CVPR2016, CVPR2017

iMorpheus

自动驾驶论文讲坛：

YOLO, YOLO9000:
Unified, Real-Time Object Detection

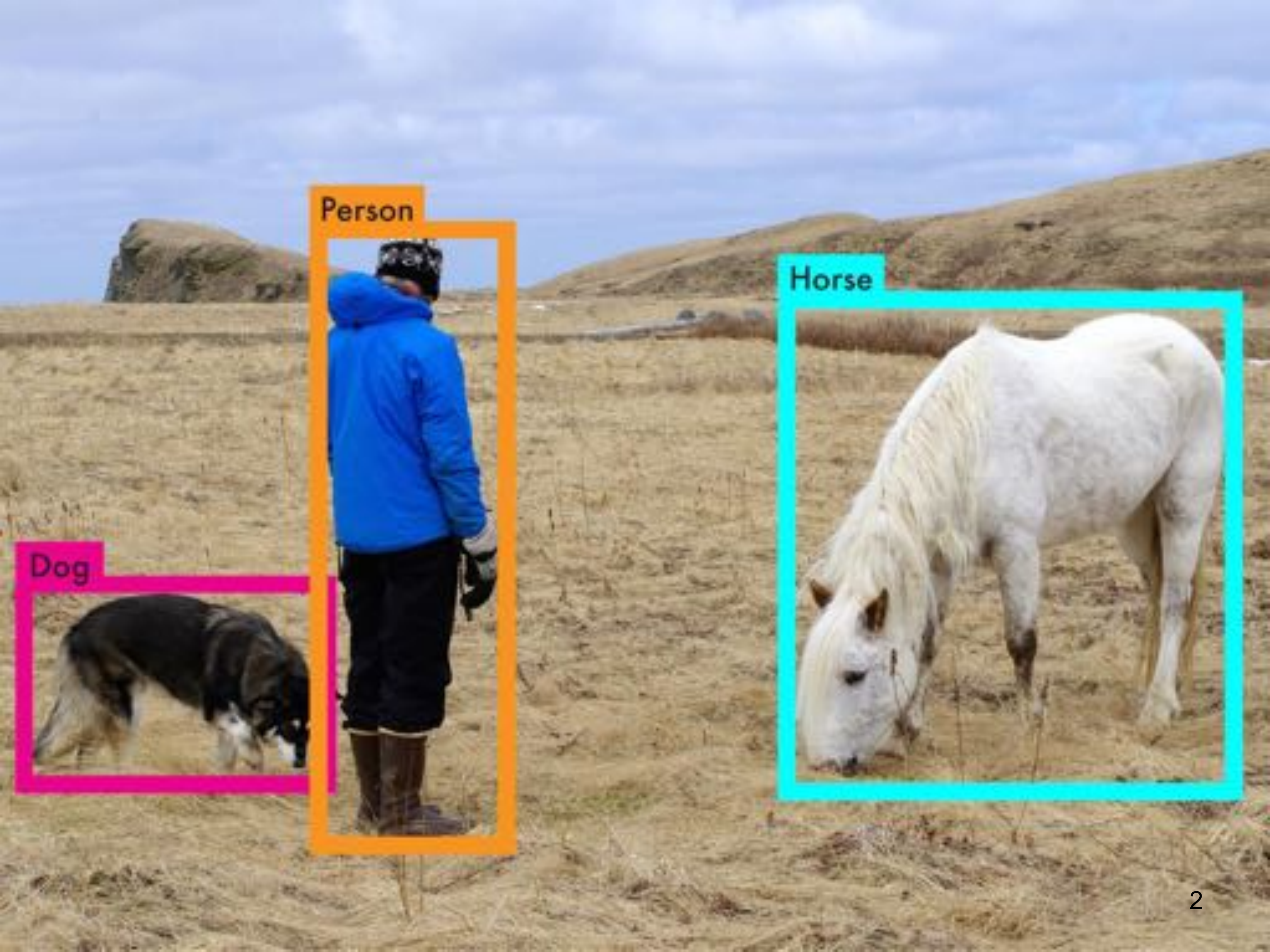
Authors: Joseph Redmon, Ali Farhadi, Santosh Divvala, Ross Girshick

Speaker: Yuehong Huang

Friday 17/11/2017

12:00 PM (GMT+8)

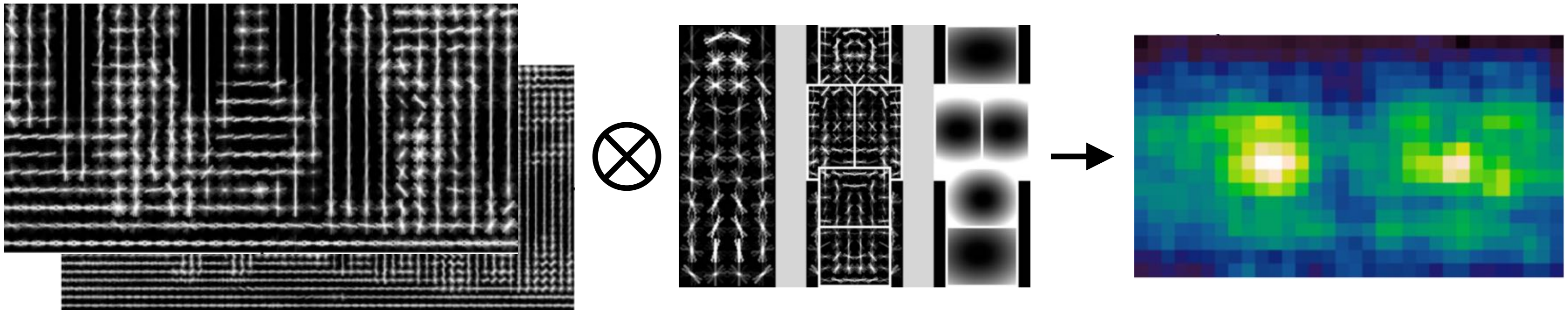
Zoom.us Webinar


Person
Horse
Dog

Comparison to Other Detection System -- Accurate object detection is slow!

	Pascal 2007 mAP	**Speed**	
DPM v5	33.7	.07 FPS	14 s/img

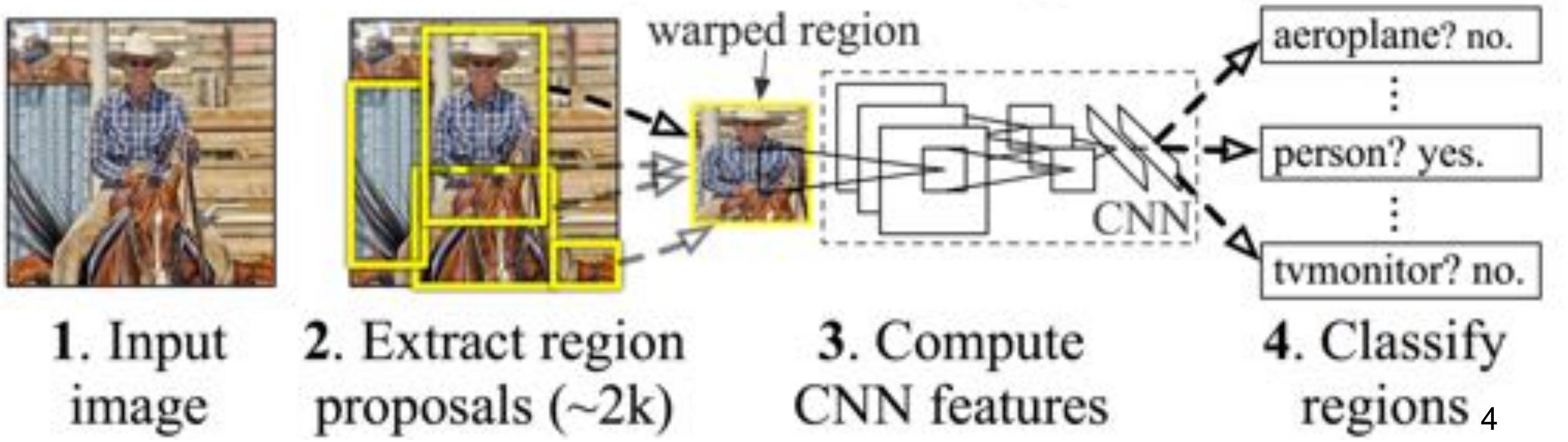
DPM: ***Deformable Part Models***



Accurate object detection is slow!

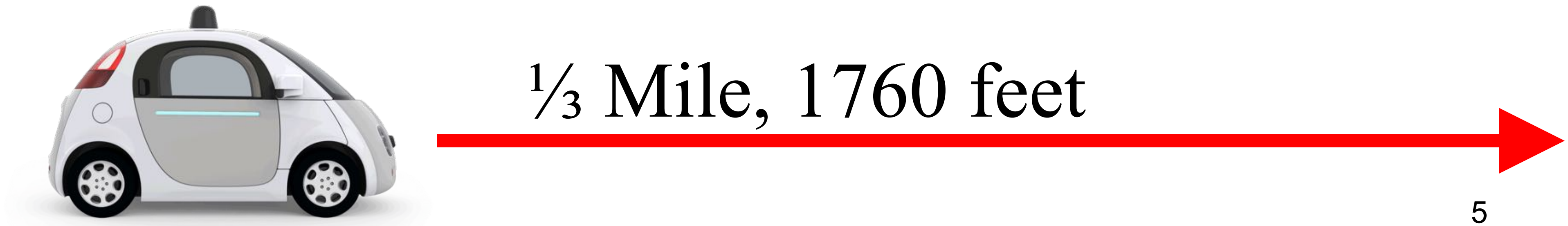
	Pascal 2007 mAP	Speed	
DPM v5	33.7	.07 FPS	14 s/img
R-CNN	66.0	.05 FPS	20 s/img

R-CNN: *Regions with CNN features*



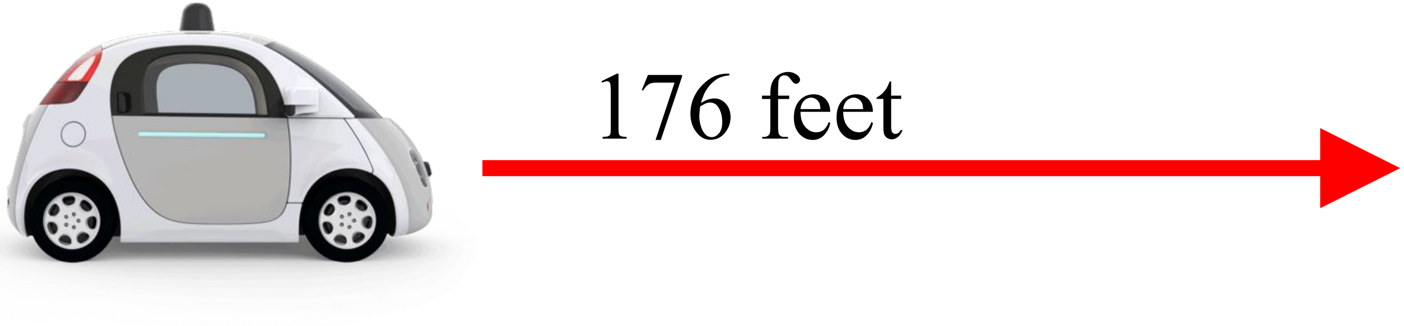
Accurate object detection is slow!

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



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

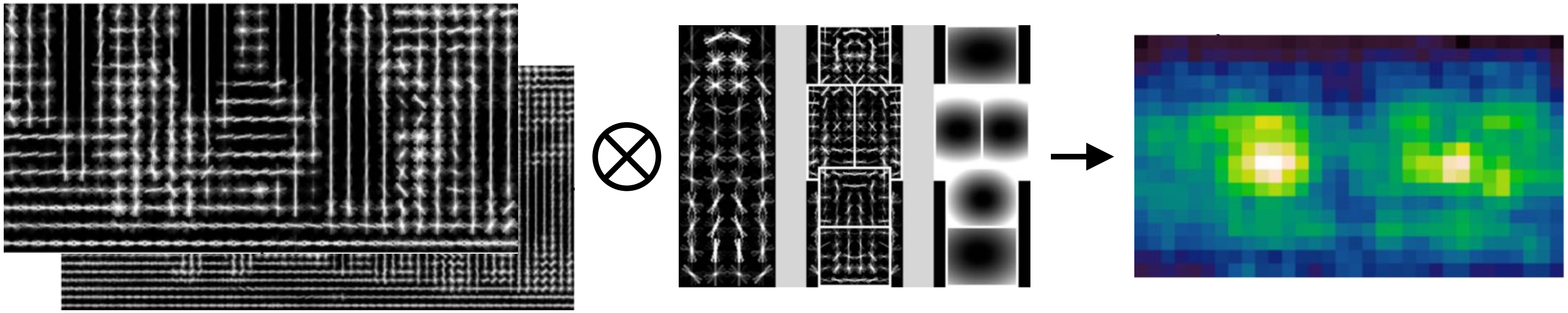
Accurate object detection is slow!

	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

8 feet
12 feet

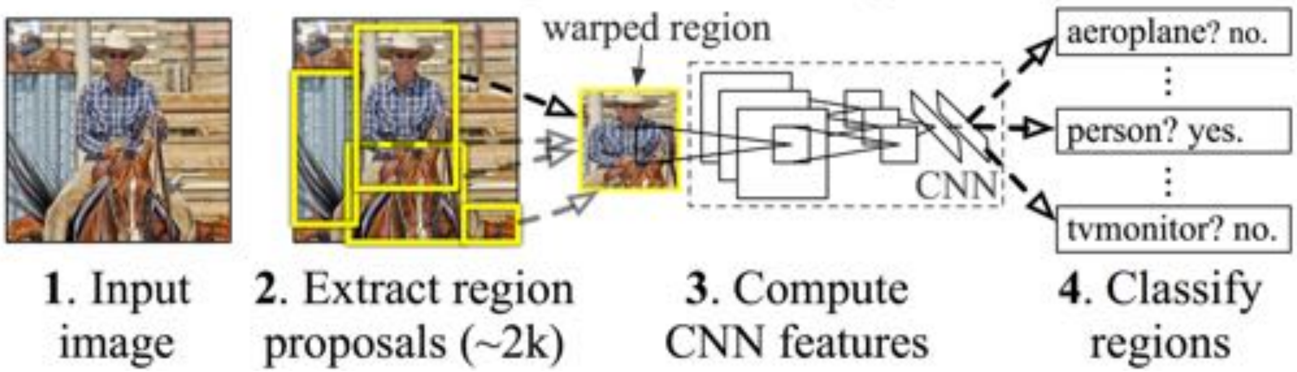
Sliding window, DPM, R-CNN all train region-based classifiers to perform detection

DPM: ***Deformable Part Models***



Complex Pipeline

R-CNN: ***Regions with CNN features***



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	63.4	45 FPS	22 ms/img

2 feet

YOLO can be better!

	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	63.4	45 FPS	22 ms/img

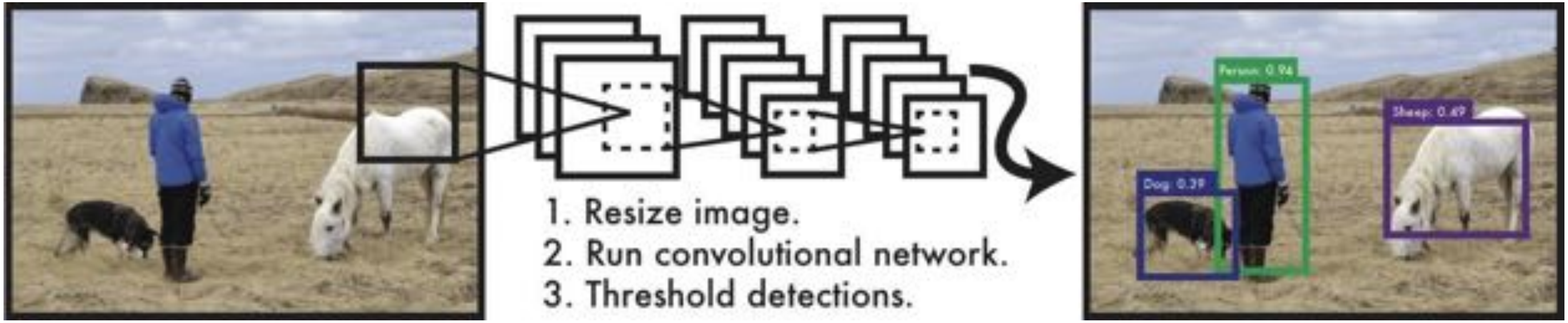
🙁 ↗

↖ 😁

YOLOv2, YOLO9000

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

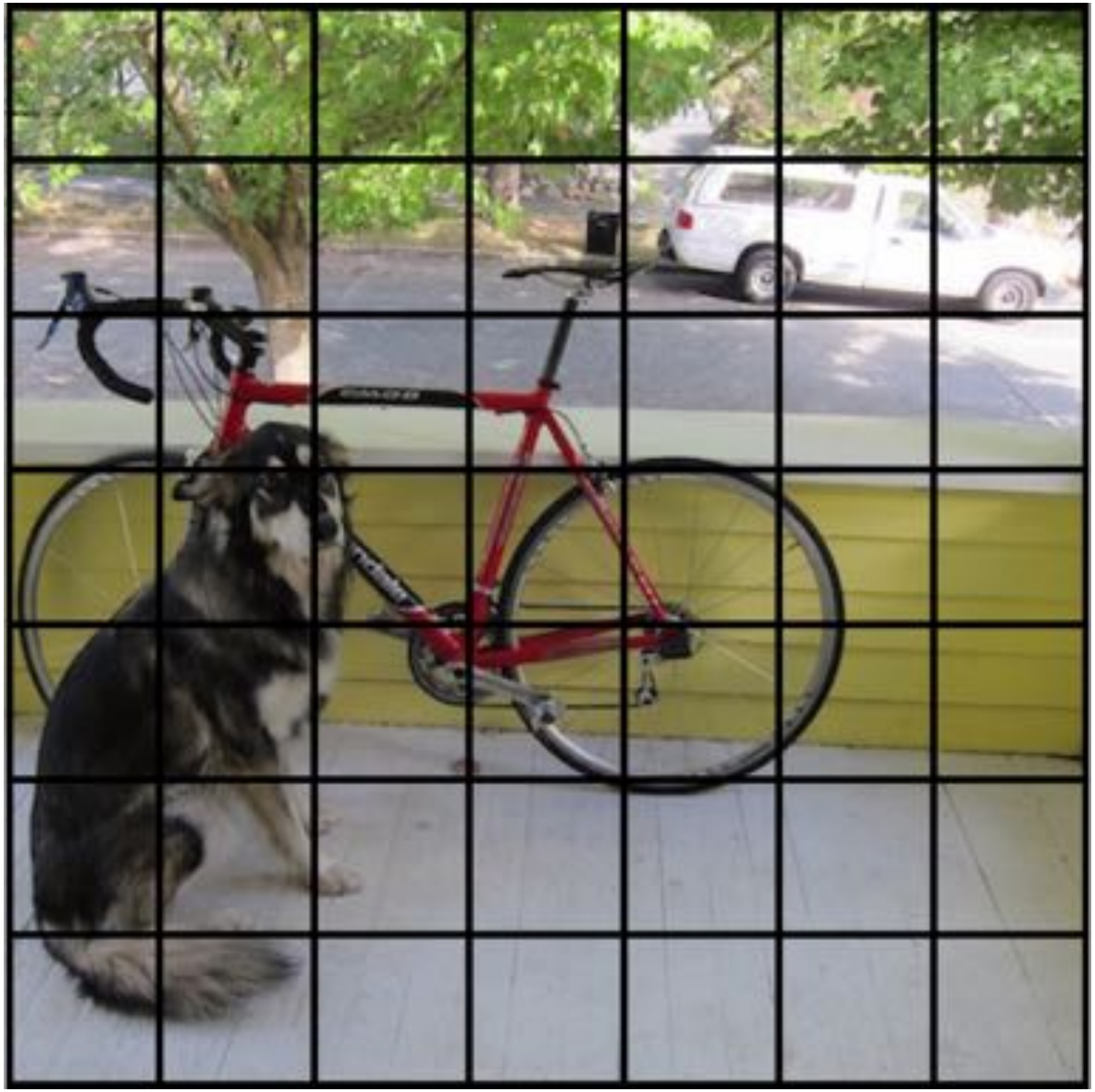
YOLO: ***You Only Look Once***



Unified Model:

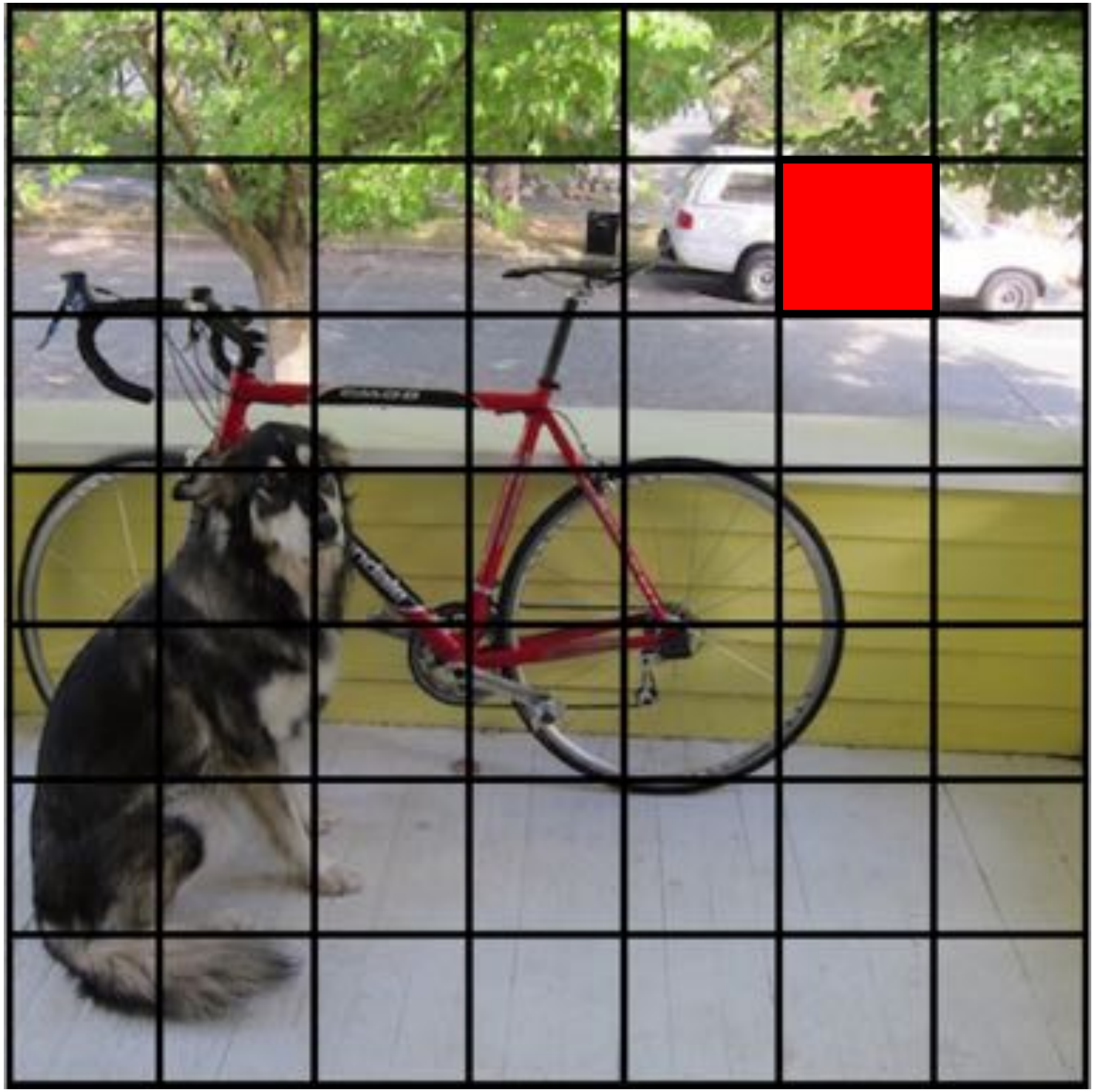
1. YOLO is extremely fast -- no complex pipeline
2. Twice the mean average precision of other real-time systems
3. YOLO reasons globally about the image – less background errors
4. YOLO learns generalizable representations of objects – new domain and unexpected input (art works).

Unified Detection -- we split the image into a grid



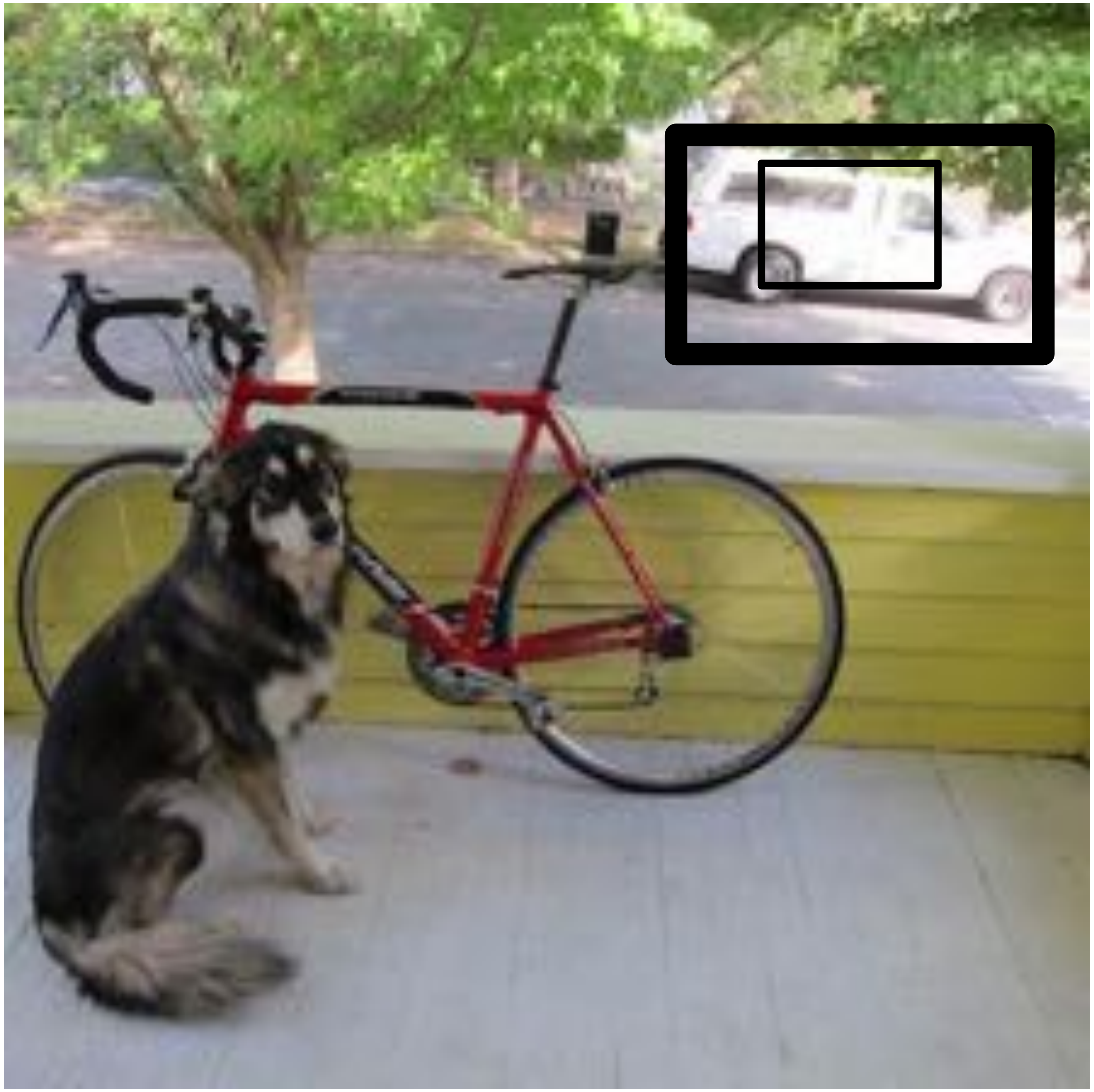
iMorpheus

Each cell predicts boxes and confidences: P(Object)



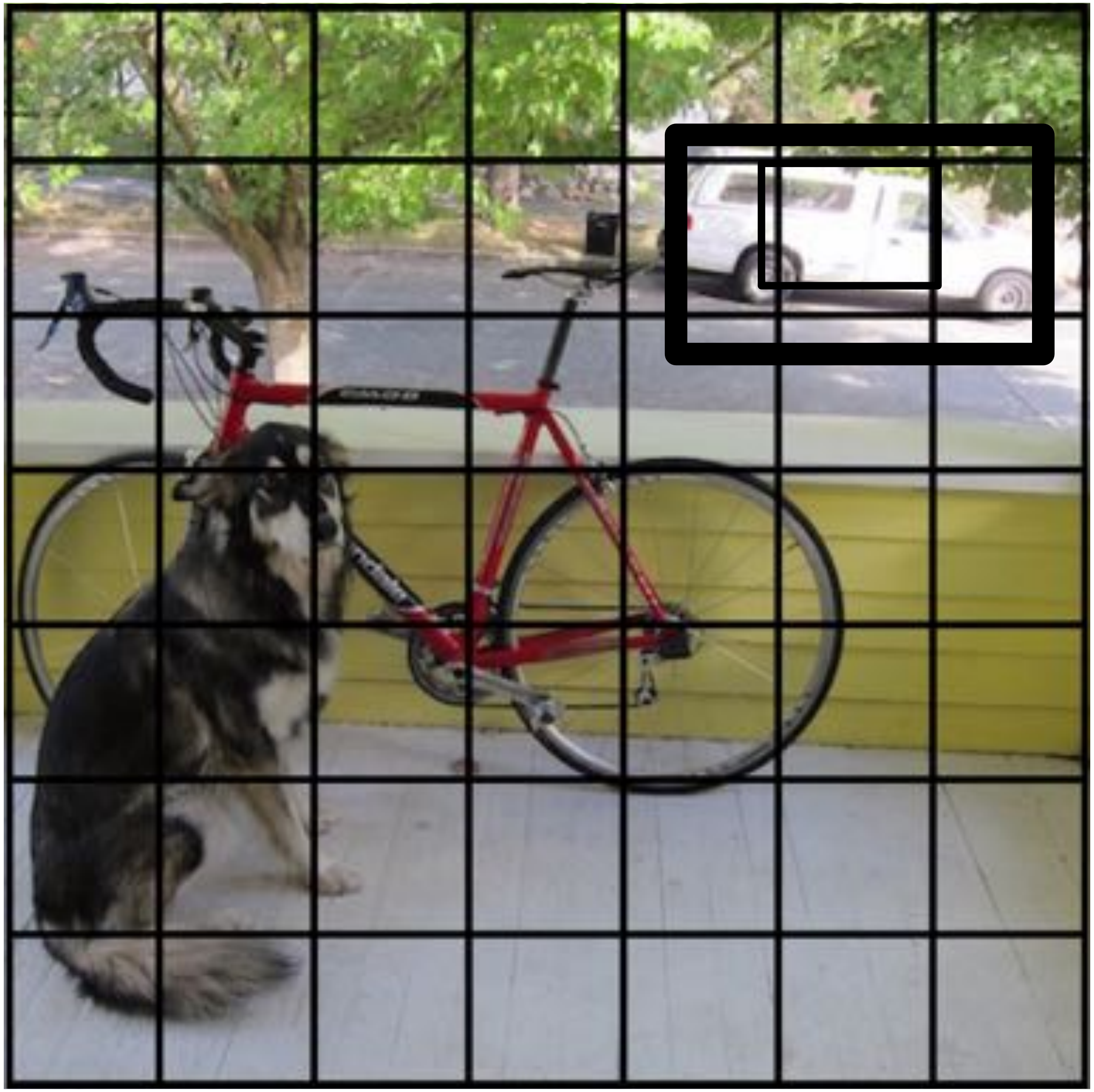
iMorpheus

Each cell predicts boxes and confidences: P(Object)



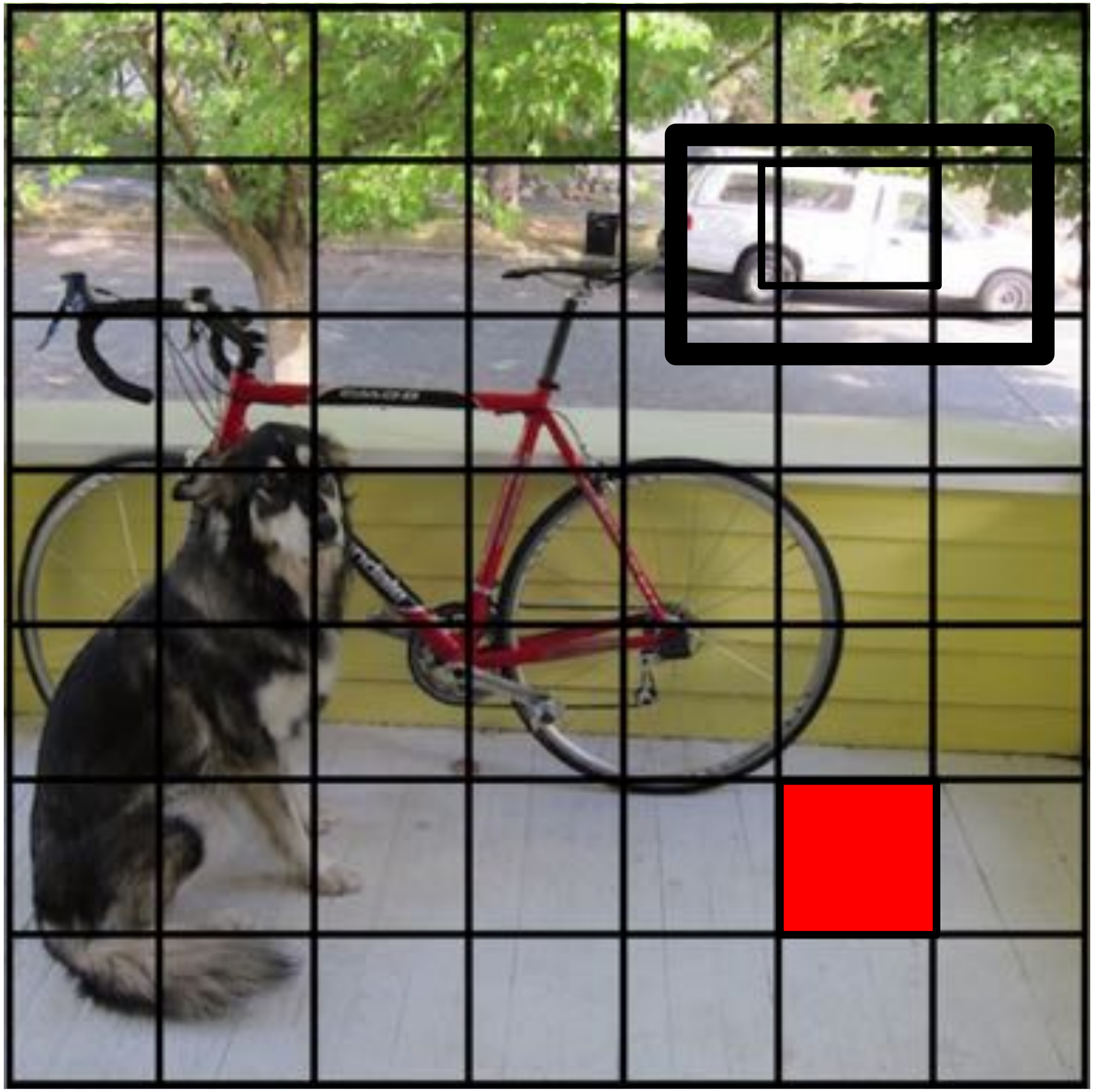
iMorpheus

Each cell predicts boxes and confidences: P(Object)

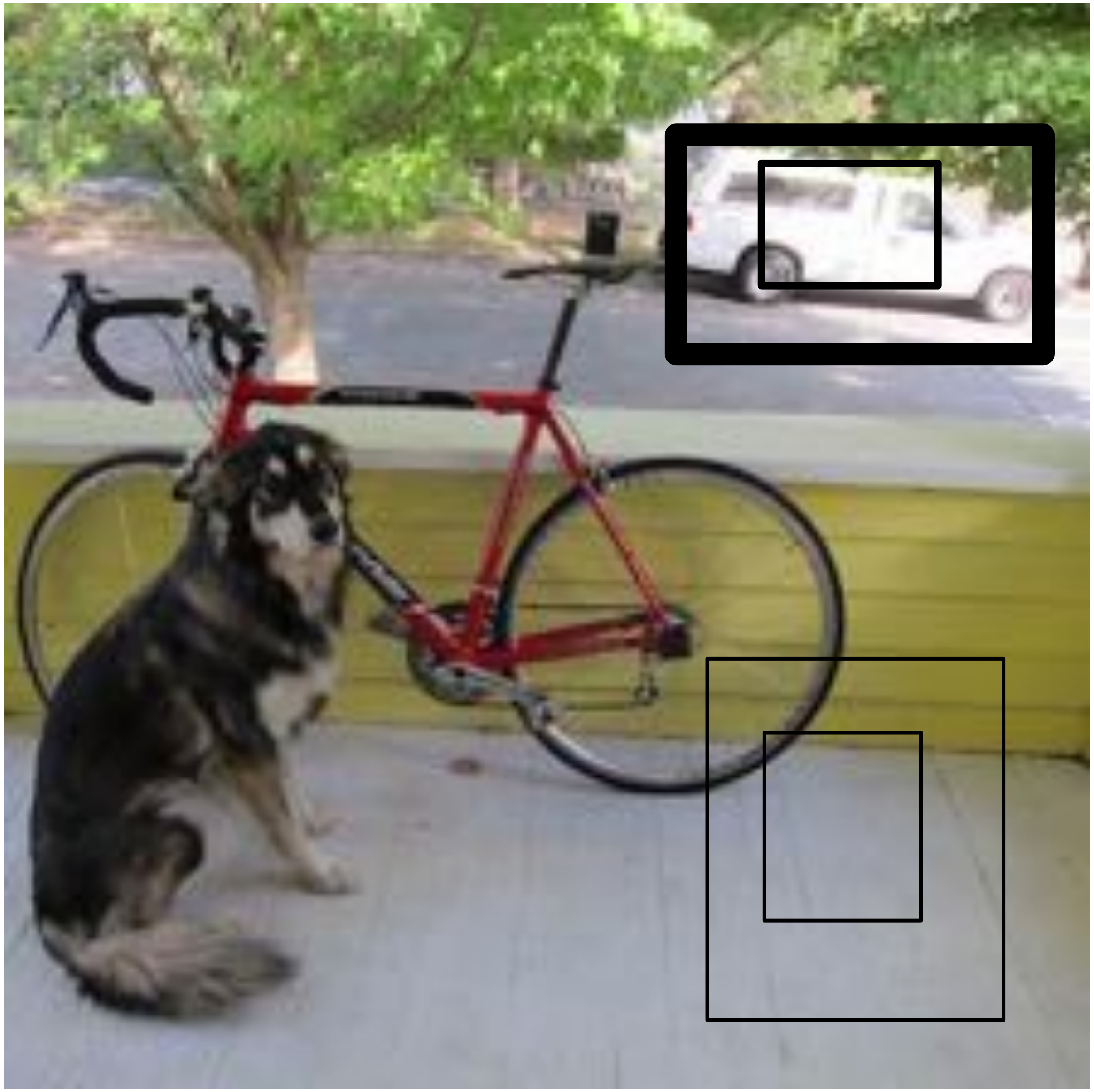


iMorpheus

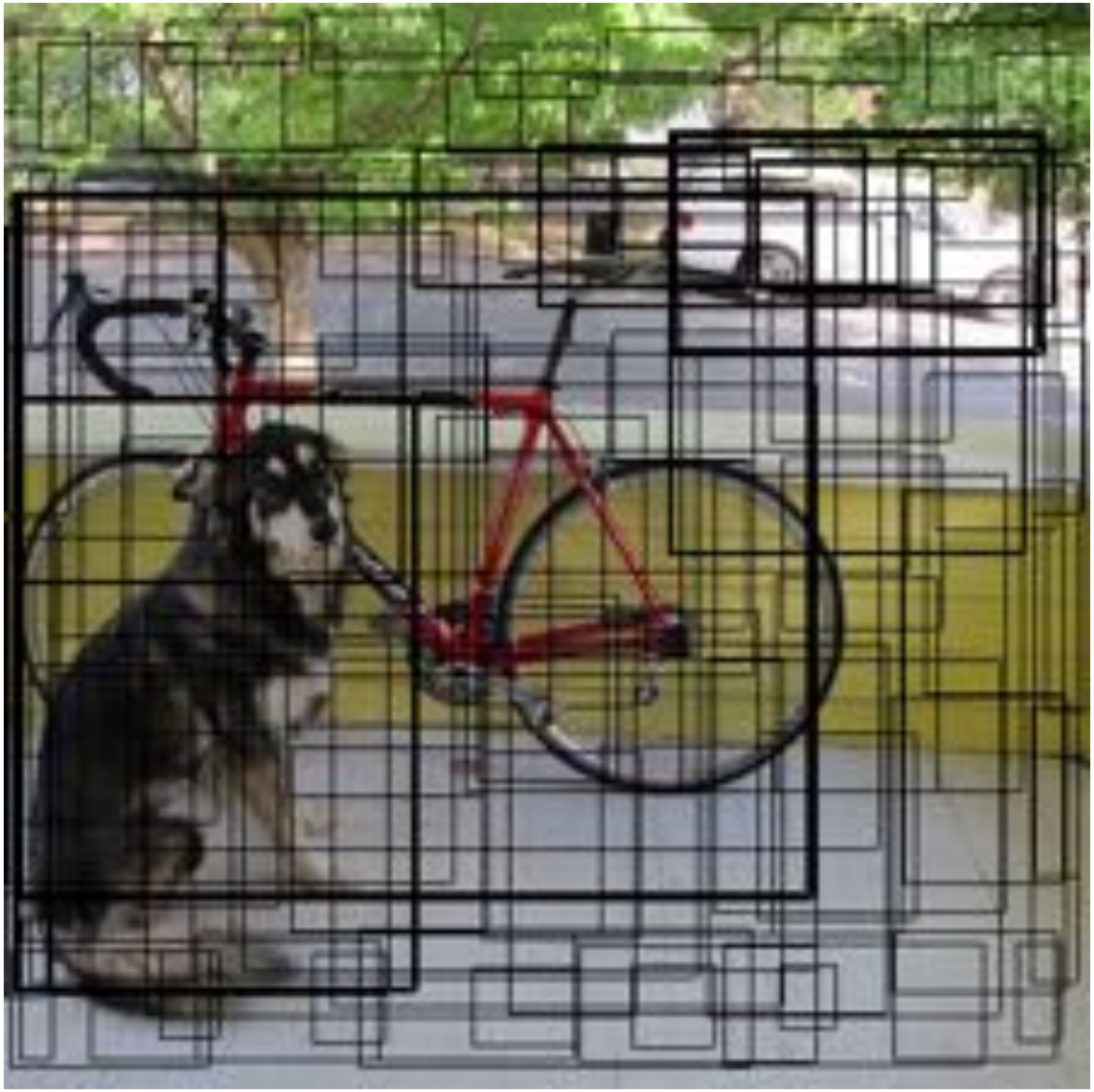
Each cell predicts boxes and confidences: P(Object)



Each cell predicts boxes and confidences: P(Object)



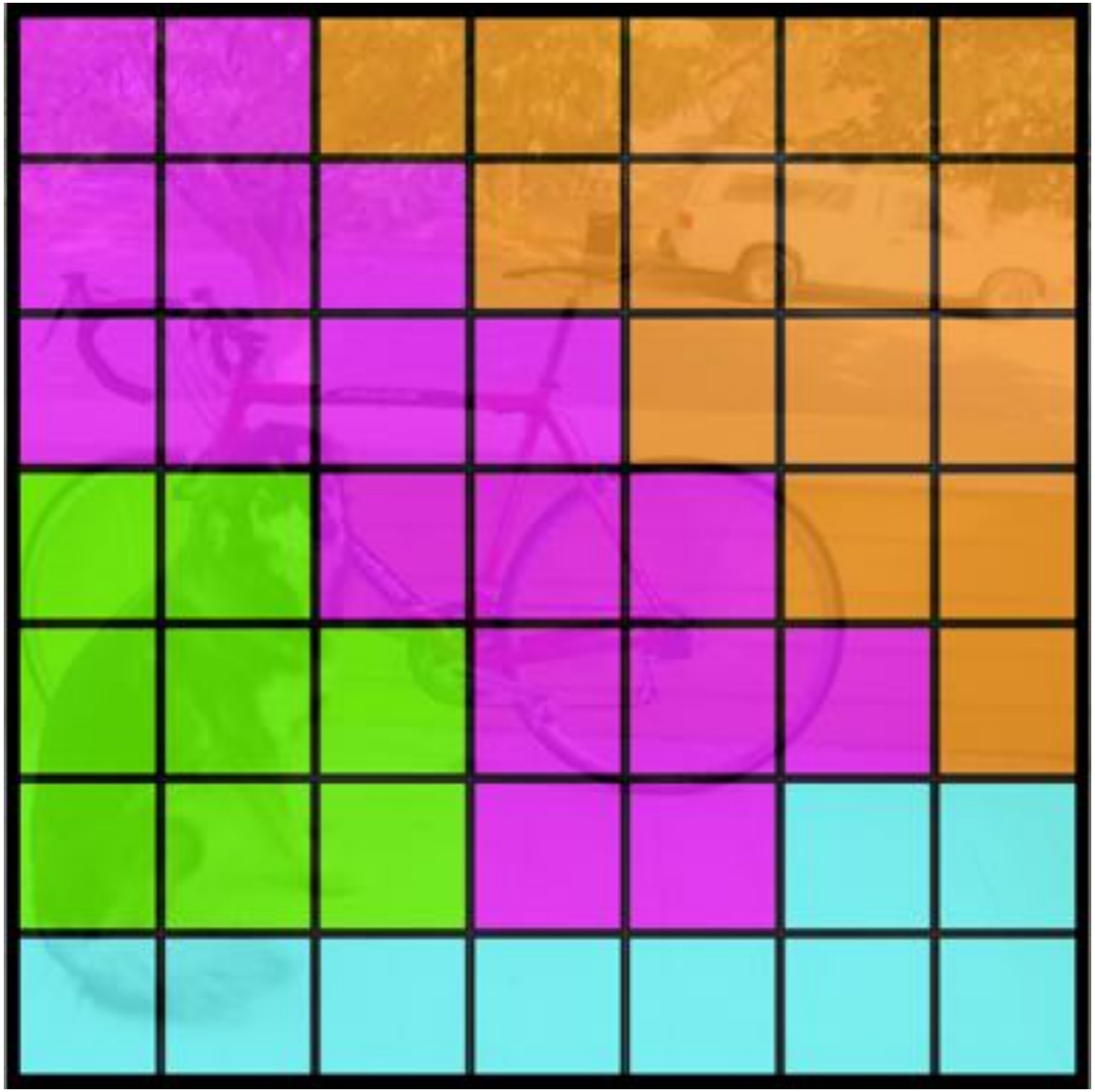
Each cell predicts boxes and confidences: P(Object)



iMorpheus

Each cell also predicts a class probability.

Conditioned on object: P(Car | Object)



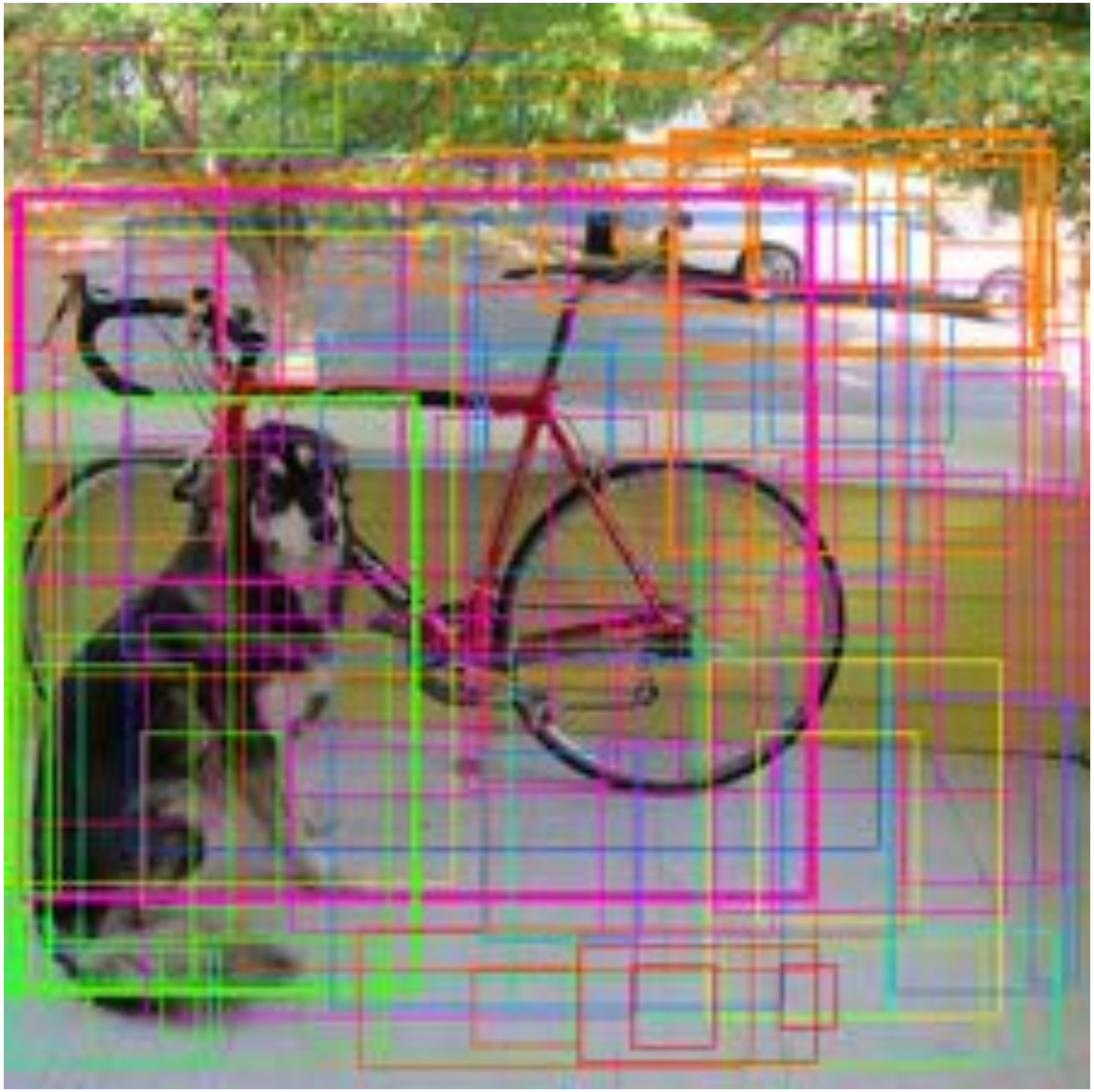
Bicycle

Car

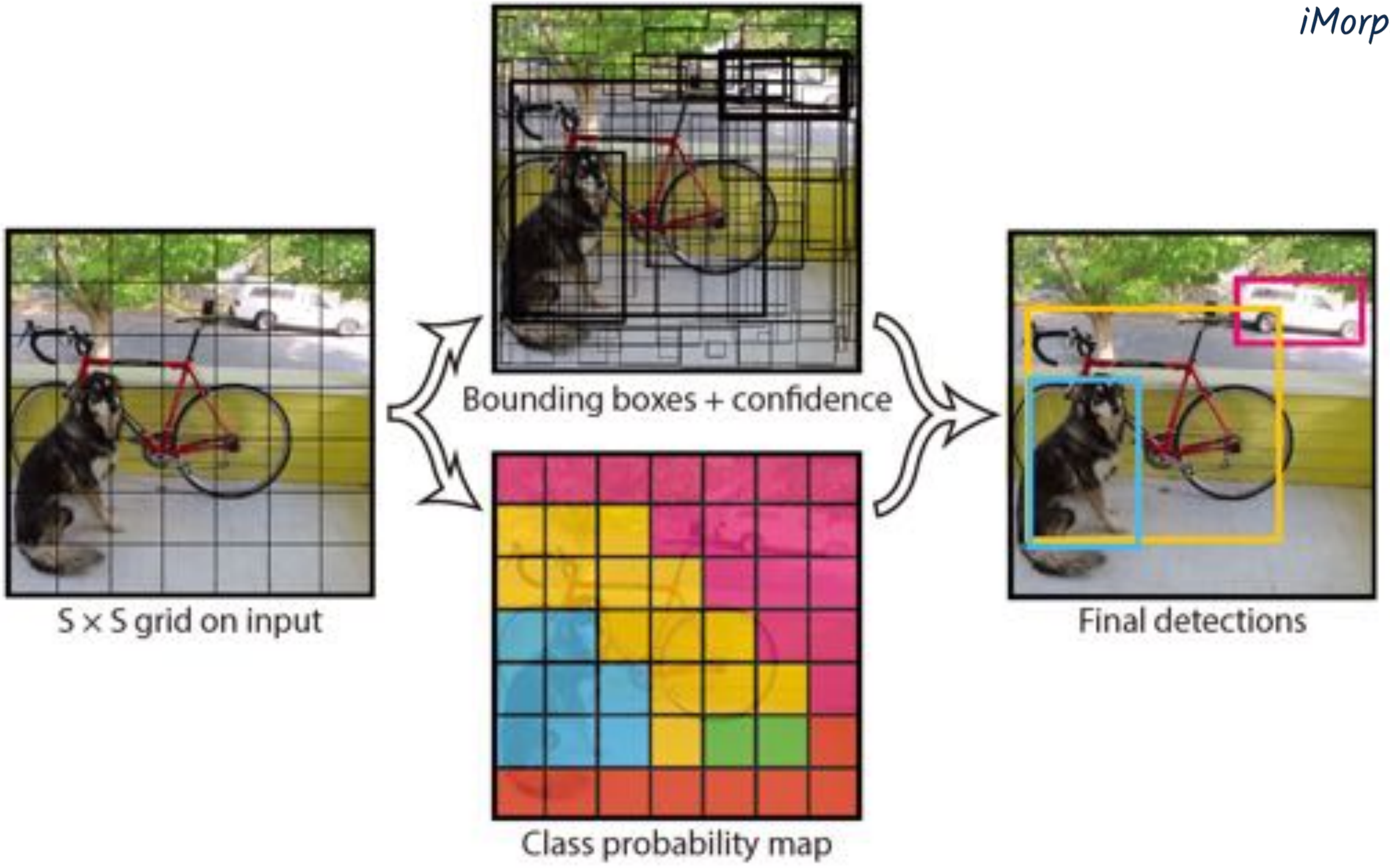
Dog

Dining Table

Then we combine the box and class predictions.



Finally we do threshold detections

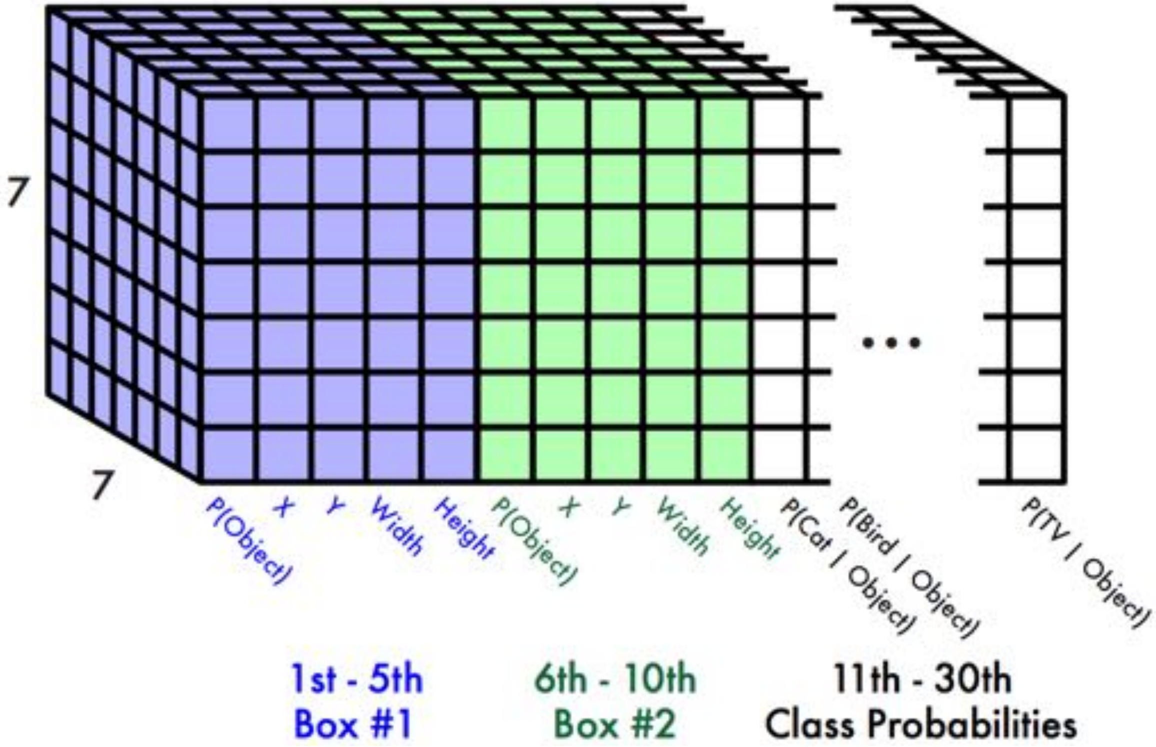


iMorpheus

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



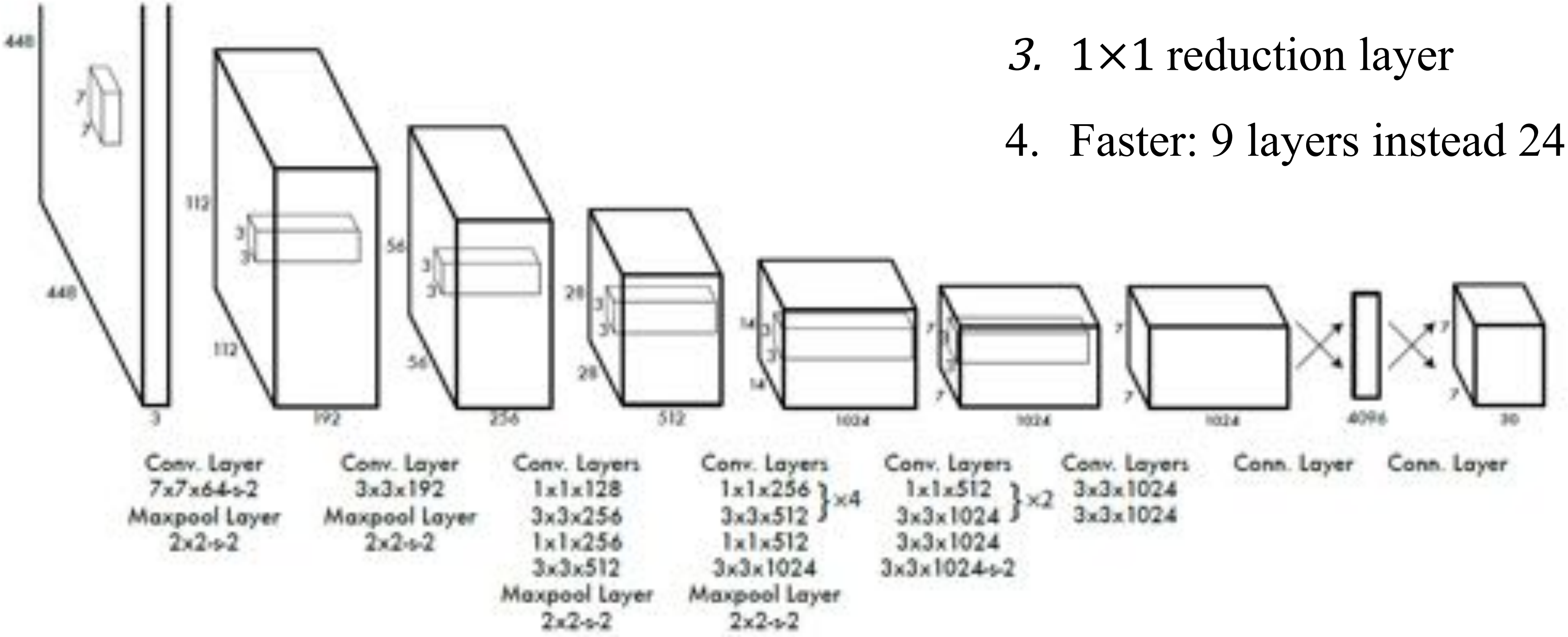
For Pascal VOC:

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

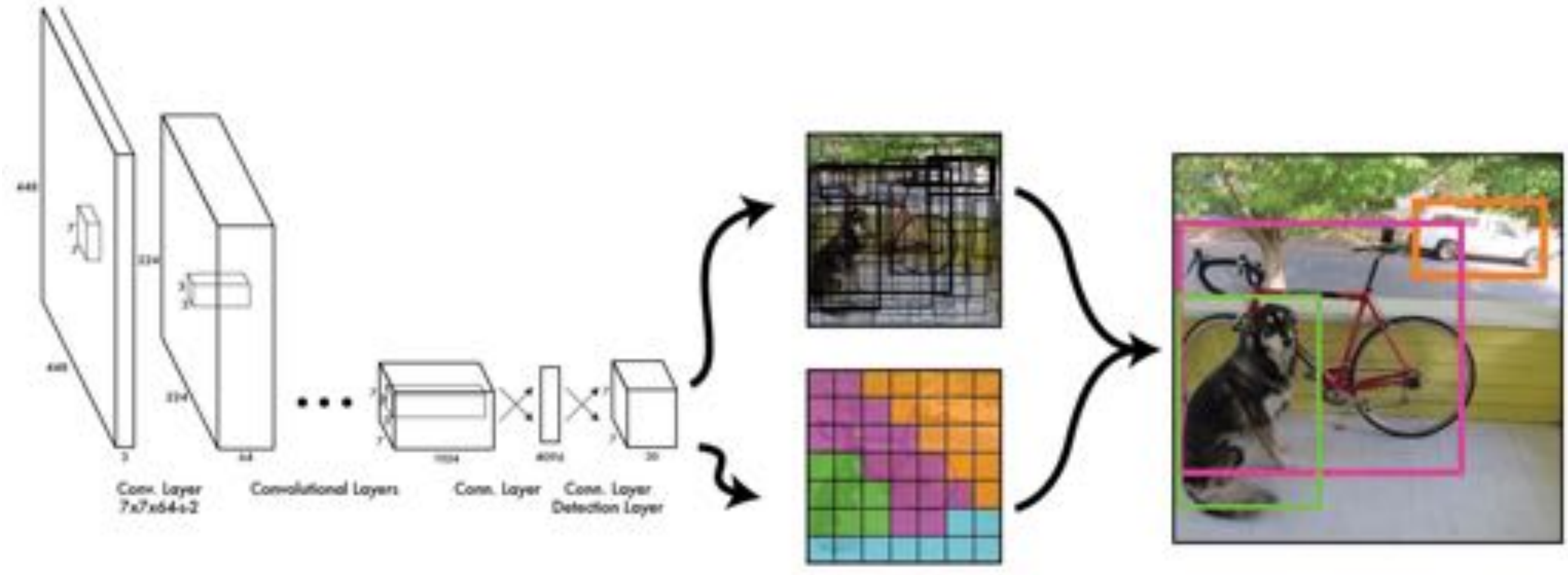
7 x 7 x (2 x 5 + 20) = 7 x 7 x 30 tensor = **1470 outputs**

The architecture of network

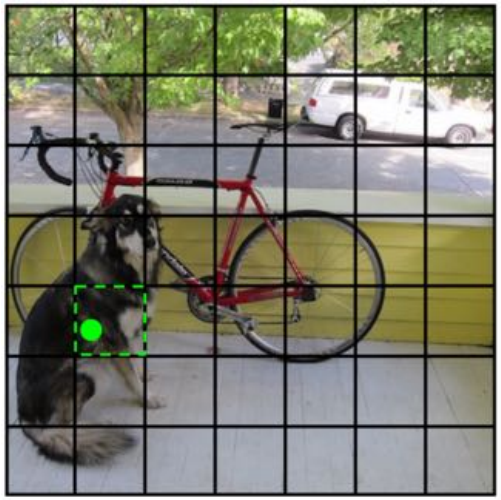
1. 24 convolutional layers
2. 2 fully connected layers
3. 1×1 reduction layer
4. Faster: 9 layers instead 24



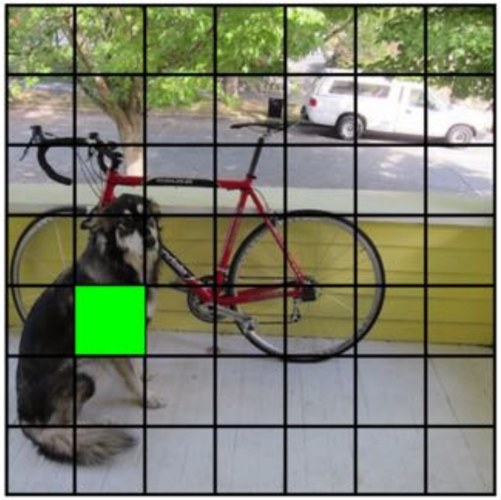
Thus we can train one neural network to be a whole detection pipeline

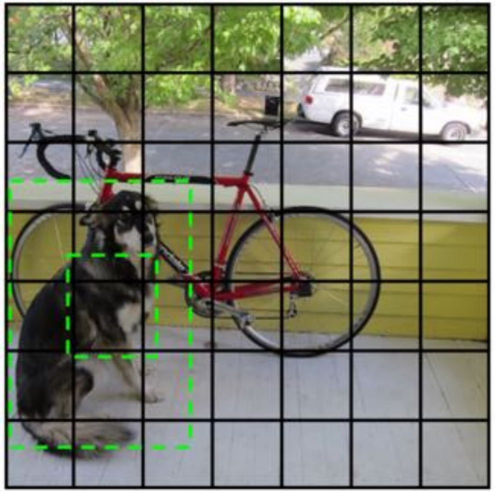


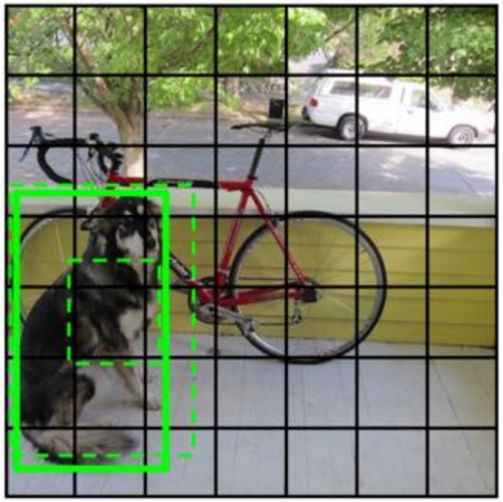
During training, match example to the right cell

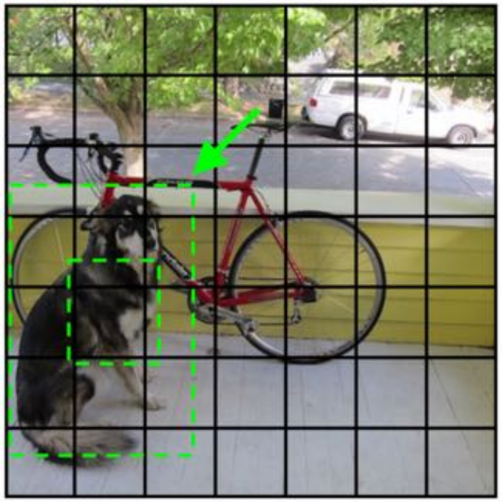


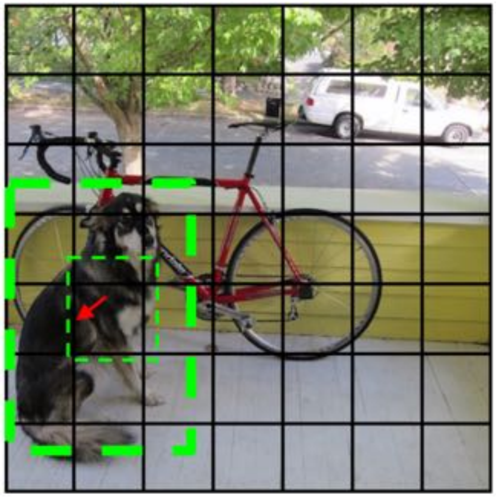
Dog = 1
Cat = 0
Bike = 0
...



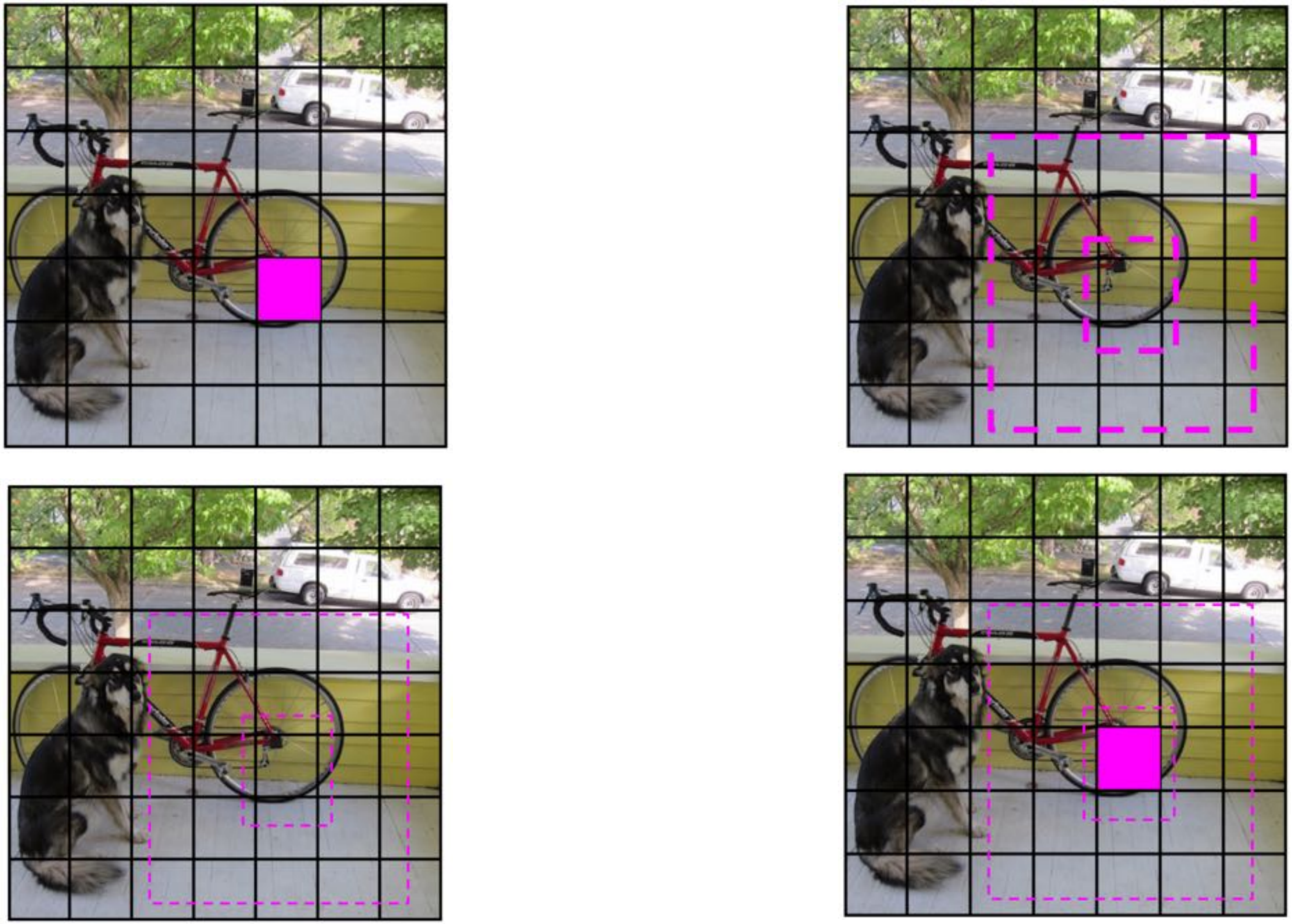






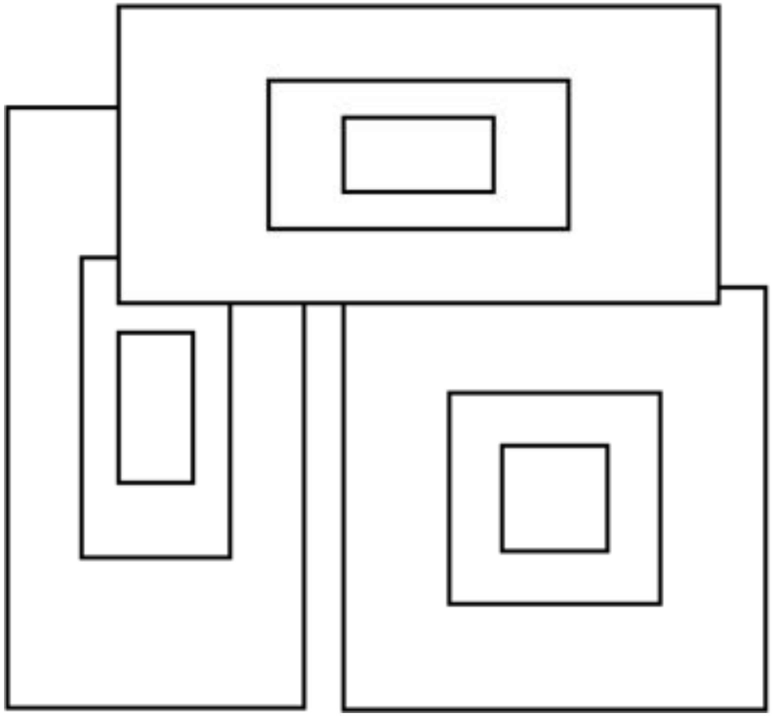


Some cells don’t have any ground truth detections!



YOLOv2, YOLO9000

Anchor boxes use static initialization



Sizes:

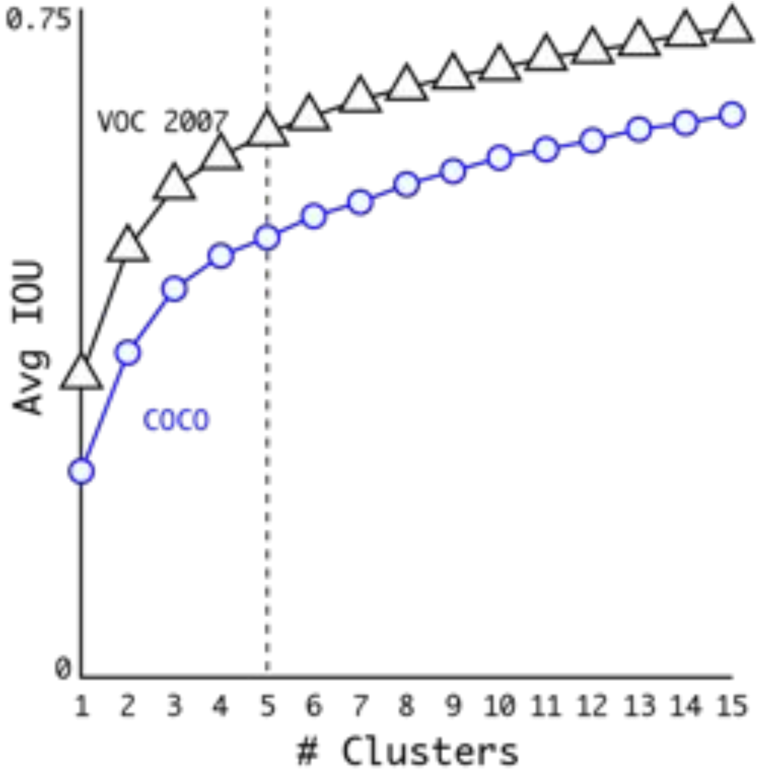
- 128
- 256
- 512

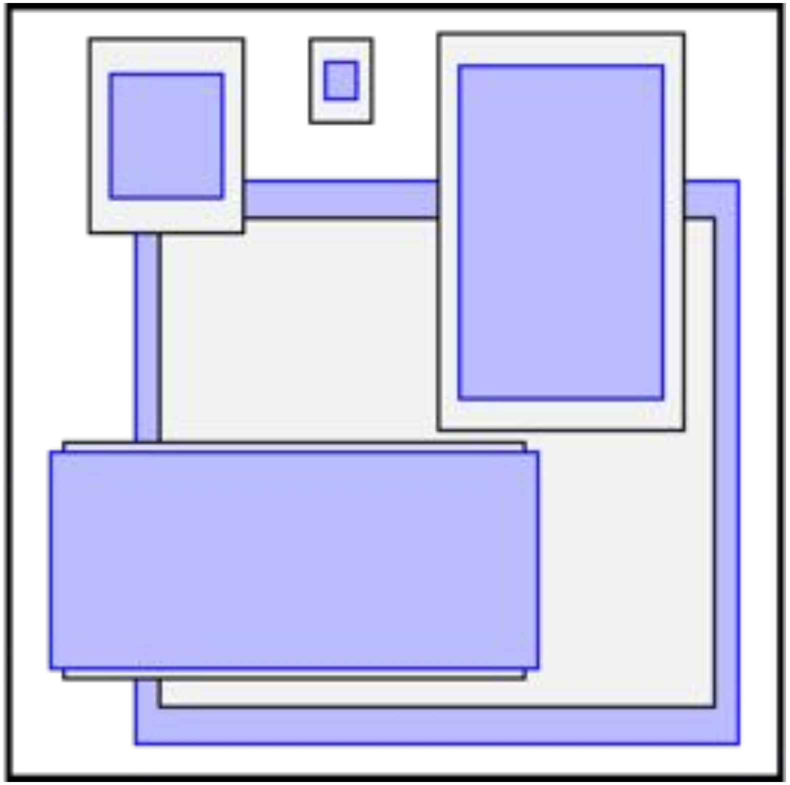
Ratios:

- 1x1
- 1x2
- 2x1

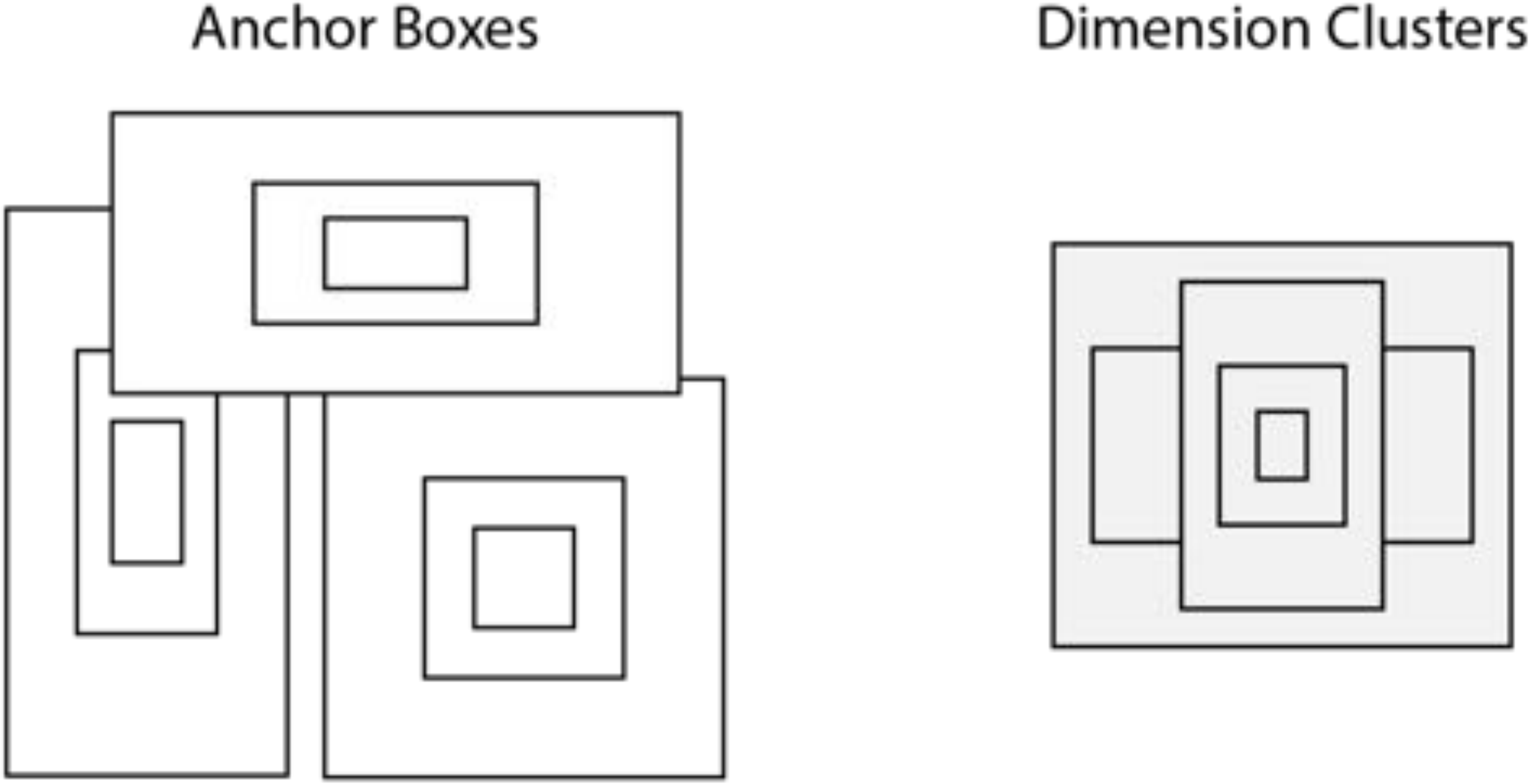
YOLOv2, YOLO9000

We use k-means to find better initializations



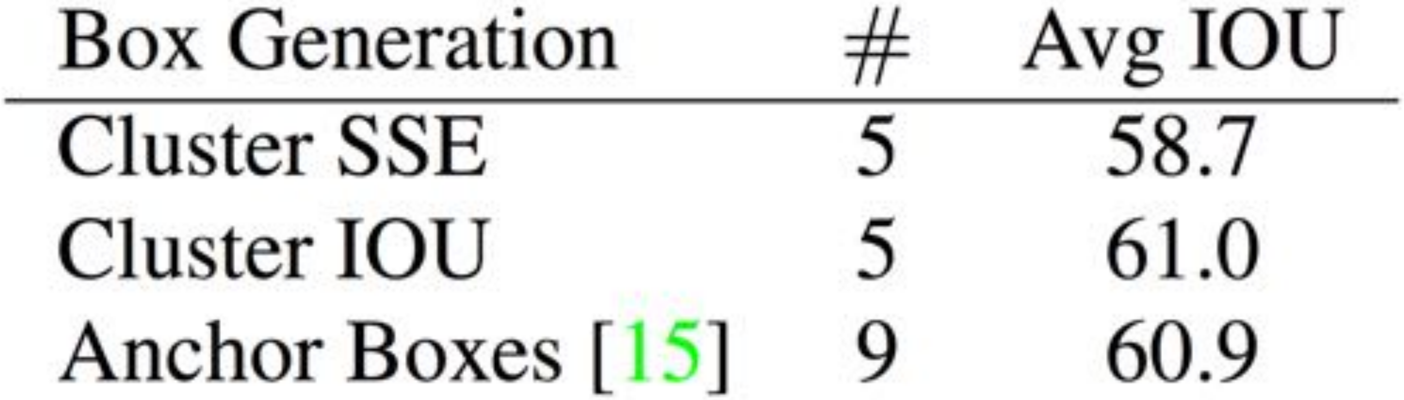


YOLOv2, YOLO9000



YOLOv2, YOLO9000

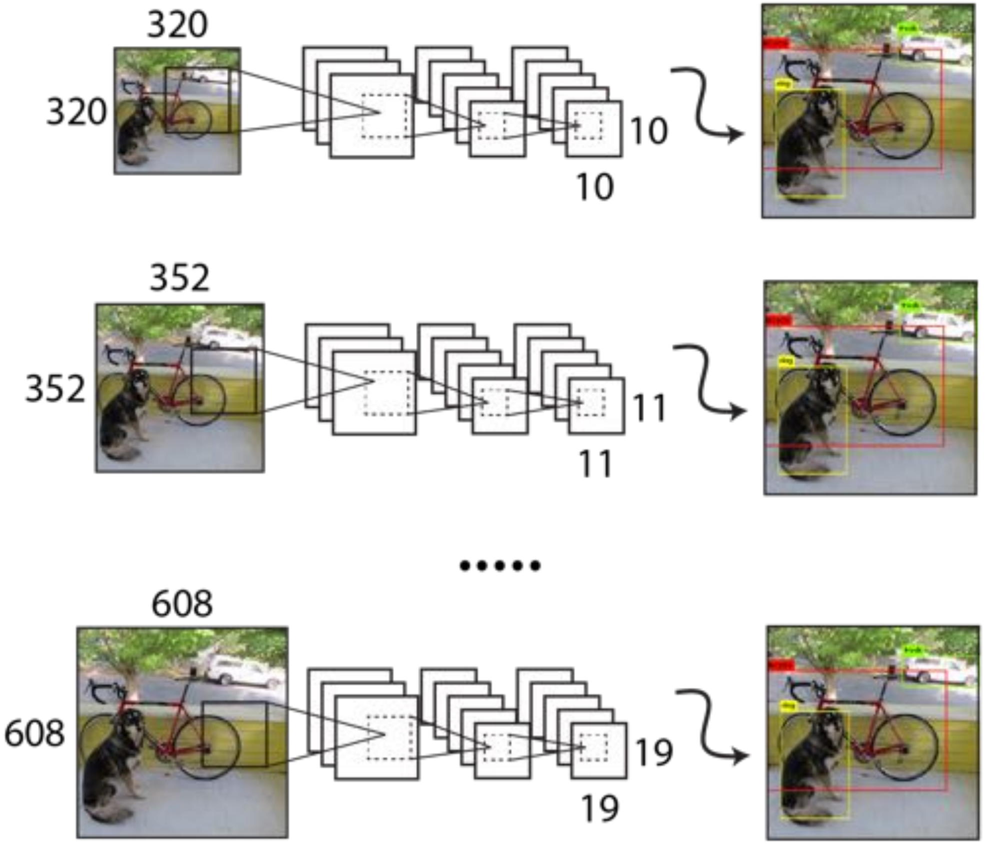
Dimension Clusters: +5% mAP



Box Generation	#	Avg IOU
Cluster SSE	5	58.7
Cluster IOU	5	61.0
Anchor Boxes [15]	9	60.9

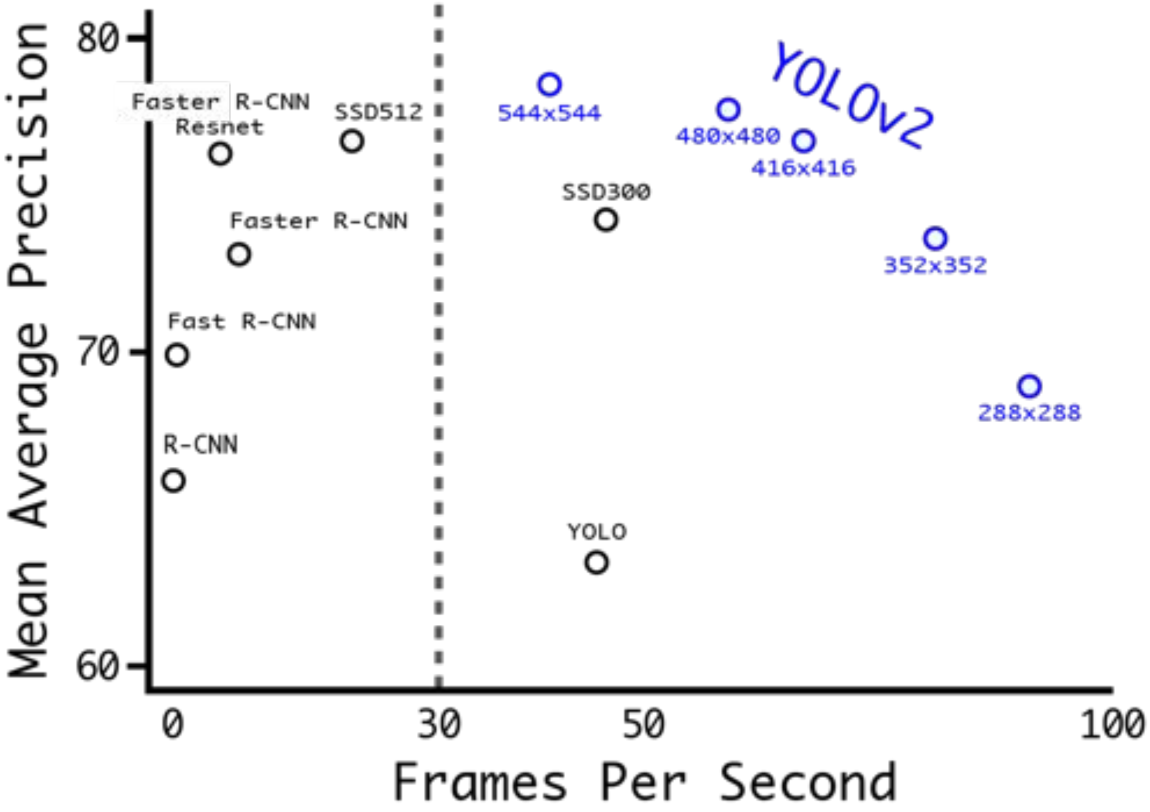
YOLOv2, YOLO9000

Multi-scale training: +1.5% mAP



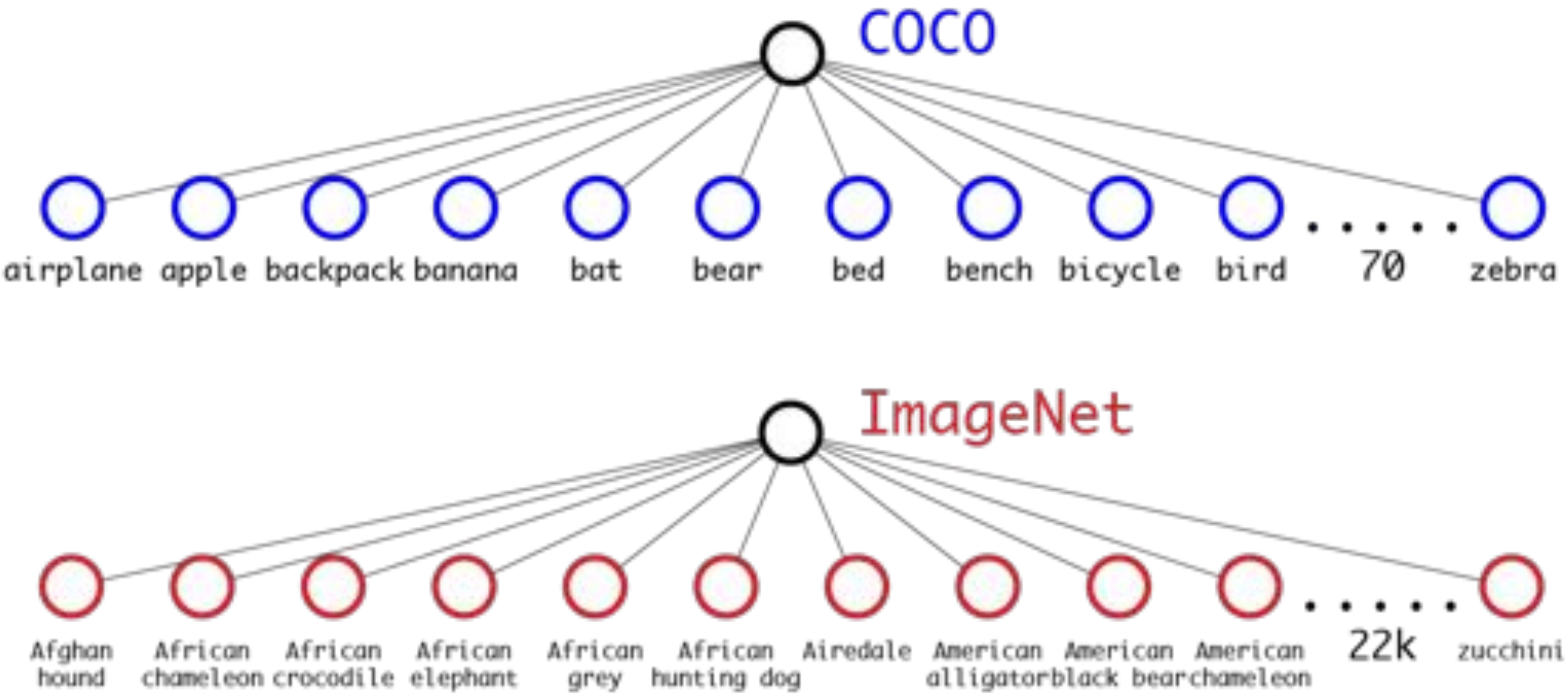
YOLOv2: Fast, Accurate Detection

iMorpheus



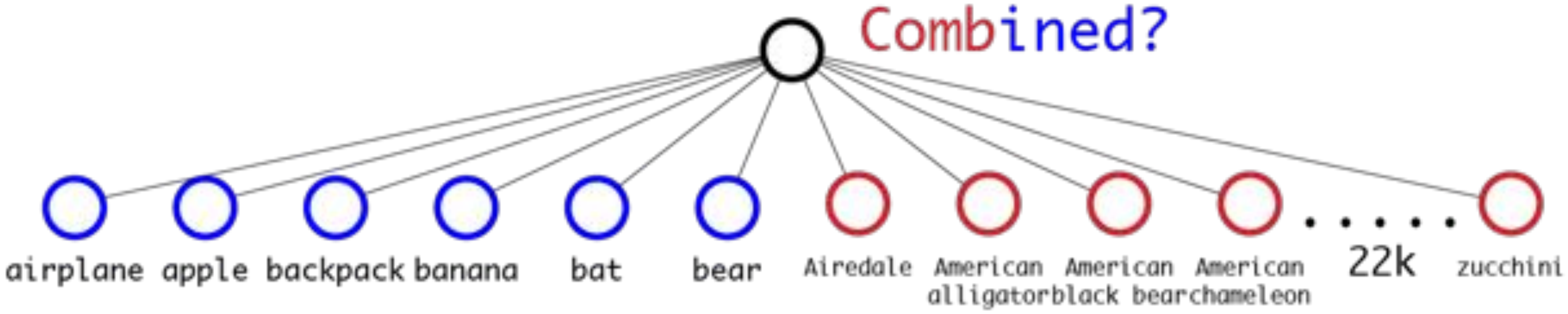
YOLOv2, YOLO9000

Typically use softmax over all classes



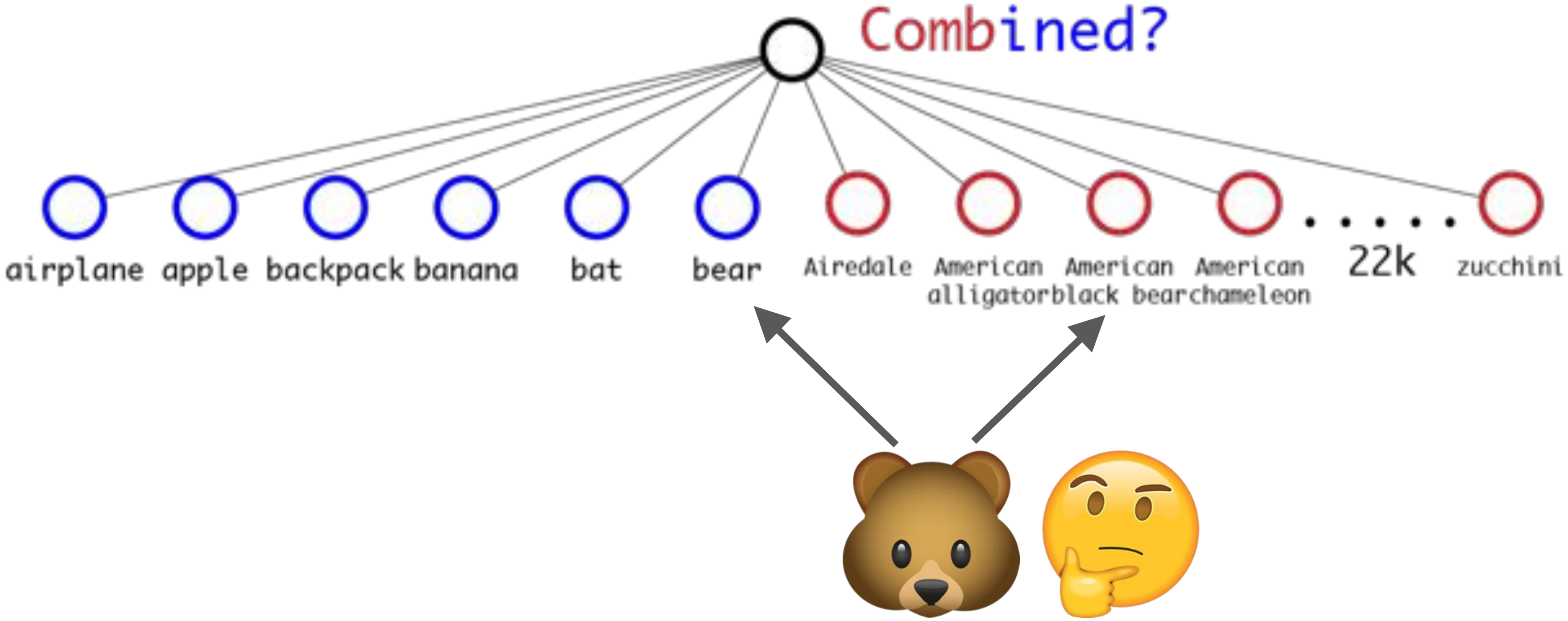
YOLOv2, YOLO9000

Can’t just mash classes together...



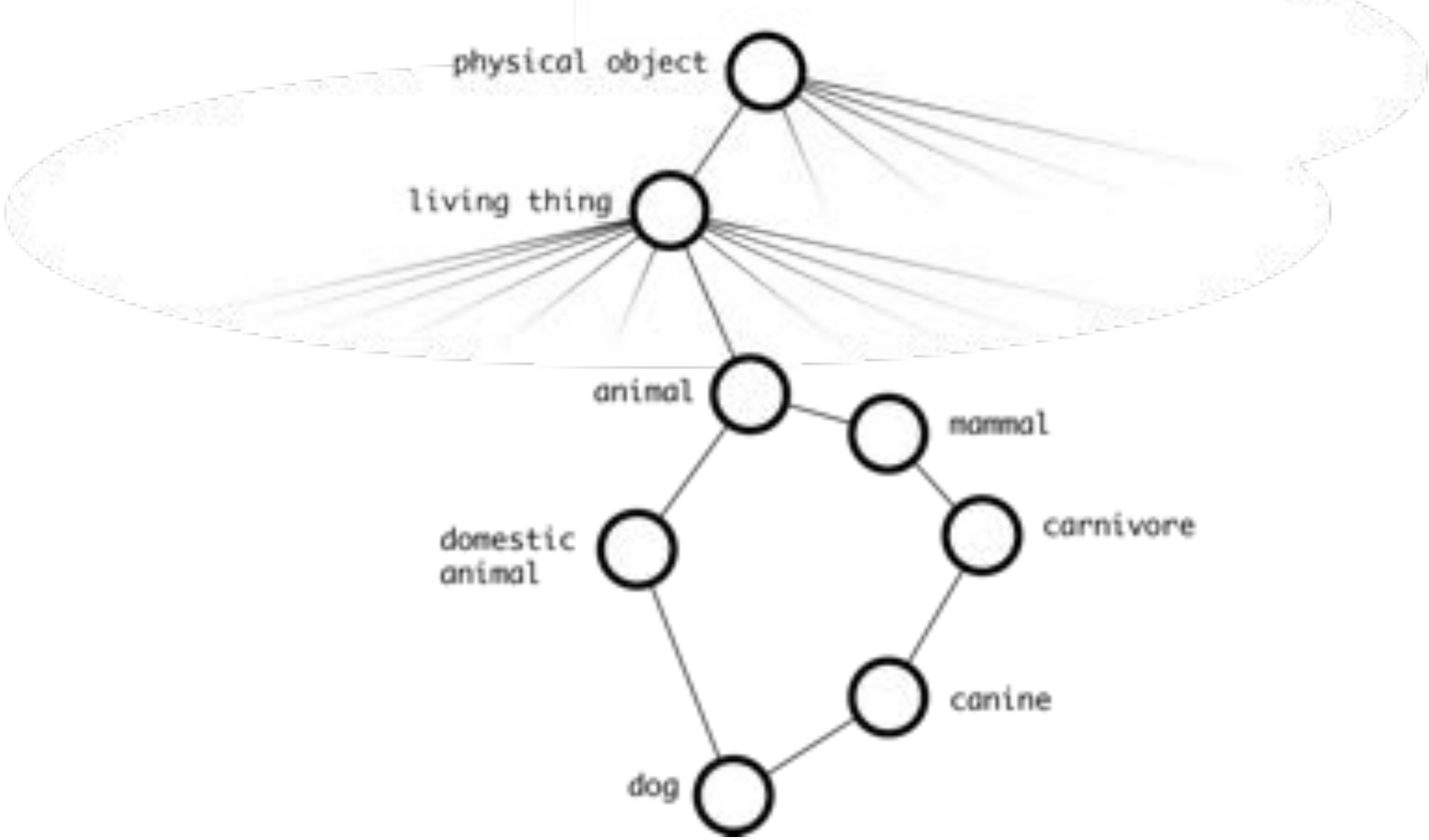
YOLOv2, YOLO9000

Can’t just mash classes together...

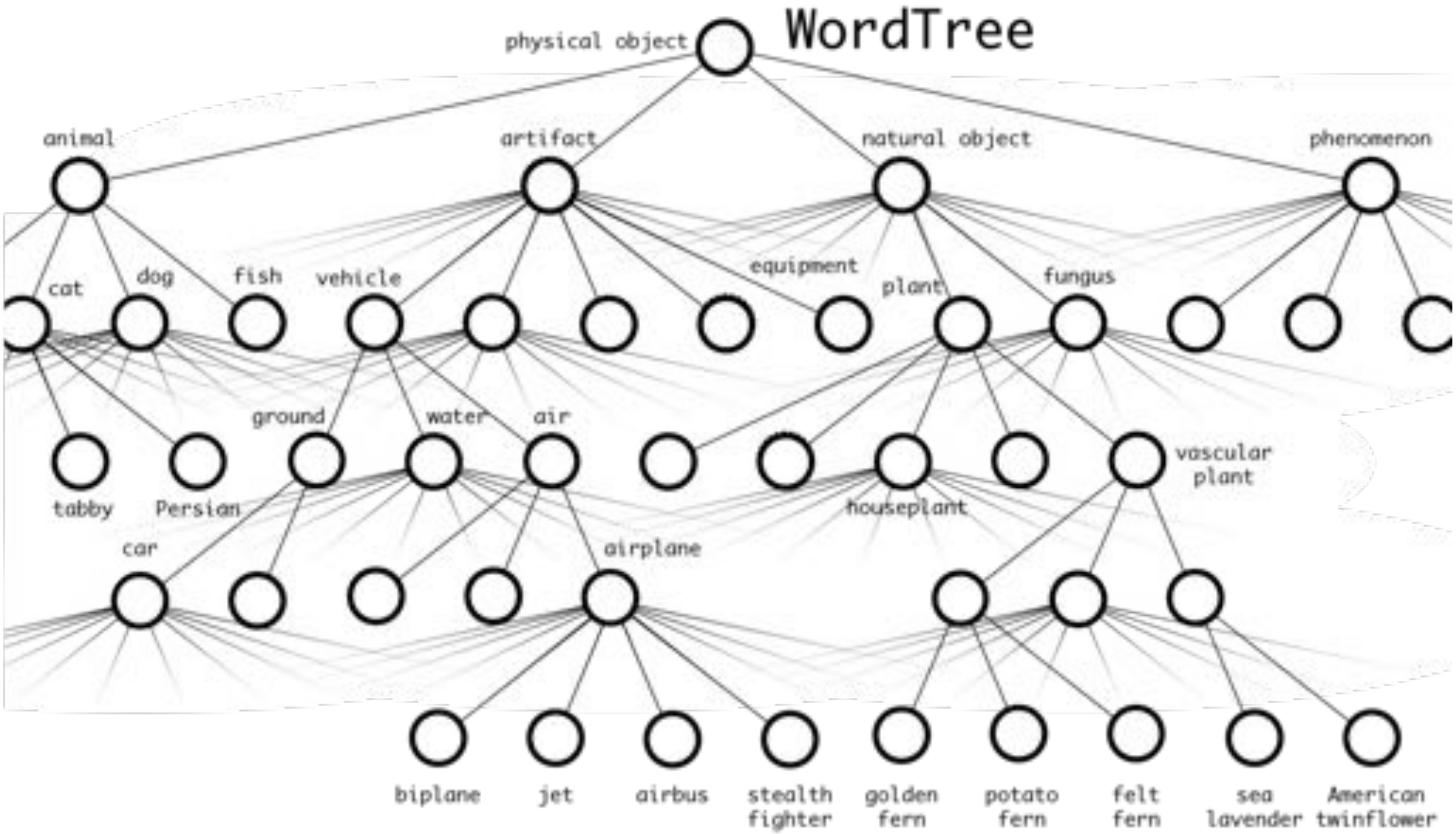


YOLOv2, YOLO9000

WordNet has structure but it's messy

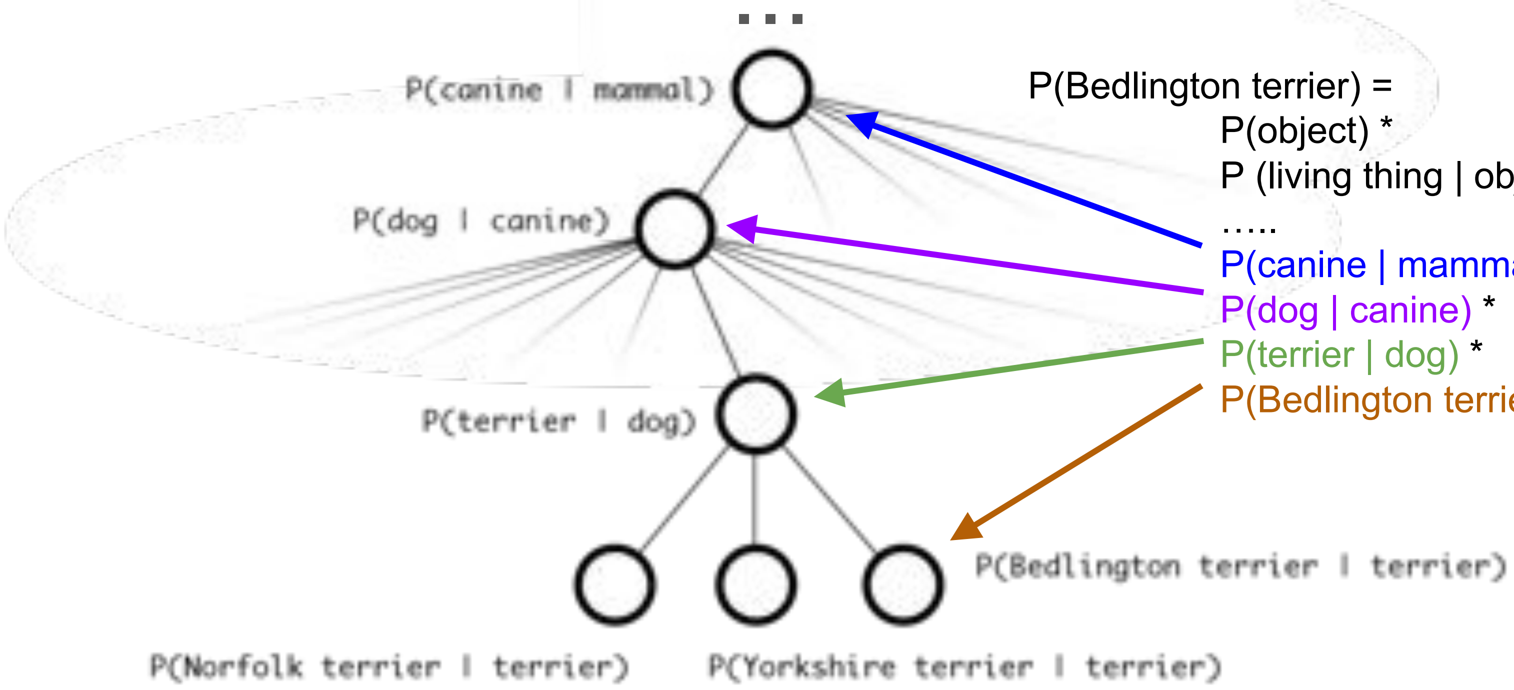


YOLOv2, YOLO9000



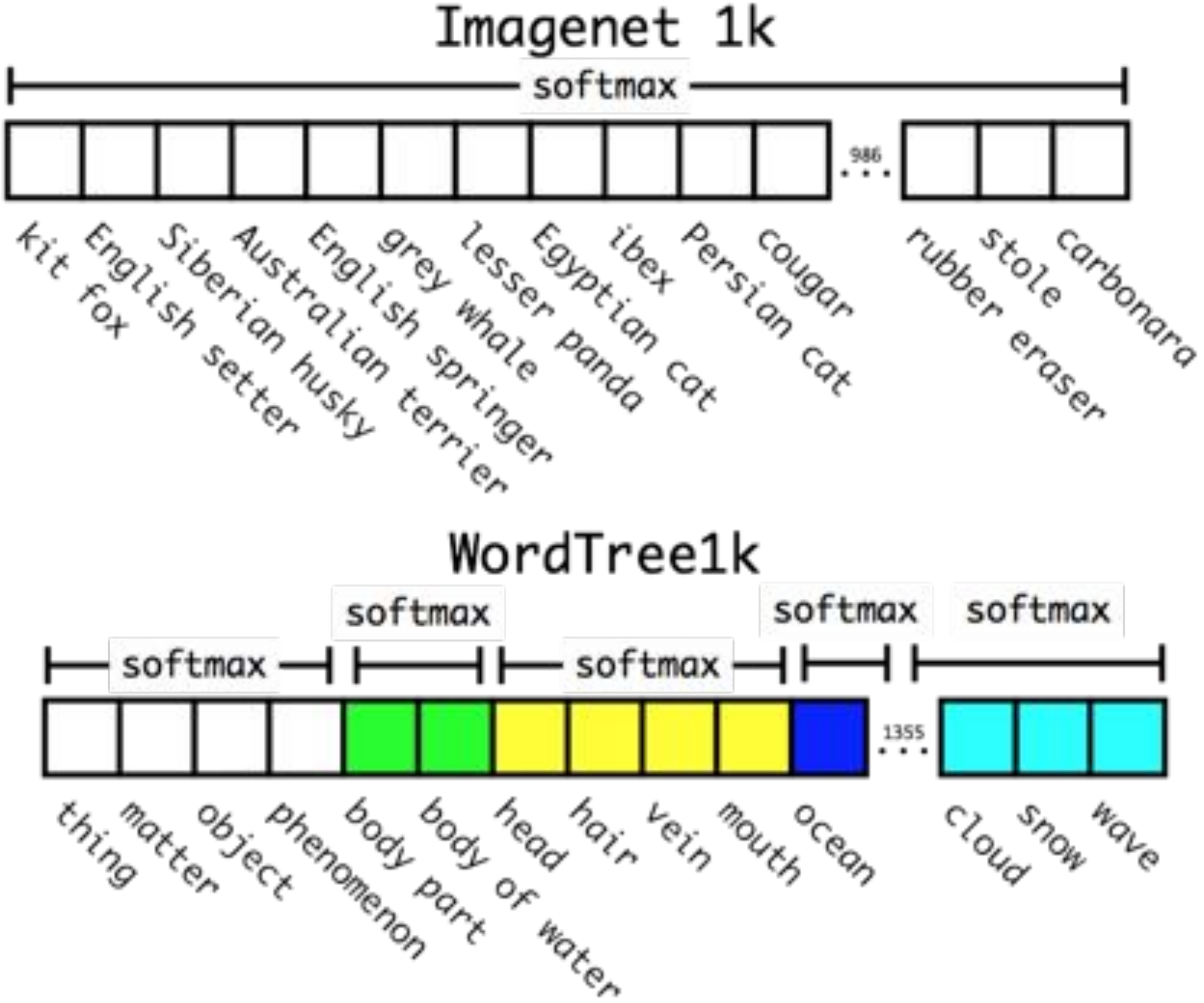
YOLOv2, YOLO9000

Each node is a conditional probability

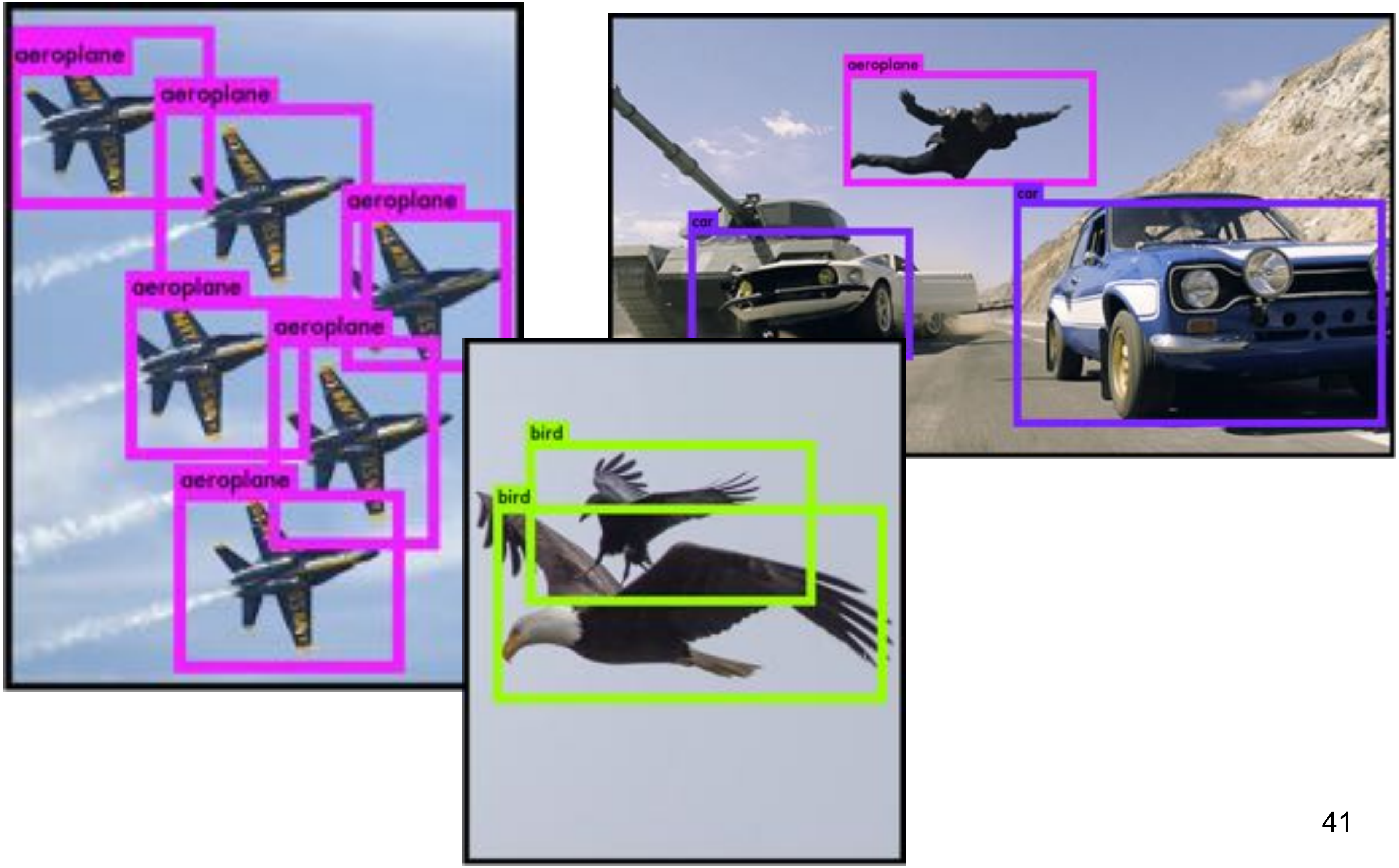

...
P(canine | mammal)
P(dog | canine)
P(terrier | dog)
P(Bedlington terrier | terrier)
P(Norfolk terrier | terrier)
P(Yorkshire terrier | terrier)

P(Bedlington terrier) =
P(object) *
P (living thing | object) *
…..
P(canine | mammal) *
P(dog | canine) *
P(terrier | dog) *
P(Bedlington terrier | terrier)

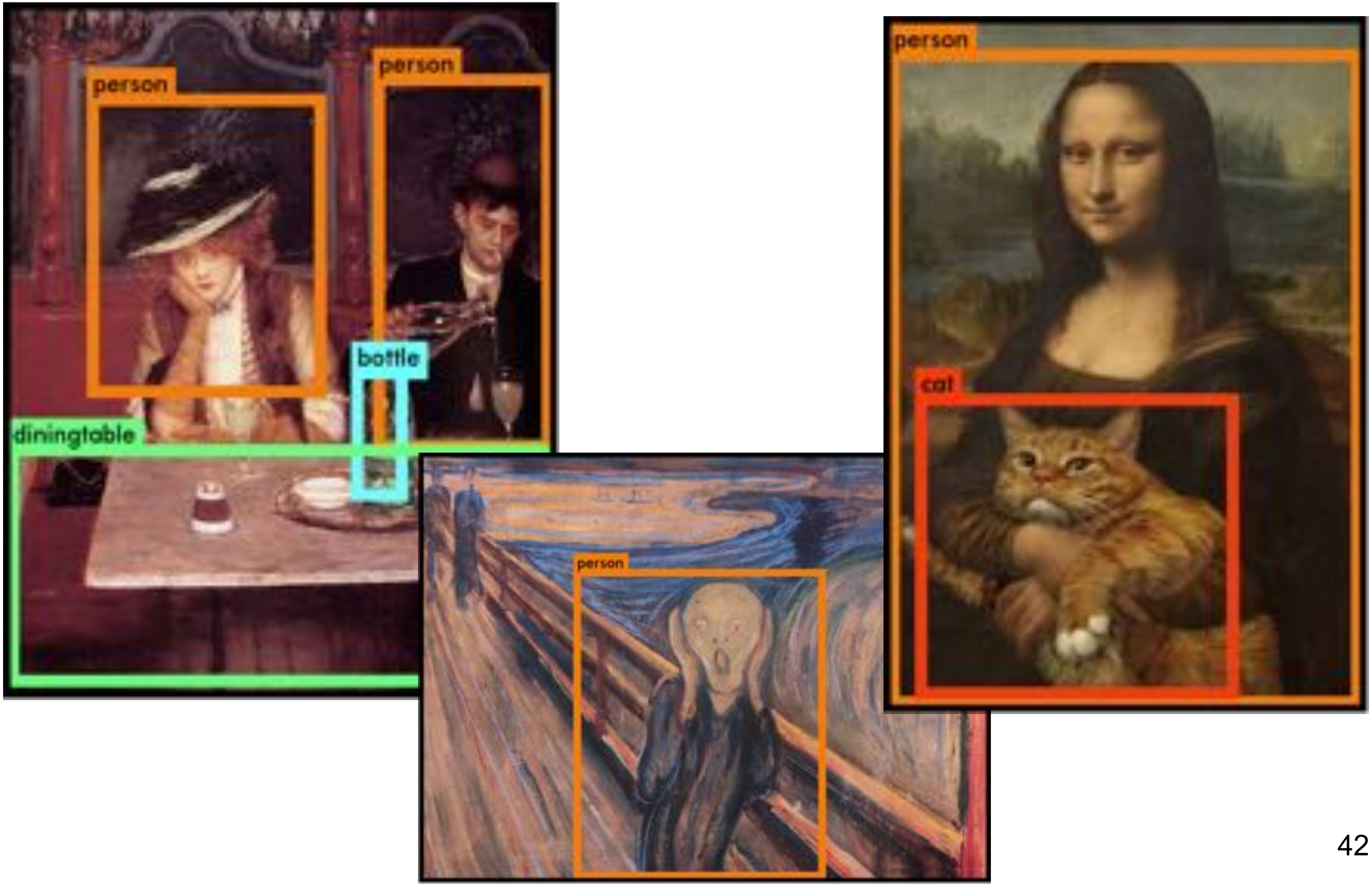
YOLOv2, YOLO9000



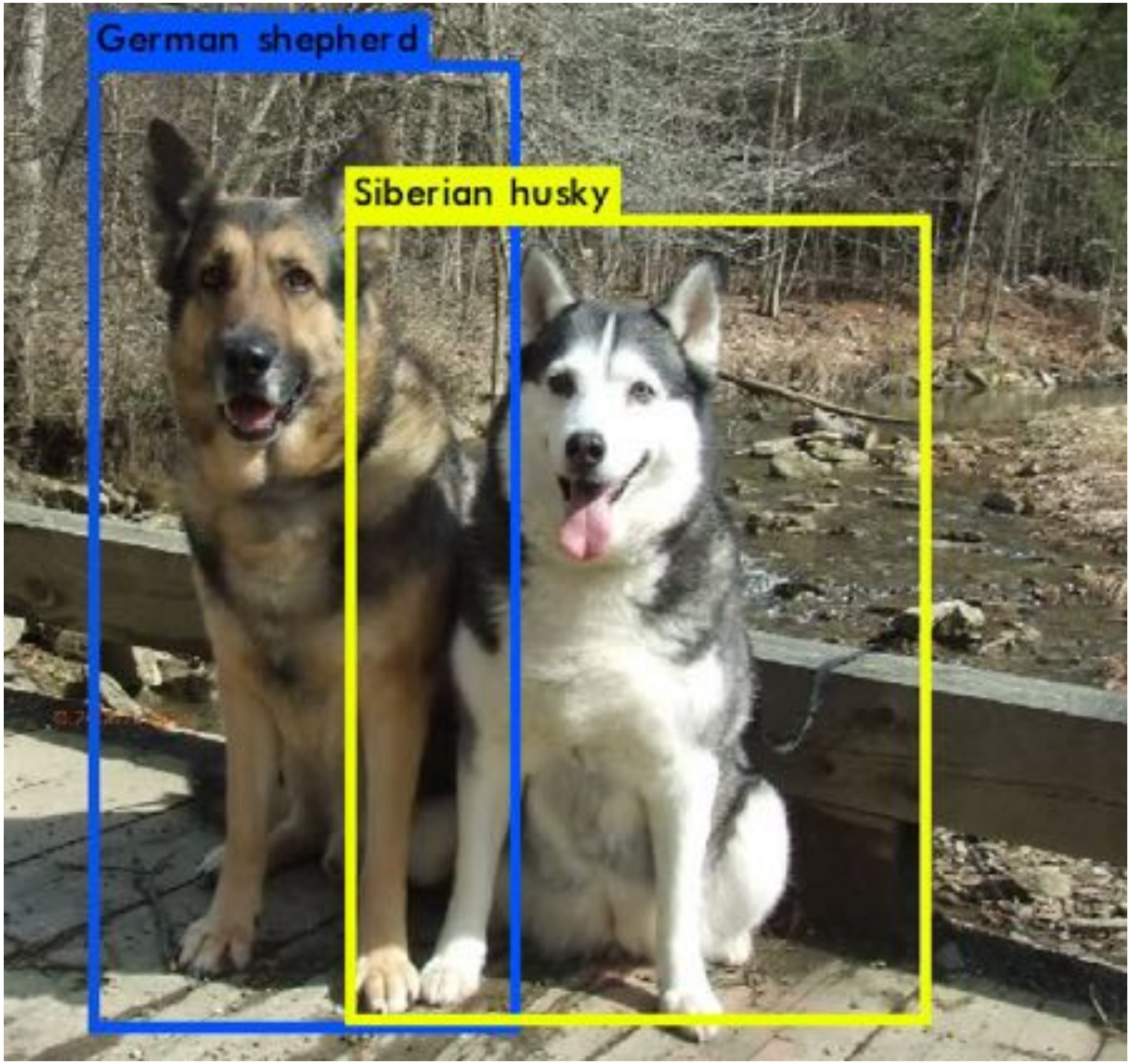
Experiments -- YOLO works across a variety of natural images

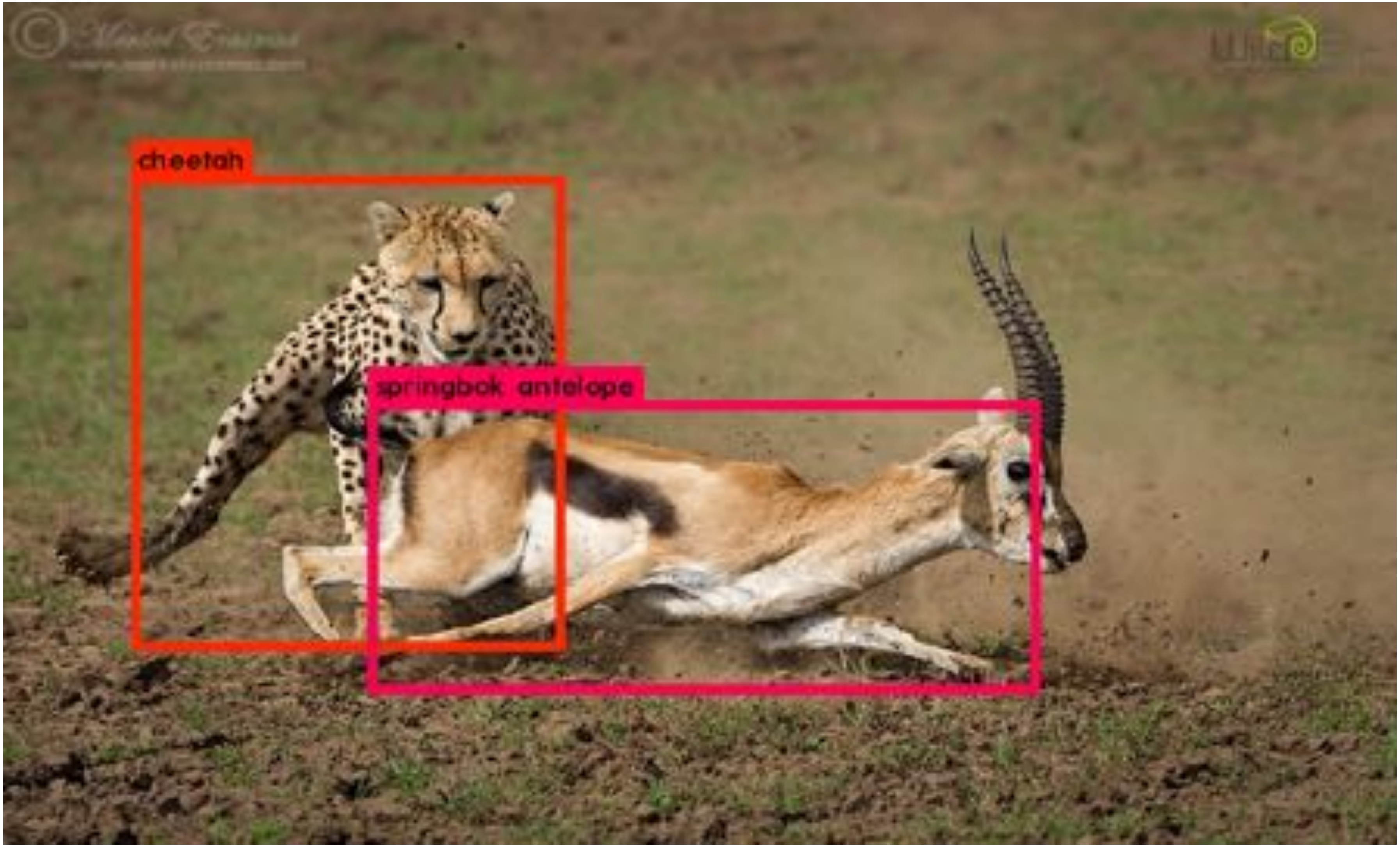


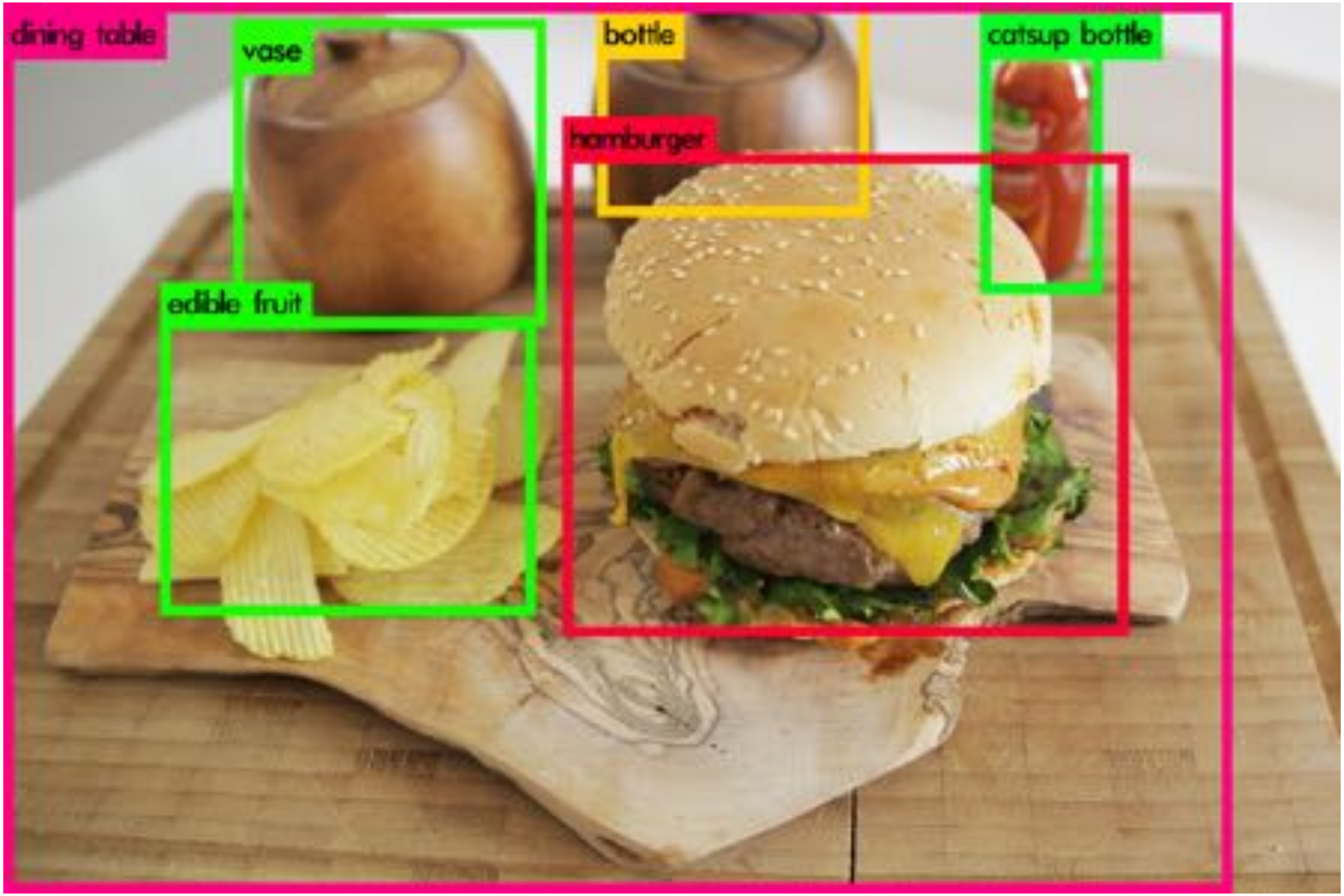
It also generalizes well to new domains (like art)

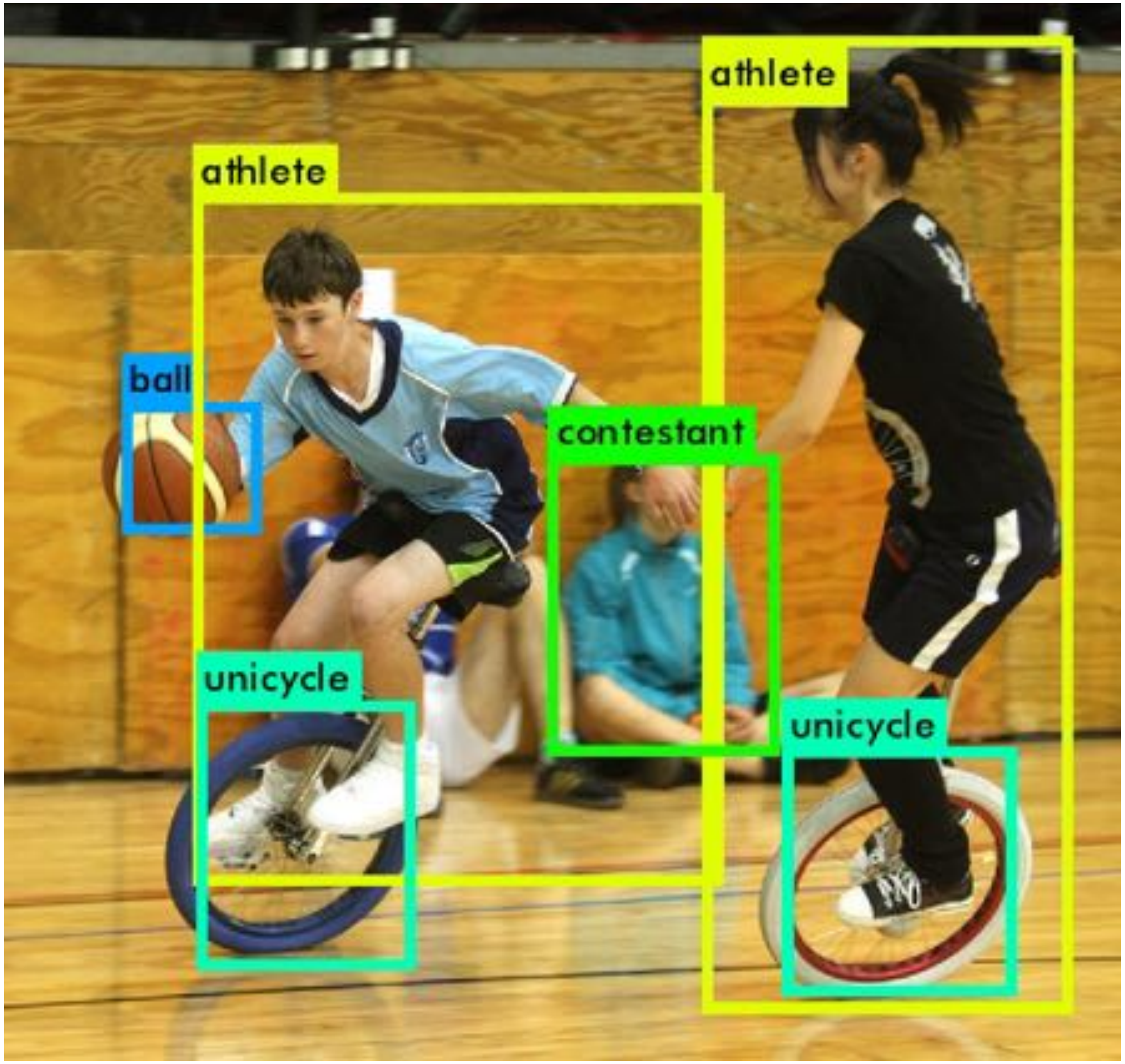


YOLOv2, YOLO9000

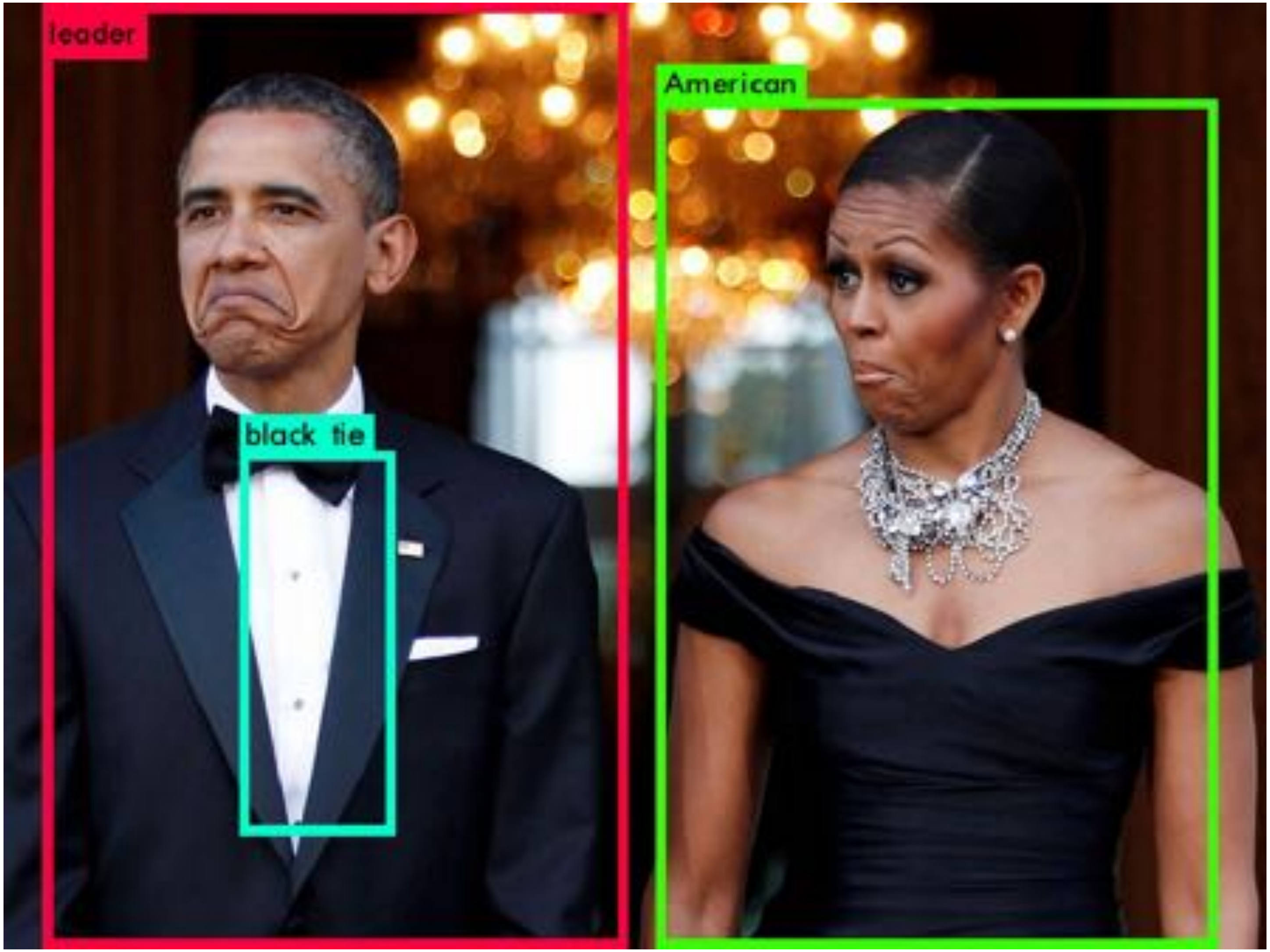


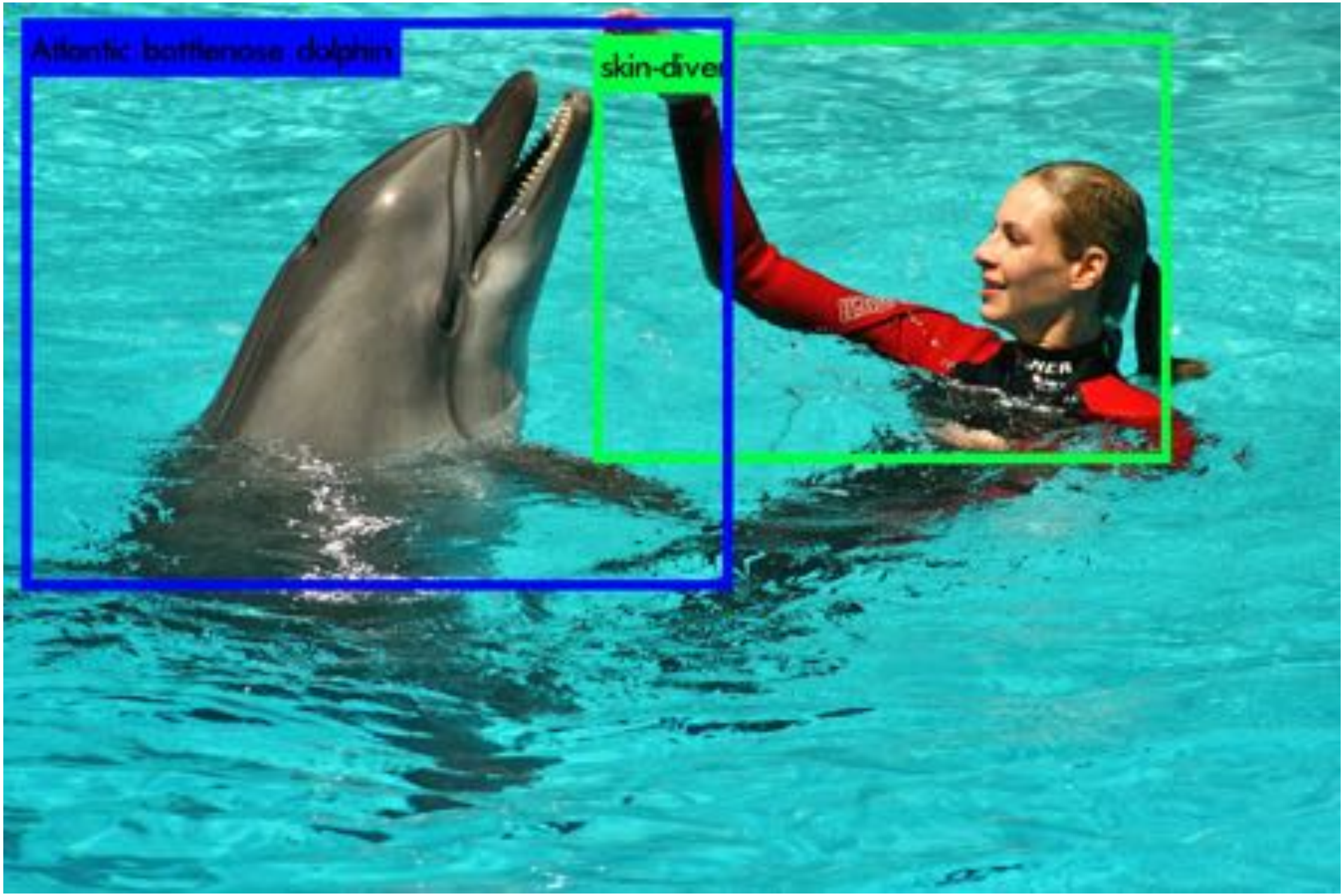

cheetah
springbok antelope

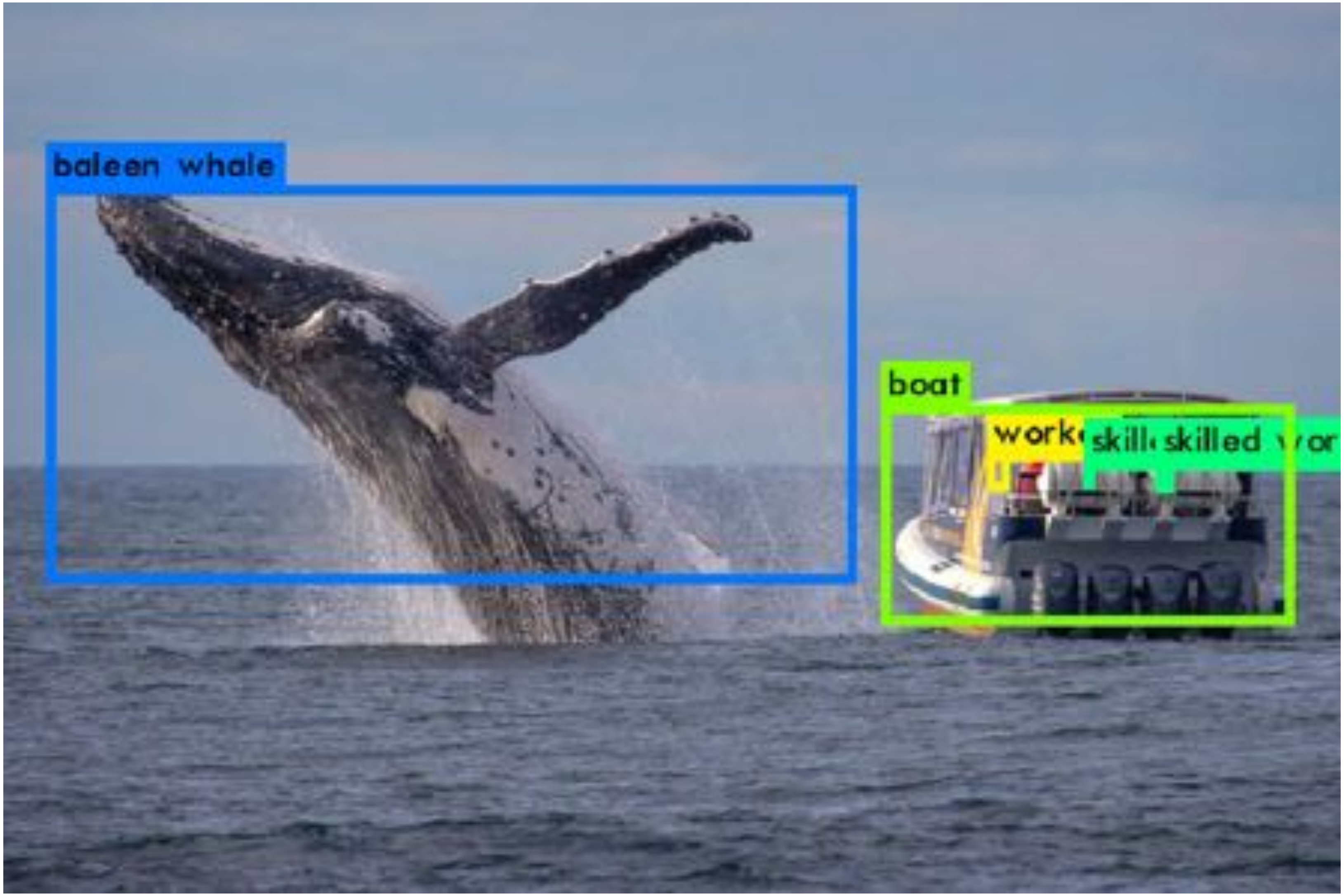

dining table
vase
bottle
catsup bottle
hamburger
edible fruit

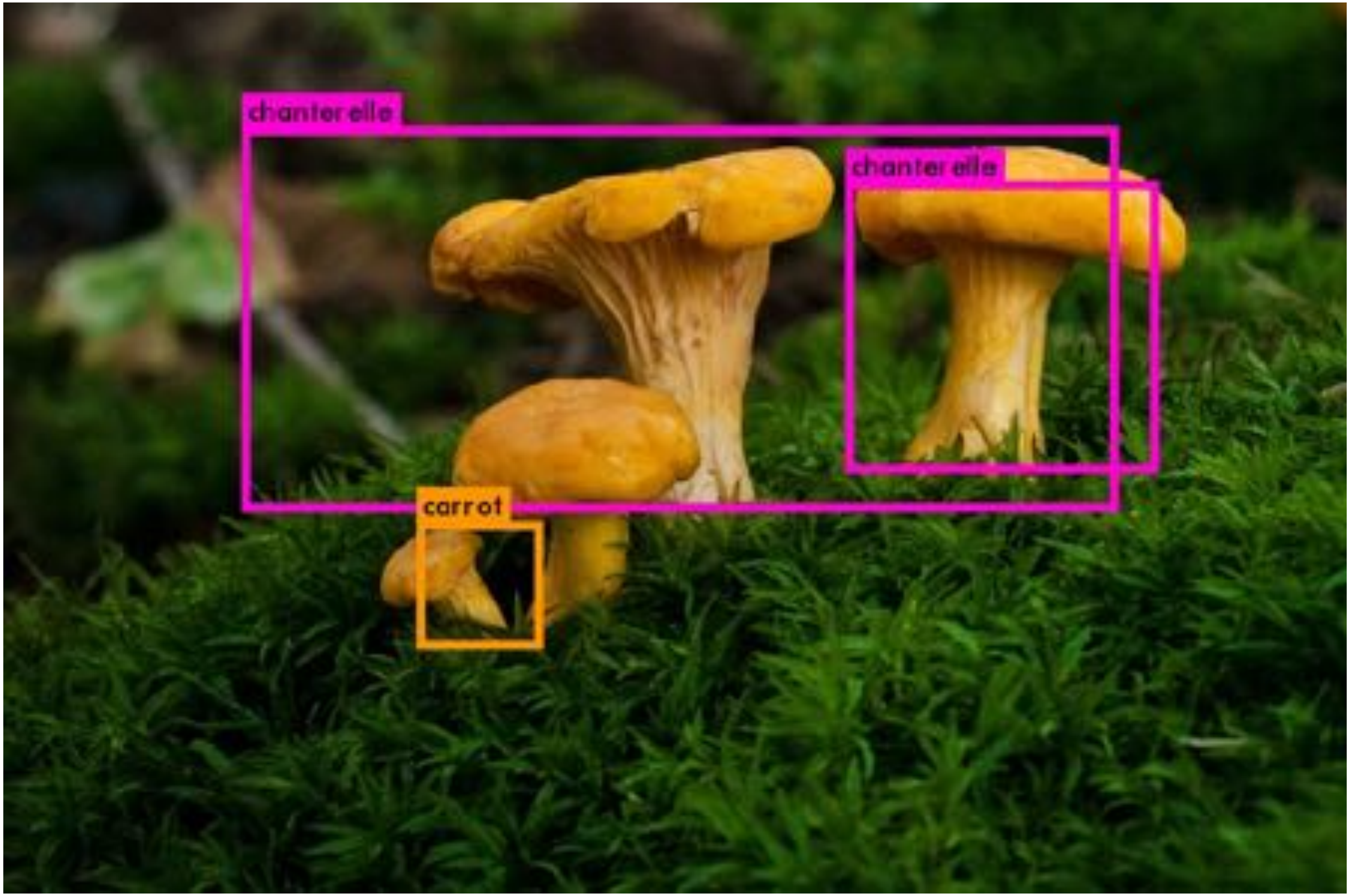


iMorpheus


leader
American
black tie

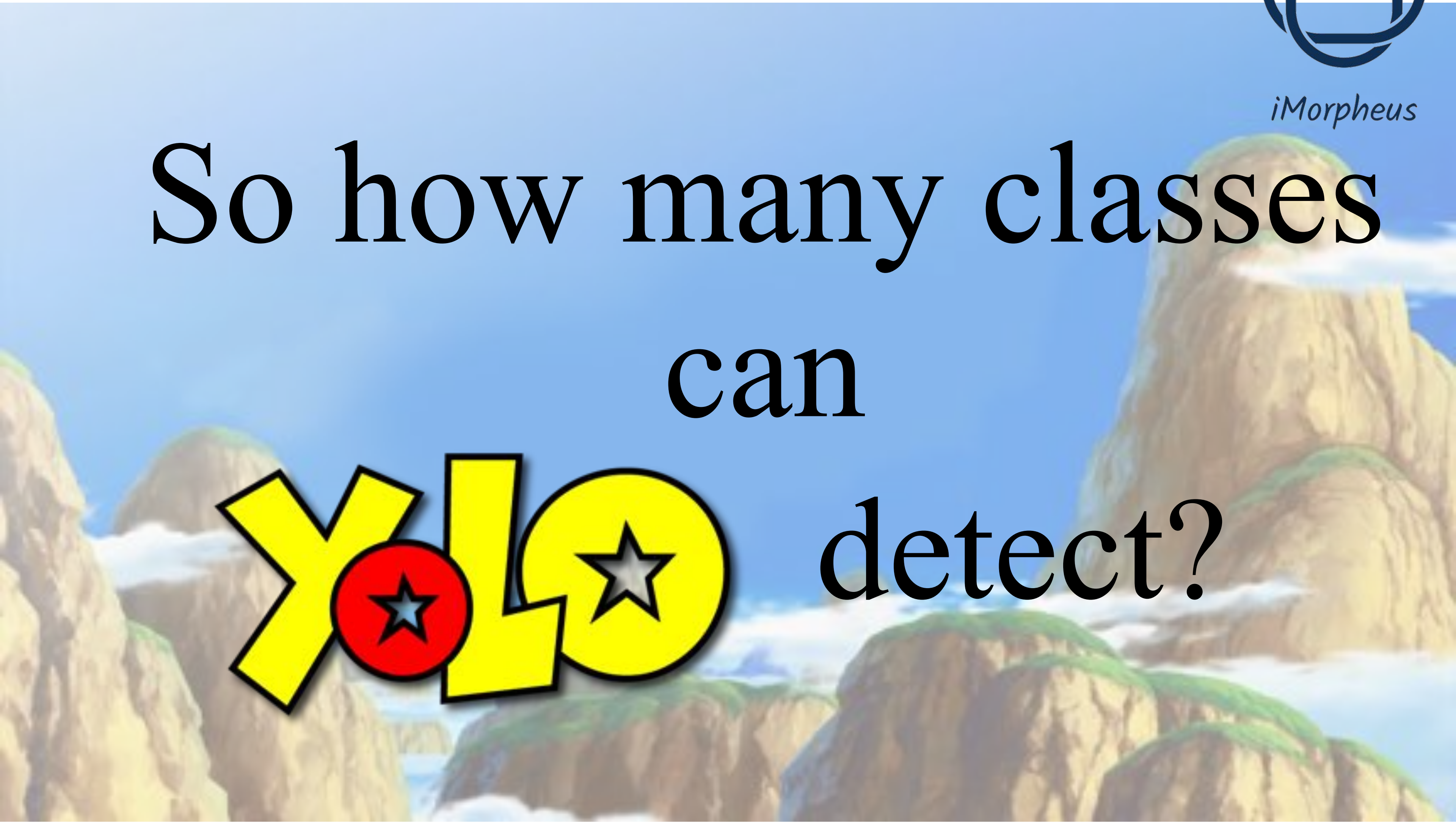

Atlantic bottlenose dolphin
skin-diver


baleen whale
boat
work
skill
skilled wor



chanterelle
chanterelle
carrot

iMorpheus



So how many classes can YOLO detect?

iMorpheus

iMorpheus

Code, models, and updates:

https://pjreddie.com/yolo/

XNOR.AI

iMorpheus

iMorpheus Journal Club (Friday 12:00PM GMT+8, Weekly)

每周五 下午12点 （北京时间）

微信 Wechat

iMorpheus website : www.imorpheus.ai

Email Address : live@imorpheus.ai