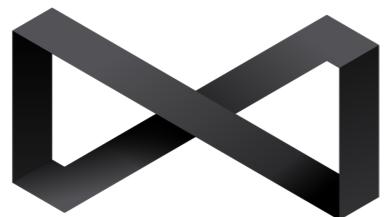
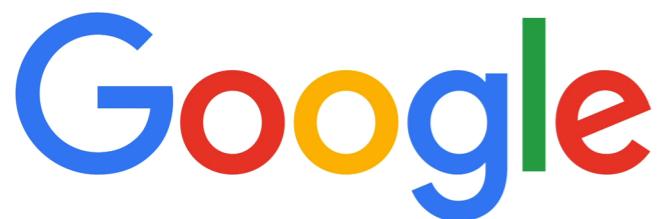


# SSD: Single Shot MultiBox Detector

Wei Liu(1), **Dragomir Anguelov(2)**, Dumitru Erhan(3), Christian Szegedy(3),  
Scott Reed(4), Cheng-Yang Fu(1), Alexander C. Berg(1)

UNC Chapel Hill(1), **Zoox Inc.(2)**, Google Inc.(3),  
University of Michigan(4)



THE UNIVERSITY  
*of* NORTH CAROLINA  
at CHAPEL HILL

FPS: 0.00

WR

1.53

OR

1.53

person: 0.83



Rio 2016

9

person: 0



person: (person: 0.37)



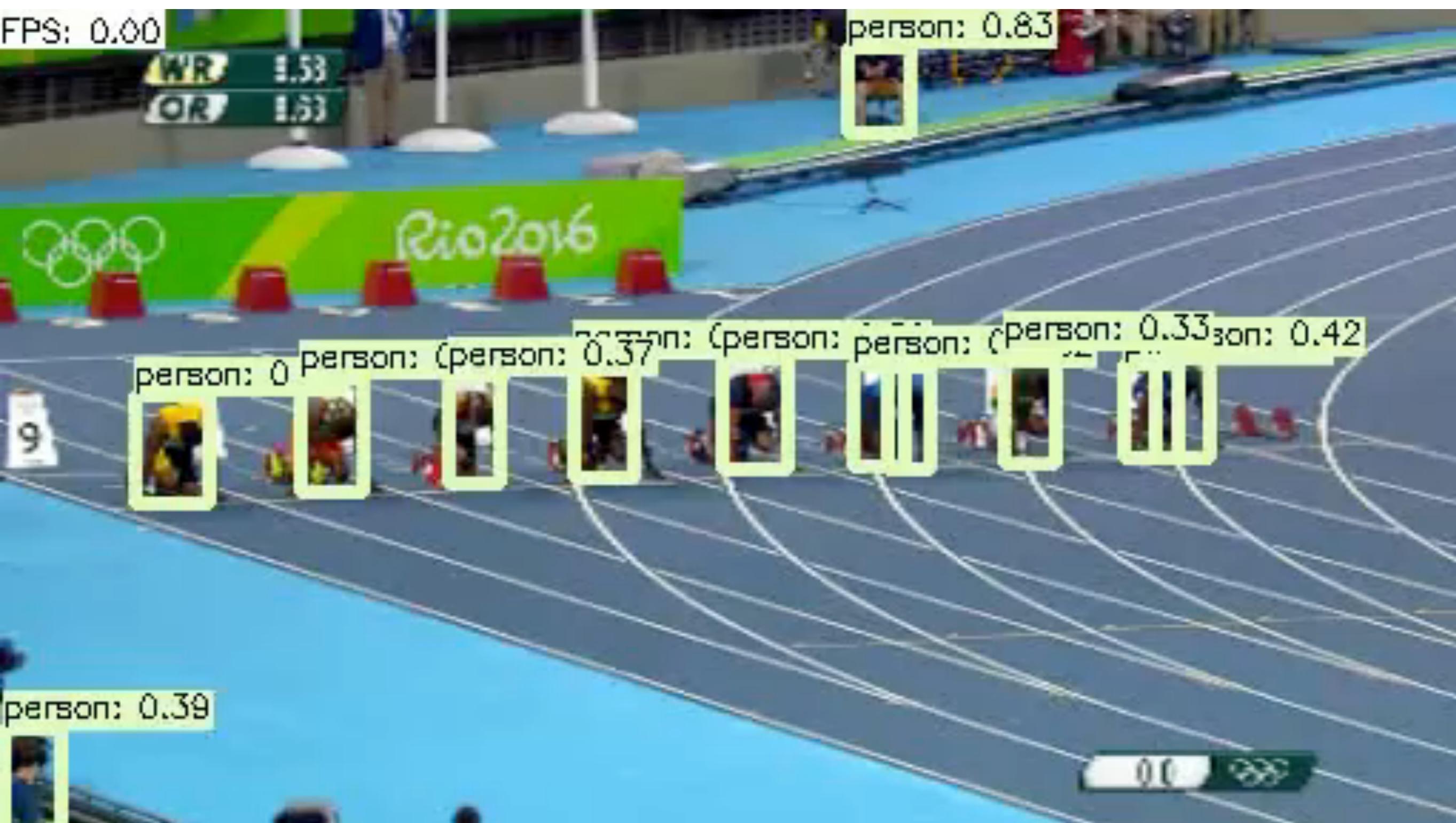
person: 0.39



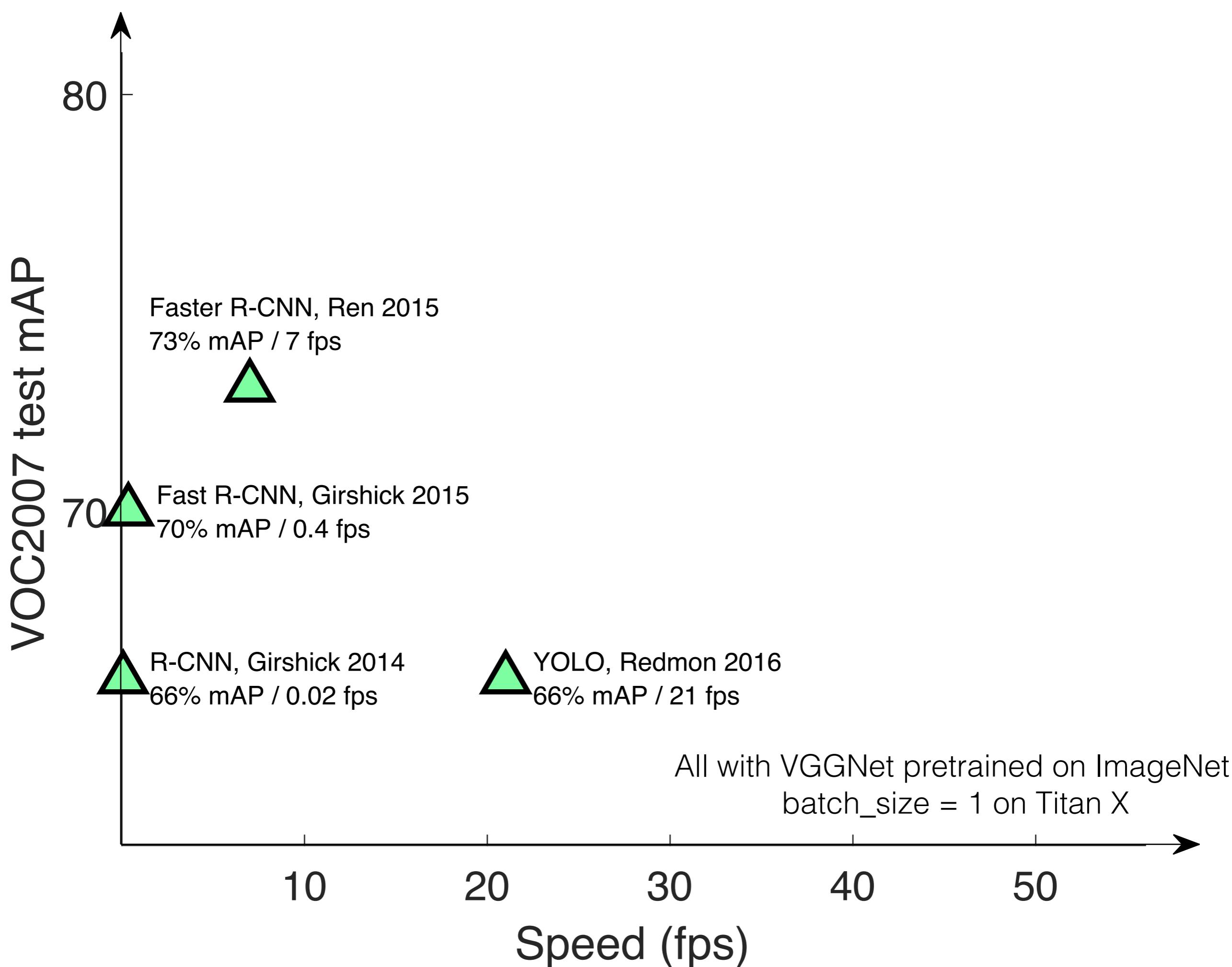
00

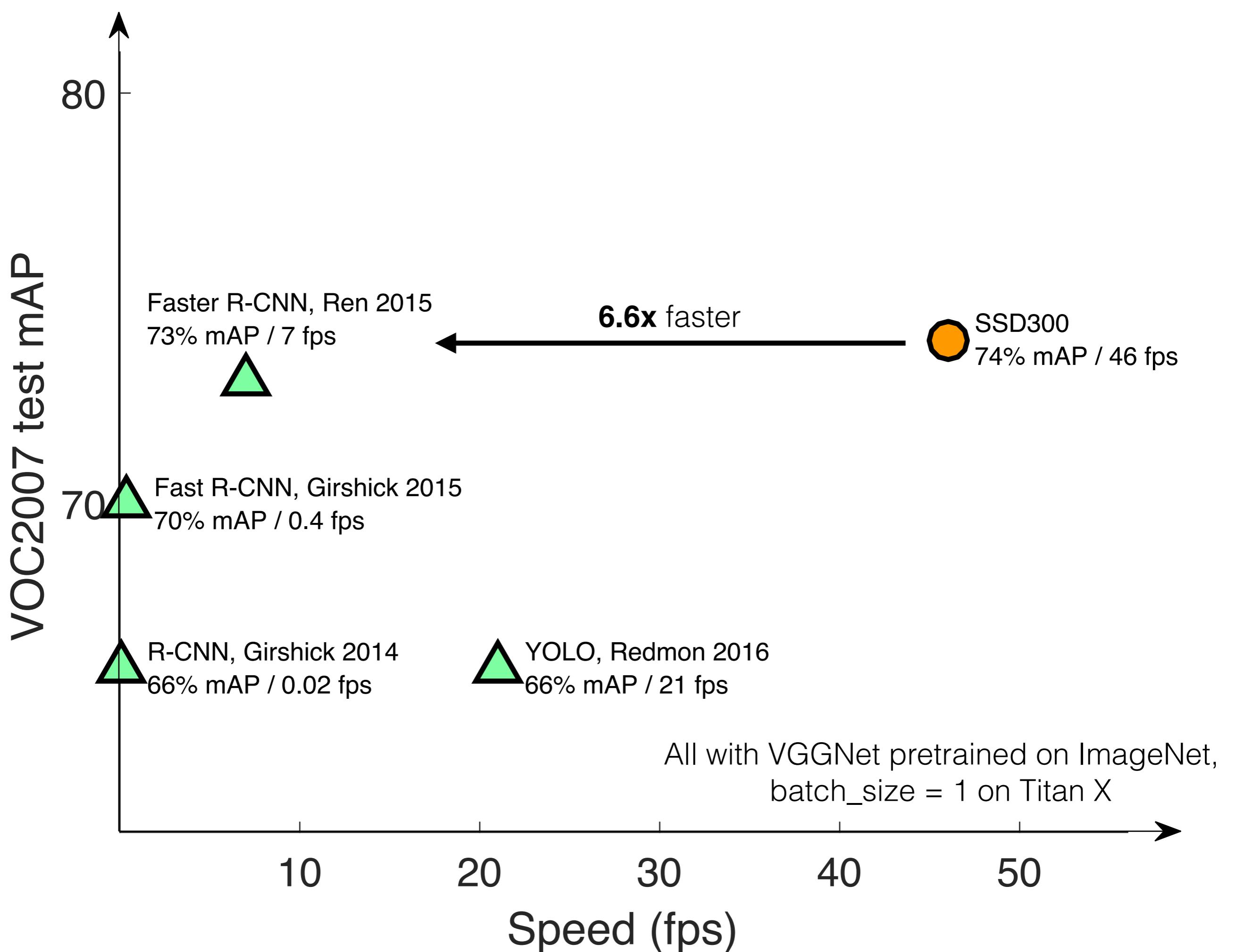


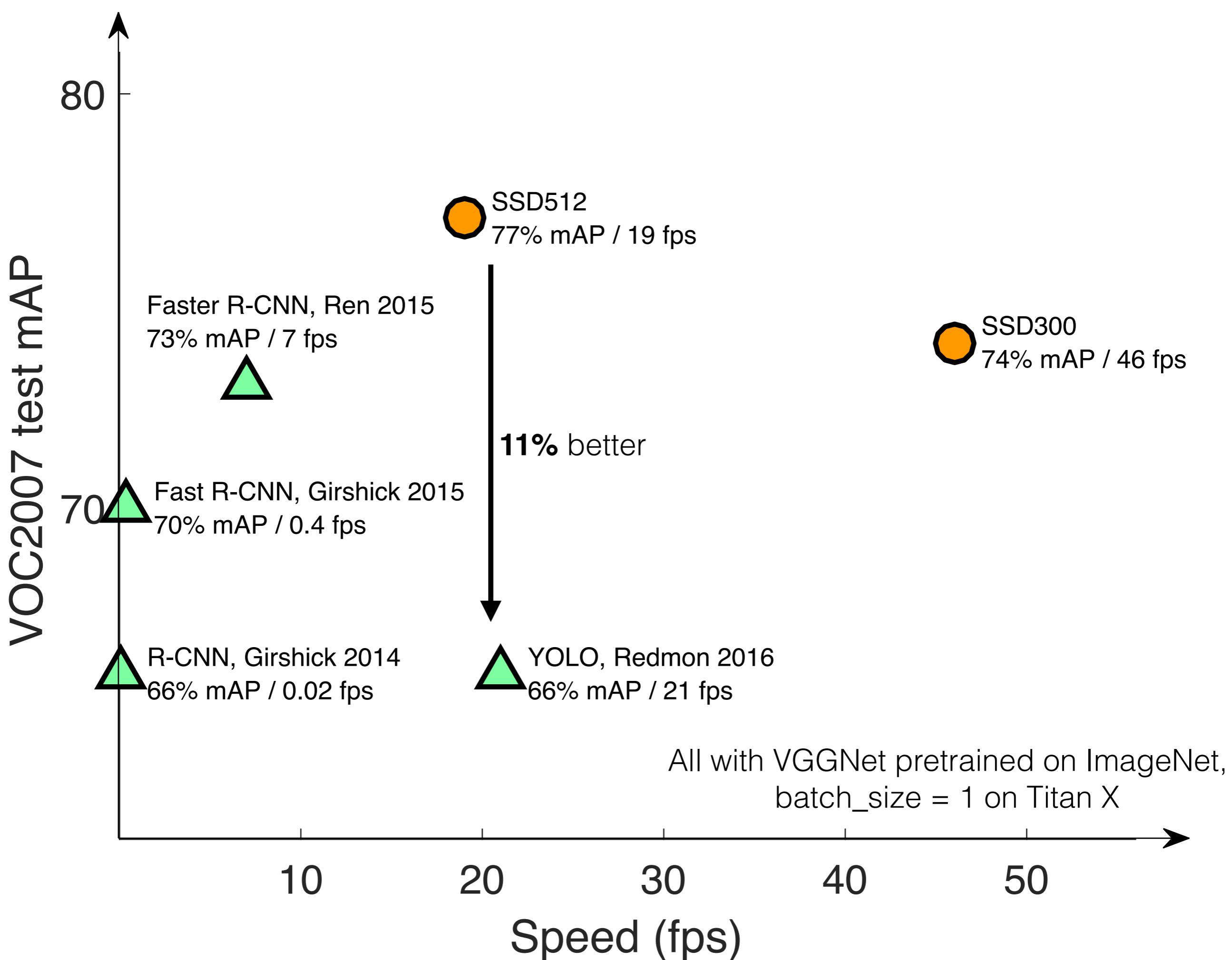
VGGNet  
Titan X Pascal

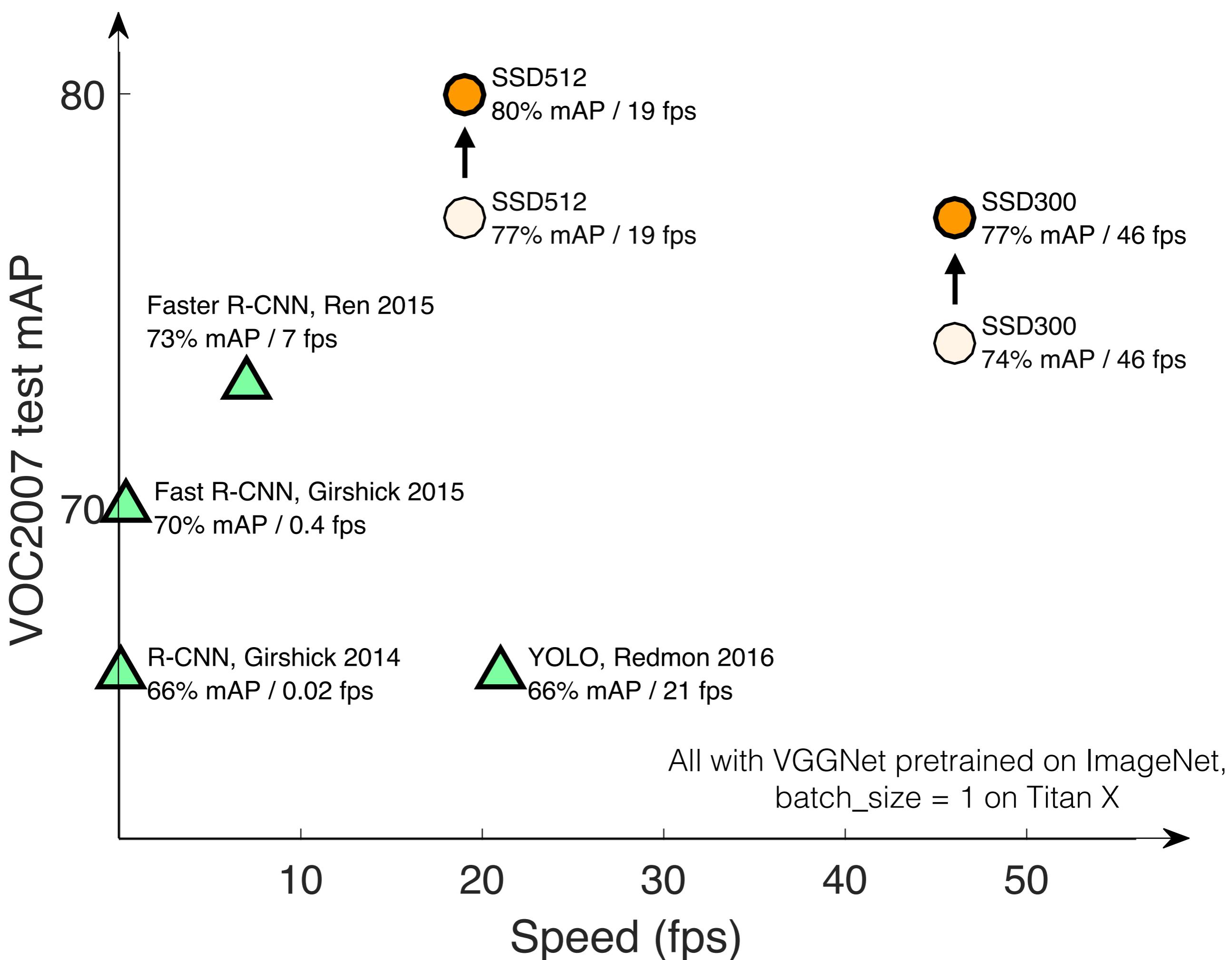


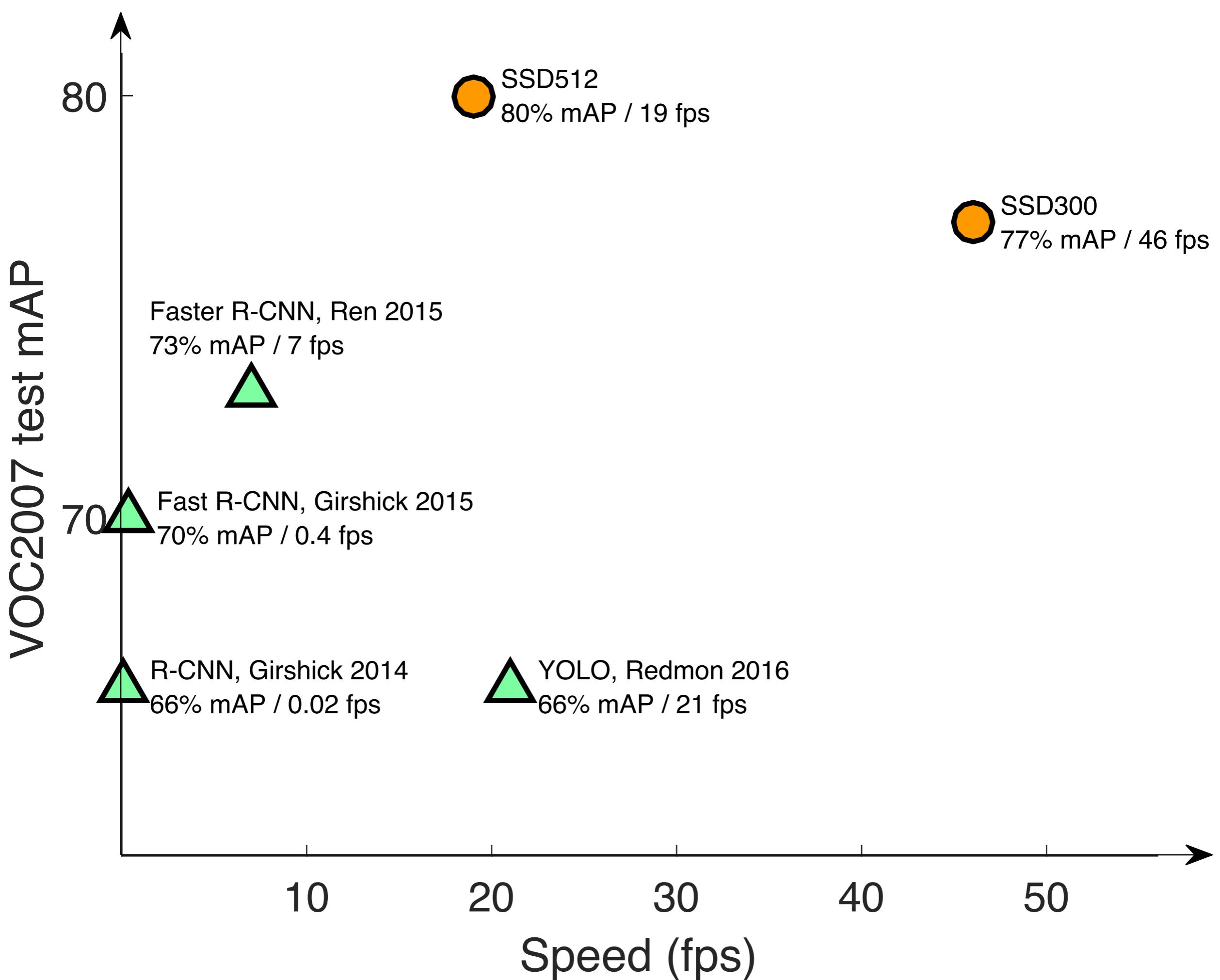
VGGNet  
Titan X Pascal

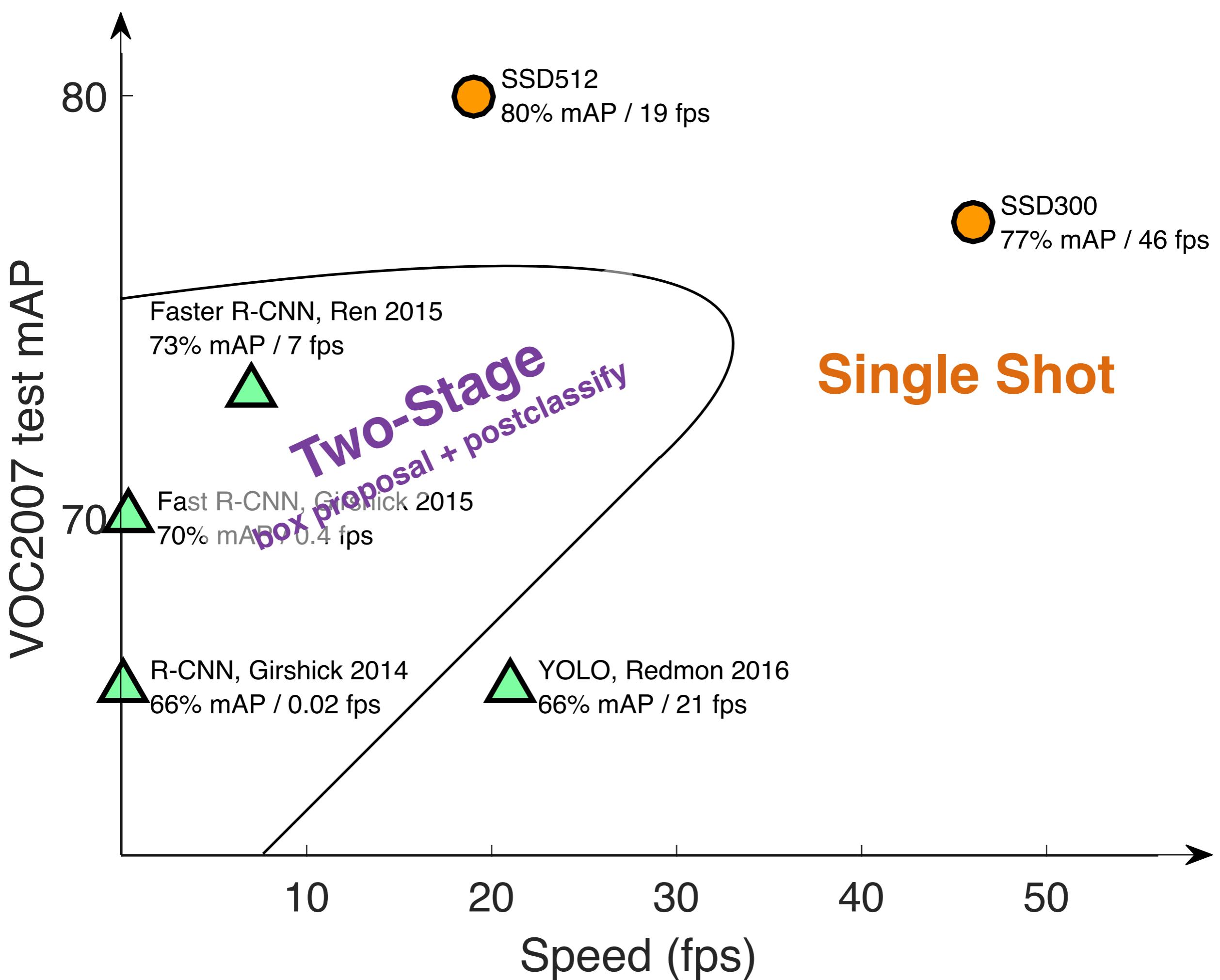






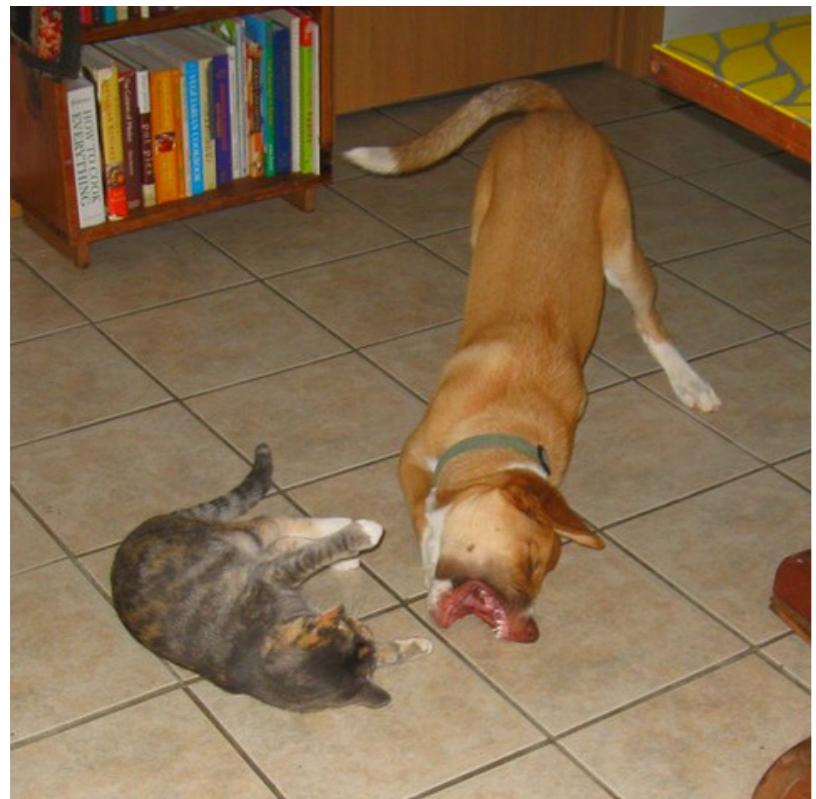






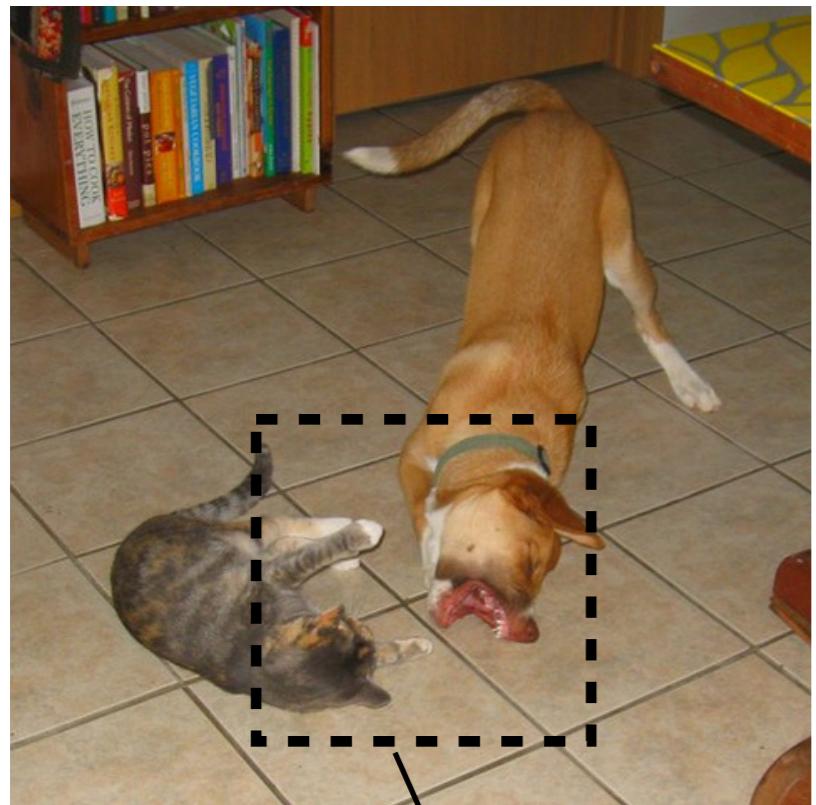
# Bounding Box Prediction

Classical sliding  
windows



# Bounding Box Prediction

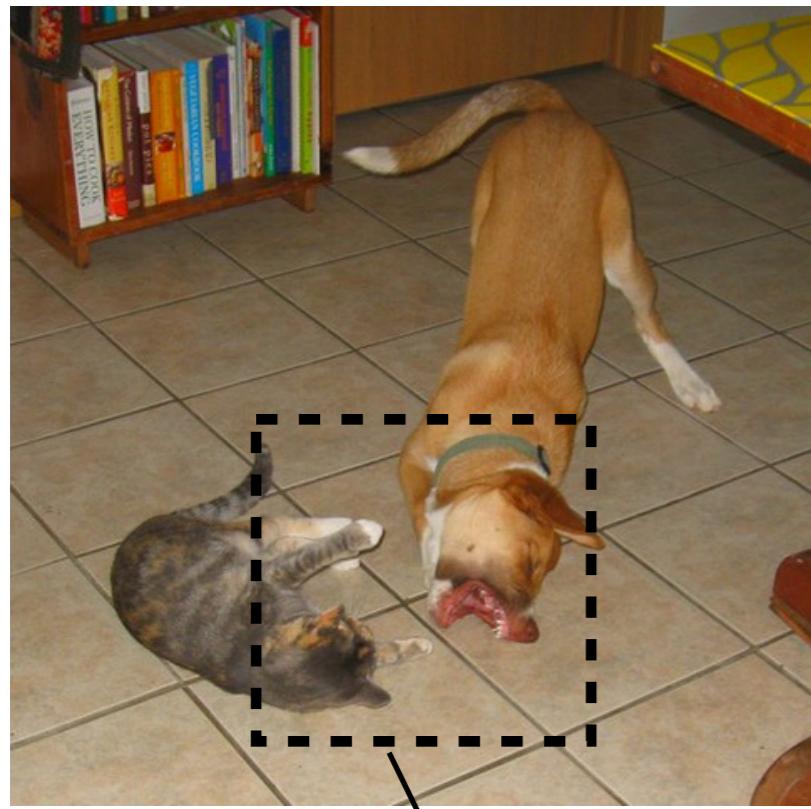
Classical sliding  
windows



Is it a cat? **No**

# Bounding Box Prediction

Classical sliding  
windows

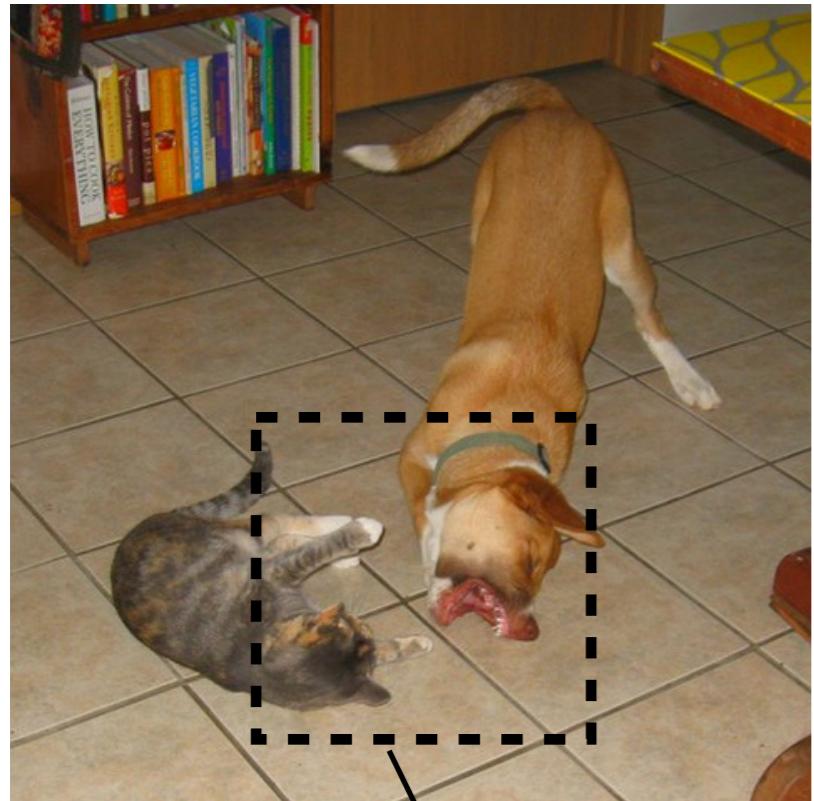


Is it a cat? **No**

Discretize the box space **densely**

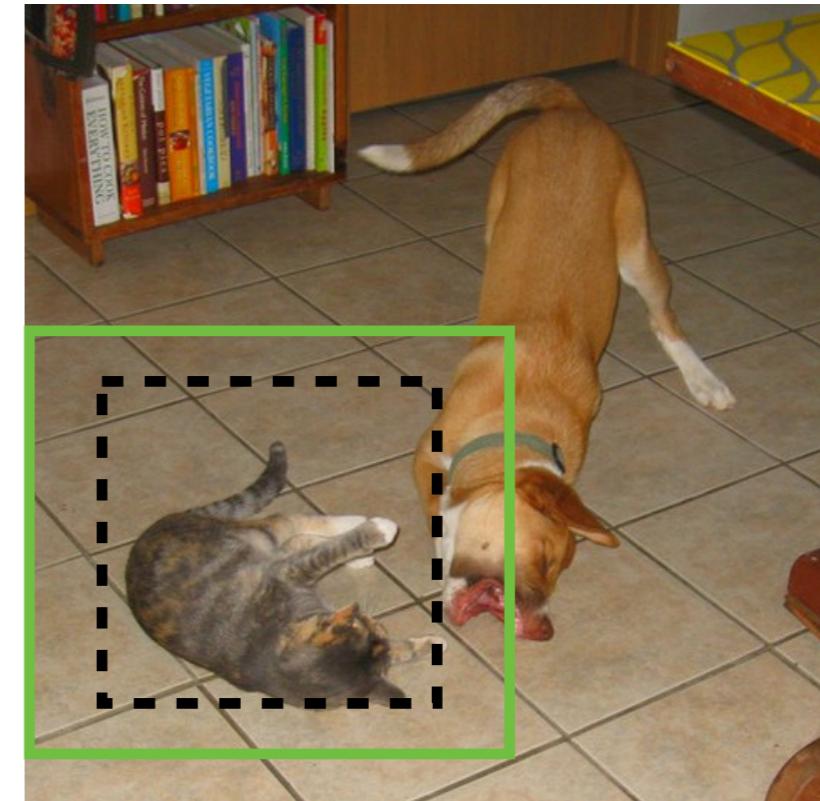
# Bounding Box Prediction

Classical sliding  
windows



Is it a cat? **No**

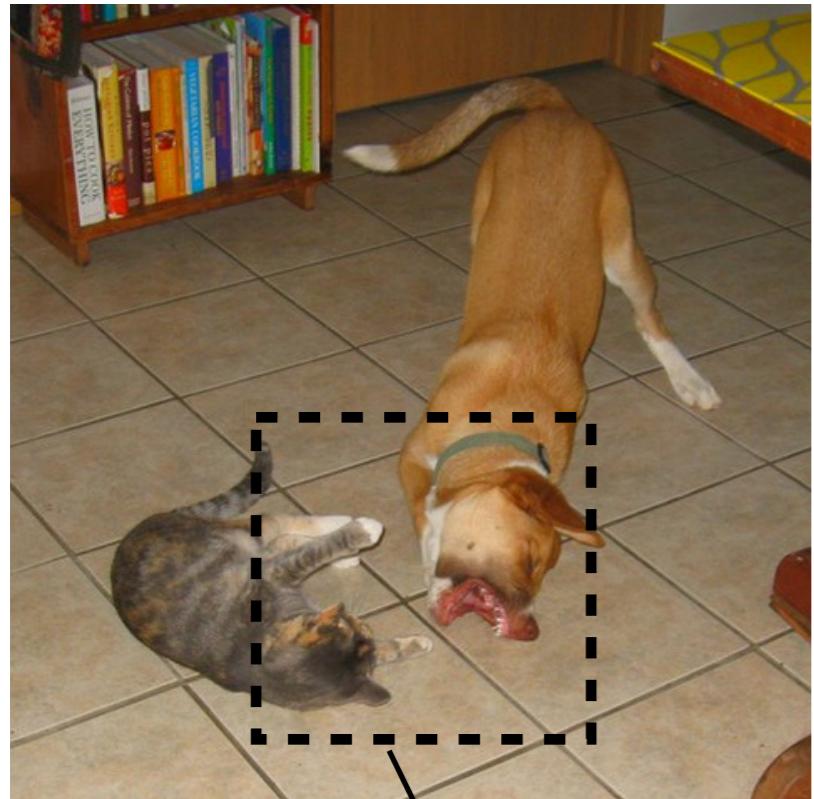
SSD and other deep  
approaches



Discretize the box space **densely**

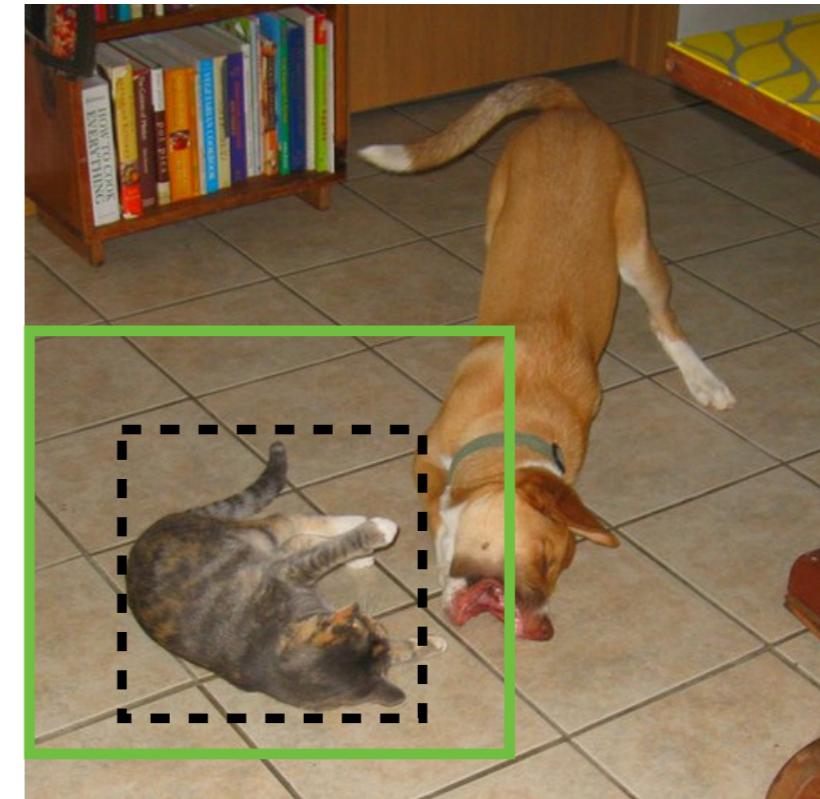
# Bounding Box Prediction

Classical sliding  
windows



Is it a cat? **No**

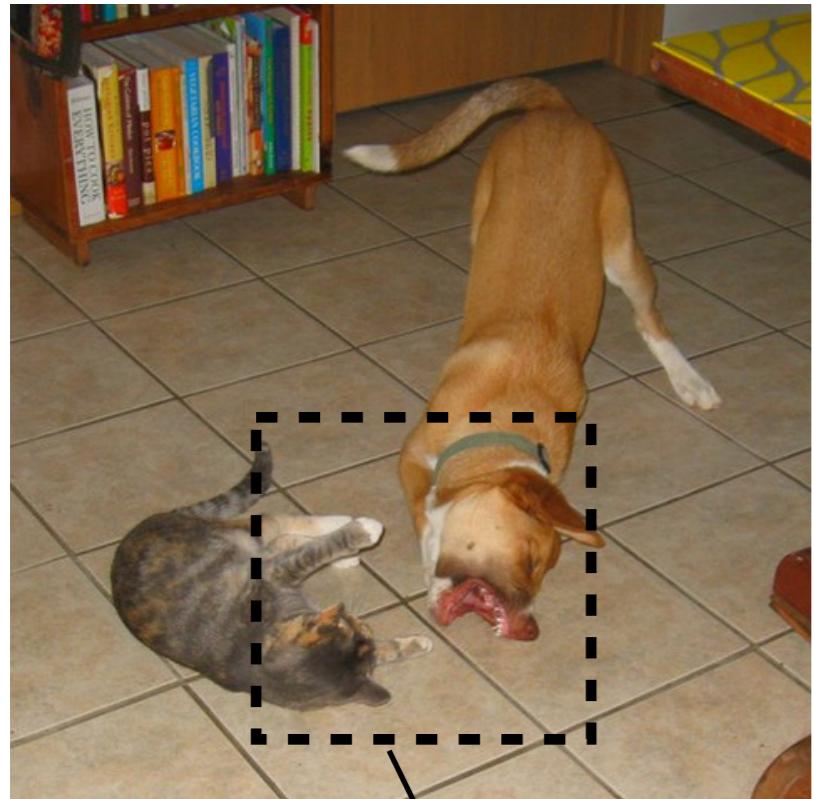
SSD and other deep  
approaches



Discretize the box space **densely**

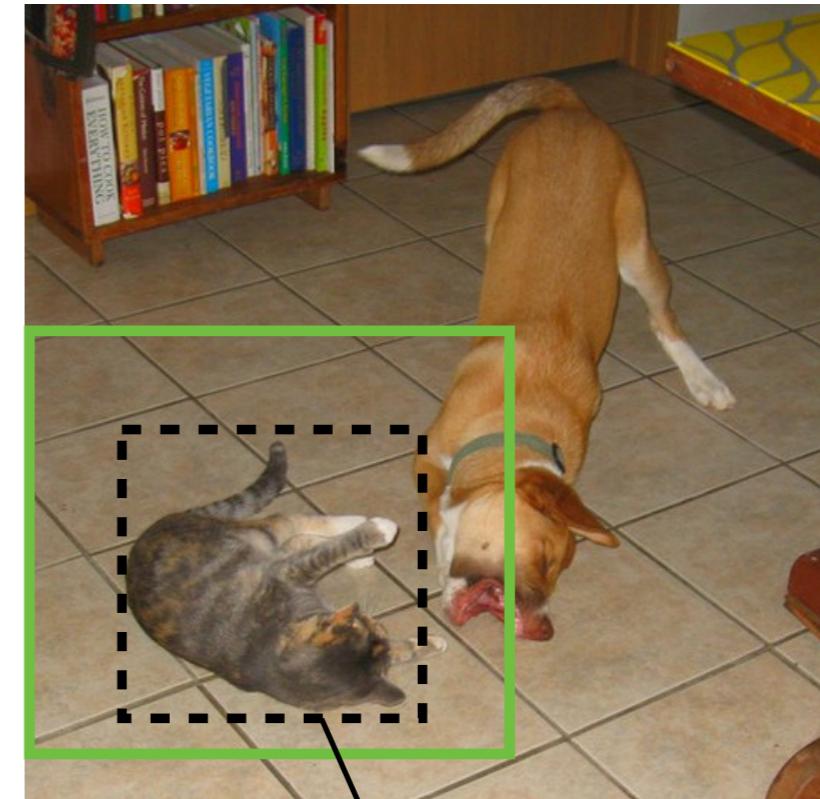
# Bounding Box Prediction

Classical sliding windows



Is it a cat? **No**

SSD and other deep approaches

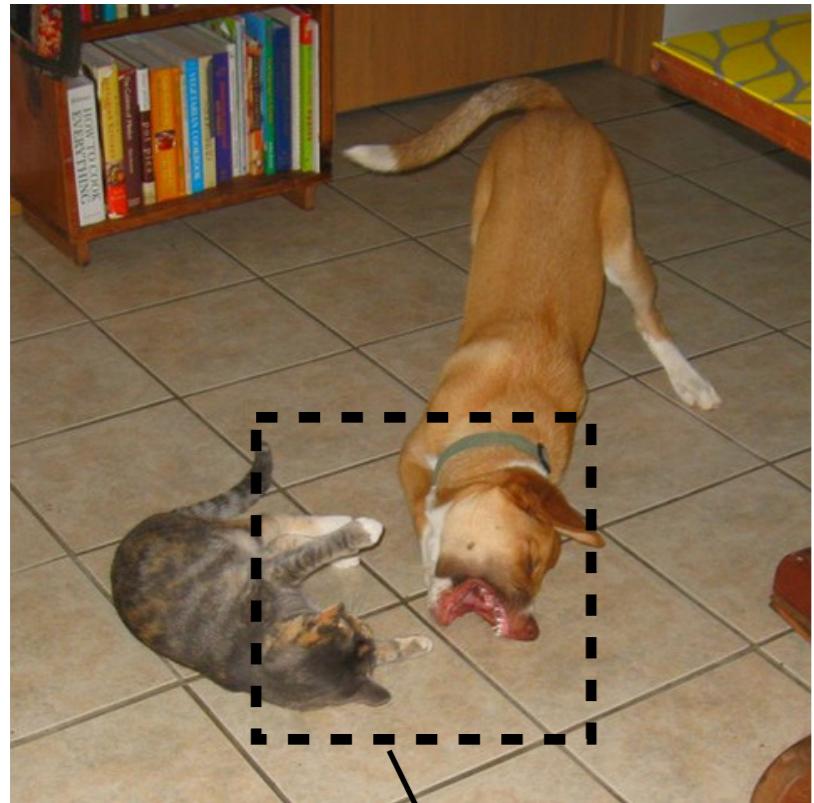


cat: 0.8 dog: 0.1

Discretize the box space **densely**

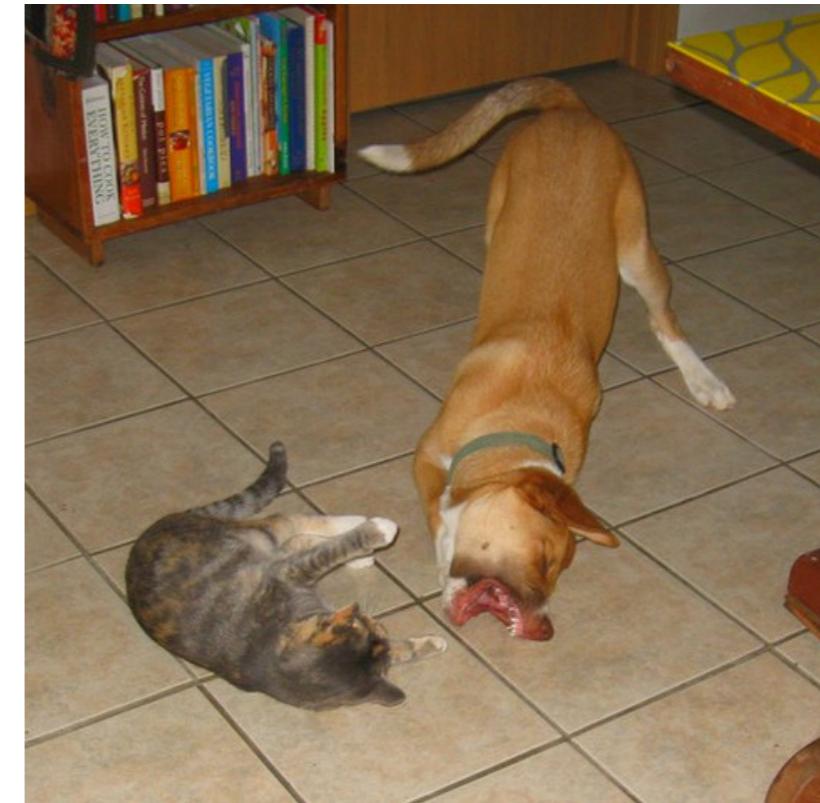
# Bounding Box Prediction

Classical sliding  
windows



Is it a cat? **No**

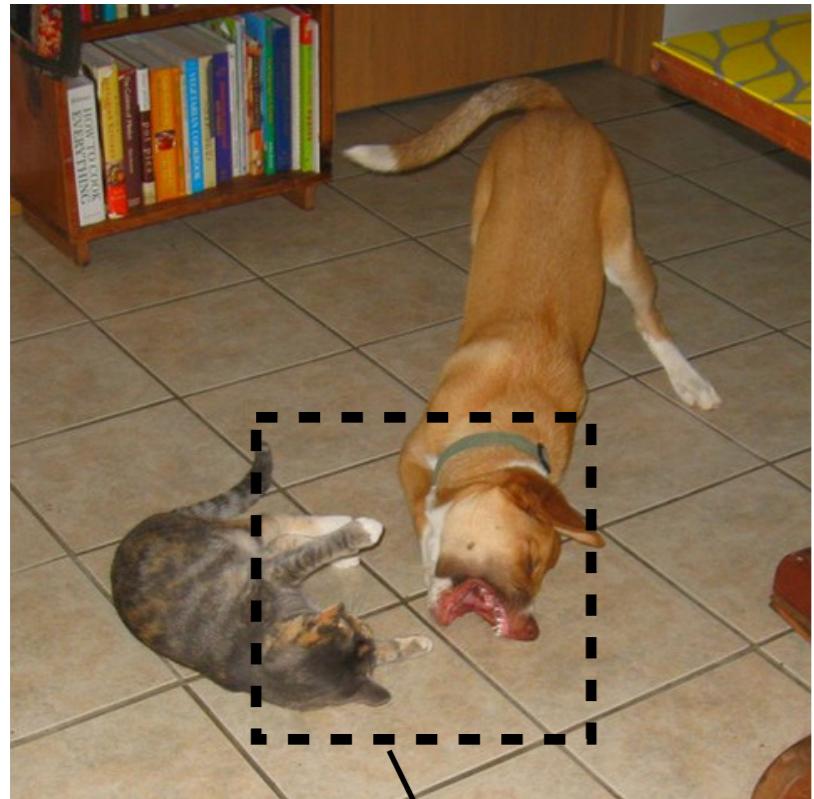
SSD and other deep  
approaches



Discretize the box space **densely**

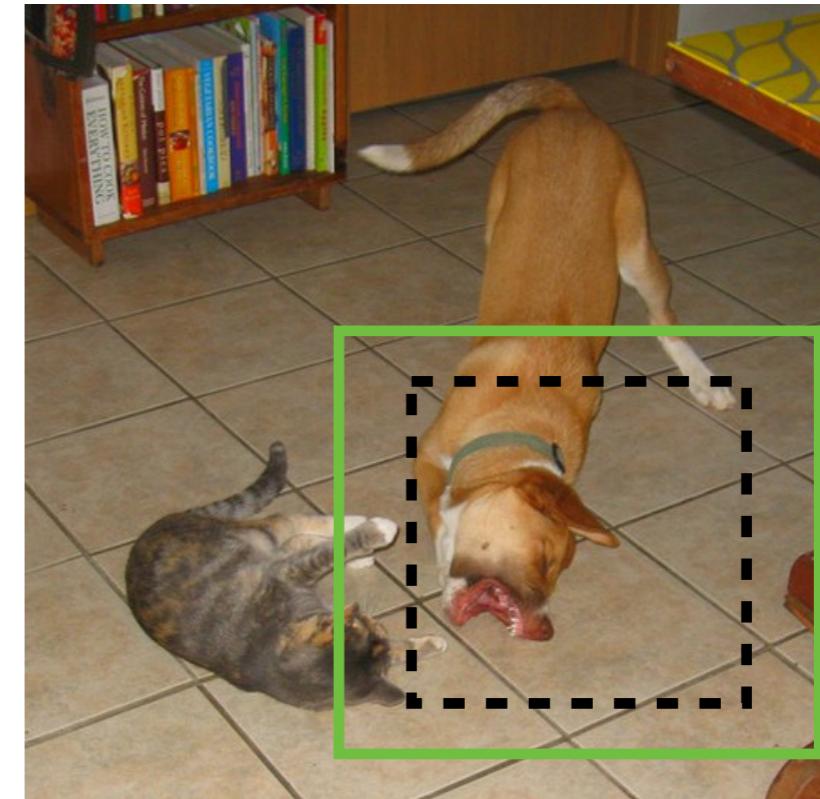
# Bounding Box Prediction

Classical sliding  
windows



Is it a cat? **No**

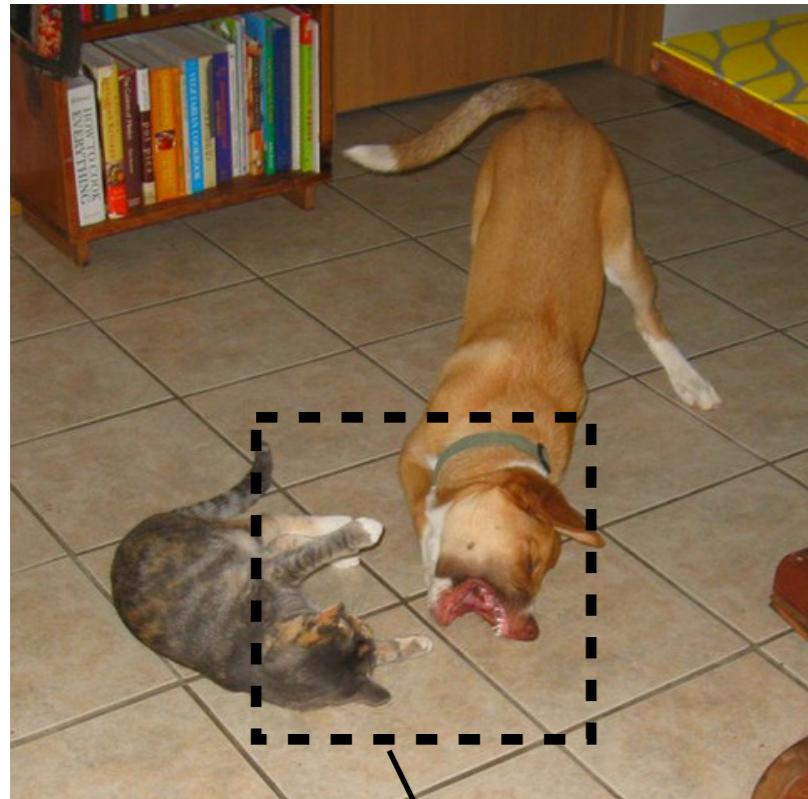
SSD and other deep  
approaches



Discretize the box space **densely**

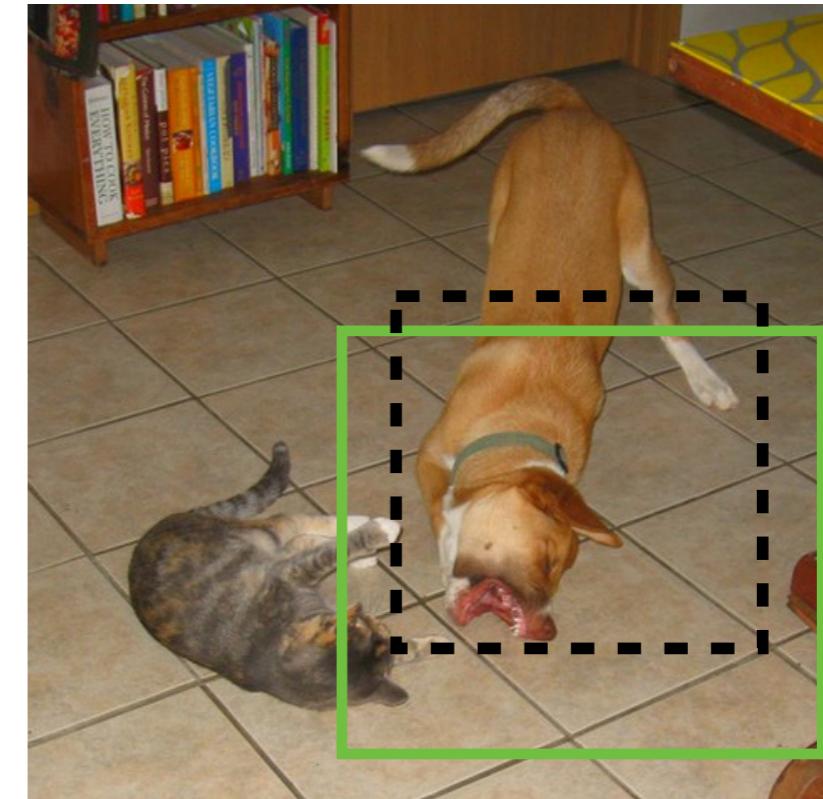
# Bounding Box Prediction

Classical sliding  
windows



Is it a cat? **No**

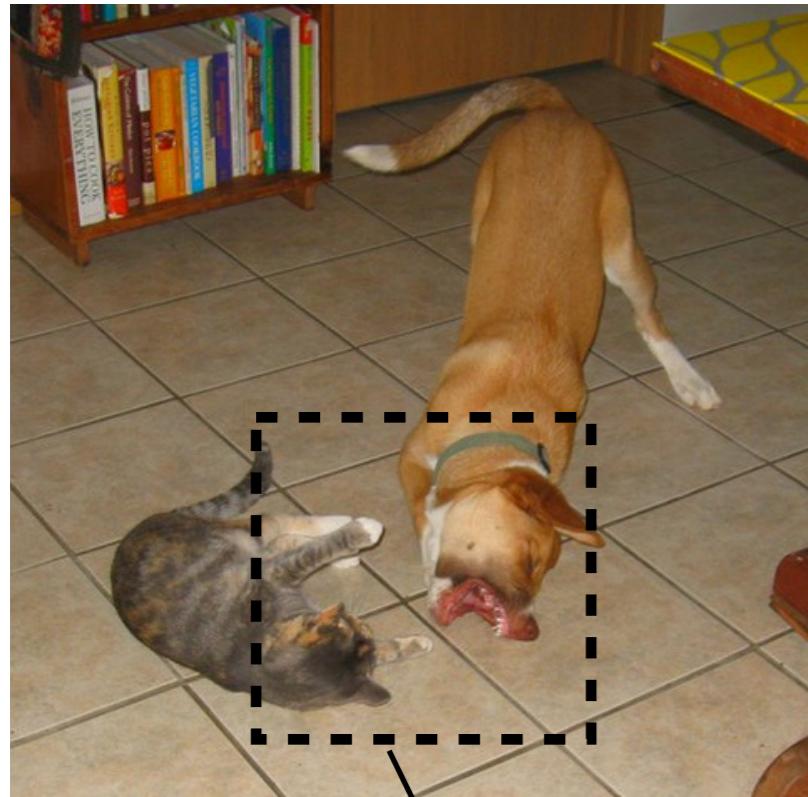
SSD and other deep  
approaches



Discretize the box space **densely**

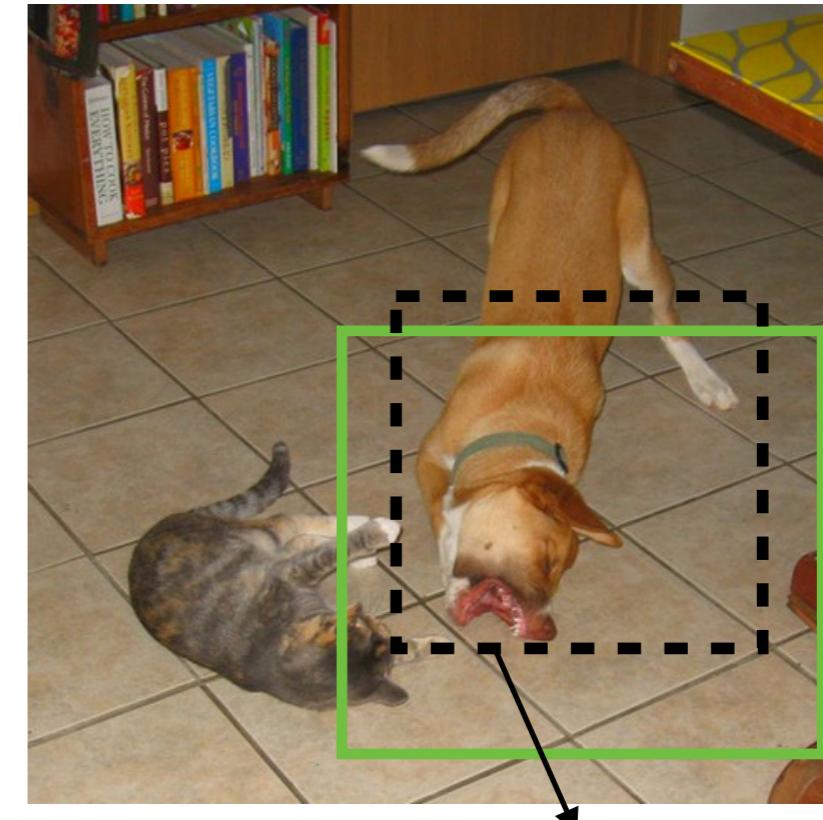
# Bounding Box Prediction

Classical sliding windows



Is it a cat? **No**

SSD and other deep approaches

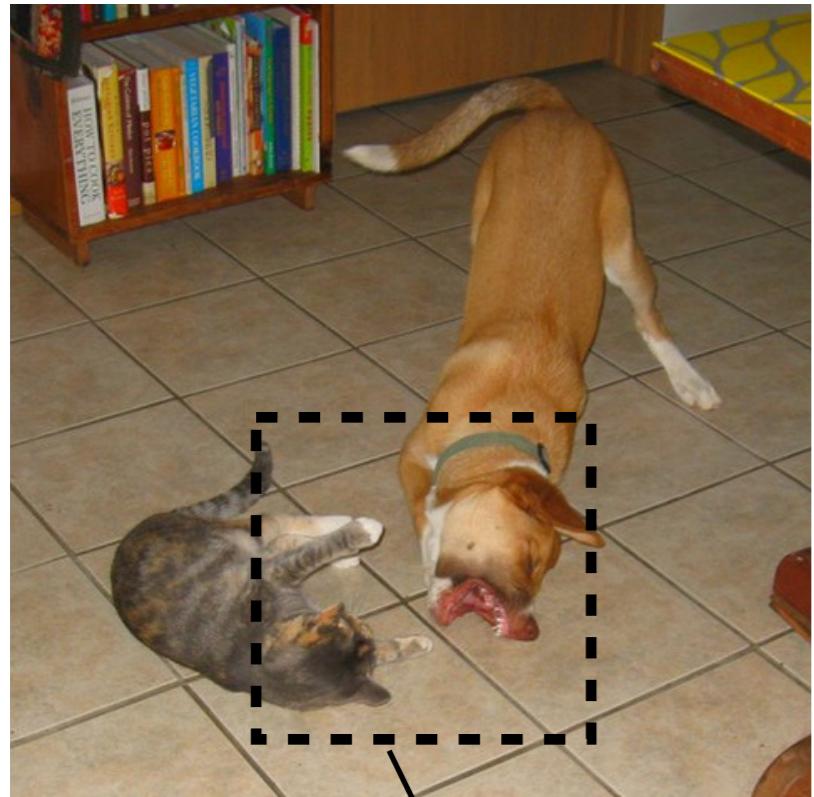


dog: 0.4 cat: 0.2

Discretize the box space **densely**

# Bounding Box Prediction

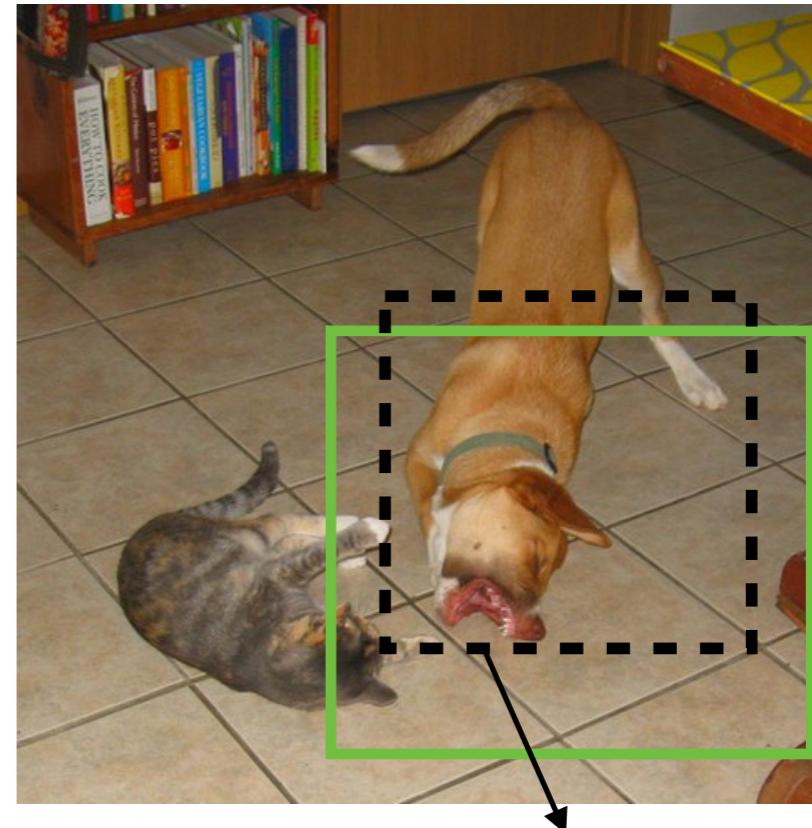
Classical sliding windows



Is it a cat? **No**

Discretize the box space **densely**

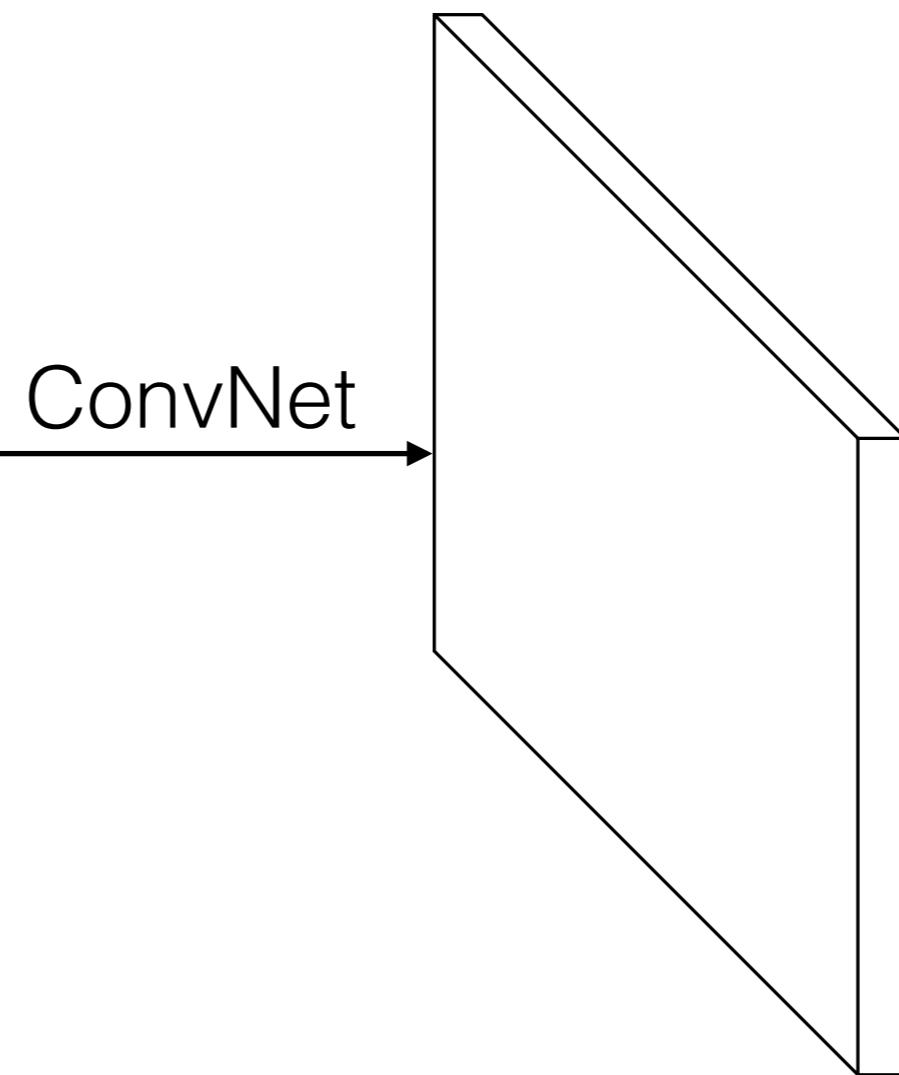
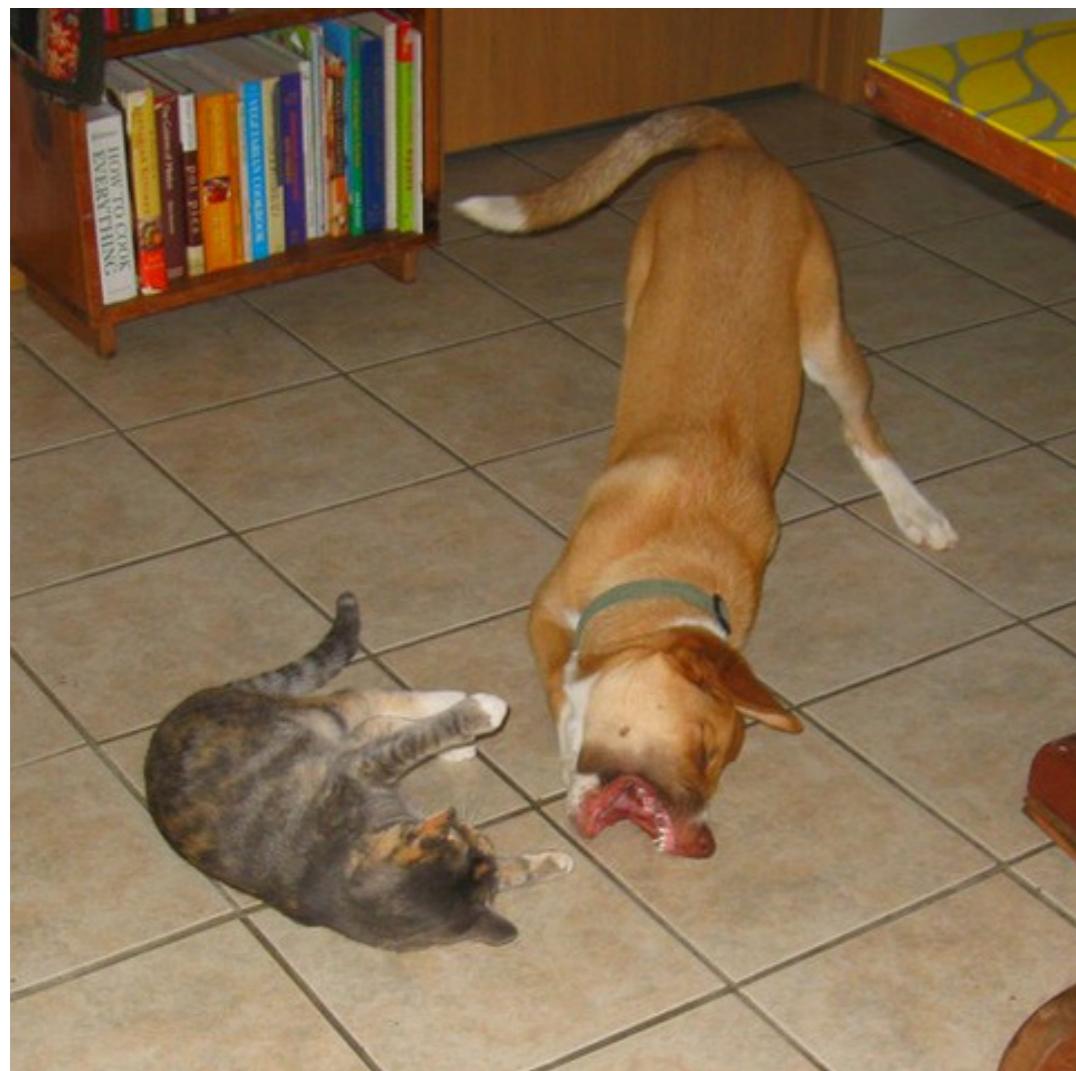
SSD and other deep approaches



dog: 0.4 cat: 0.2

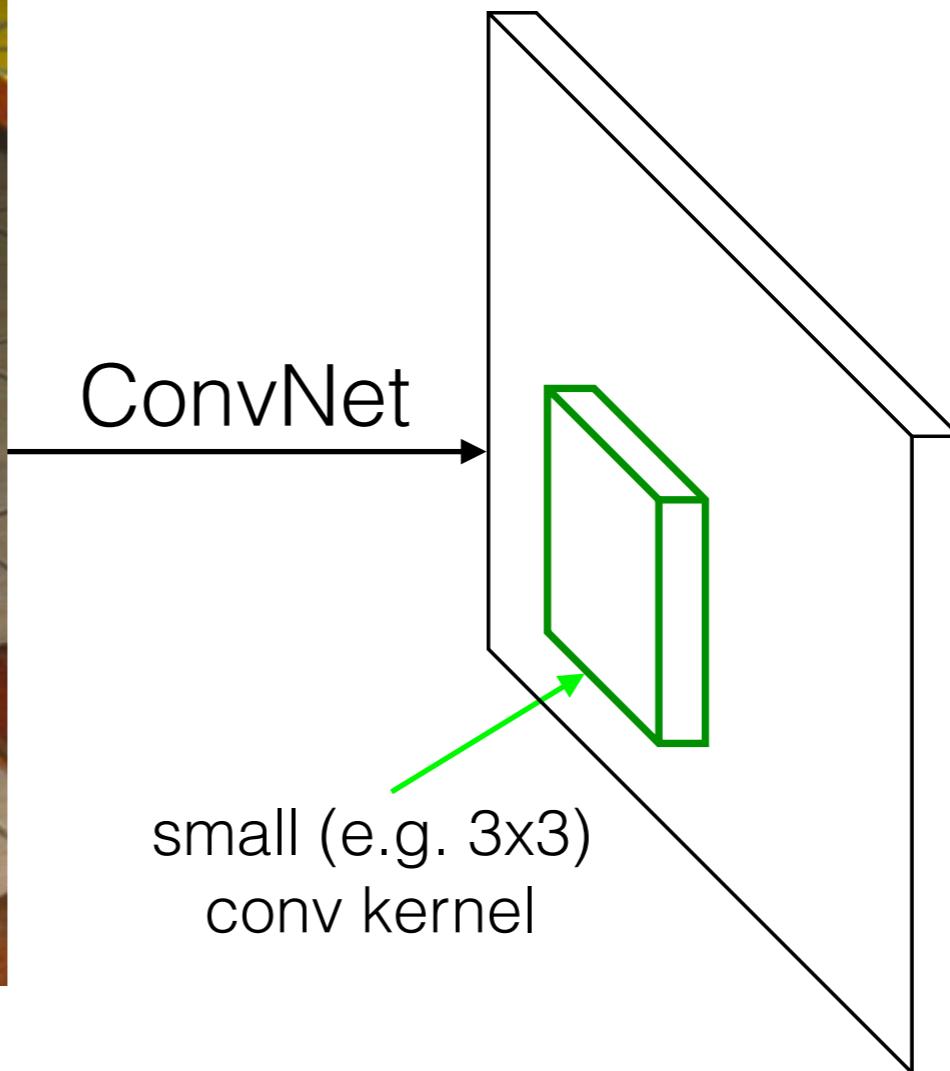
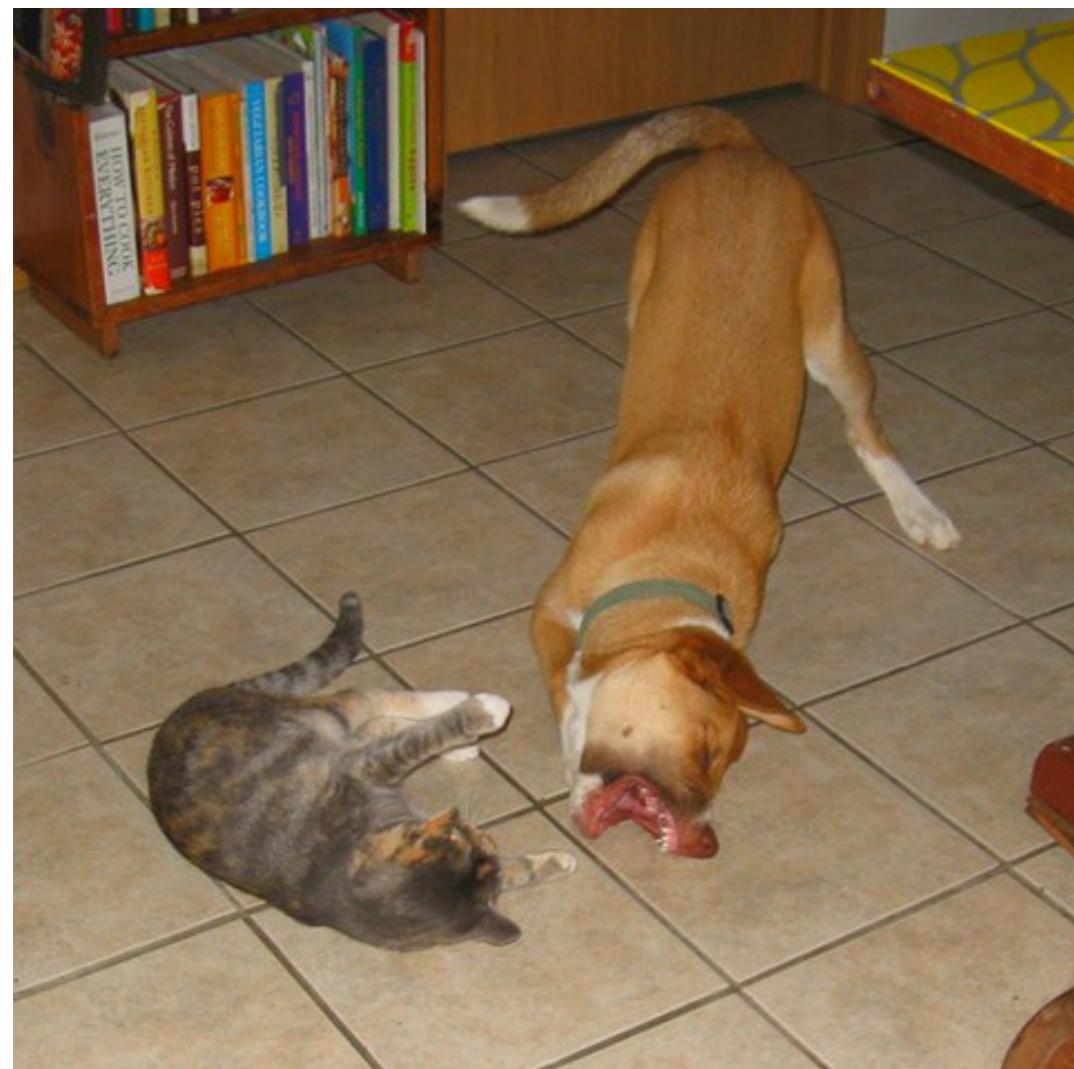
Discretize the box space more **coarsely**  
**Refine** the coordinates of each box

# SSD Output Layer



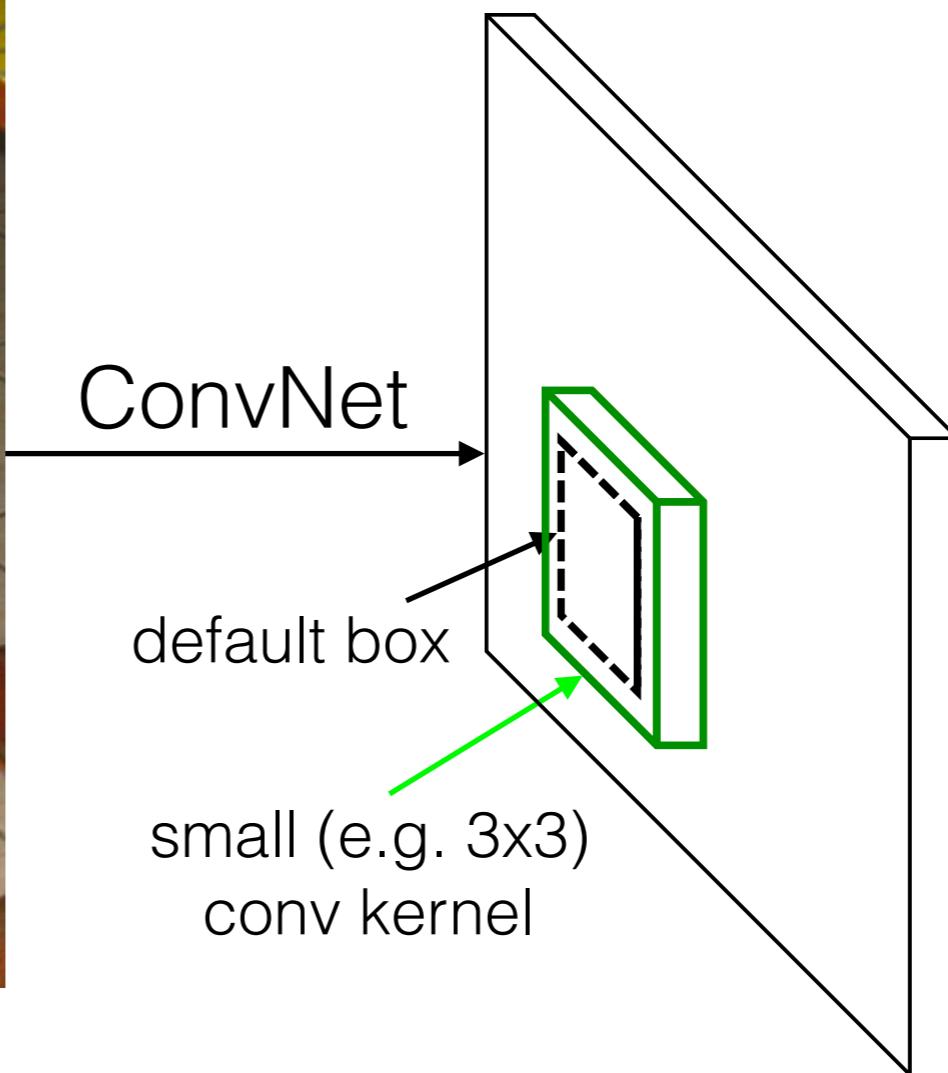
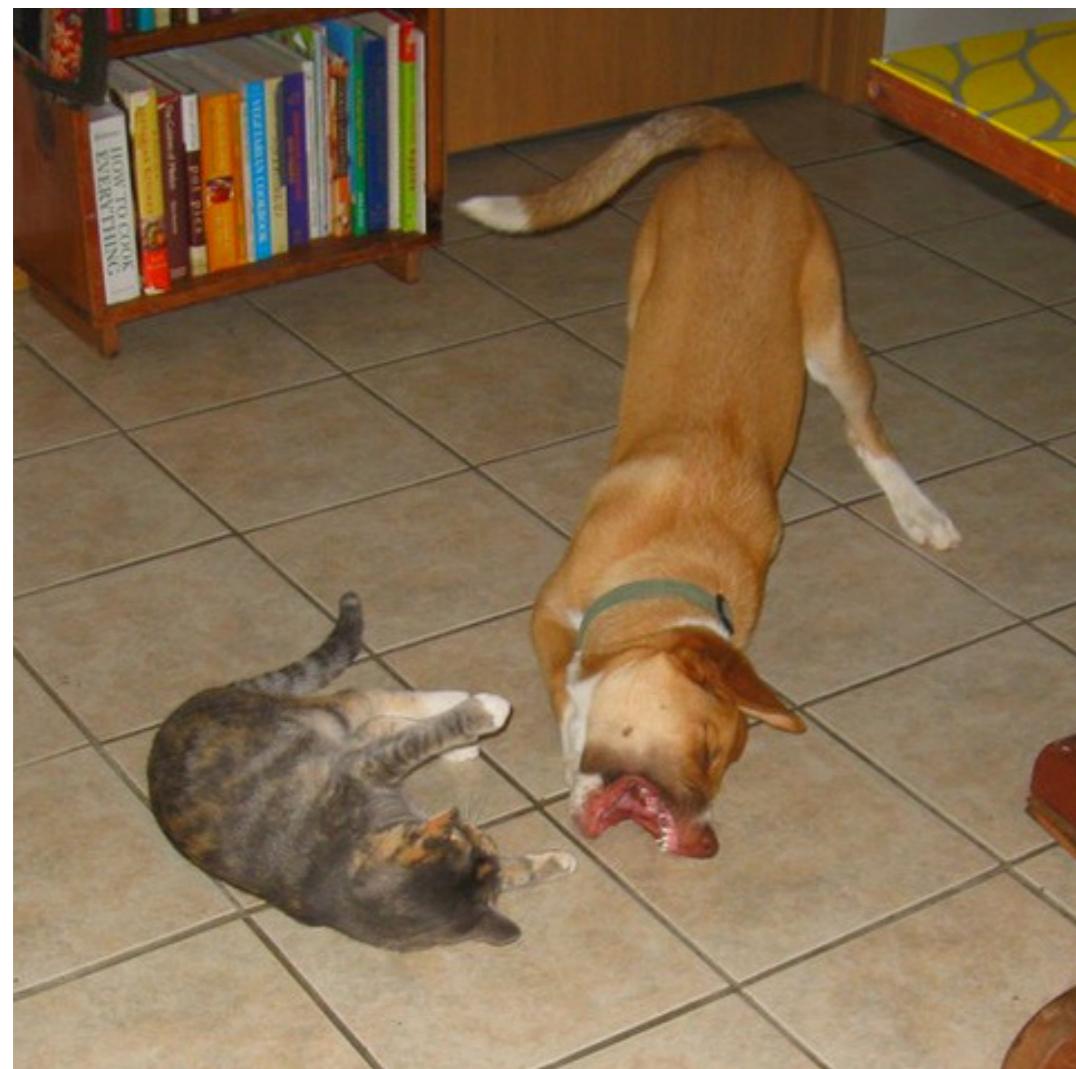
feature map

# SSD Output Layer



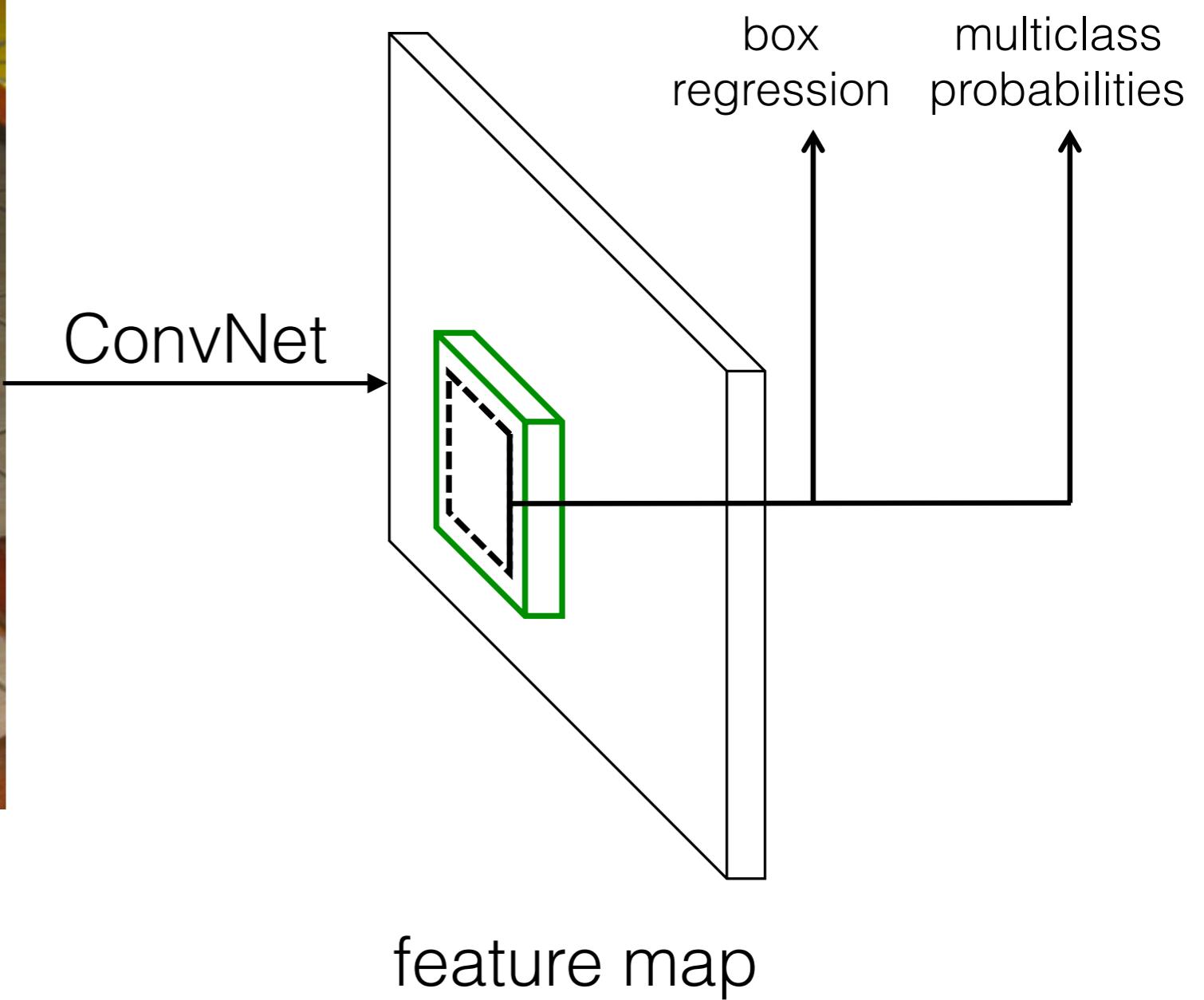
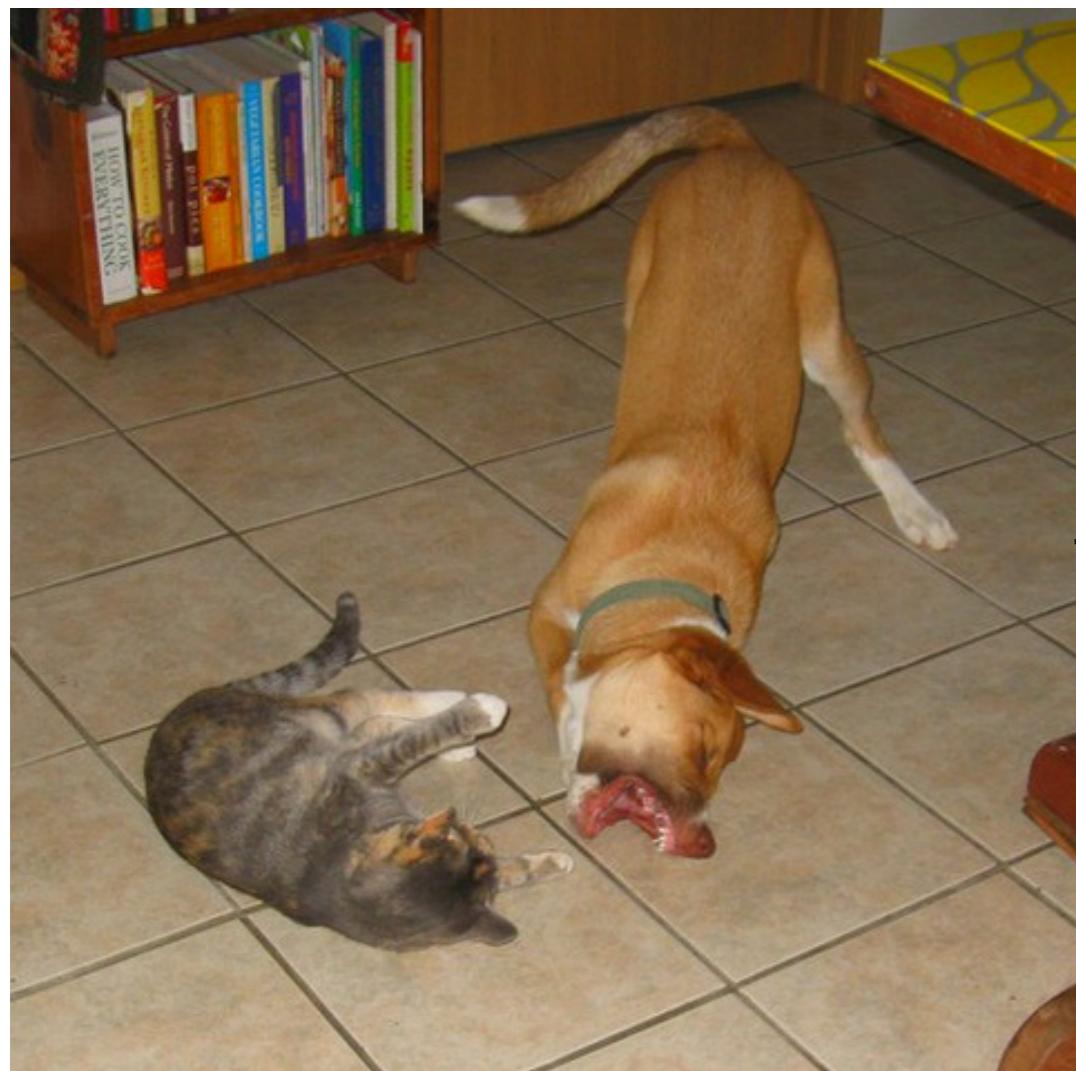
feature map

# SSD Output Layer



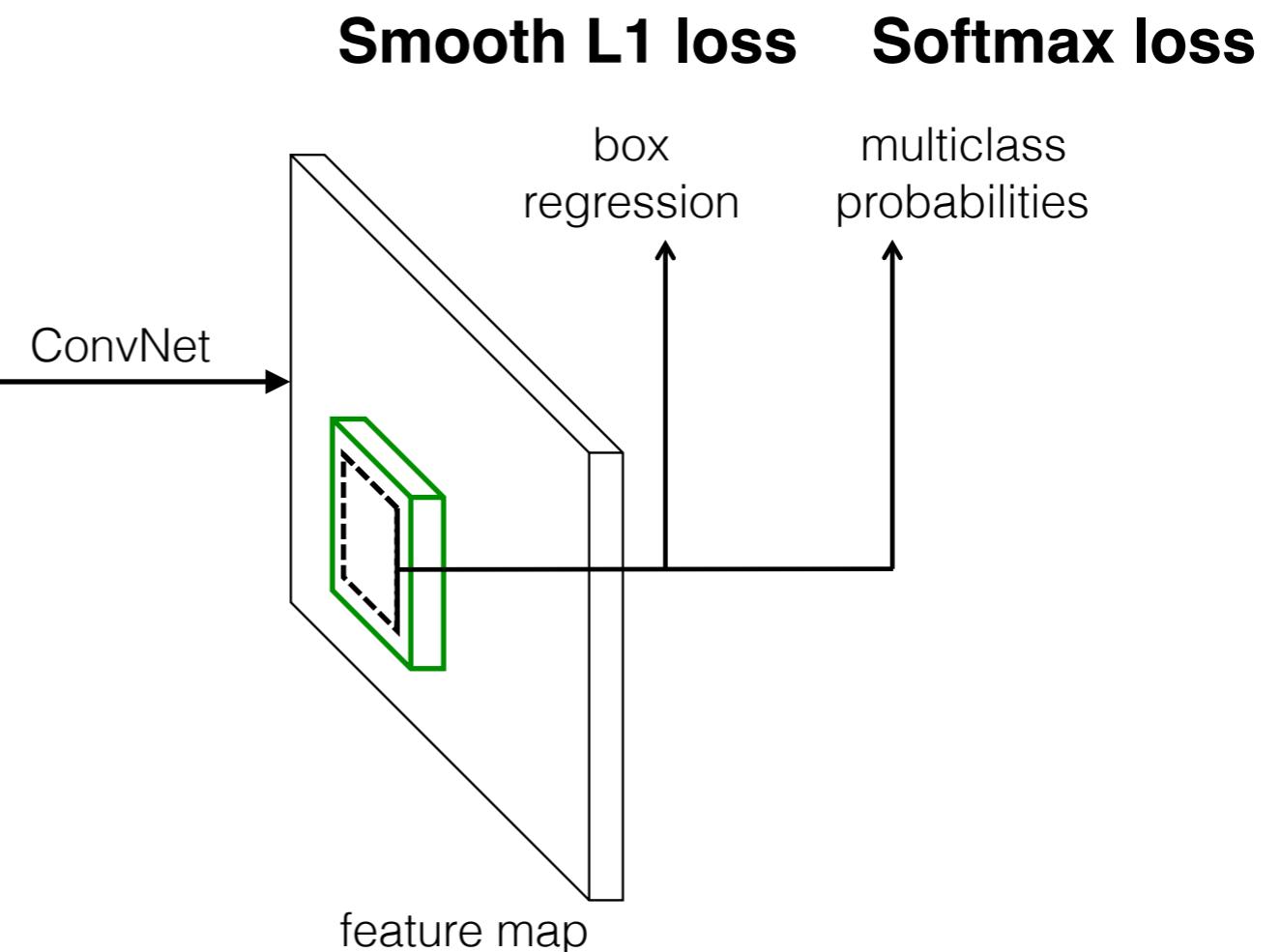
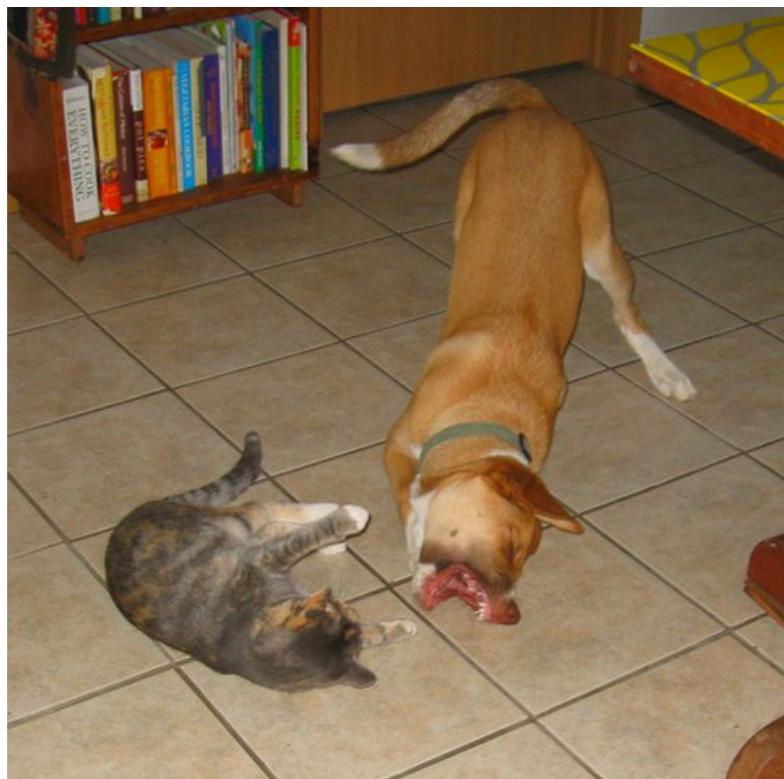
feature map

# SSD Output Layer



# SSD Training

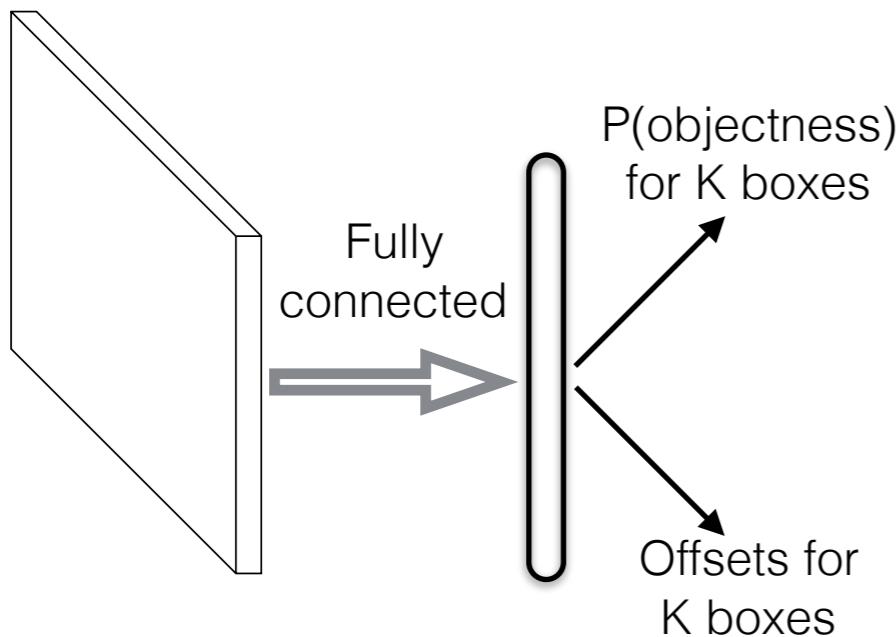
- Match default boxes to ground truth boxes to determine true/false positives.
- Loss = **SmoothL1**(box param) + **Softmax**(class prob)



# Related Work

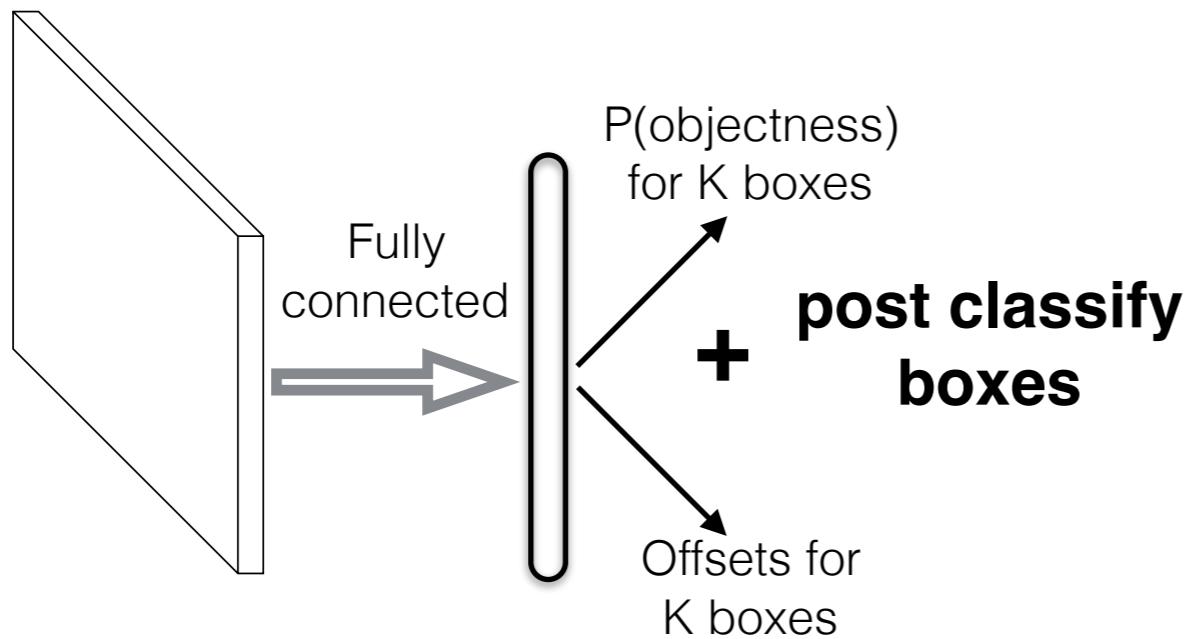
# Related Work

## **MultiBox** [Erhan et al. CVPR14]



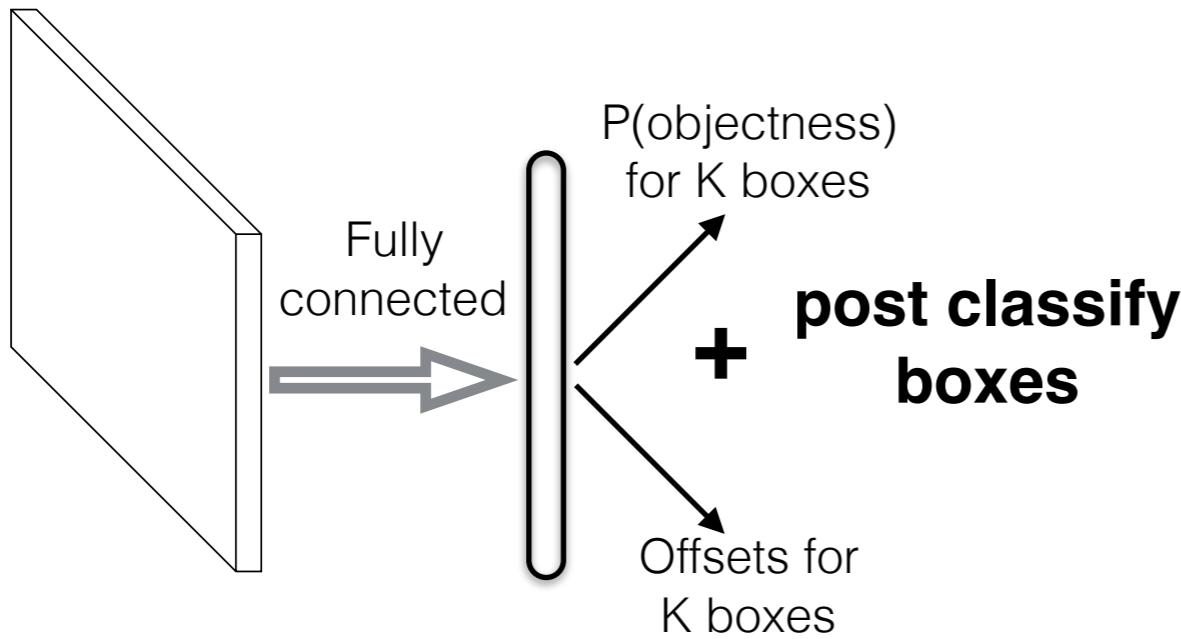
# Related Work

## **MultiBox** [Erhan et al. CVPR14]

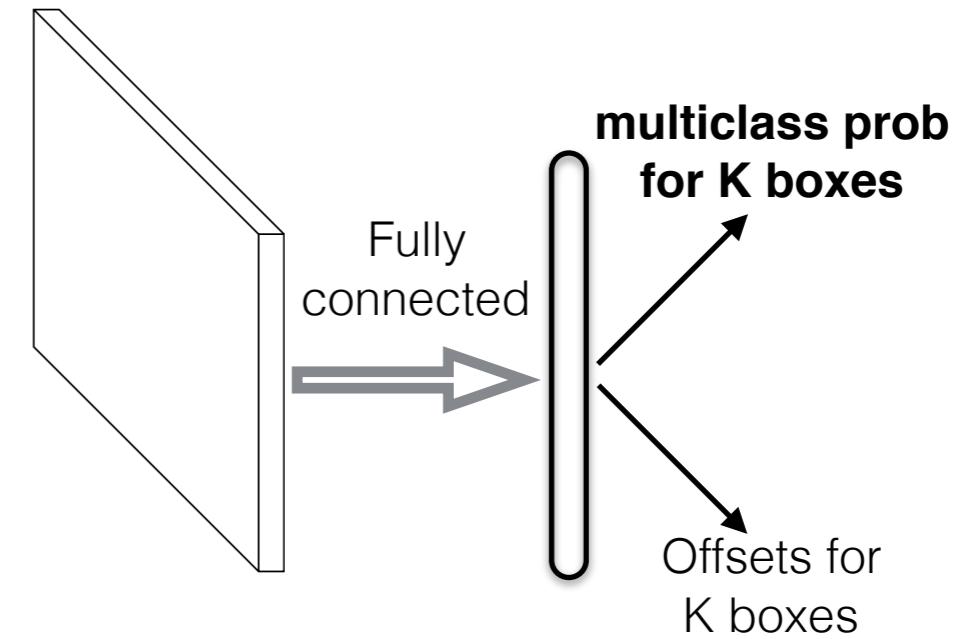


# Related Work

**MultiBox** [Erhan et al. CVPR14]

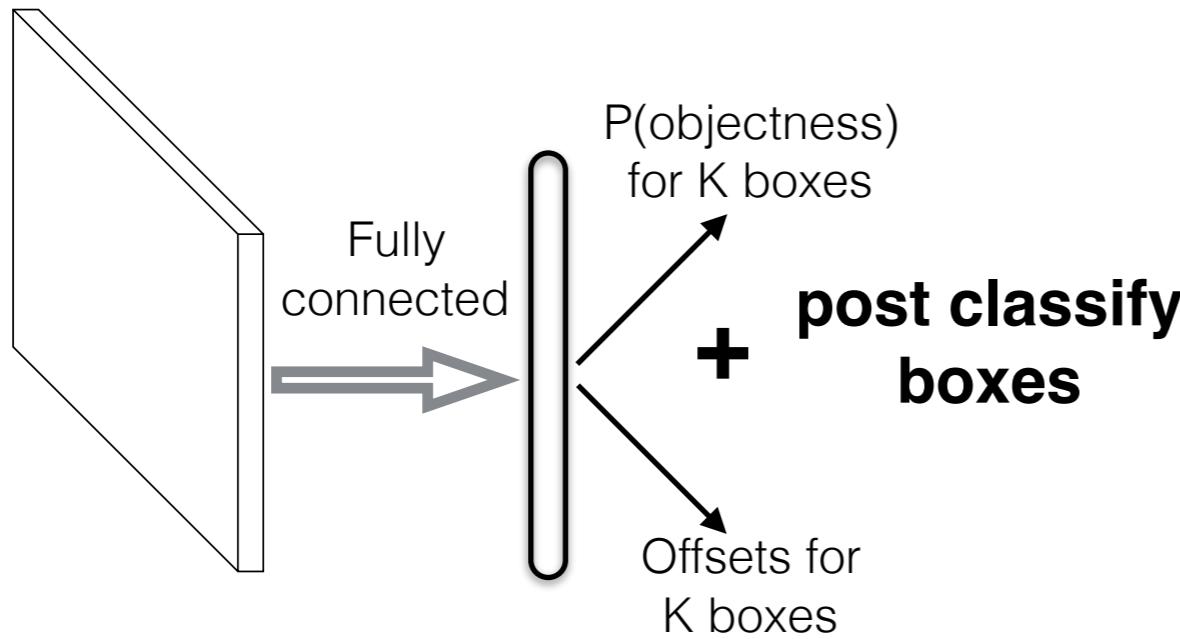


**YOLO** [Redmon et al. CVPR16]

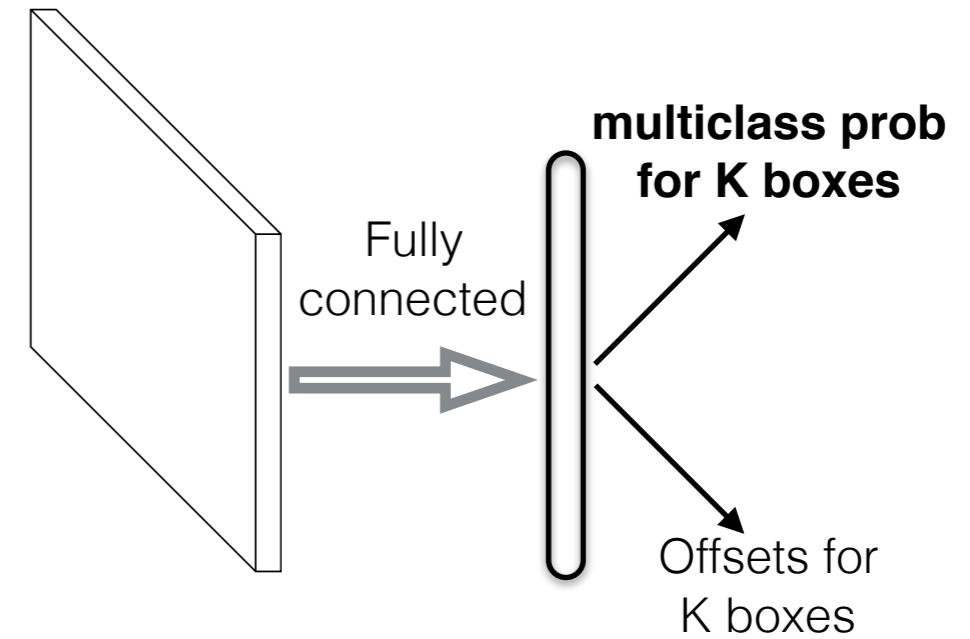


# Related Work

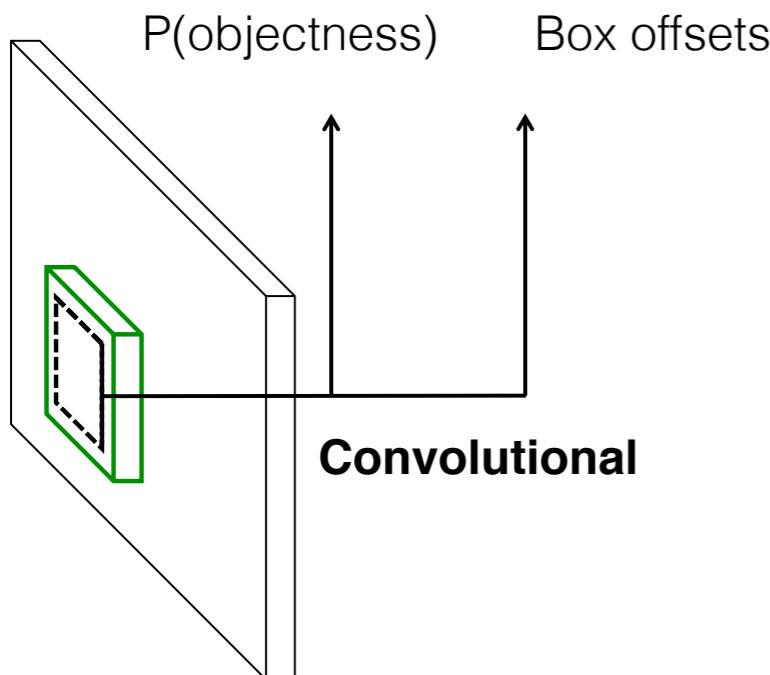
**MultiBox** [Erhan et al. CVPR14]



**YOLO** [Redmon et al. CVPR16]

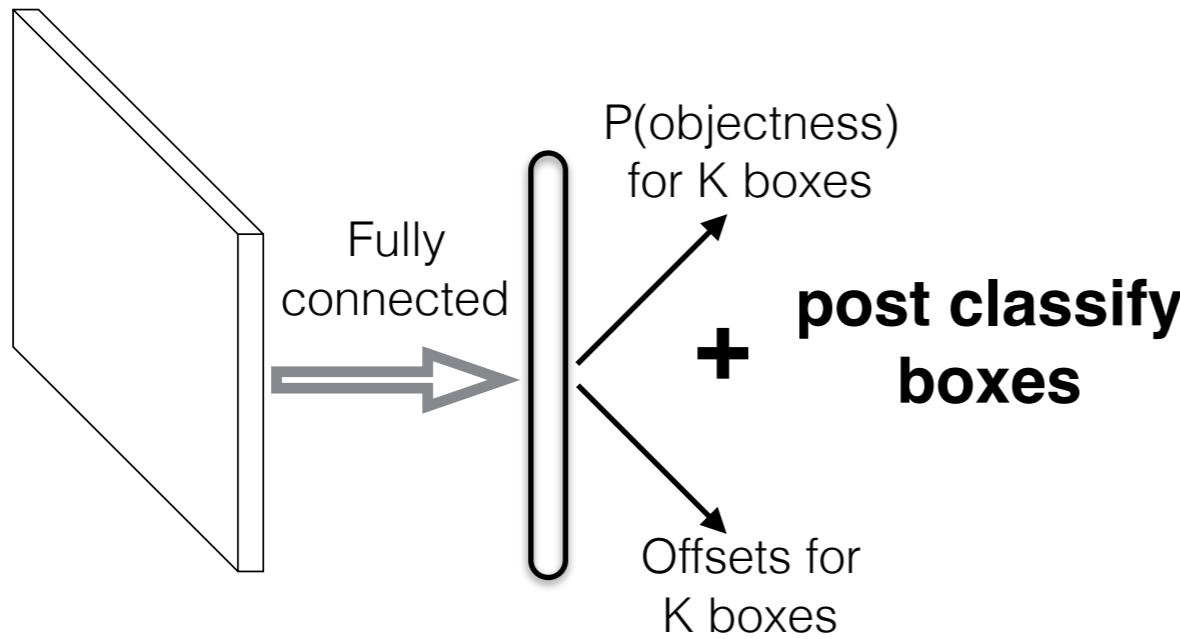


**Faster R-CNN** [Ren et al. NIPS15]

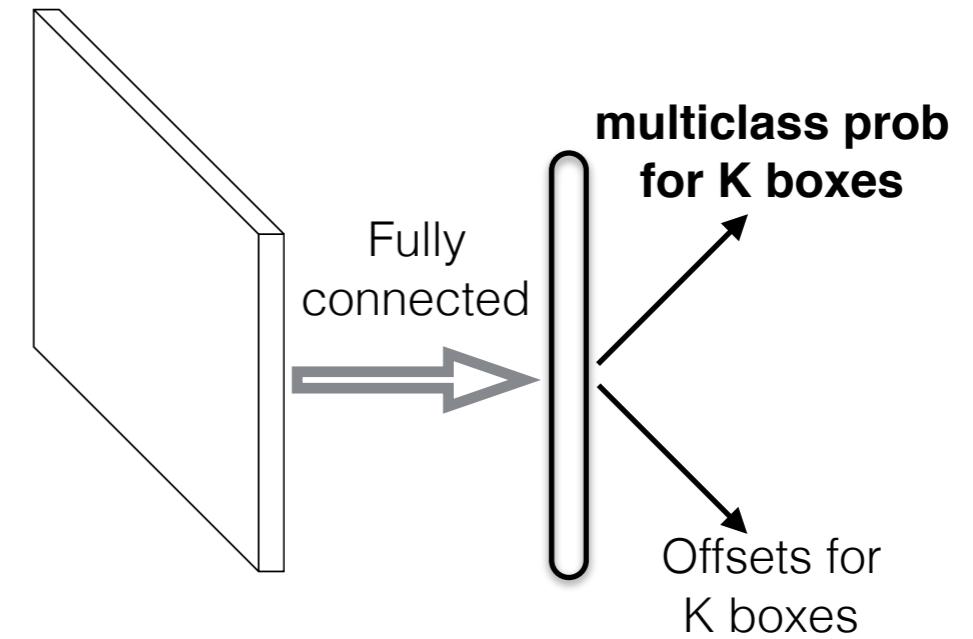


# Related Work

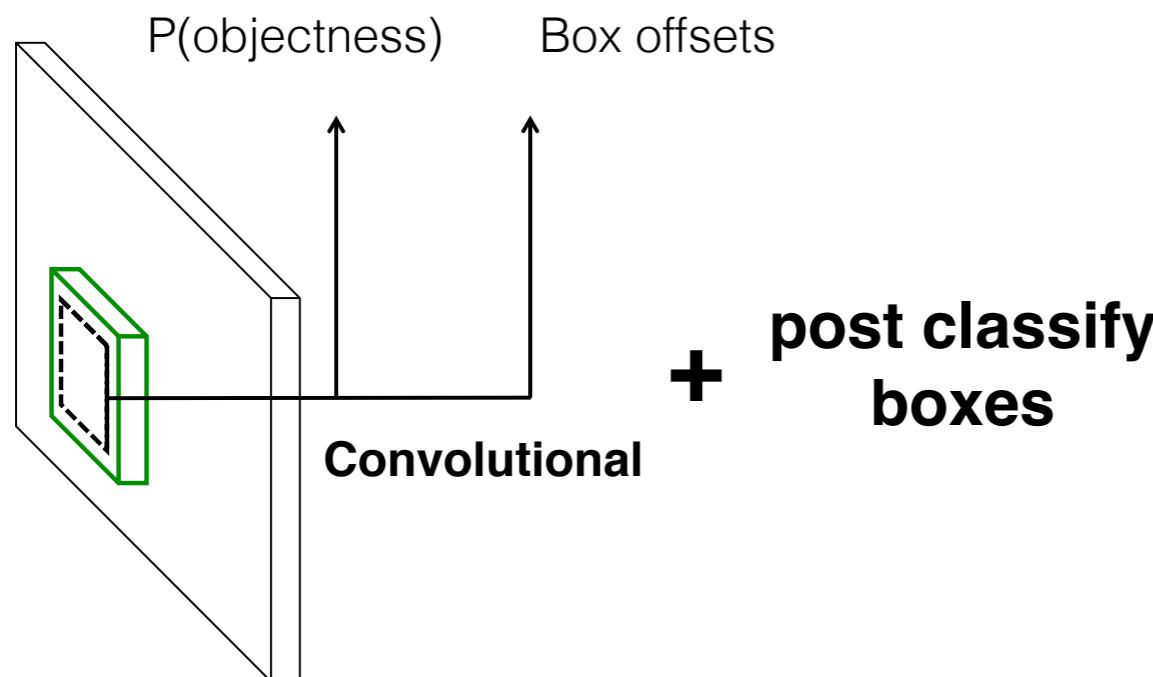
**MultiBox** [Erhan et al. CVPR14]



**YOLO** [Redmon et al. CVPR16]

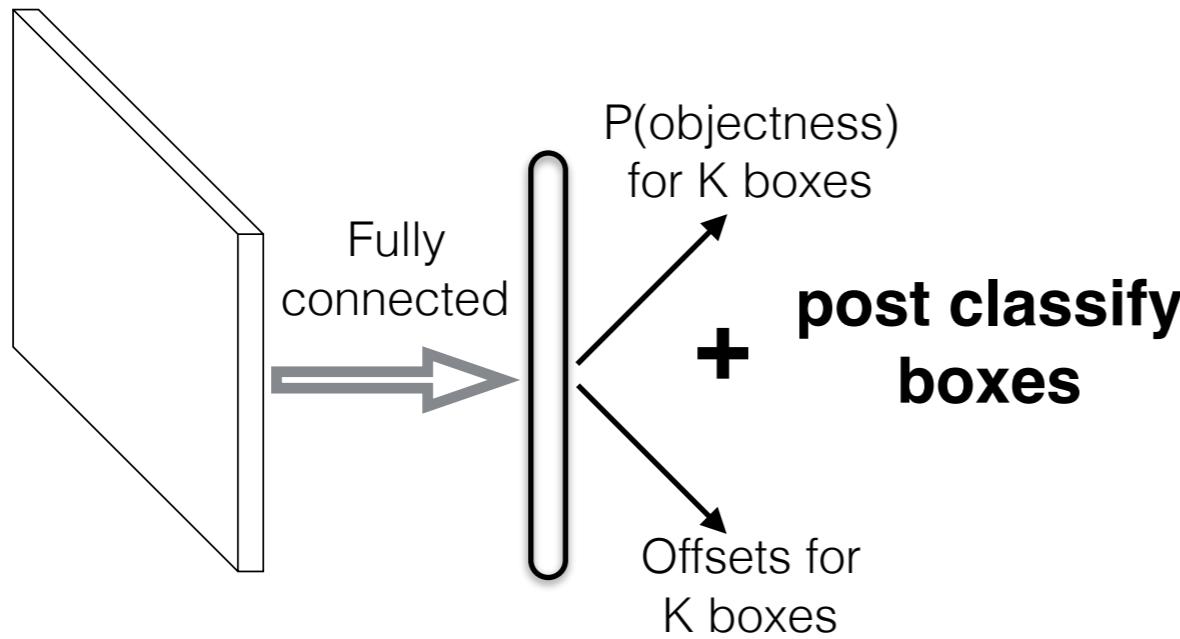


**Faster R-CNN** [Ren et al. NIPS15]

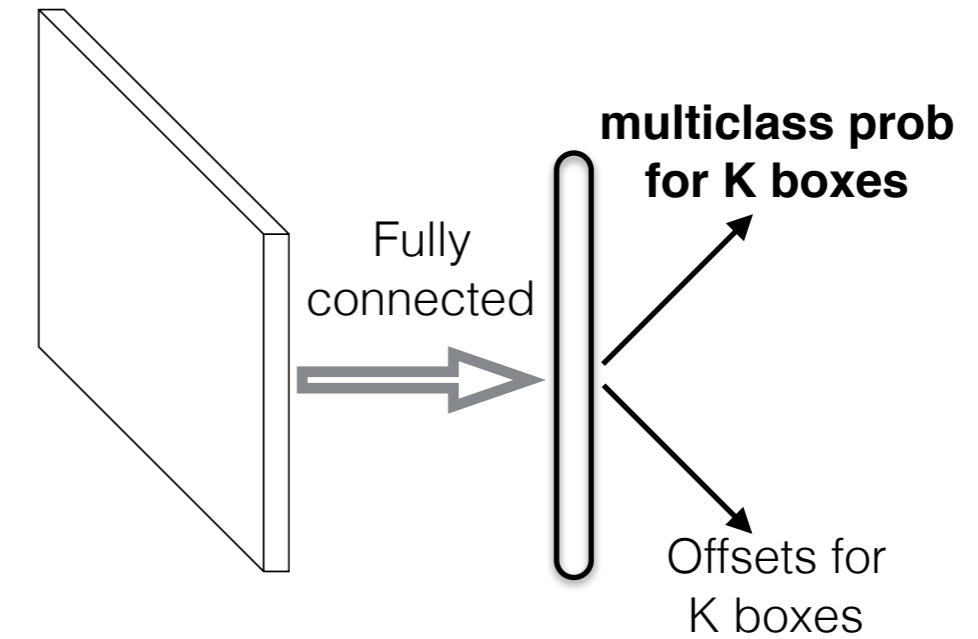


# Related Work

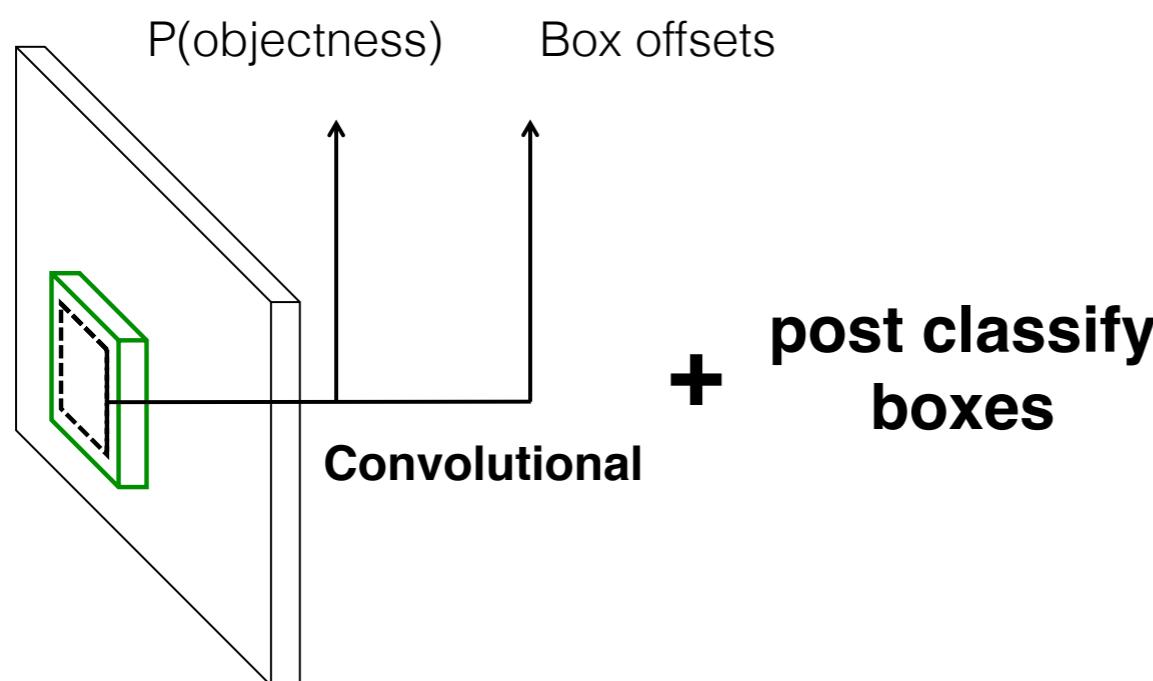
**MultiBox** [Erhan et al. CVPR14]



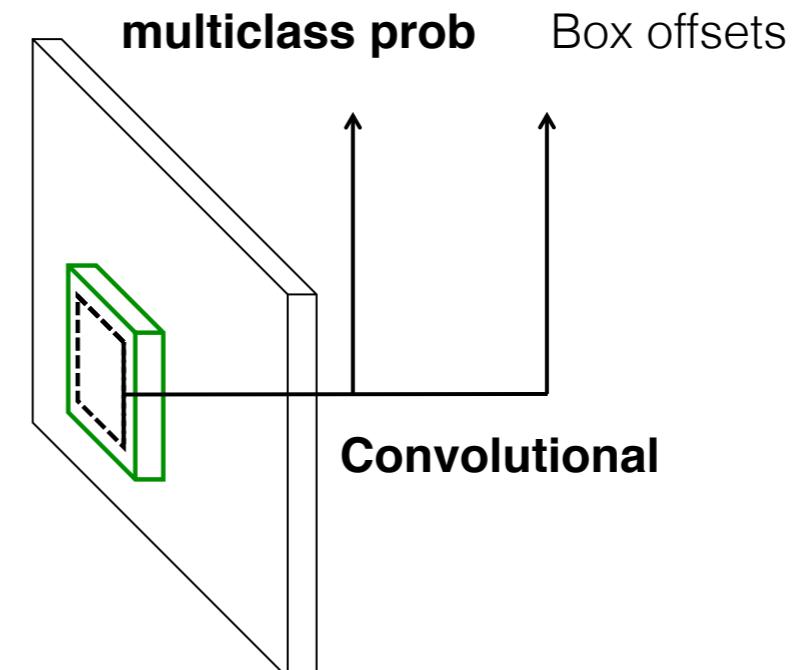
**YOLO** [Redmon et al. CVPR16]



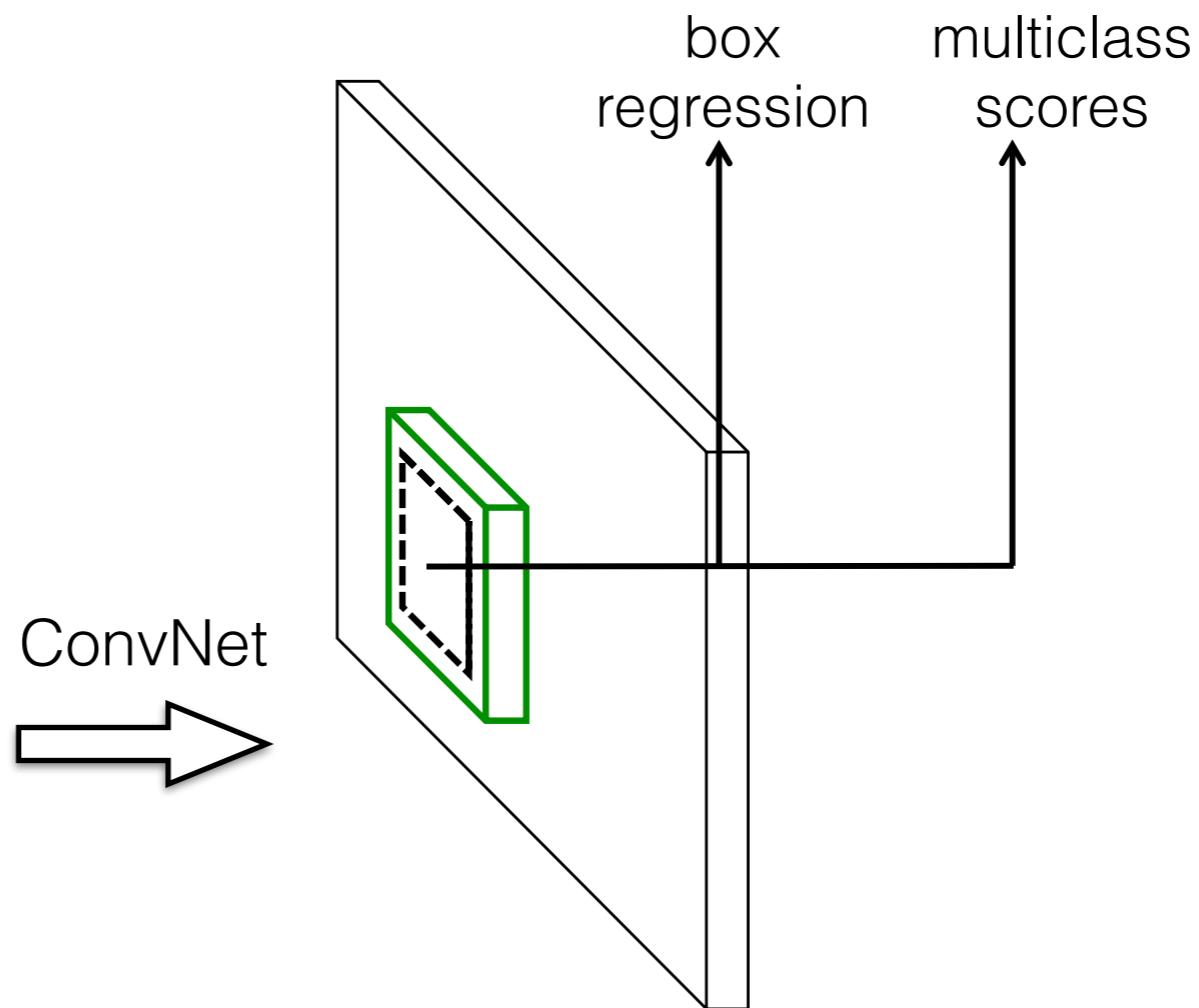
**Faster R-CNN** [Ren et al. NIPS15]



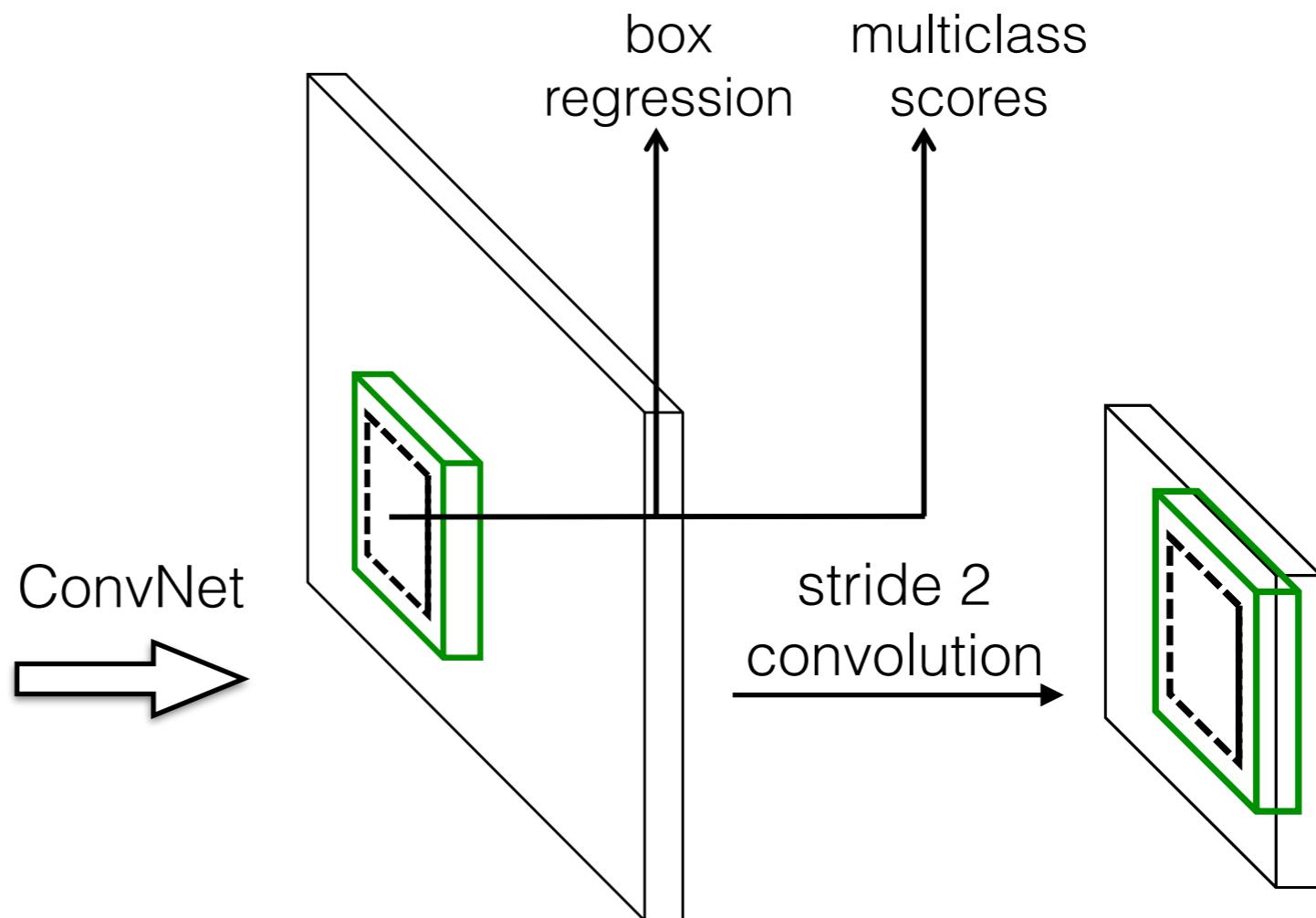
**SSD**



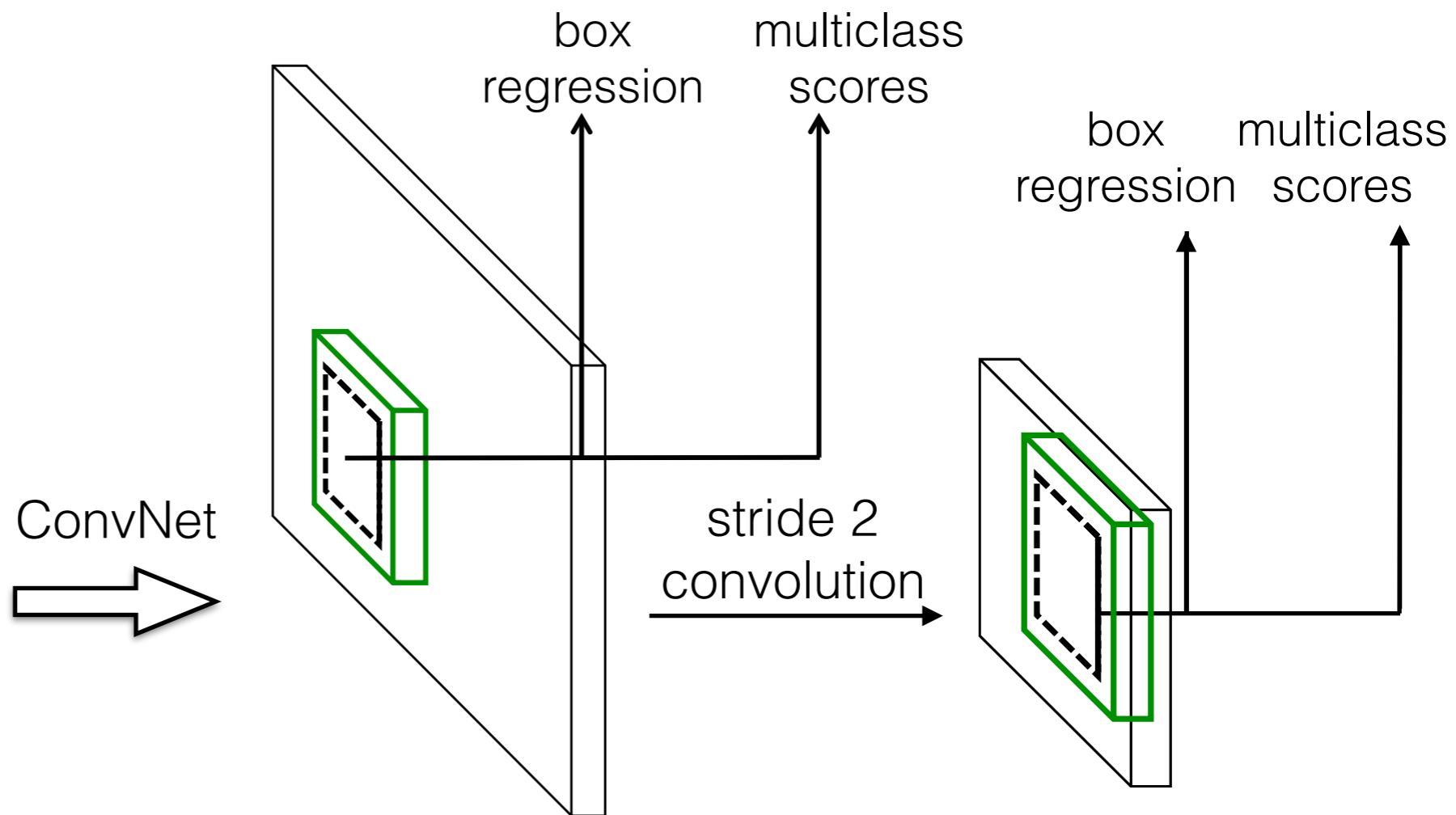
# Contribution #1: Multi-Scale Feature Maps



# Contribution #1: Multi-Scale Feature Maps

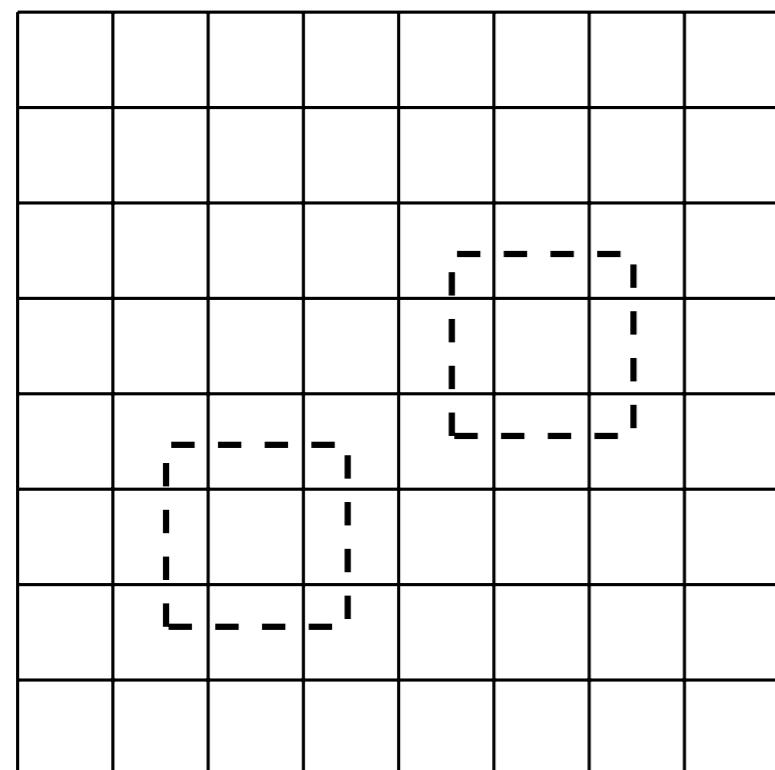


# Contribution #1: Multi-Scale Feature Maps

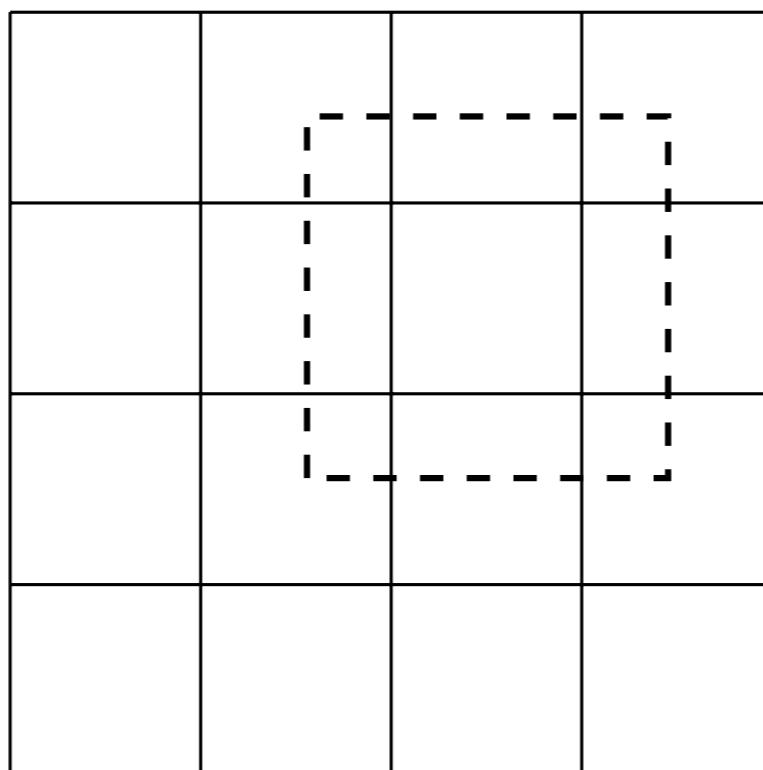


# Multi-Scale Feature Maps

SSD



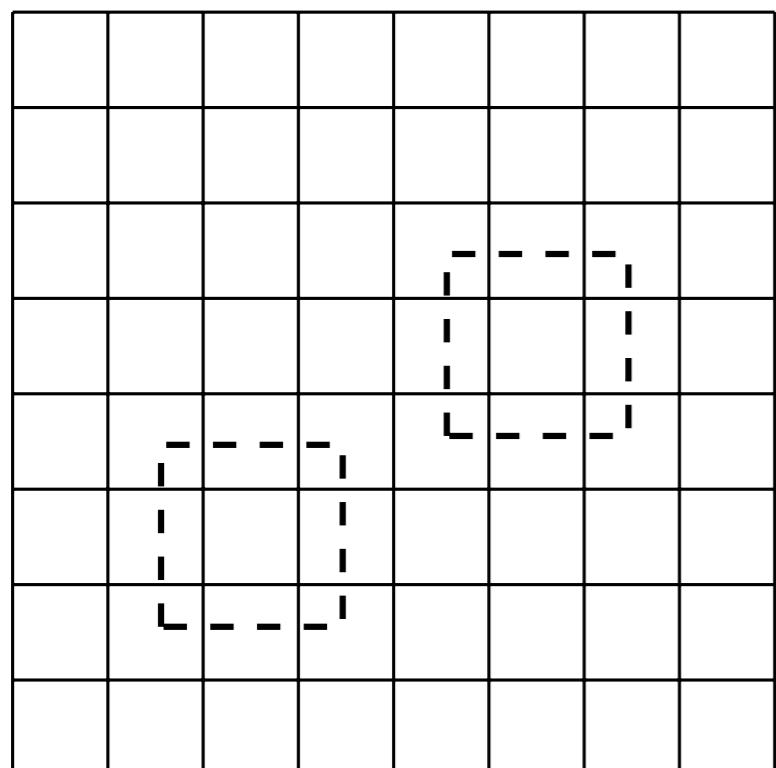
$8 \times 8$  feature map



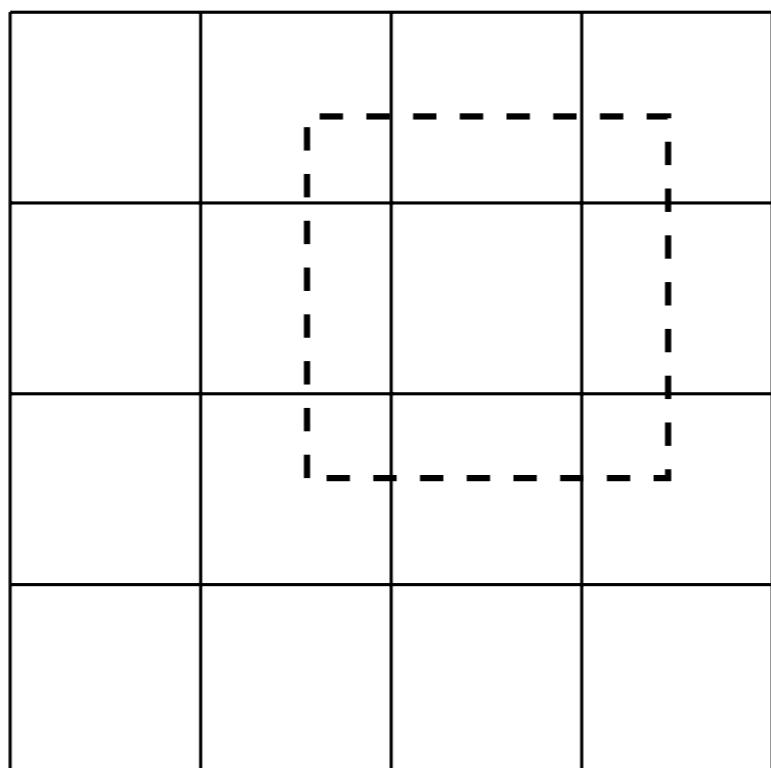
$4 \times 4$  feature map

# Multi-Scale Feature Maps

SSD



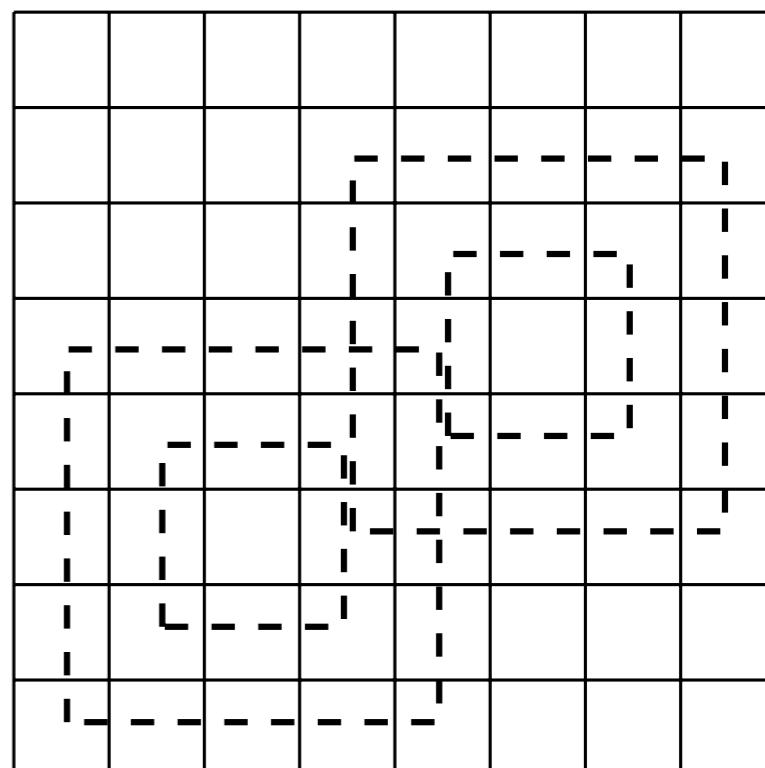
$8 \times 8$  feature map



$4 \times 4$  feature map

Faster R-CNN Objectness  
Proposal, Ren 2015

vs.



$8 \times 8$  feature map

# Multi-Scale Feature Maps Experiment

Prediction source layers from:						mAP		# Boxes
38 × 38	19 × 19	10 × 10	5 × 5	3 × 3	1 × 1	use boundary boxes?		
✓	✓	✓	✓	✓	✓	Yes	No	
						74.3	63.4	8732
						70.7	69.2	9864
						62.4	64.0	8664

# Multi-Scale Feature Maps Experiment

Prediction source layers from:						use boundary boxes?	mAP	# Boxes
38 × 38	19 × 19	10 × 10	5 × 5	3 × 3	1 × 1			
✓	✓	✓	✓	✓	✓	Yes	74.3	8732
✓	✓	✓				No	63.4	
✓						70.7	69.2	9864
						62.4	64.0	8664

# Multi-Scale Feature Maps Experiment

Prediction source layers from:						mAP		# Boxes
38 × 38	19 × 19	10 × 10	5 × 5	3 × 3	1 × 1	use boundary boxes?		
✓	✓	✓	✓	✓	✓	Yes	No	
						74.3	63.4	8732
						70.7	69.2	9864
						62.4	64.0	8664

# Multi-Scale Feature Maps Experiment

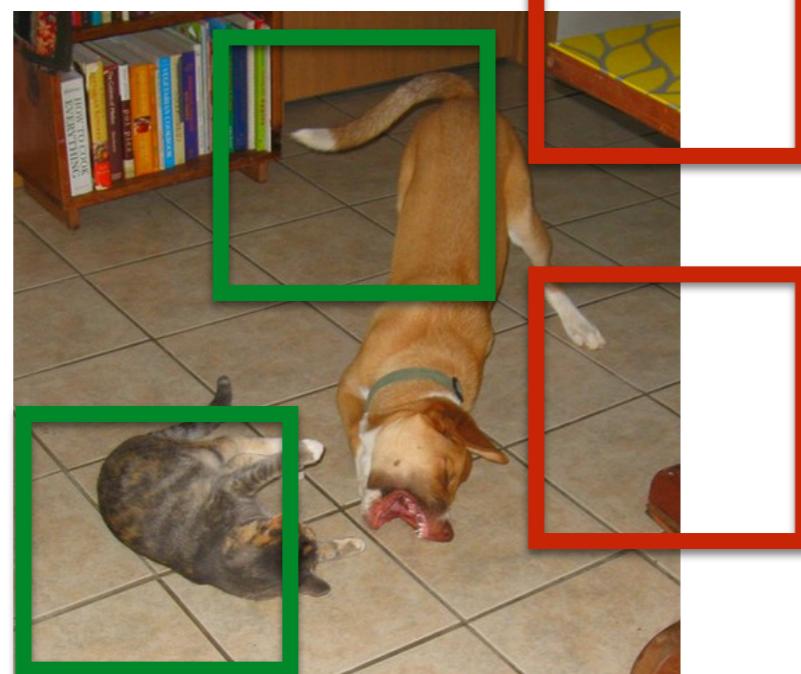
Prediction source layers from:						mAP		# Boxes
38 × 38	19 × 19	10 × 10	5 × 5	3 × 3	1 × 1	use boundary boxes?		
✓	✓	✓	✓	✓	✓	Yes	No	
						74.3	63.4	8732
						70.7	69.2	9864
						62.4	64.0	8664

# Multi-Scale Feature Maps Experiment

Prediction source layers from:						mAP		# Boxes
38 × 38	19 × 19	10 × 10	5 × 5	3 × 3	1 × 1	use boundary boxes?		
✓	✓	✓	✓	✓	✓	Yes	No	
						74.3	63.4	8732
						70.7	69.2	9864
						62.4	64.0	8664

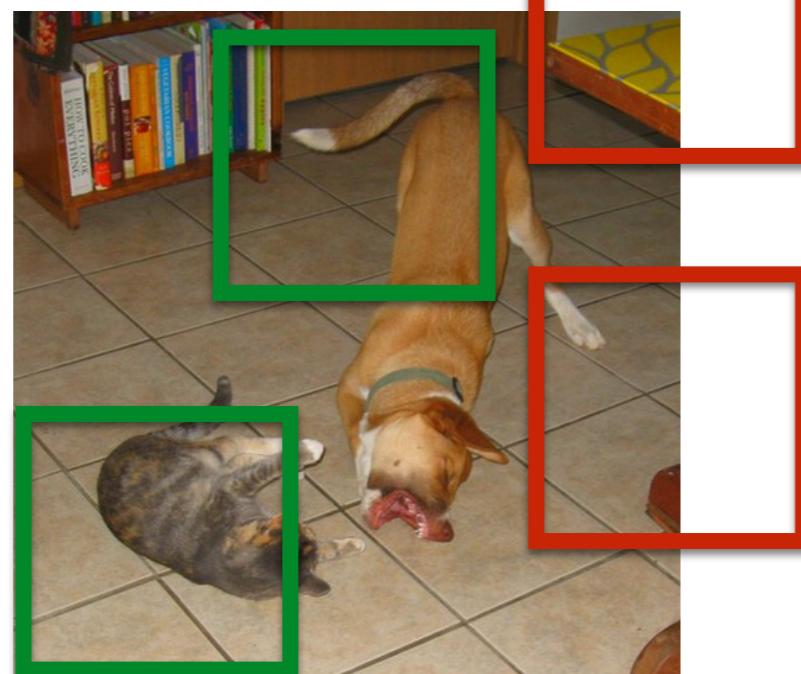
# Multi-Scale Feature Maps Experiment

Prediction source layers from:						mAP		# Boxes
$38 \times 38$	$19 \times 19$	$10 \times 10$	$5 \times 5$	$3 \times 3$	$1 \times 1$	use boundary boxes?		
✓	✓	✓	✓	✓	✓	Yes	No	
✓	✓	✓	✓	✓	✓	74.3	63.4	8732
✓	✓	✓				70.7	69.2	9864
✓						62.4	64.0	8664



# Multi-Scale Feature Maps Experiment

Prediction source layers from:						mAP		# Boxes
$38 \times 38$	$19 \times 19$	$10 \times 10$	$5 \times 5$	$3 \times 3$	$1 \times 1$	use boundary boxes?		
✓	✓	✓	✓	✓	✓	Yes	No	
						74.3	63.4	8732
						70.7	69.2	9864
						62.4	64.0	8664



boundary boxes

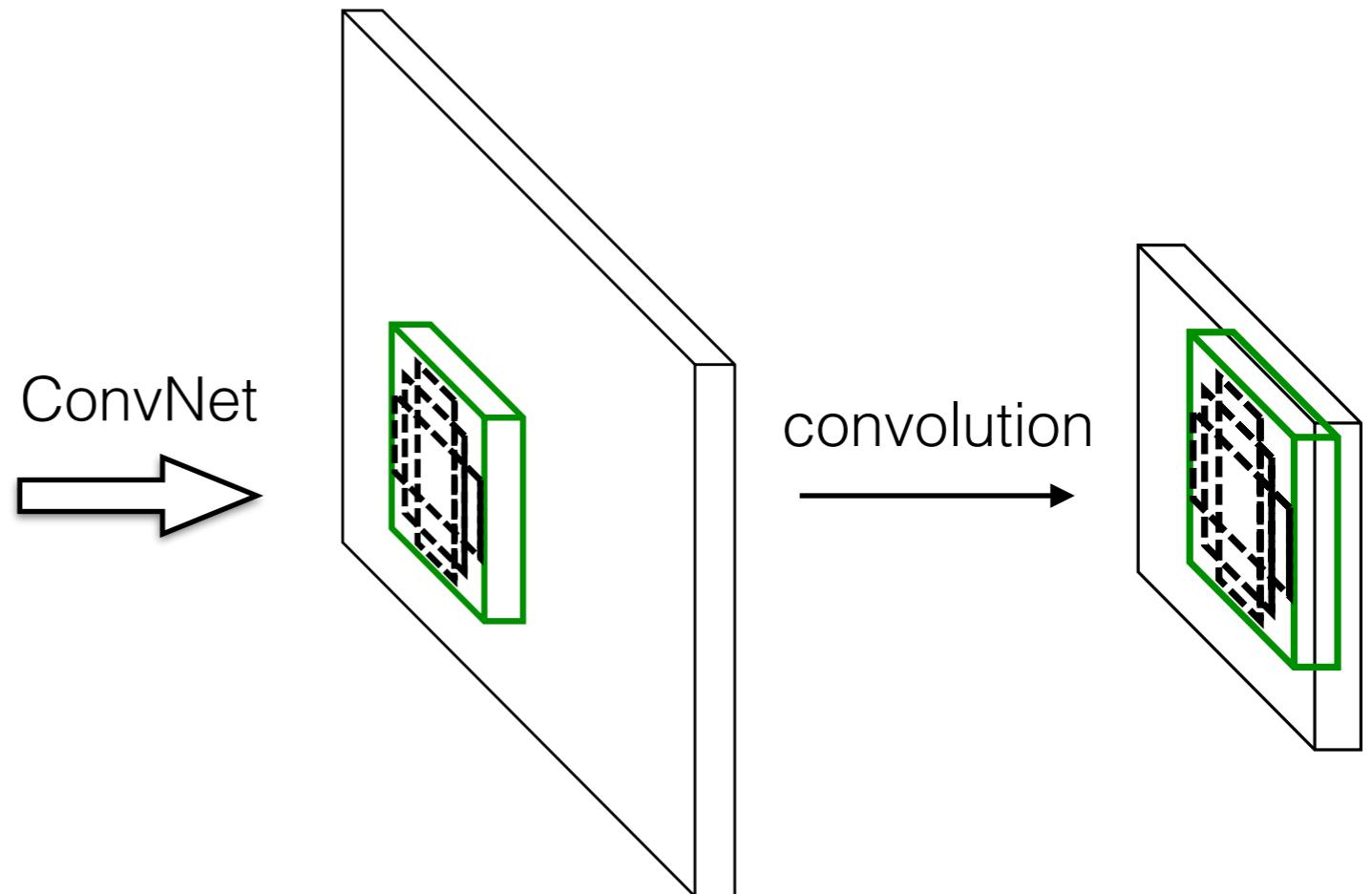
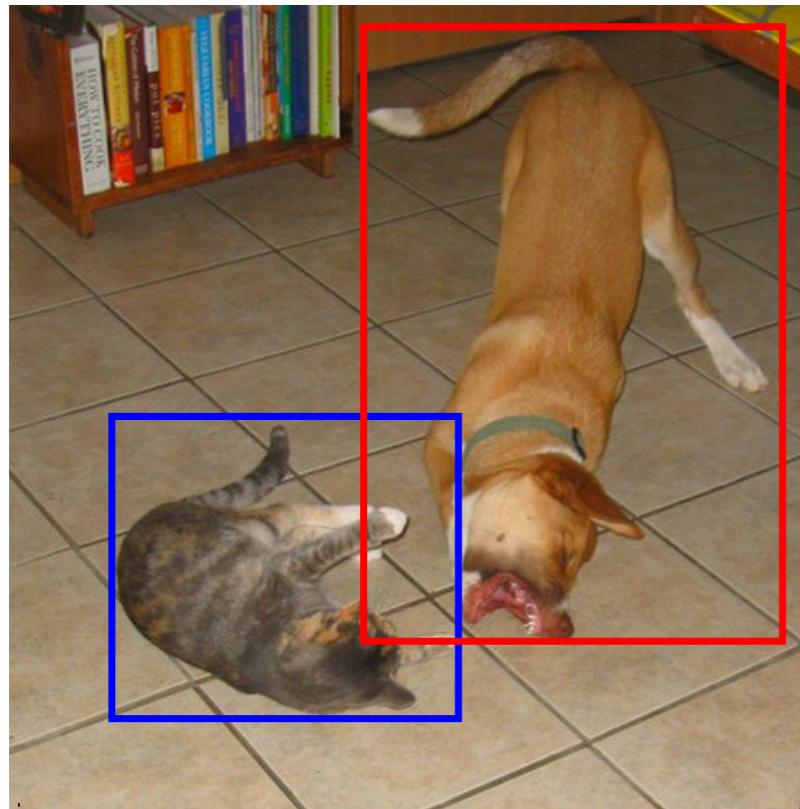
# Multi-Scale Feature Maps Experiment

Prediction source layers from:						use boundary boxes?	mAP	# Boxes
38 × 38	19 × 19	10 × 10	5 × 5	3 × 3	1 × 1			
✓	✓	✓	✓	✓	✓	Yes	74.3	8732
✓	✓	✓				No	63.4	
						70.7	69.2	9864
						62.4	64.0	8664

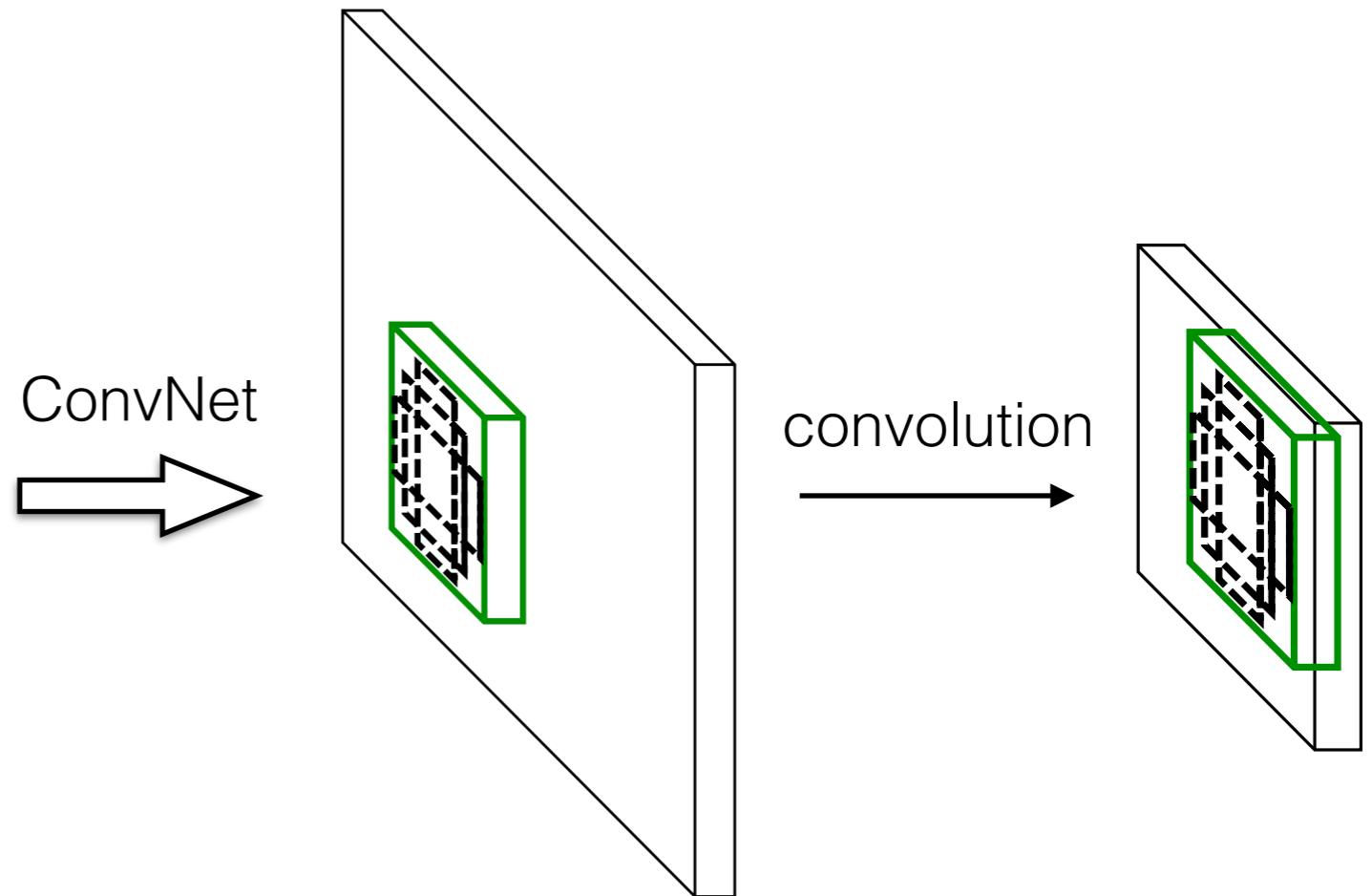
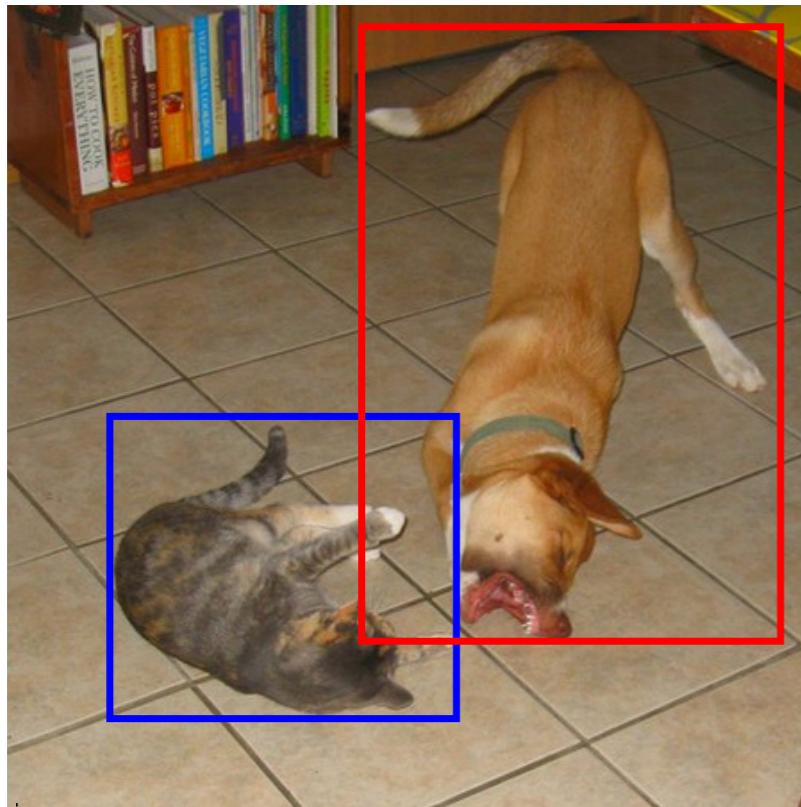
# Multi-Scale Feature Maps Experiment

Prediction source layers from:						mAP	use boundary boxes?	# Boxes
38 × 38	19 × 19	10 × 10	5 × 5	3 × 3	1 × 1	Yes		
✓	✓	✓	✓	✓	✓	74.3	63.4	8732
✓	✓	✓				70.7	69.2	9864
	✓					62.4	64.0	8664

# Contribution #2: Splitting the Region Space

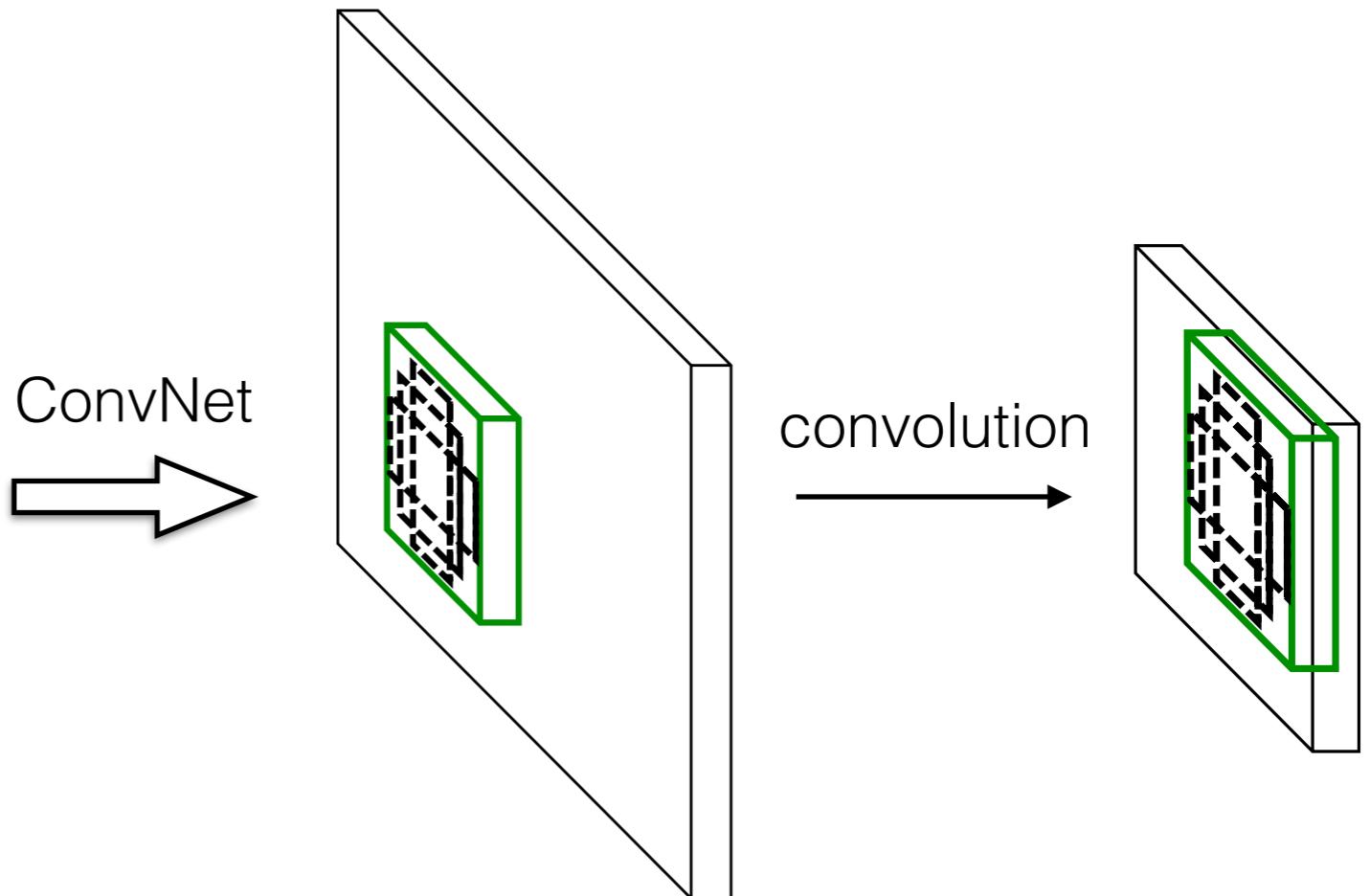
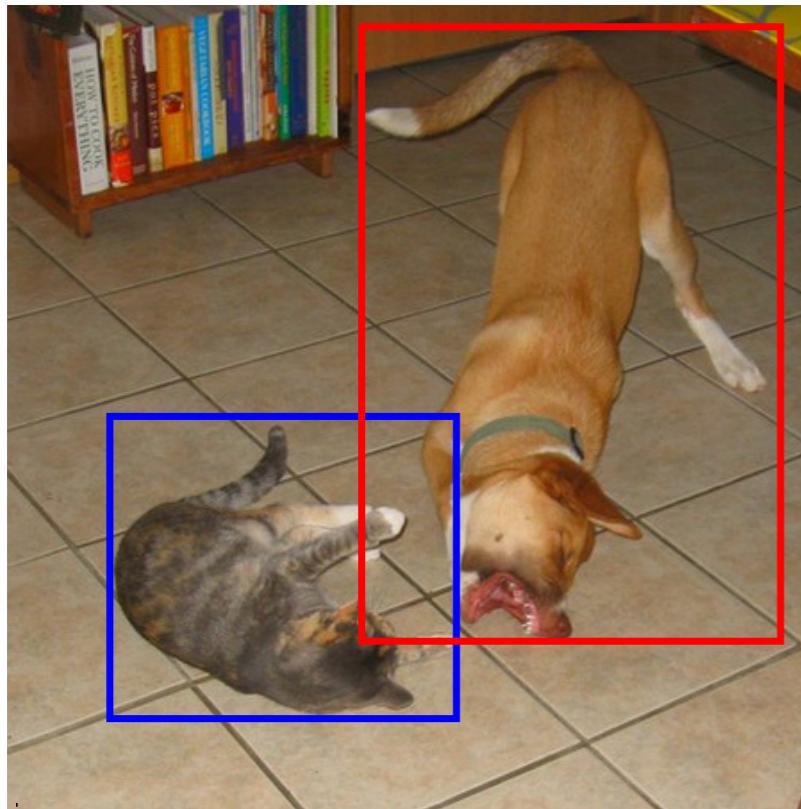


# Contribution #2: Splitting the Region Space



	SSD300	
include $\{\frac{1}{2}, 2\}$ box?	✓	✓
include $\{\frac{1}{3}, 3\}$ box?		✓
number of Boxes	3880	7760
VOC2007 test mAP	71.6	73.7
		<b>74.3</b>

# Contribution #2: Splitting the Region Space



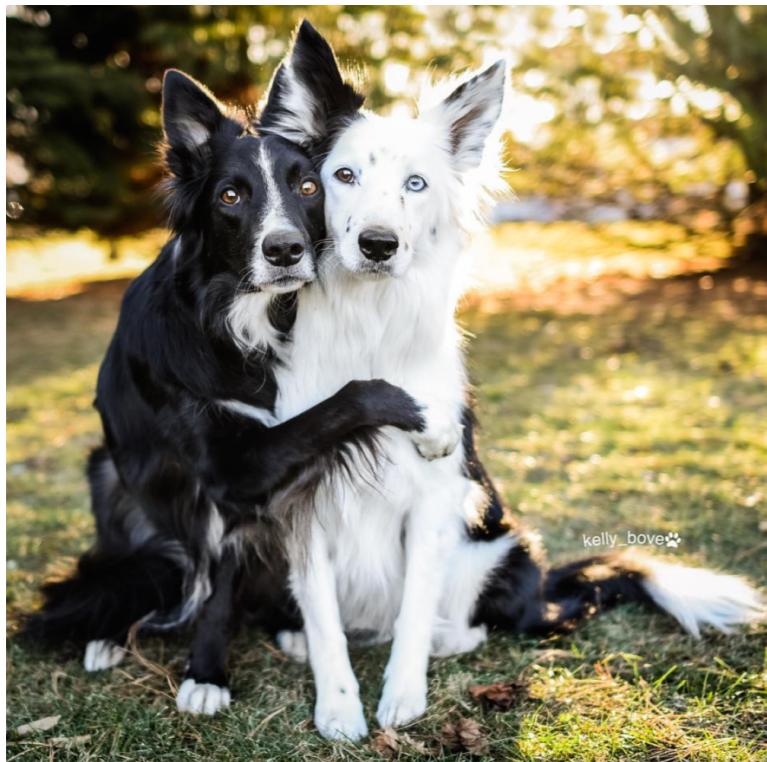
Use 38x38 feature map : **+2.5 mAP**  
(conv4\_3)

# Why So Many Default Boxes?

	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512

# Why So Many Default Boxes?

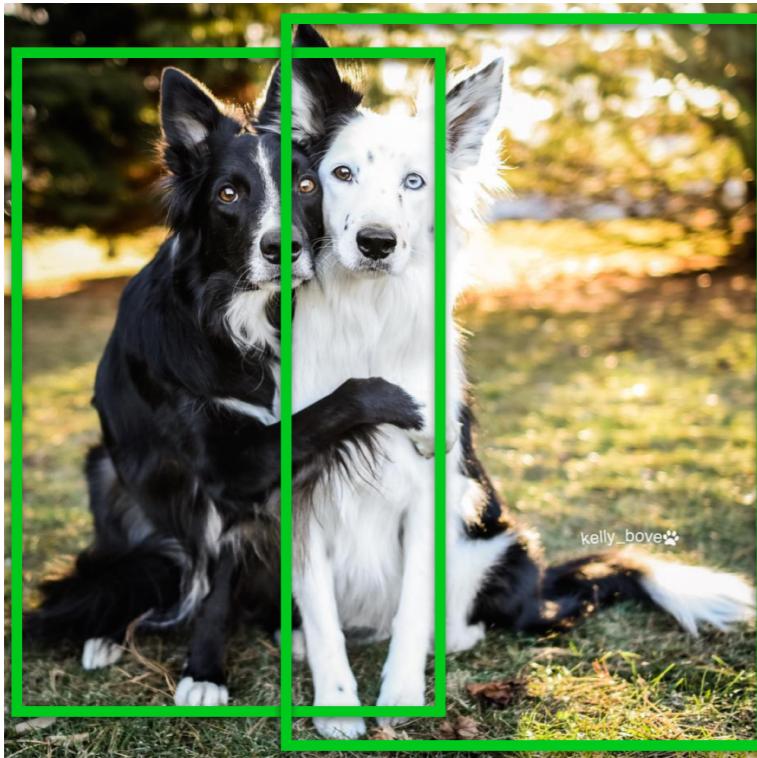
	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512



# Why So Many Default Boxes?

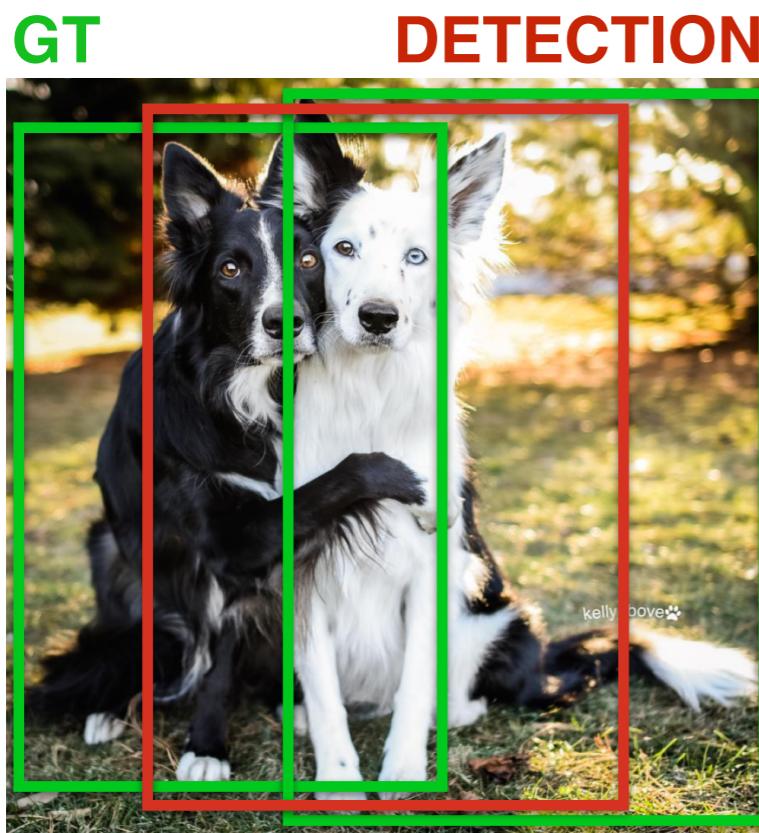
	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512

GT



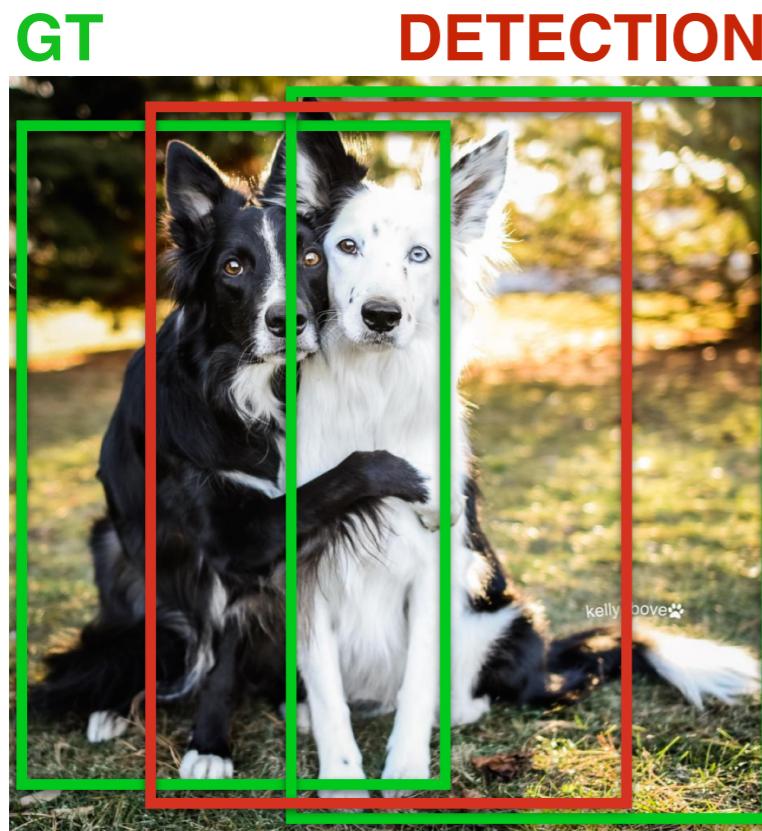
# Why So Many Default Boxes?

	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512



# Why So Many Default Boxes?

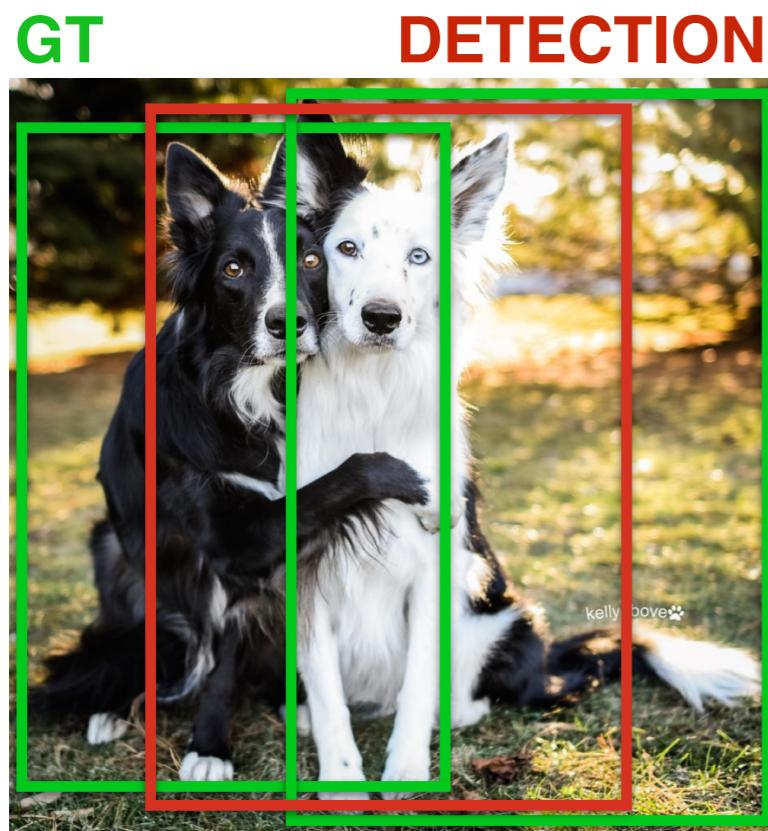
	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512



- SmoothL1 or L2 loss for box shape averages among likely hypotheses

# Why So Many Default Boxes?

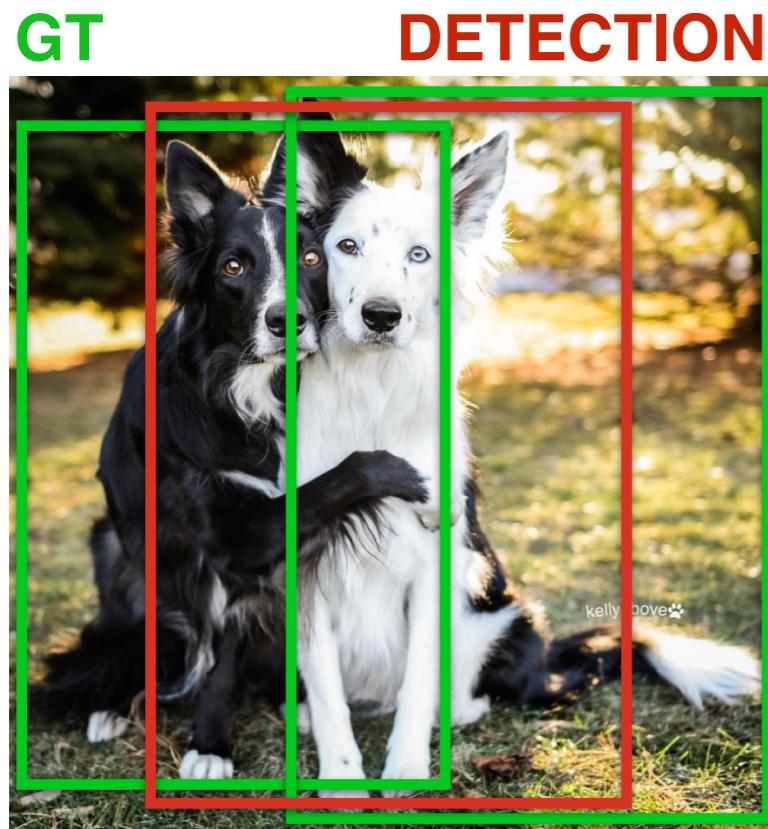
	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512



- SmoothL1 or L2 loss for box shape averages among likely hypotheses
- Need to have enough default boxes (discrete bins) to do accurate regression in each

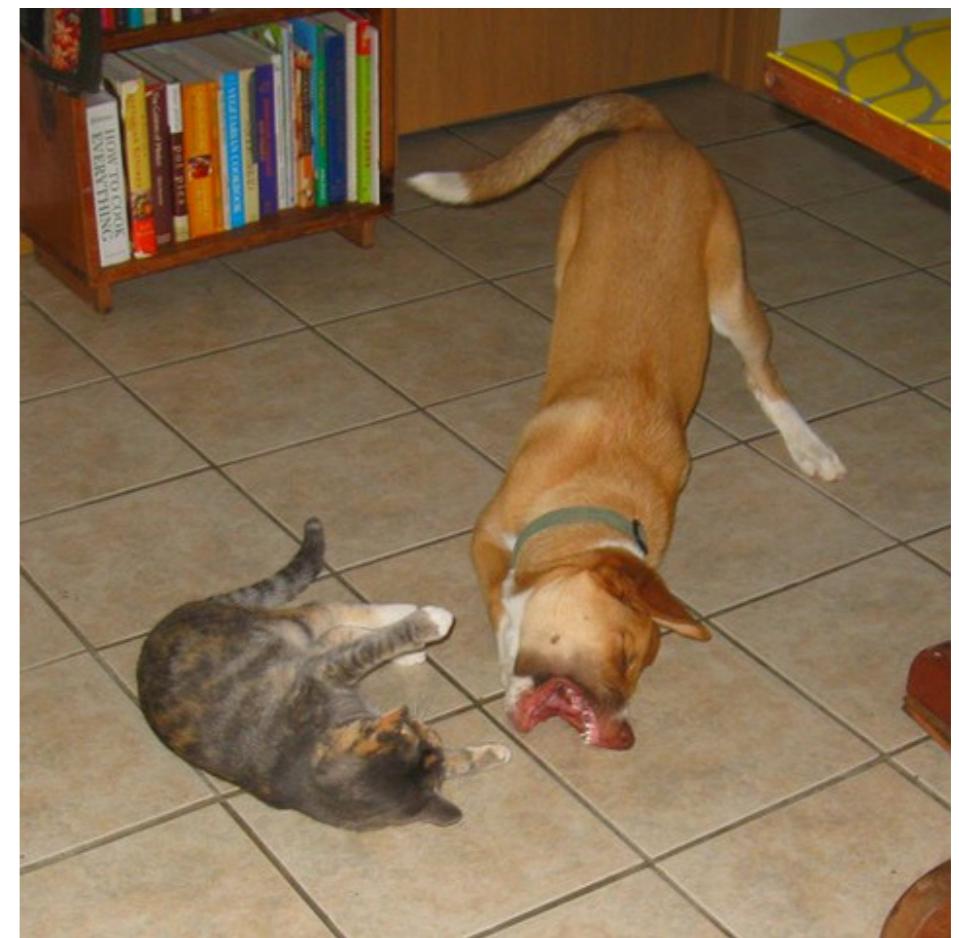
# Why So Many Default Boxes?

	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512



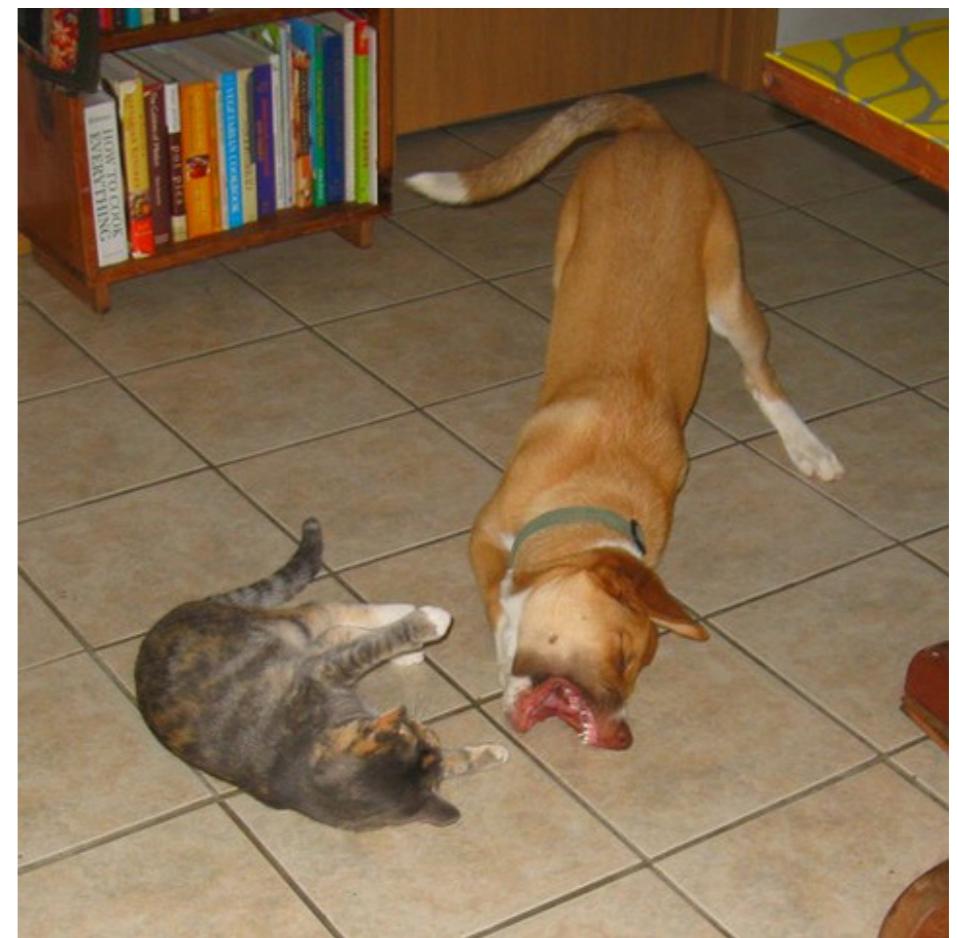
- SmoothL1 or L2 loss for box shape averages among likely hypotheses
- Need to have enough default boxes (discrete bins) to do accurate regression in each
- General principle for regressing complex continuous outputs with deep nets

# Handling Many Default Boxes



# Handling Many Default Boxes

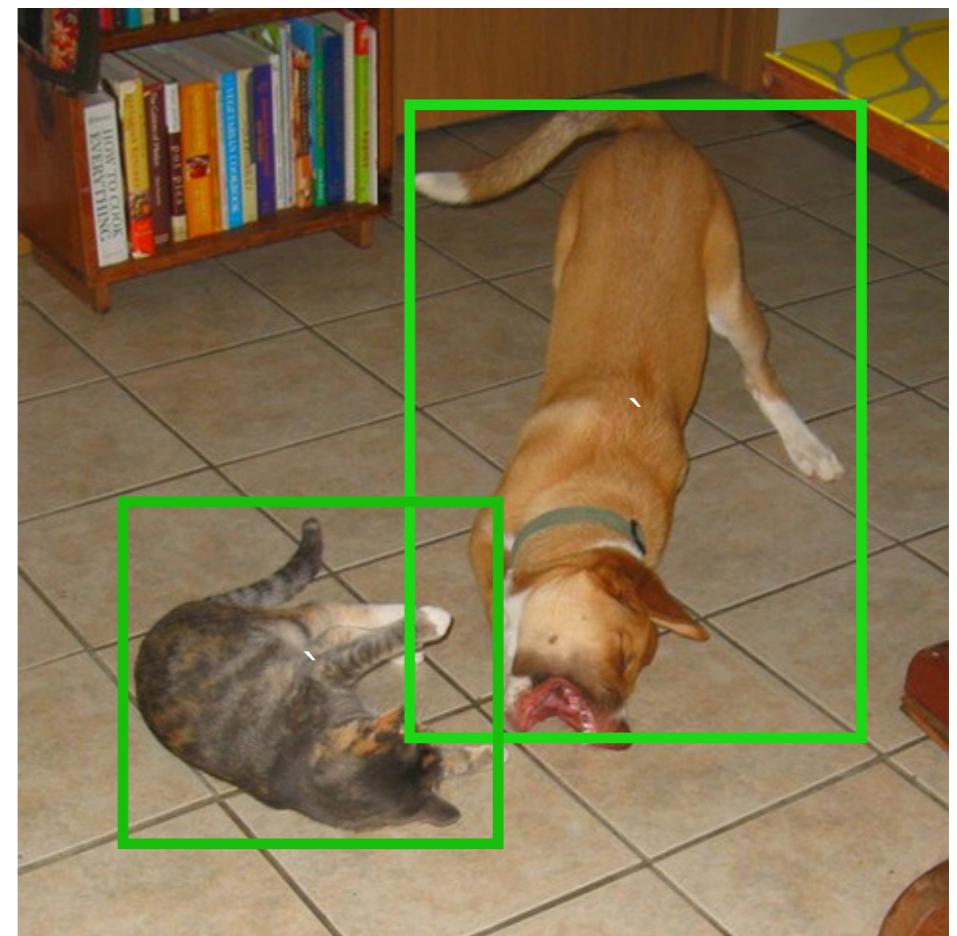
- Matching ground truth and default boxes



# Handling Many Default Boxes

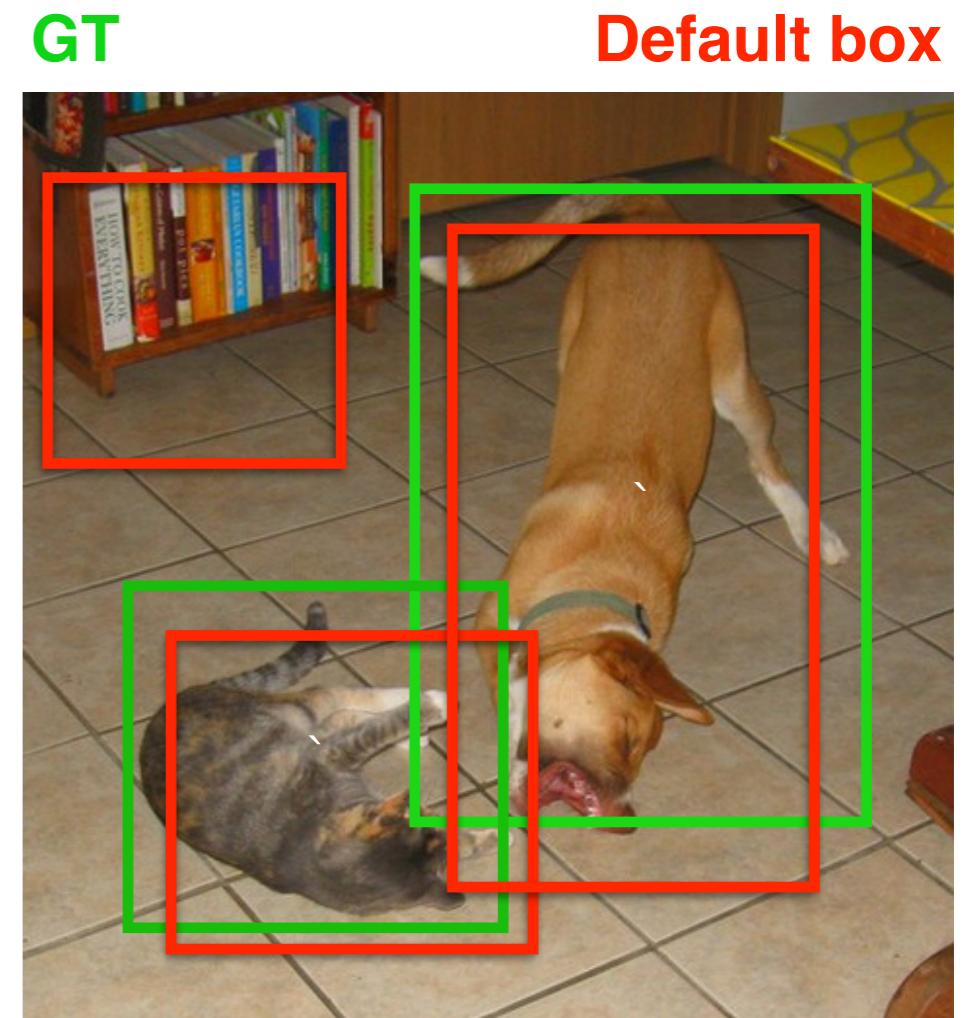
- Matching ground truth and default boxes

GT



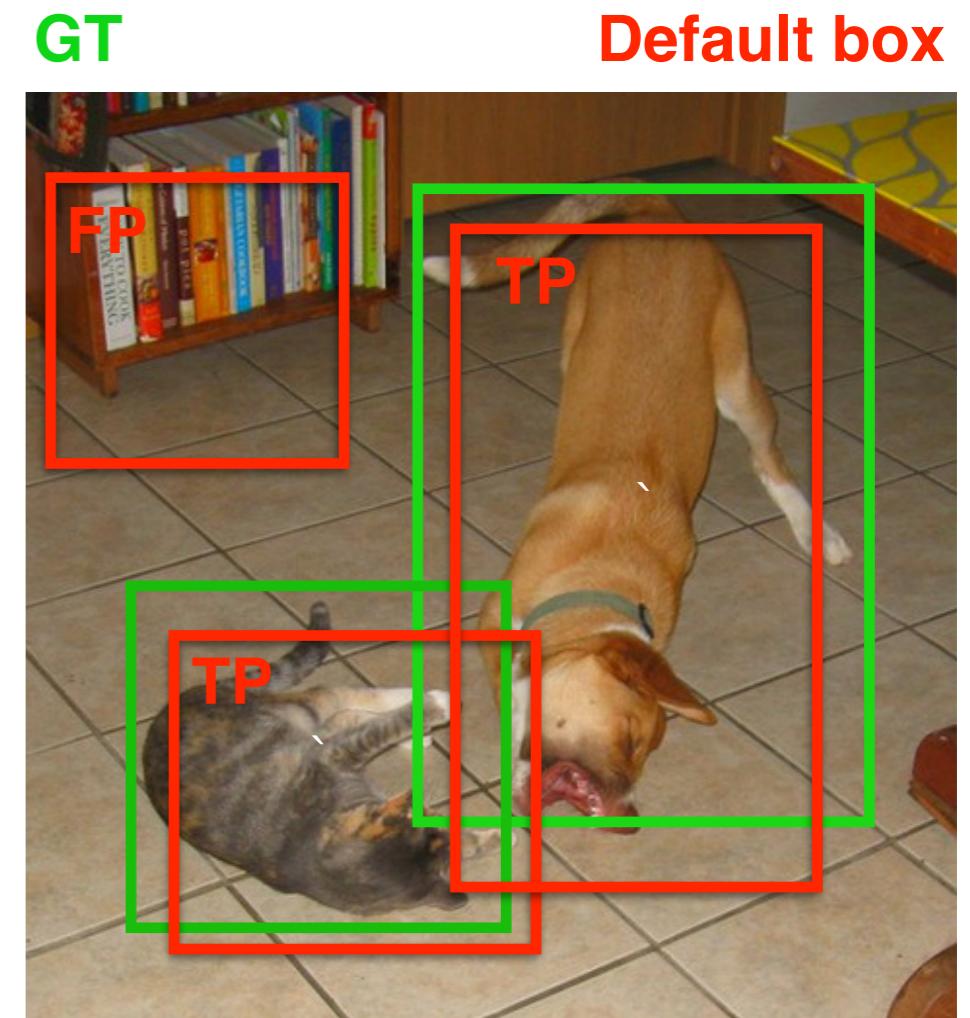
# Handling Many Default Boxes

- Matching ground truth and default boxes



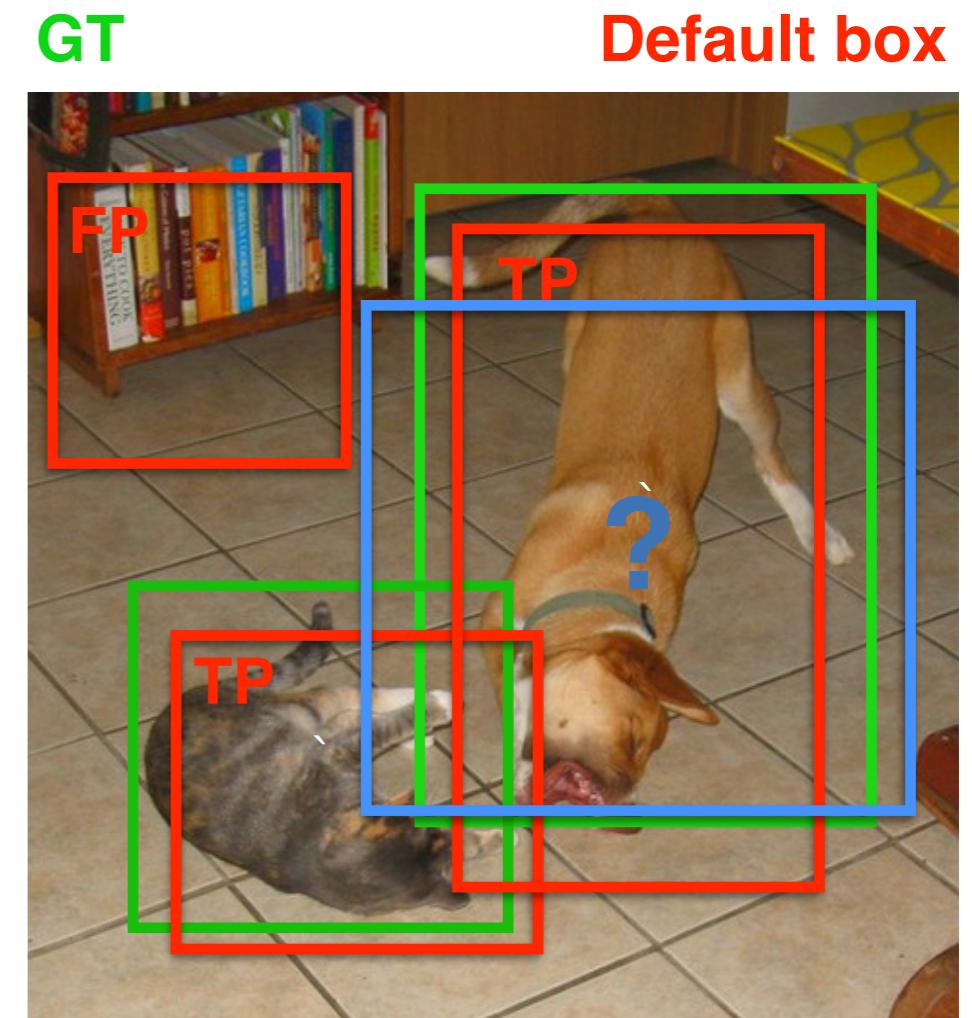
# Handling Many Default Boxes

- Matching ground truth and default boxes



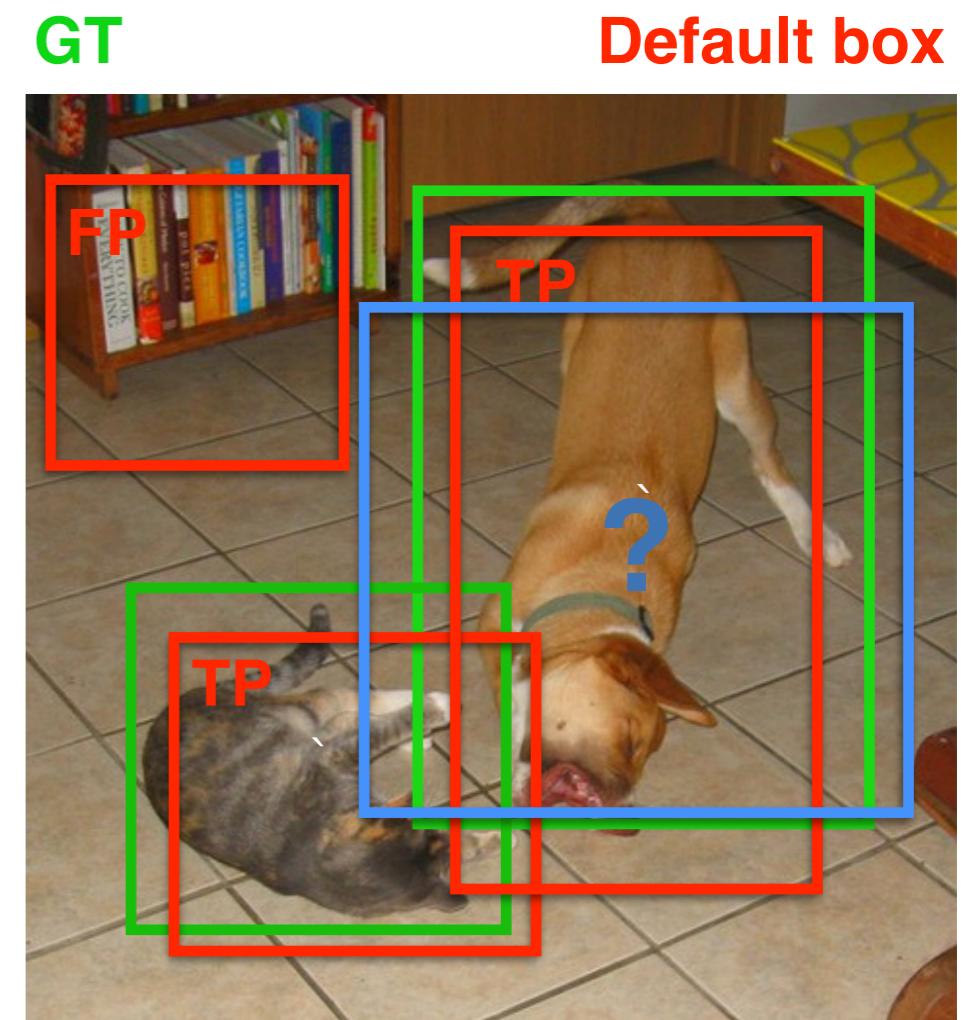
# Handling Many Default Boxes

- Matching ground truth and default boxes



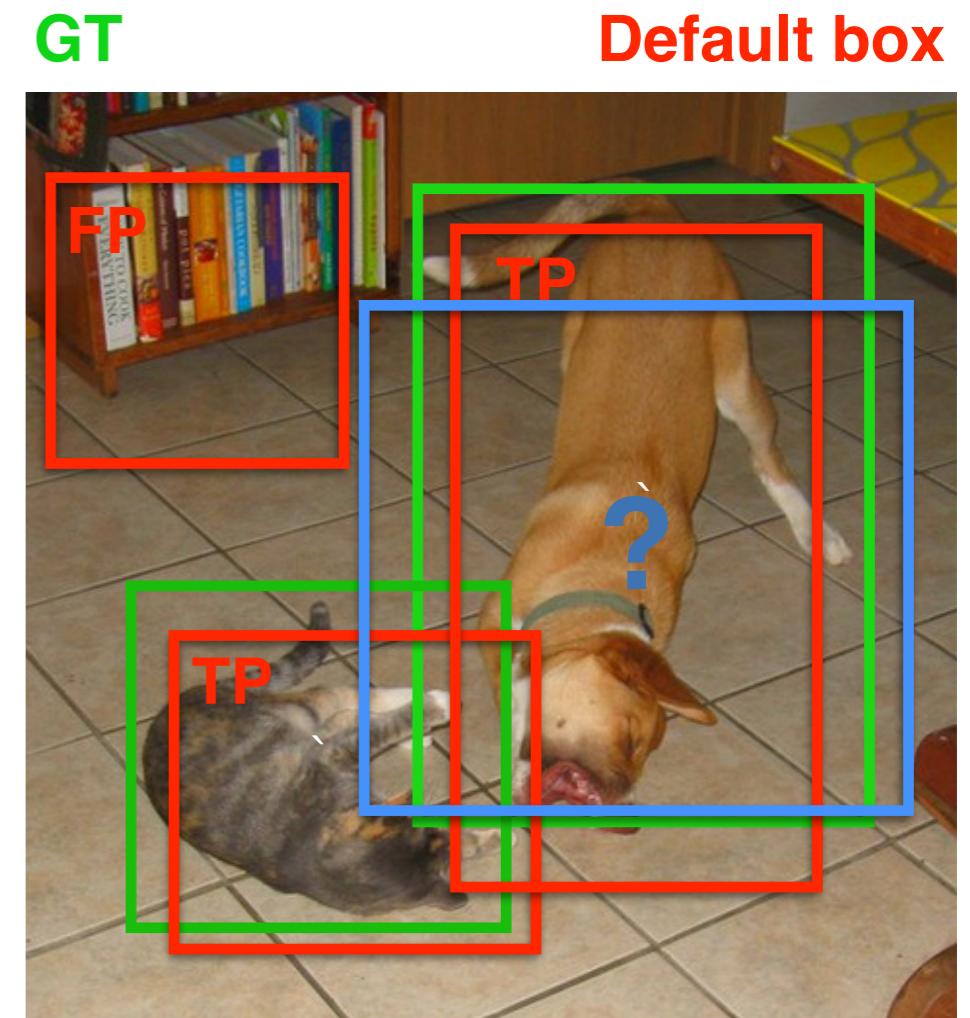
# Handling Many Default Boxes

- Matching ground truth and default boxes
  - Match each GT box to closest default box



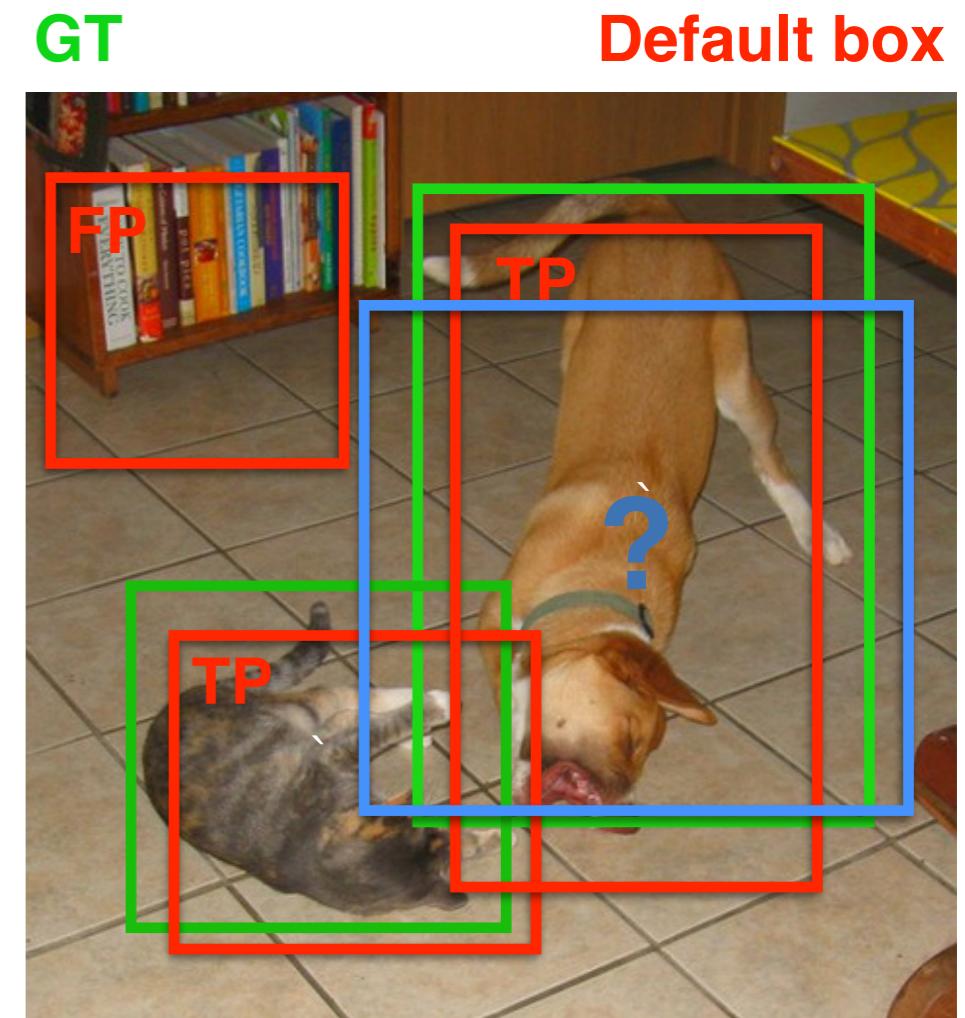
# Handling Many Default Boxes

- Matching ground truth and default boxes
  - Match each GT box to closest default box
  - Also match each GT box to all unassigned default boxes with  $\text{IoU} > 0.5$



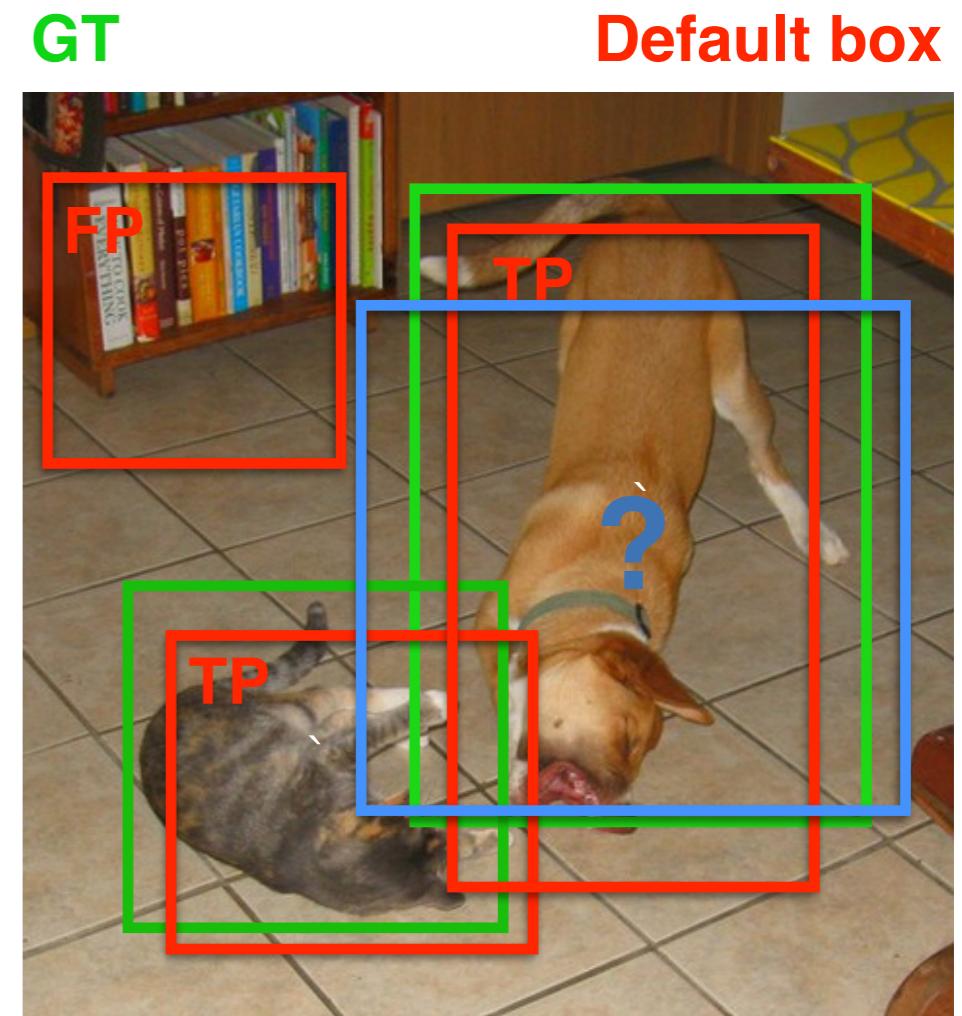
# Handling Many Default Boxes

- Matching ground truth and default boxes
  - Match each GT box to closest default box
  - Also match each GT box to all unassigned default boxes with  $\text{IoU} > 0.5$
- Hard negative mining



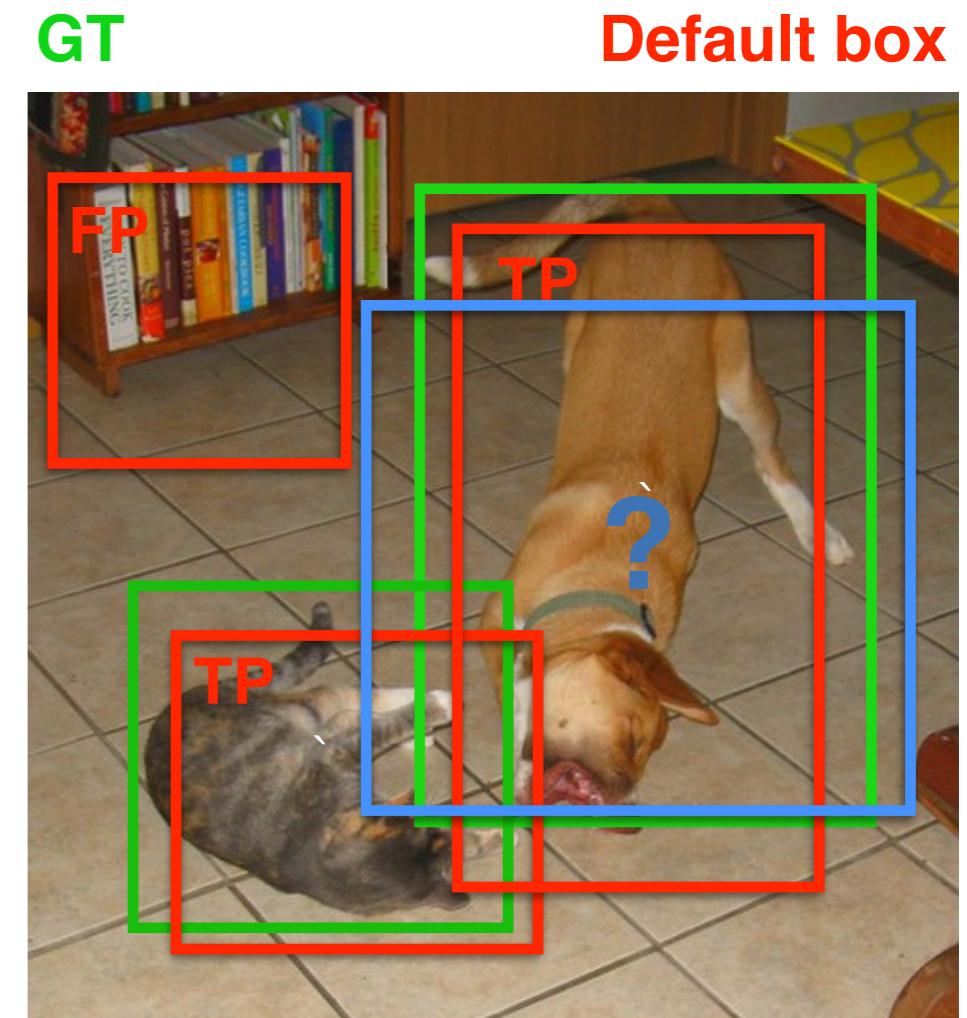
# Handling Many Default Boxes

- Matching ground truth and default boxes
  - Match each GT box to closest default box
  - Also match each GT box to all unassigned default boxes with  $\text{IoU} > 0.5$
- Hard negative mining
  - Unbalanced training: 1-30 TP, 8k-25k FP

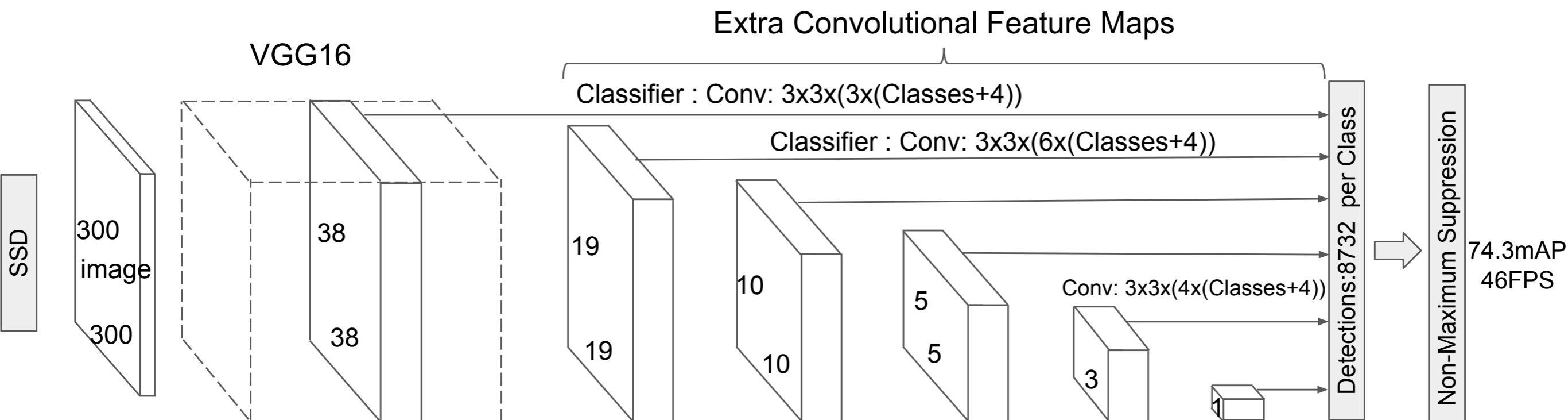


# Handling Many Default Boxes

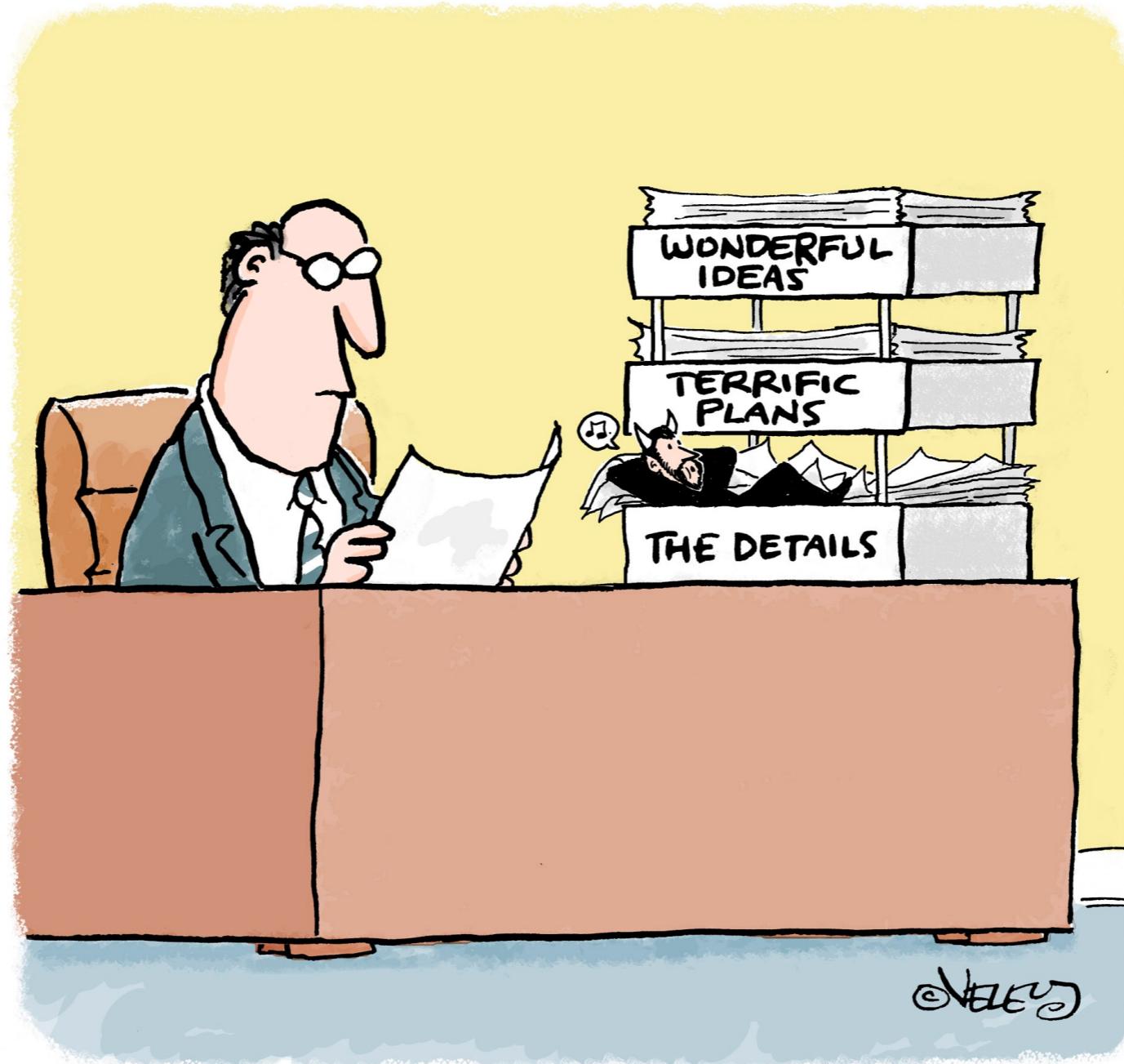
- Matching ground truth and default boxes
  - Match each GT box to closest default box
  - Also match each GT box to all unassigned default boxes with  $\text{IoU} > 0.5$
- Hard negative mining
  - Unbalanced training: 1-30 TP, 8k-25k FP
  - Keep TP:FP ratio fixed (1:3), use worst-misclassified FPs.



# SSD Architecture

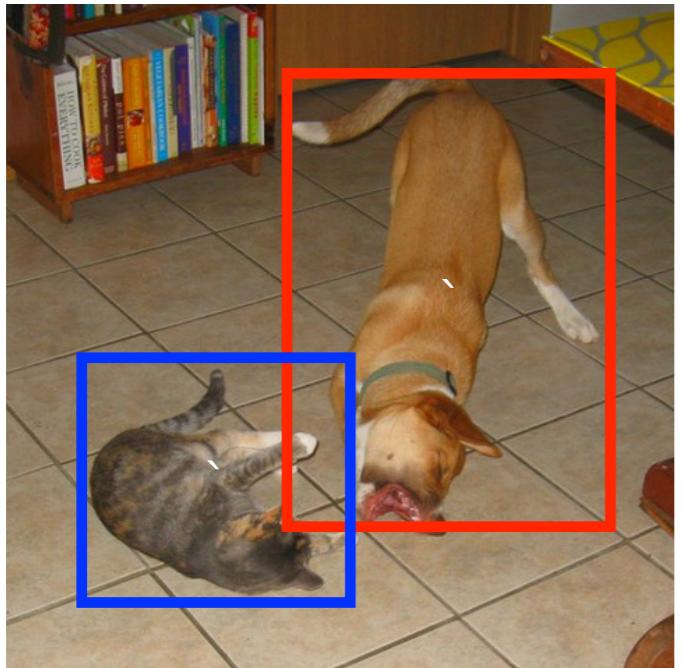


# Contribution #3: The Devil is in the Details

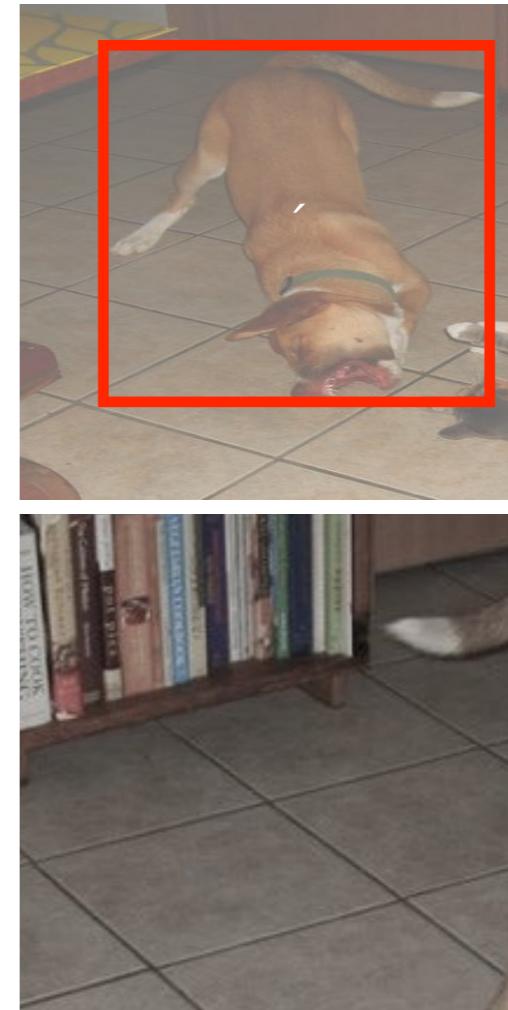
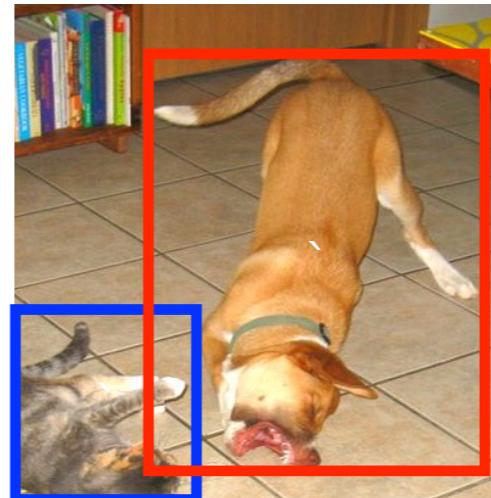
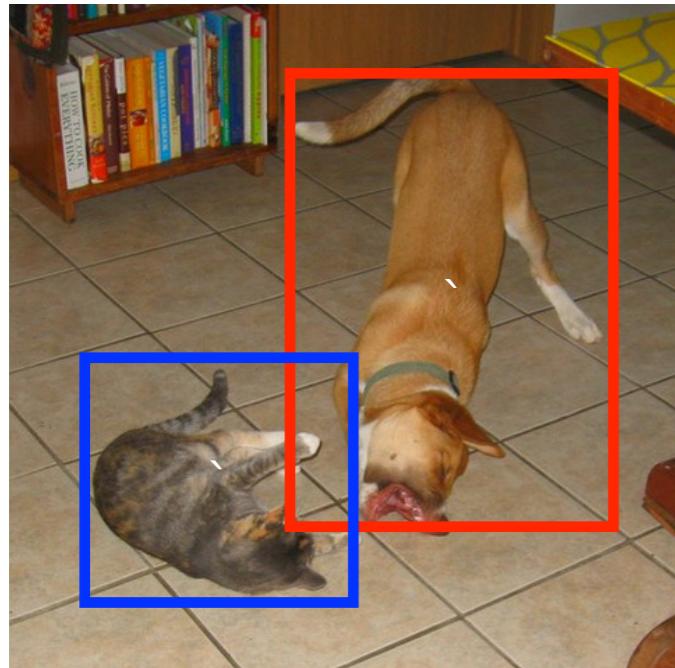


# Data Augmentation

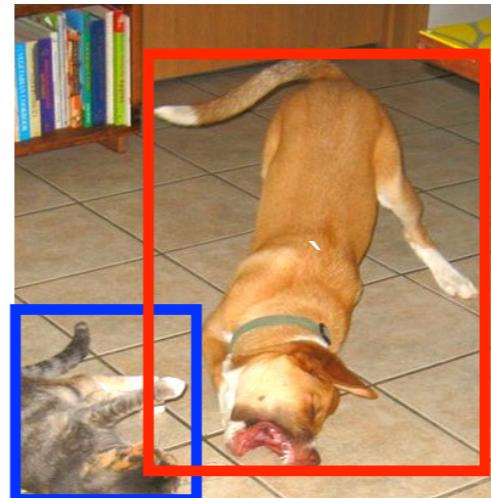
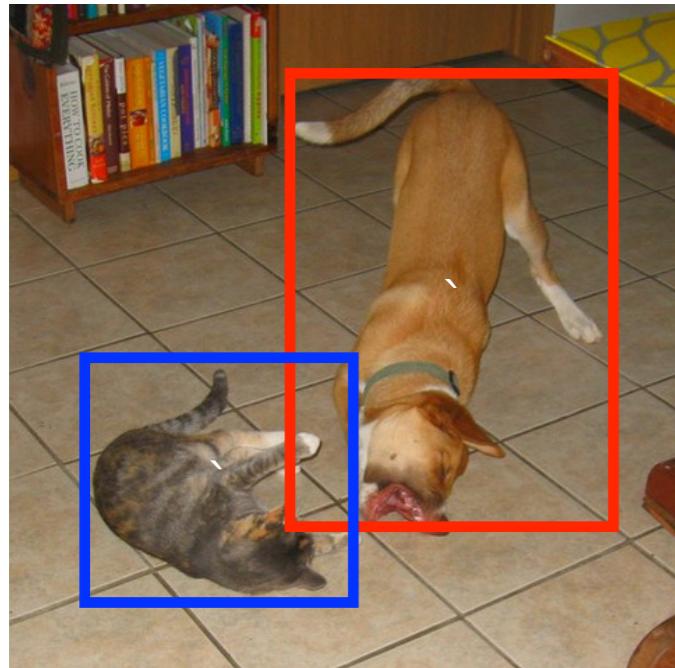
# Data Augmentation



# Data Augmentation



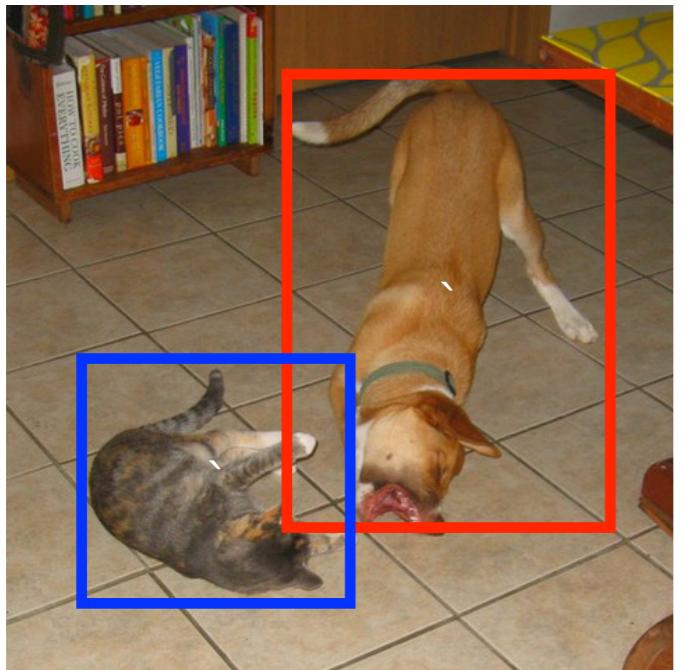
# Data Augmentation



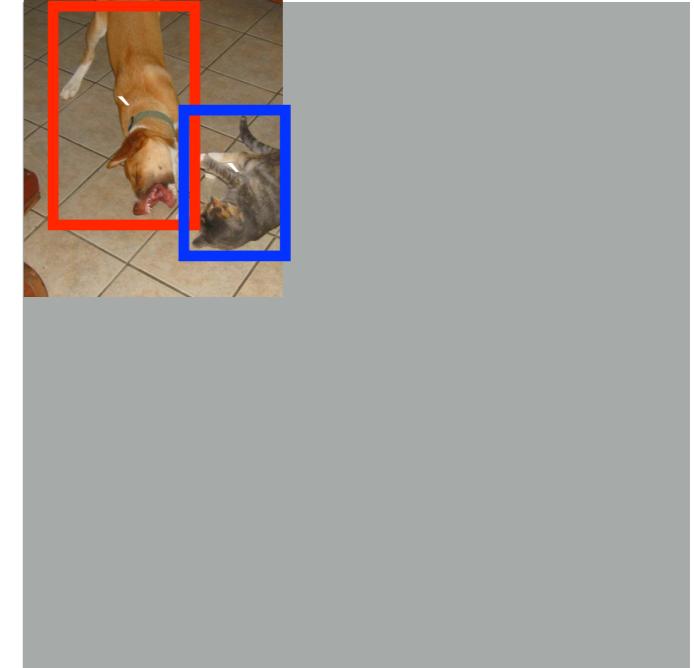
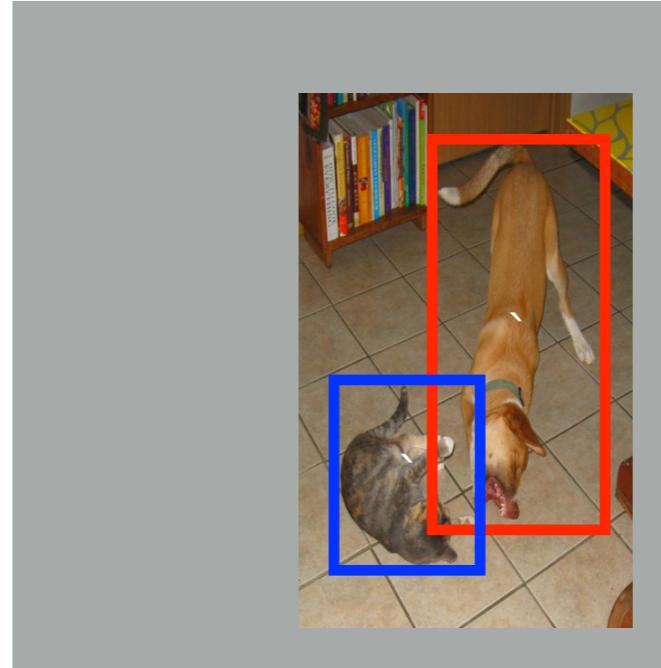
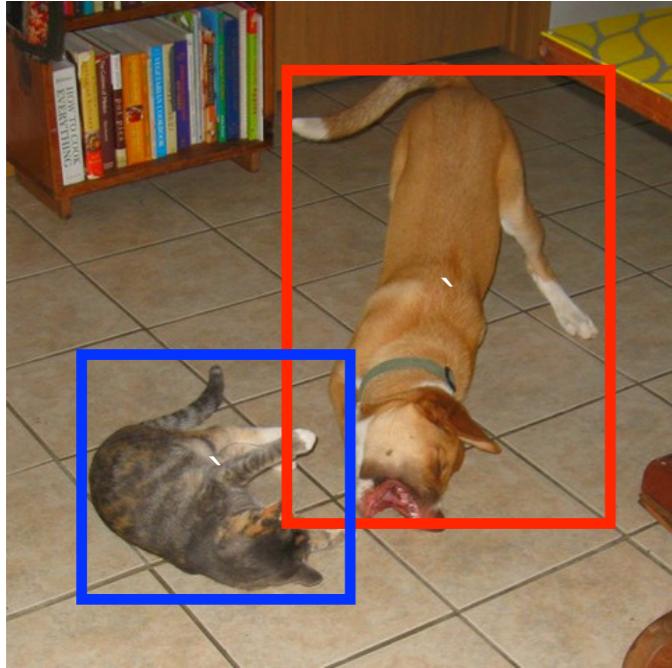
data augmentation	SSD300	
horizontal flip	✓	✓
random crop & color distortion		✓
VOC2007 test mAP	65.5	<b>74.3</b>

# Data Augmentation

# Data Augmentation

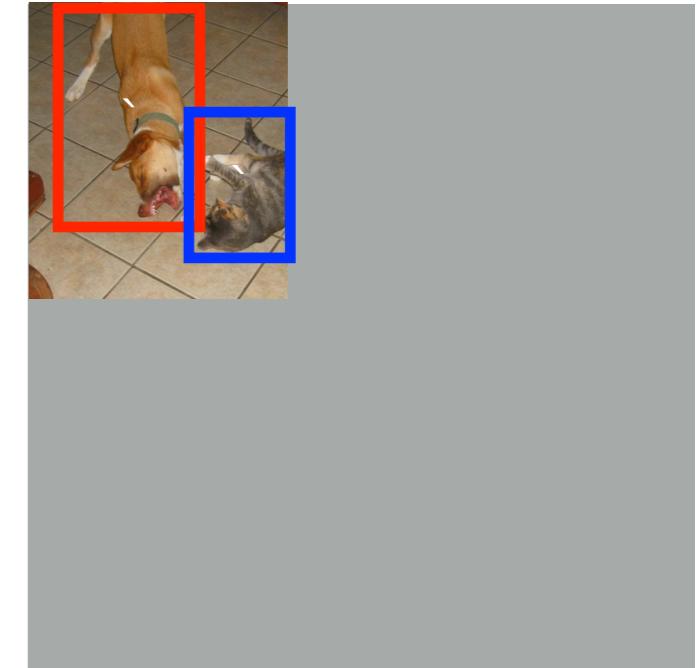
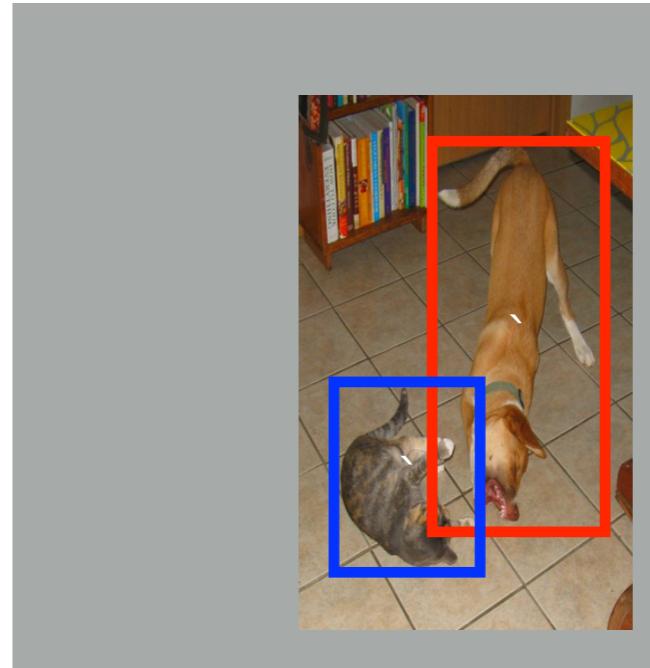
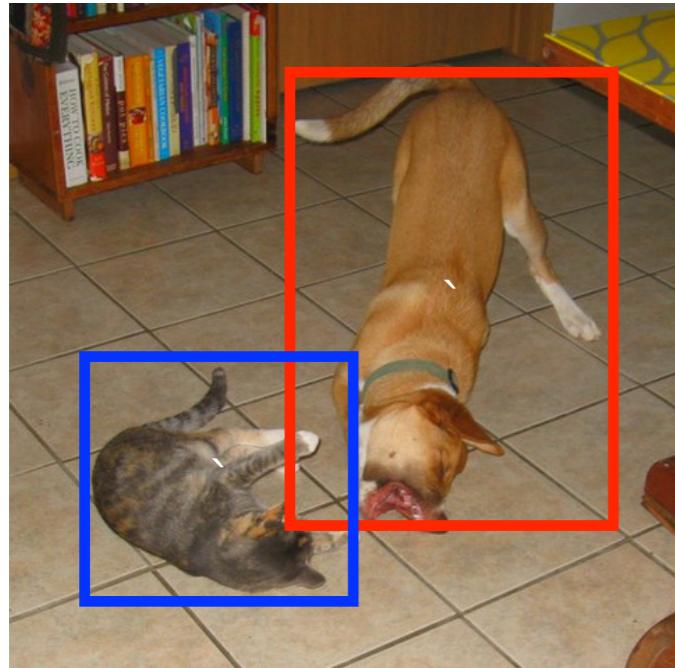


# Data Augmentation



Random expansion creates more  
**small** training examples

# Data Augmentation



Random expansion creates more  
**small** training examples

data augmentation	SSD300		
horizontal flip	✓	✓	✓
random crop & color distortion		✓	✓
random expansion			✓
VOC2007 test mAP	65.5	74.3	<b>77.2</b>

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

**6.6x↑**

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

**10%↑**

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	74.3	46	1	8732	300 × 300
SSD512	76.8	19	1	24564	512 × 512
SSD300	74.3	59	8	8732	300 × 300
SSD512	76.8	22	8	24564	512 × 512

# Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	~ 1000 × 600
Fast YOLO	52.7	155	1	98	448 × 448
YOLO (VGG16)	66.4	21	1	98	448 × 448
SSD300	77.2	74.3	1	8732	300 × 300
SSD512	79.8	76.8	1	24564	512 × 512
SSD300	77.2	74.3	8	8732	300 × 300
SSD512	79.8	76.8	8	24564	512 × 512

# Results on More Datasets

# Results on More Datasets

Method	VOC2007	VOC2012	MS COCO	ILSVRC2014
	test	test	test-dev	val2
Fast R-CNN	70.0	68.4	19.7	N/A
Faster R-CNN	<b>73.2</b>	<b>70.4</b>	<b>21.9</b>	N/A
YOLO	63.4	57.9	N/A	N/A

# Results on More Datasets

Method	VOC2007 test	VOC2012 test	MS COCO test-dev	ILSVRC2014 val2
Fast R-CNN	70.0	68.4	19.7	N/A
Faster R-CNN	73.2	70.4	21.9	N/A
YOLO	63.4	57.9	N/A	N/A
<b>SSD300</b>	<b>74.3</b>	<b>72.4</b>	<b>23.2</b>	<b>43.4</b>

# Results on More Datasets

Method	VOC2007 test	VOC2012 test	MS COCO test-dev	ILSVRC2014 val2
Fast R-CNN	70.0	68.4	19.7	N/A
Faster R-CNN	73.2	70.4	21.9	N/A
YOLO	63.4	57.9	N/A	N/A
SSD300	74.3	72.4	23.2	43.4
SSD512	<b>76.8</b>	<b>74.9</b>	<b>26.8</b>	<b>46.4</b>

# Results on More Datasets

Method	VOC2007 test	VOC2012 test	MS COCO test-dev	ILSVRC2014 val2
Fast R-CNN	70.0	68.4	19.7	N/A
Faster R-CNN	73.2	70.4	21.9	N/A
YOLO	63.4	57.9	N/A	N/A
SSD300*	77.2	75.8	25.1	N/A
SSD512*	<b>79.8</b>	<b>78.5</b>	<b>28.8</b>	N/A

# COCO Bounding Box precision

# COCO Bounding Box precision

mAP @ IoU	0.5	0.75	0.5:0.95
Faster R-CNN	45.3	23.5	24.2
SSD512*	48.5	30.3	28.8
gain	+3.2	+6.8	+4.6

# Future Work

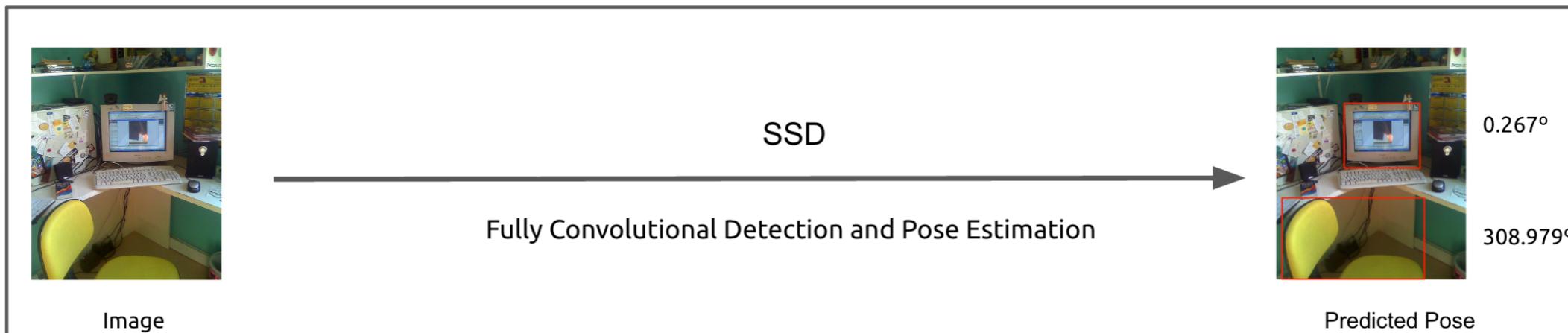
# Future Work

- Object detection + pose estimation

# Future Work

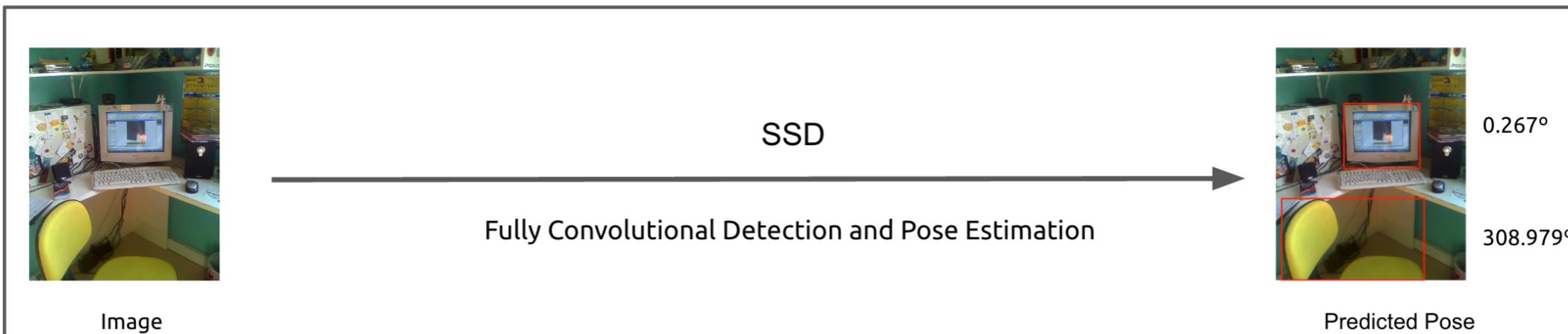
- Object detection + pose estimation

[Poirson et al, coming out at 3DV, 2016]



# Future Work

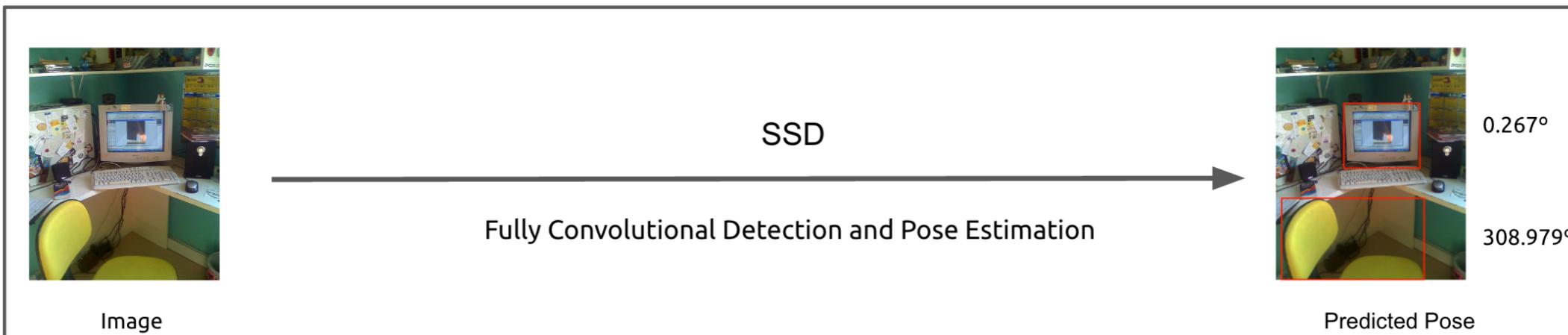
- Object detection + pose estimation  
[Poirson et al, coming out at 3DV, 2016]



- Single shot 3D bounding box detection

# Future Work

- Object detection + pose estimation  
[Poirson et al, coming out at 3DV, 2016]



- Single shot 3D bounding box detection
- Joint object detection + tracking model

# Check out the code/models



<https://github.com/weiliu89/caffe/tree/ssd>

**Thank you!**  
Come by our poster O-1A-02