Neural Object Detection

Hyun Soo Park

IMAGE CLASSIFICATION VS. OBJECT DETECTION

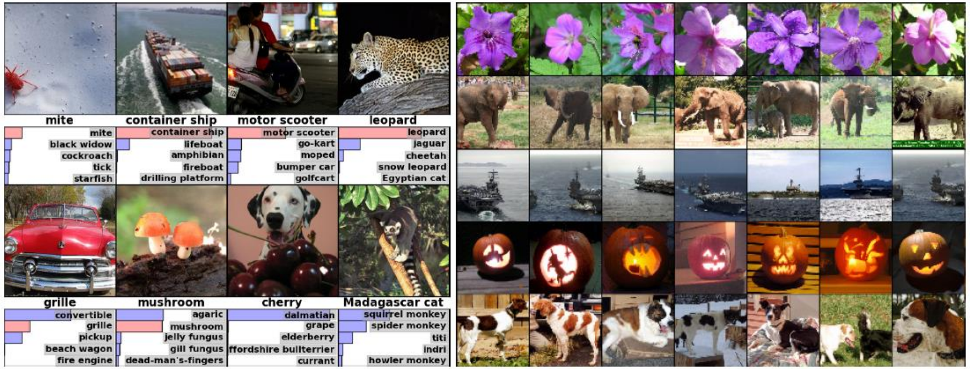
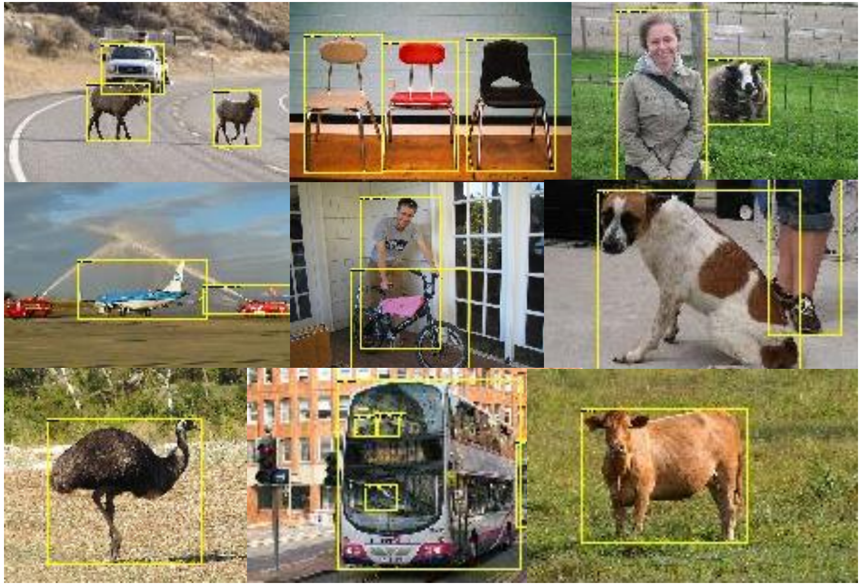


Image classification



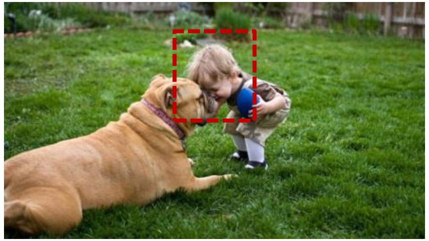
Object detection

Object Detection Pipeline

Input image

Object Detection Pipeline

Input image

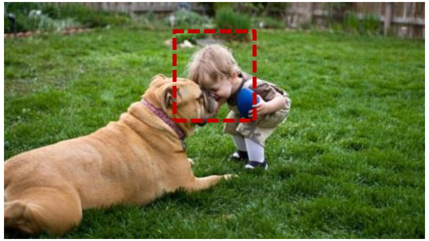


Localization

Sliding window

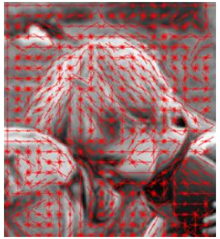
Object Detection Pipeline

Input image



Localization

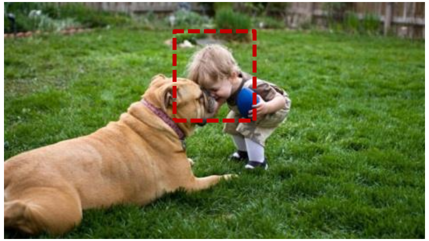
Sliding window

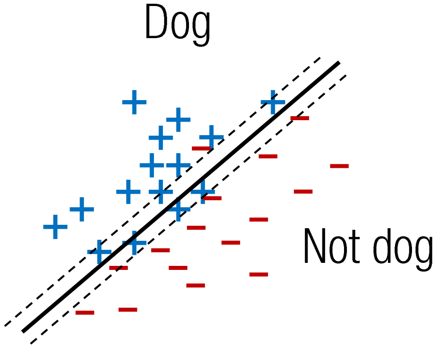


Feature extraction

HOG/SIFT/BoW

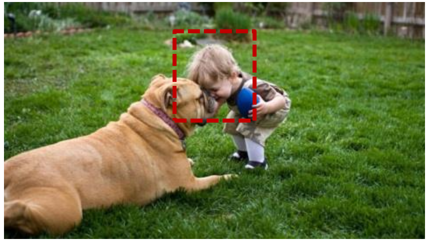
OBJECT DETECTION PIPELINE

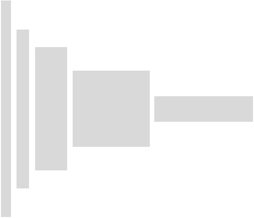


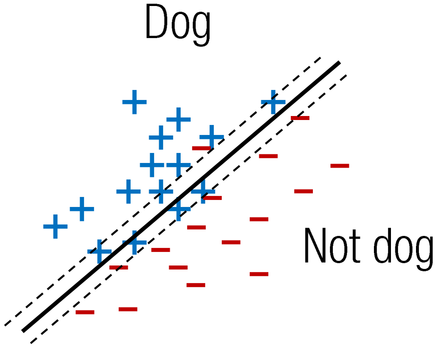


Input image	Localization	Feature extraction	Classification
	Sliding window	HOG/SIFT/BoW	SVM
Limitations	Slow ~ 1M of evaluations	Shallow	n-classifiers

R-CNN (Girshick et al.)

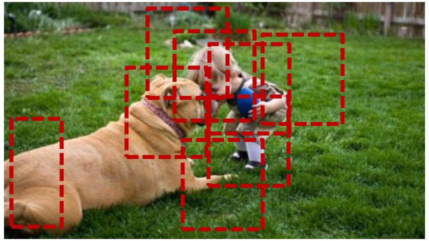


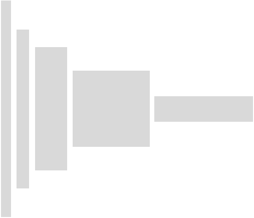


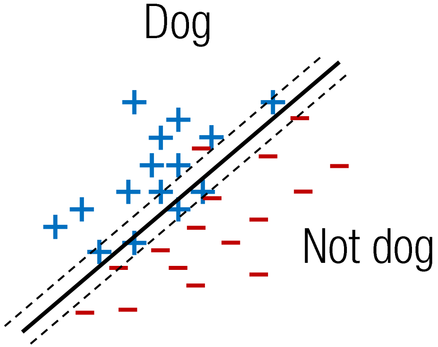


Input image	Localization	Feature extraction	Classification
	Sliding window	**CNN**	SVM
Limitations	Slow ~ 1M of evaluations	**Deep**	n-classifiers

R-CNN (GIRSHICK ET AL.)



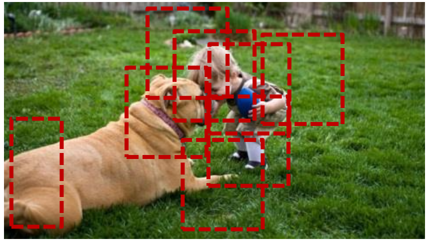




Input image	Localization	Feature extraction	Classification
	Region proposal	**CNN**	SVM
Limitations	**2000 evluations**	**Deep**	n-classifiers

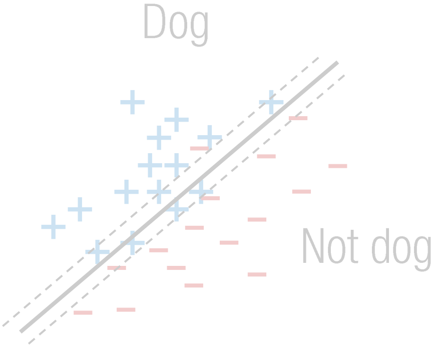
R-CNN (Girshick et al.)

Input image



Localization

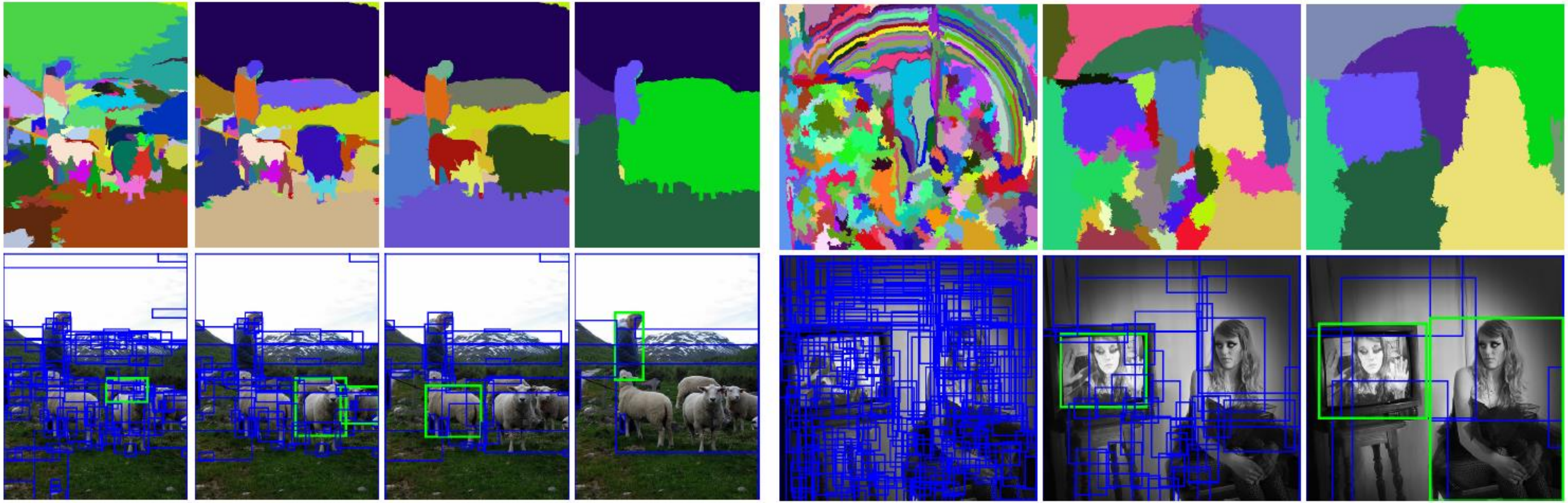
Feature extraction



Classification

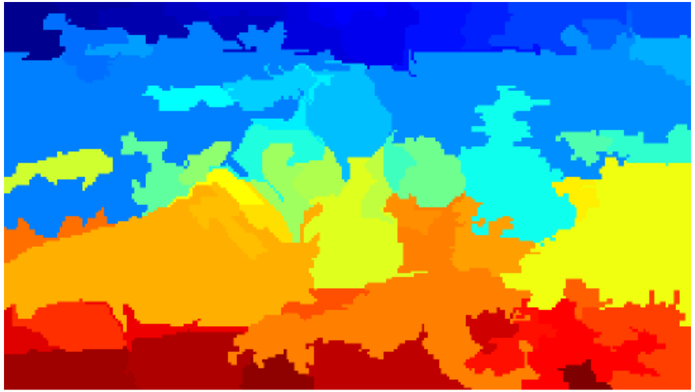
	Region proposal	CNN	SVM
Limitations	2000 evluations	Deep	n-classifiers

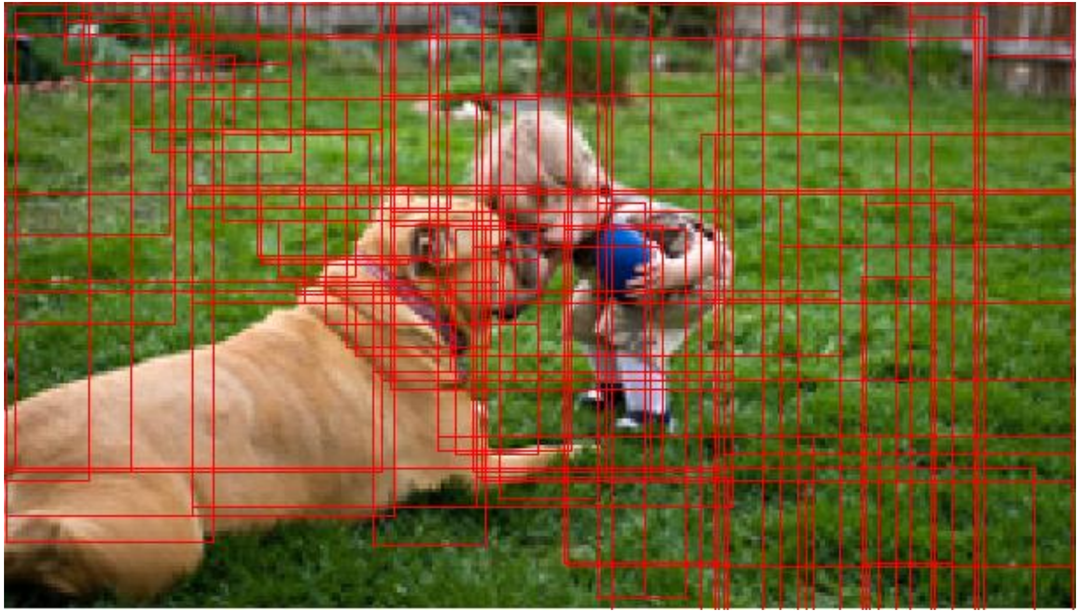
Objectness (Selective Search~Uijlings et al.)



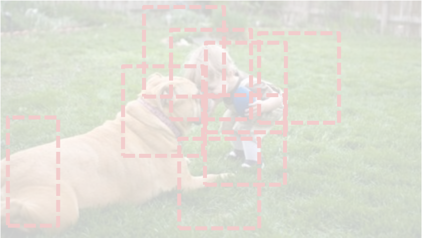
- Merging regions from over-segmentation
- Objectness classification via BoW on merged regions

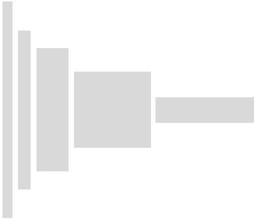
Objectness (Selective Search~Uijlings et al.)

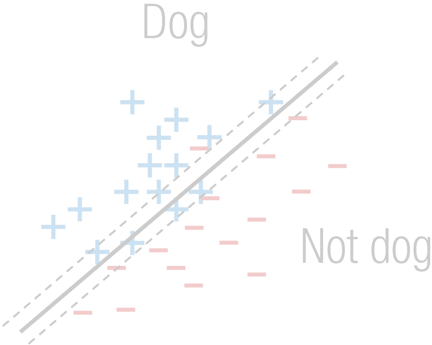




R-CNN (Girshick et al.)



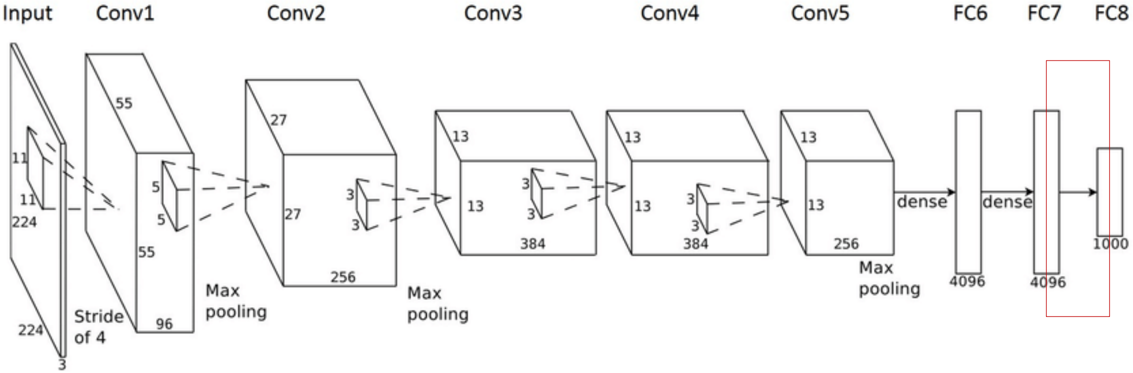




Input image	Localization	Feature extraction	Classification
	Region proposal	CNN	SVM
Limitations	2000 evluations	Deep	n-classifiers

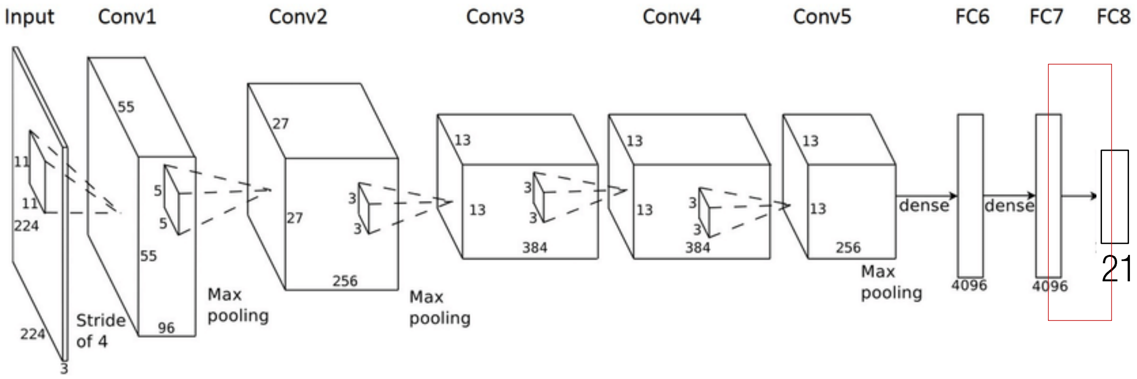
DOMAIN ADAPTATION (IMAGE CLASS. → OBJECT DET.)

1000 image classes (~15M)



4096x1000

Domain Adaptation (Image Class. → Object Det.)



1000 image classes (~15M)

↓

20 object classes (~20K)

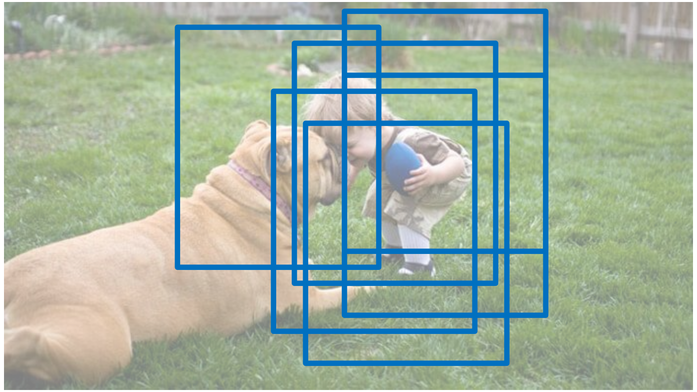
4096x21

←

$$\frac{\partial L}{\partial \mathbf{w}_n}$$

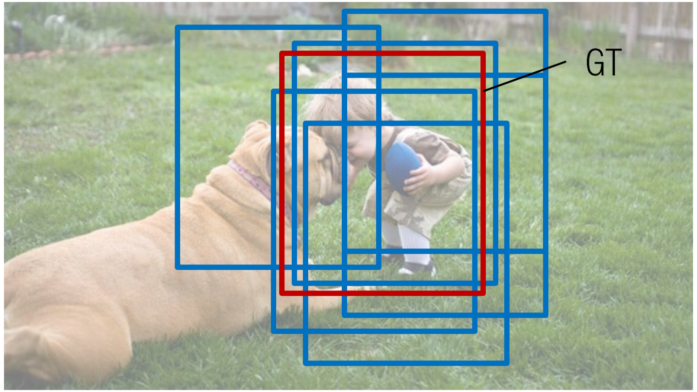
w/ small learning rate

REGION-CNN



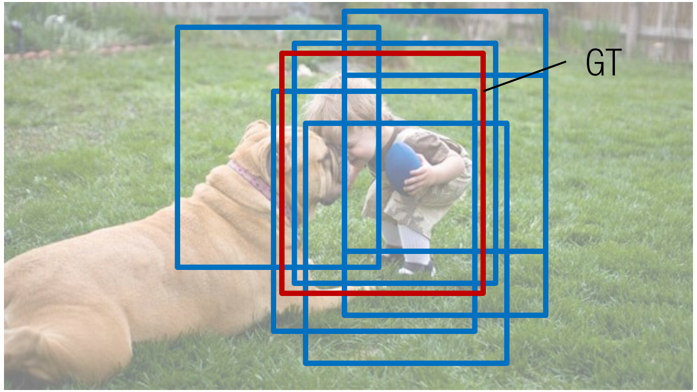
Region proposal

Region-CNN



Region proposal

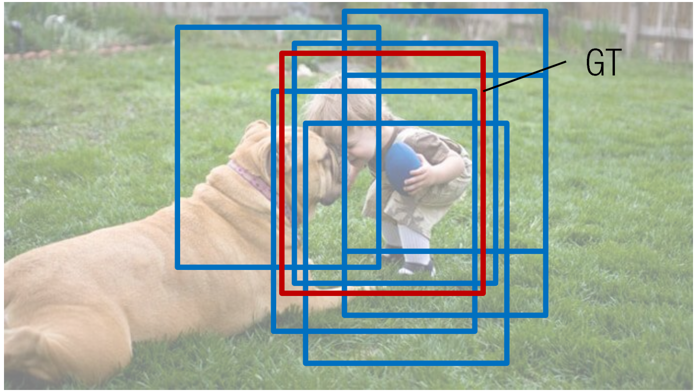
TRAINING DATA



Region proposal

GT

Region-CNN



Region proposal

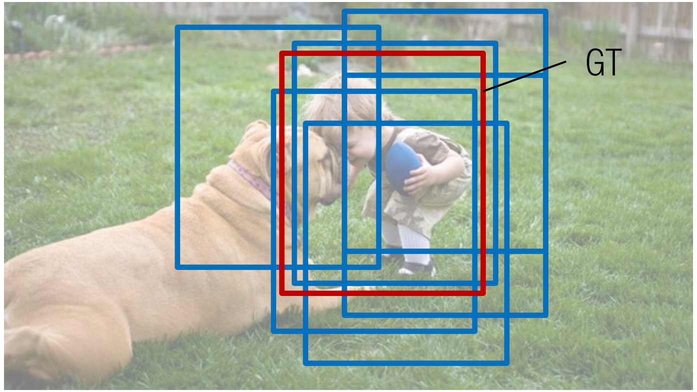
GT

Region proposal

+ if

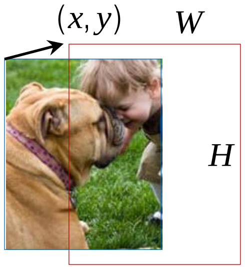
intersection of union (IoU) > 0.5

Region-CNN



Region proposal

GT

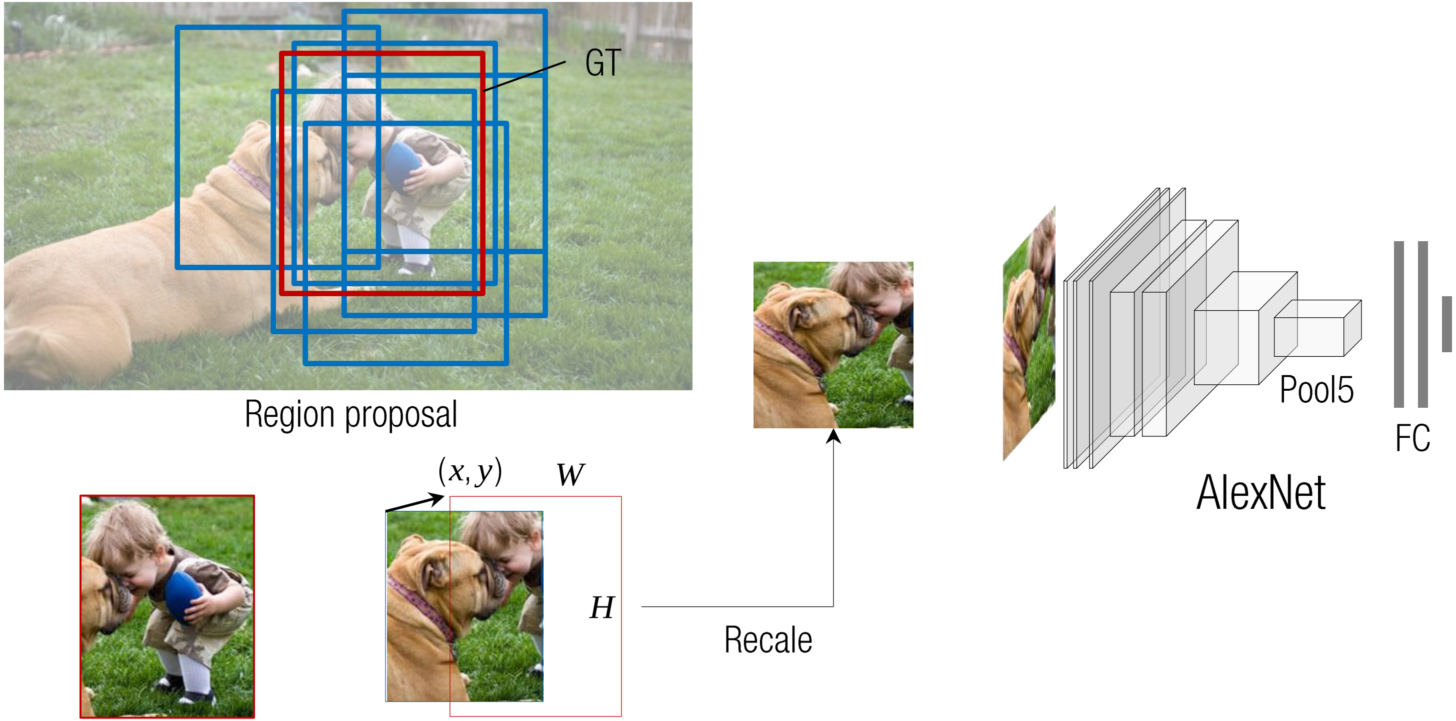


Detection offset

Region proposal

+ if
intersection of union (IoU) > 0.5

REGION-CNN



GT

Region proposal

(x, y) W

H

Recale

Pool5

FC

AlexNet

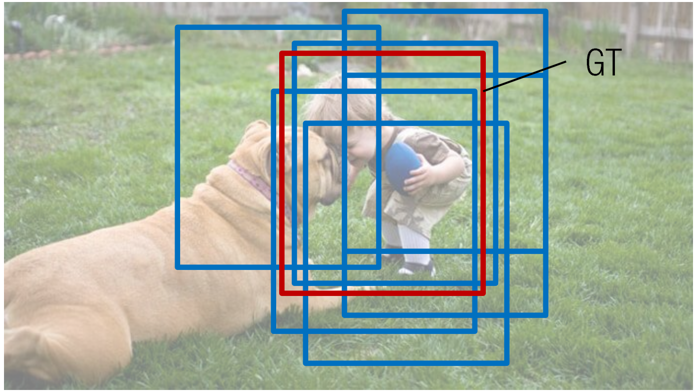
GT

Region proposal

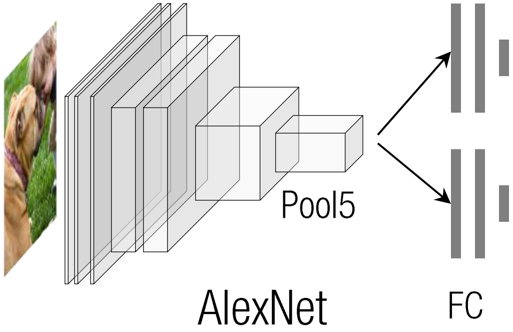
+ if

intersection of union (IoU) > 0.5

Region-CNN



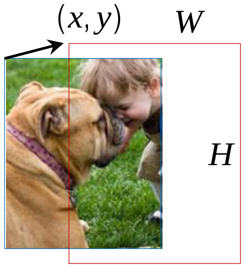
Region proposal



Region classification
person?

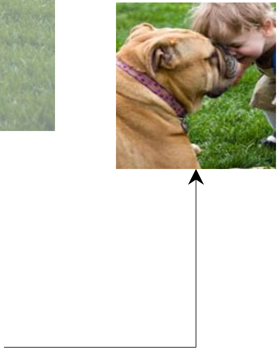
Bounding box regression
(x,y,W,H)

GT



Region proposal

+ if
intersection of union (IoU) > 0.5



Recale

Relative offset:

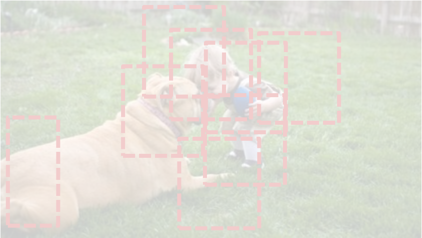
$$x=\frac{P_x-G_x}{P_W},y=\frac{P_y-G_y}{P_H},W=\log(\frac{G_W}{P_W}),H=\log(\frac{G_H}{P_H})$$

$$\boxed{(P_x,P_y,P_W,P_H)}$$

$$\boxed{(G_x,G_y,G_W,G_H)}$$

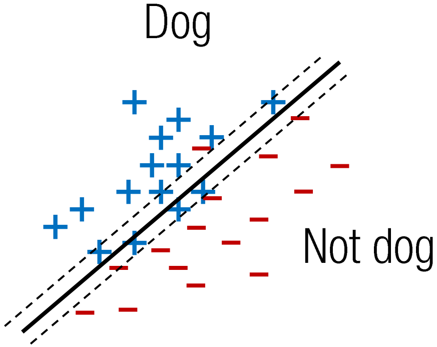
R-CNN (GIRSHICK ET AL.)

Input image



Localization

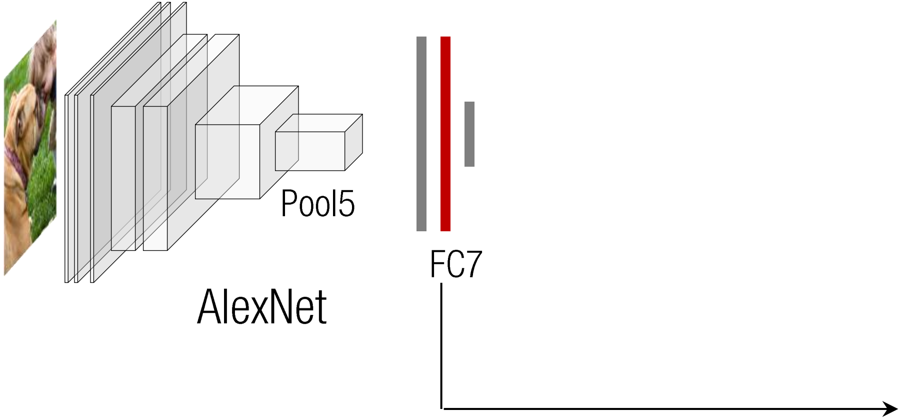
Feature extraction

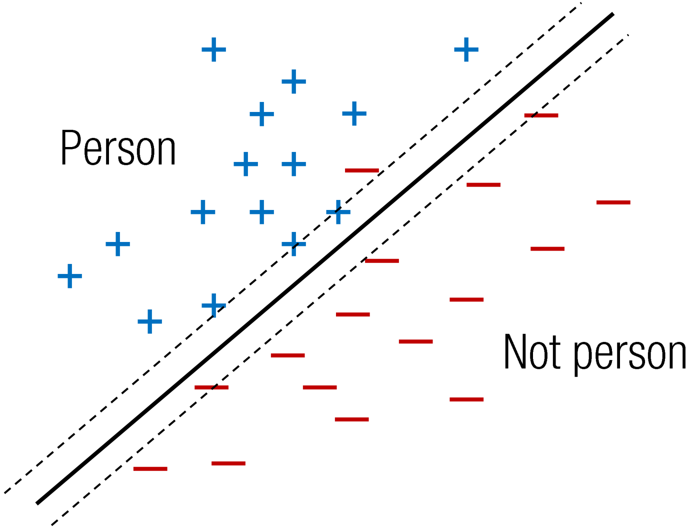


Classification

	Region proposal	CNN	SVM
Limitations	**2000 evluations**	Deep	n-classifiers

CLASSIFICATION

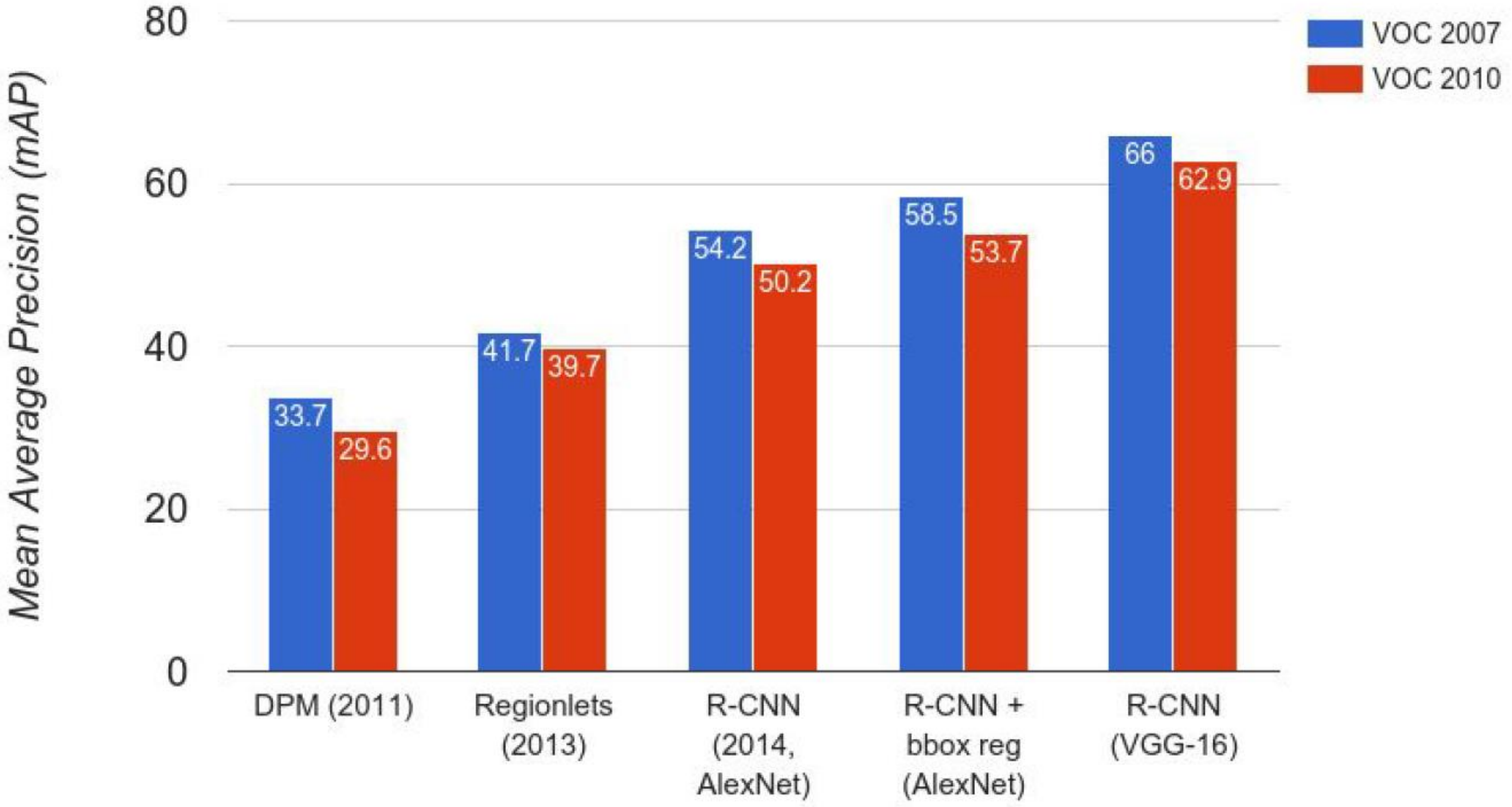


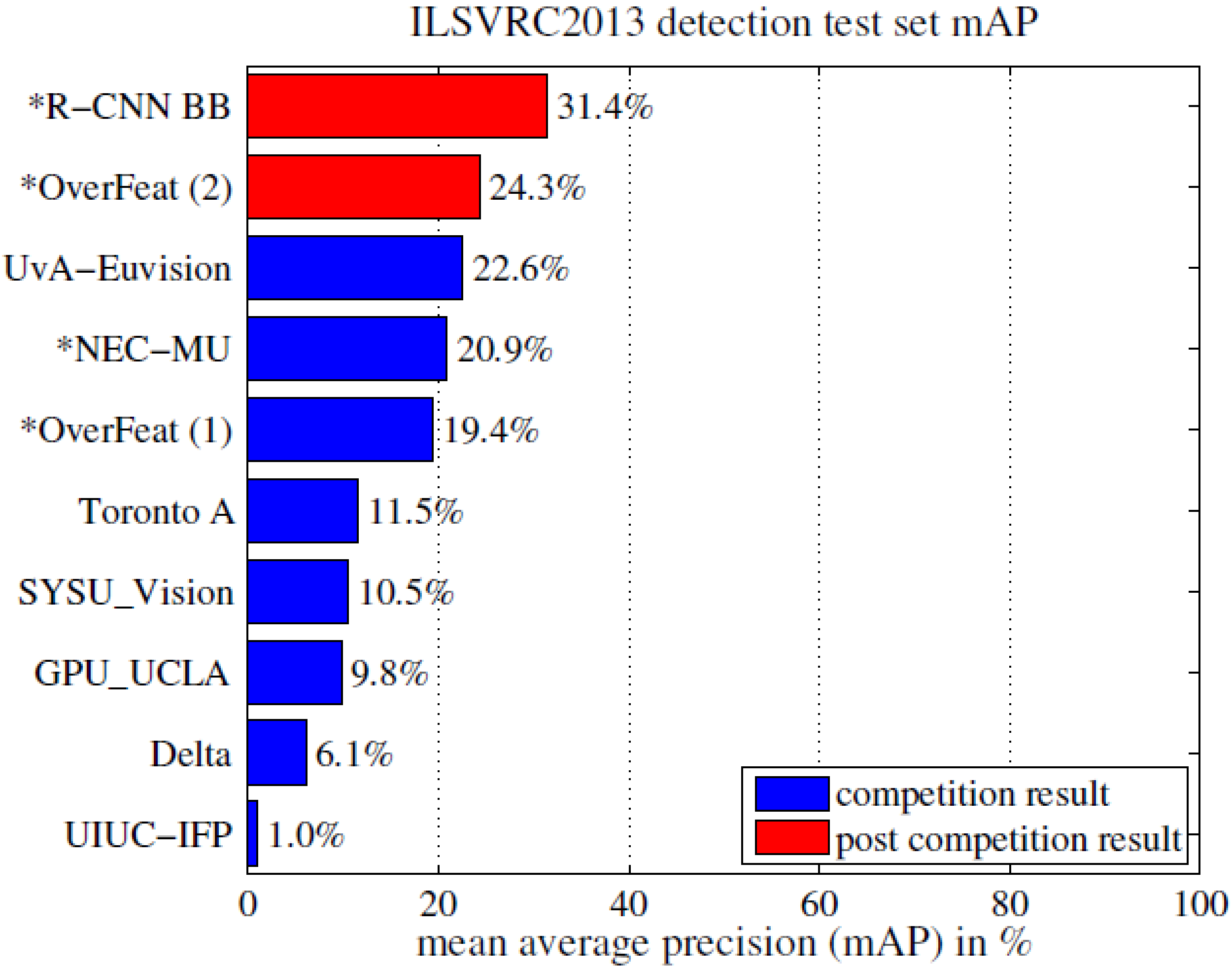


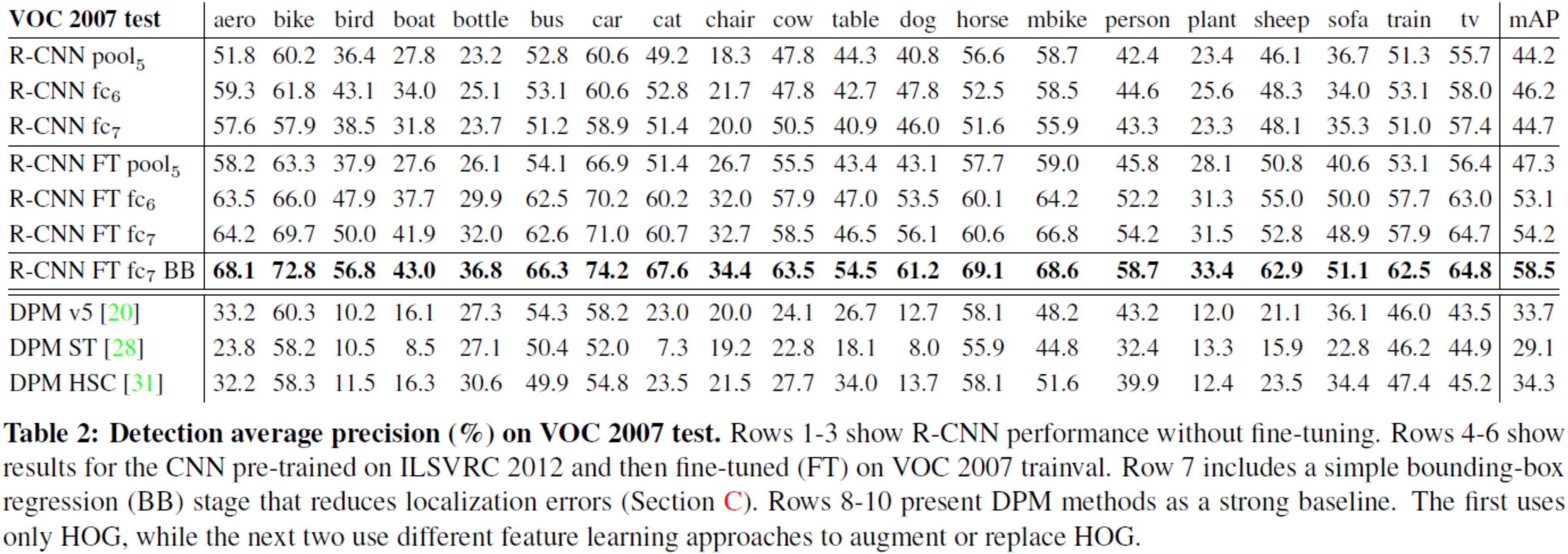
Max margin SVM classifier

$\mathbf{x} \cdot \mathbf{w} + \mathbf{b} > 0$ Positive D.

$\mathbf{x} \cdot \mathbf{w} + \mathbf{b} < 0$ Negative D.

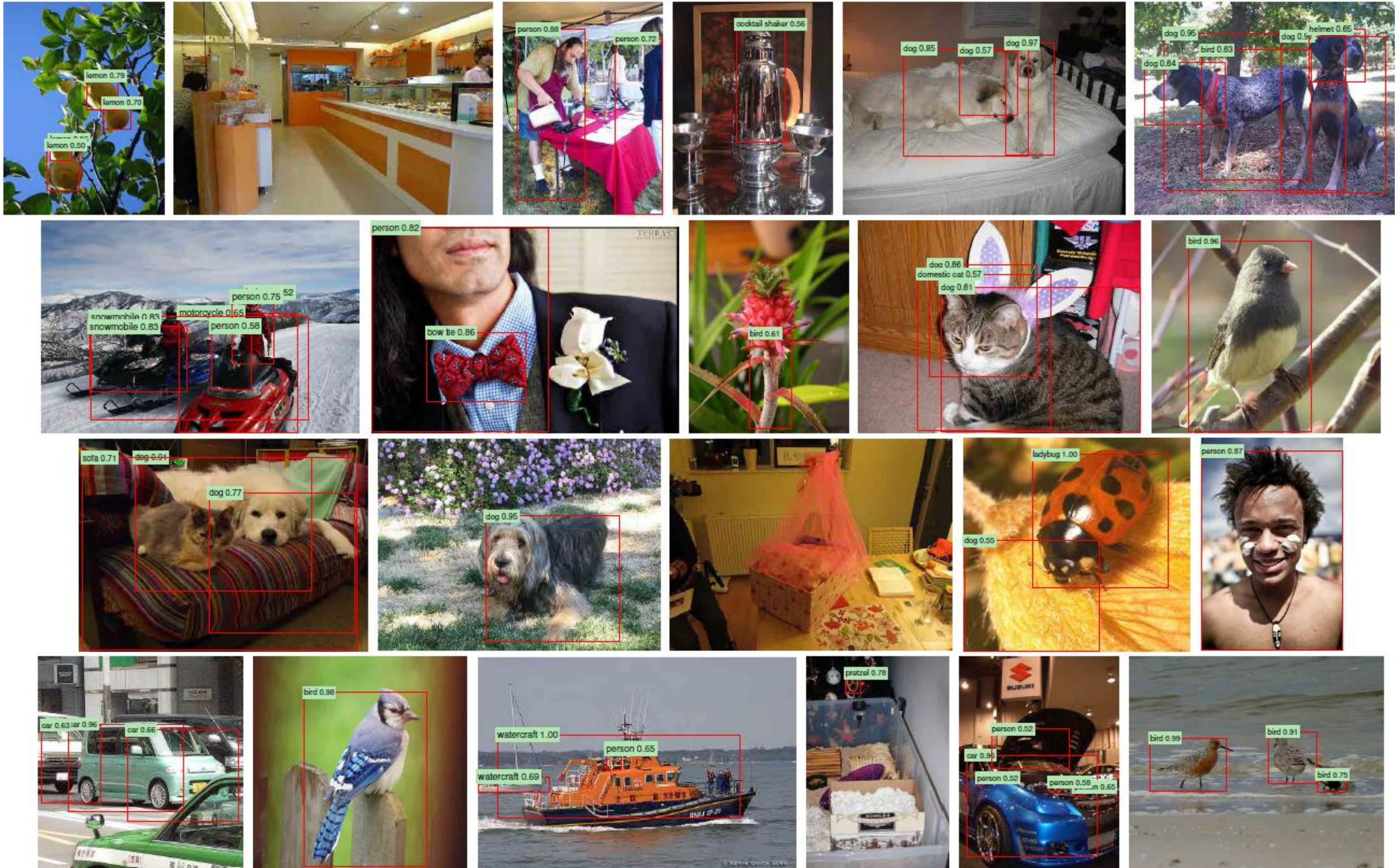

80
60
40
20
0
Mean Average Precision (mAP)
VOC 2007
VOC 2010
33.7
29.6
41.7
39.7
54.2
50.2
58.5
53.7
66
62.9
DPM (2011)
Regionlets (2013)
R-CNN (2014, AlexNet)
R-CNN + bbox reg (AlexNet)
R-CNN (VGG-16)

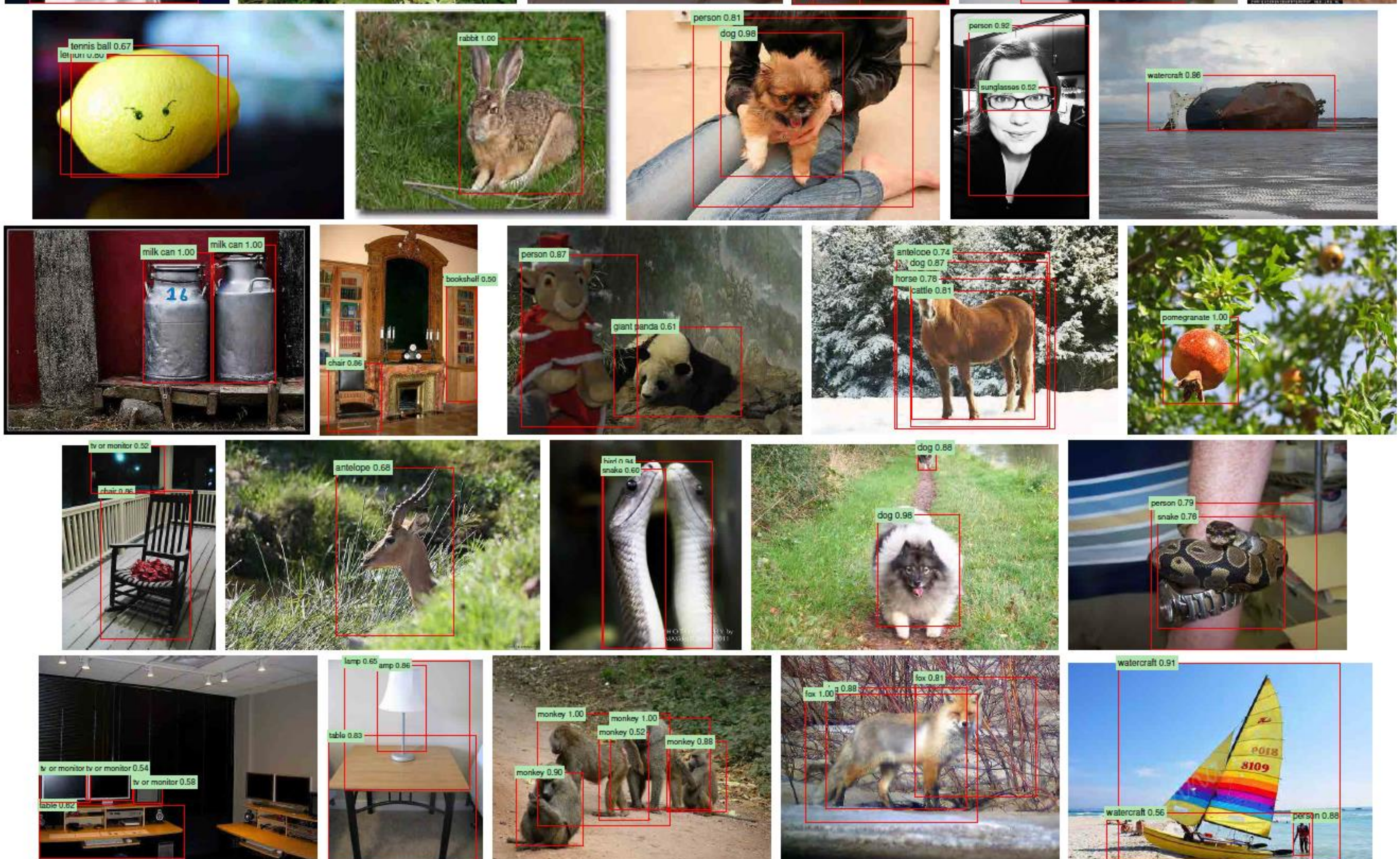

ILSVRC2013 detection test set mAP
*R−CNN BB
31.4%
*OverFeat (2)
24.3%
UvA−Euvision
22.6%
*NEC−MU
20.9%
*OverFeat (1)
19.4%
Toronto A
11.5%
SYSU_Vision
10.5%
GPU_UCLA
9.8%
Delta
6.1%
UIUC−IFP
1.0%
🟦 competition result
🟥 post competition result
0
20
40
60
80
100
mean average precision (mAP) in %



VOC 2007 test	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	mAP
R-CNN pool$_5$	51.8	60.2	36.4	27.8	23.2	52.8	60.6	49.2	18.3	47.8	44.3	40.8	56.6	58.7	42.4	23.4	46.1	36.7	51.3	55.7	44.2
R-CNN fc$_6$	59.3	61.8	43.1	34.0	25.1	53.1	60.6	52.8	21.7	47.8	42.7	47.8	52.5	58.5	44.6	25.6	48.3	34.0	53.1	58.0	46.2
R-CNN fc$_7$	57.6	57.9	38.5	31.8	23.7	51.2	58.9	51.4	20.0	50.5	40.9	46.0	51.6	55.9	43.3	23.3	48.1	35.3	51.0	57.4	44.7
R-CNN FT pool$_5$	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
R-CNN FT fc$_6$	63.5	66.0	47.9	37.7	29.9	62.5	70.2	60.2	32.0	57.9	47.0	53.5	60.1	64.2	52.2	31.3	55.0	50.0	57.7	63.0	53.1
R-CNN FT fc$_7$	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
R-CNN FT fc$_7$ BB	**68.1**	**72.8**	**56.8**	**43.0**	**36.8**	**66.3**	**74.2**	**67.6**	**34.4**	**63.5**	**54.5**	**61.2**	**69.1**	**68.6**	**58.7**	**33.4**	**62.9**	**51.1**	**62.5**	**64.8**	**58.5**
DPM v5 [20]	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
DPM ST [28]	23.8	58.2	10.5	8.5	27.1	50.4	52.0	7.3	19.2	22.8	18.1	8.0	55.9	44.8	32.4	13.3	15.9	22.8	46.2	44.9	29.1
DPM HSC [31]	32.2	58.3	11.5	16.3	30.6	49.9	54.8	23.5	21.5	27.7	34.0	13.7	58.1	51.6	39.9	12.4	23.5	34.4	47.4	45.2	34.3

Table 2: Detection average precision (%) on VOC 2007 test. Rows 1-3 show R-CNN performance without fine-tuning. Rows 4-6 show results for the CNN pre-trained on ILSVRC 2012 and then fine-tuned (FT) on VOC 2007 trainval. Row 7 includes a simple bounding-box regression (BB) stage that reduces localization errors (Section C). Rows 8-10 present DPM methods as a strong baseline. The first uses only HOG, while the next two use different feature learning approaches to augment or replace HOG.


lemon 0.79
lemon 0.70
lemon 0.50
cocktail shaker 0.56
person 0.88
person 0.72
dog 0.85
dog 0.57
dog 0.97
dog 0.95
dog 0.84
bird 0.83
helmet 0.65
person 0.75
snowmobile 0.83
snowmobile 0.83
motorcycle 0.65
person 0.58
person 0.82
bow tie 0.86
bird 0.61
dog 0.66
domestic cat 0.57
dog 0.81
bird 0.96
sofa 0.71
dog 0.91
dog 0.77
dog 0.95
ladybug 1.00
dog 0.55
person 0.87
car 0.63
car 0.96
car 0.66
bird 0.98
watercraft 1.00
person 0.65
watercraft 0.69
pretzel 0.78
person 0.52
car 0.96
person 0.52
person 0.58
person 0.65
bird 0.99
bird 0.91
bird 0.75


tennis ball 0.67
lemon 0.80
rabbit 1.00
person 0.81
dog 0.98
person 0.92
sunglasses 0.52
watercraft 0.86
milk can 1.00
milk can 1.00
bookshelf 0.50
chair 0.86
person 0.87
giant panda 0.61
antelope 0.74
dog 0.87
horse 0.78
cattle 0.81
pomegranate 1.00
tv or monitor 0.52
chair 0.86
antelope 0.68
bird 0.94
snake 0.60
dog 0.88
dog 0.96
person 0.79
snake 0.76
tv or monitor
tv or monitor 0.54
tv or monitor 0.58
table 0.62
lamp 0.65
lamp 0.86
table 0.83
monkey 1.00
monkey 1.00
monkey 0.52
monkey 0.88
monkey 0.90
fox 0.81
dog 0.88
fox 1.00
watercraft 0.91
watercraft 0.56
person 0.88

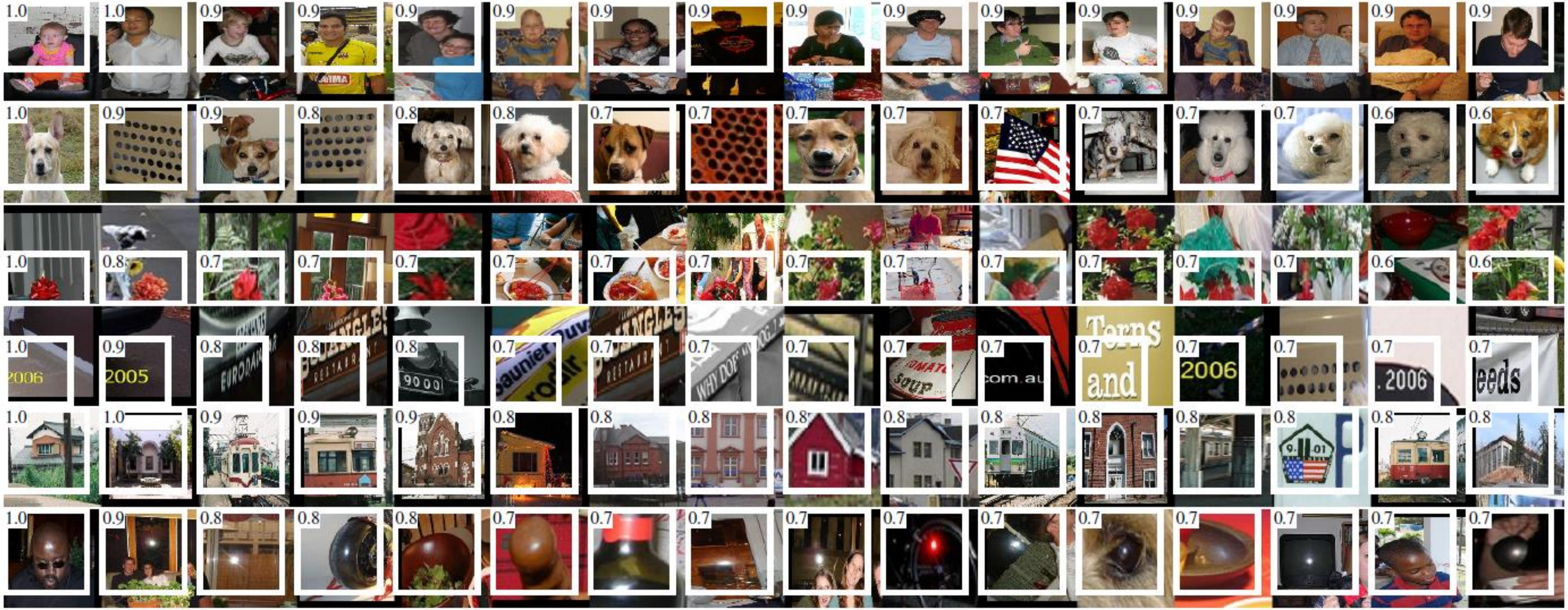
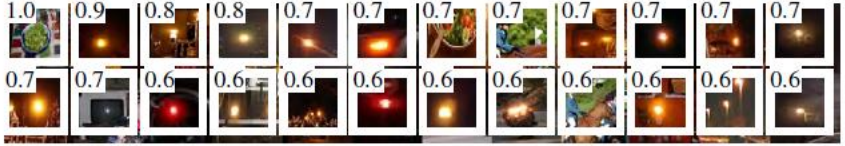
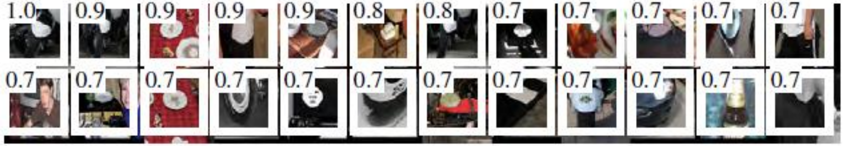


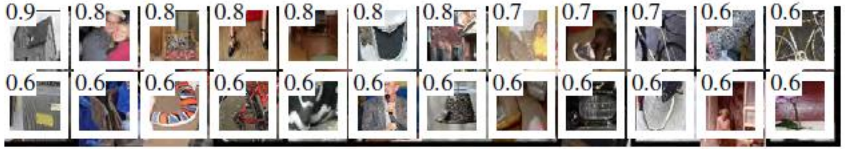
Figure 4: Top regions for six pool$_5$ units. Receptive fields and activation values are drawn in white. Some units are aligned to concepts, such as people (row 1) or text (4). Other units capture texture and material properties, such as dot arrays (2) and specular reflections (6).



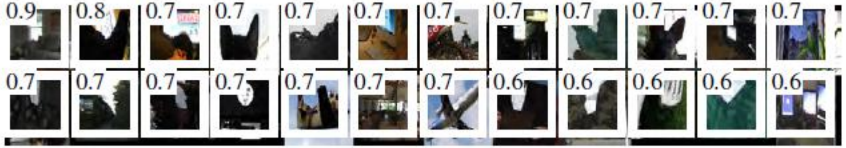
pool5 feature: (3,3,1) (top 1 – 24)



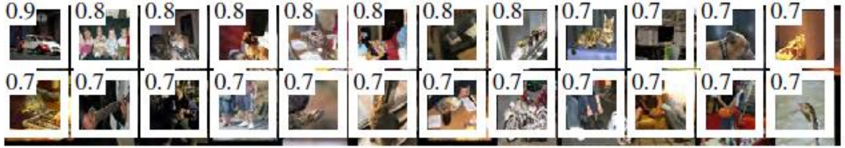
pool5 feature: (3,3,2) (top 1 – 24)



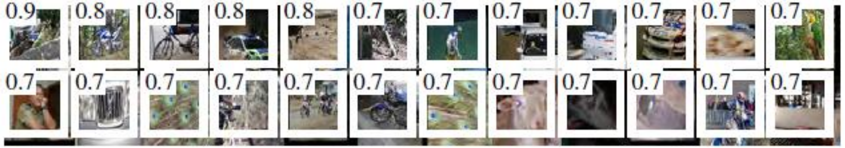
pool5 feature: (3,3,3) (top 1 – 24)



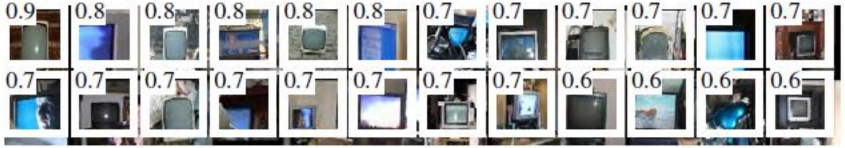
pool5 feature: (3,3,4) (top 1 – 24)



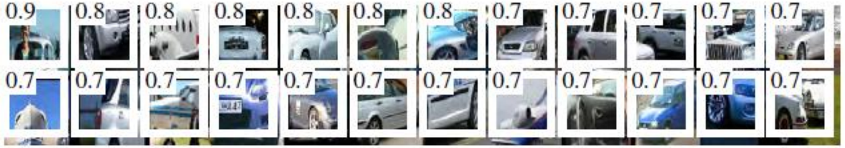
pool5 feature: (3,3,5) (top 1 – 24)



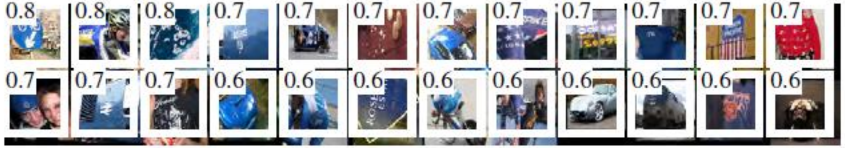
pool5 feature: (3,3,6) (top 1 – 24)



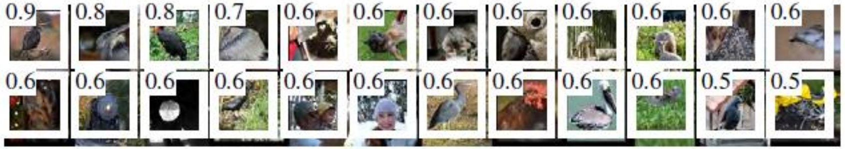
pool5 feature: (3,3,7) (top 1 – 24)



pool5 feature: (3,3,8) (top 1 – 24)

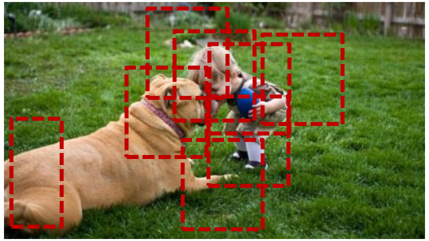


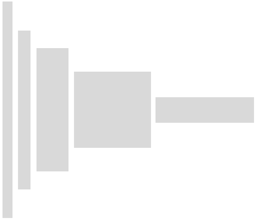
pool5 feature: (3,3,9) (top 1 – 24)

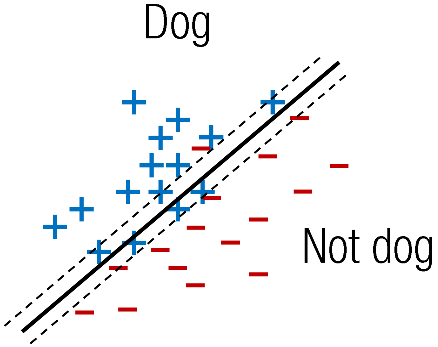


pool5 feature: (3,3,10) (top 1 – 24)

OBJECT DETECTION PIPELINE





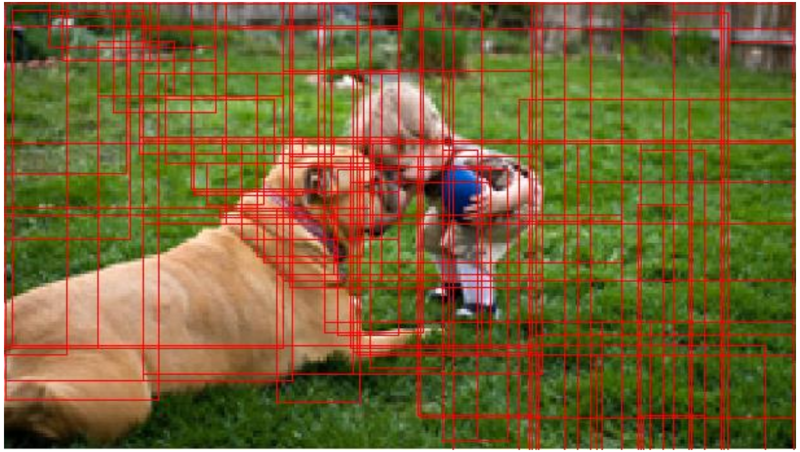


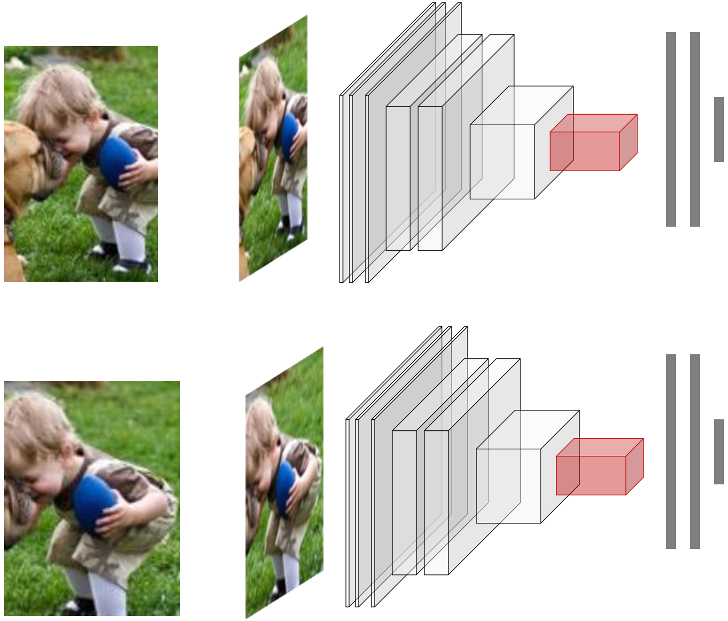
	Input image	Localization	Feature extraction	Classification
		Region proposal	CNN	SVM
Limitations		2000 evluations	Deep	n-classifiers

Too slow in testing time

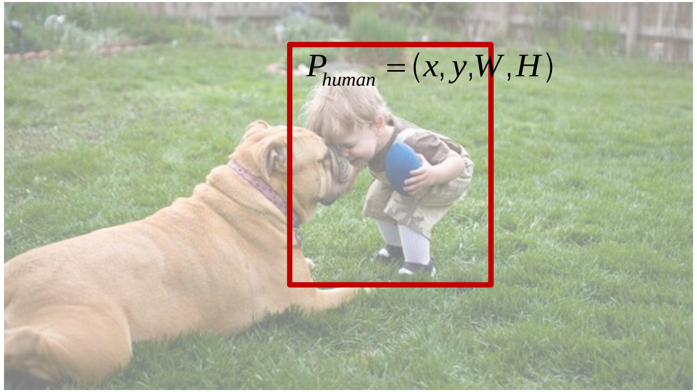
Post-hoc optimization

Redundant Feature Map

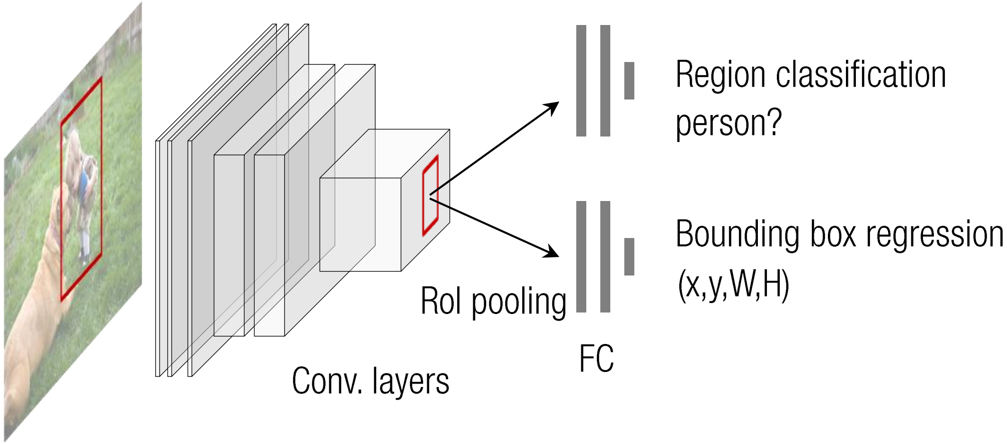




Feature Map Recycling



Groun truth region



Region classification
person?

Bounding box regression
(x,y,W,H)

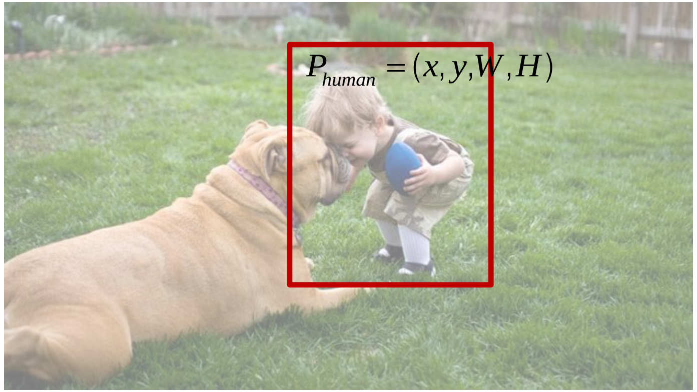
RoI pooling

FC

Conv. layers

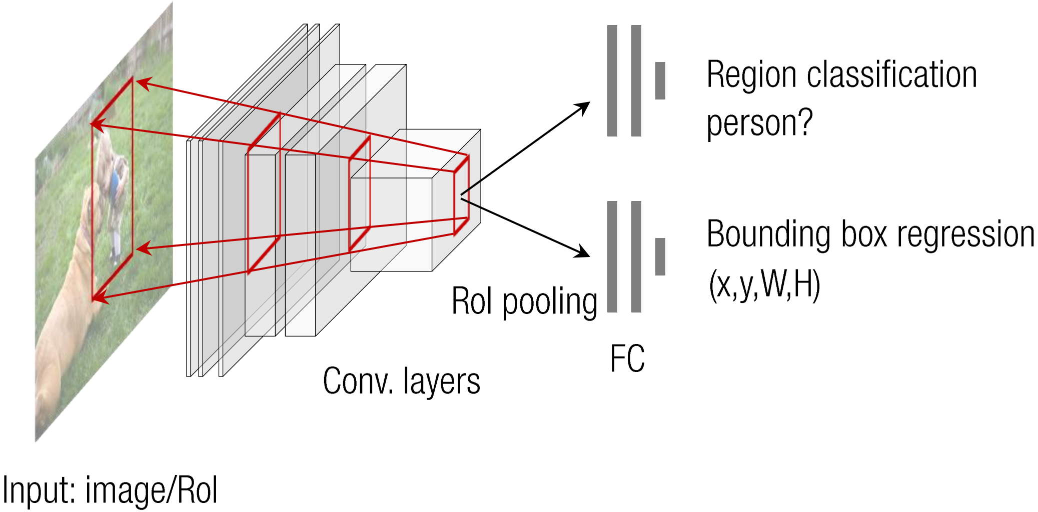
Input: image/RoI

Back-propagation



$P_{human} = (x, y, W, H)$

Groun truth region



Region classification person?

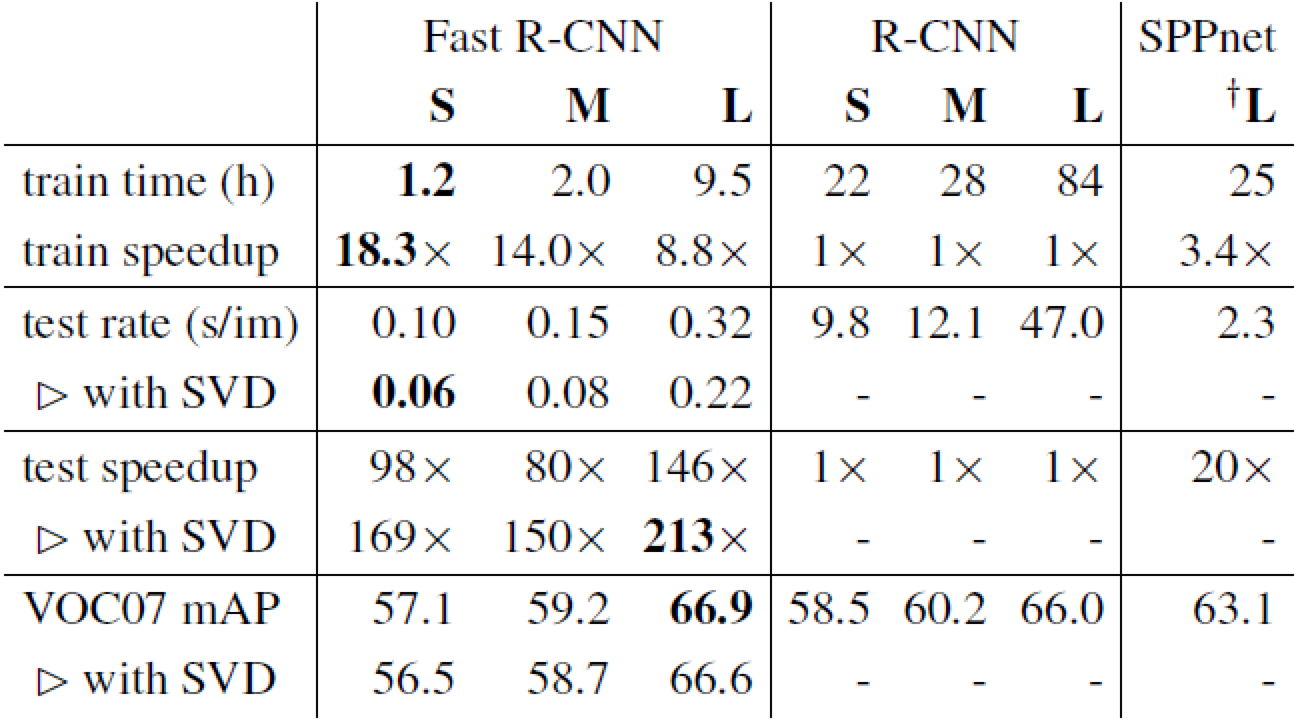
Bounding box regression (x,y,W,H)

RoI pooling

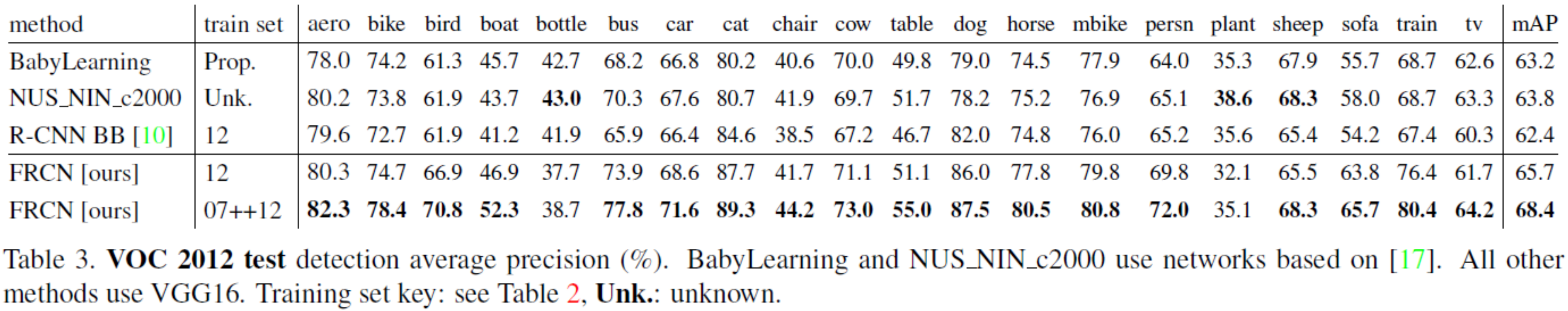
FC

Conv. layers

Input: image/RoI

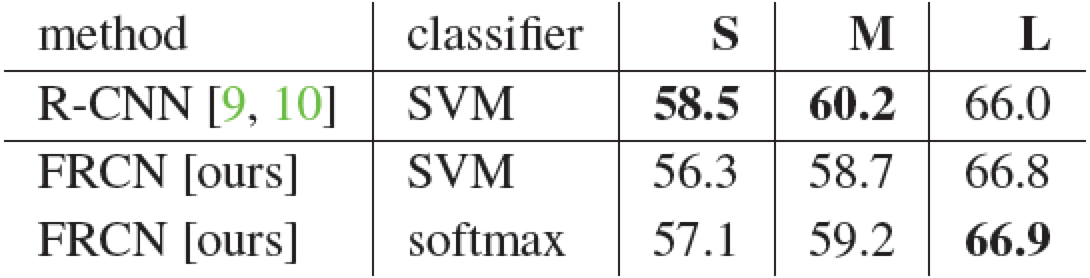


	Fast R-CNN			R-CNN			SPPnet
	S	**M**	**L**	**S**	**M**	**L**	†**L**
train time (h)	**1.2**	2.0	9.5	22	28	84	25
train speedup	**18.3**×	14.0×	8.8×	1×	1×	1×	3.4×
test rate (s/im)	0.10	0.15	0.32	9.8	12.1	47.0	2.3
▷ with SVD	**0.06**	0.08	0.22	-	-	-	-
test speedup	98×	80×	146×	1×	1×	1×	20×
▷ with SVD	169×	150×	**213**×	-	-	-	-
VOC07 mAP	57.1	59.2	**66.9**	58.5	60.2	66.0	63.1
▷ with SVD	56.5	58.7	66.6	-	-	-	-



method	train set	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	persn	plant	sheep	sofa	train	tv	mAP
BabyLearning	Prop.	78.0	74.2	61.3	45.7	42.7	68.2	66.8	80.2	40.6	70.0	49.8	79.0	74.5	77.9	64.0	35.3	67.9	55.7	68.7	62.6	63.2
NUS_NIN_c2000	Unk.	80.2	73.8	61.9	43.7	**43.0**	70.3	67.6	80.7	41.9	69.7	51.7	78.2	75.2	76.9	65.1	**38.6**	**68.3**	58.0	68.7	63.3	63.8
R-CNN BB [10]	12	79.6	72.7	61.9	41.2	41.9	65.9	66.4	84.6	38.5	67.2	46.7	82.0	74.8	76.0	65.2	35.6	65.4	54.2	67.4	60.3	62.4
FRCN [ours]	12	80.3	74.7	66.9	46.9	37.7	73.9	68.6	87.7	41.7	71.1	51.1	86.0	77.8	79.8	69.8	32.1	65.5	63.8	76.4	61.7	65.7
FRCN [ours]	07++12	**82.3**	**78.4**	**70.8**	**52.3**	38.7	**77.8**	**71.6**	**89.3**	**44.2**	**73.0**	**55.0**	**87.5**	**80.5**	**80.8**	**72.0**	35.1	**68.3**	**65.7**	**80.4**	**64.2**	**68.4**

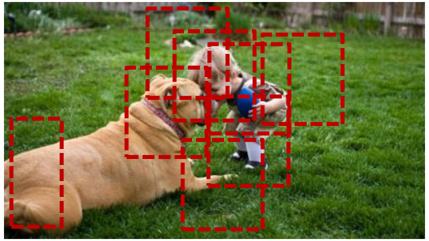
Table 3. **VOC 2012 test** detection average precision (%). BabyLearning and NUS_NIN_c2000 use networks based on [17]. All other methods use VGG16. Training set key: see Table 2, **Unk.**: unknown.



method	classifier	**S**	**M**	**L**
R-CNN [9, 10]	SVM	**58.5**	**60.2**	66.0
FRCN [ours]	SVM	56.3	58.7	66.8
FRCN [ours]	softmax	57.1	59.2	**66.9**

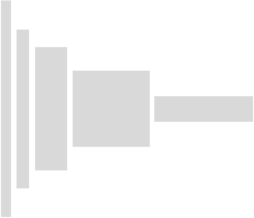
FASTER RCNN (REN ET AL.)

Input image



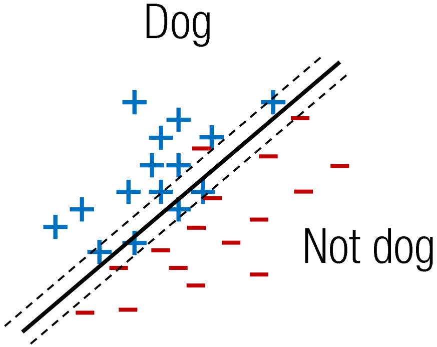
Localization

Region proposal



Feature extraction

CNN



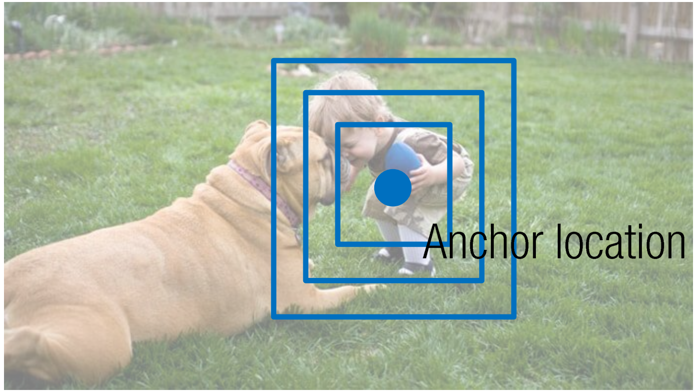
Classification

SVM

One network

Region Proposal Network

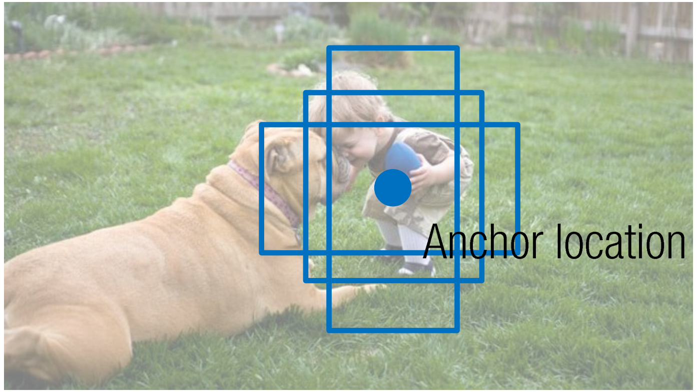
Region Proposal Network



Proposed regions

3 scales

Region Proposal Network

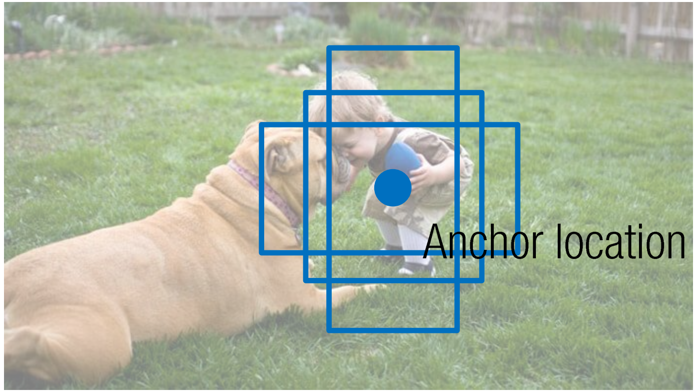


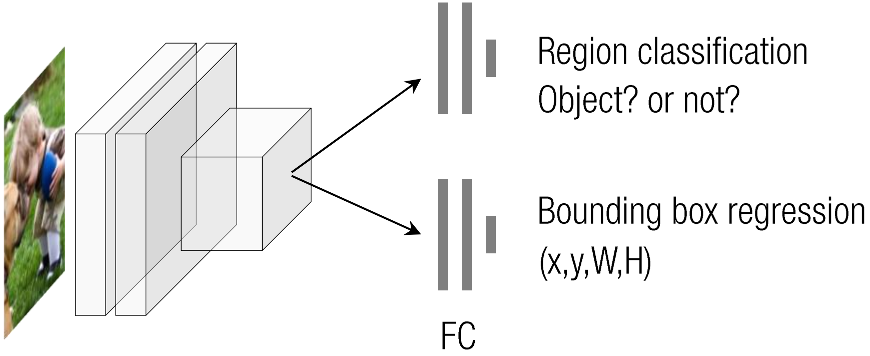
Proposed regions

3 scales
3 aspect ratio

9 proposals per anchor

Region Proposal Network





Region classification
Object? or not?

Bounding box regression
(x,y,W,H)

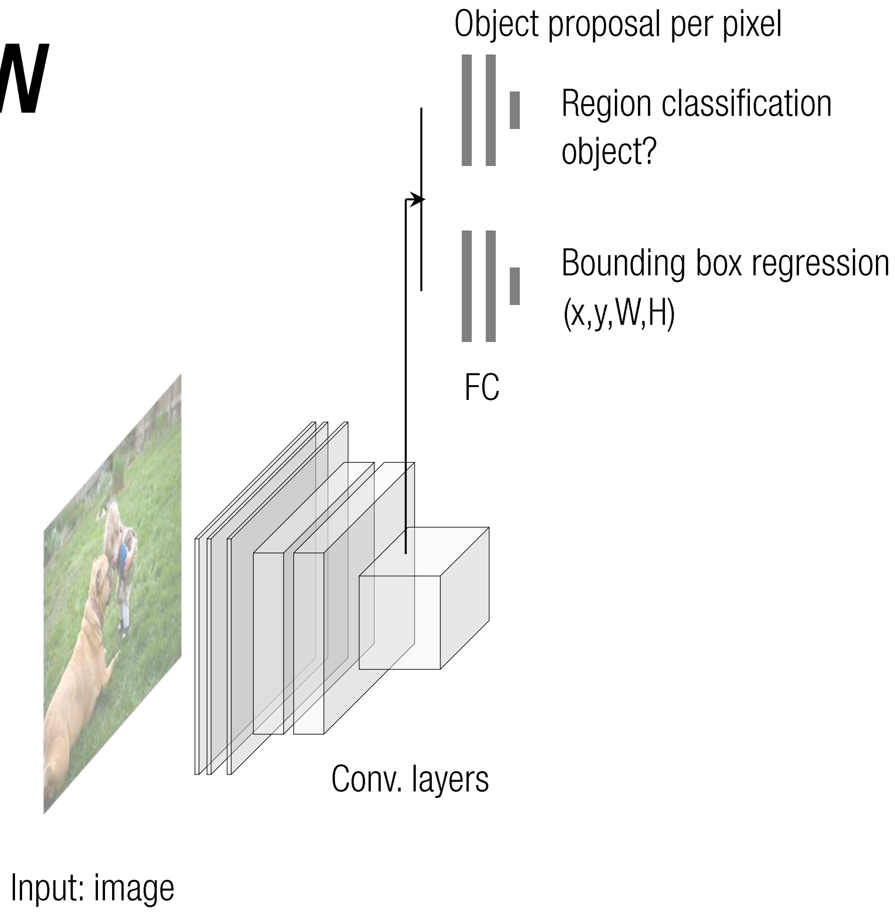
FC

Proposed regions

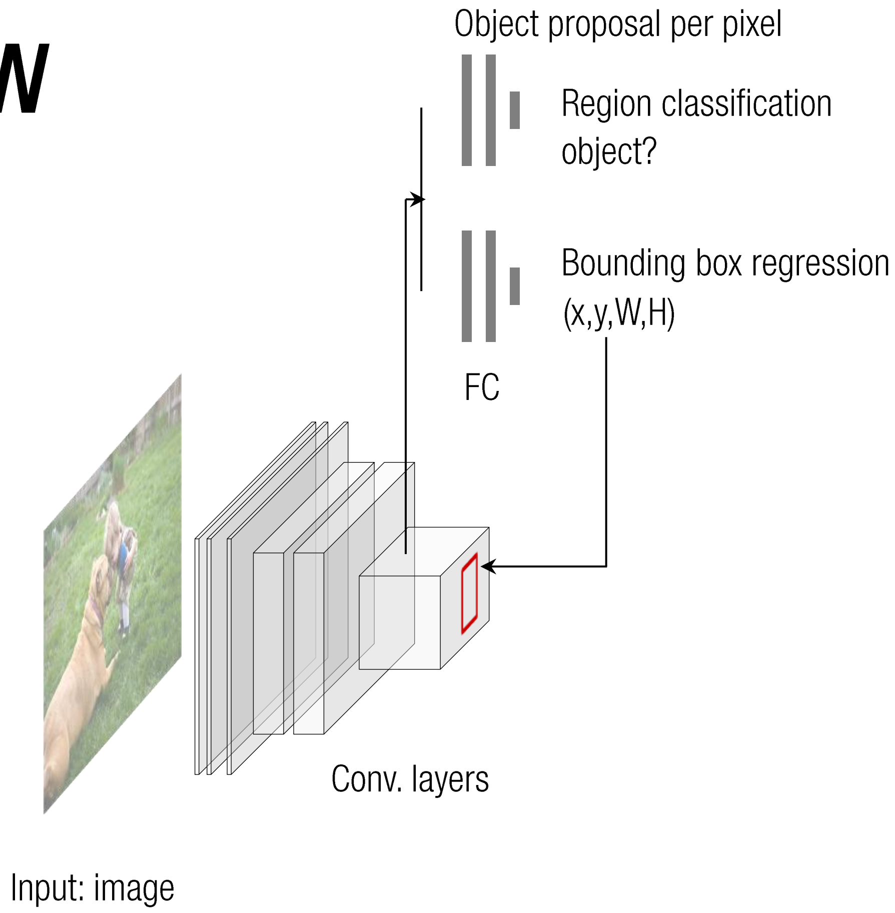
3 scales
3 aspect ratio

9 proposals per anchor

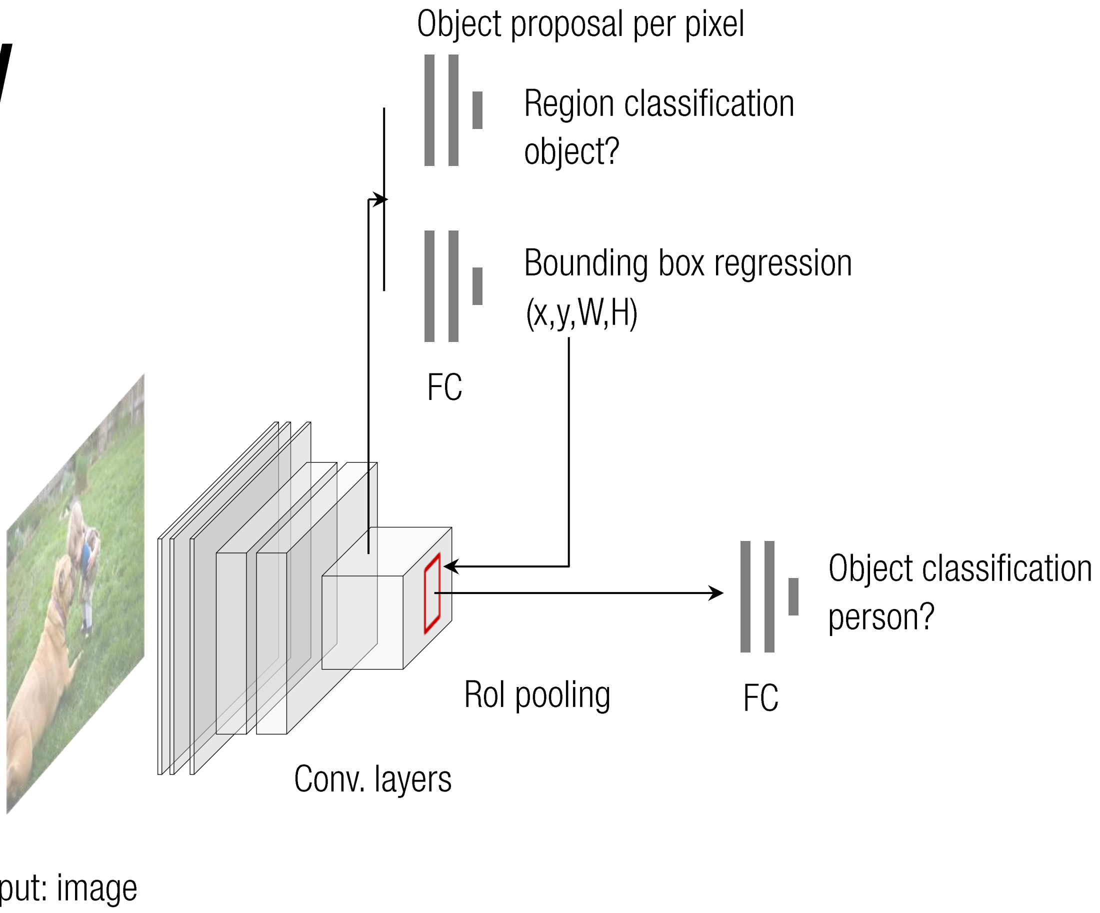
Faster RCNN

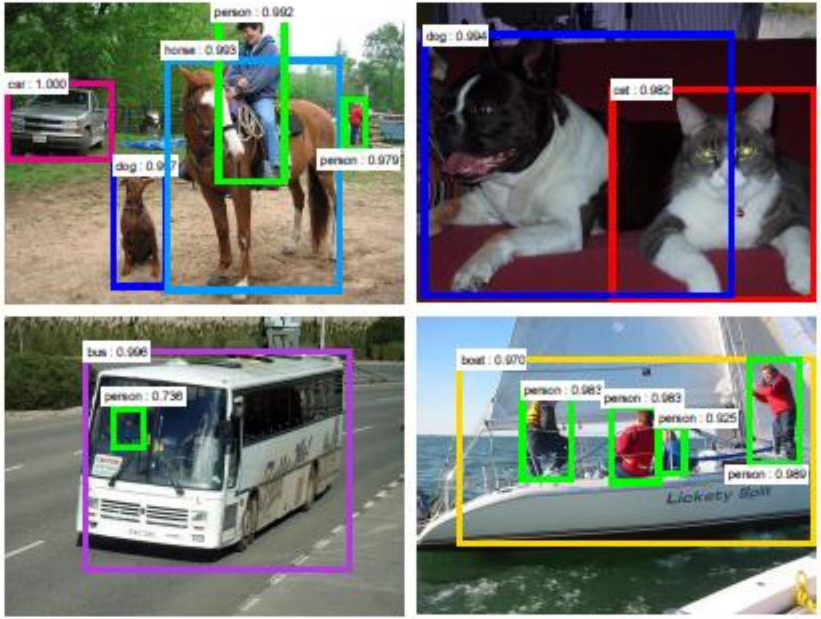

Object proposal per pixel
Region classification
object?
Bounding box regression
(x,y,W,H)
FC
Conv. layers
Input: image

Faster RCNN


Object proposal per pixel
Region classification
object?
Bounding box regression
(x,y,W,H)
FC
Conv. layers
Input: image

FASTER RCNN


Object proposal per pixel
Region classification
object?
Bounding box regression
(x,y,W,H)
FC
Object classification
person?
RoI pooling
FC
Conv. layers
Input: image


person : 0.992
horse : 0.993
car : 1.000
dog : 0.9 7
person : 0.979
dog : 0.994
cat : 0.982
bus : 0.996
person : 0.736
boat : 0.970
person : 0.983
person : 0.983
person : 0.925
person : 0.989
Lickety Split

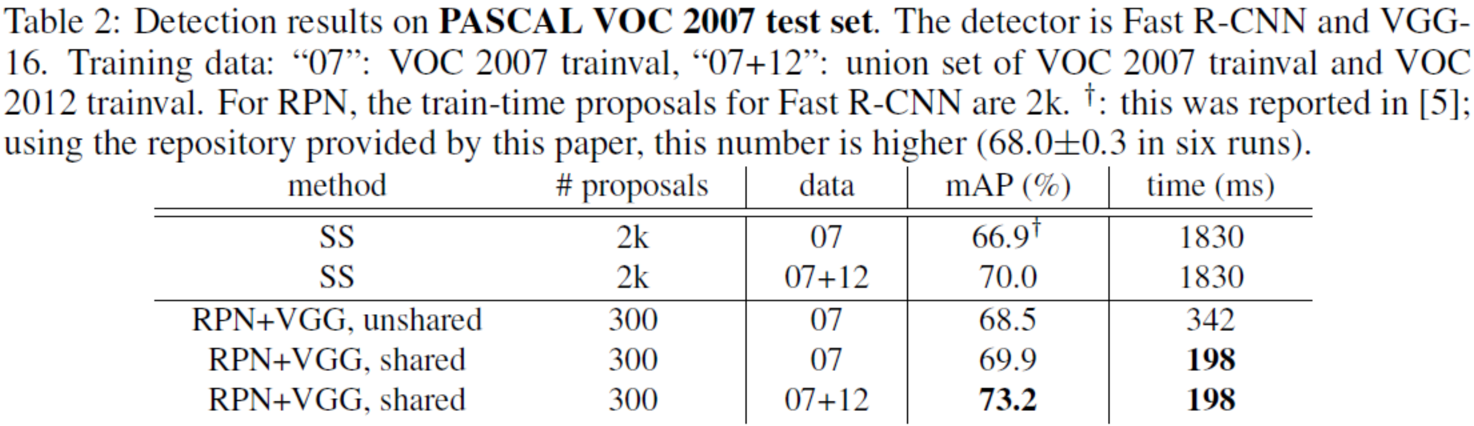
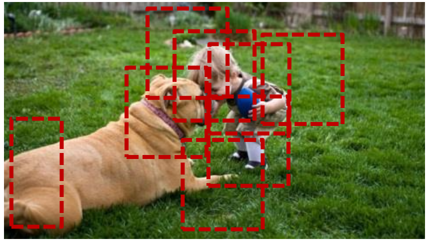


Table 2: Detection results on **PASCAL VOC 2007 test set**. The detector is Fast R-CNN and VGG-16. Training data: "07": VOC 2007 trainval, "07+12": union set of VOC 2007 trainval and VOC 2012 trainval. For RPN, the train-time proposals for Fast R-CNN are 2k. †: this was reported in [5]; using the repository provided by this paper, this number is higher (68.0±0.3 in six runs).

method	# proposals	data	mAP (%)	time (ms)
SS	2k	07	66.9†	1830
SS	2k	07+12	70.0	1830
RPN+VGG, unshared	300	07	68.5	342
RPN+VGG, shared	300	07	69.9	**198**
RPN+VGG, shared	300	07+12	**73.2**	**198**

Object Detection Pipeline

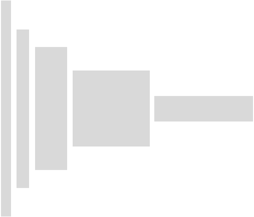
Input image



Localization

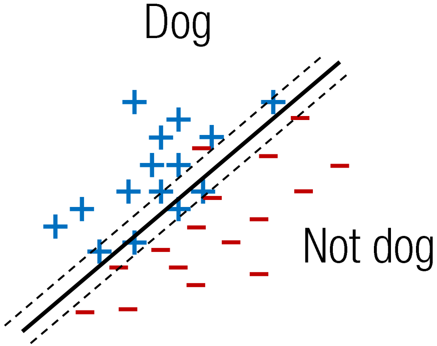
Region proposal

Per grid classification



Feature extraction

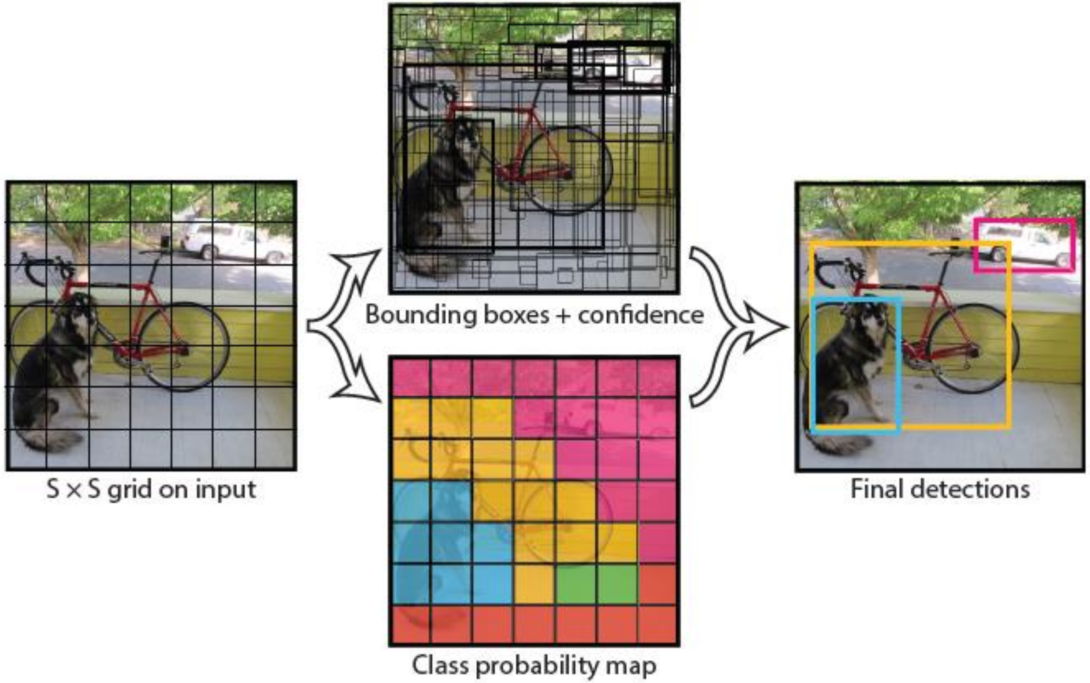
CNN



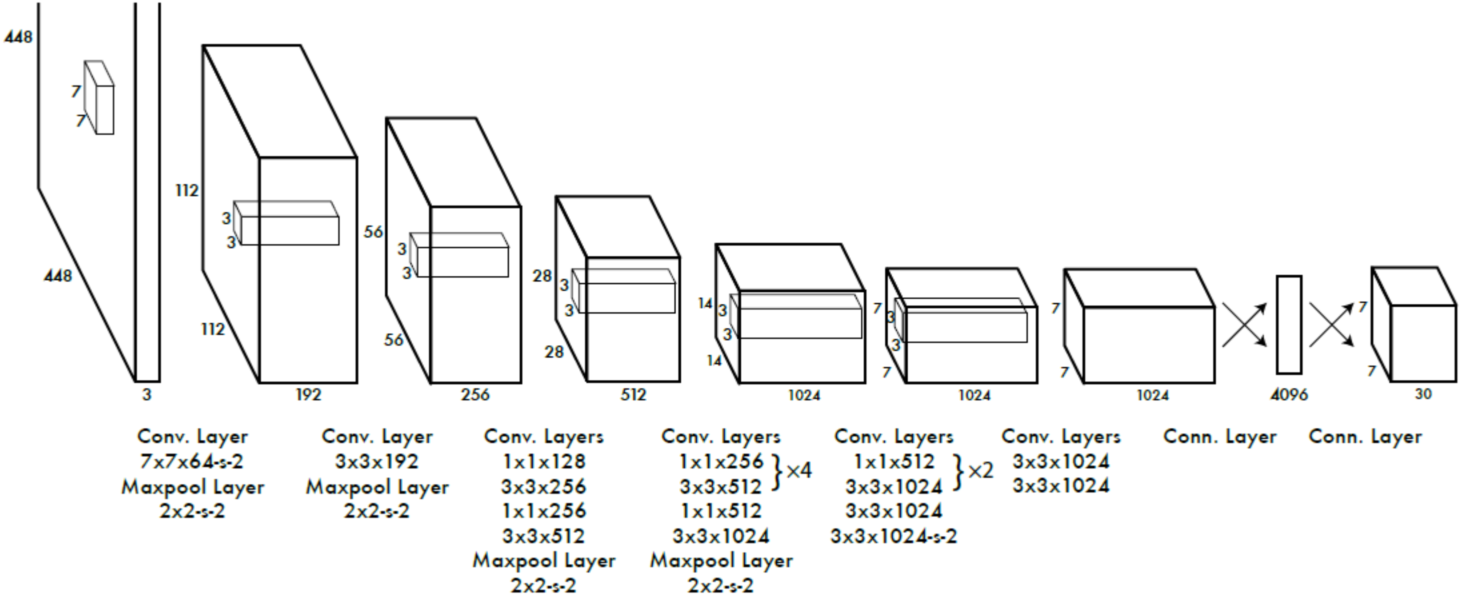
Classification

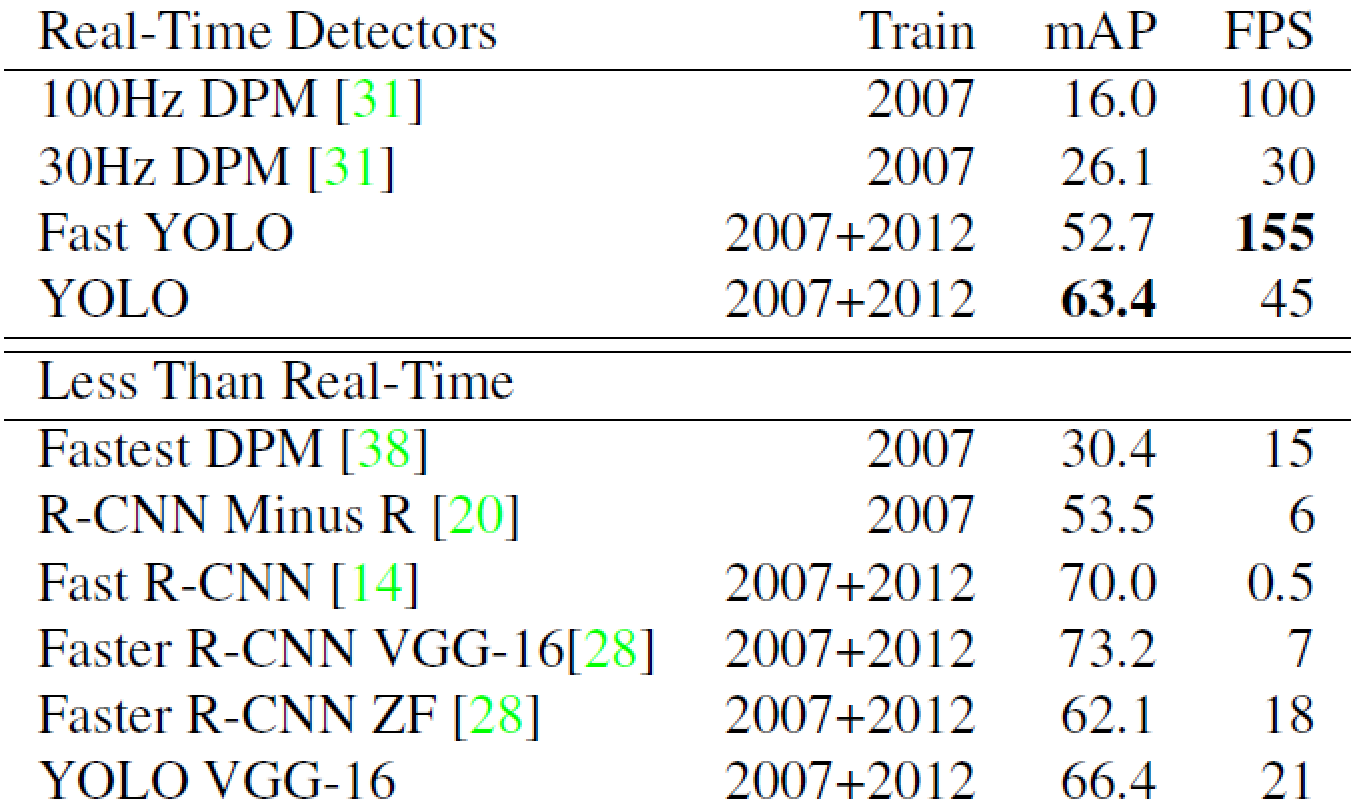
SVM

YOLO (You Only Look Once, Redmon et al.)

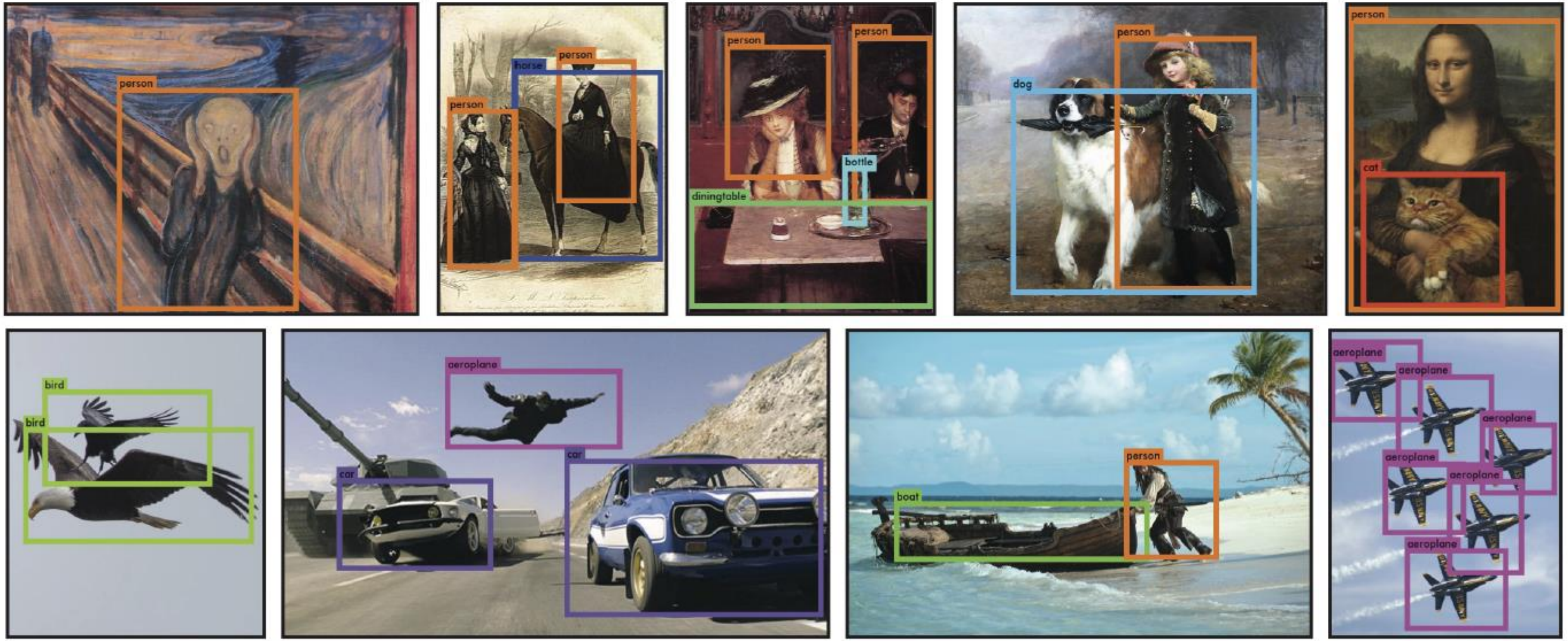


YOLO (You Only Look Once, Redmon et al.)





Real-Time Detectors	Train	mAP	FPS
100Hz DPM [31]	2007	16.0	100
30Hz DPM [31]	2007	26.1	30
Fast YOLO	2007+2012	52.7	**155**
YOLO	2007+2012	**63.4**	45
Less Than Real-Time			
Fastest DPM [38]	2007	30.4	15
R-CNN Minus R [20]	2007	53.5	6
Fast R-CNN [14]	2007+2012	70.0	0.5
Faster R-CNN VGG-16[28]	2007+2012	73.2	7
Faster R-CNN ZF [28]	2007+2012	62.1	18
YOLO VGG-16	2007+2012	66.4	21


person
horse
person
person
person
person
bottle
diningtable
person
dog
person
cat
bird
bird
aeroplane
car
car
boat
person
aeroplane
aeroplane
aeroplane
aeroplane
aeroplane
aeroplane

https://www.youtube.com/watch?time_continue=10&v=VOC3huqHrss
https://www.youtube.com/watch?time_continue=164&v=MPU2Histivl