Object Detection and Segmentation

Niv Haim

December 28th, 2022

credit: slides modified from Dolev Ofri & Shai Bagon (DL4CV 2021)

Previously

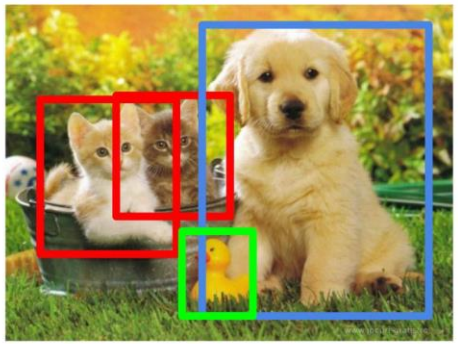
Classification

CAT

Classification + Localization

CAT

Object Detection



CAT, DOG, DUCK

Semantic Segmentation

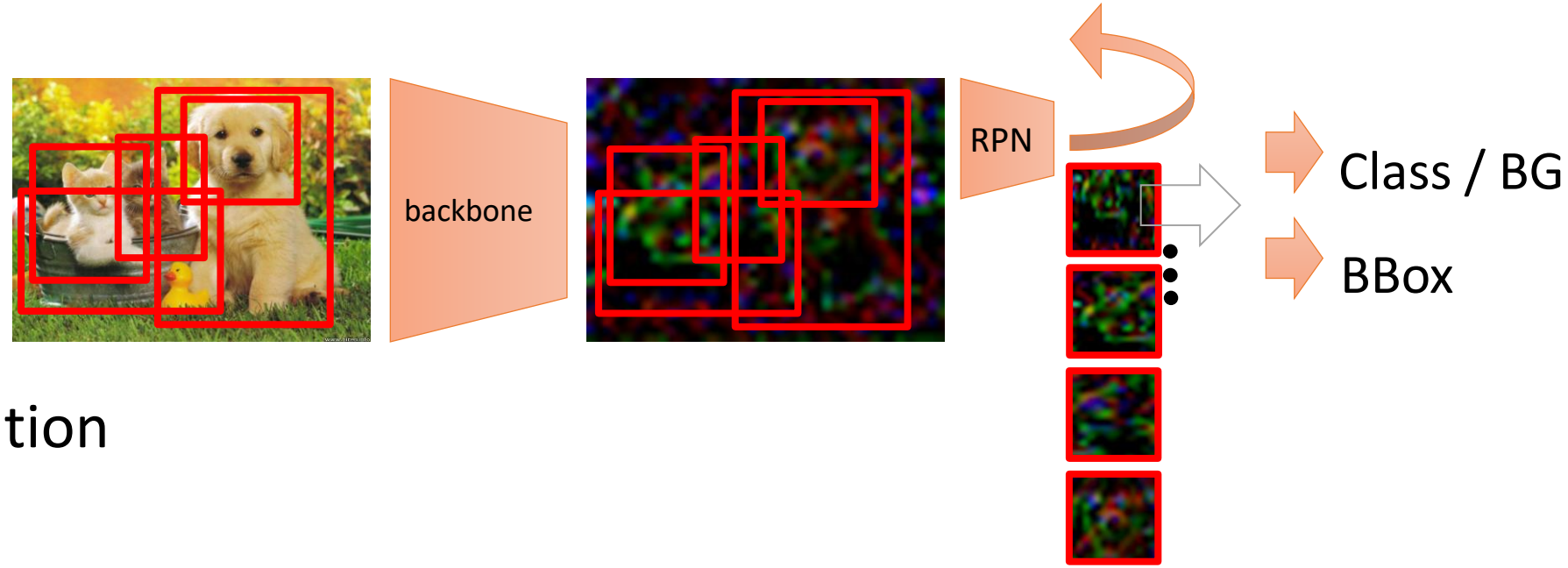
CAT, DOG, DUCK

Instance Segmentation

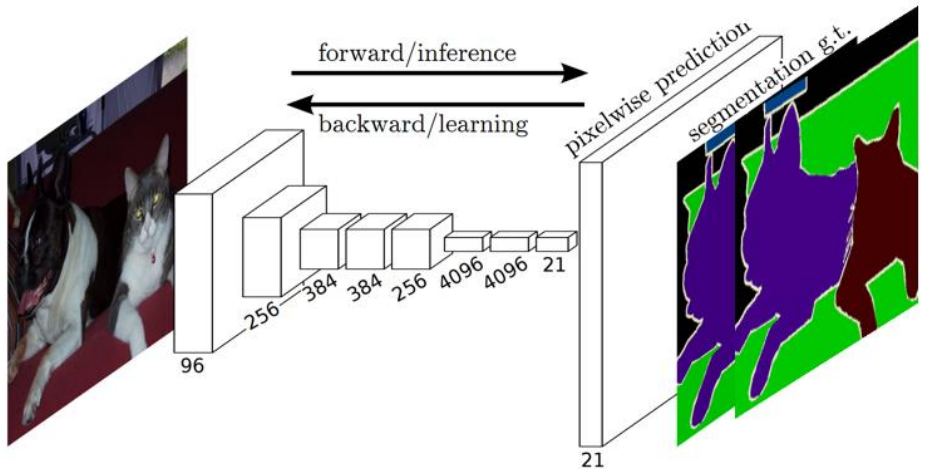
CAT, DOG, DUCK

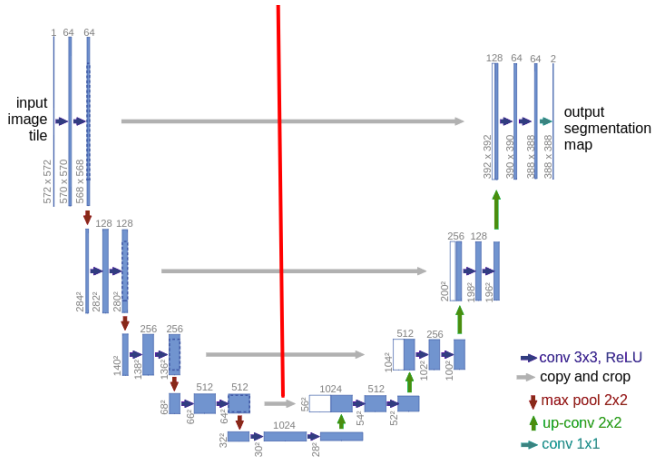
Previously

- Object detection
  - Faster RCNN



- Semantic segmentation
  - FCN / DeepLab
  - UNet





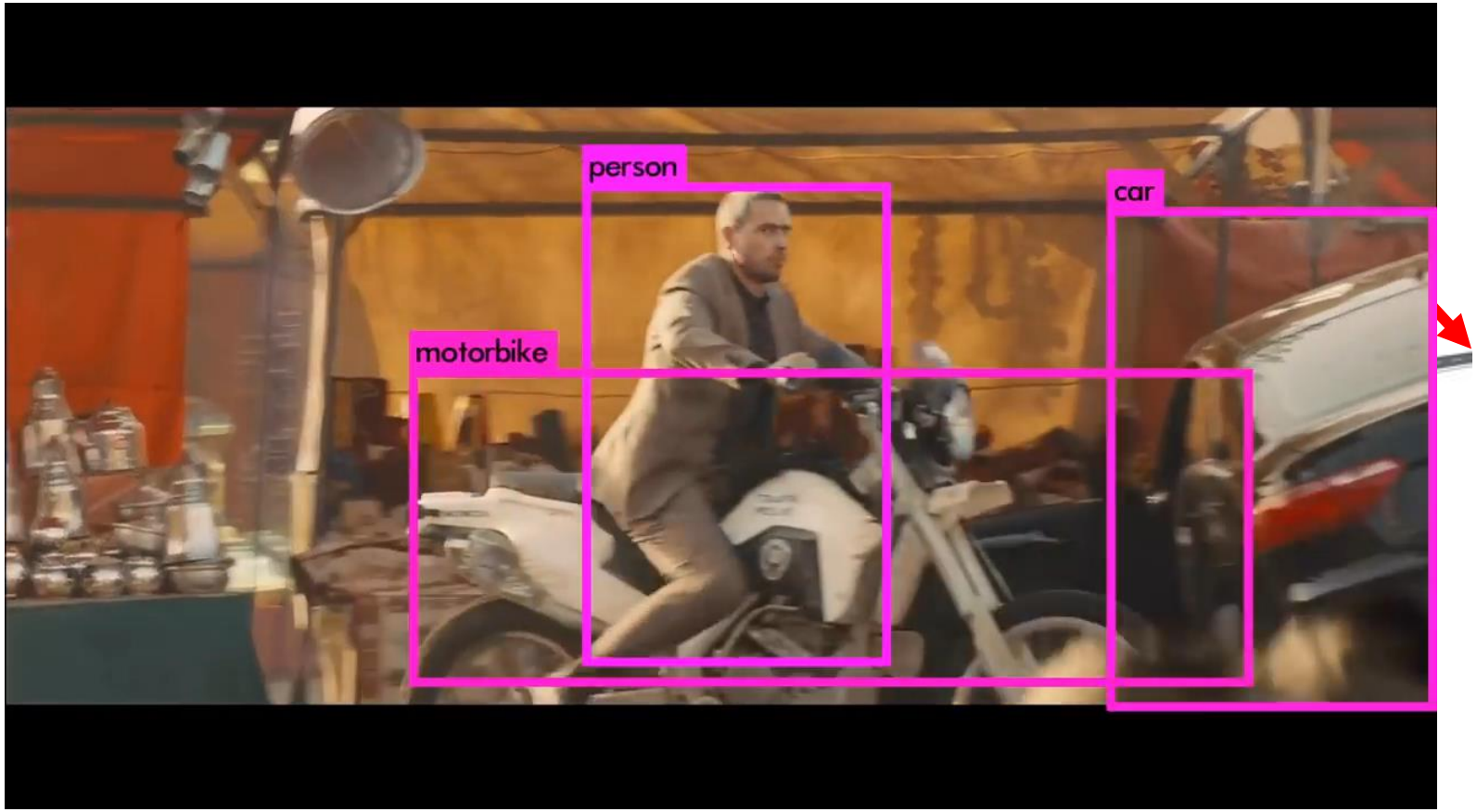
Previously

- 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



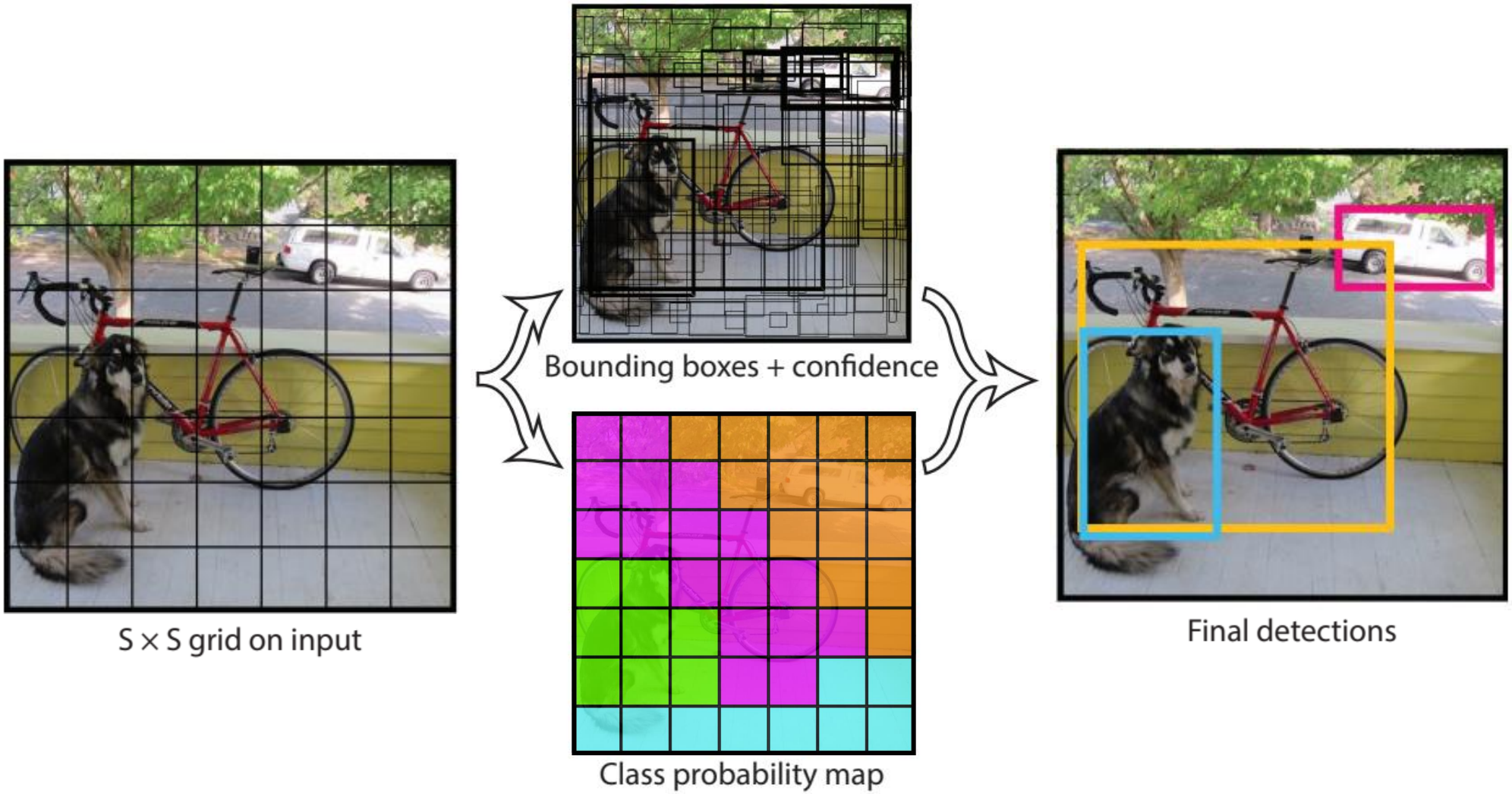
Single Shot:
SSD, YOLO …

Fast
High false rate

Video: YOLOv2

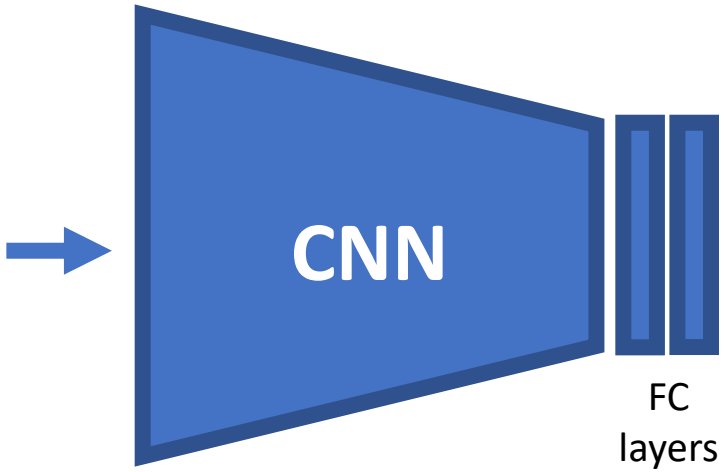
Slide credit: Shai Bagon

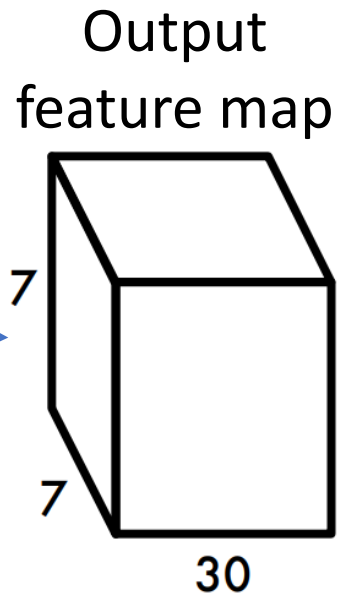
YOLO – Overview

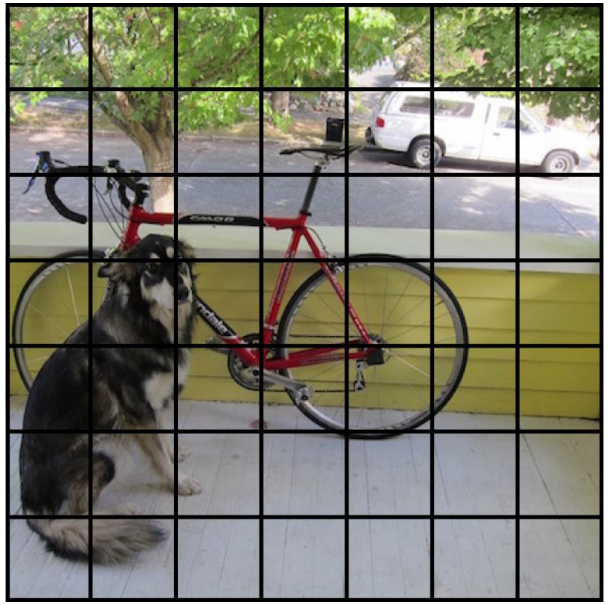


J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection, 2015 (CVPR 2016)

YOLO







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

YOLO

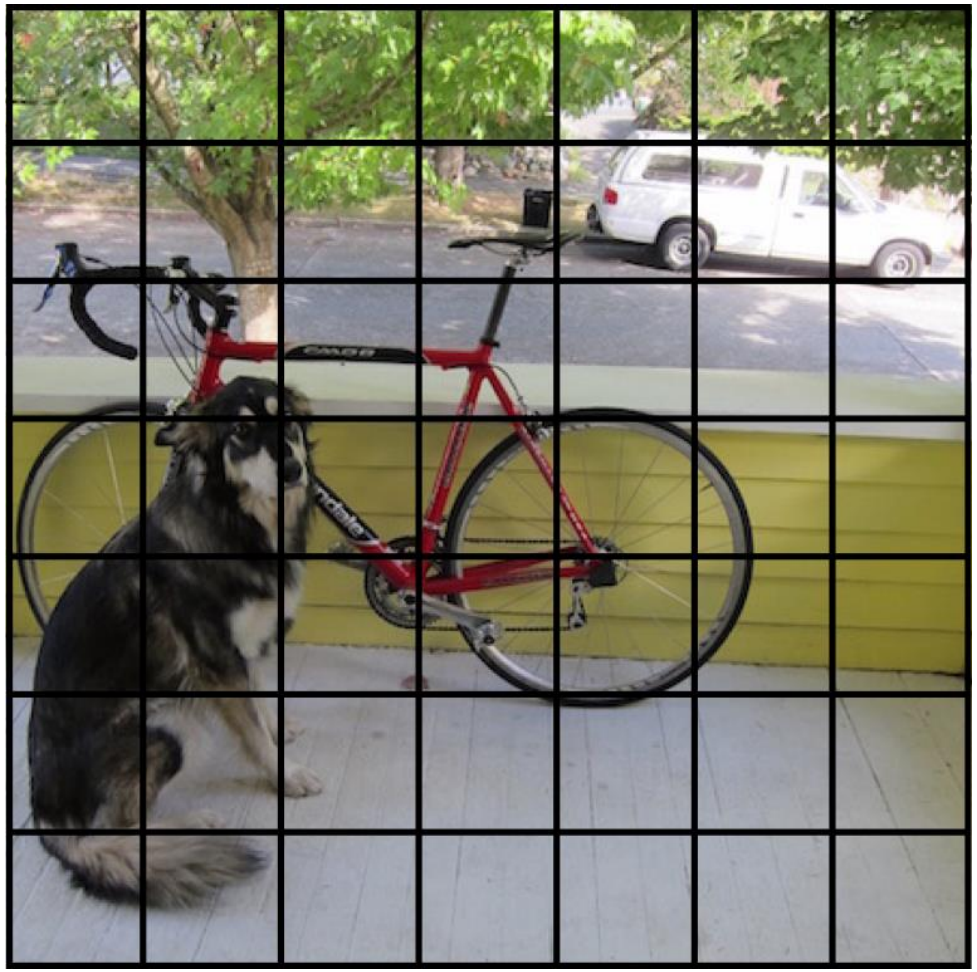


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

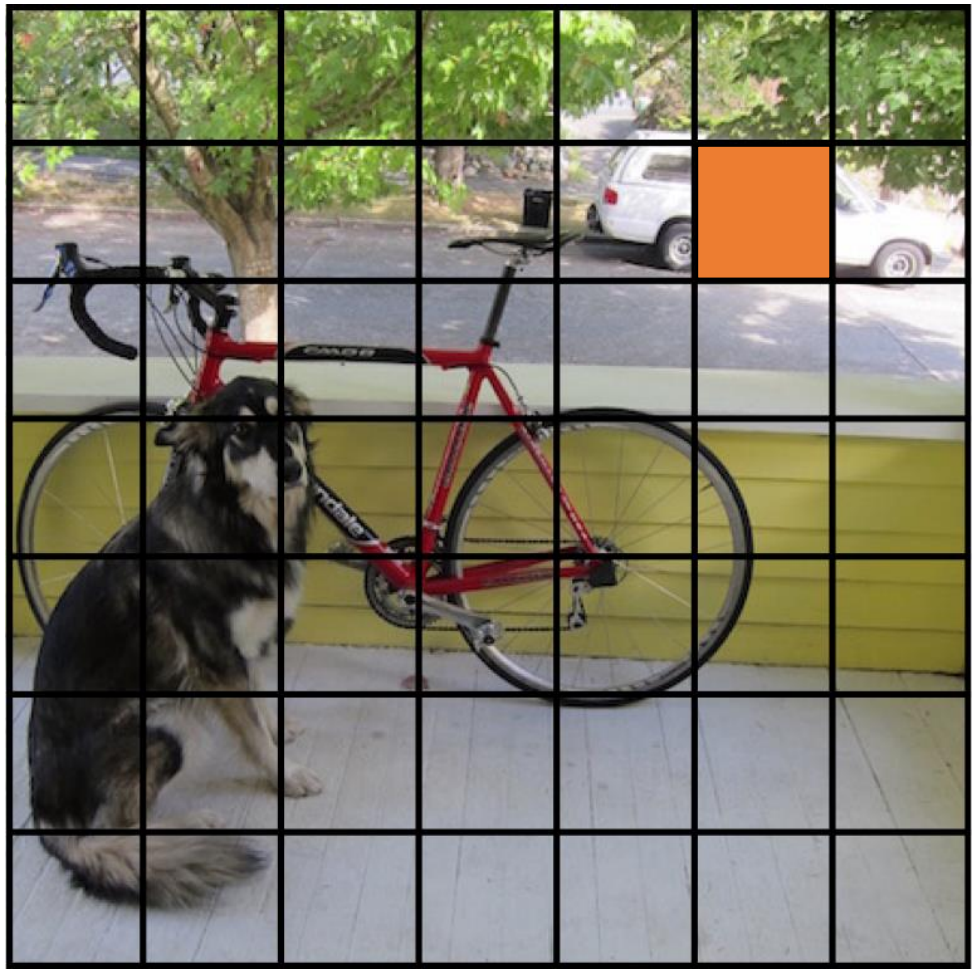


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

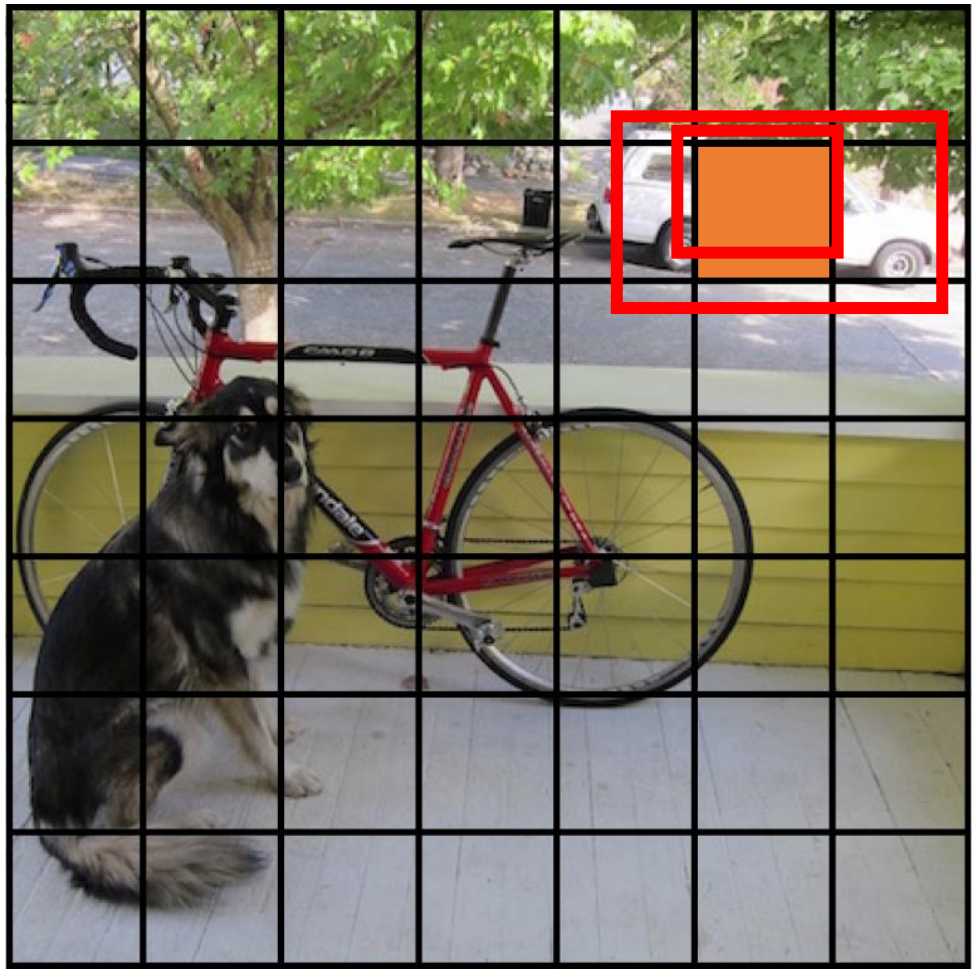


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Inference

Each cell predicts

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

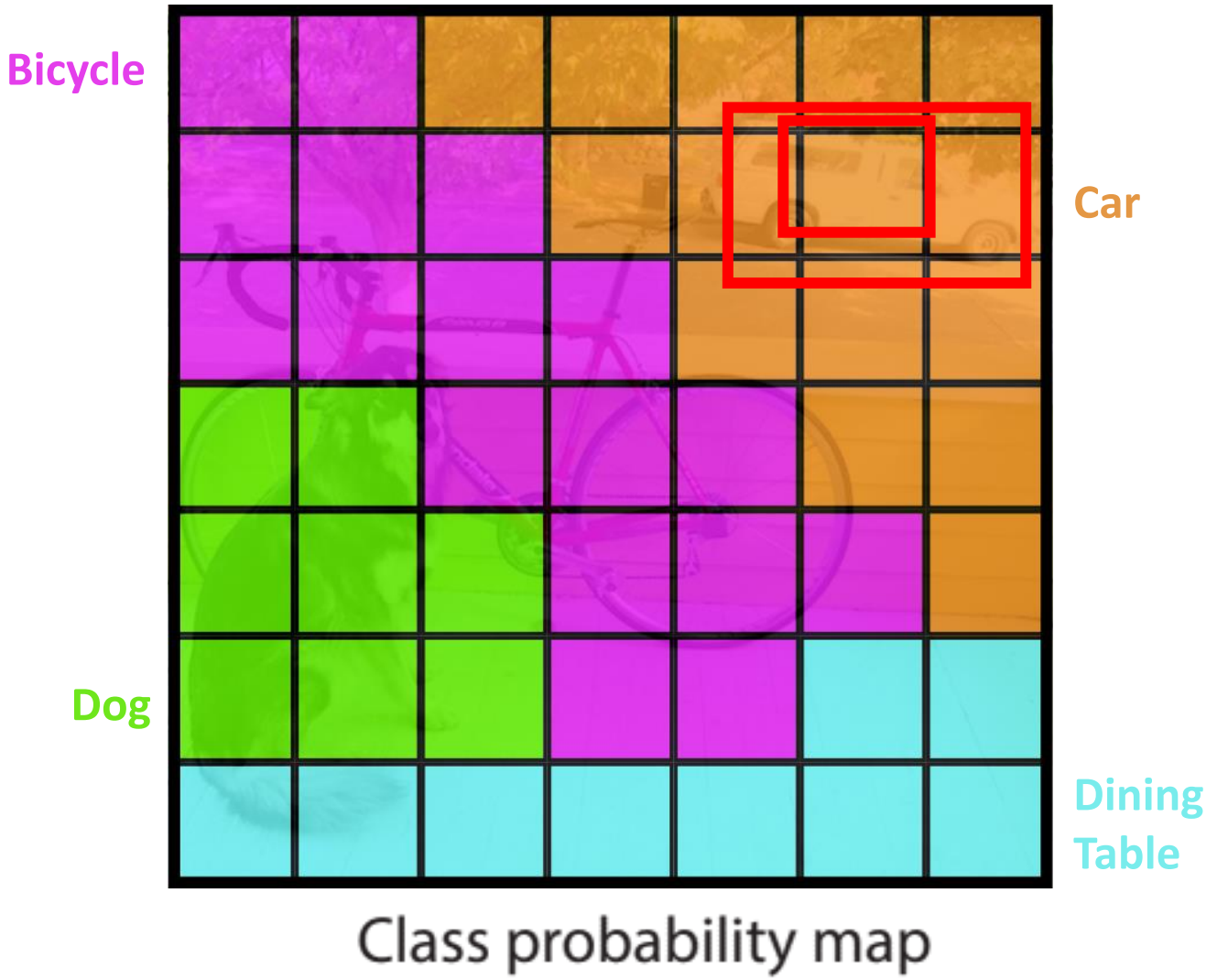
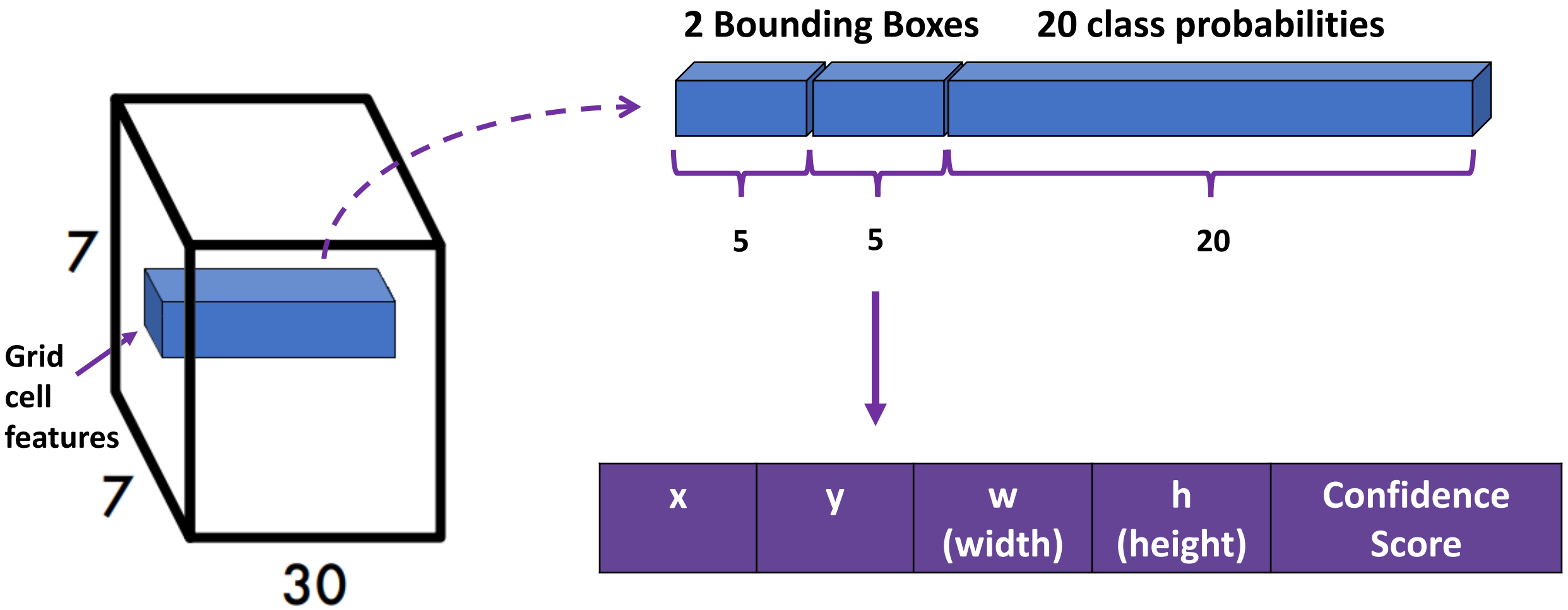


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Output feature map



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
- $C = 20$ class probabilities

SxSxB Bounding-Boxes (S=7,B=2 → 96 Bboxs)

S × S grid on input

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
- $C = 20$ class probabilities

SxSxB Bounding-Boxes (S=7,B=2 → 96 Bboxs)

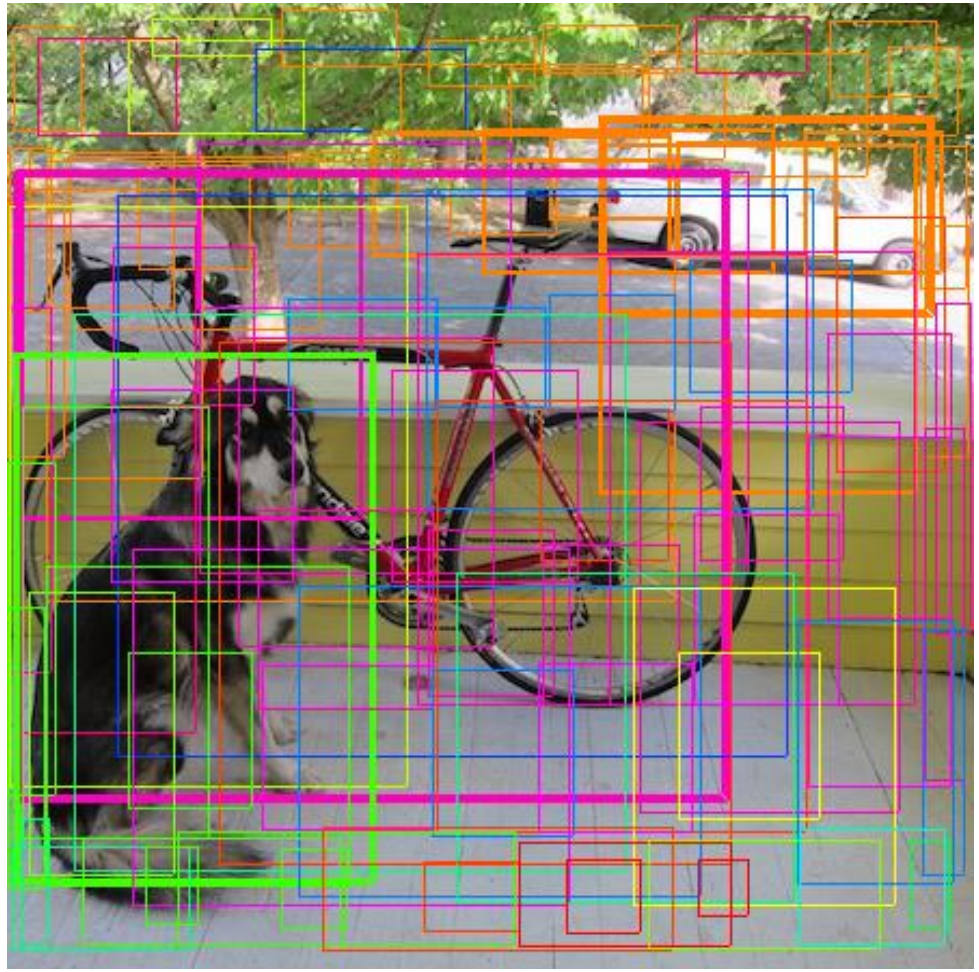


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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
- $C = 20$ class probabilities

- Apply Non-Maximum Suppression

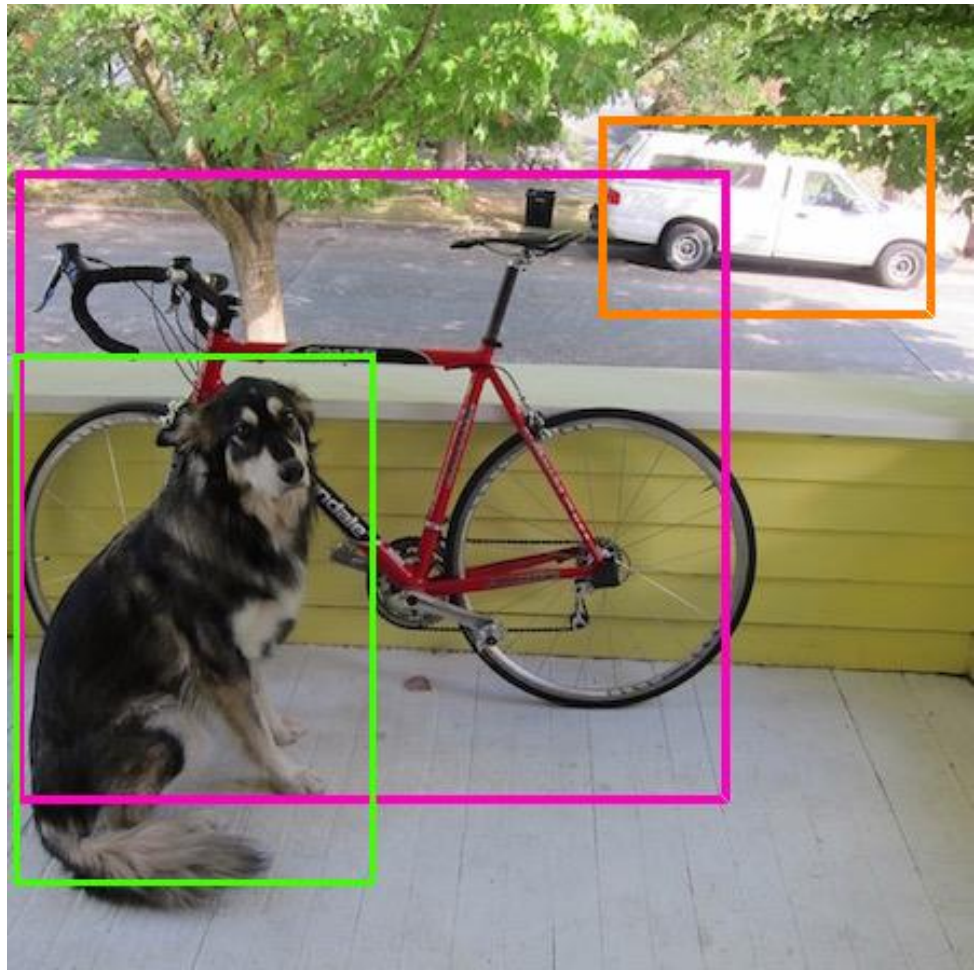
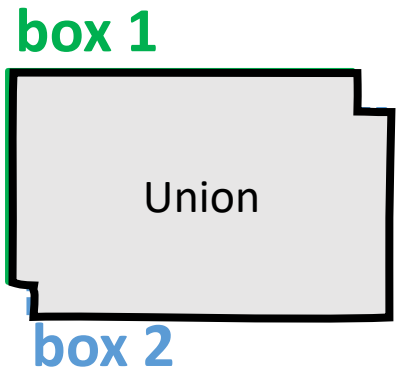


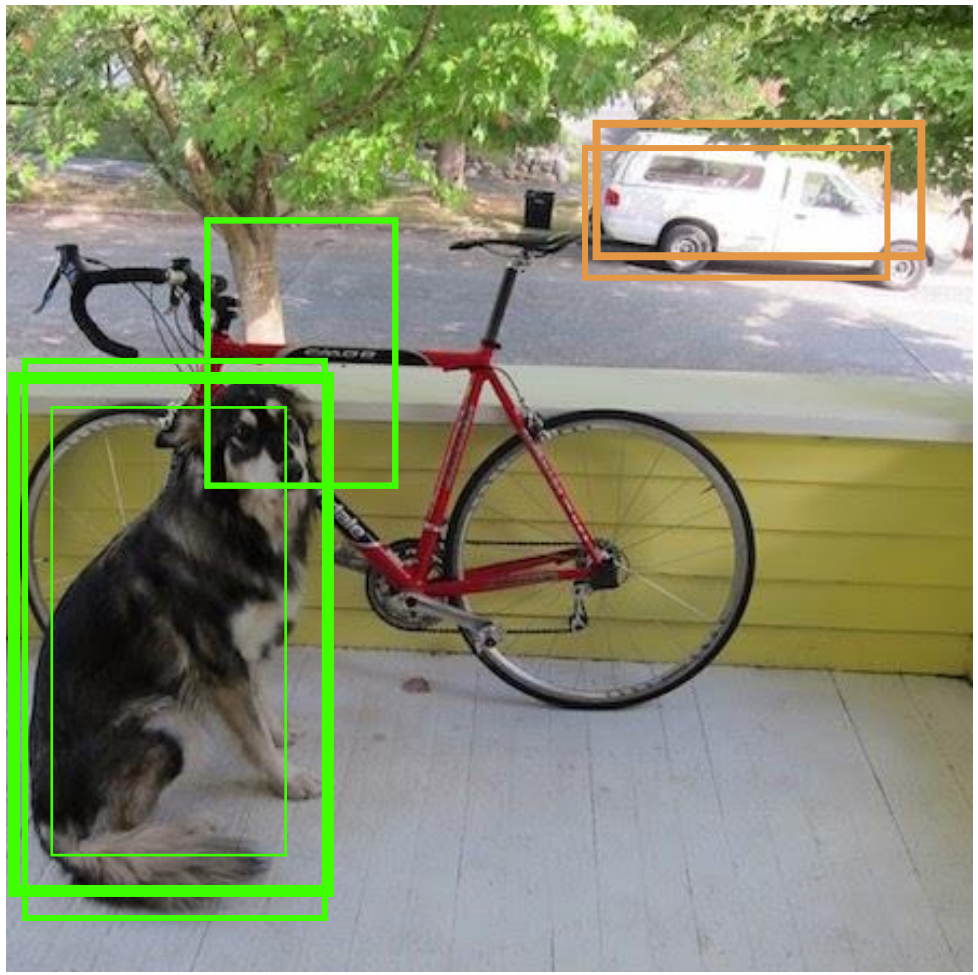
Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO - NMS



IoU = $\frac{\text{intersection}}{\text{union}}$

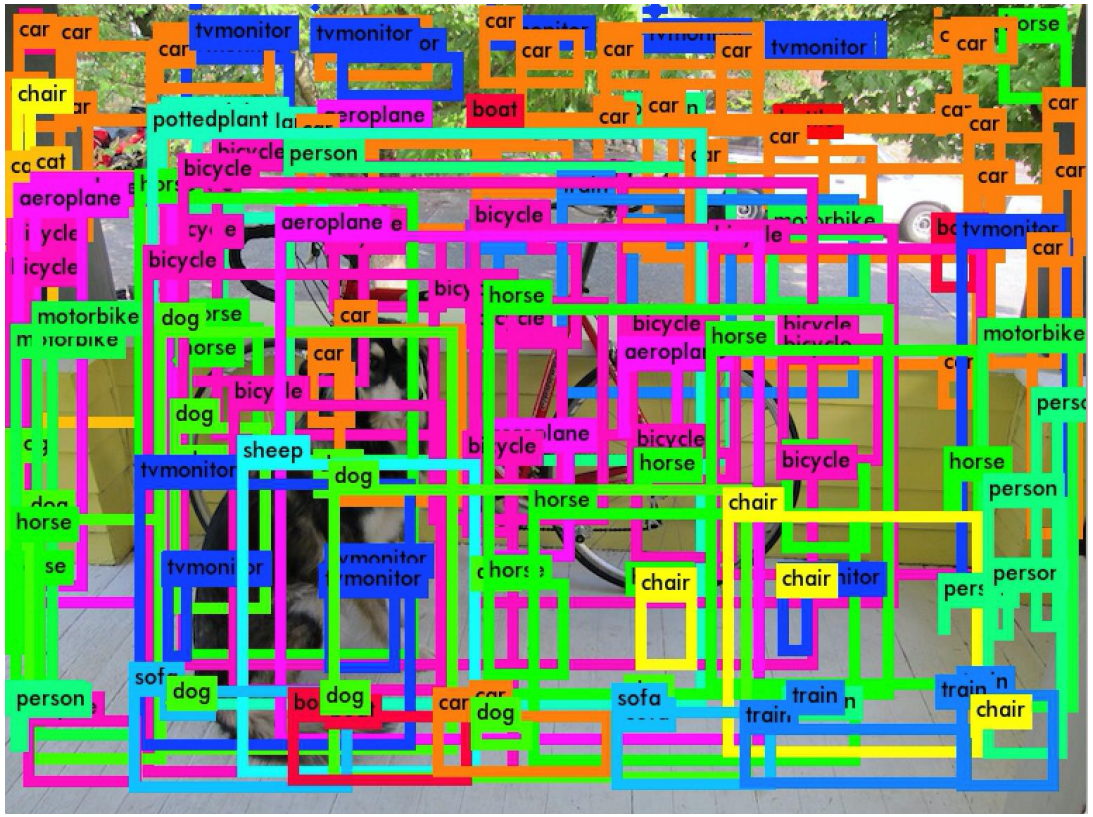


Slide idea: Shai Bagon

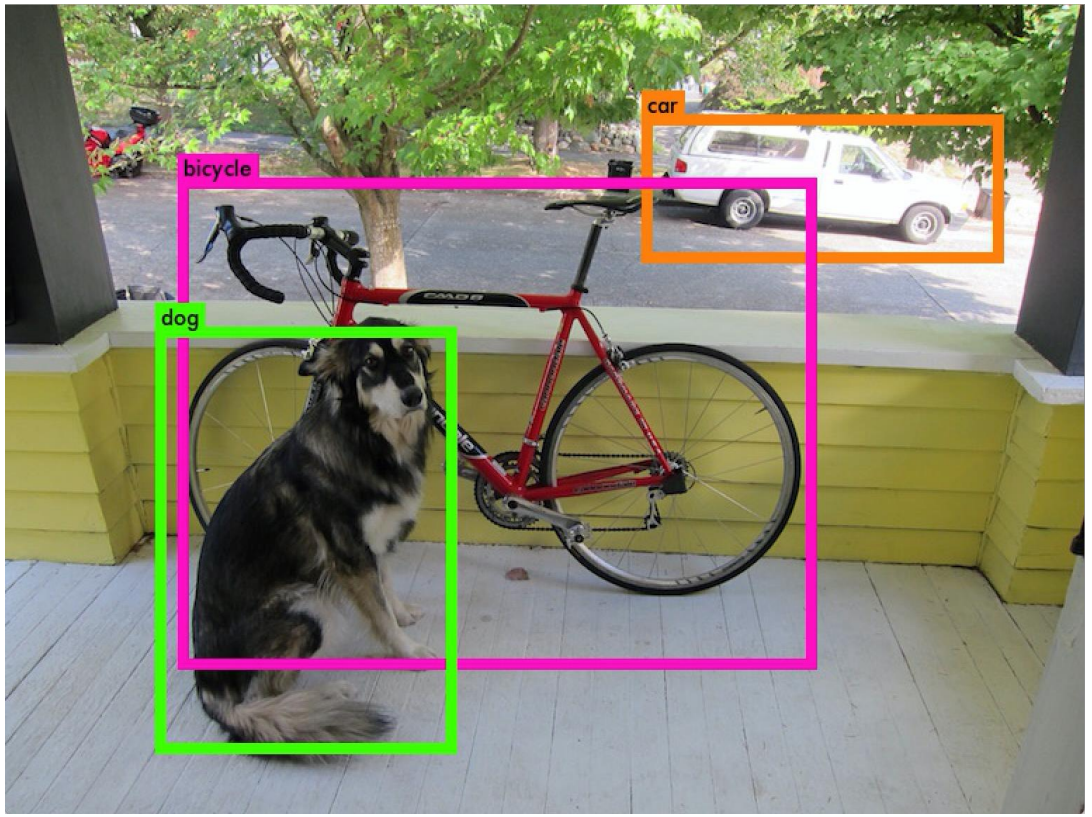
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection, 2015 (CVPR 2016)

YOLO – Role of Threshold

Low Threshold



High Threshold



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

YOLO – Training (end-to-end)

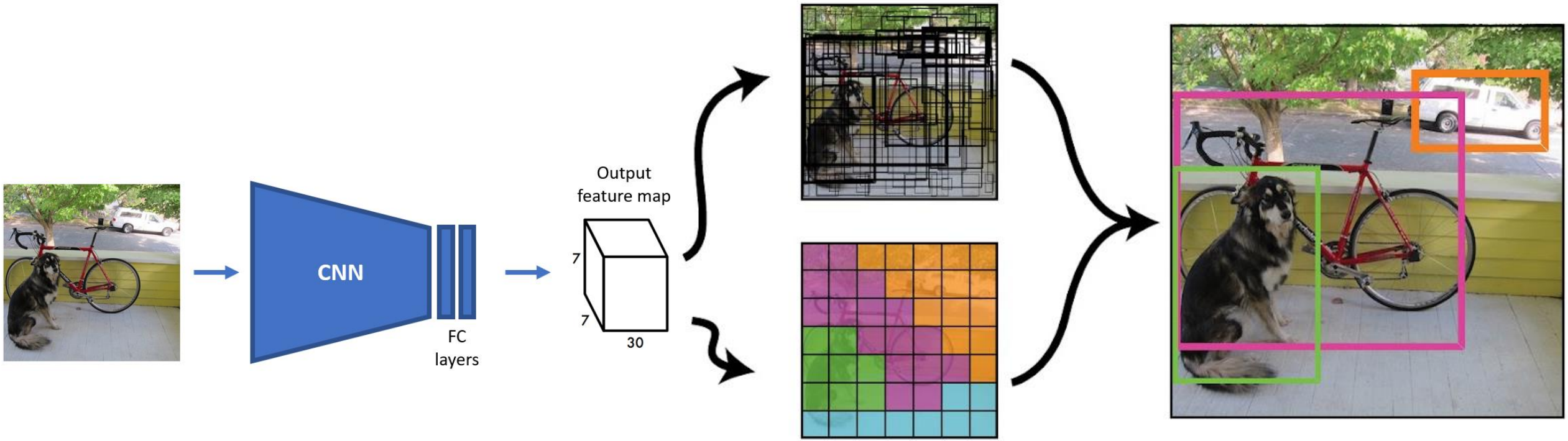


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

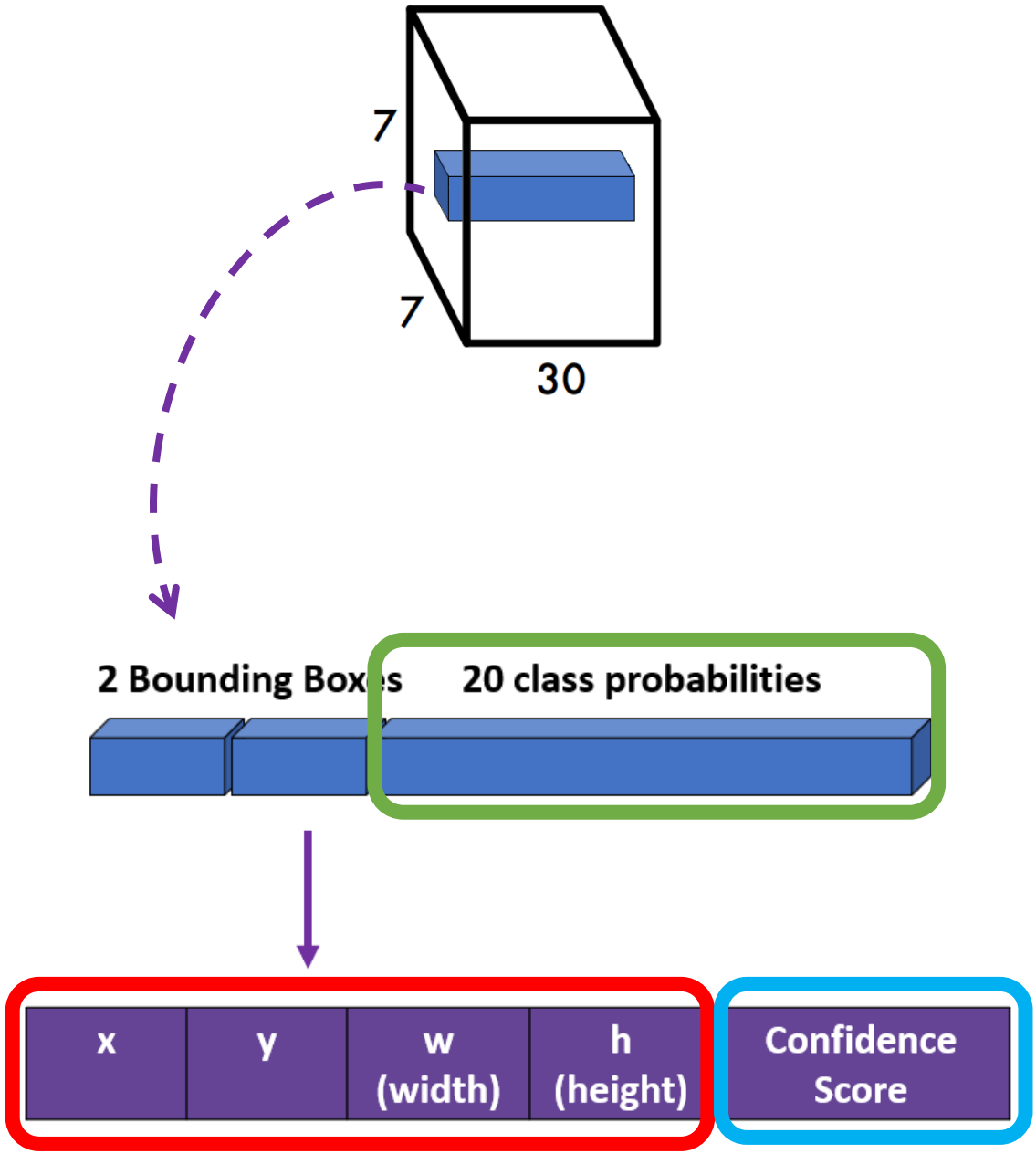
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection, 2015 (CVPR 2016)

YOLO – Loss function

$$\mathcal{L} = \mathcal{L}_{Localization\ Loss}$$

$$+ \mathcal{L}_{Confidence\ Loss}$$

$$+ \mathcal{L}_{Classification\ Loss}$$



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

YOLO – Training

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

Ground truth bounding box

Center of object

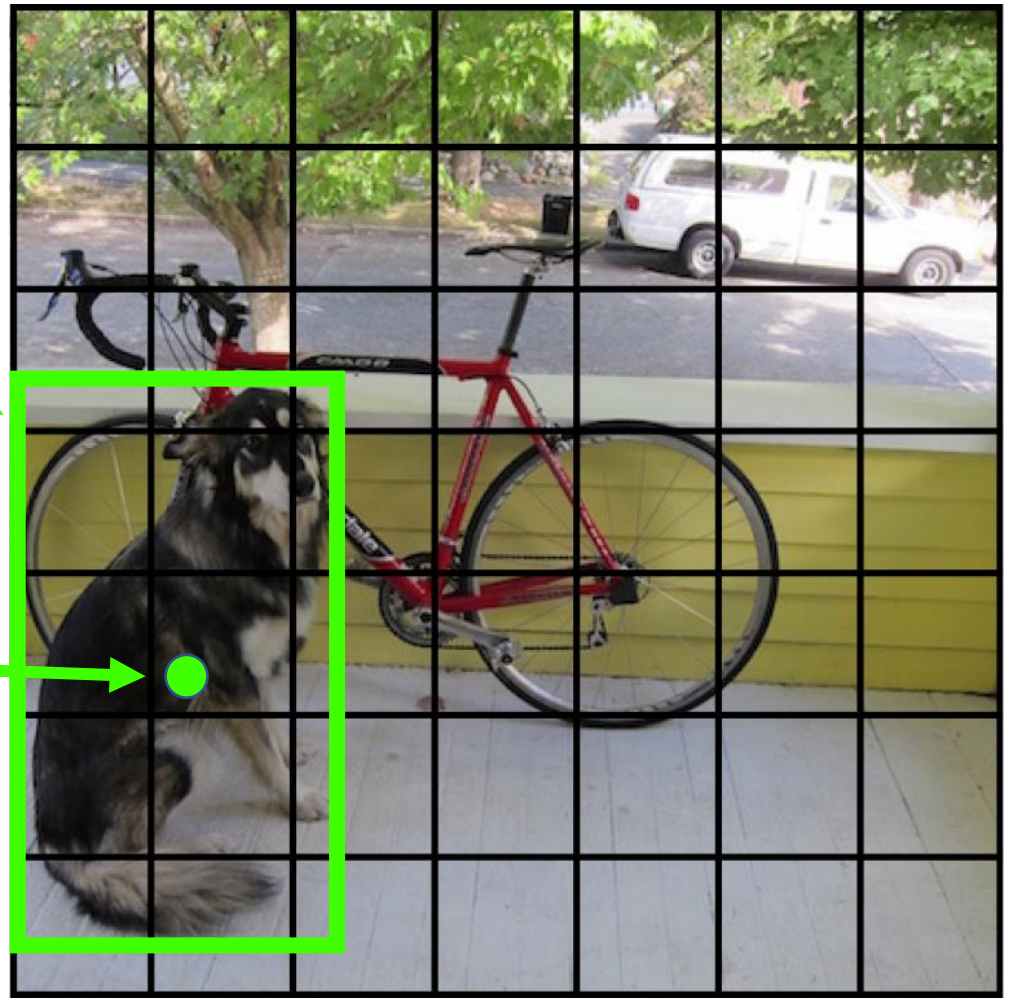


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Training

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

Ground truth bounding box

Assign to specific cell

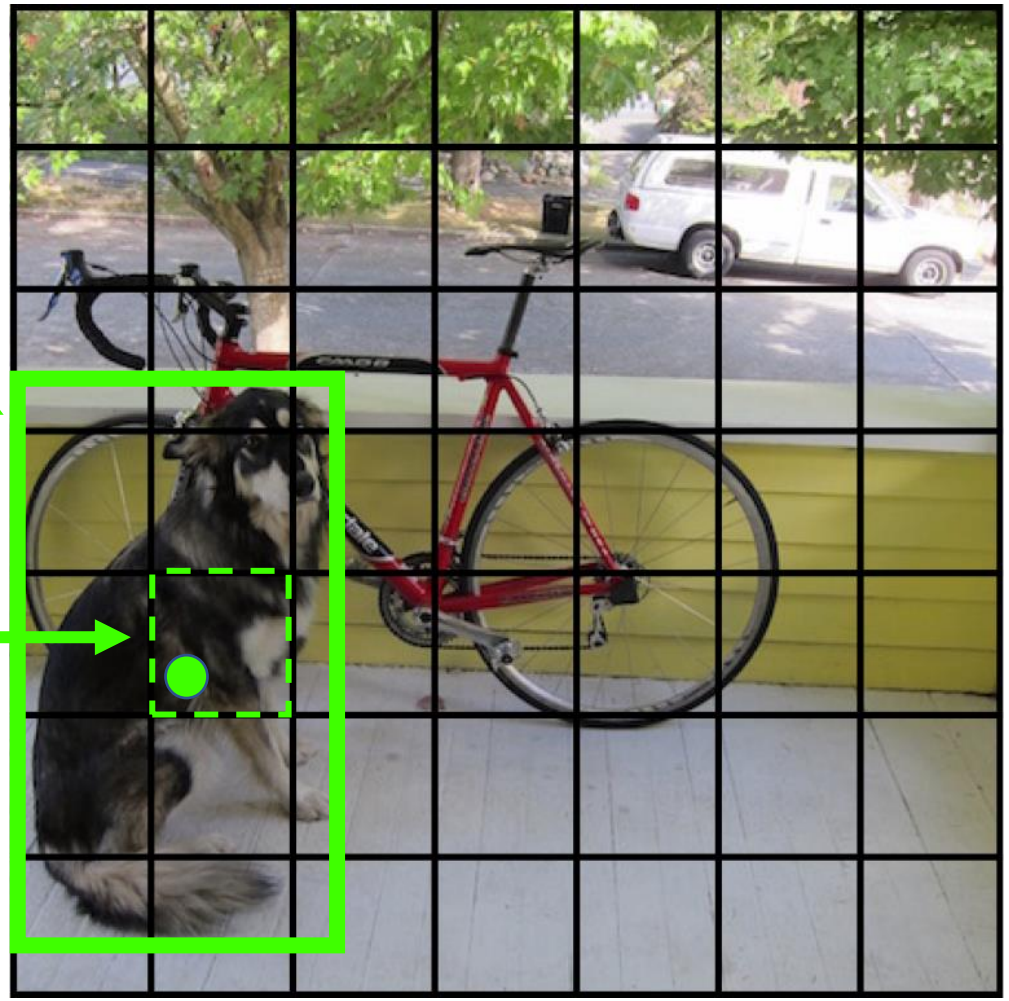


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

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
…

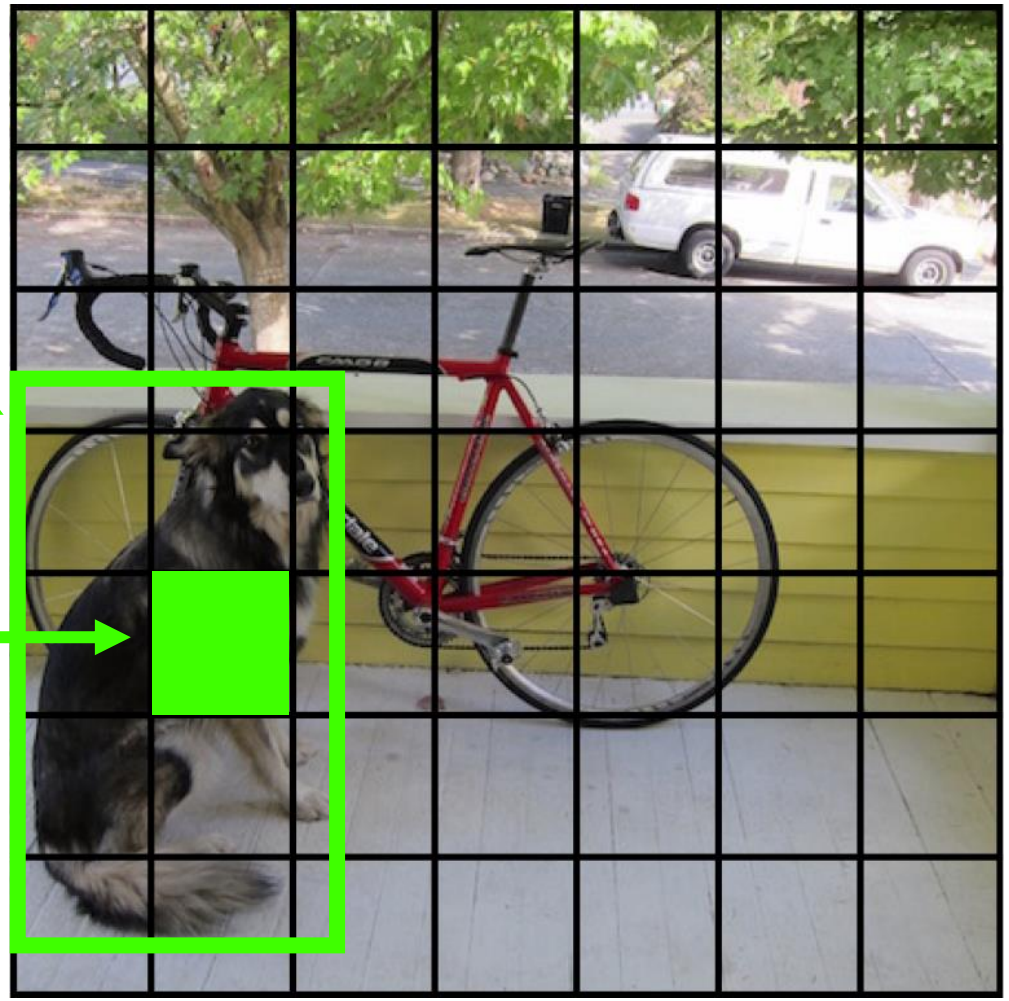


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

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

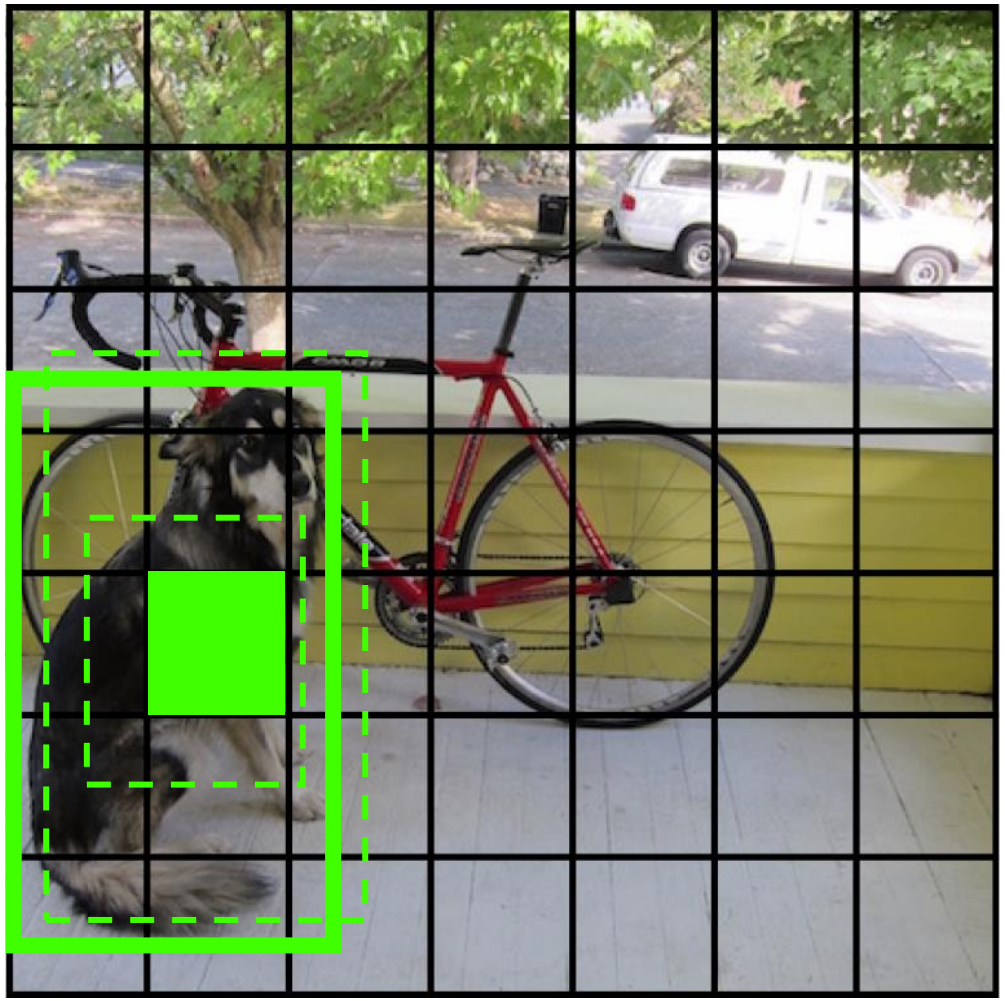


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Training

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

Ground truth bounding box

Increase confidence score

adjust Llocalization

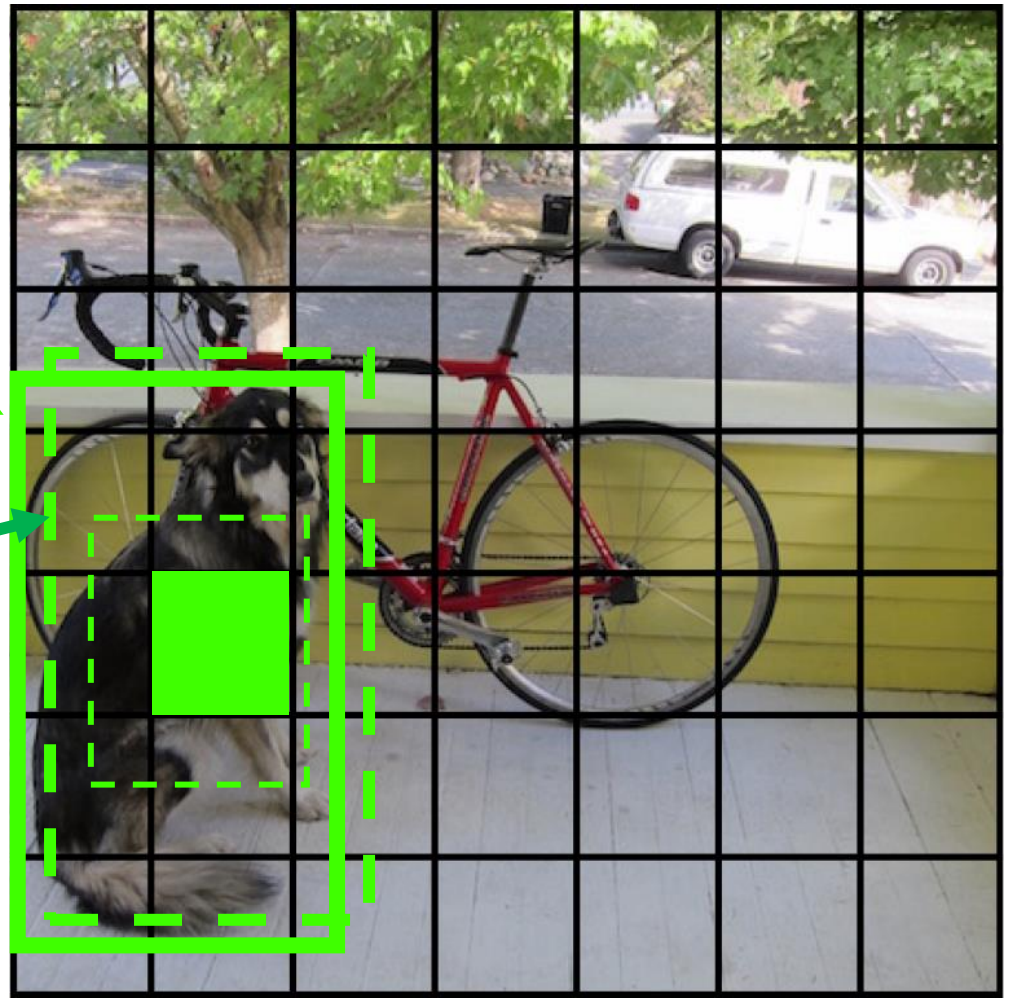


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – 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 $L_{\text{localization}}$

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

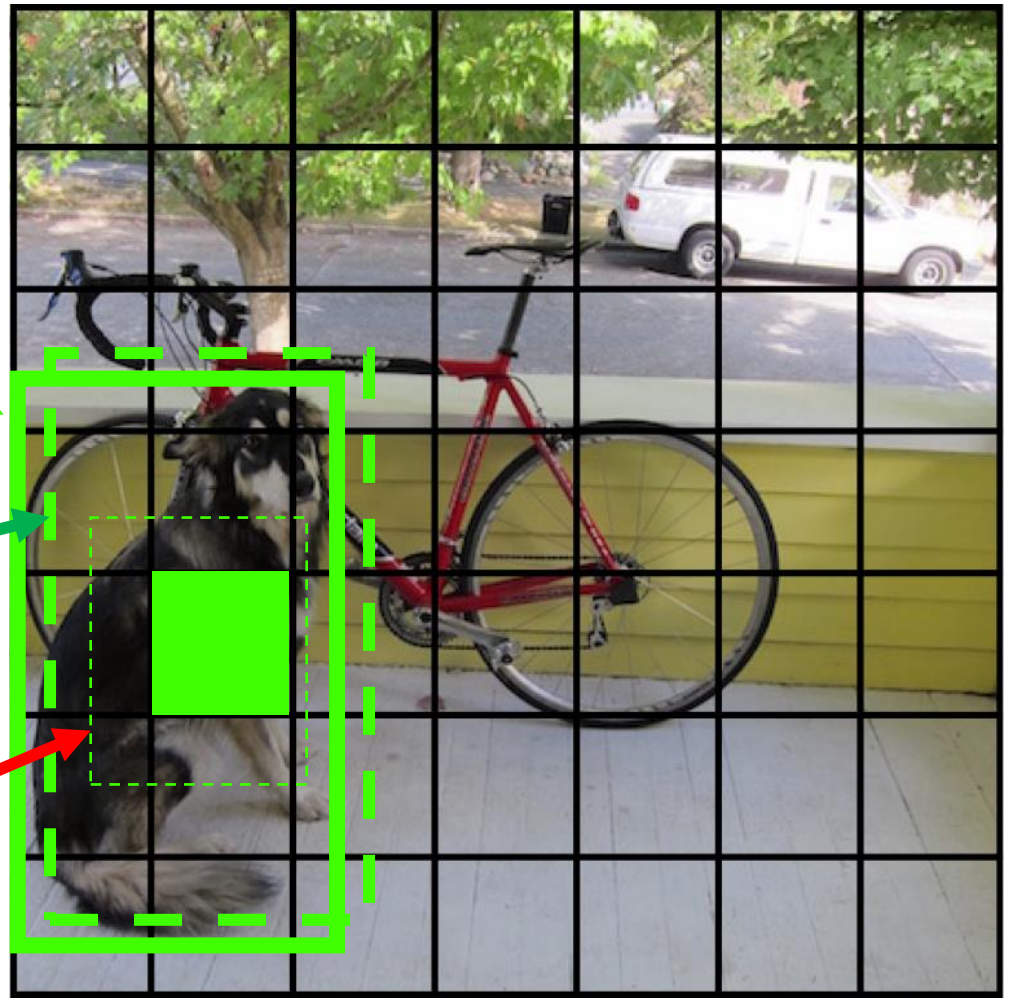


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Training

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

A cell with no ground truth detection

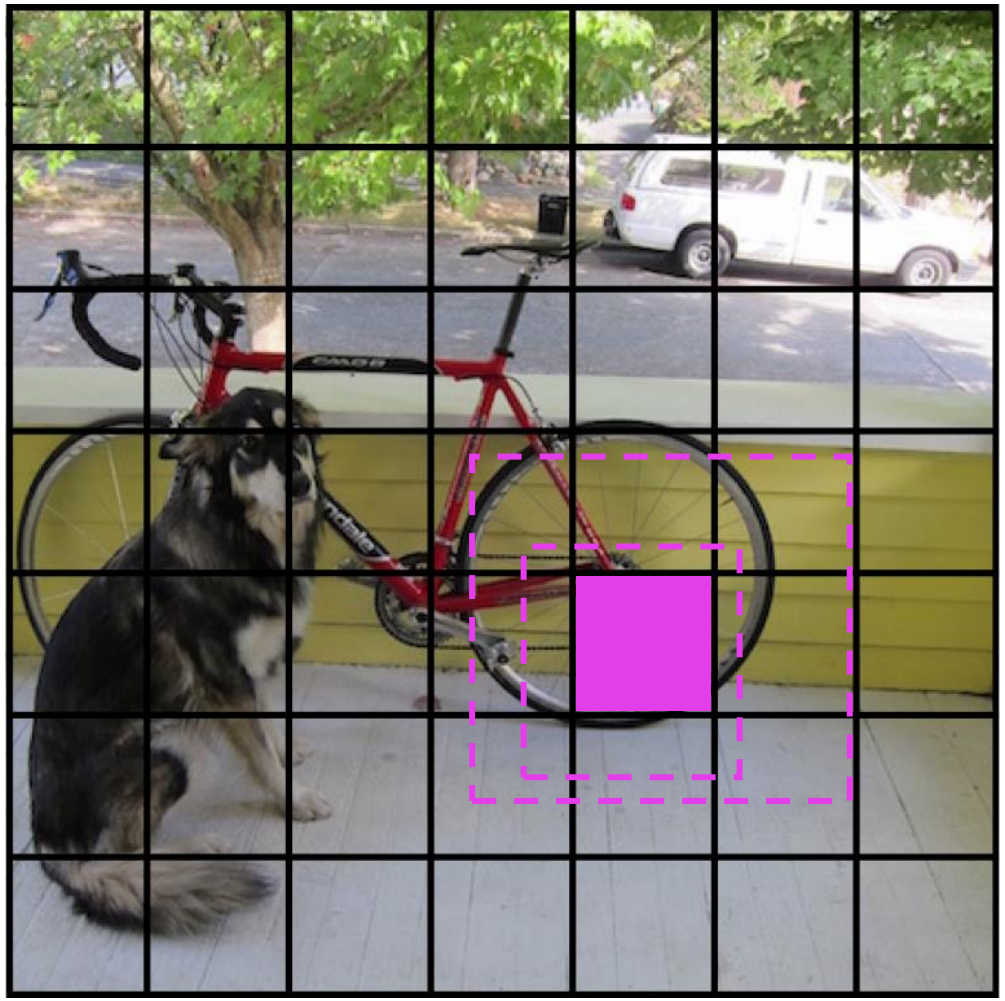


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Training

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

A cell with no ground truth detection

Decrease confidence score

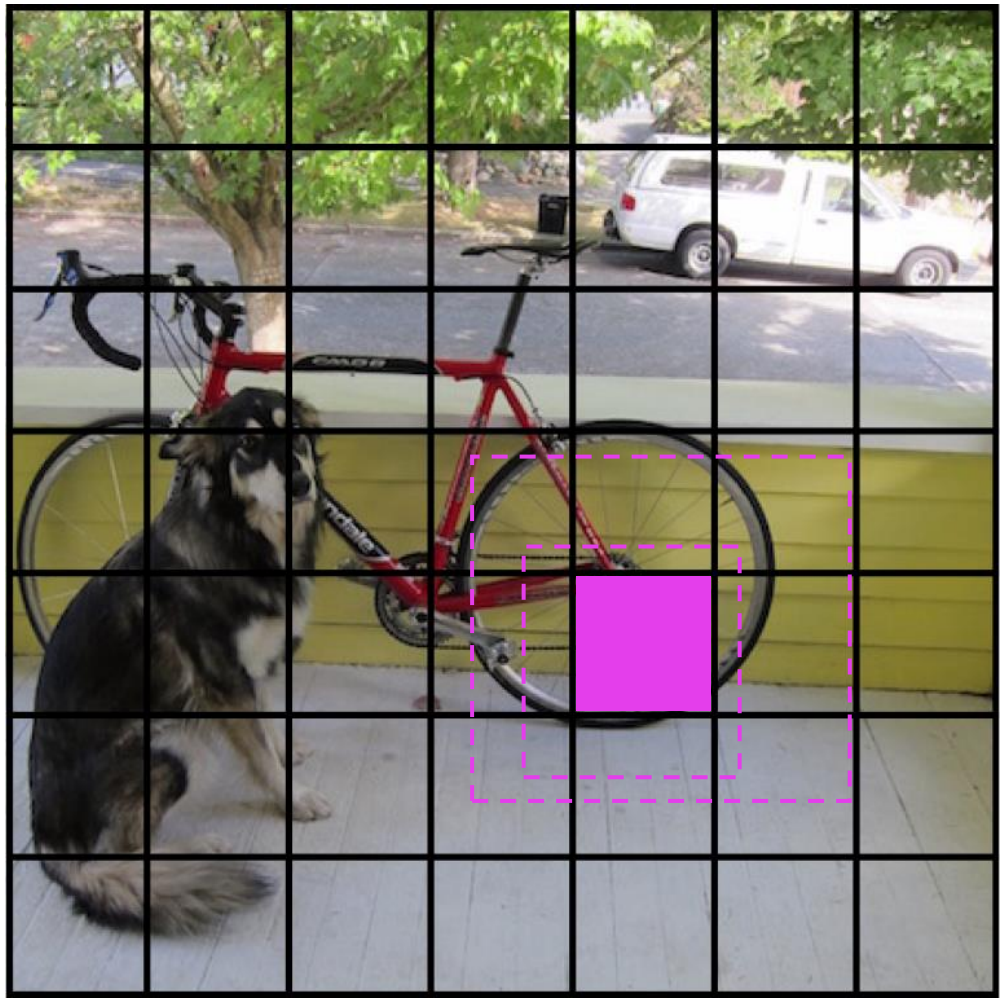


Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Training

$$\mathcal{L} = \mathcal{L}_{Localization\ Loss}$$

$$+ \mathcal{L}_{Confidence\ Loss}$$

$$+ \mathcal{L}_{Classification\ Loss}$$

A cell with no ground truth detection

Decrease confidence score

Don't adjust

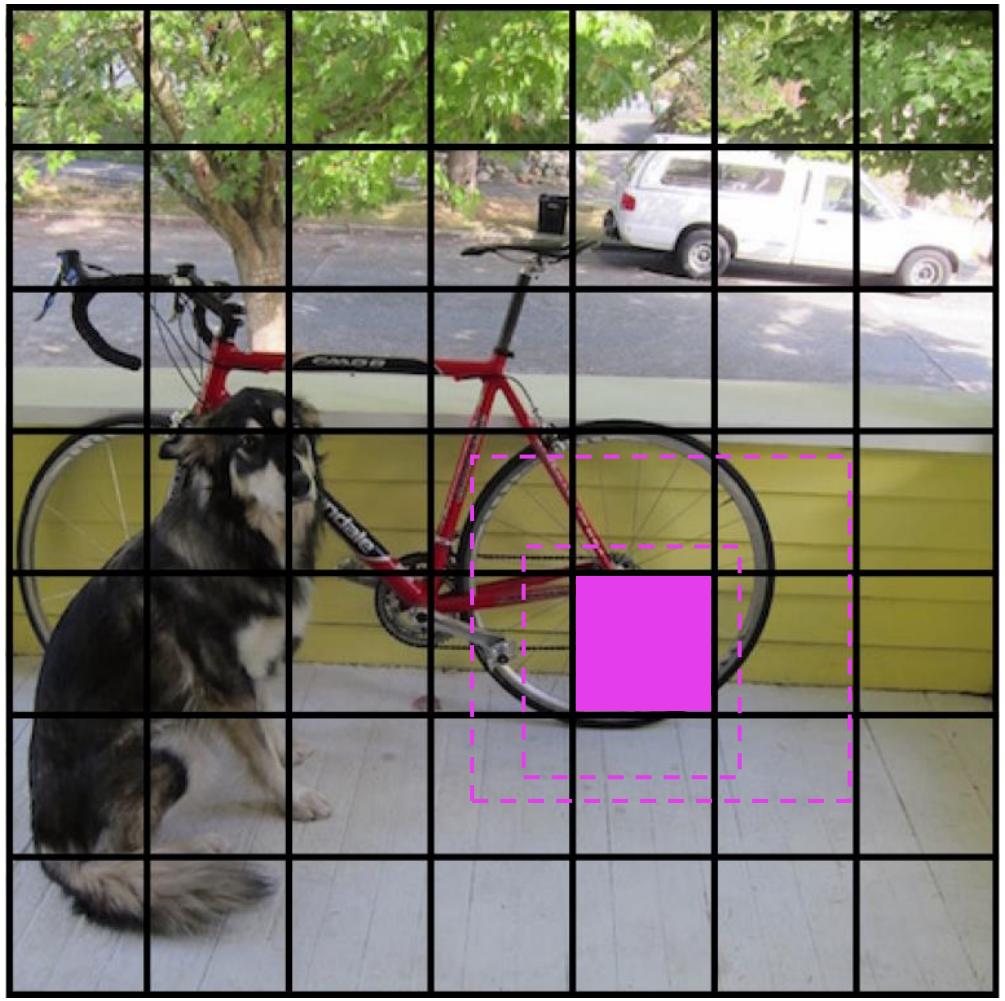
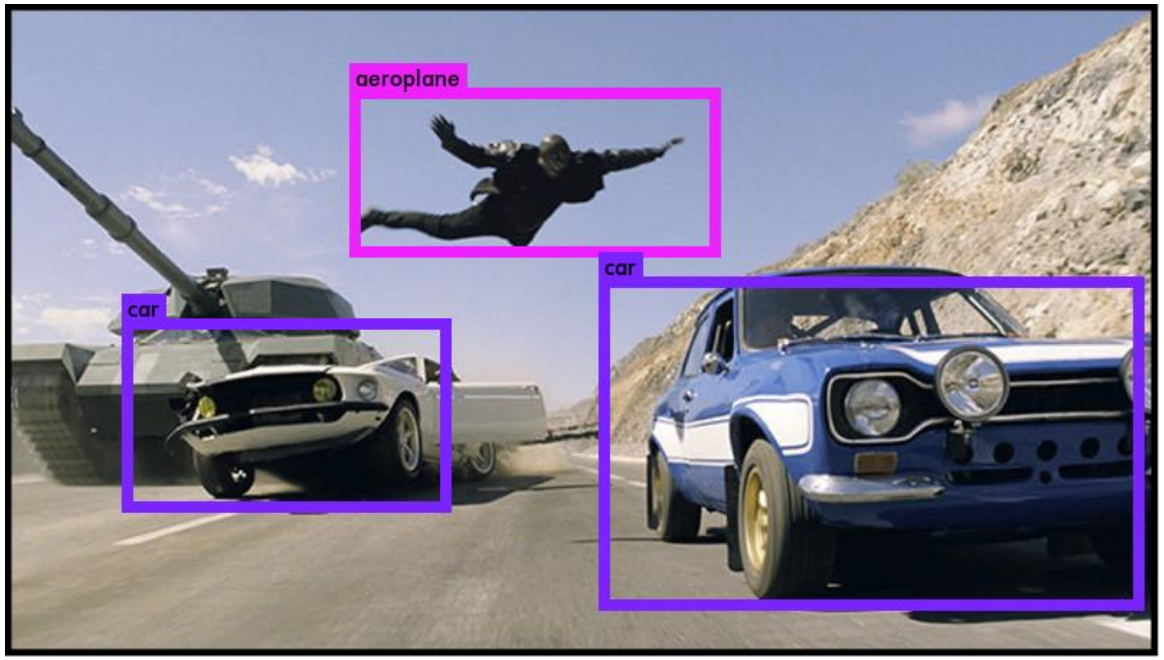


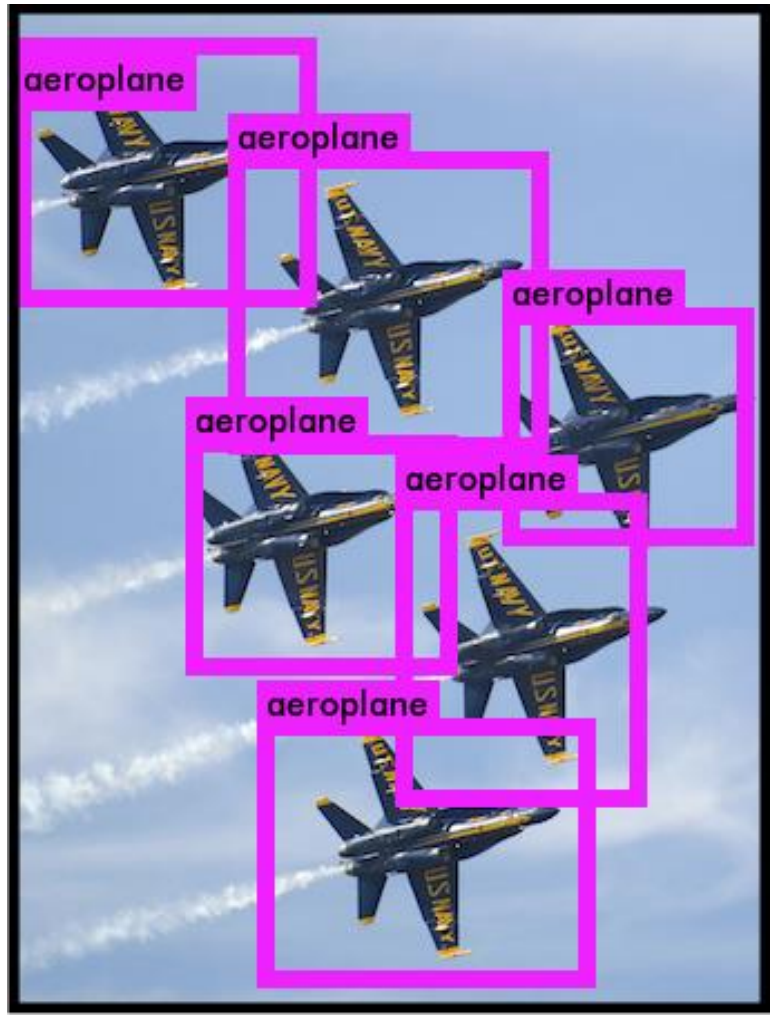
Image credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Benefits

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





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

YOLO – Limitations

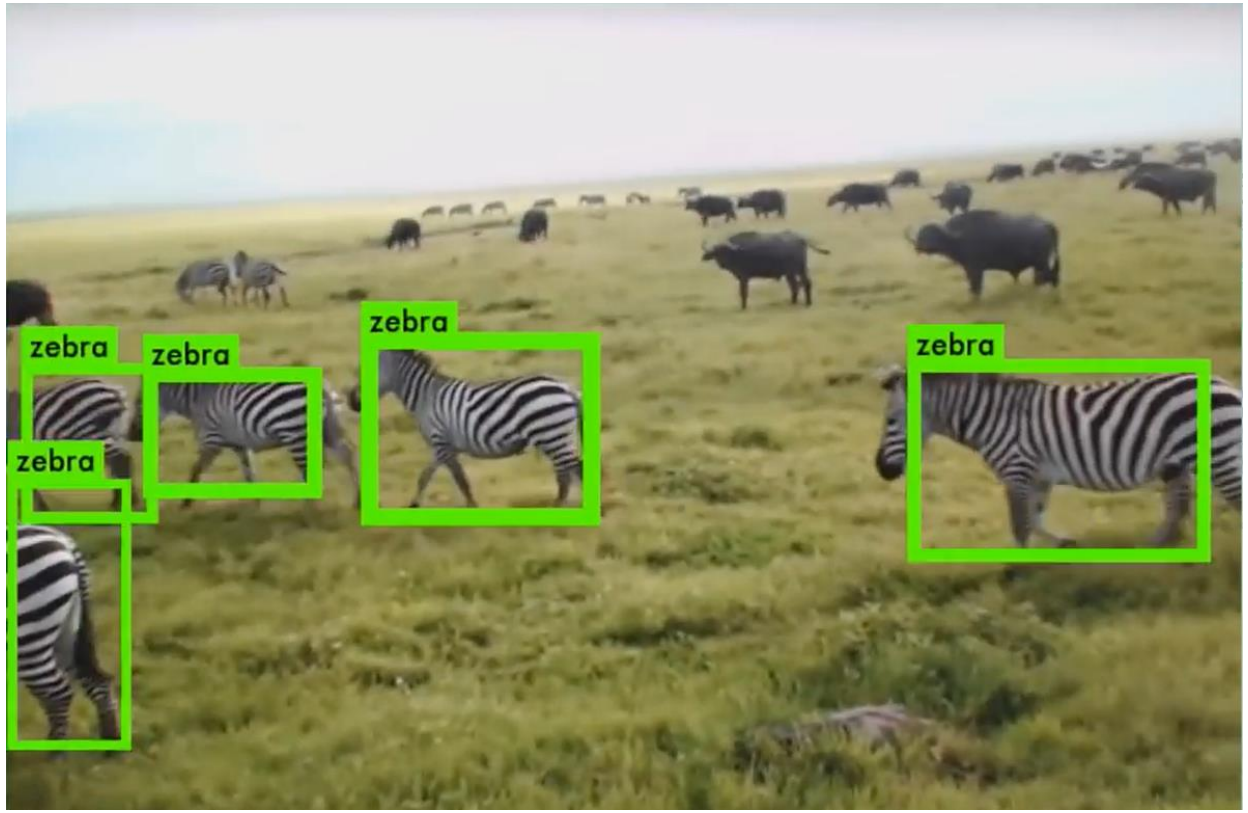


Image credit: https://pjreddie.com/darknet/yolov1/

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

YOLO – Limitations

- Difficult to detect small objects
- Coarse predictions

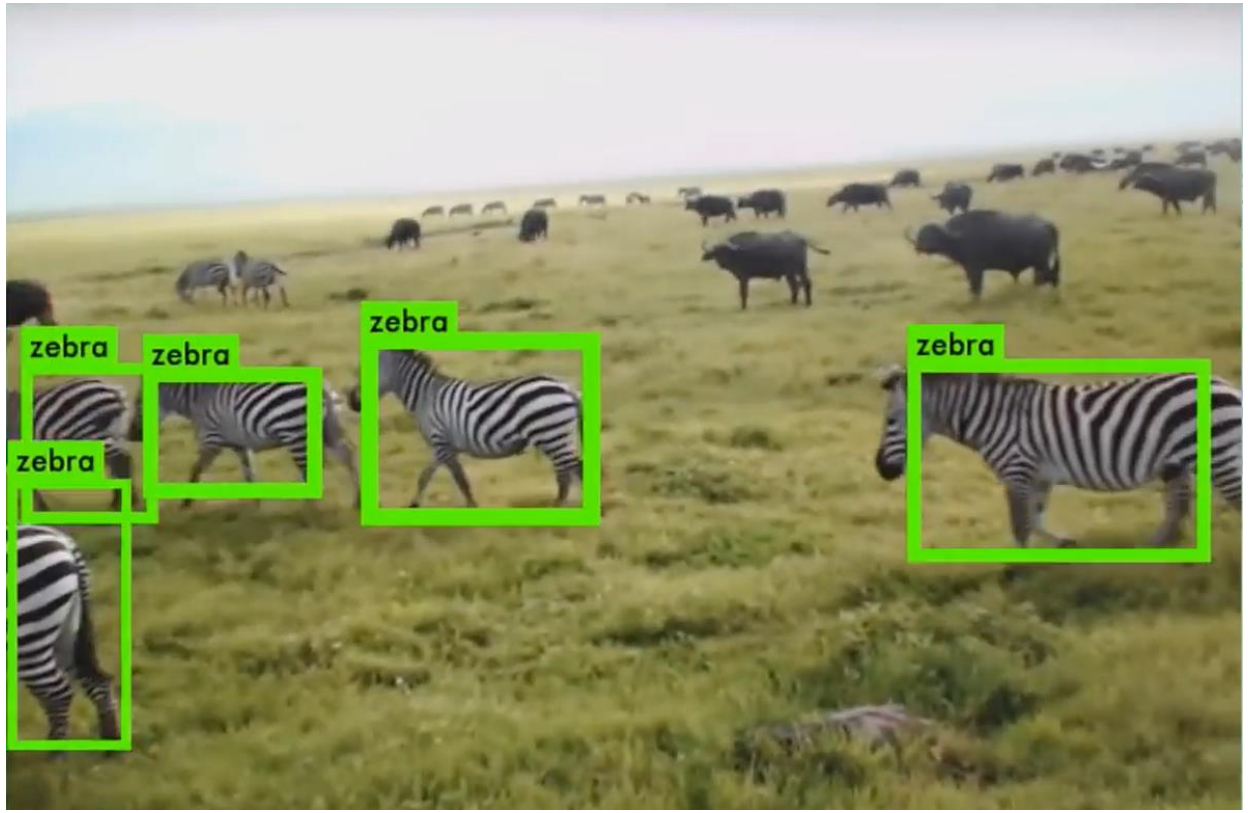
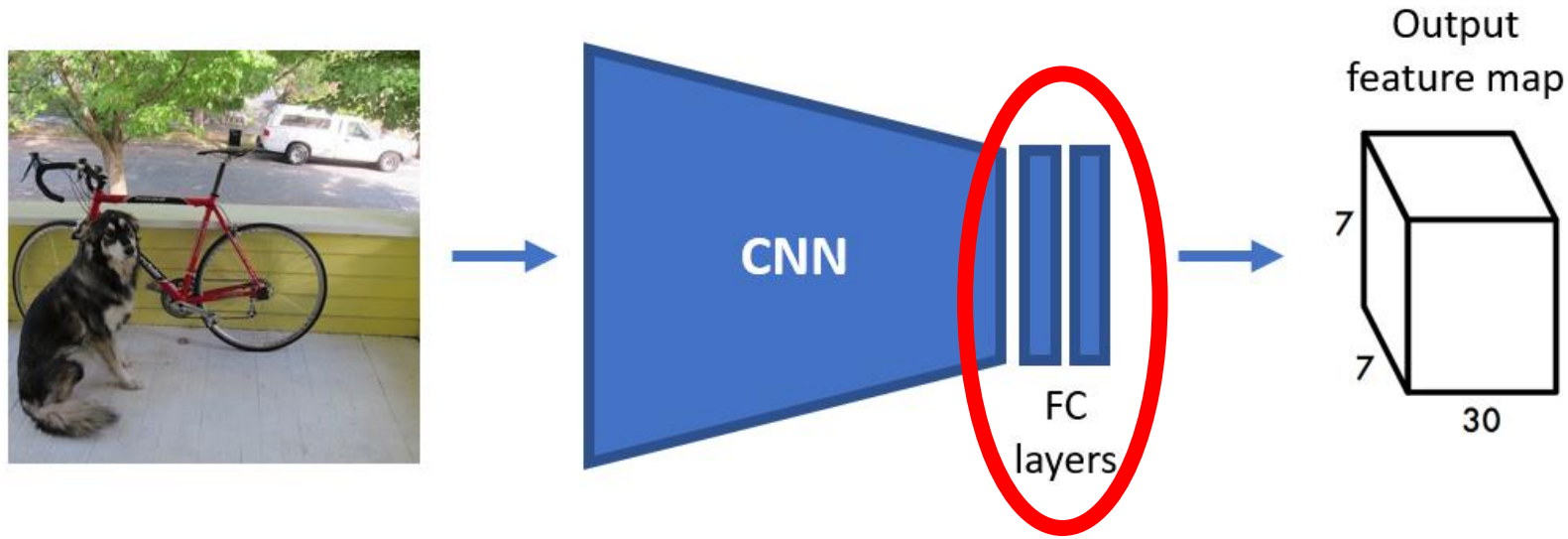


Image credit: https://pjreddie.com/darknet/yolov1/

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

YOLO – Limitations

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



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

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)

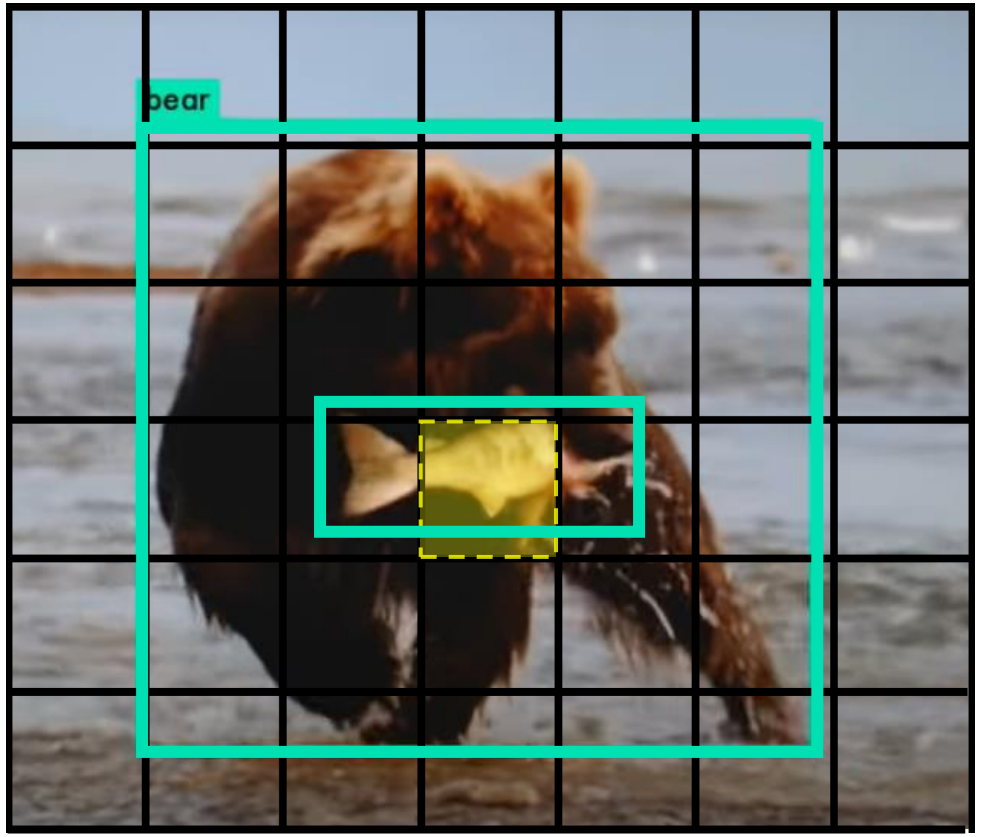


Image credit: https://pjreddie.com/darknet/yolov1/

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

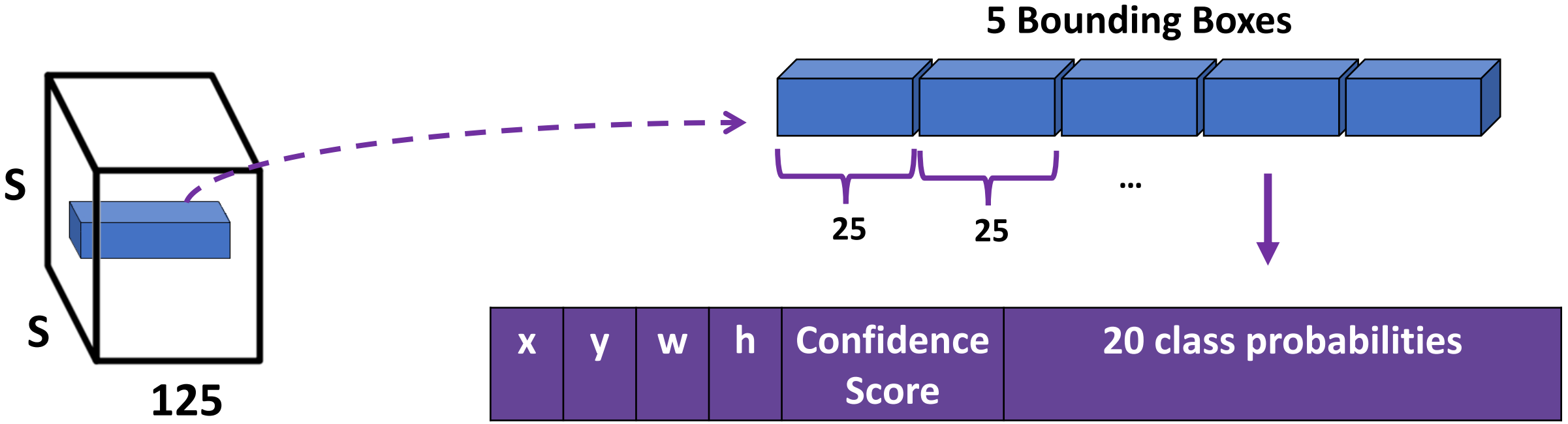
YOLOv2

- Removed fully connected layers

J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger (CVPR 2017)

YOLOv2

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



J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger (CVPR 2017)

YOLOv2

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

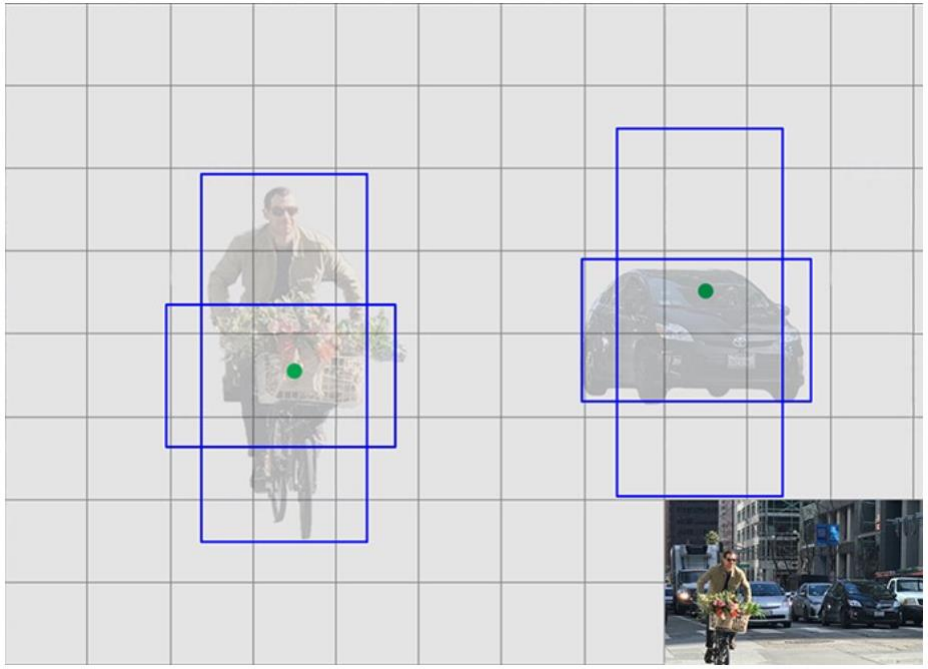
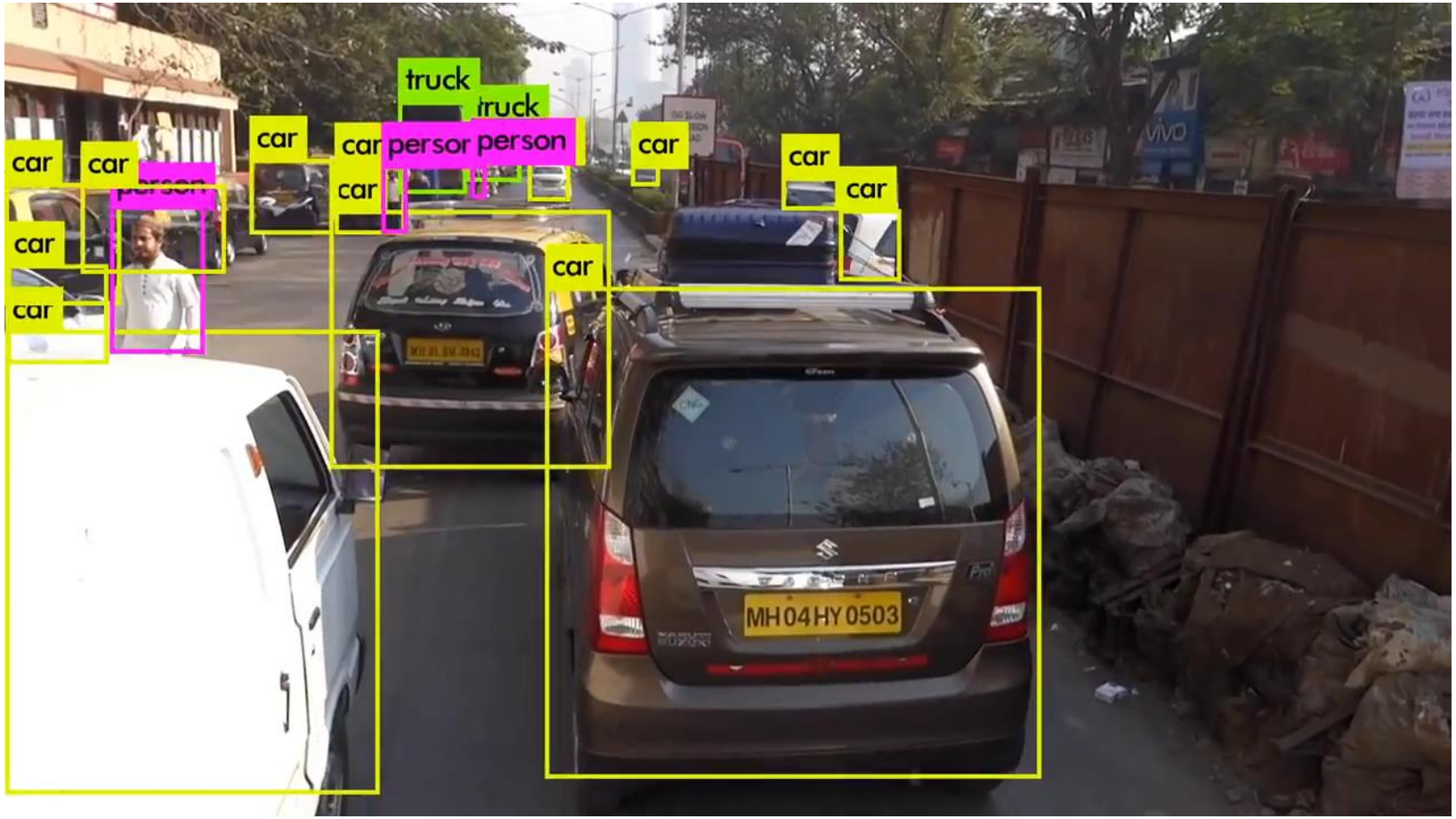


Image credit: medium

J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger (CVPR 2017)

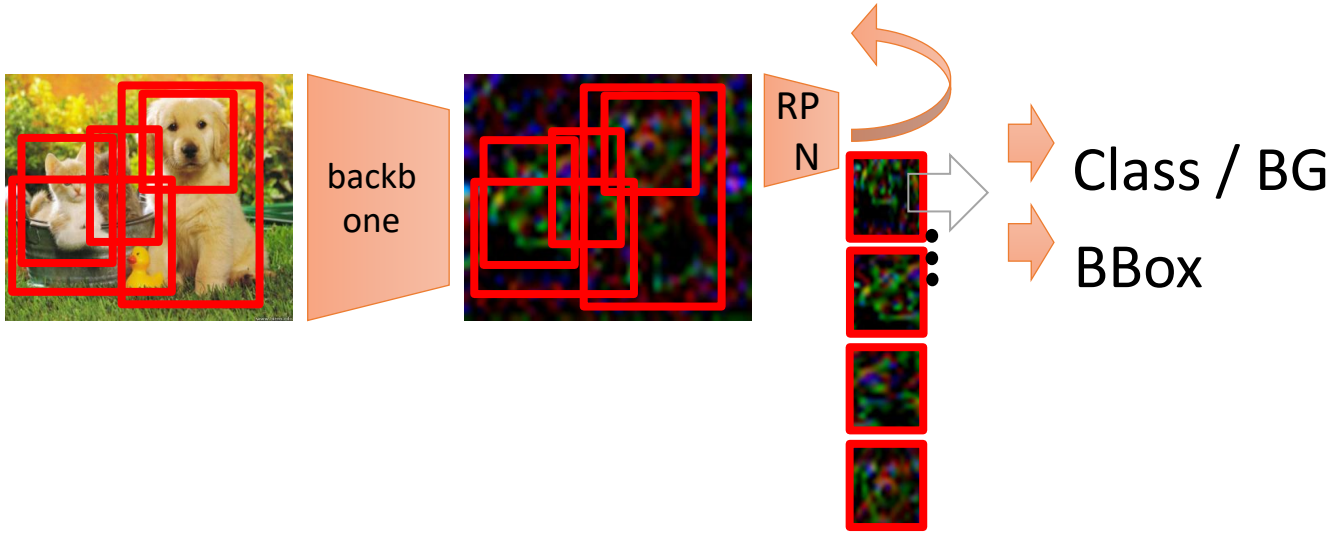
There’s always room for improvement!

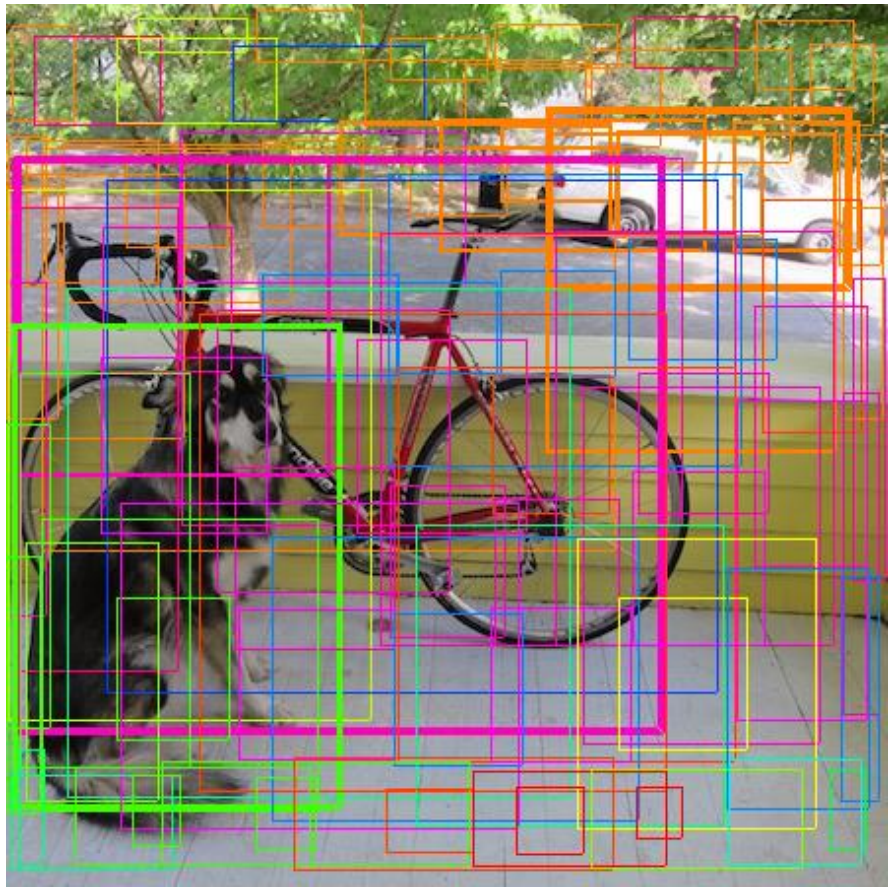
- YOLOv3
  J. Redmon, A. Farhadi. Yolov3: An incremental improvement, 2018
- YOLO v4
  A. Bochkovskiy, C. Wang, H. Liao. Yolov4: Optimal speed and accuracy of object detection (Feb. 2020)
- YOLOv5
  YOLOv5 by ultralytics (June 2020)
- PP-YOLO
  X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. 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)
- …



Video: YOLOv3

Object Detection So Far..

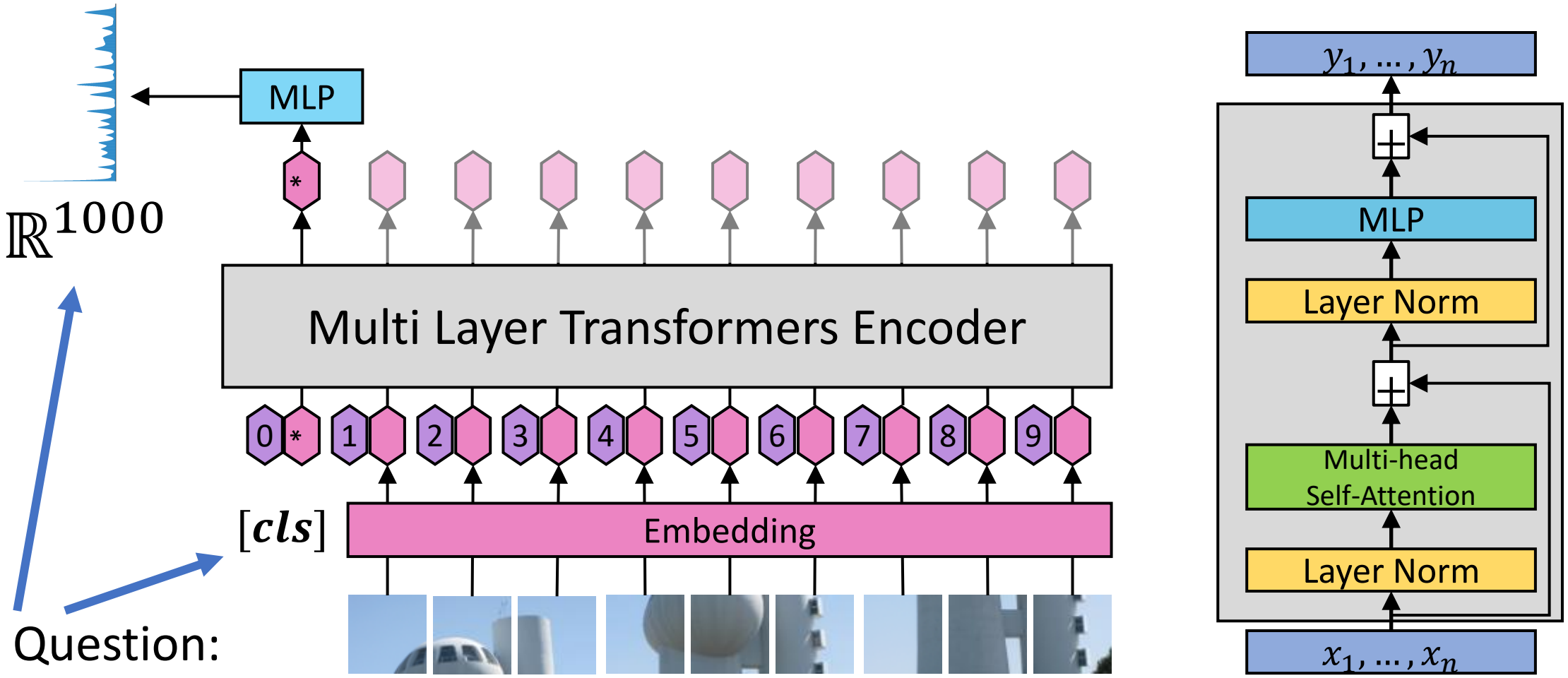




Annoying parts of Detection:
Multiple Bboxs → NMS, Threshold

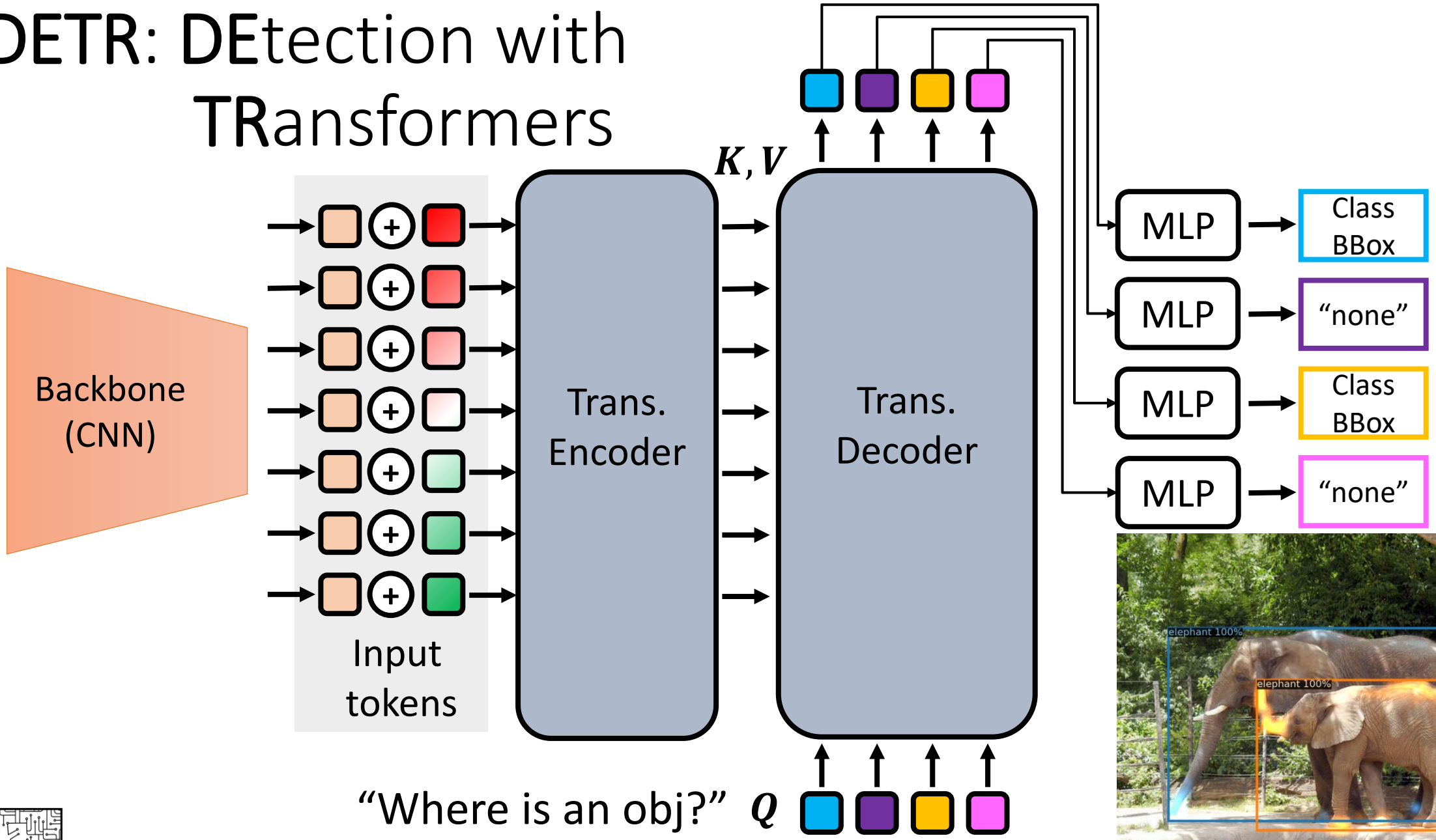
Can we do better? “Set” prediction

Vision Transformers (ViT)



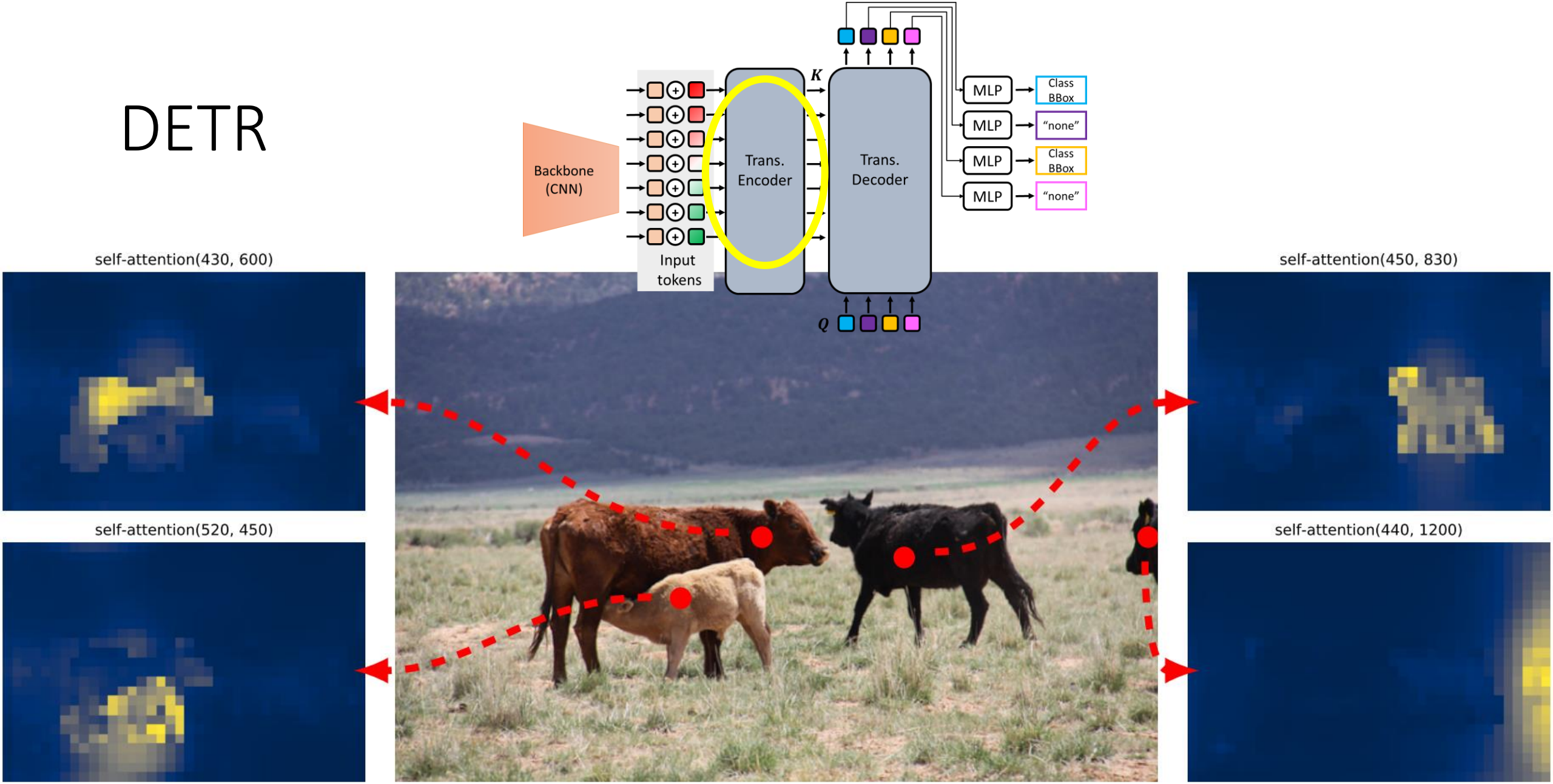
Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J. and Houlsby N. "**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**" (ICLR 2021)

DETR: DEtection with TRansformers



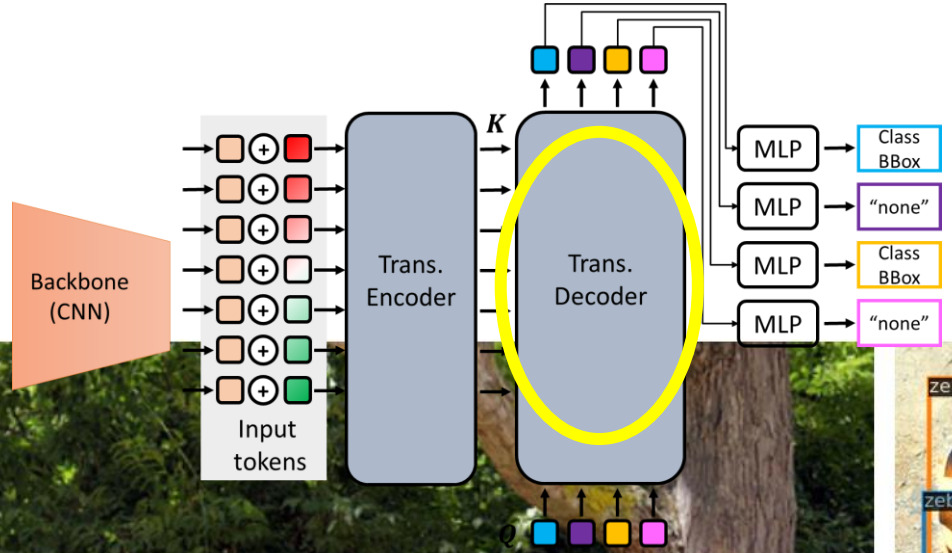
Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. DETR: End-to-End Object Detection with Transformers (ECCV 2020)

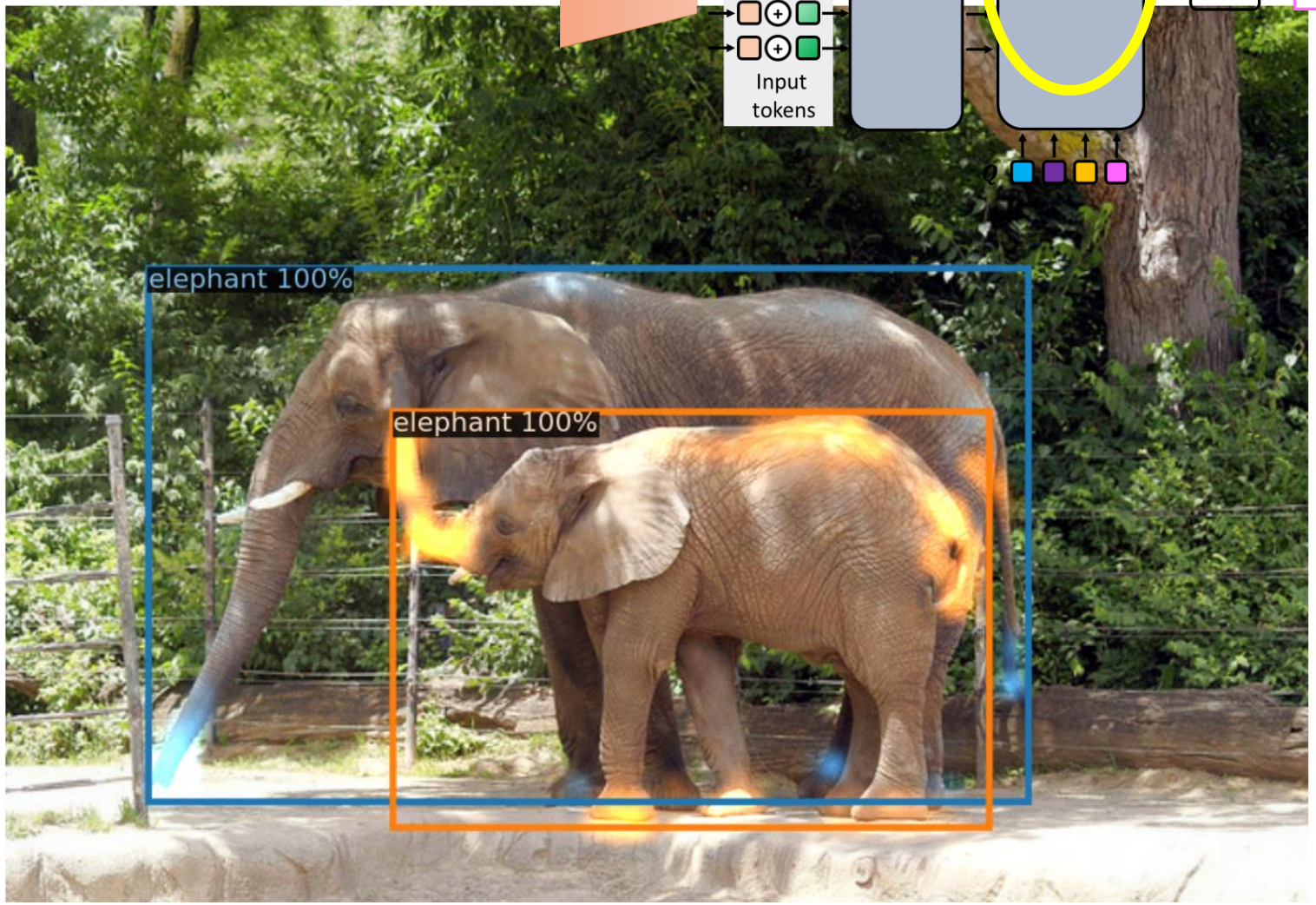
DETR

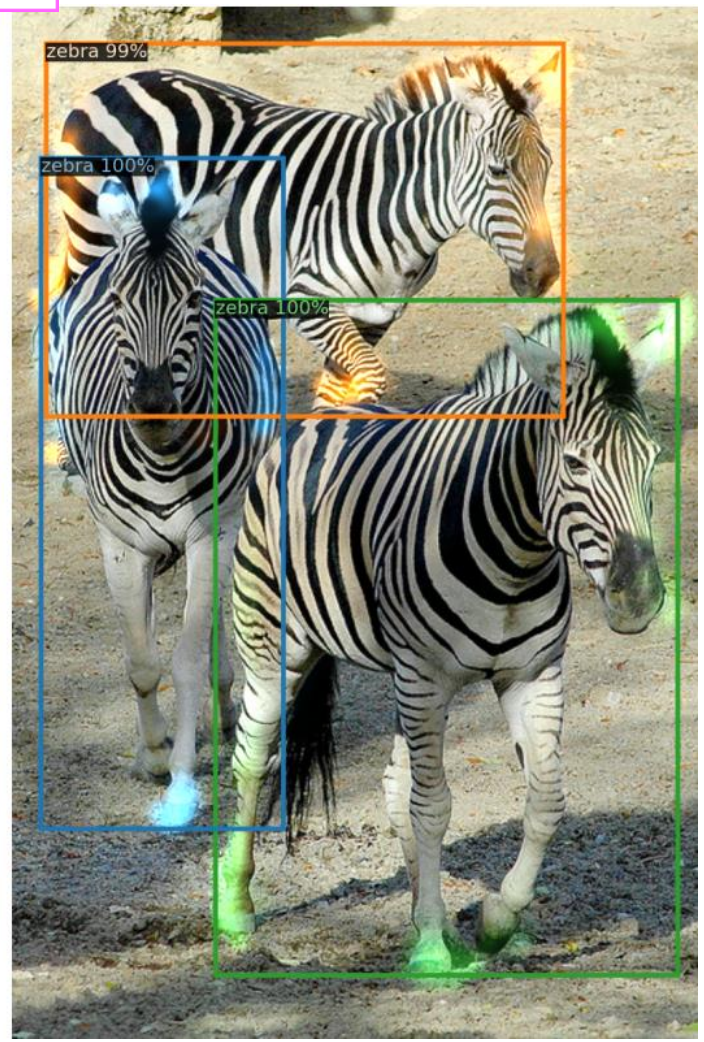


Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. [DETR: End-to-End Object Detection with Transformers](#) (ECCV 2020)

DETR



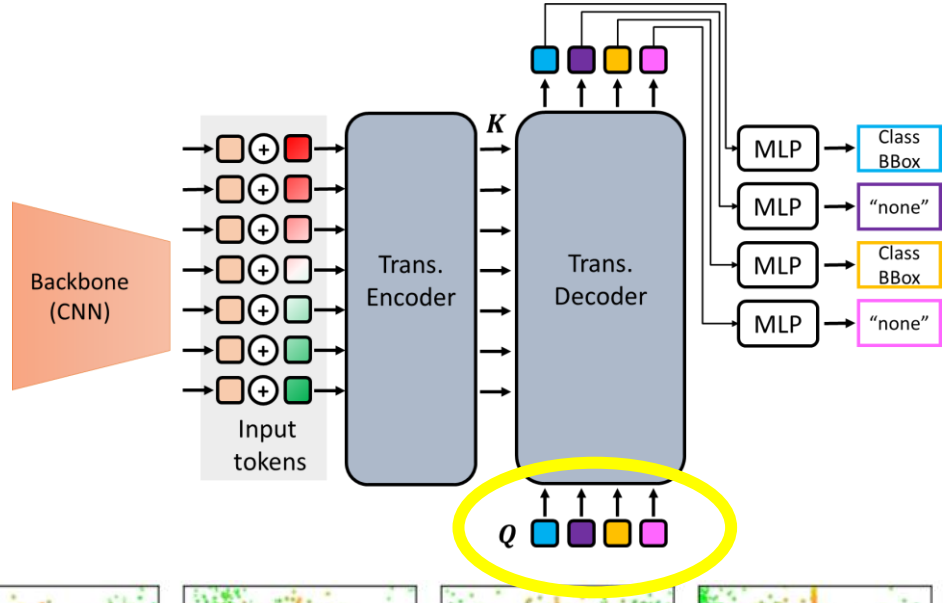


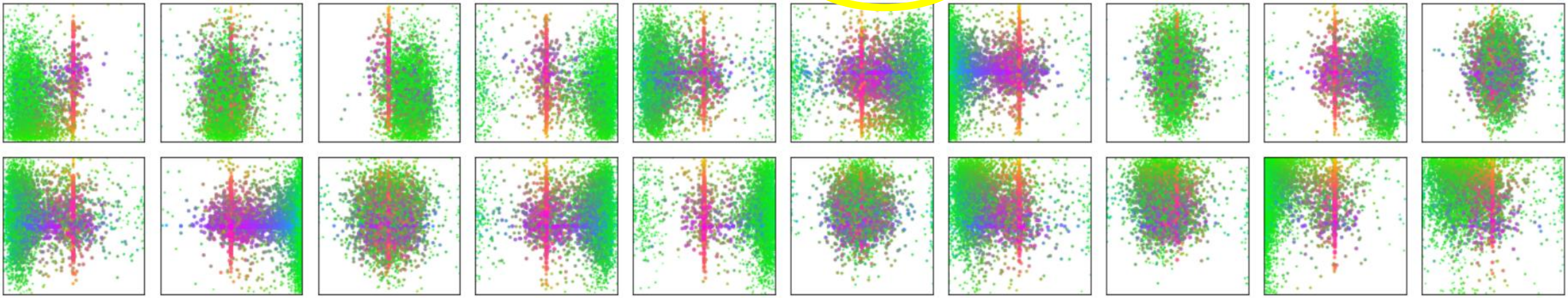


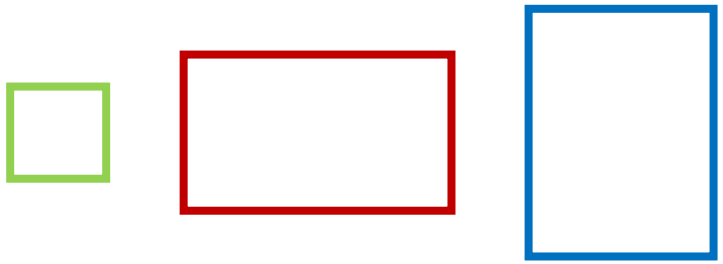
WAIC

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

DETR





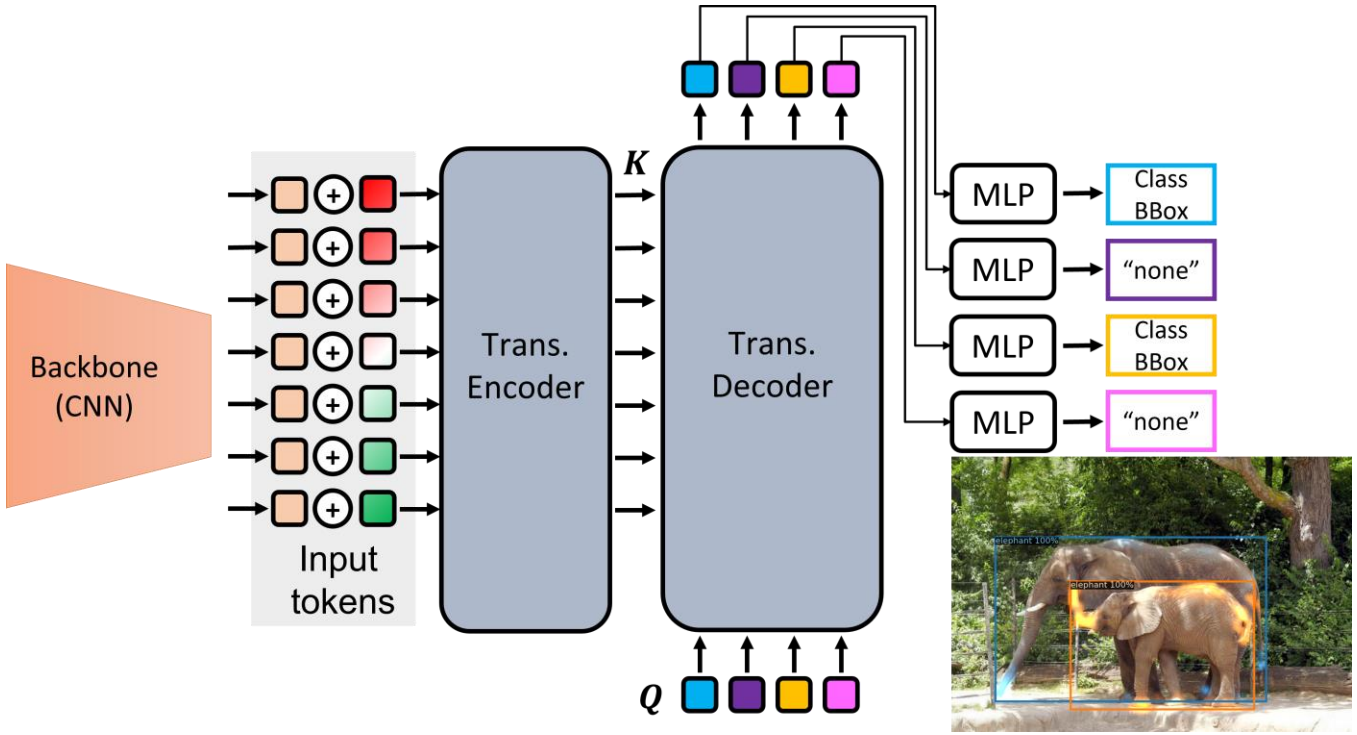


Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. [DETR: End-to-End Object Detection with Transformers](#) (ECCV 2020)

DETR

- End-to-end training
- No NMS (!)

- Simple approach
- On par with Faster-R-CNN



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

Previously

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

Today

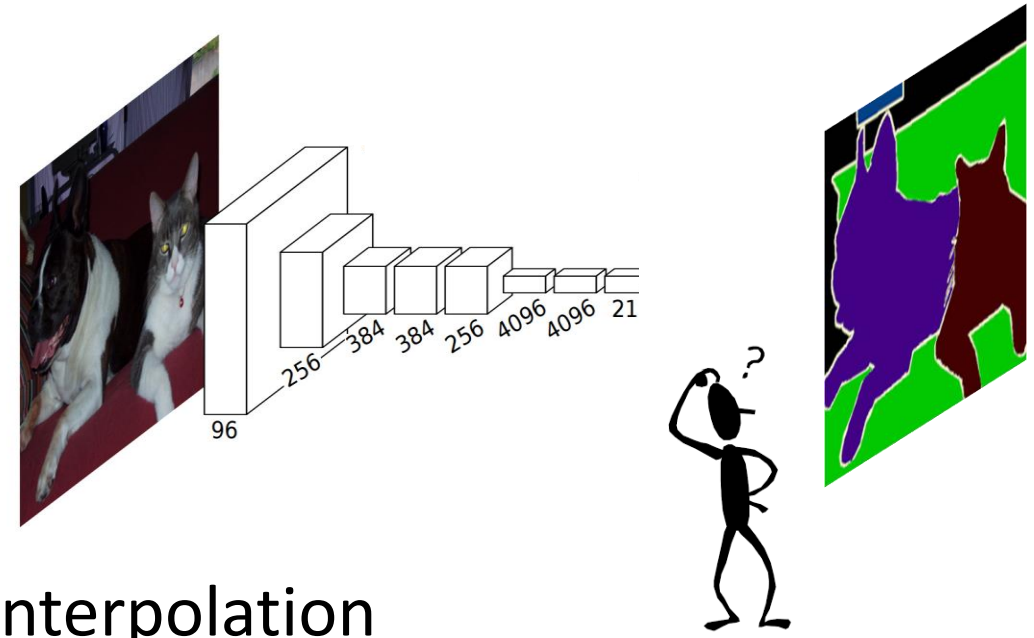
- (Multiple) Object detection
  - YOLO (You Only Look Once)
  - DETR (DEtection TRansformer)
- Semantic Segmentation
  - Segmenter

Semantic Segmentation

 Slide credit: Shai Bagon

Semantic Segmentation

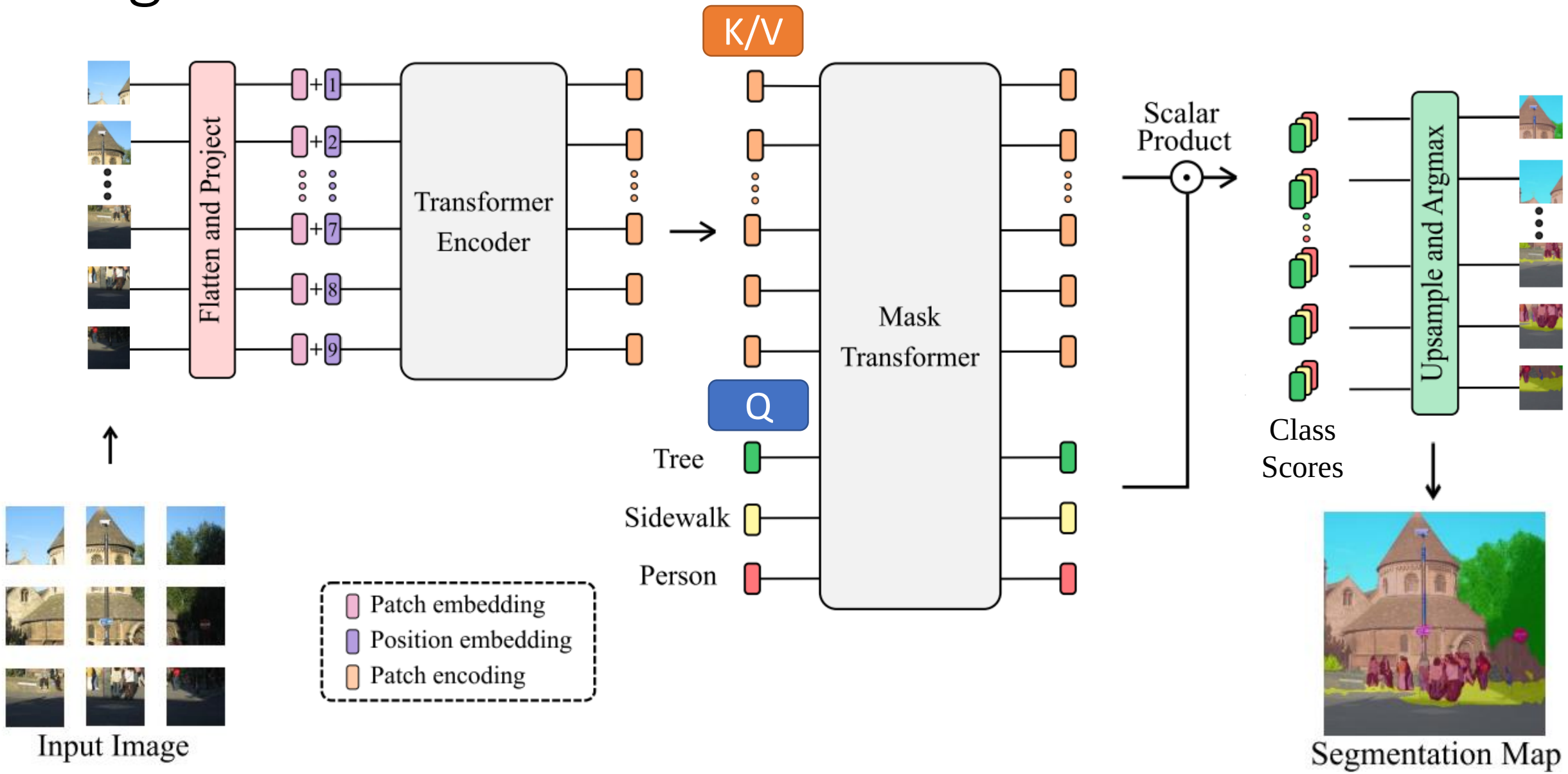
Resolution vs. Semantic information



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

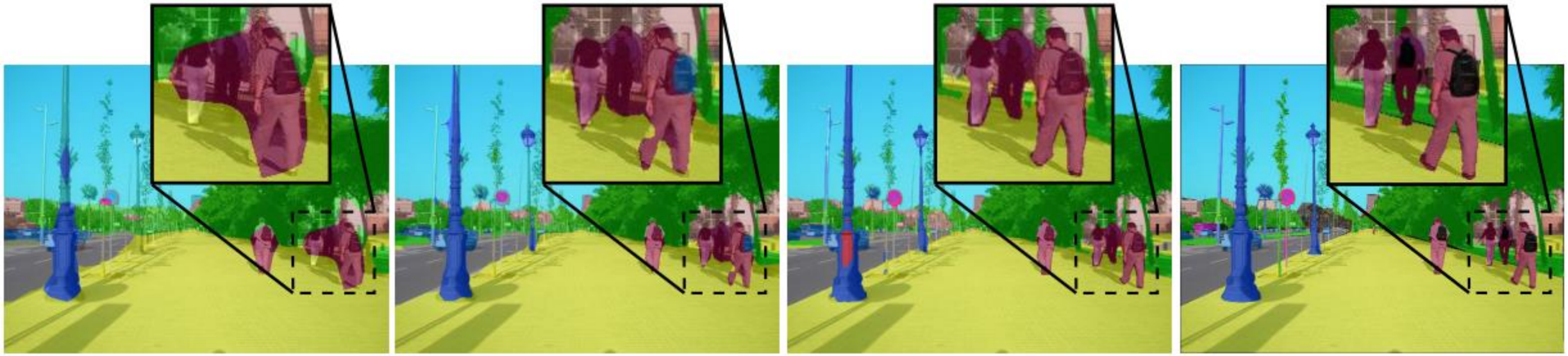
 Slide credit: Shai Bagon

Segmenter



R. Strudel, R. Garcia, I. Laptev, C. Schmid. Segmenter: Transformer for semantic segmentation (ICCV 2021)

Segmenter – Effect of Patch Size



(a) Patch size 32×32 (b) Patch size 16×16 (c) Patch size 8×8 (d) Ground Truth

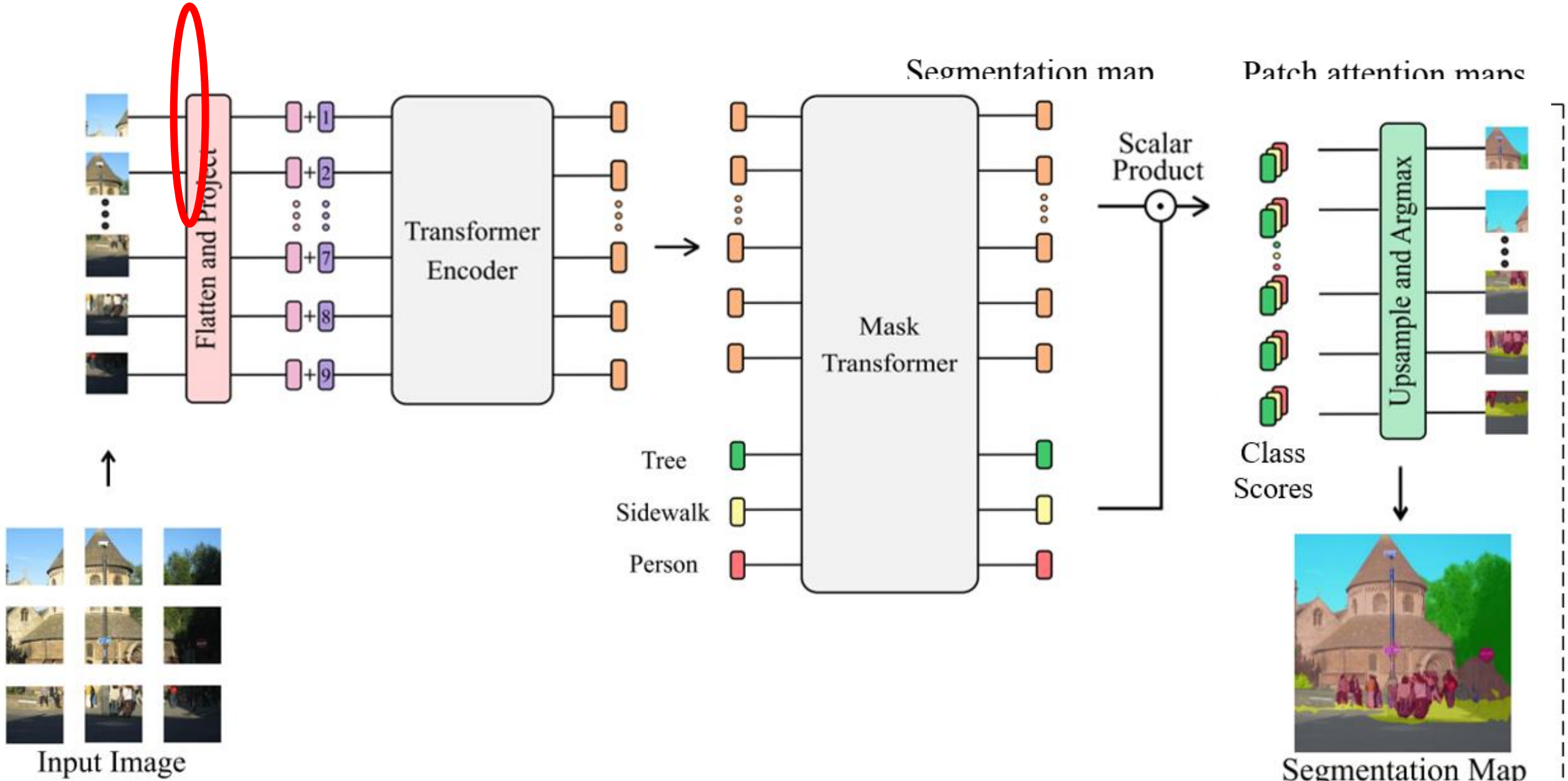
Fast

Low accuracy

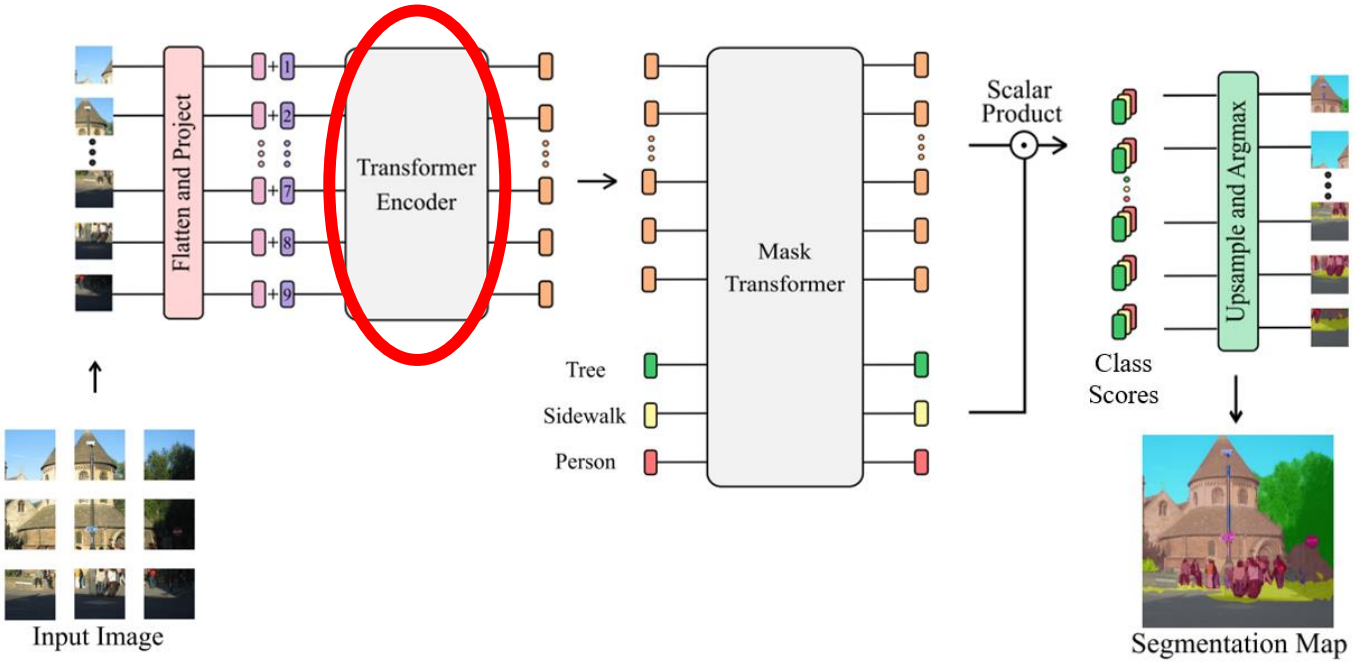
High computational cost

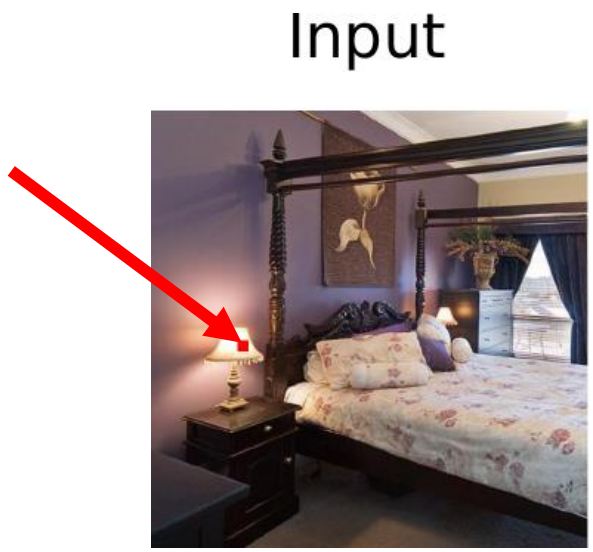
High accuracy

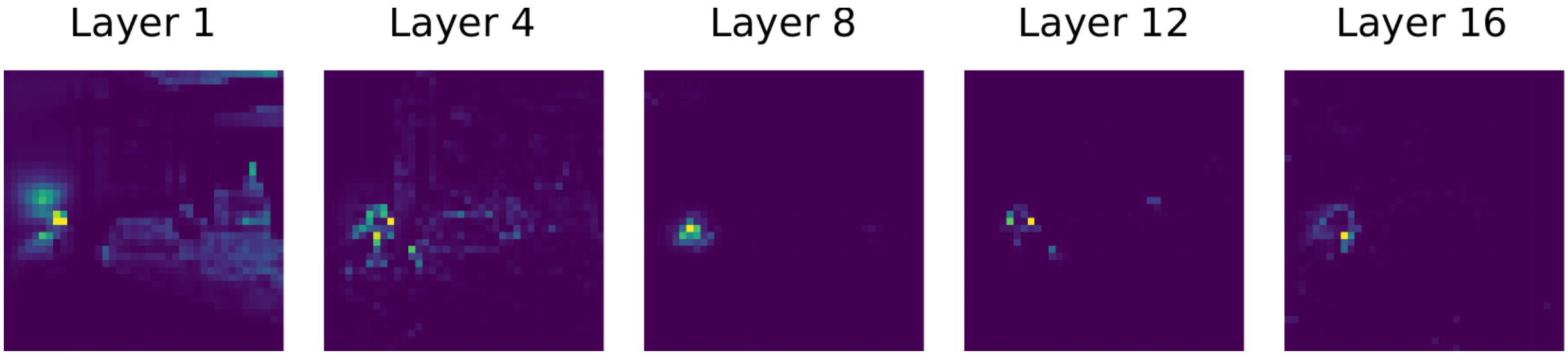
R. Strudel, R. Garcia, I. Laptev, C. Schmid. Segmenter: Transformer for semantic segmentation (ICCV 2021)



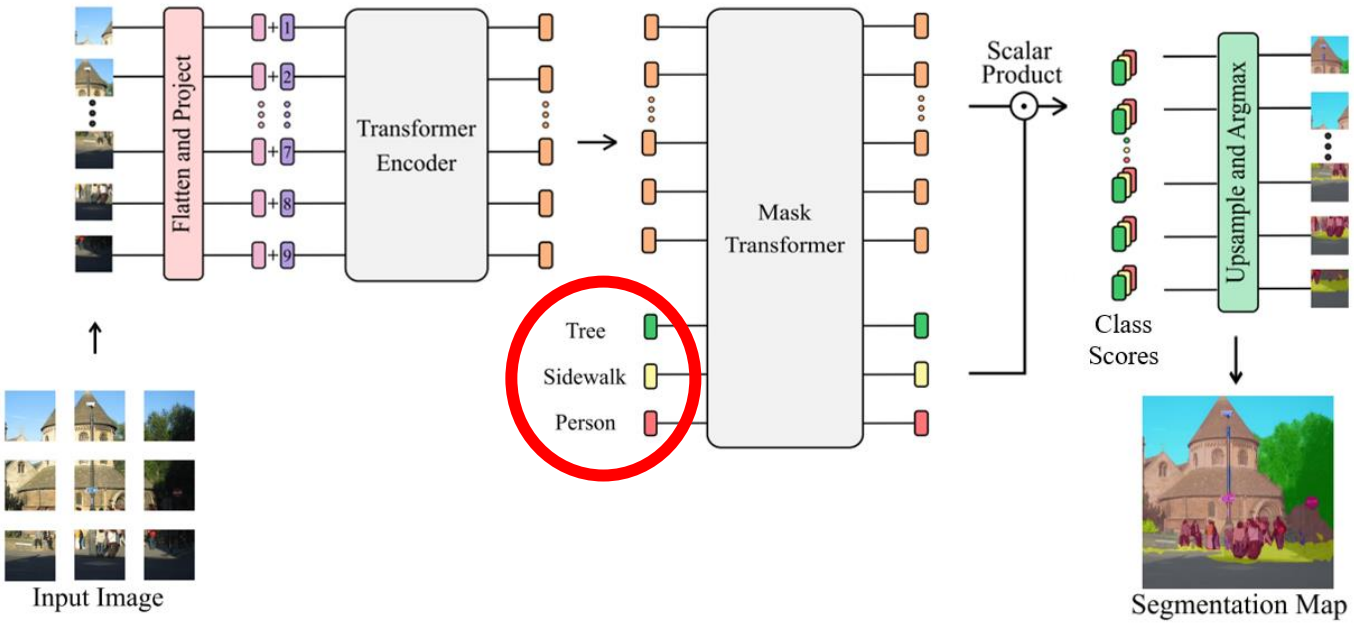
R. Strudel, R. Garcia, I. Laptev, C. Schmid. Segmenter: Transformer for semantic segmentation (ICCV 2021)





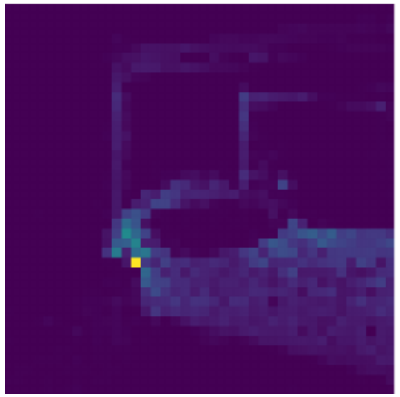


R. Strudel, R. Garcia, I. Laptev, C. Schmid. Segmenter: Transformer for semantic segmentation (ICCV 2021)

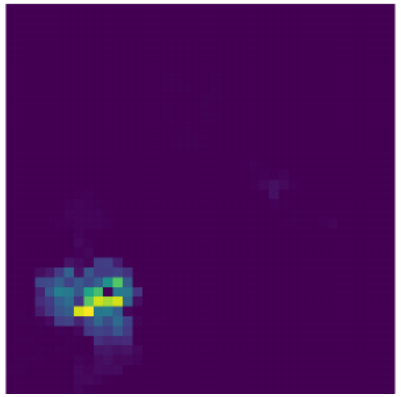


Prediction

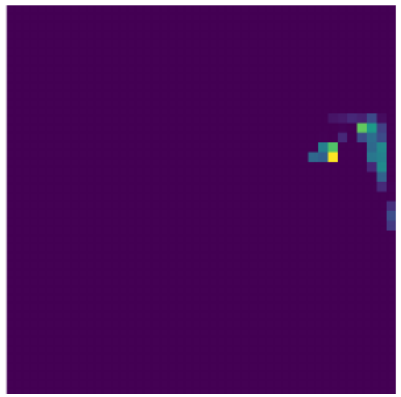
CLS 7 (bed)



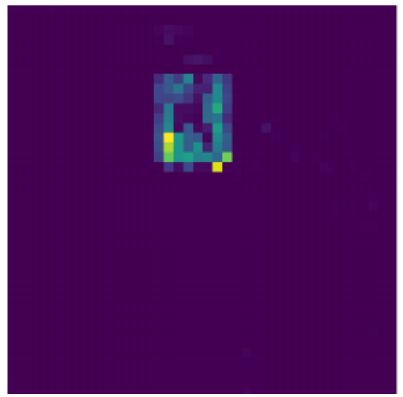
CLS 15 (table)



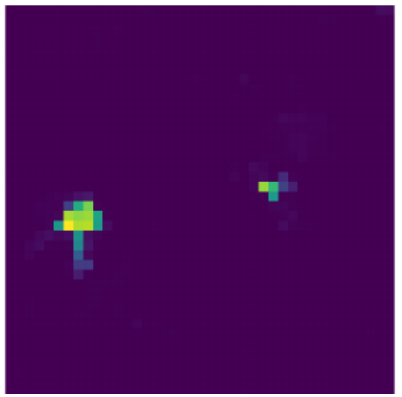
CLS 18 (curtain)



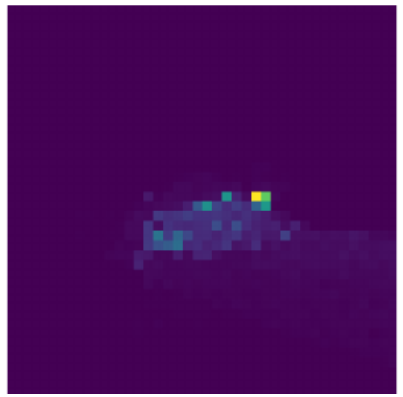
CLS 22 (painting)



CLS 36 (lamp)

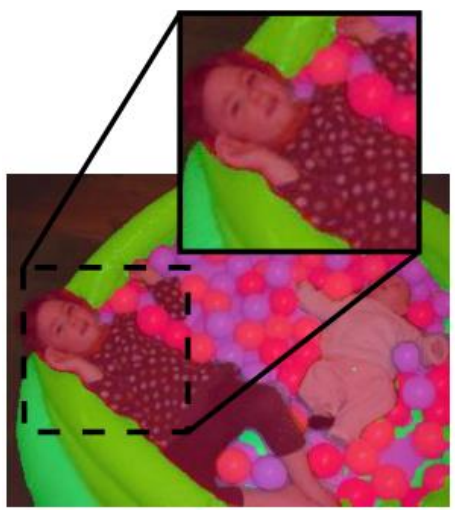


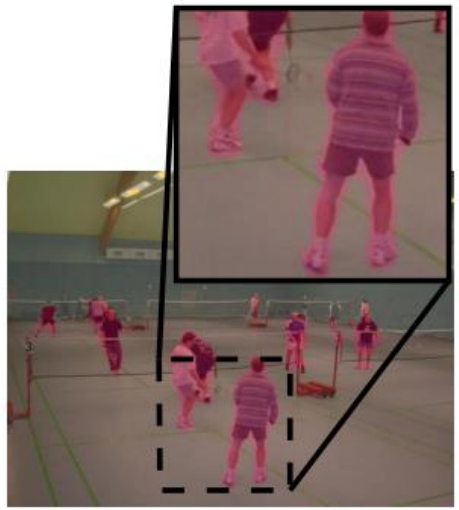
CLS 57 (pillow)

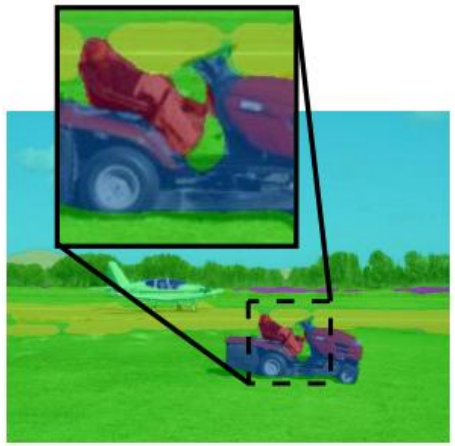


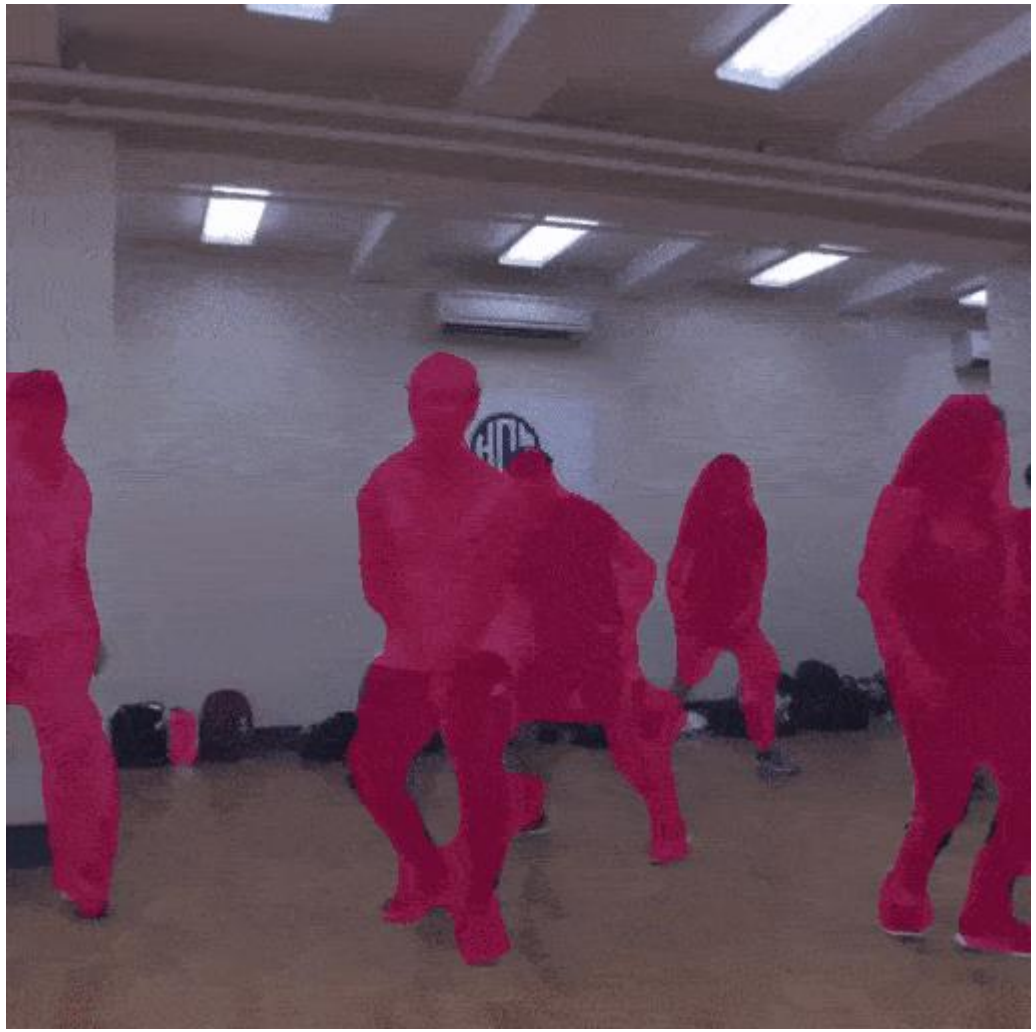
R. Strudel, R. Garcia, I. Laptev, C. Schmid. Segmenter: Transformer for semantic segmentation (ICCV 2021)

Segmenter









video

R. Strudel, R. Garcia, I. Laptev, C. Schmid. Segmenter: Transformer for semantic segmentation (ICCV 2021)

Today

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

Next Week

- Advanced Generative Models