Project Admin

Someone from every team must come and see me!

If you chose:

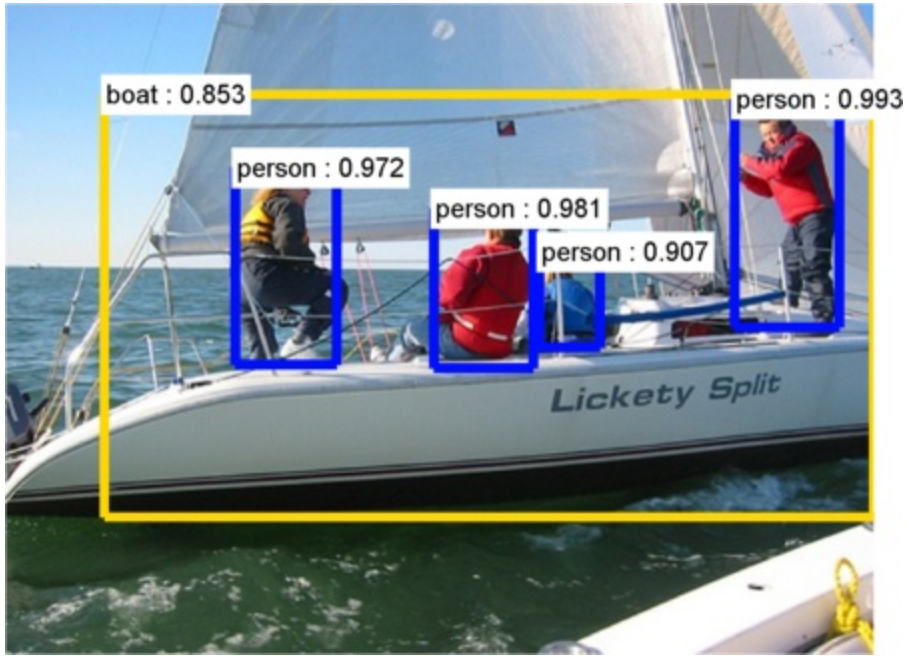
- Depth Prediction
- Neural Style Transfer
- Image Captioning

I will outline what project involves IMMEADIATELY after this class

For other projects, come and see me during office hours.

Object Detection

Image Classification
(what?)

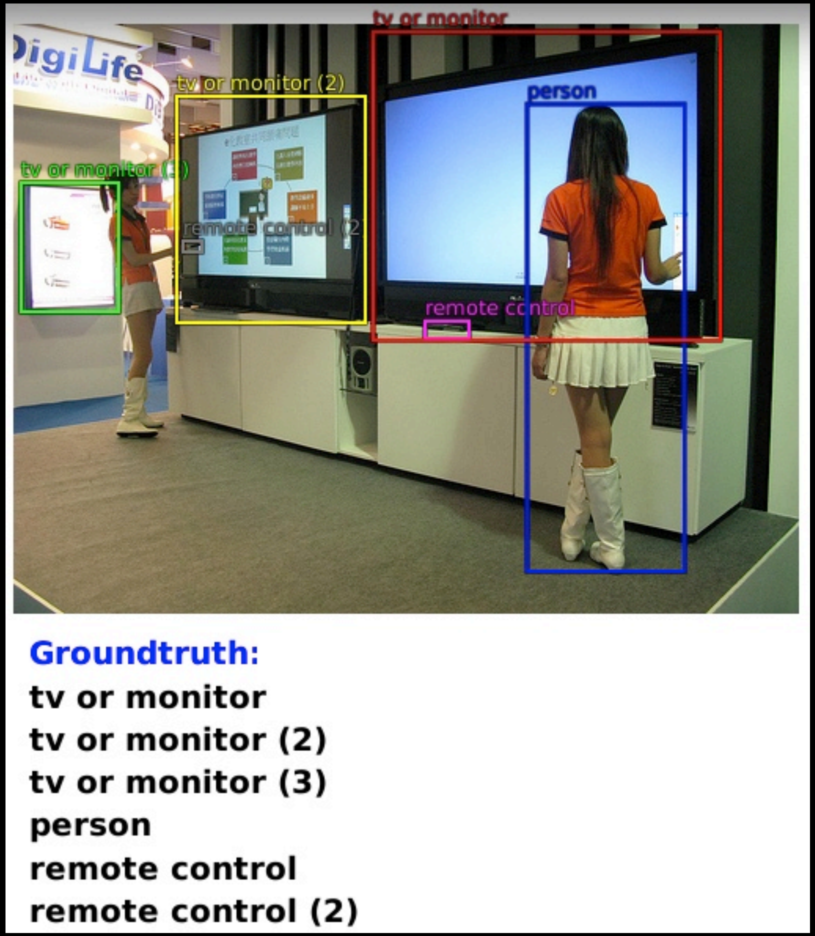


Object Detection
(what + where?)

Detection with ConvNets

- So far, all about classification

- What about localizing objects within the scene?

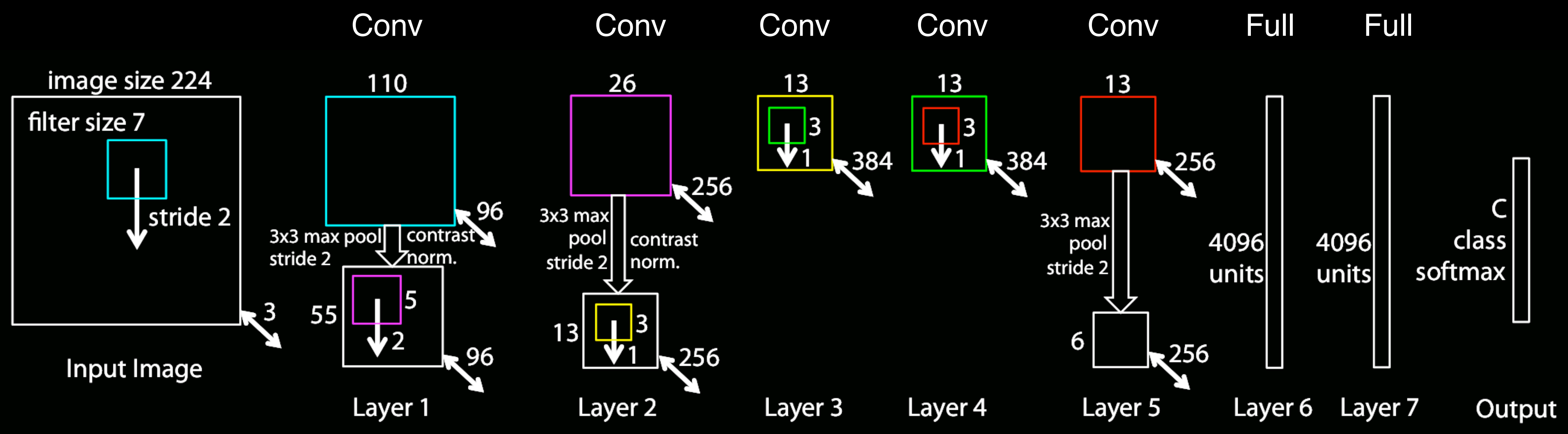


Two General Approaches

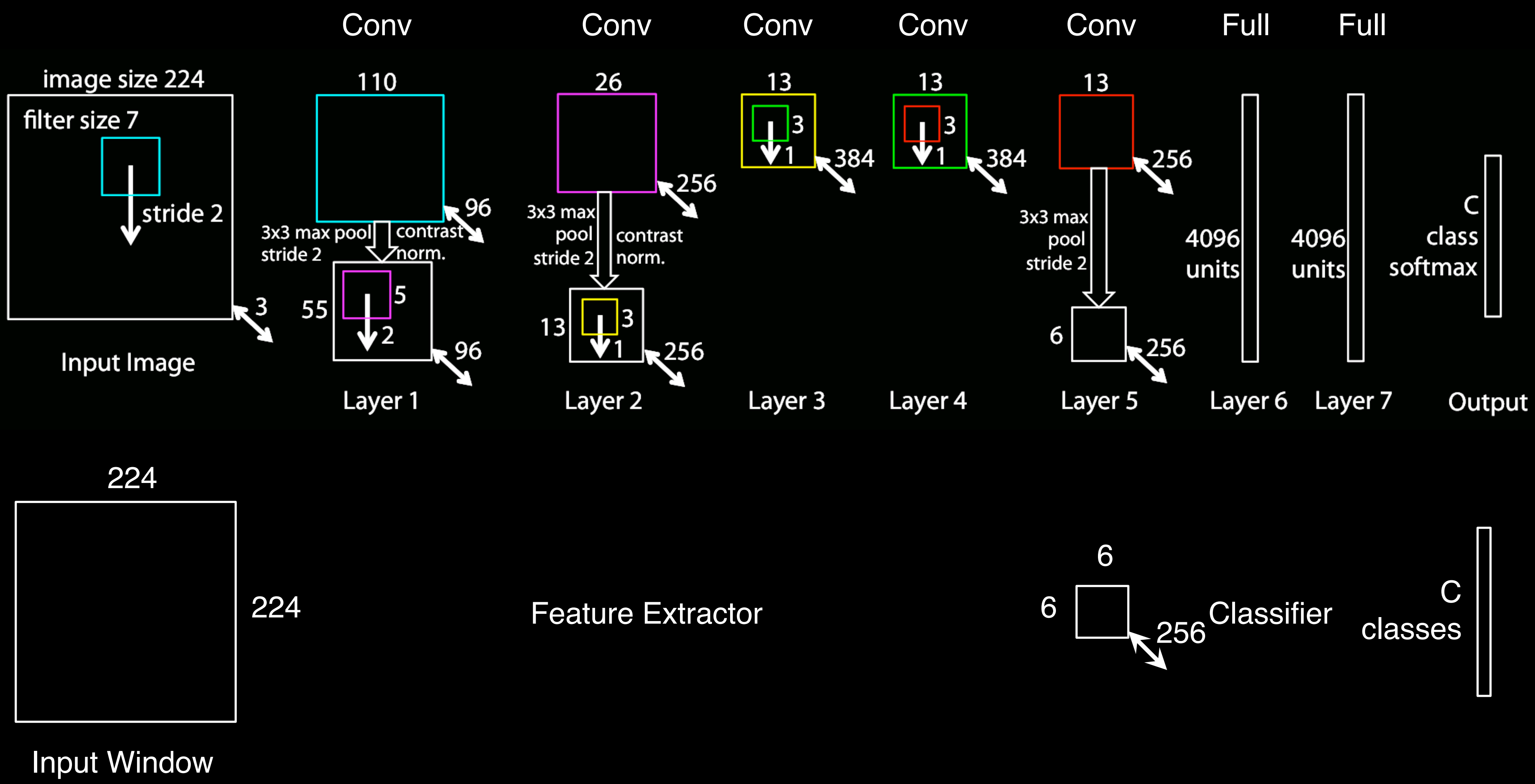
1. Examine very position / scale
   – E.g. Overfeat: Integrated recognition, localization and detection using convolutional networks, Sermanet et al., ICLR 2014

2. Use some kind of proposal mechanism to attend to a set of possible regions
   – E.g. Region-CNN [Rich feature hierarchies for accurate object detection and semantic segmentation, Girshick et al., CVPR 2014]

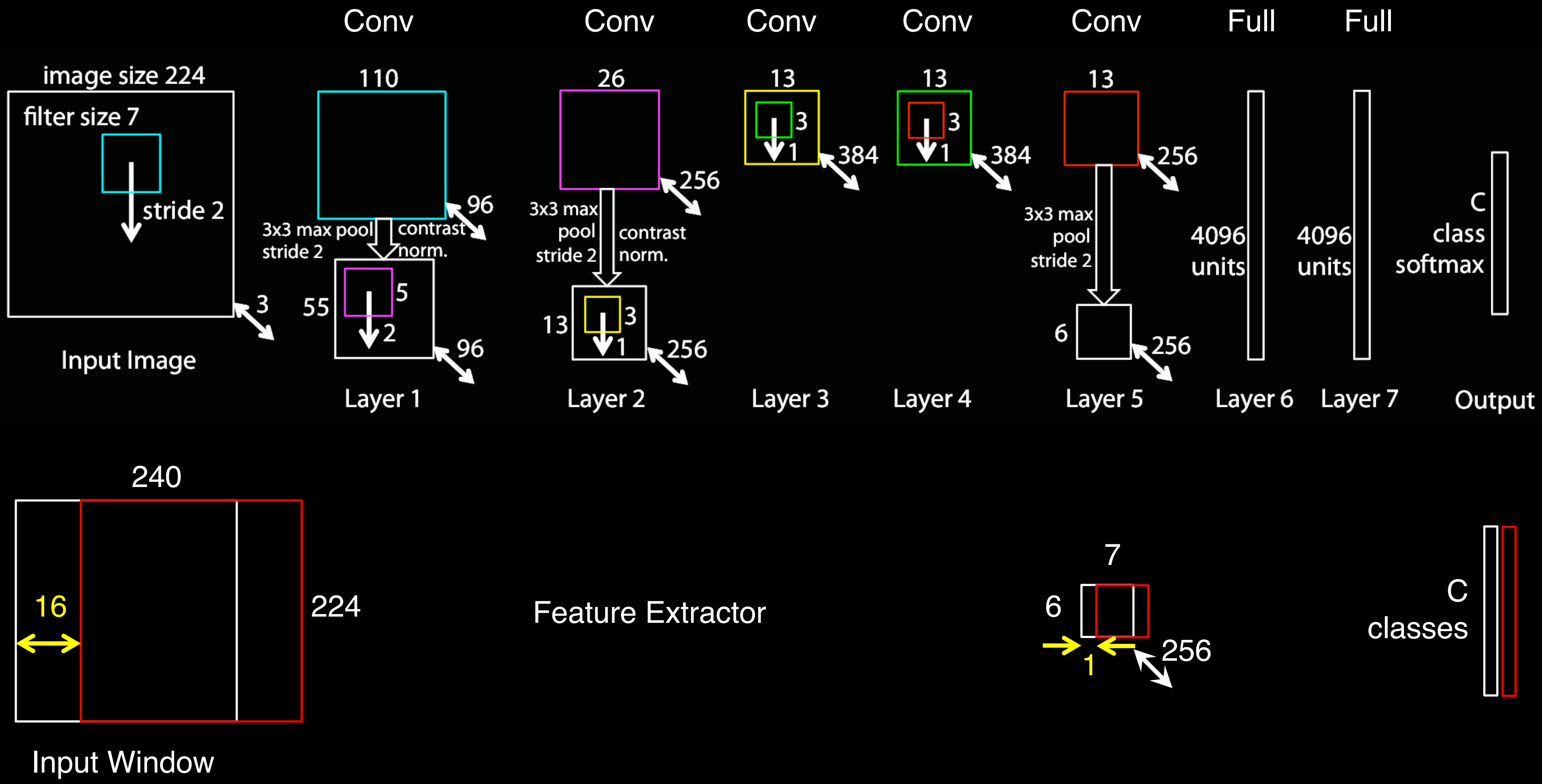
Sliding Window with ConvNet



Sliding Window with ConvNet

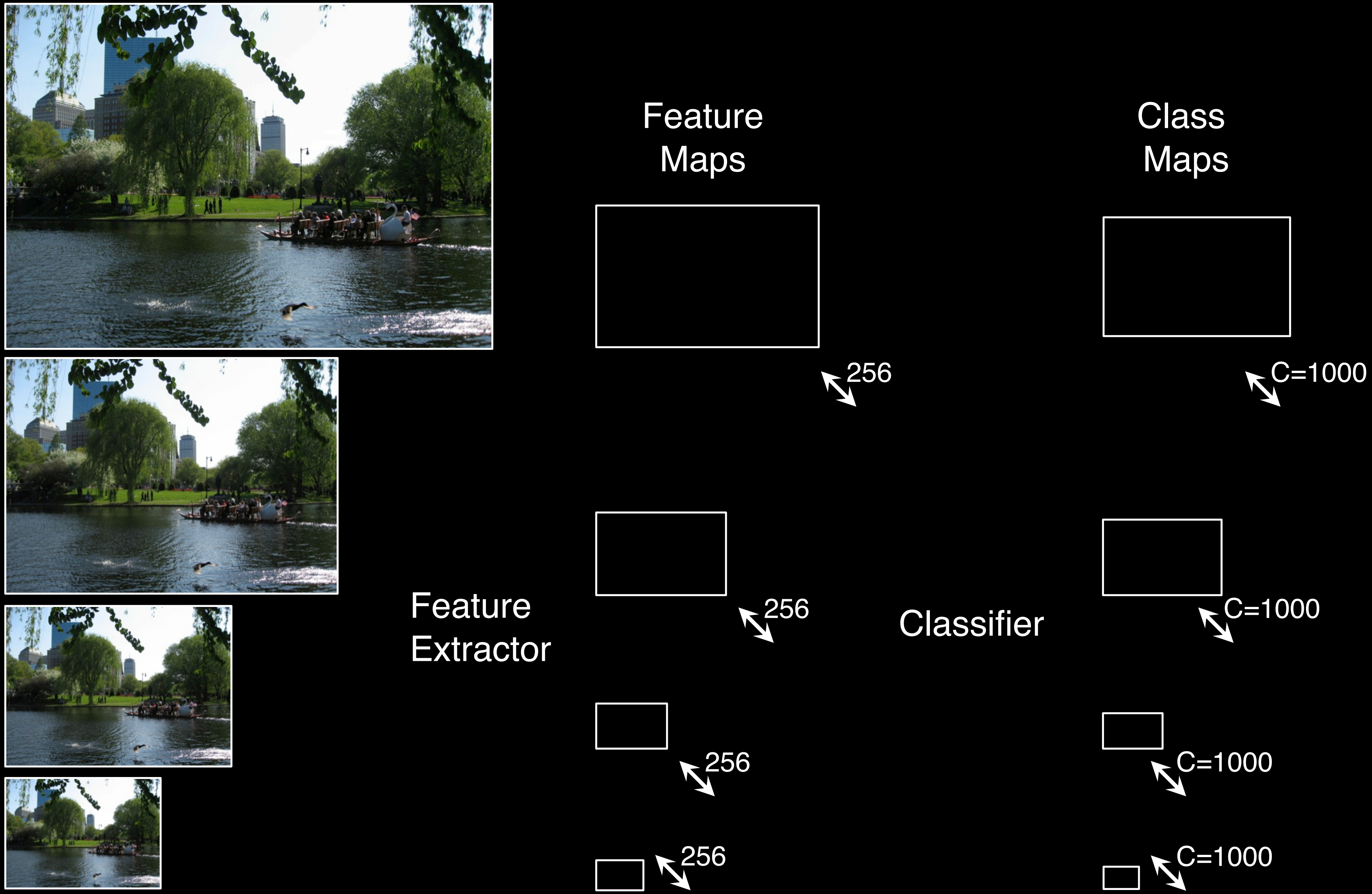


Sliding Window with ConvNet

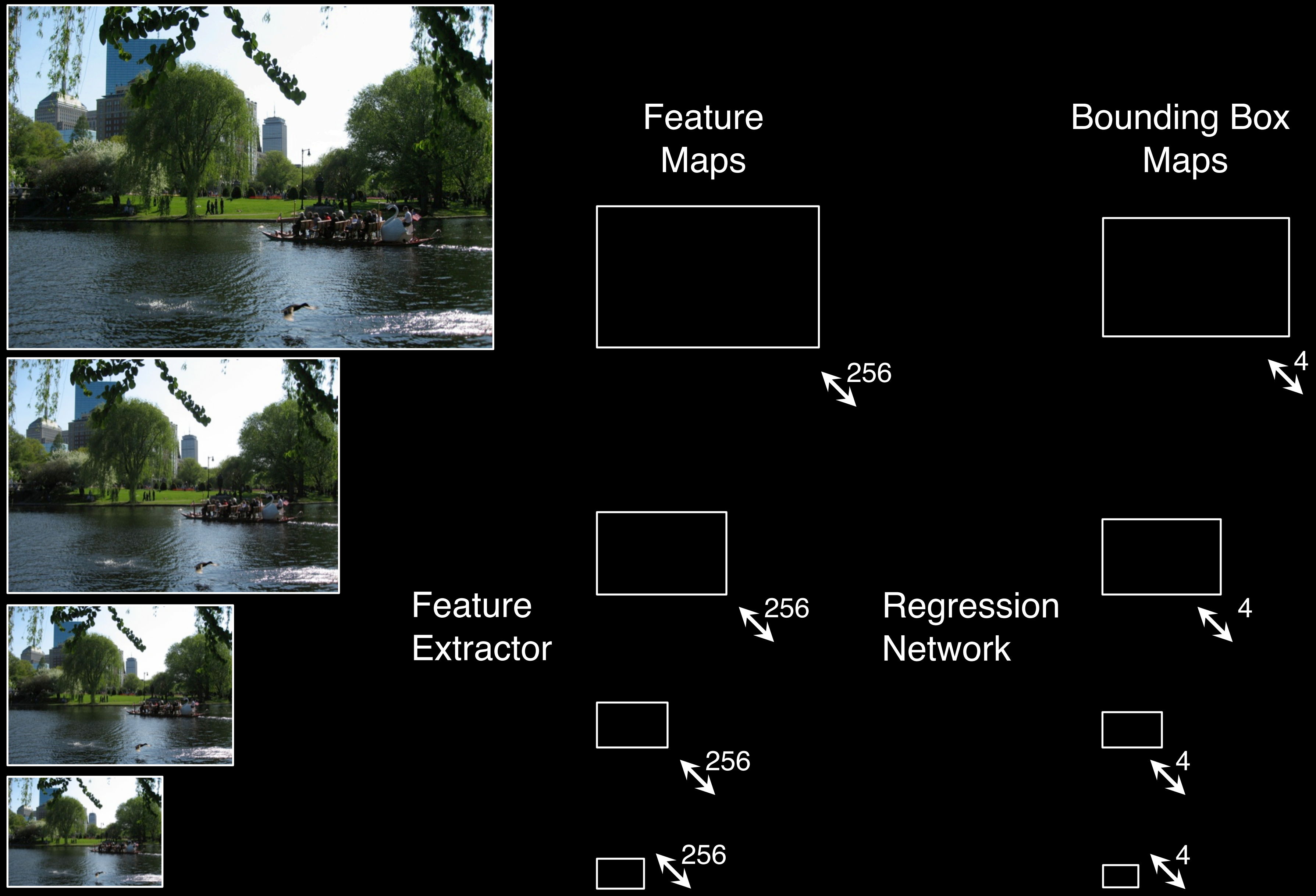


No need to compute two separate windows --- Just one big input window

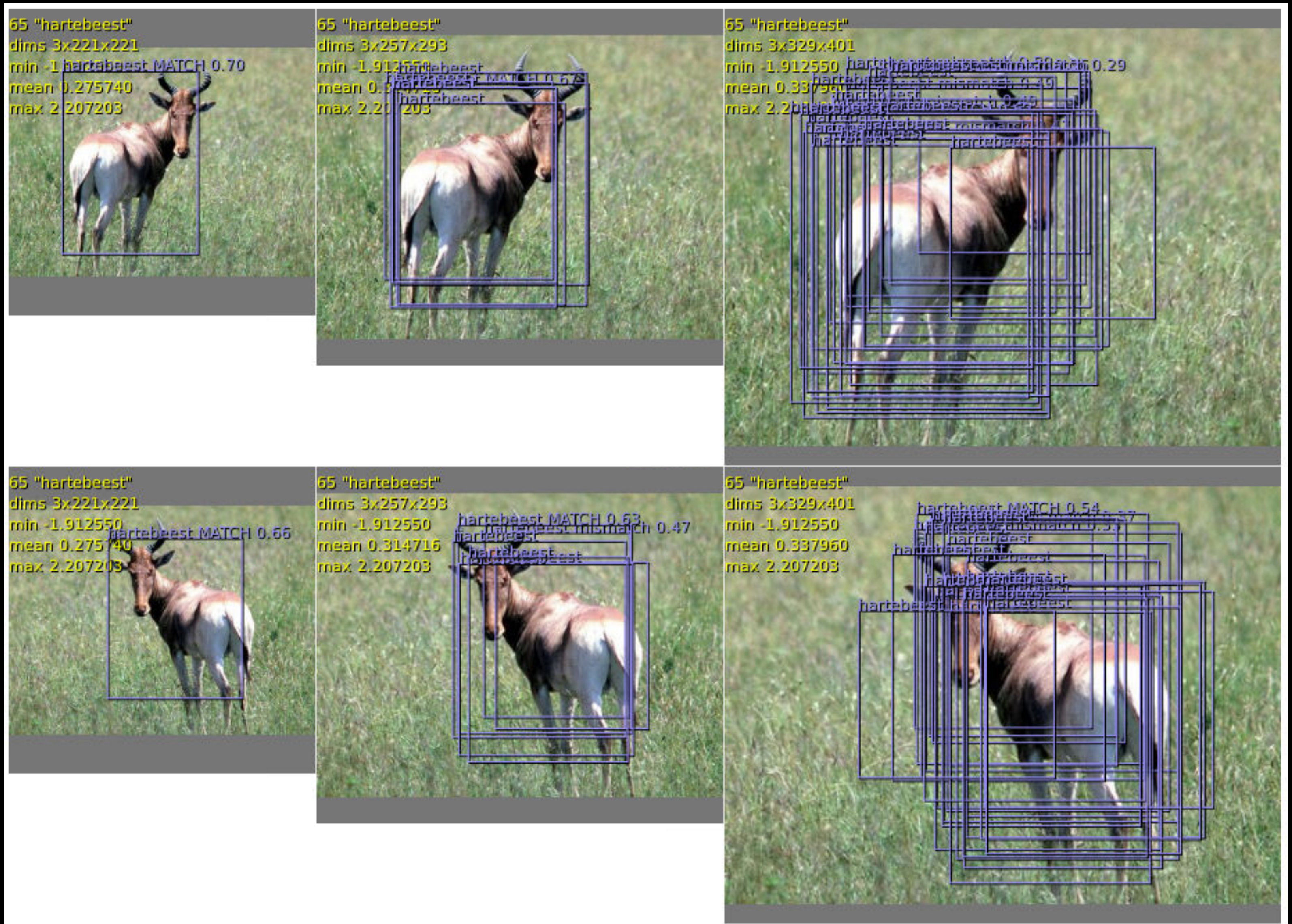
Multi-Scale Sliding Window ConvNet



Multi-Scale Sliding Window ConvNet

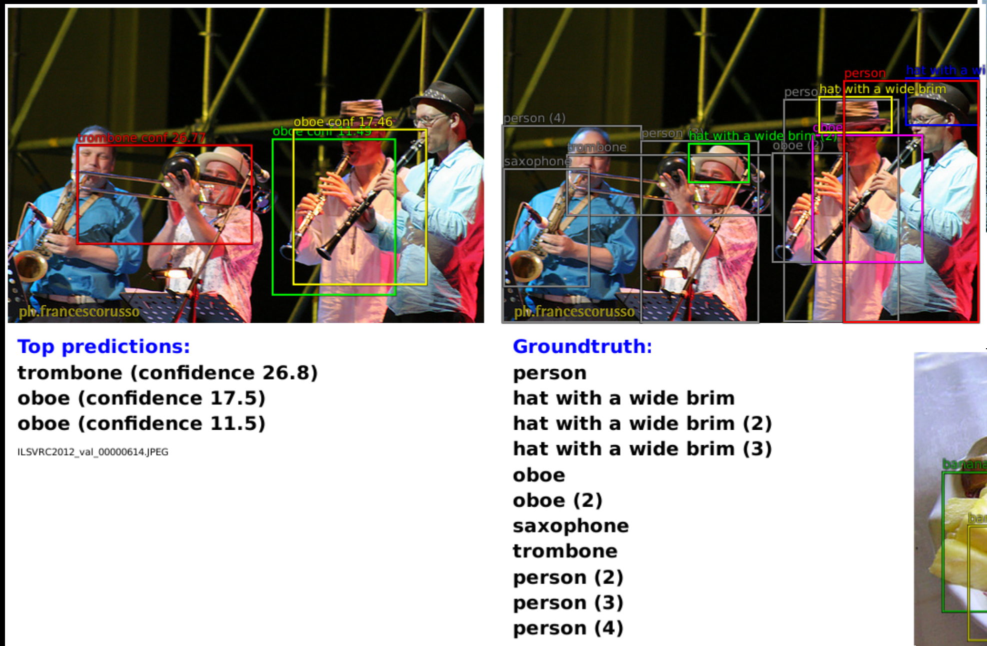


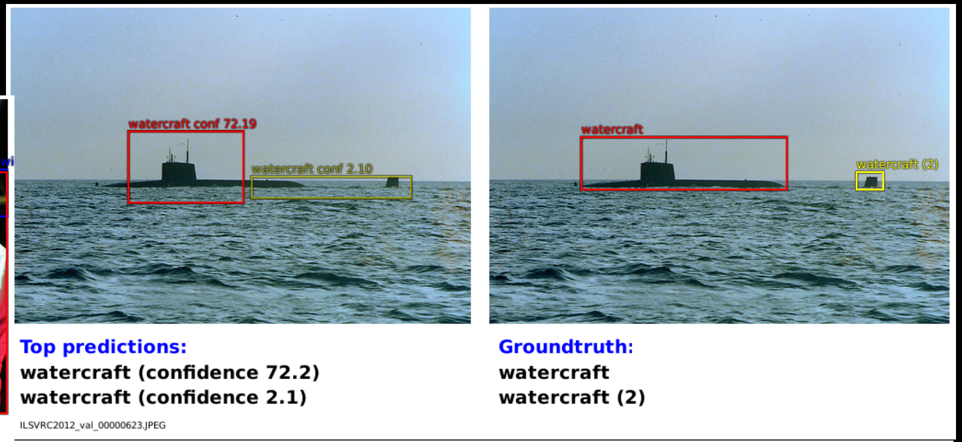
OverFeat – Output before NMS

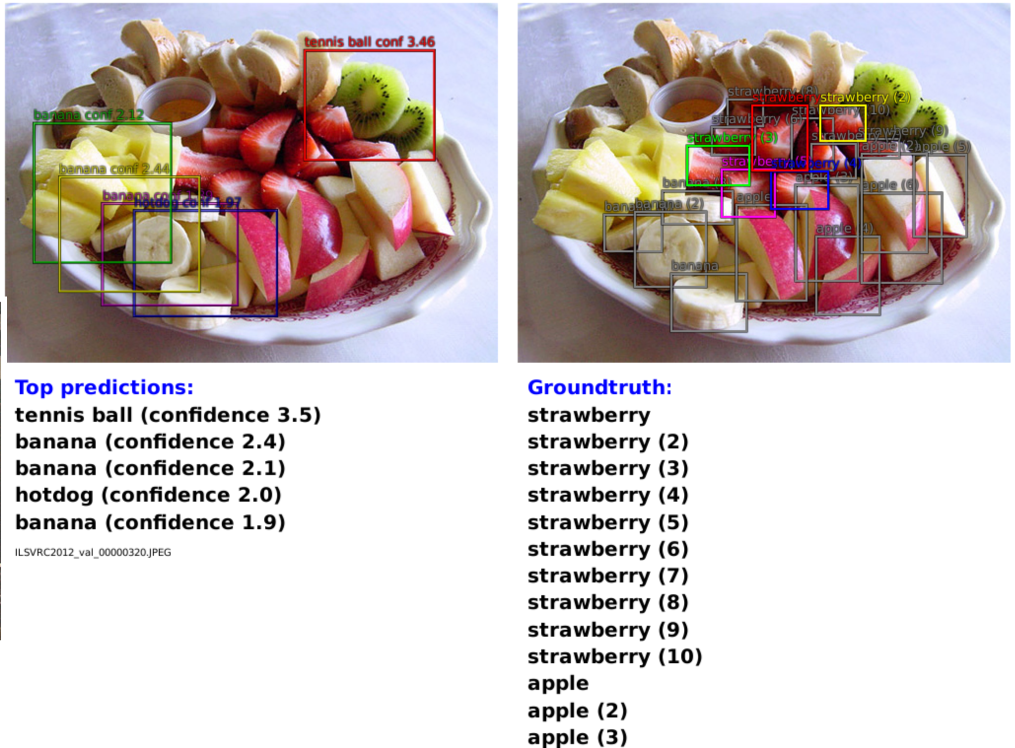


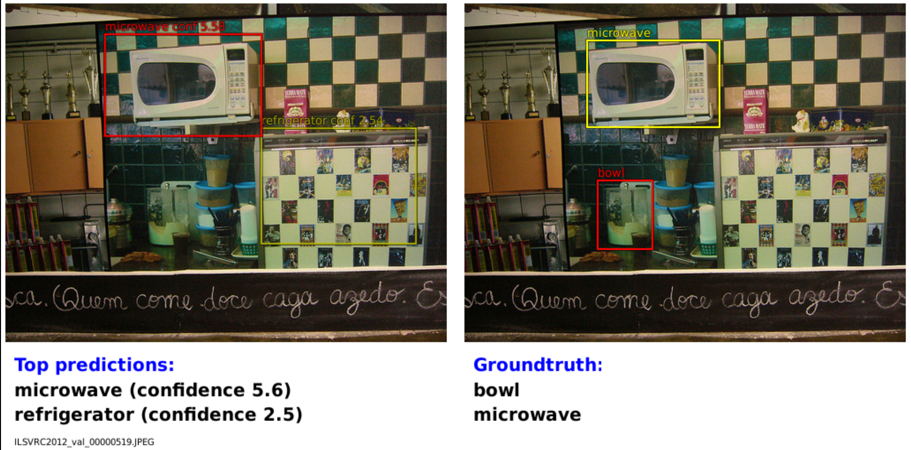
Overfeat Detection Results

[Sermanet et al. ICLR 2014]









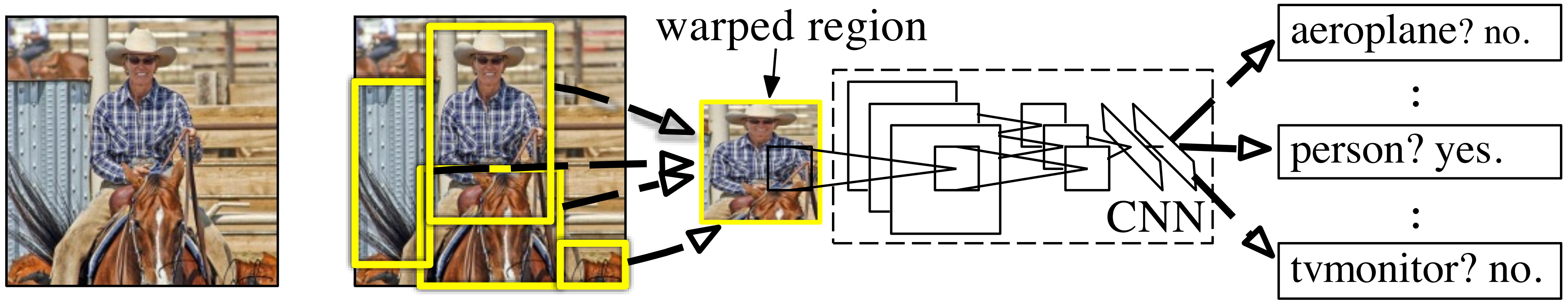
Two General Approaches

1. Examine very position / scale
   – E.g. Overfeat: Integrated recognition, localization and detection using convolutional networks, Sermanet et al., ICLR 2014

2. Use some kind of proposal mechanism to attend to a set of possible regions
   – E.g. Region-CNN [Rich feature hierarchies for accurate object detection and semantic segmentation, Girshick et al., CVPR 2014]

Object Detection: R-CNN

figure credit: R. Girshick et al.



input image

region proposals
~2,000

1 CNN for each region

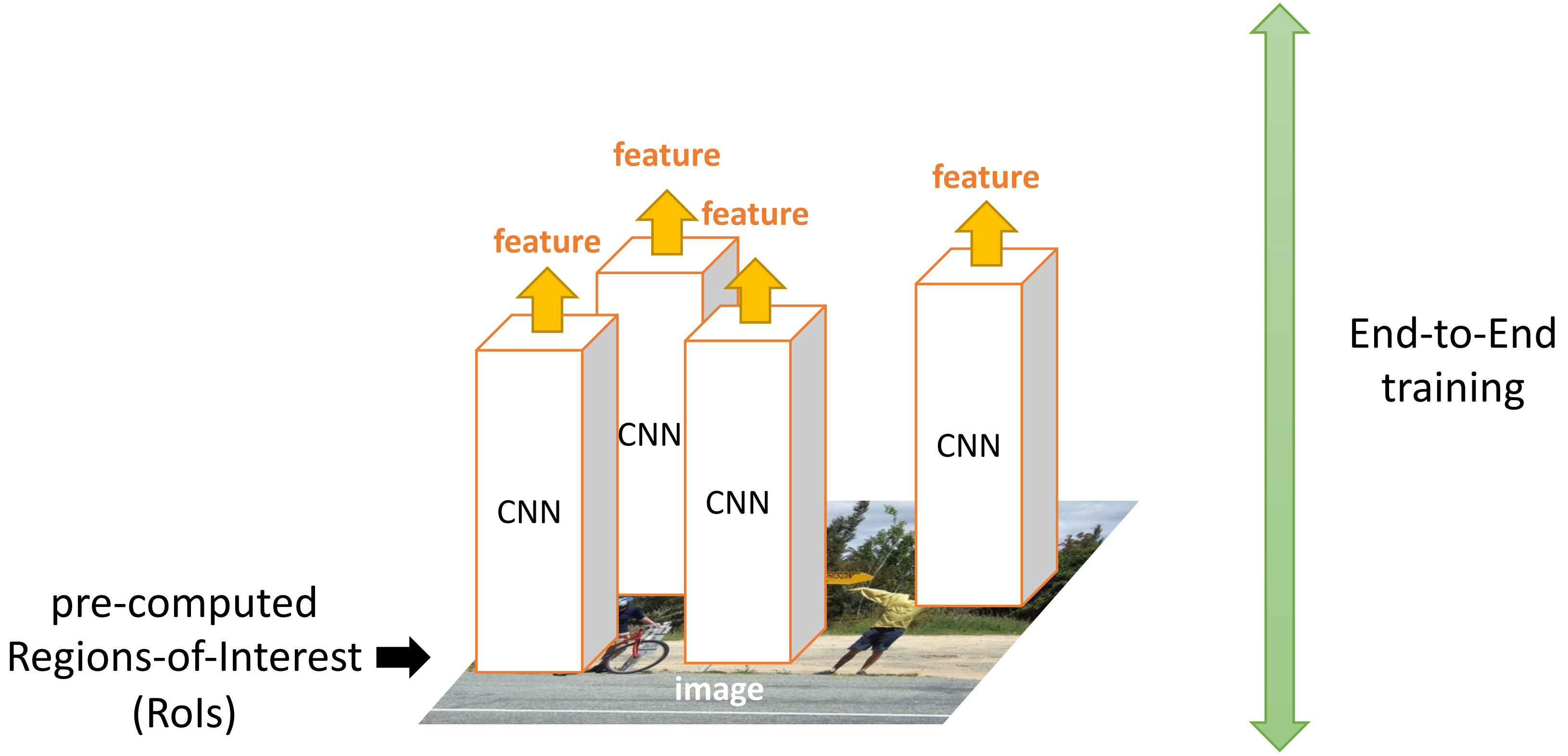
classify regions

Region-based **CNN** pipeline

Girshick, Donahue, Darrell, Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR 2014

Object Detection: R-CNN

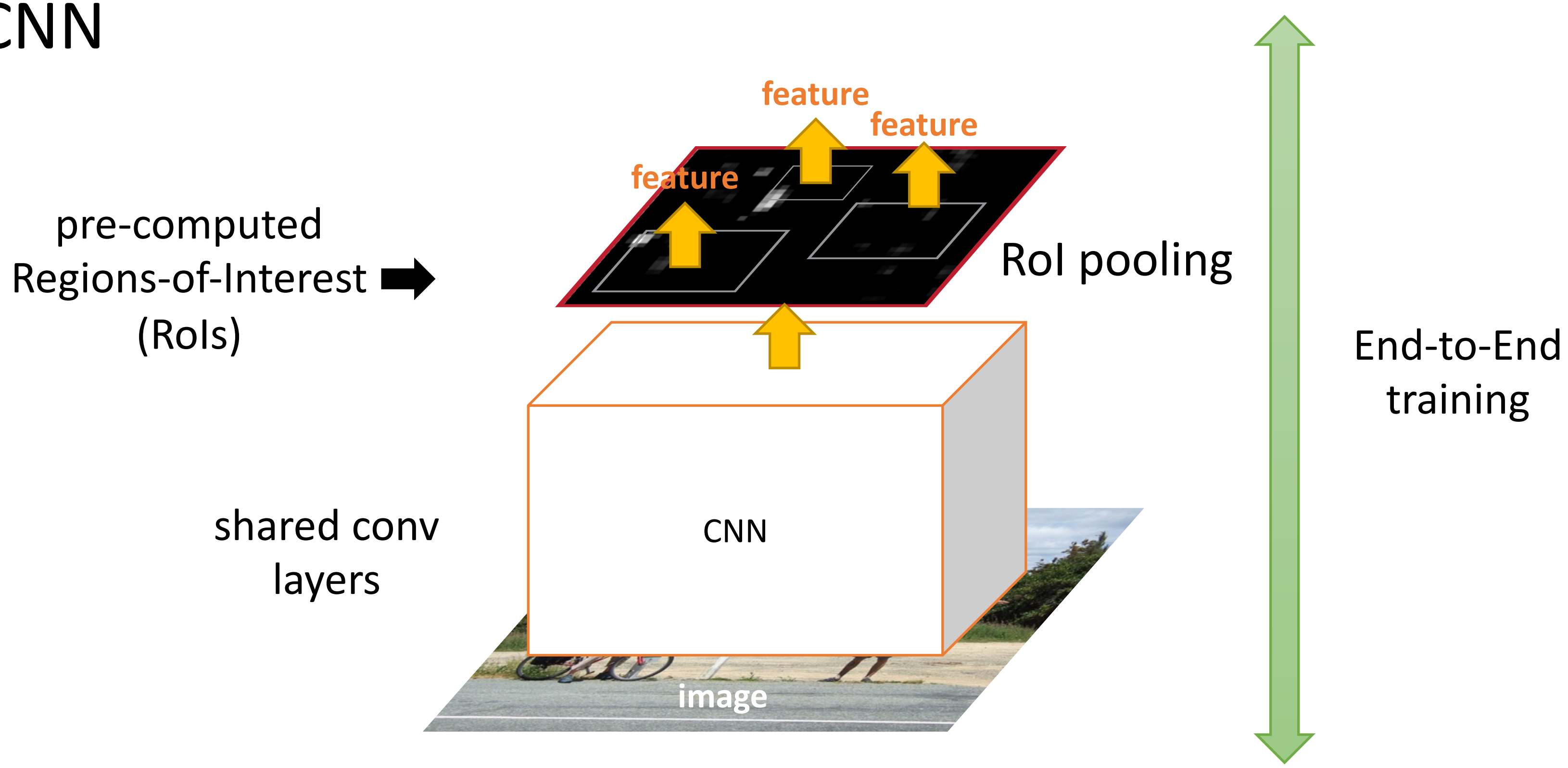
- R-CNN



Girshick, Donahue, Darrell, Malik. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. CVPR 2014

Object Detection: Fast R-CNN

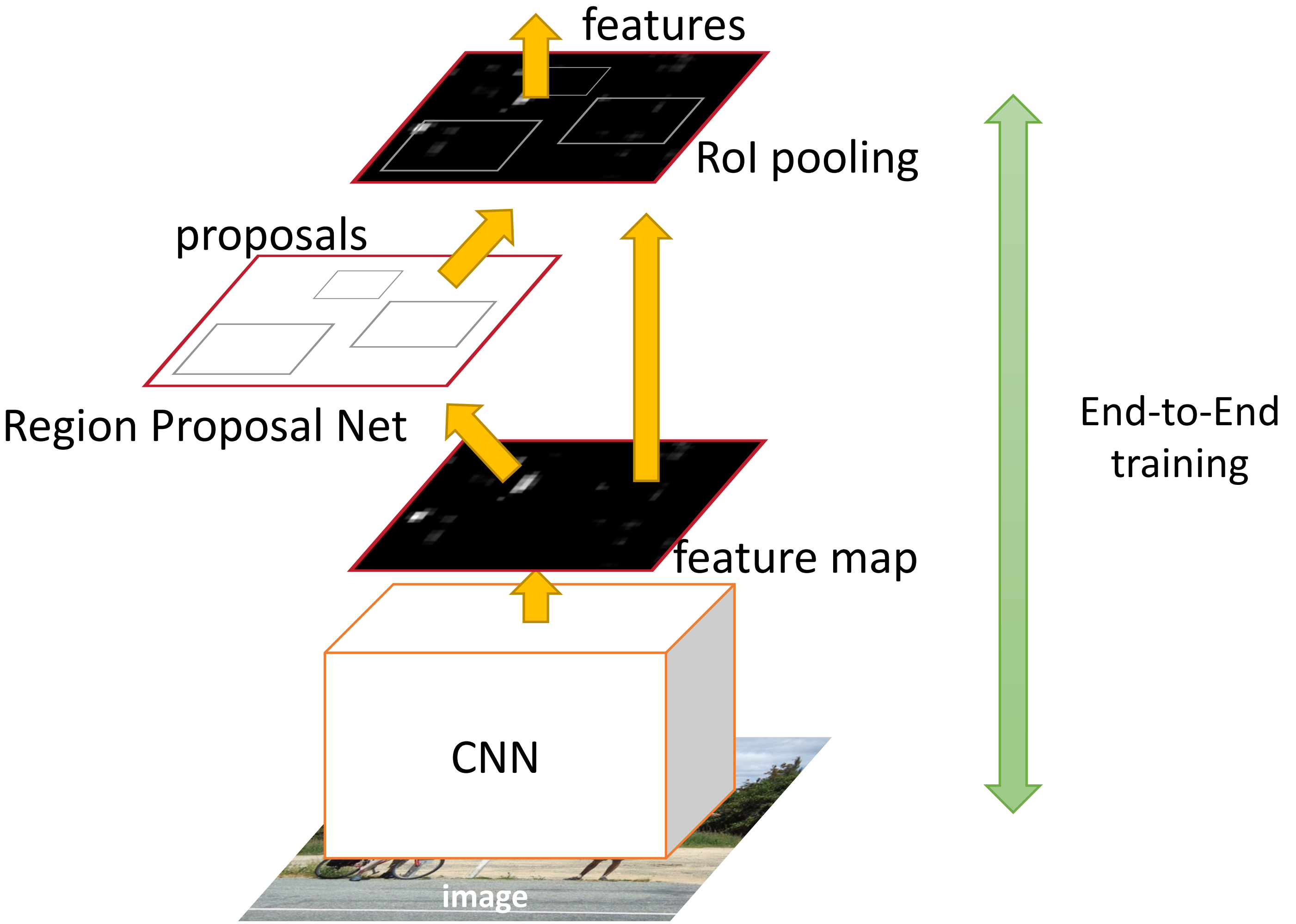
- Fast R-CNN



Girshick. Fast R-CNN. ICCV 2015

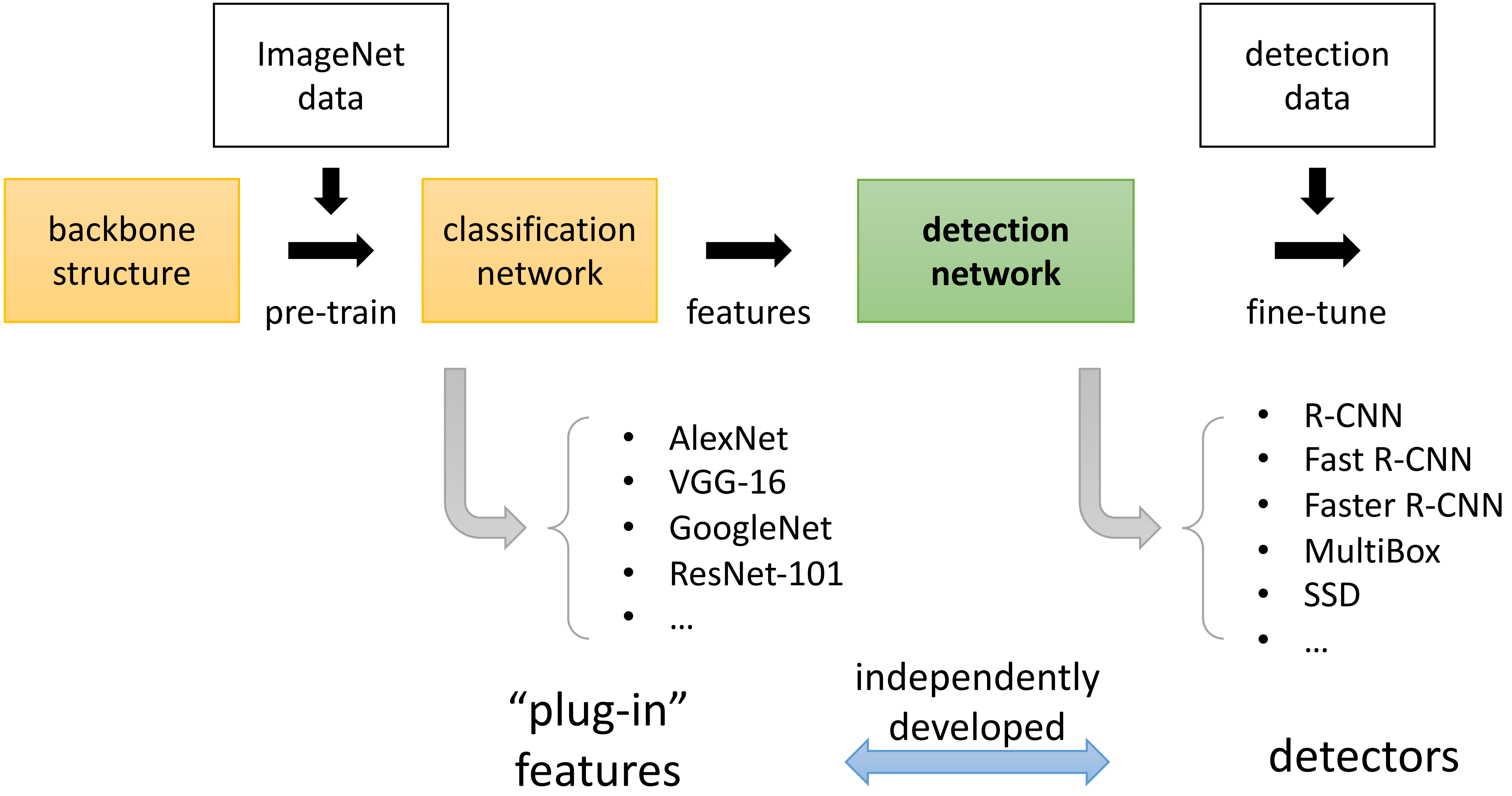
Object Detection: Faster R-CNN

- Faster R-CNN
  - Solely based on CNN
  - No external modules
  - Each step is end-to-end



Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.

Object Detection



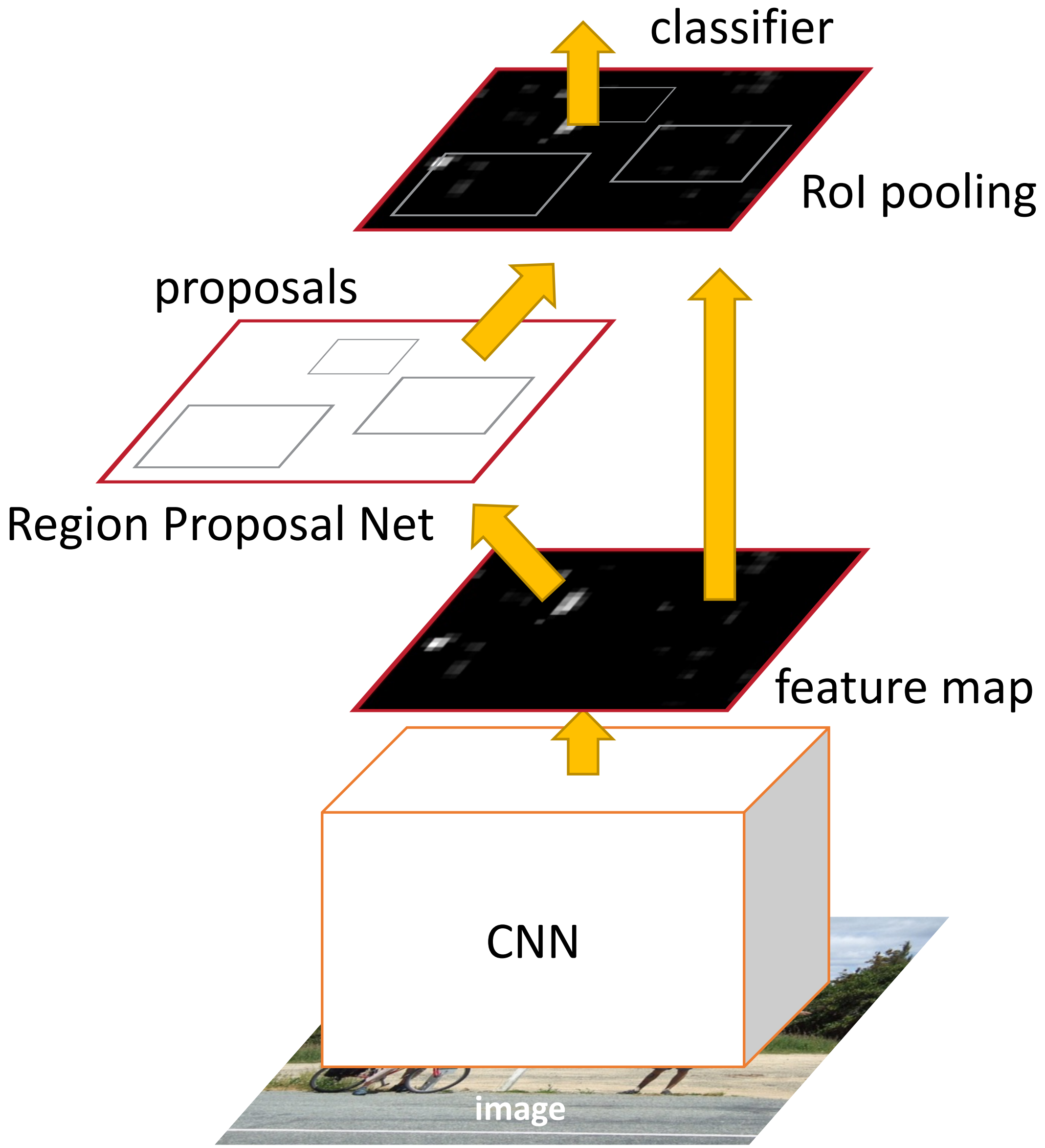
Object Detection

- Simply "Faster R-CNN + ResNet"

Faster R-CNN baseline	mAP@.5	mAP@.5:.95
VGG-16	41.5	21.5
ResNet-101	**48.4**	**27.2**

COCO detection results

ResNet-101 has 28% relative gain vs VGG-16



Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. "Deep Residual Learning for Image Recognition". CVPR 2016.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". NIPS 2015.

Object Detection

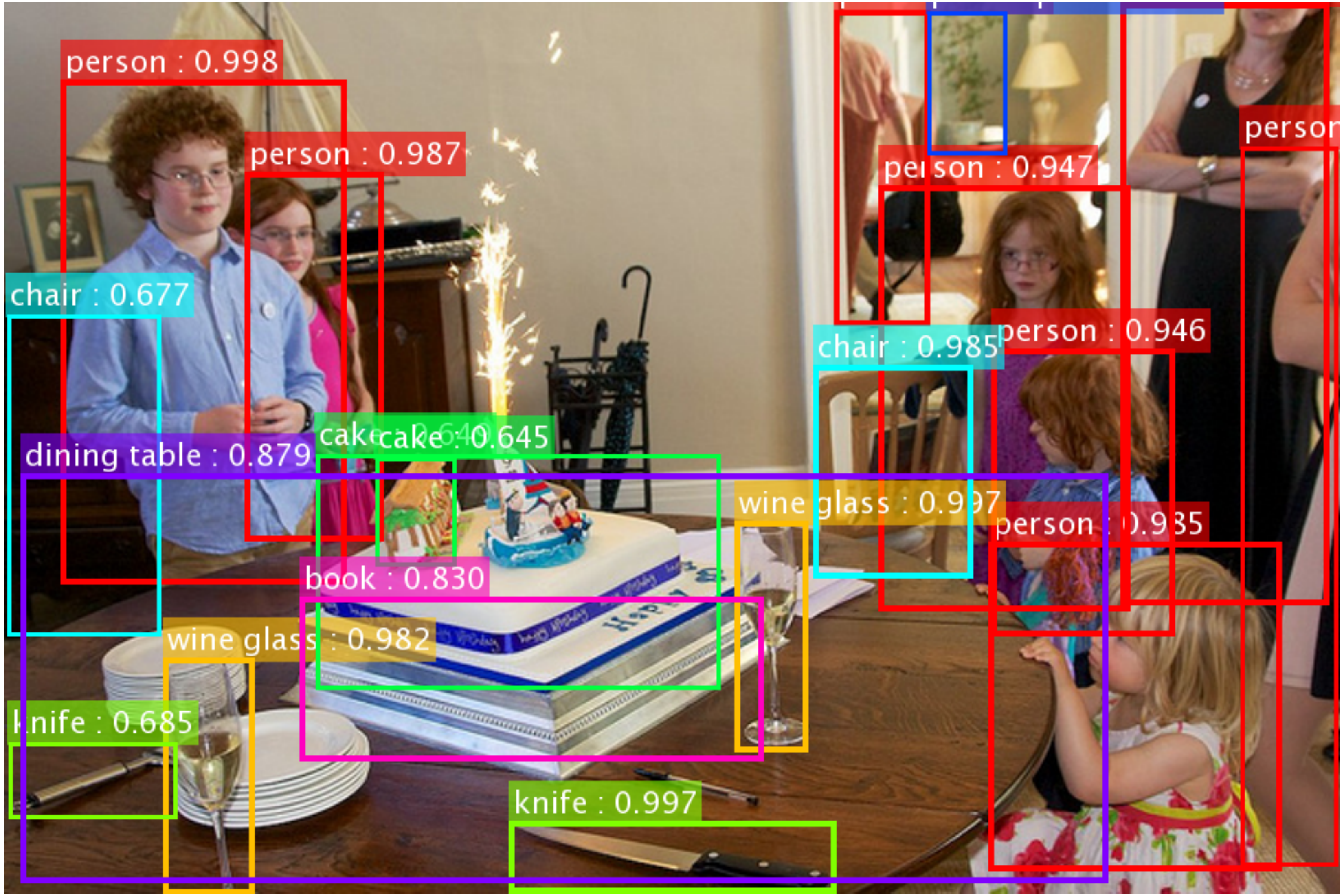
- RPN learns proposals by extremely deep nets
  - We use only 300 proposals (no hand-designed proposals)

- Add components:
  - Iterative localization
  - Context modeling
  - Multi-scale testing

- All components are based on CNN features; all steps are end-to-end

- All benefit more from deeper features – cumulative gains!

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. CVPR 2016.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.

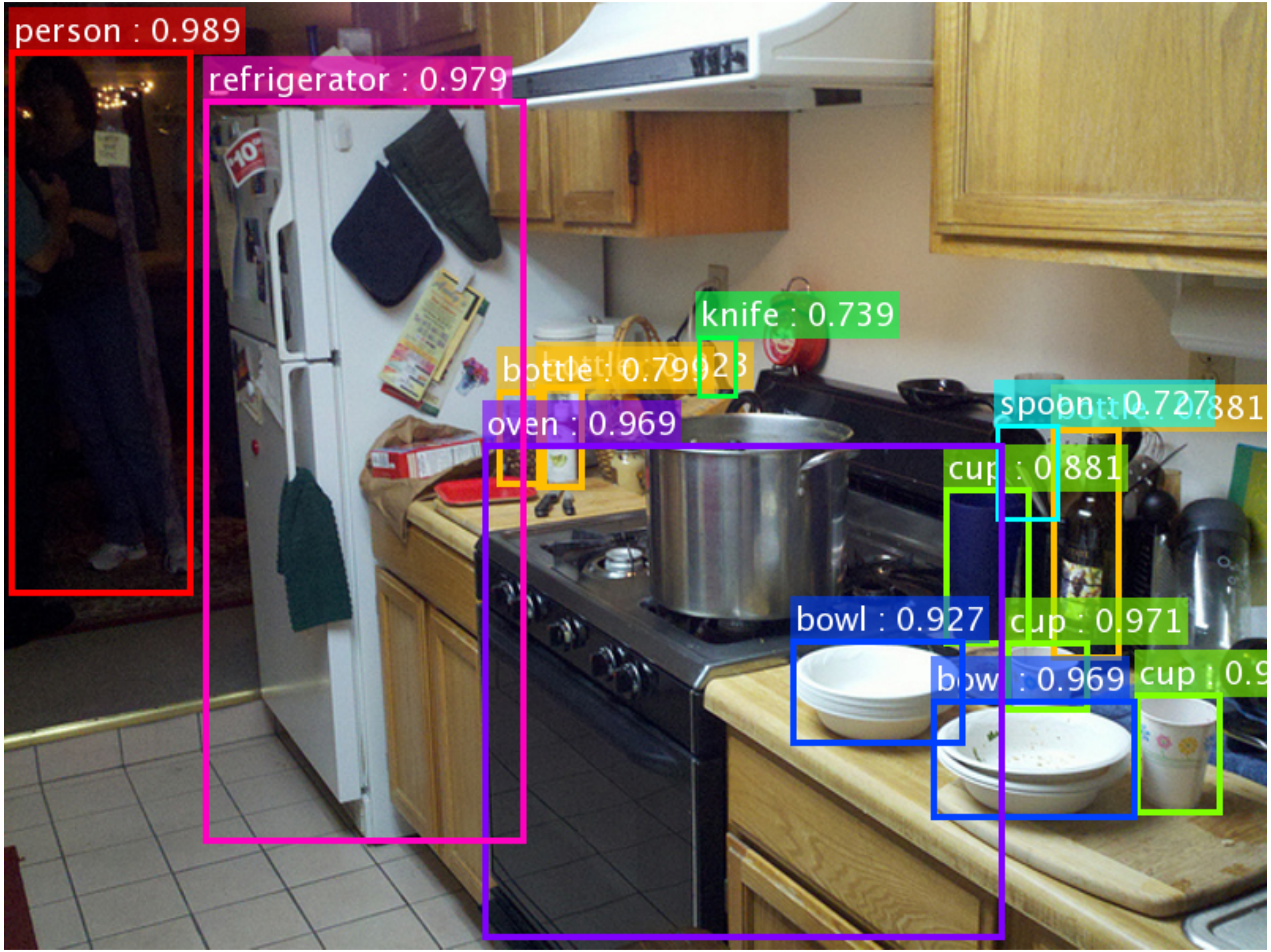


ResNet’s object detection result on COCO

*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

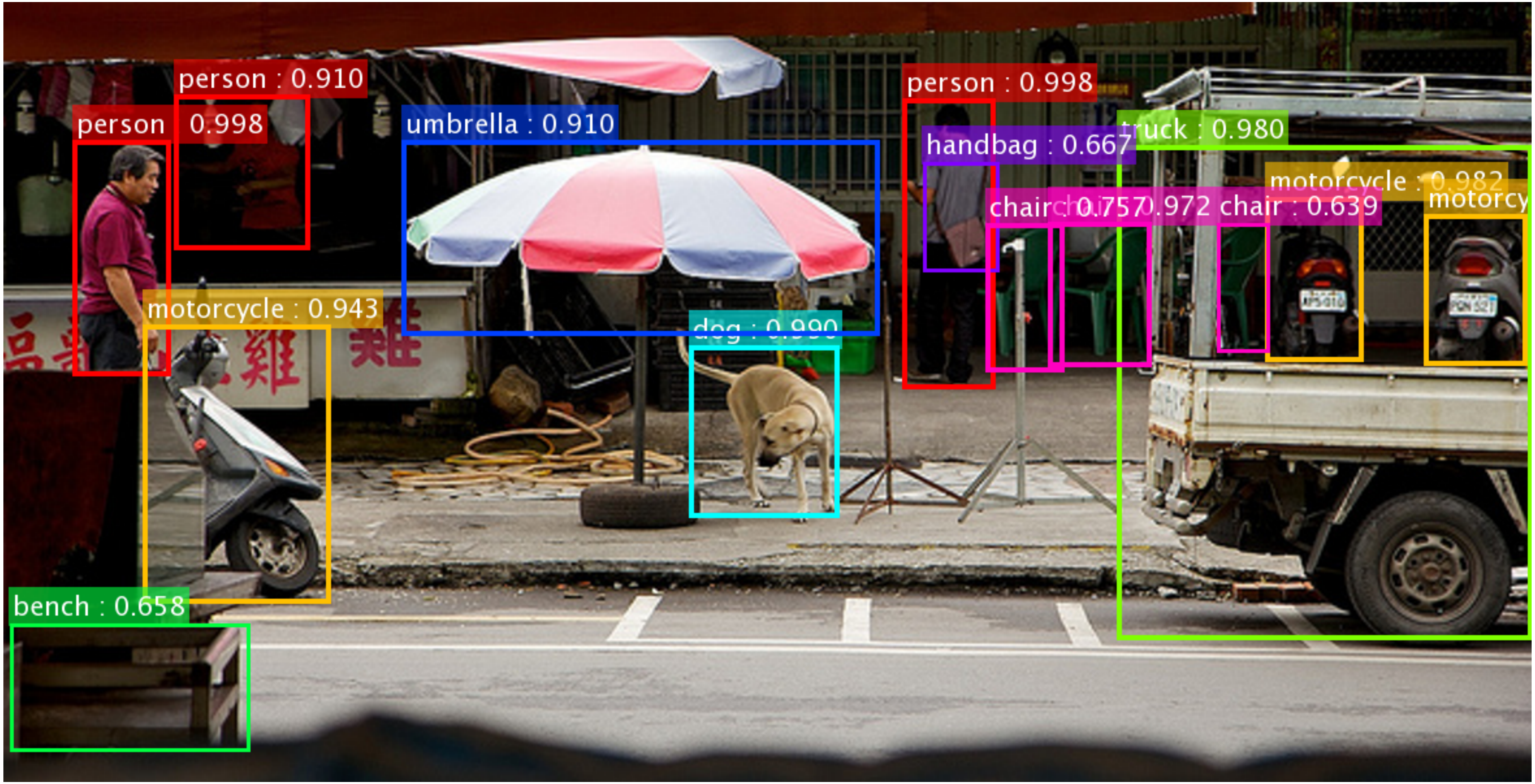
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.



*the original image is from the COCO dataset

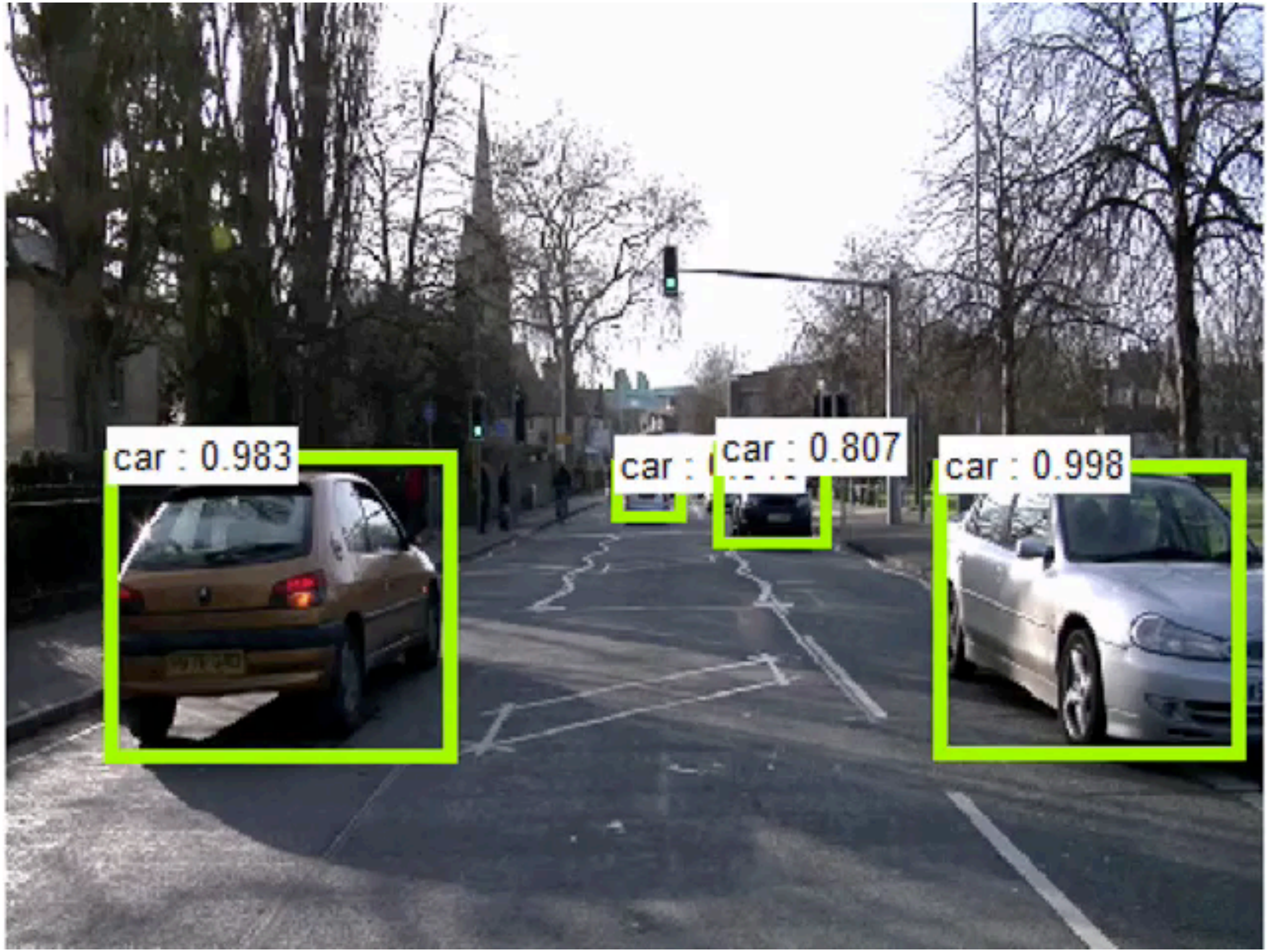
Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.

Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.



*the original image is from the COCO dataset

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.



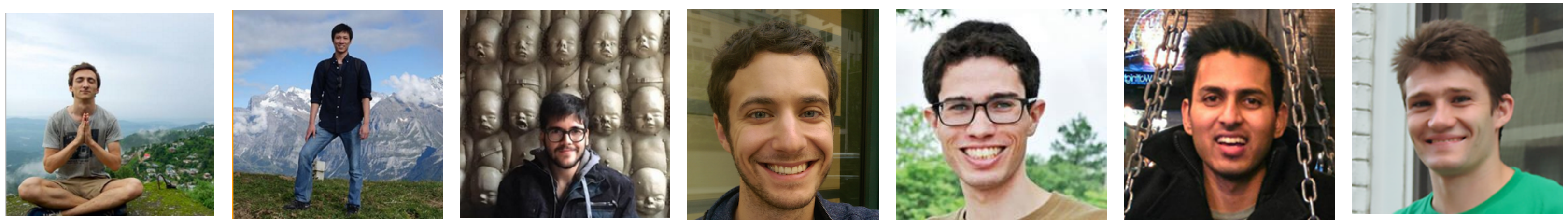
Results on real video. Models trained on MS COCO (80 categories).
(frame-by-frame; no temporal processing)

Kaiming He, Xiangyu Zhang, Shaoqing Ren, & Jian Sun. “Deep Residual Learning for Image Recognition”. arXiv 2015.
Shaoqing Ren, Kaiming He, Ross Girshick, & Jian Sun. “Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”. NIPS 2015.

Facebook AI Research

FAIR COCO Object Detection

Sergey Zagoruyko*, Tsung-Yi Lin*, Pedro Pinheiro*, Adam Lerer, Sam Gross, Soumith Chintala, Piotr Dollár



(*equal contribution)

Results

	AP bbox	AP small	AP medium	AP large	AR max=100	**AP segm**
MSRA	0.373	0.183	0.419	0.524	0.491	0.282
FAIRCNN	0.335	0.139	0.378	0.477	0.485	0.251
ION	0.310	0.123	0.332	0.447	0.457	
FastRCNN	0.197	0.035	0.188	0.346	0.298	

66% improvement over FastRCNN baseline

Overview

I. DeepMask segmentation proposals [Pinheiro NIPS 15]

+ iterative localization

+ top-down refinement

II. Fast R-CNN object detector [Girshick ICCV 15]

+ foveal context regions

+ modified loss function

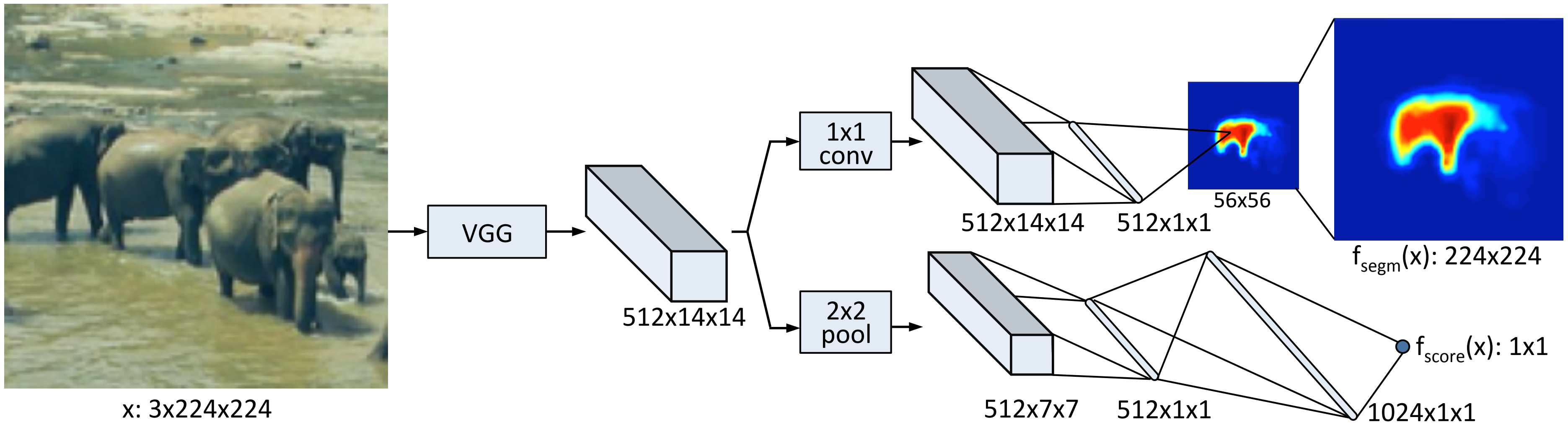
+ skip connections

+ ensembling

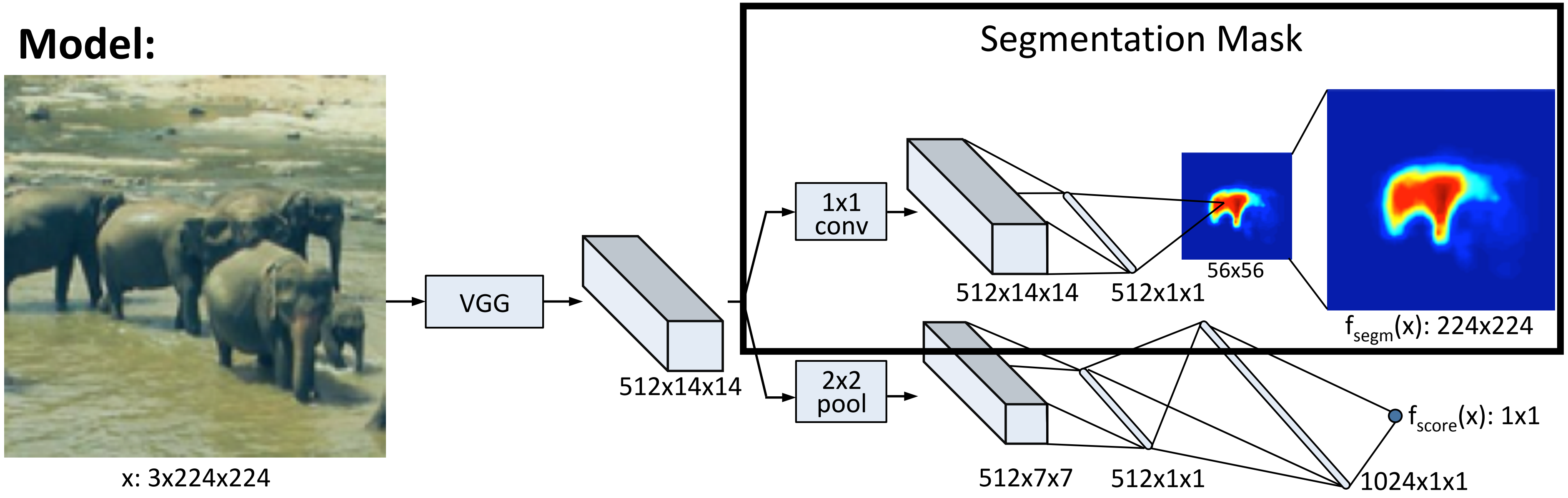
I. Deep MASK Object Proposals

DeepMask Framework

Model:

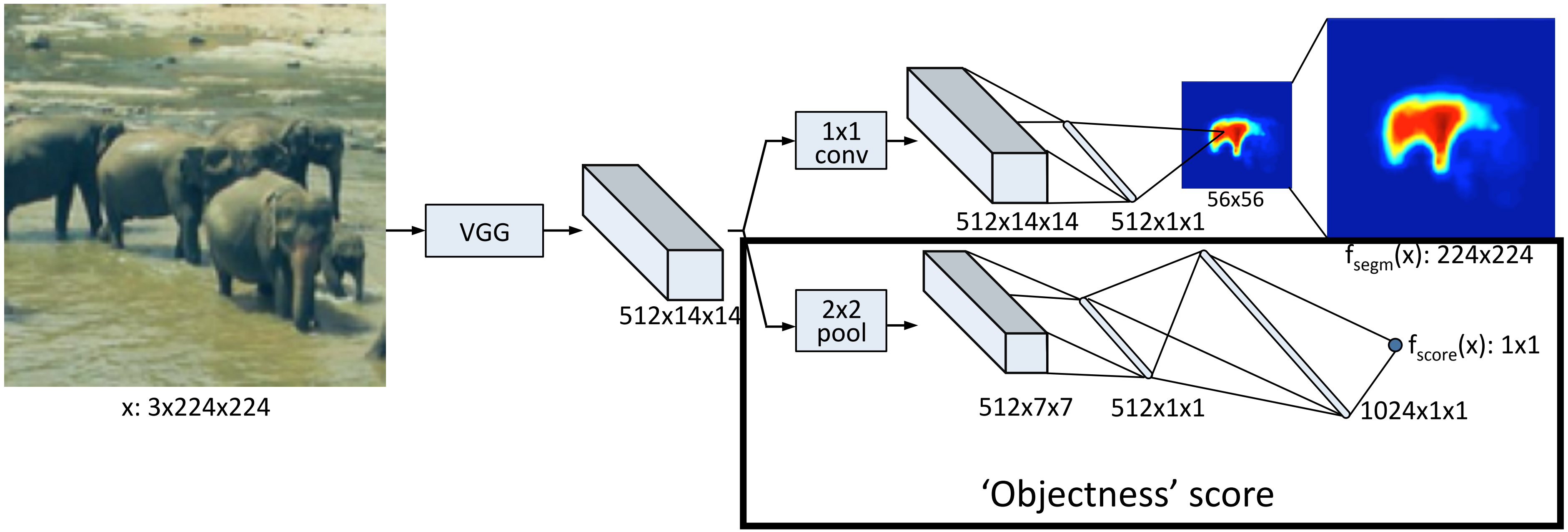


DeepMask Framework



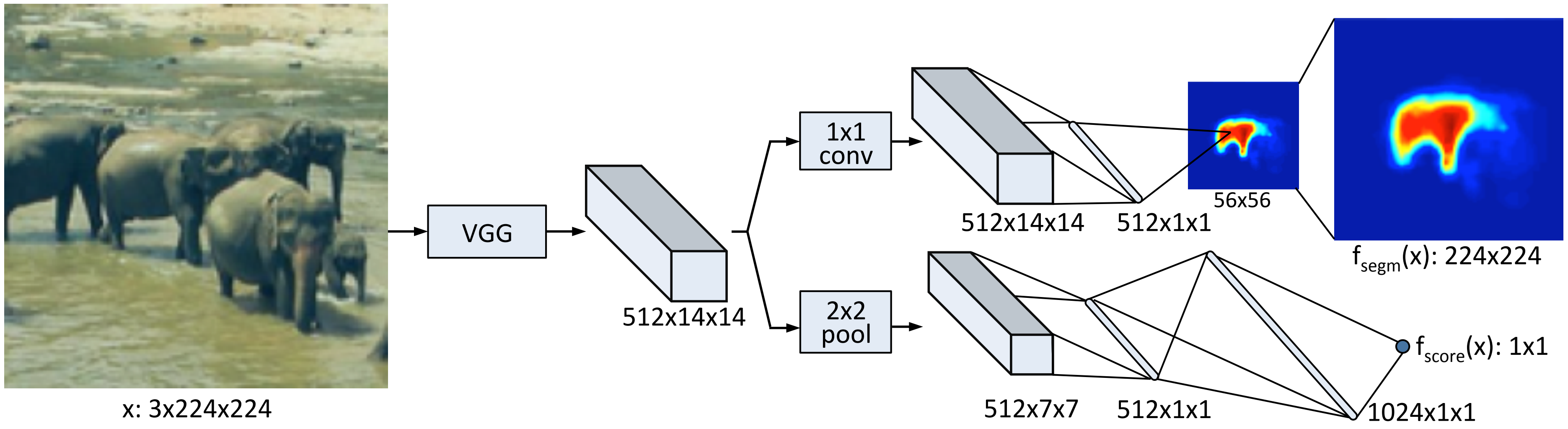
DeepMask Framework

Model:

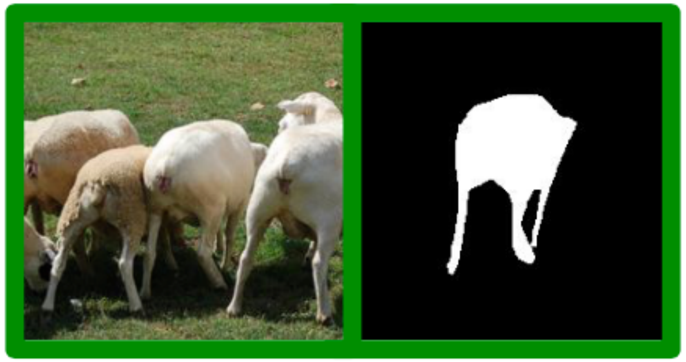


DeepMask Framework

Model:

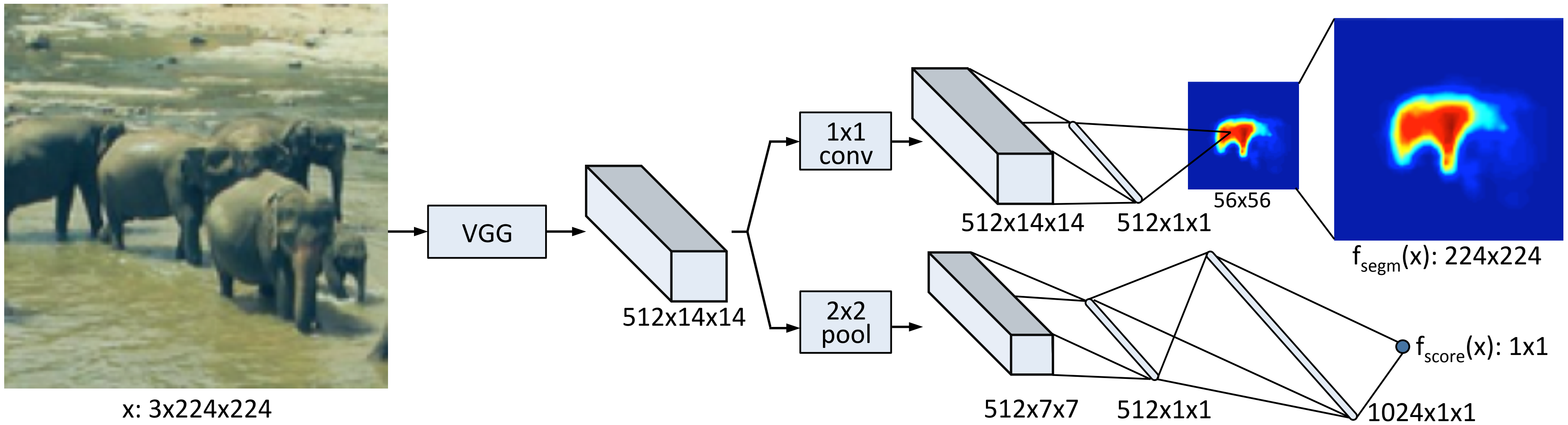


x: 3x224x224



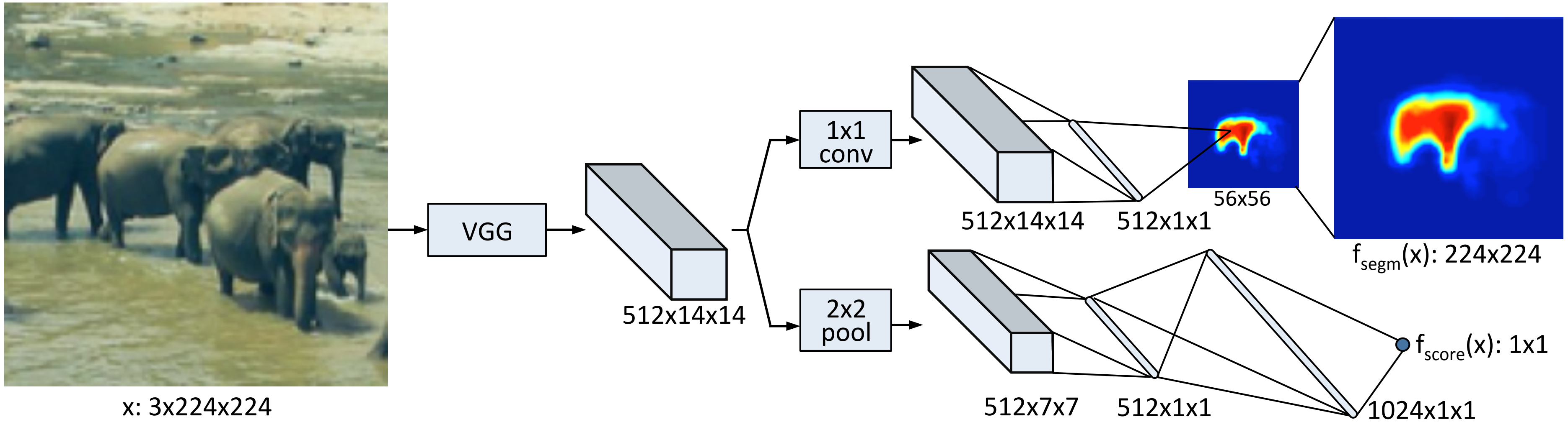
DeepMask Framework

Model:



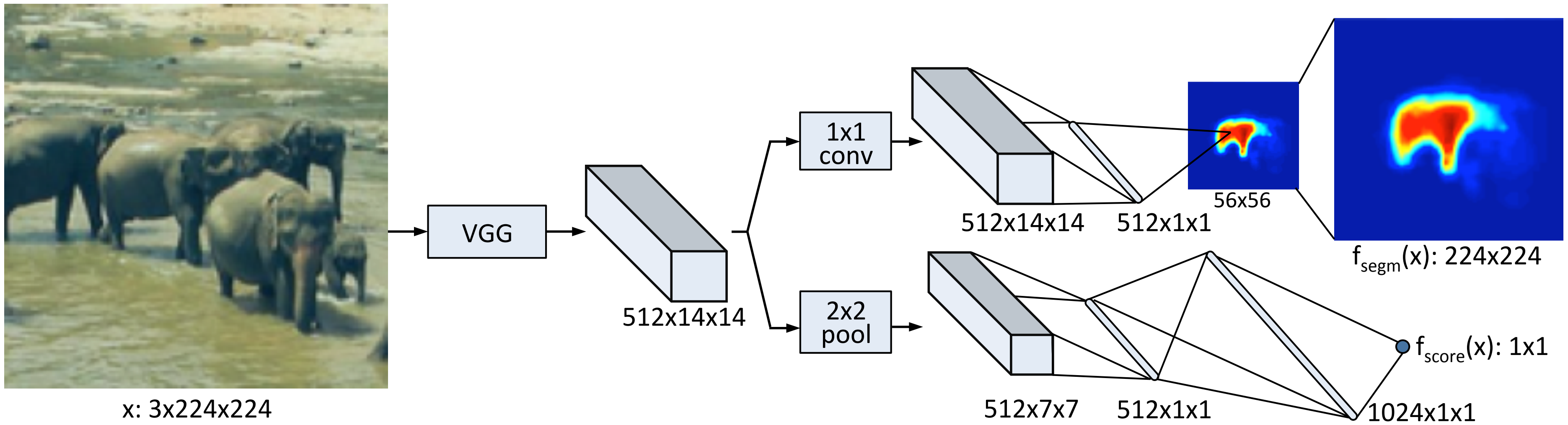
DeepMask Framework

Model:



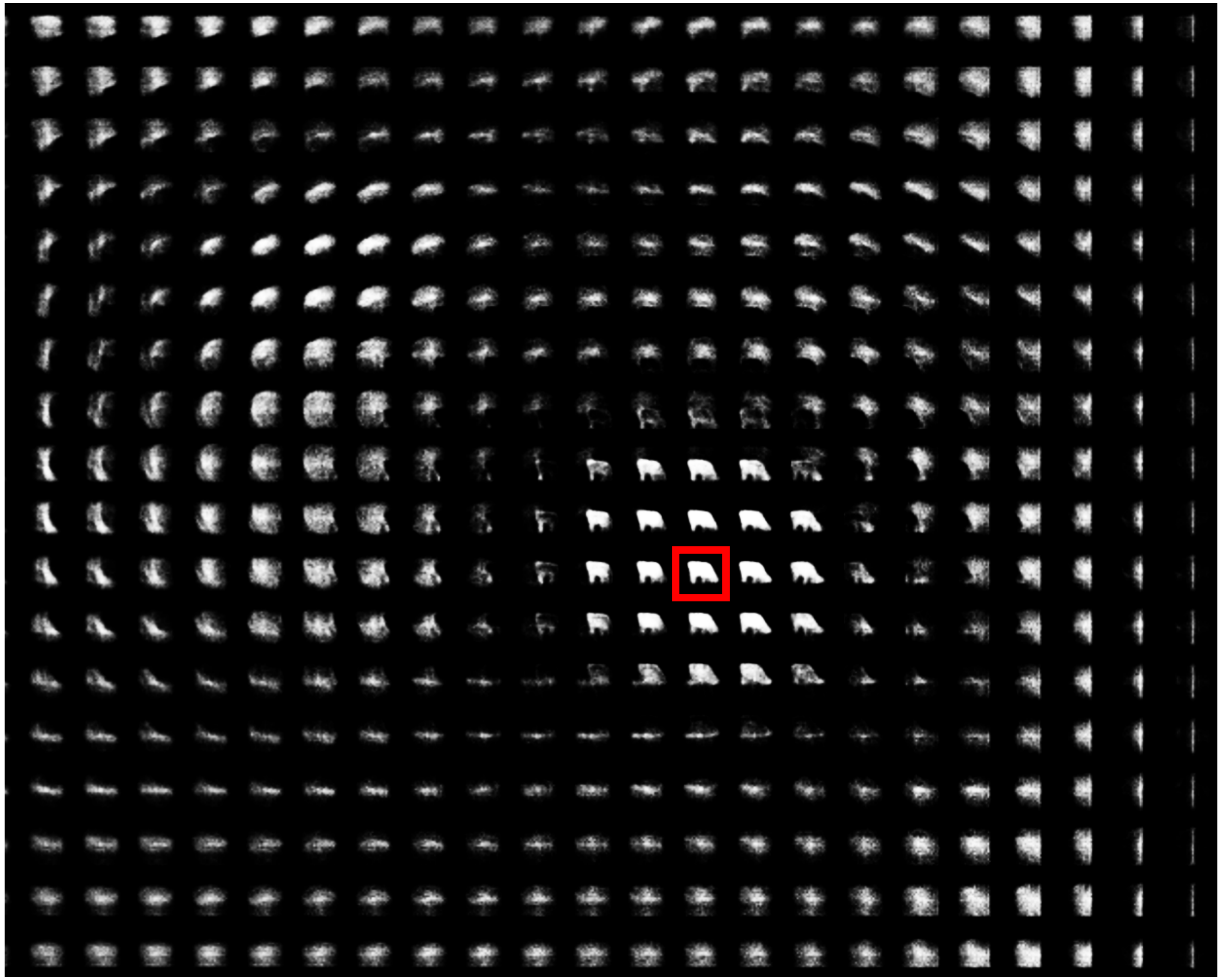
DeepMask Framework

Model:



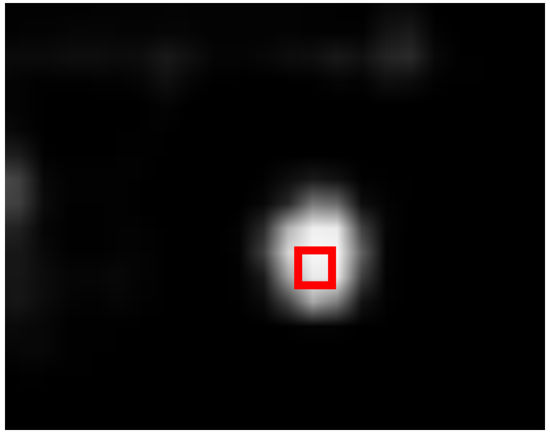
Single Scale Inference

masks



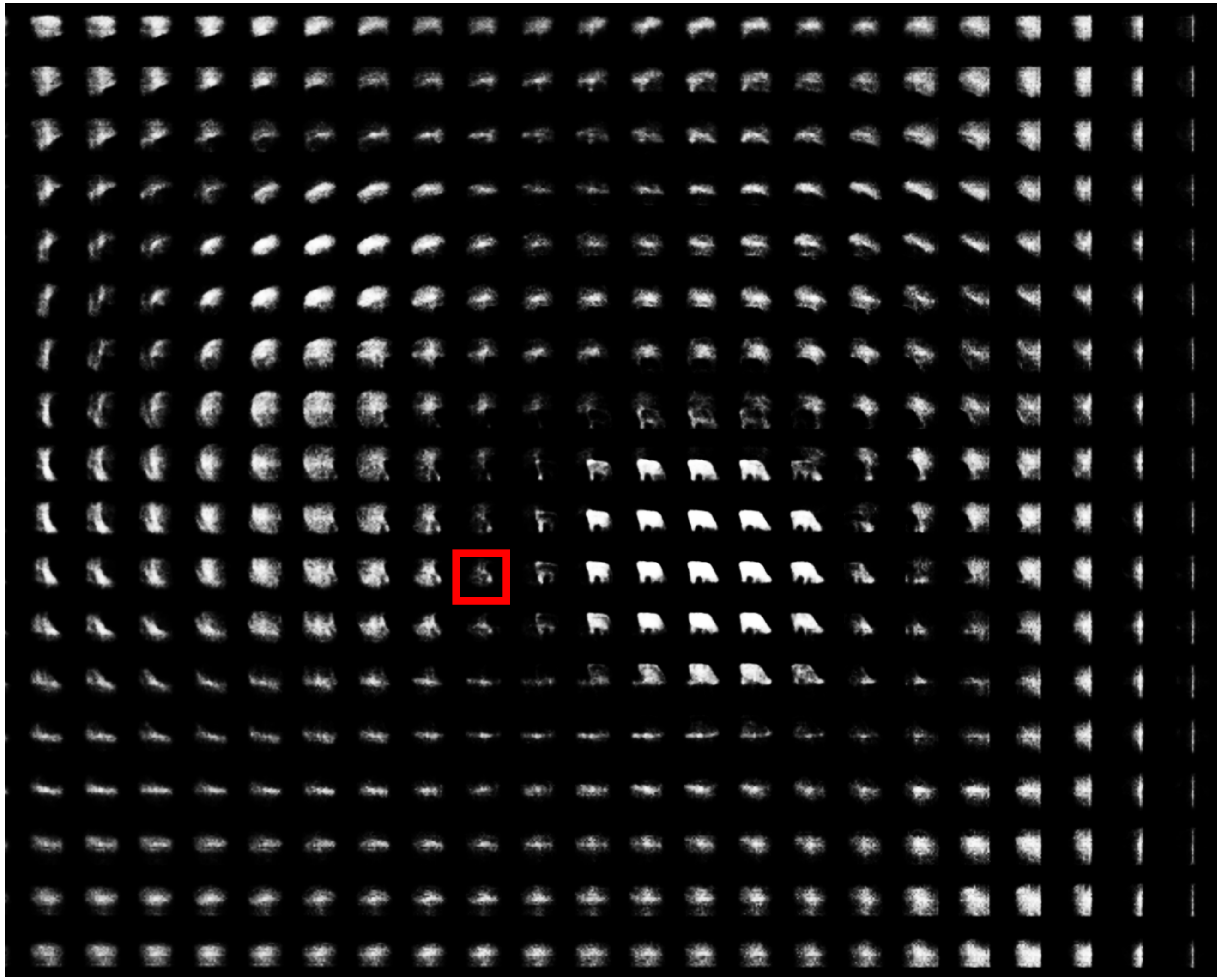
image

scores



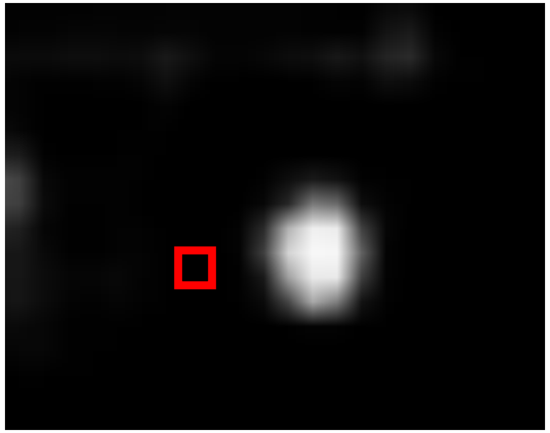
Single Scale Inference

masks

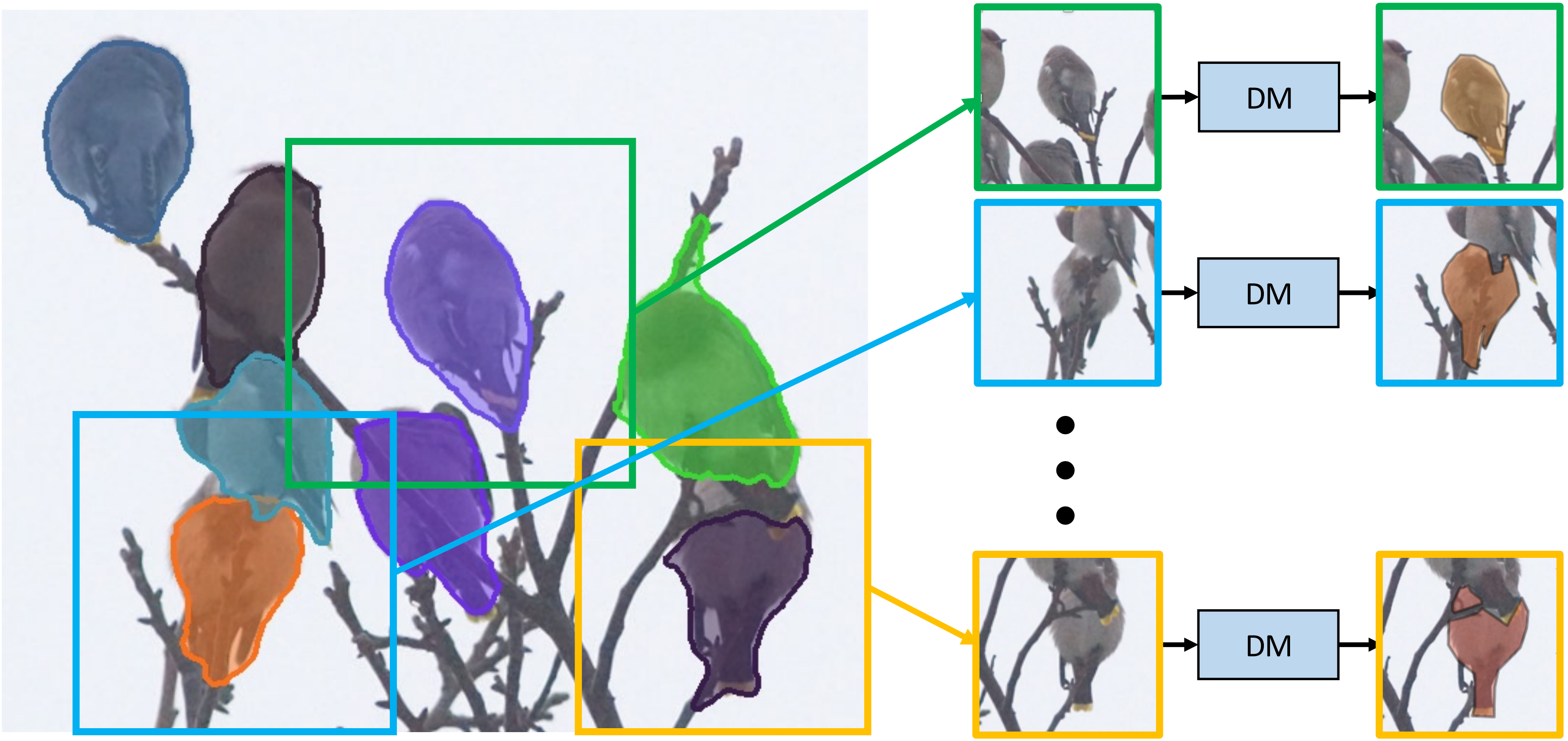


image

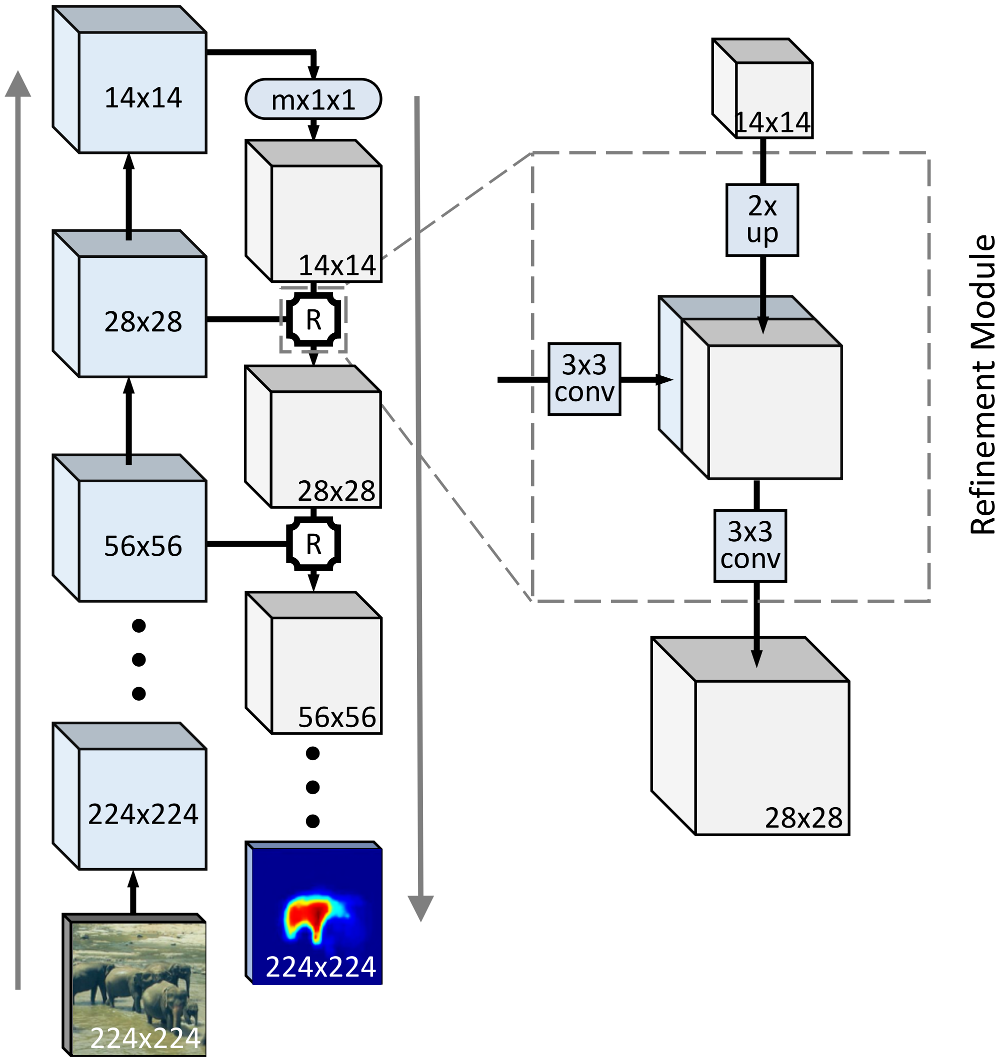
scores



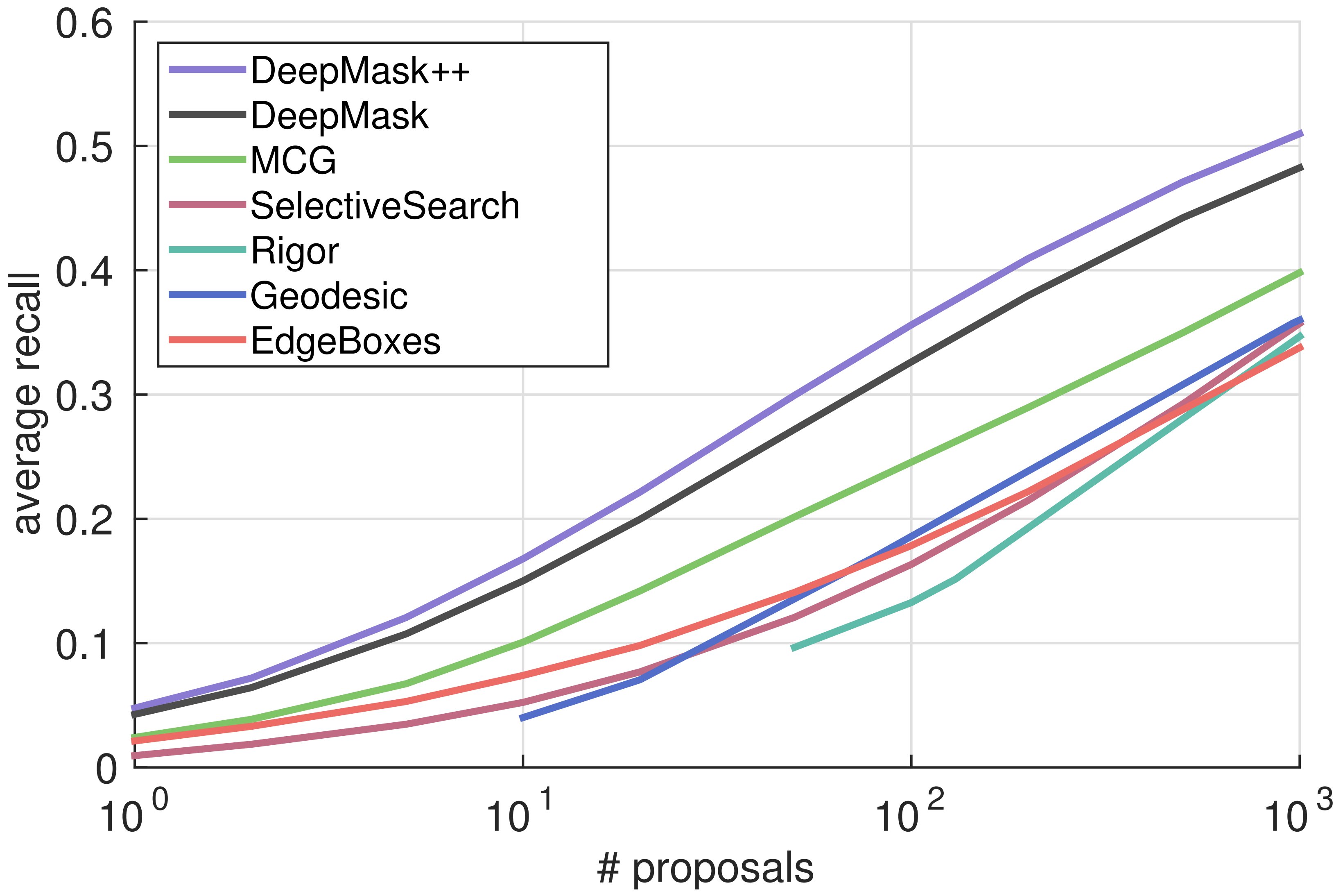
New: Iterative Localization (+1.0 AP)



New: Top-Down Refinement (+0.7 AP)



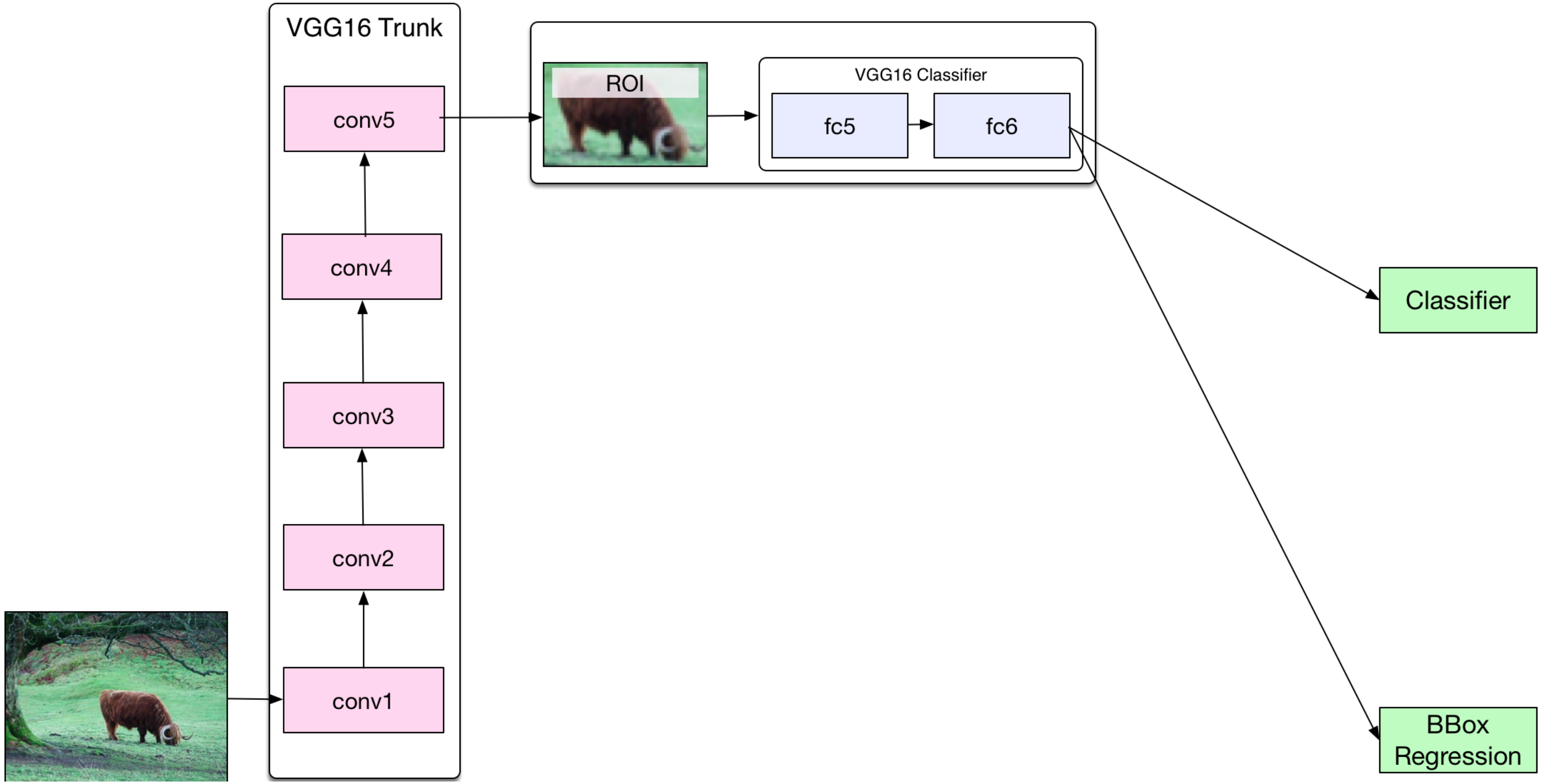
Proposal Quality (boxes)



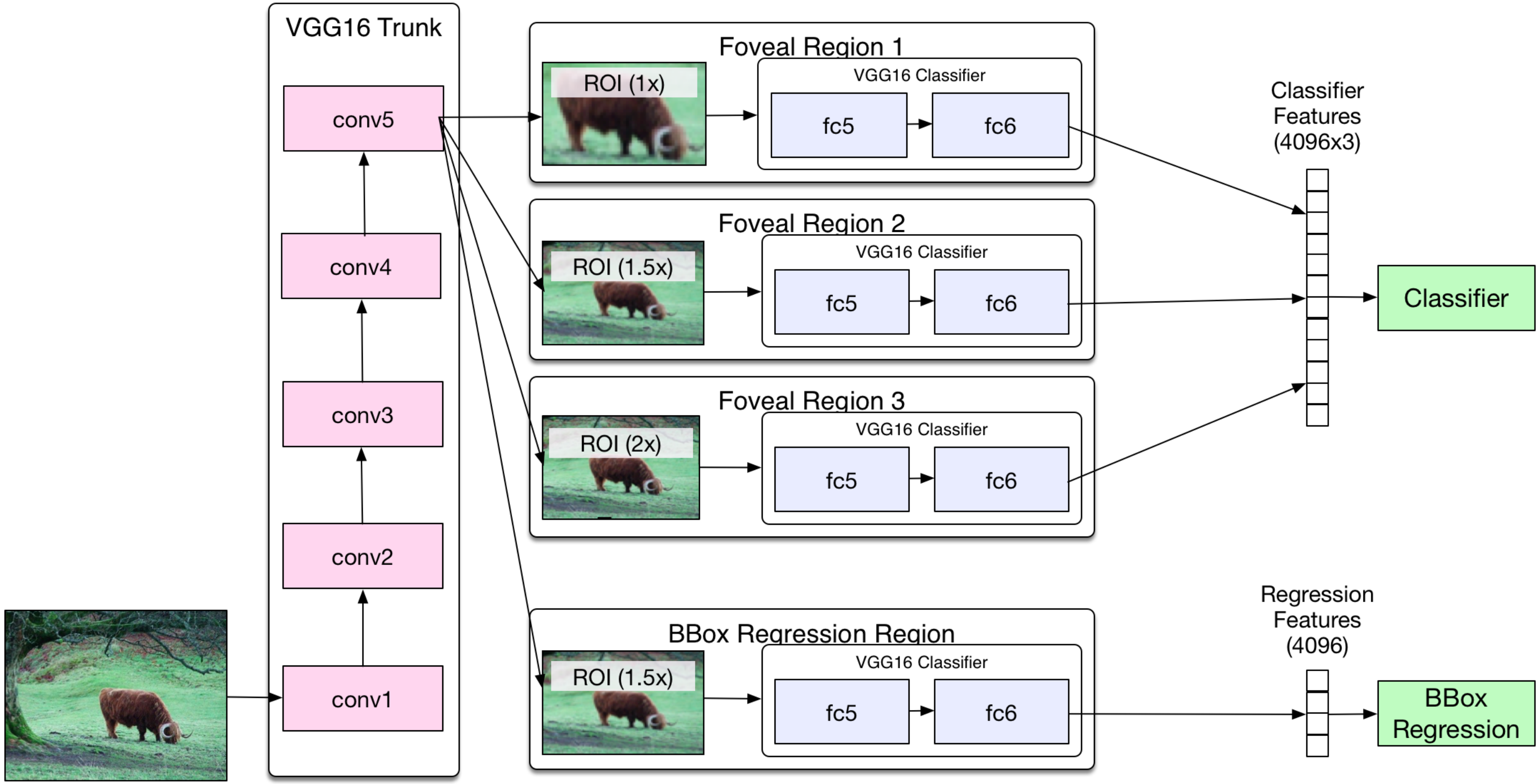
DeepMask Object Proposals

II. Classification framework

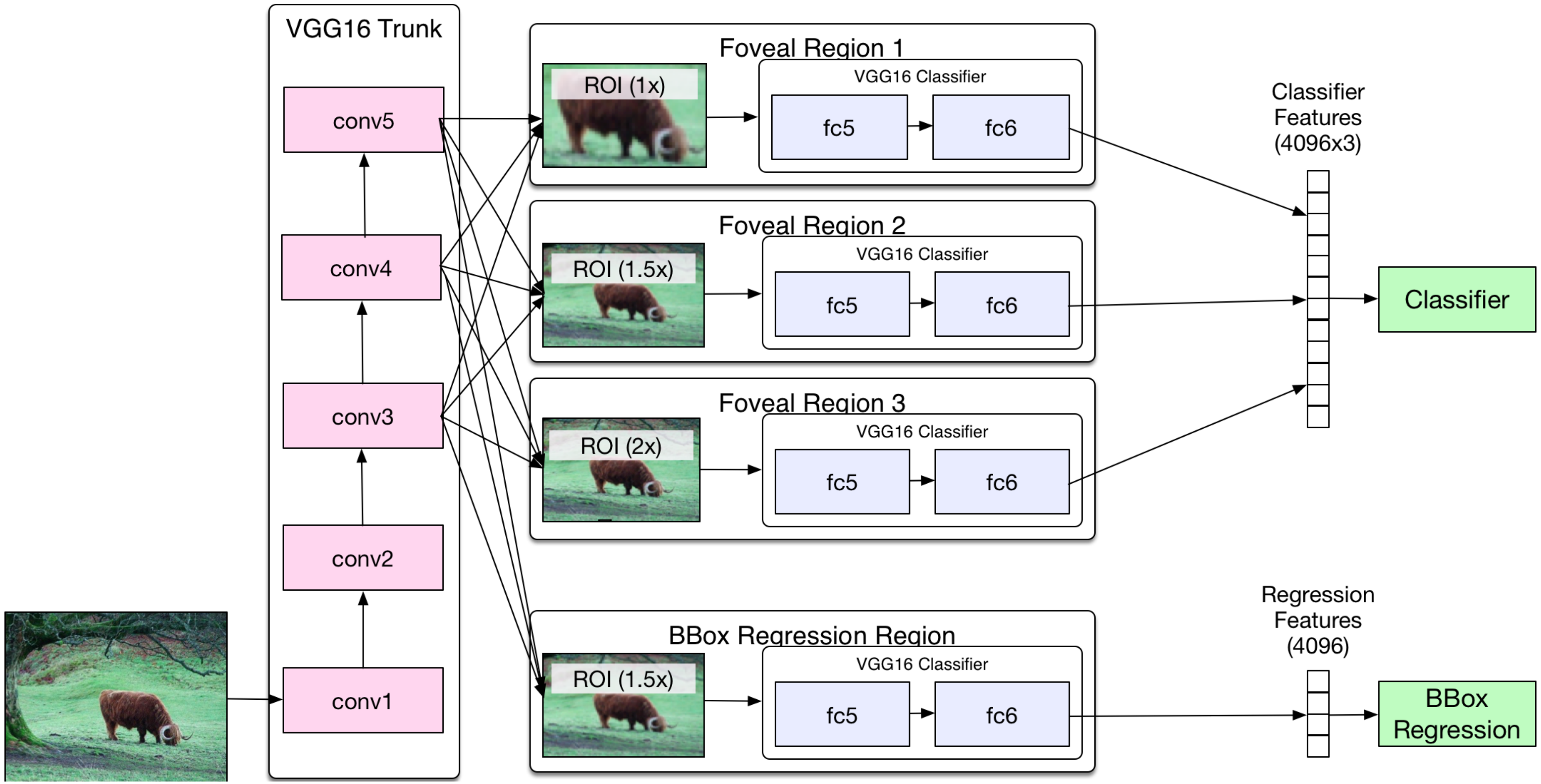
- Fast R-CNN setup [Girshick, ICCV15]



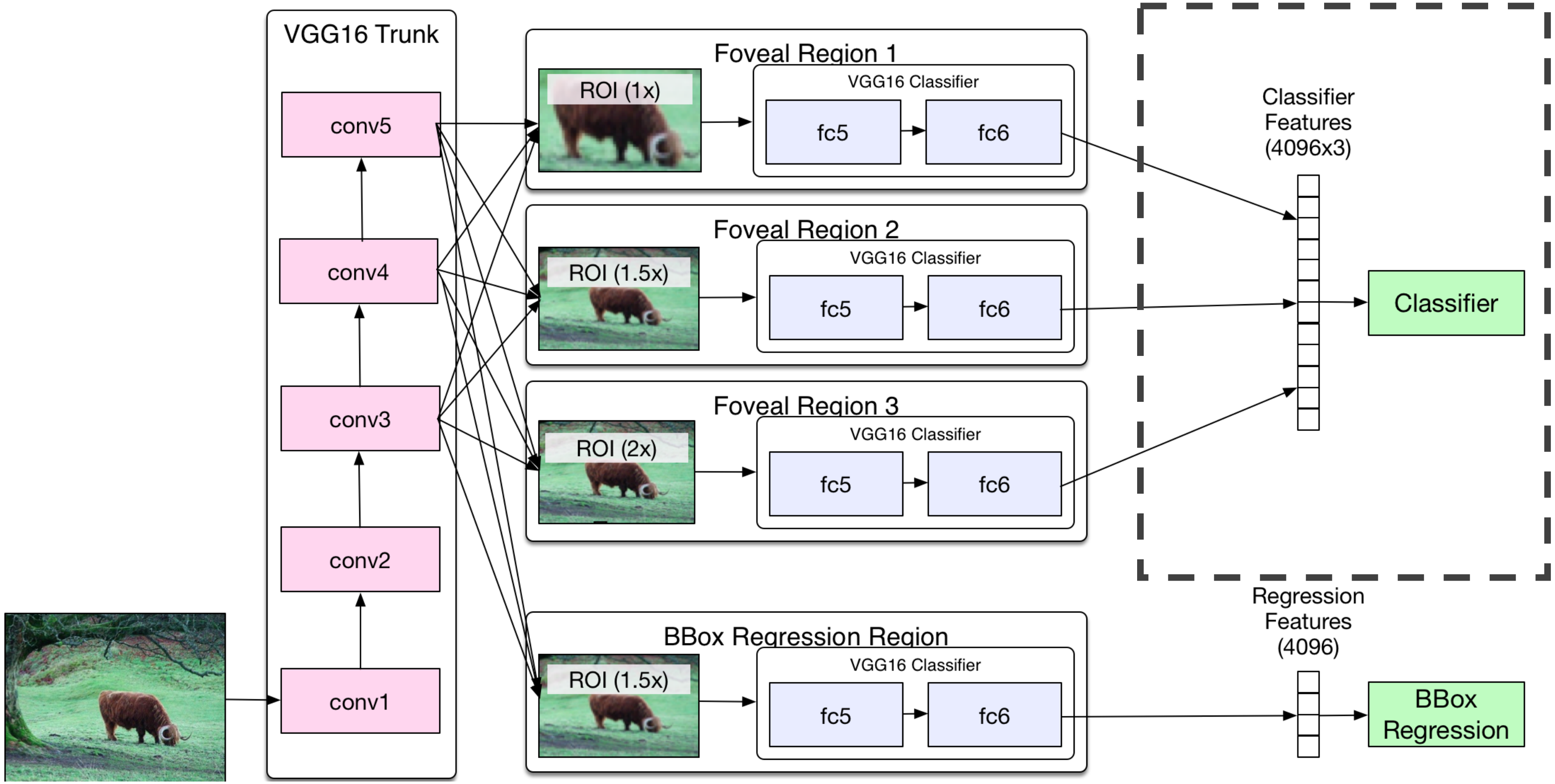
- Fast R-CNN setup [Girshick, ICCV15]
- Foveal structure [inspired by Gidaris & Komodakis, ICCV15] (+2 AP)



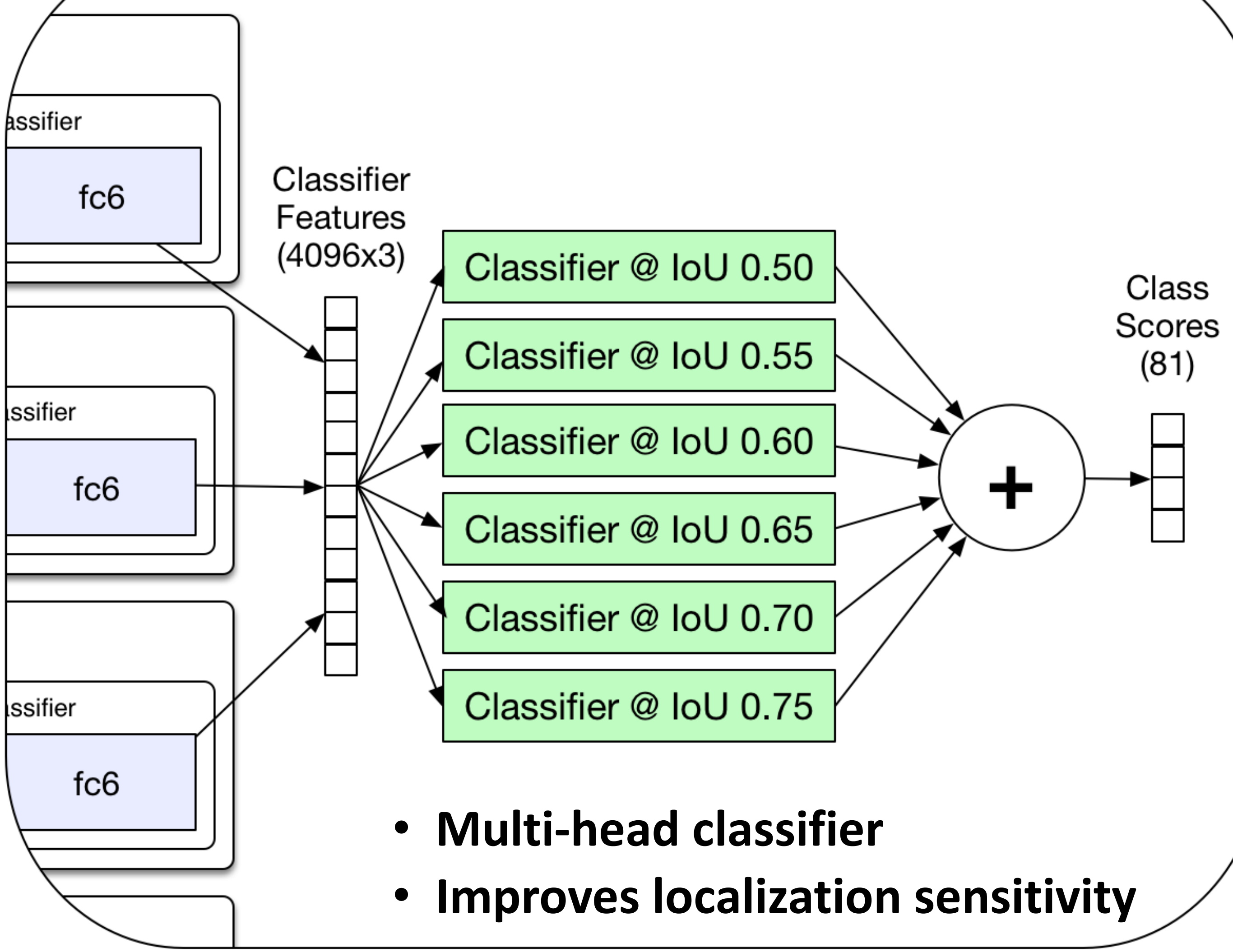
- Fast R-CNN setup [Girshick, ICCV15]
- Foveal structure [inspired by Gidaris & Komodakis, ICCV15] (+2 AP)
- Skip connections (+1 AP)



- Fast R-CNN setup [Girshick, ICCV15]
- Foveal structure [inspired by Gidaris & Komodakis, ICCV15] (+2 AP)
- Skip connections (+1 AP)



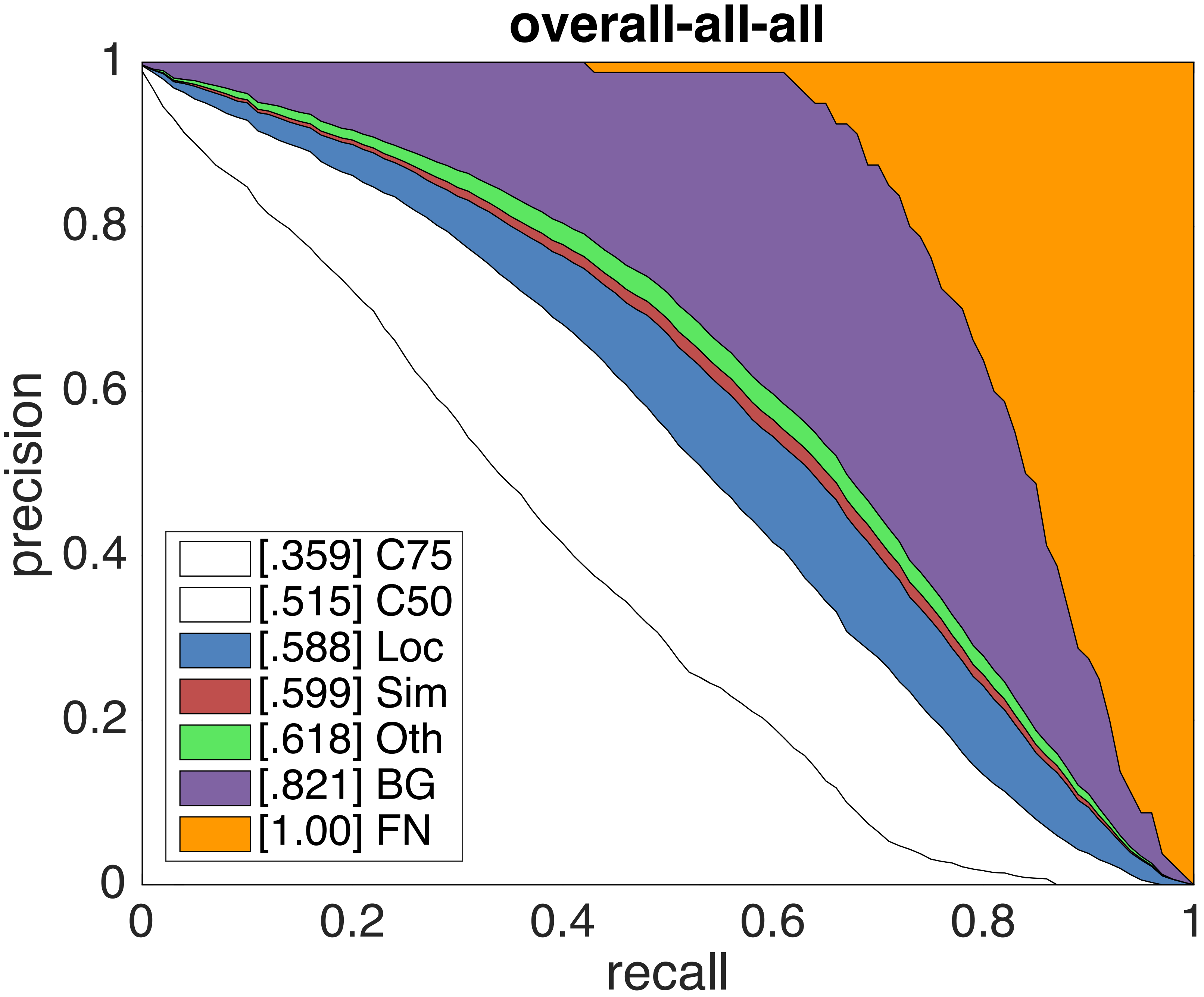
Multi-threshold Loss (+1.5 AP)



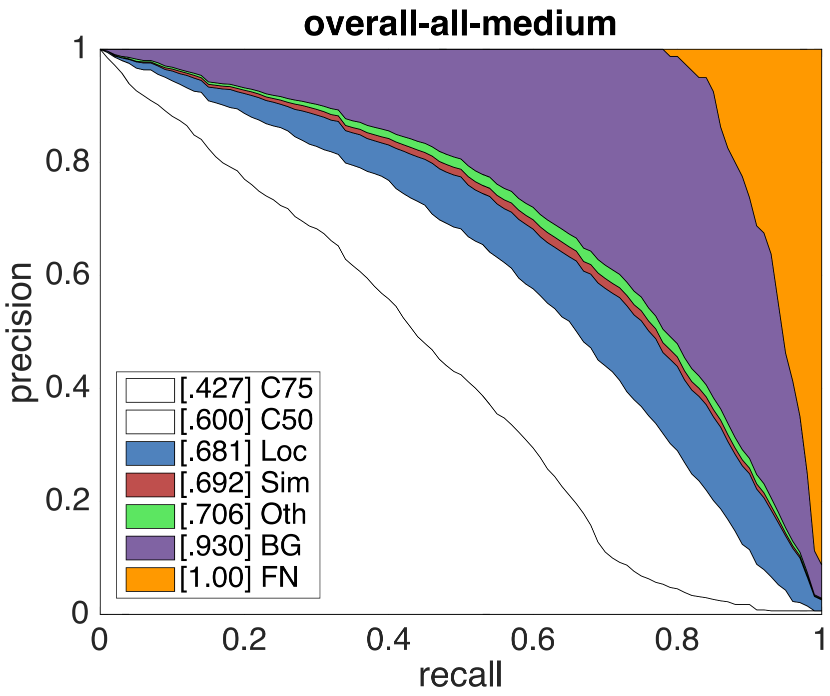
- **Multi-head classifier**
- **Improves localization sensitivity**

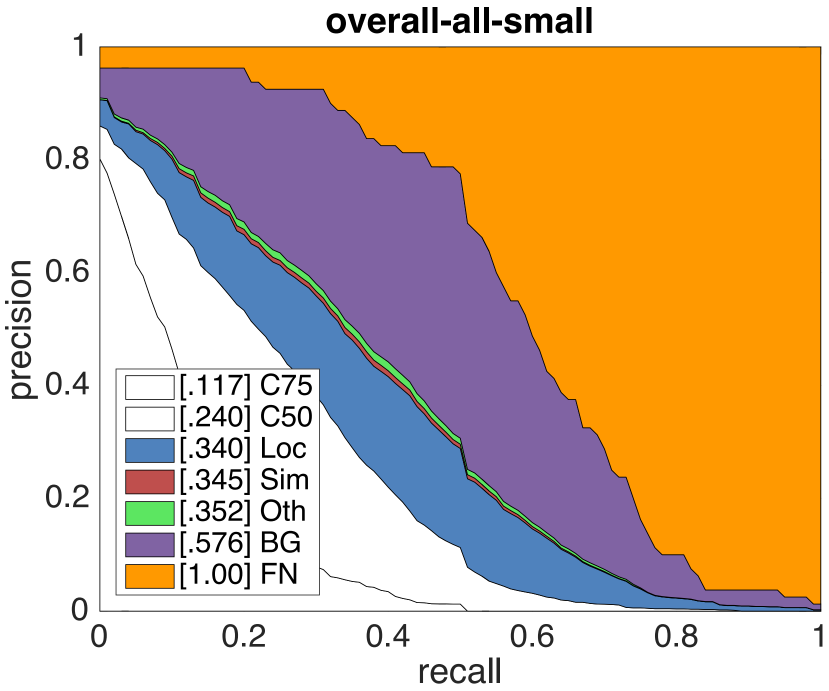
Inference

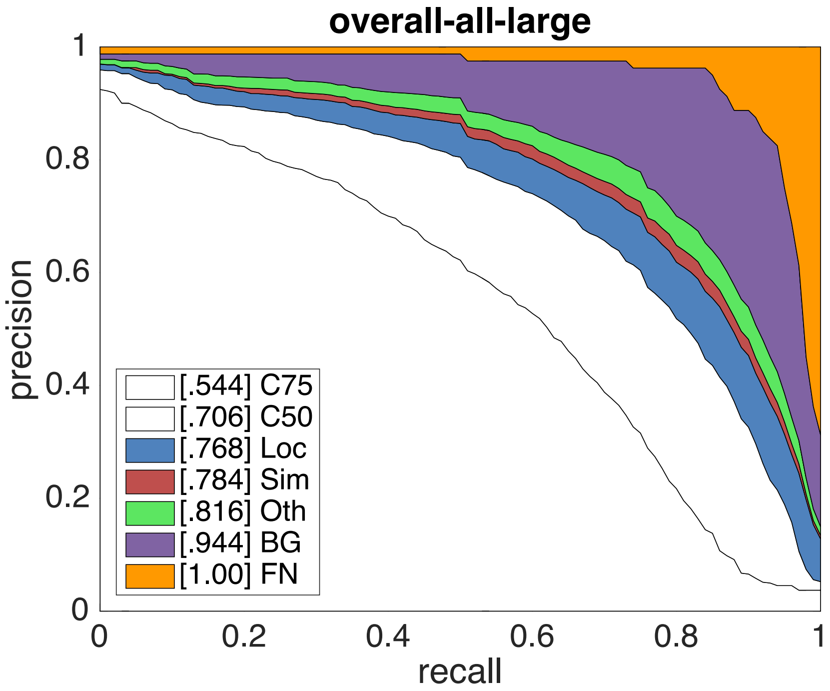
Base Model	**30.1 AP**
+ horizontal flip	**31.1 AP**
+ ROI Pooling ‘2 crop’	**32.1 AP**
+ 7-model Ensemble	**33.5 AP**


overall-all-all
precision
recall
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
[.359] C75
[.515] C50
[.588] Loc
[.599] Sim
[.618] Oth
[.821] BG
[1.00] FN

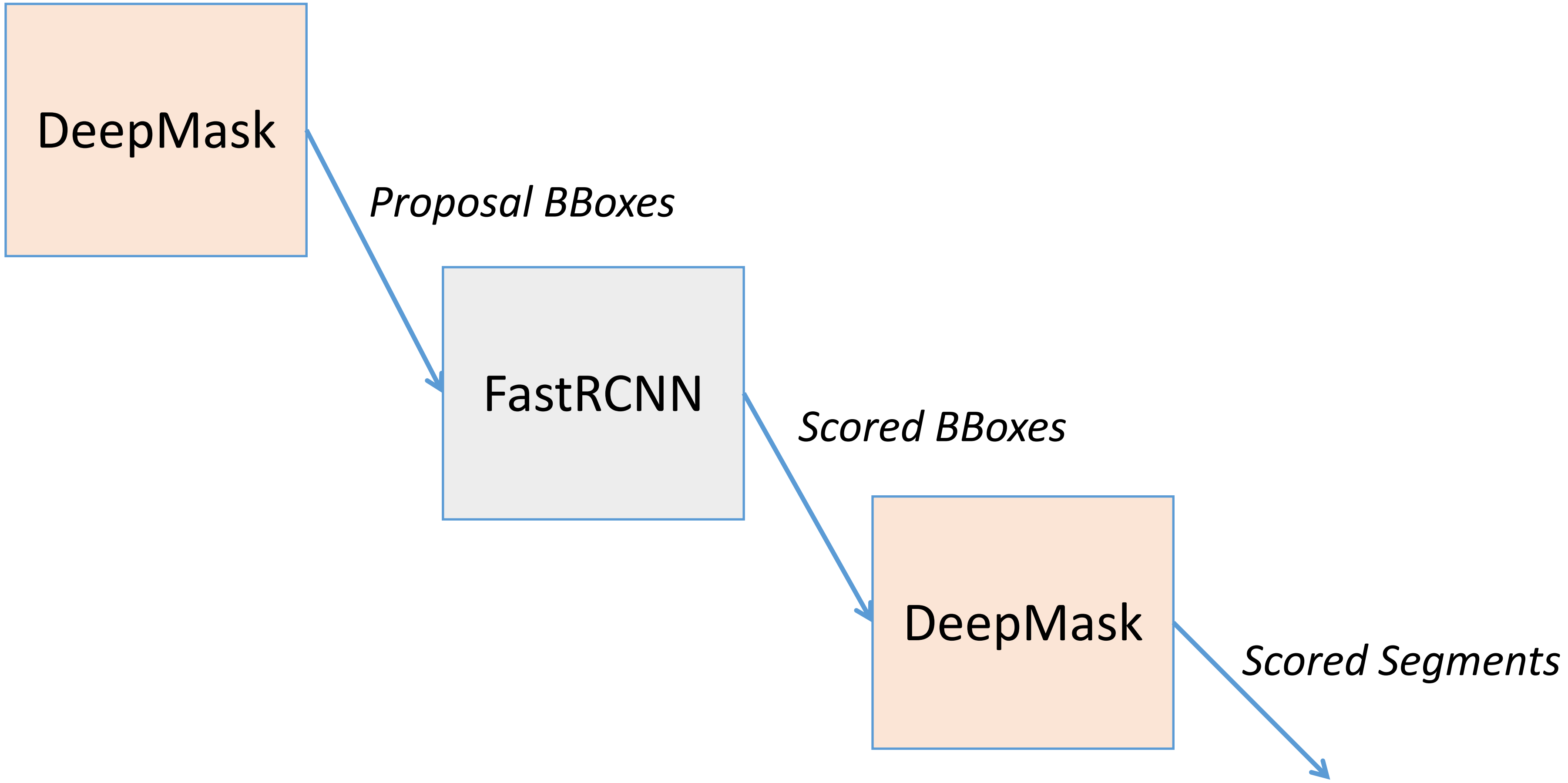
	AP
Small	0.139
Medium	0.378
Large	0.477



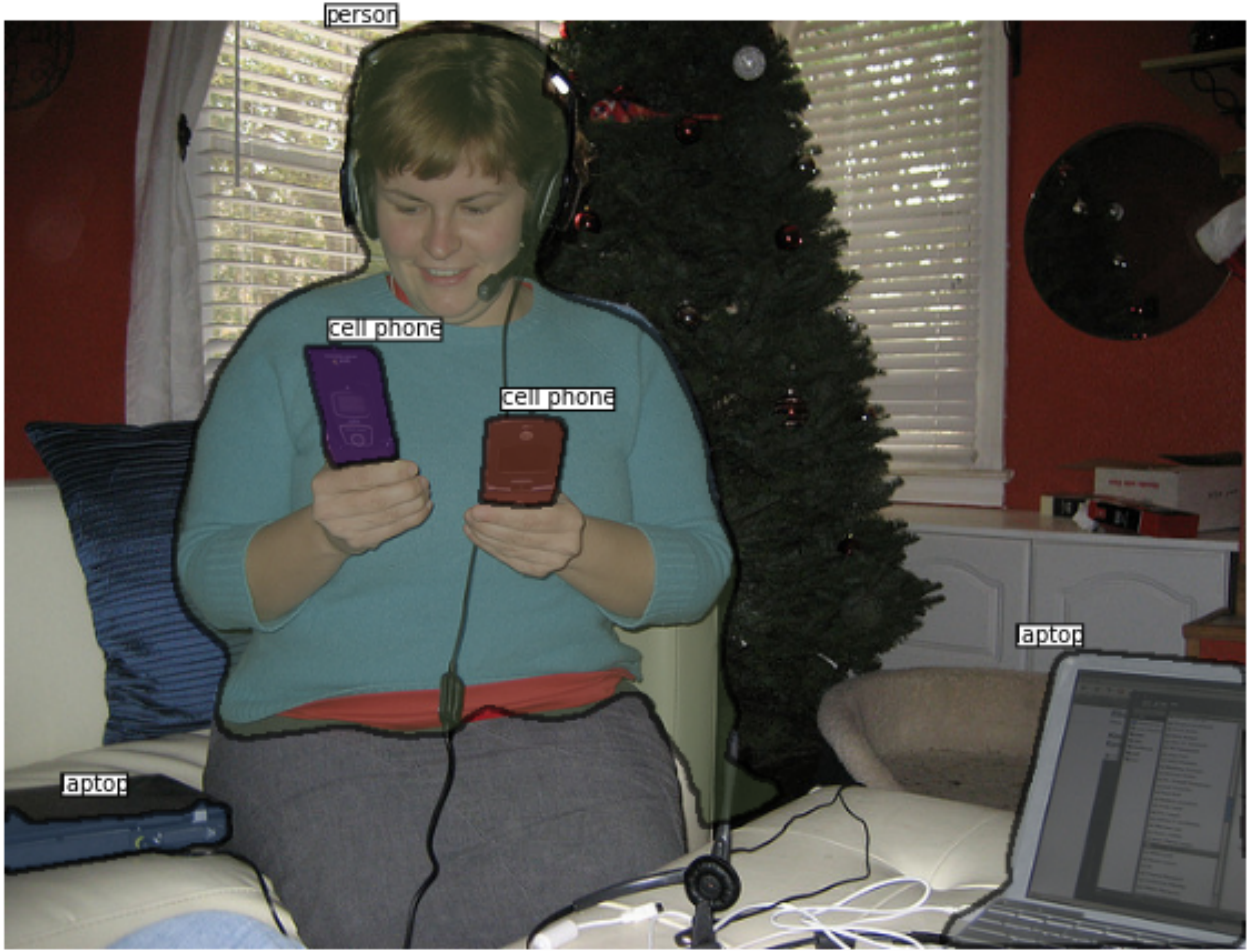


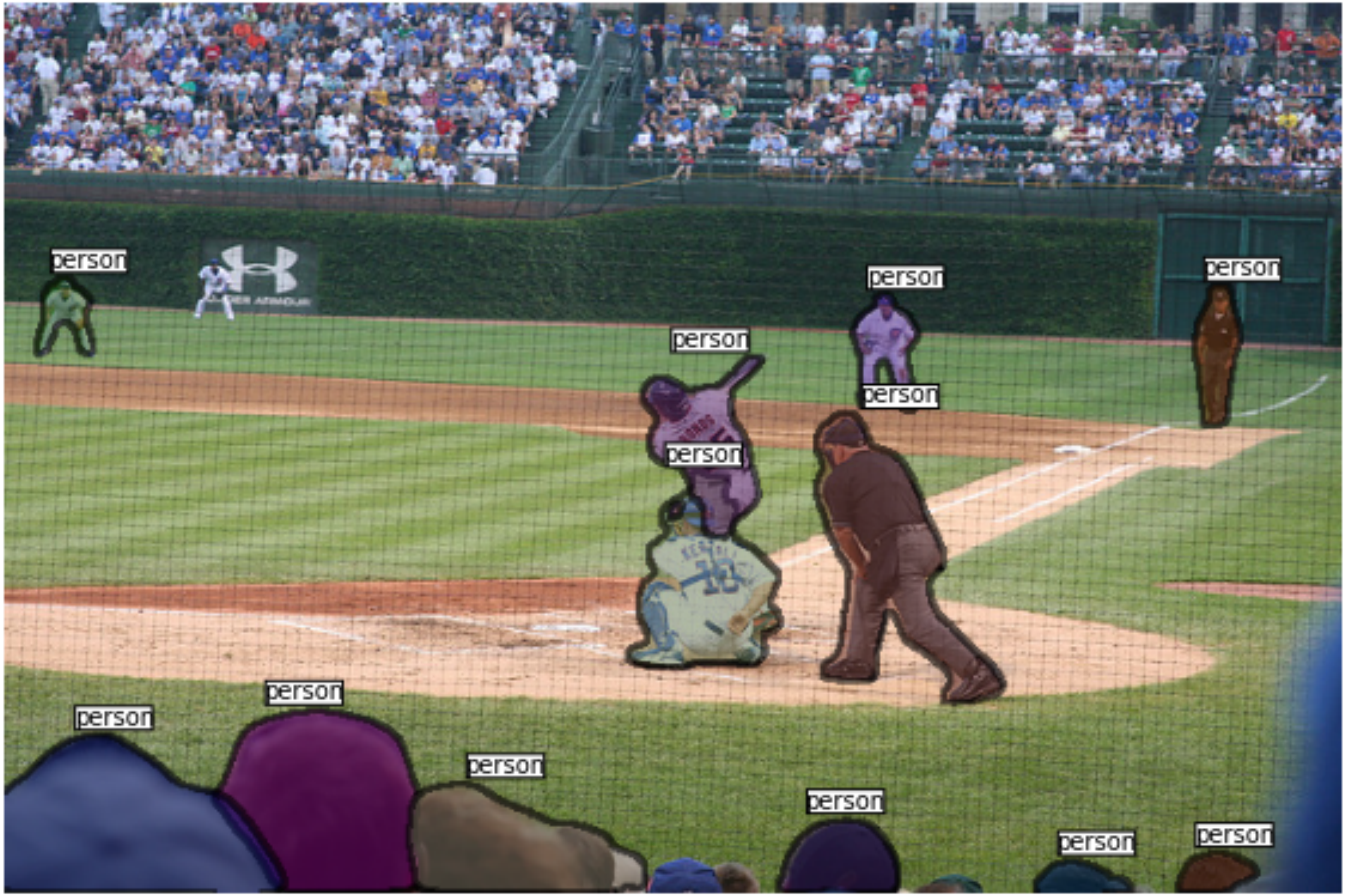


Segmentation Examples

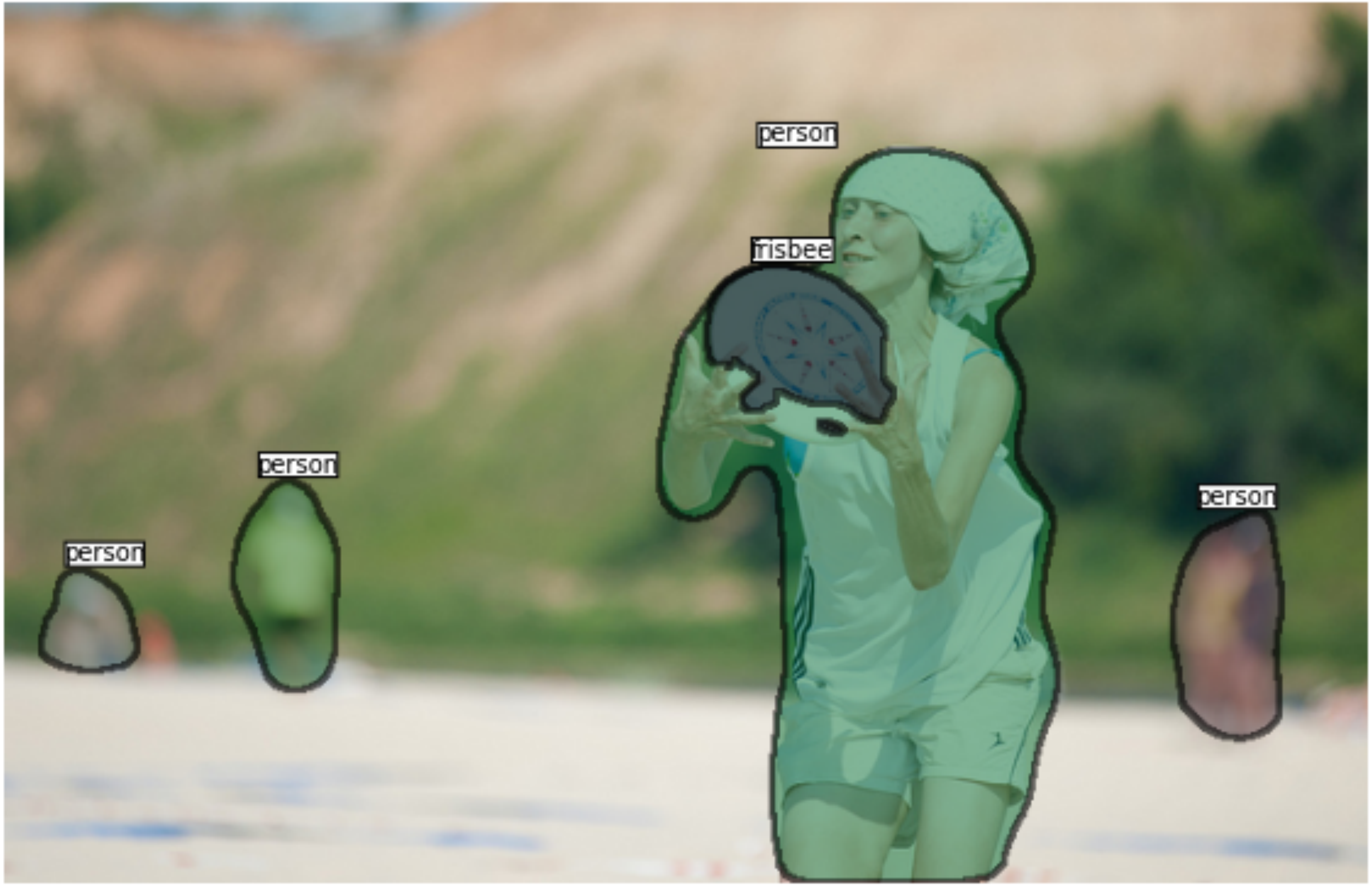


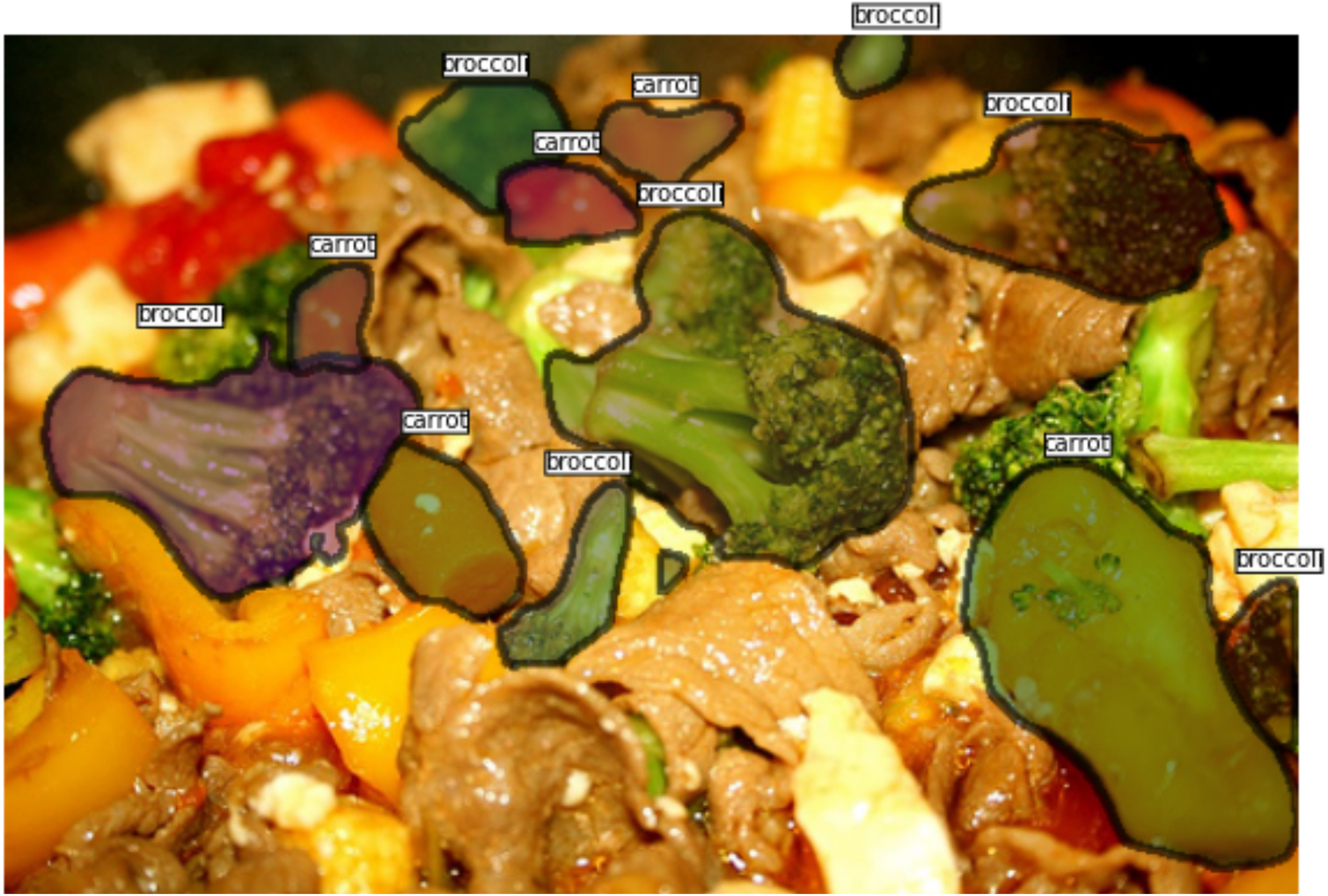

person
boat
boat
person
skateboard


person
cell phone
cell phone
laptop
laptop

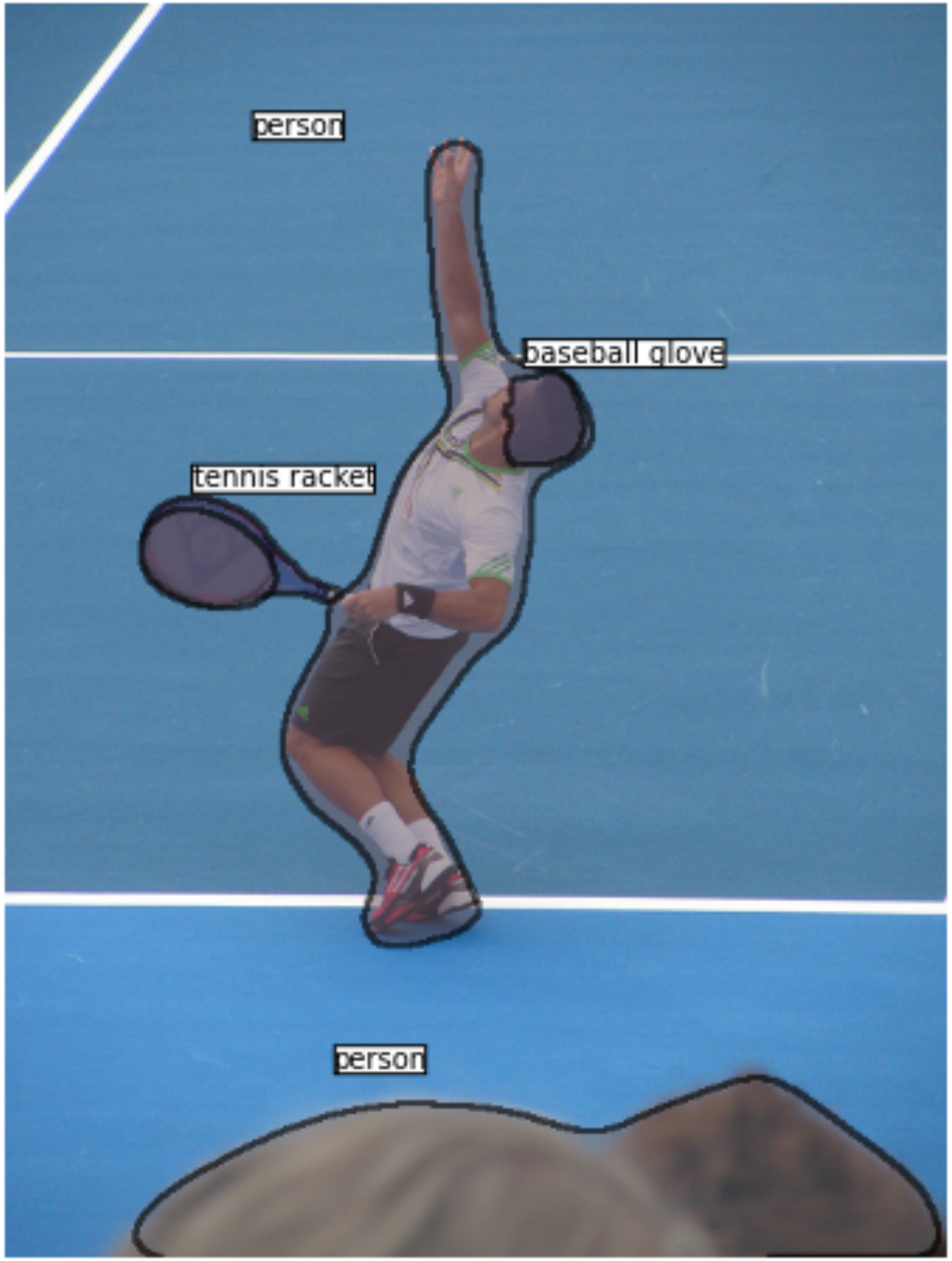

person
person
person
person
person
person
person
person
person
person
person
person
person

557556




broccol
broccol
carrot
carrot
broccol
broccol
carrot
broccol
carrot
broccol
carrot
broccol


traffic light
traffic light
car
car
person
car
car


person
baseball glove
tennis racket
person


bear
frisbee
frisbee


sink
sink


mouse
refrigerator
refrigerator
sink
sink

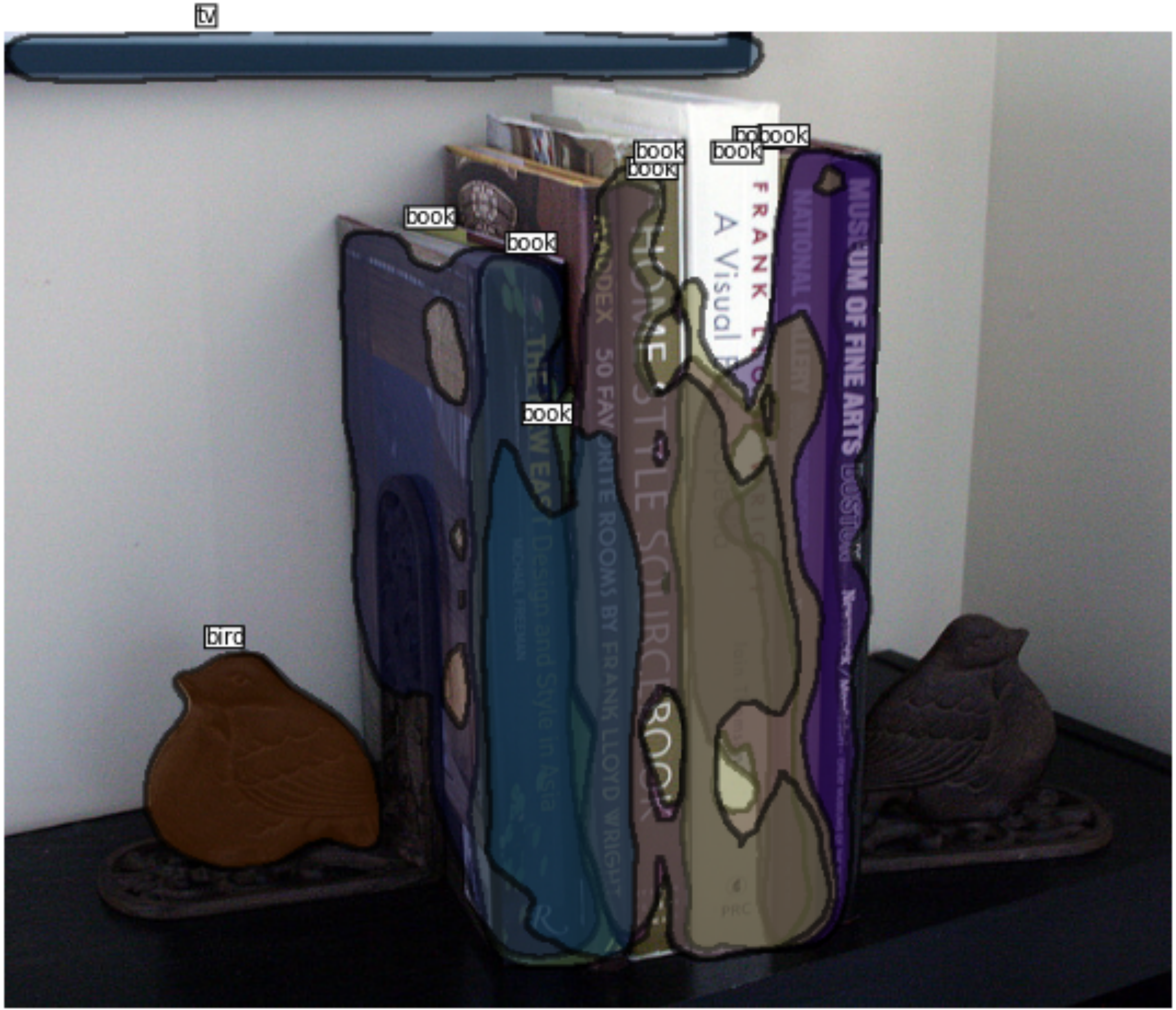

sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep


sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep
sheep

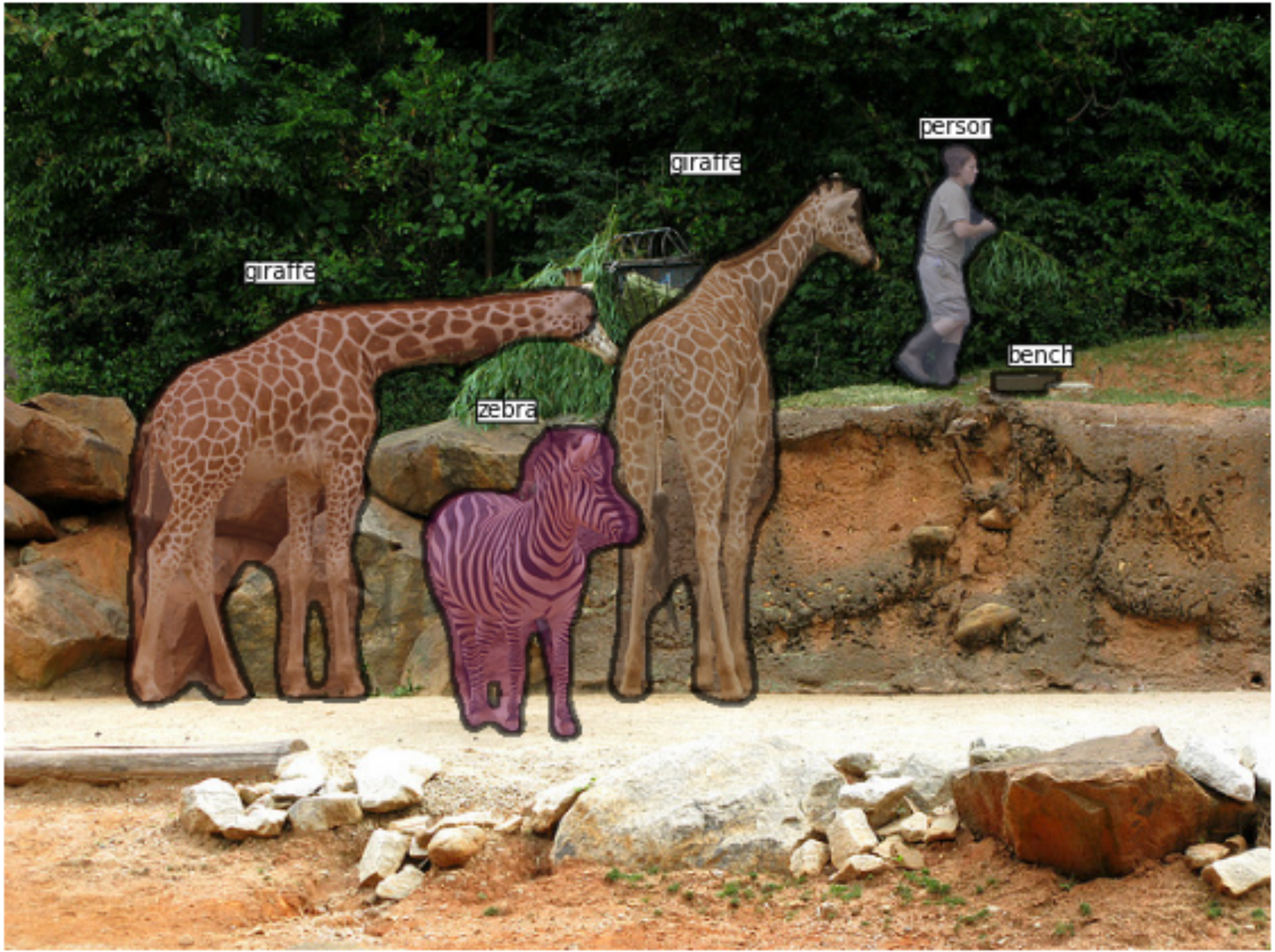

vase

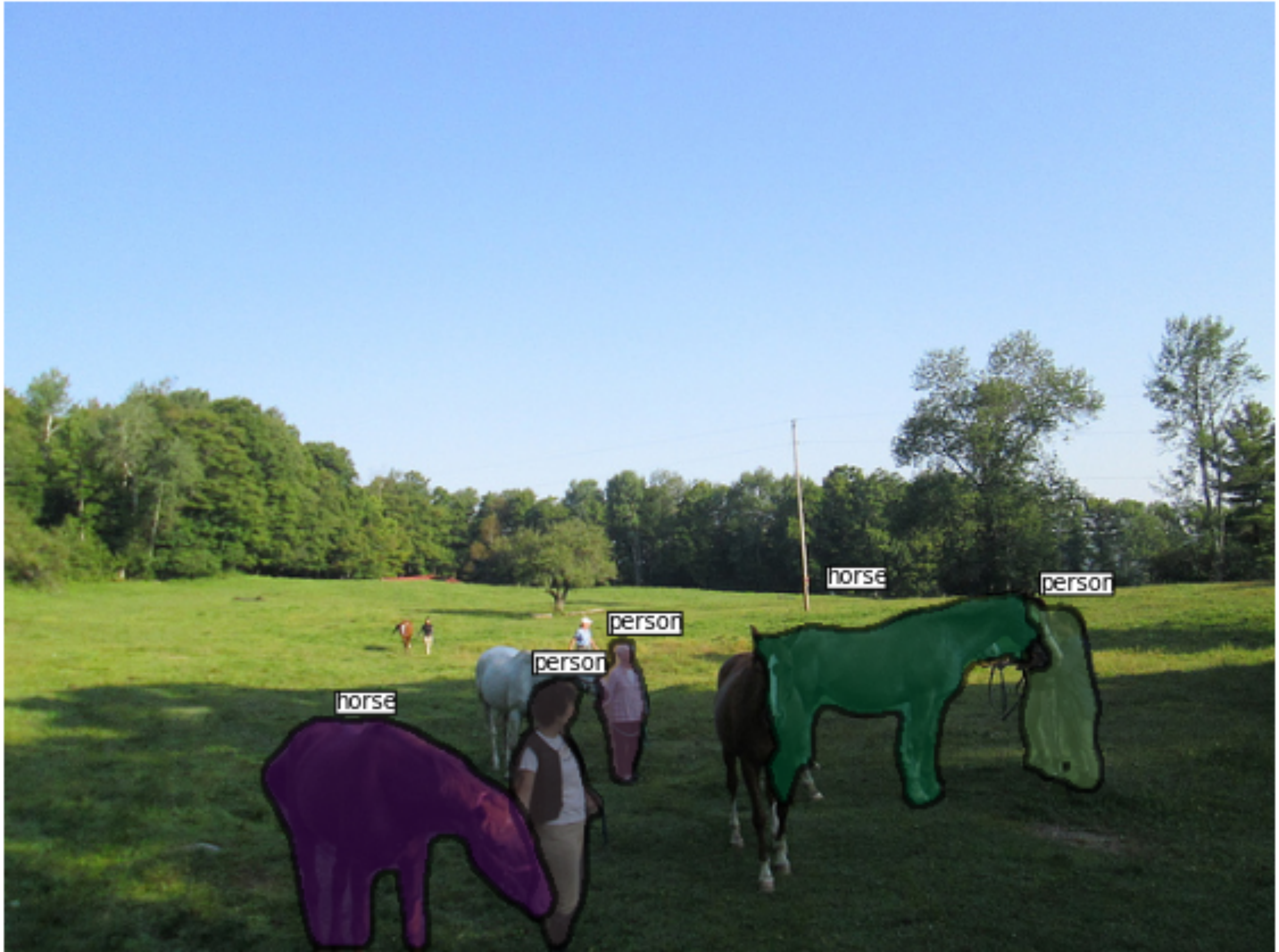
sandwich
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut
donut

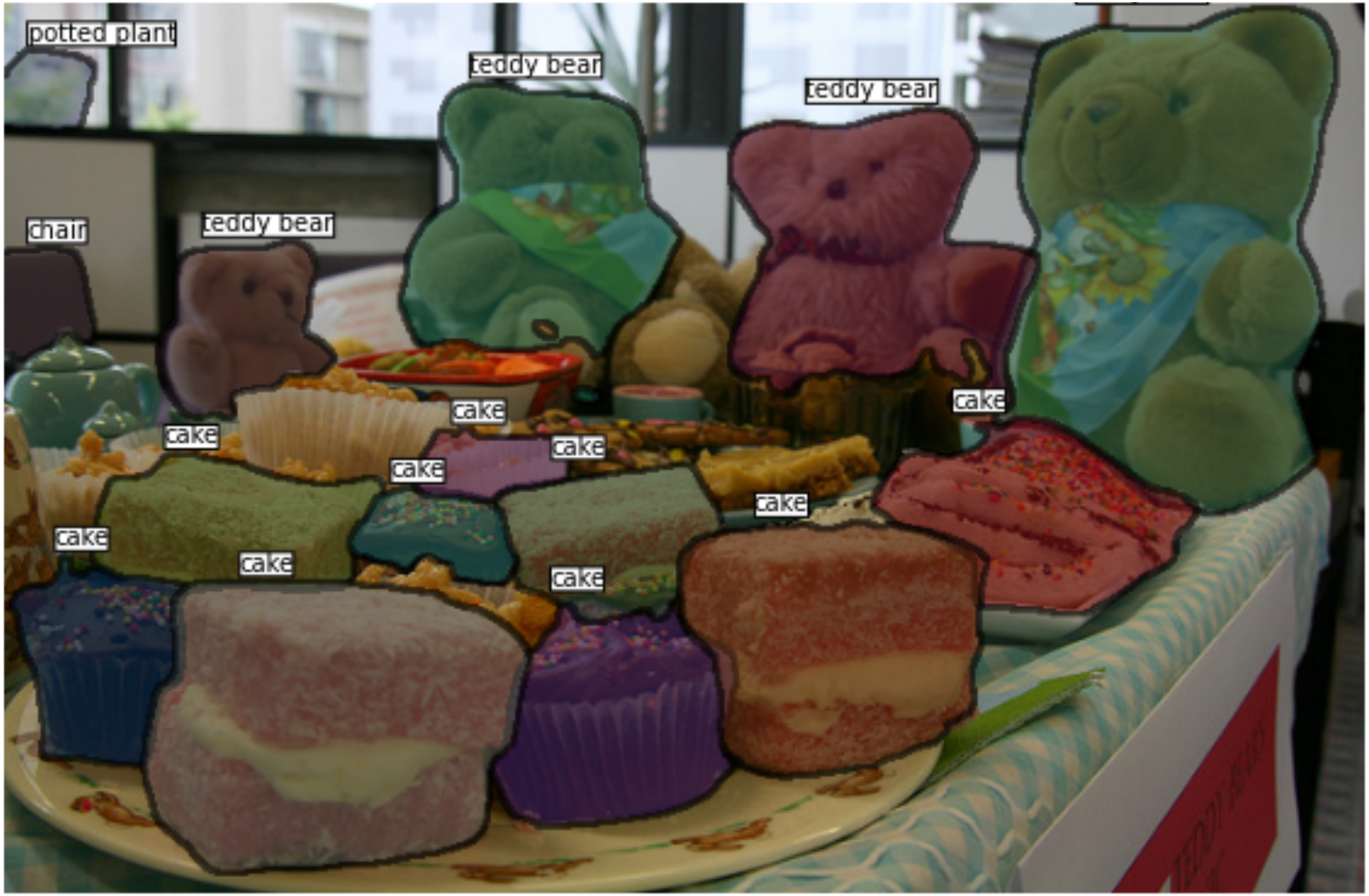

zebra
zebra

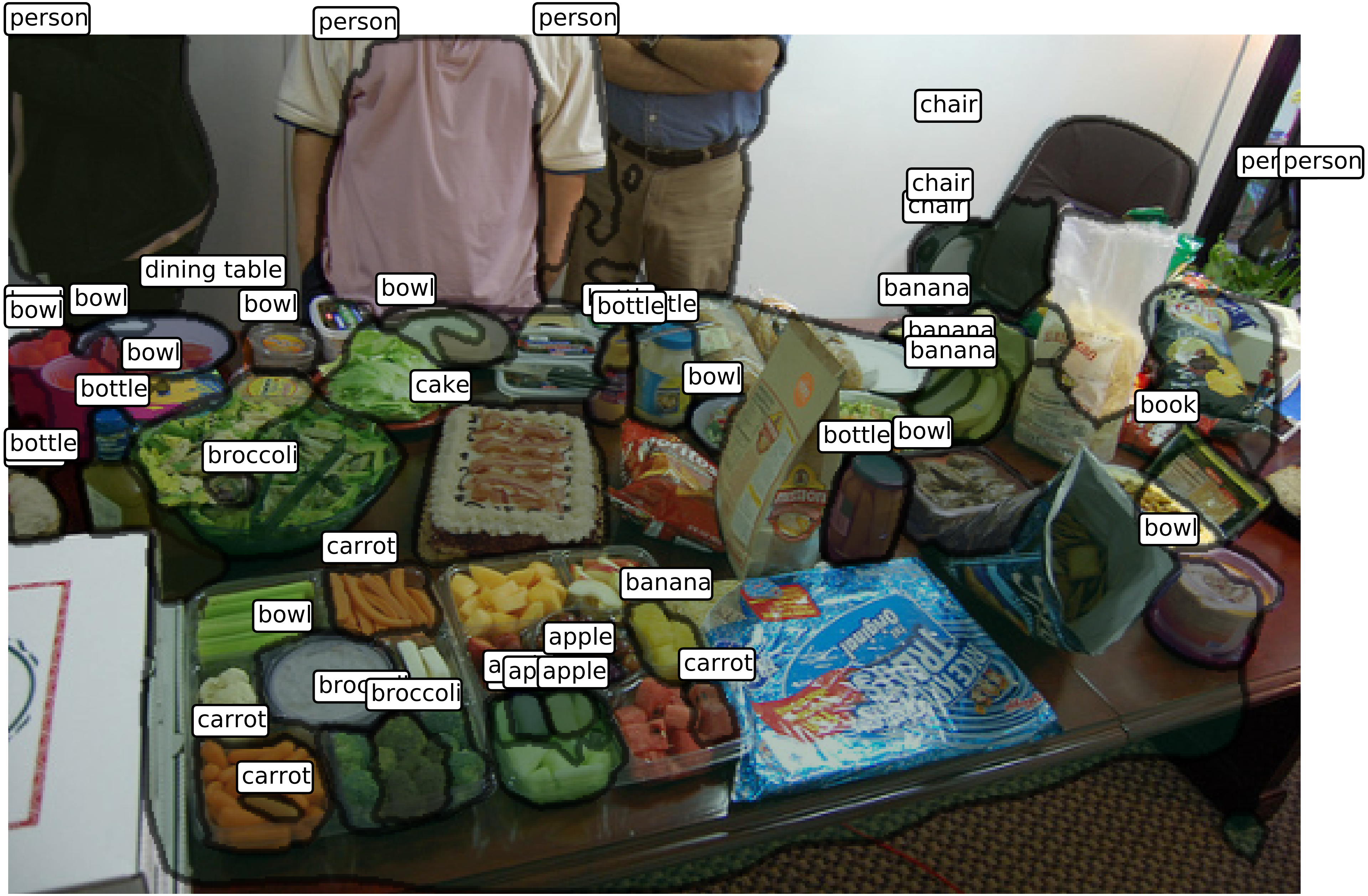

tv
book
book
book
book
book
book
book
book
bird
MUSEUM OF FINE ARTS BOSTON
A Visual
50 FAVORITE ROOMS BY FRANK LLOYD WRIGHT


bottle
toilet
toilet


person
giraffe
giraffe
bench
zebra


horse
person
person
person
horse
horse
person


potted plant
teddy bear
teddy bear
chair
teddy bear
cake
cake
cake
cake
cake
cake
cake
cake
cake


person
person
person
chair
chair
chair
per
person
dining table
bowl
bowl
bowl
bowl
bottle
bottle
bowl
bowl
banana
banana
banana
cake
bowl
bottle
bottle
bowl
broccoli
book
carrot
banana
bowl
bowl
apple
apple
ap
carrot
broccoli
broccoli
carrot
carrot

Future Directions

- most room for improvement:
  - background confusion (FP/FN)
  - small objects
- more effective use of context
- fast / proposal-free detection

Questions?


cup
bowl
clock
person
person
bowl
clock
clock
vase
person
vase
bowl
bottle
clock
vase
vase
vase
cup
vase
cup
cup
vase
vase
vase
vase
cup
voyageAnatolia.blogspot.com