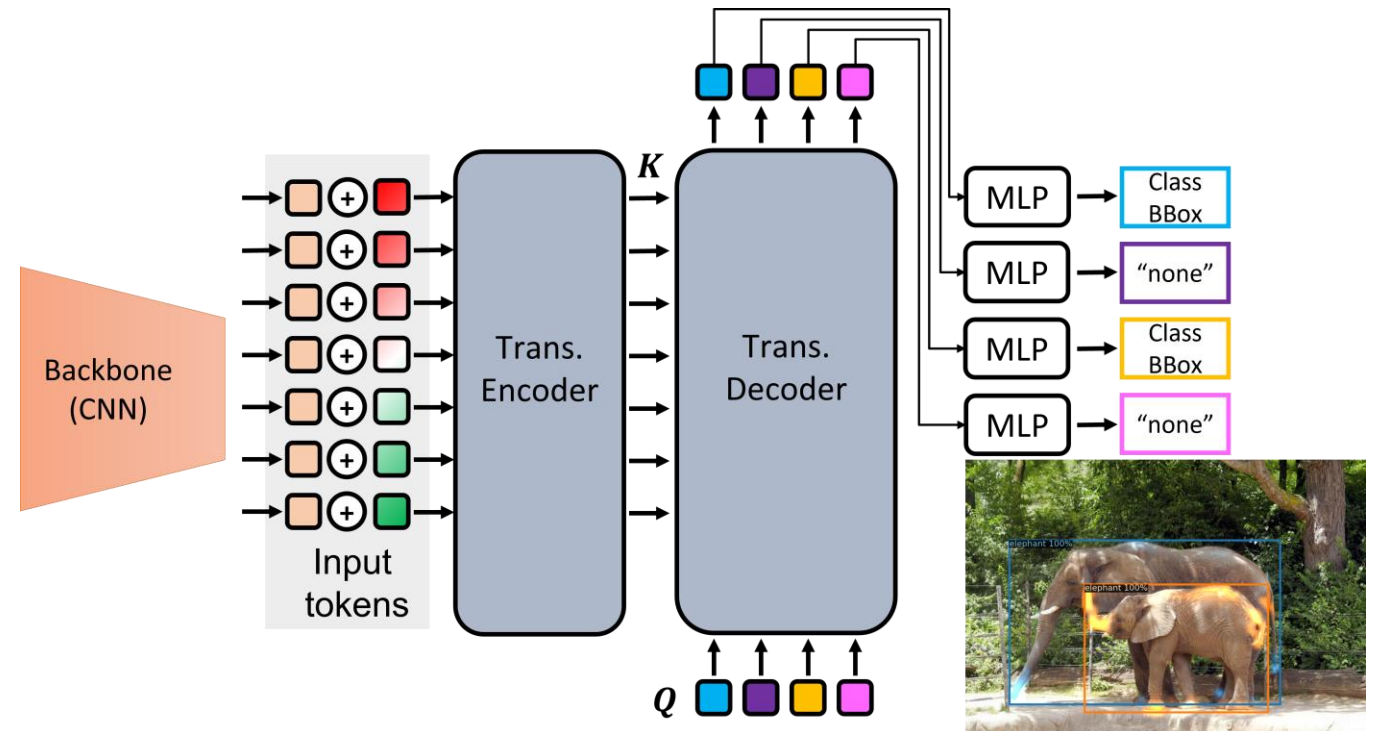
Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. [DETR: End-to-End Object Detection with Transformers](#) (ECCV 2020)

DETR



DETR

- End-to-end training
- No NMS (!)
- Simple approach
- On par with Faster-R-CNN



Previously

- Object detection
 - Faster RCNN
- Semantic segmentation
 - FCN / DeepLab
 - UNet

Today

- (Multiple) Object detection
 - YOLO (You Only Look Once)
 - DETR (DEtection TRansformer)
- Semantic Segmentation
 - Segmenter

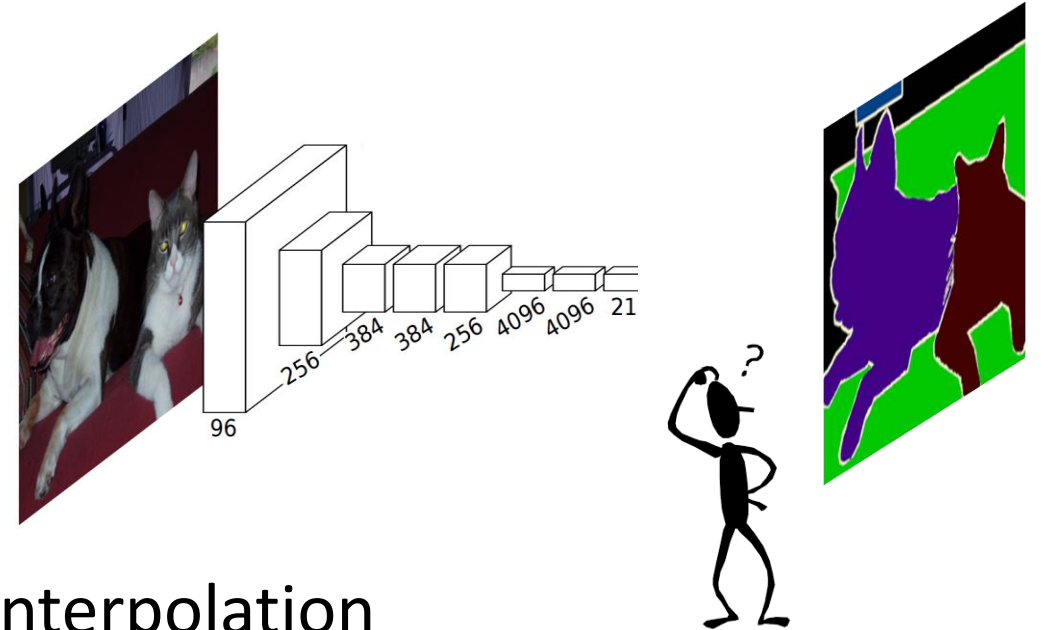
Semantic Segmentation



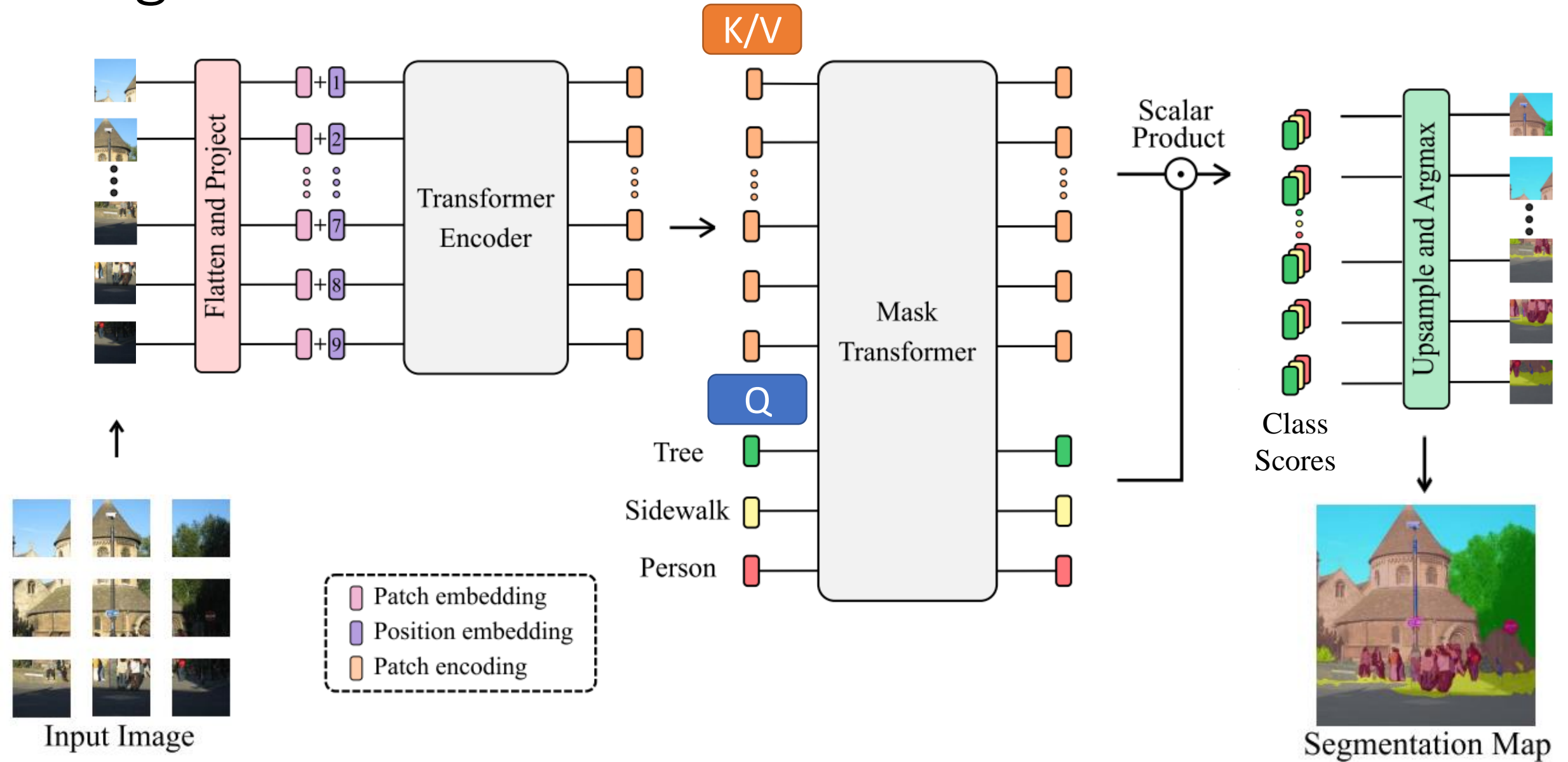
Semantic Segmentation

Resolution vs. Semantic information

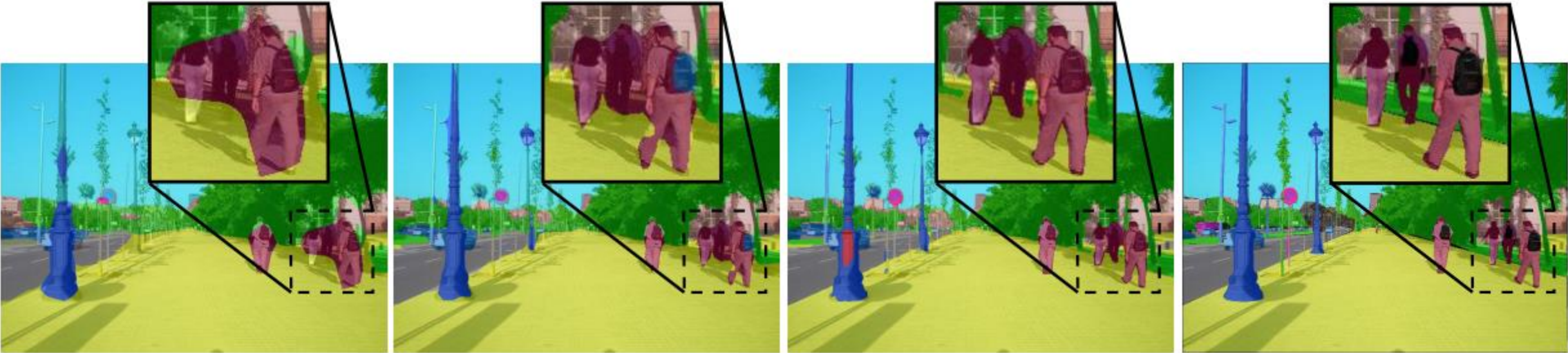
- FCN: using “transposed convolution”
- DeepLab: dilated convolution + simple interpolation
- U-net: skip connections



Segmenter



Segmenter – Effect of Patch Size



(a) Patch size 32×32

(b) Patch size 16×16

(c) Patch size 8×8

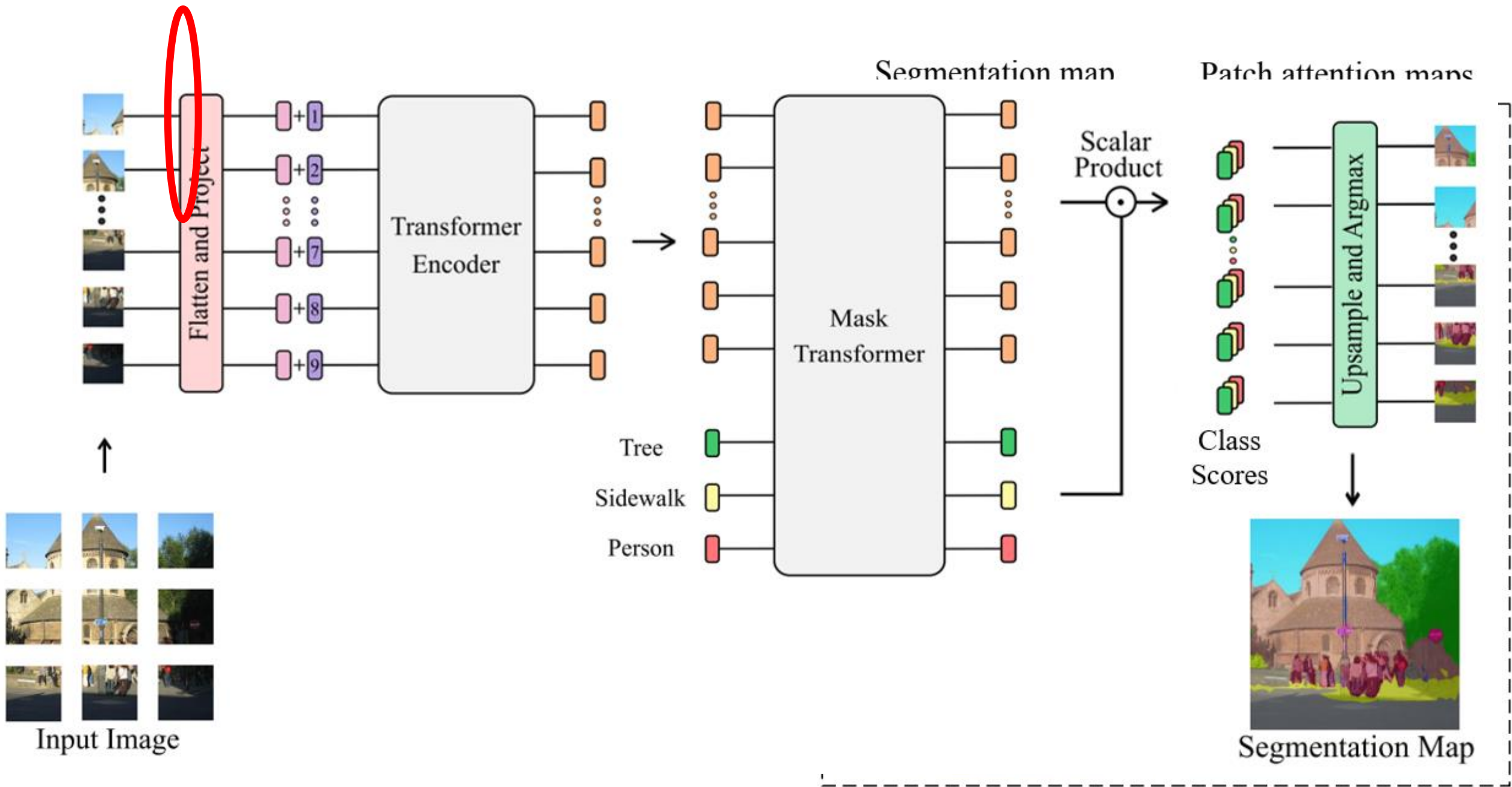
(d) Ground Truth

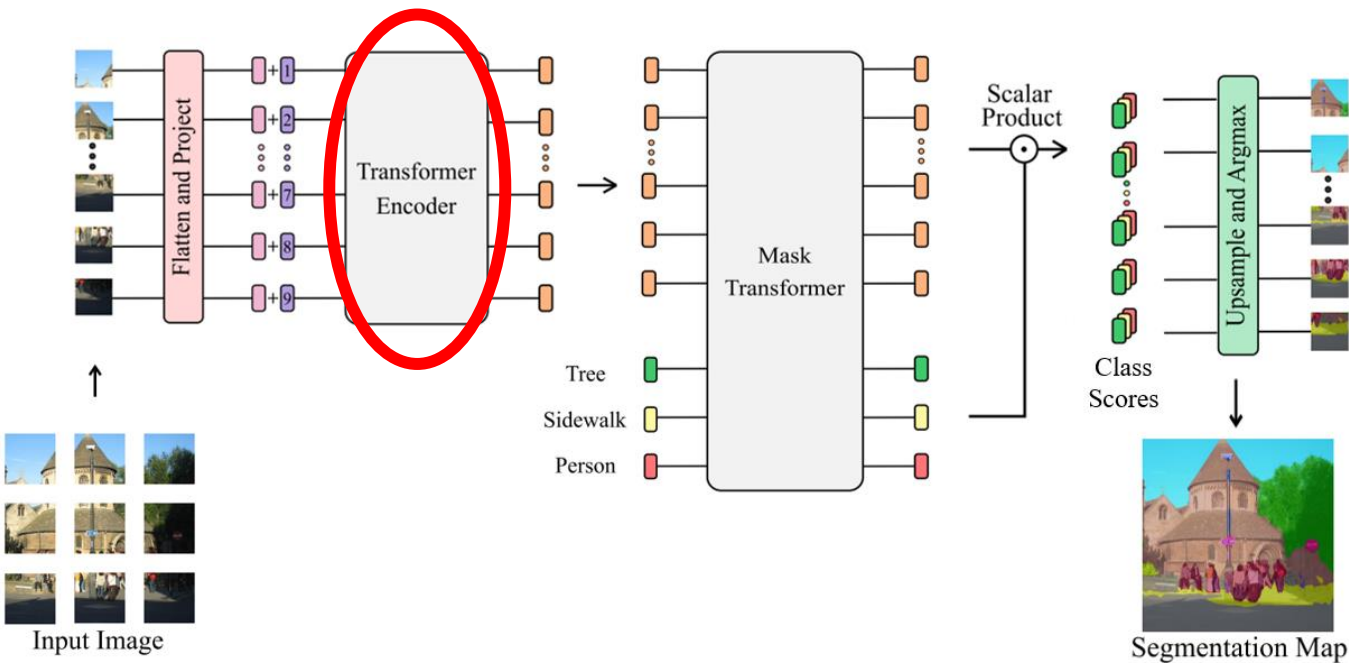
Fast
Low accuracy



High computational cost
High accuracy







Input



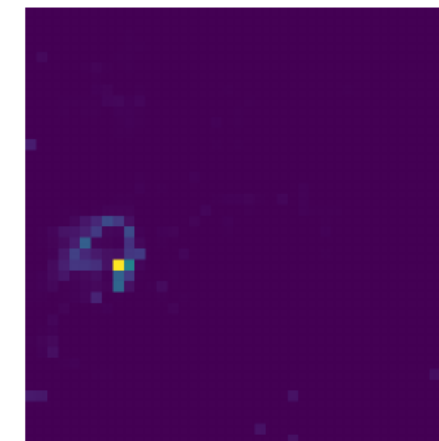
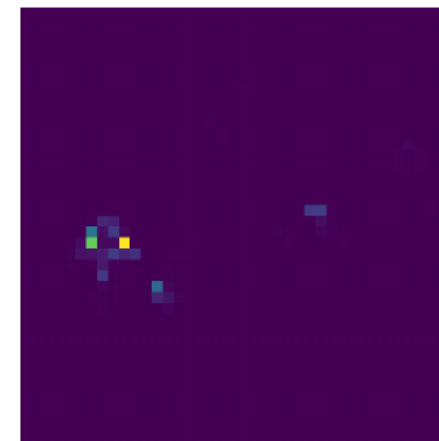
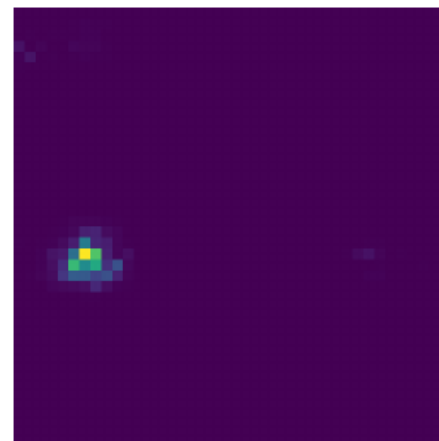
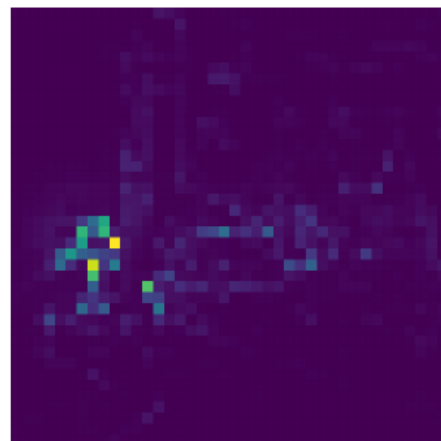
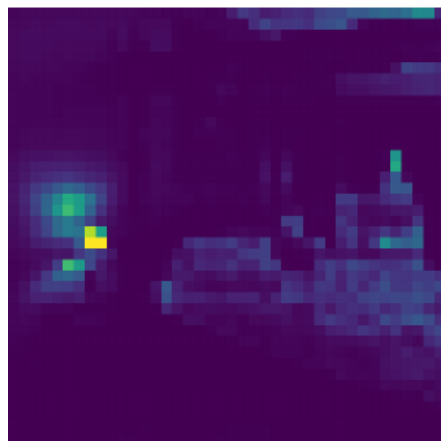
Layer 1

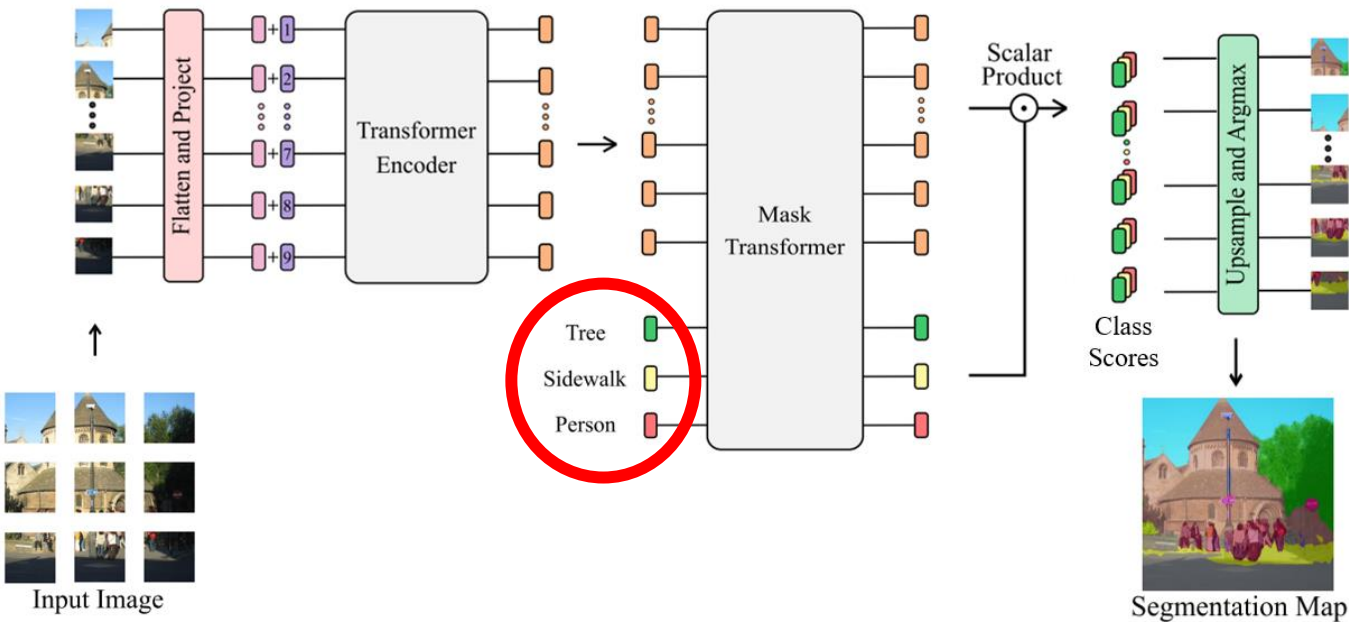
Layer 4

Layer 8

Layer 12

Layer 16

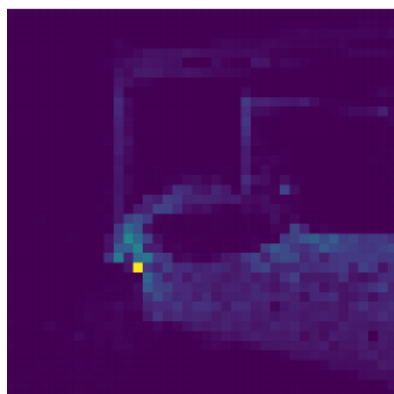




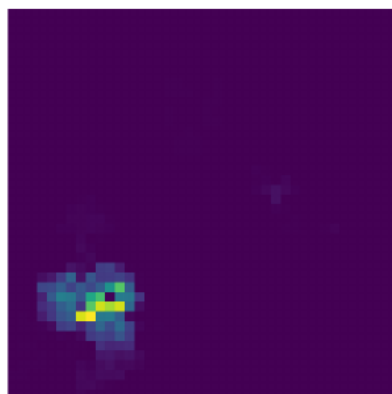
Prediction



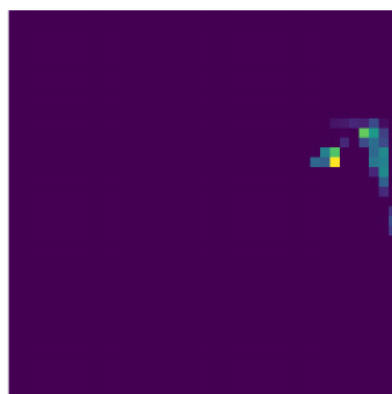
CLS 7
(bed)



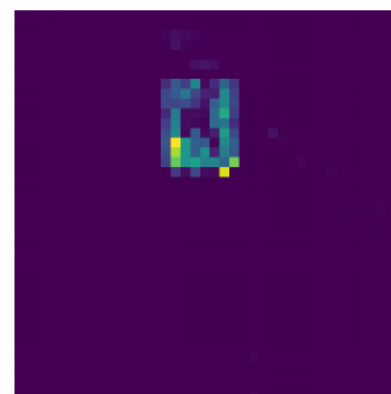
CLS 15
(table)



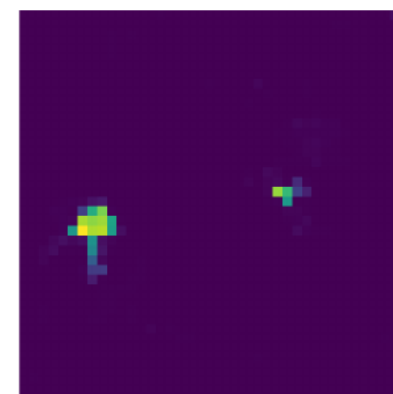
CLS 18
(curtain)



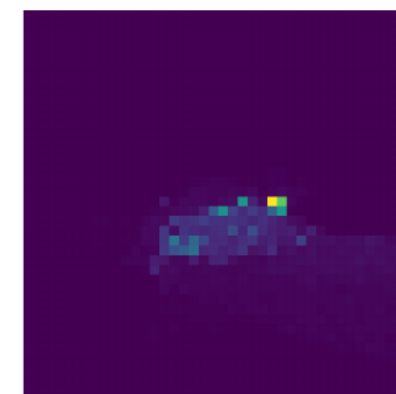
CLS 22
(painting)



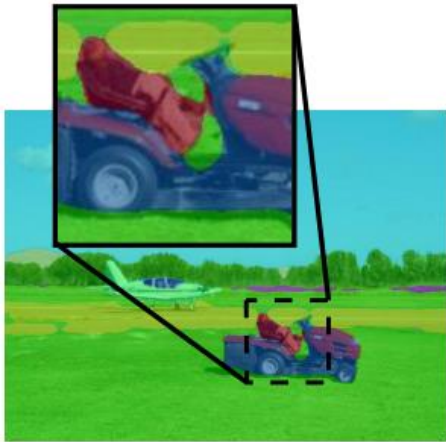
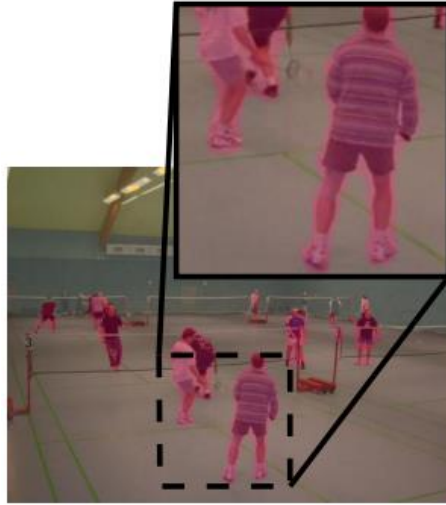
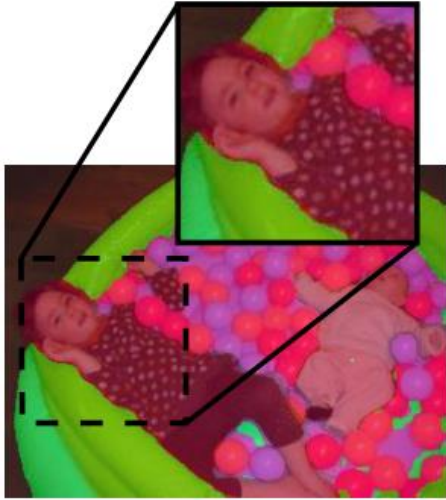
CLS 36
(lamp)



CLS 57
(pillow)



Segmenter



[video](#)



Today

- (Multiple) Object detection
 - YOLO (You Only Look Once)
 - DETR (DEtection TRansformer)
- Semantic Segmentation
 - Segmenter

Next Week

- Final project expo!