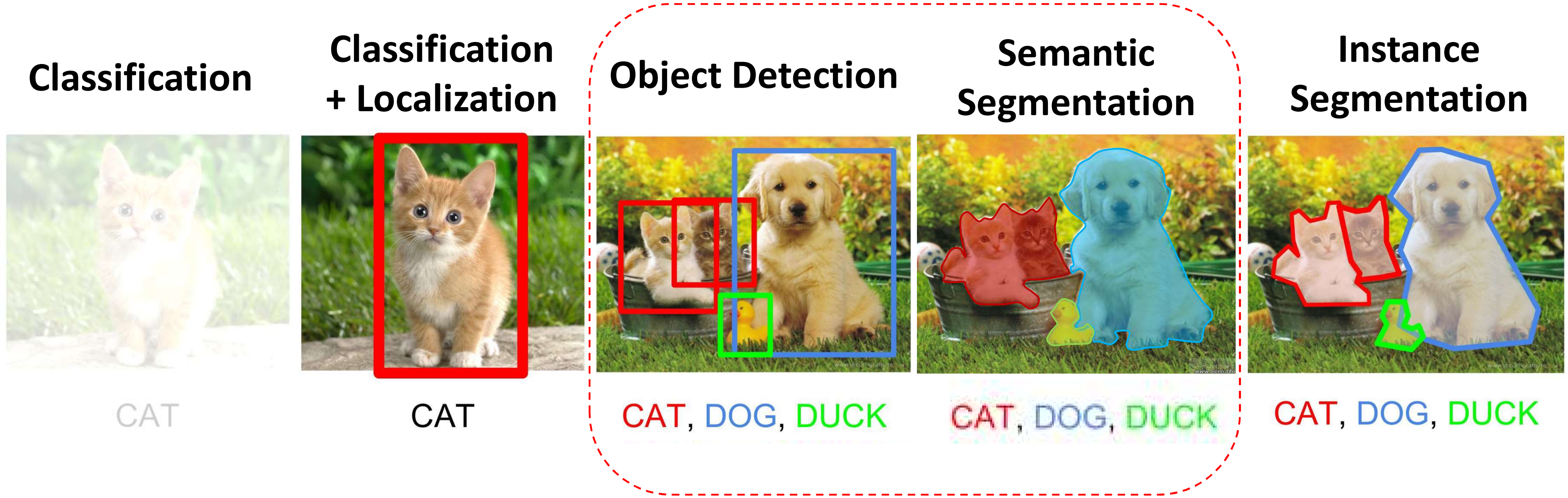
Object Detection and Segmentation

Dolev Ofri

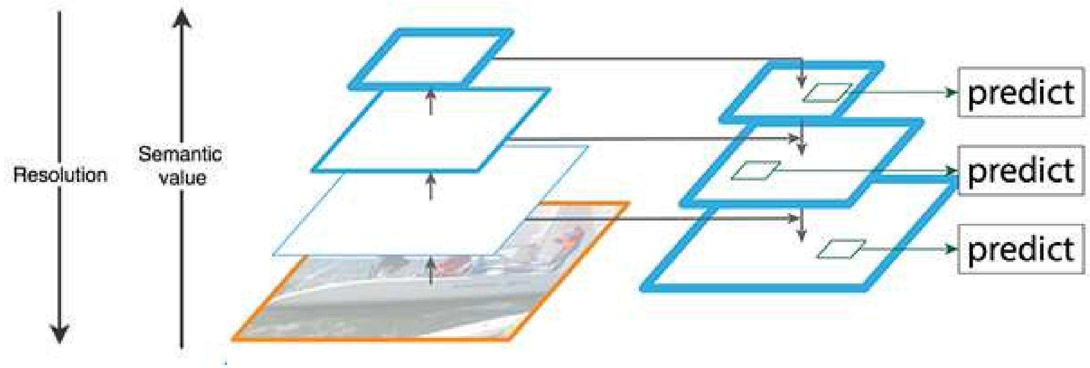
December 16th, 2021

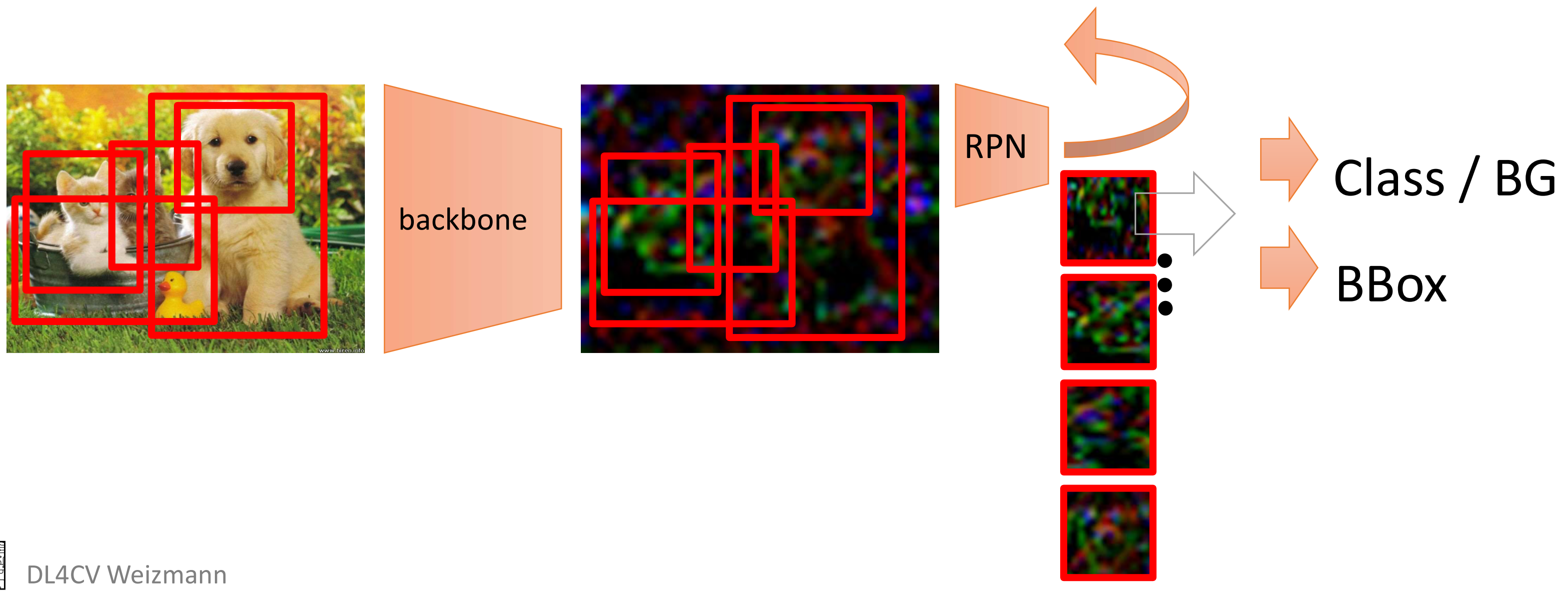
Last week


Classification
Classification + Localization
Object Detection
Semantic Segmentation
Instance Segmentation
CAT
CAT
CAT, DOG, DUCK
CAT, DOG, DUCK
CAT, DOG, DUCK

Last week

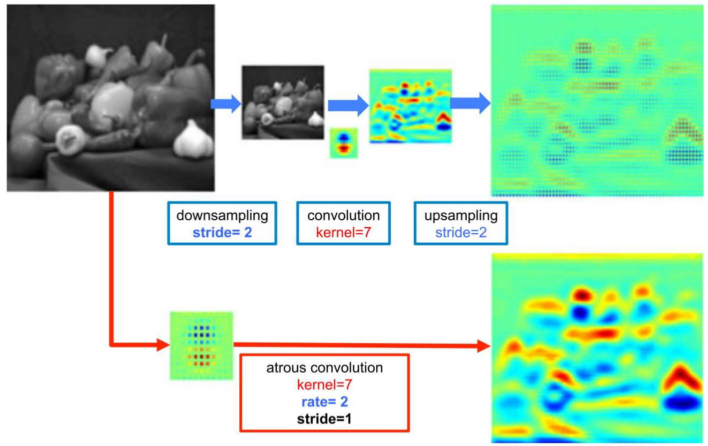
- (Multiple) Object detection
  - Faster RCNN
  - FPN (Feature Pyramid Network)

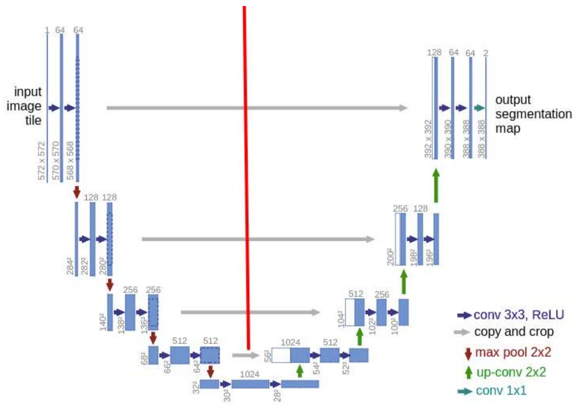


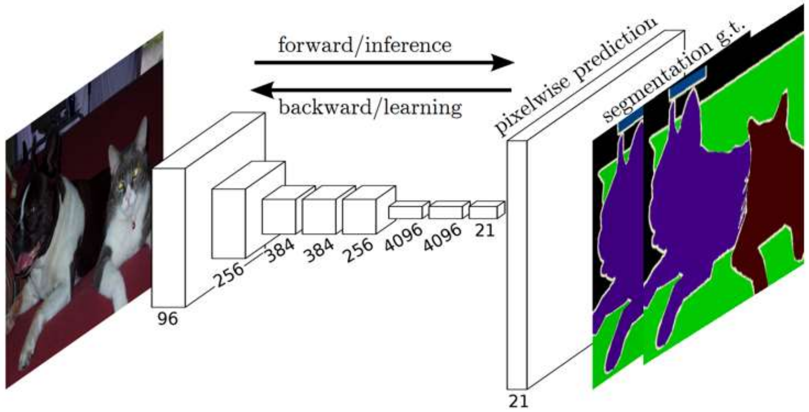


Last week

- (Multiple) Object detection
    - Faster RCNN
    - FPN (Feature Pyramid Network)
- Semantic segmentation
    - FCN
    - DeepLab
    - U-net







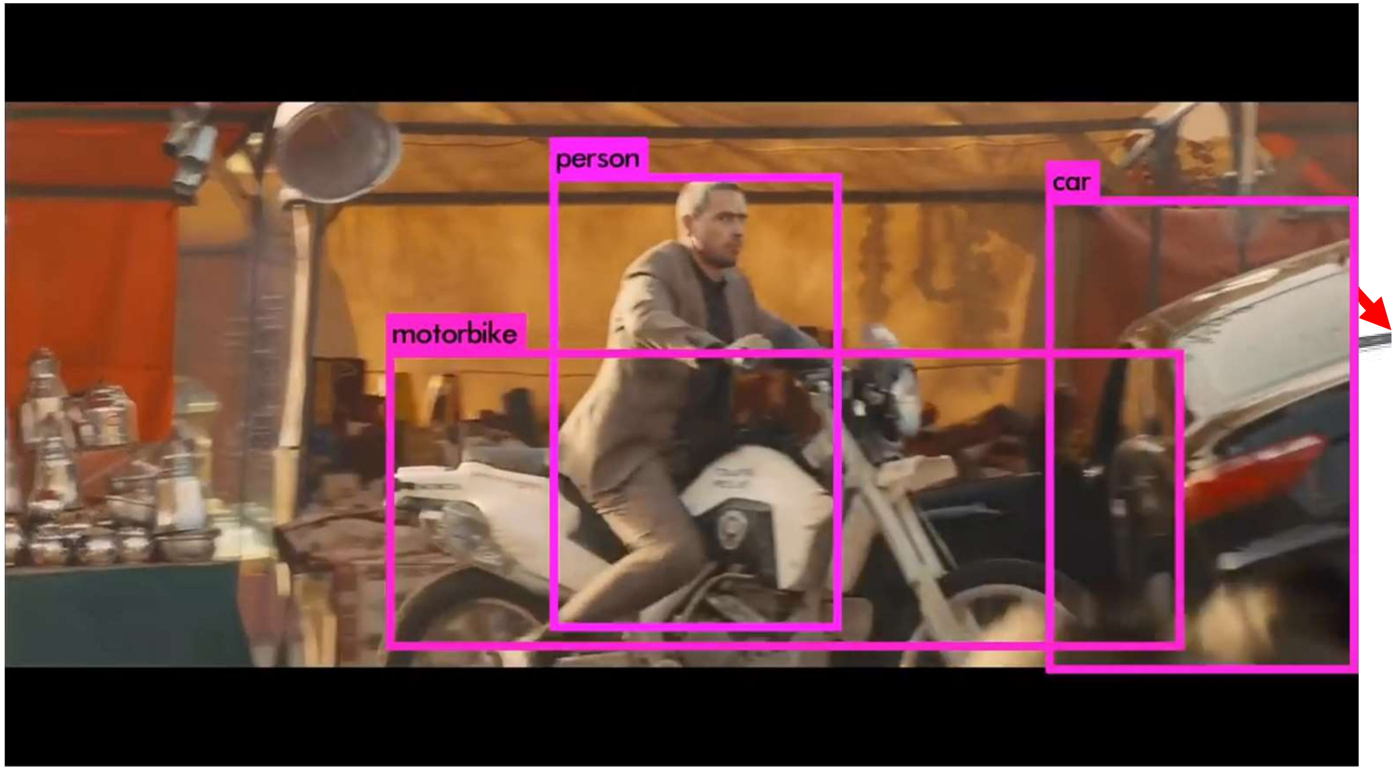
Last week

- (Multiple) Object detection
  - Faster RCNN
  - FPN (Feature Pyramid Network)
- Semantic segmentation
  - FCN
  - DeepLab
  - U-net

Today

- (Multiple) Object detection
  - YOLO (You Only Look Once)
- Semantic Segmentation
  - Segmenter: Transformer for Semantic Segmentation

Object Detection

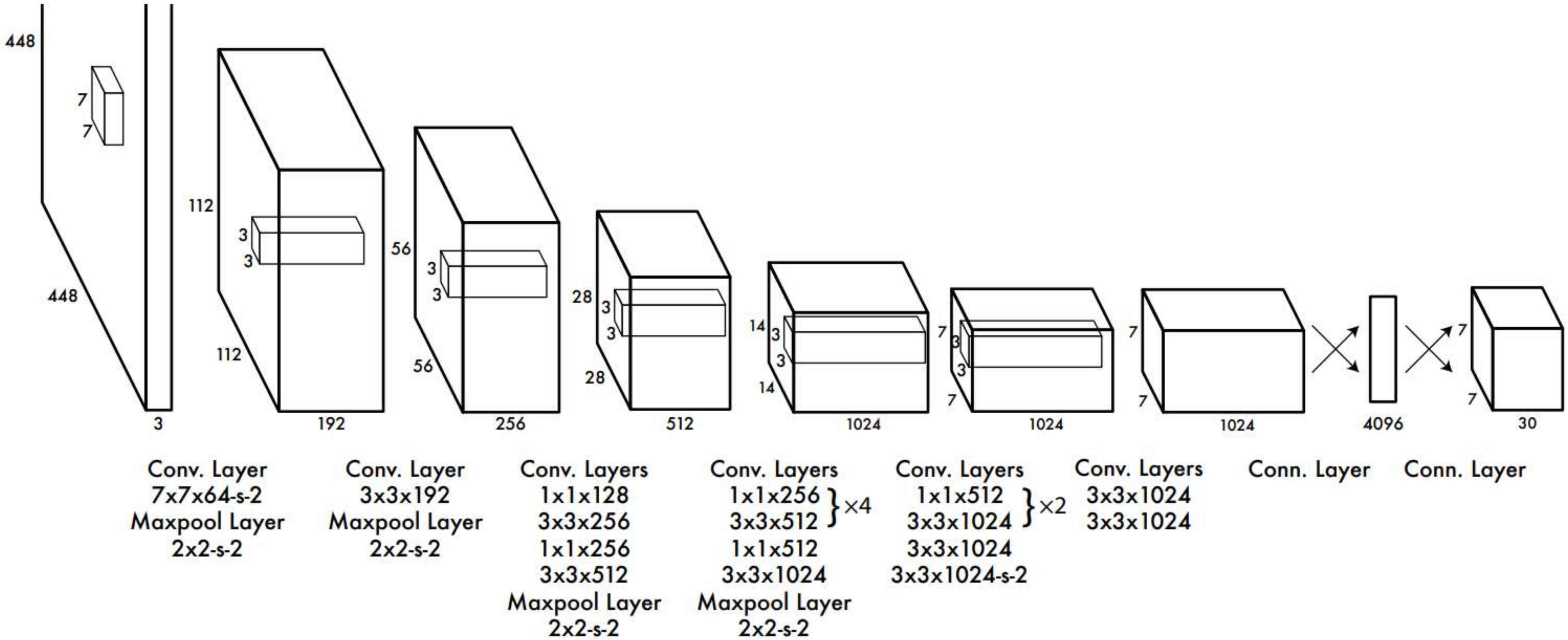


Single Shot:
SSD, YOLO …

Fast
High false rate

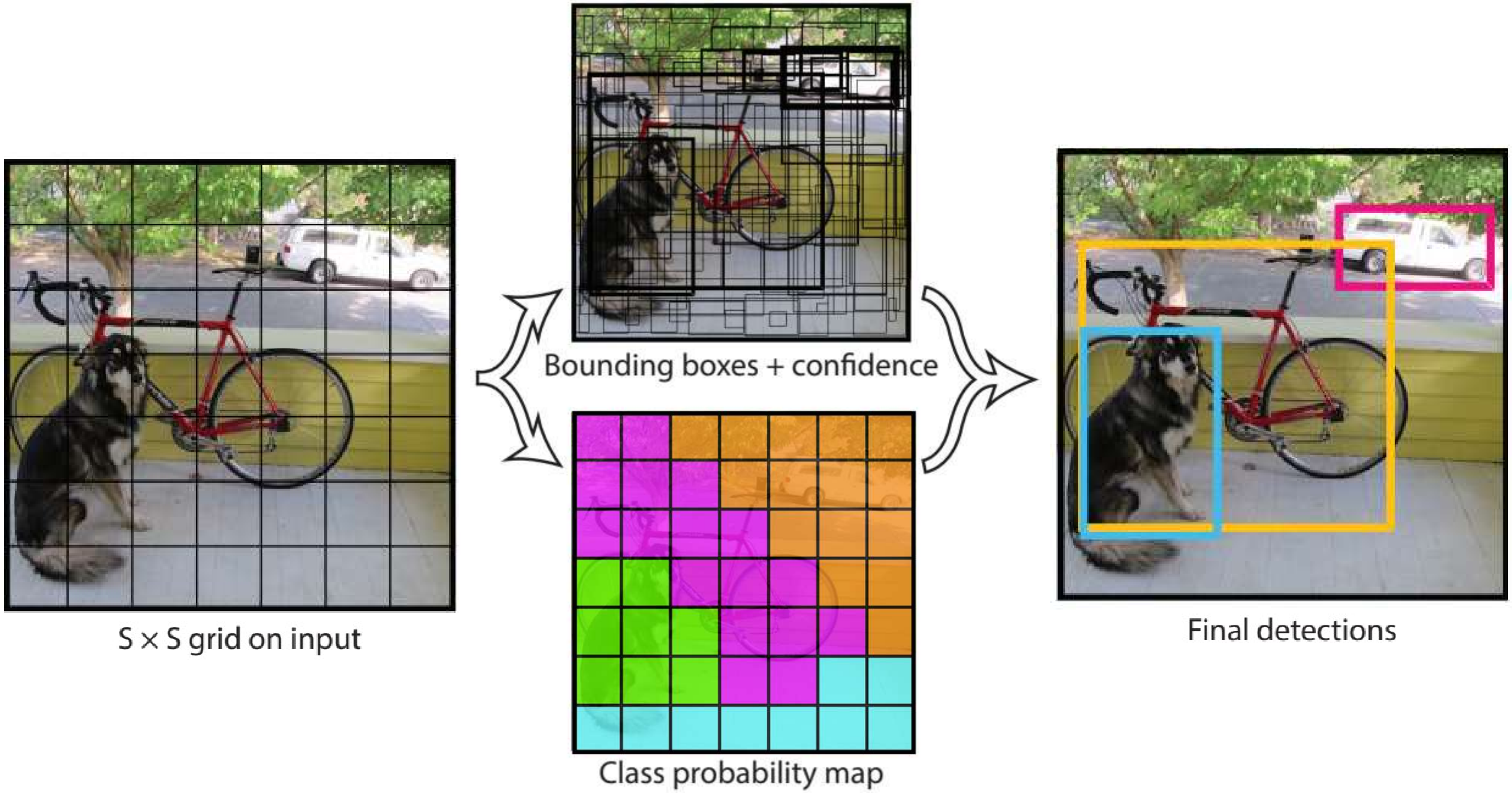
 Video: [YOLOv2]

YOLO (You Only Look Once)



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

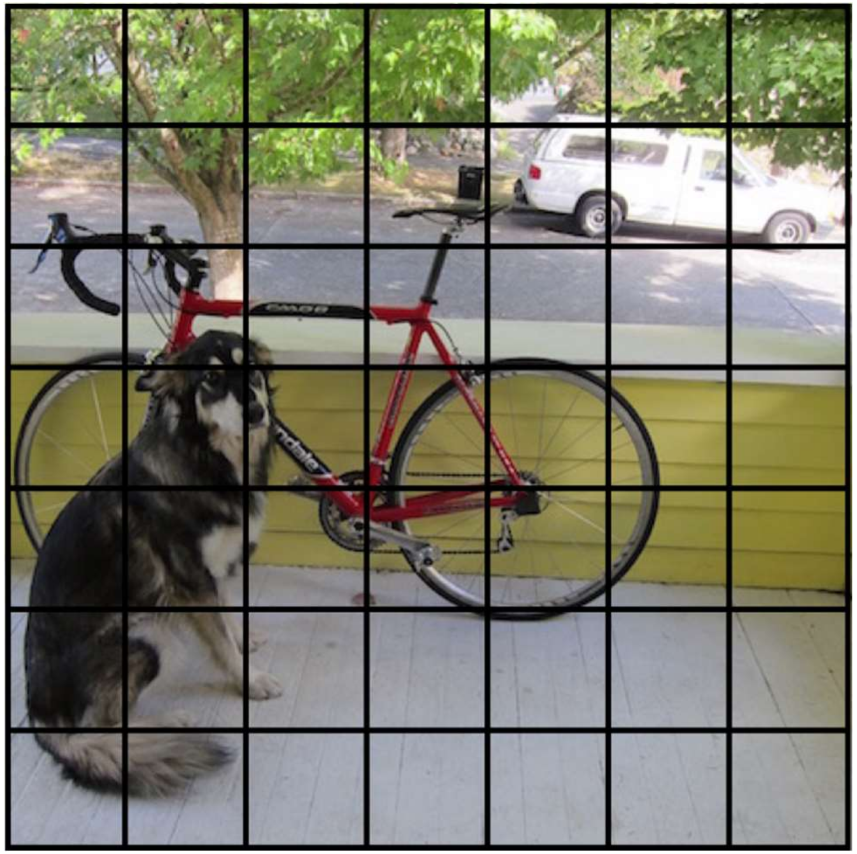
YOLO – Overview



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

YOLO – Inference

Divide input image into a grid



$S \times S$ grid on input

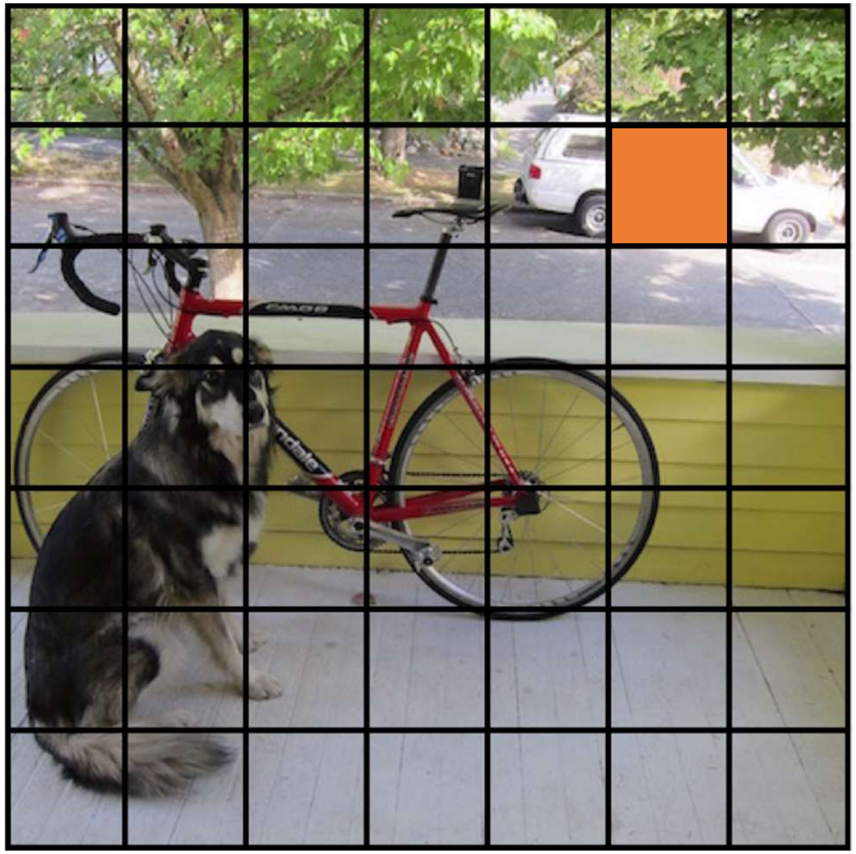
Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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



S × S grid on input

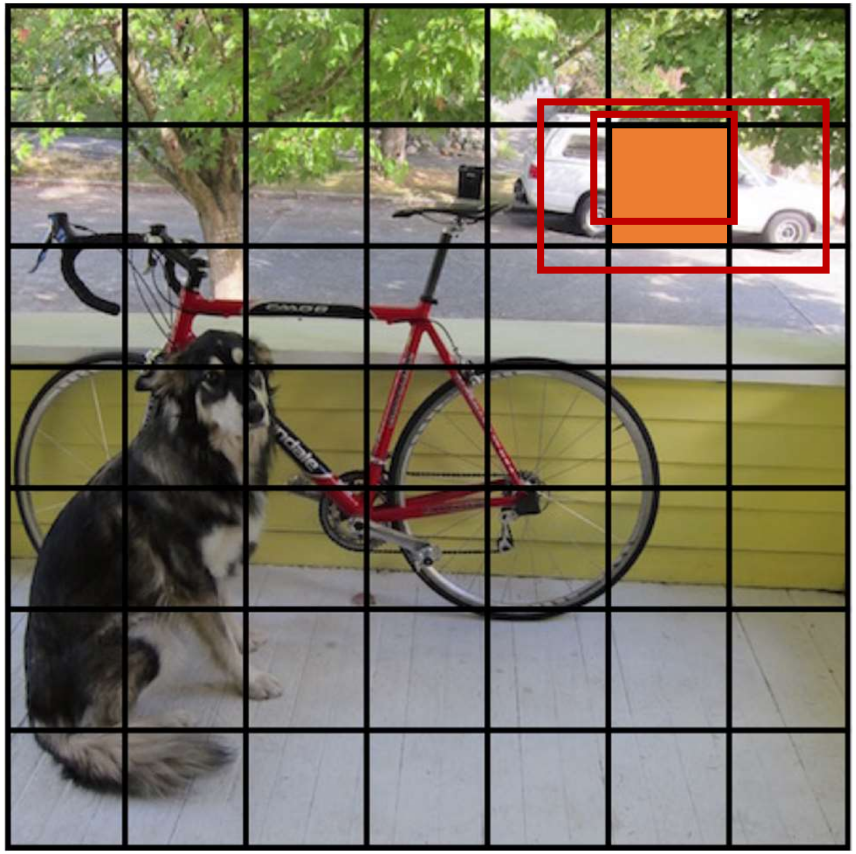
Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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



S × S grid on input

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

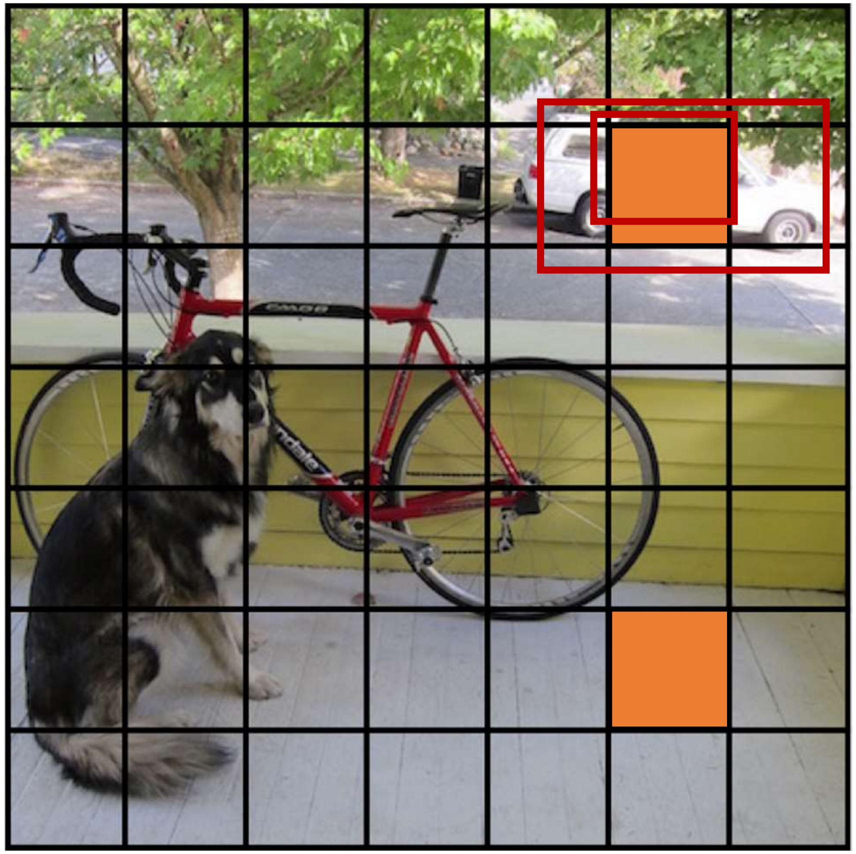
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection, 2015 (CVPR 2016)

YOLO – Inference

Each cell predicts

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

Cells with no object
→ low confidence score



S × S grid on input

Slide 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

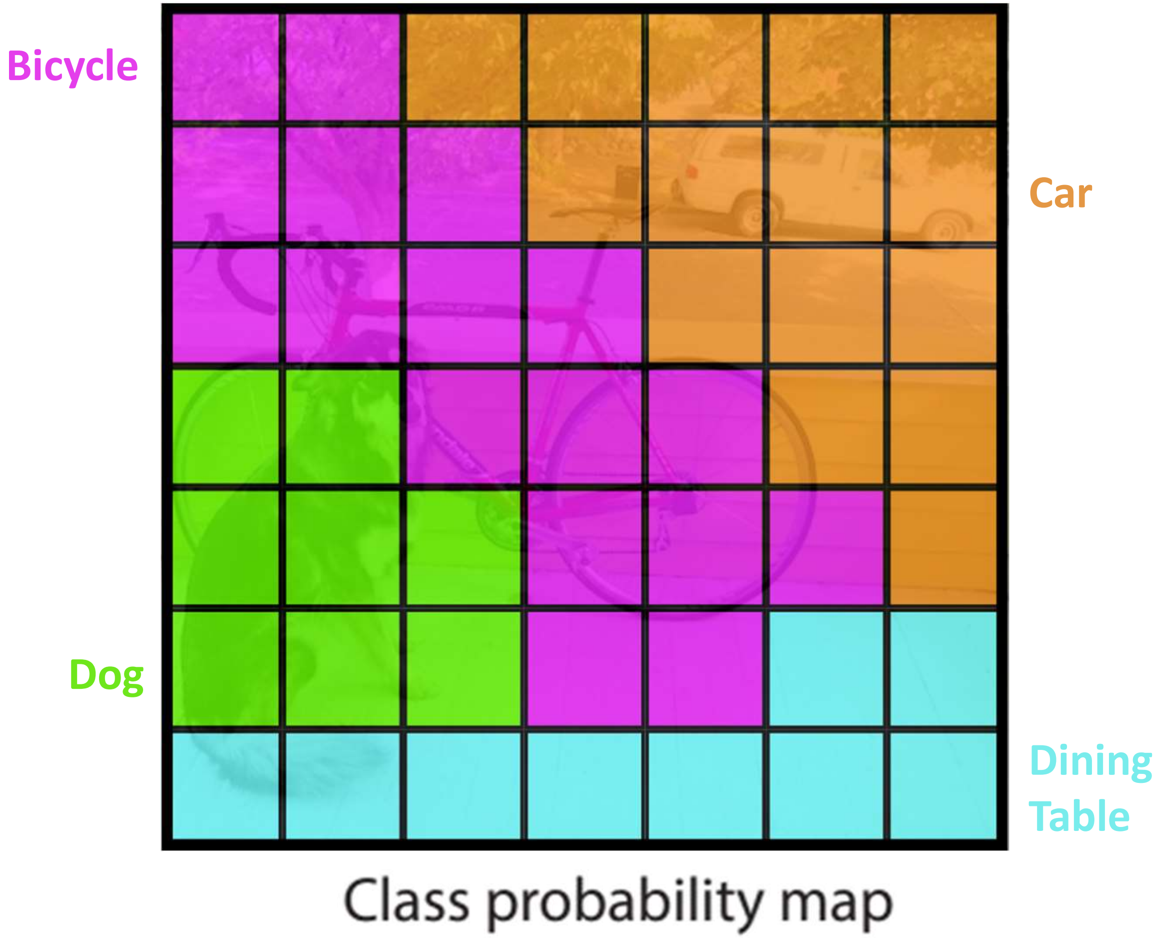
Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

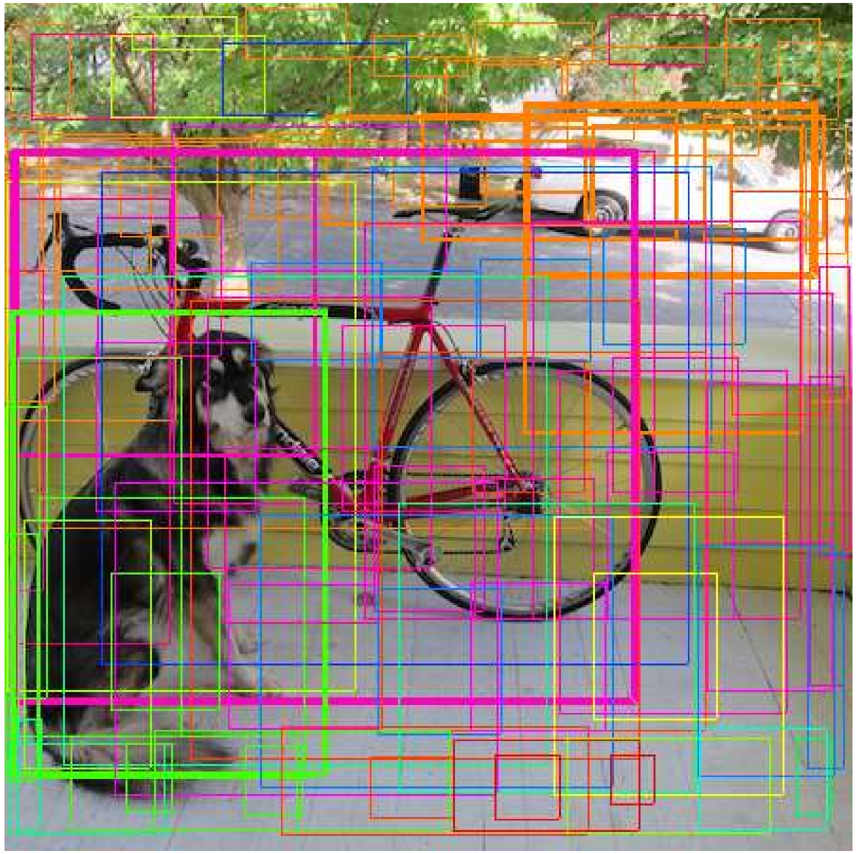


Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Inference

Combine predictions



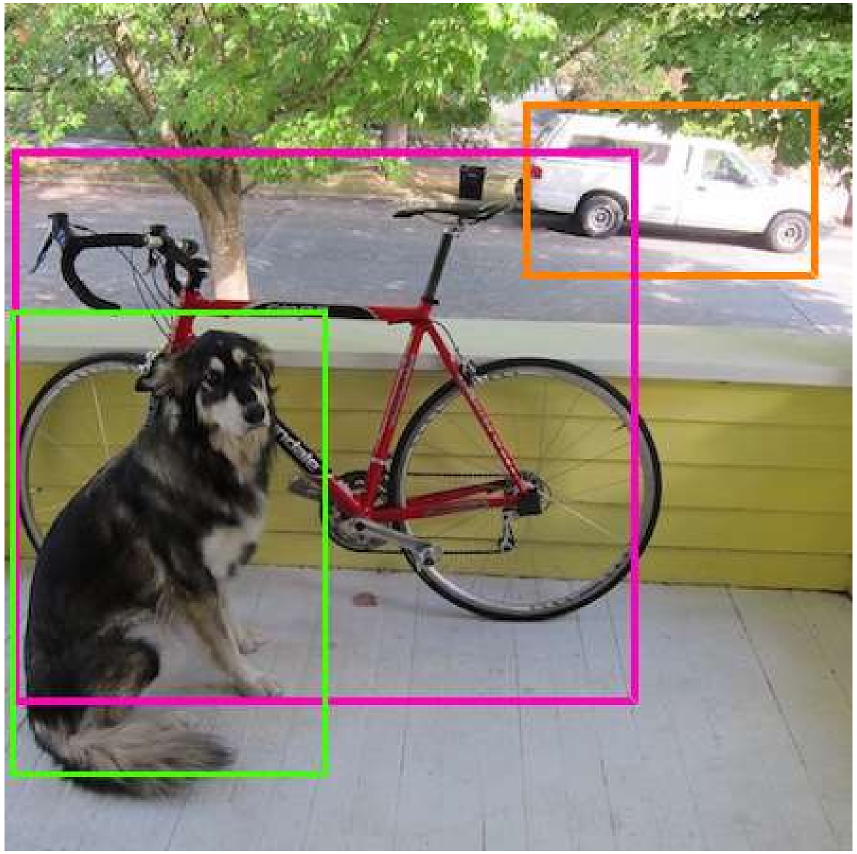
Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

YOLO – Inference

Apply

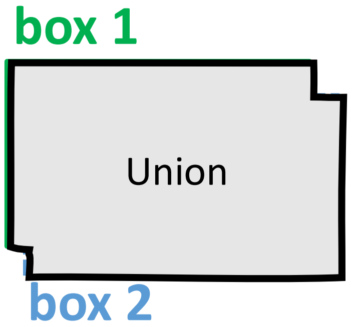
- Non-maximal suppression (NMS)
- Threshold



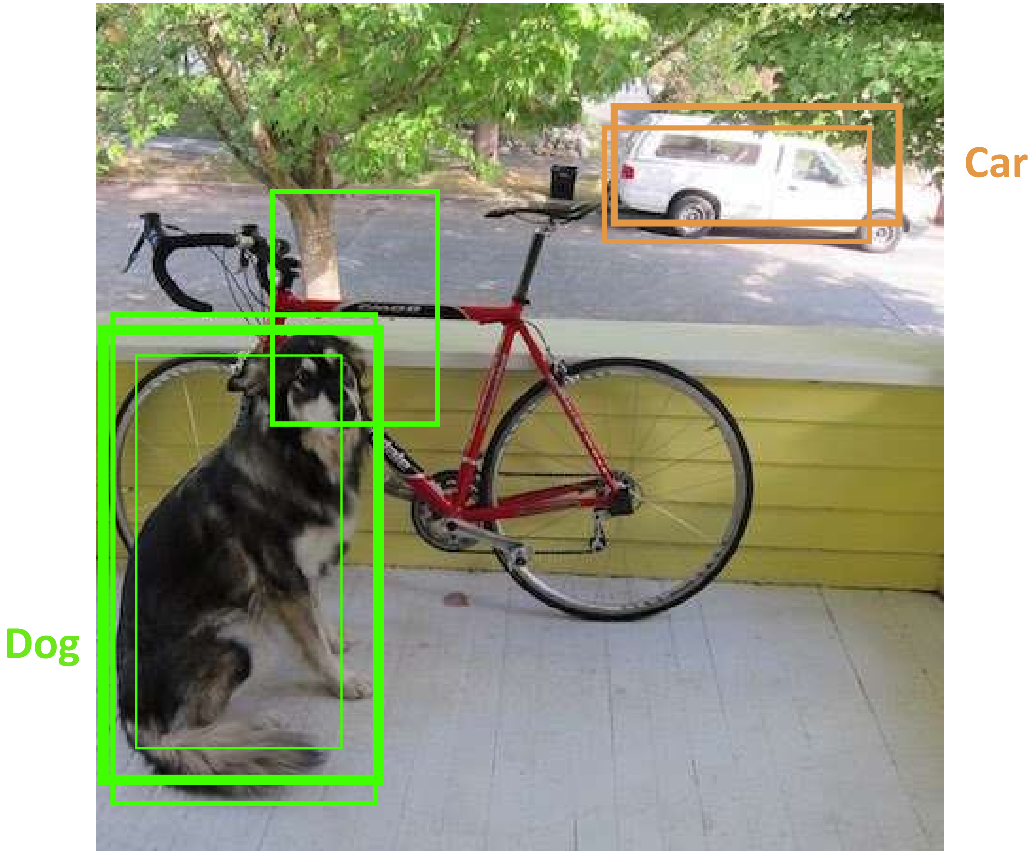
Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

Inference – Non Maximal Suppression



IoU =

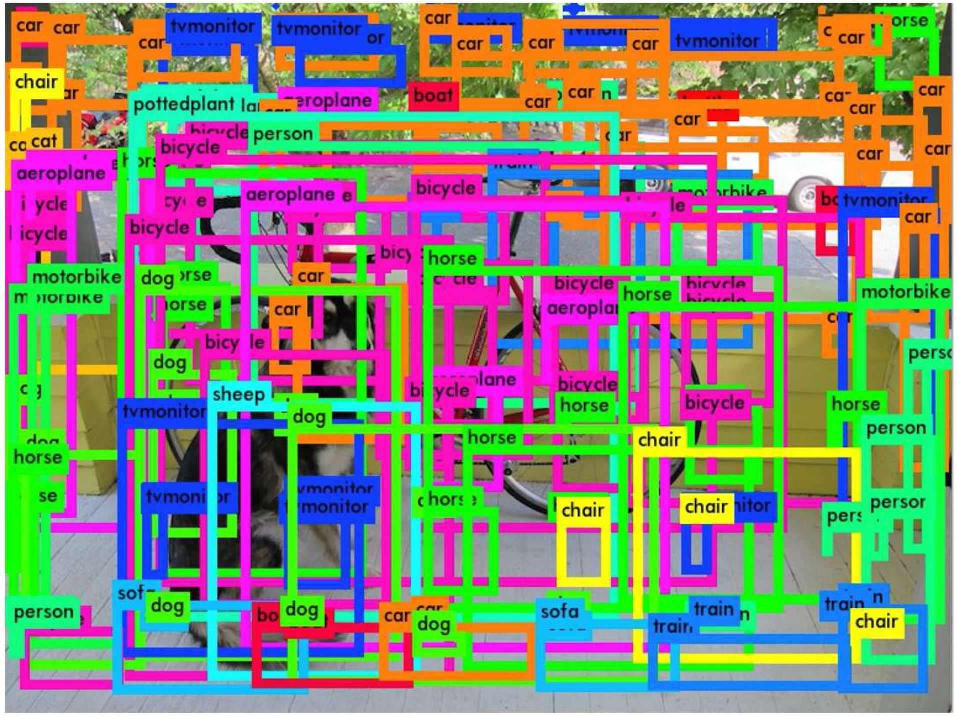


Slide idea: Shai Bagon

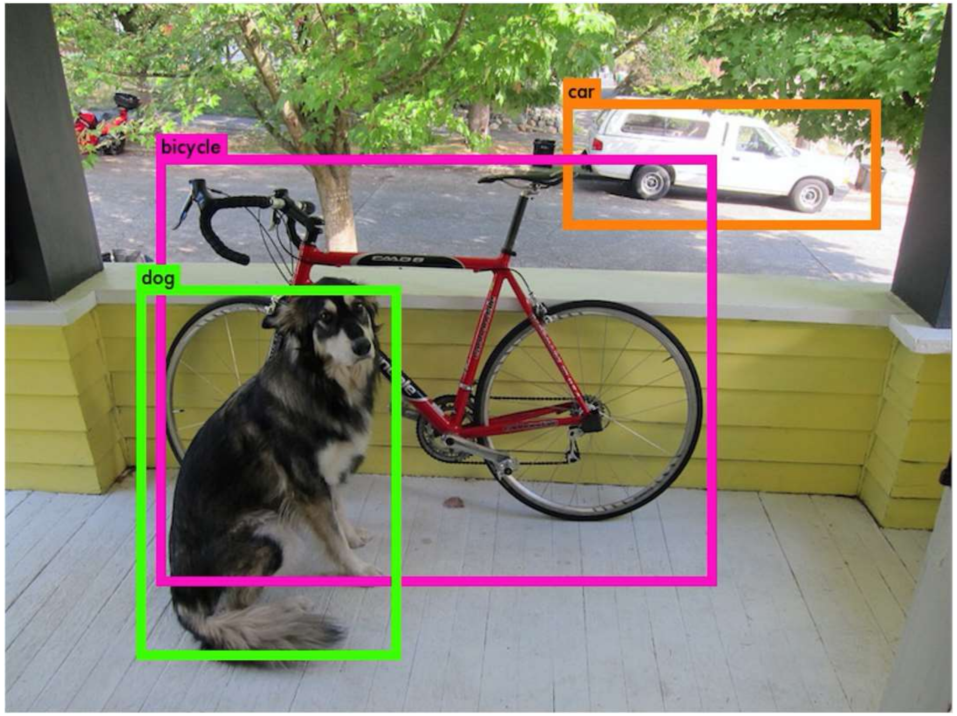
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](), 2015 (CVPR 2016)

Inference – Threshold

Low Threshold

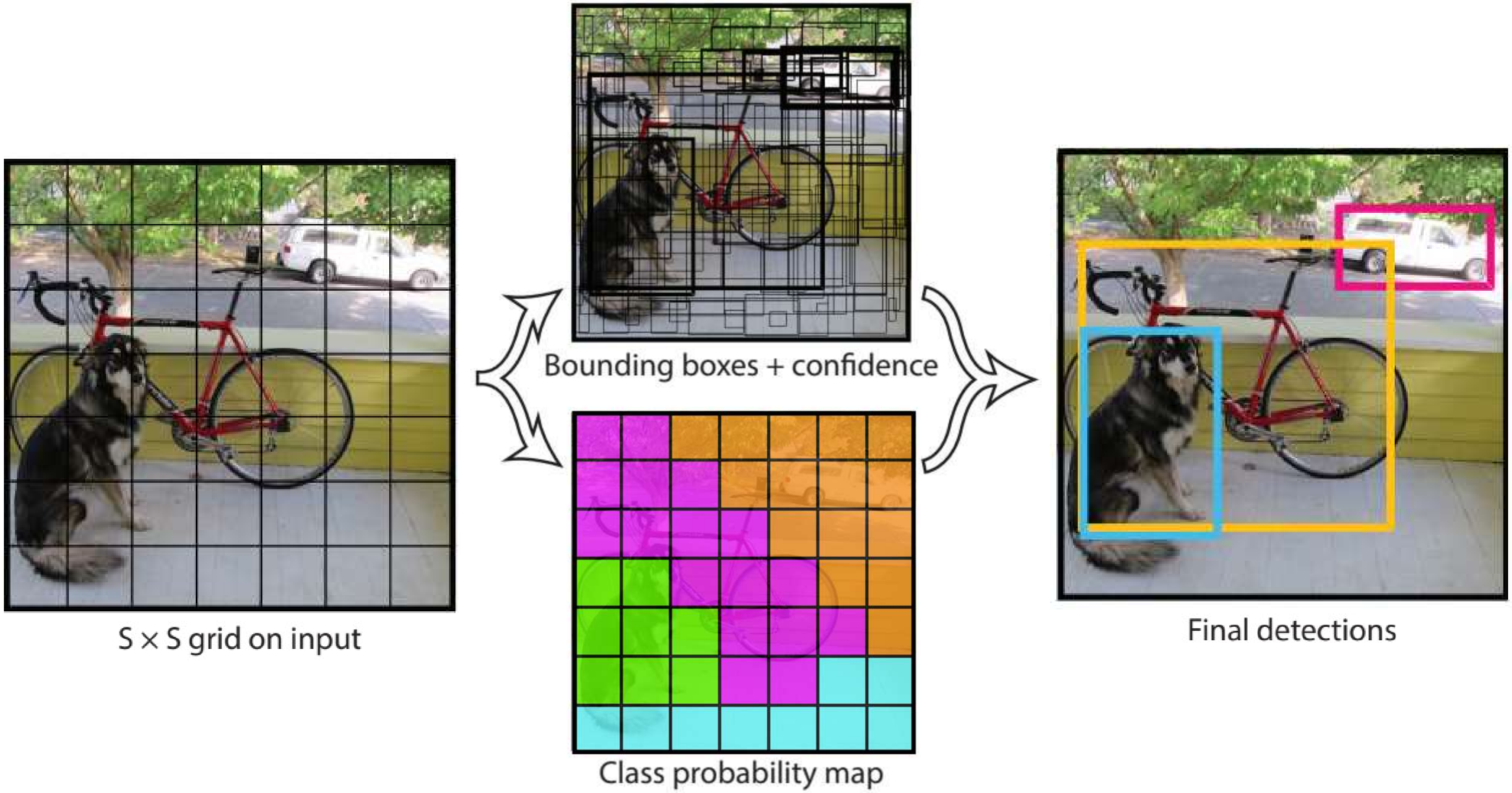


High Threshold



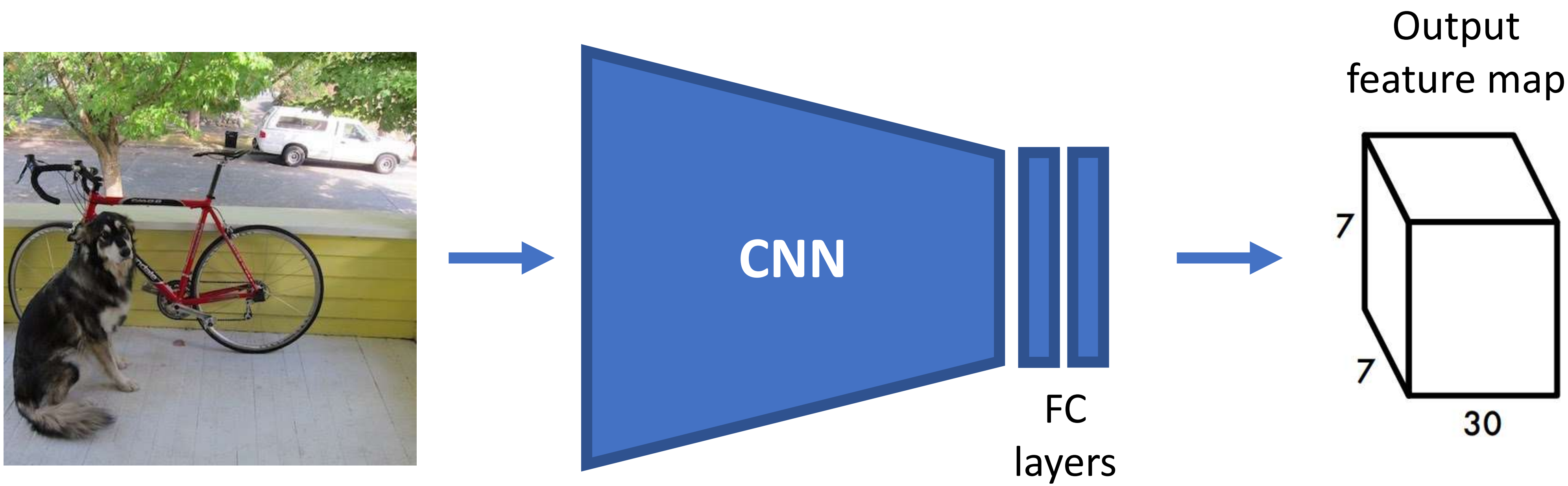
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](), 2015 (CVPR 2016)

YOLO – Overview



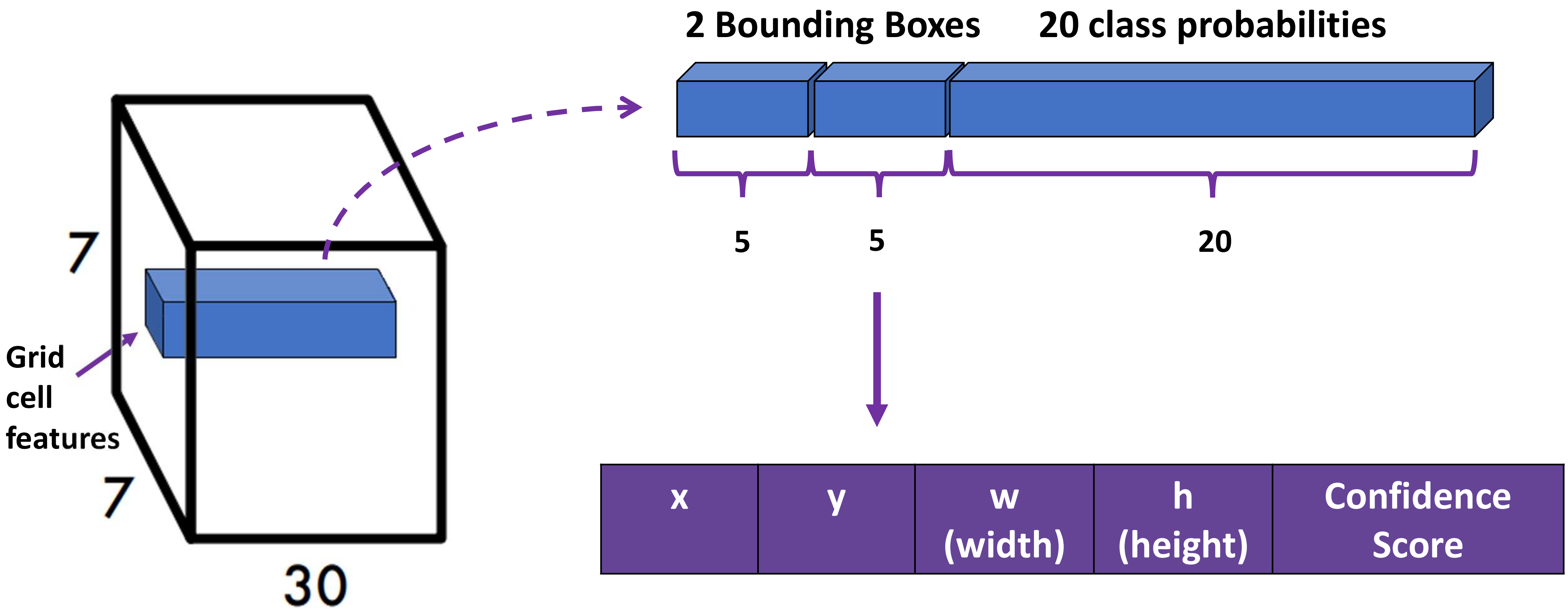
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection, 2015 (CVPR 2016)

YOLO – Output feature map



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

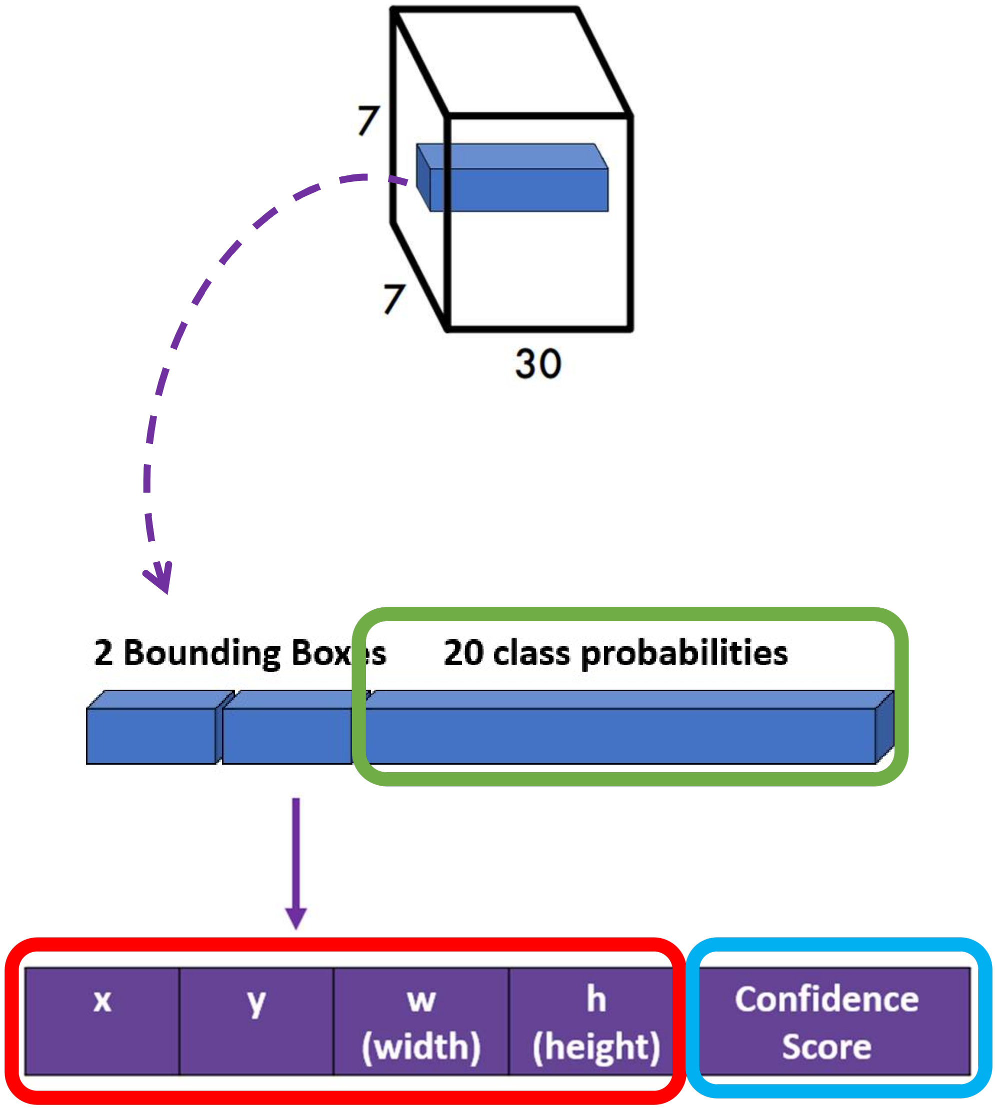
YOLO – Output feature map



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

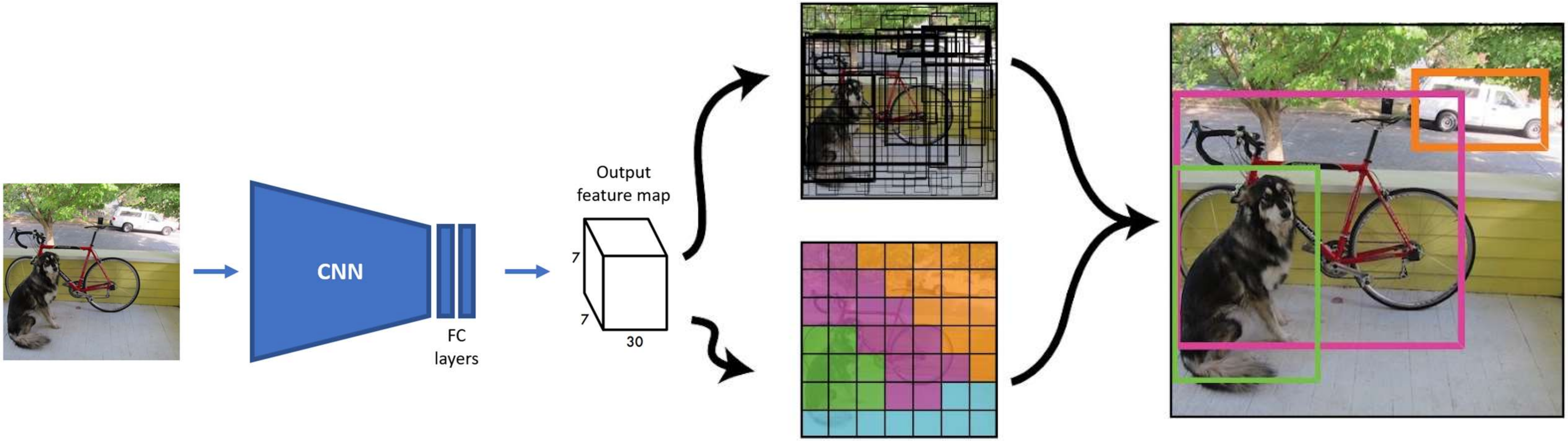
YOLO – Loss function

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



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

YOLO – end-to-end training!



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

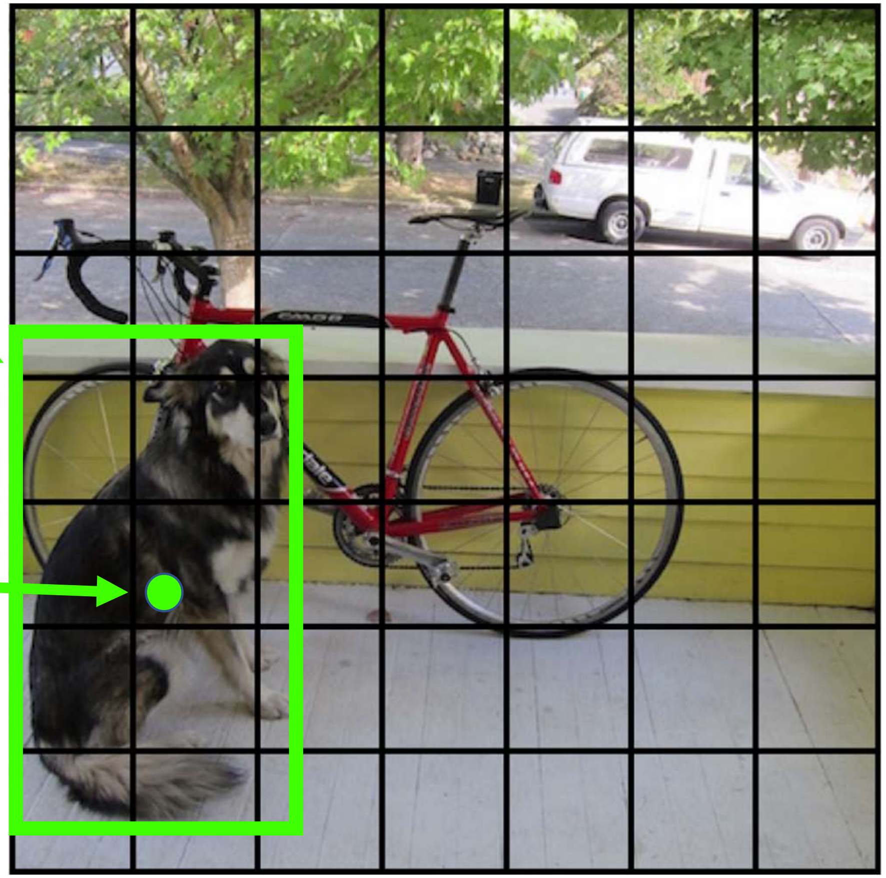
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](), 2015 (CVPR 2016)

YOLO – Training

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

Ground truth bounding box

Center of object



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

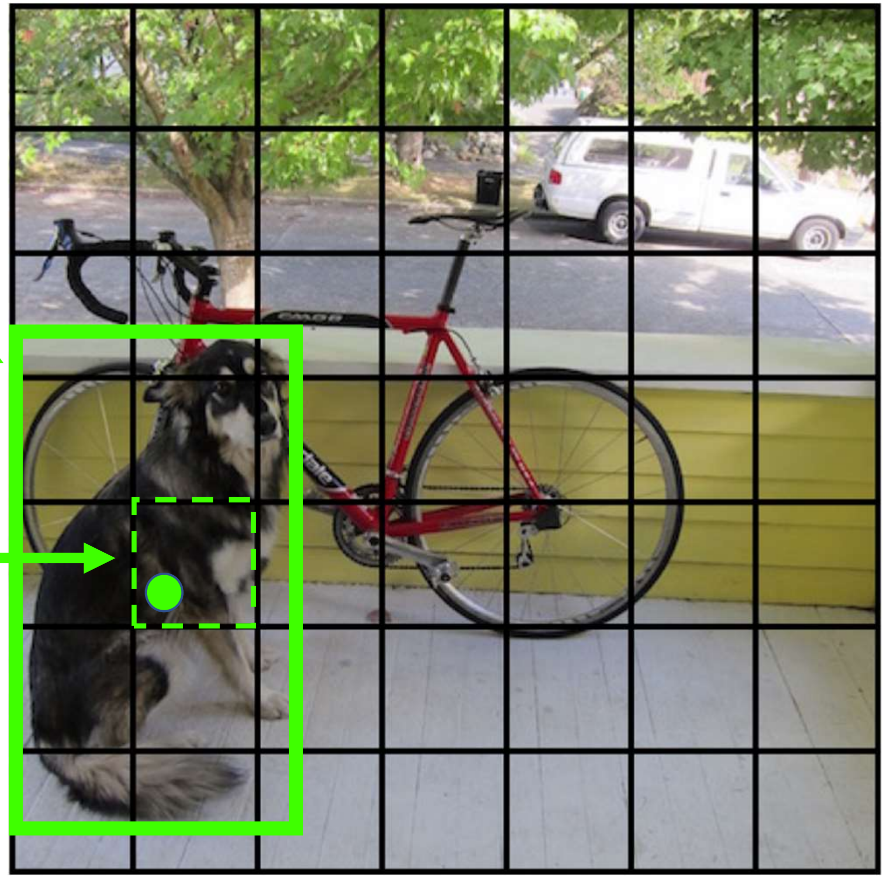
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](), 2015 (CVPR 2016)

YOLO – Training

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

Ground truth bounding box

Assign to specific cell



Slide 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)

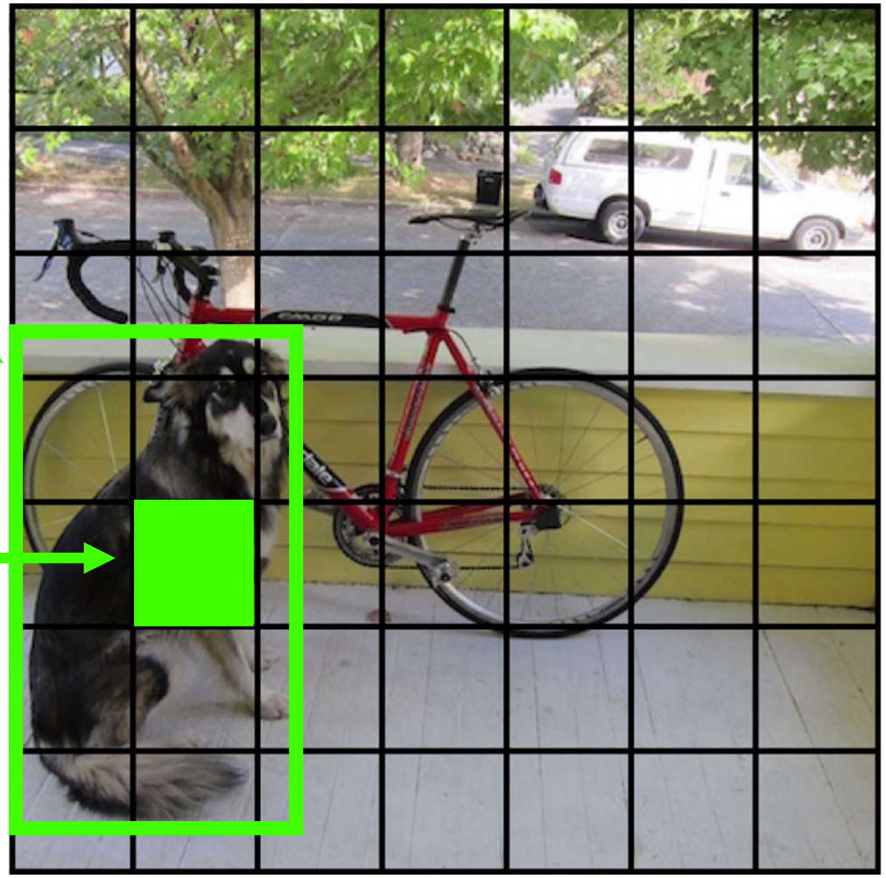
YOLO – Training

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

Ground truth bounding box

Supervisory signal:

Dog: 1
Cat: 0
Bike: 0
…



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

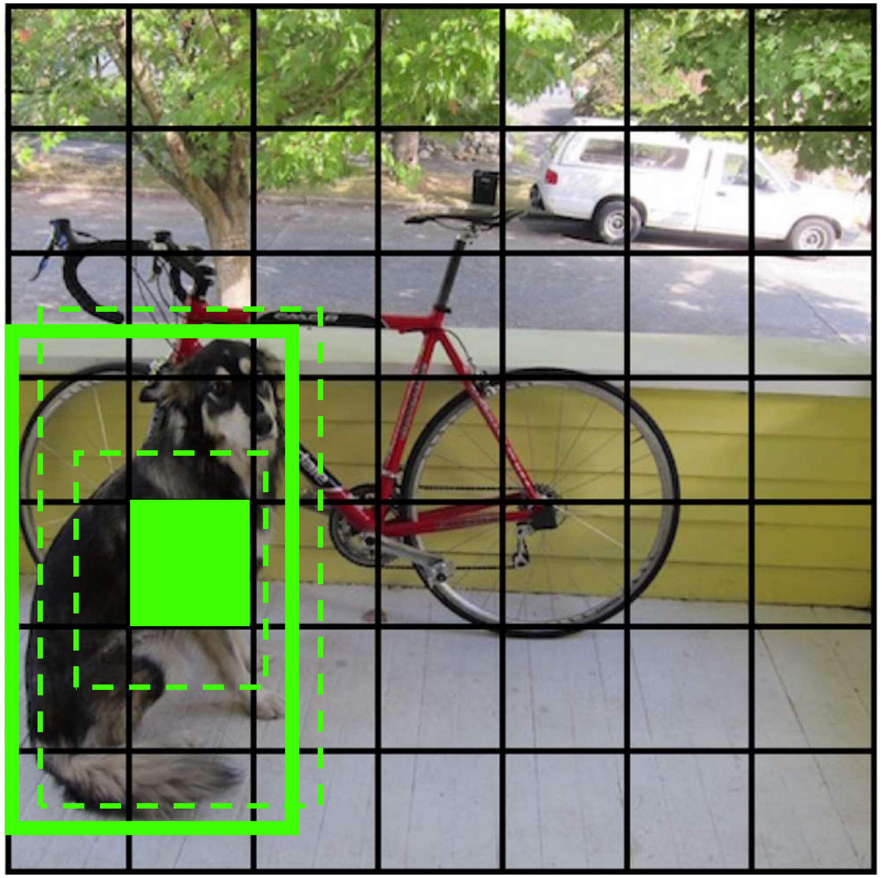
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](), 2015 (CVPR 2016)

YOLO – Training

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

Ground truth bounding box

Look at cell's predicted boxes



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

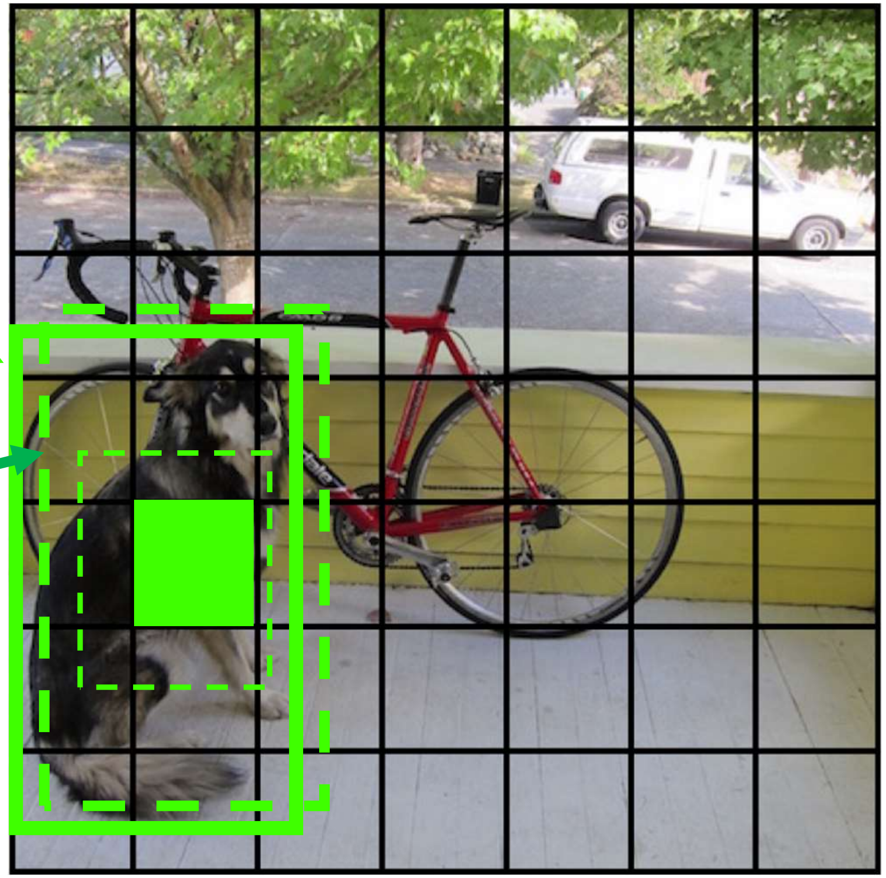
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](), 2015 (CVPR 2016)

YOLO – Training

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

Ground truth bounding box

Increase confidence score adjust



Slide 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)

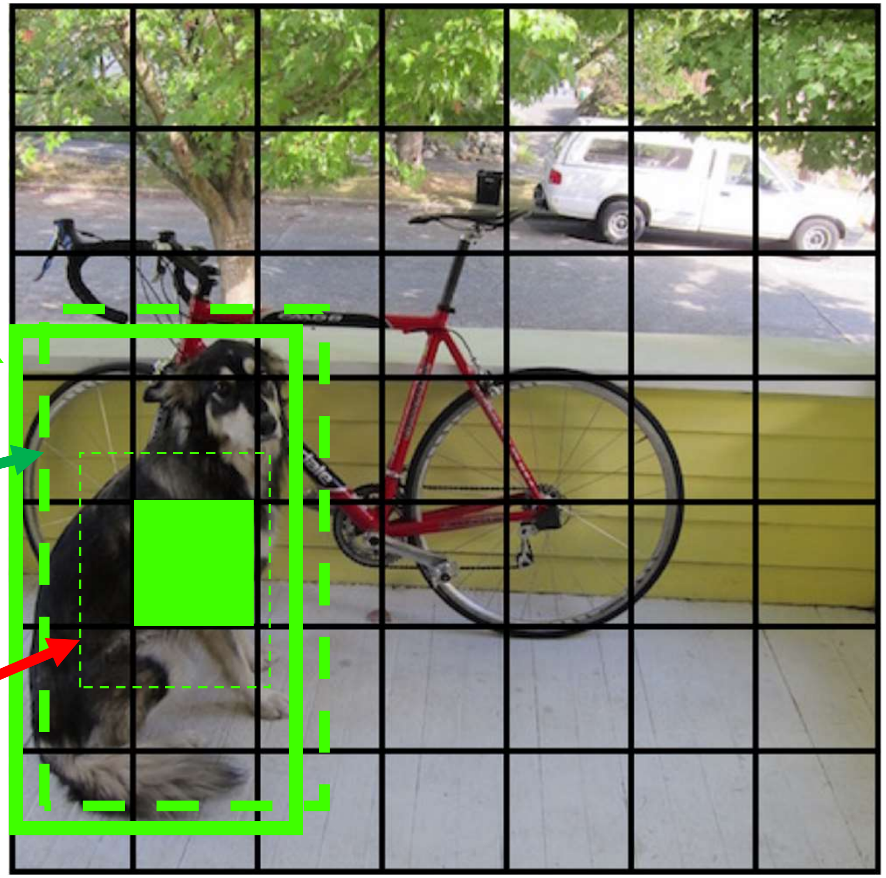
YOLO – Training

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

Ground truth bounding box

Increase confidence score adjust

Decrease confidence score don't adjust



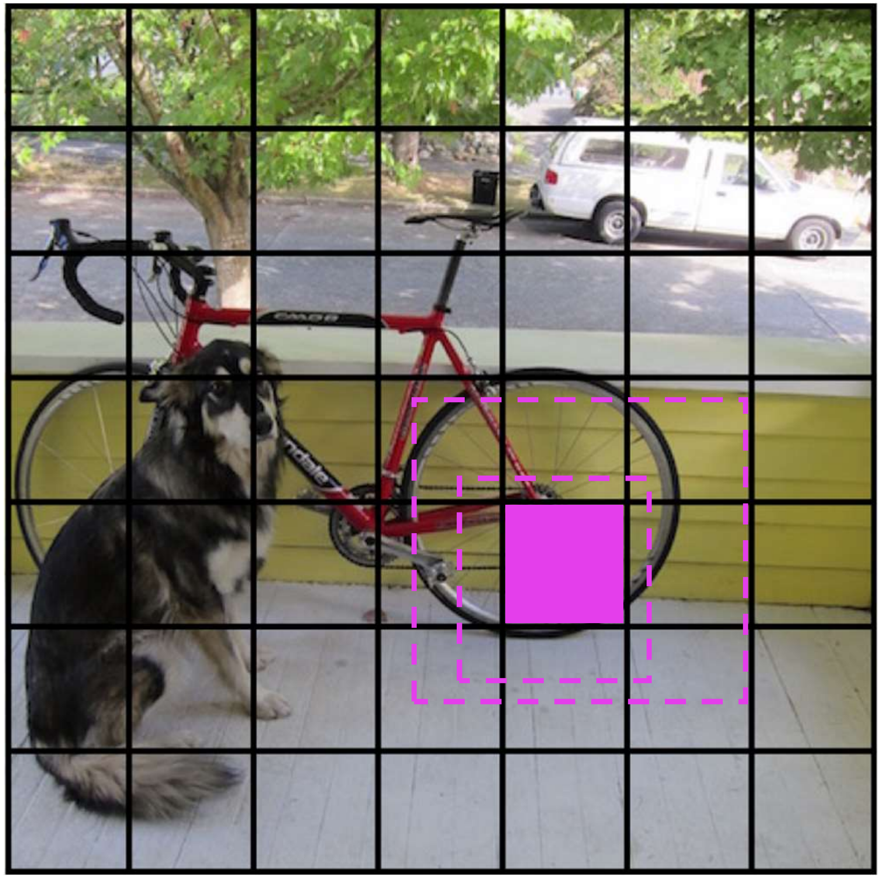
Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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



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

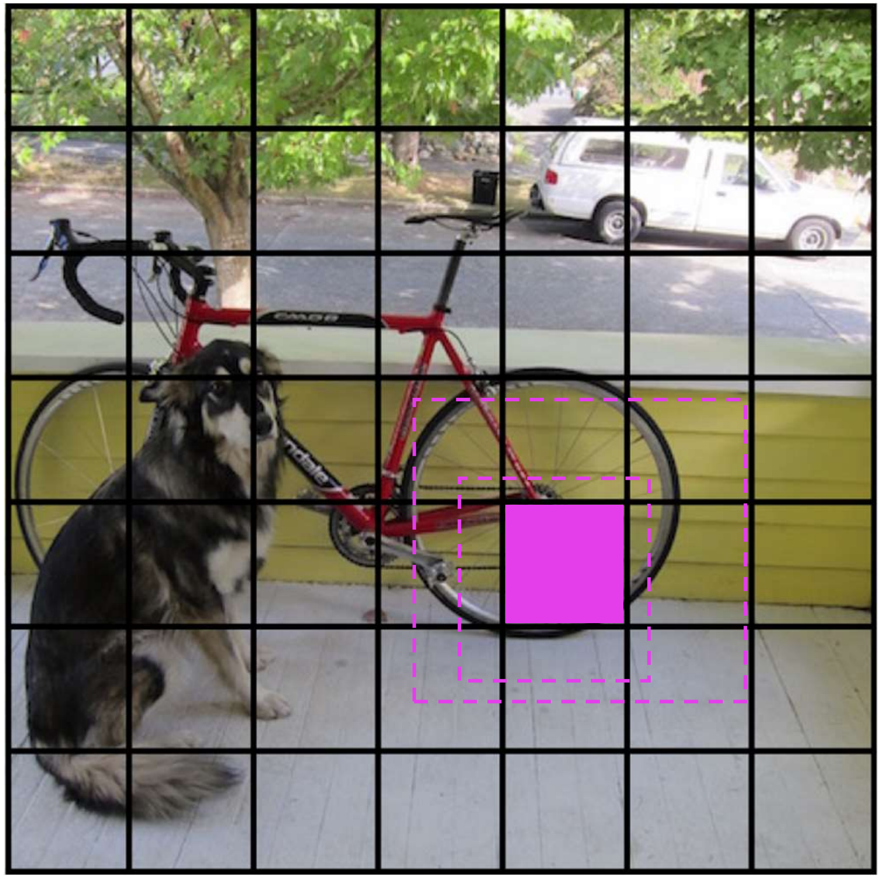
J. Redmon, S. Divvala, R. Girshick, A. Farhadi. [You only look once: Unified, real-time object detection](), 2015 (CVPR 2016)

YOLO – Training

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

A cell with no ground truth detection

Decrease confidence score



Slide 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)

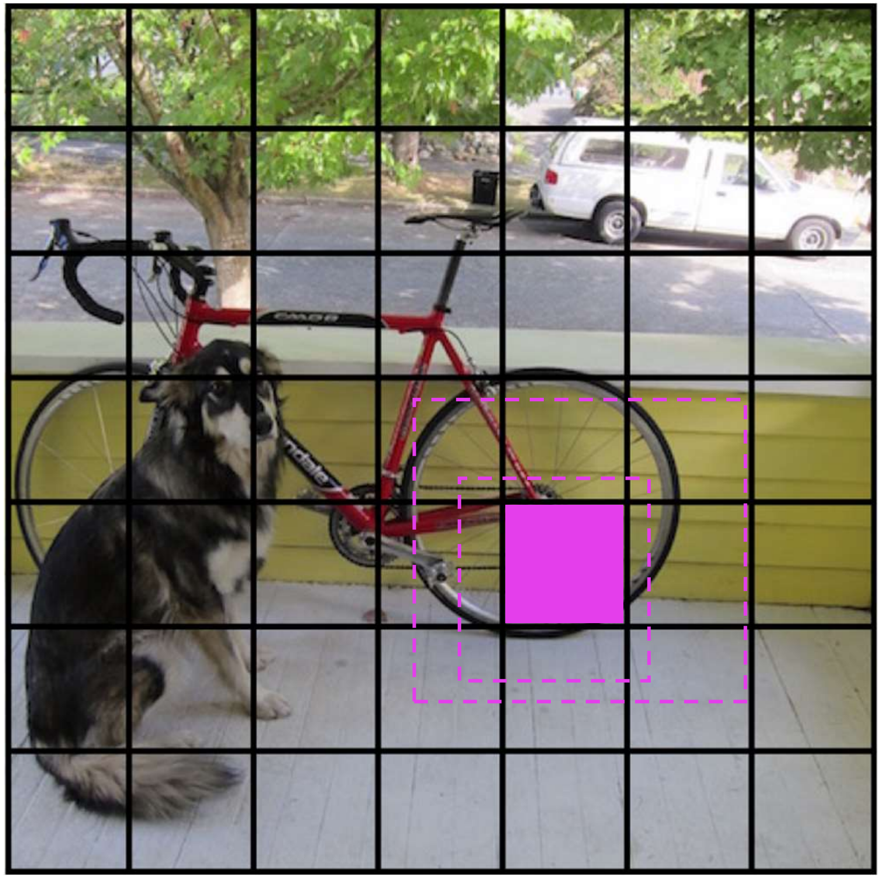
YOLO – Training

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

A cell with no ground truth detection

Decrease confidence score

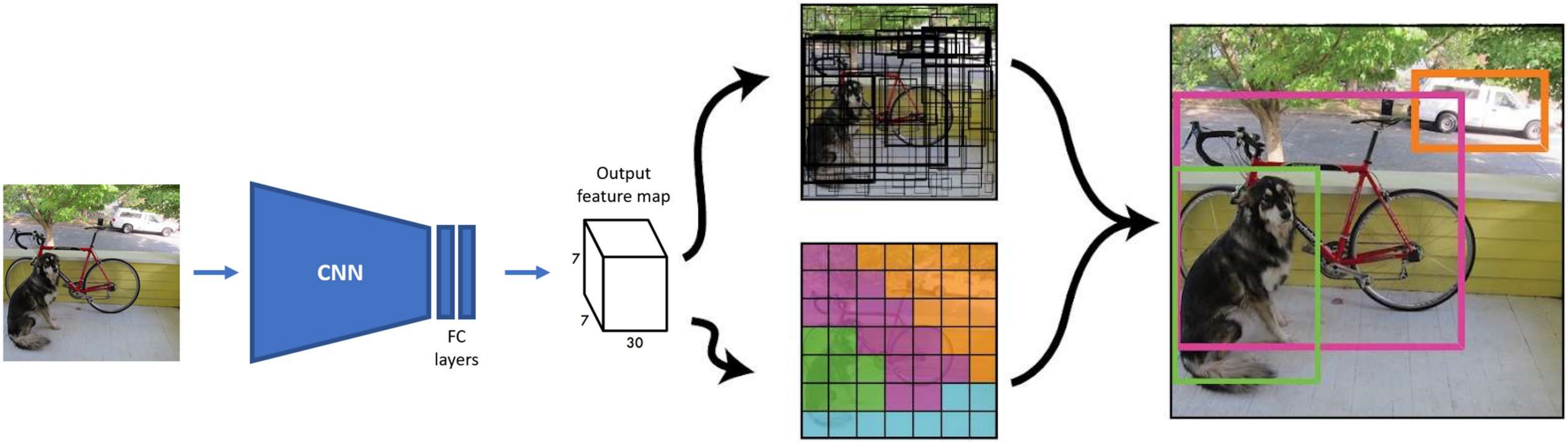
Don't adjust



Slide 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 – end-to-end training!

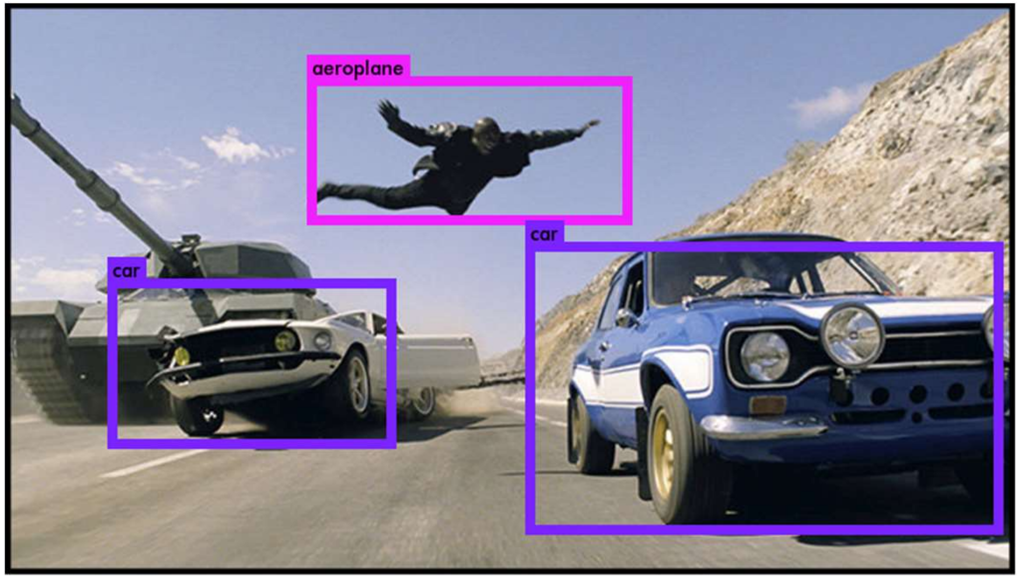


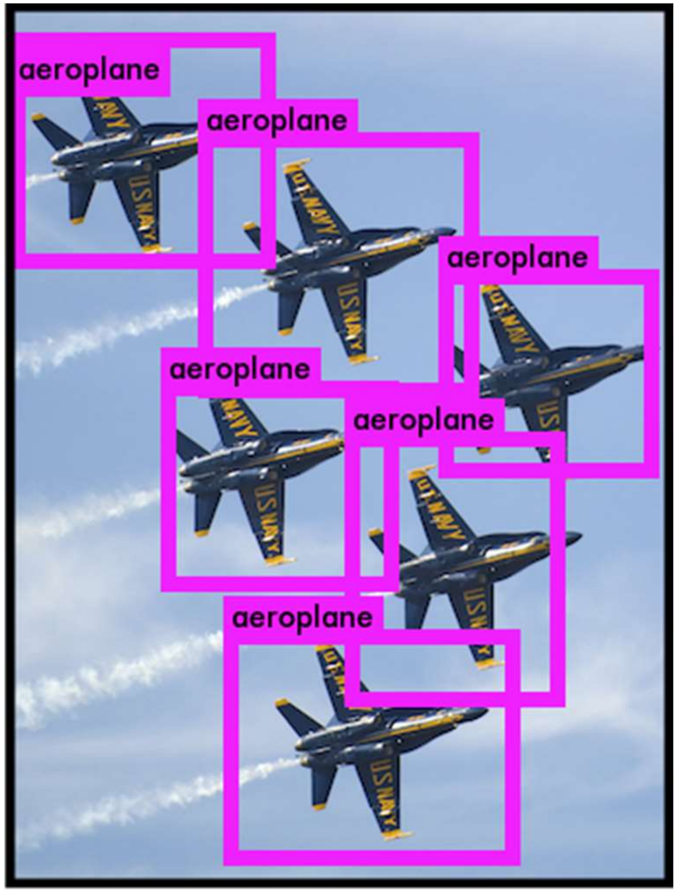
Slide credit: J. Redmon, S. Divvala, R. Girshick, A. Farhadi

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

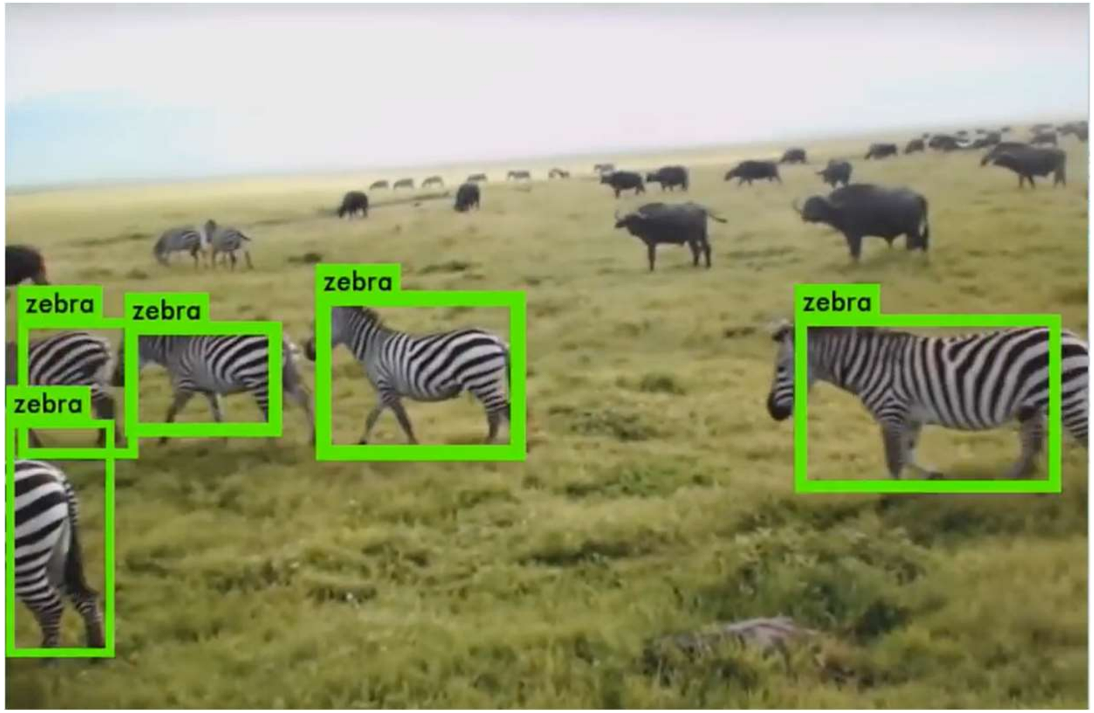


Image credit: https://pjreddie.com/darknet/yolov1/

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

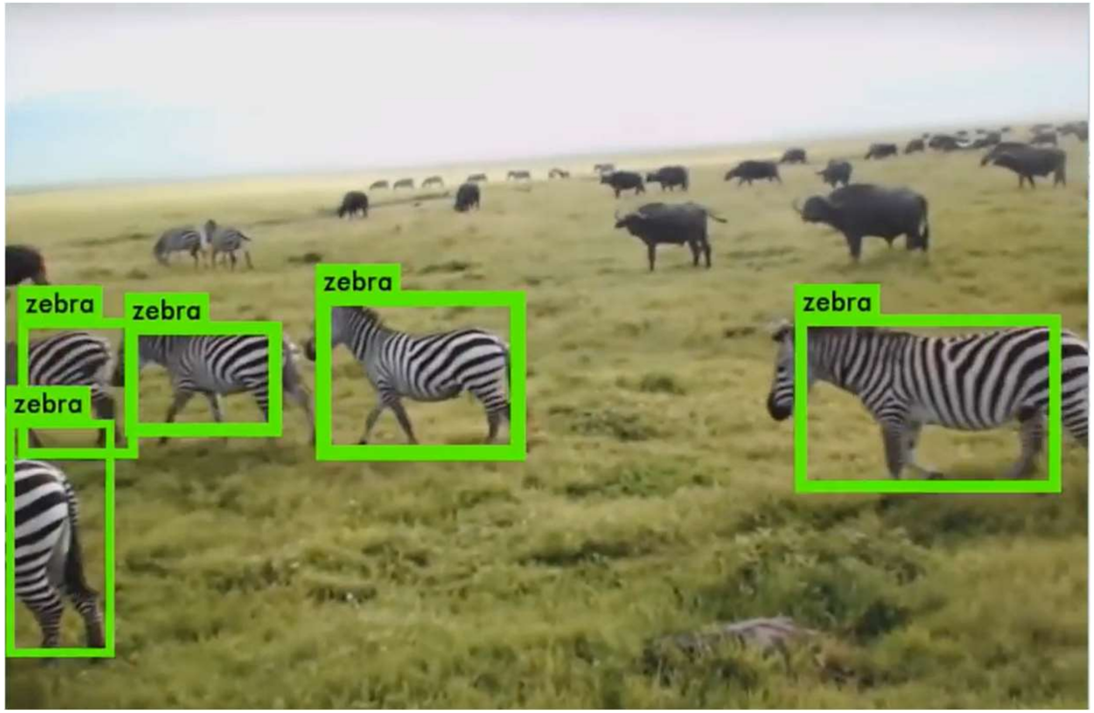
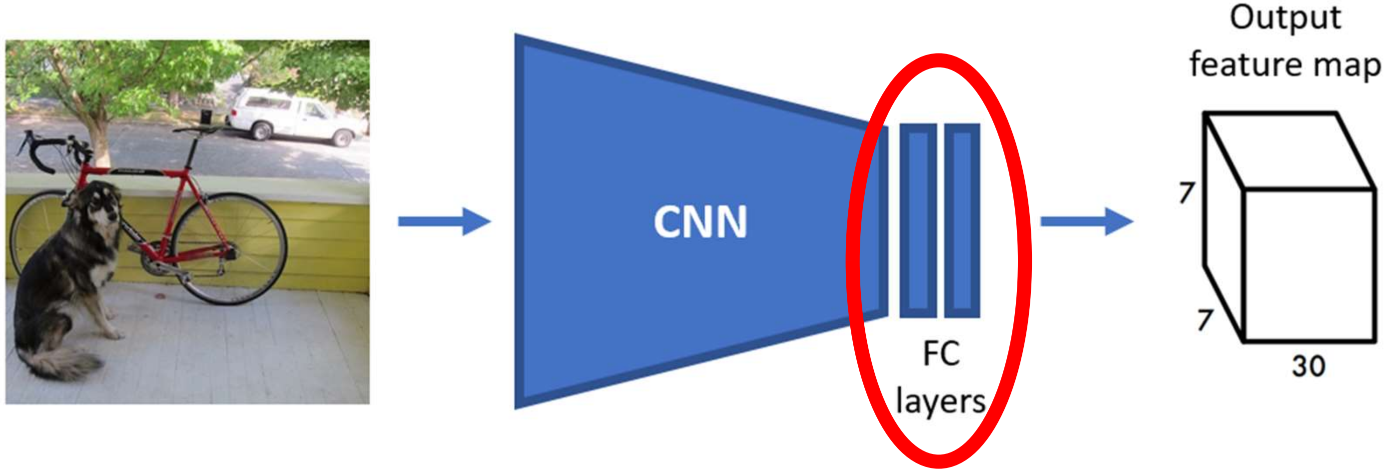


Image credit: https://pjreddie.com/darknet/yolov1/

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:

Change localization method!

Remove fc layers!

Increase features per grid cell!

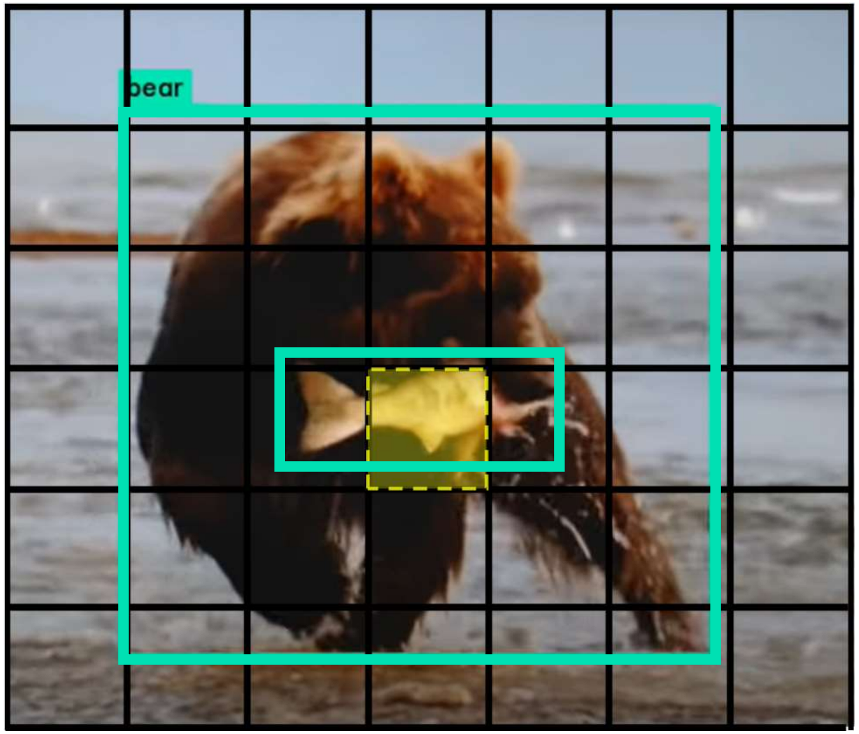


Image credit: https://pjreddie.com/darknet/yolov1/

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

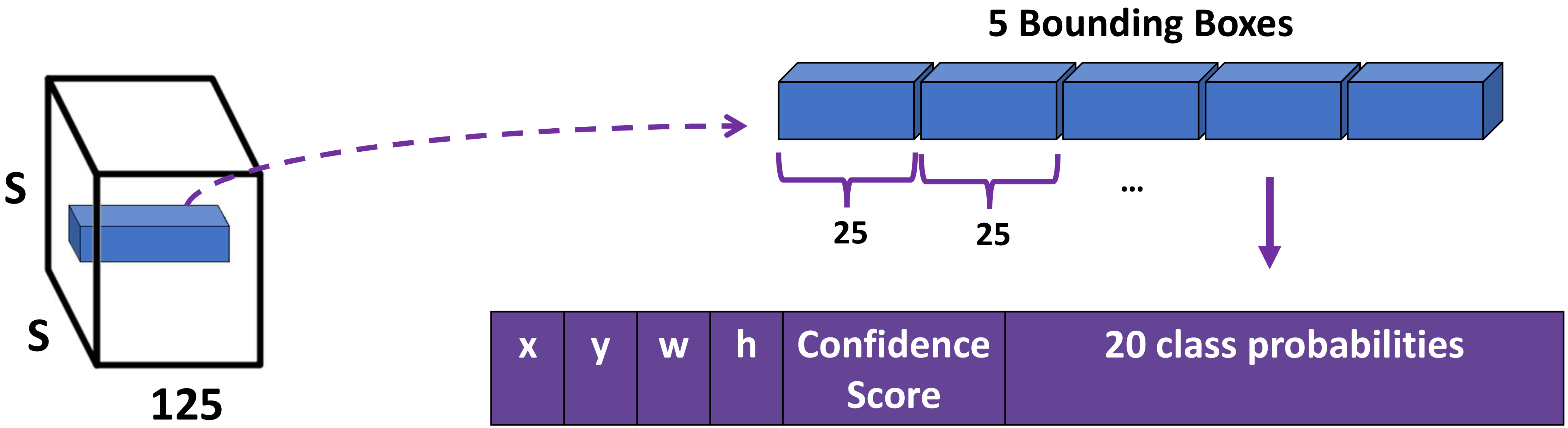
YOLOv2

- 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)

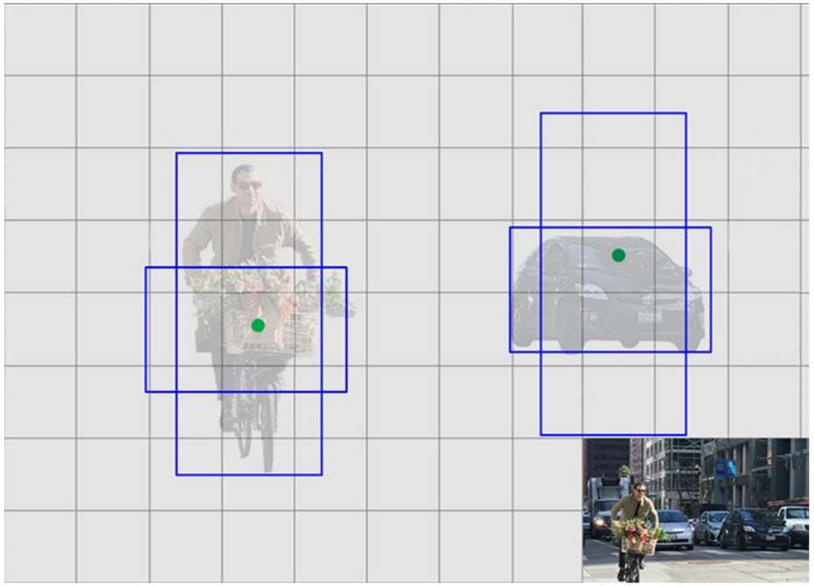
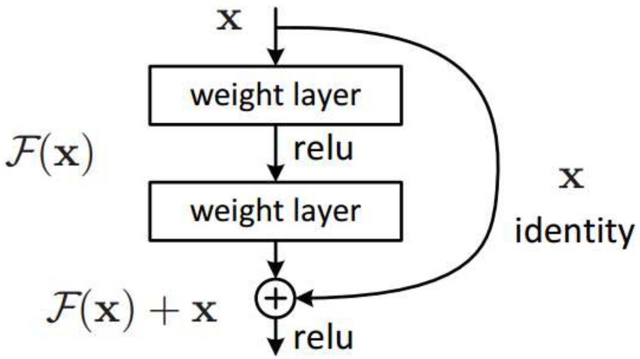


Image credit: medium

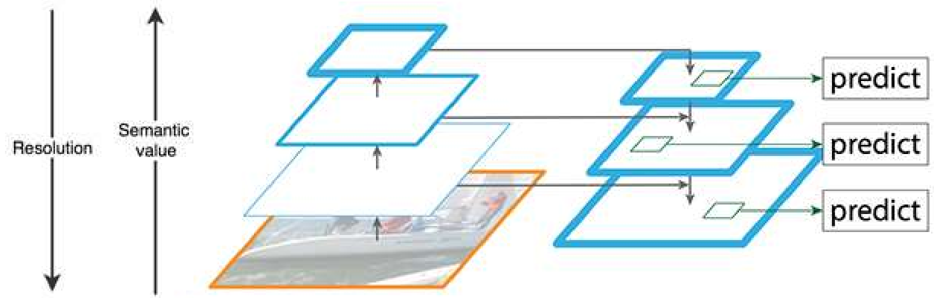
J. Redmon and A. Farhadi. Yolo9000: Better, faster, stronger (CVPR 2017)

There's always room for improvement!

- YOLOv3
  J. Redmon, A. Farhadi. Yolov3: An incremental improvement, 2018



Residual connections



FPN / Predicting in different scales

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)

There’s always room for improvement!

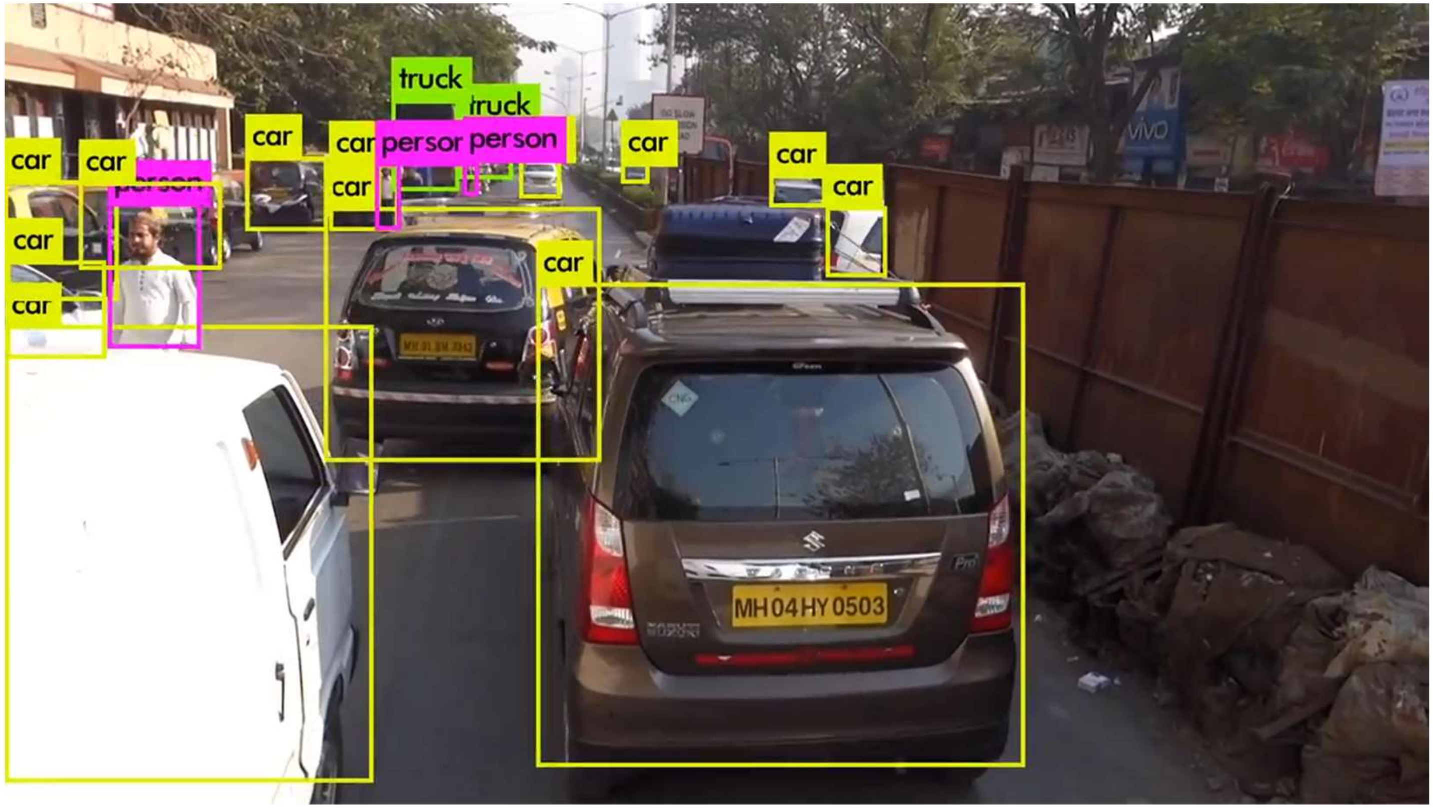
- YOLOv3
  J. Redmon, A. Farhadi. Yolov3: An incremental improvement, 2018
- YOLO v4
  A. Bochkovskiy, C. Wang, H. Liao. Yolov4: Optimal speed and accuracy of object detection (Feb. 2020)
- YOLOv5
  YOLOv5 by ultralytics (June 2020)

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)

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)
- ...



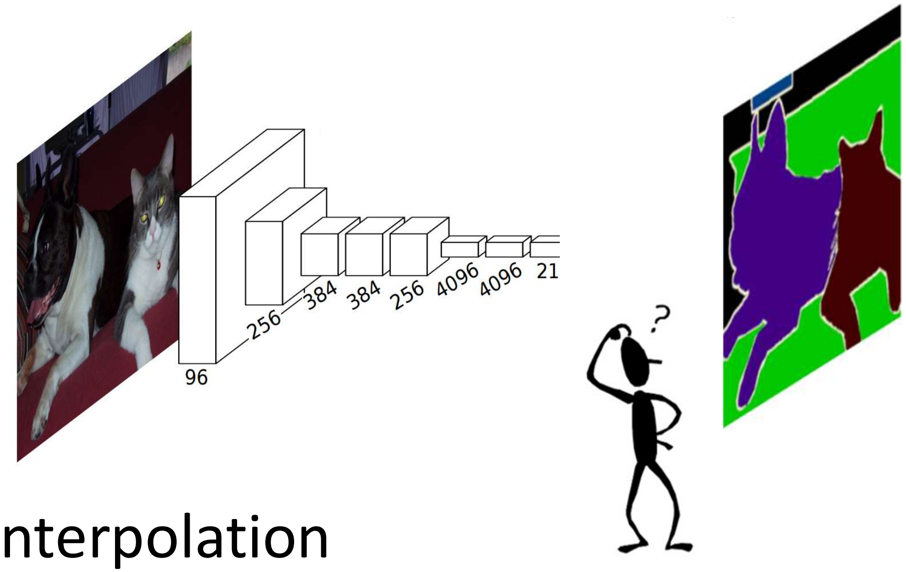
Video: YOLOv3

Semantic Segmentation

 Slide credit: Shai Bagon

Semantic Segmentation

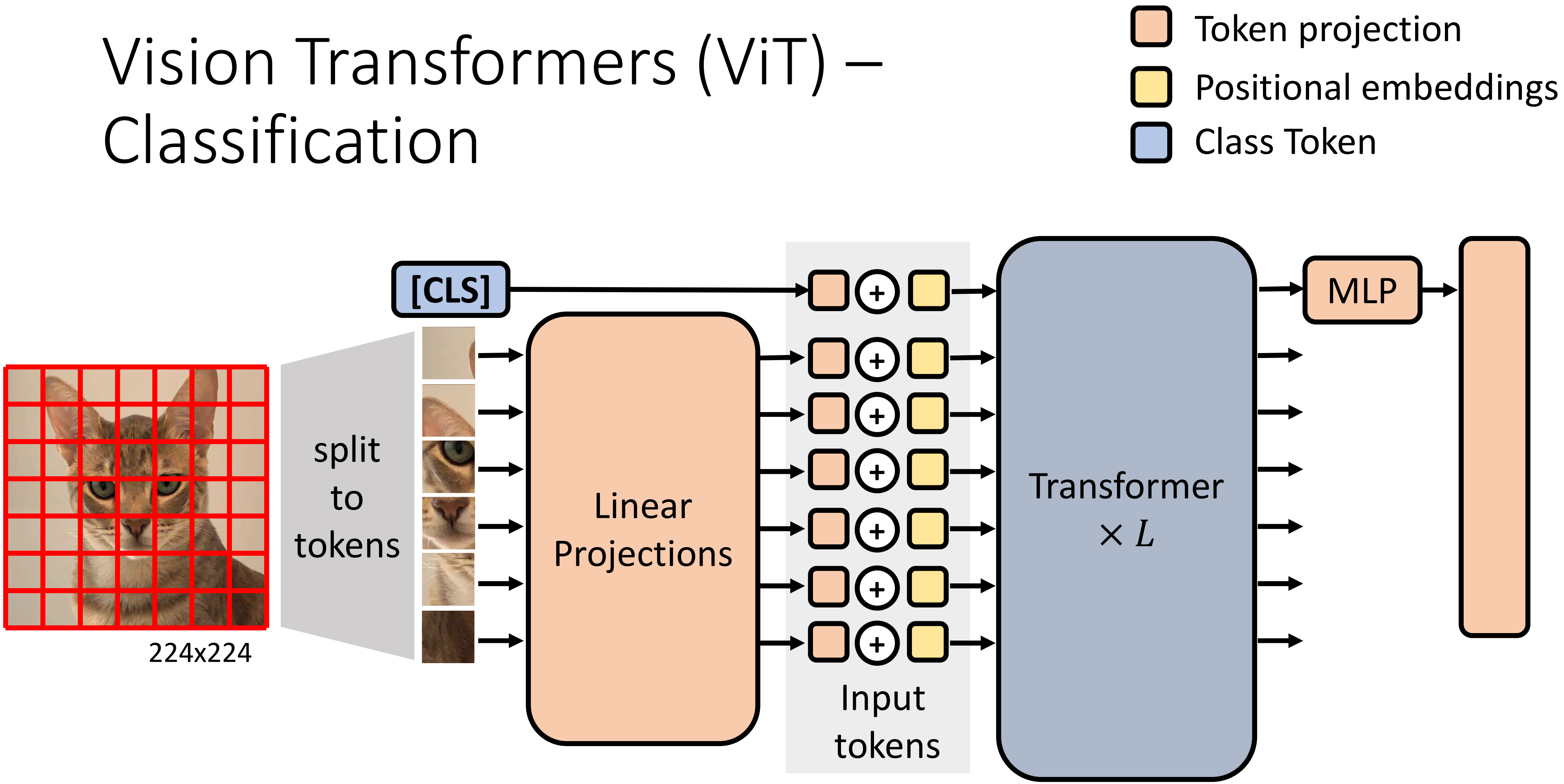
Resolution vs. Semantic information



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

 Slide credit: Shai Bagon

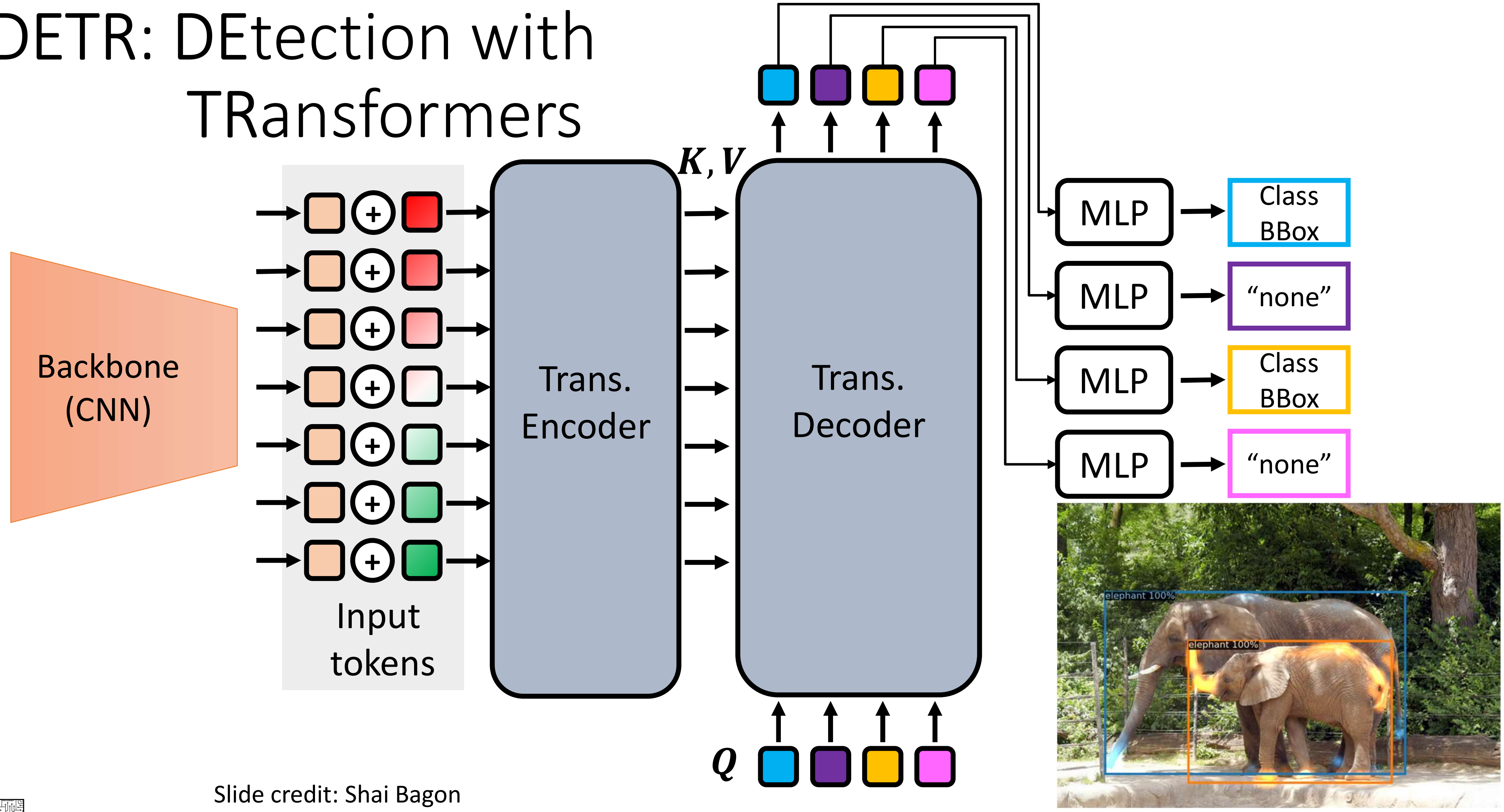
Vision Transformers (ViT) – Classification



DL4CV Weizmann Slide credit: Shir Amir

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)

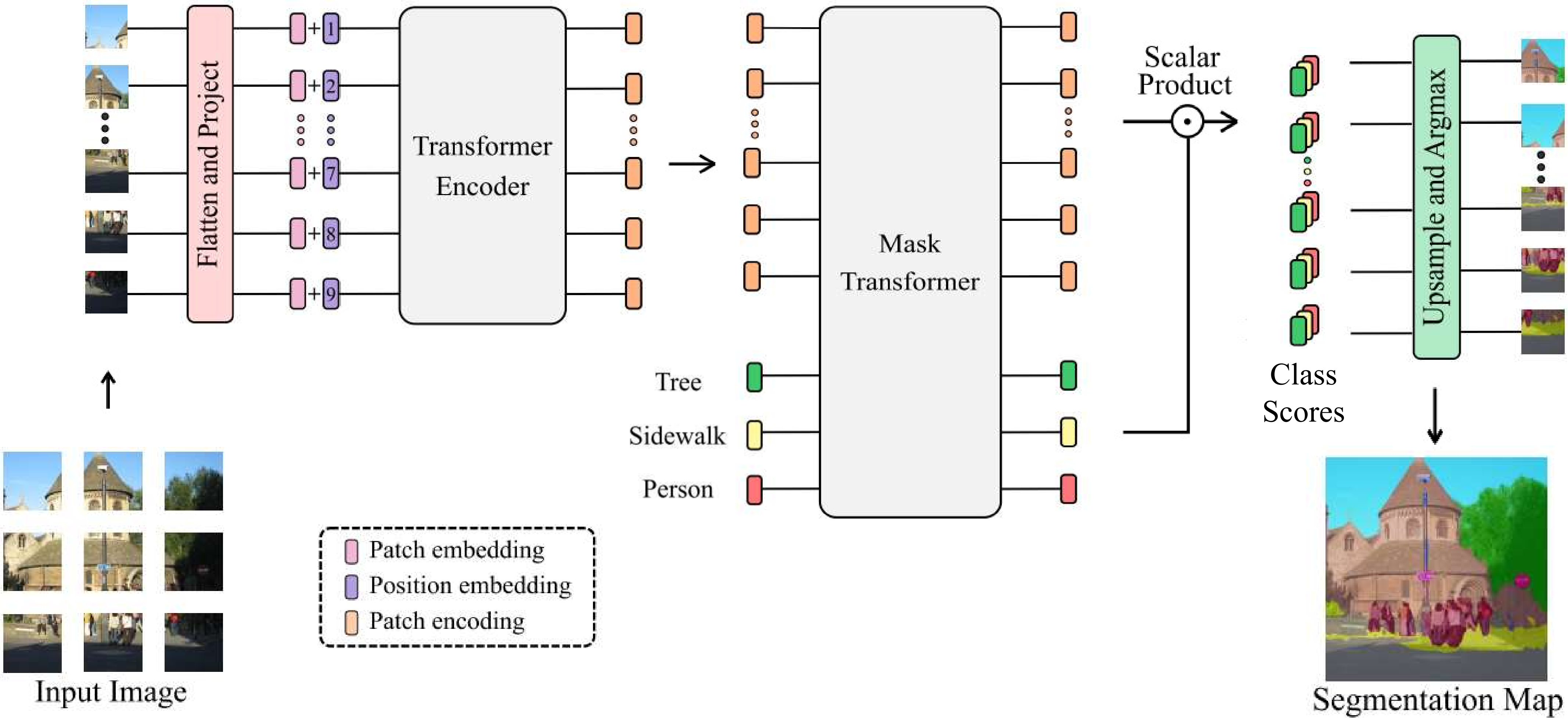
DETR: DEtection with TRansformers



Slide credit: Shai Bagon

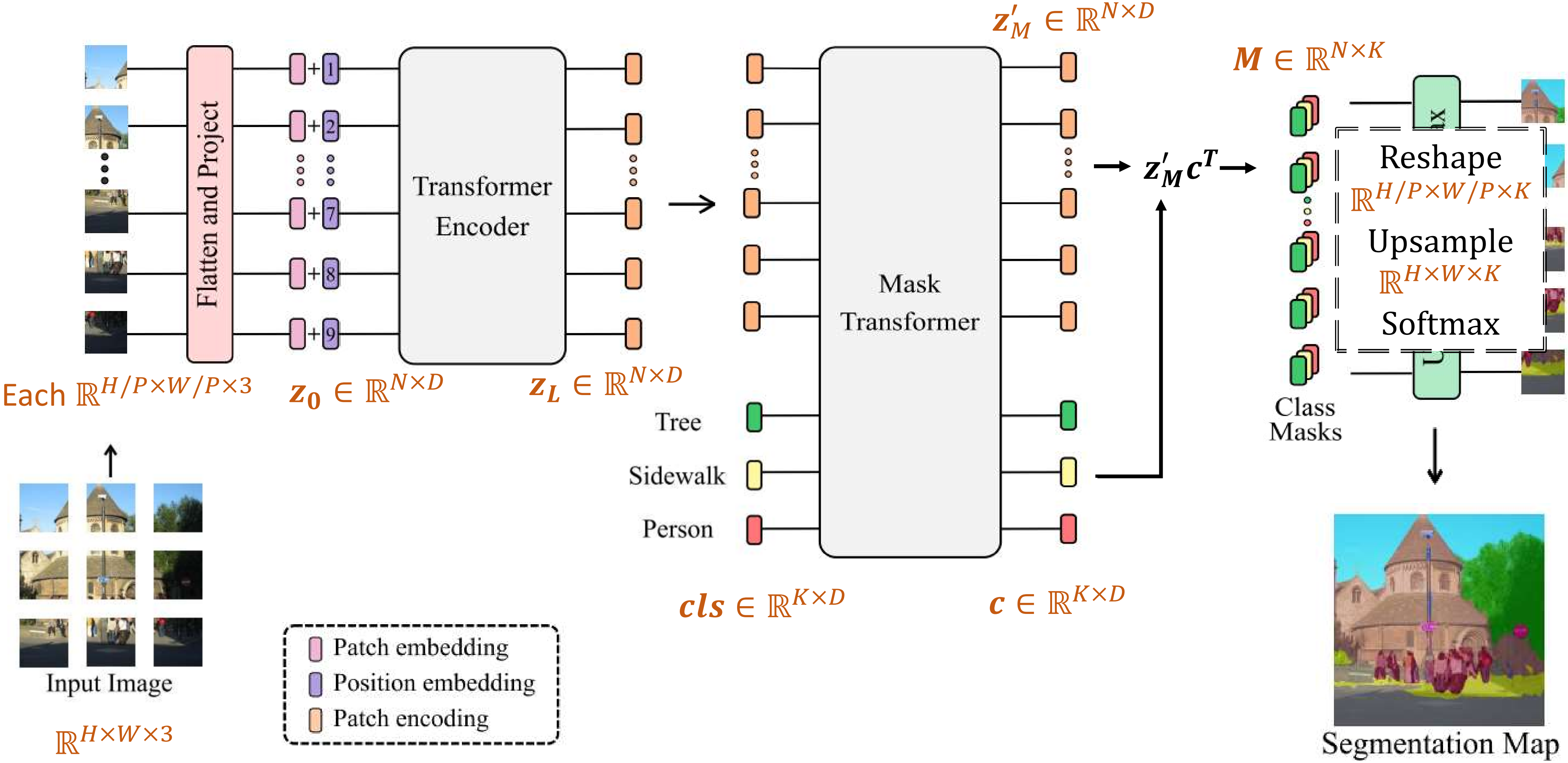
Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S. **DETR: End-to-End Object Detection with Transformers** (ECCV 2020)

Segmenter



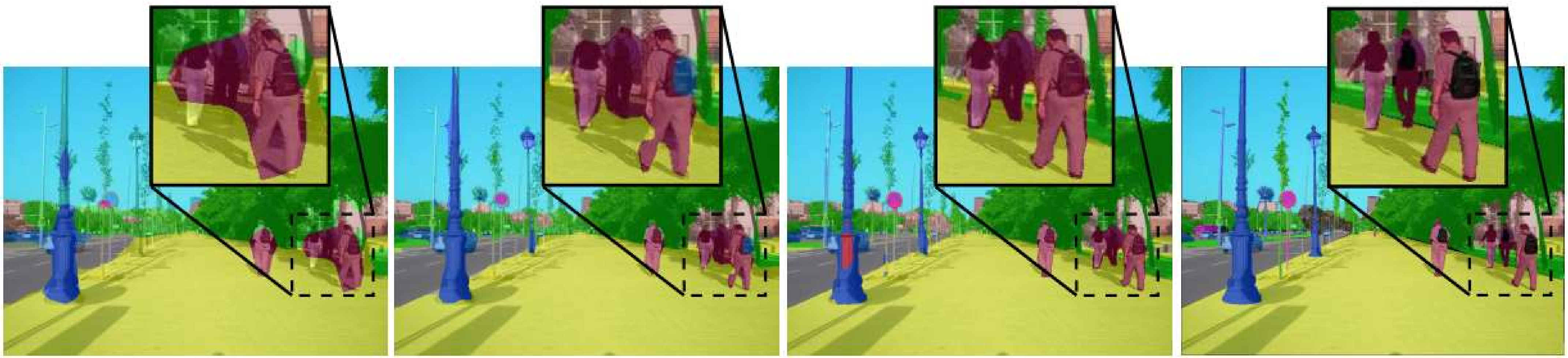
R. Strudel, R. Garcia, I. Laptev, C. Schmid. [Segmenter: Transformer for semantic segmentation](#) (ICCV 2021)

Segmenter



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

Segmenter



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

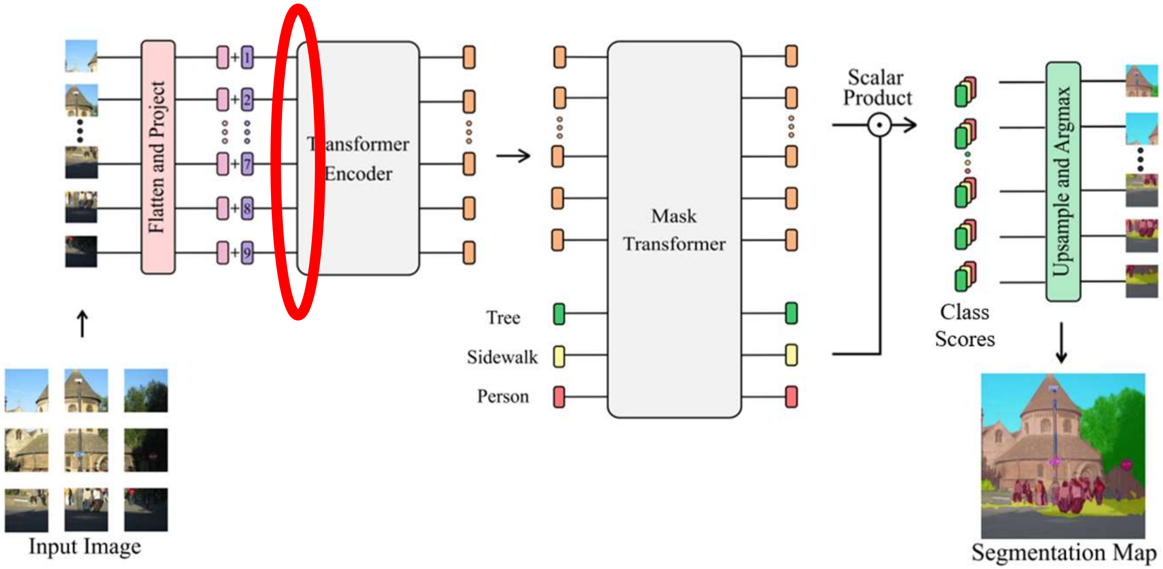
Fast

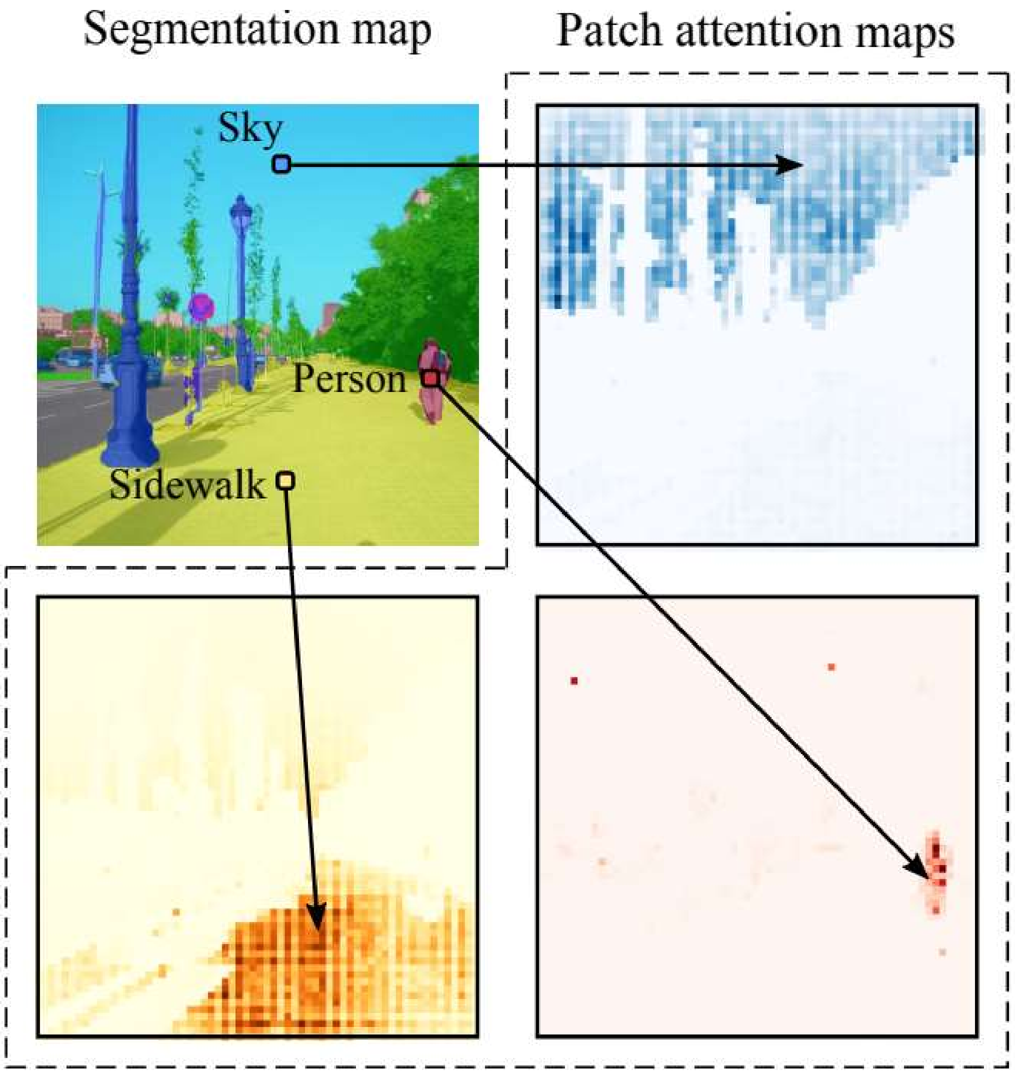
Low accuracy

High computational cost

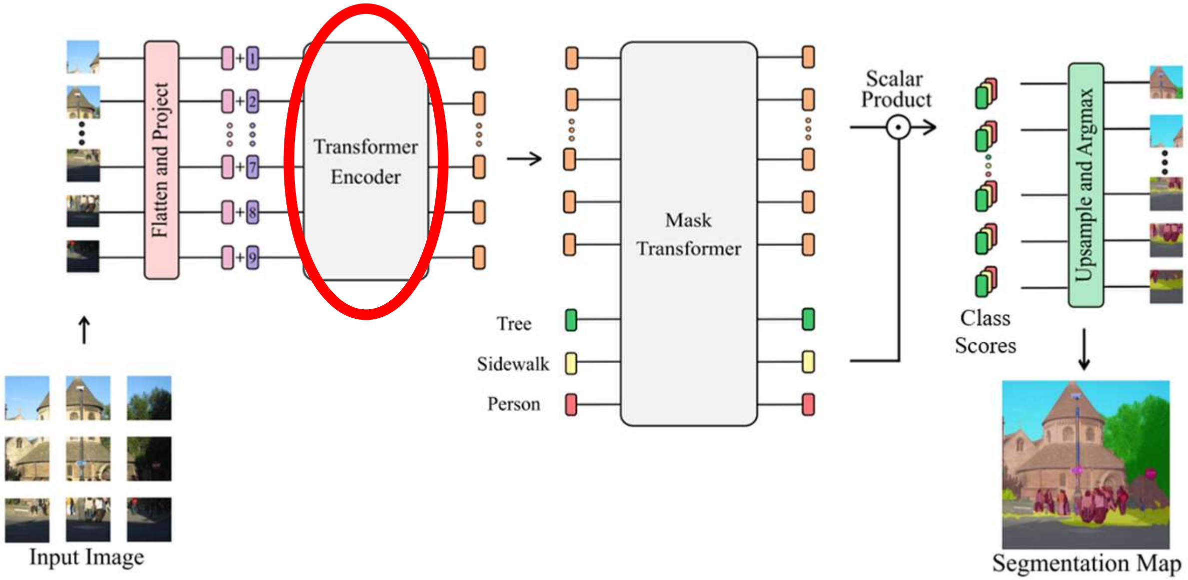
High accuracy

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

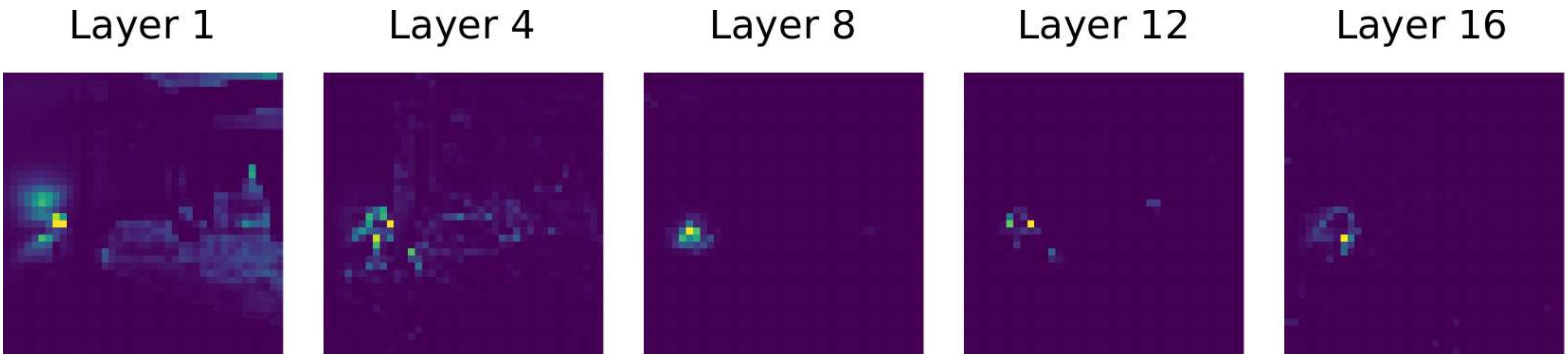




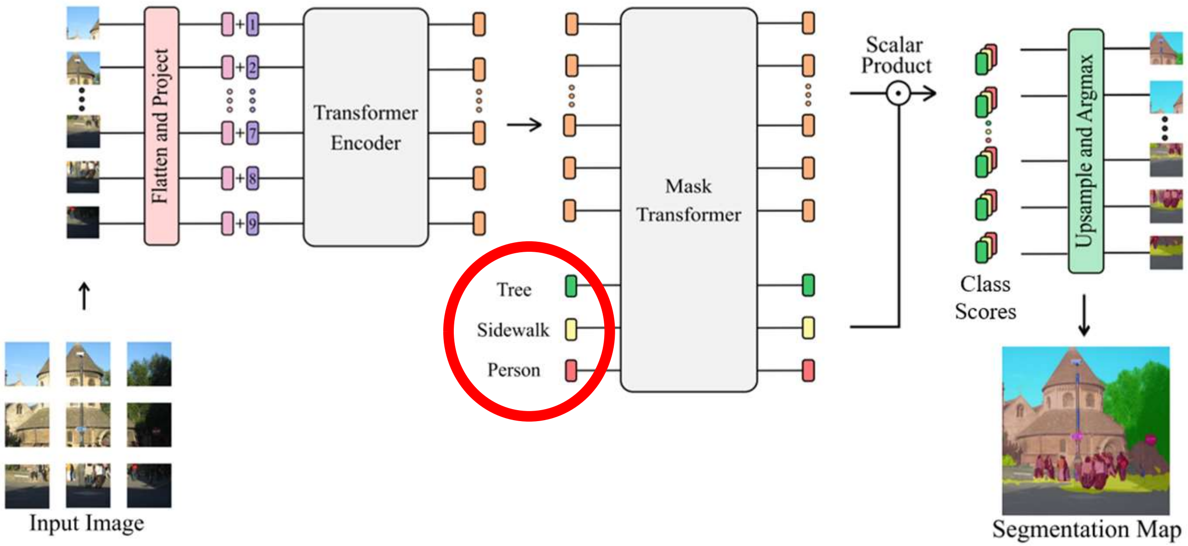
R. Strudel, R. Garcia, I. Laptev, C. Schmid. [Segmenter: Transformer for semantic segmentation] (ICCV 2021)



Input

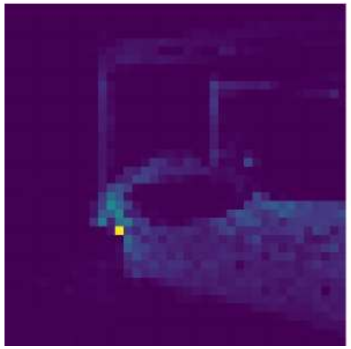


R. Strudel, R. Garcia, I. Laptev, C. Schmid. Segmenter: Transformer for semantic segmentation (ICCV 2021)

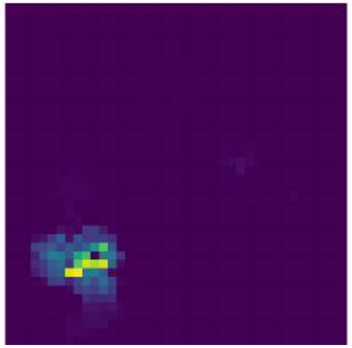


Prediction

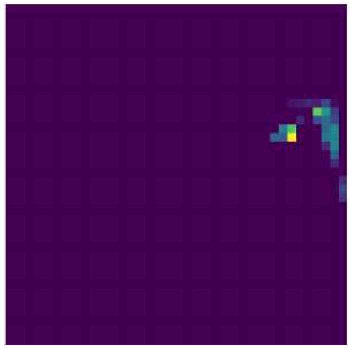
CLS 7 (bed)



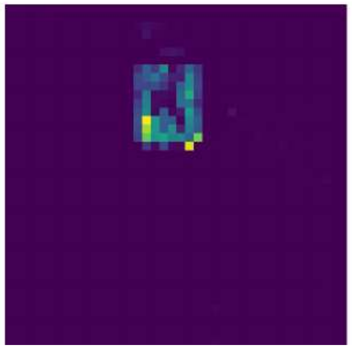
CLS 15 (table)



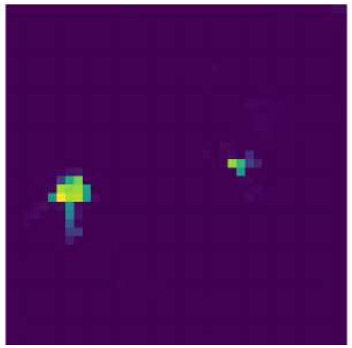
CLS 18 (curtain)



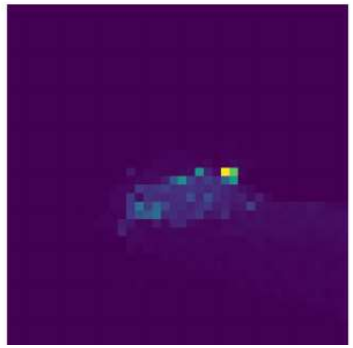
CLS 22 (painting)



CLS 36 (lamp)

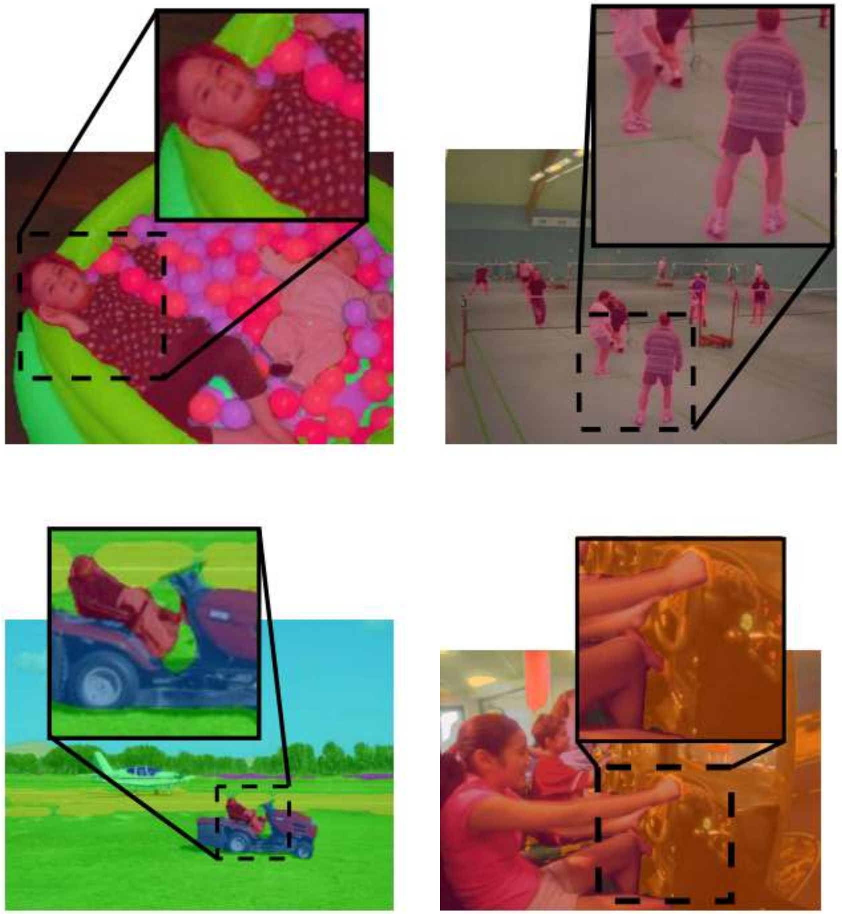


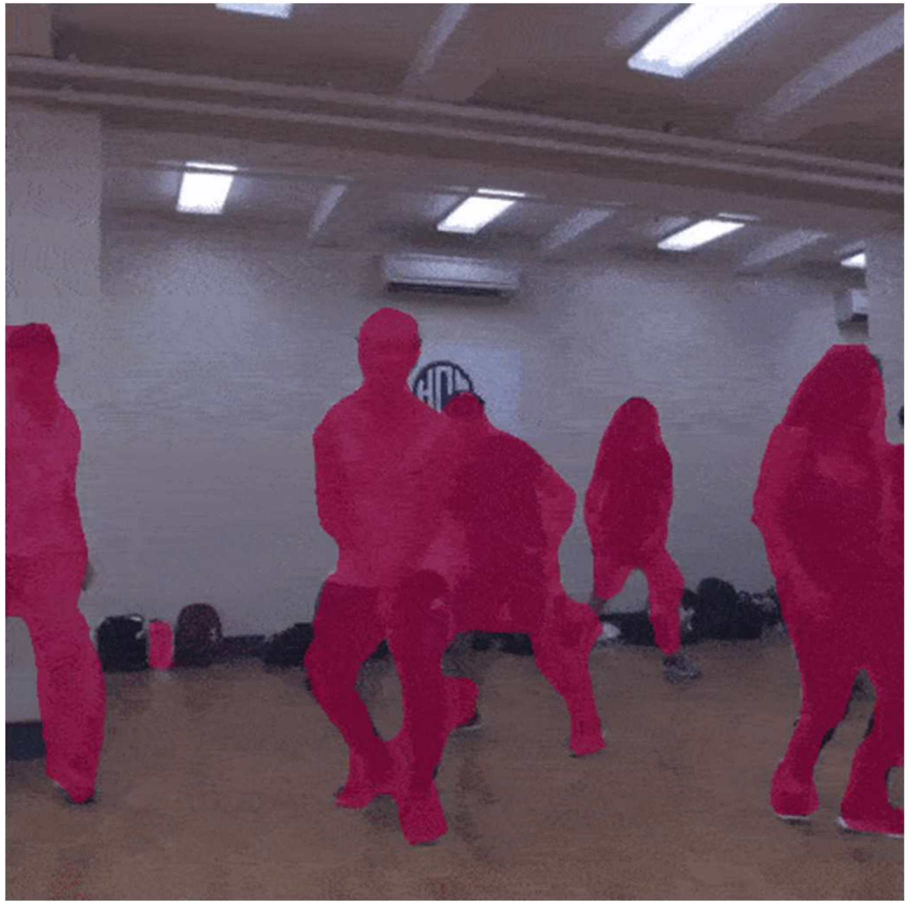
CLS 57 (pillow)



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)

Next Week:

Self Supervision

Tali Dekel