



Large Scale Visual Recognition Challenge 2015 (ILSVRC2015)

# Cascade Region Regression for Robust Object Detection

Jiankang Deng, Shaoli Huang, Jing Yang, Hui Shuai, Zhengbo Yu, Zongguang Lu, Qiang Ma, Yali Du,  
Yi Wu, Qingshan Liu, Dacheng Tao

Centre for Quantum Computation & Intelligent Systems (QCIS), University of Technology Sydney (UTS)

Jiangsu Key Laboratory of Big Data Analysis Technology (B-DAT), Nanjing University of Information Science & Technology (NUIST)



# Submission Brief

## (With Additional Training Data)

- Object detection (DET)  
rank 1# (mAP: 0.57848)
- Object localization (LOC)  
rank 2# (Loc error: 0.14574, CIs error: 0.04354)
- Object detection from video (VID)  
rank 1# (mAP: 0.730746)

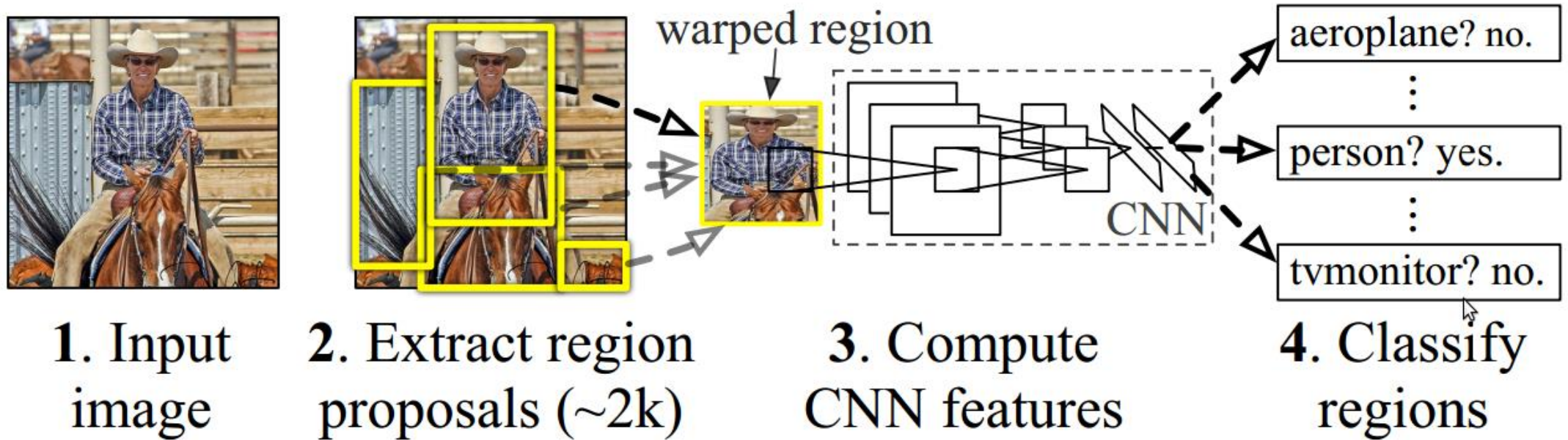
Key idea: **Cascade Region Regression**

“Where” from a former layer, and “What” from a later layer  
Answering “where” more accurately helps answer “what”

[1] P. Dollár, P. Welinder, and P. Perona, “Cascaded pose regression,” in *CVPR*, 2010.

[2] X. Xiong and F. D. la Torre, “Supervised Descent Method and its Applications to Face Alignment,” in *CVPR*, 2013.

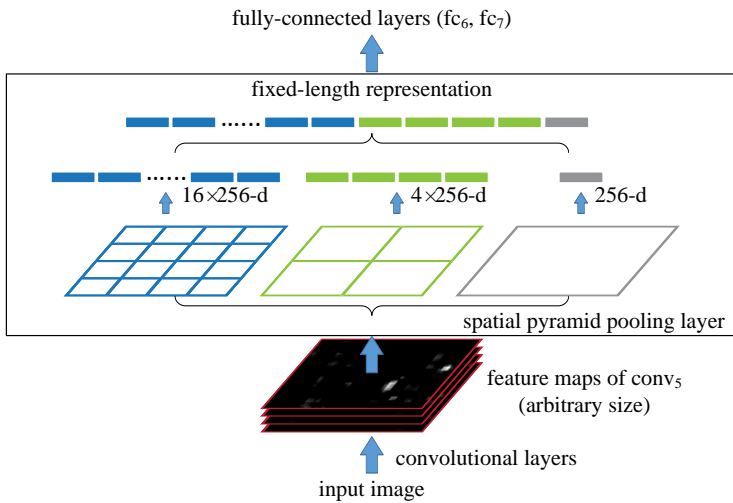
# R-CNN



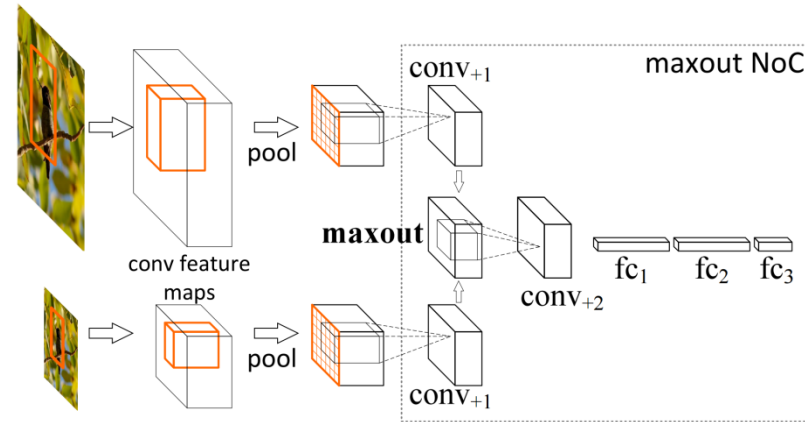
**General framework: Region proposal + DCNN based region classification**

**Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation**, R. Girshick, J. Donahue, T. Darrell, J. Malik, in CVPR 2014

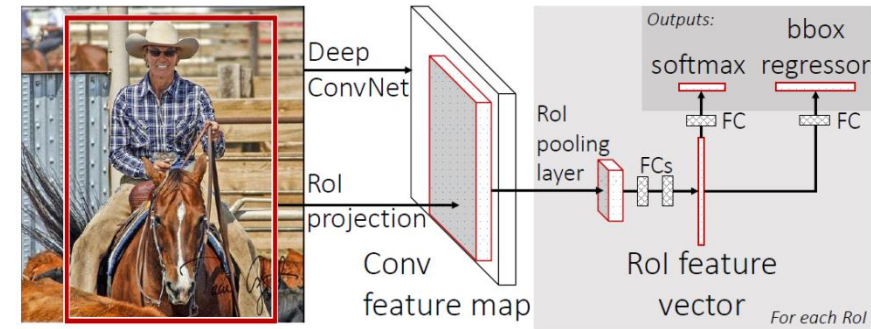
# Improving R-CNN



**SPP-net**



**NoC**



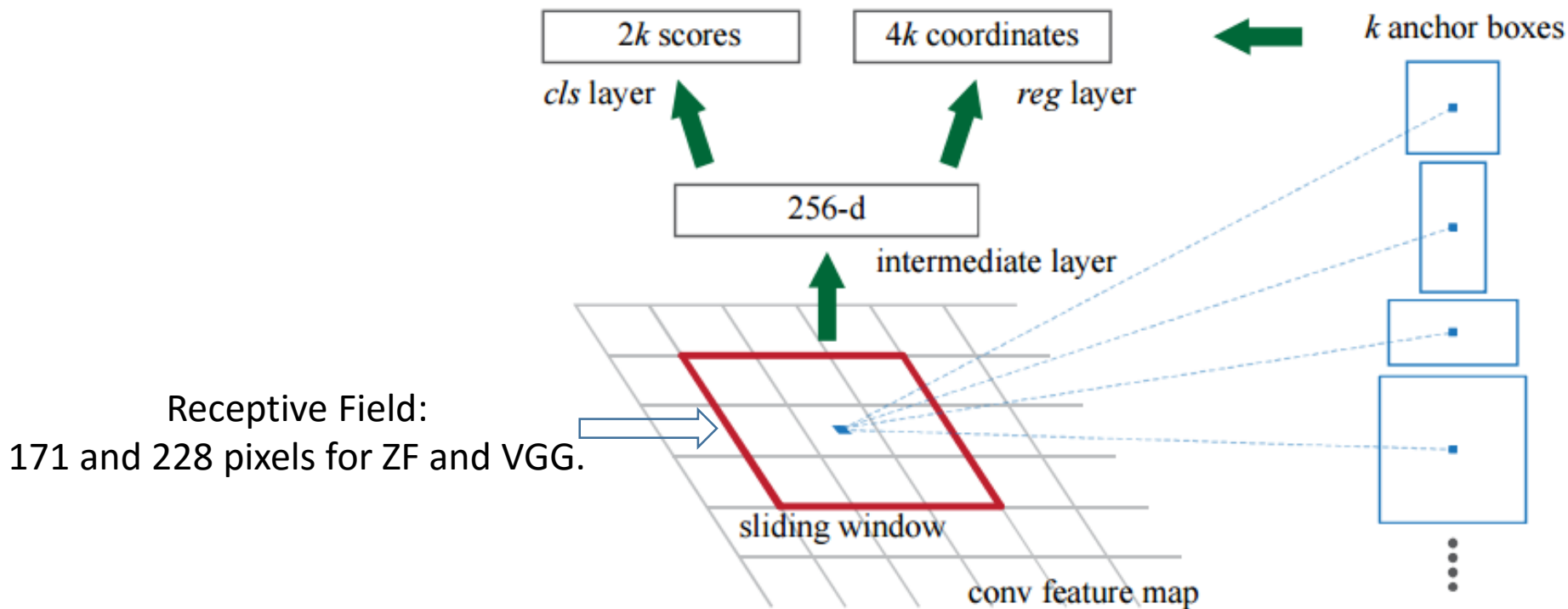
**Fast R-CNN**

1. *Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition*, Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, in ECCV 2014

2. *Object Detection Networks on Convolutional Feature Maps*, Shaoqing Ren, Kaiming He, Ross Girshick, Xiangyu Zhang, Jian Sun, in arXiv 2015

3. *Fast R-CNN*, Ross Girshick, in ICCV 2015

# Improving R-CNN



RPN (Faster R-CNN)

## Observations:

1. More accurate and less number of proposal boxes improve the region classification performance. (Fast R-CNN vs Faster R-CNN)
2. High capacity model usually leads to high performance. (ZF vs VGG)

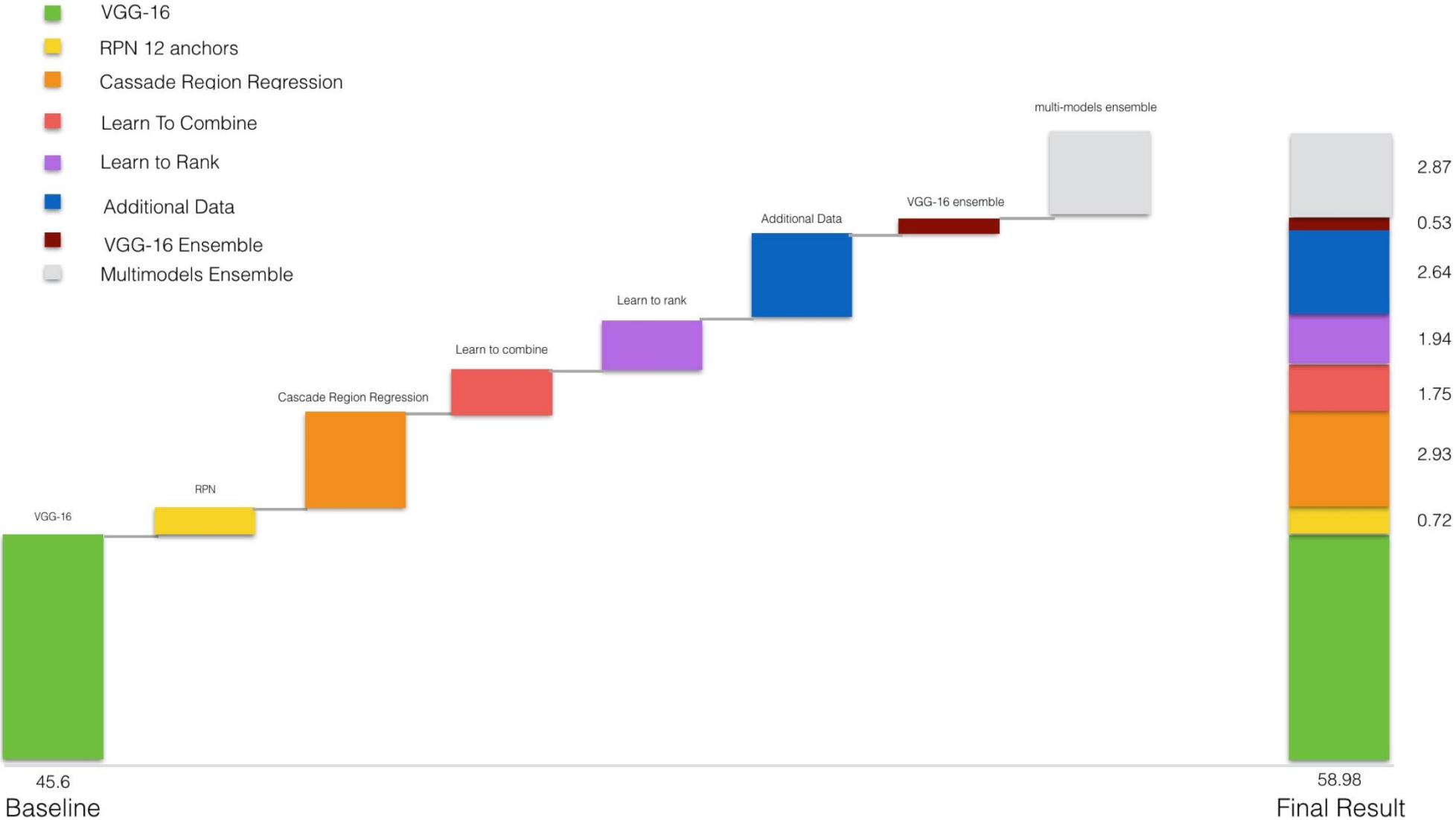
## Question:

Location indexed features are able to regress more accurate boxes.

**What's the condition?**

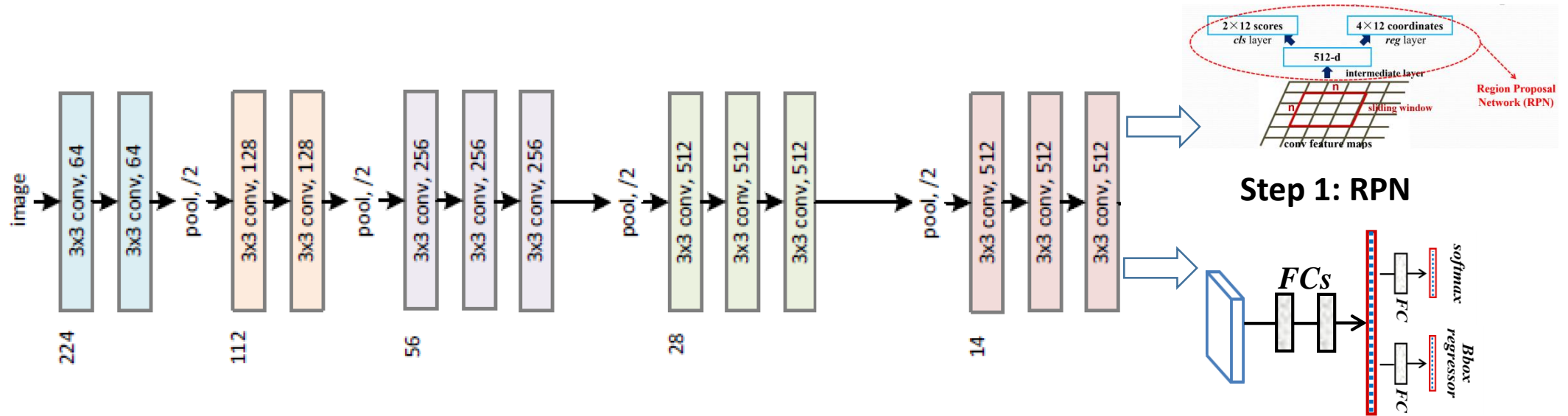
0.7IoU? 0.5IoU? 0.4IoU?

# Our Method



Diagnosis experiments on val2

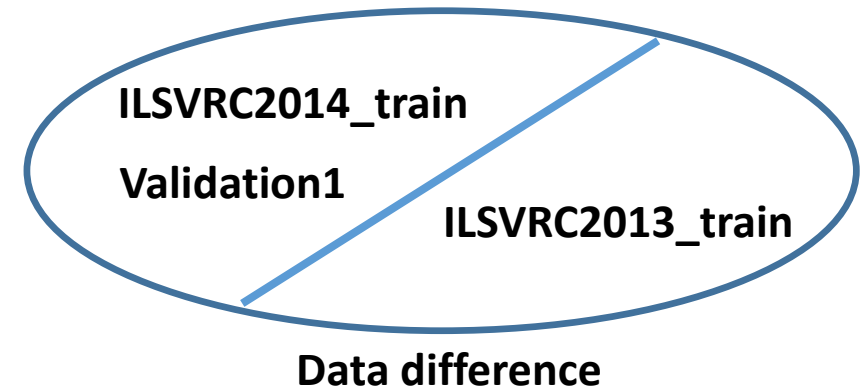
# Faster R-CNN Baseline



## Training procedure:

1. Train Faster R-CNN on ILSVRC2014\_train and Validation1.
2. Get the scores of the annotation boxes on all training data.
3. Remove the wrong annotation at low score.
4. Add leak annotation at high score.
5. Test the model on ILSVRC2013\_train data set.
6. Easy training data (too salient, single object) is removed.
7. Train Faster R-CNN on the refined training data.

## Step 2: Fast R-CNN



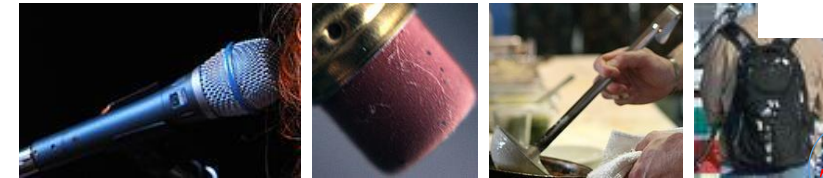
# Easiest and hardest categories



It's easy



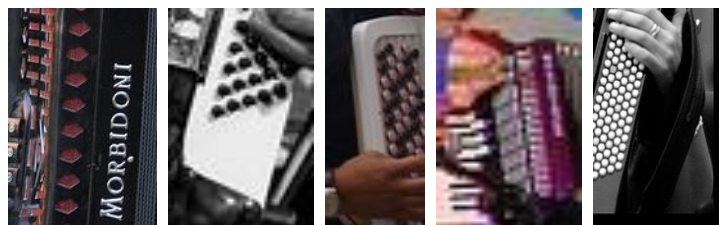
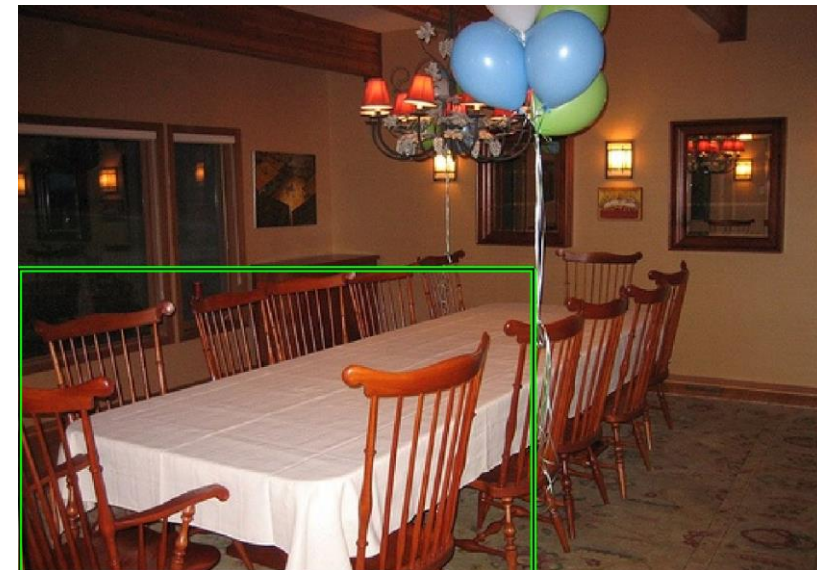
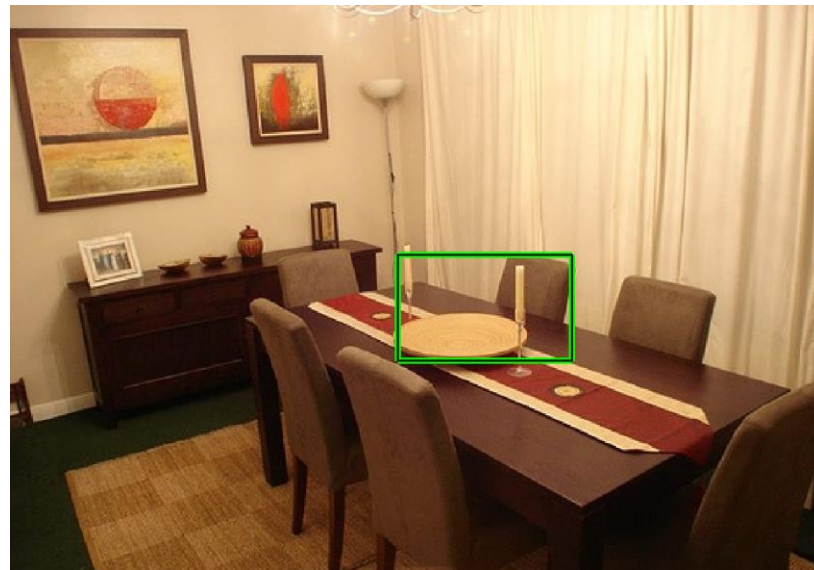
- Large object area within box
- discriminative appearance or shape
- Small variance
- More training data



Too difficult

- Very small object area within box
- Thin objects
- large variance

# False Positive examples



The box is too small.

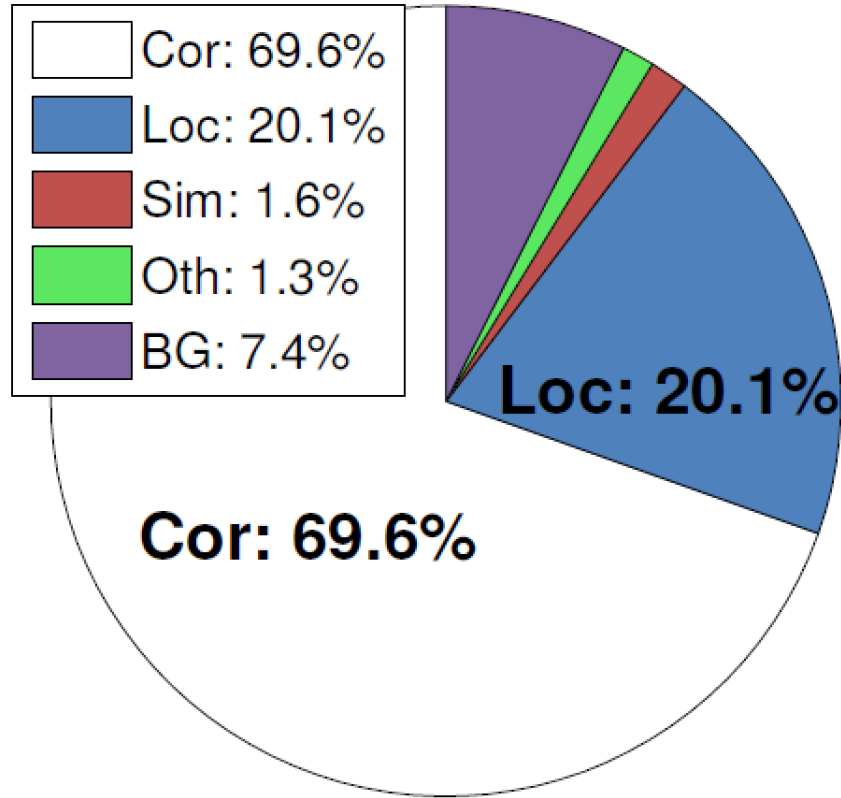
The box is too large.

The box covers dense objects.

Many false positives result from inaccurate localization.



# False Positive Analysis

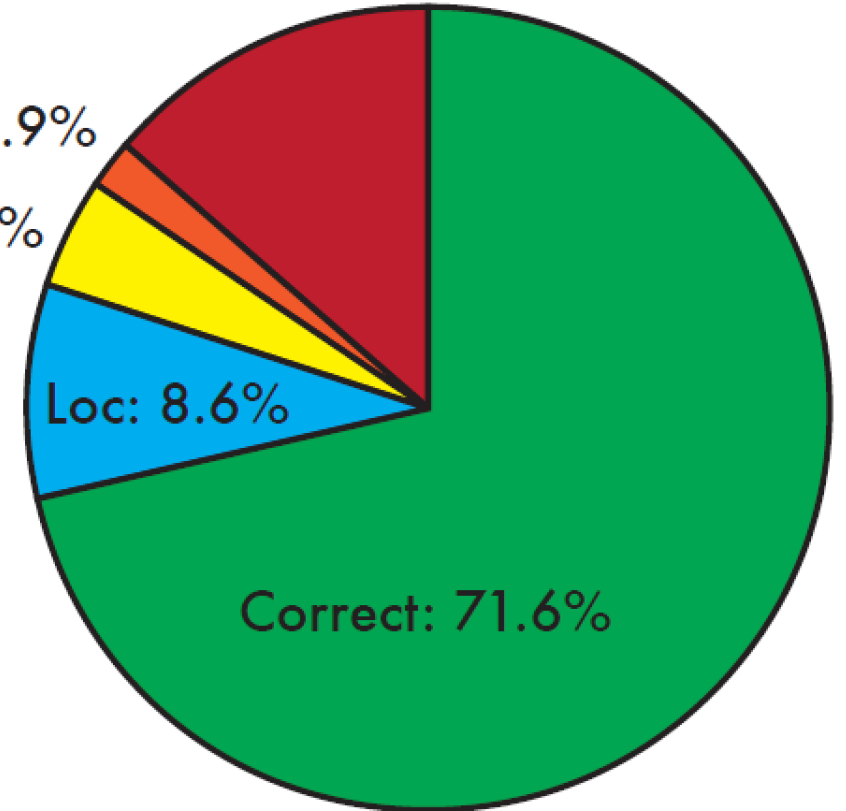


NoC (region based training)

Background: 13.6%

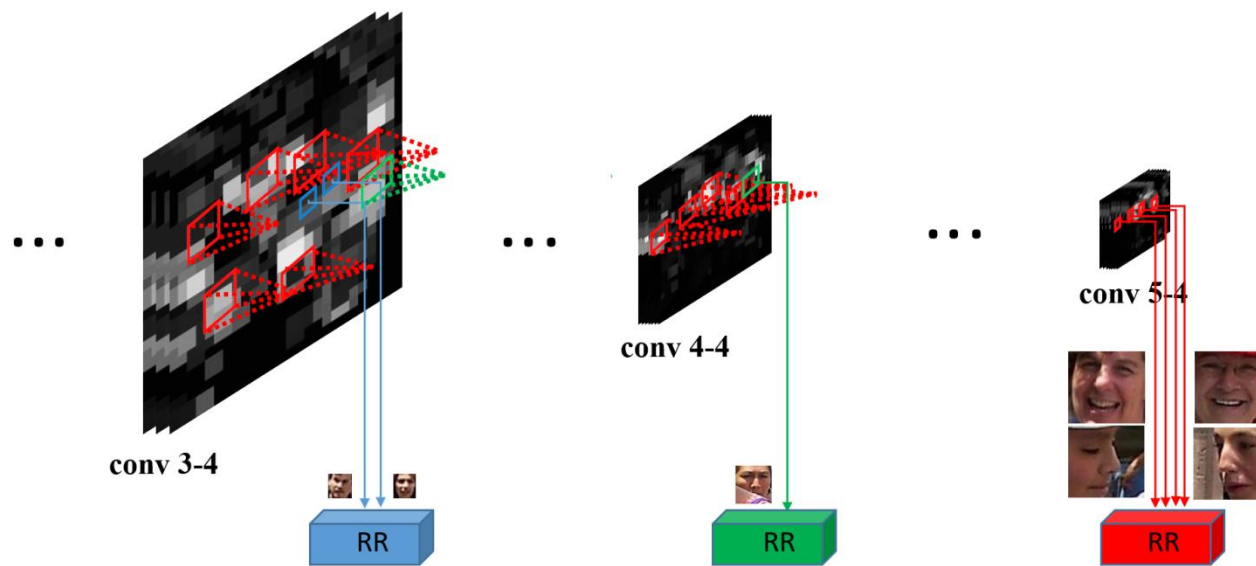
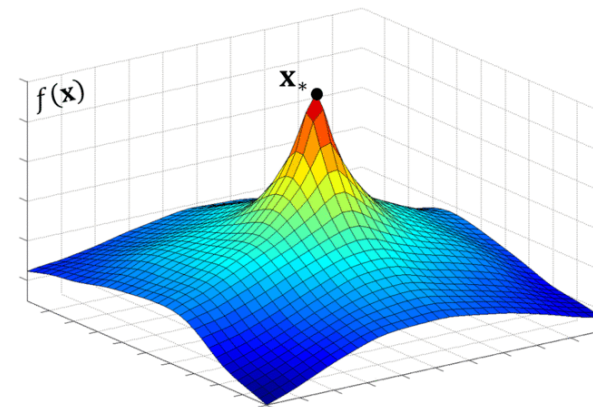
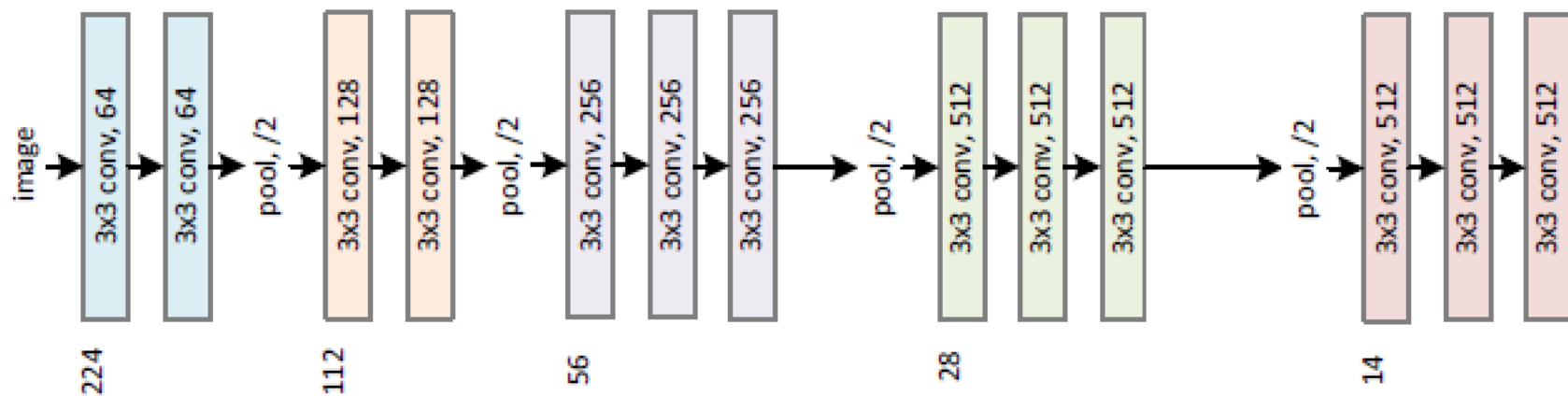
Other: 1.9%

Sim: 4.3%

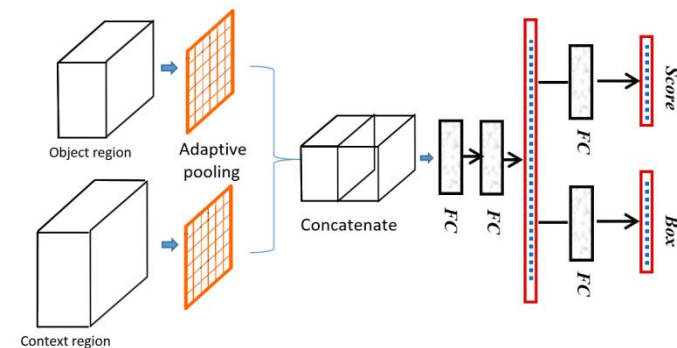


Fast R-CNN (image based training)

# Cascade Region Regression

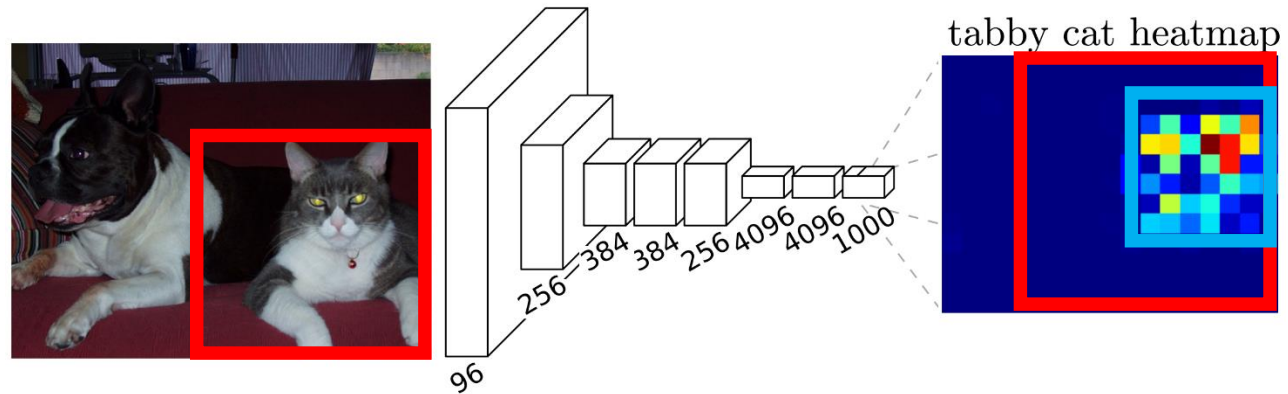


**Multi-layer Conv Feature  
(region size specific)**

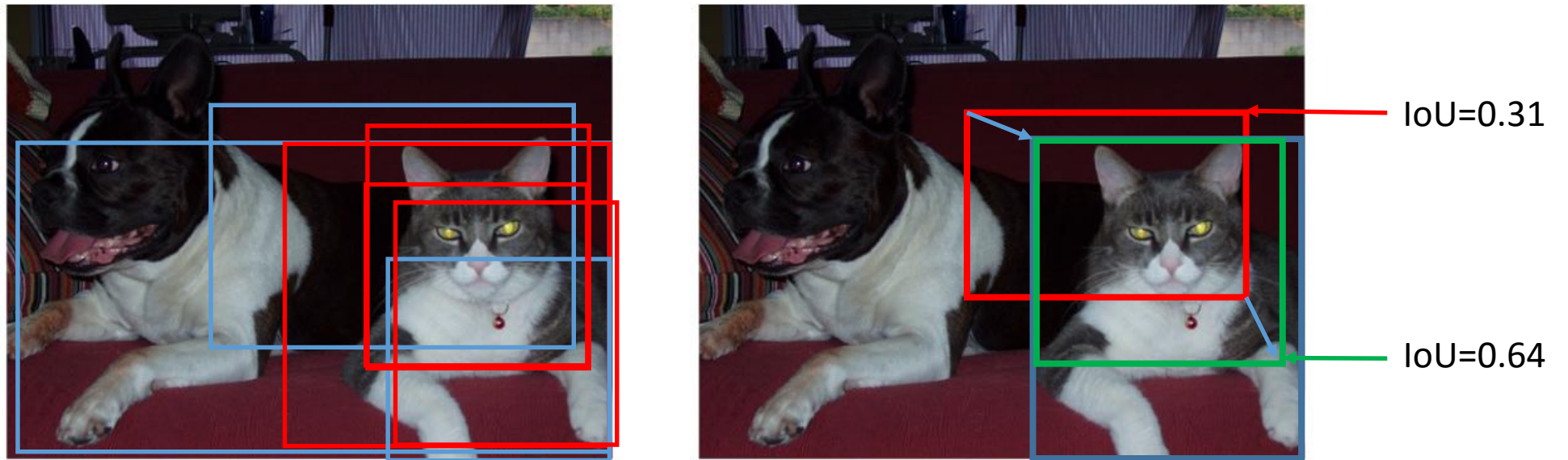


**Multi-scale Conv Feature  
(object + around context)**

# Conditions of Initial location

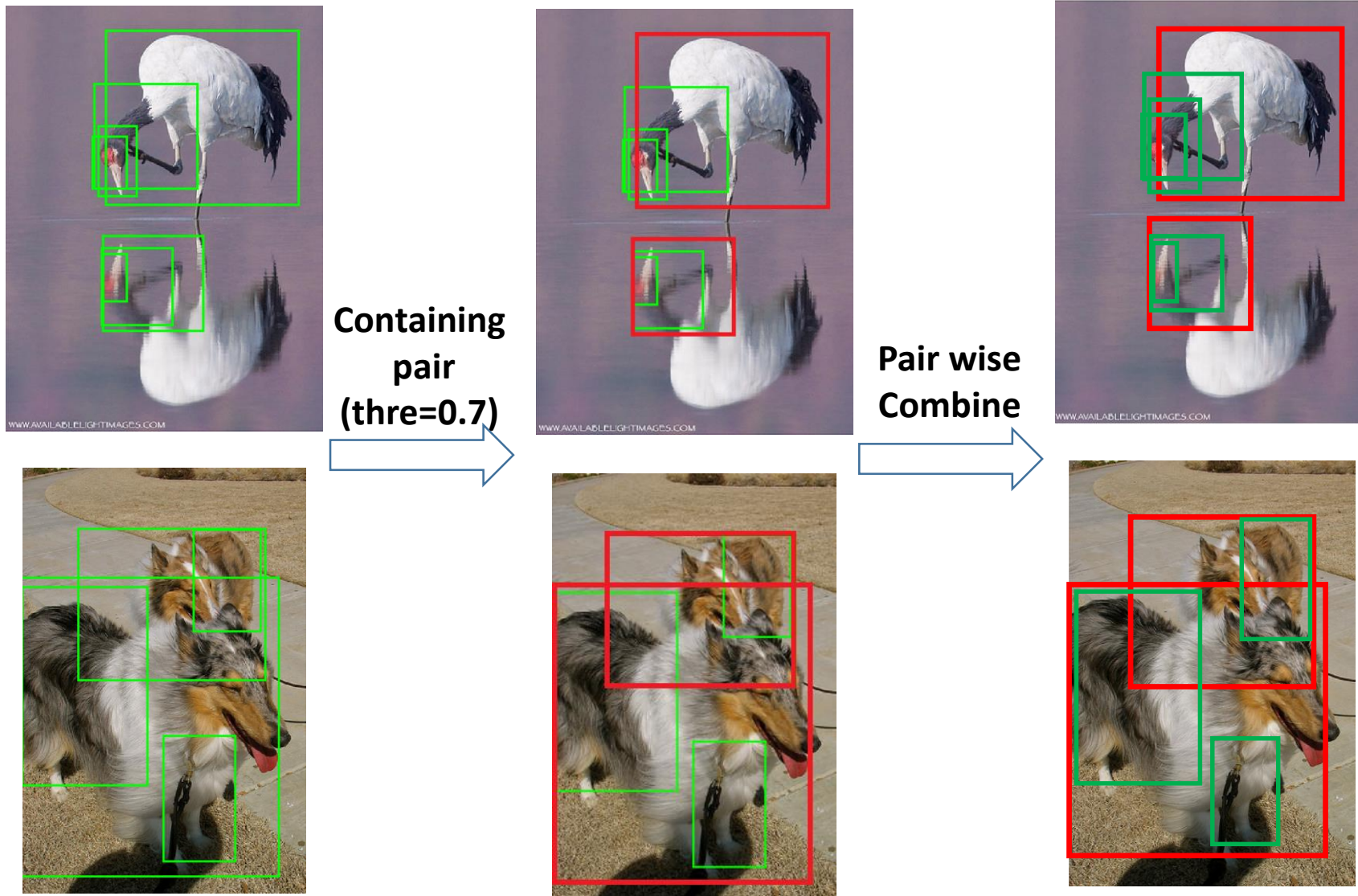


Class-wise energy / box receptive field energy is highly related to the probability of convergence.



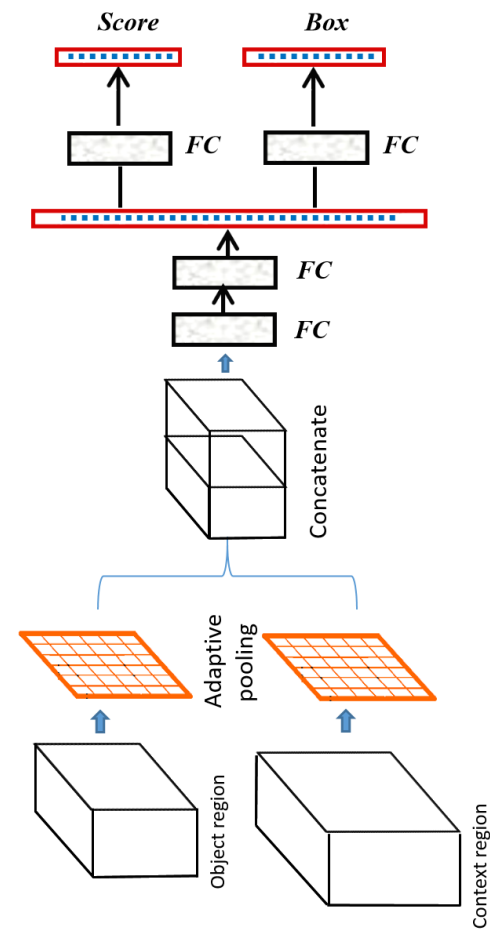
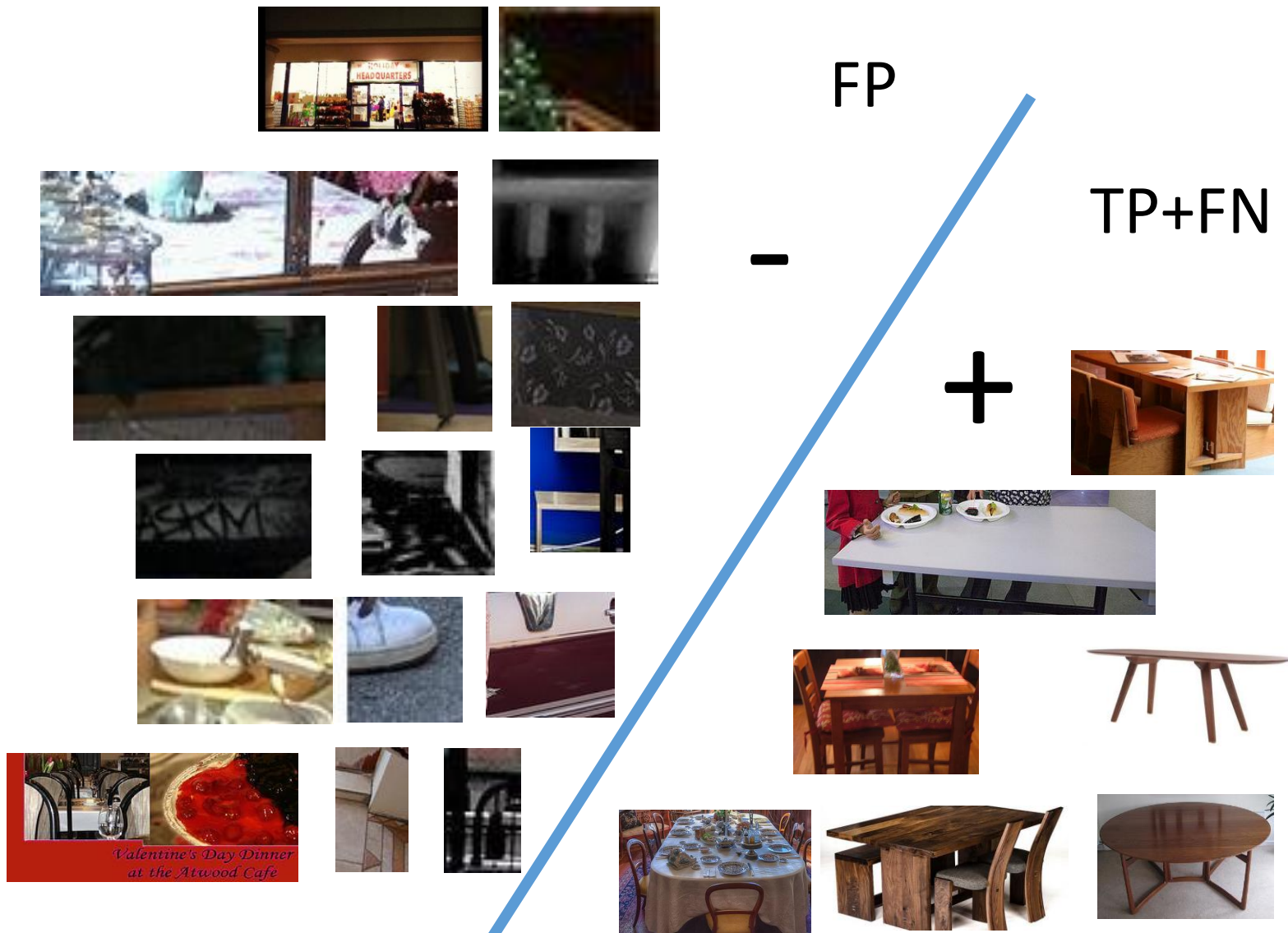
In practice, we define positive examples which can regress better locations (or keep).

# Learning to Combine



**Object detection via a multi-region & semantic segmentation-aware CNN model, Spyros Gidaris, Nikos Komodakis, in ICCV 2015**

# Learning to rank



**Class-specific classifier is trained with SPP-net (multi-scale).**

**Suppress false positives from background.**

# Additional Training Data

ClassName(86)	mAP ↑
accordion	4.27%
ant	5.64%
armadillo	3.93%
balance beam	7.33%
banjo	15.46%
baseball	4.05%
bee	4.72%
binder	2.32%
bow tie	3.54%
bow	3.63%
.....	.....

Add training data



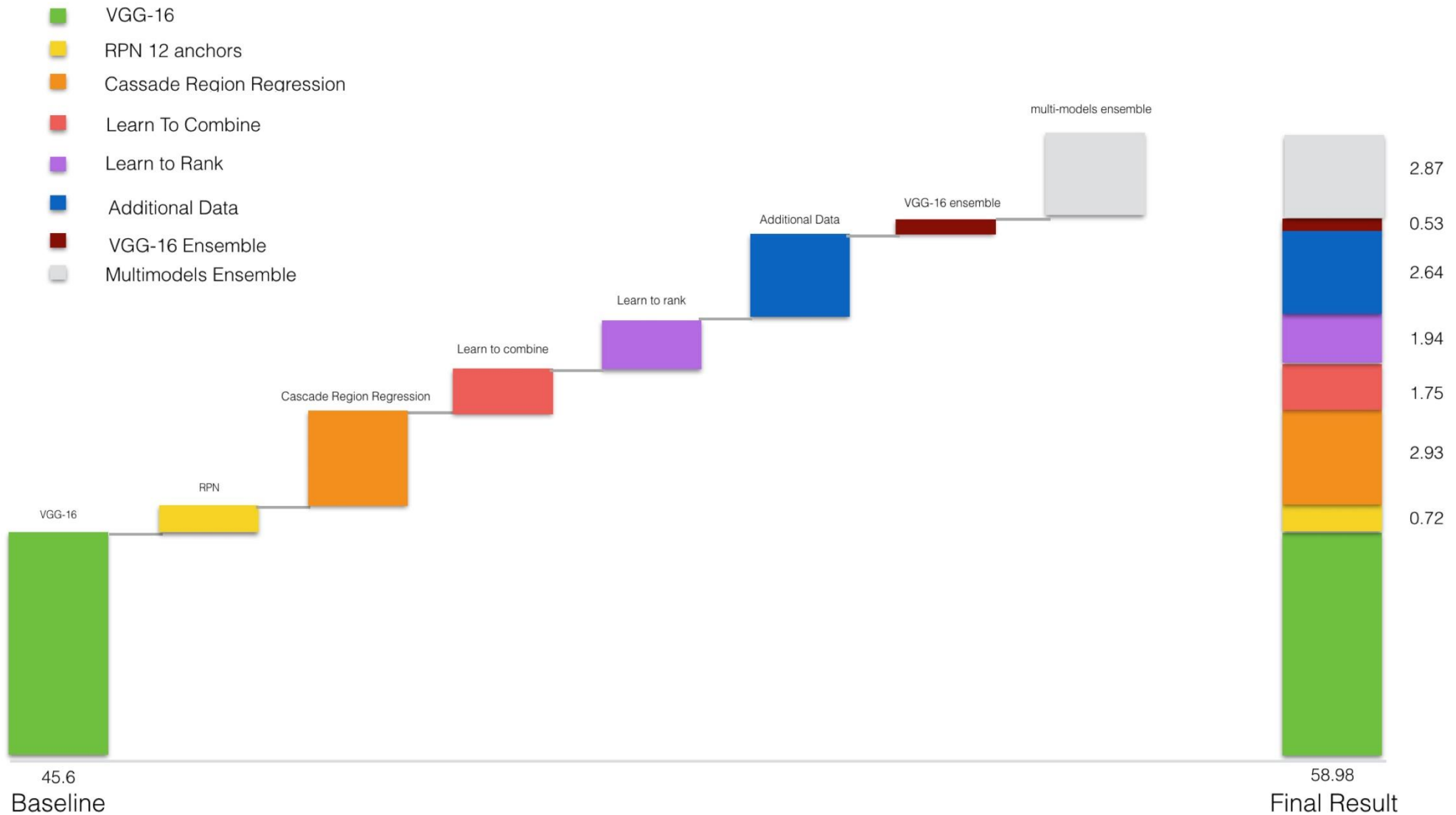
Detection (thre=0.5)



Remove FP, Add FN, Refine boxes

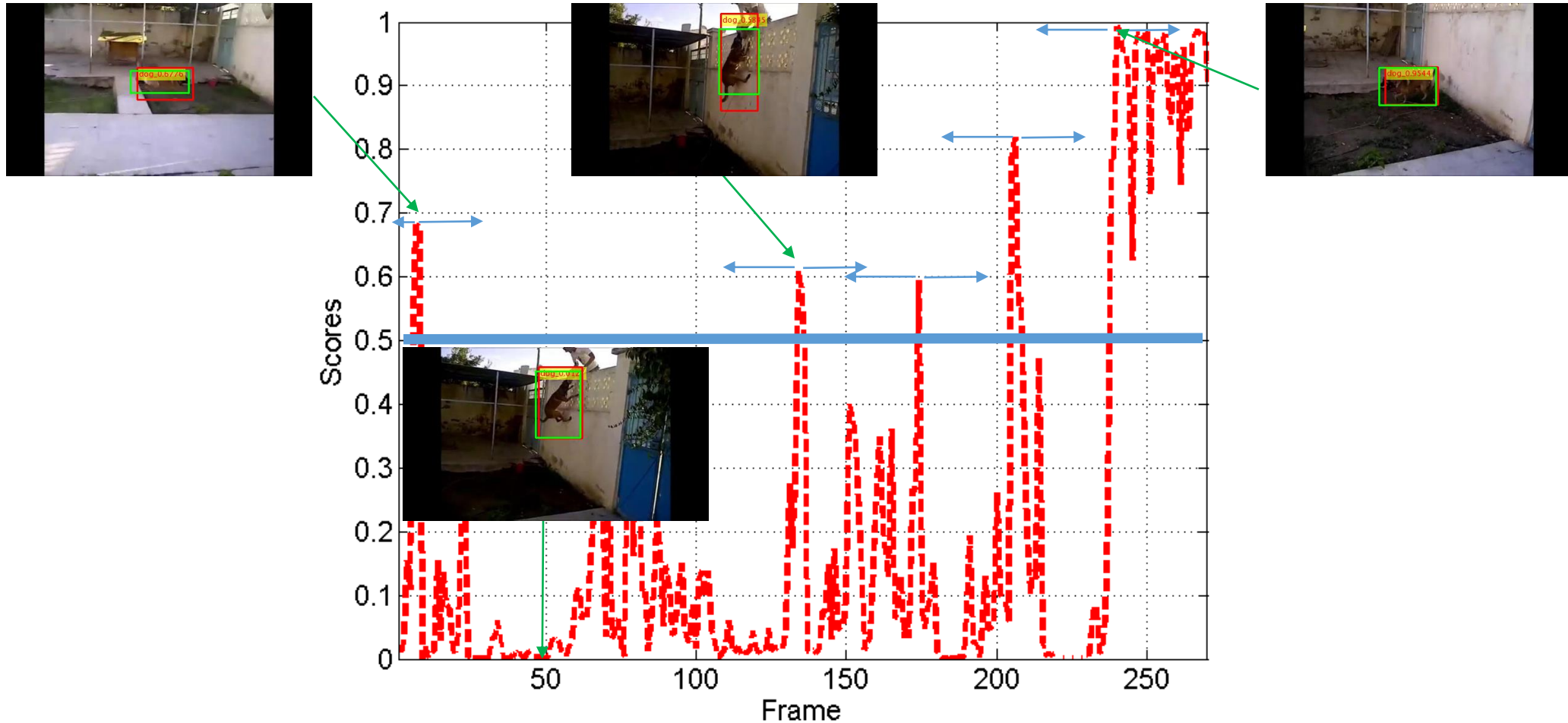


# Trick Validation



Diagnosis experiments on val2

# Object detection from Video



**Object detection on each frame**

**Tracking from the high score frame (temporal smooth)**

**Class-wise box regression and NMS on each frame**

# Object detection from Video



**Scene Cluster (object detection + similarity scene)  
Scene Context is helpful to suppress FP.**

