See.4C Spatio-temporal Series Hackathon
February 14, 2017

Weakly supervised learning from images and video

Ivan Laptev

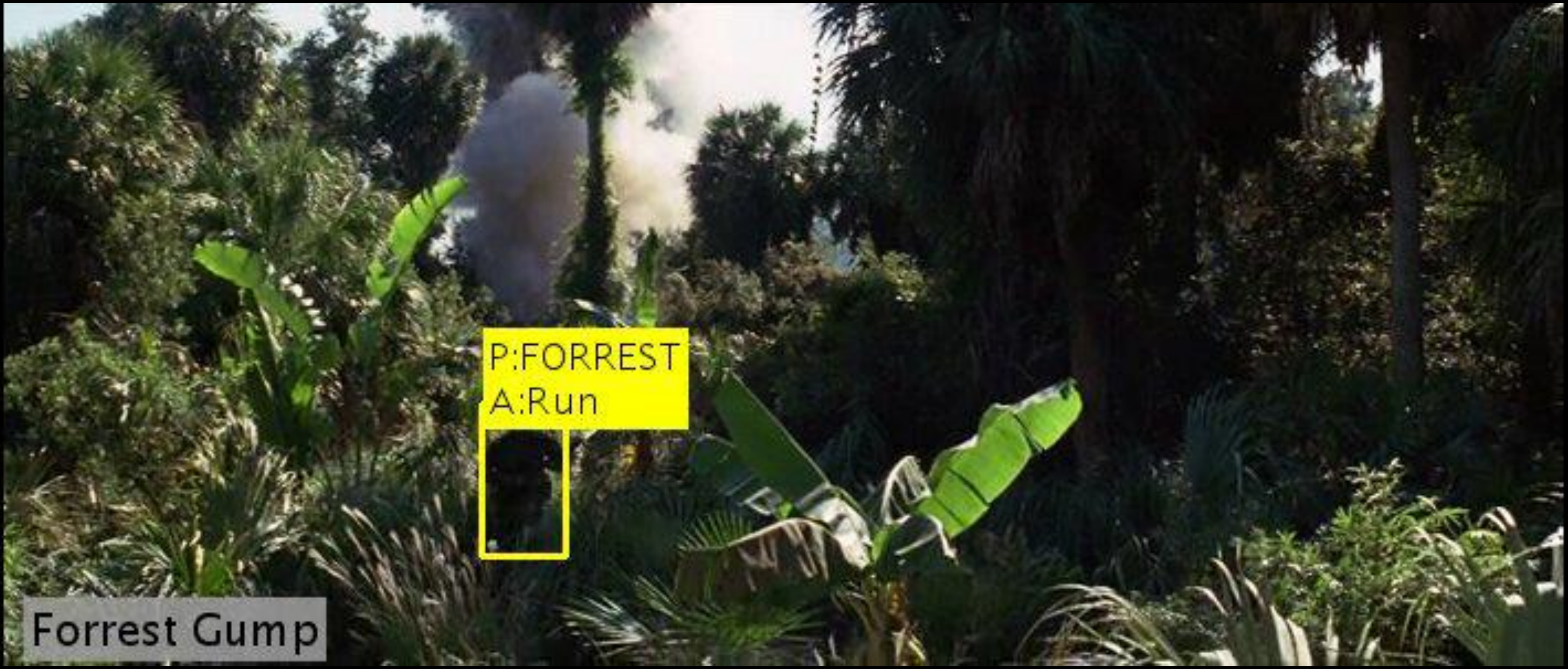
ivan.laptev@inria.fr

WILLOW, INRIA/ENS/CNRS, Paris

Joint work with: Maxime Oquab – Piotr Bojanowski – Rémi Lajugie –
Jean-Baptiste Alayrac – Leon Bottou – Francis Bach –
Simon Lacoste-Julien – Jean Ponce – Cordelia Schmid – Josef Sivic

What is Computer Vision?

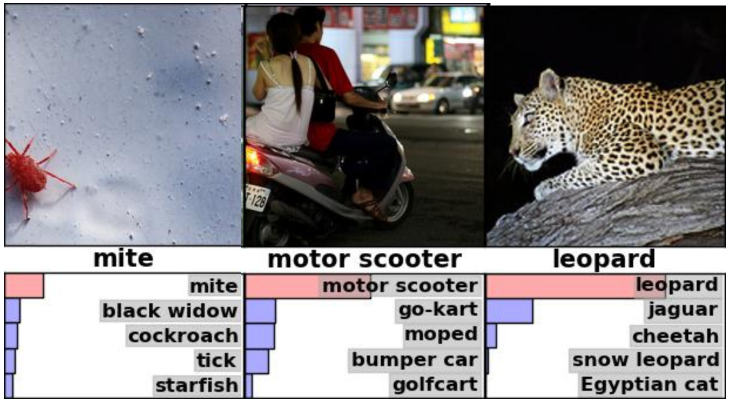
Computer vision works



Recent Progress: Convolutional Neural Networks

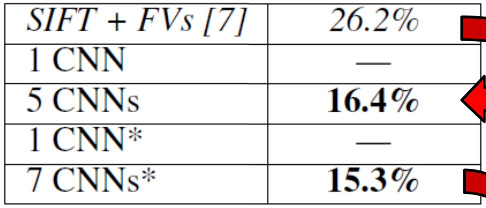
Object classification

ILSVRC'12: 1.2M images, 1K classes



Top 5 error:

2012:



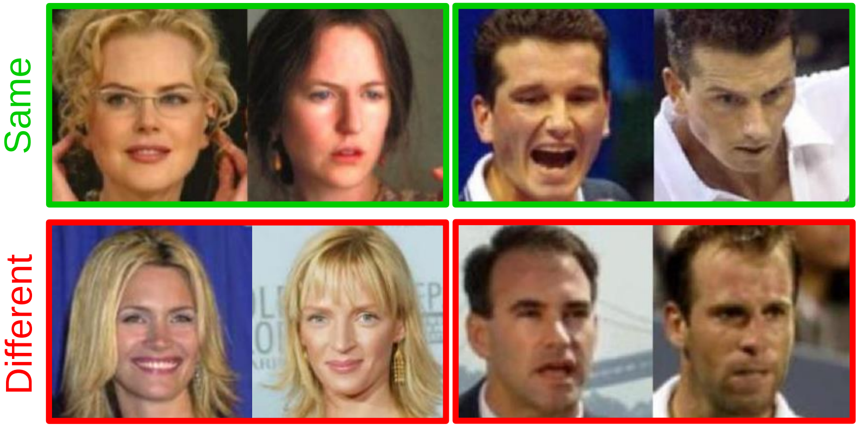
SIFT + FVs [7]	*26.2%*
1 CNN	—
5 CNNs	**16.4%**
1 CNN*	—
7 CNNs*	**15.3%**

2014-2015:

VGG:	6.8%
GoogLeNet:	6.6%
BAIDU	5.3%
Human	*5.1%*
ResNet	3.6%

Face Recognition

LFW



Accuracy:

--2013:

LBP	87.3%
FVF	93.0%

2014-2016:

DeepFace	97.3%
VGG	99.1%
Human	*99.2%*
VisionLabs	99.3%
FaceNet	99.6%
BAIDU	99.7%

How does it work?

AlexNet [Krizhevsky et al. 2012]
~60M parameters

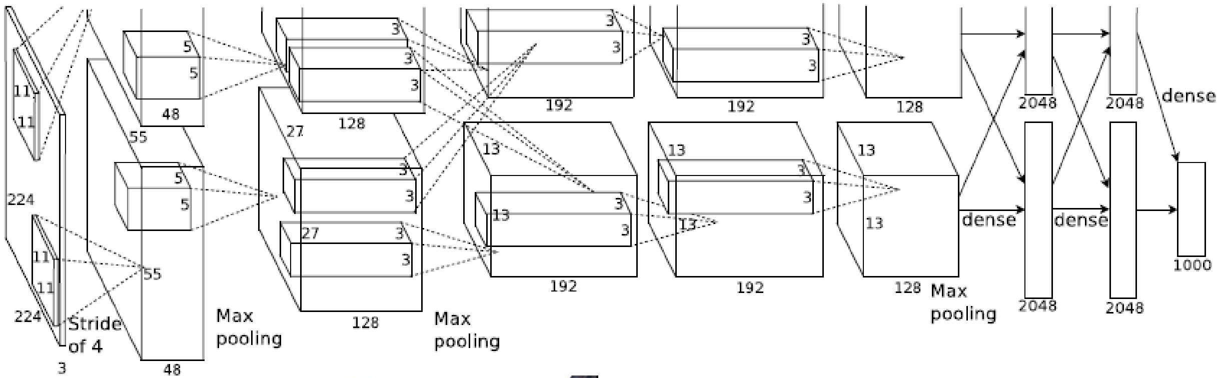
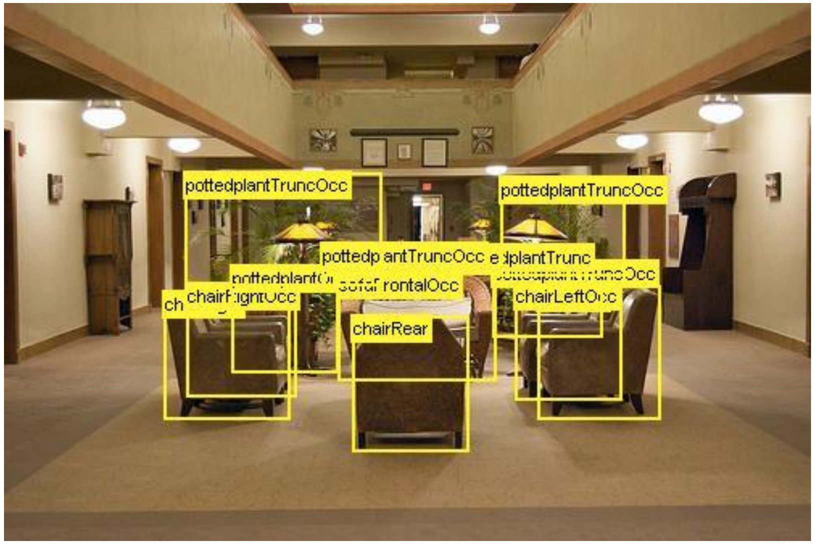


Image annotation

Problems with annotation

- Expensive

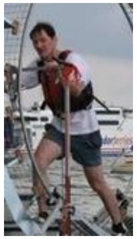


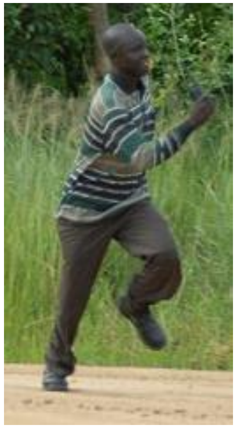
- Ambiguous

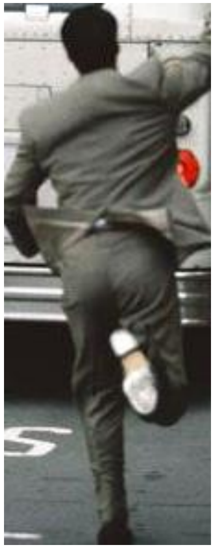
Table? Dining table? Desk? …

Problems with annotation

What action class?

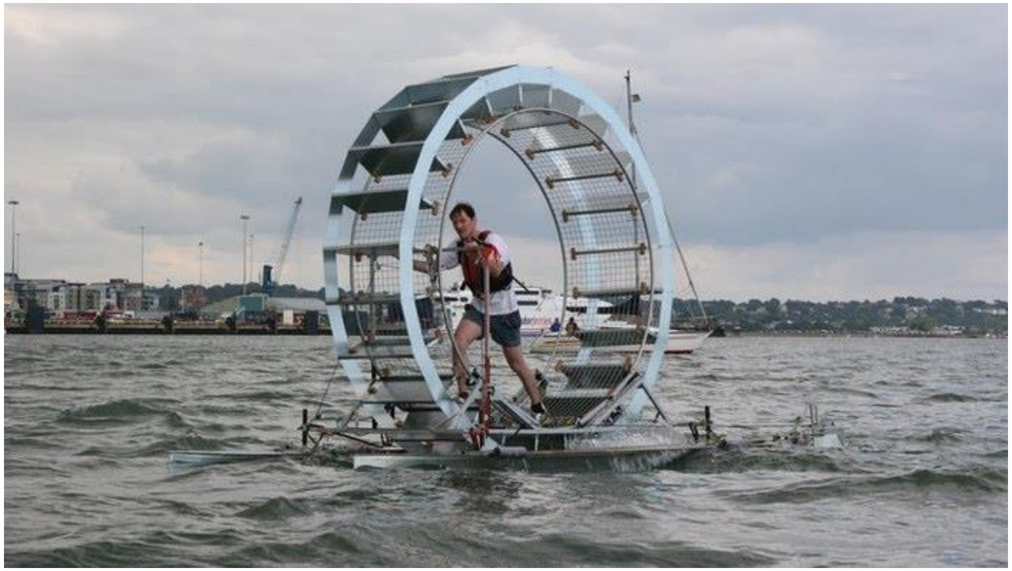


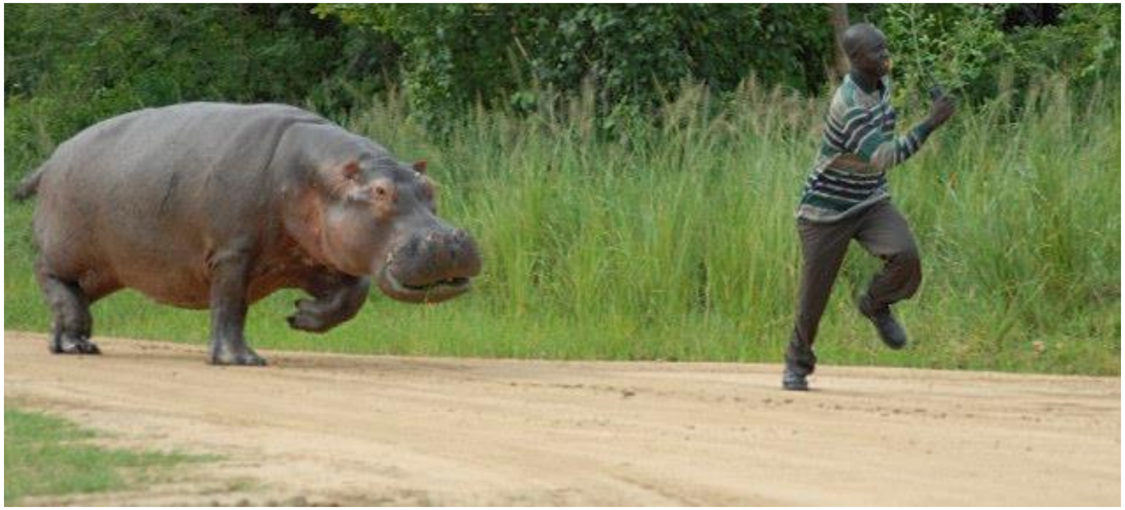


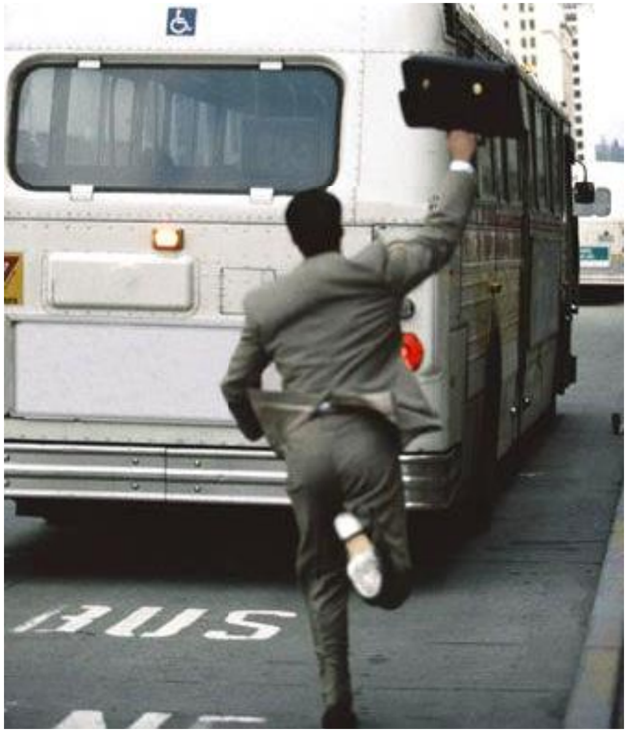


Problems with annotation

What action class?



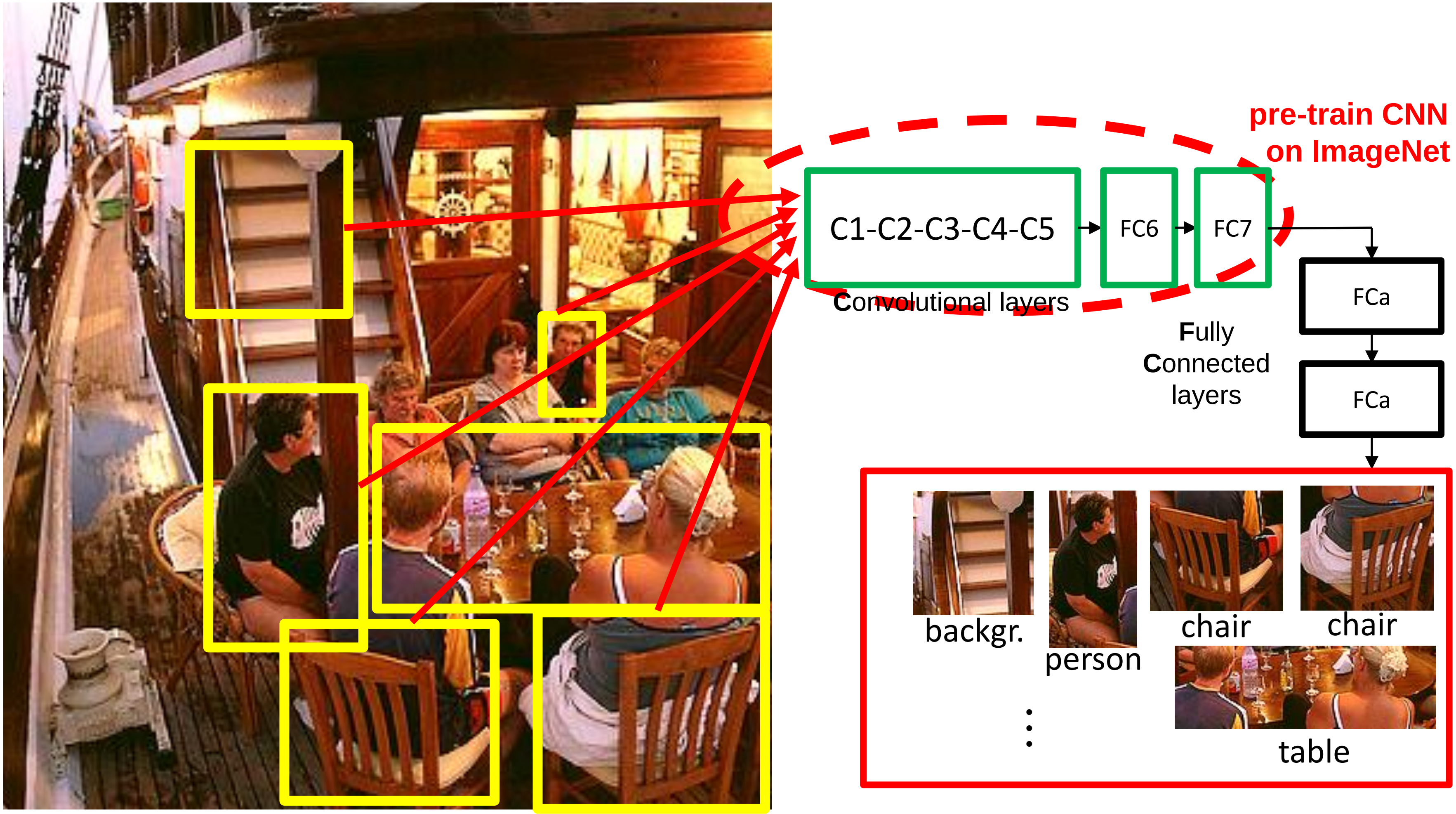




How to avoid manual supervision?

Weakly-supervised learning from images and video

Train CNNs for object detection

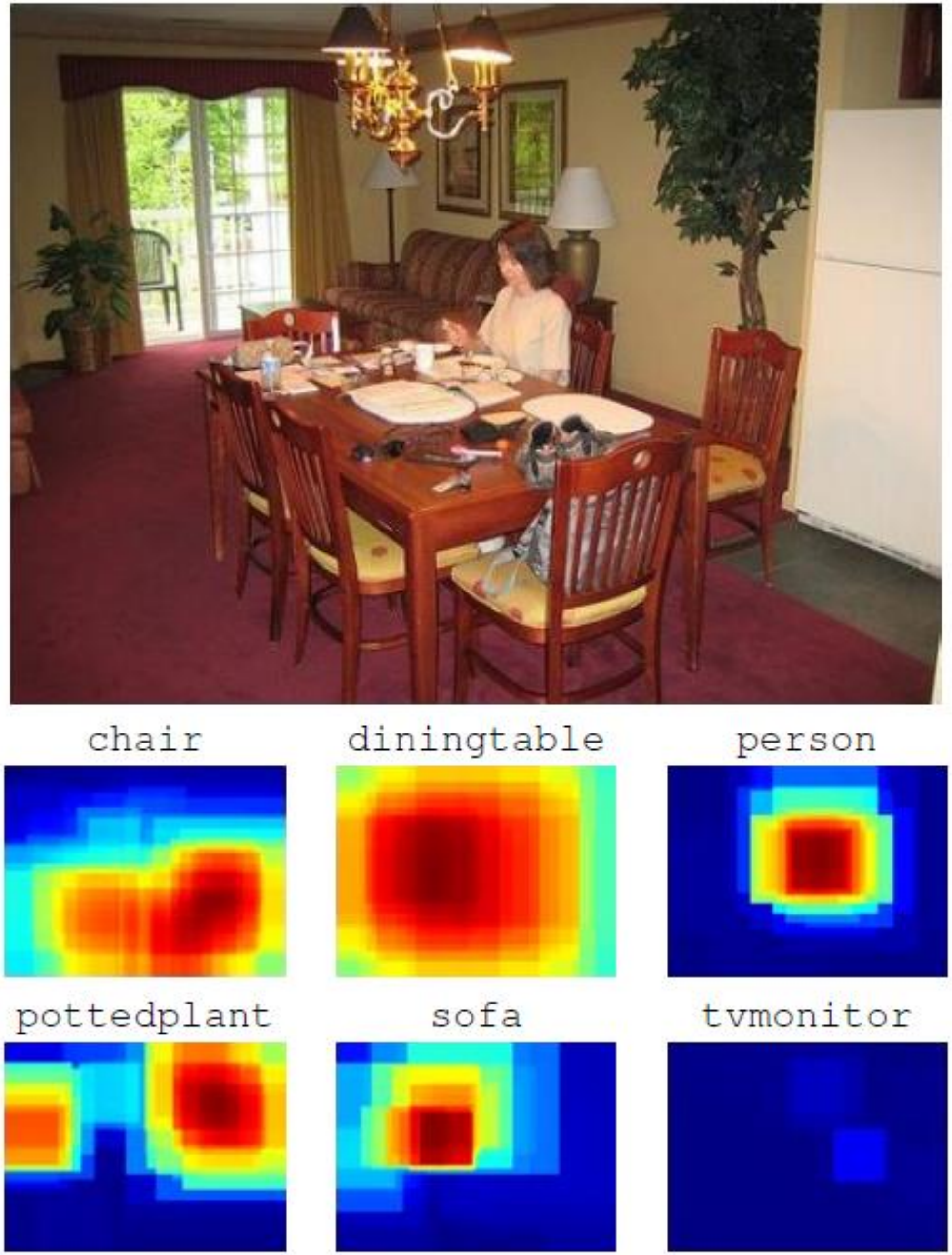


[Girshick'15], [Girshick et al.'14], [Oquab et al.'14], [Sermanet et al.'13], [Donahue et al. '13], [Zeiler & Fergus '13] ...

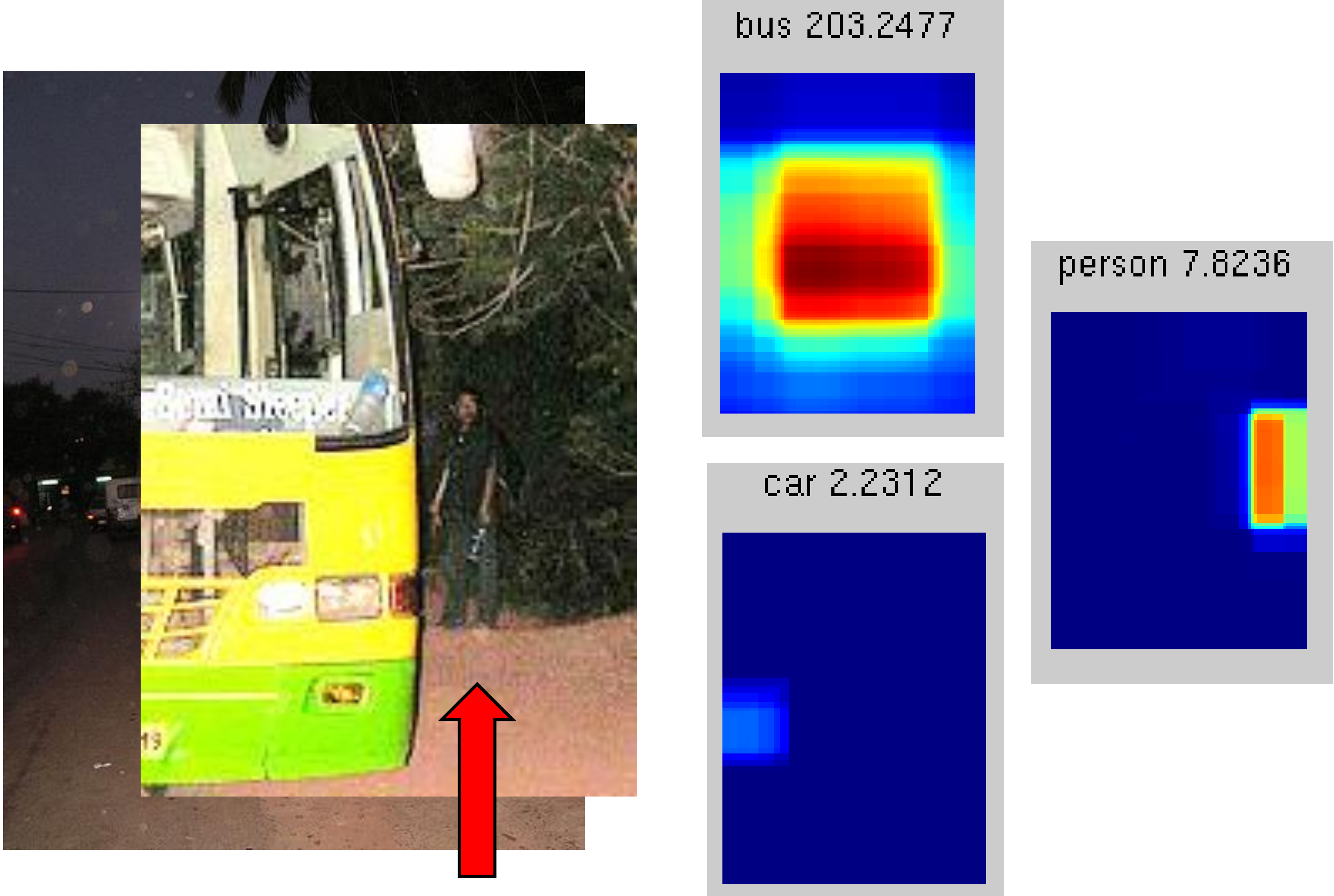
Results

Pascal VOC

Oquab, Bottou, Laptev and Sivic
CVPR 2014

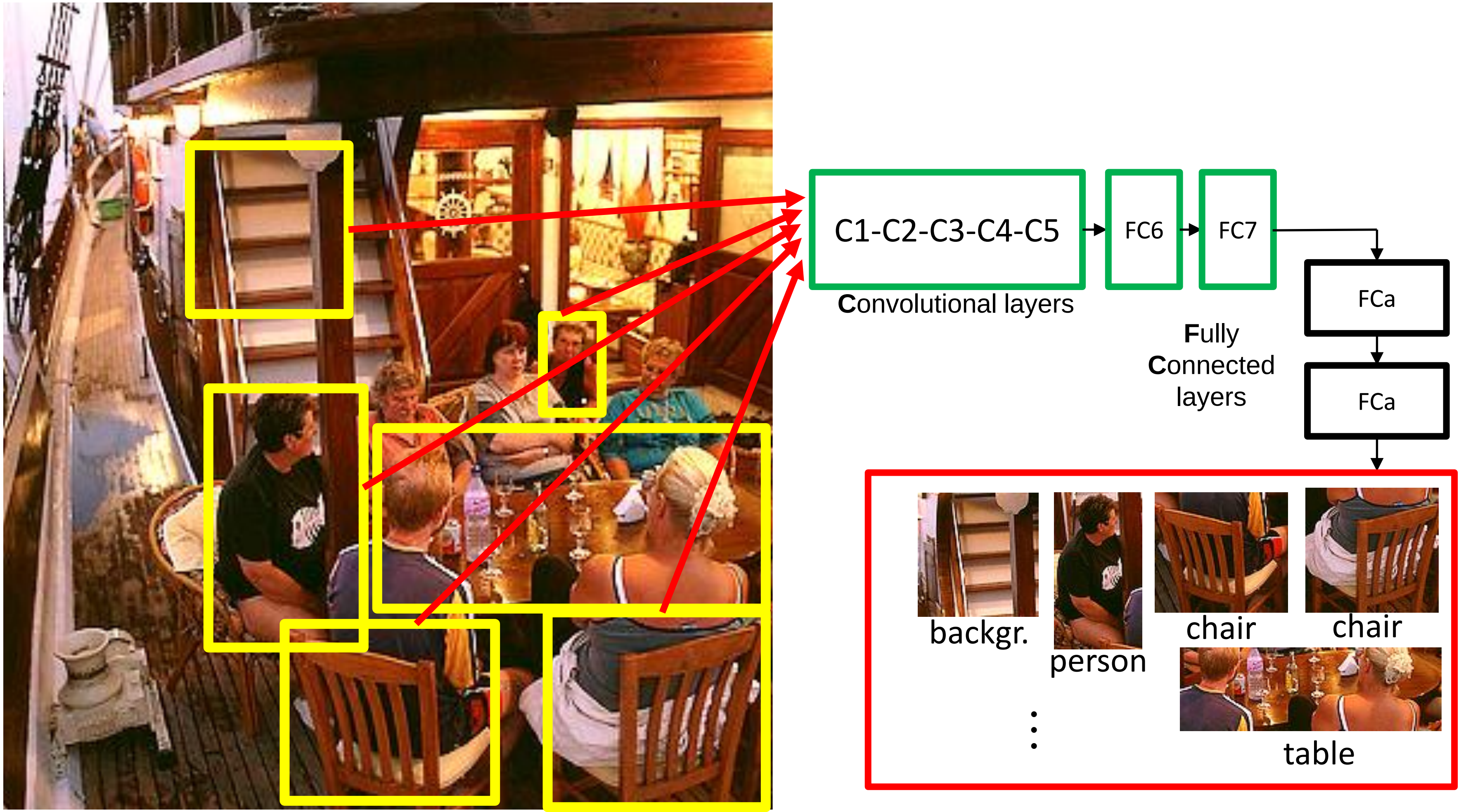


Results



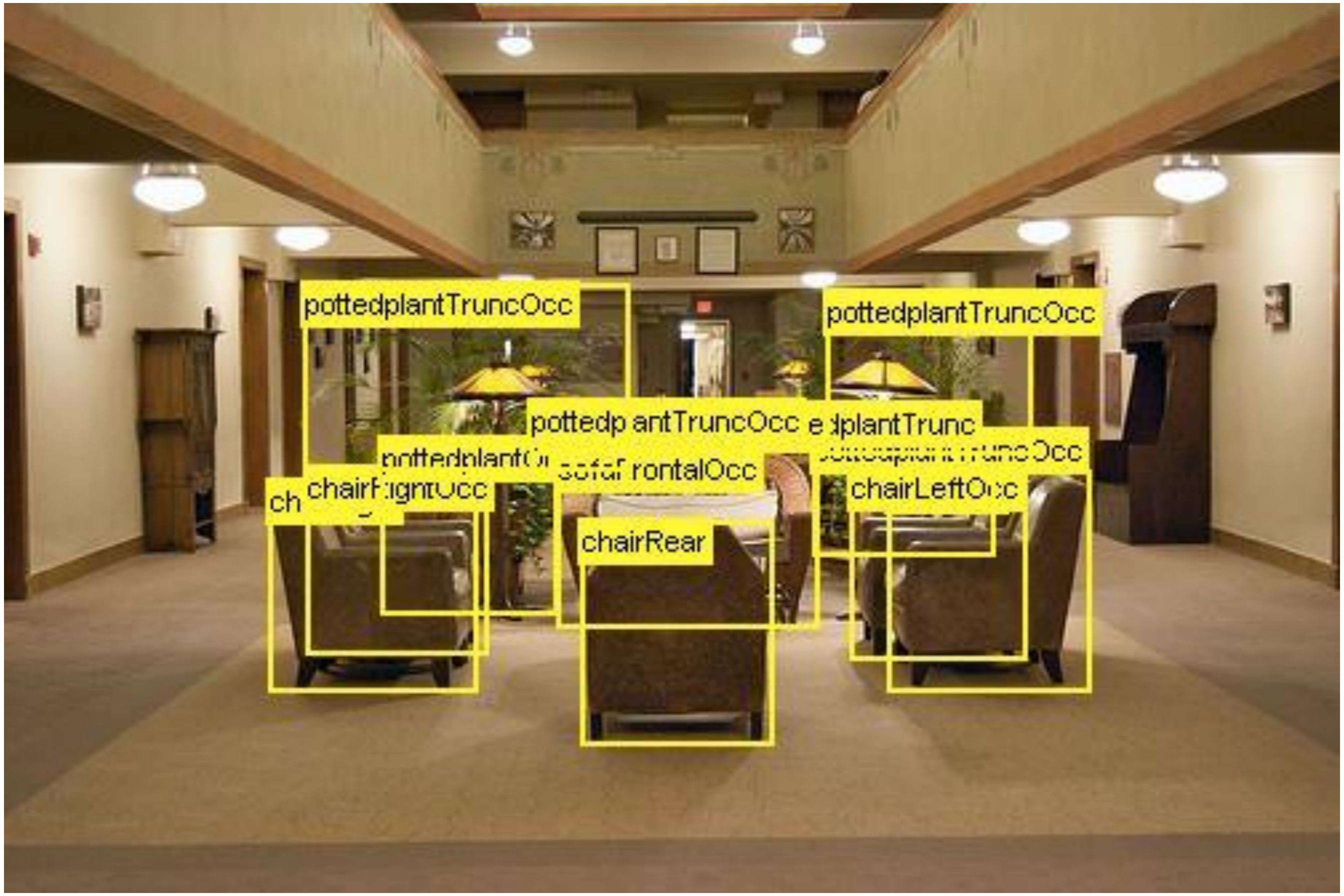
[Oquab, Bottou, Laptev and Sivic, CVPR 2014]

How to use CNNs for cluttered scenes?



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

Motivation: labeling bounding boxes is tedious



Are bounding boxes needed for training CNNs?

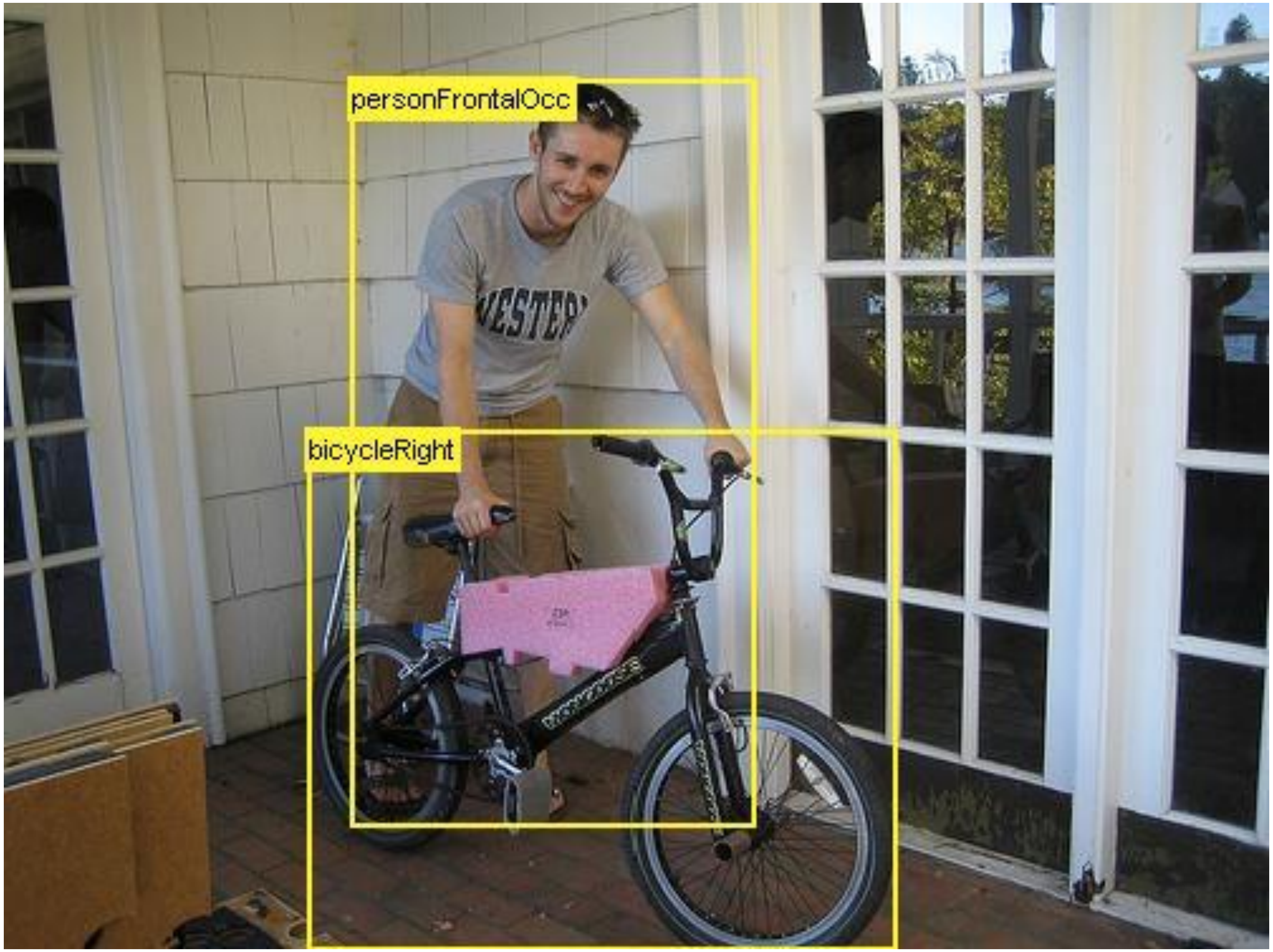
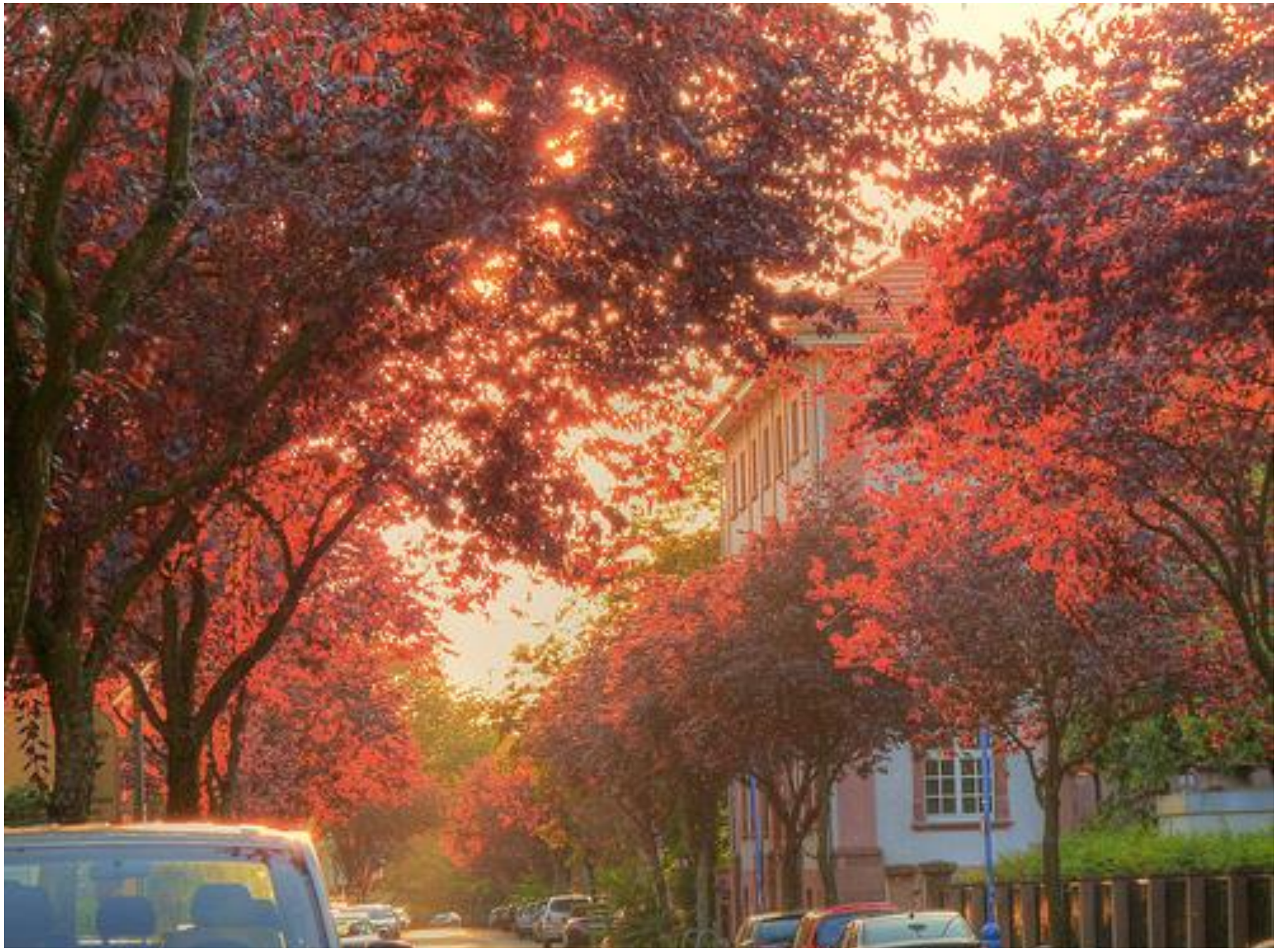


Image-level labels: Bicycle, Person

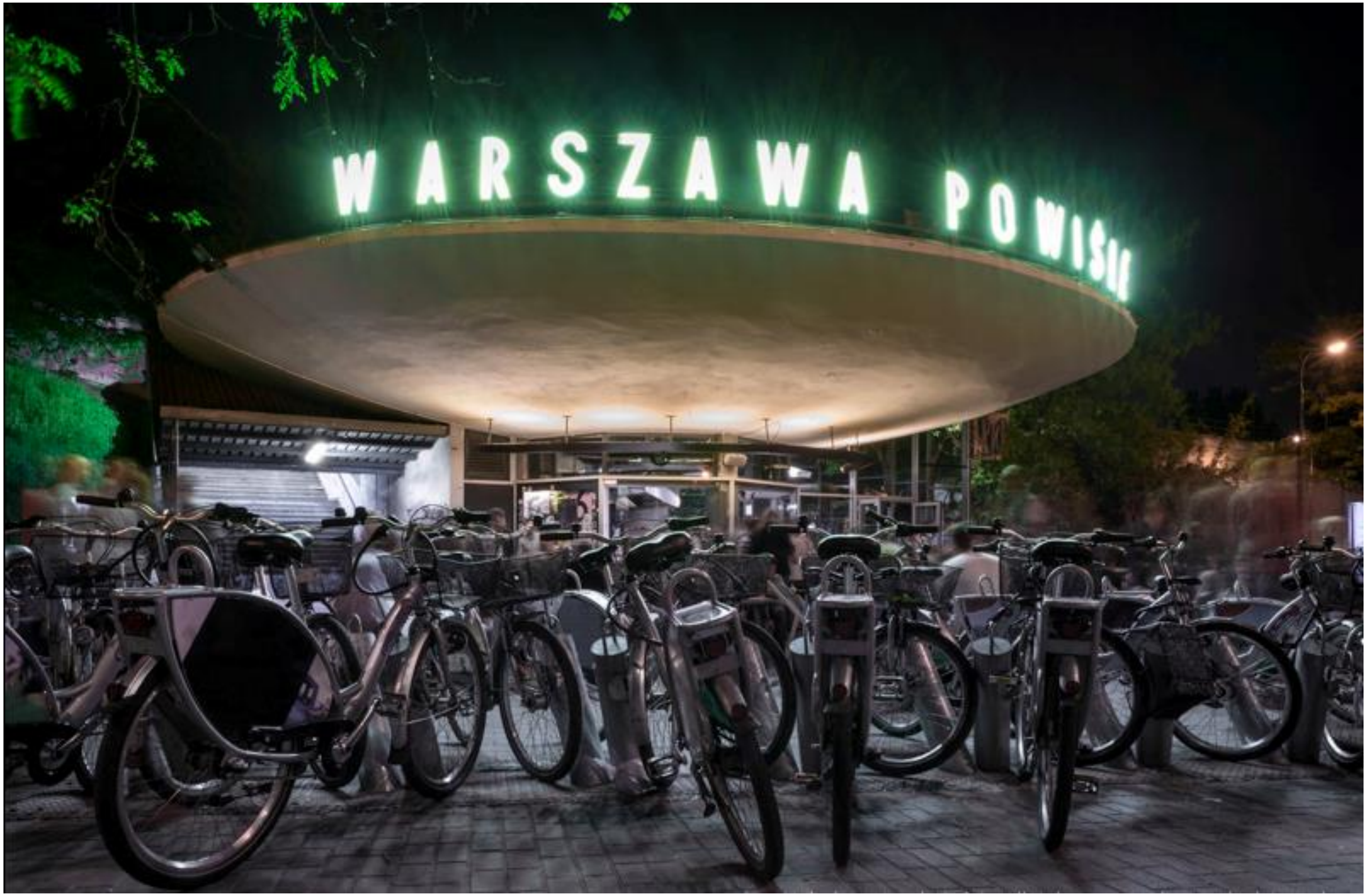
Motivation: image-level labels are plentiful



“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

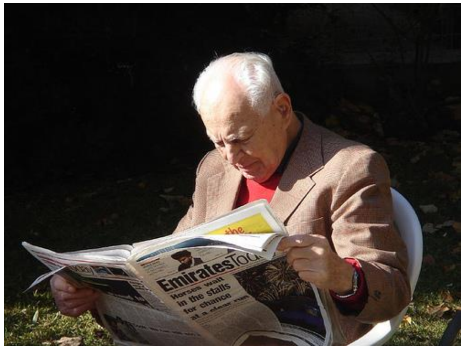


“Public bikes in Warsaw during night”

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

Goal

Training input



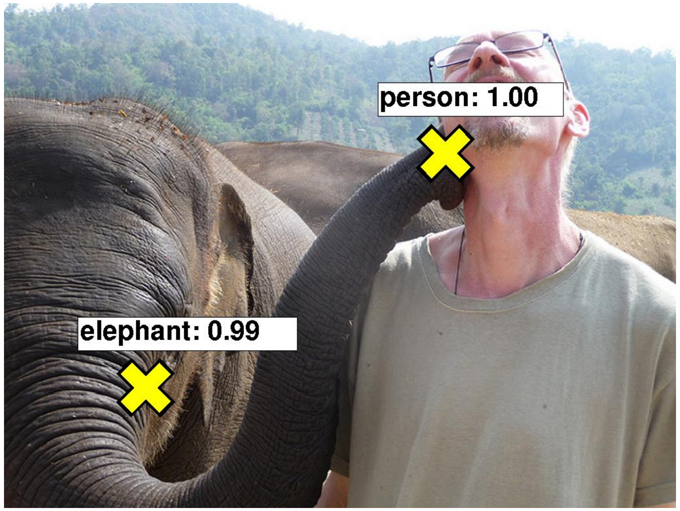
+

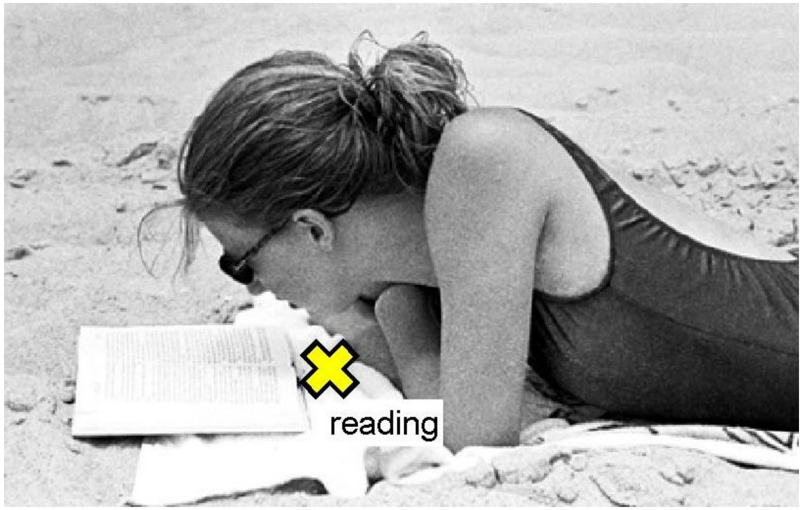
image-level labels:

- ✓ Person
- ✓ Chair
- ✗ Airplane
- …

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

Test output

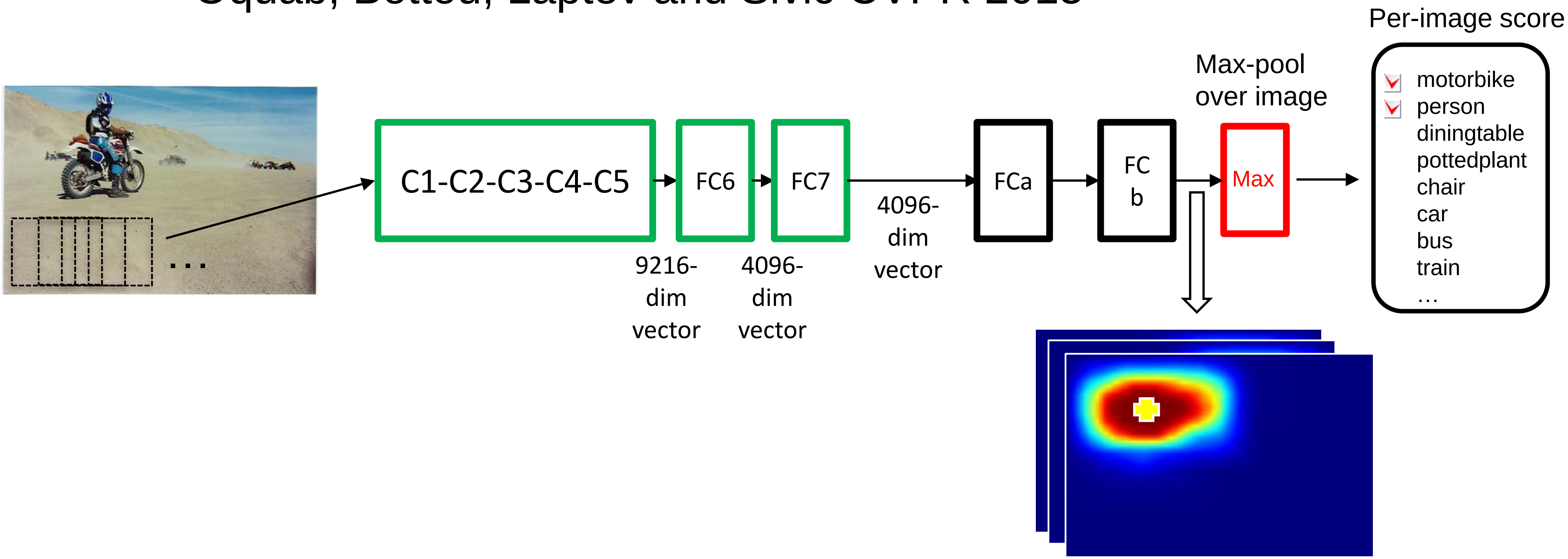




More details in http://www.di.ens.fr/willow/research/weakcnn/

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

Oquab, Bottou, Laptev and Sivic CVPR 2015

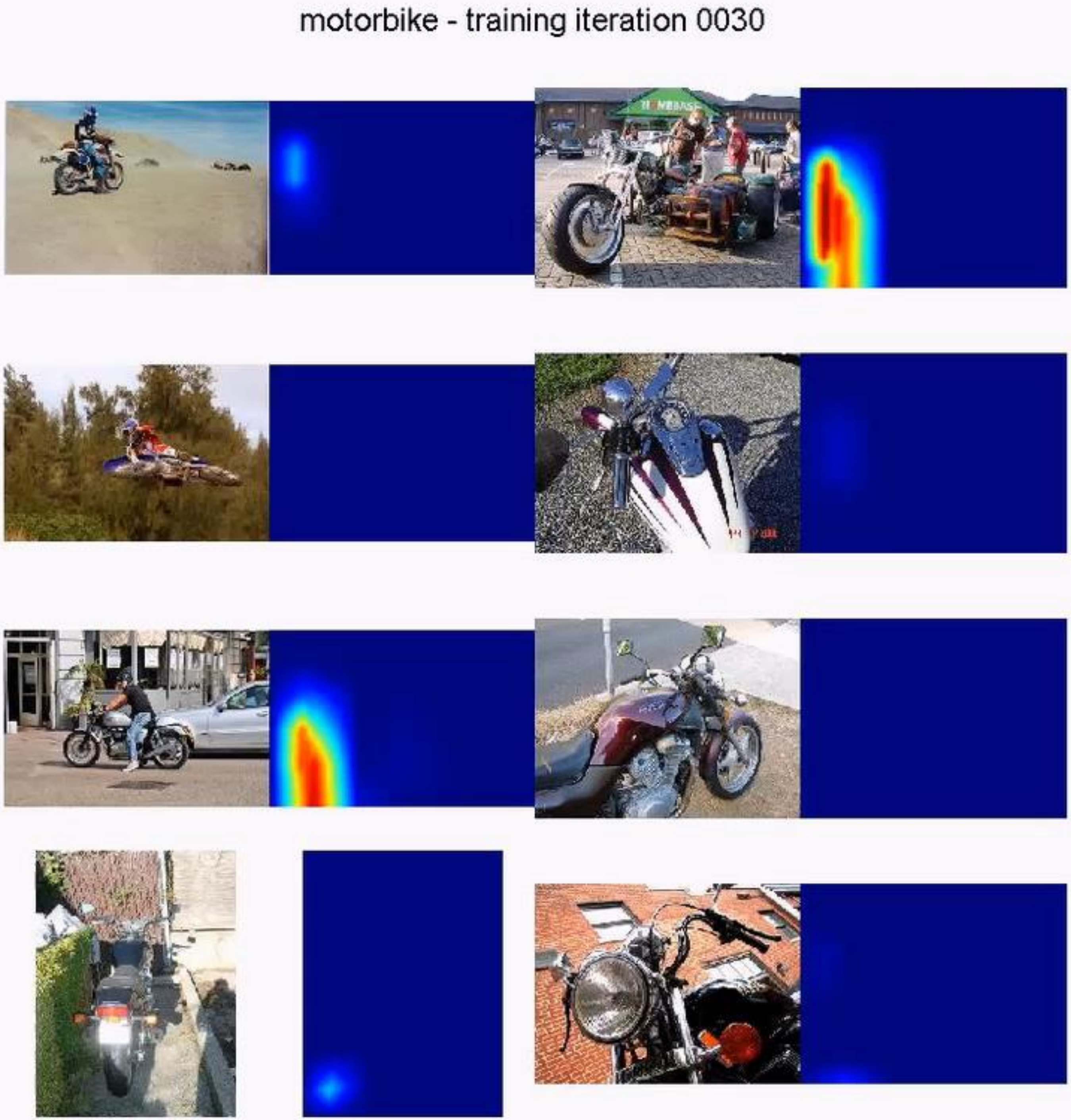


1. Fully convolutional network
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

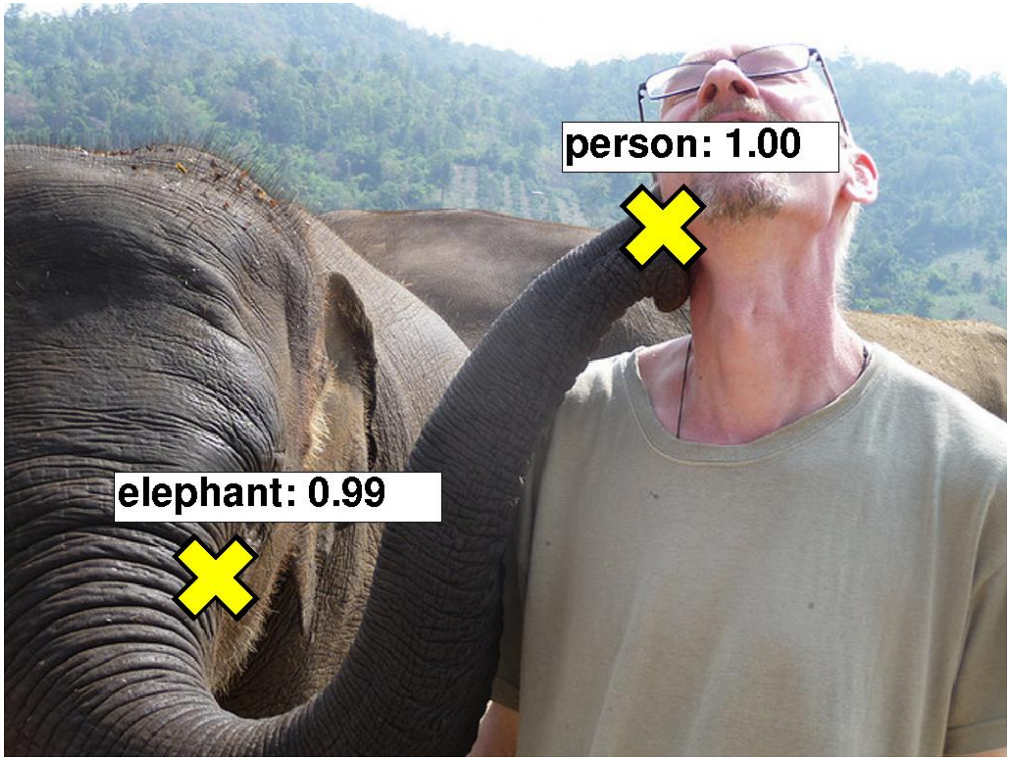
See also [Papandreou et al. '15, Sermanet et al. '14, Chaftield et al.'14]

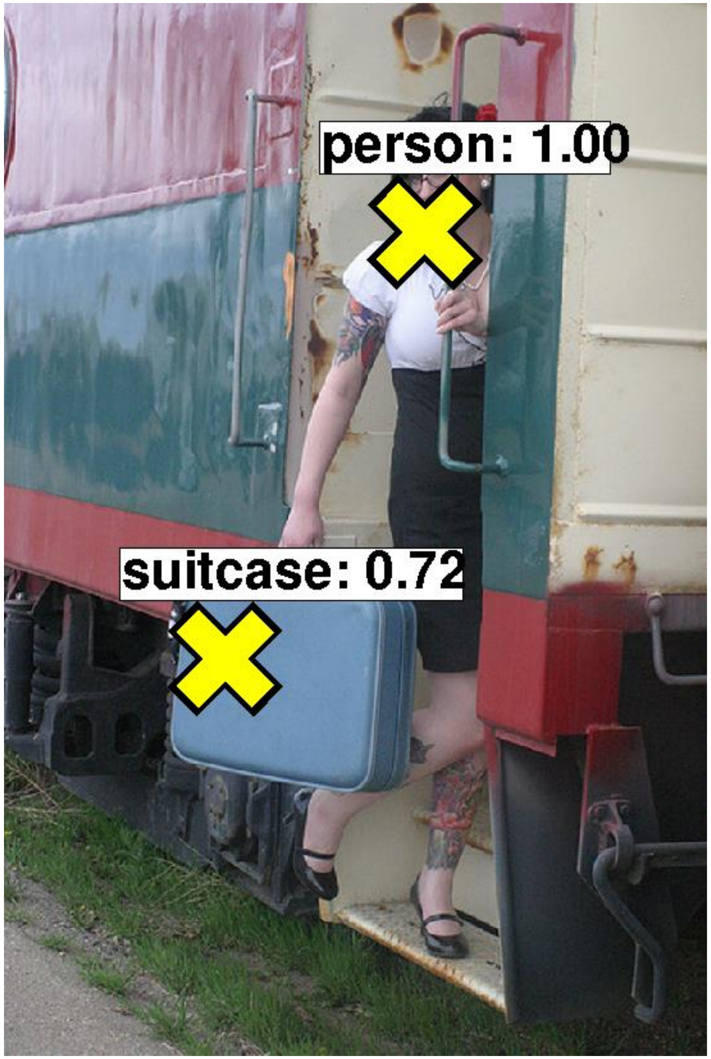
Training Motorbikes

Evolution of localization score maps over training epochs

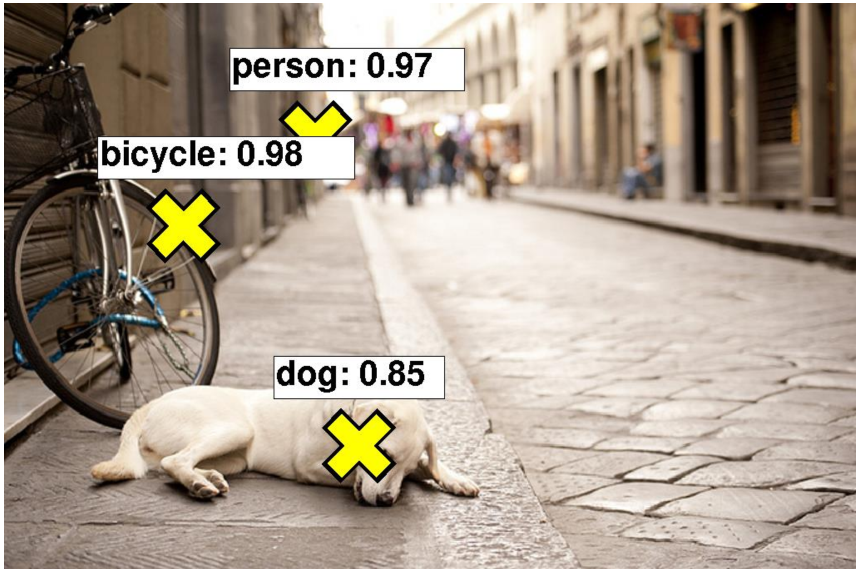


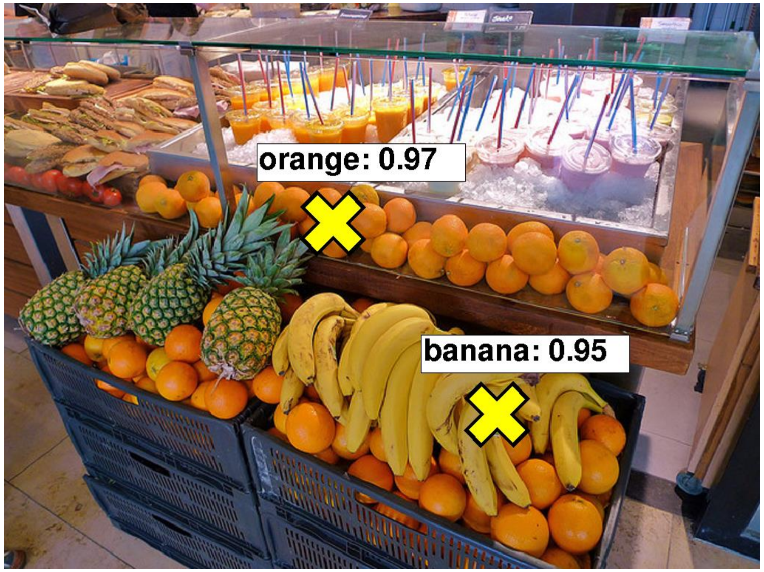
Test results on 80 classes in Microsoft COCO dataset

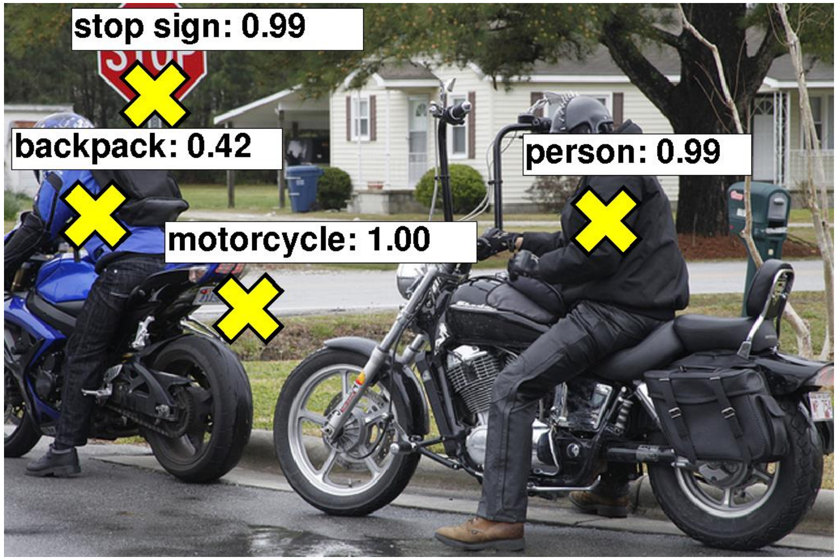


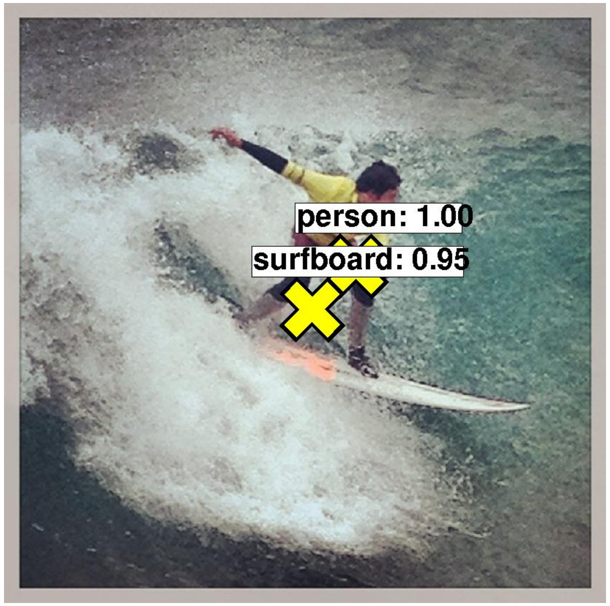


Test results on 80 classes in Microsoft COCO dataset

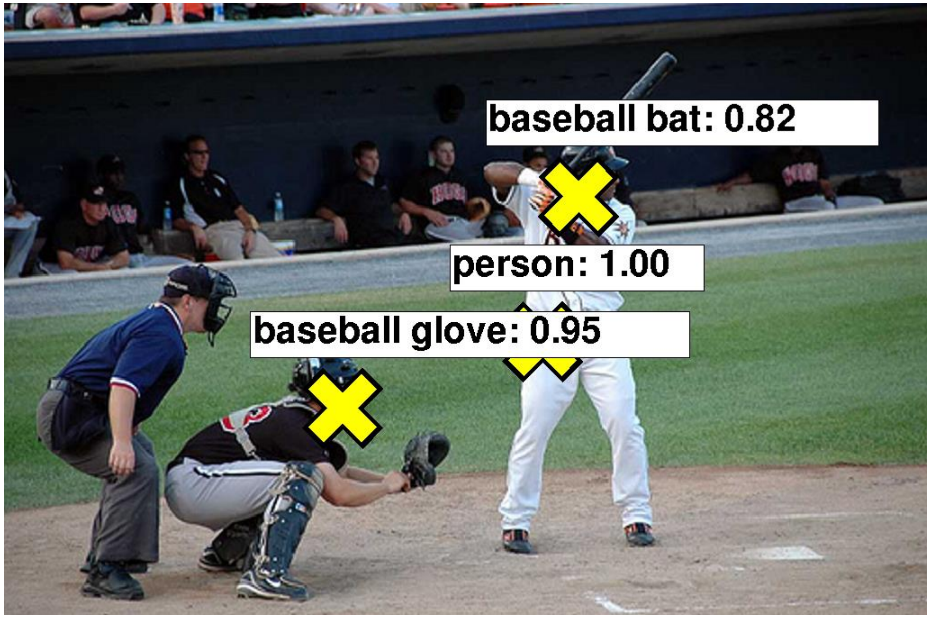


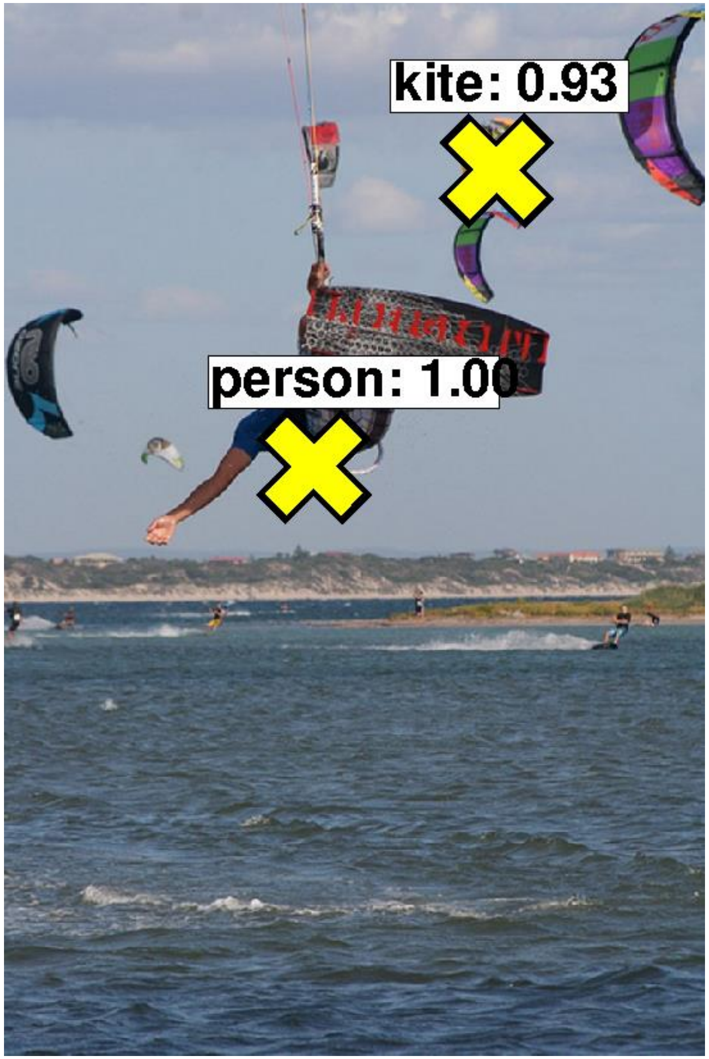


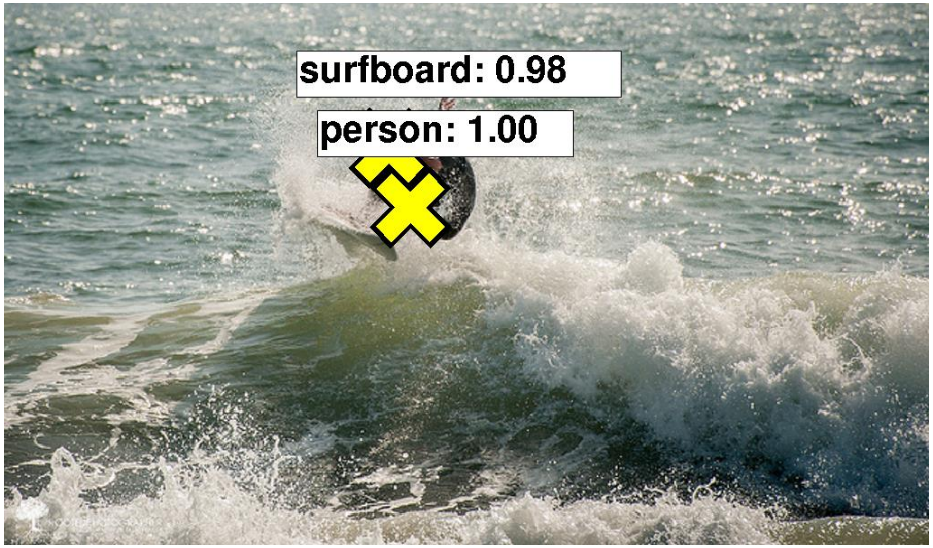




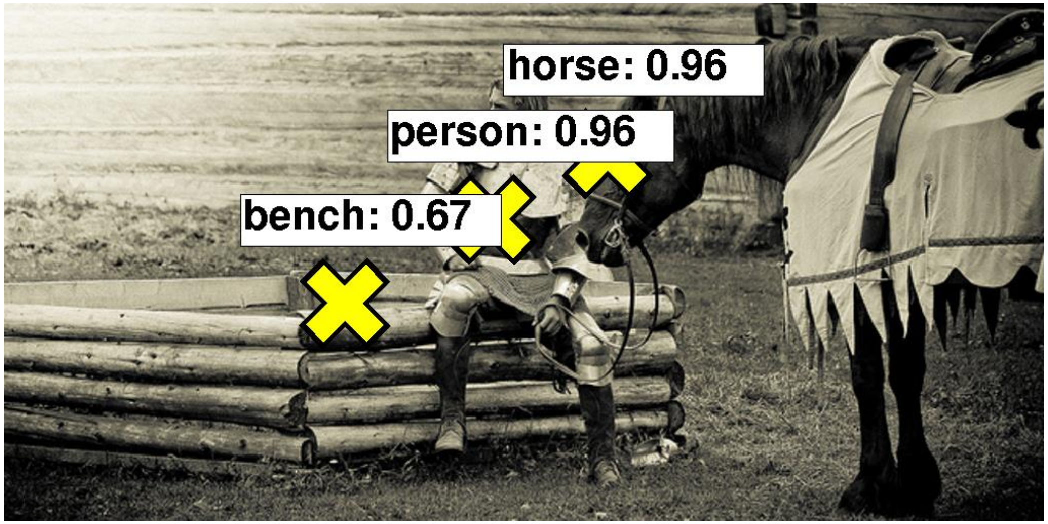
Test results on 80 classes in Microsoft COCO dataset

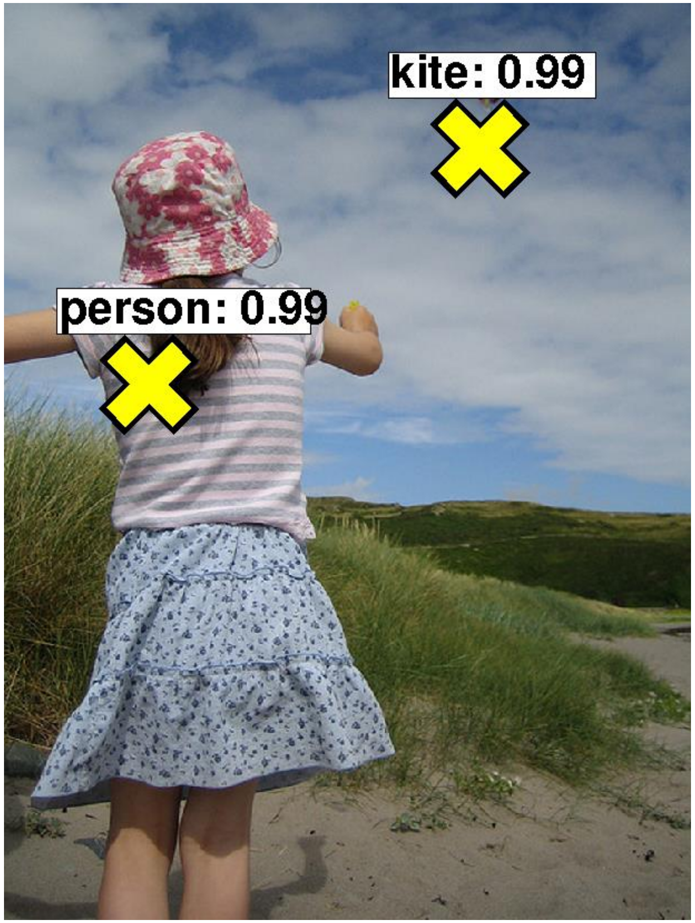


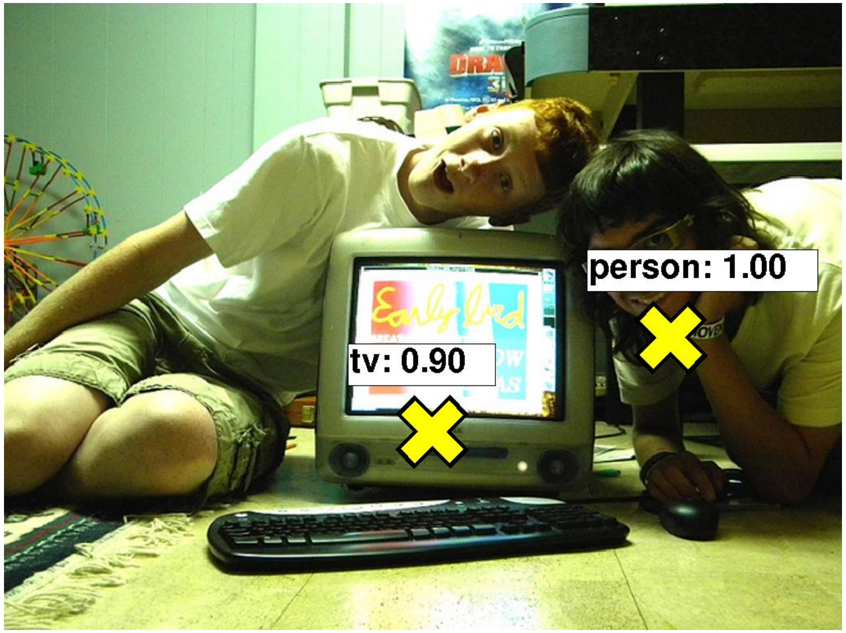




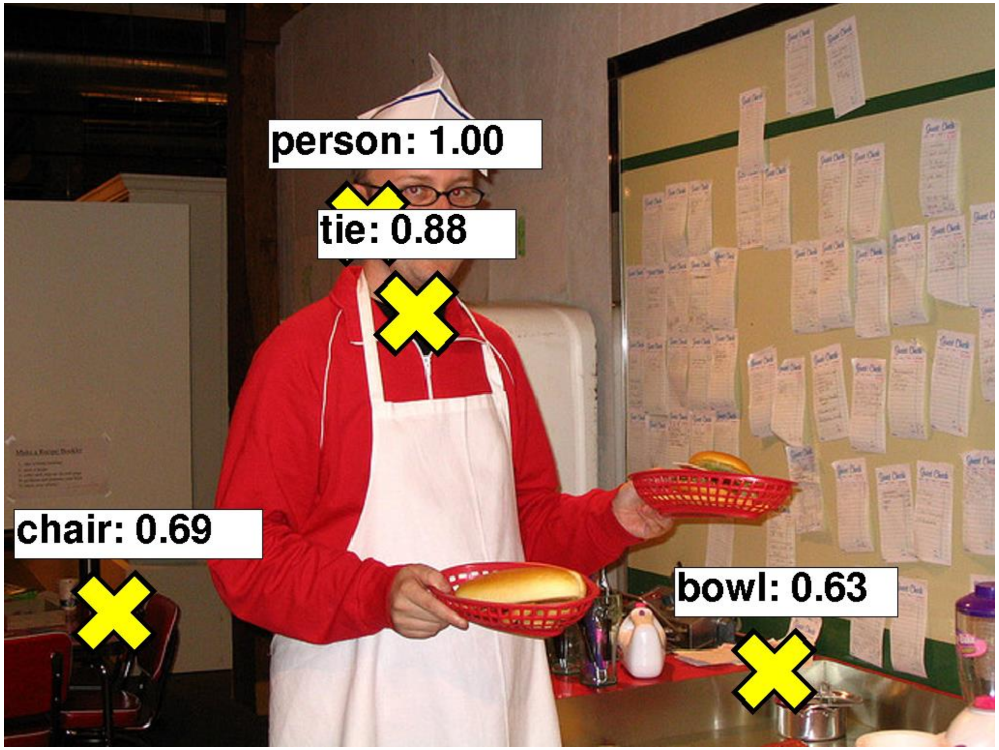
Test results on 80 classes in Microsoft COCO dataset

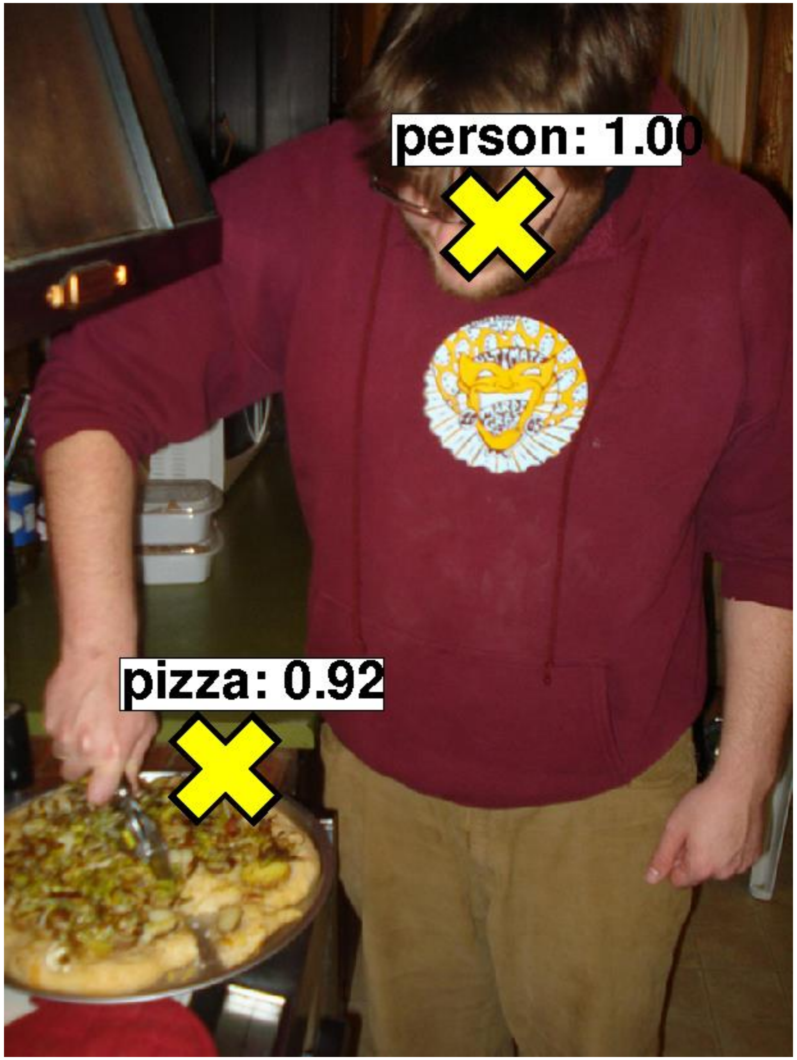






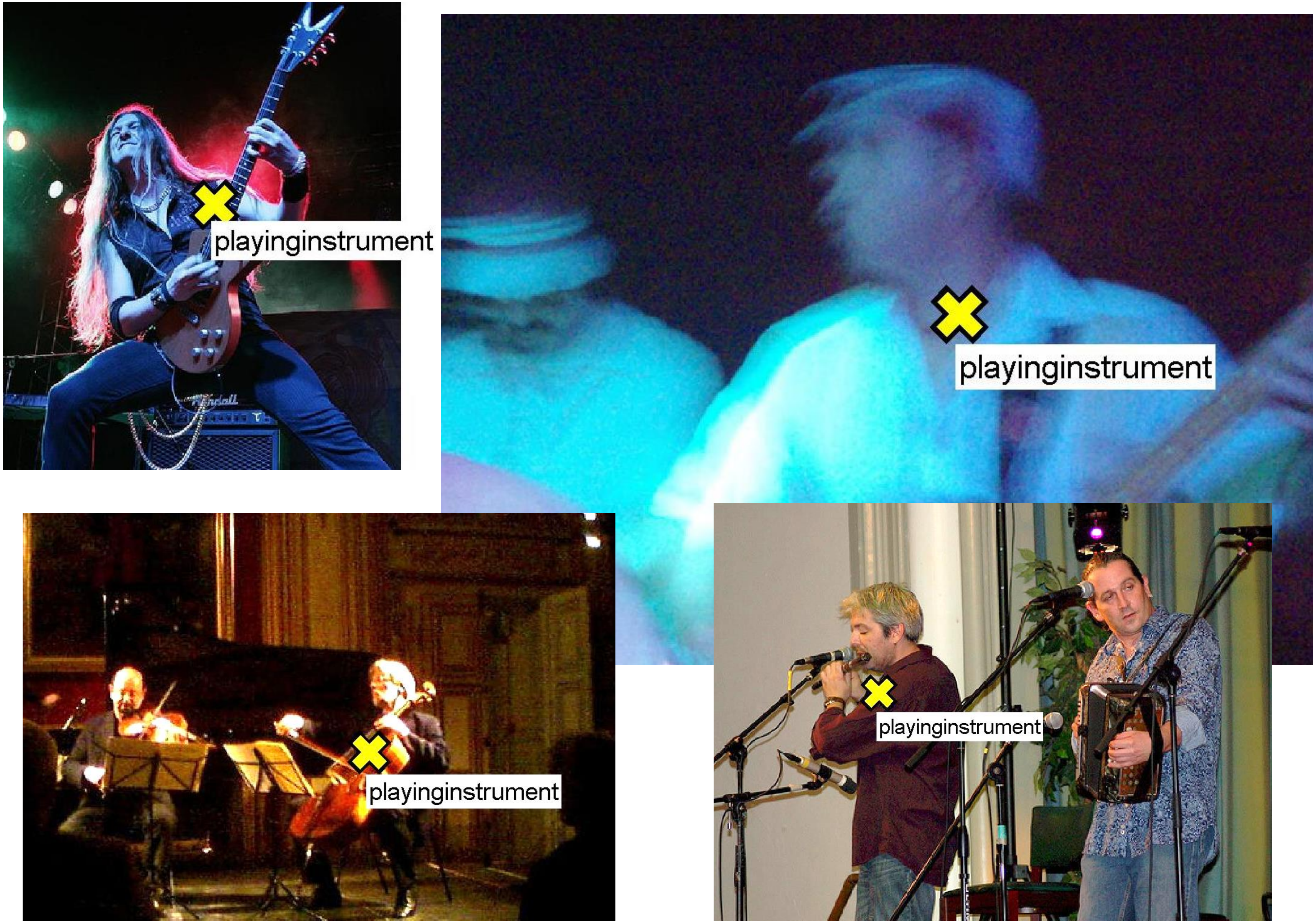
Test results on 80 classes in Microsoft COCO dataset



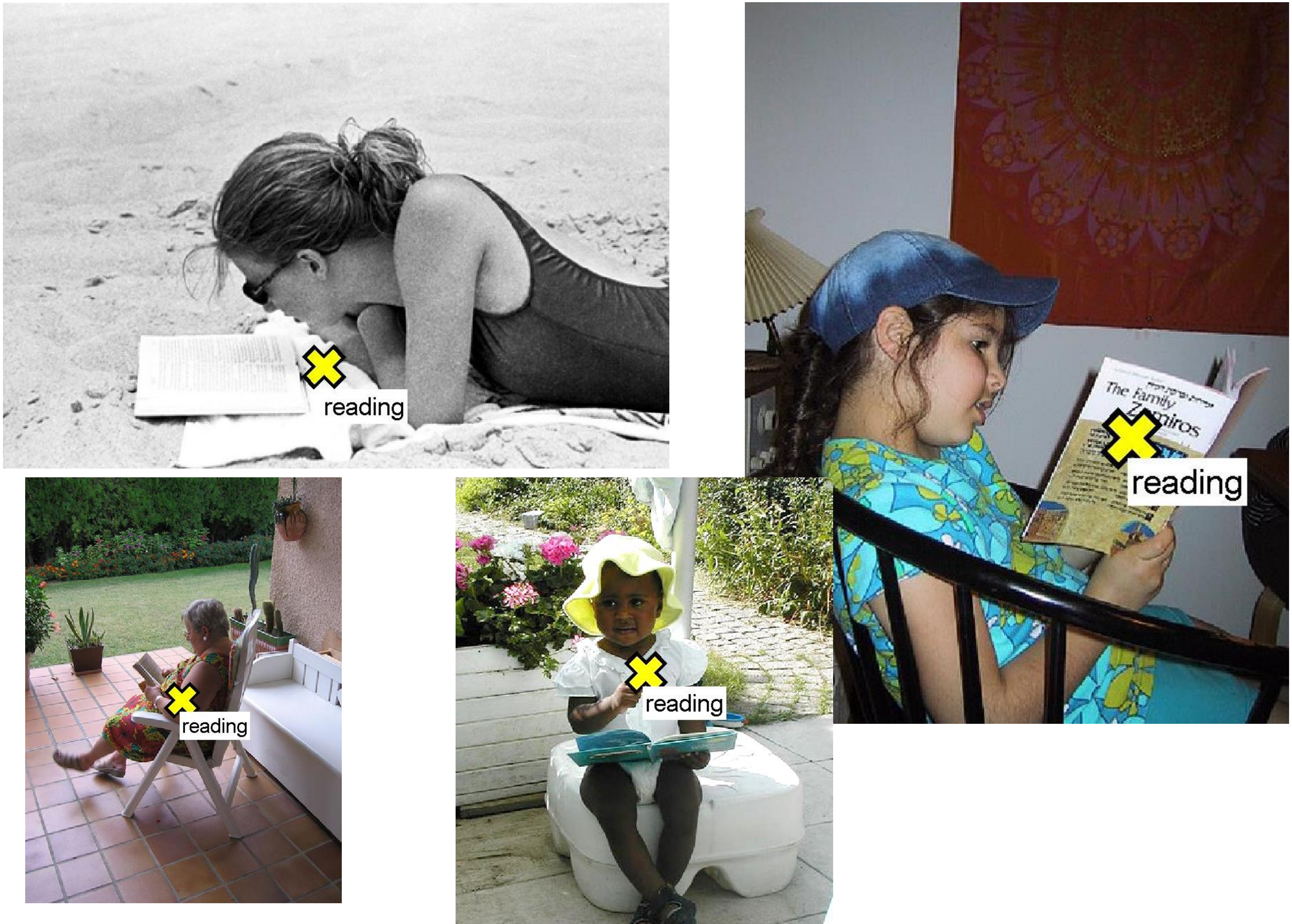


Results for weakly-supervised *action* recognition in Pascal VOC’12 dataset

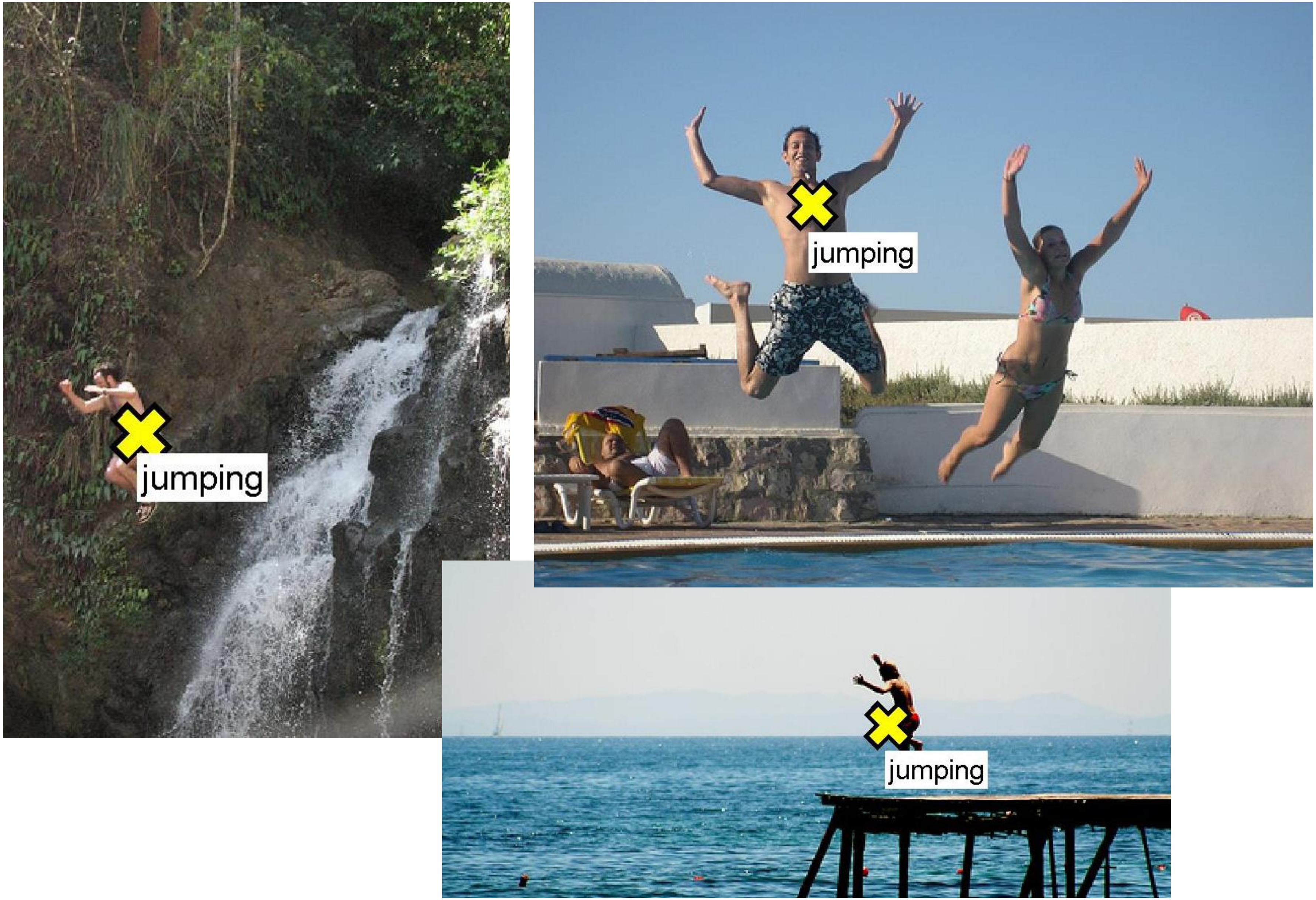
Test results for **10 action classes** in Pascal VOC12



Test results for **10 action classes** in Pascal VOC12

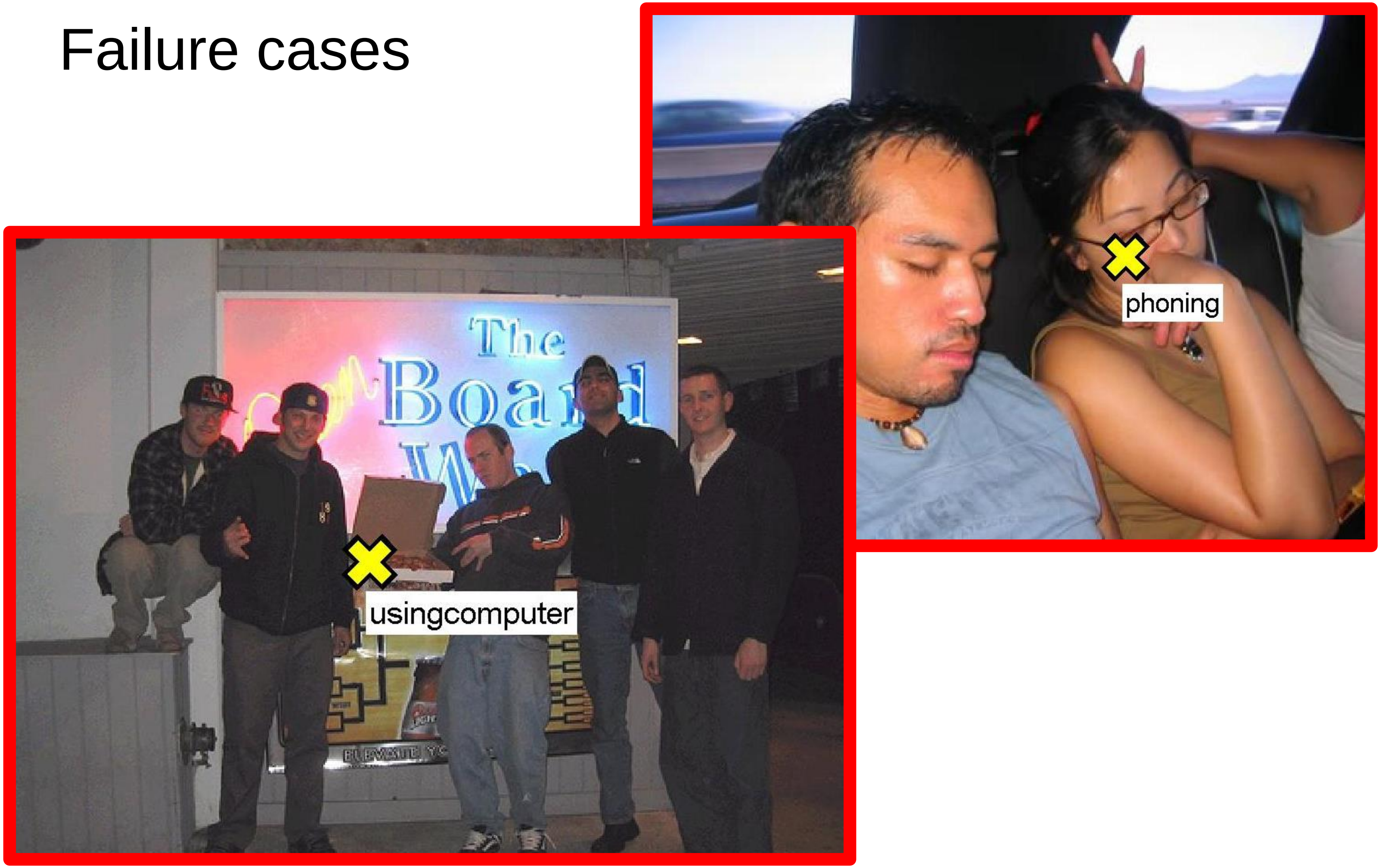


Test results for **10 action classes** in Pascal VOC12



Test results for **10 action classes** in Pascal VOC12

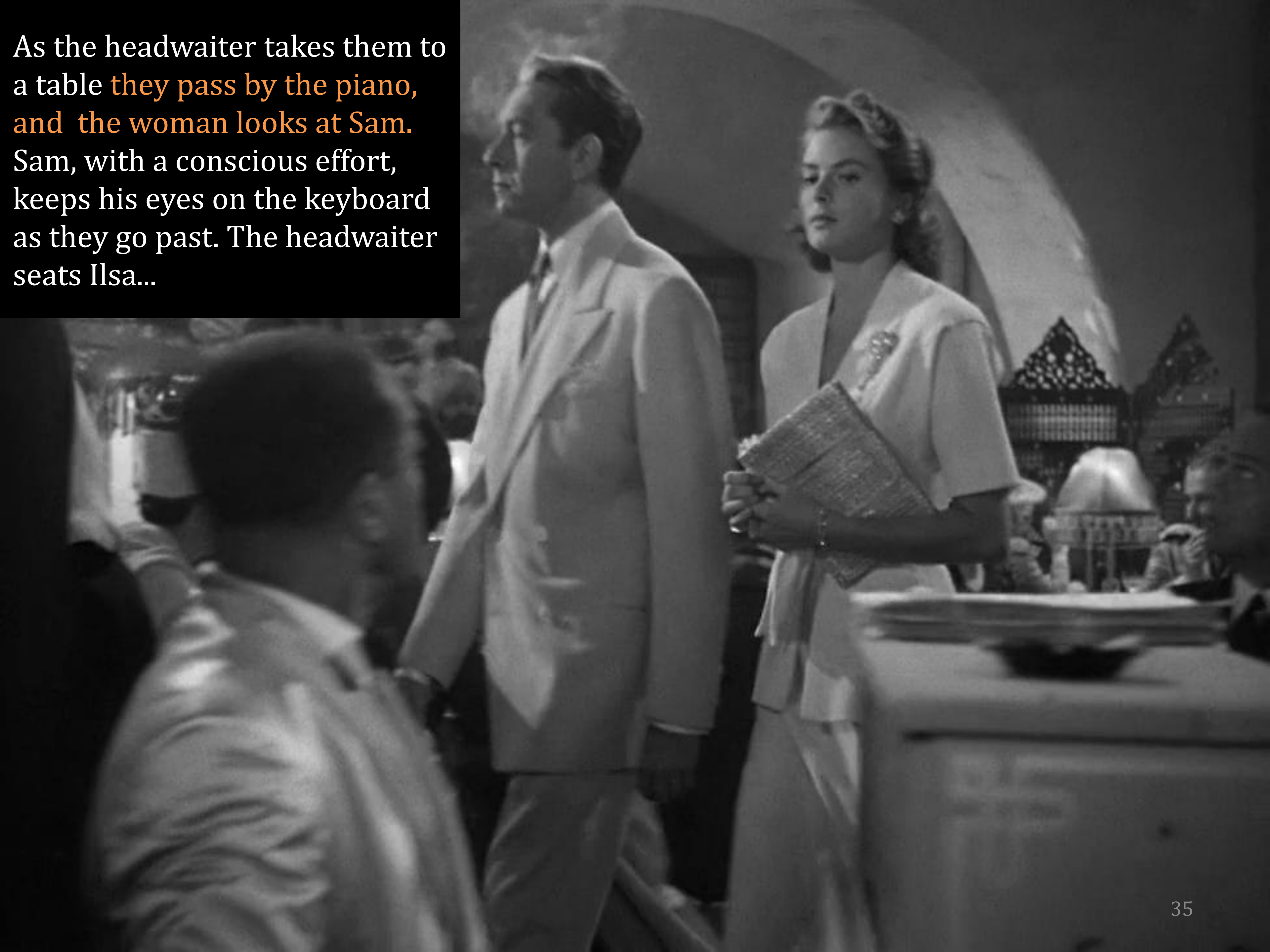
Failure cases



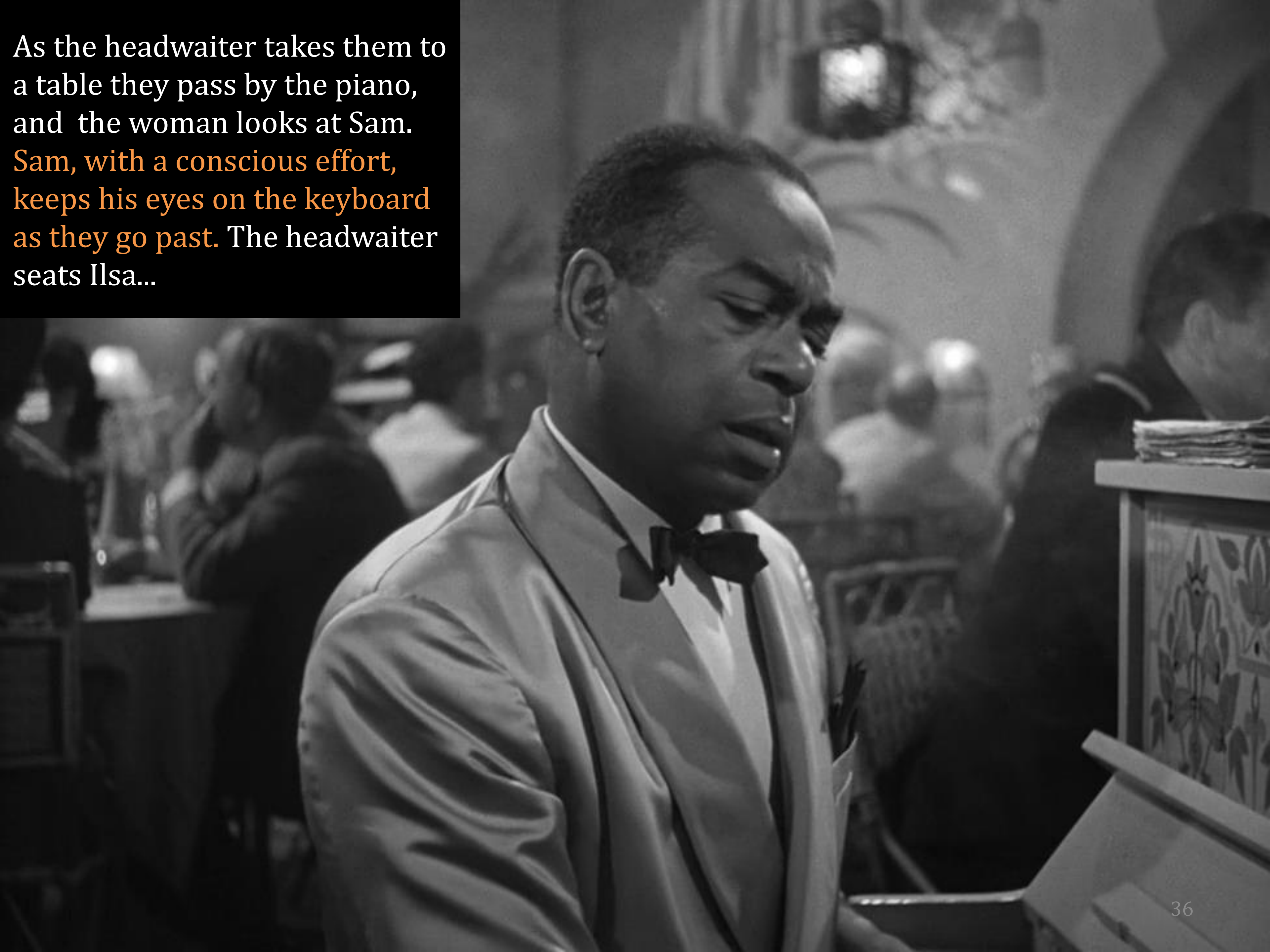
Weakly-supervised learning of actions *in video* from scripts and narrations

As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...

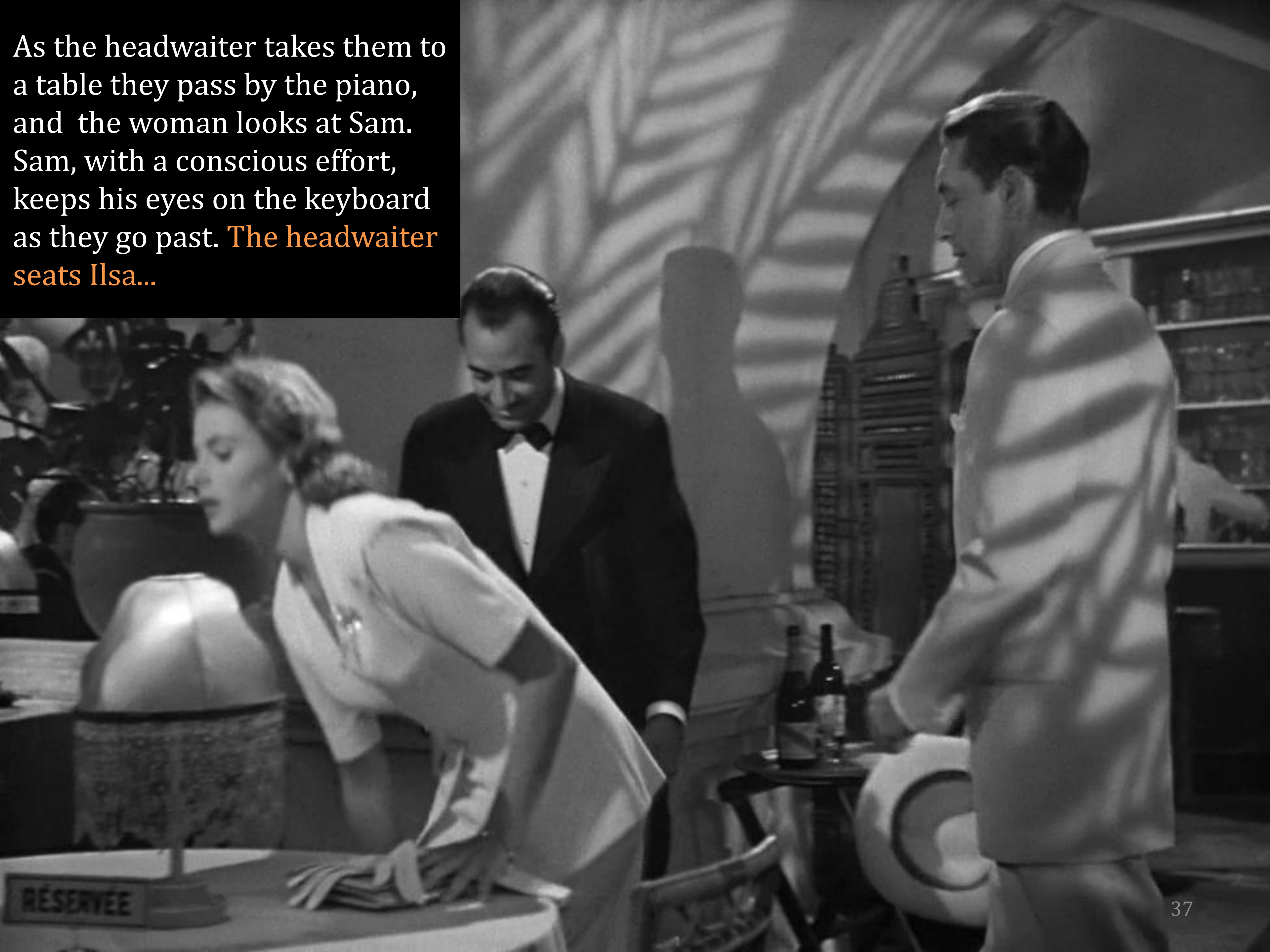
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



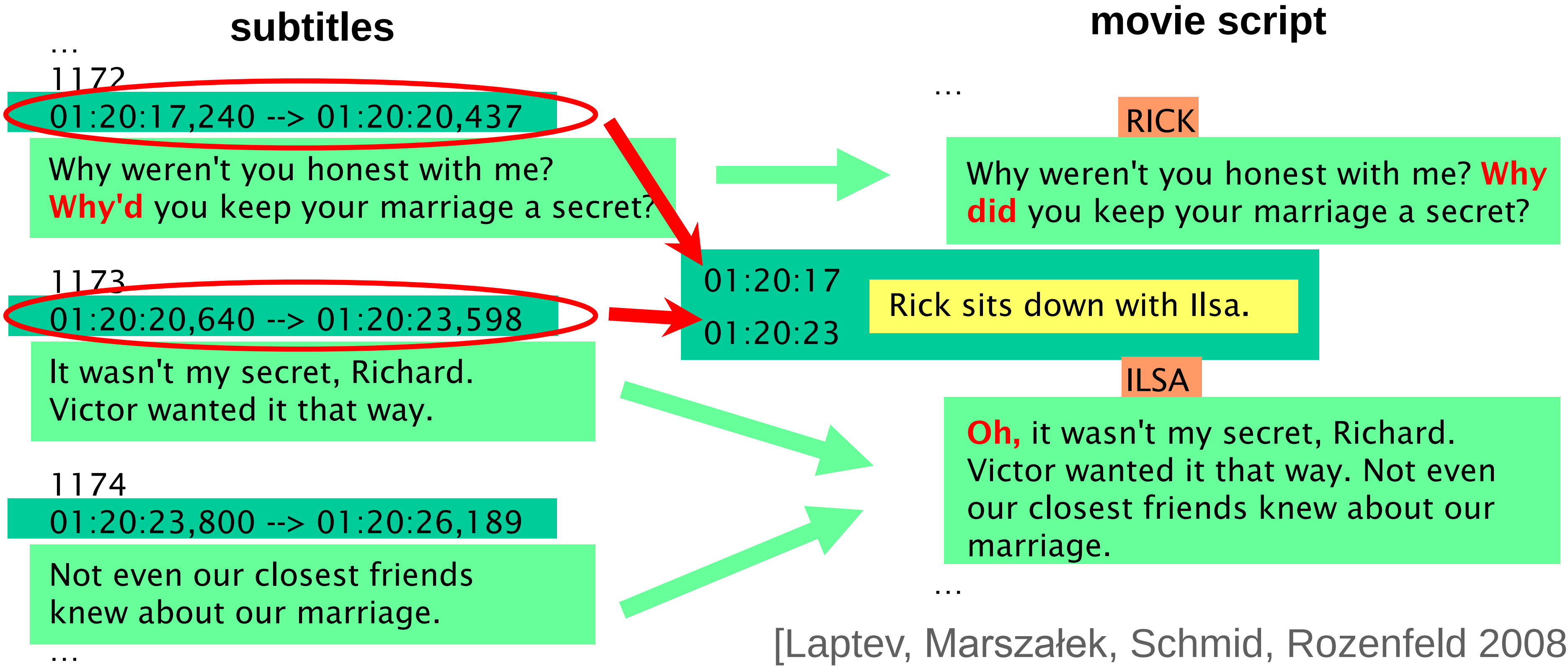
As the headwaiter takes them to a table they pass by the piano, and the woman looks at Sam. Sam, with a conscious effort, keeps his eyes on the keyboard as they go past. The headwaiter seats Ilsa...



Script-based video annotation

- Scripts available for >500 movies (no time synchronization)

  www.dailyscript.com, www.movie-page.com, www.weeklyscript.com …
- Subtitles (with time info.) are available for the most of movies
- Can transfer time to scripts by text alignment



subtitles

…

1172

01:20:17,240 --> 01:20:20,437

Why weren't you honest with me? Why'd you keep your marriage a secret?

1173

01:20:20,640 --> 01:20:23,598

It wasn't my secret, Richard. Victor wanted it that way.

1174

01:20:23,800 --> 01:20:26,189

Not even our closest friends knew about our marriage.

…

movie script

…

RICK

Why weren't you honest with me? Why did you keep your marriage a secret?

01:20:17

01:20:23

Rick sits down with Ilsa.

ILSA

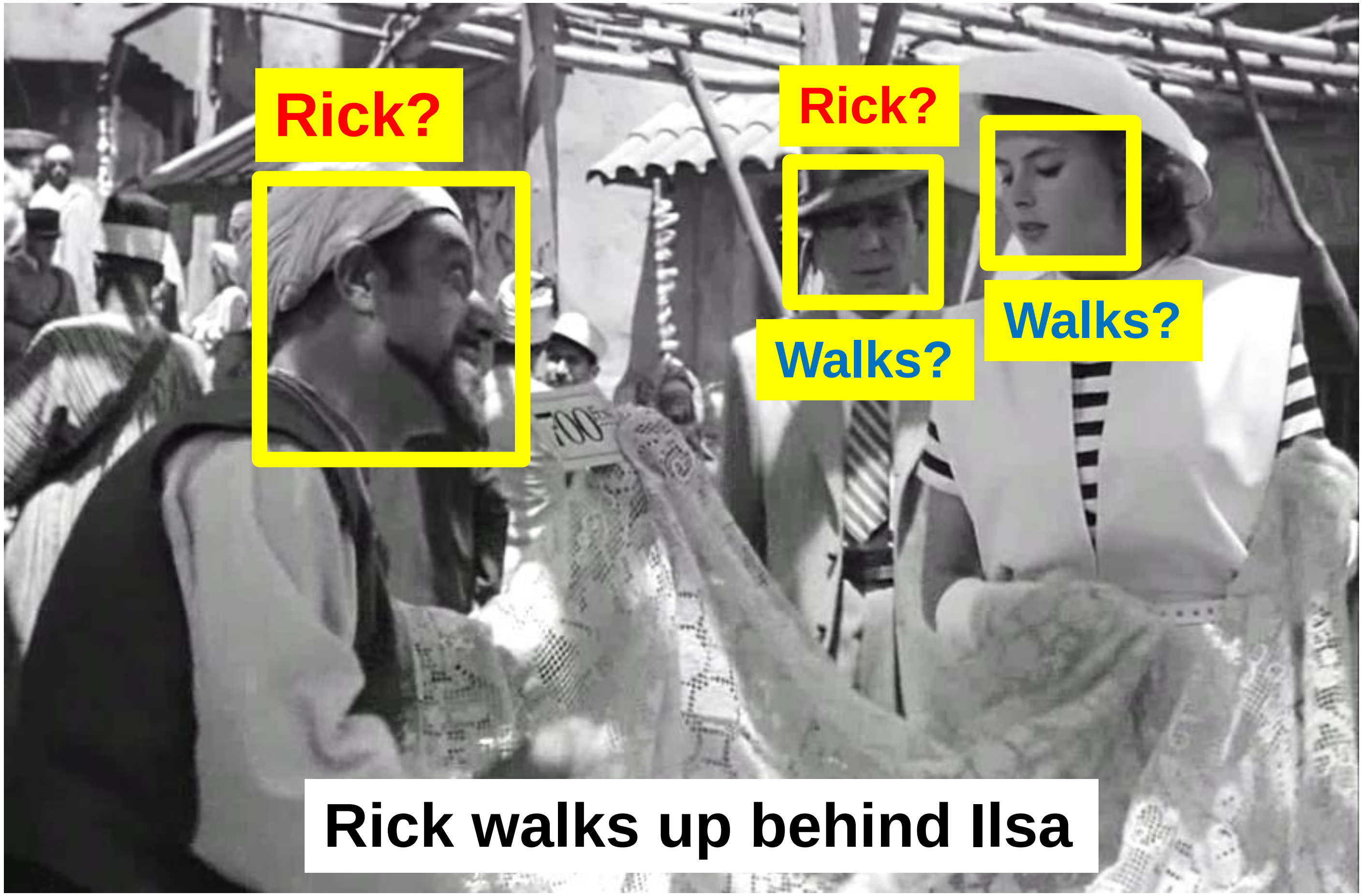
Oh, it wasn't my secret, Richard. Victor wanted it that way. Not even our closest friends knew about our marriage.

…

[Laptev, Marszałek, Schmid, Rozenfeld 2008]

Joint Learning of Actors and Actions

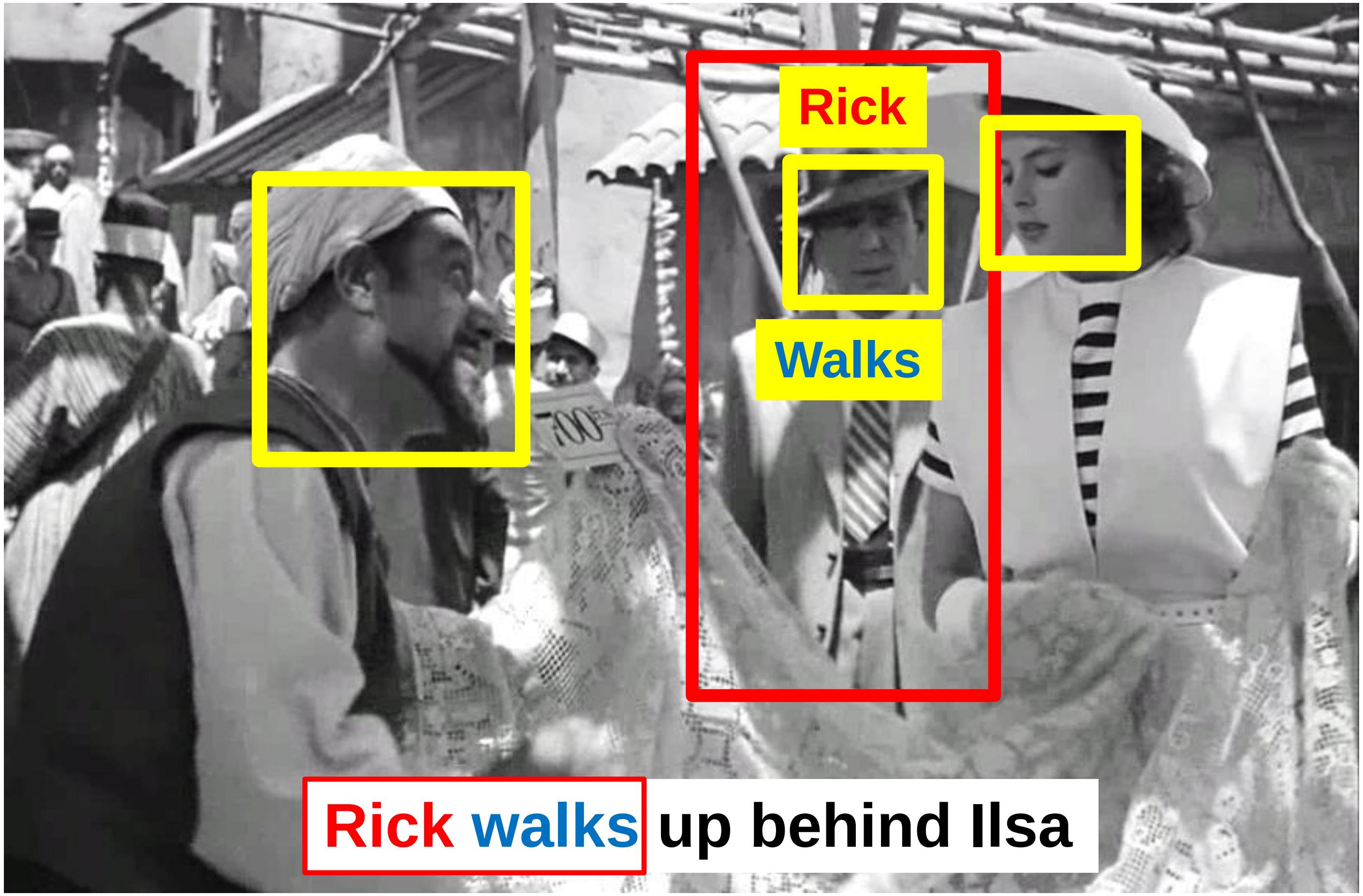
[Bojanowski et al. ICCV 2013]



Rick walks up behind Ilsa

Joint Learning of Actors and Actions

[Bojanowski et al. ICCV 2013]



Formulation: Cost function

$$\frac{1}{N}\|Z - \boxed{\phi(X)}w - b\|_F^2 + \lambda_1 \, Tr(w^T \, w)$$

Actor classifier

Actor labels

Actor image features

Rick
Ilsa
Sam

Formulation: Cost function

$$\frac{1}{N}\|Z - \phi(X)w - b\|_F^2 + \lambda_1 \; Tr(w^T \; w)$$

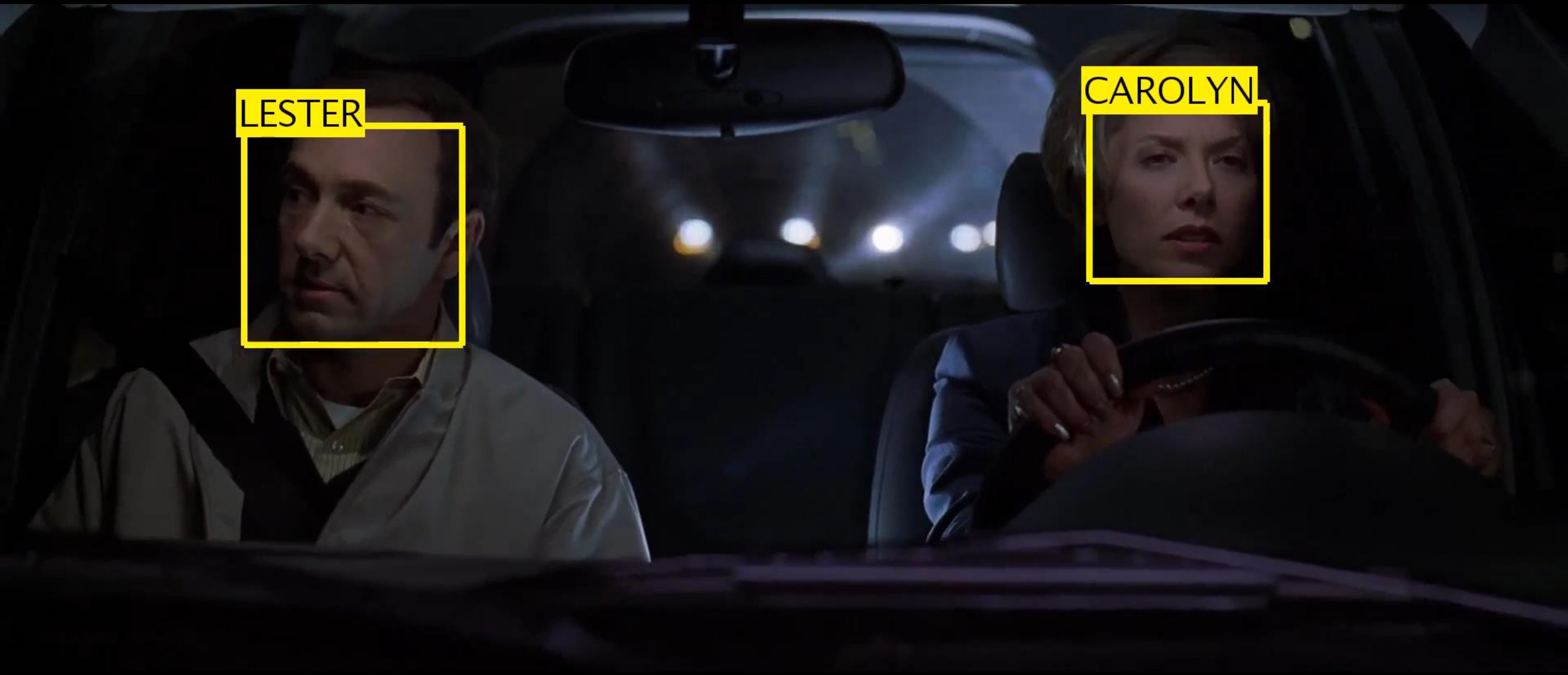
$$\begin{bmatrix} z_{11} & \cdots & z_{1p} & \cdots & z_{1P} \\ \vdots & & \vdots & & \vdots \\ z_{n_1 1} & \cdots & z_{n_1 p} & \cdots & z_{n_1 P} \\ z_{n_2 1} & \cdots & z_{n_2 p} & \cdots & z_{n_2 P} \\ z_{n_3 1} & \cdots & z_{n_3 p} & \cdots & z_{n_3 P} \\ \vdots & & \vdots & & \vdots \\ z_{N1} & \cdots & z_{Np} & \cdots & z_{NP} \end{bmatrix}$$

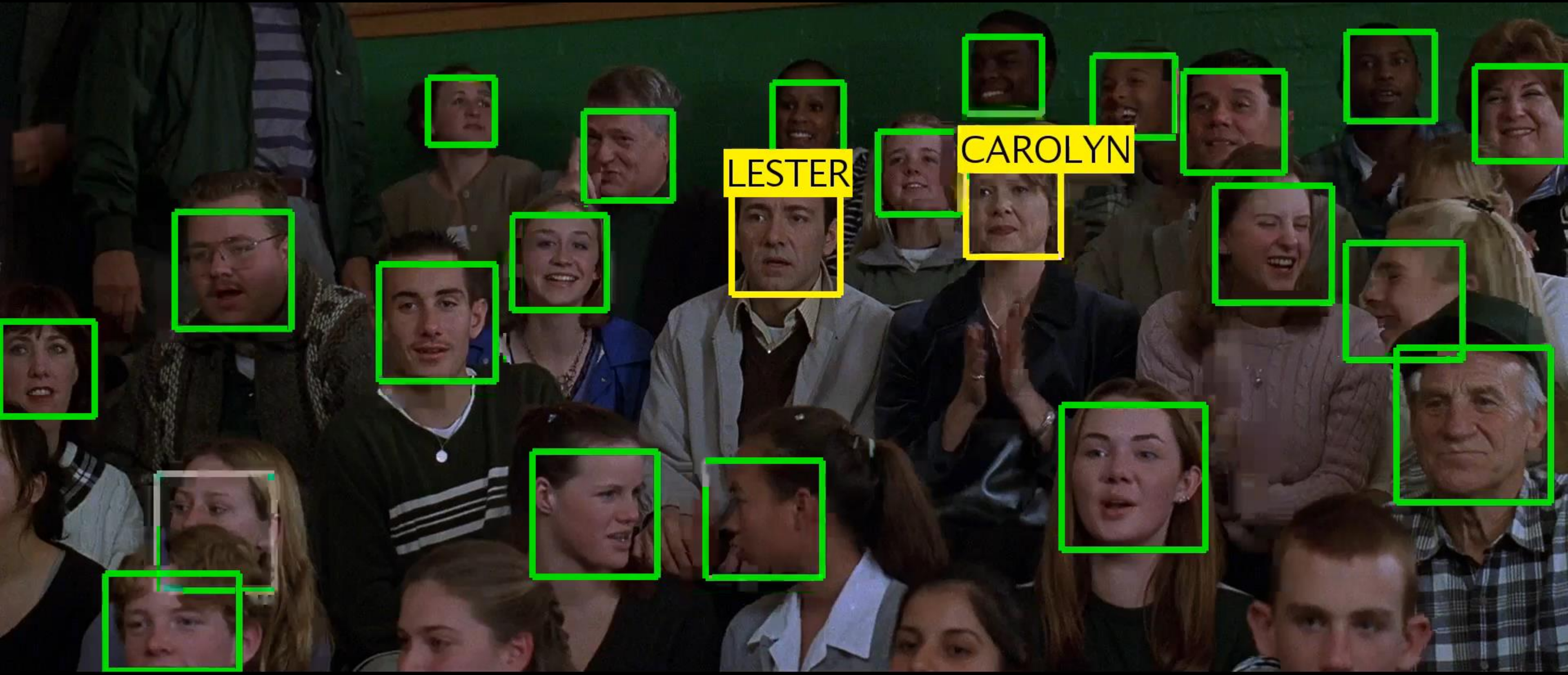
p = Rick

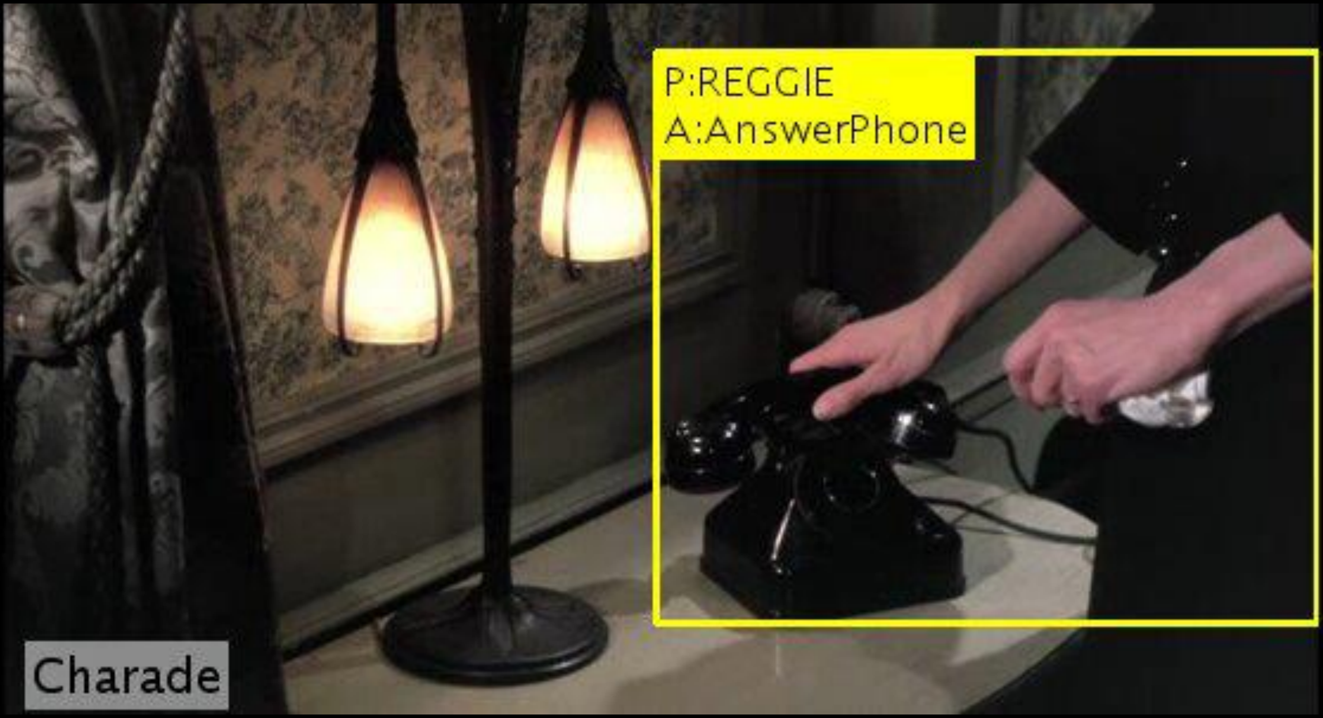
Weak supervision from scripts:

Person p appears at least once in clip N :

$$\sum_{n \in \mathcal{N}_i} z_{np} \geq 1$$

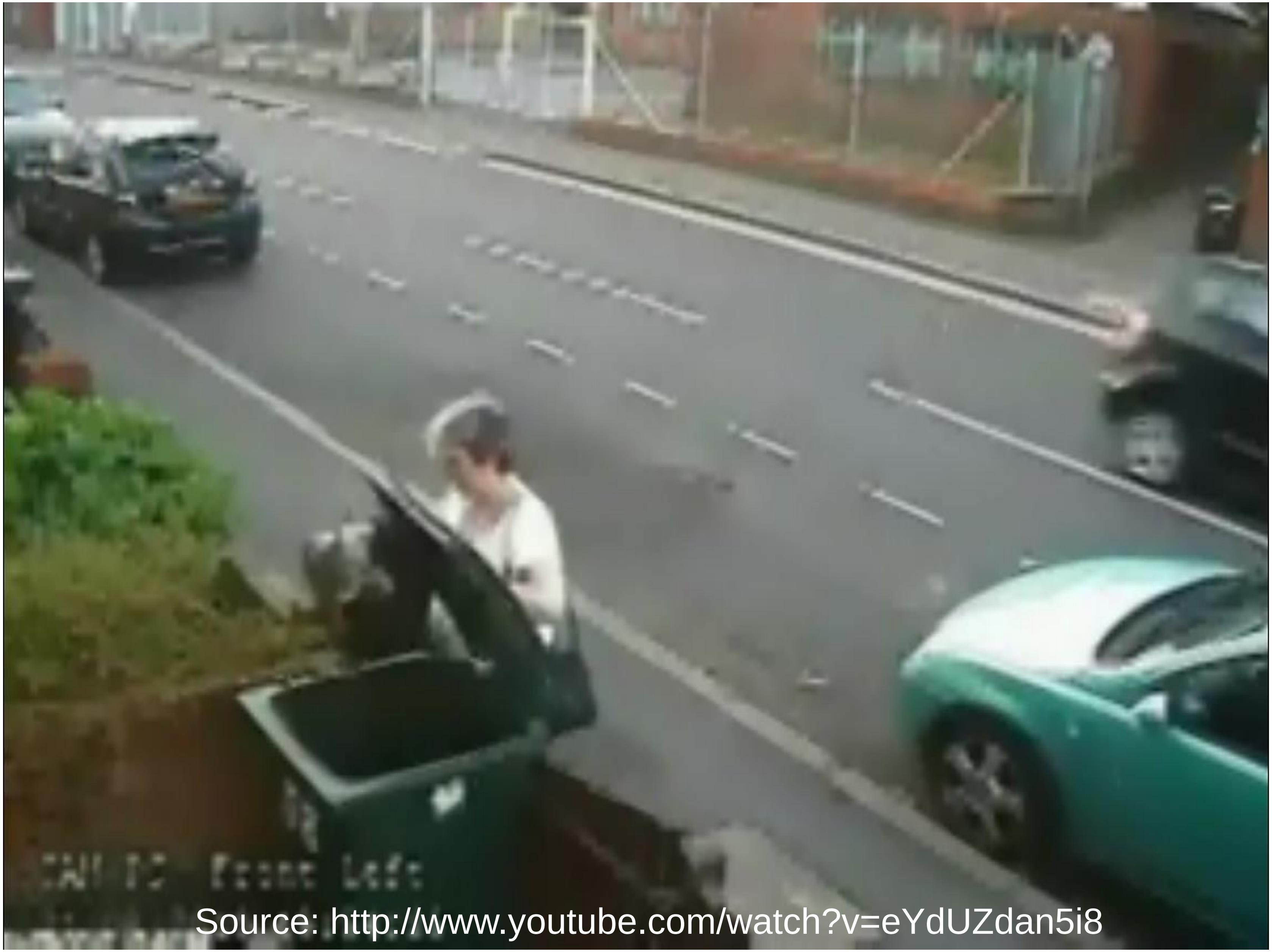

LESTER
CAROLYN


LESTER
CAROLYN


P:REGGIE
A:AnswerPhone
Charade

All problems solved?

CAM-EC Front Left
21/08/2010 19:54:18



Source: http://www.youtube.com/watch?v=eYdUZdan5i8

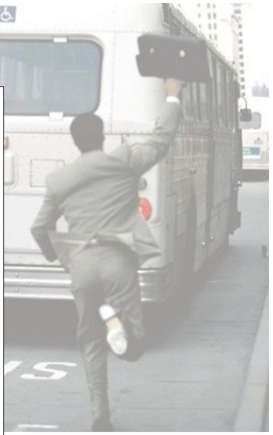
Current solution: learn *person-throws-cat-into-trash-bin* classifier

Limitations of Current Methods

What is unusual in this scene?

Is this scene dangerous?

What is intention of this person?



What is unusual in this scene?