GDR-ISIS, Paris
March 20, 2015

Weakly Supervised Object Recognition with Convolutional Neural Networks

Ivan Laptev

ivan.laptev@inria.fr

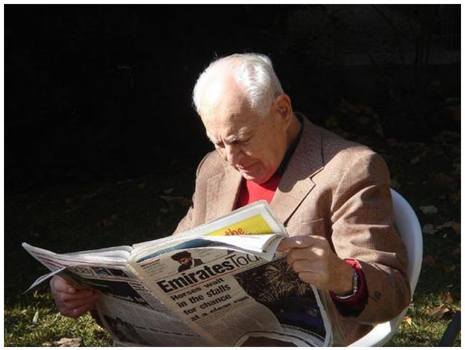
WILLOW, INRIA/ENS/CNRS, Paris



Joint work with: Maxime Oquab – Leon Bottou – Josef Sivic

Summary

Training input

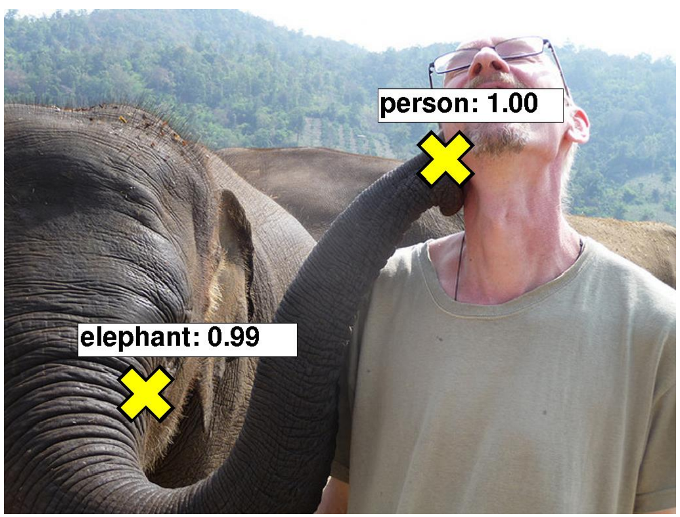


+

image-level labels:

✓ Person	✓ Reading
✓ Chair	× Riding bike
× Airplane	× Running
…	…

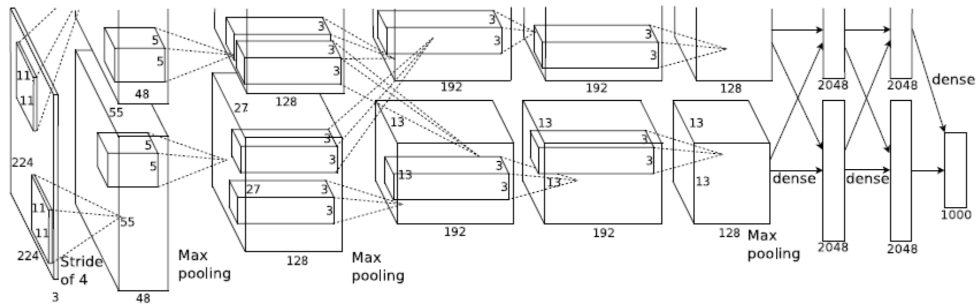
Test output



More details in http://www.di.ens.fr/willow/research/weakcnn/

Recent Progress: Convolutional Neural Networks

- Success in character recognition [LeCun'88].
- Limited performance on natural images until 2012.



- [Krizhevsky et al. 2012]: break-through in ImageNet object classification.

ILSVRC'12: 1.2M images, 1K classes

2012

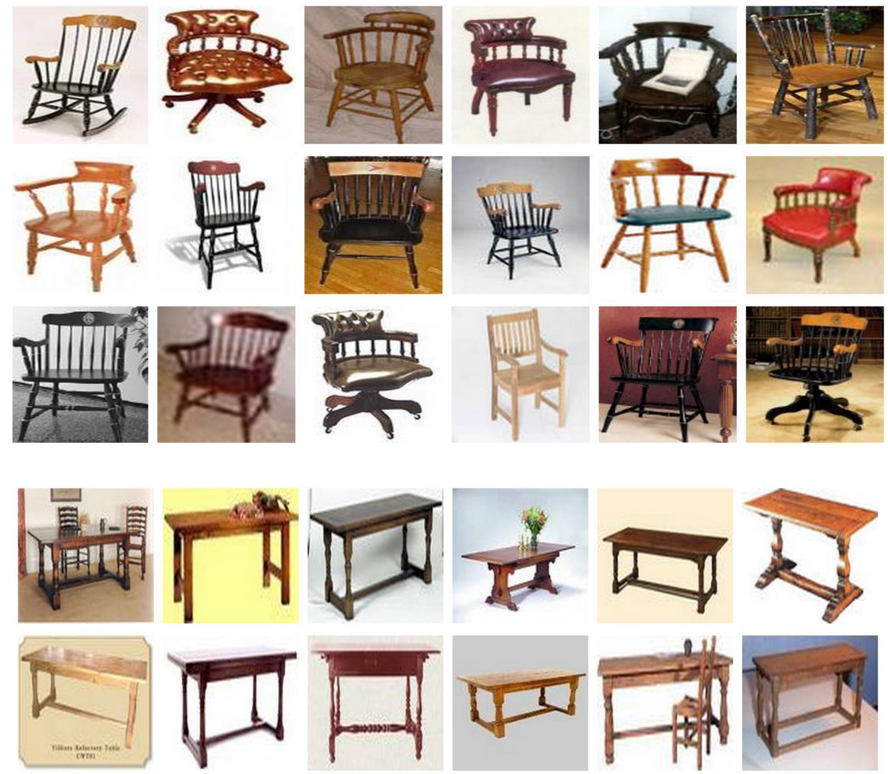
Method:	Top 5 error:
SIFT + FVs [7]	*26.2%*
1 CNN	—
5 CNNs	**16.4%**
1 CNN*	—
7 CNNs*	**15.3%**

2014-2015

GoogLeNet:	6.6%
VGG:	6.8%
BAIDU	5.3%
Human	*5.1%*

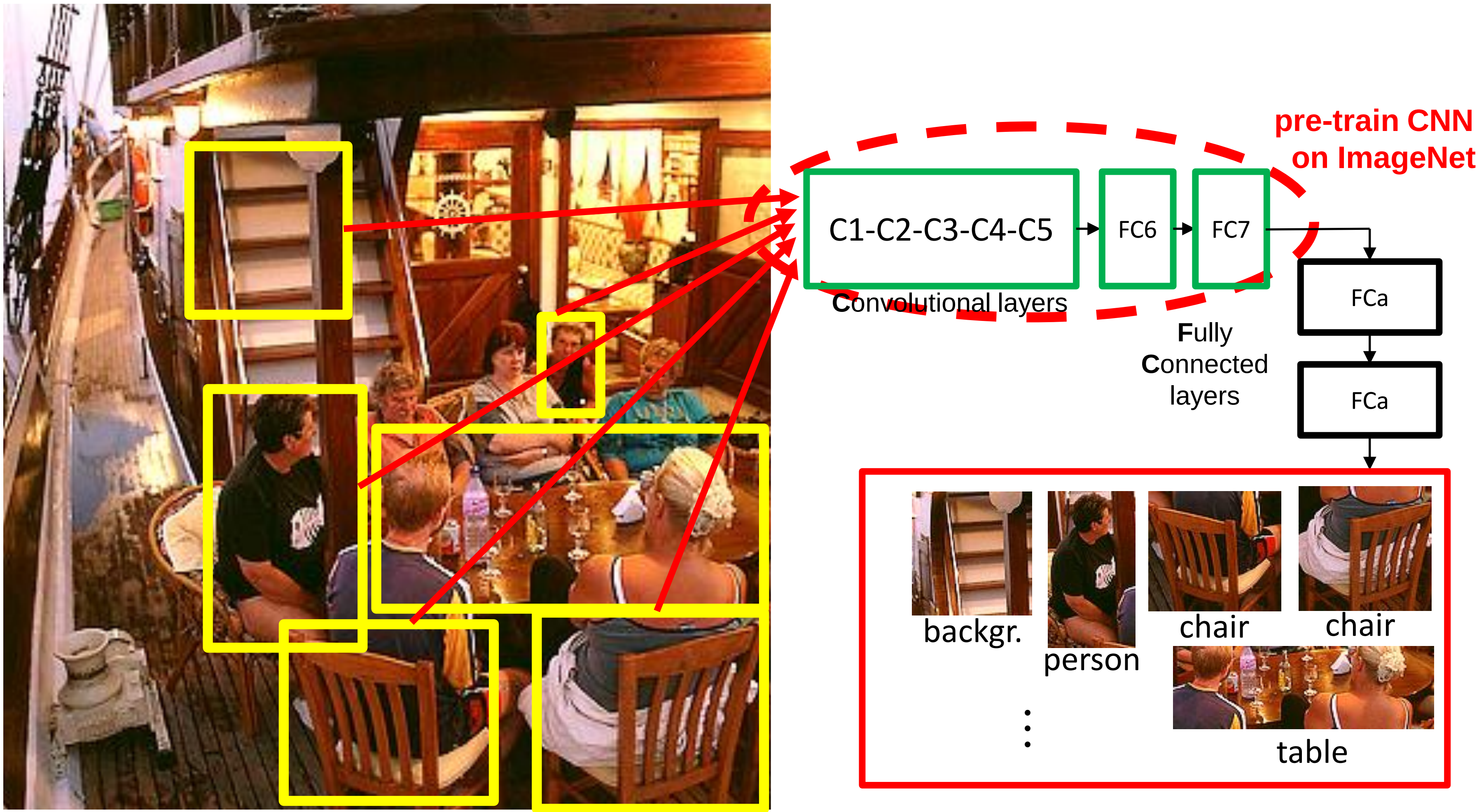
Let’s look at the data

A typical image with chairs and tables on Flickr.com



Images of chairs and tables in ImageNet

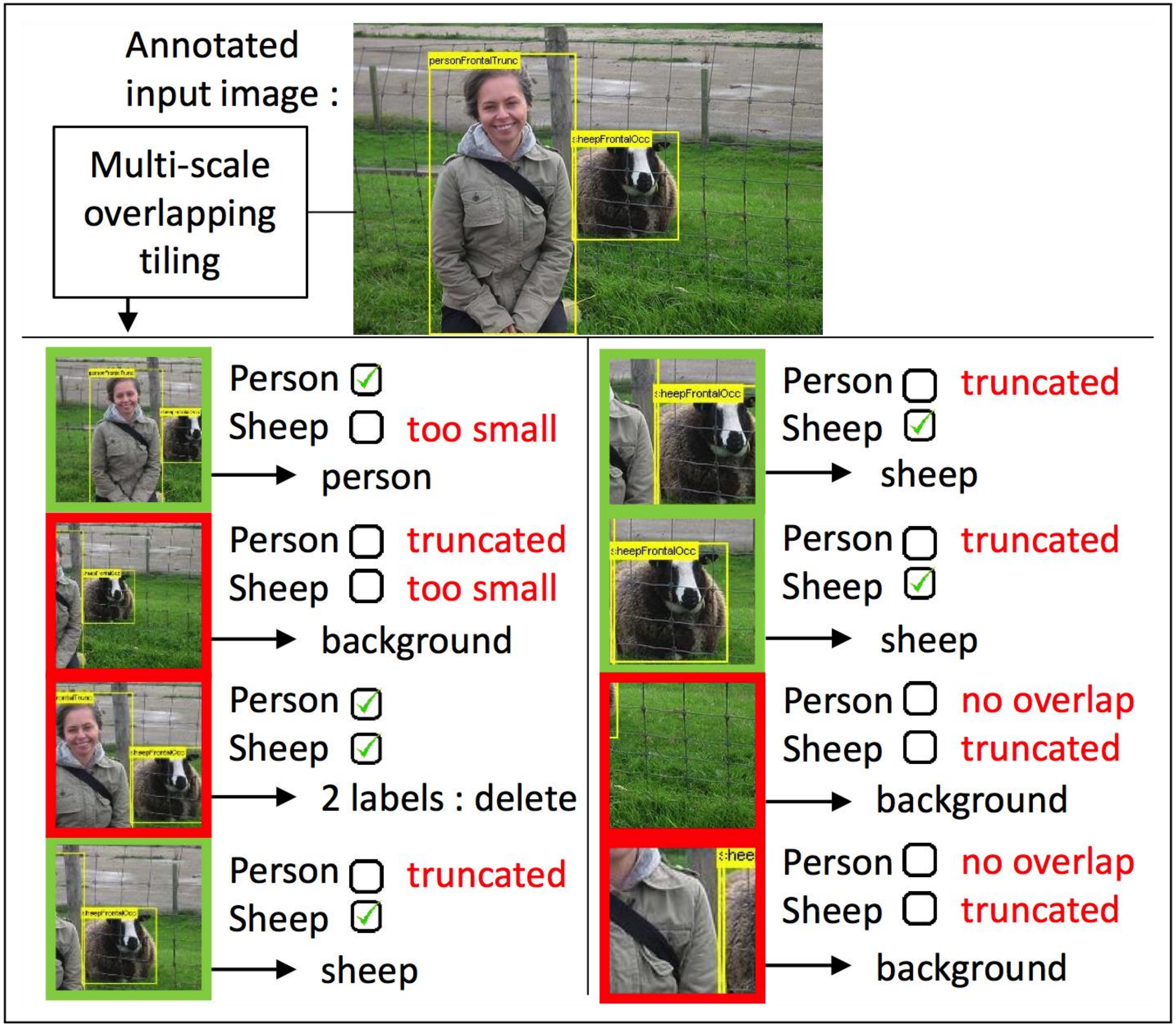
How to use CNNs for cluttered scenes?



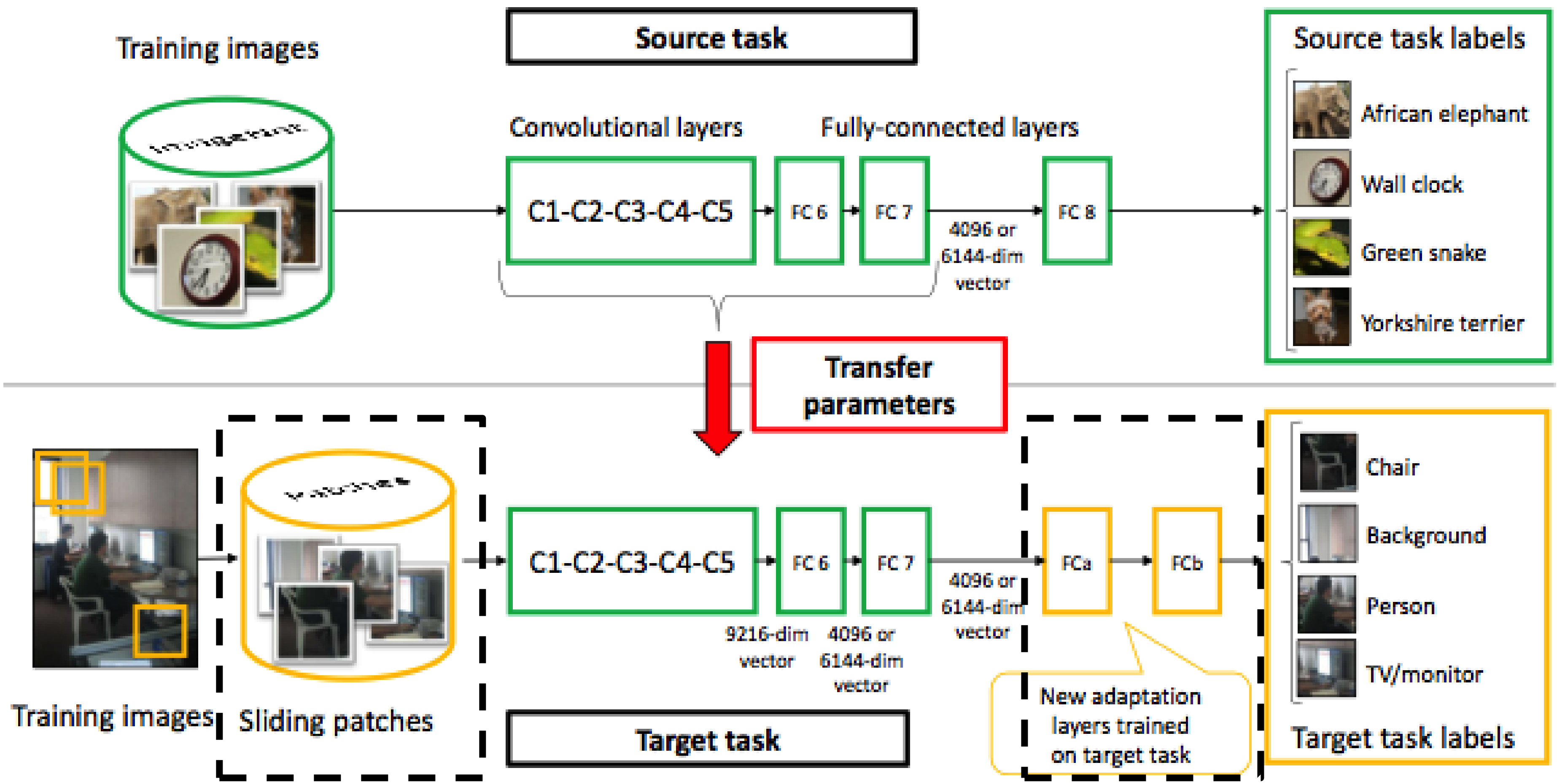
Use ImageNet pre-trained CNN → Post-train on the new task

[Girshick et al.'14], [Oquab et al.'14], [Sermanet et al.'13], [Donahue et al. '13], [Zeiler & Fergus '13] ...

Approach – sliding window training / testing



Approach

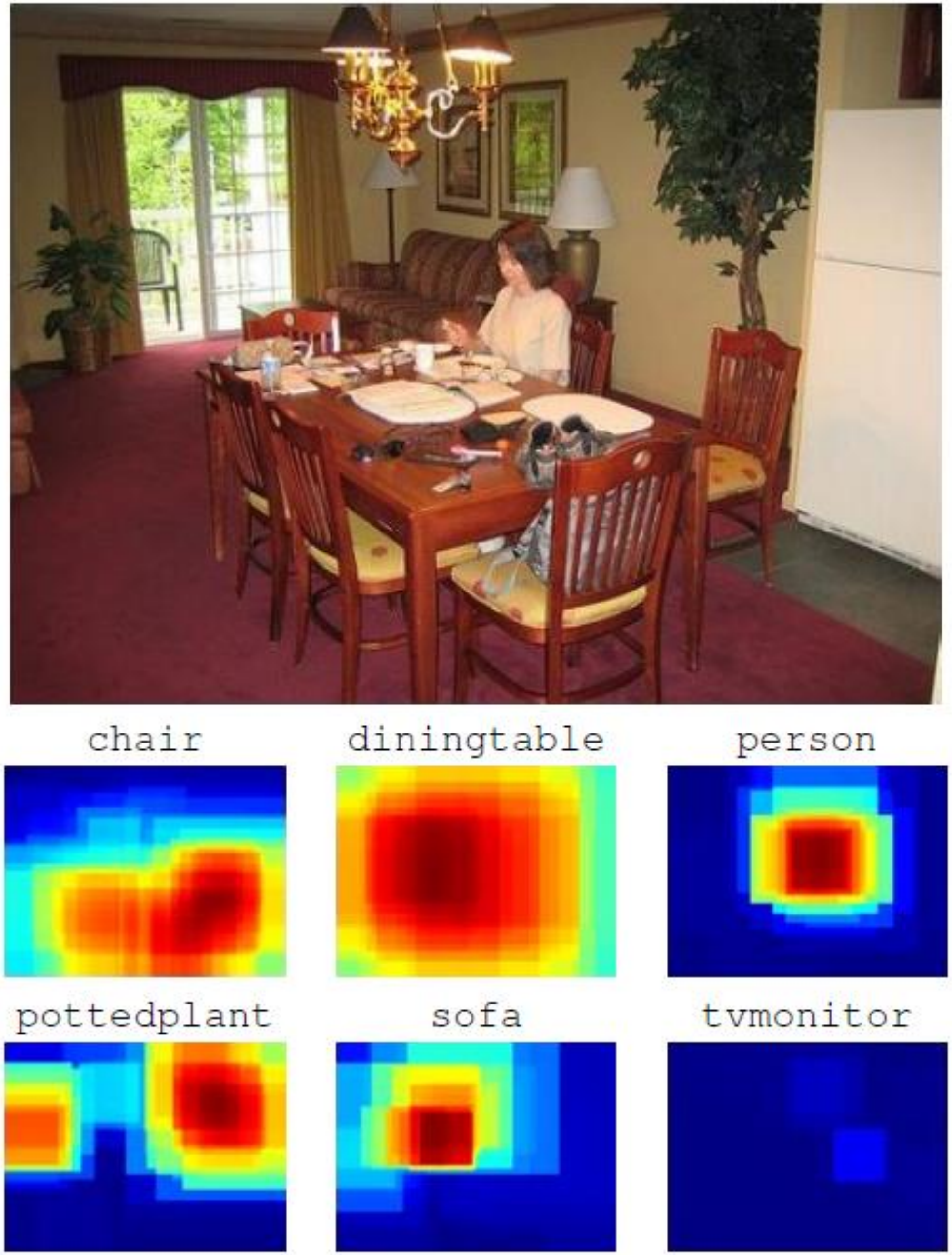


1. Design training/test procedure using sliding windows
2. Train adaptation layers to map labels

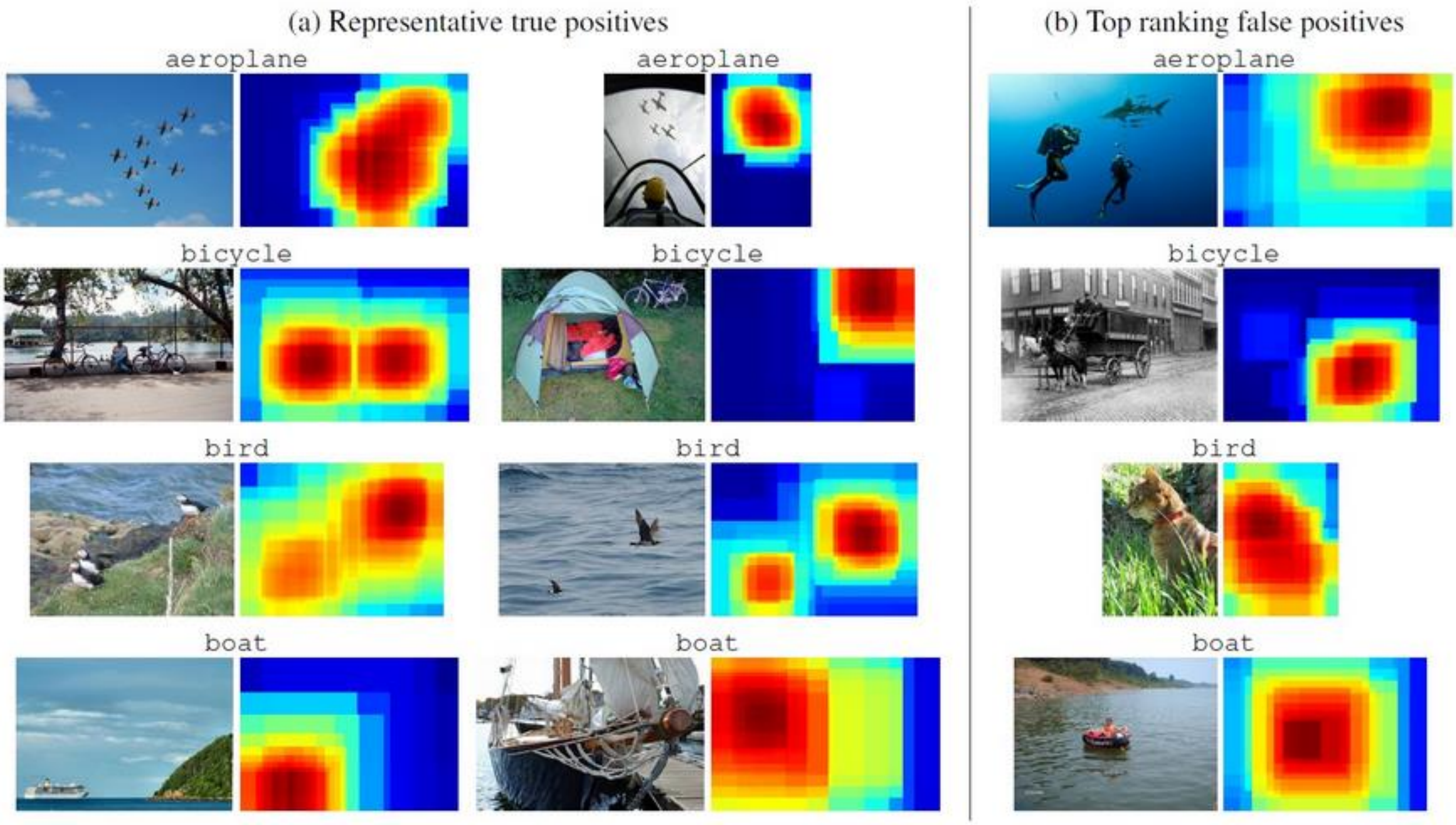
Results

Pascal VOC

Oquab, Bottou, Laptev and Sivic
CVPR 2014

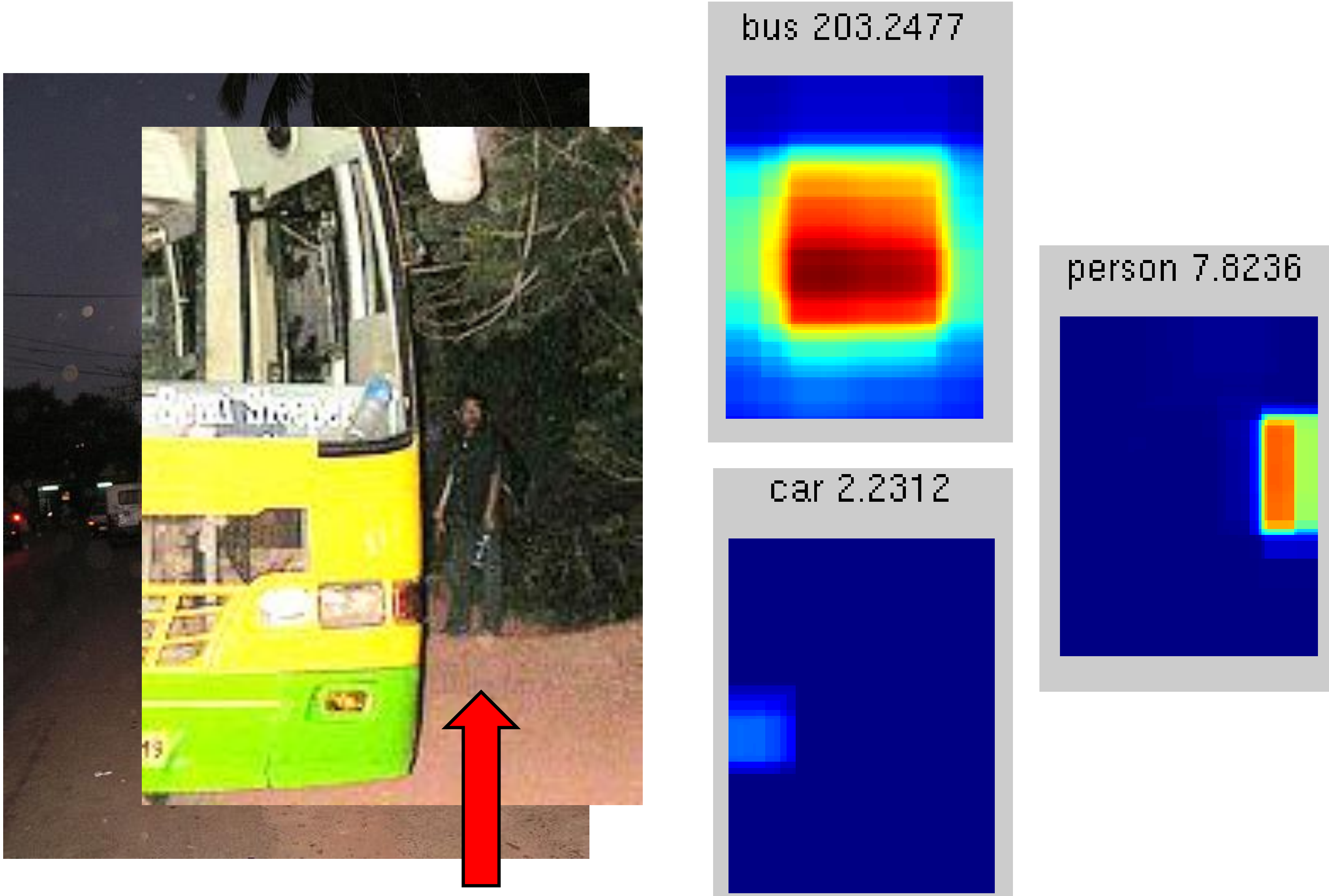


Results



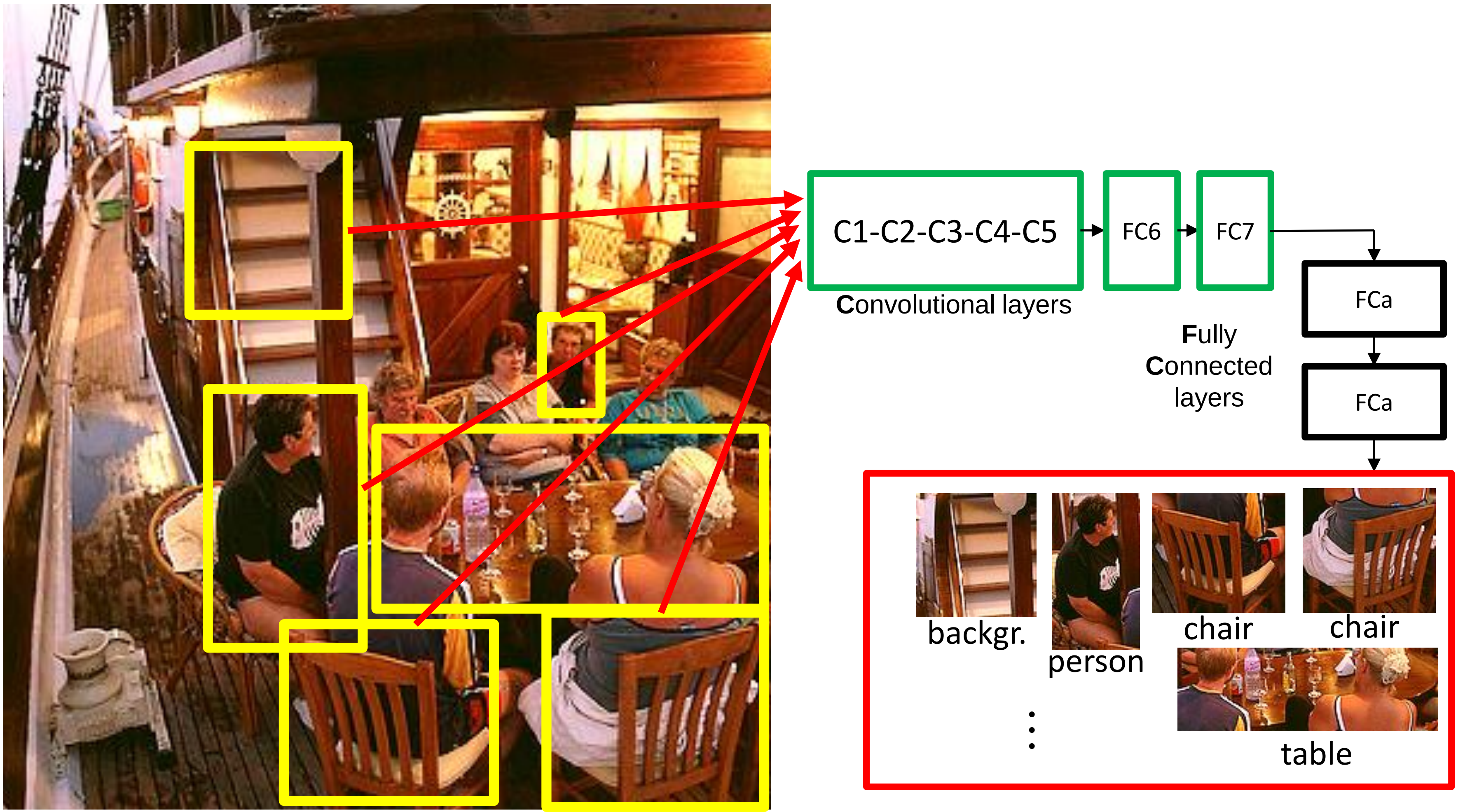
[Oquab, Bottou, Laptev and Sivic, CVPR 2014]

Results



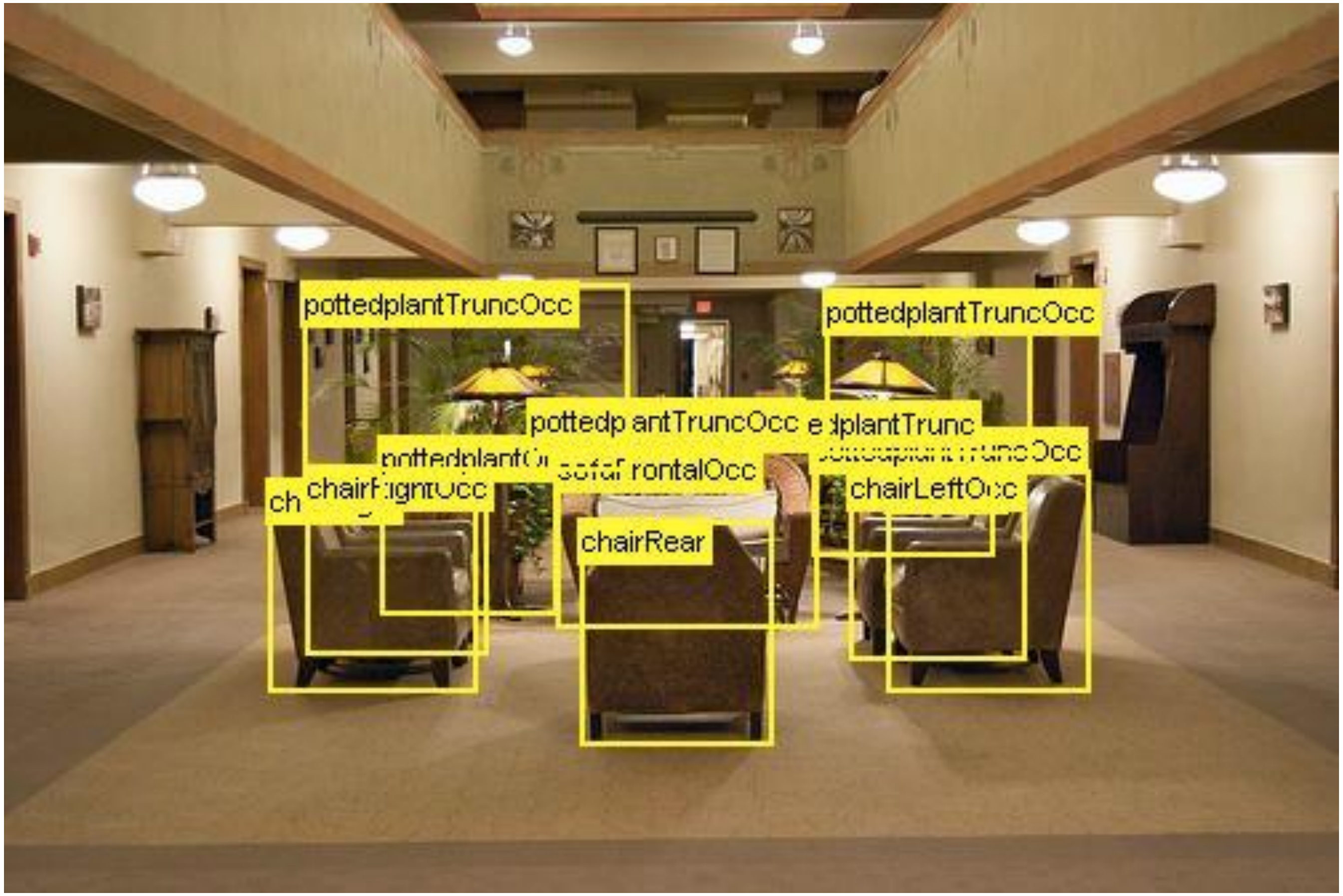
[Oquab, Bottou, Laptev and Sivic, CVPR 2014]

How to use CNNs for cluttered scenes?



Problem: Annotation of bounding boxes is (a): subjective (b): expensive

Motivation: labeling bounding boxes is tedious



Are bounding boxes needed for training CNNs?

Image-level labels: Bicycle, Person

Motivation: image-level labels are plentiful

"Beautiful red leaves in a back street of Freiburg"

[Kuznetsova et al., ACL 2013]
http://www.cs.stonybrook.edu/~pkuznetsova/imgcaption/captions1K.html

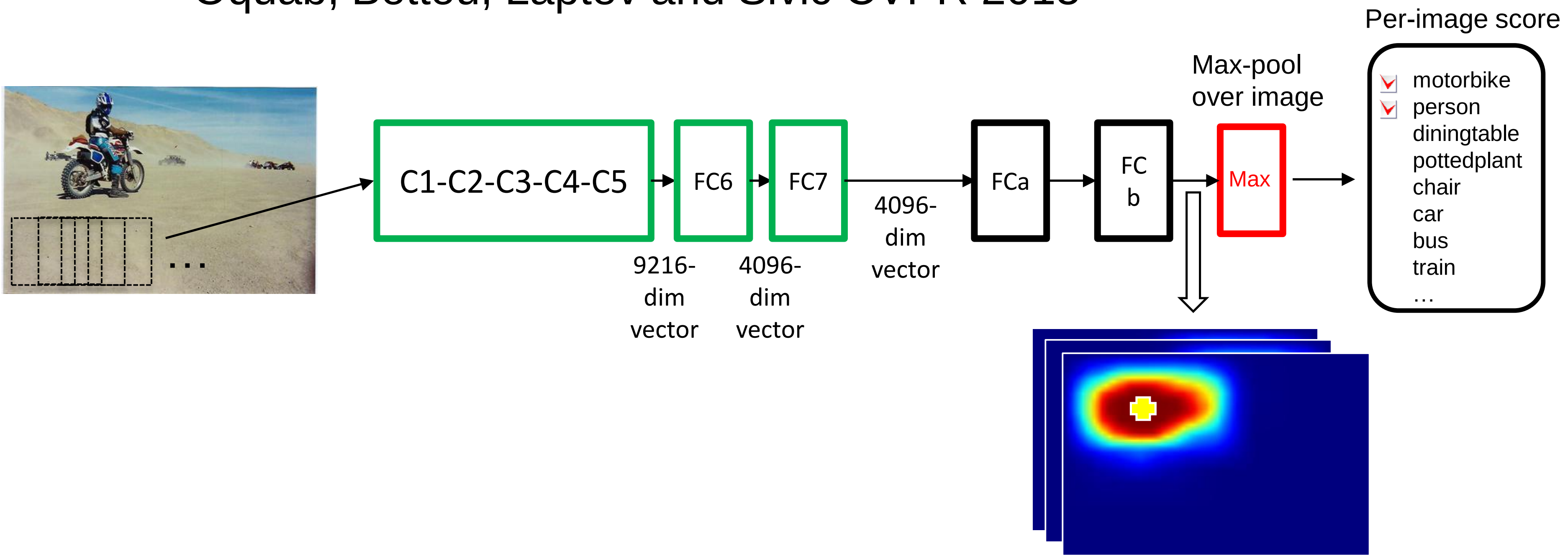
Motivation: image-level labels are plentiful

"Public bikes in Warsaw during night"

https://www.flickr.com/photos/jacek_kadaj/8776008002/in/photostream/

Approach: search over object's location at the *training time*

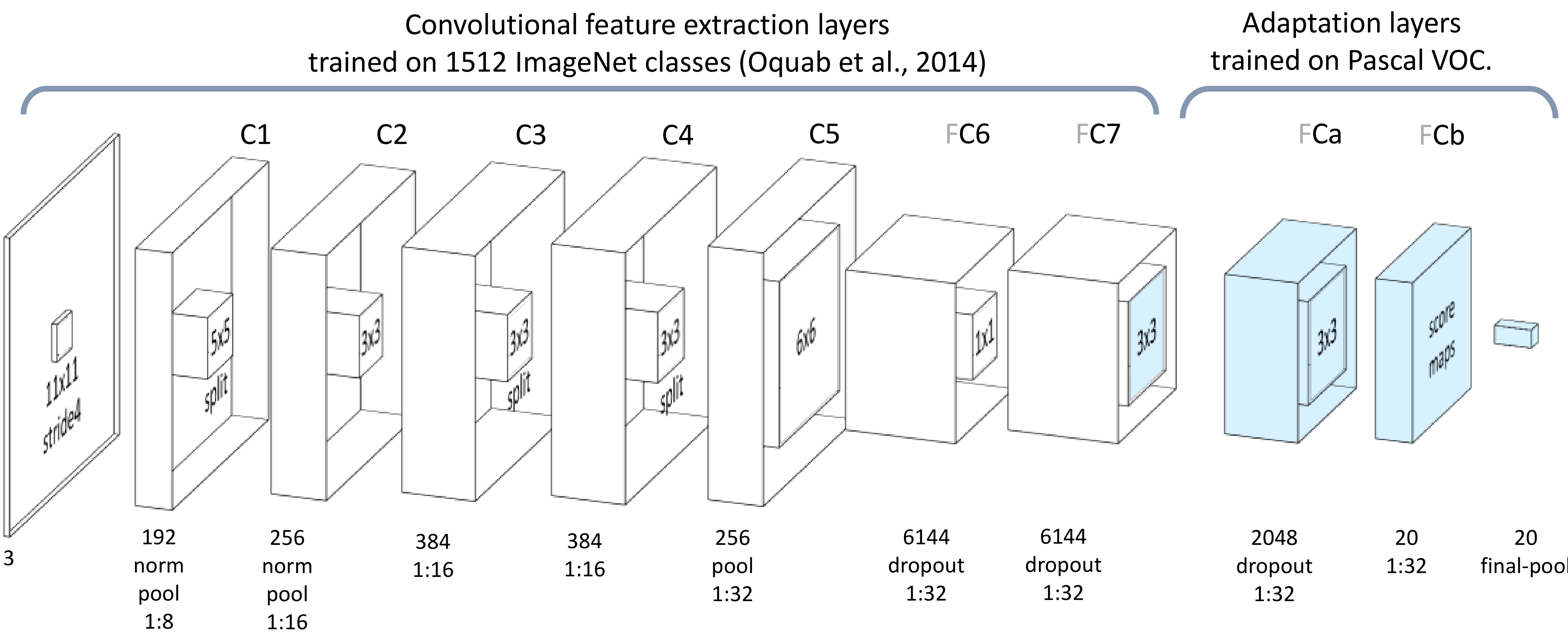
Oquab, Bottou, Laptev and Sivic CVPR 2015

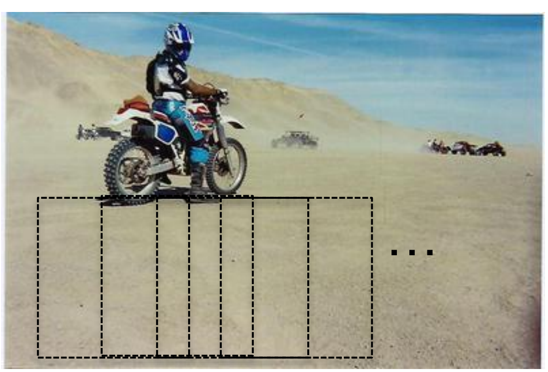


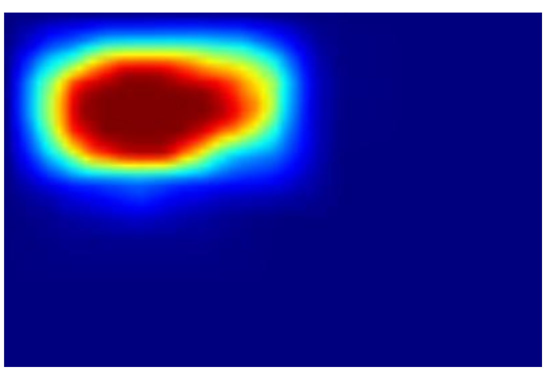
1. Efficient window sliding to find object location hypothesis
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

See also [Kokkinos et al. '15, Sermanet et al. '14, Chaftield et al.'14]

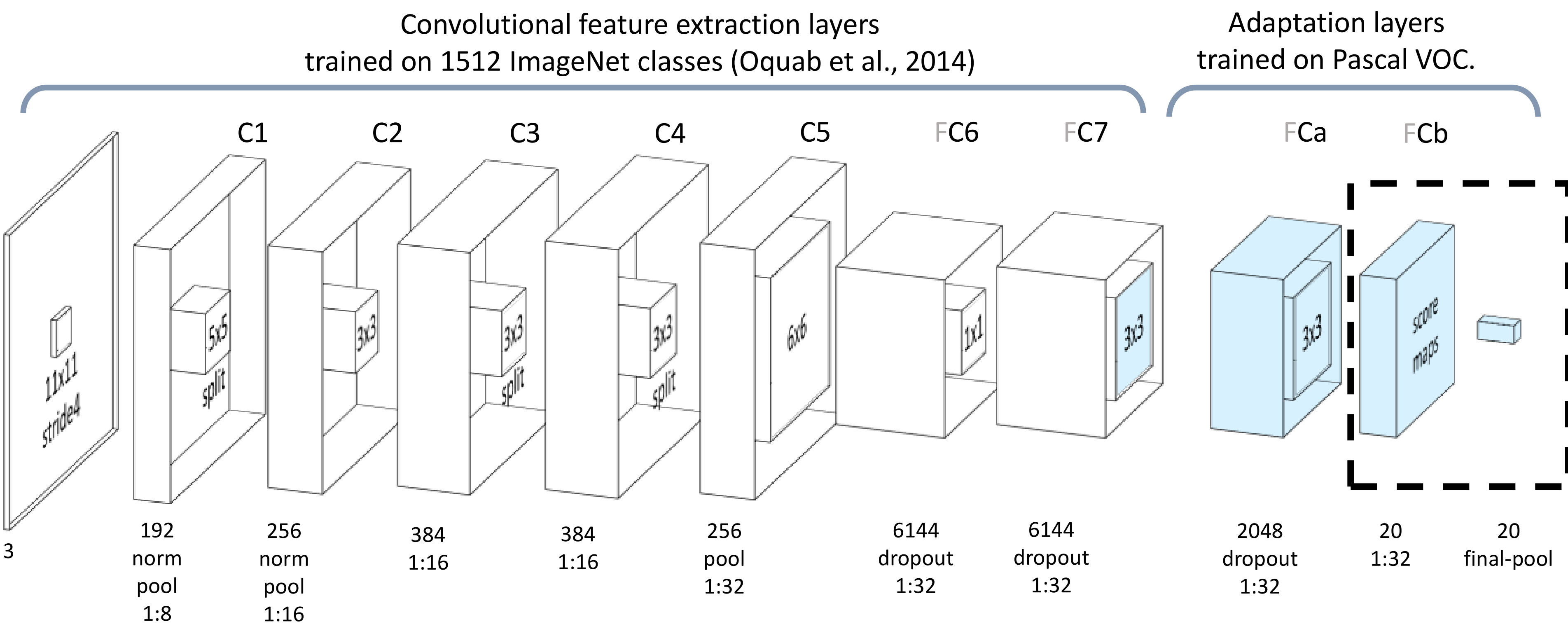
1. Efficient window sliding to find object location

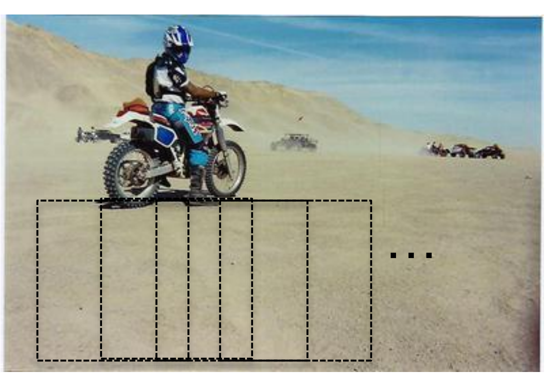


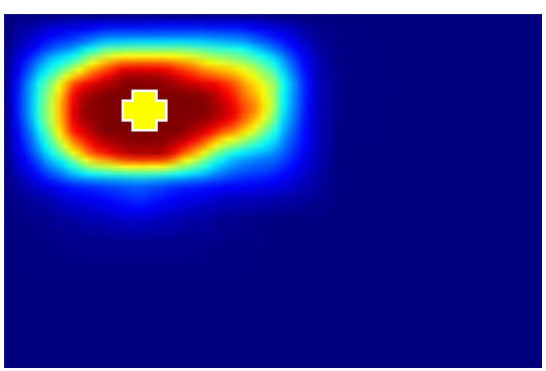




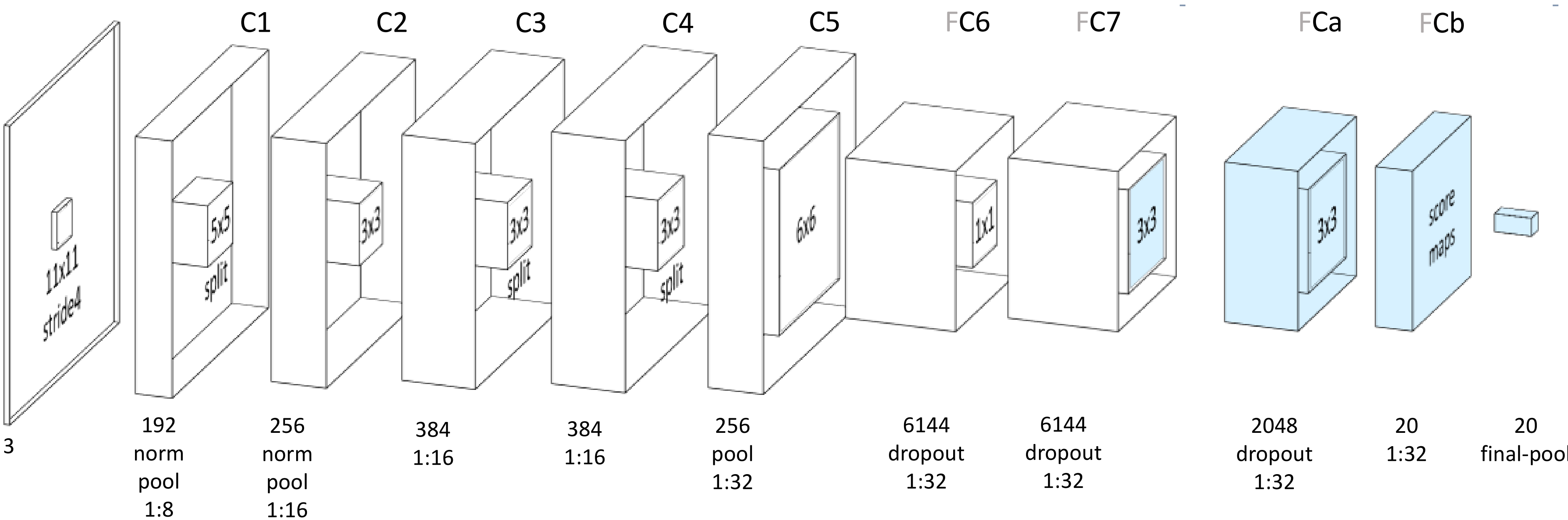
2. Image-level aggregation using global max-pool







3. Multi-label loss function (to allow for multiple objects in image)



Cost function: Sum of log-loss functions over K classes:

$$\ell(f(\mathbf{x}), y) = \sum_k \log(1 + e^{-y_k f_k(\mathbf{x})})$$

Training with global max-pooling

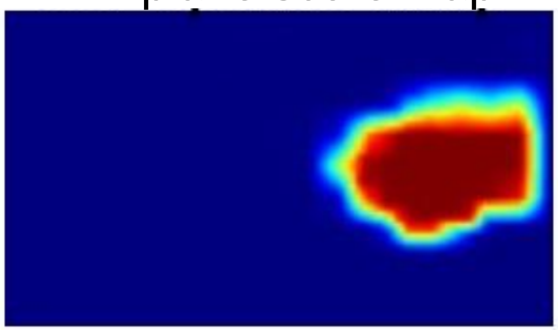
Training input:

+

image-level labels:

- ✓ Airplane
- × Car
- × Chair …

Airplane score map

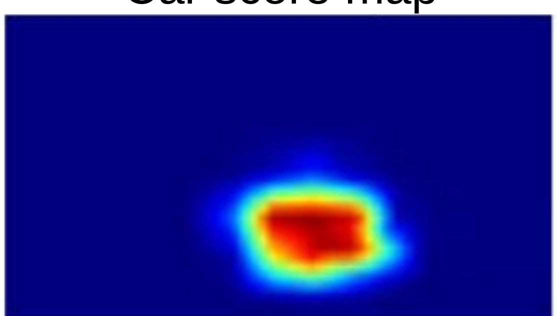


max-pool →

Correct label:
increase score

Learn discriminative
object parts

Car score map



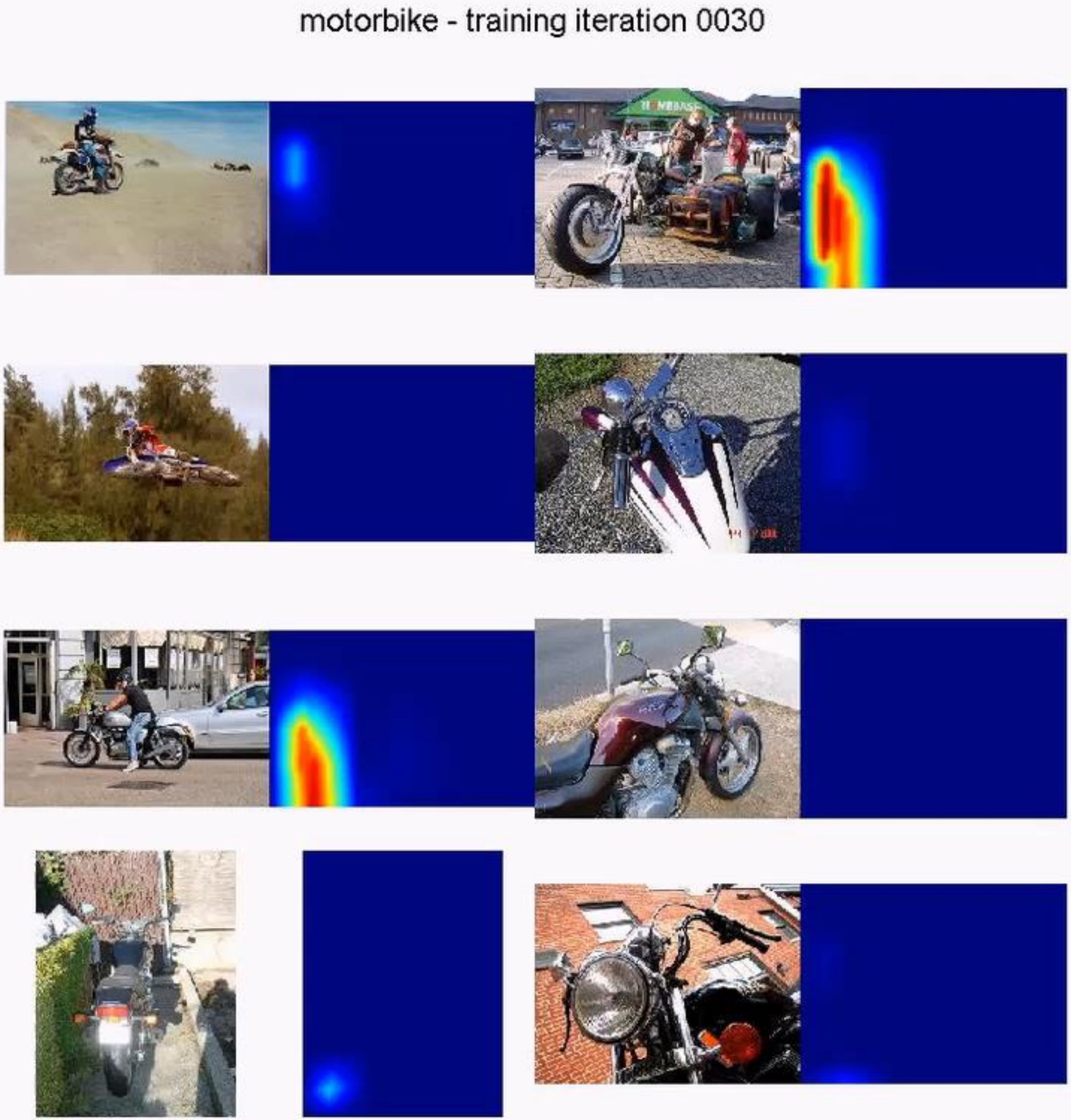
max-pool →

Incorrect label:
decrease score

Suppress *Hard Negatives*

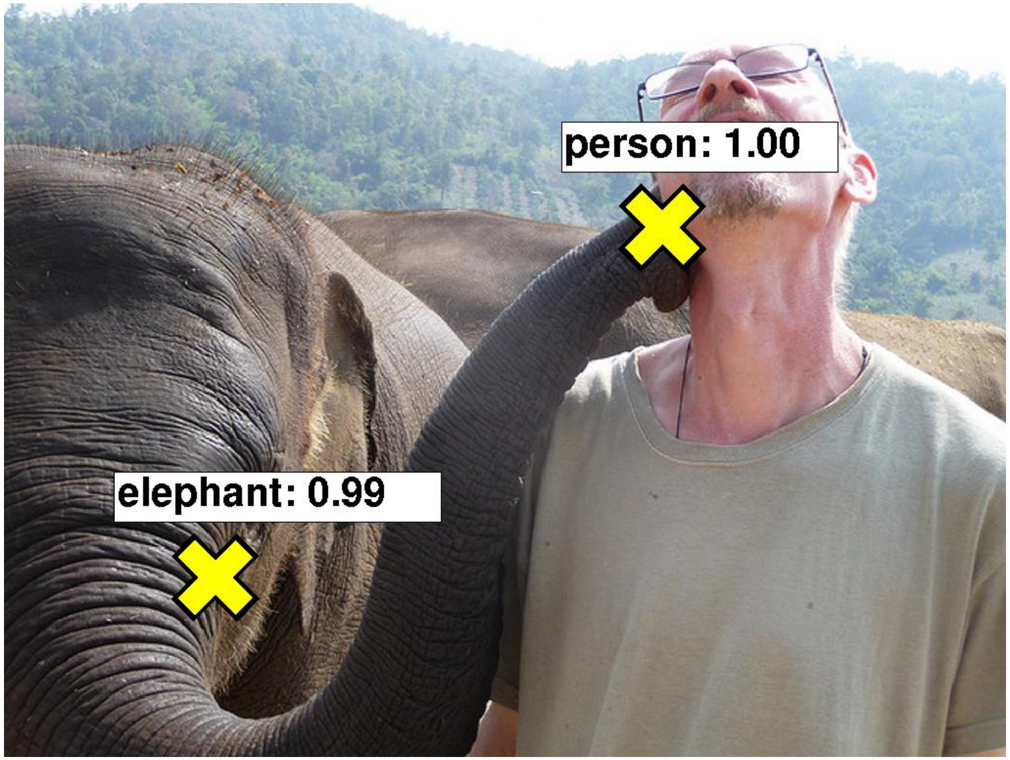
Training Motorbikes

Evolution of localization score maps over training epochs

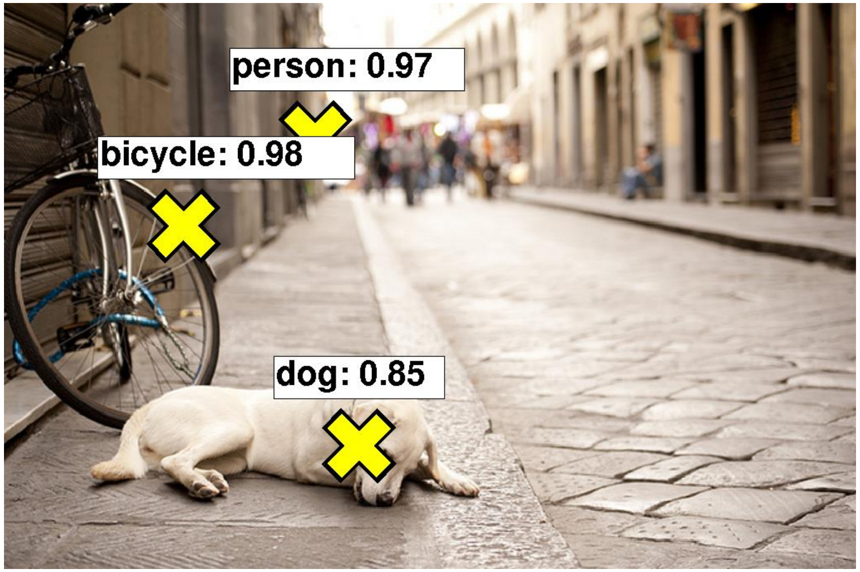


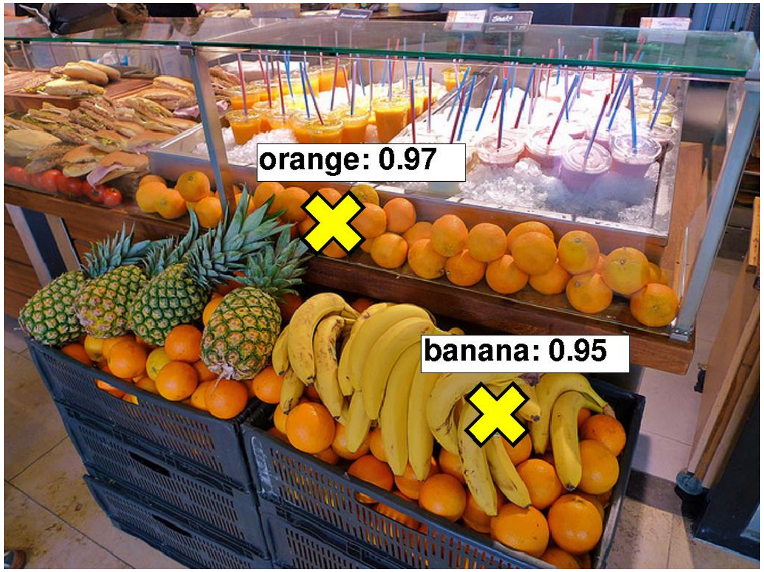
Results for weakly-supervised object recognition in Microsoft COCO dataset

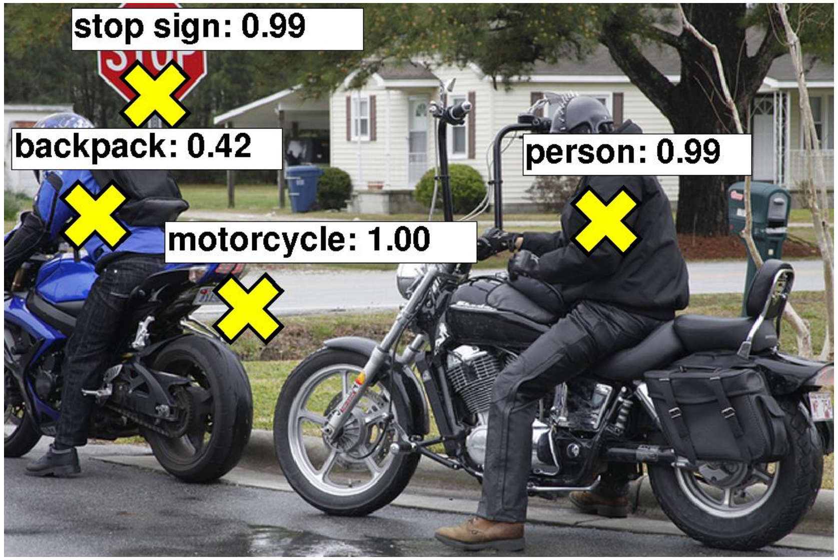
Test results in Microsoft COCO: **80 object classes**



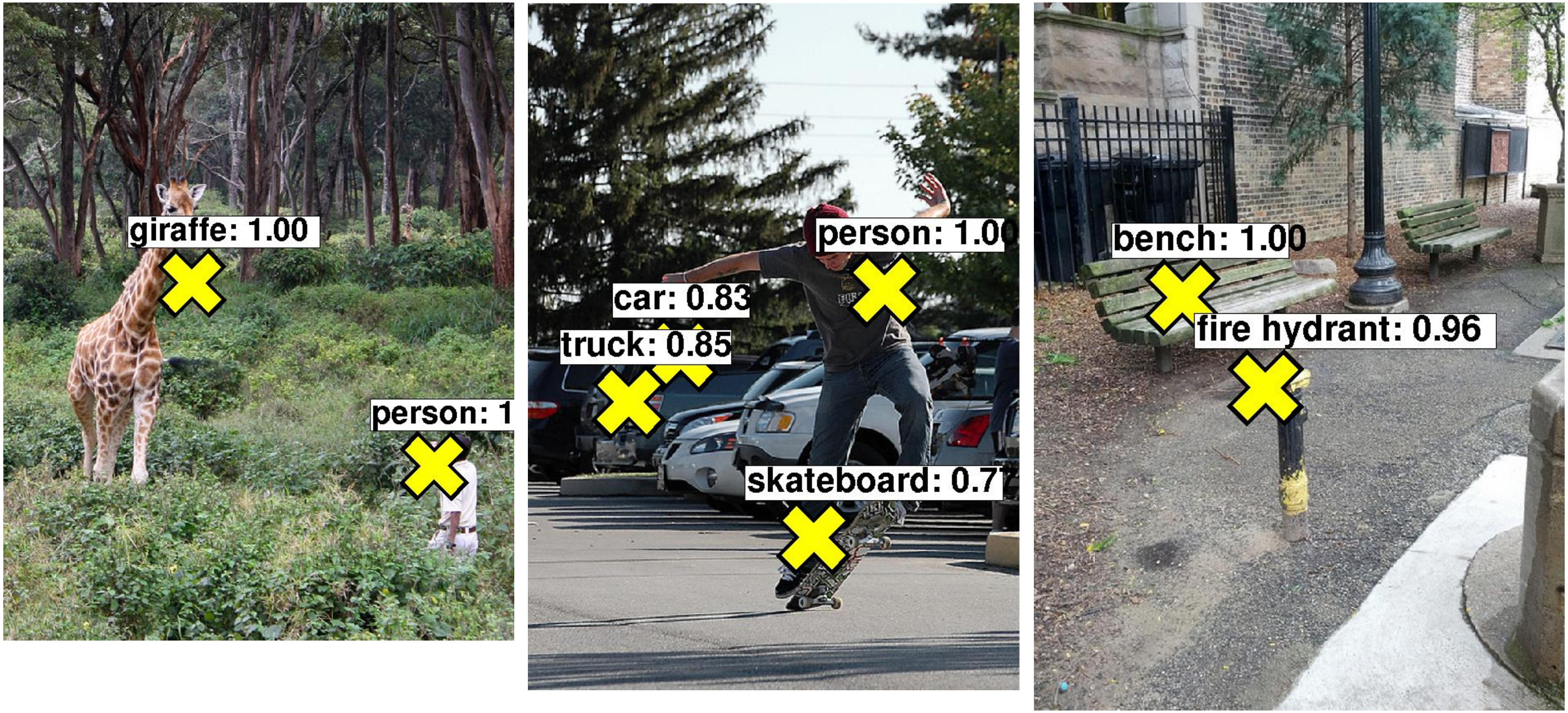
Test results in Microsoft COCO: **80 object classes**



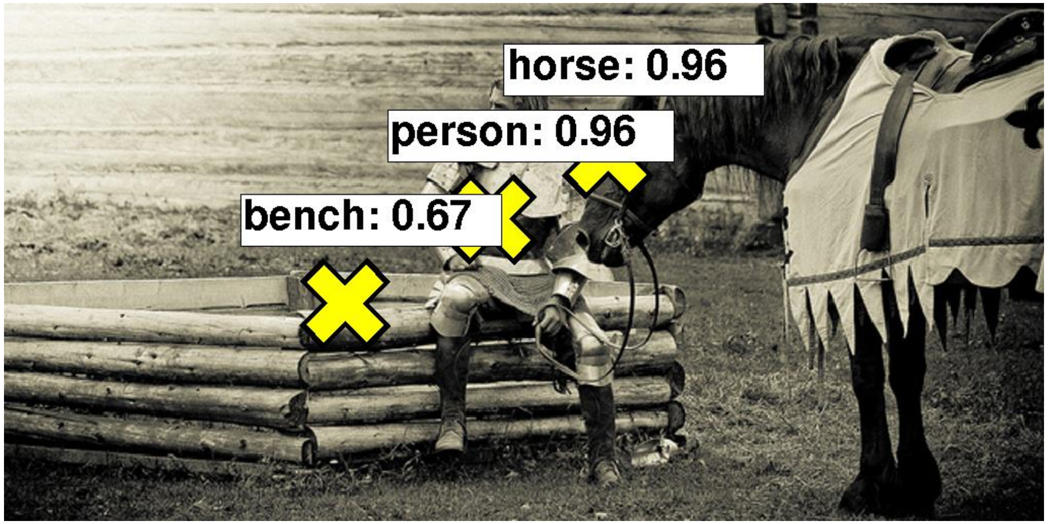


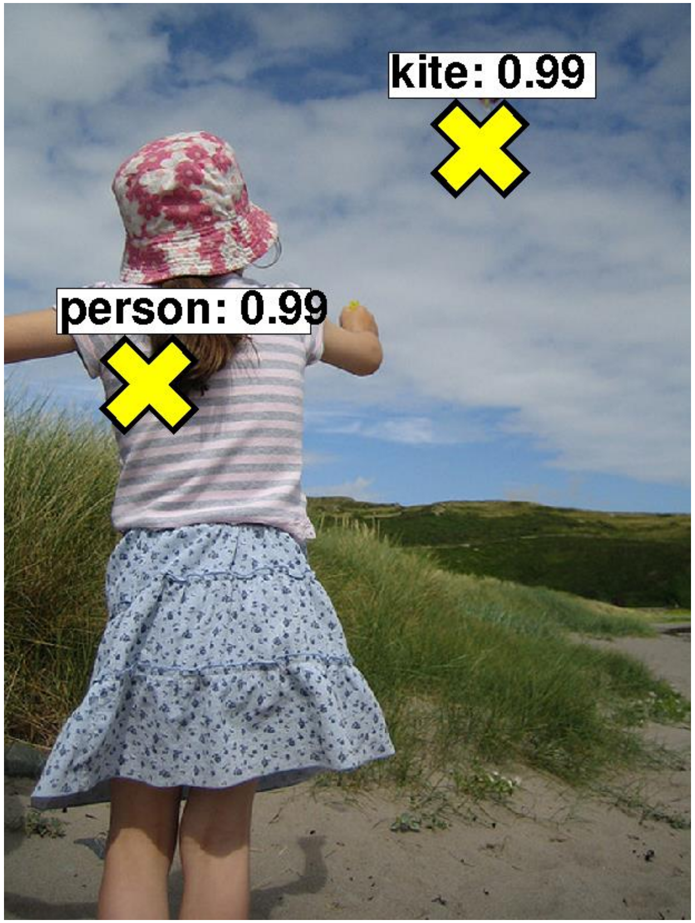


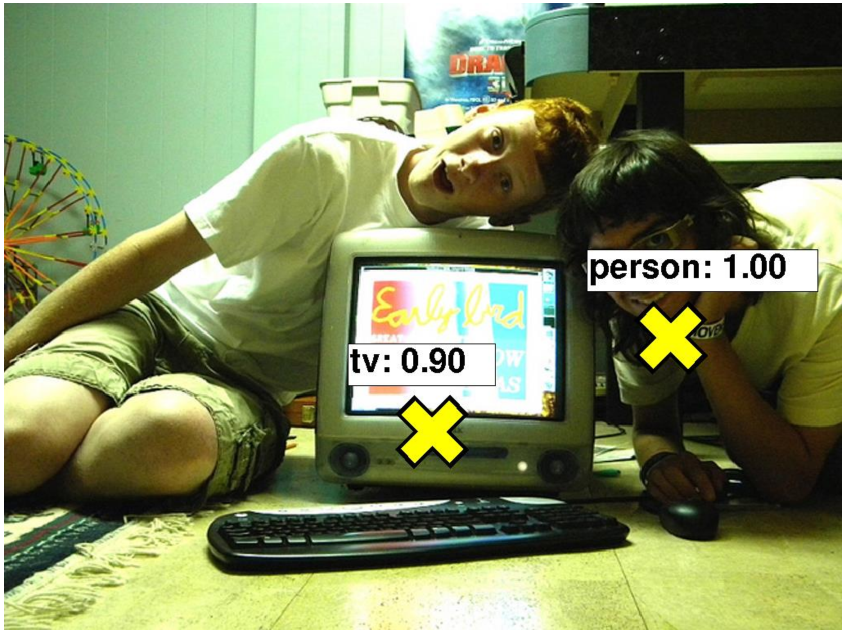
Test results in Microsoft COCO: **80 object classes**



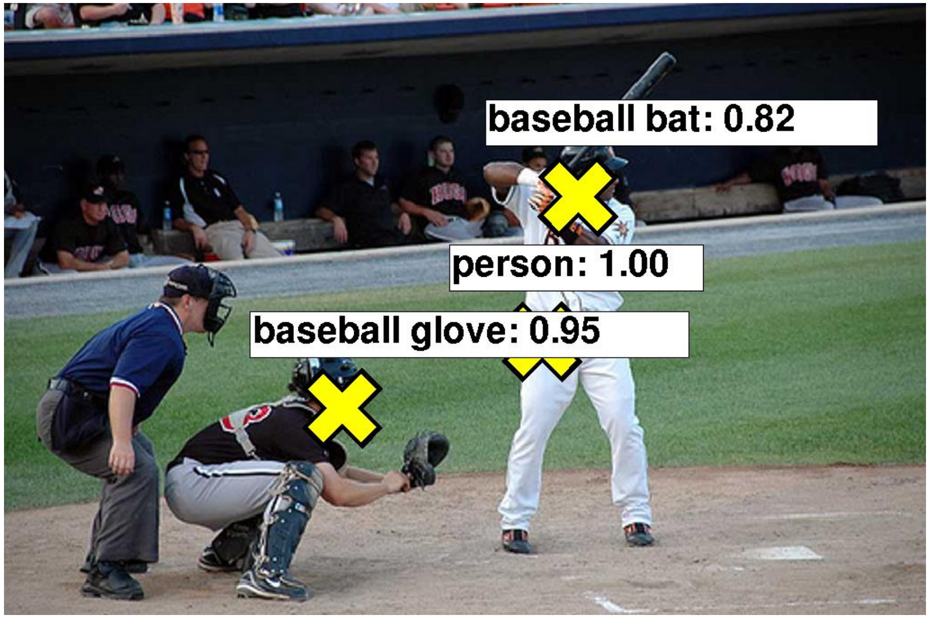
Test results in Microsoft COCO: **80 object classes**

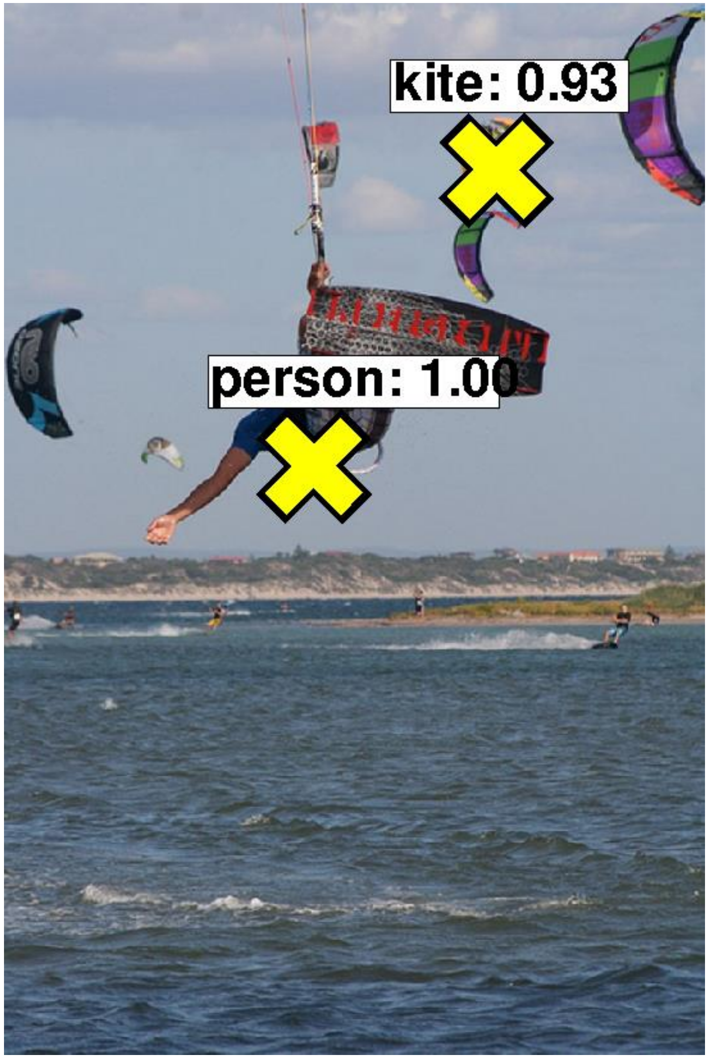


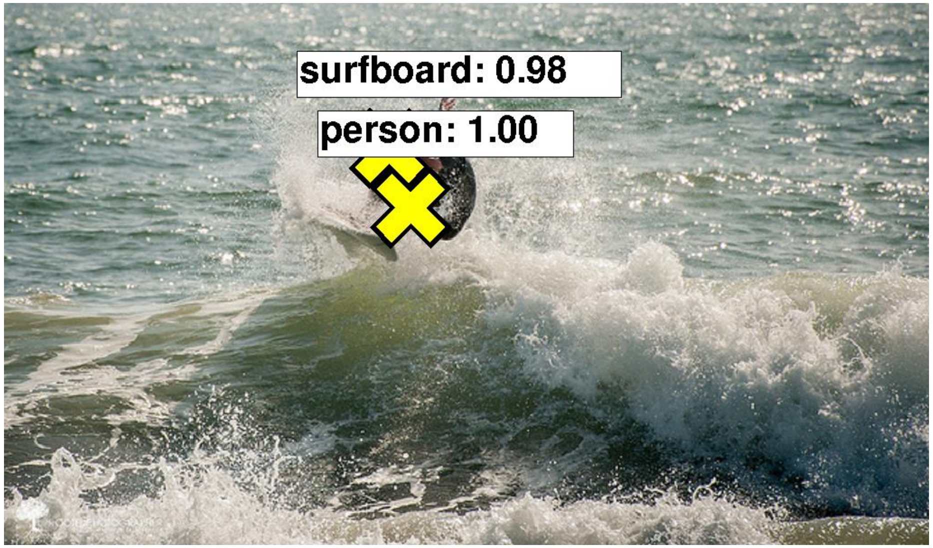




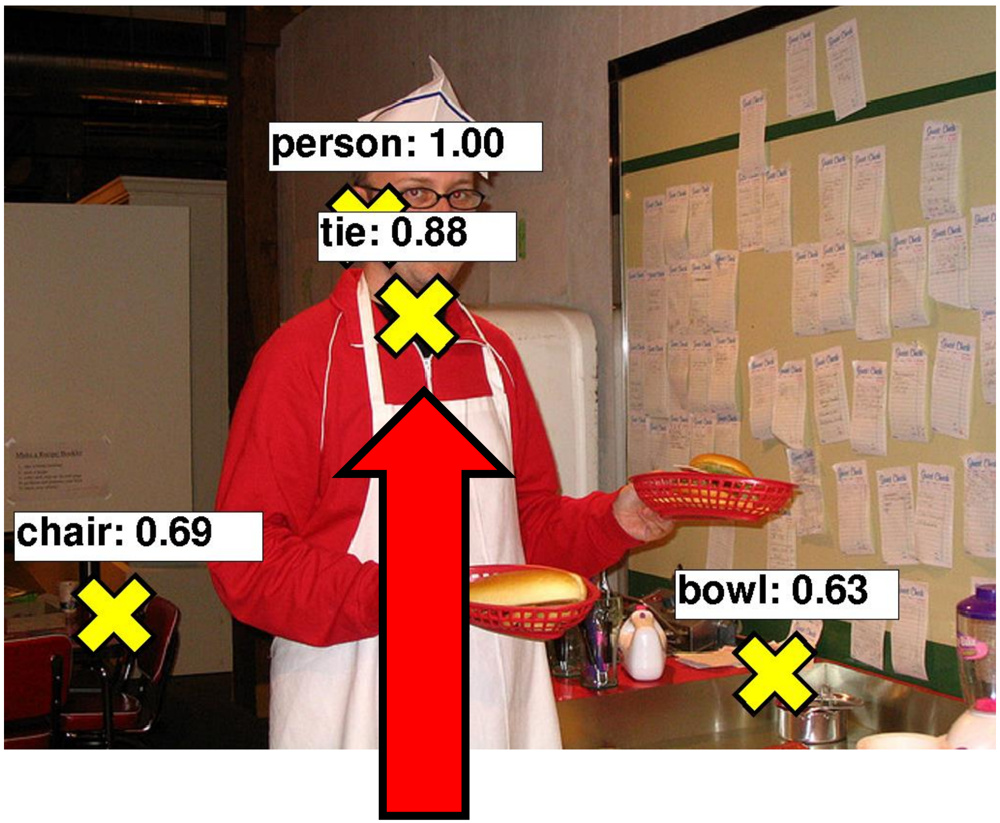
Test results in Microsoft COCO: **80 object classes**

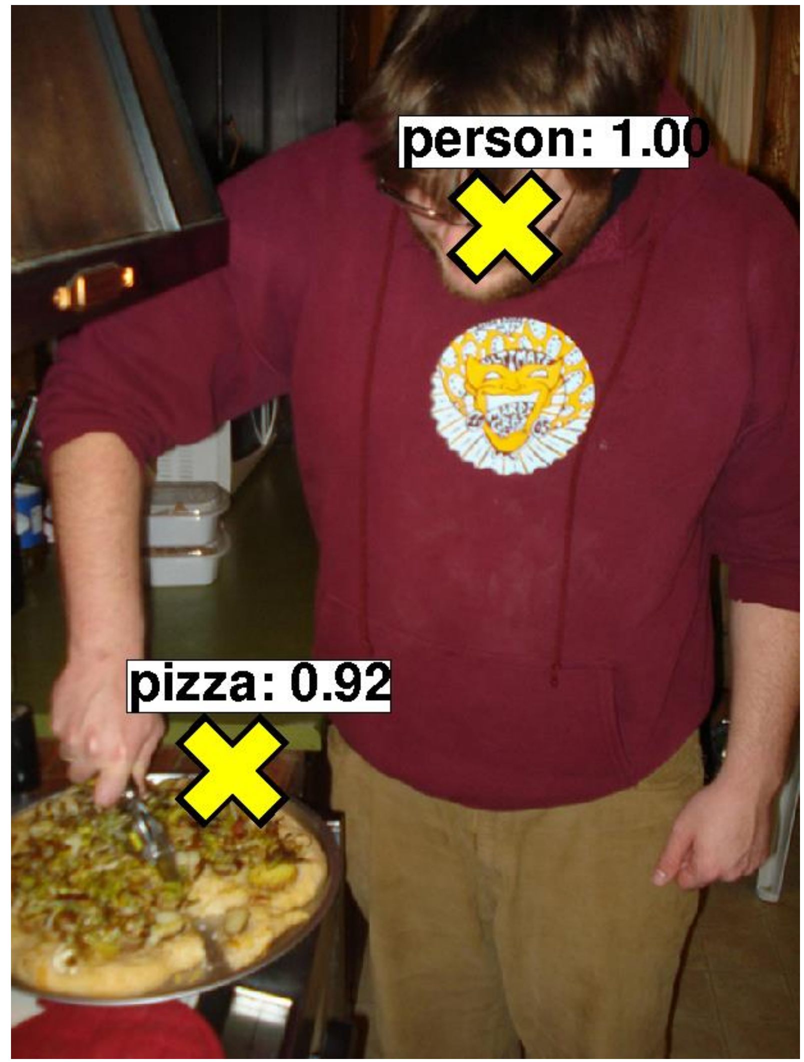






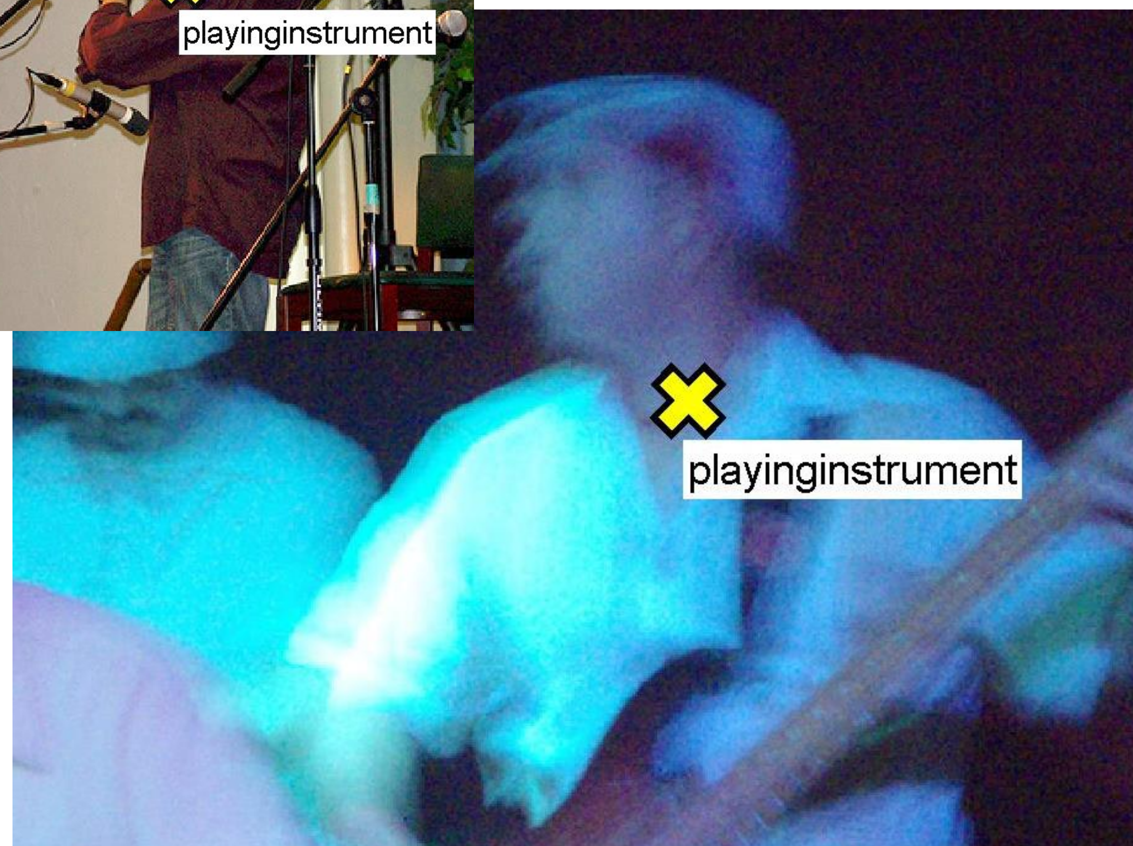
Test results in Microsoft COCO: **80 object classes**



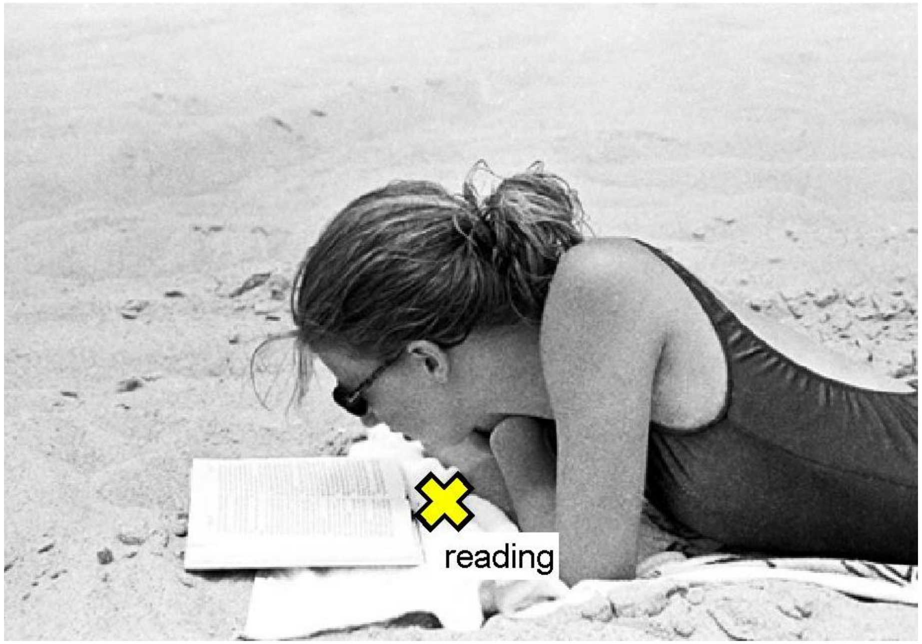


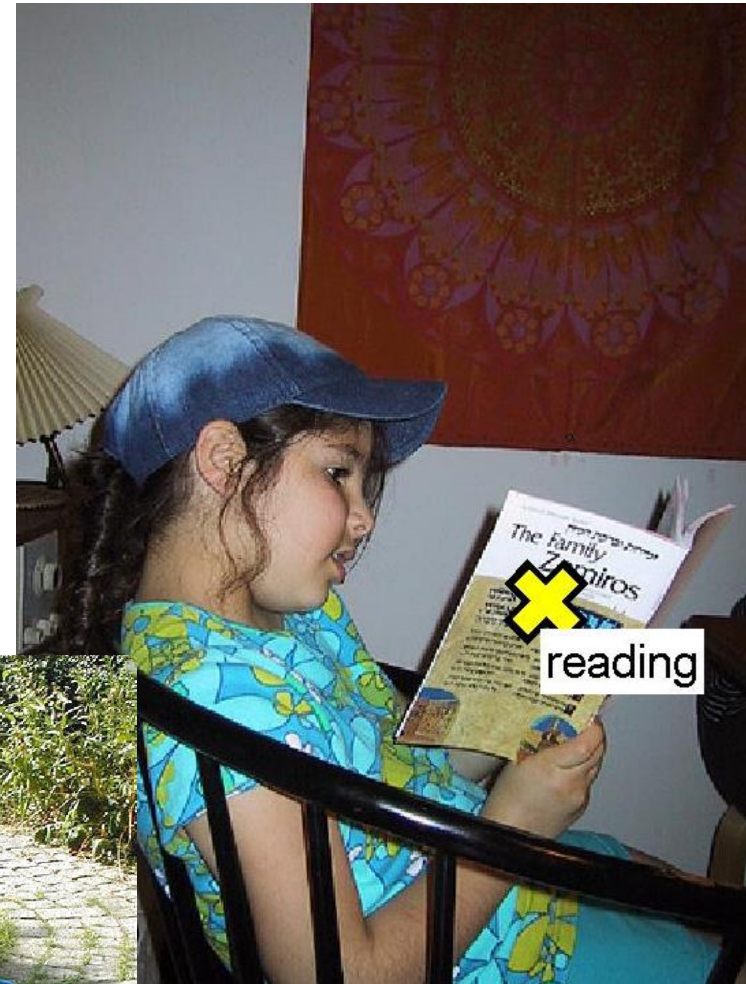
Results for weakly-supervised *action* recognition in Pascal VOC'12 dataset

Test results in Pascal VOC'12: **10 action classes**



Test results in Pascal VOC'12: **10 action classes**





Test results in Pascal VOC'12: **10 action classes**

Test results in Pascal VOC’12: **10 action classes**

Test results in Pascal VOC'12: **10 action classes**

Test results in Pascal VOC'12: **10 action classes**

Failure cases

Results PASCAL VOC 2012

Object classification

Object-level sup.	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	**mAP**
A.NUS-SCM [43]	97.3	84.2	80.8	85.3	60.8	89.9	86.8	89.3	**75.4**	77.8	75.1	83.0	87.5	90.1	95.0	57.8	79.2	**73.4**	**94.5**	80.7	82.2
B.OQUAB [31]	94.6	82.9	88.2	84.1	60.3	89.0	84.4	90.7	72.1	86.8	69.0	92.1	93.4	88.6	96.1	64.3	86.6	62.3	91.1	79.8	82.8

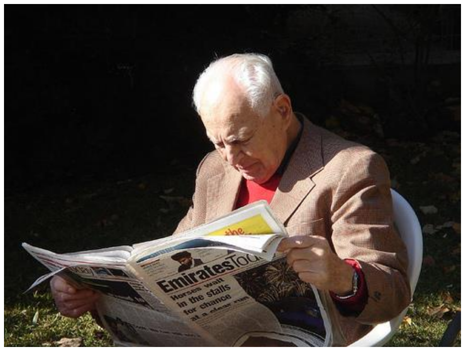
Image-level sup.	plane	bike	bird	boat	btl	bus	car	cat	chair	cow	table	dog	horse	moto	pers	plant	sheep	sofa	train	tv	**mAP**
C.Z&F [51]	96.0	77.1	88.4	85.5	55.8	85.8	78.6	91.2	65.0	74.4	67.7	87.8	86.0	85.1	90.9	52.2	83.6	61.1	91.8	76.1	79.0
D.CHATFIELD [4]	96.8	82.5	91.5	88.1	62.1	88.3	81.9	**94.8**	70.3	80.2	76.2	92.9	90.3	89.3	95.2	57.4	83.6	66.4	93.5	81.9	83.2
E.NUS-HCP [47]	**97.5**	84.3	**93.0**	**89.4**	62.5	90.2	84.6	**94.8**	69.7	**90.2**	74.1	93.4	93.7	88.8	93.2	59.7	90.3	61.8	94.4	78.0	84.2
F.FULL IMAGES	95.3	77.4	85.6	83.1	49.9	86.7	77.7	87.2	67.1	79.4	73.5	85.3	90.3	85.6	92.7	47.8	81.5	63.4	91.4	74.1	78.7
G.WEAK SUP	96.7	**88.8**	92.0	87.4	**64.7**	**91.1**	**87.4**	94.4	74.9	89.2	**76.3**	**93.7**	**95.2**	**91.1**	**97.6**	**66.2**	**91.2**	70.0	**94.5**	**83.7**	**86.3**

VGG 89.3

[Oquab, Bottou, Laptev and Sivic, CVPR 2015]

Summary

Training input



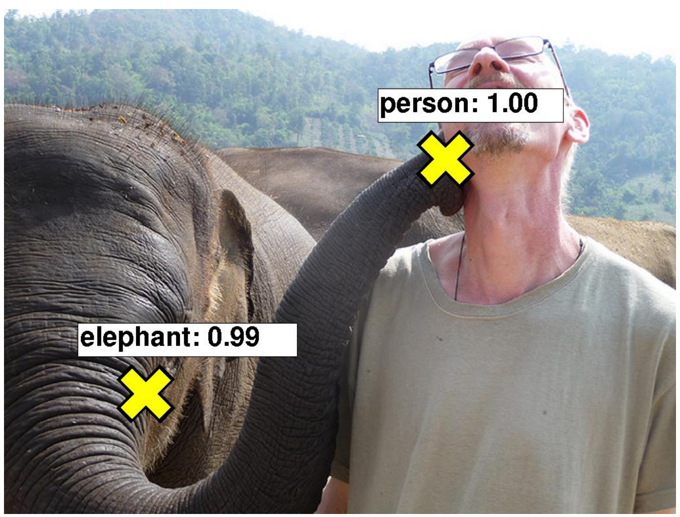
+

image-level labels:

- ✓ Person
- ✓ Chair
- ✗ Airplane
- …

- ✓ Reading
- ✗ Riding bike
- ✗ Running
- …

Test output



More details in http://www.di.ens.fr/willow/research/weakcnn/

What's next?

a dog **sitting beside** a red fire hydrant in a dog park.

a dog **holding** a skateboard trotting down a street.