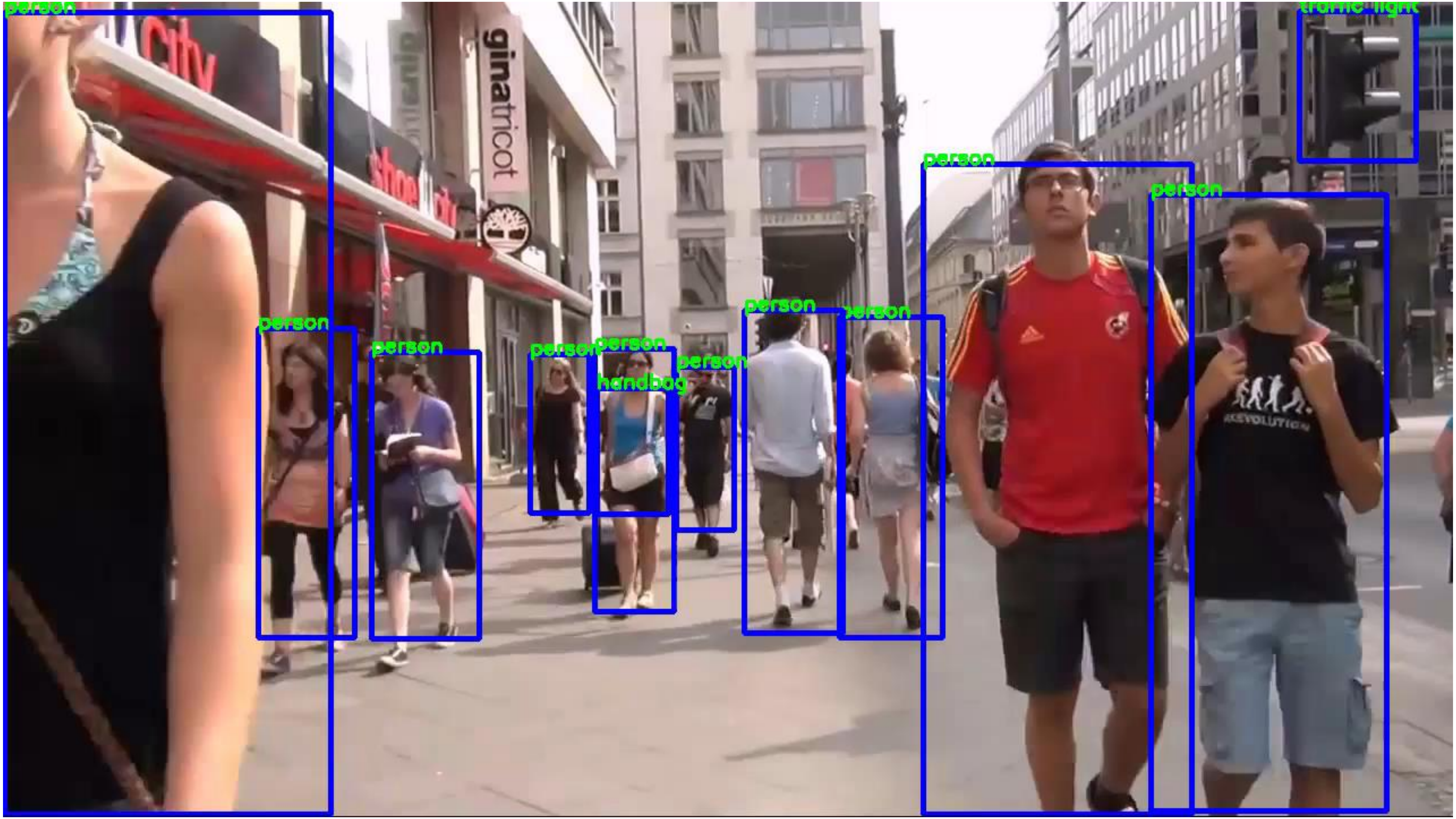
RON: Reverse Connection with Objectness Prior Networks for Object Detection

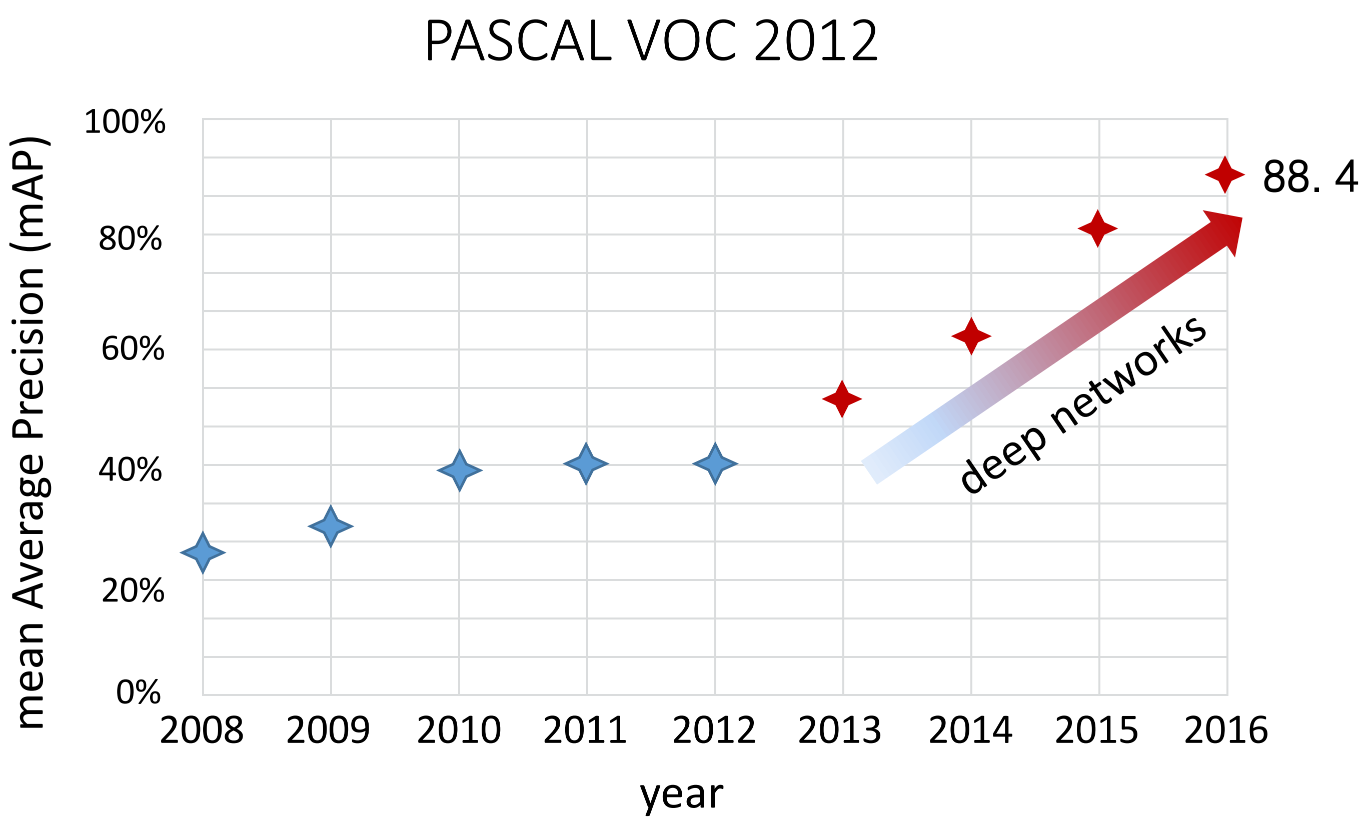
Tao Kong[1], Fuchun Sun[1], Anbang Yao[2], Huaping Liu[1], Ming Lu[3], Yurong Chen[2]

[1]Department of CST, Tsinghua University, [2]Intel Labs China
[3]Department of EE, Tsinghua University



person
traffic light
person
person
person
person
person
person
person
handbag
person
person
person
City
ginatricot

The progress of object detection



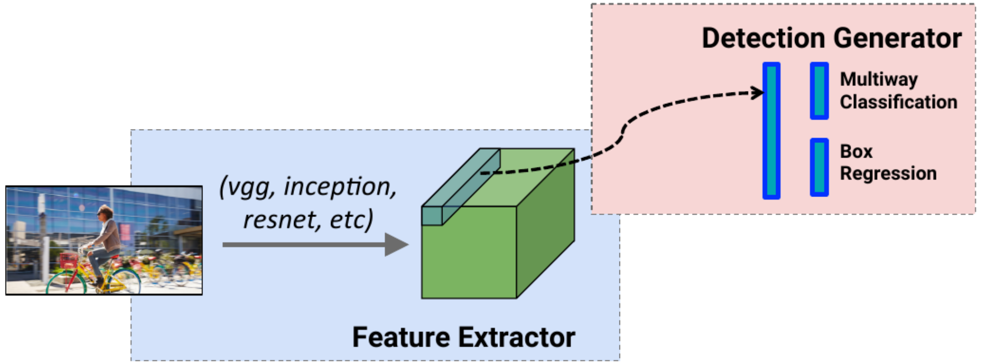
Region Based.

R-CNN, Fast(er) R-CNN, R-FCN
ION, HyperNet, OHEM, MR-CNN, etc.

Region Free.

YOLO, SSD etc.

Two object detection architectures



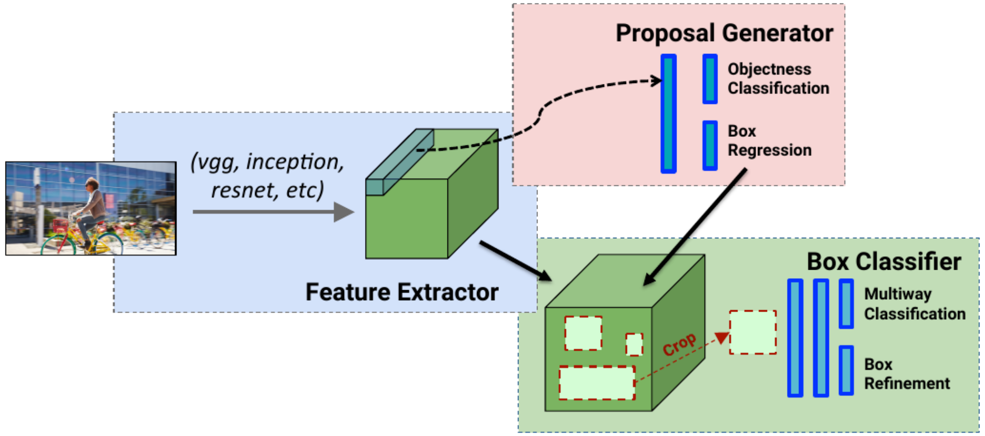
SSD

a) featurized image pyramid

b) Single-shot detector with on CNN

c) Multiple anchors at one level feature

1. High-speed
2. No repeated computations
3. Not easy to train
4. Strugle with small instances



Faster R-CNN

a) Region proposal network

b) Region-wise object detection sub-network

1. High-accuracy, easy to train
2. Easy to follow
3. Resource/time consuming
4. Repeated computation with region-wise computing

So, what is the lesson?

- √ Feature pyramid works better in locating all scales of objects
  from: SSD, HyperNet, ION, MR-CNN
- √ Using region proposal network to reduce searching space
  from: R-FCN, Faster R-CNN, Fast R-CNN
- √ A fully CNN pipeline with no repeated computation can achieve high detection performance.
  from: SSD, R-FCN

RON

So, what is the lesson?

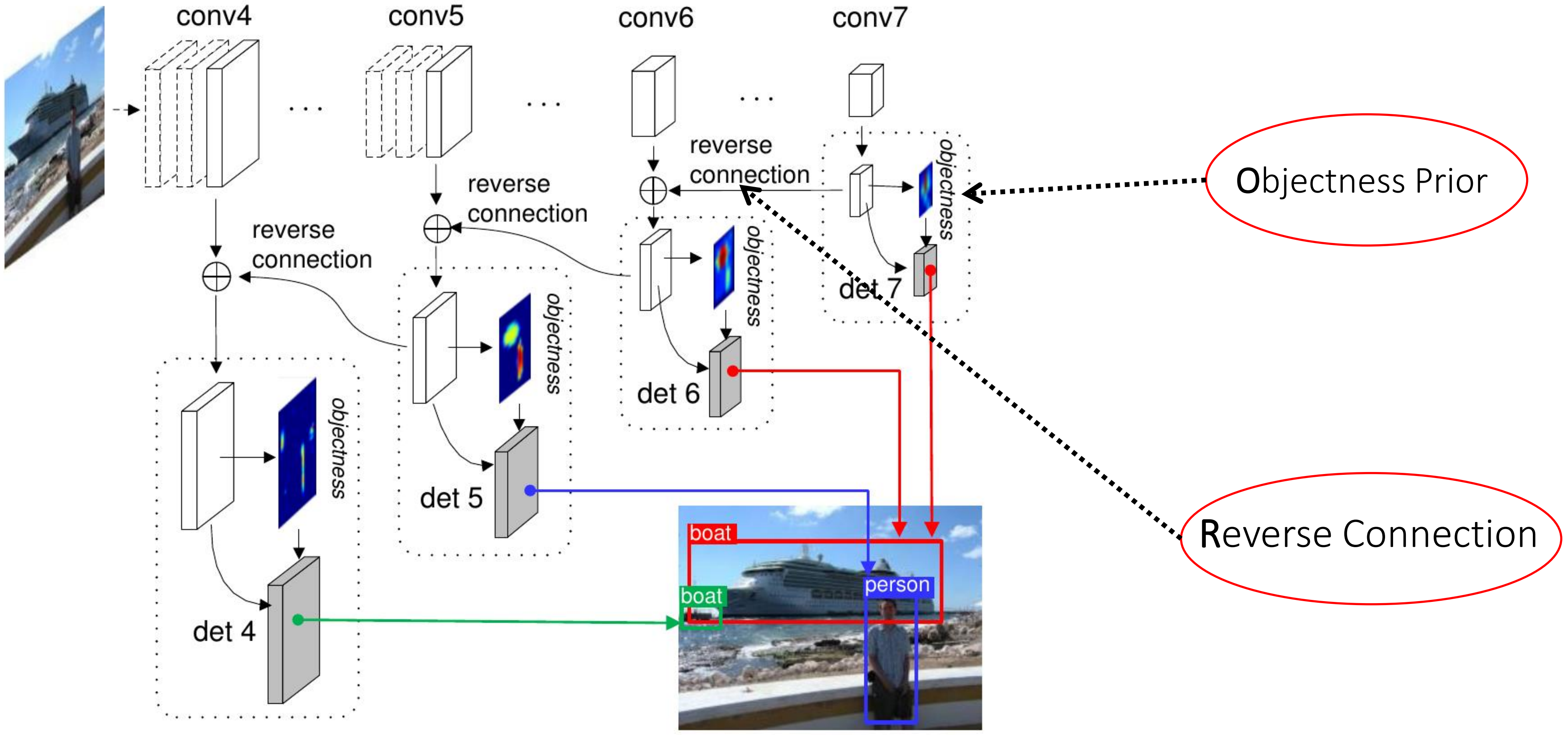
- √ Feature pyramid works better in locating all scales of objects
  from: SSD, HyperNet, ION, MR-CNN

- √ Using region proposal network to reduce searching space
  from: R-FCN, Faster R-CNN, Fast R-CNN

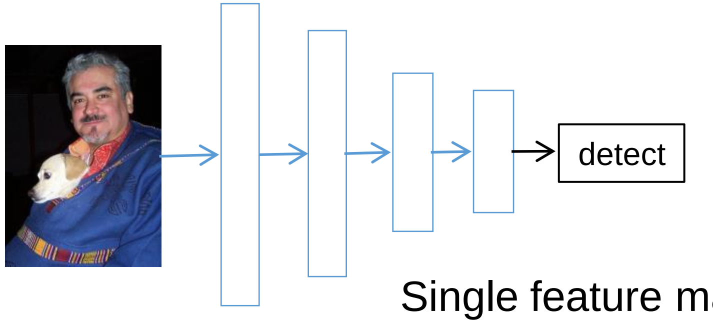
- √ A fully CNN pipeline with no repeated computation can achieve high detection performance.
  from: SSD, R-FCN

RON

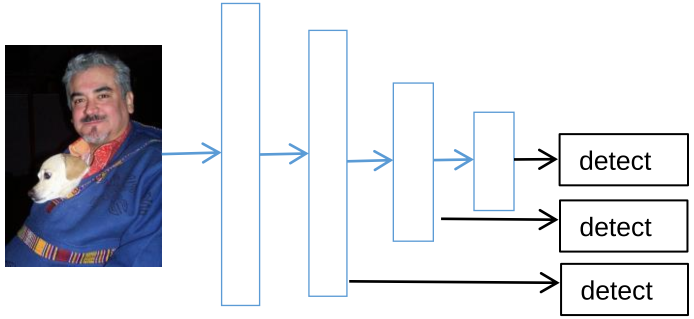
RON: Reverse Connection with Objectness Prior Networks



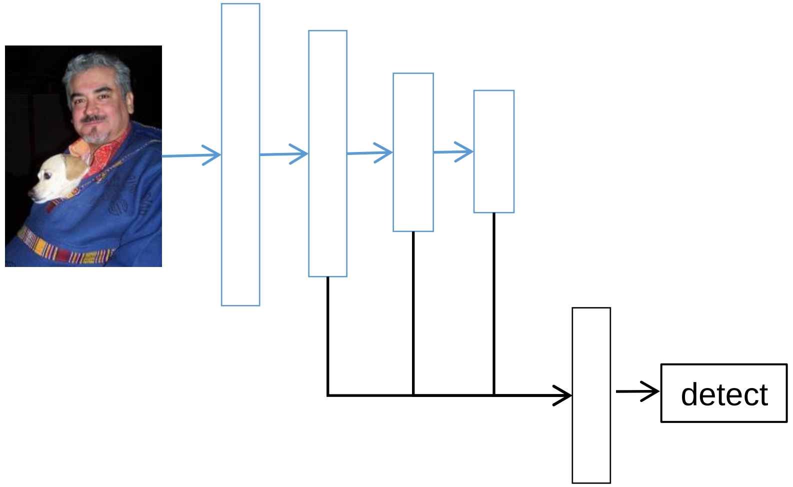
What is reverse connection and why?



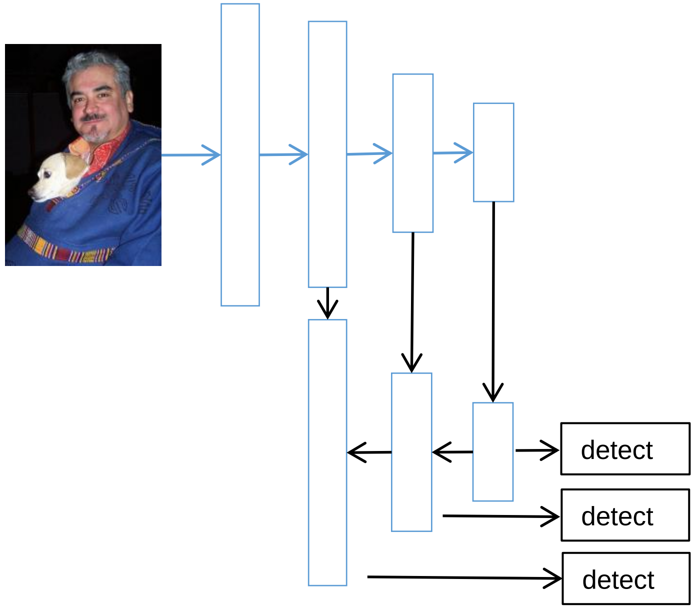
Single feature map
((Fast(er) R-CNN)



Exploring multiple feature directly
(SSD)

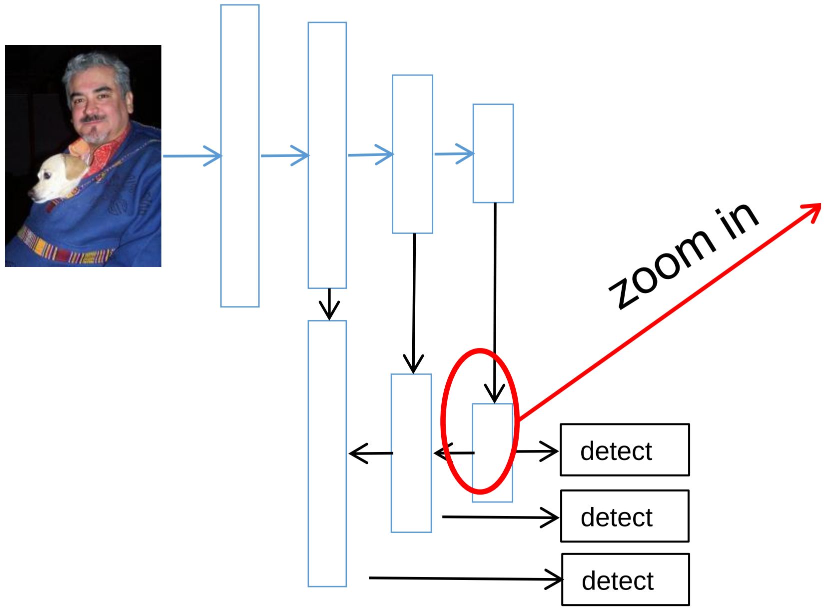


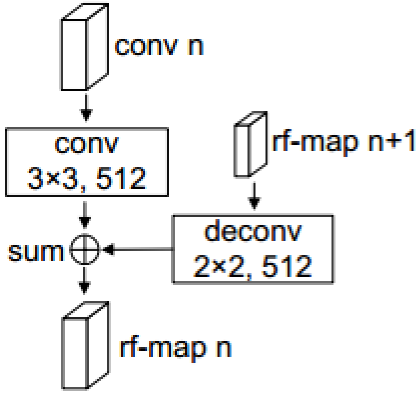
Multiple feature concatenation
(ION, HyperNet)



Multiple feature with
reverse connection

What is reverse connection and why?





A reverse connection block

a) Simple design

b) The semantic information of former layers can be significantly enriched

c) Keep the spatial sizes

d) Easy to do multiple level detection

So, what is the lesson?

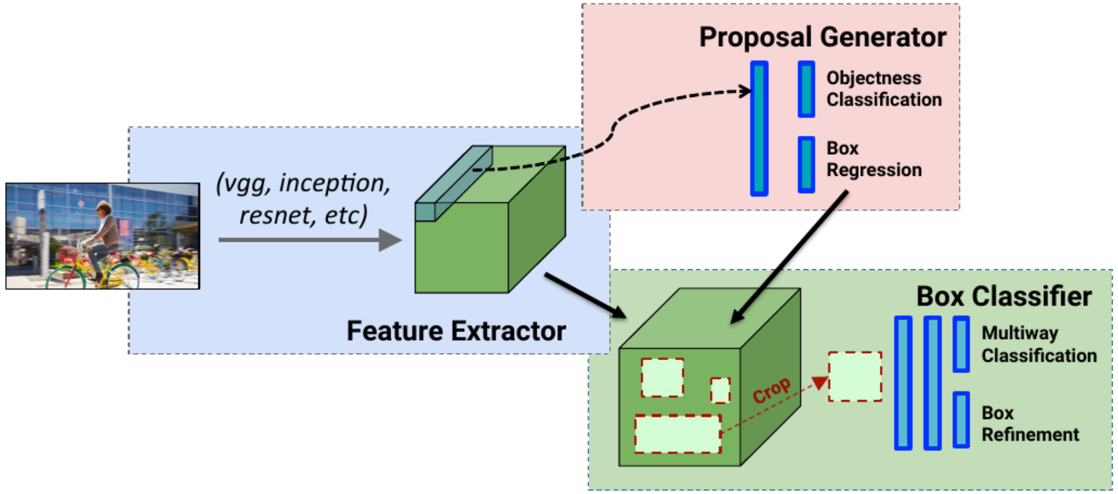
√ Feature pyramid works better in locating all scales of objects
from: SSD, HyperNet, ION, MR-CNN

√ Using region proposal network to reduce searching space
from: R-FCN, Faster R-CNN, Fast R-CNN

√ A fully CNN pipeline with no repeated computation can achieve high detection performance.
from: SSD, R-FCN

RON

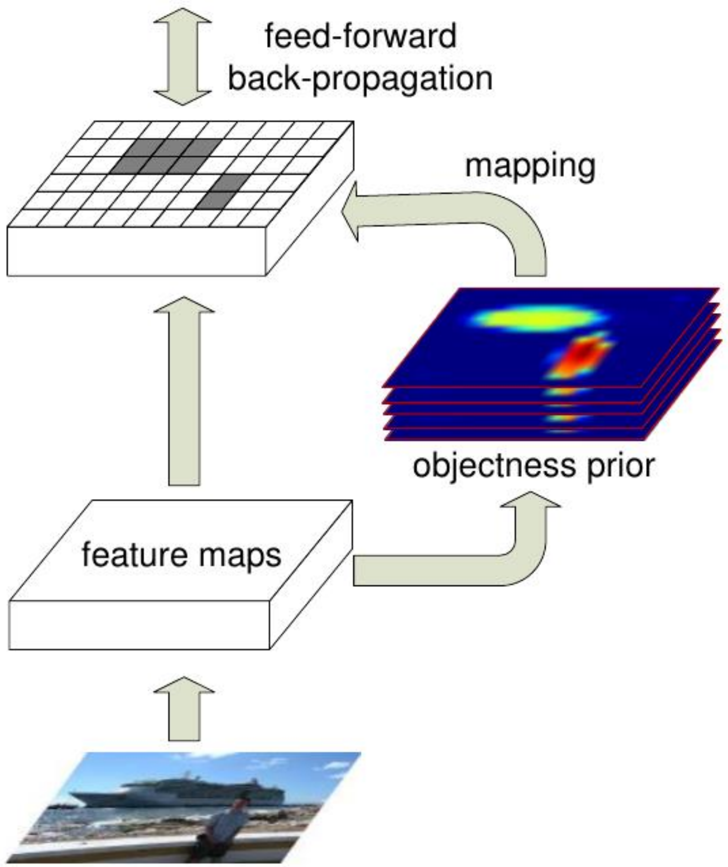
From *region propsoal boxes* to *region proposal maps*



Faster R-CNN:
Region proposal network (RPN) is fully convolutinal, but detection sub-network is with repeated computation (ROI-Pooling).

Why?
The bbox regression in RPN changes the spatial locations of all boxes, which breaks the anchor's reletionship with its corresponding kernel.

From *region propsoal boxes* to *region proposal maps*



Objectness pior:

Share anchors between RPN and detector, make it posible to detect objects with fully ConvNet.

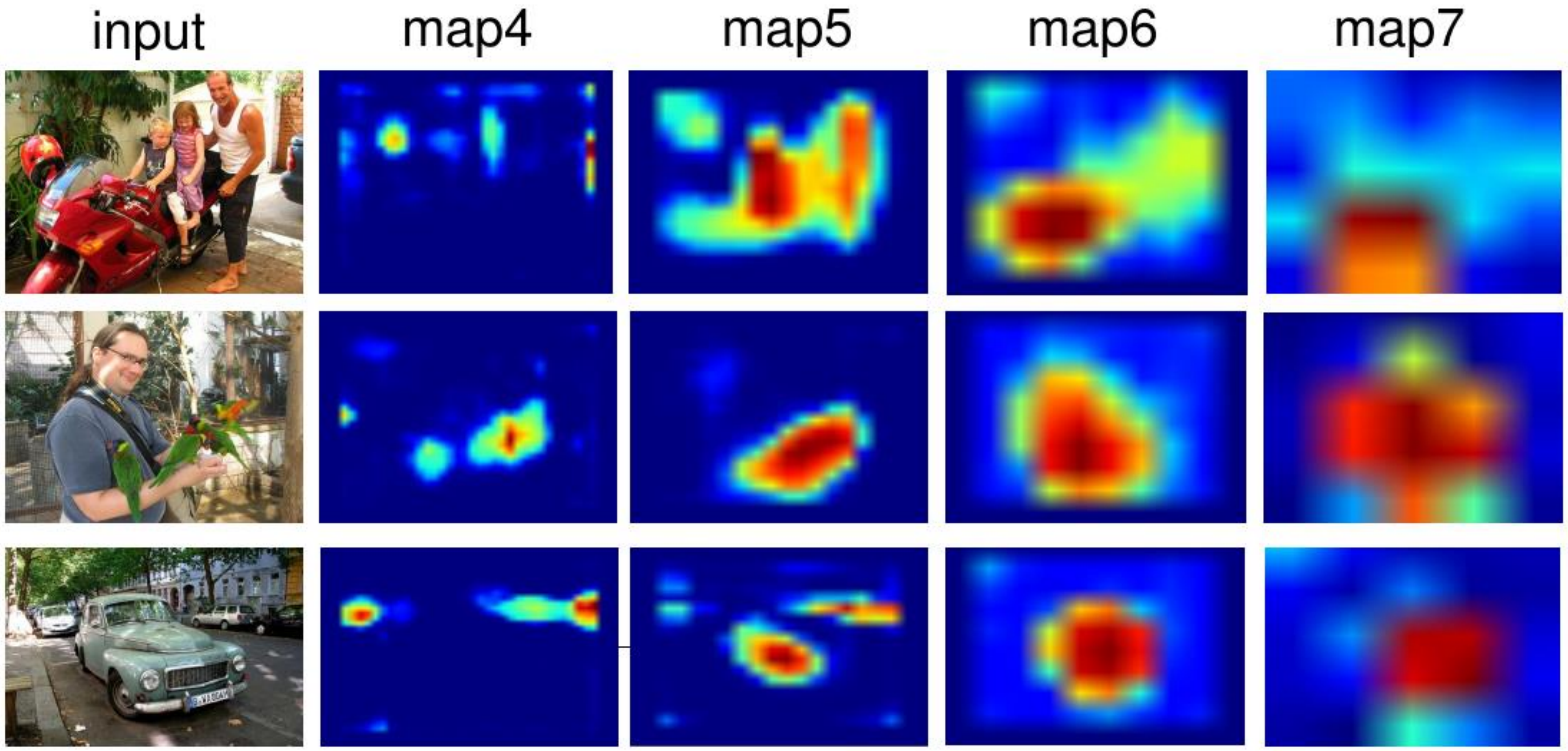
√ No repeated computations, much faster

√ The total network is fully convolutional

√ There are one map for each type of anchors

different from these mask-based methods.

Region proposal maps

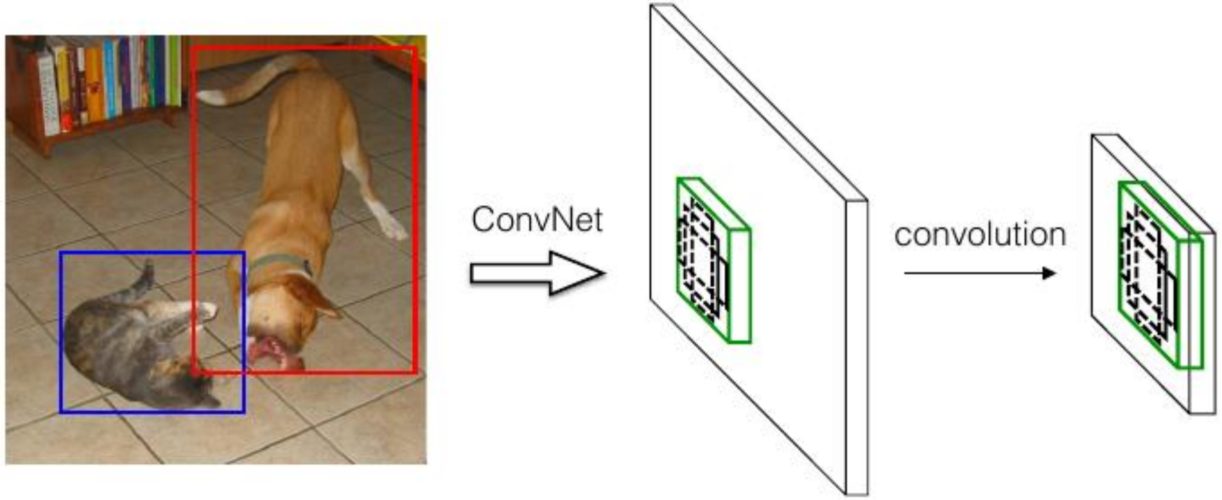


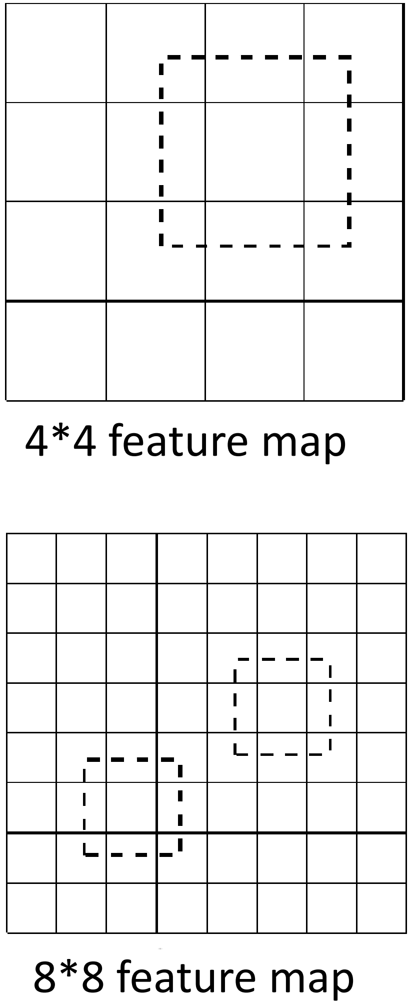
So, what is the lesson?

- √ Feature pyramid works better in locating all scales of objects
  from: SSD, HyperNet, ION, MR-CNN
- √ Using region proposal network to reduce searching space
  from: R-FCN, Faster R-CNN, Fast R-CNN
- √ A fully CNN pipeline with no repeated computation can achieve high detection performance.
  from: SSD, R-FCN

RON

RON bounding box generation





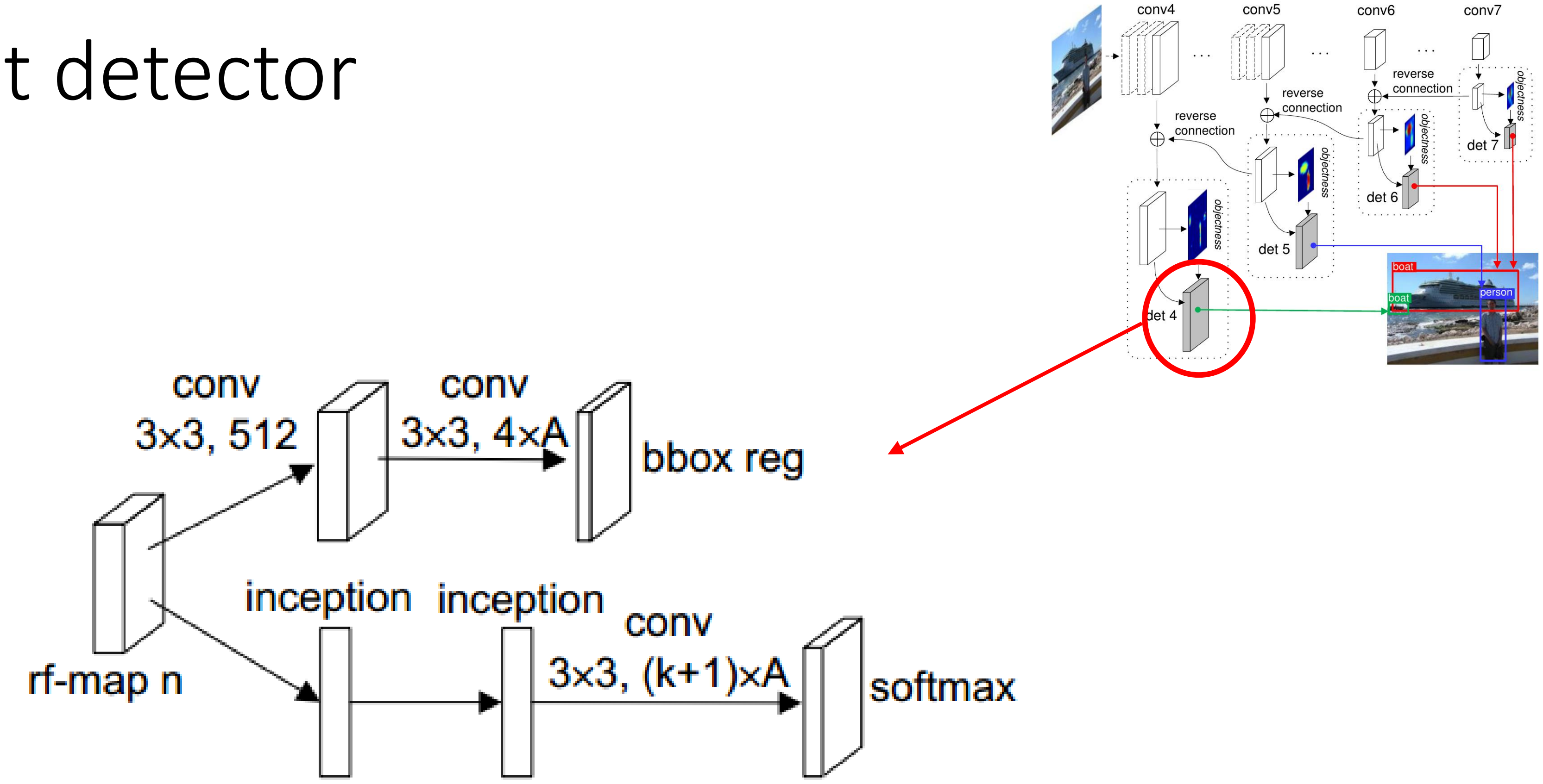
Convolutinal output in this position

a) sub-network for objectness

b) sub-network for detection with a)

c) sub-network for bounding box regression

RON object detector



Object detection and bounding box regression modules. Top: bounding box regression; Bottom: object classification

RON optimization

$$L = \alpha \frac{1}{N_{obj}} L_{obj} + \beta \frac{1}{N_{loc}} L_{loc} + (1 - \alpha - \beta) \frac{1}{N_{cls|obj}} L_{cls|obj}$$

objectness prior

bbox location

detection

optimize the network jointly

Main results

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast R-CNN[10]	70.0	77.0	78.1	69.3	59.4	38.3	81.6	78.6	86.7	42.8	78.8	68.9	84.7	82.0	76.6	69.9	31.8	70.1	74.8	80.4	70.4
Faster R-CNN[23]	73.2	76.5	79.0	70.9	65.5	52.1	83.1	84.7	86.4	52.0	**81.9**	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83.0	72.6
SSD300[19]	72.1	75.2	79.8	70.5	62.5	41.3	81.1	80.8	86.4	51.5	74.3	72.3	83.5	84.6	80.6	74.5	46.0	71.4	73.8	83.0	69.1
SSD500[19]	75.1	79.8	79.5	74.5	63.4	51.9	84.9	**85.6**	87.2	56.6	80.1	70.0	85.4	84.9	80.9	78.2	49.0	**78.4**	72.4	84.6	75.5
RON320	74.2	75.7	79.4	74.8	66.1	53.2	83.7	83.6	85.8	55.8	79.5	69.5	84.5	81.7	83.1	76.1	49.2	73.8	75.2	80.3	72.5
RON384	75.4	78.0	82.4	76.7	67.1	56.9	85.3	84.3	86.1	55.5	80.6	71.4	84.7	84.8	82.4	76.2	47.9	75.3	74.1	83.8	74.5
RON320++	76.6	79.4	**84.3**	75.5	**69.5**	56.9	83.7	84.0	**87.4**	57.9	81.3	**74.1**	84.1	85.3	**83.5**	77.8	49.2	76.7	**77.3**	**86.7**	77.2
RON384++	**77.6**	**86.0**	82.5	**76.9**	69.1	**59.2**	**86.2**	85.5	87.2	**59.9**	81.4	73.3	**85.9**	**86.8**	82.2	**79.6**	**52.4**	78.2	76.0	86.2	**78.0**

+2.5%

Table 1. Detection results on PASCAL VOC 2007 test set. The entries with the best APs for each object category are bold-faced.

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Fast R-CNN[10]	68.4	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
OHEM[26]	71.9	83.0	81.3	72.5	55.6	49.0	78.9	74.7	89.5	52.3	75.0	61.0	**87.9**	80.9	82.4	76.3	47.1	72.5	67.3	80.6	**71.2**
Faster R-CNN[23]	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
HyperNet[16]	71.4	84.2	78.5	73.6	55.6	53.7	78.7	79.8	87.7	49.6	74.9	52.1	86.0	81.7	83.3	**81.8**	48.6	73.5	59.4	79.9	65.7
SSD300[19]	70.3	84.2	76.3	69.6	53.2	40.8	78.5	73.6	88.0	50.5	73.5	61.7	85.8	80.6	81.2	77.5	44.3	73.2	66.7	81.1	65.8
SSD500[19]	73.1	84.9	82.6	74.4	55.8	50.0	80.3	78.9	88.8	53.7	76.8	59.4	87.6	83.7	82.6	81.4	47.2	75.5	65.6	84.3	68.1
RON320	71.7	84.1	78.1	71.0	56.8	46.9	79.0	74.7	87.5	52.5	75.9	60.2	84.8	79.9	82.9	78.6	47.0	75.7	66.9	82.6	68.4
RON384	73.0	85.4	80.6	71.9	56.3	49.8	80.6	76.8	88.2	53.6	78.1	60.4	86.4	81.5	83.8	79.4	48.6	77.4	67.7	83.4	69.5
RON320++	74.5	**87.1**	81.0	74.6	58.8	51.7	**82.1**	77.0	89.7	57.2	79.9	**62.6**	87.2	83.2	**85.0**	80.5	51.4	76.7	68.5	**84.8**	70.4
RON384++	**75.4**	86.5	**82.9**	**76.6**	**60.9**	**55.8**	81.7	**80.2**	**91.1**	**57.3**	**81.1**	60.4	87.2	**84.8**	84.9	81.7	**51.9**	**79.1**	**68.6**	84.1	70.3

+2.3%

Table 2. Results on PASCAL VOC 2012 test set. All methods are based on the pre-trained VGG-16 networks.

Main results

Method	Train Data	Average Precision		
		0.5	0.75	0.5:0.95
Fast R-CNN[10]	train	35.9	-	19.7
OHEM[26]	trainval	42.5	22.2	22.6
OHEM++[26]	trainval	45.9	26.1	25.5
Faster R-CNN[23]	trainval	42.7	-	21.9
SSD300[19]	trainval35k	38.0	20.5	20.8
SSD500[19]	trainval35k	43.7	24.7	24.4
RON320	trainval	44.7	22.7	23.6
RON384	trainval	46.5	25.0	25.4
RON320++	trainval	47.5	25.9	26.2
$\boxed{\text{RON384++}}$	trainval	**49.5**	**27.1**	**27.4**

+3.7%

Table 3. MS COCO test-dev2015 detection results.

Main results

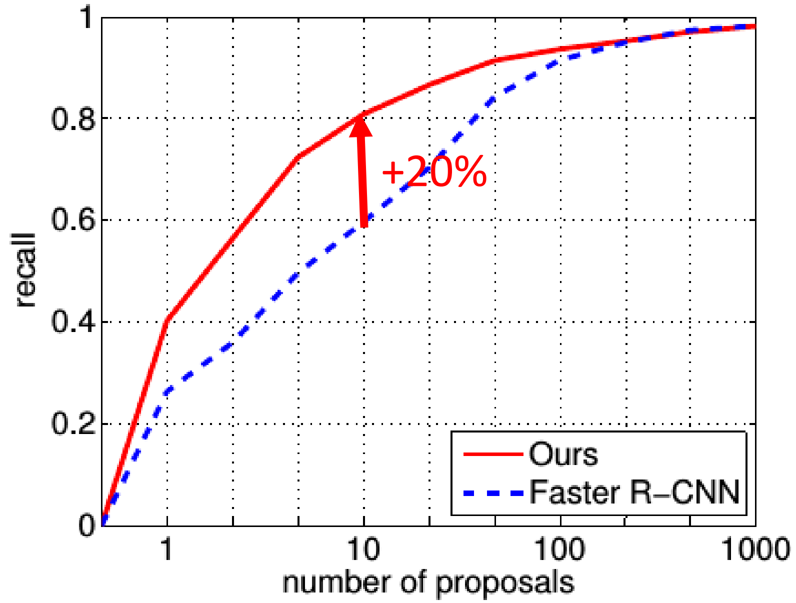
Method	2007 test	2012 test
Faster R-CNN[23]	78.8	75.9
OHEM++[26]	-	80.1
SSD512[19]	-	80.0
RON320	78.7	76.3
RON384	80.2	79.0
RON320++	80.3	78.7
RON384++	**81.3**	**80.7**

Table 4. The performance on PASCAL VOC datasets. All models are pre-trained on MS COCO, and fine-tuned on PASCAL VOC.

Main results

detection from layer 4	5	6	7	mAP
			✓	65.6
		✓	✓	68.3
	✓	✓	✓	72.5
✓	✓	✓	✓	74.2

Table 5. Combining features from different layers.



- Tao Kong, Fuchun Sun, Anbang Yao, Huaping Liu, Ming Lu, Yurong Chen. RON: Reverse Connection with Objectness Prior Networks for Object Detection, In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- Paper: https://arxiv.org/abs/1707.01691

Check out the code/models

https://github.com/taokong/RON

Thanks