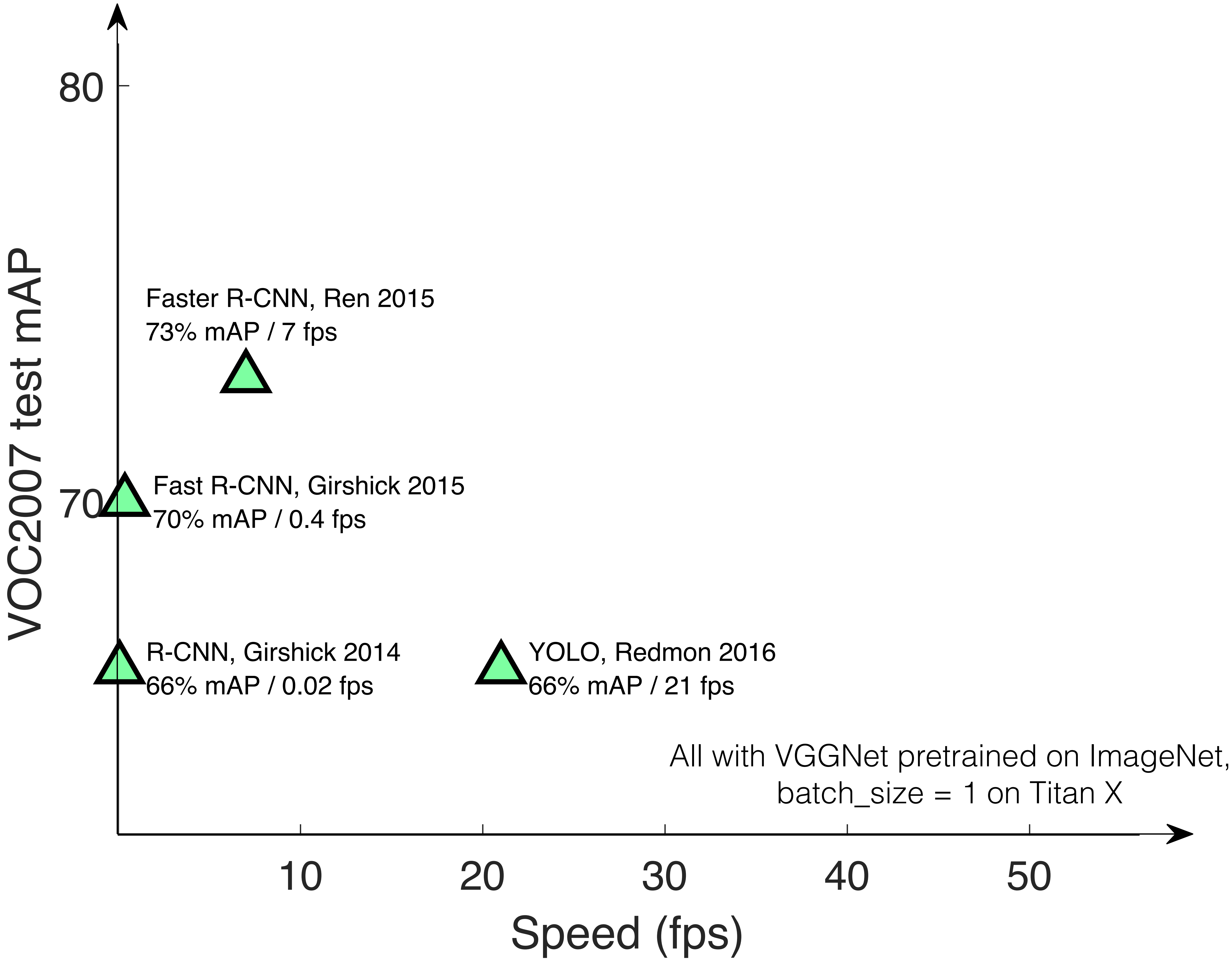
SSD: Single Shot MultiBox Detector

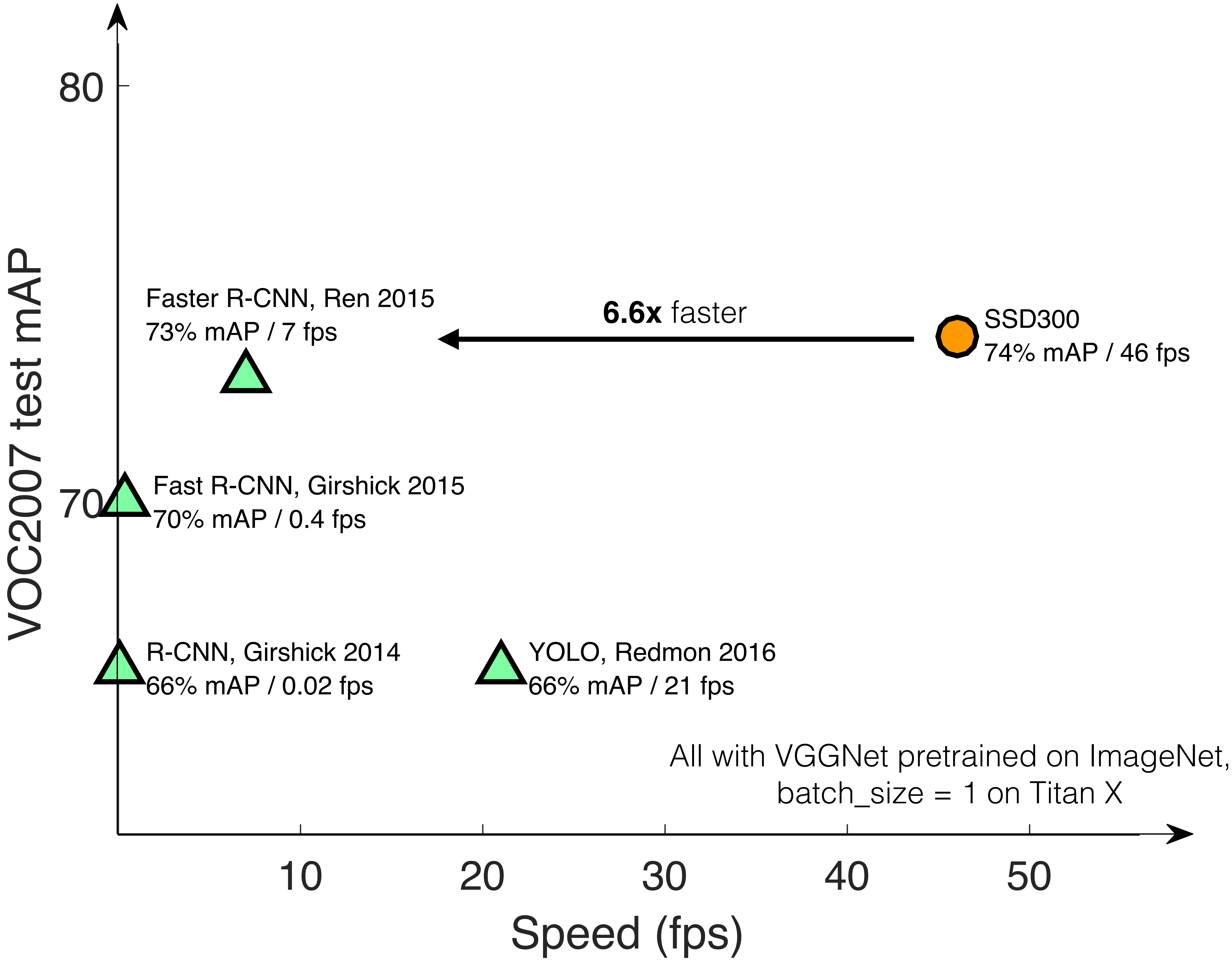
Wei Liu(1), **Dragomir Anguelov(2)**, Dumitru Erhan(3), Christian Szegedy(3),
Scott Reed(4), Cheng-Yang Fu(1), Alexander C. Berg(1)

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

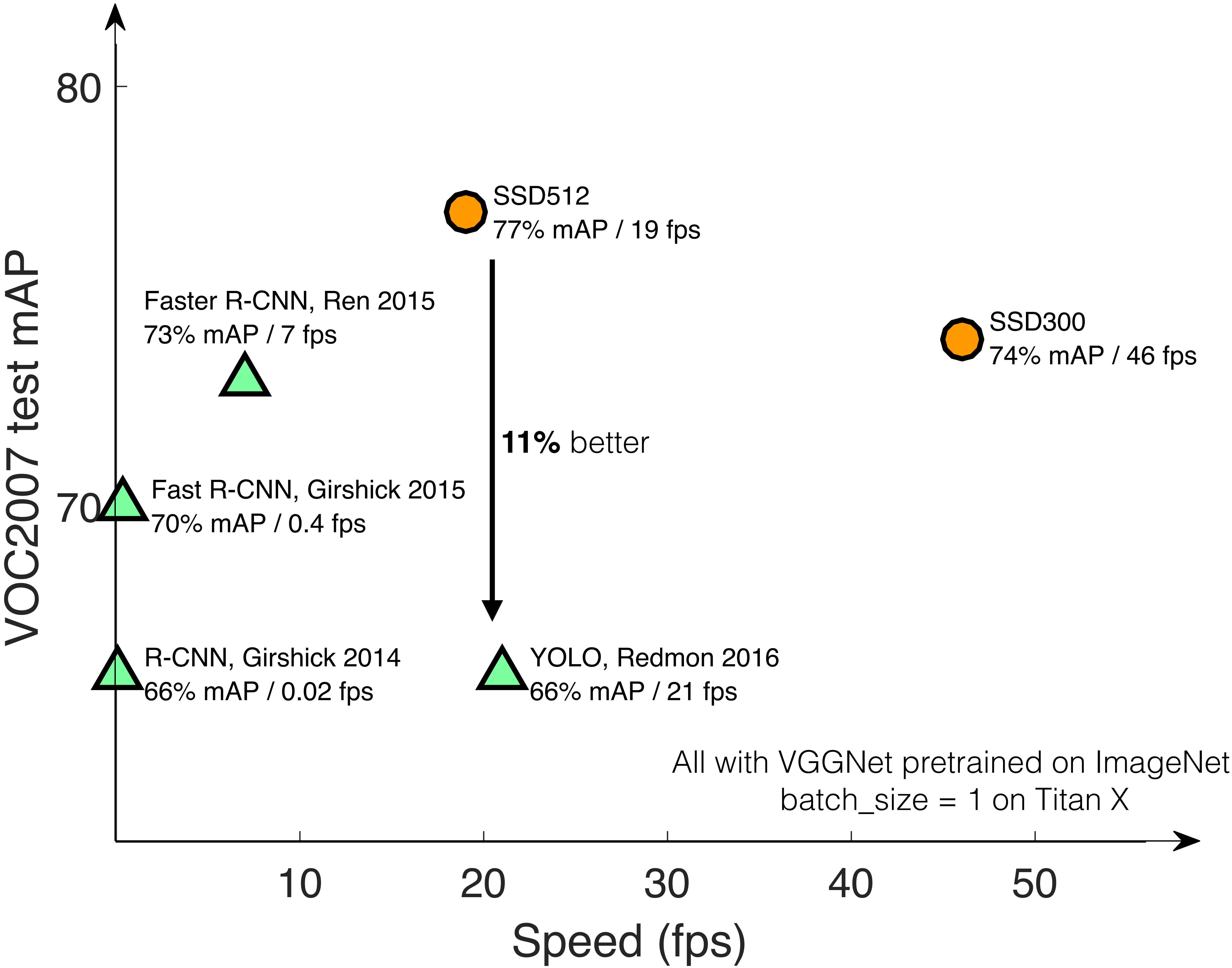
VGGNet
Titan X Pascal

VGGNet
Titan X Pascal

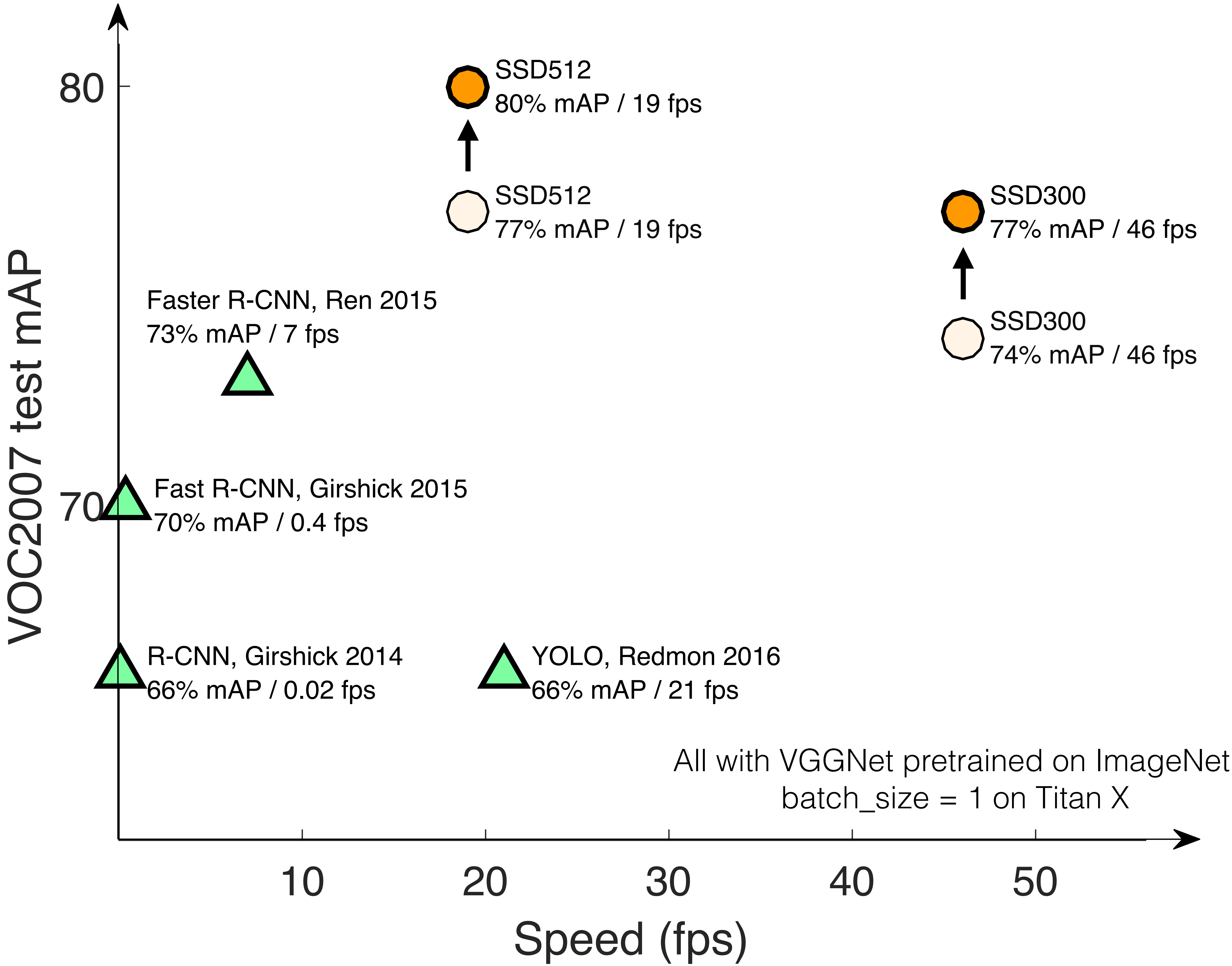

VOC2007 test mAP
80
70
Faster R-CNN, Ren 2015
73% mAP / 7 fps
Fast R-CNN, Girshick 2015
70% mAP / 0.4 fps
R-CNN, Girshick 2014
66% mAP / 0.02 fps
YOLO, Redmon 2016
66% mAP / 21 fps
All with VGGNet pretrained on ImageNet,
batch_size = 1 on Titan X
10
20
30
40
50
Speed (fps)



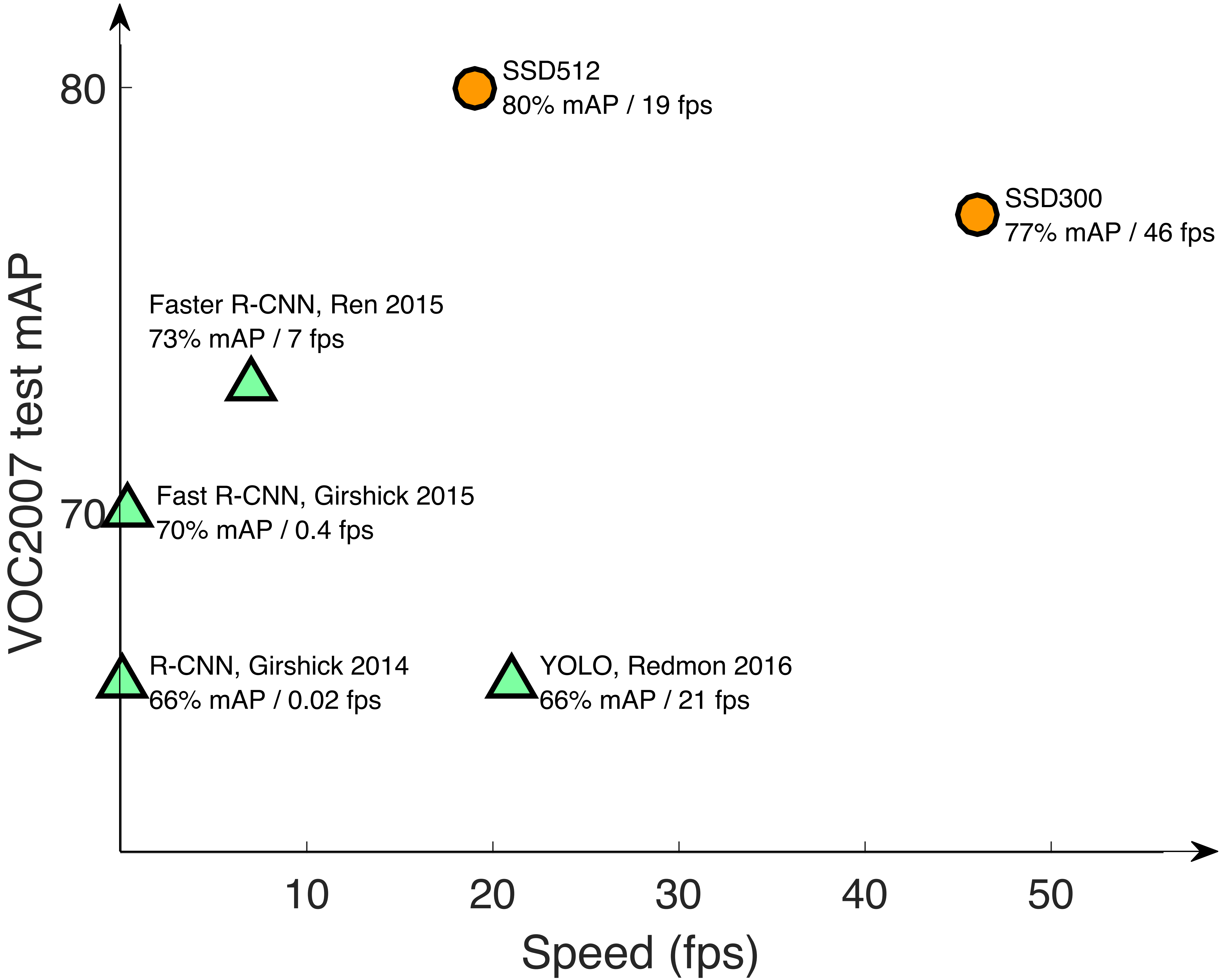
VOC2007 test mAP
80
70
Faster R-CNN, Ren 2015
73% mAP / 7 fps
6.6x faster
SSD300
74% mAP / 46 fps
Fast R-CNN, Girshick 2015
70% mAP / 0.4 fps
R-CNN, Girshick 2014
66% mAP / 0.02 fps
YOLO, Redmon 2016
66% mAP / 21 fps
All with VGGNet pretrained on ImageNet,
batch_size = 1 on Titan X
10
20
30
40
50
Speed (fps)



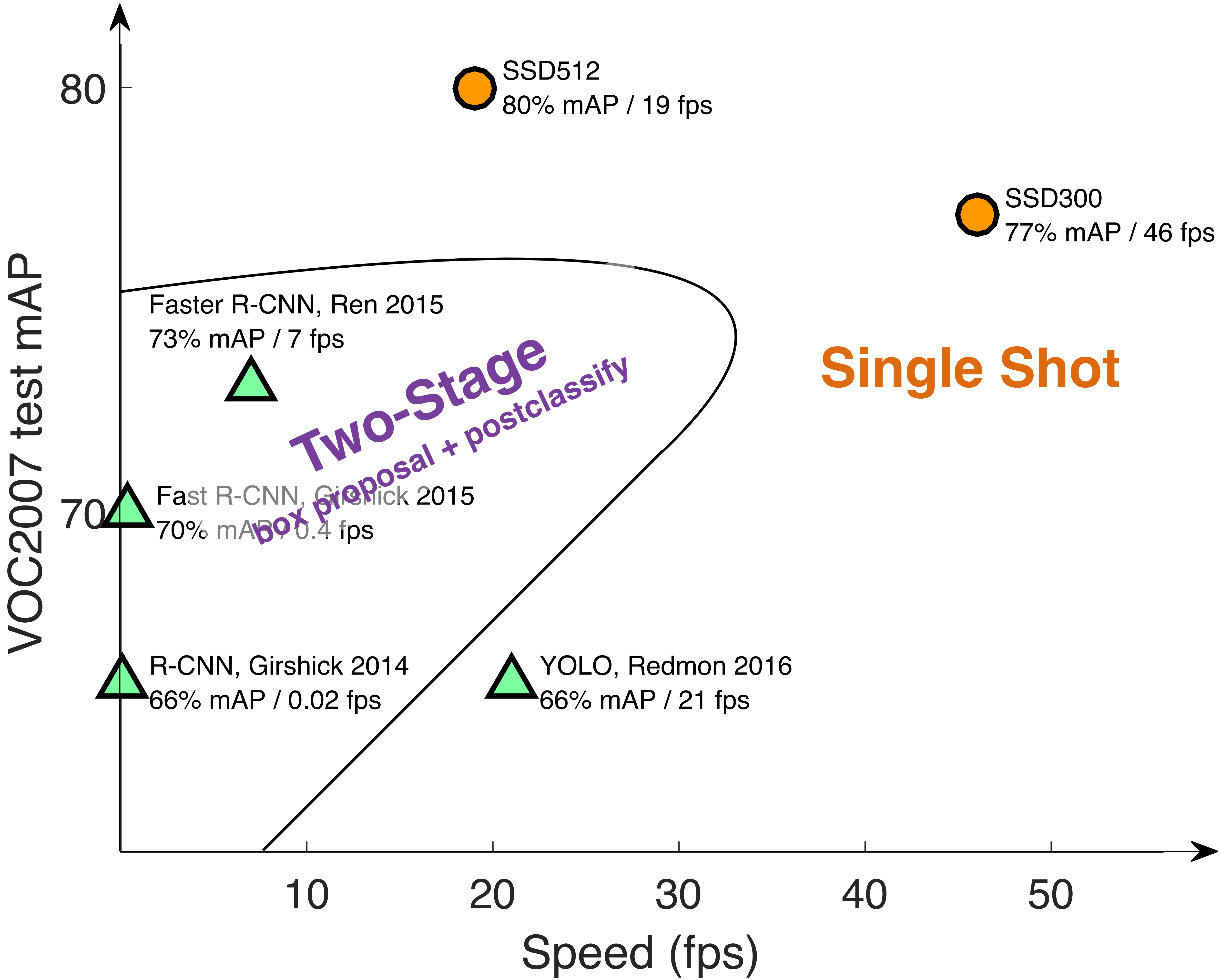
VOC2007 test mAP
80
70
SSD512
77% mAP / 19 fps
Faster R-CNN, Ren 2015
73% mAP / 7 fps
SSD300
74% mAP / 46 fps
11% better
Fast R-CNN, Girshick 2015
70% mAP / 0.4 fps
R-CNN, Girshick 2014
66% mAP / 0.02 fps
YOLO, Redmon 2016
66% mAP / 21 fps
All with VGGNet pretrained on ImageNet,
batch_size = 1 on Titan X
10
20
30
40
50
Speed (fps)



VOC2007 test mAP
80
70
SSD512
80% mAP / 19 fps
SSD512
77% mAP / 19 fps
SSD300
77% mAP / 46 fps
SSD300
74% mAP / 46 fps
Faster R-CNN, Ren 2015
73% mAP / 7 fps
Fast R-CNN, Girshick 2015
70% mAP / 0.4 fps
R-CNN, Girshick 2014
66% mAP / 0.02 fps
YOLO, Redmon 2016
66% mAP / 21 fps
All with VGGNet pretrained on ImageNet,
batch_size = 1 on Titan X
10
20
30
40
50
Speed (fps)



VOC2007 test mAP
80
70
SSD512
80% mAP / 19 fps
SSD300
77% mAP / 46 fps
Faster R-CNN, Ren 2015
73% mAP / 7 fps
Fast R-CNN, Girshick 2015
70% mAP / 0.4 fps
R-CNN, Girshick 2014
66% mAP / 0.02 fps
YOLO, Redmon 2016
66% mAP / 21 fps
10
20
30
40
50
Speed (fps)

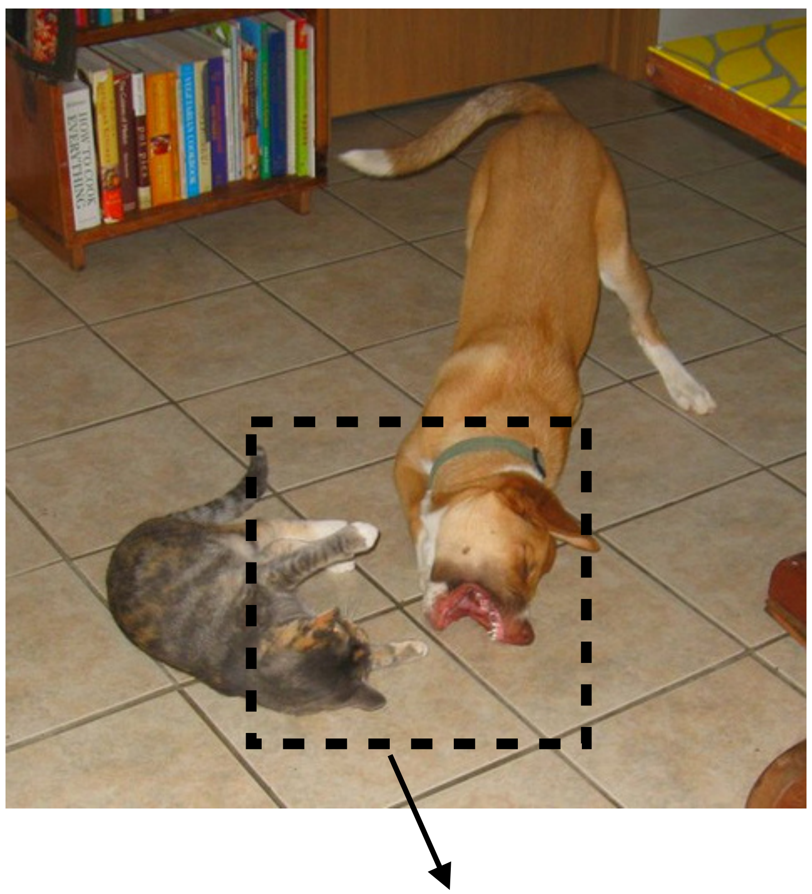

SSD512
80% mAP / 19 fps
SSD300
77% mAP / 46 fps
Faster R-CNN, Ren 2015
73% mAP / 7 fps
Two-Stage
box proposal + postclassify
Single Shot
Fast R-CNN, Girshick 2015
70% mAP / 0.4 fps
R-CNN, Girshick 2014
66% mAP / 0.02 fps
YOLO, Redmon 2016
66% mAP / 21 fps
VOC2007 test mAP
80
70
10
20
30
40
50
Speed (fps)

Bounding Box Prediction

Classical sliding windows

Bounding Box Prediction

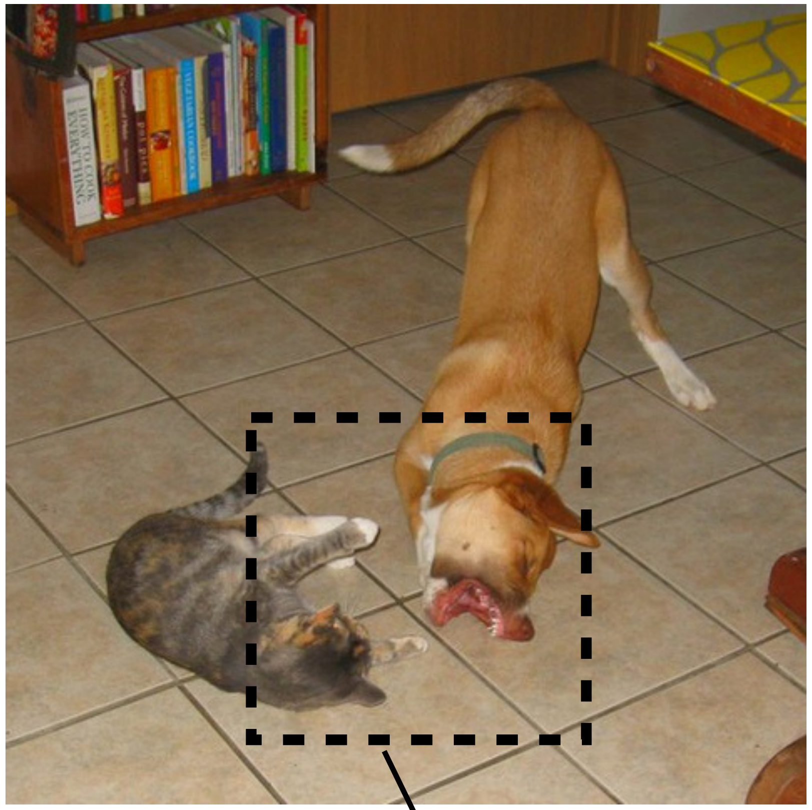
Classical sliding windows



Is it a cat? **No**

Bounding Box Prediction

Classical sliding windows

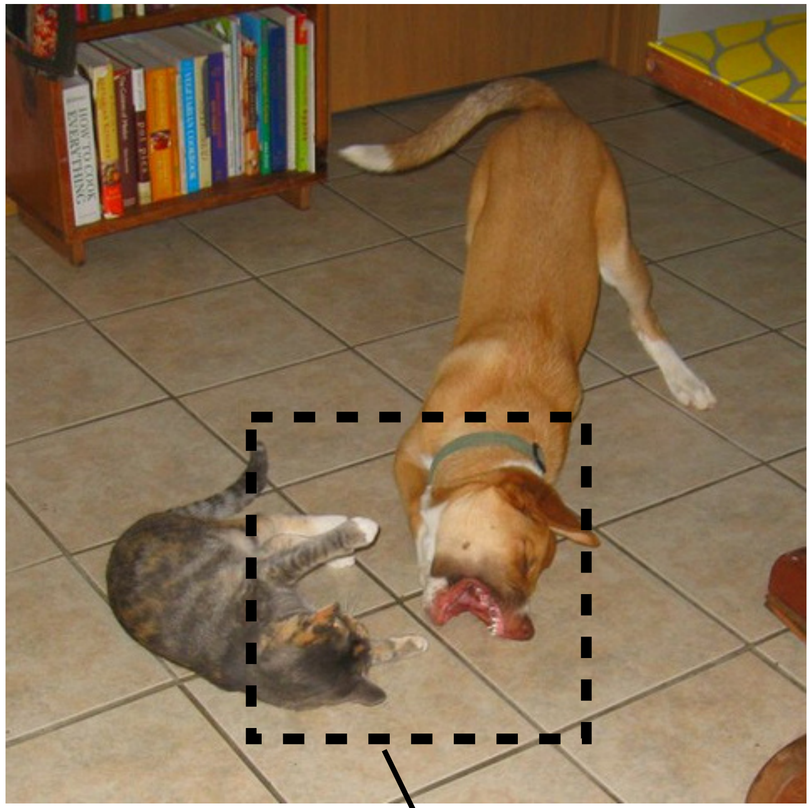


Is it a cat? **No**

Discretize the box space **densely**

Bounding Box Prediction

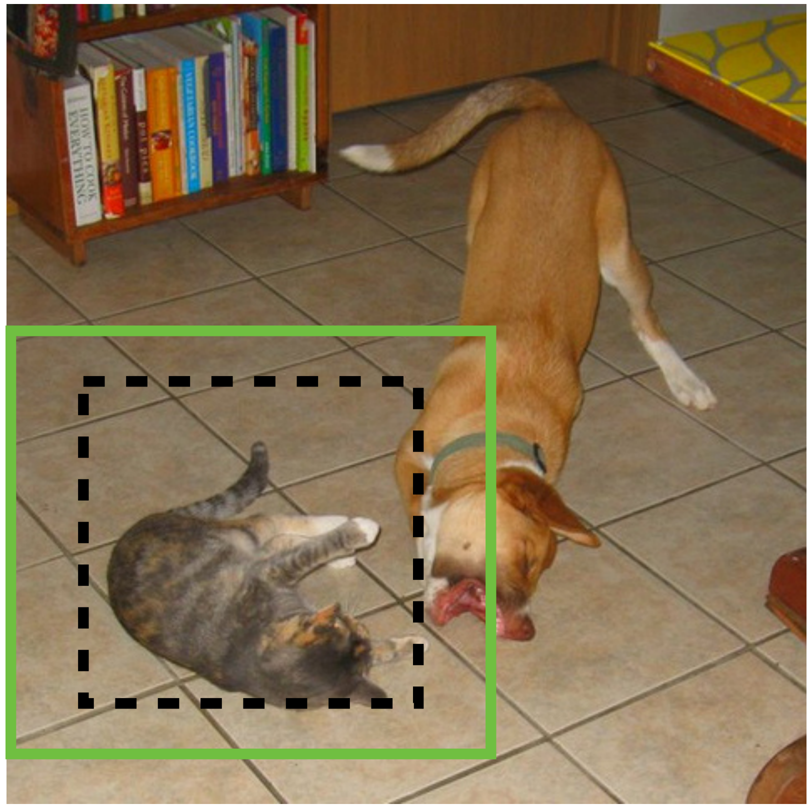
Classical sliding windows



Is it a cat? **No**

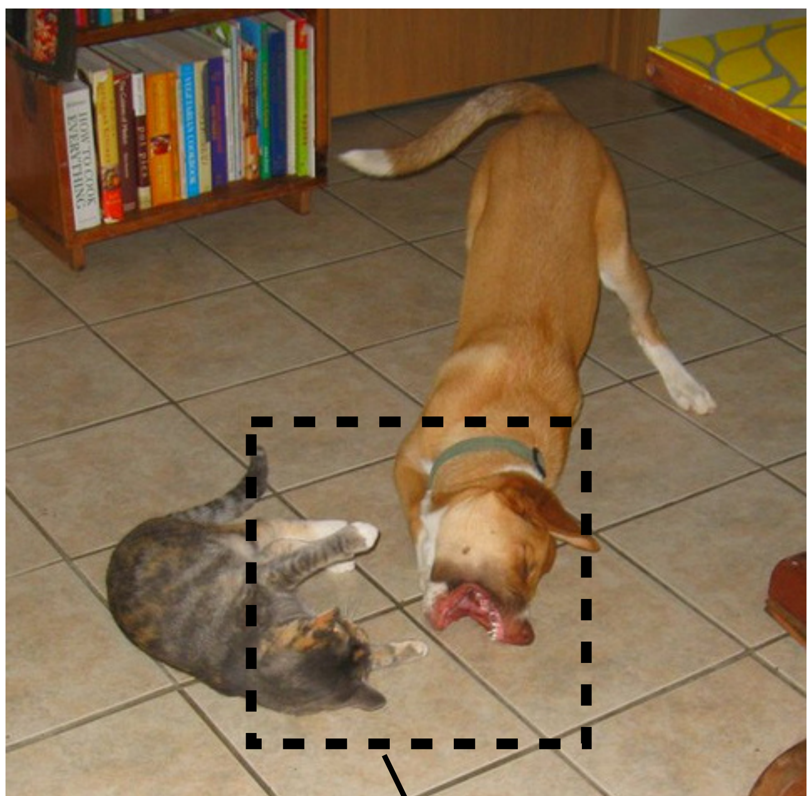
Discretize the box space **densely**

SSD and other deep approaches



Bounding Box Prediction

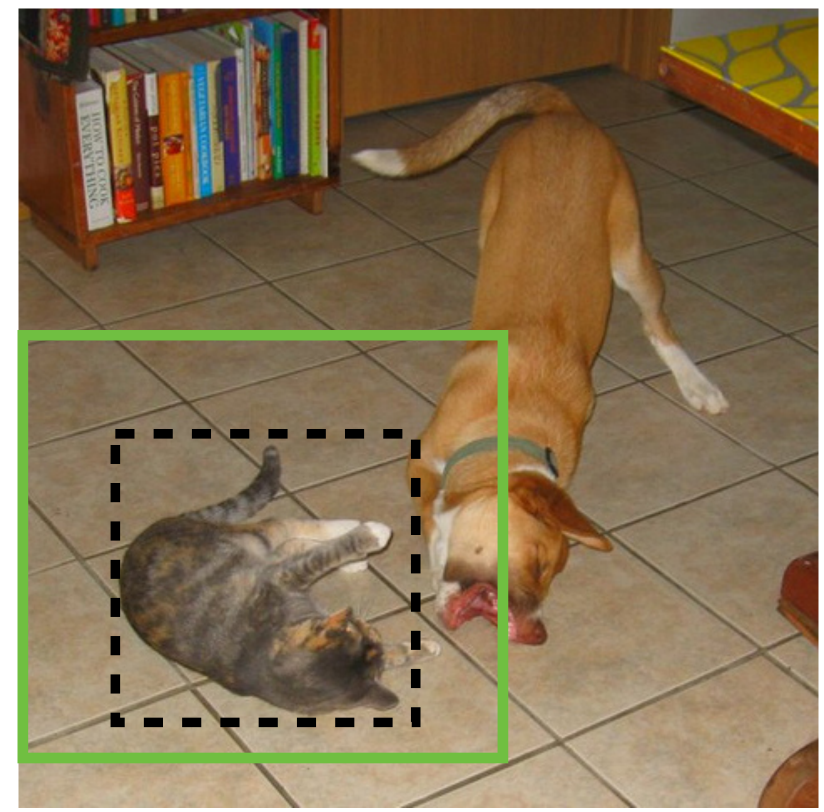
Classical sliding windows



Is it a cat? **No**

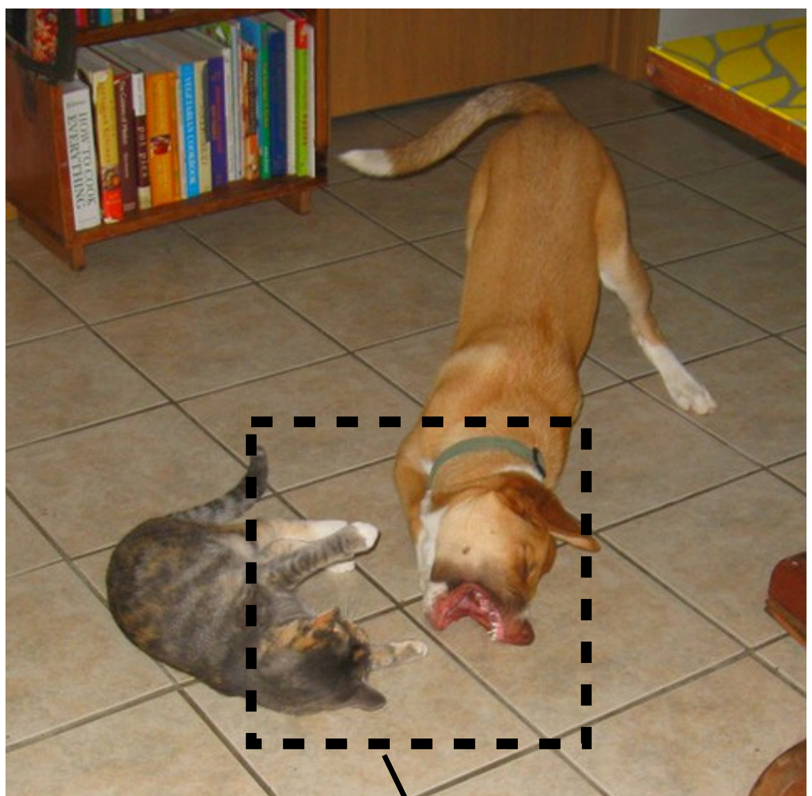
Discretize the box space **densely**

SSD and other deep approaches



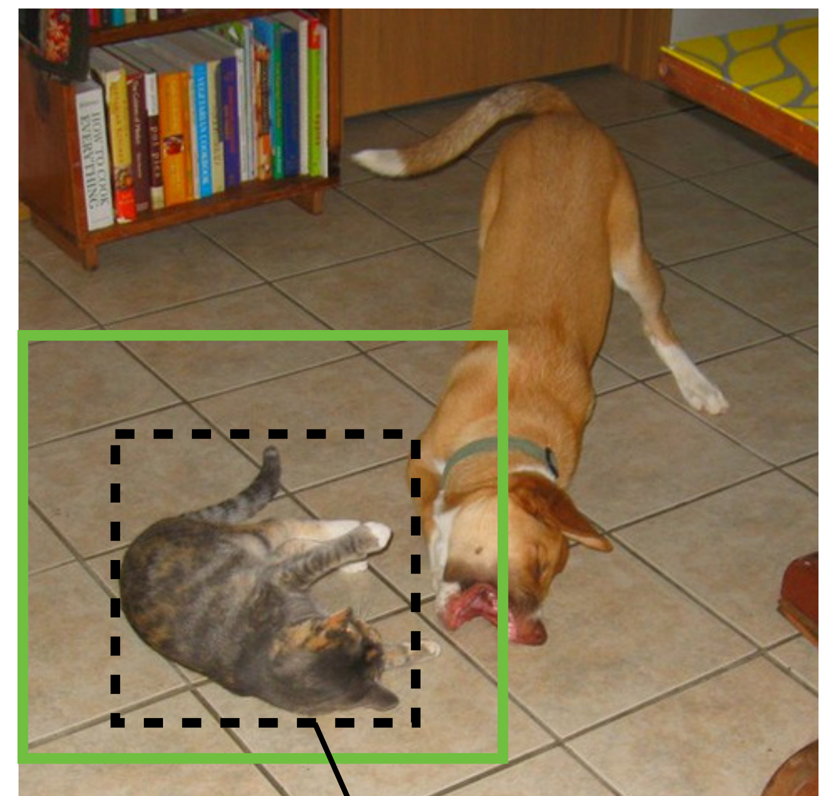
Bounding Box Prediction

Classical sliding windows



Is it a cat? **No**

SSD and other deep approaches

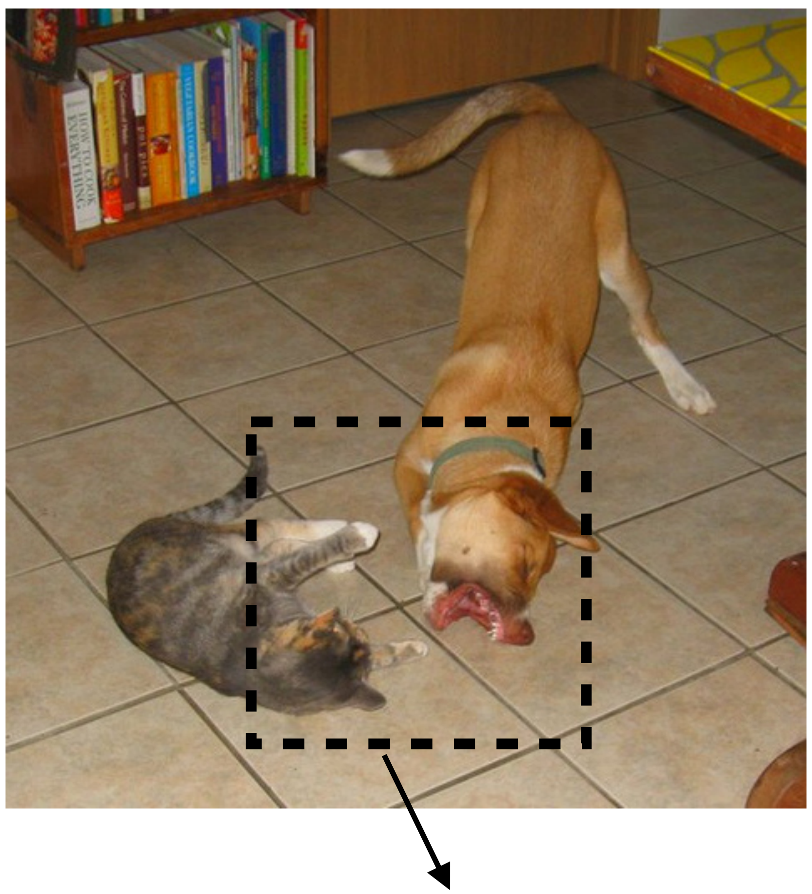


cat: 0.8 dog: 0.1

Discretize the box space **densely**

Bounding Box Prediction

Classical sliding windows



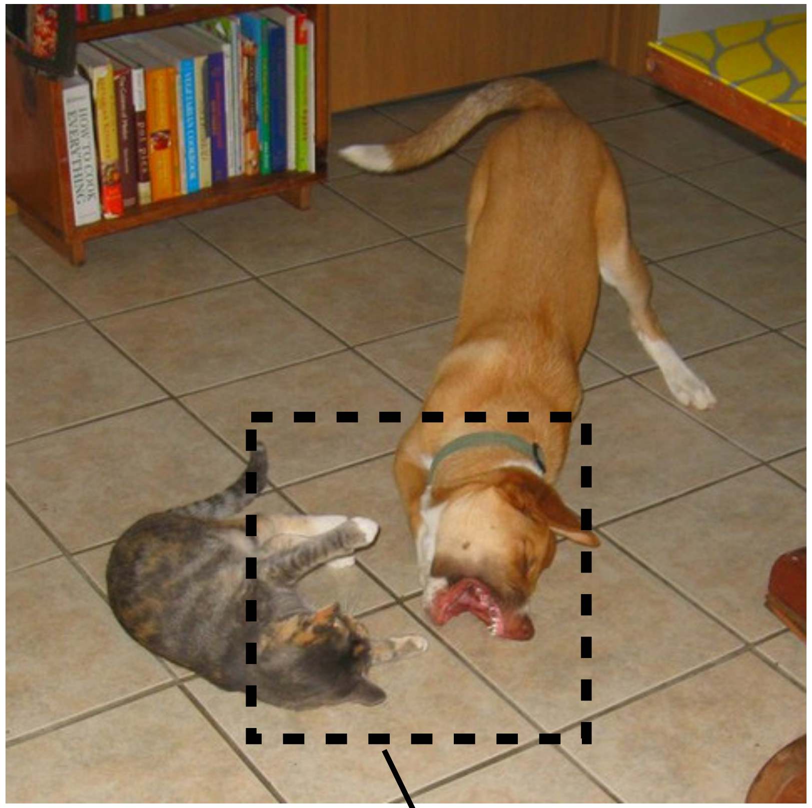
Is it a cat? **No**

Discretize the box space **densely**

SSD and other deep approaches

Bounding Box Prediction

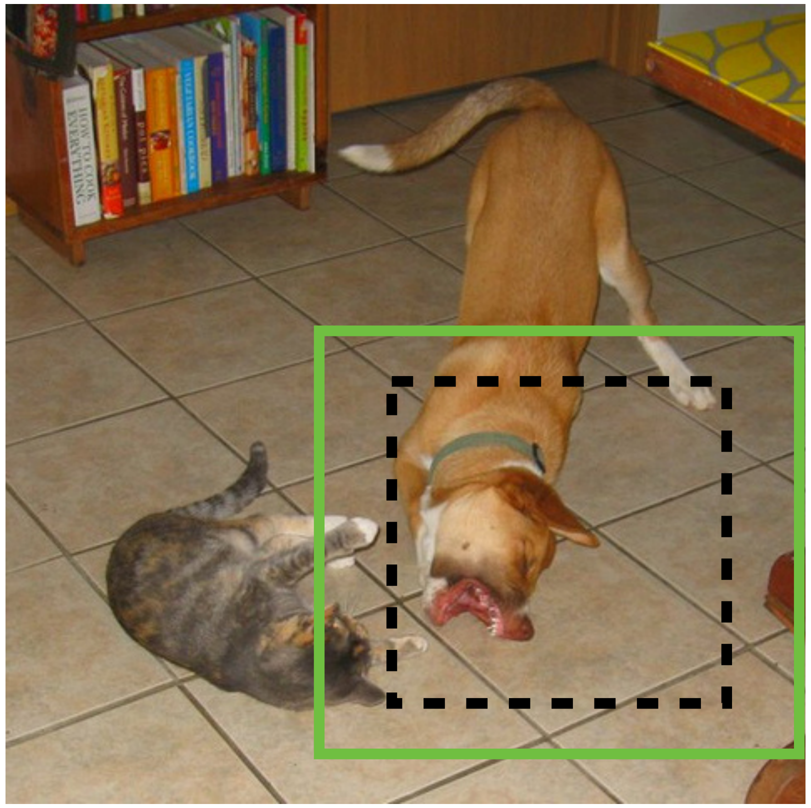
Classical sliding windows



Is it a cat? No

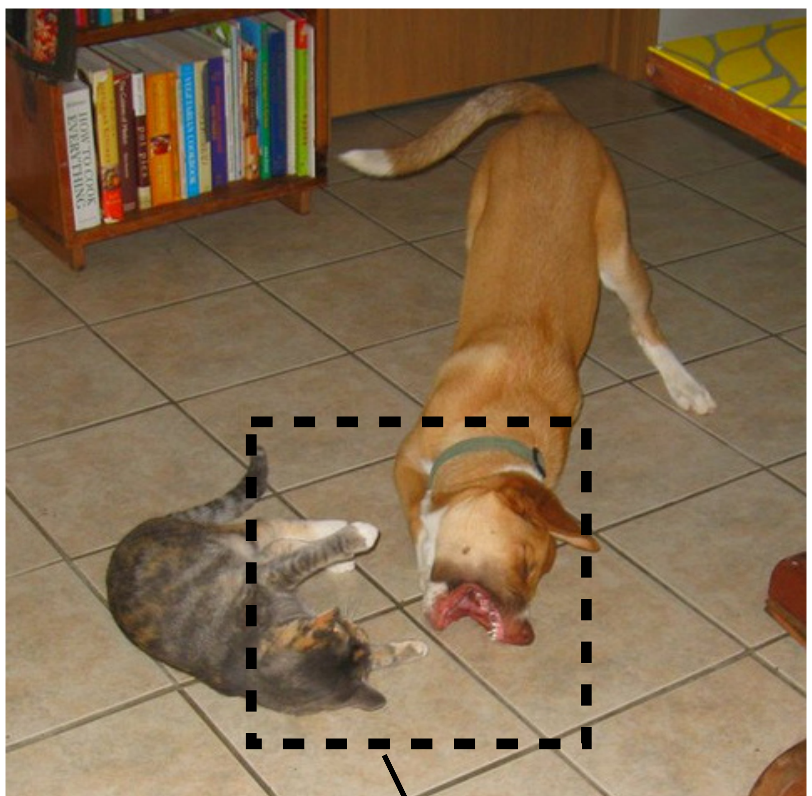
Discretize the box space **densely**

SSD and other deep approaches



Bounding Box Prediction

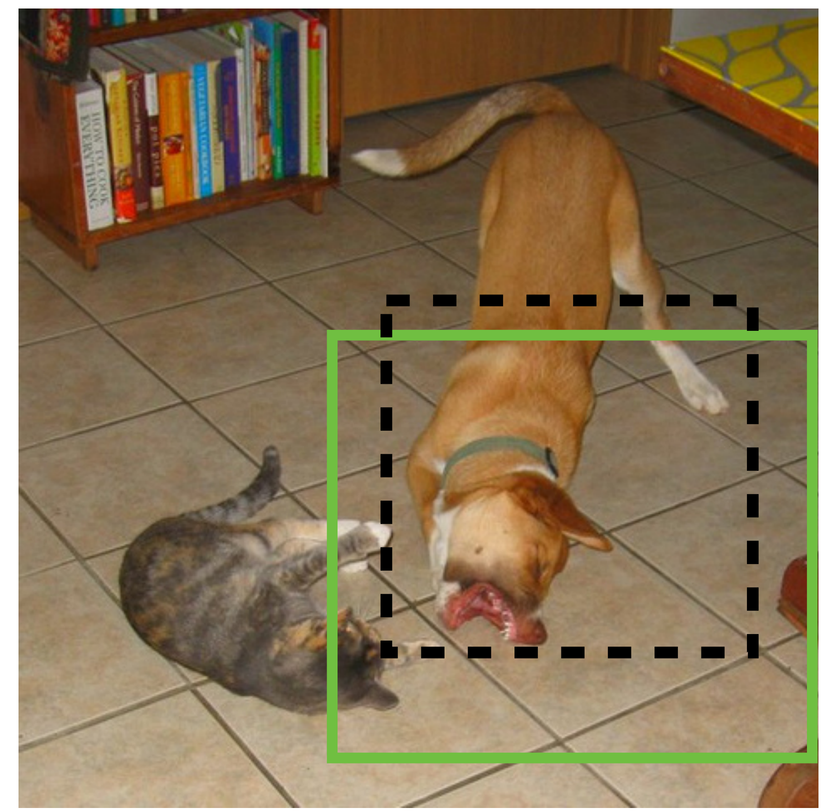
Classical sliding windows



Is it a cat? **No**

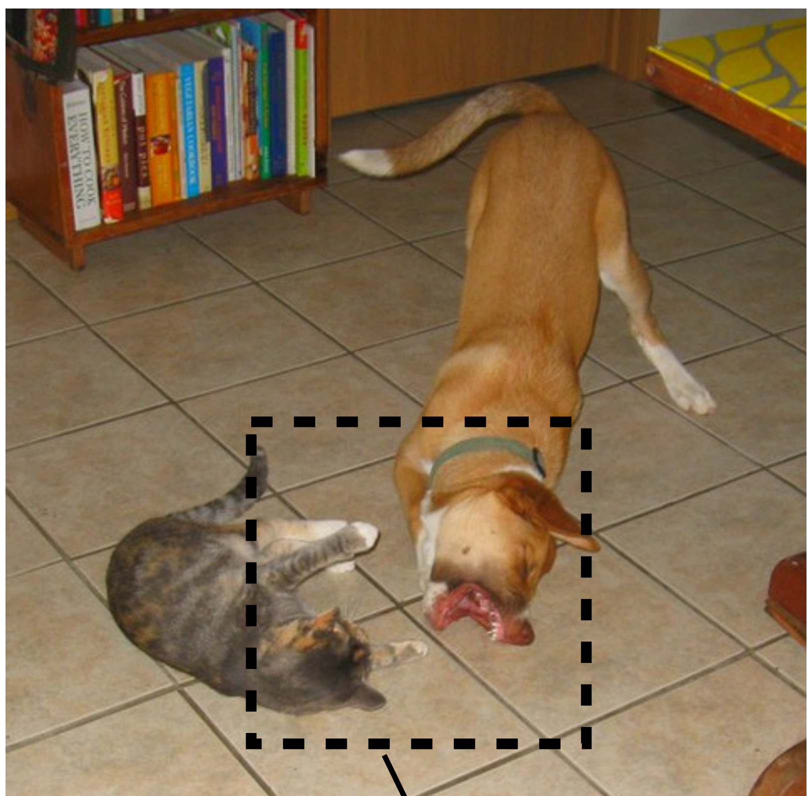
Discretize the box space **densely**

SSD and other deep approaches



Bounding Box Prediction

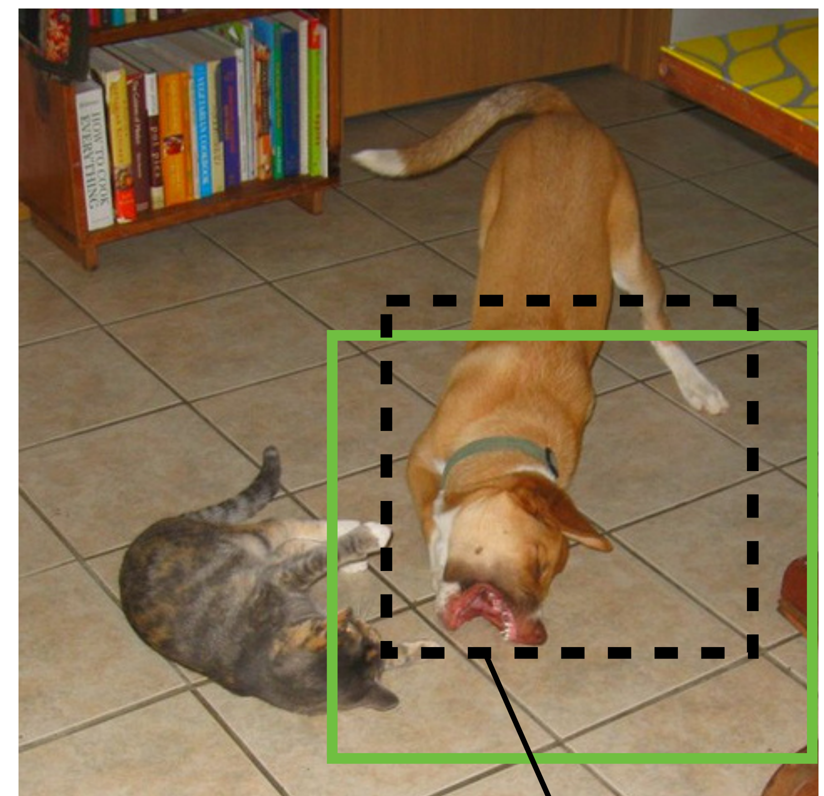
Classical sliding windows



Is it a cat? **No**

Discretize the box space **densely**

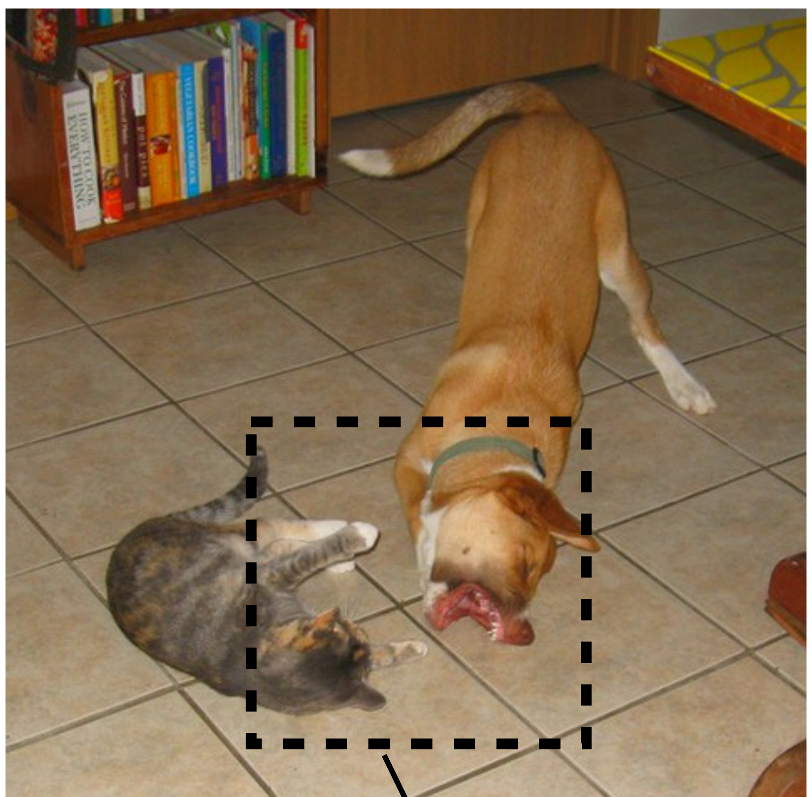
SSD and other deep approaches



dog: 0.4 cat: 0.2

Bounding Box Prediction

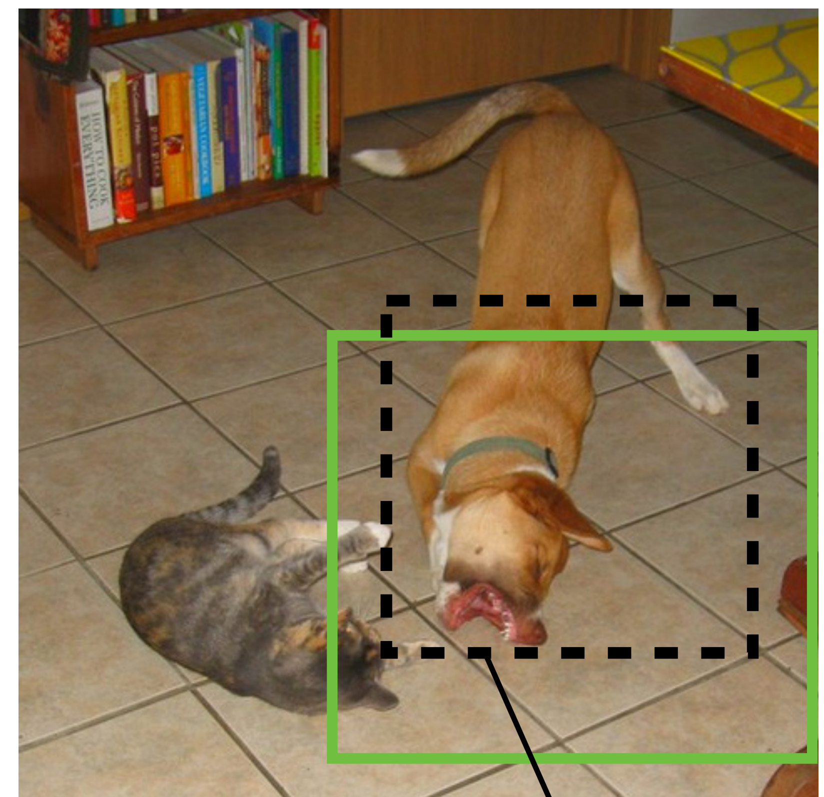
Classical sliding windows



Is it a cat? **No**

Discretize the box space **densely**

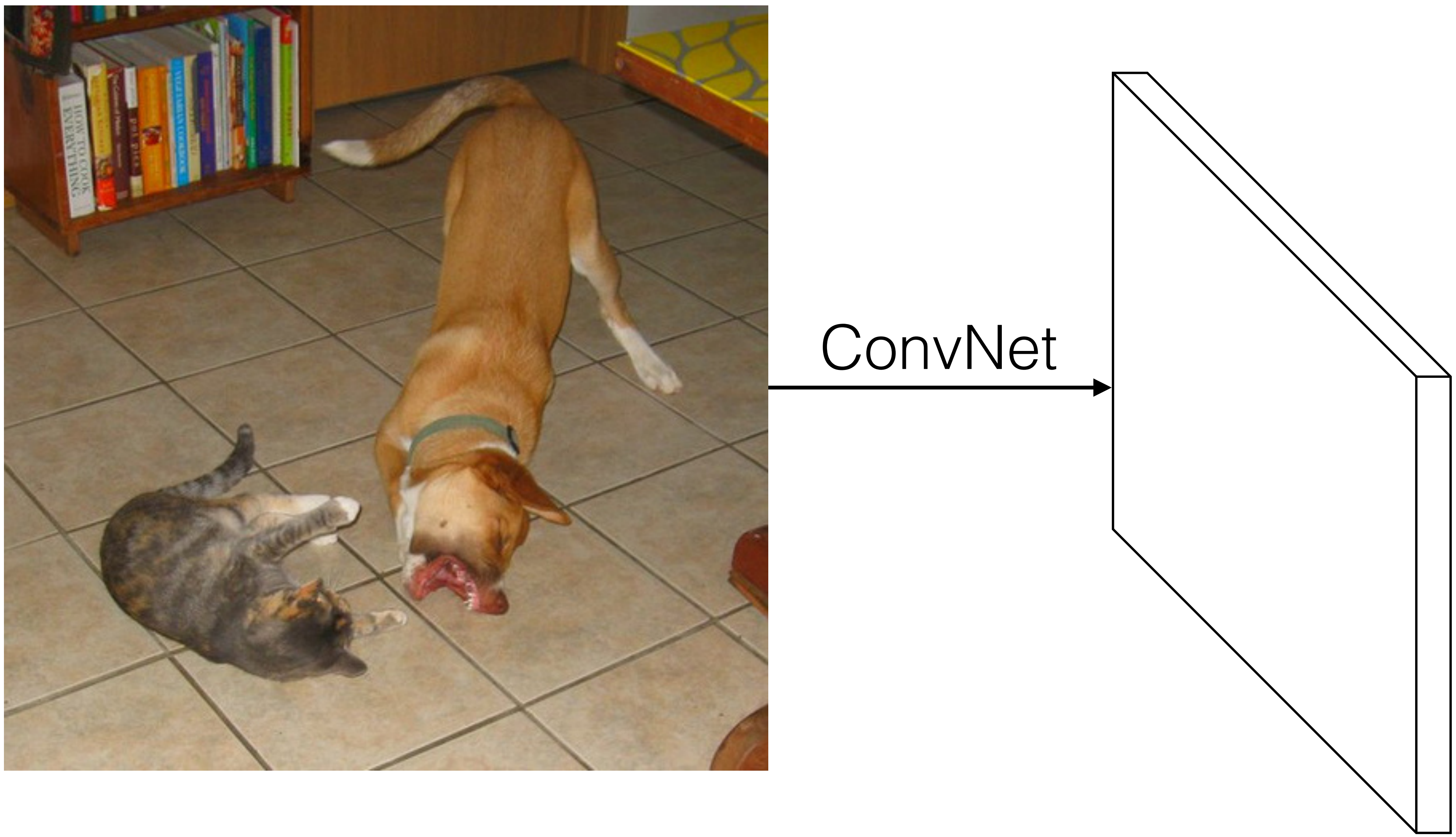
SSD and other deep approaches



dog: 0.4 cat: 0.2

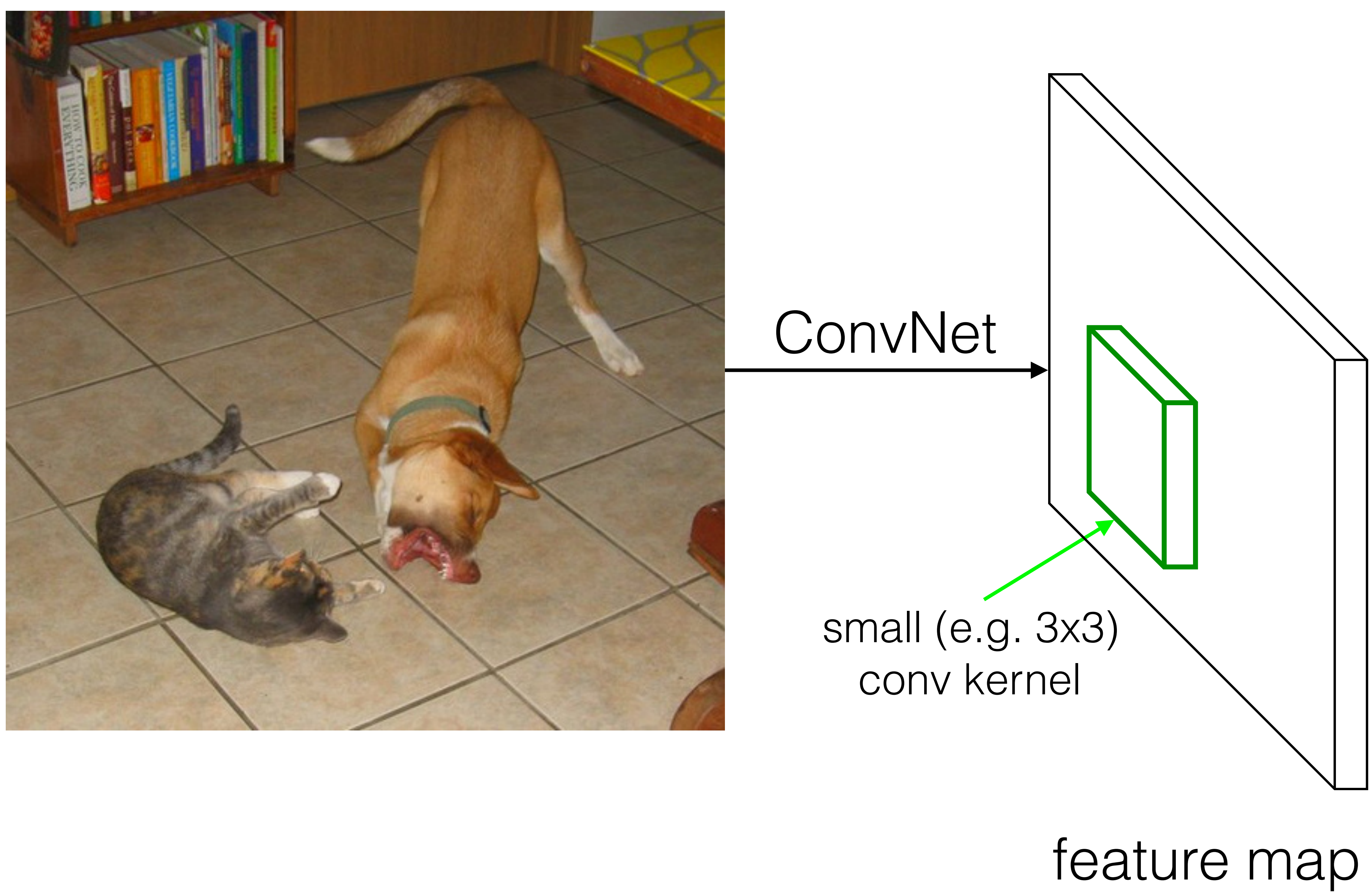
Discretize the box space more **coarsely**
Refine the coordinates of each box

SSD Output Layer

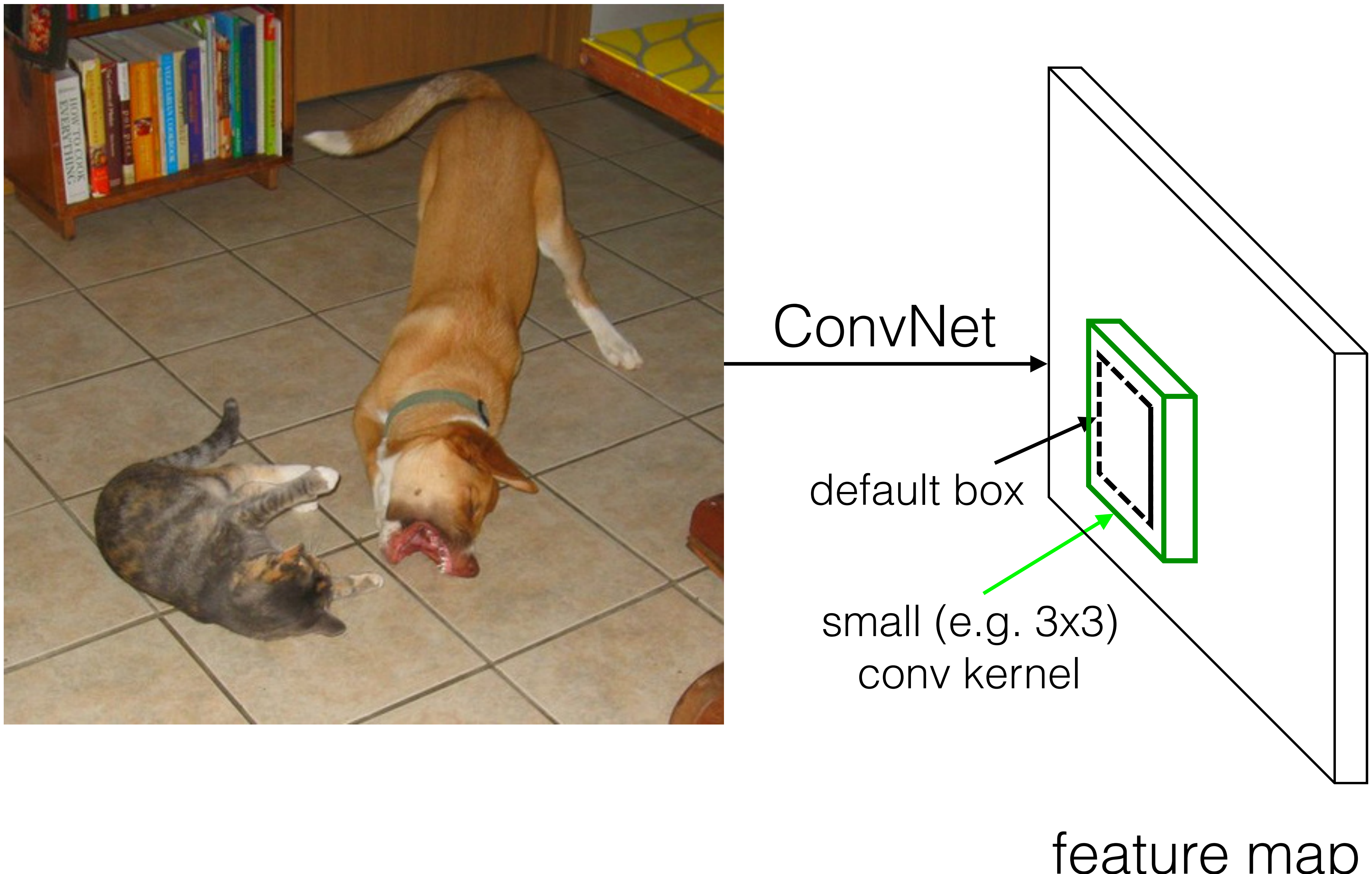


feature map

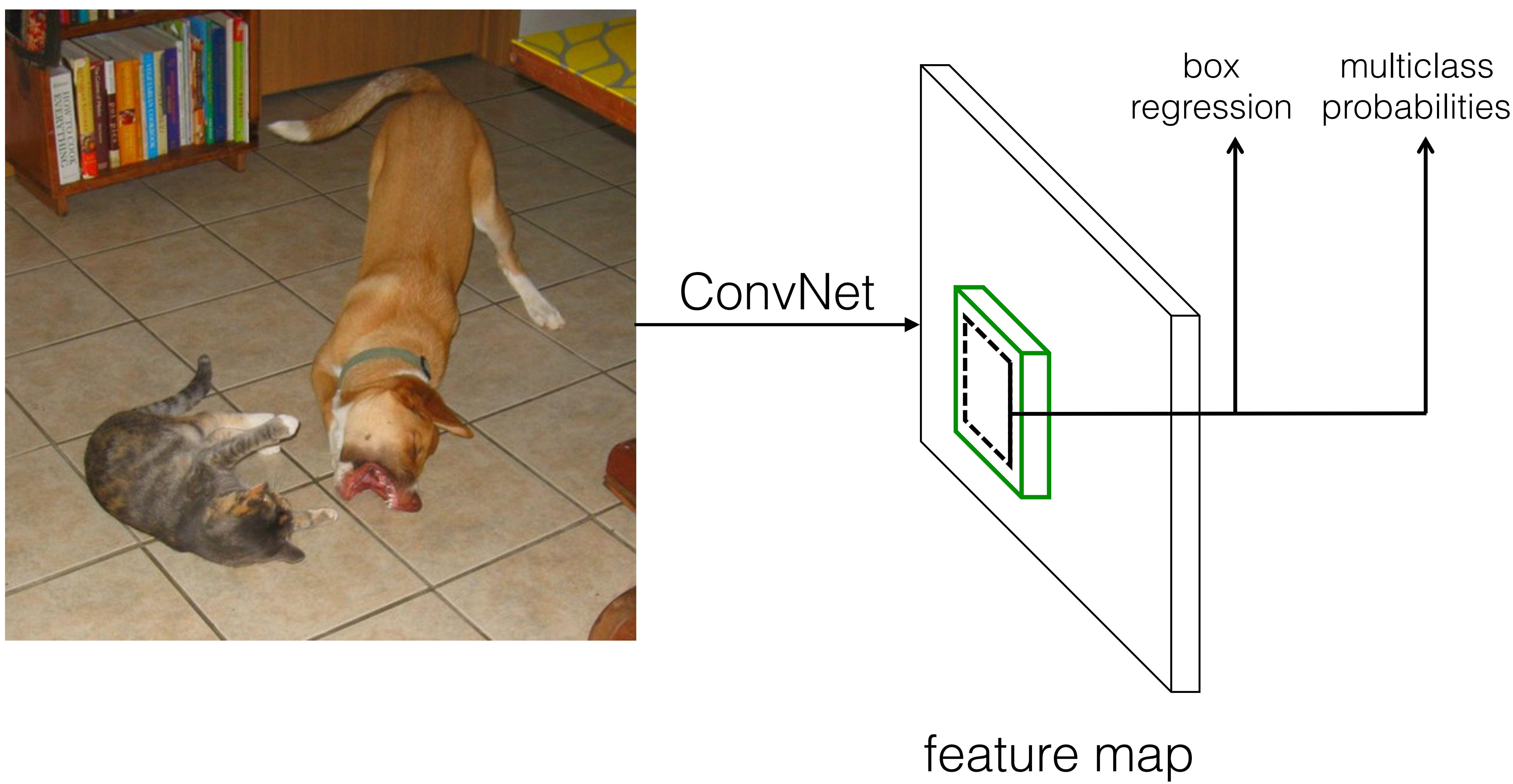
SSD Output Layer



SSD Output Layer



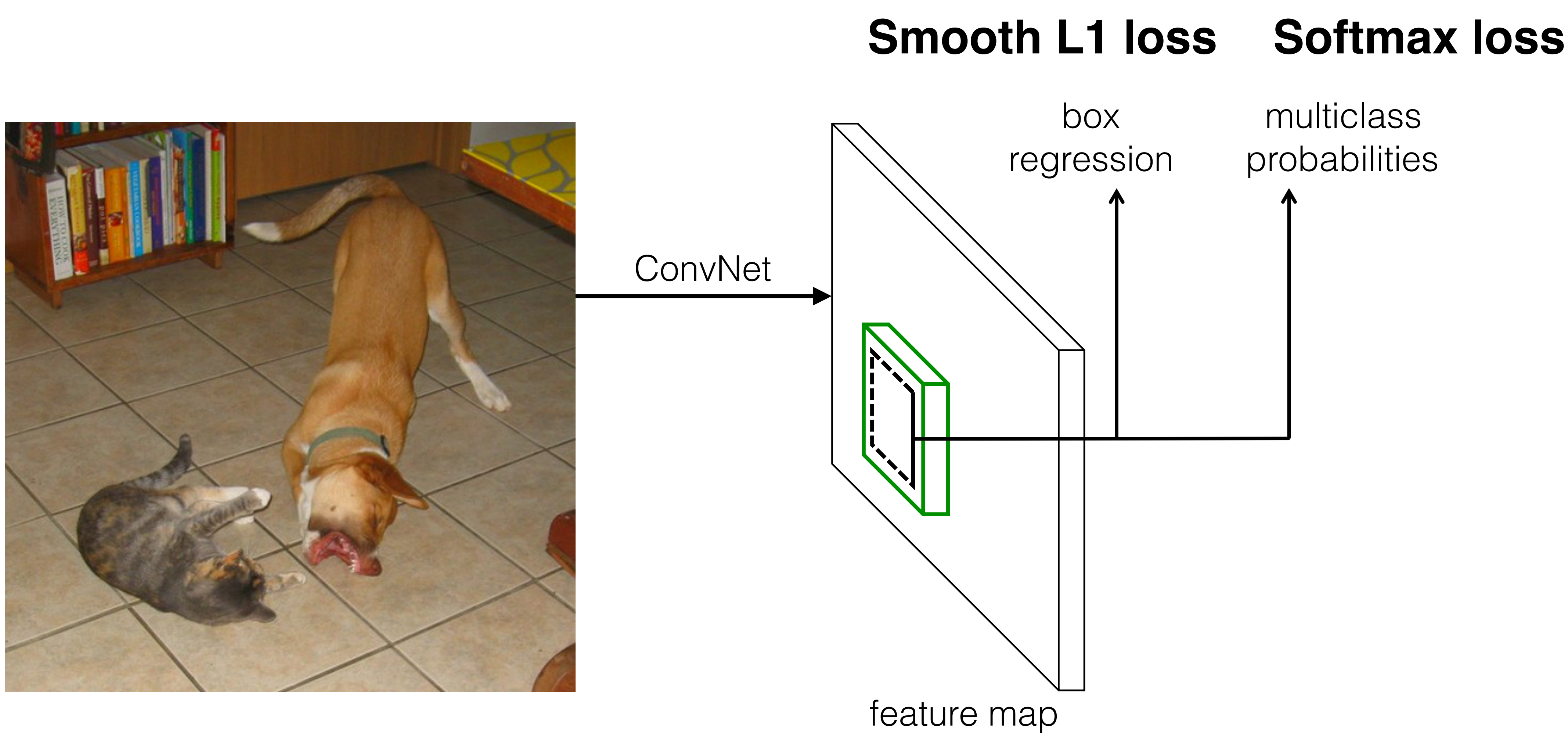
SSD Output Layer



SSD Training

- Match default boxes to ground truth boxes to determine true/false positives.

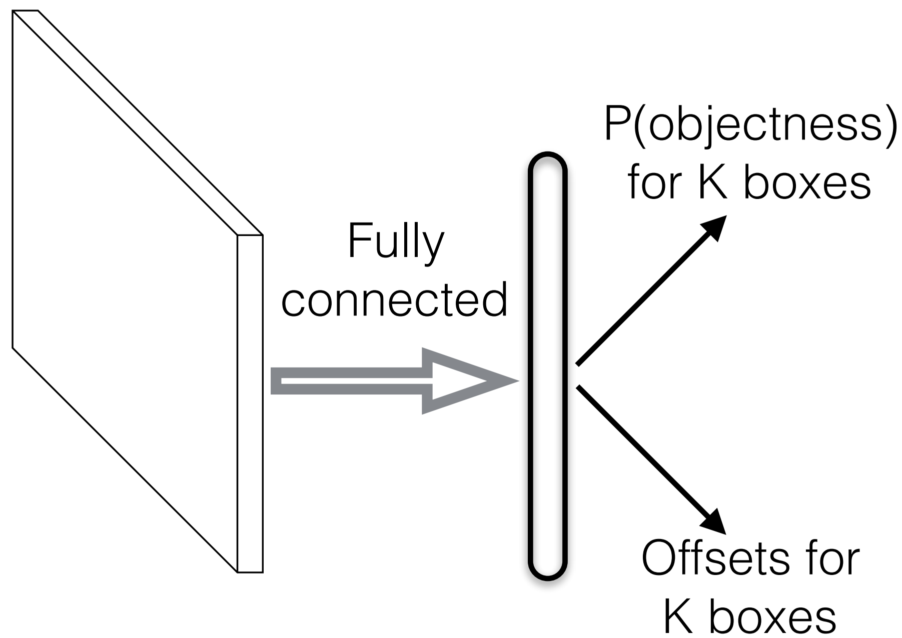
- Loss = **SmoothL1**(box param) + **Softmax**(class prob)



Related Work

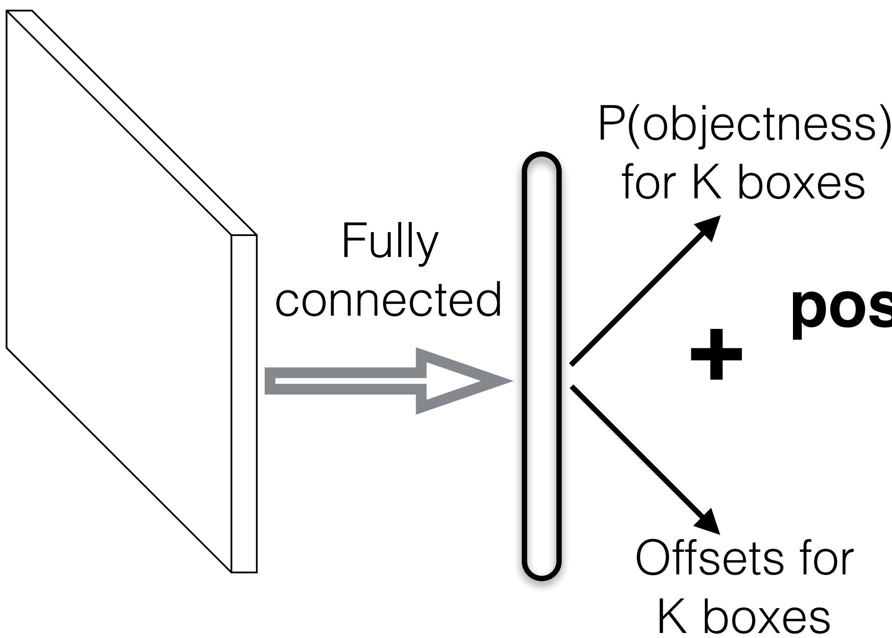
Related Work

MultiBox [Erhan et al. CVPR14]



Related Work

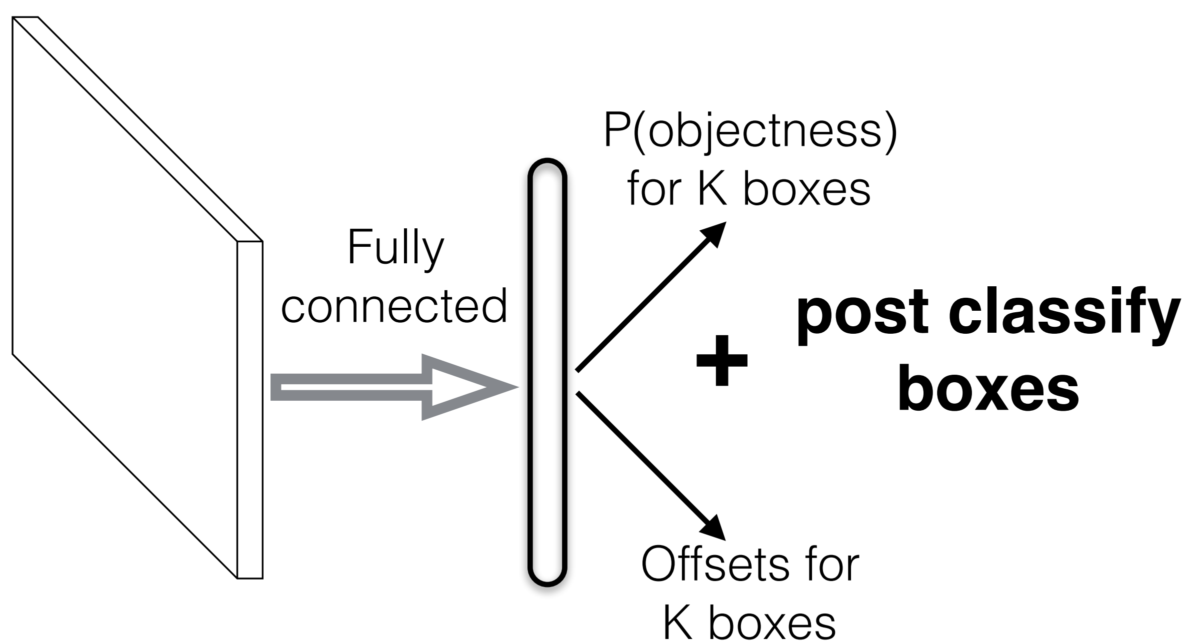
MultiBox [Erhan et al. CVPR14]



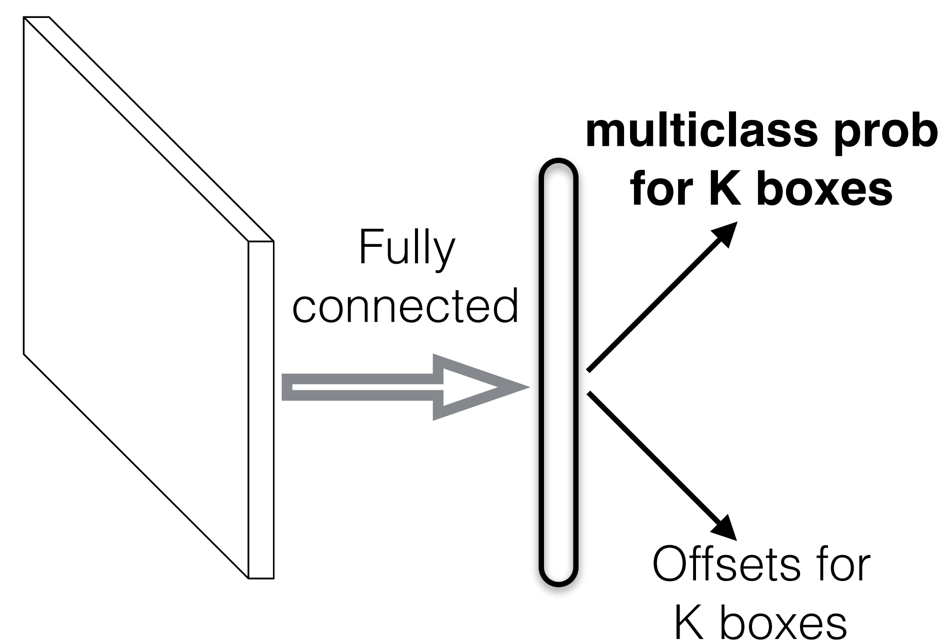
+ post classify boxes

Related Work

MultiBox [Erhan et al. CVPR14]

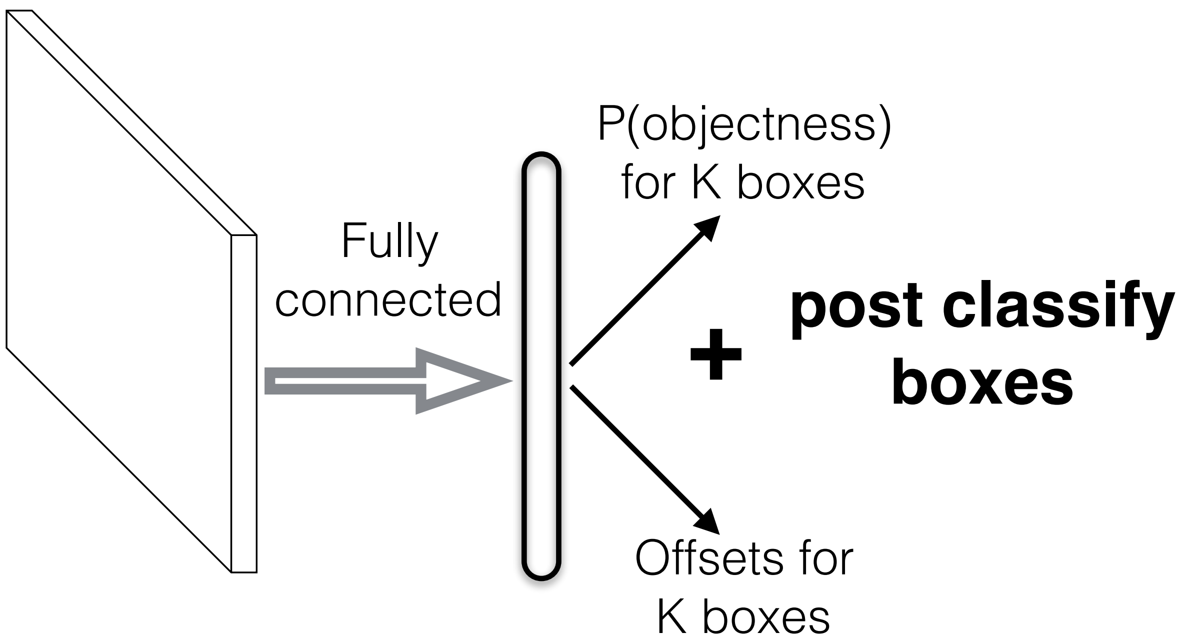


YOLO [Redmon et al. CVPR16]

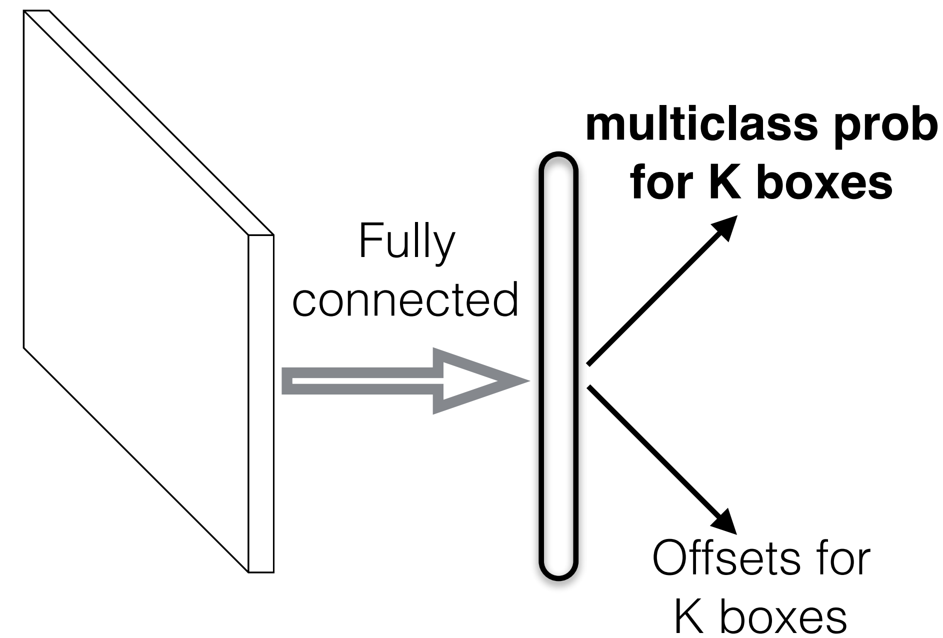


Related Work

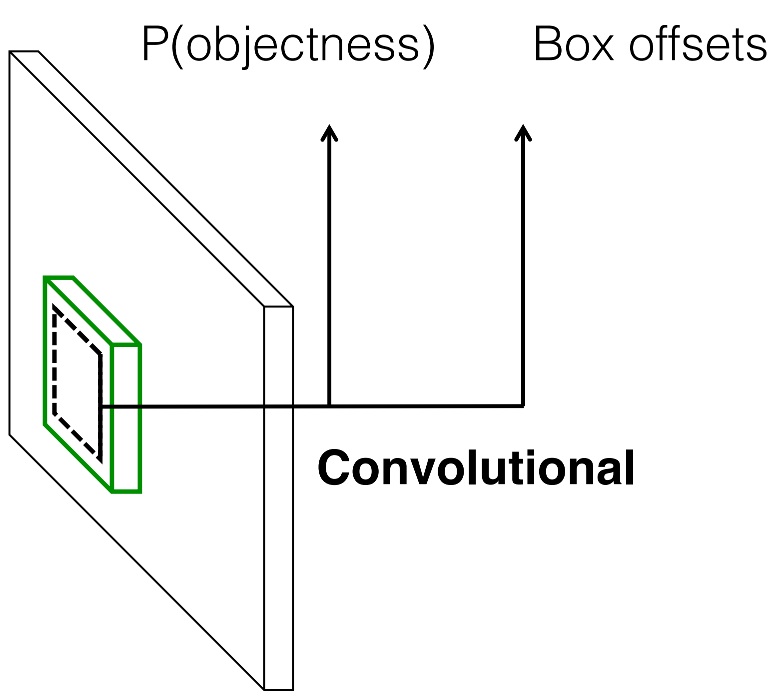
MultiBox [Erhan et al. CVPR14]



YOLO [Redmon et al. CVPR16]

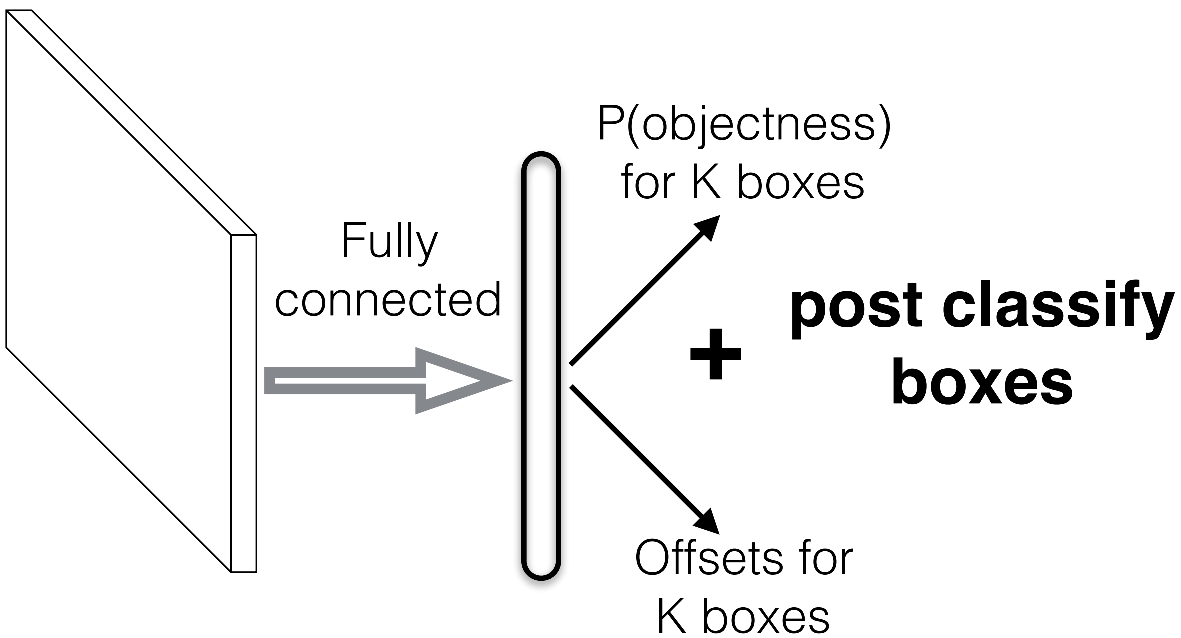


Faster R-CNN [Ren et al. NIPS15]

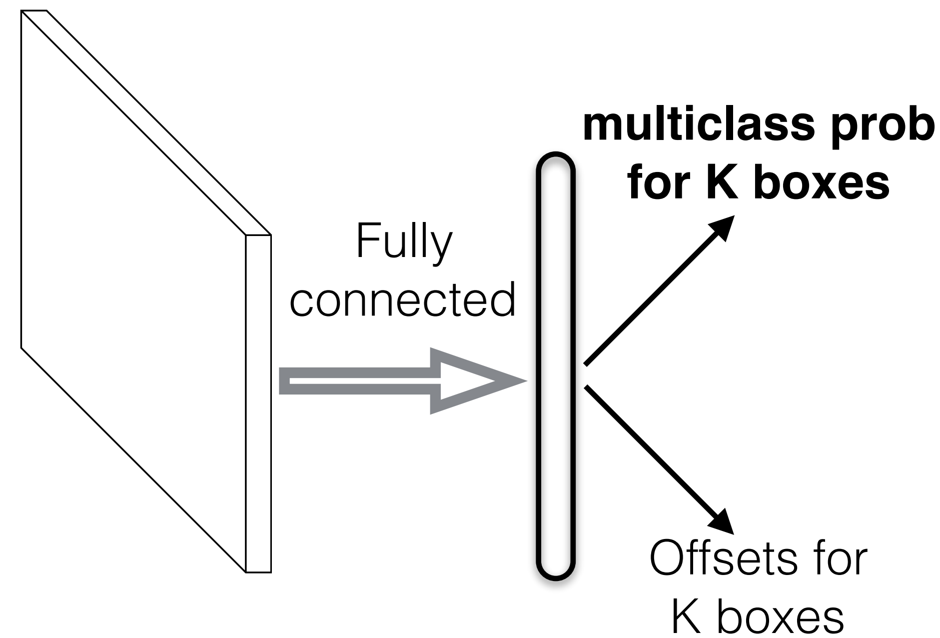


Related Work

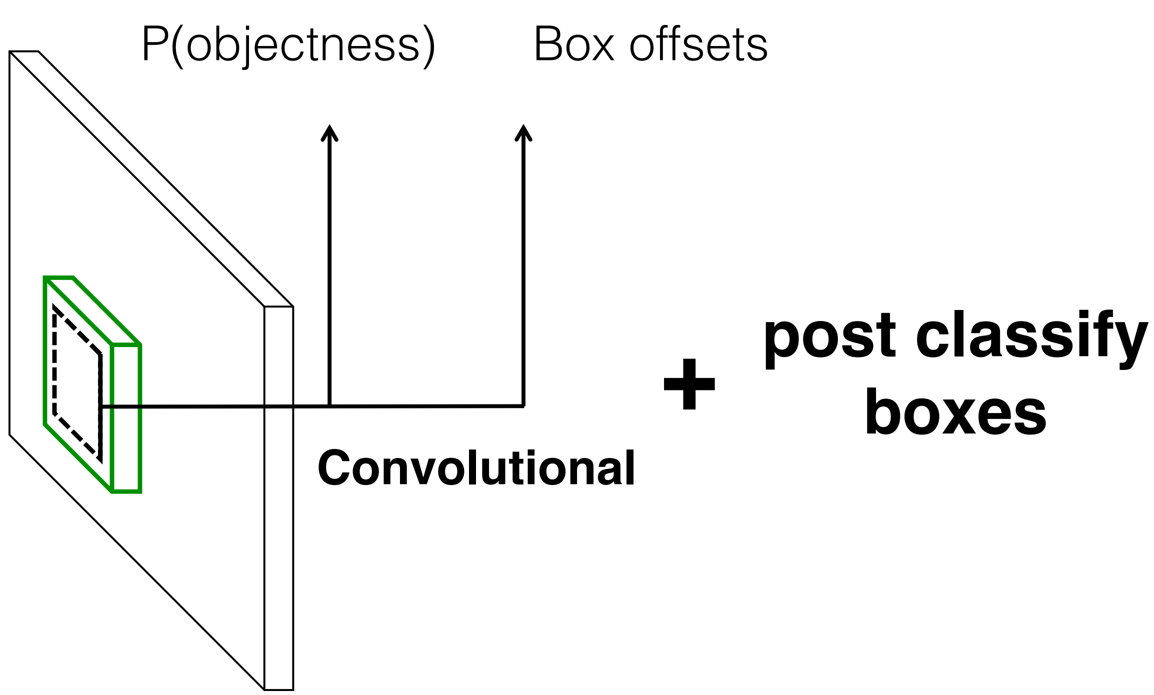
MultiBox [Erhan et al. CVPR14]



YOLO [Redmon et al. CVPR16]

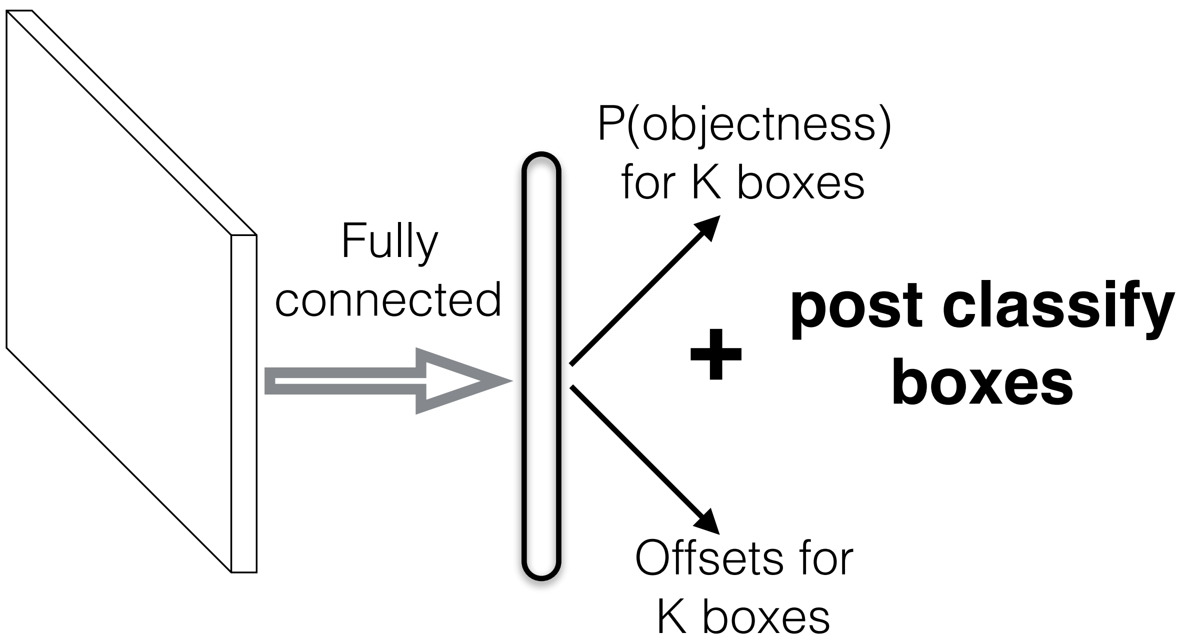


Faster R-CNN [Ren et al. NIPS15]

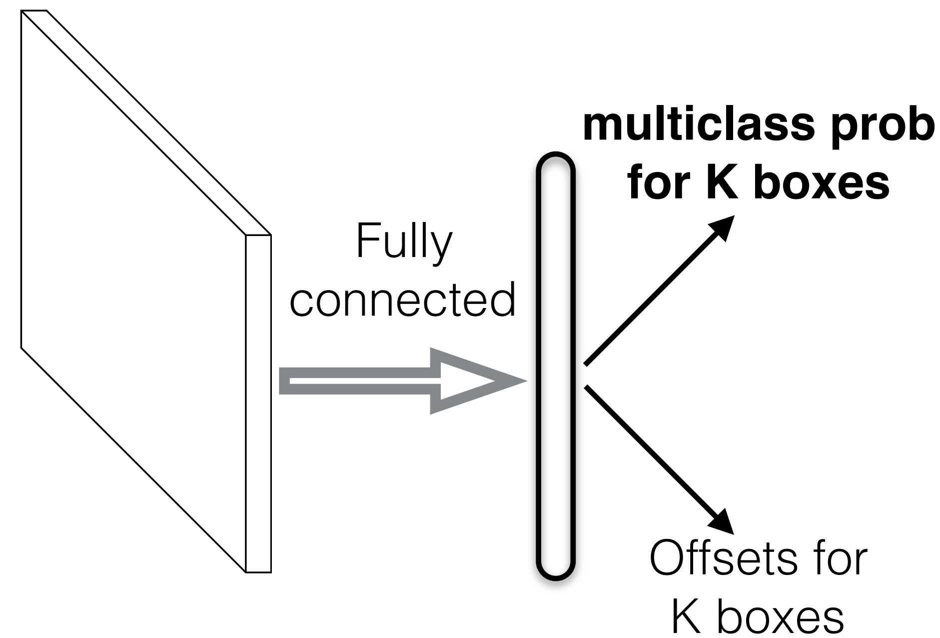


Related Work

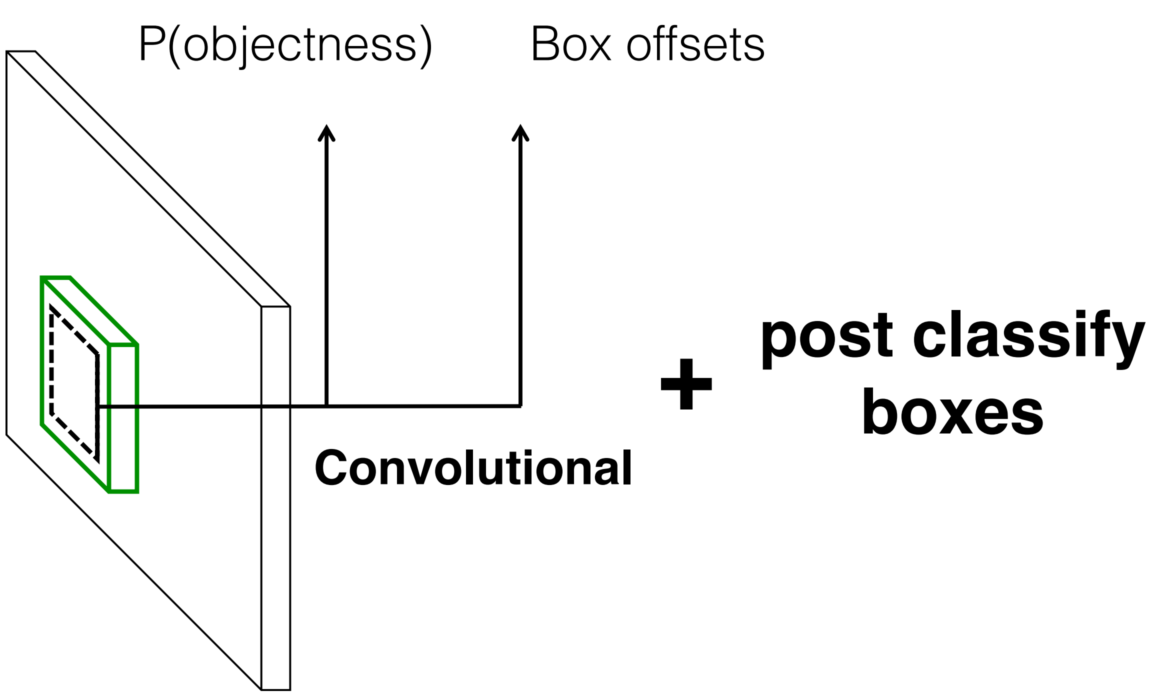
MultiBox [Erhan et al. CVPR14]



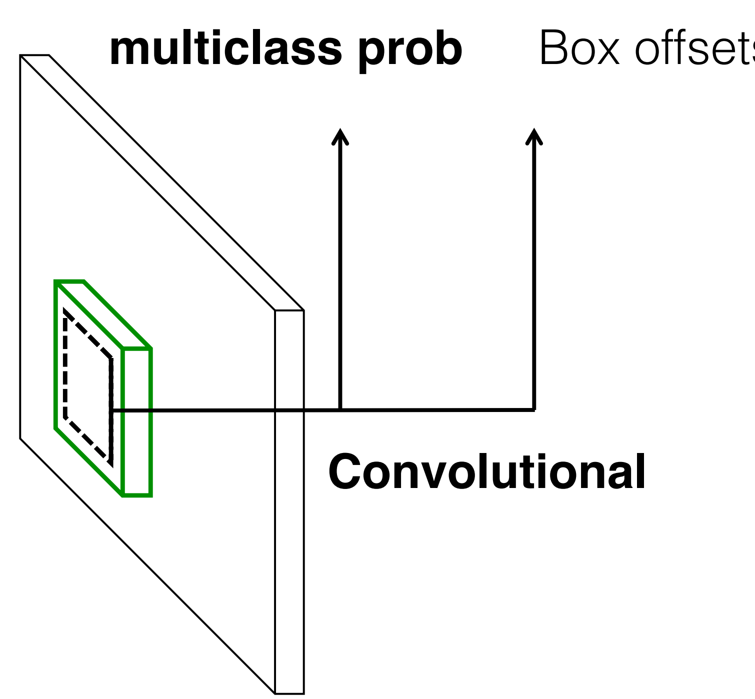
YOLO [Redmon et al. CVPR16]



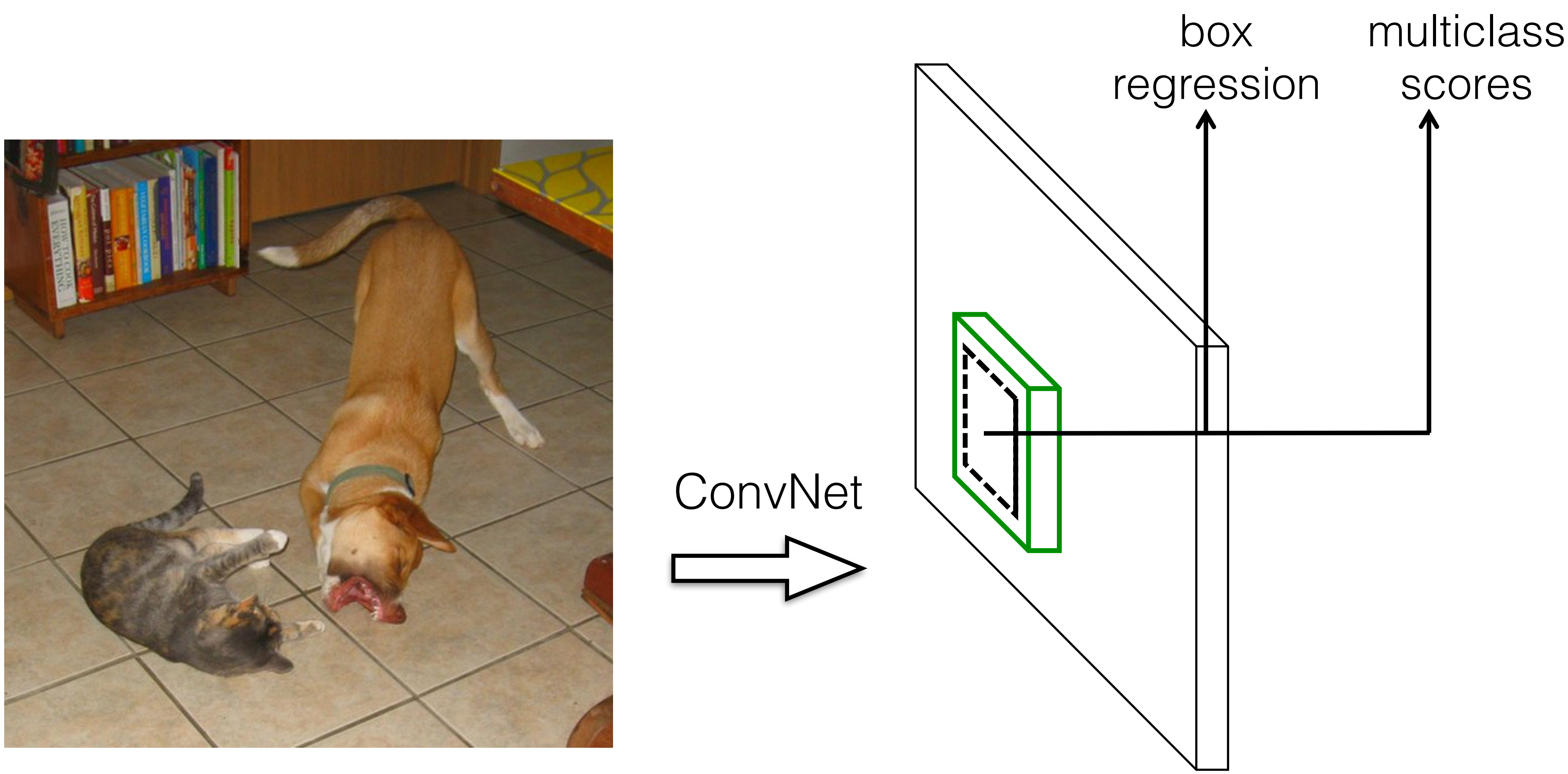
Faster R-CNN [Ren et al. NIPS15]



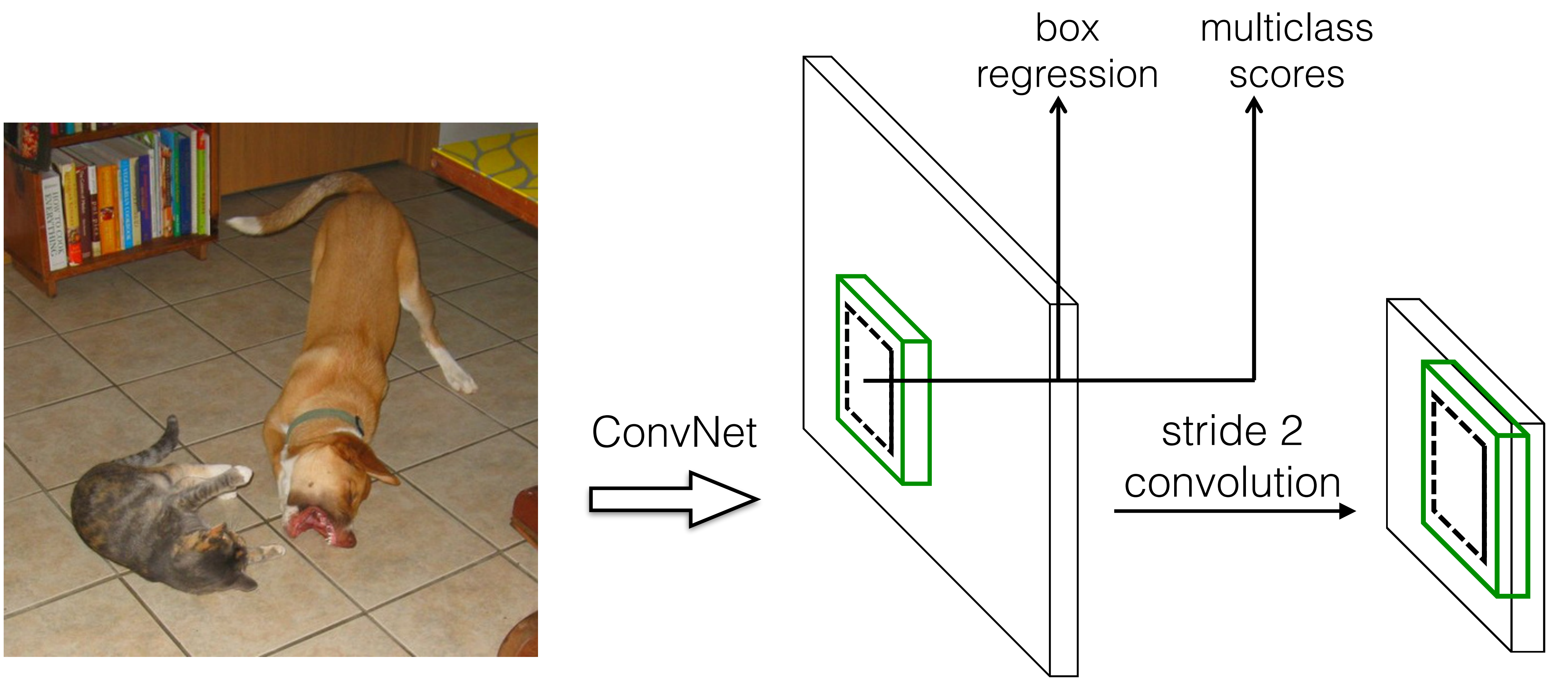
SSD



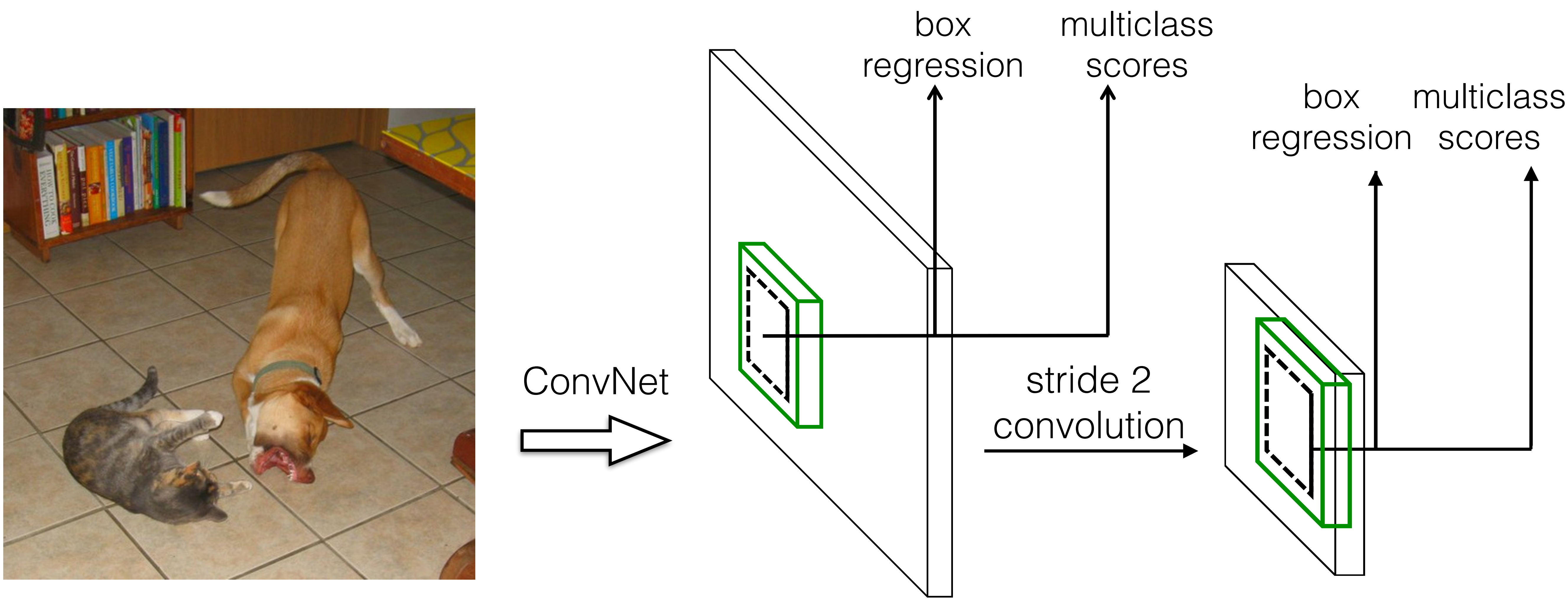
Contribution #1: Multi-Scale Feature Maps



Contribution #1: Multi-Scale Feature Maps

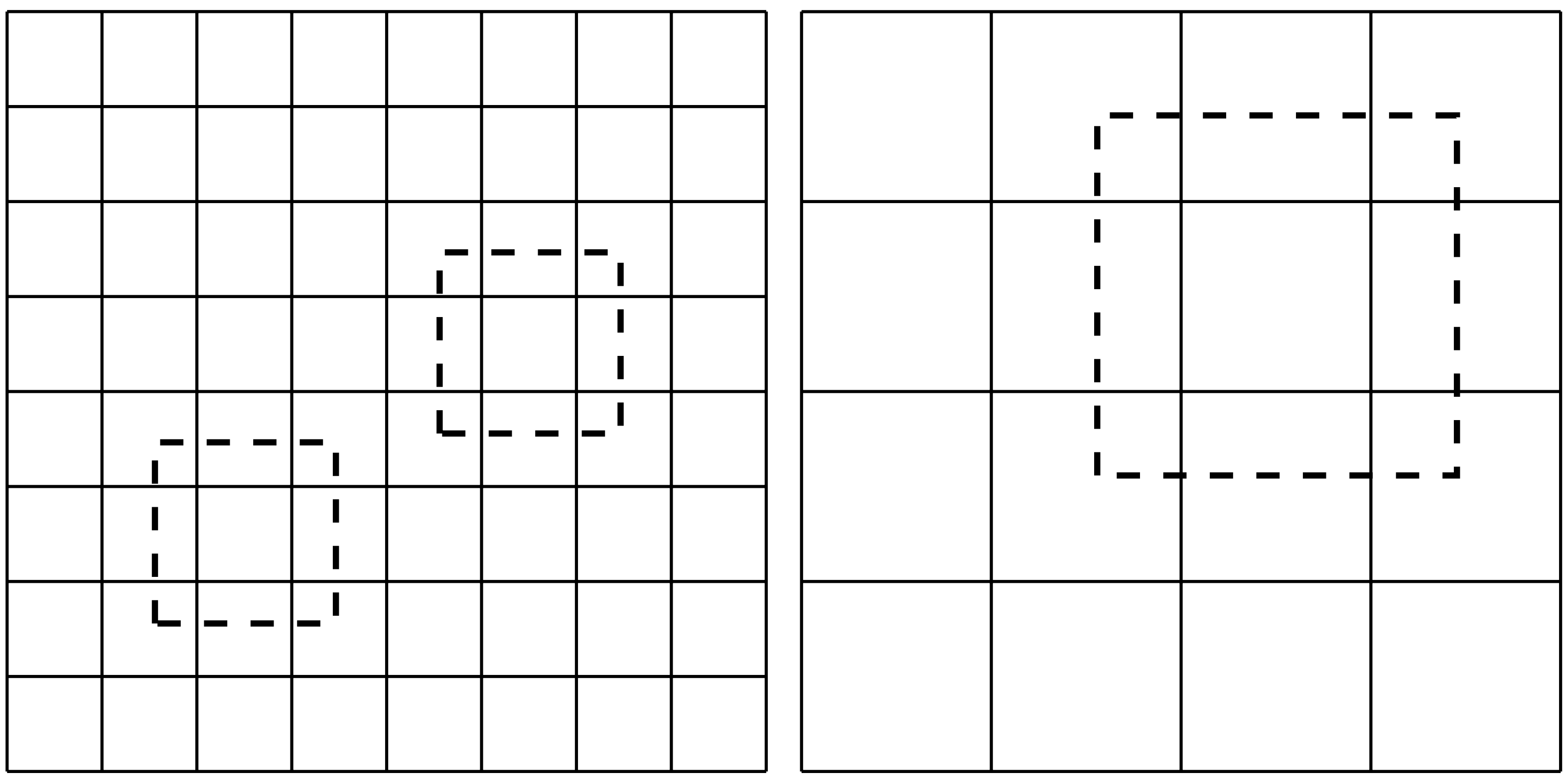


Contribution #1: Multi-Scale Feature Maps



Multi-Scale Feature Maps

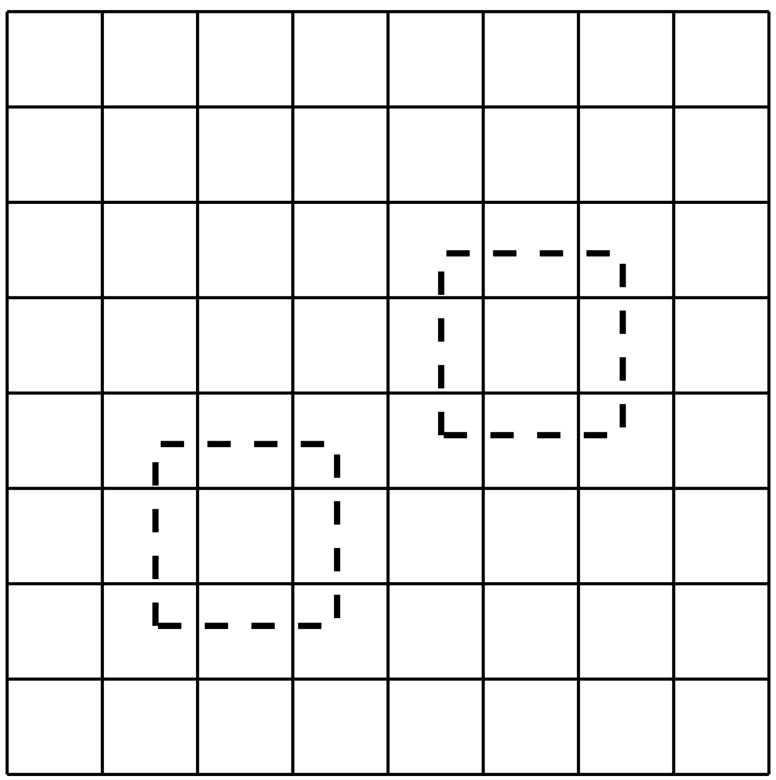
SSD



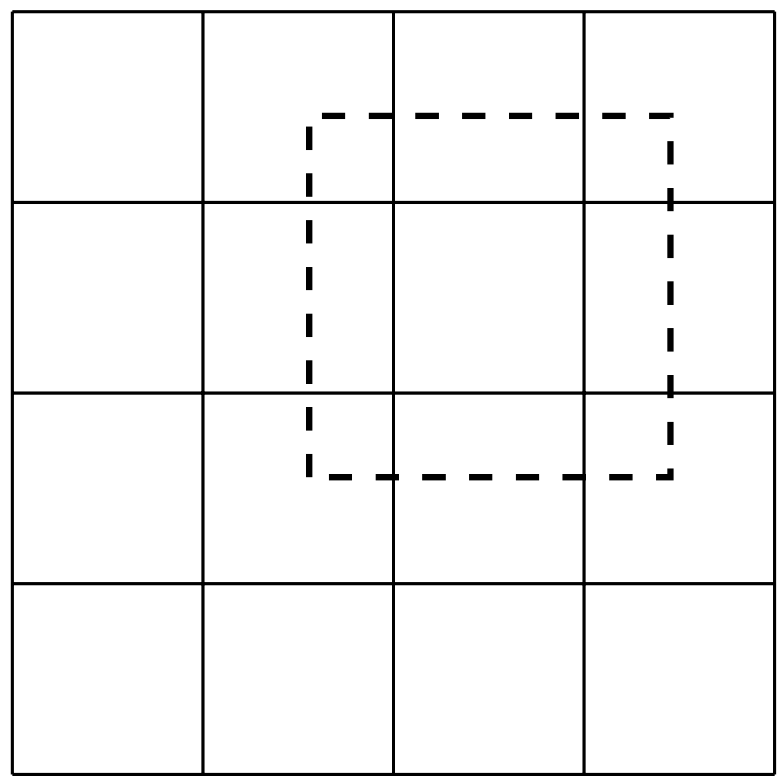
8×8 feature map $\quad$ 4×4 feature map

Multi-Scale Feature Maps

SSD



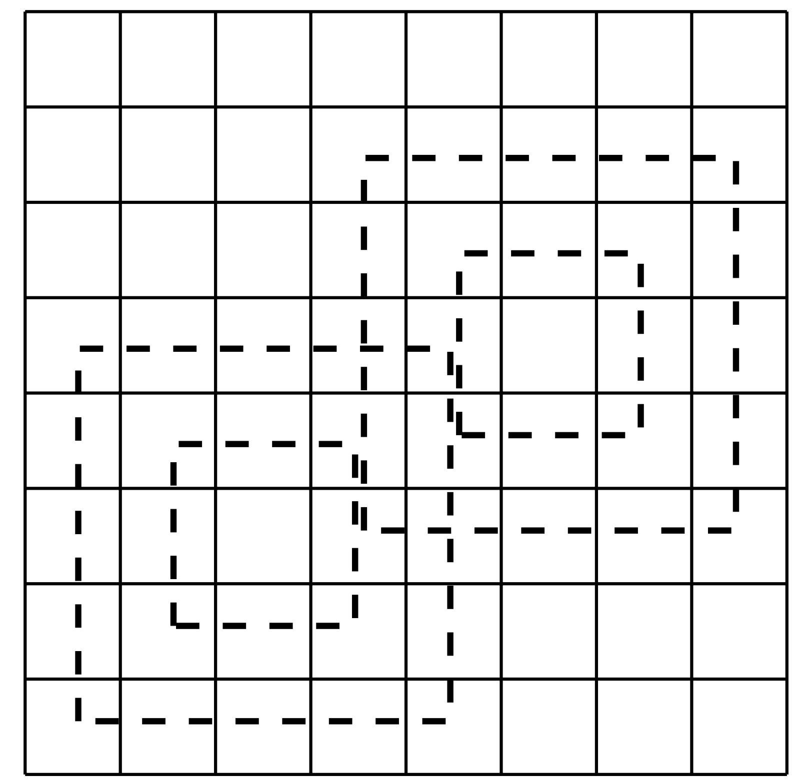
8×8 feature map



4×4 feature map

vs.

Faster R-CNN Objectness Proposal, Ren 2015



8×8 feature map

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	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 use boundary boxes?		# Boxes
38×38	19×19	10×10	5×5	3×3	1×1	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 use boundary boxes?		# Boxes
38×38	19×19	10×10	5×5	3×3	1×1	Yes	No	
✔	✔	✔	✔	✔	✔	74.3	63.4	8732
✔	✔	✔				70.7	69.2	9864
	✔					62.4	64.0	8664

Multi-Scale Feature Maps Experiment

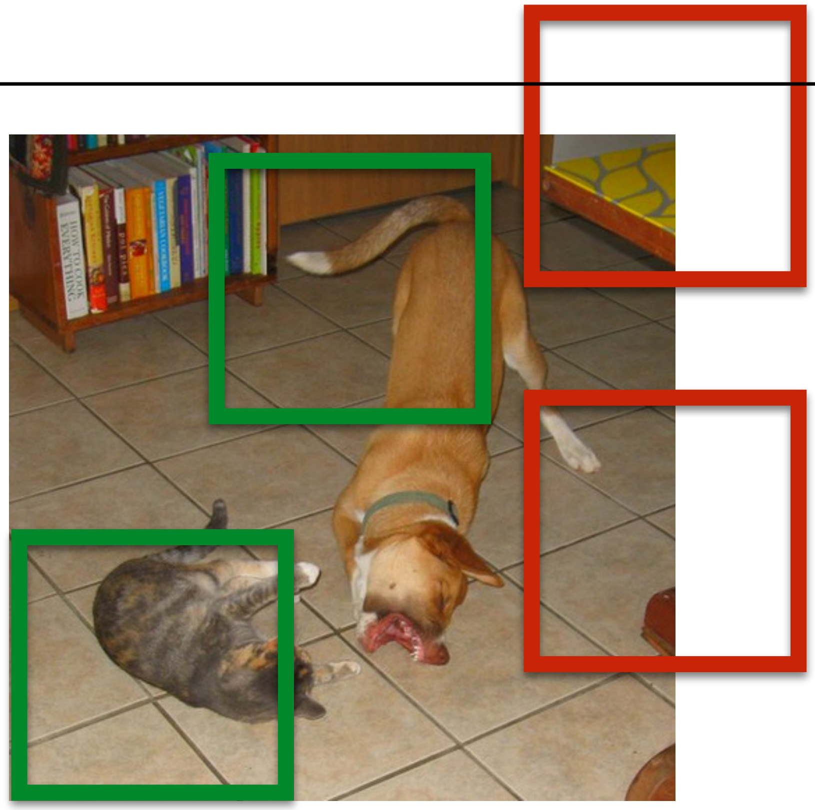
Prediction source layers from:						mAP use boundary boxes?		# Boxes
38×38	19×19	10×10	5×5	3×3	1×1	Yes	No	
✔	✔	✔	✔	✔	✔	74.3	63.4	8732
✔	✔	✔				70.7	69.2	9864
	✔					62.4	64.0	8664

Multi-Scale Feature Maps Experiment

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

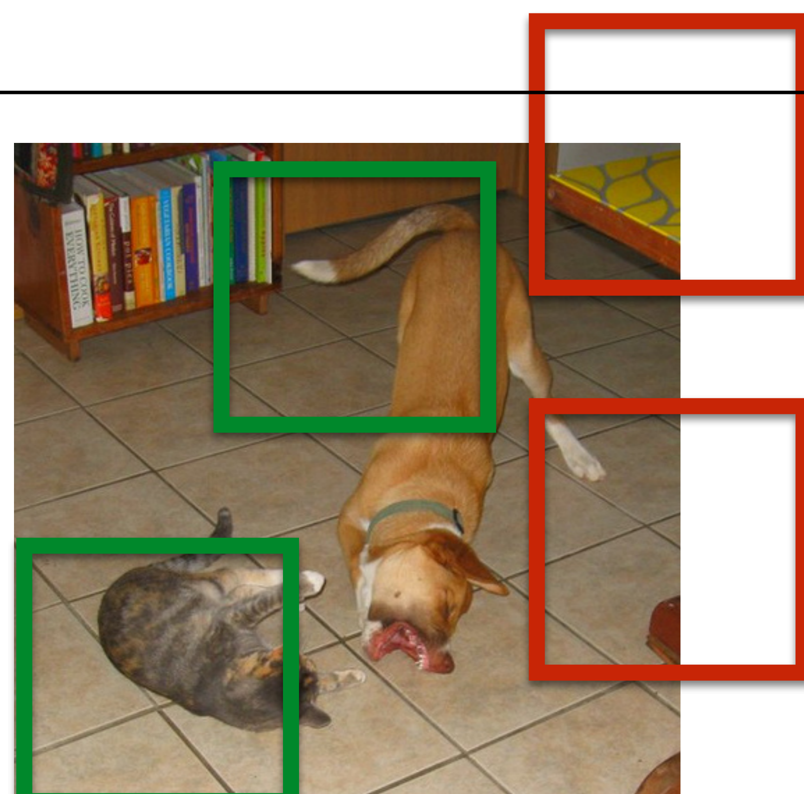
Multi-Scale Feature Maps Experiment

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



boundary boxes

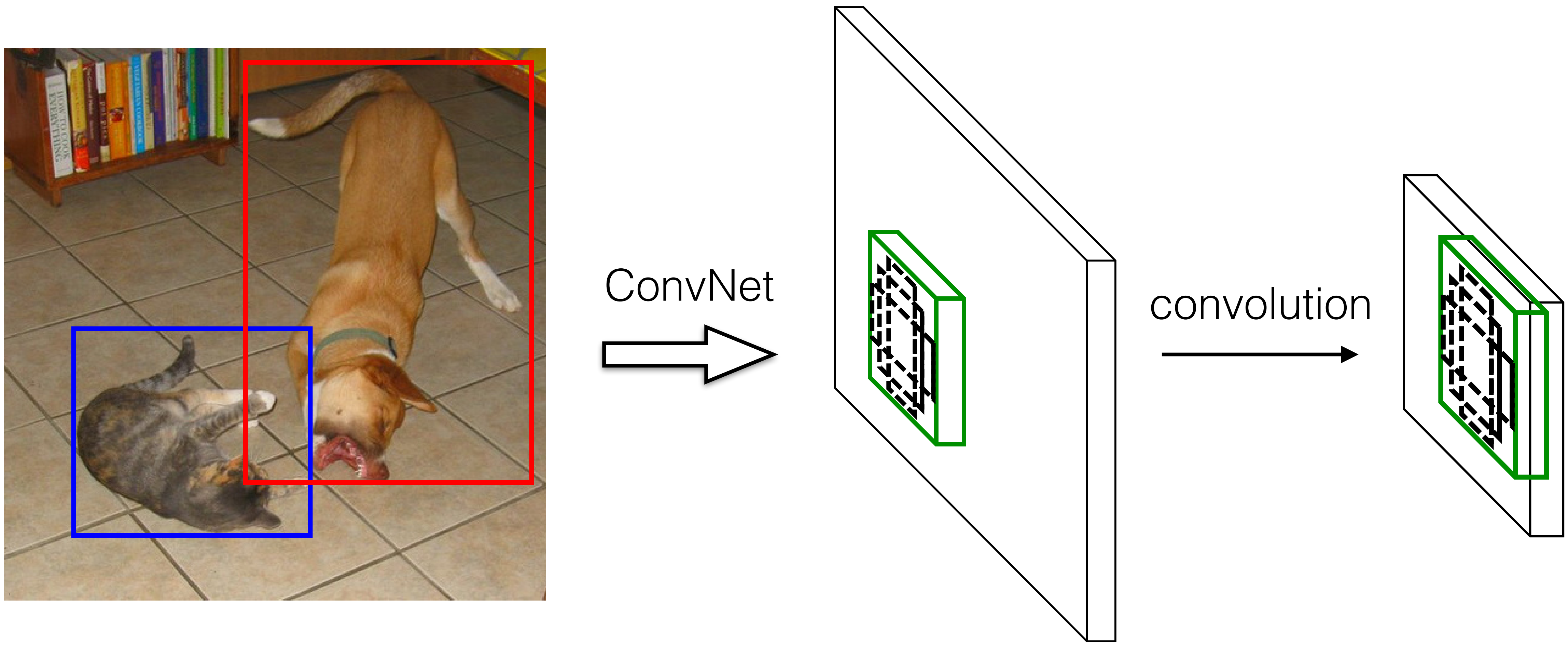
Multi-Scale Feature Maps Experiment

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

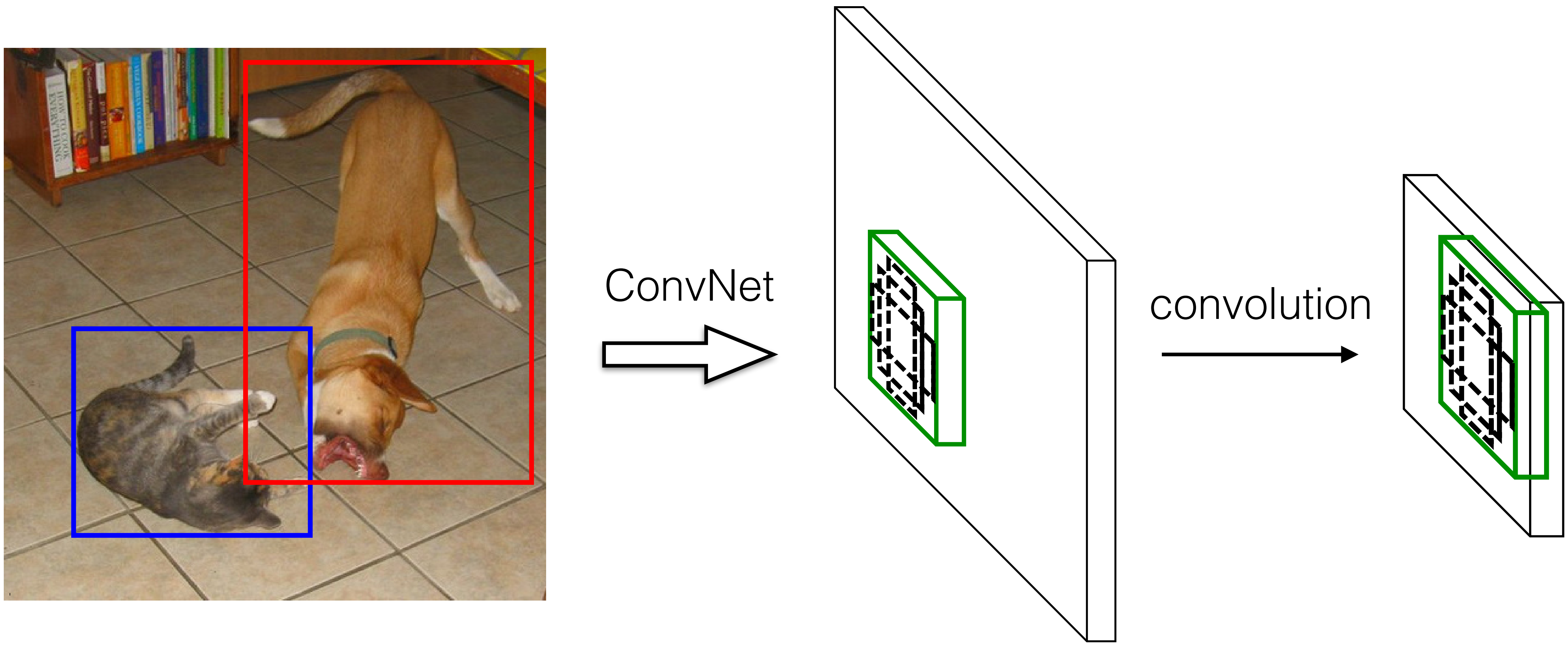
Multi-Scale Feature Maps Experiment

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

Contribution #2: Splitting the Region Space

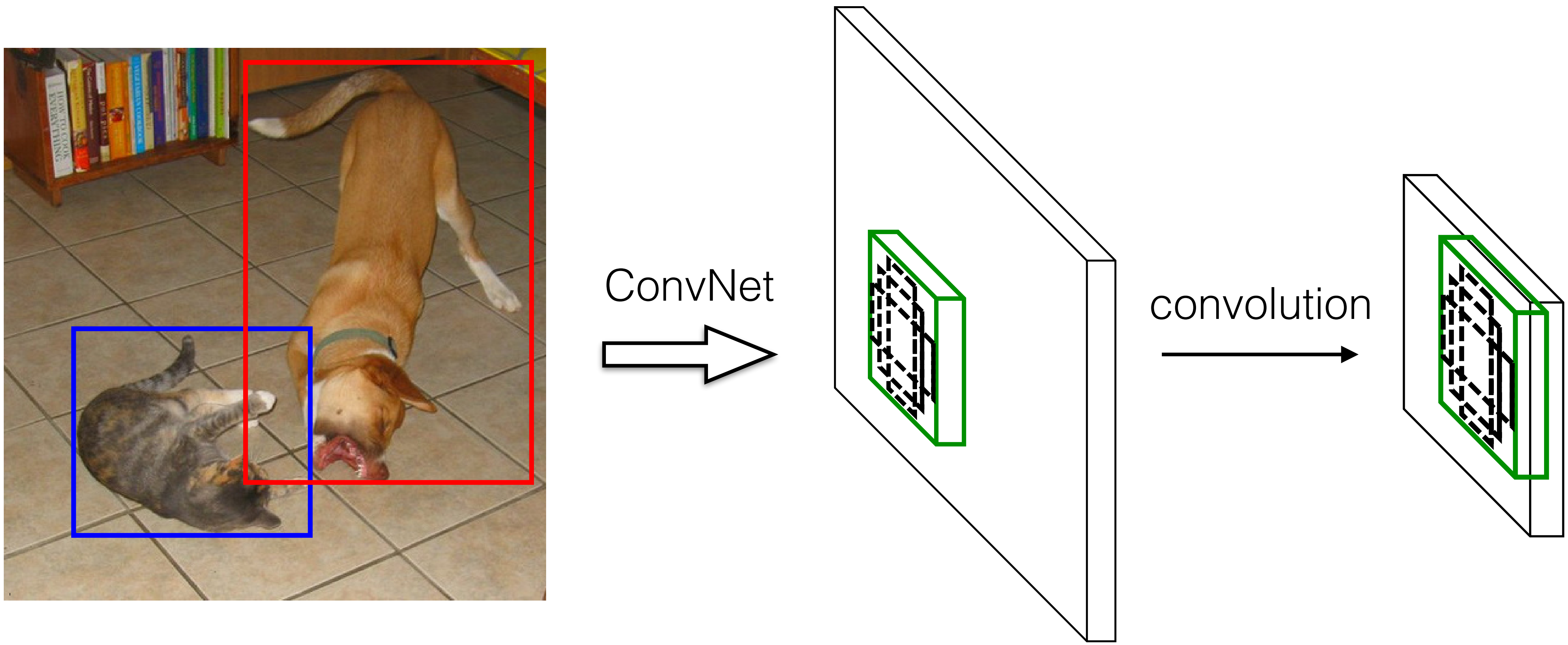


Contribution #2: Splitting the Region Space



	SSD300		
include $\{\frac{1}{2}, 2\}$ box?		✔	✔
include $\{\frac{1}{3}, 3\}$ box?			✔
number of Boxes	3880	7760	**8732**
VOC2007 `test` mAP	71.6	73.7	**74.3**

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

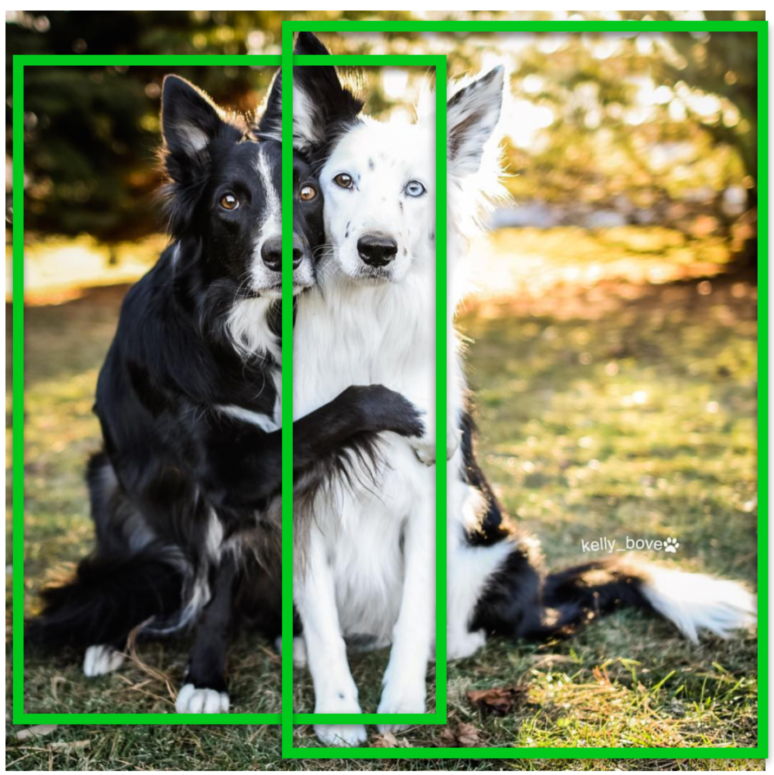
Why So Many Default Boxes?

	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

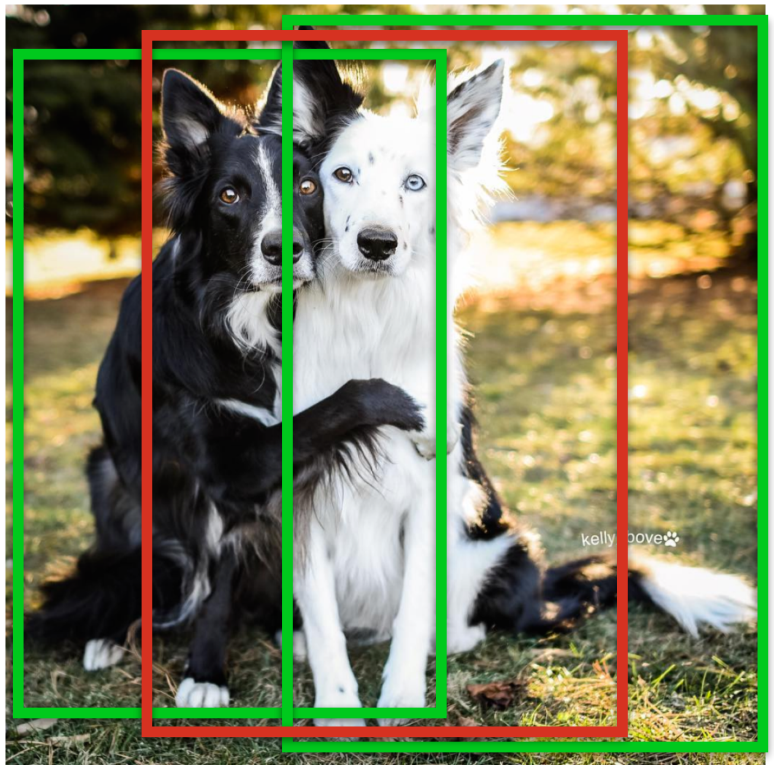
GT



Why So Many Default Boxes?

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

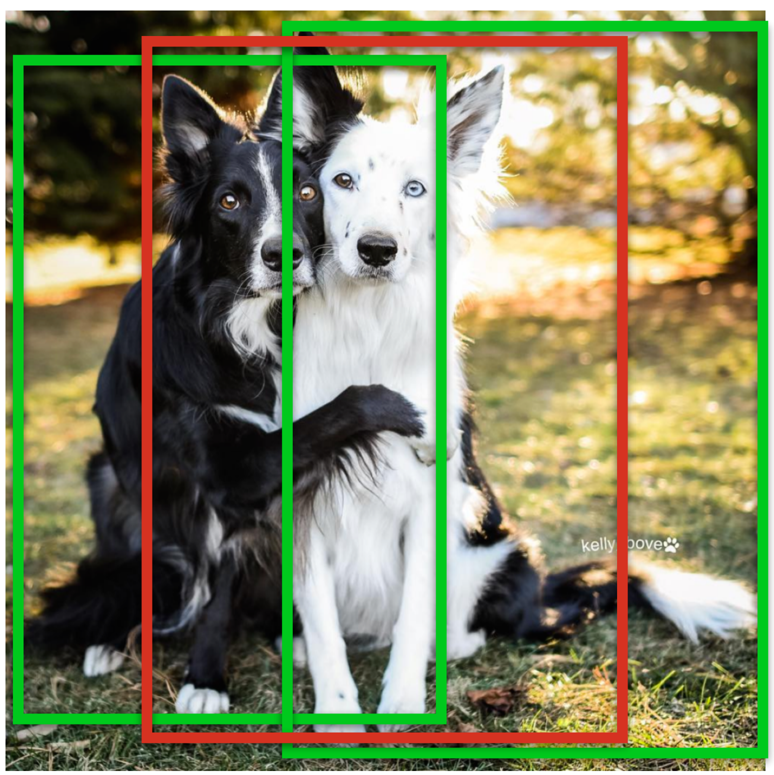
GT DETECTION



Why So Many Default Boxes?

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

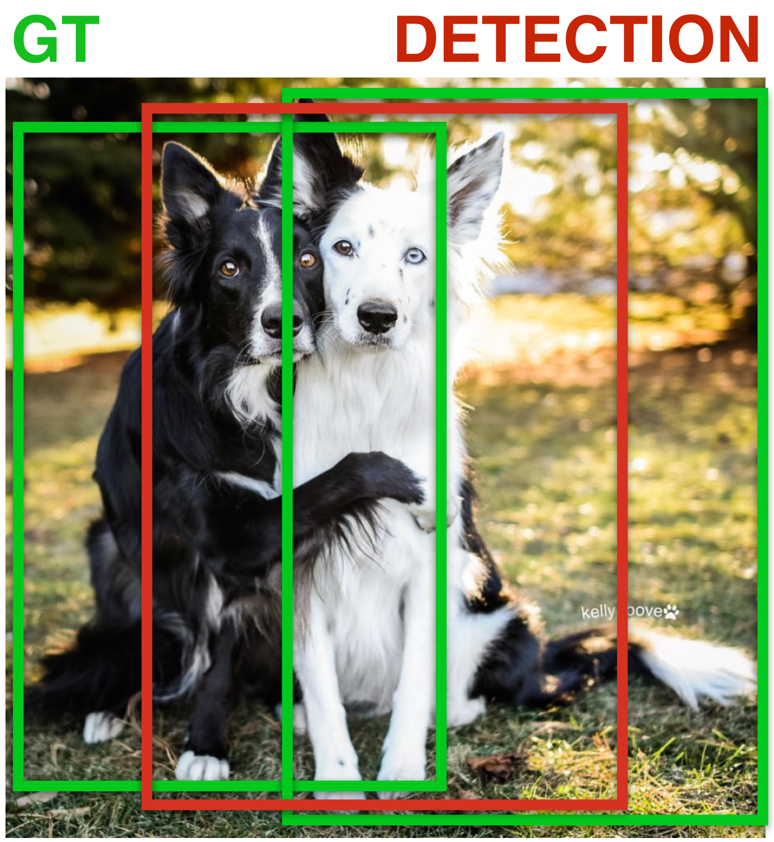
GT DETECTION



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

Why So Many Default Boxes?

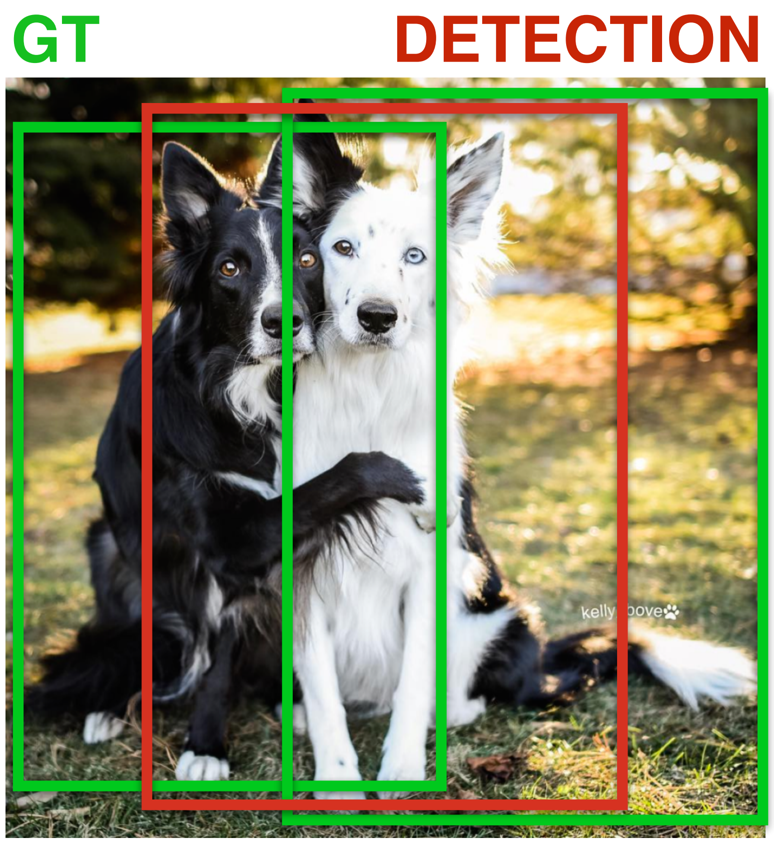
	Faster R-CNN	YOLO	SSD300	SSD512
# Default Boxes	6000	98	8732	24564
Resolution	1000x600	448x448	300x300	512x512



- 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

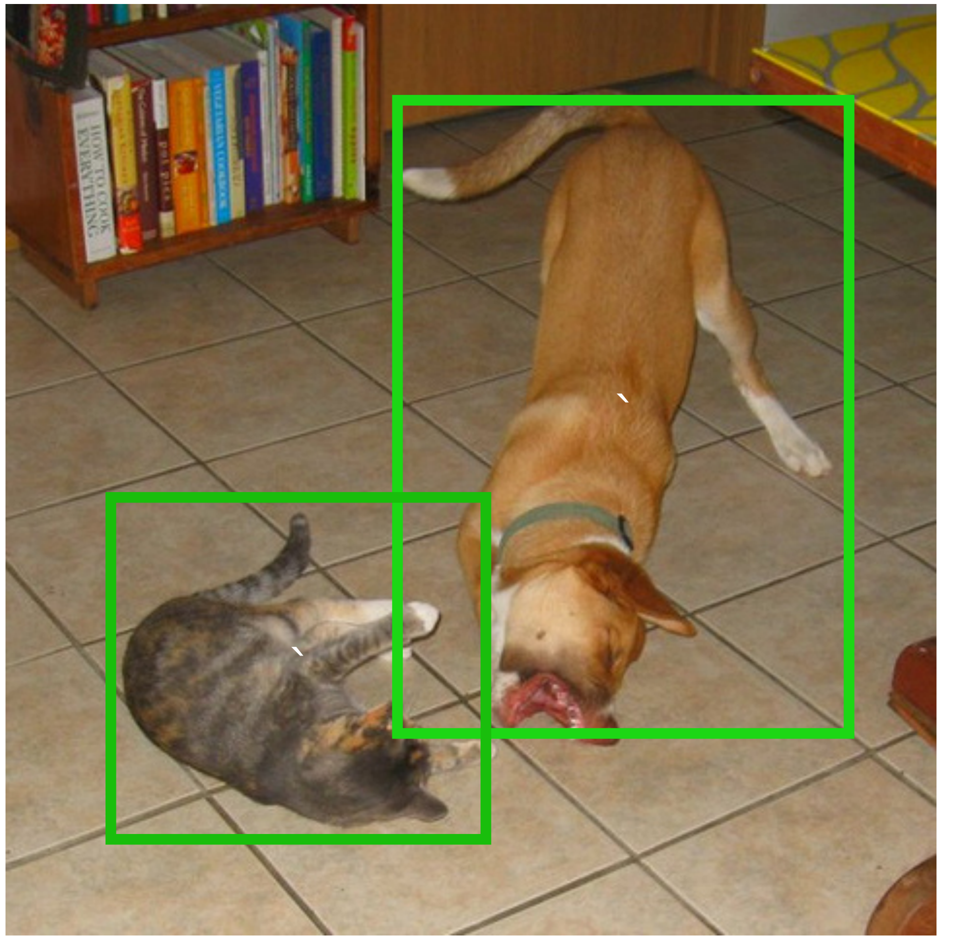
Handling Many Default Boxes

- Matching ground truth and default boxes

Handling Many Default Boxes

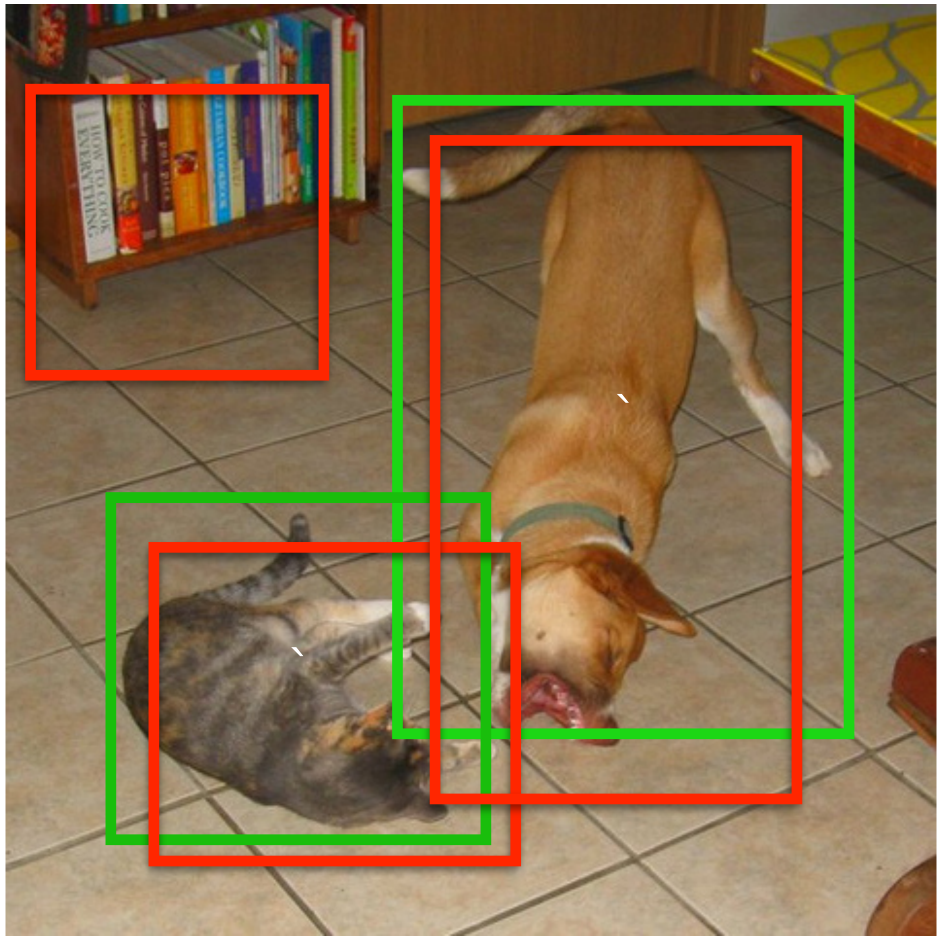
- Matching ground truth and default boxes

GT



Handling Many Default Boxes

- Matching ground truth and default boxes

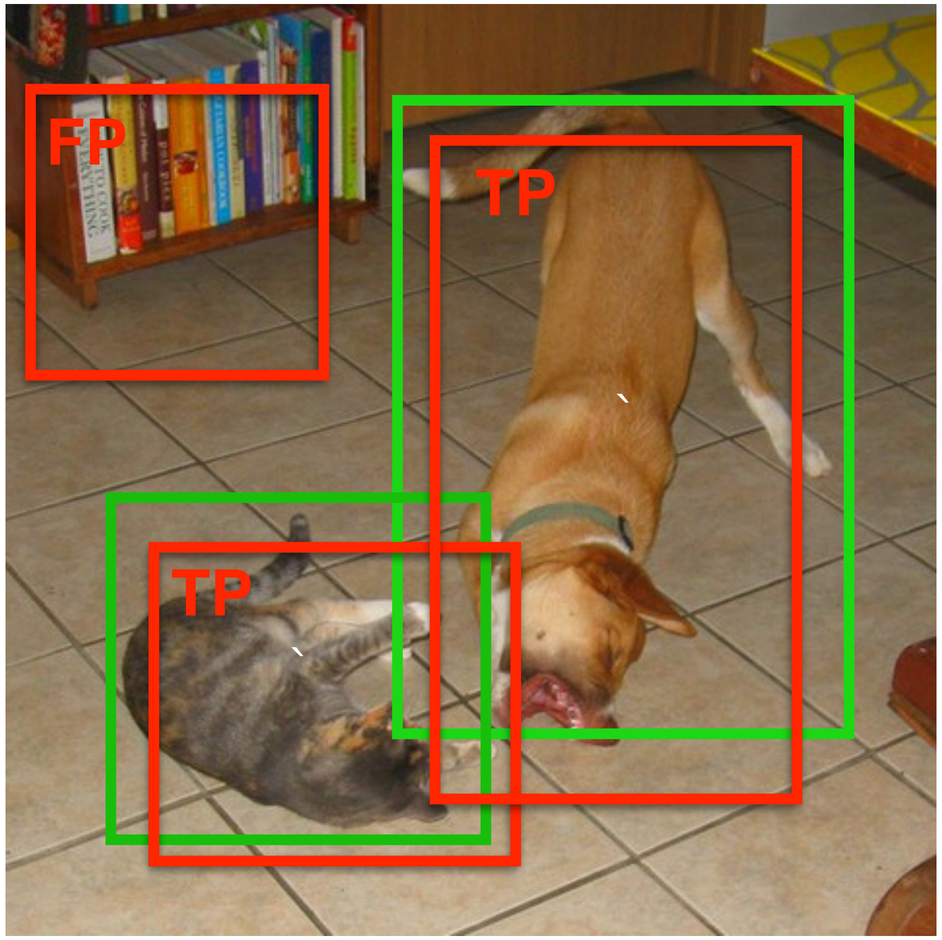


Handling Many Default Boxes

- Matching ground truth and default boxes

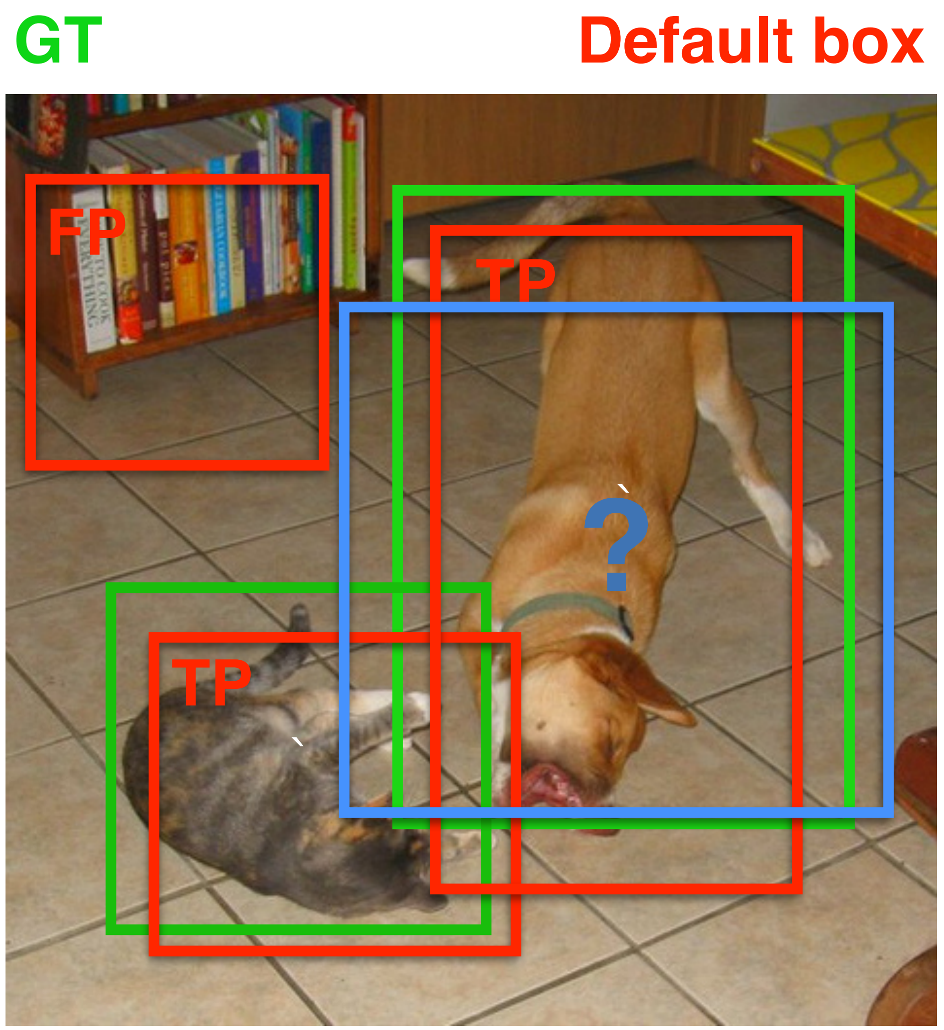
GT

Default box



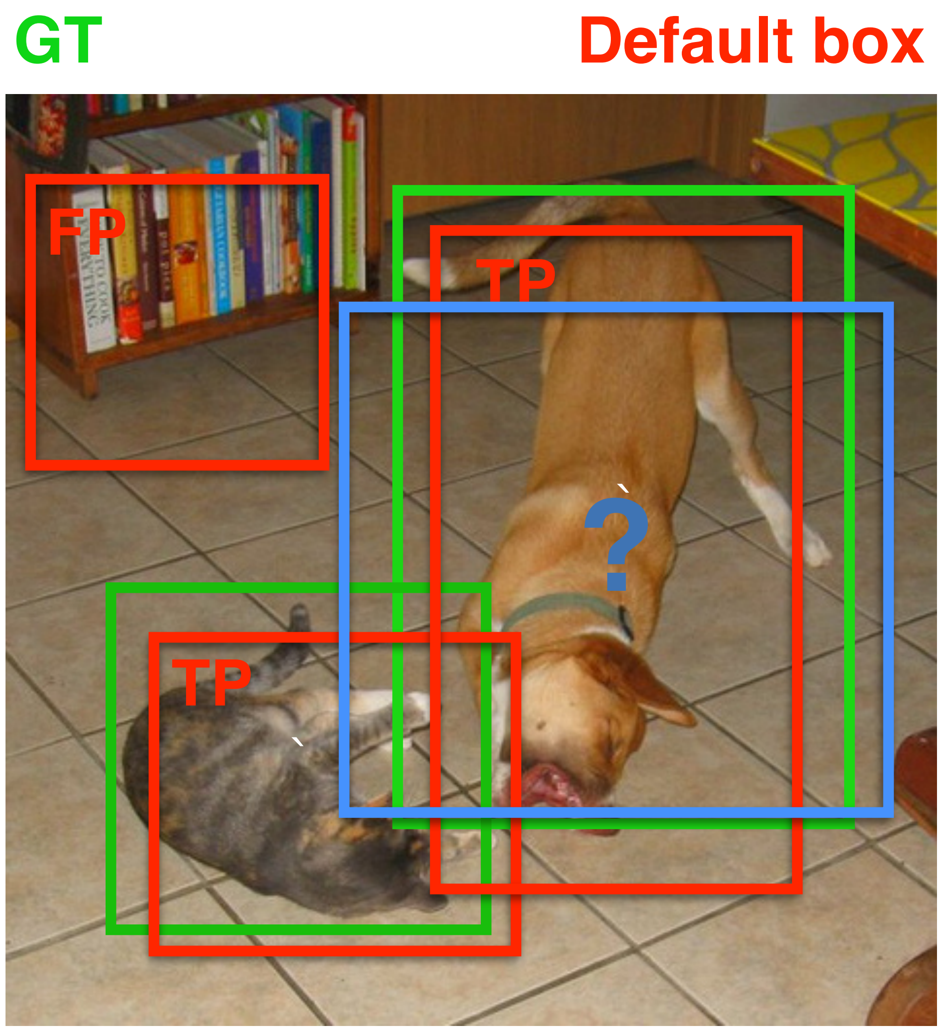
Handling Many Default Boxes

- Matching ground truth and default boxes



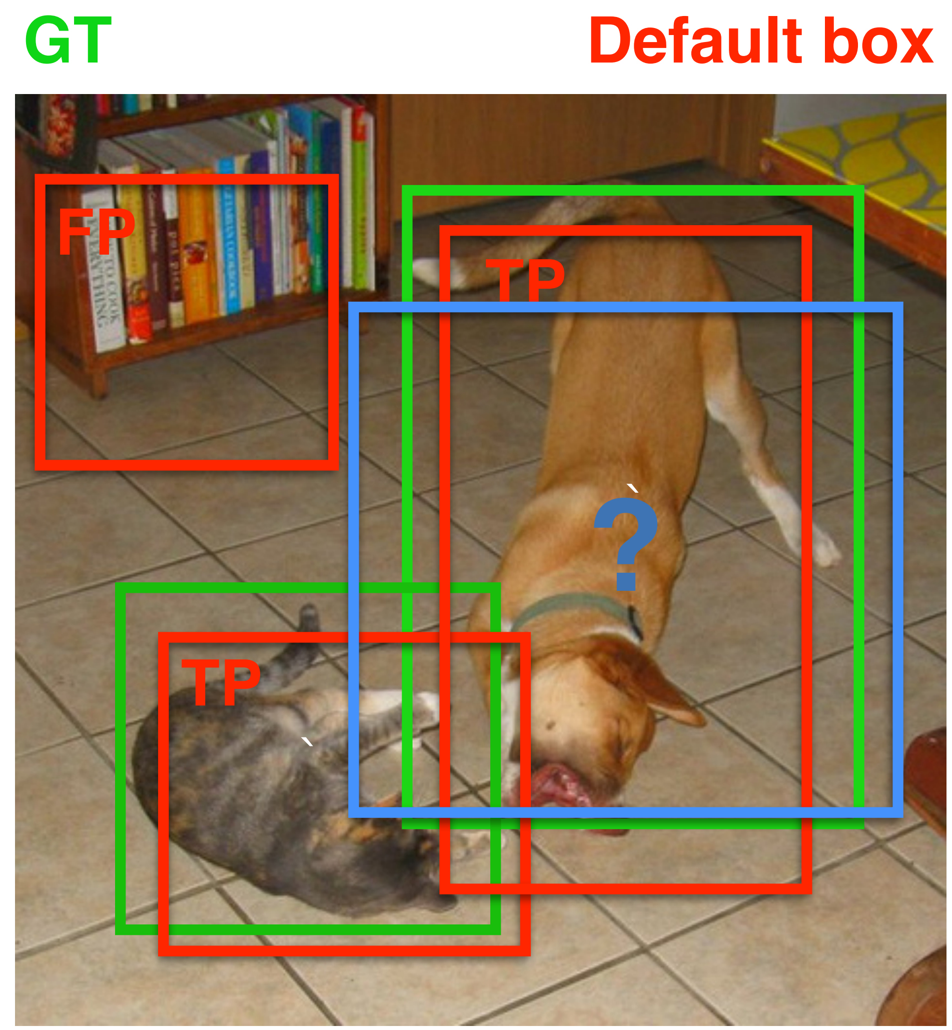
Handling Many Default Boxes

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



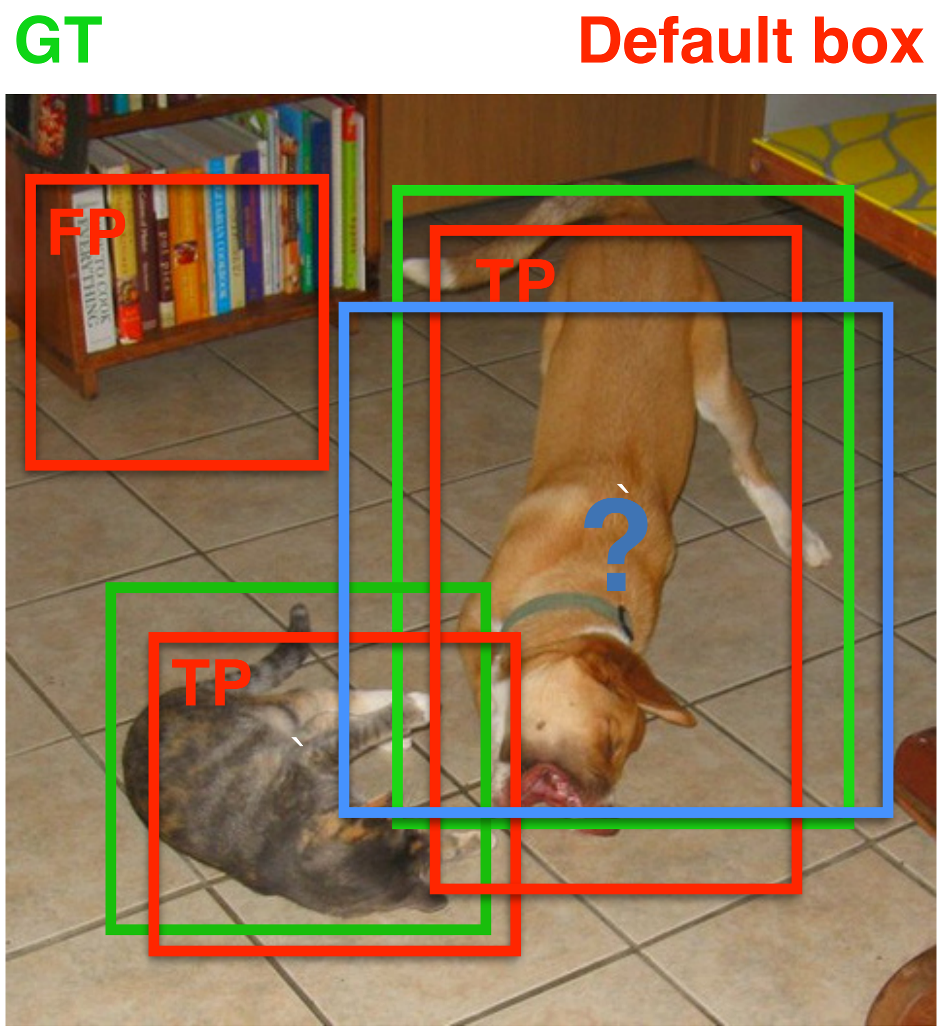
Handling Many Default Boxes

- 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 IoU > 0.5



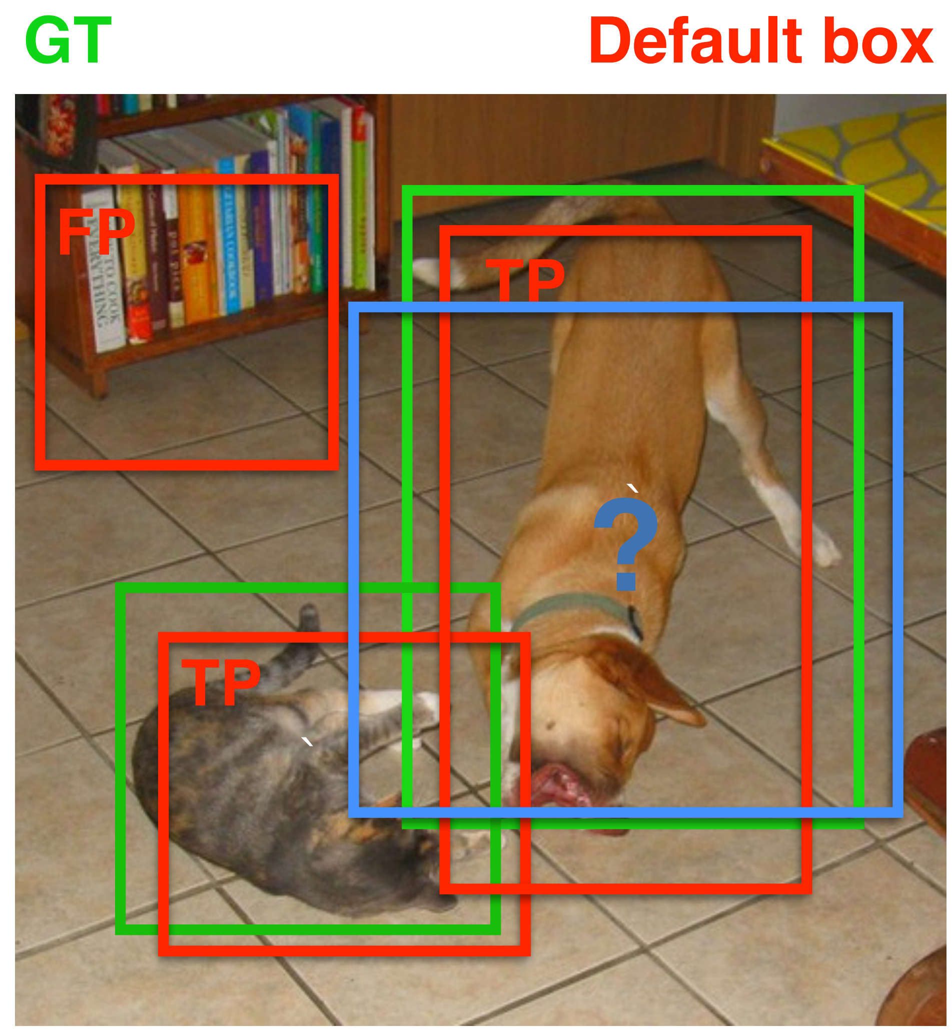
Handling Many Default Boxes

- 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 IoU > 0.5
- Hard negative mining



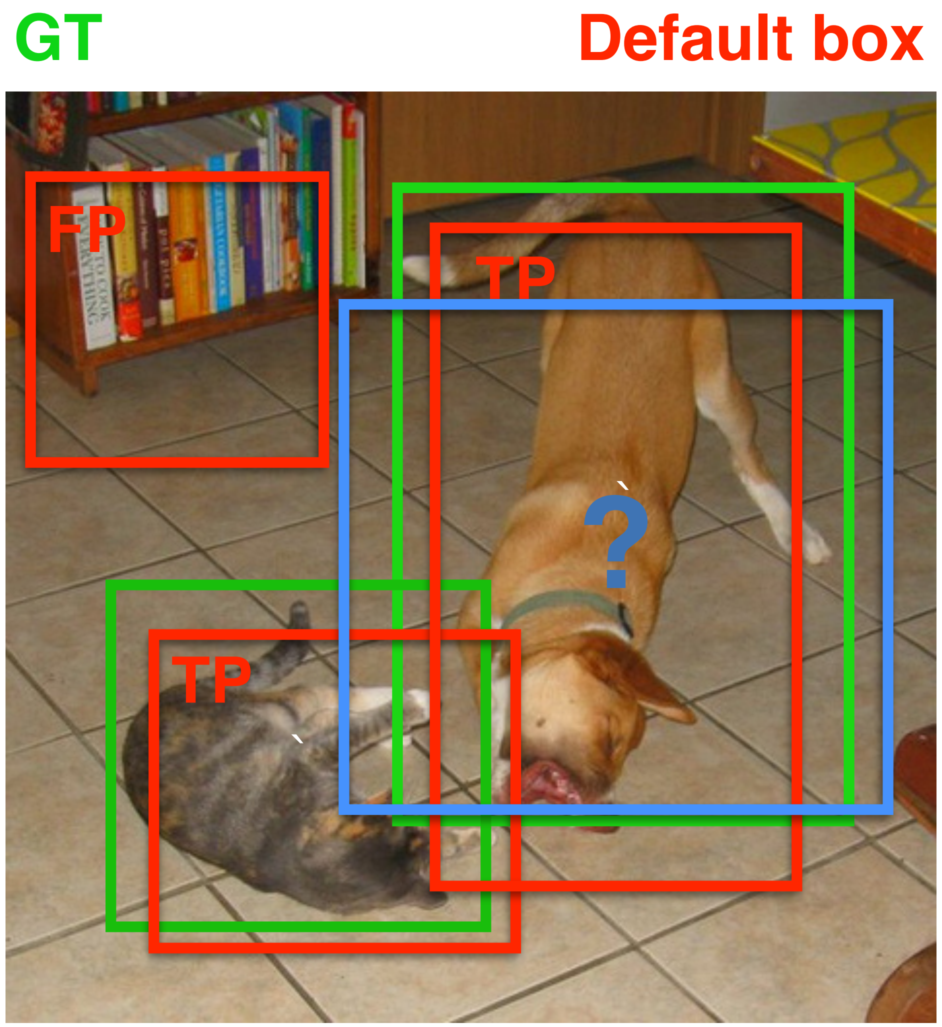
Handling Many Default Boxes

- 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 IoU > 0.5
- Hard negative mining
  - Unbalanced training: 1-30 TP, 8k-25k FP

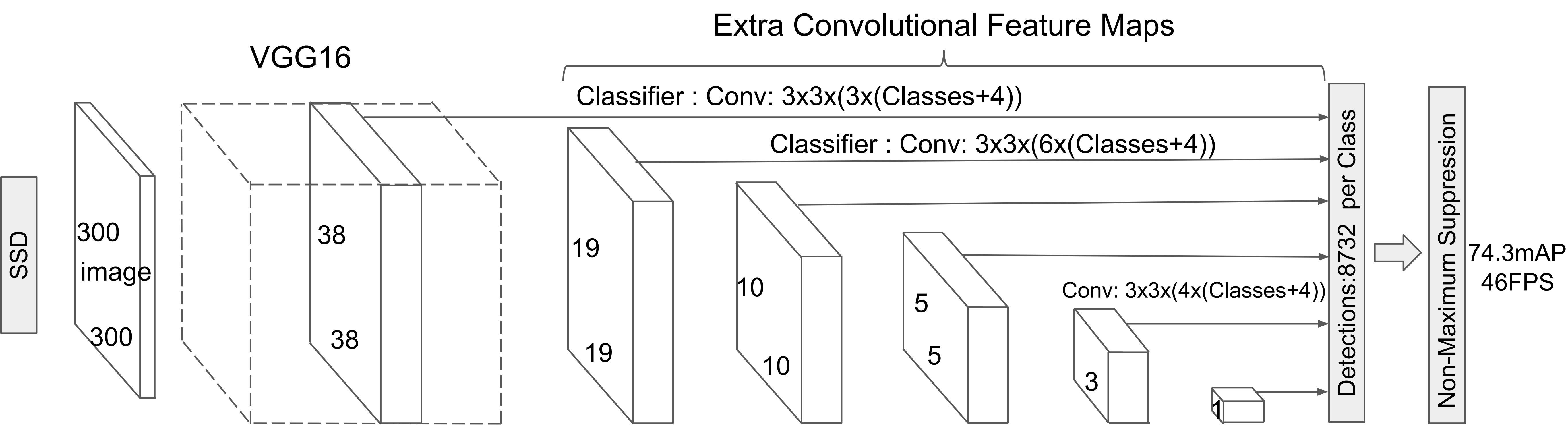


Handling Many Default Boxes

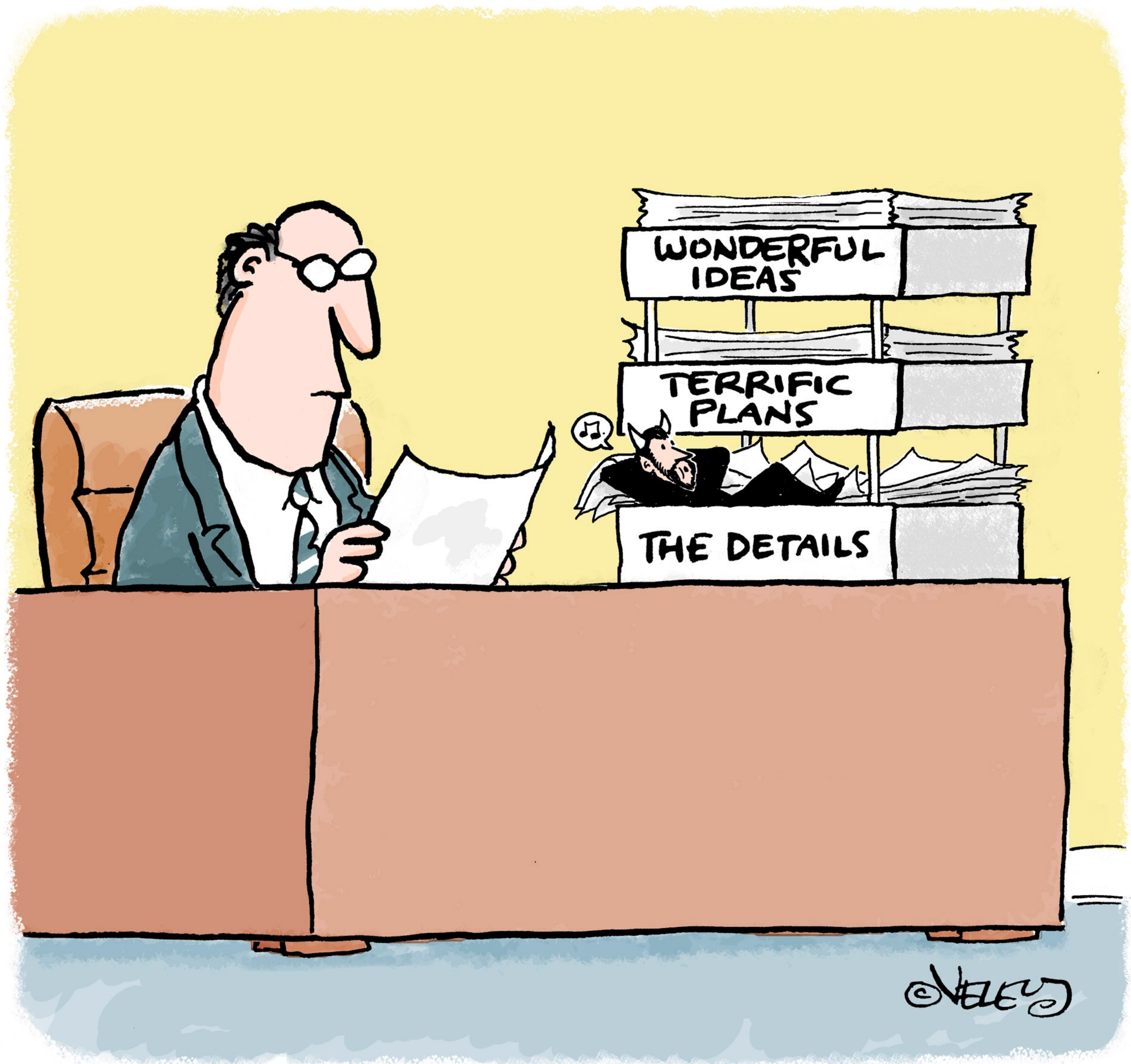
- Matching ground truth and default boxes
  - Match each GT box to closest default box
  - Also match each GT box to all unassigned default boxes with IoU > 0.5
- Hard negative mining
  - Unbalanced training: 1-30 TP, 8k-25k FP
  - Keep TP:FP ratio fixed (1:3), use worst-misclassified FPs.



SSD Architecture

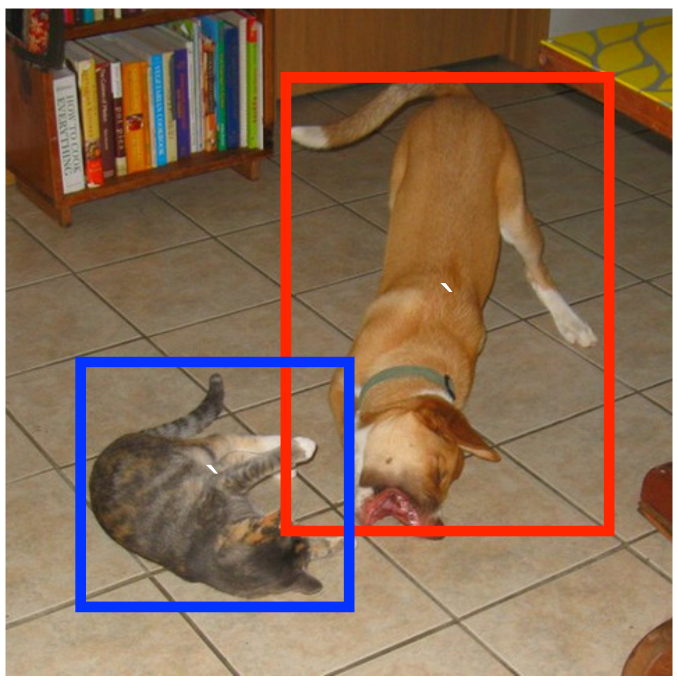


Contribution #3:
The Devil is in the Details

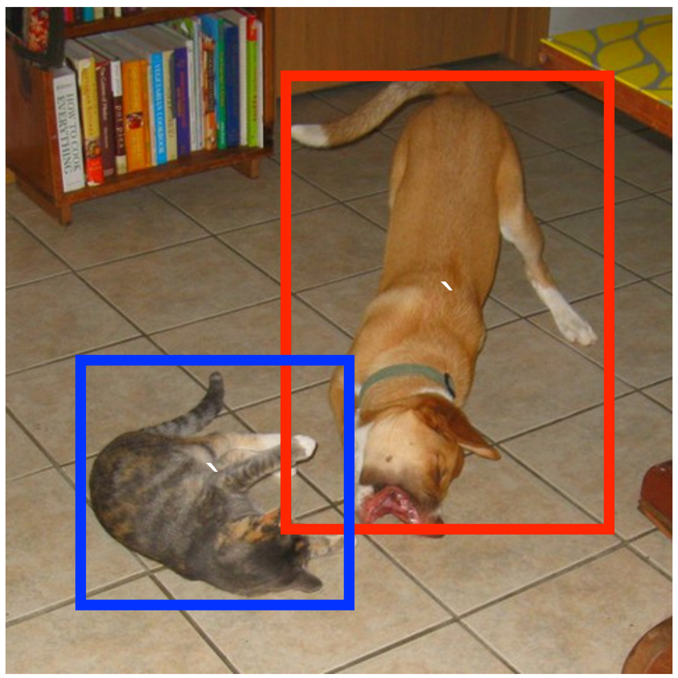


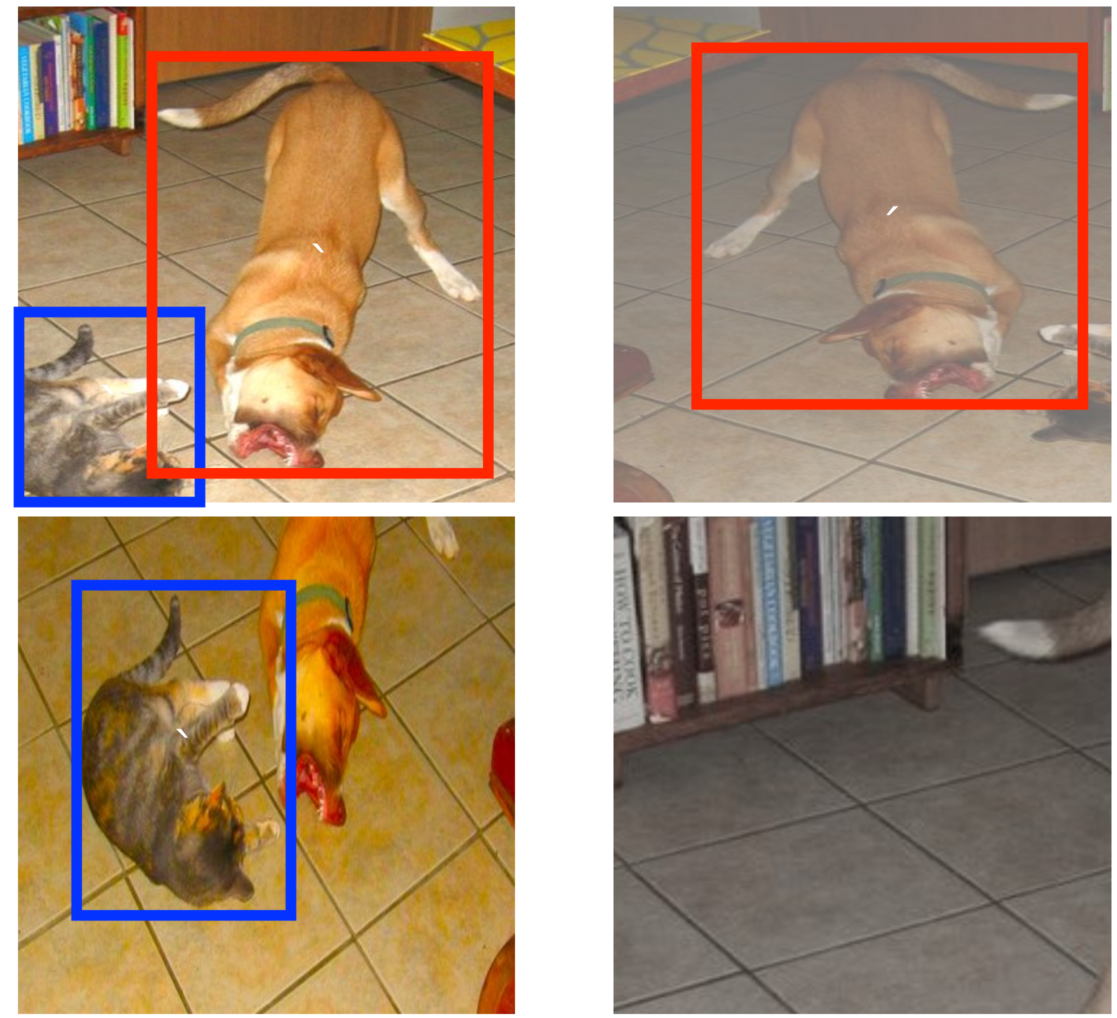
Data Augmentation

Data Augmentation

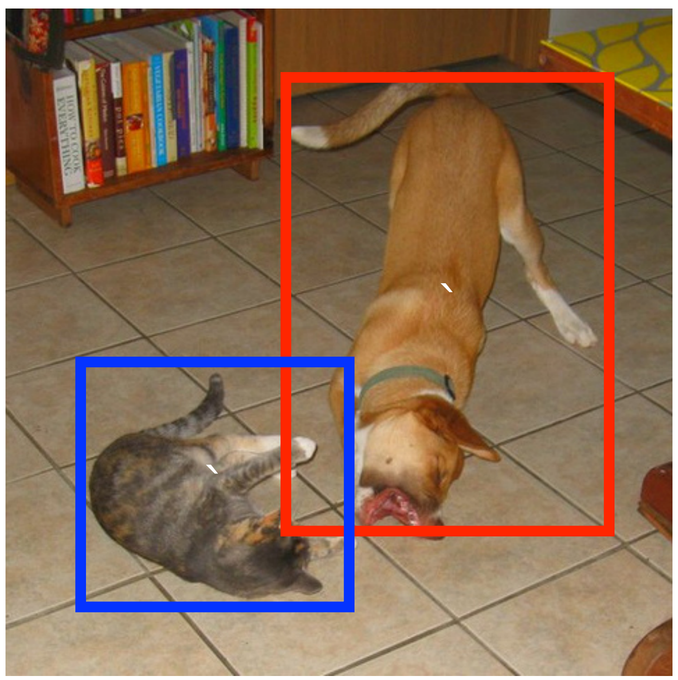


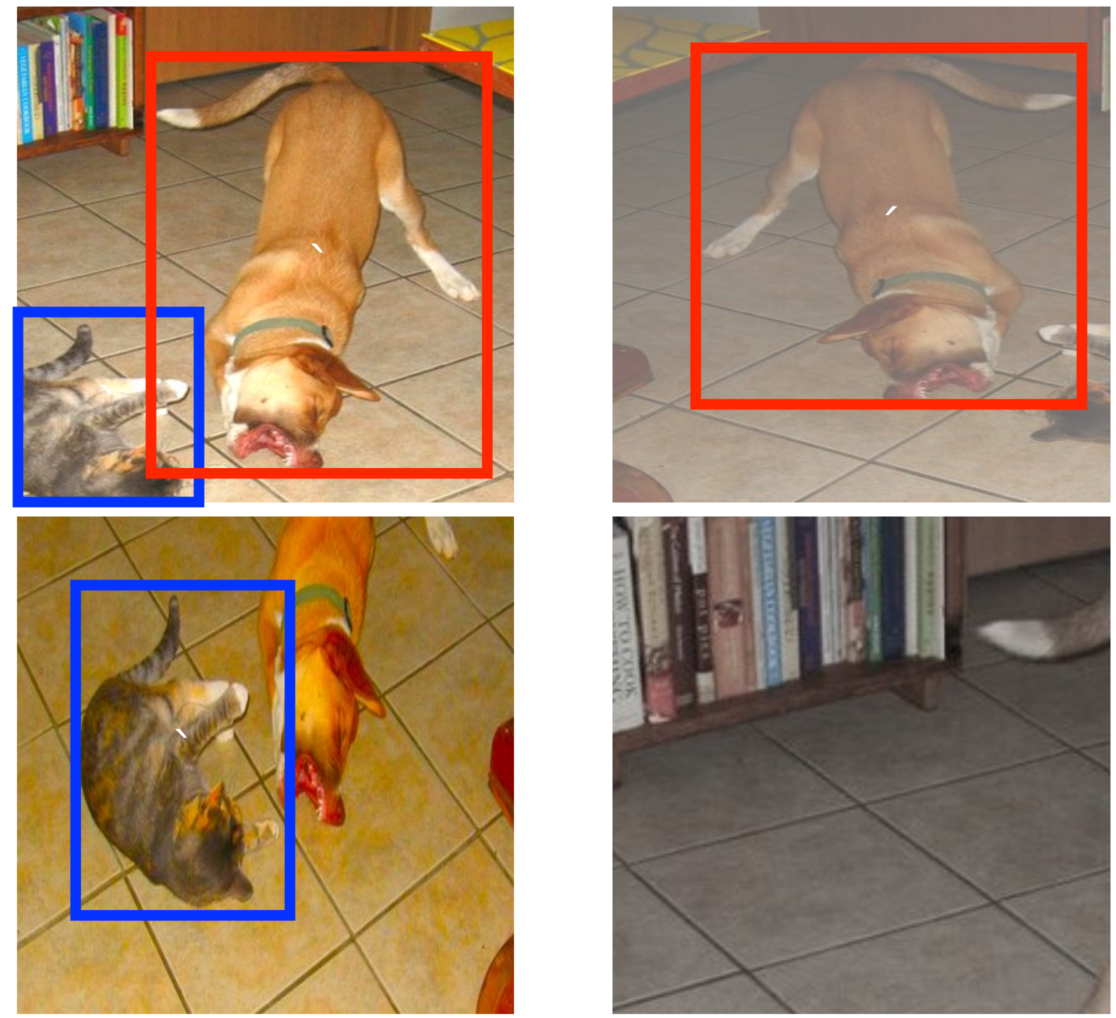
Data Augmentation





Data Augmentation

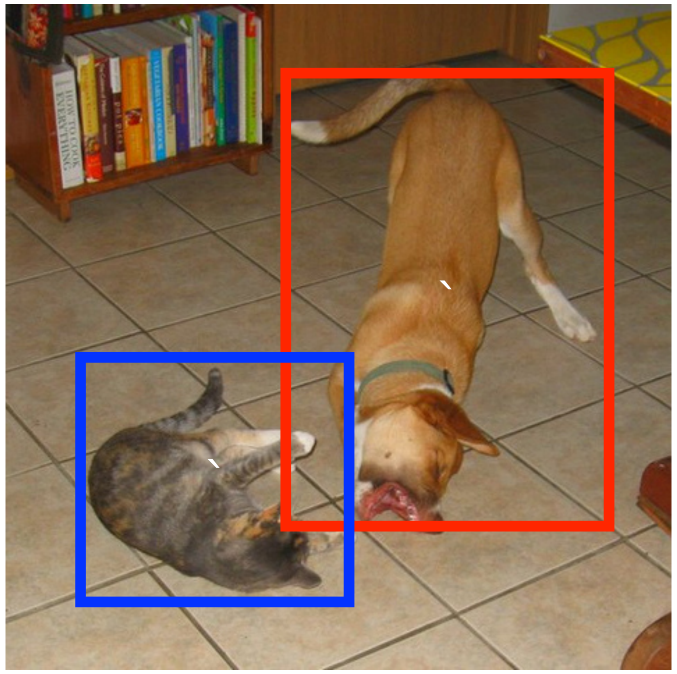




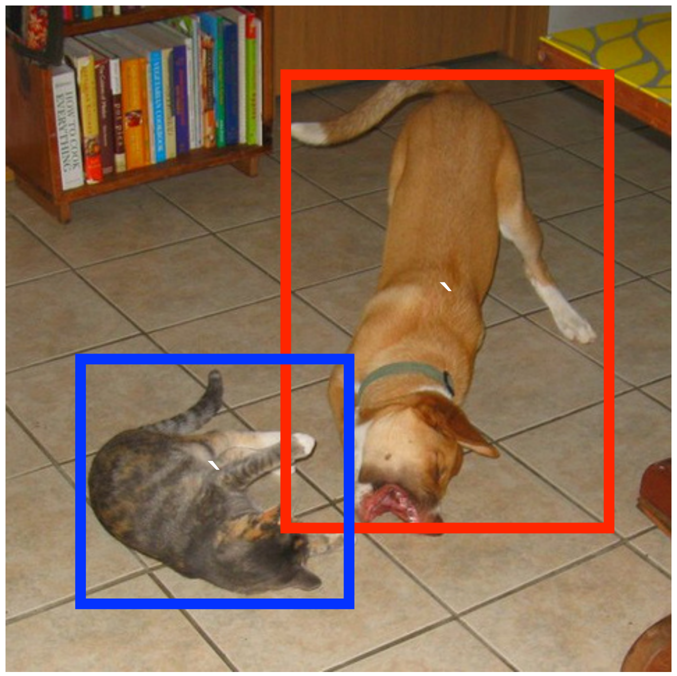
data augmentation	SSD300	
horizontal flip	✔	✔
random crop & color distortion		✔
VOC2007 `test` mAP	65.5	**74.3**

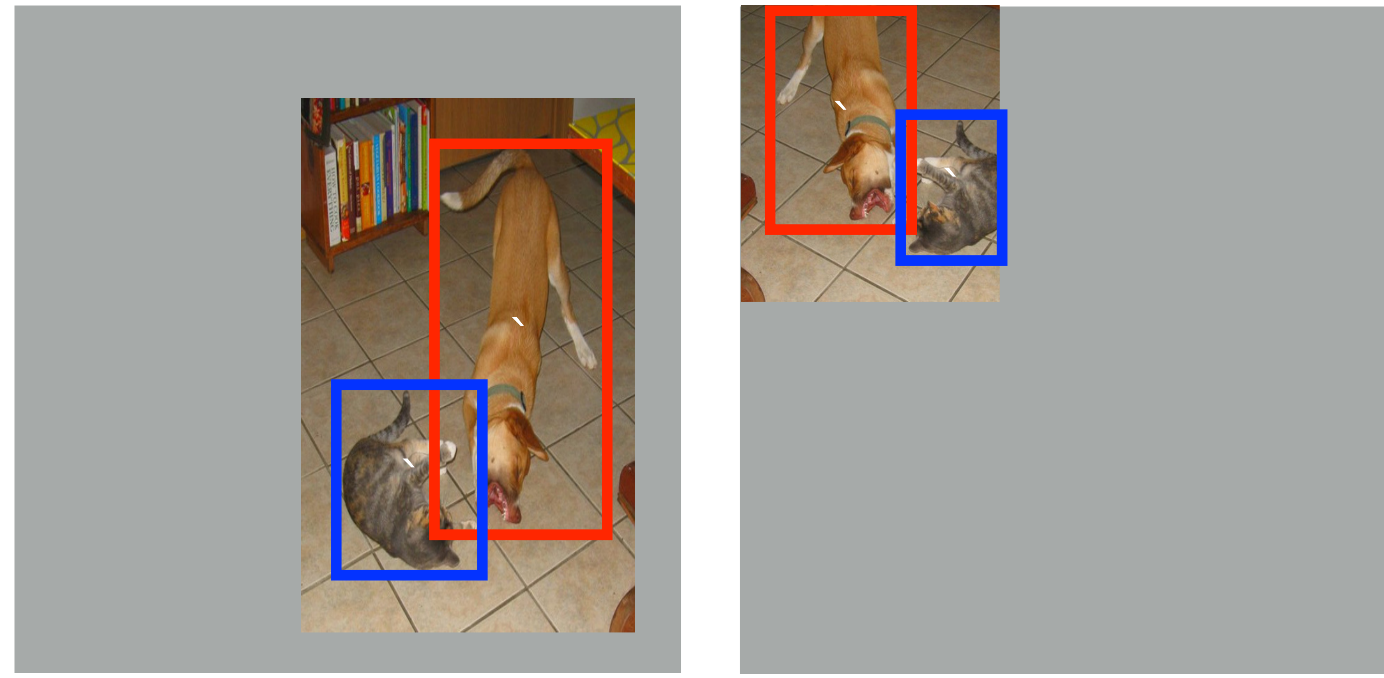
Data Augmentation

Data Augmentation



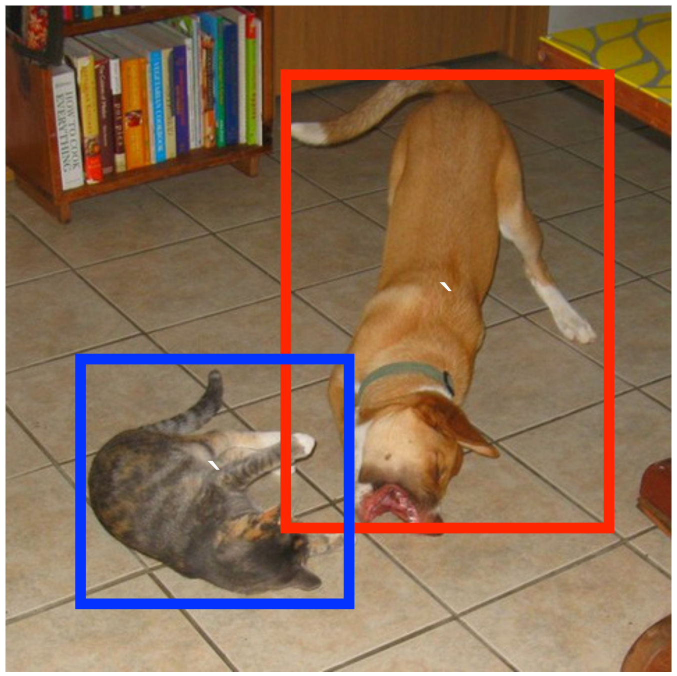
Data Augmentation

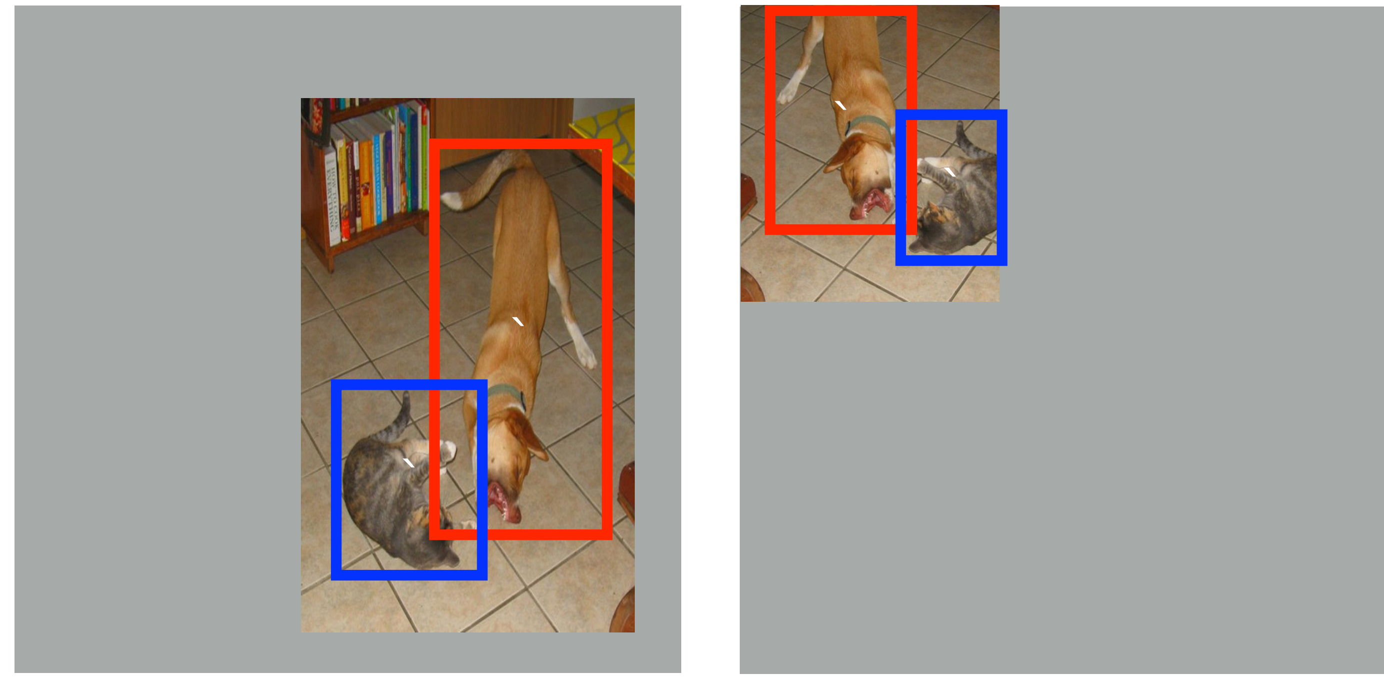




Random expansion creates more **small** training examples

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	**77.2**

Results on VOC2007 test

Method	mAP	FPS	batch size	# Boxes	Input resolution
Faster R-CNN (VGG16)	73.2	7	1	~ 6000	$\sim 1000 \times 600$
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	$\sim$ 6000	$\sim 1000 \times 600$
Fast YOLO	52.7	155	1	98	448×448
YOLO (VGG16)	66.4	21	1	98	448×448
SSD300 **6.6x**	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	$\sim 1000 \times 600$
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	$\sim 1000 \times 600$
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 **10%↑**	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	$\sim 1000 \times 600$
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	$\sim 1000 \times 600$
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	$\sim 1000 \times 600$
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	$\sim 1000 \times 600$
Fast YOLO	52.7	155	1	98	448×448
YOLO (VGG16)	66.4	21	1	98	448×448
SSD300	77.2 ~~74.3~~	46	1	8732	300×300
SSD512	79.8 ~~76.8~~	19	1	24564	512×512
SSD300	77.2 ~~74.3~~	59	8	8732	300×300
SSD512	79.8 ~~76.8~~	22	8	24564	512×512

Results on More Datasets

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

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**

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	**76.8**	**74.9**	**26.8**	**46.4**

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*	**79.8**	**78.5**	**28.8**	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

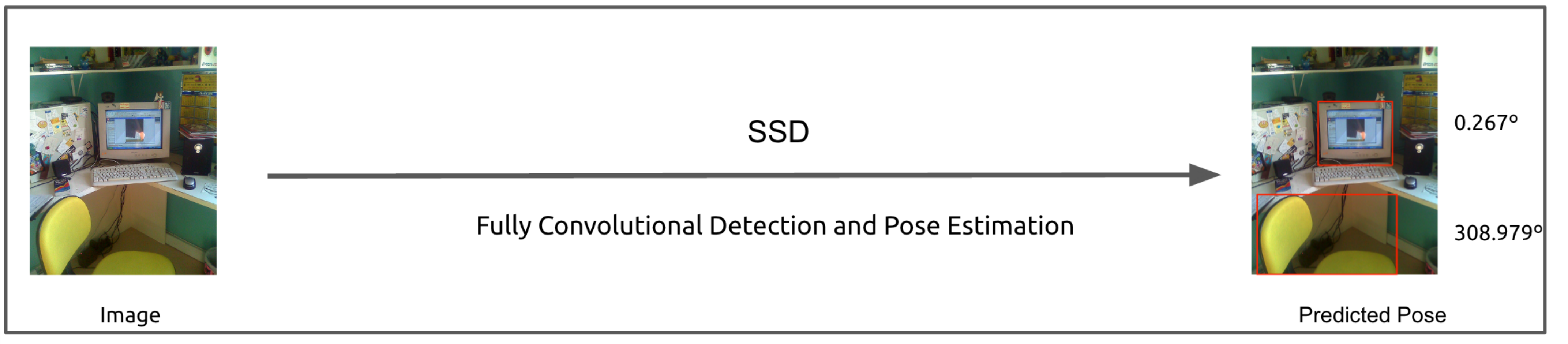
Future Work

- Object detection + pose estimation

Future Work

- Object detection + pose estimation

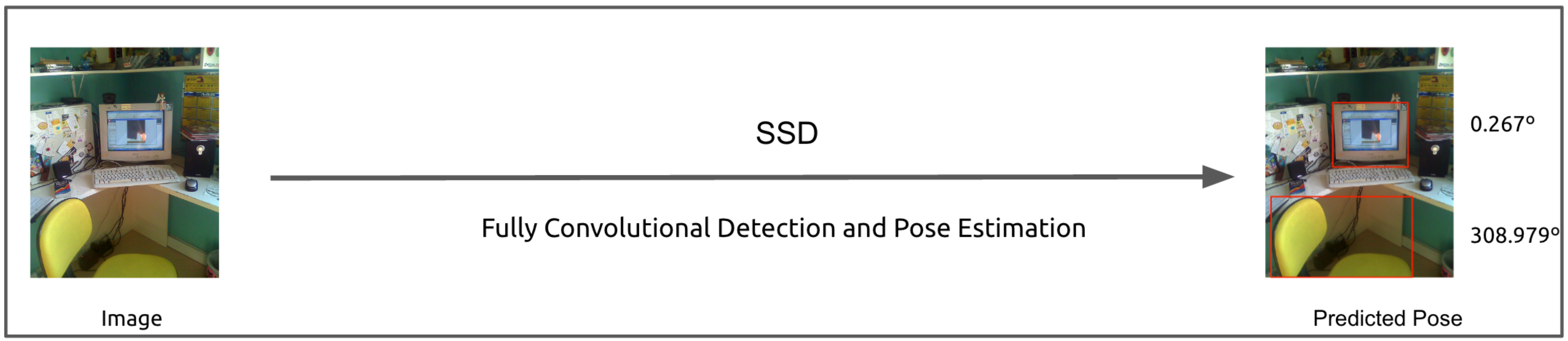
[Poirson et al, coming out at 3DV, 2016]



Future Work

- Object detection + pose estimation

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

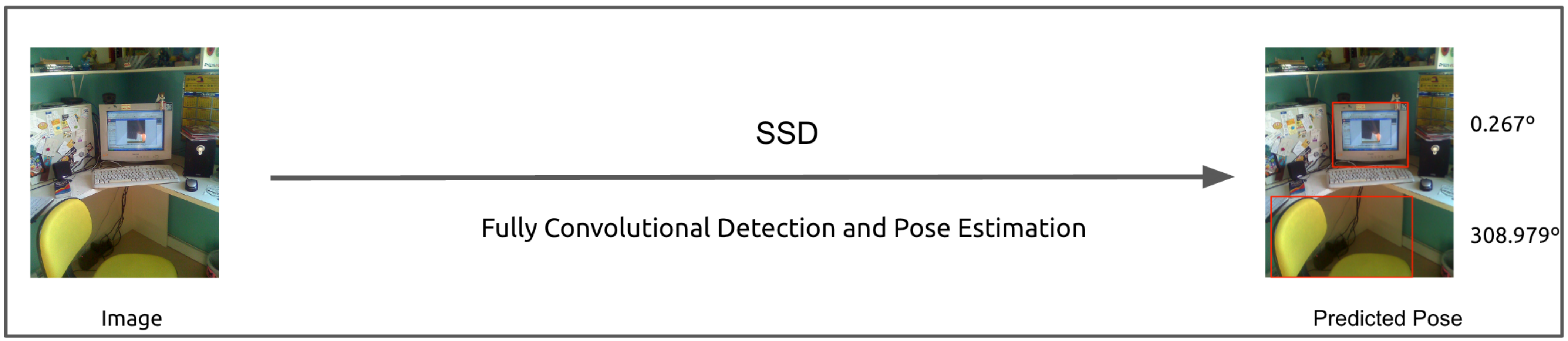


- Single shot 3D bounding box detection

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