Object Detection With YOLO

DICE ANALYTICS

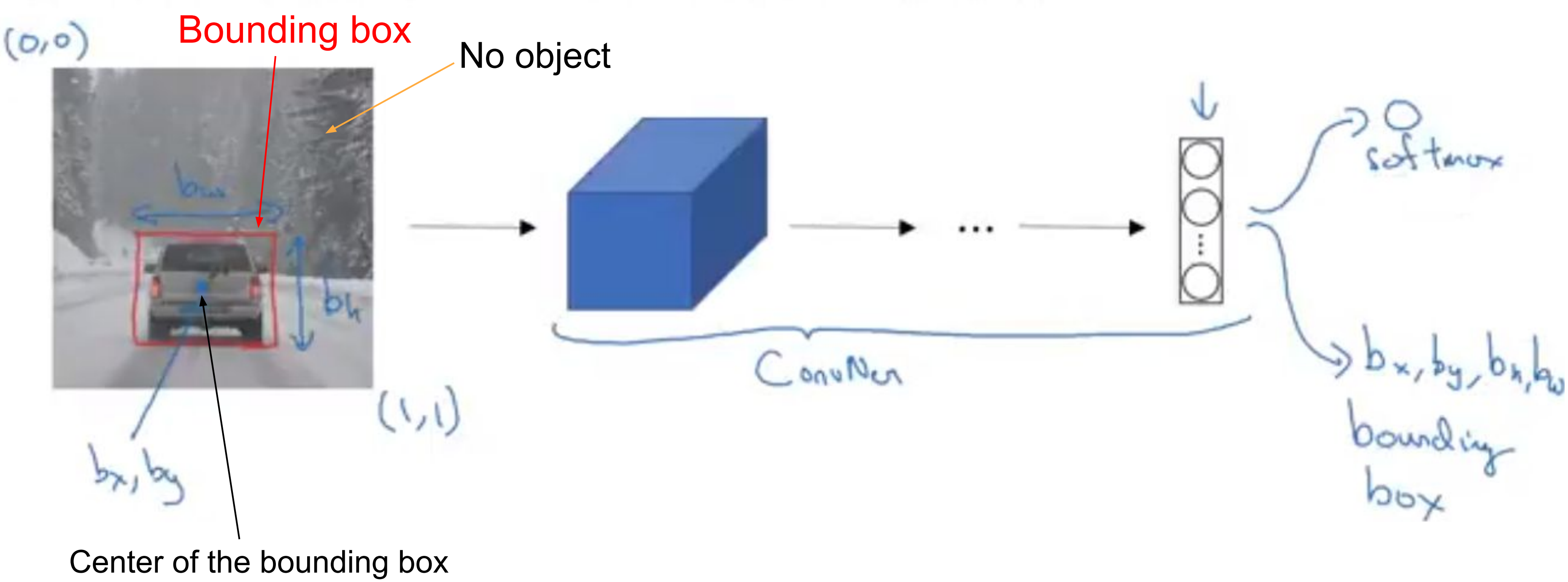
Classification With Localization

Classification

Classification with
localization

Andrew Ng

Classification With Localization



Andrew Ng

Classification With Localization

What if there is no object present in an image?

Classification With Localization

What if there is no object present in an image?

Class probability (Pc): If the object is one of the classes, then this value should be 1, otherwise this value should be zero (i.e. for background).

Then, for a single object, the label vector **y** becomes something like this:

$$\mathbf{y} = [Pc, bx, by, bw, bh, \underbrace{class1, class2, \ldots}_{\text{softmax}}]$$

Classification With Localization

$$\begin{bmatrix} 1 \\ b_x \\ b_y \\ b_h \\ b_w \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} 0 \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \\ ? \end{bmatrix}$$

← Don't care

Object detection

One step further: Detecting multiple objects in a single image

Andrew Ng

Object detection

Sliding windows detection algorithm

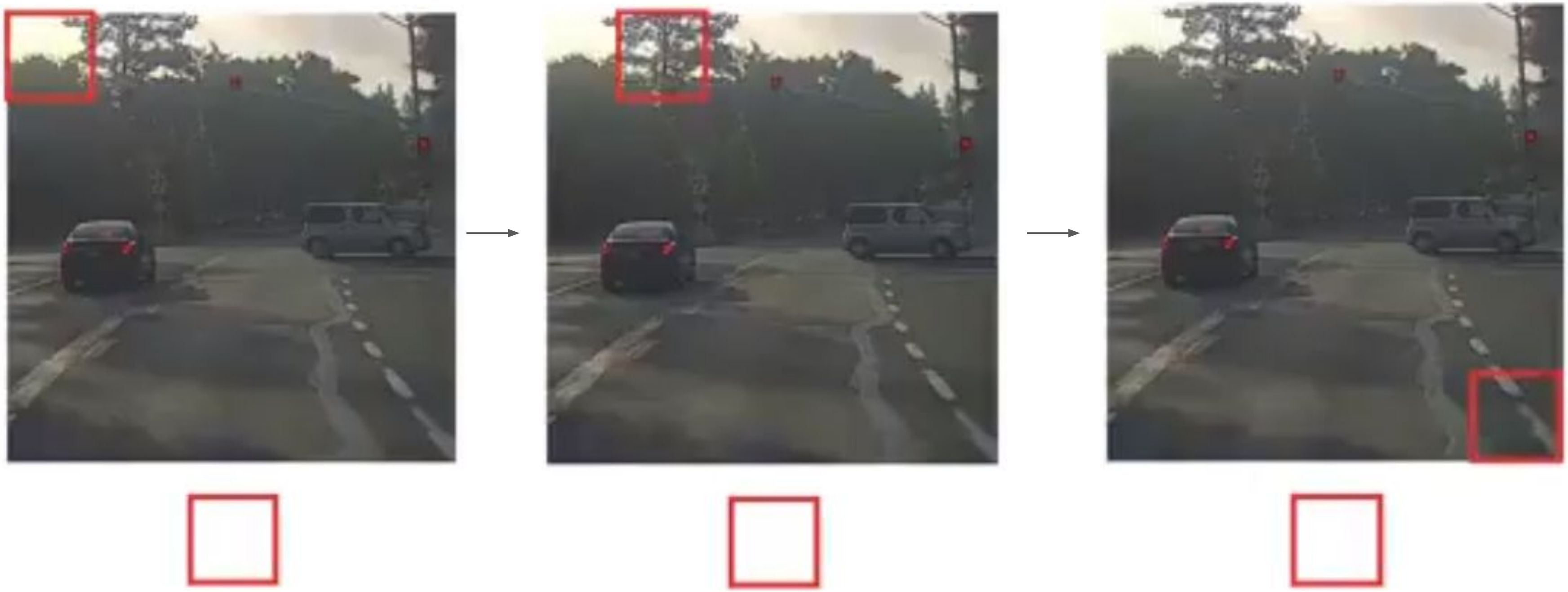
Training set:

x	y
	1
	1
	1
	0
	0

Andrew Ng

Object detection

Sliding windows detection algorithm



 → ConvNet → y

Andrew Ng

Object detection

Sliding windows detection algorithm

Andrew Ng

Object detection

Sliding windows detection algorithm

Problem with Sliding window detection algorithm?

Object detection

Sliding windows detection algorithm

Problem with Sliding window detection algorithm?

- Fix sized bounding boxes (inaccurate)
- Slow speed

Object detection

HOG: Histograms of oriented gradients

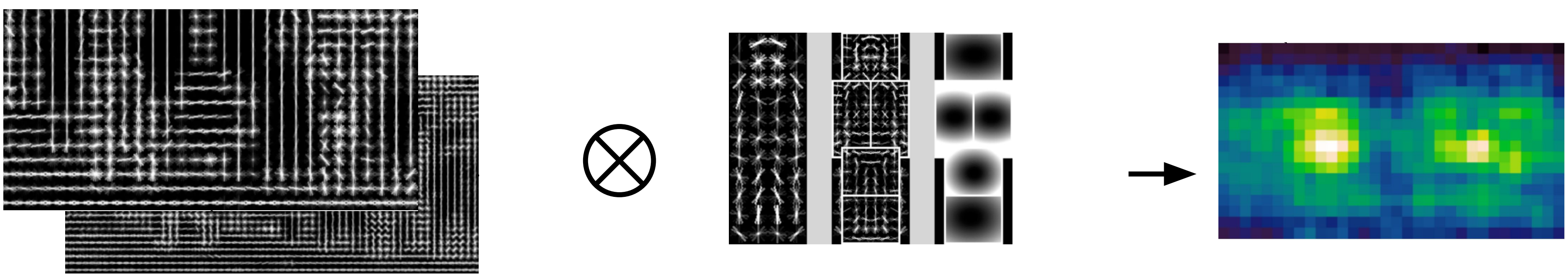
Direction of maximum
color gradient

Siraj Raval

Object detection

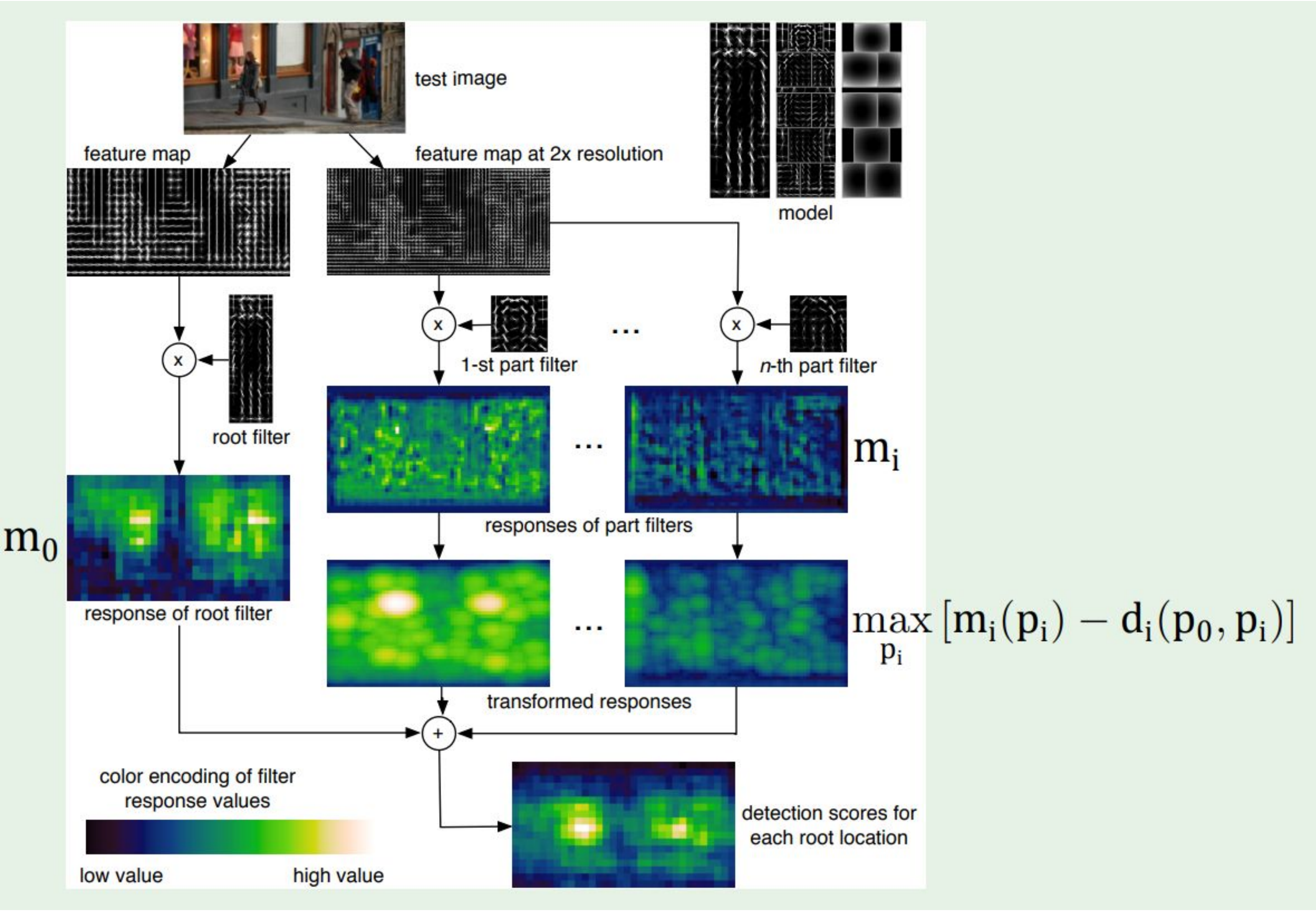
DPM: Deformable Parts Model

HOG feature maps of an image are convolved with part filters to give heat maps of object occurrences.



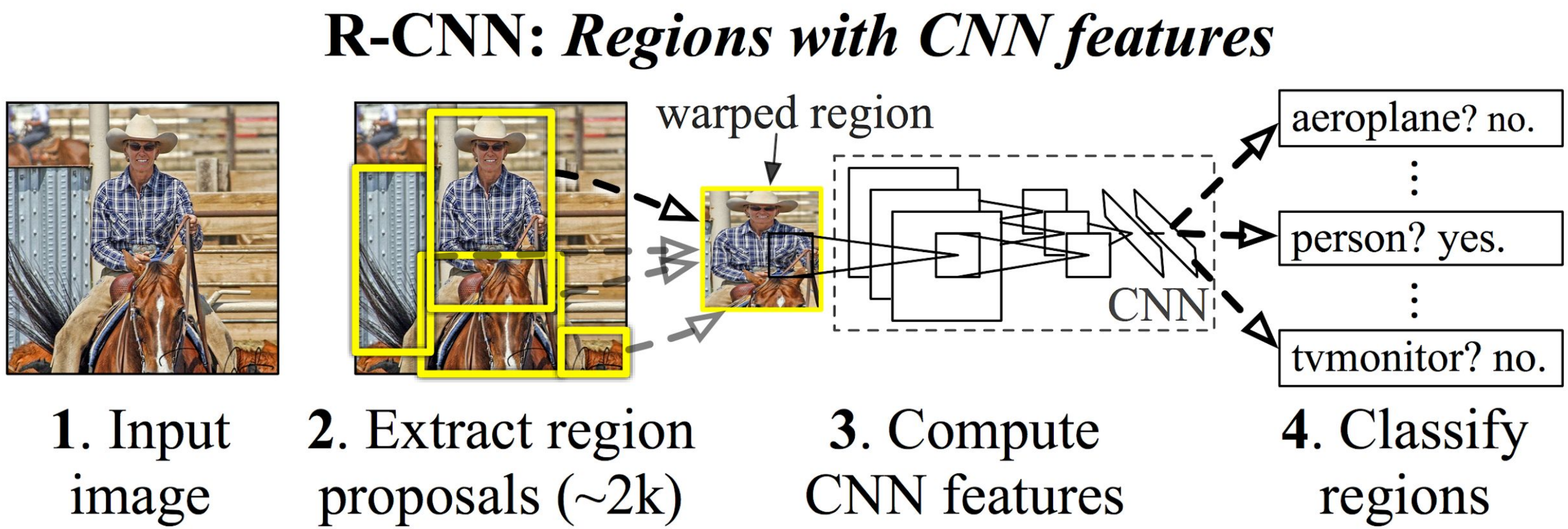
Object detection

DPM: Deformable Parts Model



Ross Girshick

Object detection



Joesph Redmon

Accurate object detection is slow!

	*Pascal 2007 **mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img

*Visual Object Classes dataset; **mean Average Precision

Accurate object detection is slow!

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img

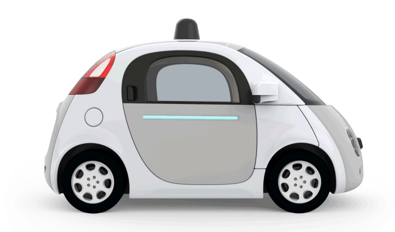
⅓ Mile, 1760 feet

Joesph Redmon

DICE ANALYTICS

Accurate object detection is slow!

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img

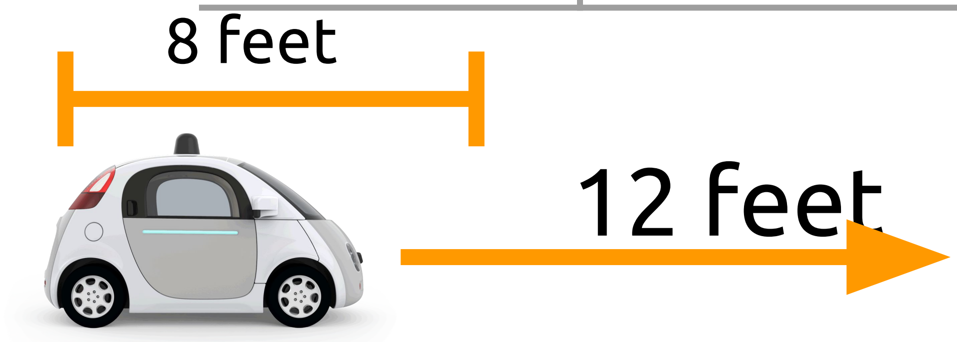


176 feet

Joesph Redmon

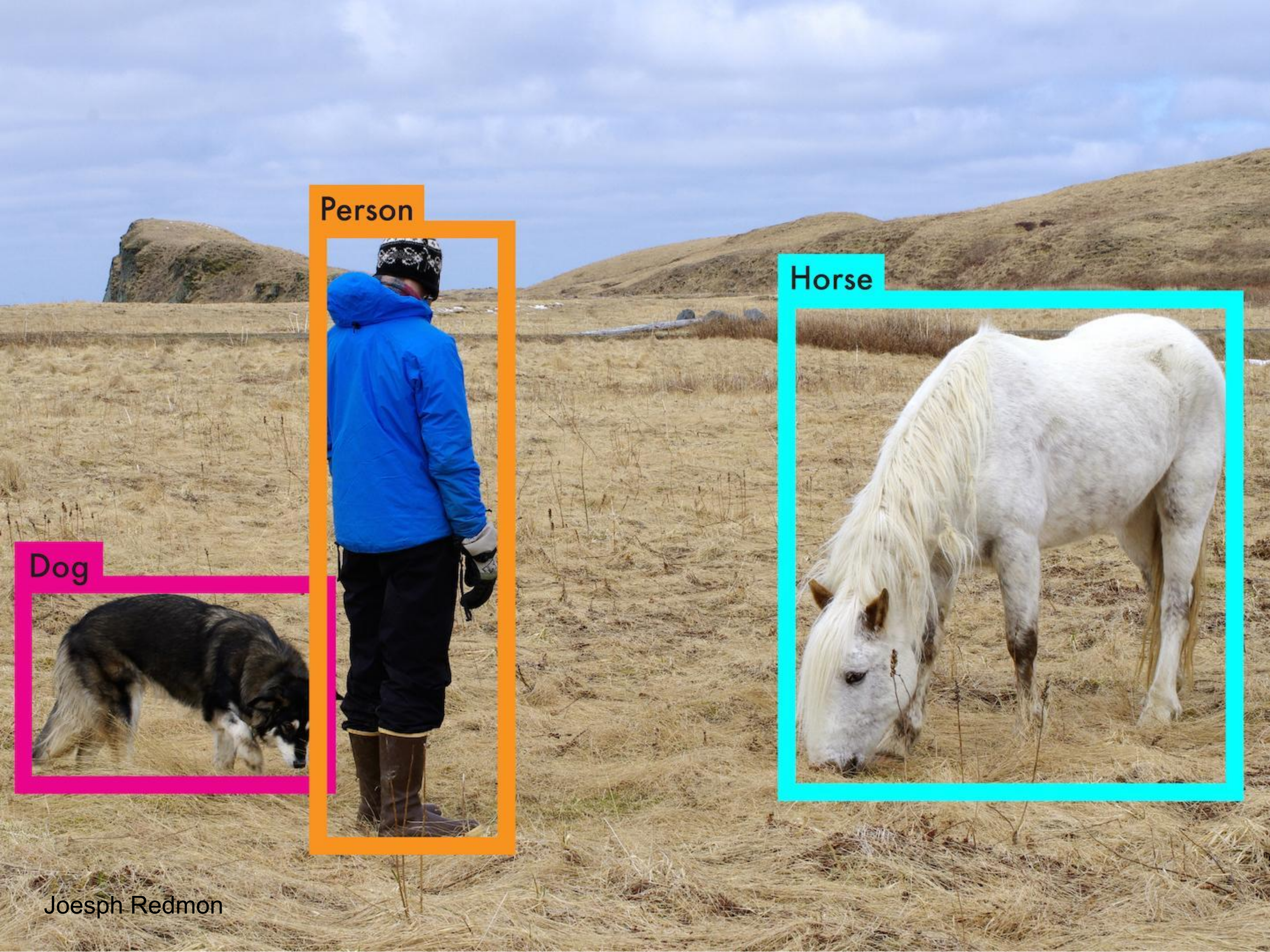
Accurate object detection is slow!

	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img



Joesph Redmon


Joseph Redmon
Ross Girshick
Santosh Divvala
Ali Farhadi
Dog
Most Accurate Real-Time Detector 2016
Fastest object detector in the literature 2016
"You Only Look Once"
Real-Time Detection



Joesph Redmon

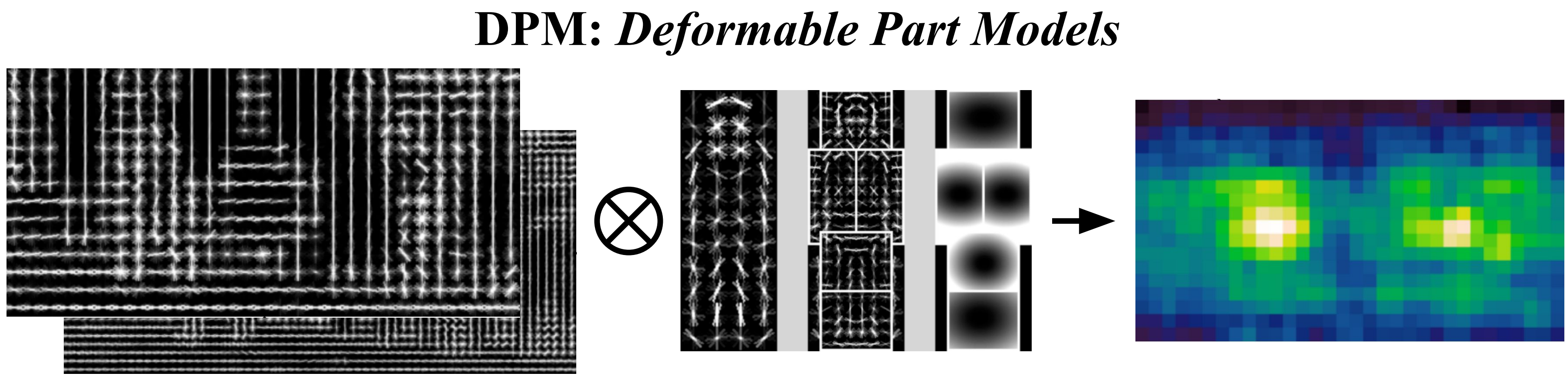
Accurate object detection is slow!

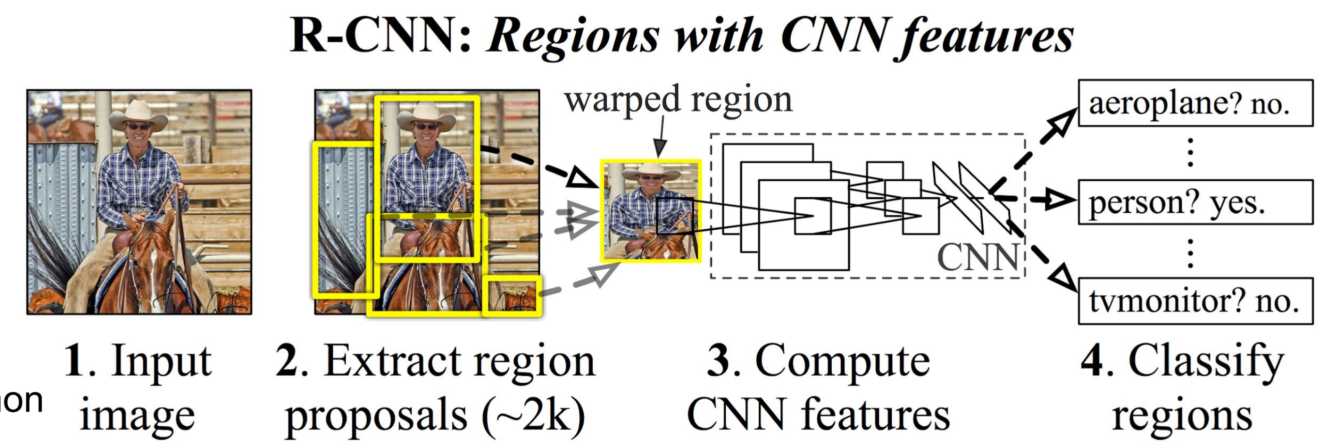
	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img
Fast R-CNN	70.0	.5 FPS	2 s/img
Faster R-CNN	73.2	7 FPS	140 ms/img
YOLO	69.0	45 FPS	22 ms/img

2 feet

Joesph Redmon

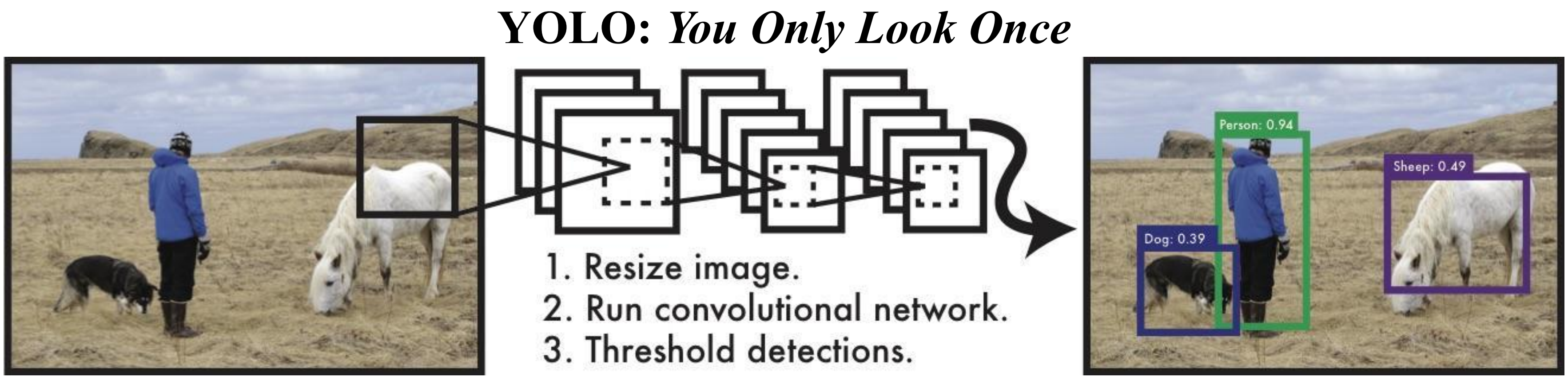
Sliding window, DPM, R-CNN all train region-based classifiers to perform detection





Joesph Redmon

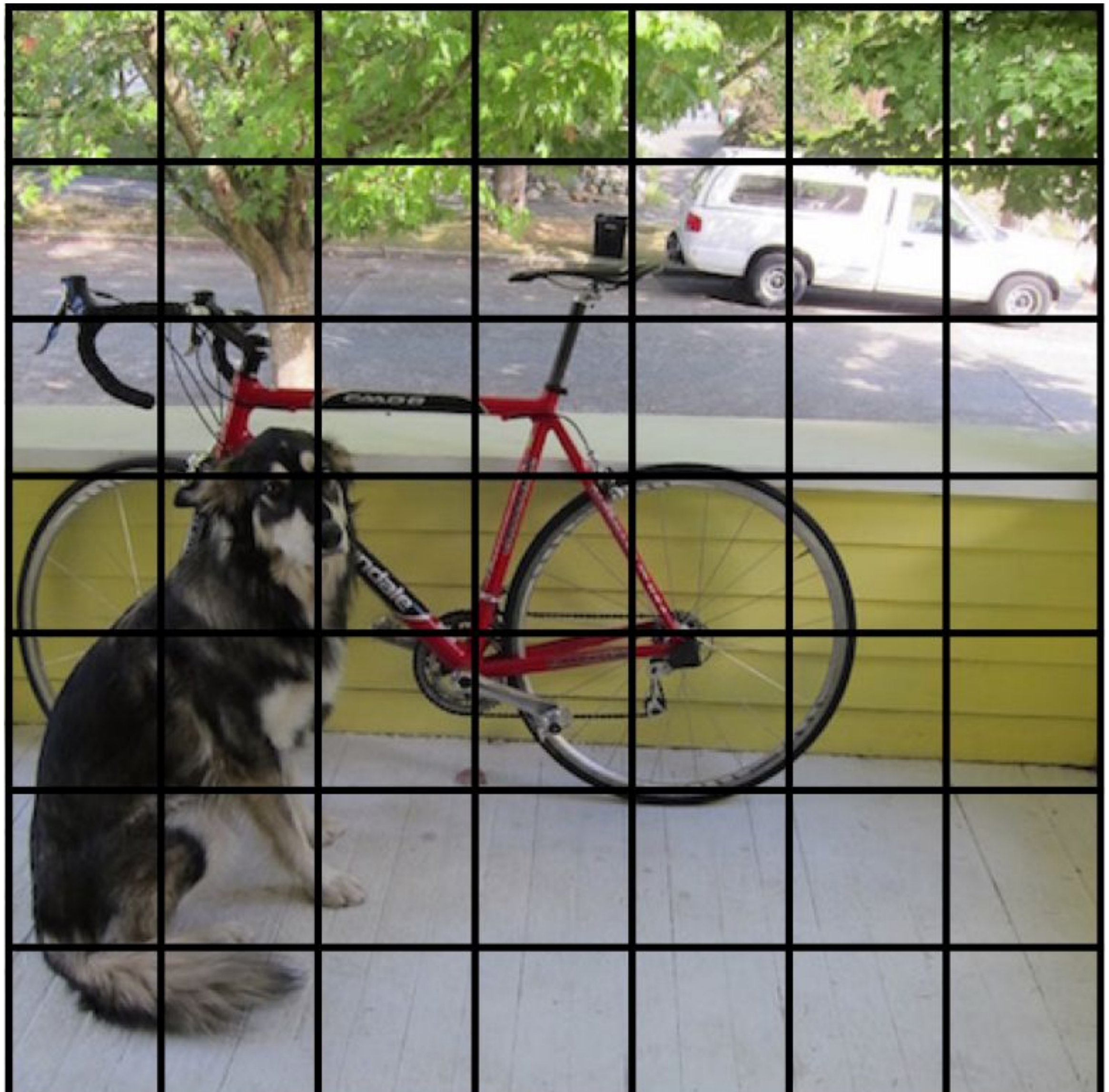
With YOLO, you only look once at an image to perform detection



Joesph Redmon

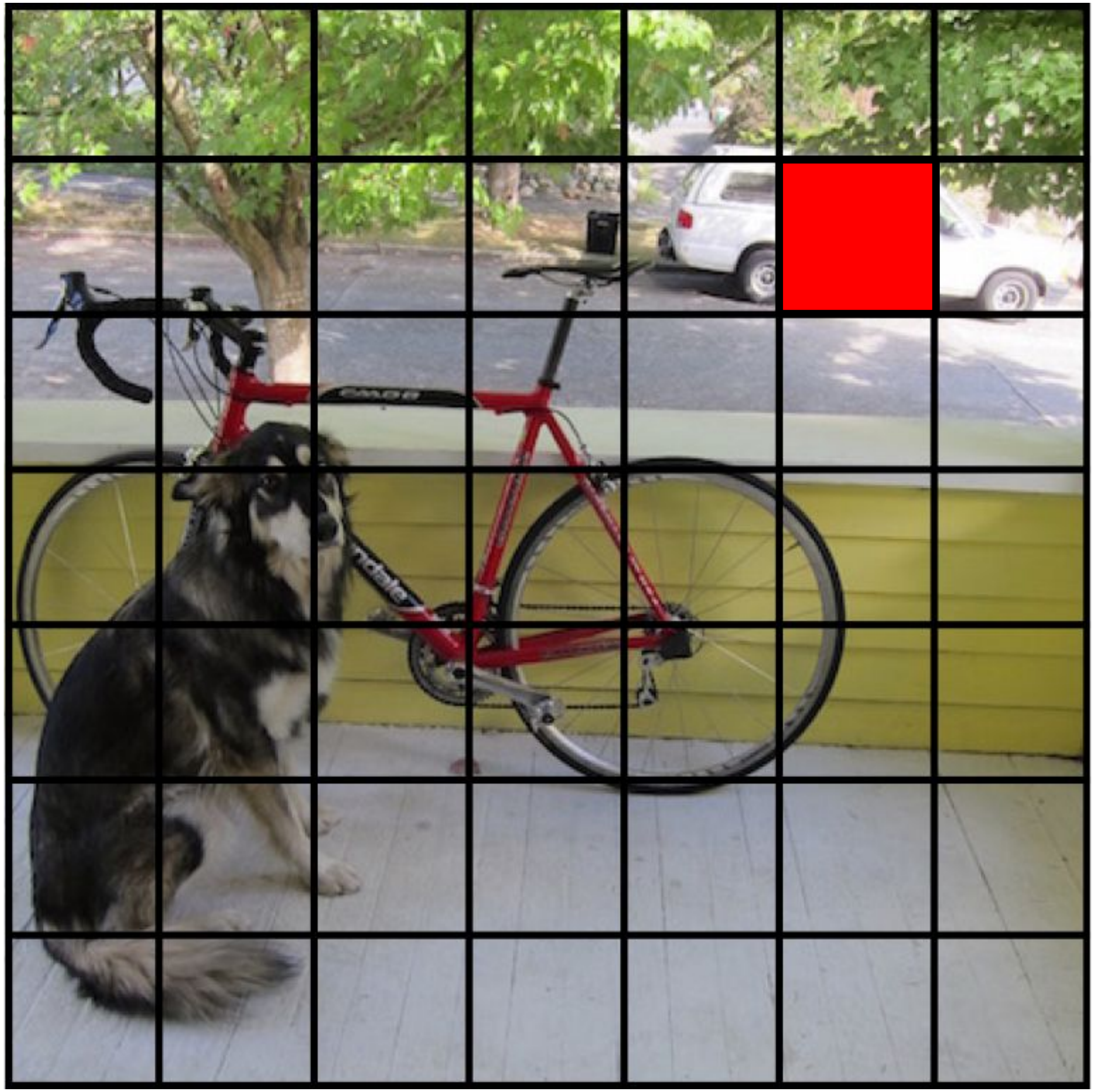
Joesph Redmon

We split the image into a grid



Joesph Redmon

Each cell predicts boxes and confidences: P(Object)

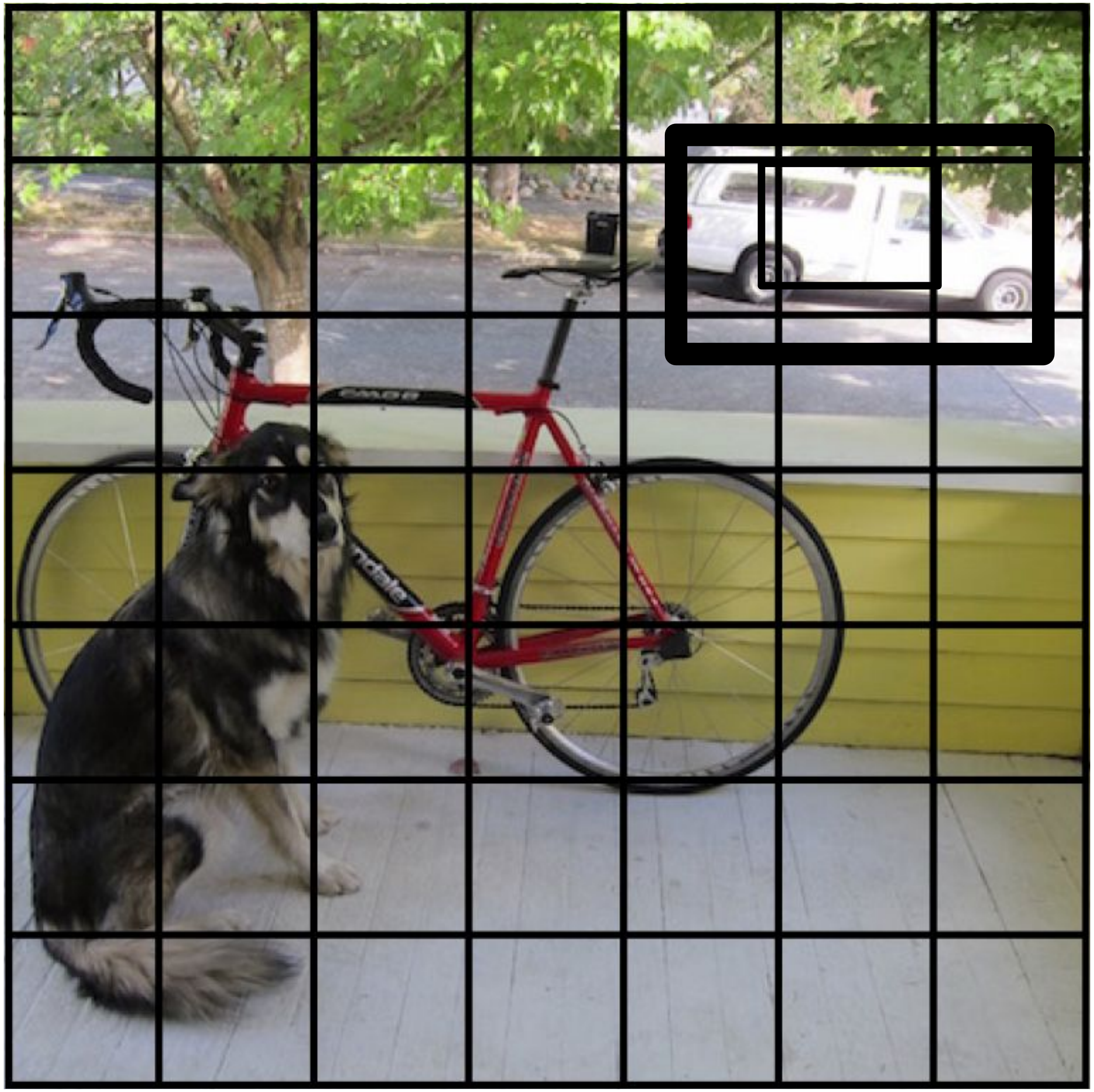


Joesph Redmon

Each cell predicts boxes and confidences: P(Object)

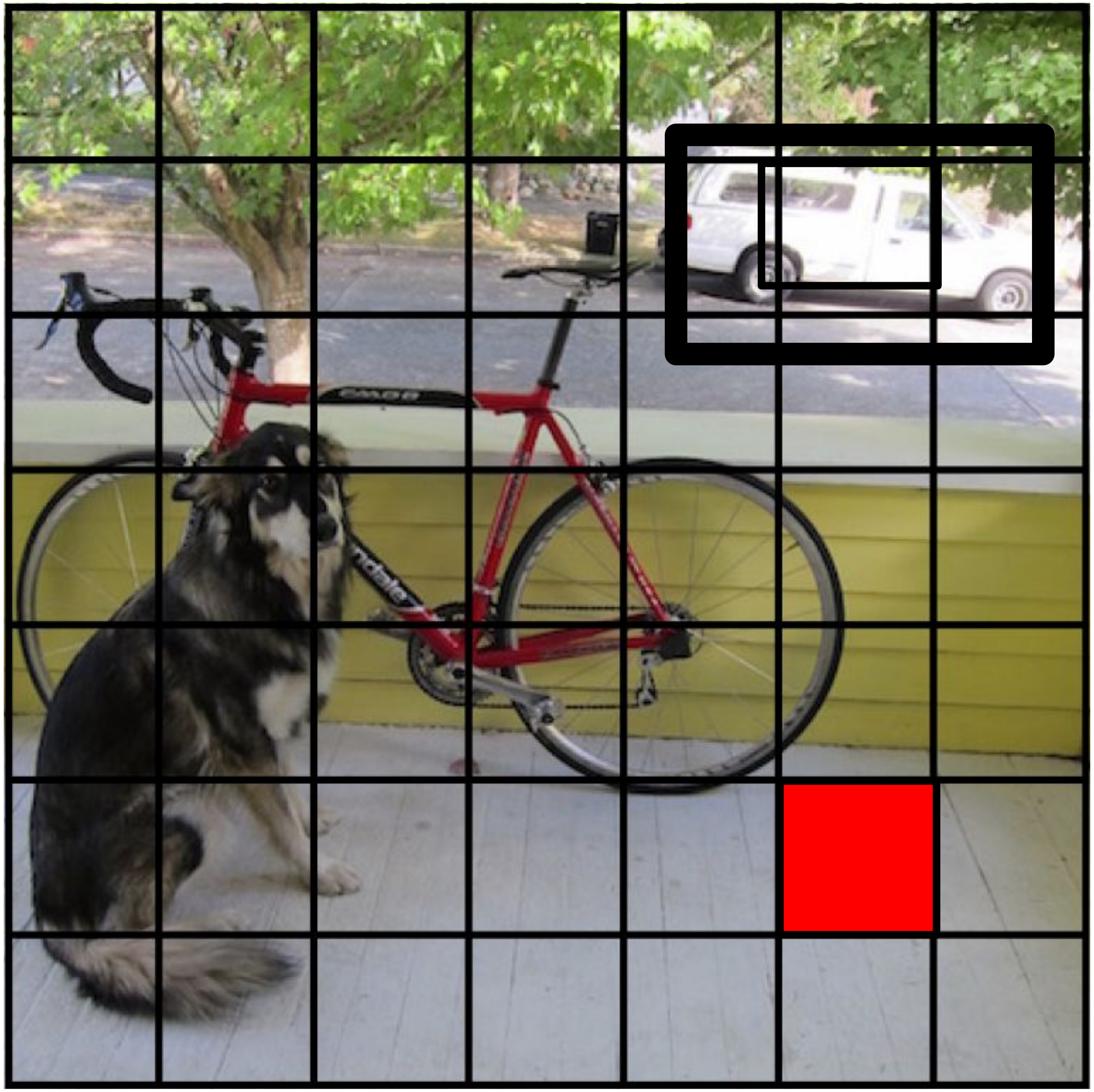
Joesph Redmon

Each cell predicts boxes and confidences: P(Object)



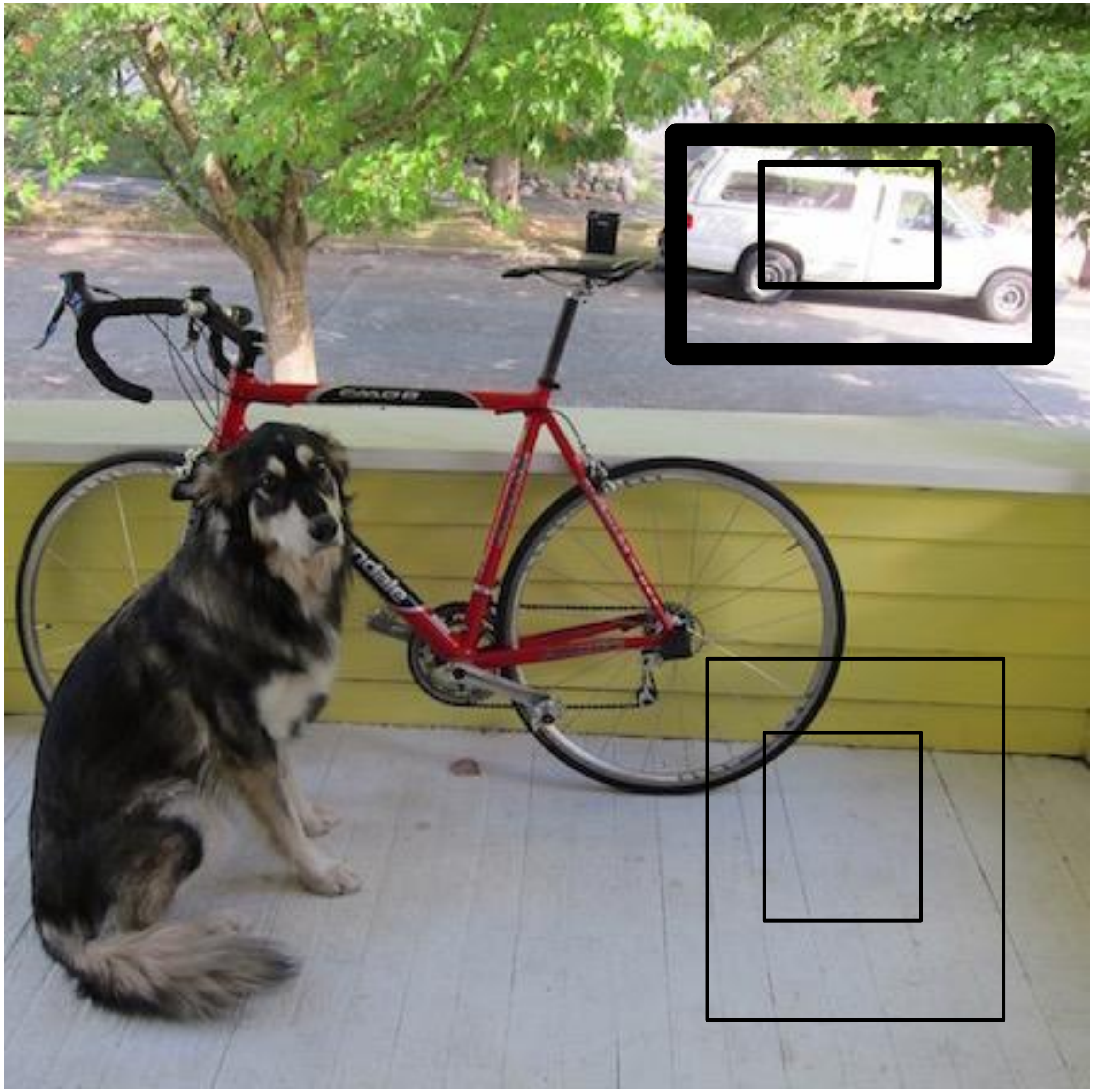
Joesph Redmon

Each cell predicts boxes and confidences: P(Object)



Joesph Redmon

Each cell predicts boxes and confidences: P(Object)



Joesph Redmon

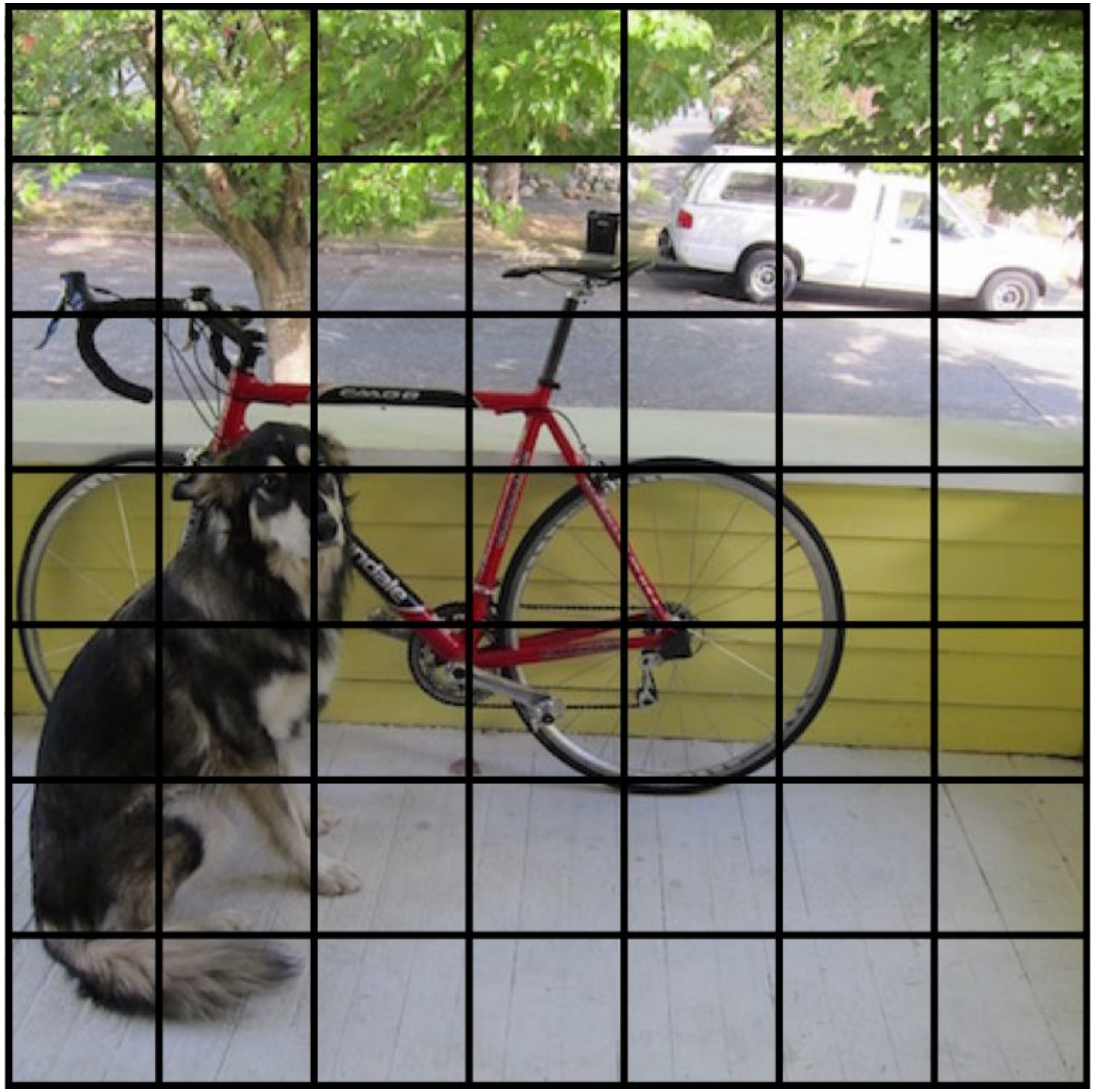
DICE ANALYTICS

Each cell predicts boxes and confidences: P(Object)

Joesph Redmon

DICE ANALYTICS

Each cell also predicts a class probability.

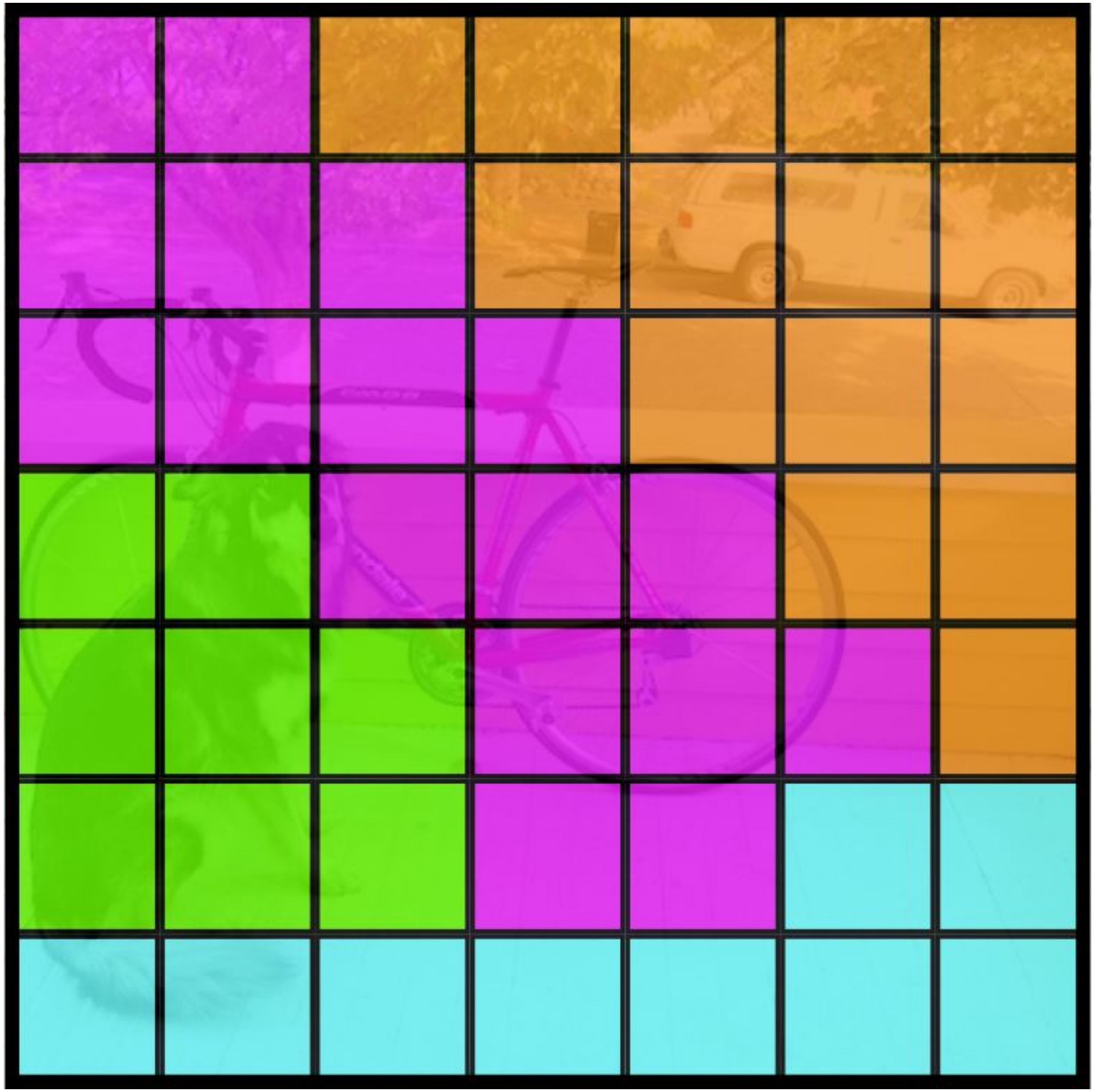


Joesph Redmon

Each cell also predicts a class probability.

Bicycle

Car

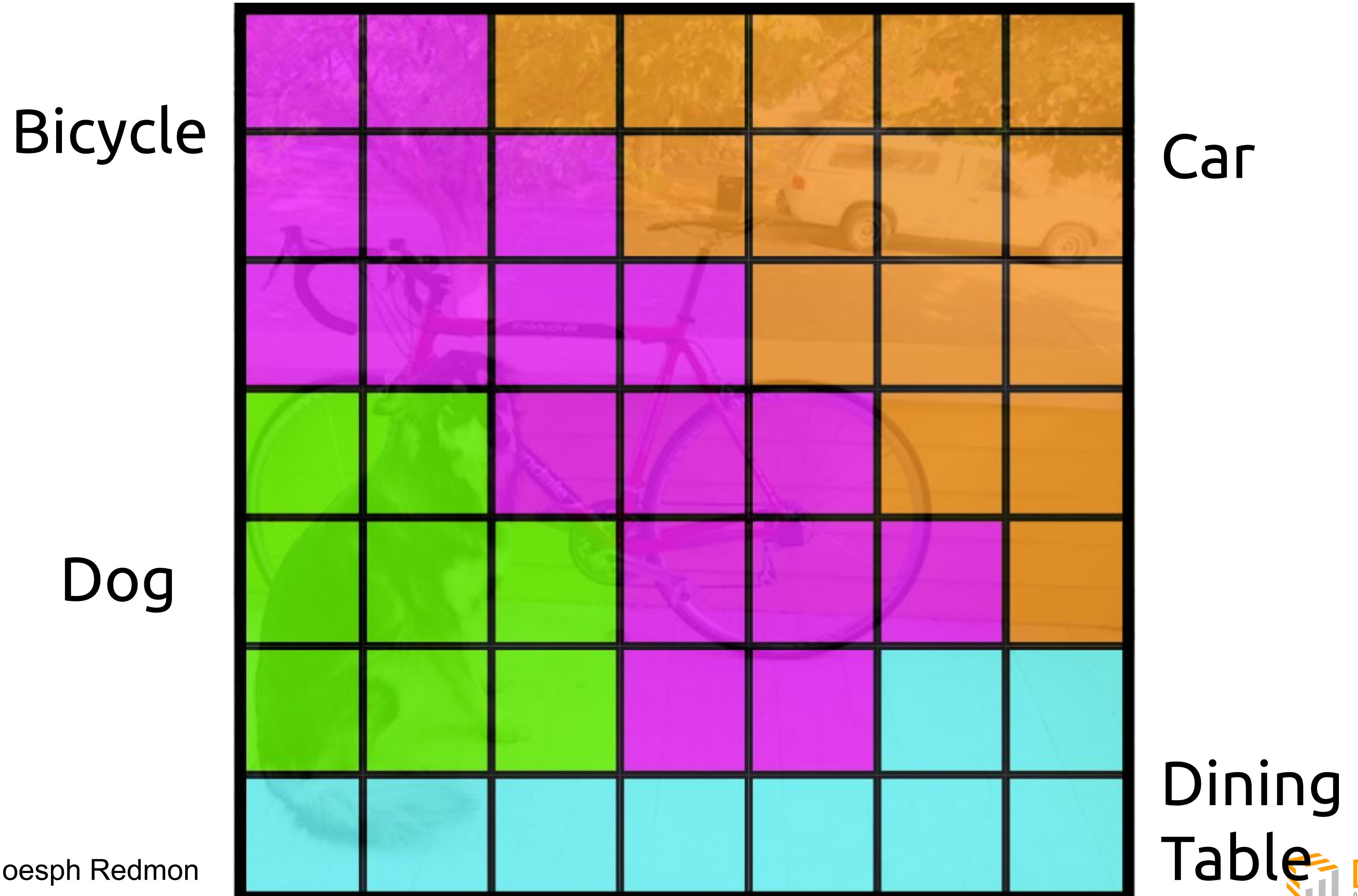


Dog

Dining Table

Joesph Redmon

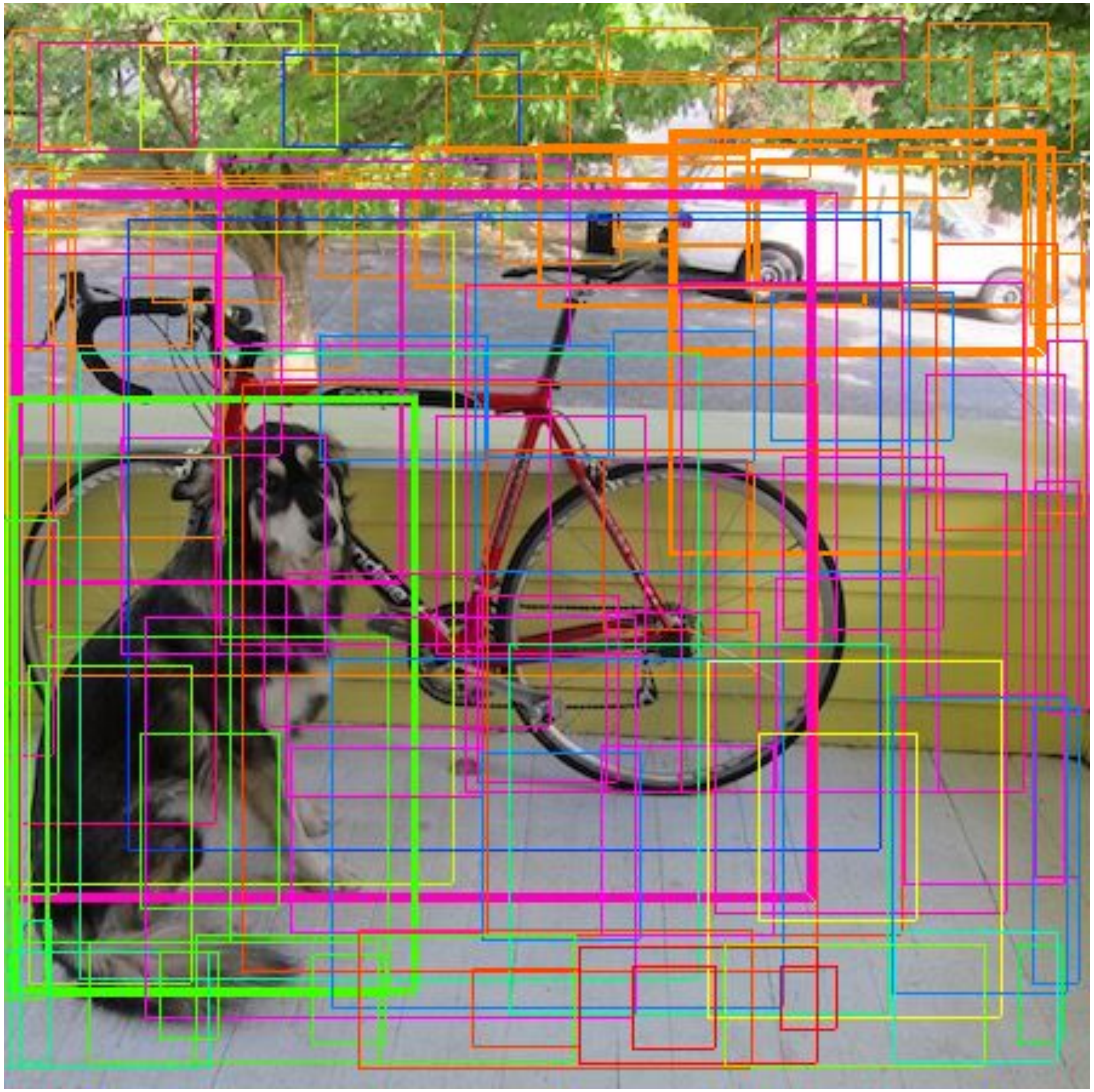
Conditioned on object: P(Car | Object)



Joesph Redmon

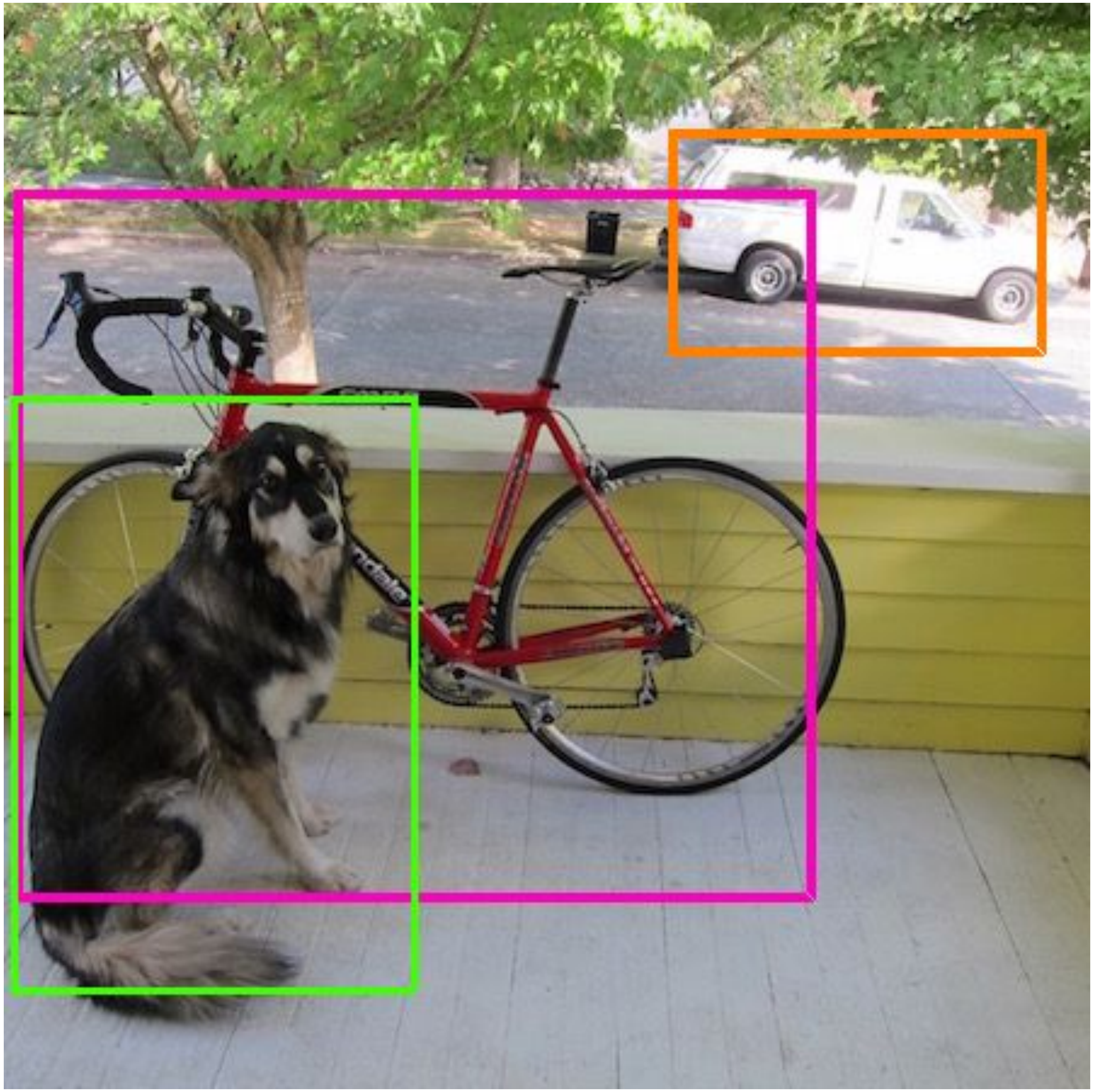
DICE ANALYTICS

Then we combine the box and class predictions.



Joesph Redmon

Finally we do NMS and threshold detections



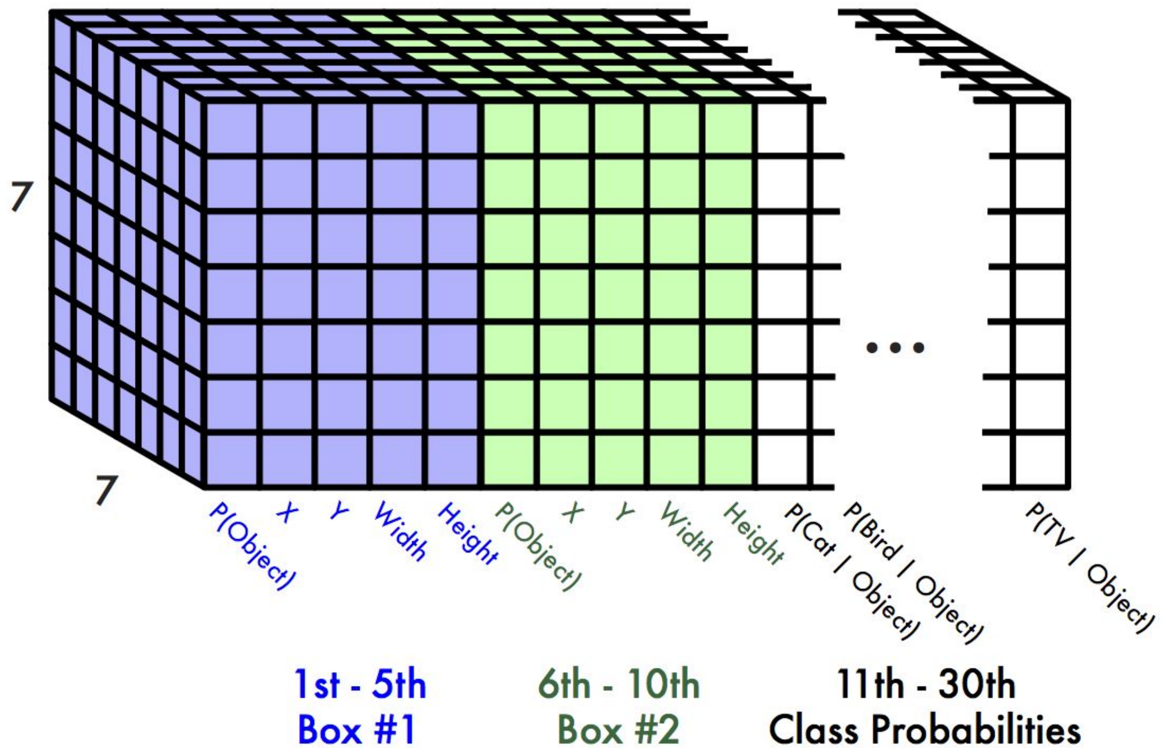
Joesph Redmon

DICE ANALYTICS

This parameterization fixes the output size

Each cell predicts:

- For each bounding box:
  - 4 coordinates (x, y, w, h)
  - 1 confidence value
- Some number of class probabilities



For Pascal VOC:

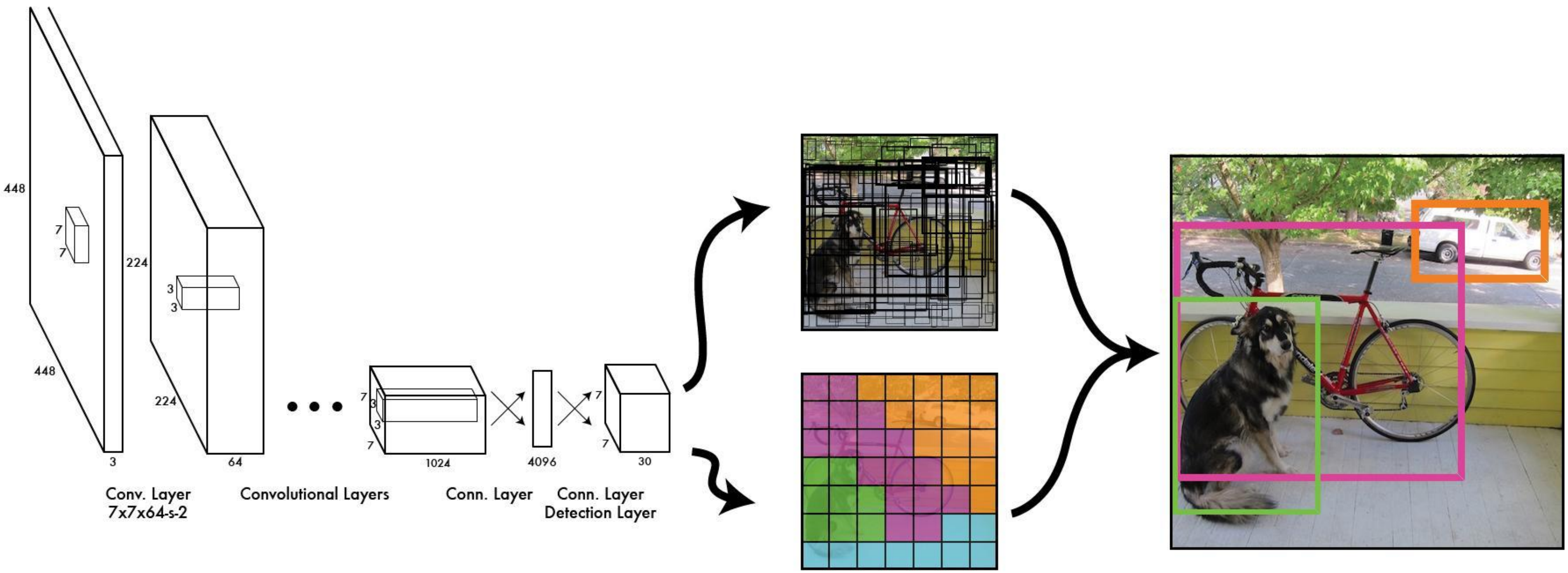
- 7x7 grid
- 2 bounding boxes / cell
- 20 classes

7 x 7 x (2 x 5 + 20) = 7 x 7 x 30 tensor = **1470 outputs**

Joesph Redmon

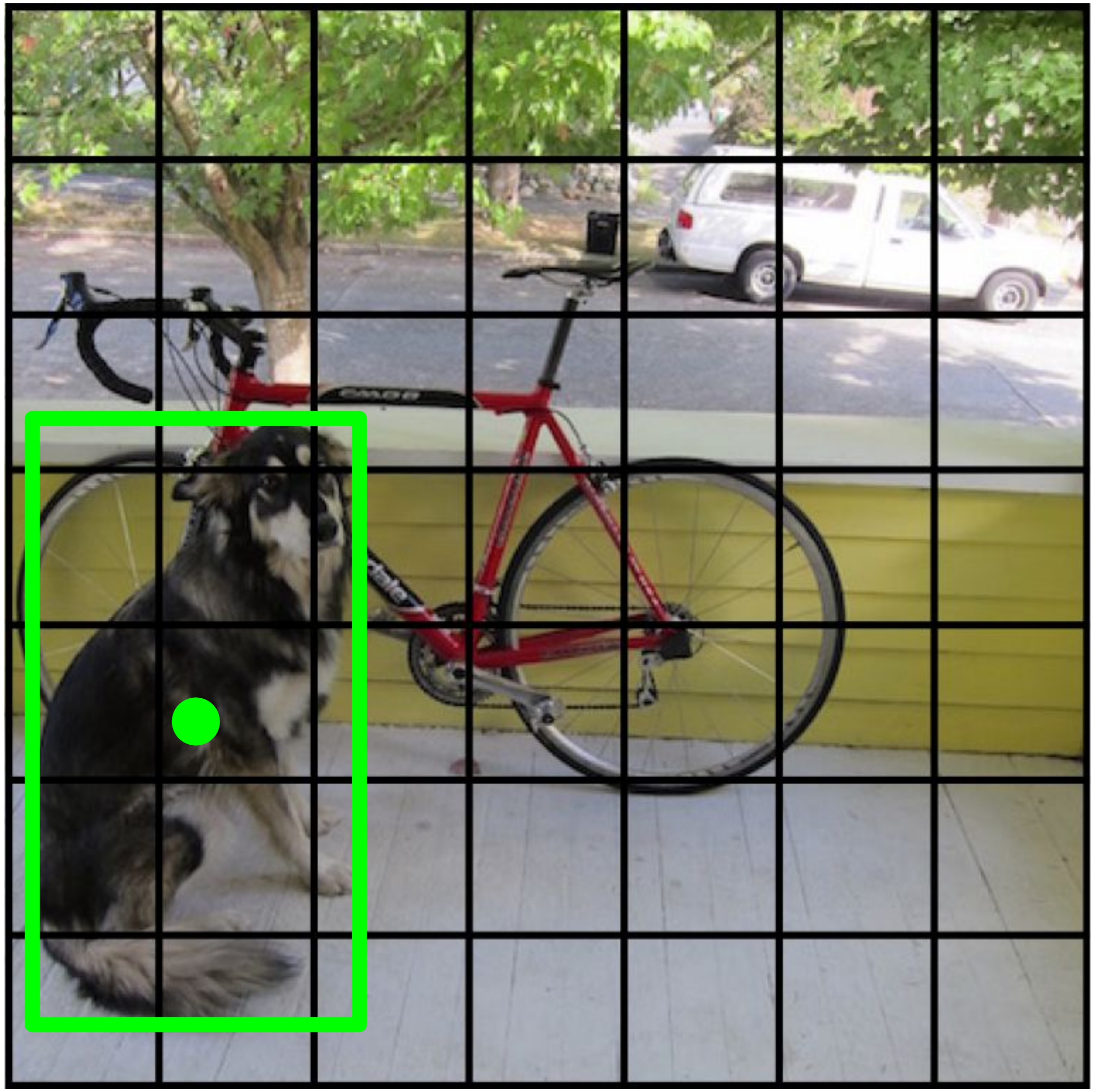
DICE ANALYTICS

Thus we can train one neural network to be a whole detection pipeline



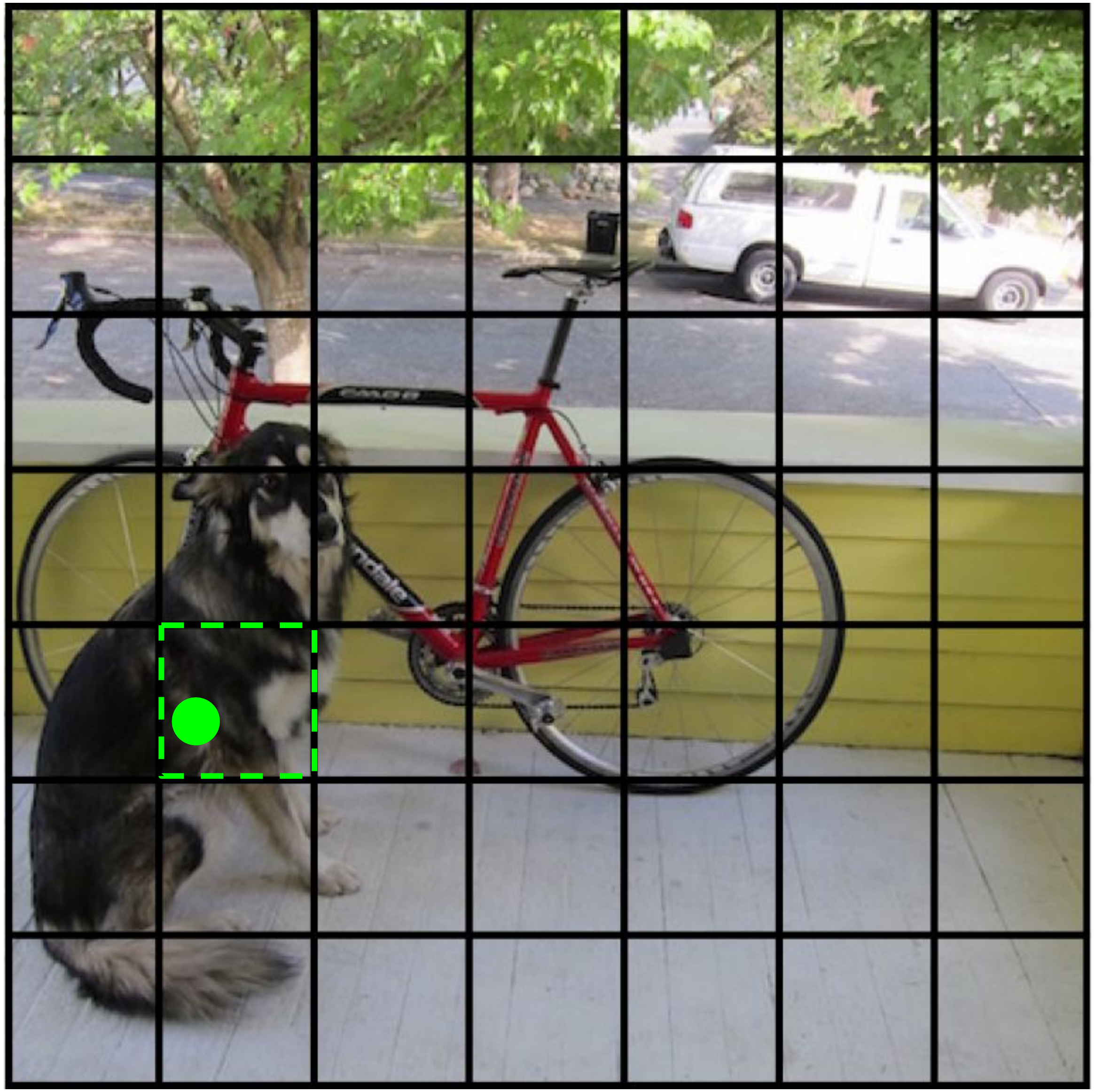
Joesph Redmon

During training, match example to the right cell



Joesph Redmon

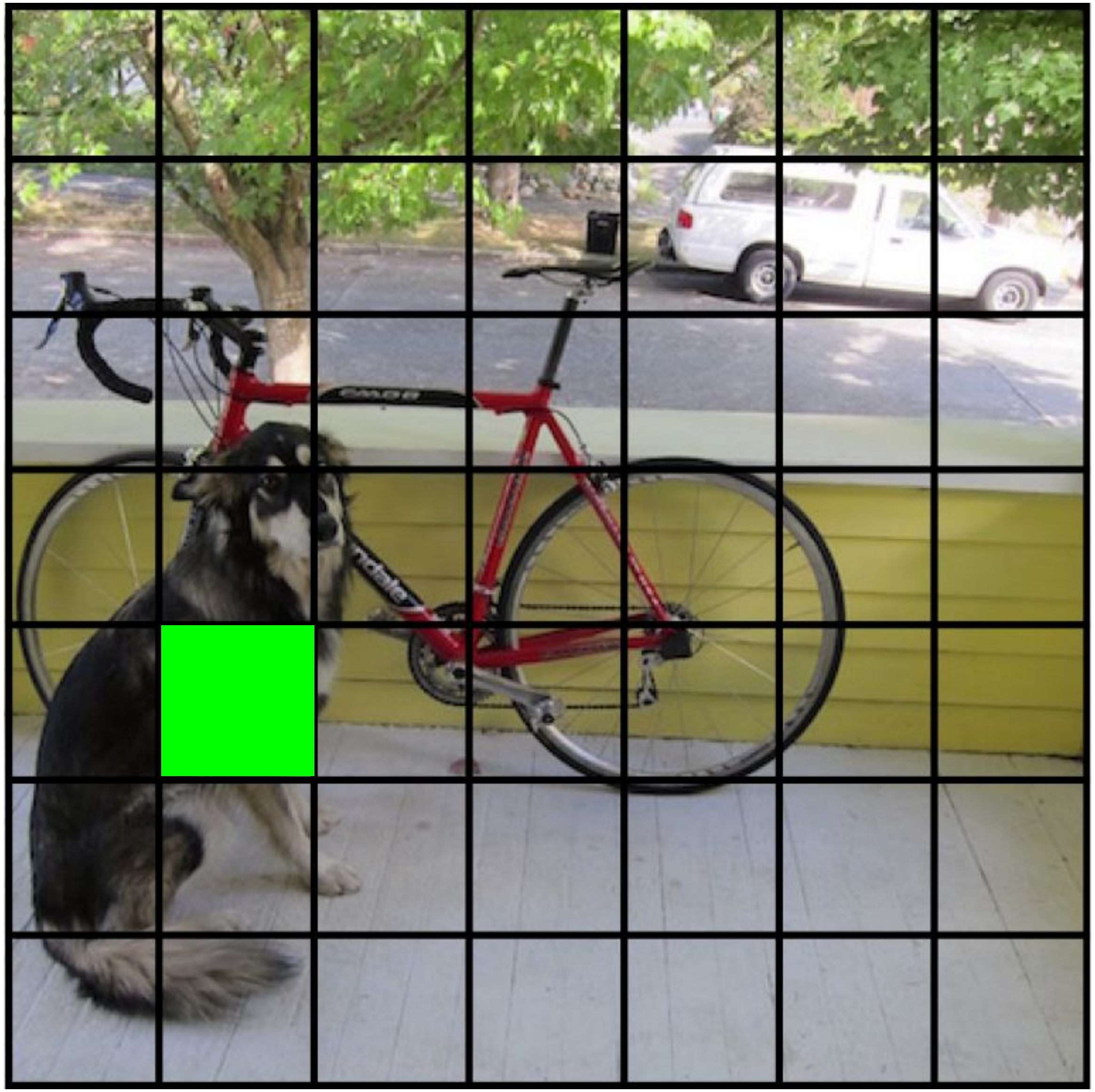
During training, match example to the right cell



Joesph Redmon

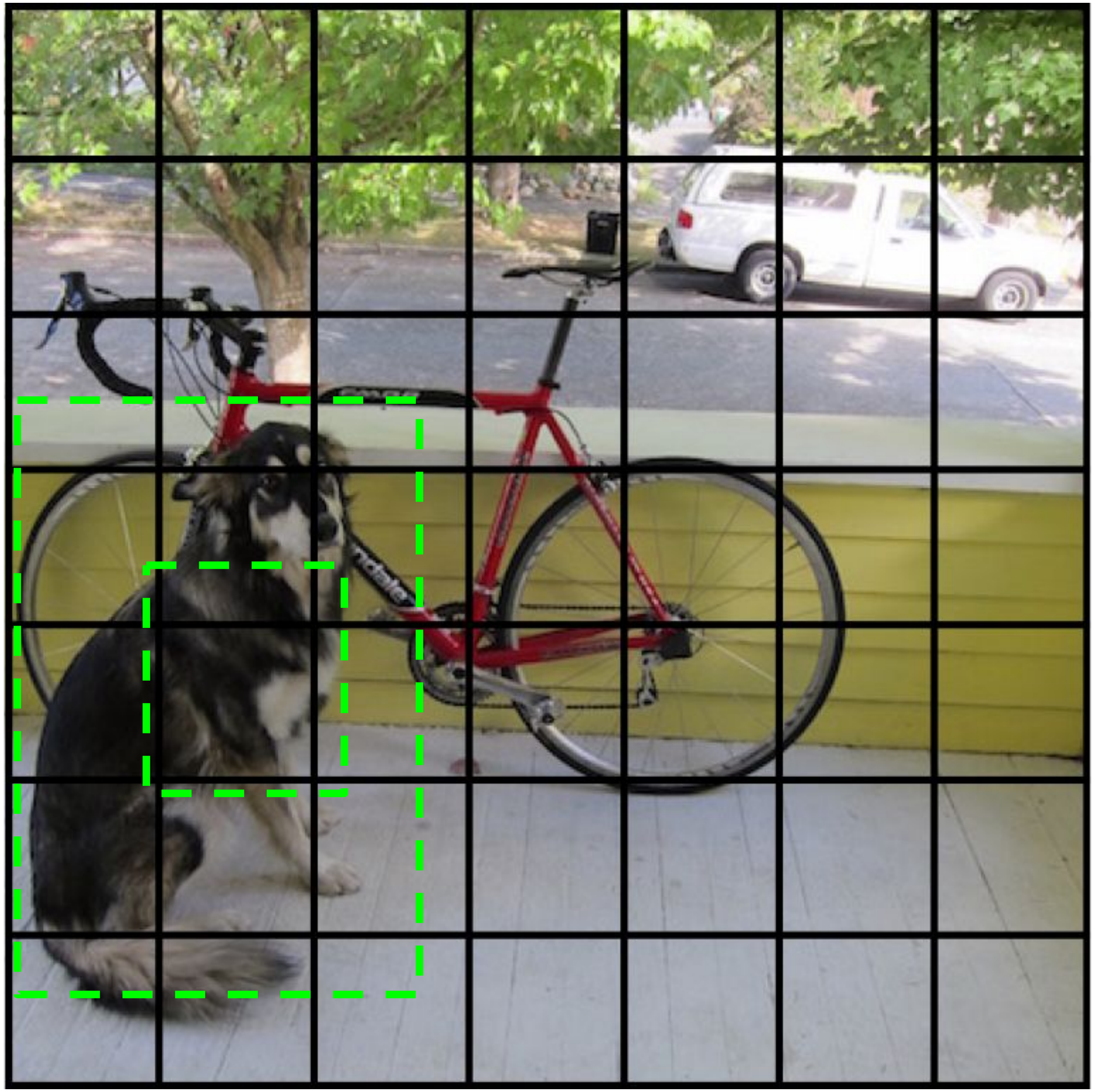
Adjust that cell's class prediction

Dog = 1
Cat = 0
Bike = 0
...



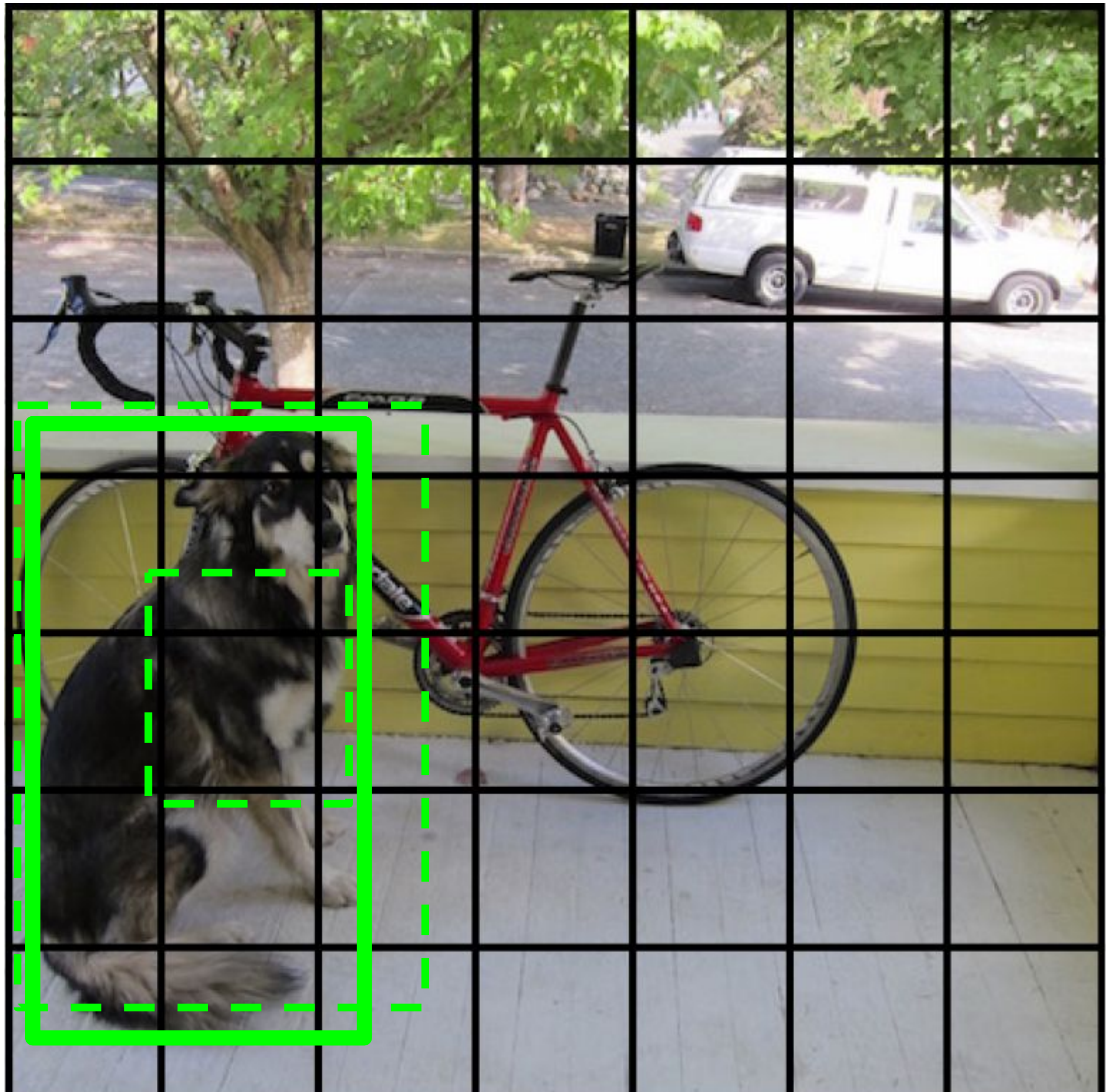
Joesph Redmon

Look at that cell's predicted boxes



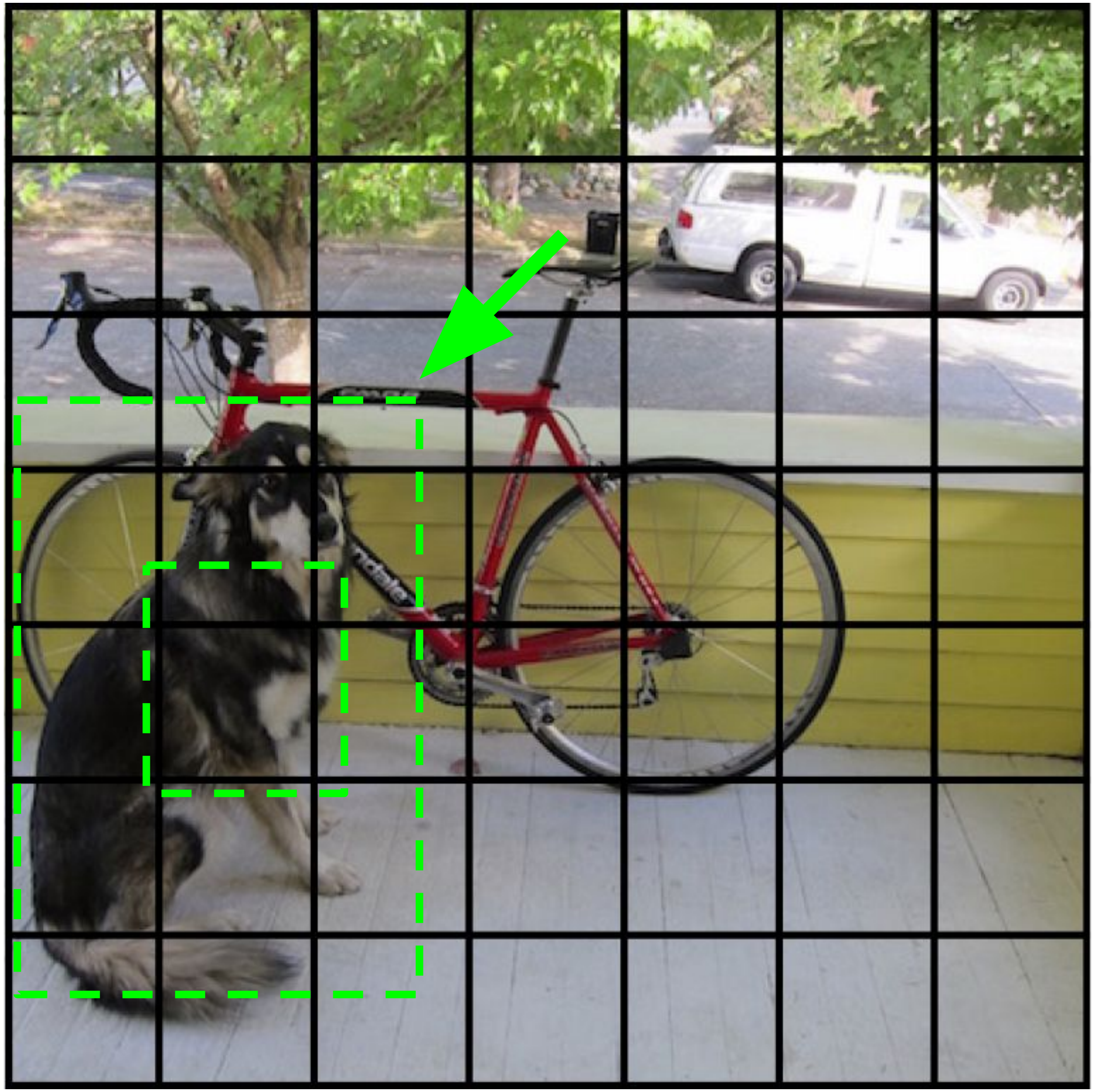
Joesph Redmon

Find the best one, adjust it, increase the confidence



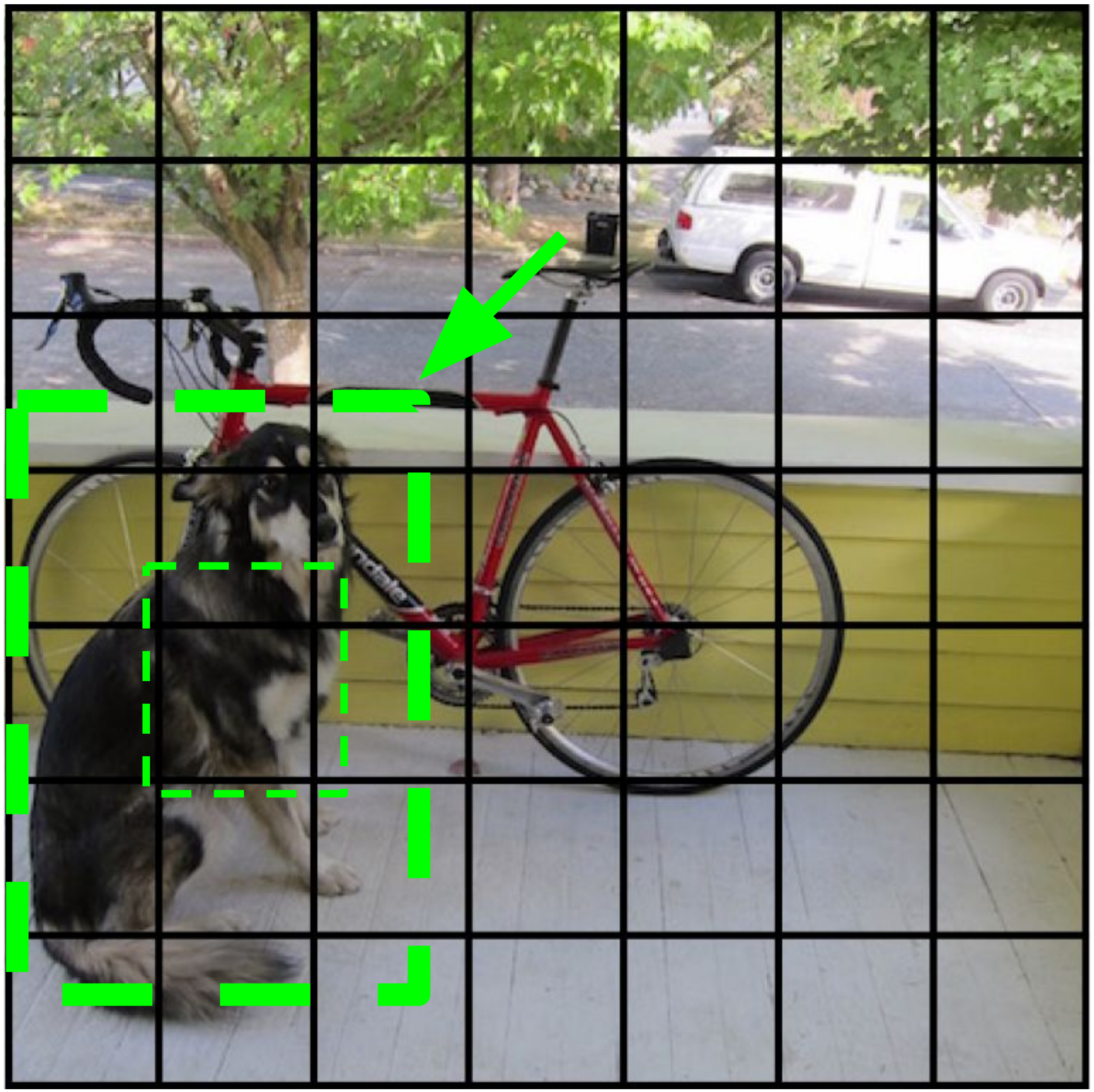
Joesph Redmon

Find the best one, adjust it, increase the confidence



Joesph Redmon

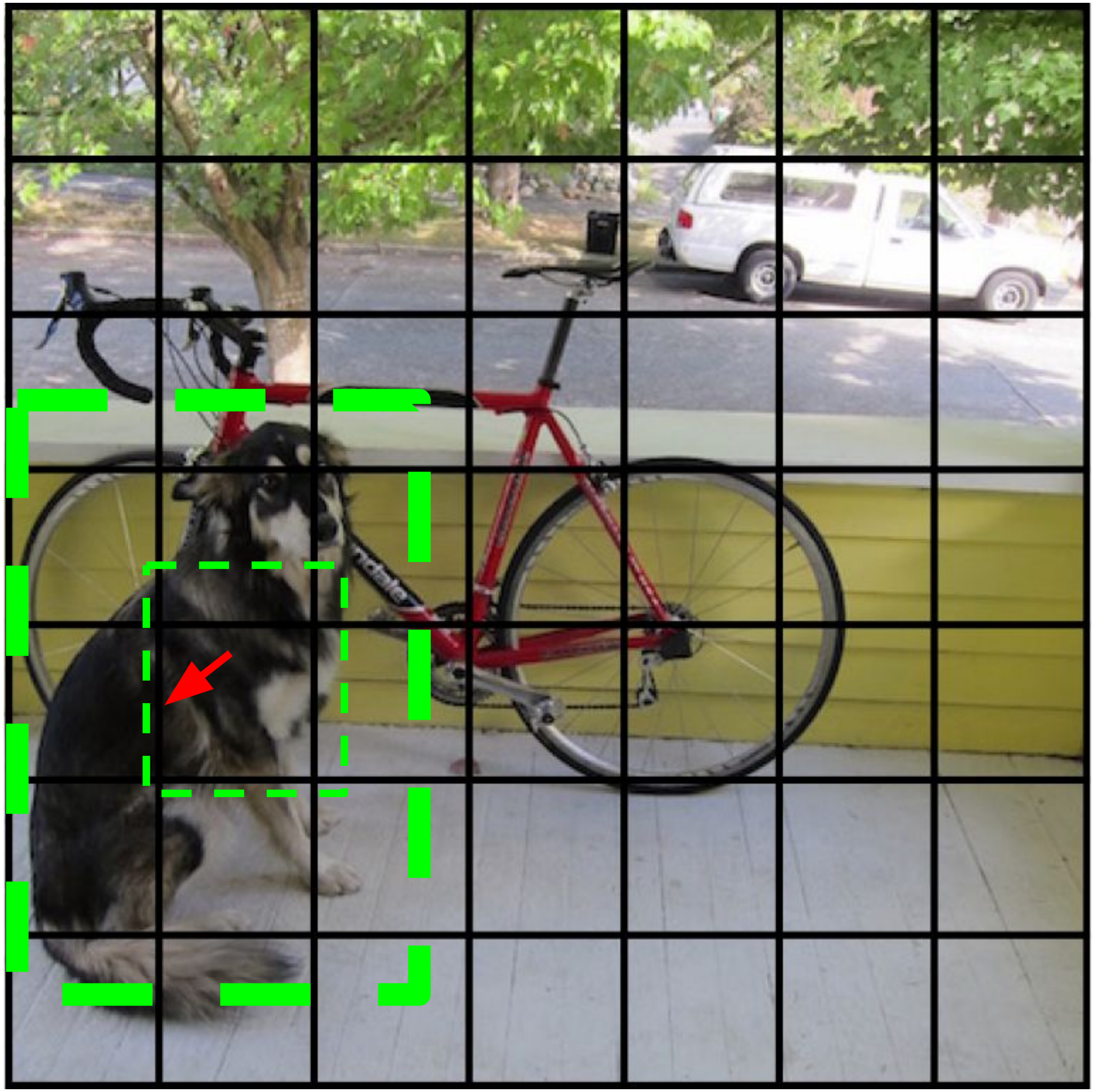
Find the best one, adjust it, increase the confidence



Joesph Redmon

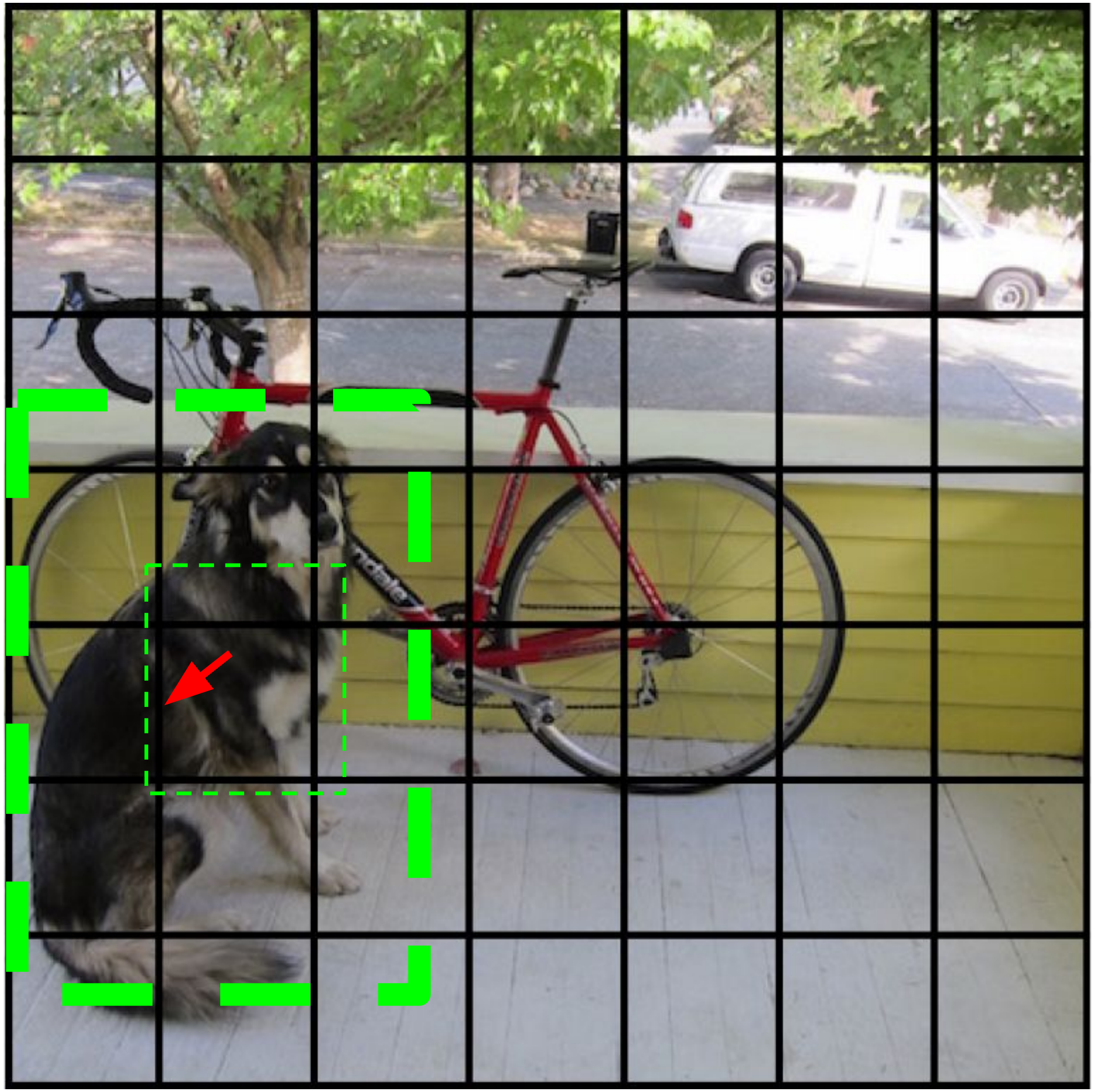
DICE ANALYTICS

Decrease the confidence of other boxes



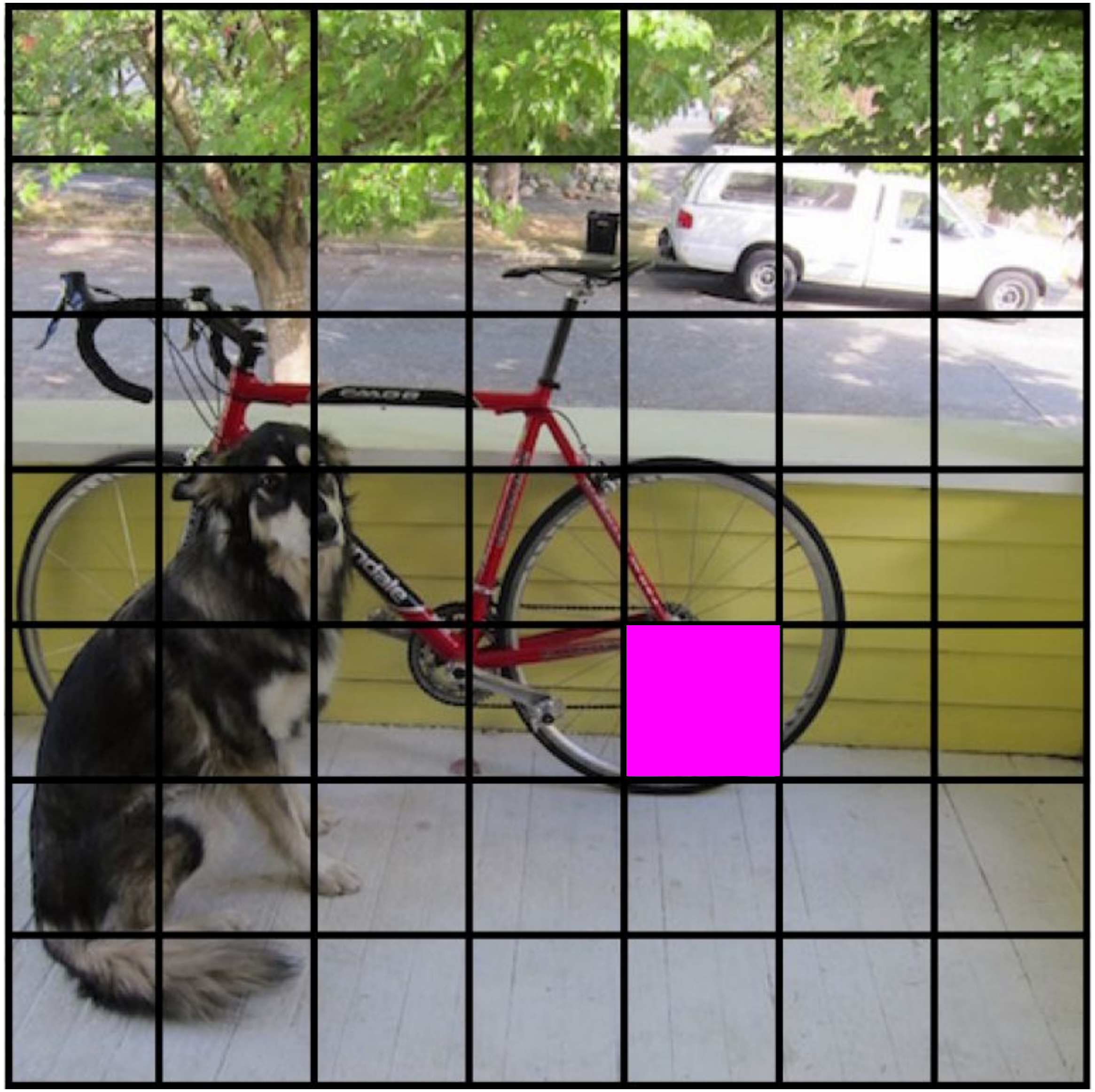
Joesph Redmon

Decrease the confidence of other boxes



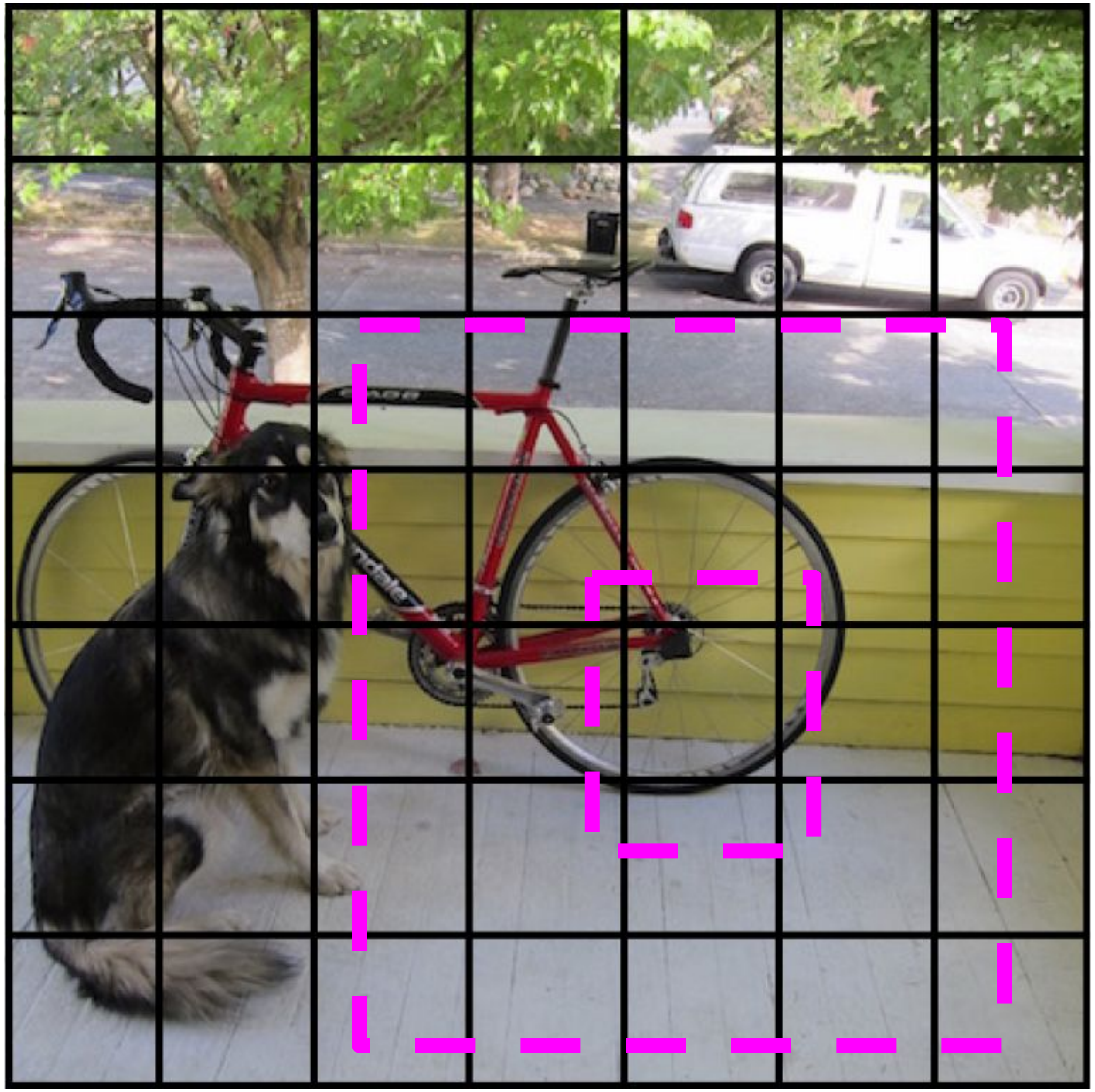
Joesph Redmon

Some cells don’t have any ground truth detections!



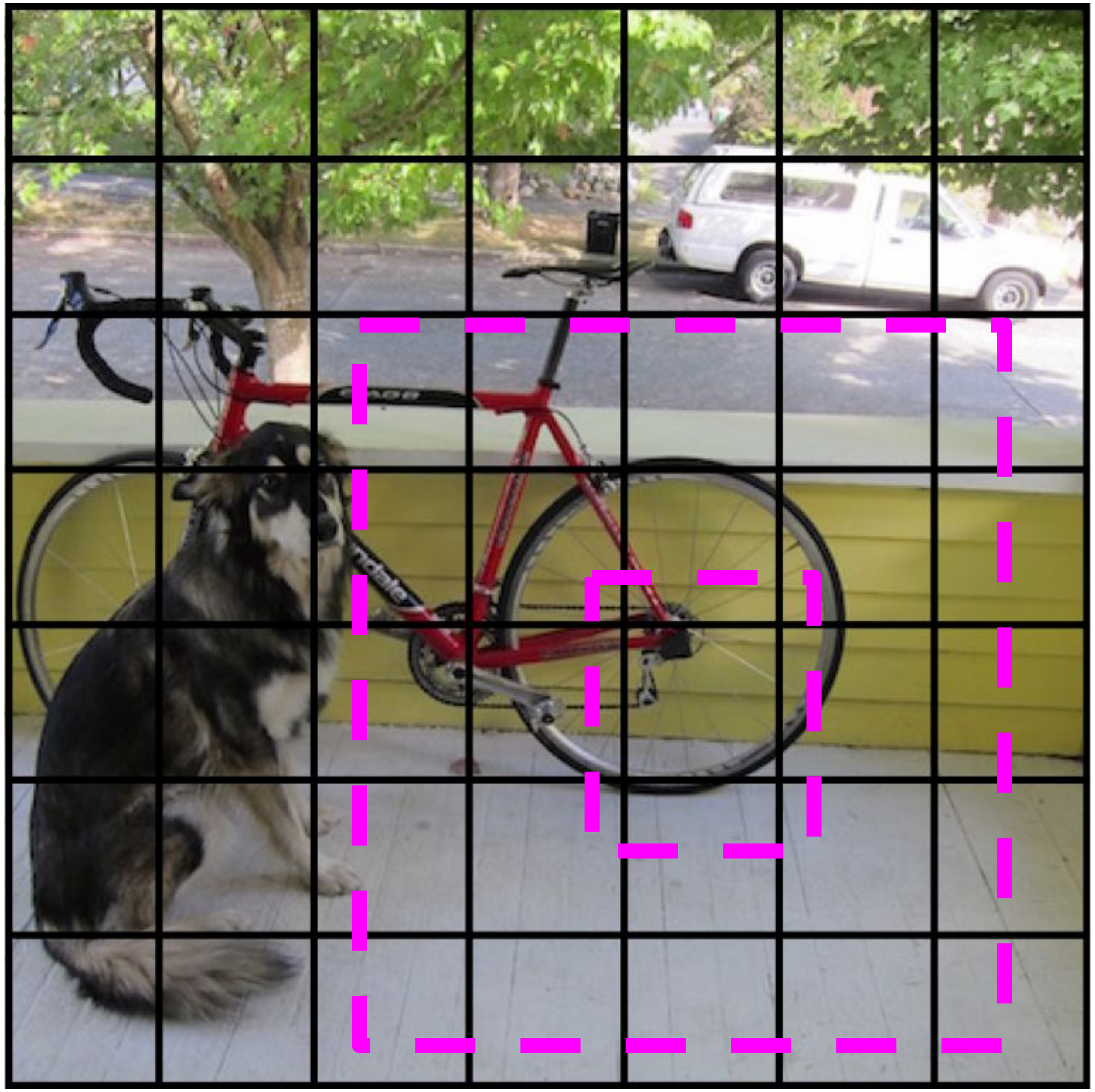
Joesph Redmon

Some cells don’t have any ground truth detections!



Joesph Redmon

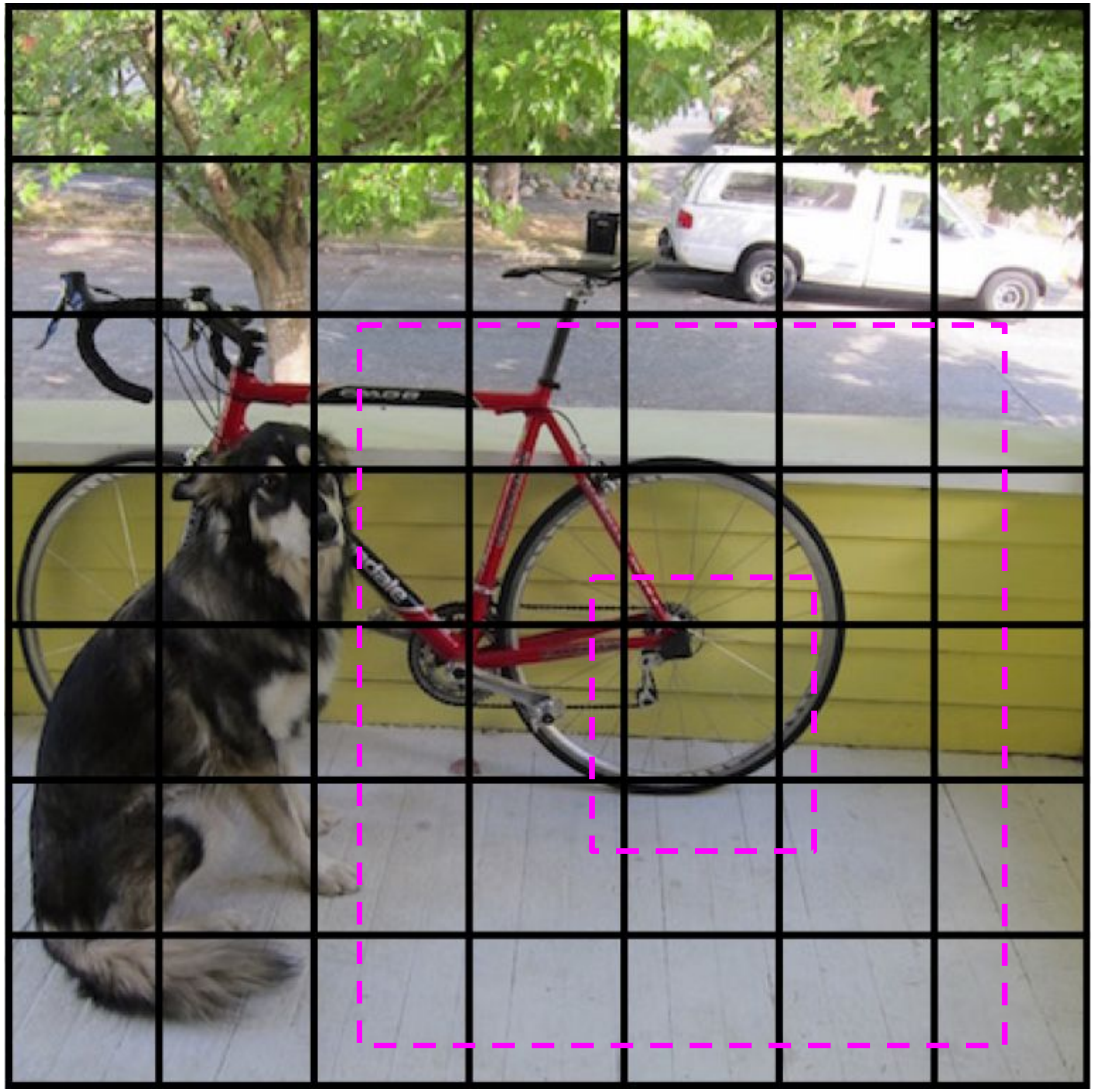
Decrease the confidence of these boxes



Joesph Redmon

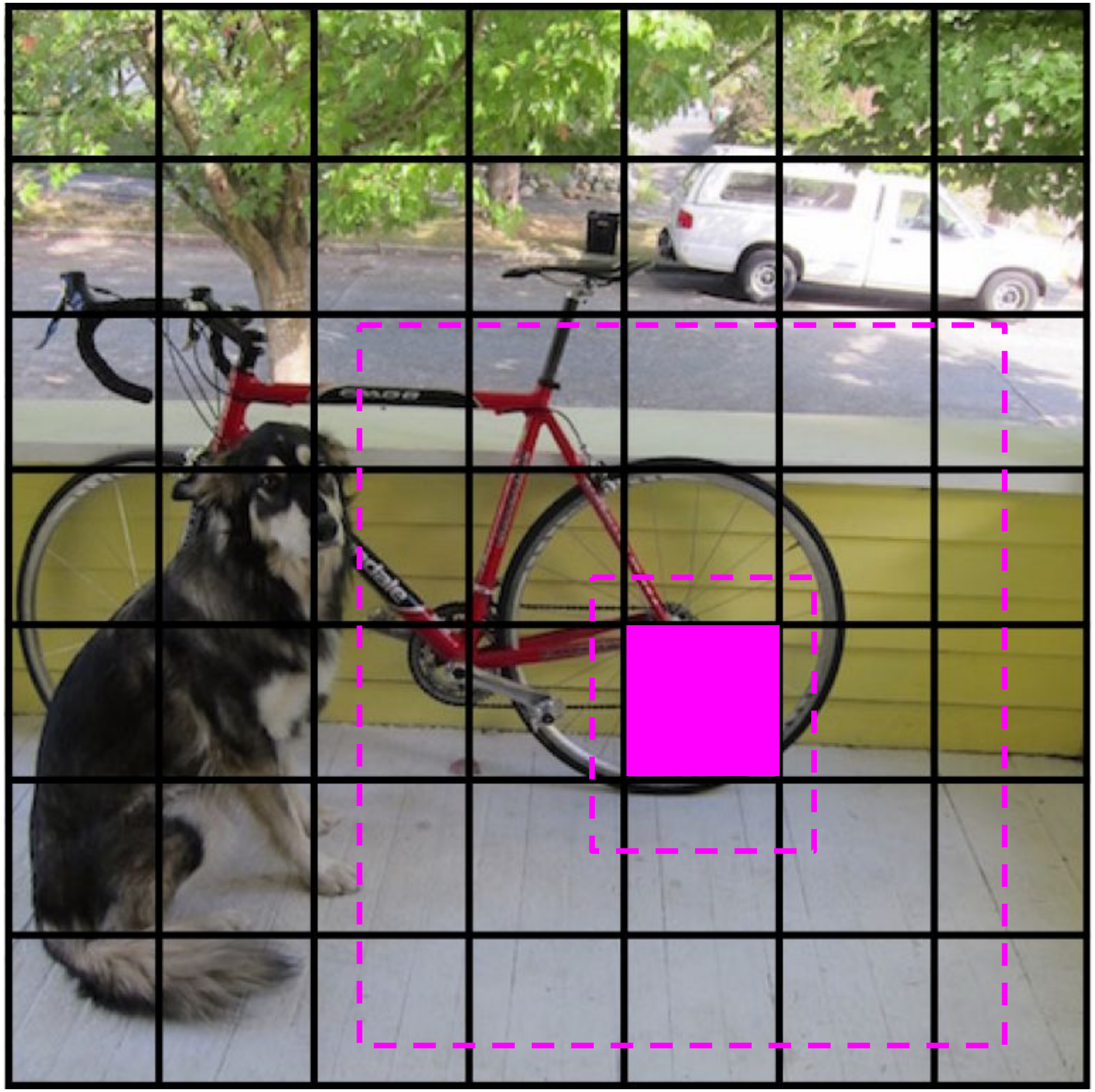
DICE ANALYTICS

Decrease the confidence of these boxes



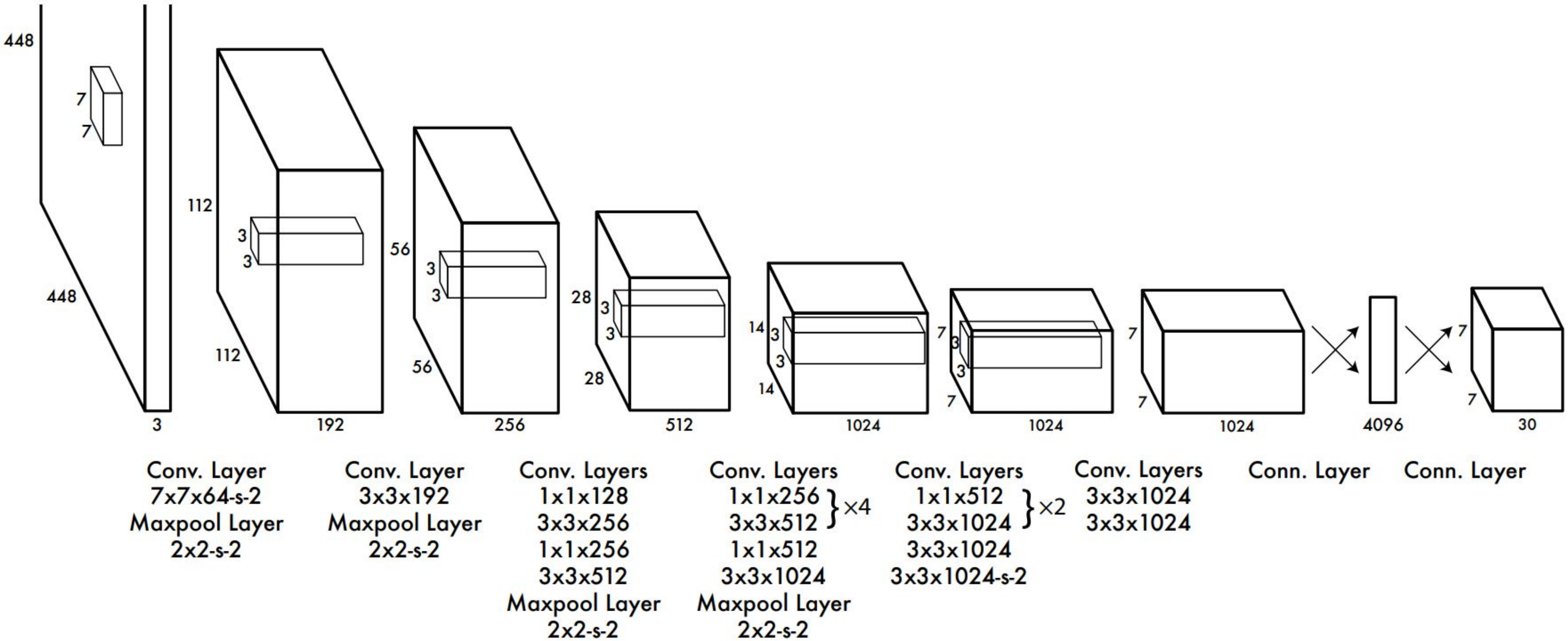
Joesph Redmon

Don’t adjust the class probabilities or coordinates



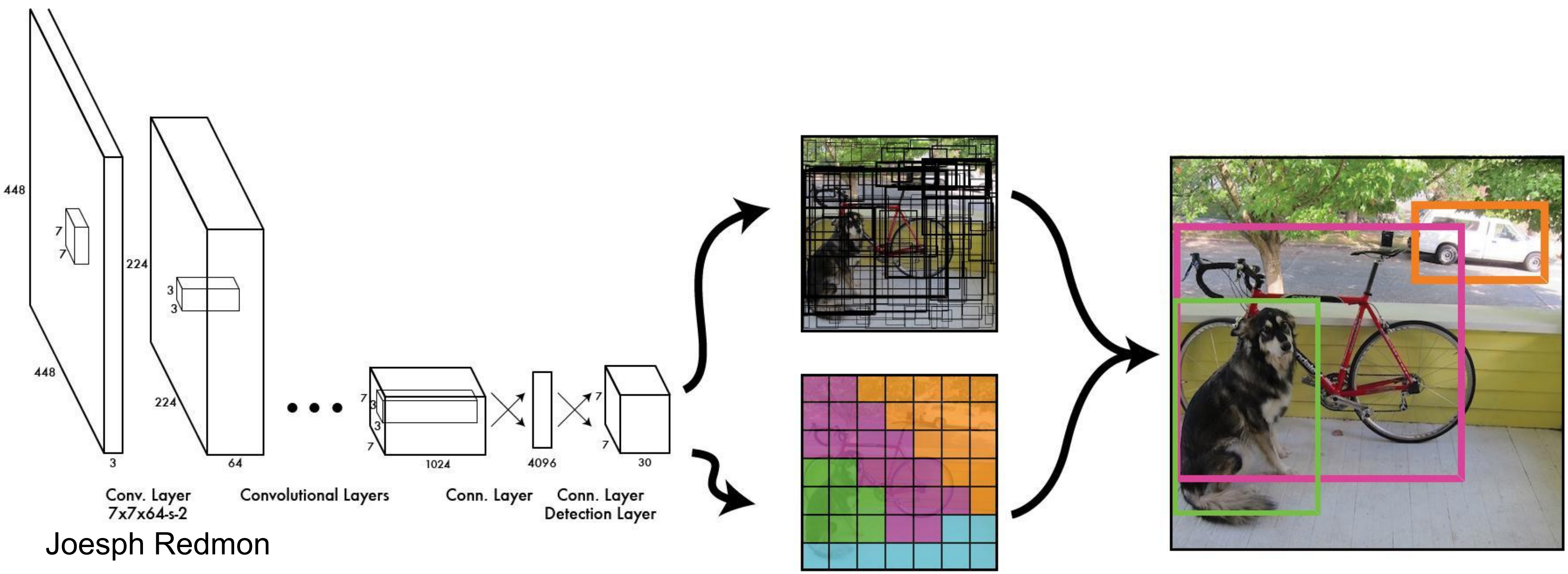
Joesph Redmon

The architecture



Yolo is trained with standard tricks:

- Pretraining on Imagenet
- SGD with decreasing learning rate
- Extensive data augmentation
- For details, you can go here



Joesph Redmon

The working parts

Intersection over union (IoU) of the bounding boxes

Andrew Ng

DICE ANALYTICS

The working parts

Intersection over union (IoU) of the bounding boxes:

Predicted

Ground truth

IoU = (size of intersection) / (size of union)

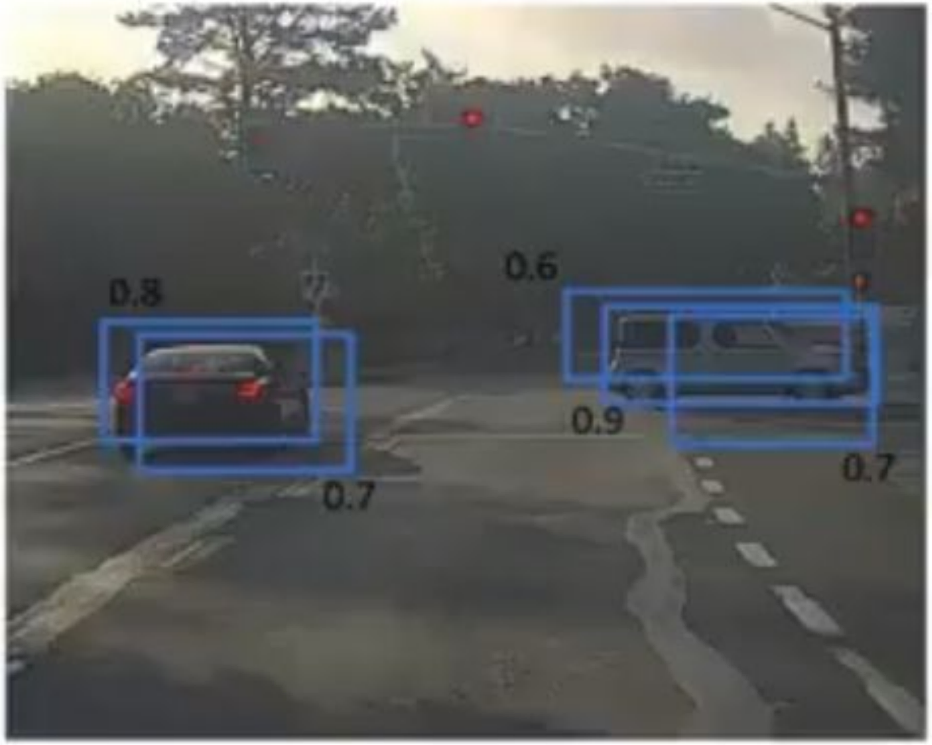
- Correctness of prediction depends on IoU having a value above a certain threshold value (hyperparameter)
- The range of IoU is from 0 (totally disjoint) to 1 (identical)

Andrew Ng

The working parts

Non-Maximum suppression:

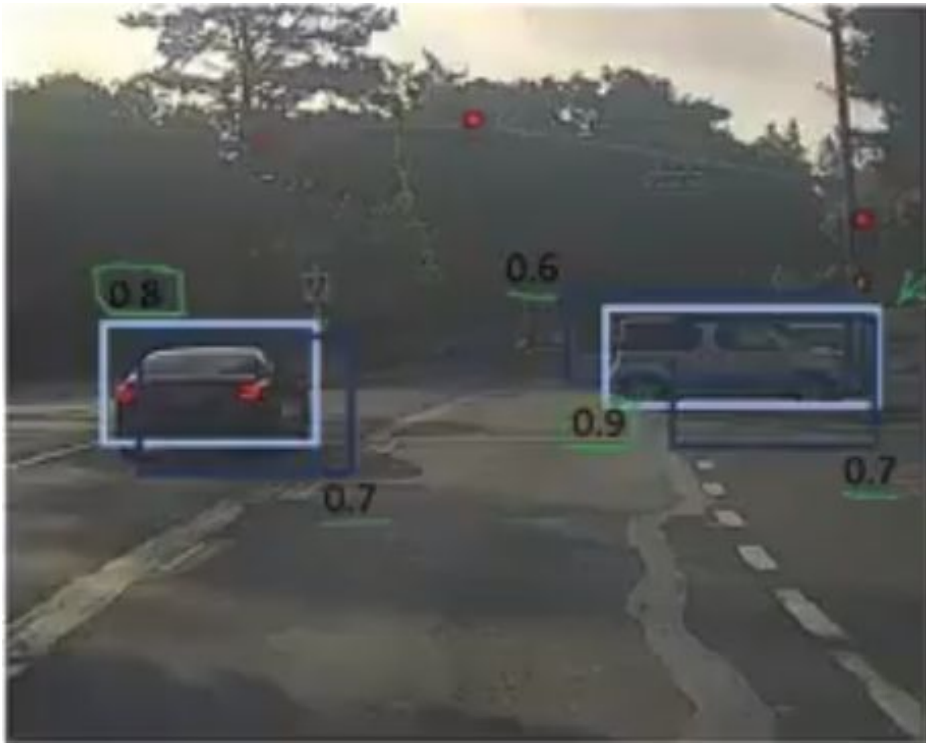
To prevent multiple detections of the same object in an image.



The working parts

Non-Maximum suppression:

To prevent multiple detections of the same object in an image.



The working parts

Non-Maximum suppression:

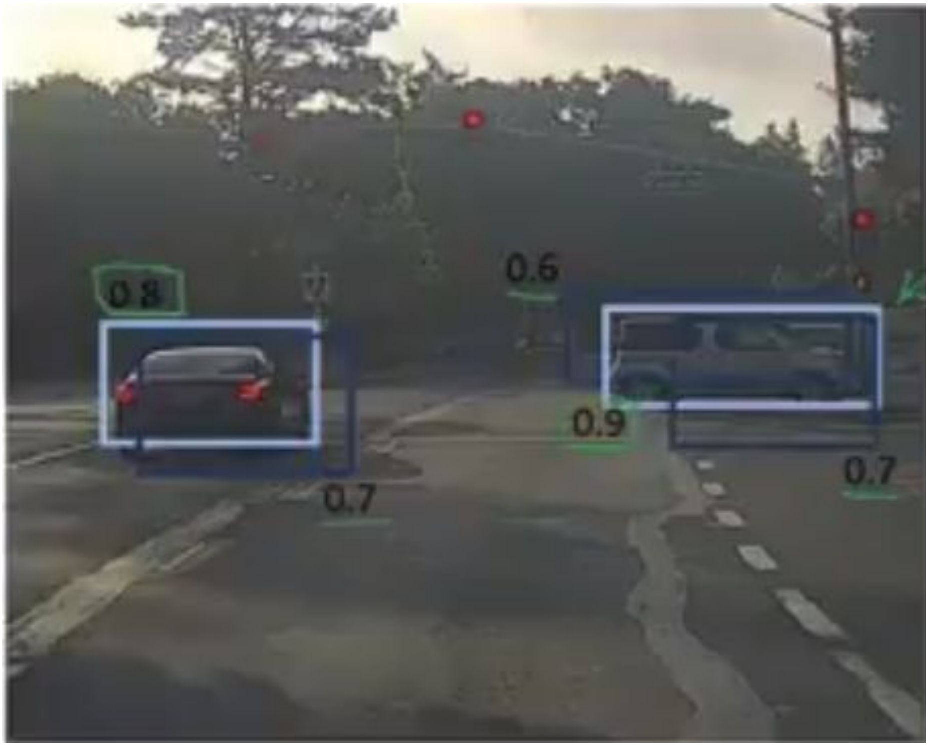
For each output prediction:

1. Discard all boxes having a confidence value below a threshold (let's say 0.6)
2. While there are remaining boxes:
   a. Pick the box with the largest confidence value and output that as prediction
   b. Discard any remaining box with an IoU above a certain threshold (let’s say 0.5) with the box output in the previous step.

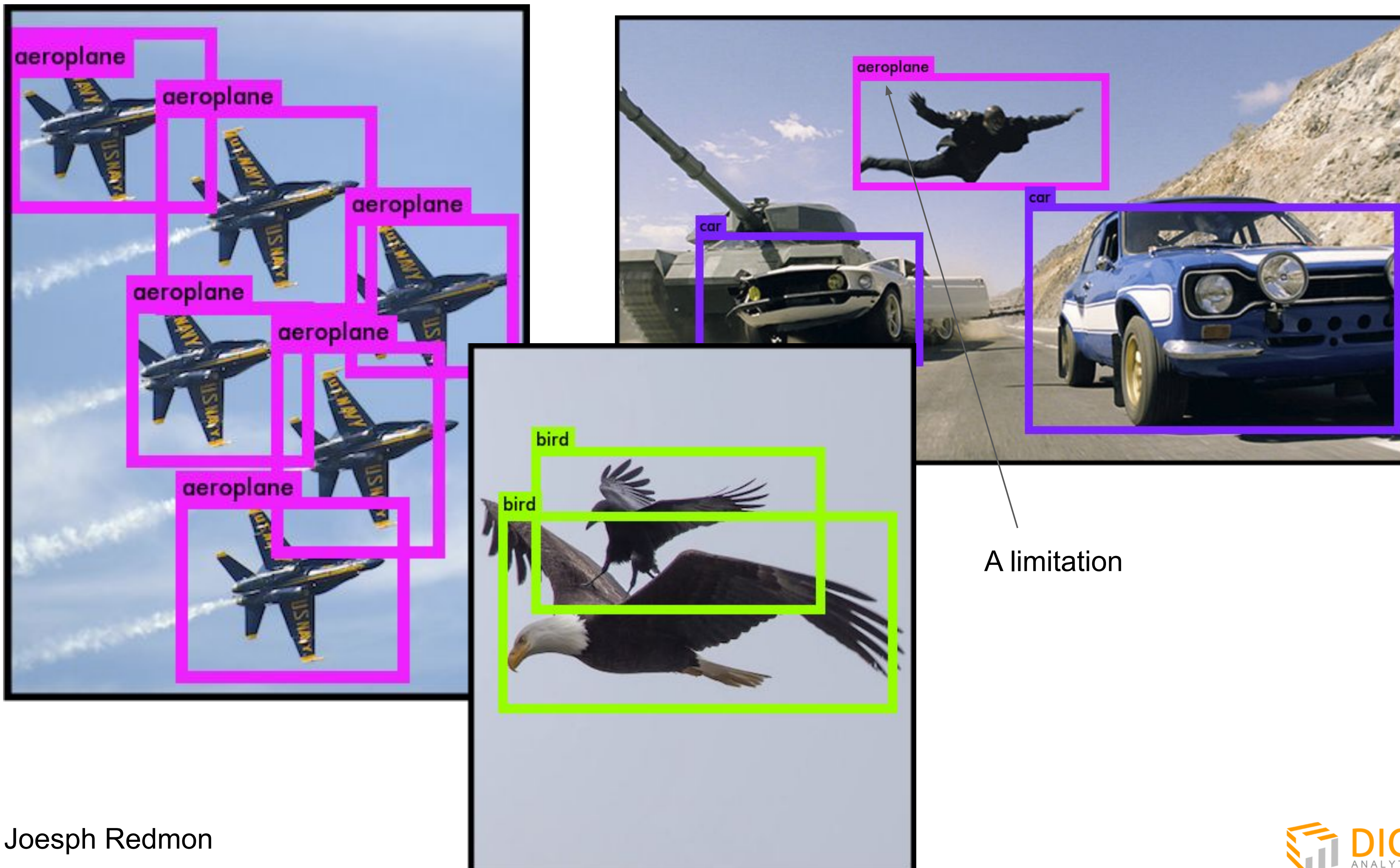
The working parts

Non-Maximum suppression:

To prevent multiple detections of the same object in an image.



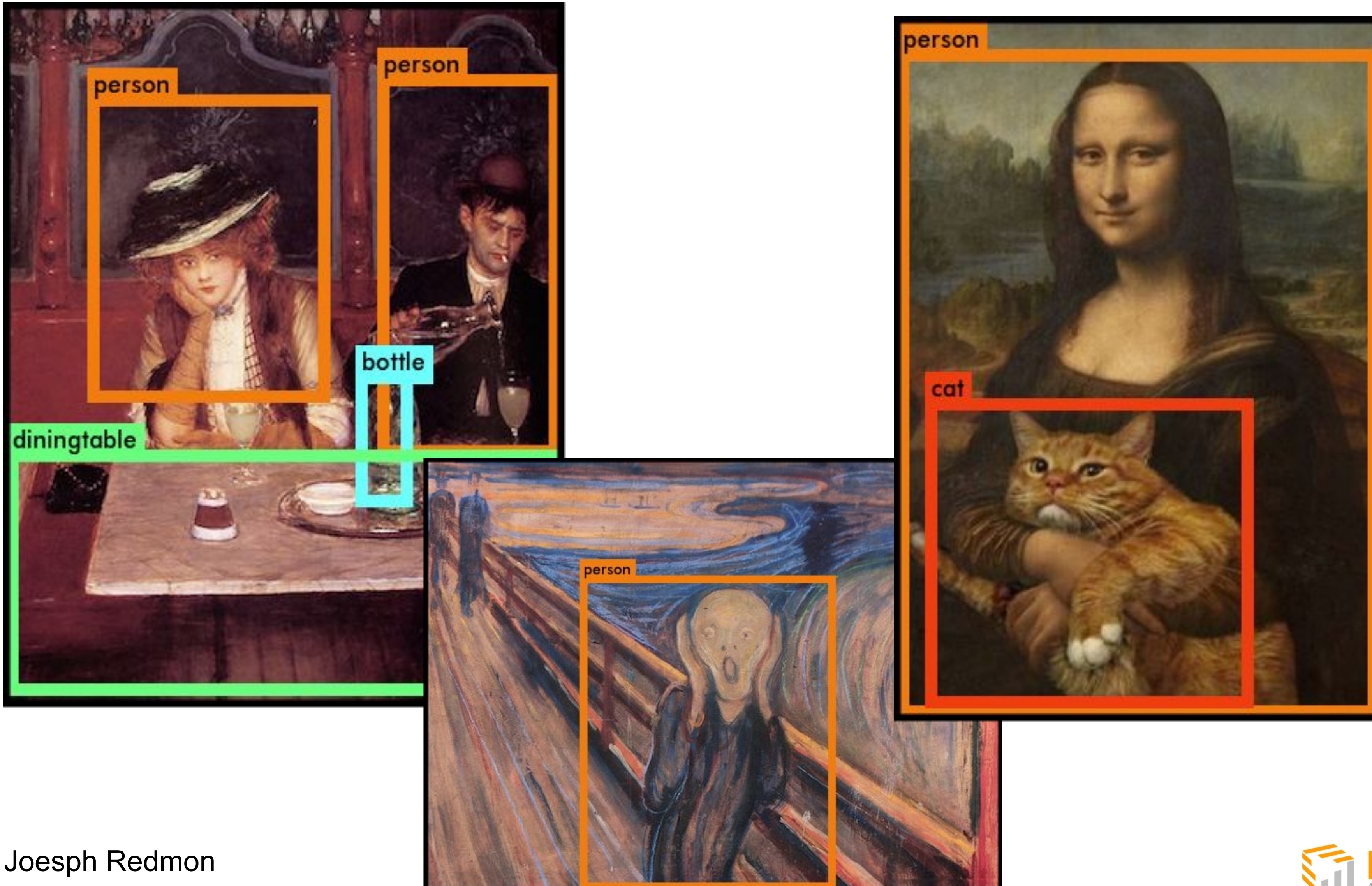
YOLO works across a variety of natural images



Joesph Redmon

DICE ANALYTICS

It also generalizes well to new domains (like art)



Joesph Redmon