Deep Machine Intelligence and its Applications
SkolTech, Moscow, June 4, 2016

Computer vision in CNN era: New challenges and opportunities

Ivan Laptev

ivan.laptev@inria.fr

WILLOW, INRIA/ENS/CNRS, Paris
VisionLabs, Moscow

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

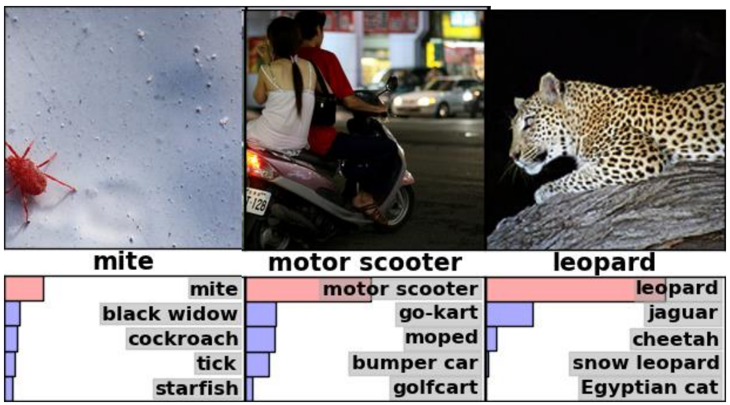
Computer Vision Grand Challenge: Dynamic scene understanding

Computer Vision Grand Challenge: Dynamic scene understanding

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:

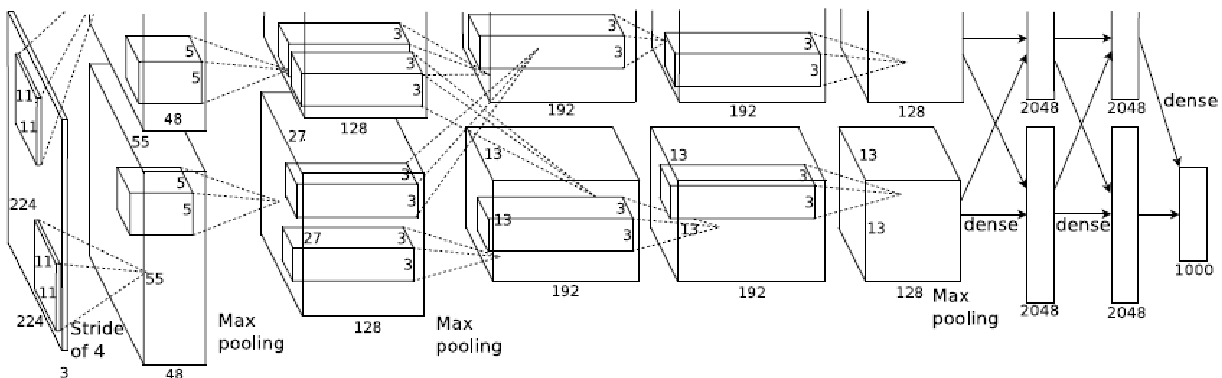
LBP	87.3%
FVF	93.0%

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



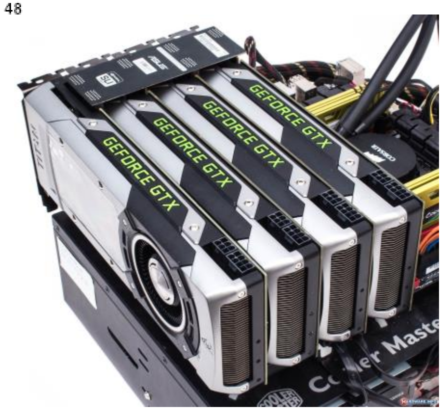
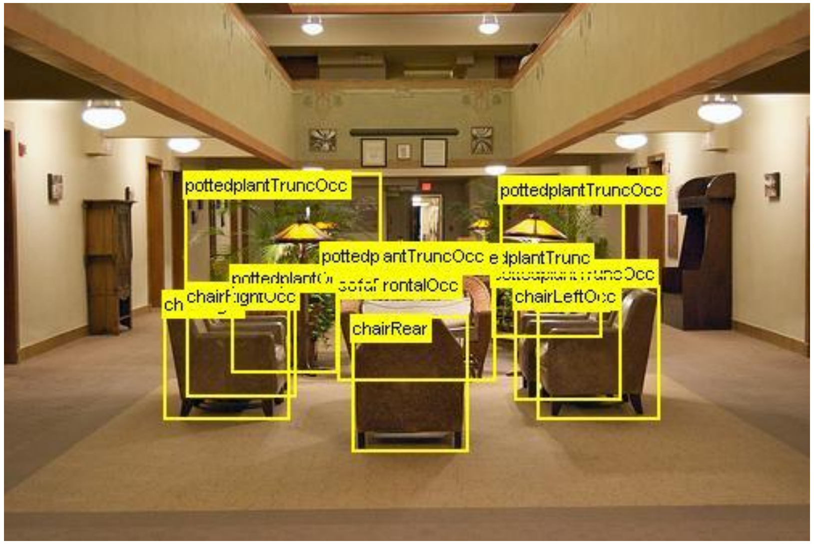


Image
annotation

Problems with annotation

- Expensive



- Ambiguous

Table? Dining table? Desk? …

Problems with annotation

What action class?

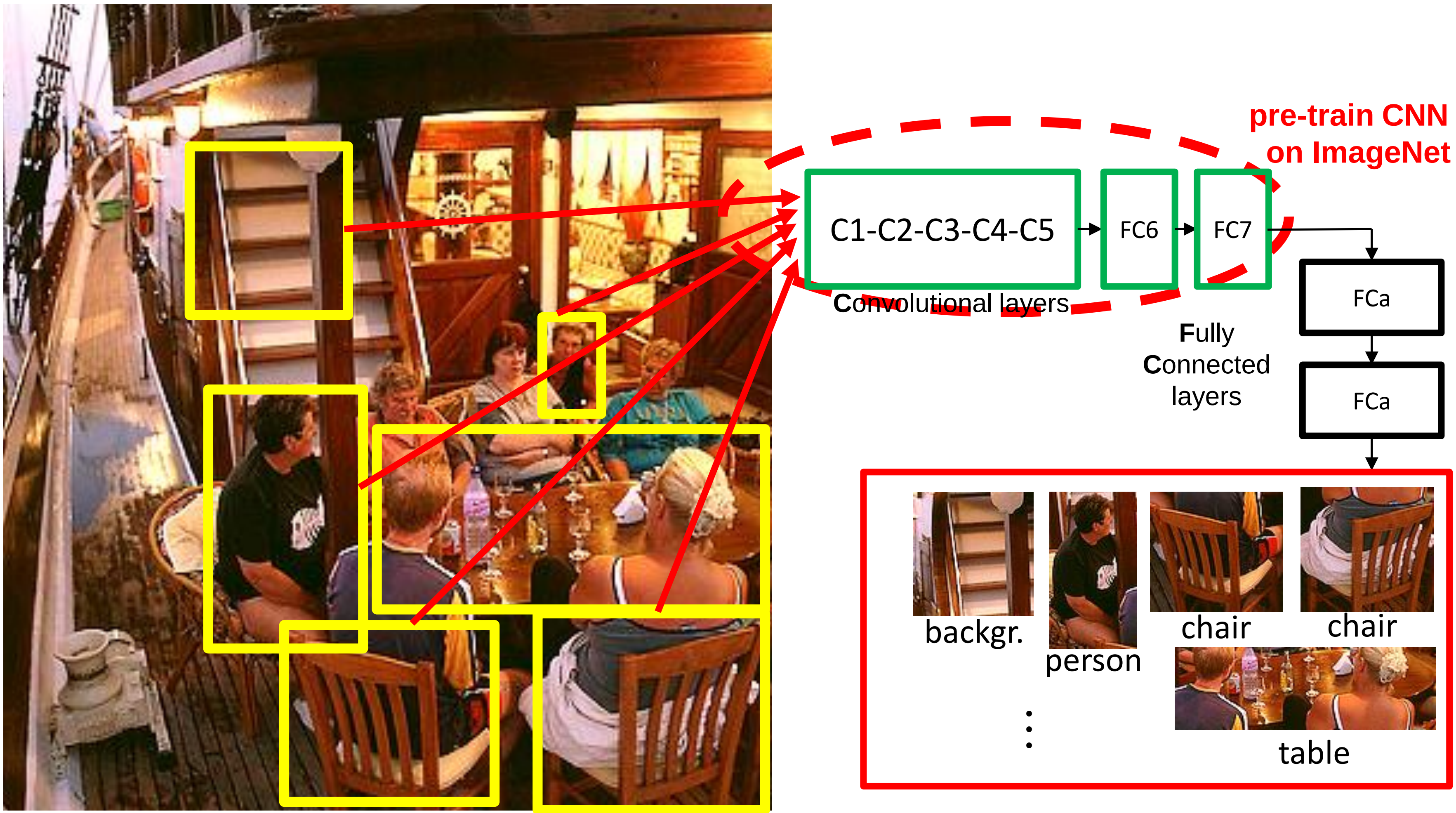
Problems with annotation
What action class?

This talk:

How to avoid manual annotation?

Weakly-supervised learning from images and video

Train CNNs for object detection

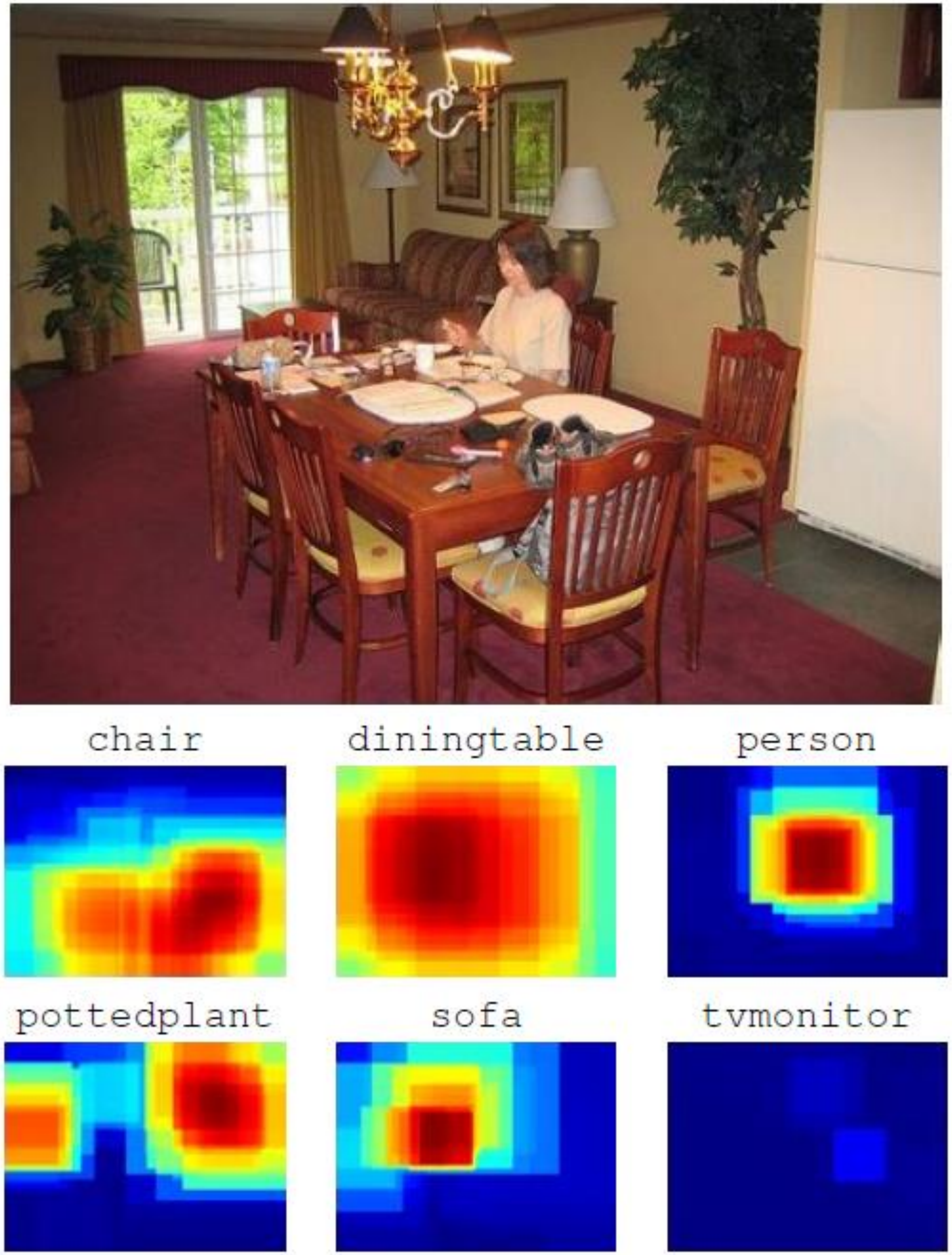


[Girshick’15], [Girshick et al.’14], [Oquab et al.’14], [Sermanet et al.’13], [Donahue et al. ’13], [Zeiler & Fergus ’13] ...

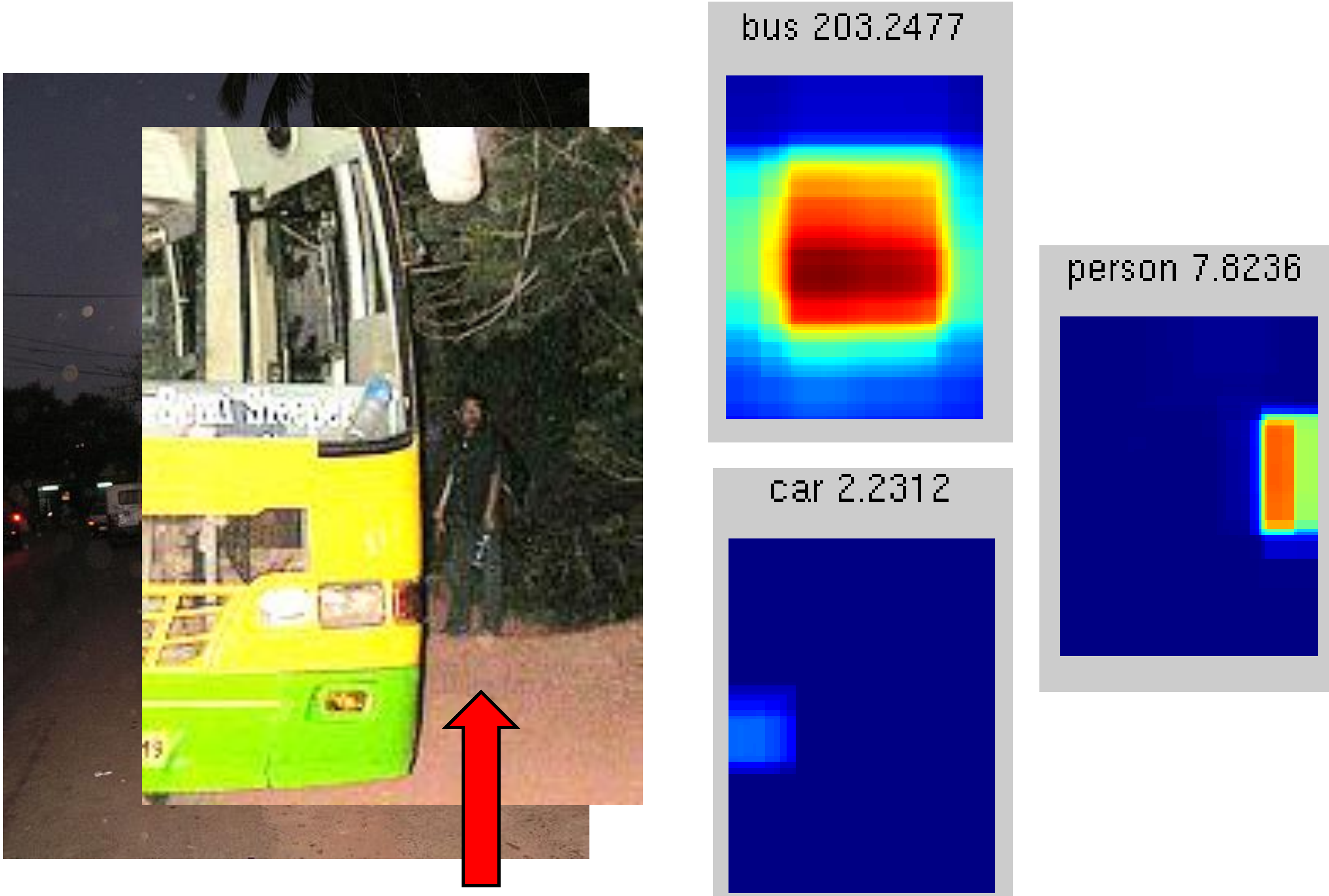
Results

Pascal VOC

Oquab, Bottou, Laptev and Sivic
CVPR 2014



Results



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

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

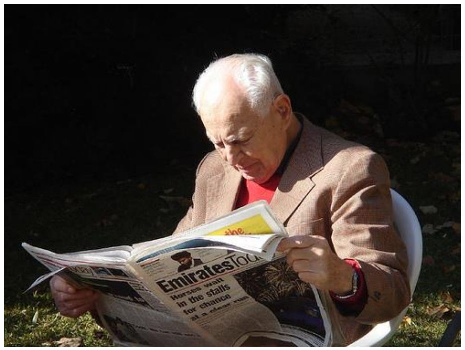
Motivation: image-level labels are plentiful

"Public bikes in Warsaw during night"

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

Goal

Training input



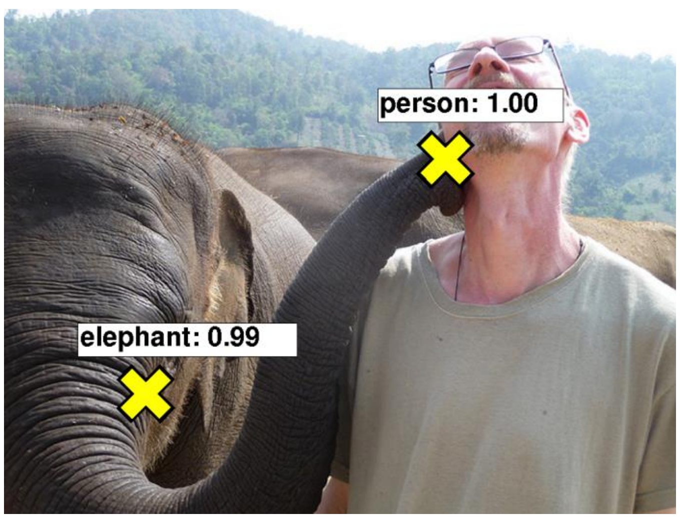
+

image-level labels:

- ✓ Person
- ✓ Chair
- ✗ Airplane
- …

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

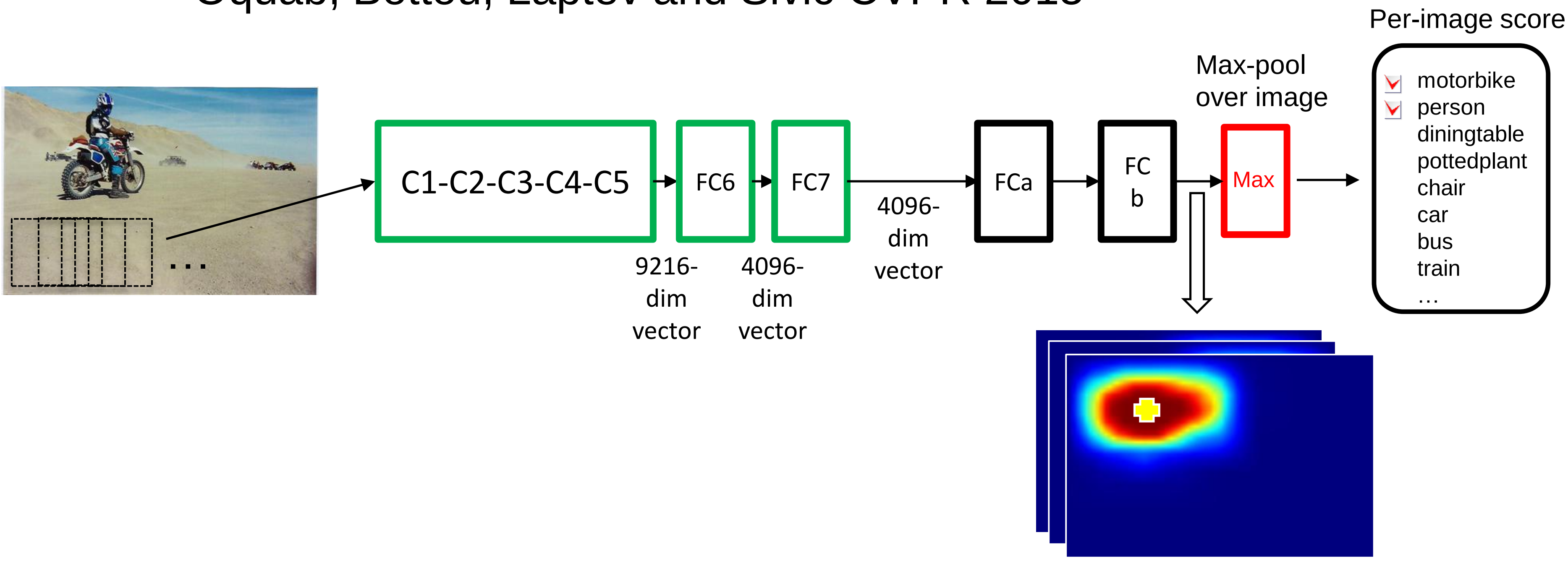
Test output



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

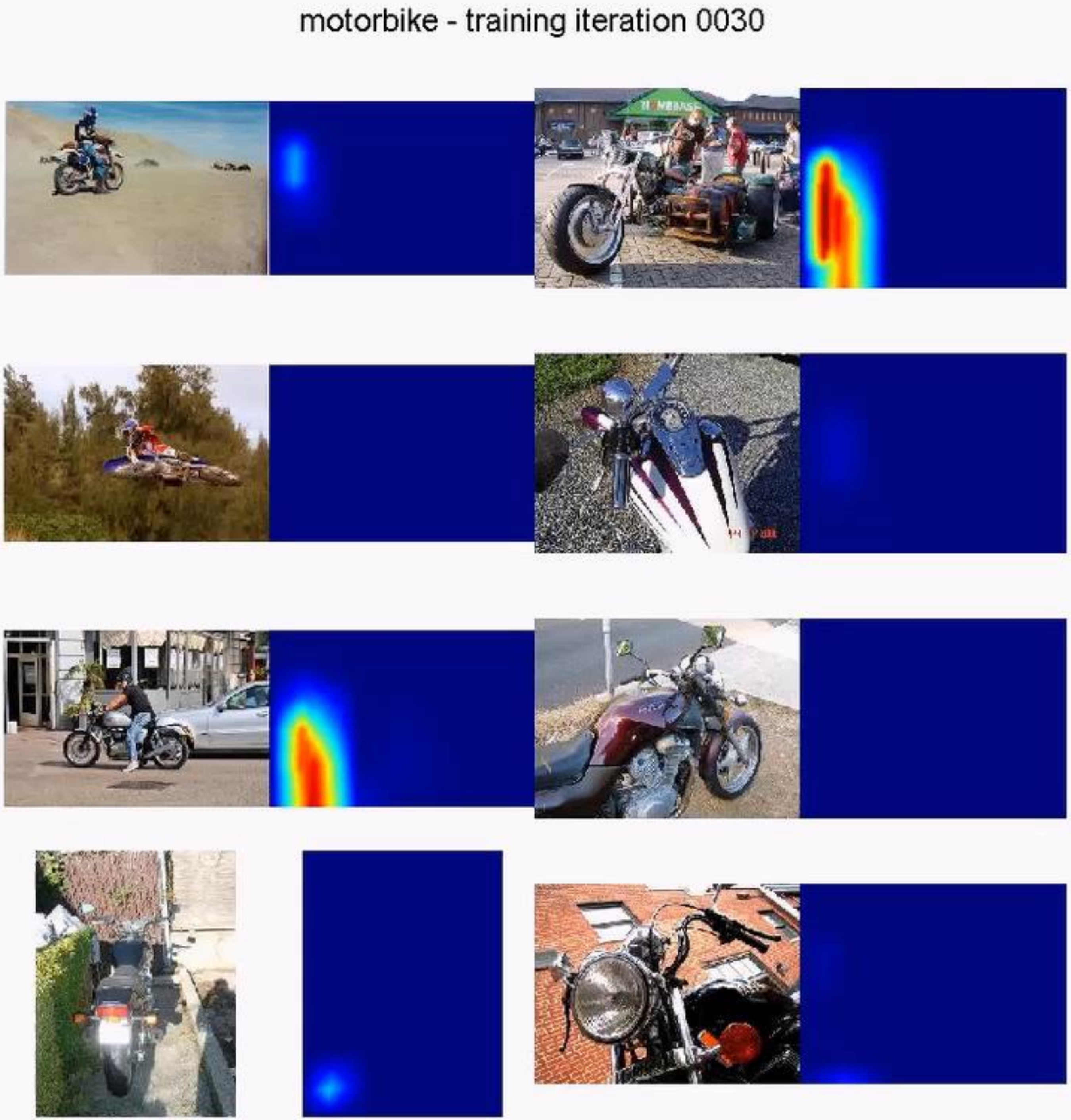


1. Fully convolutional network
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

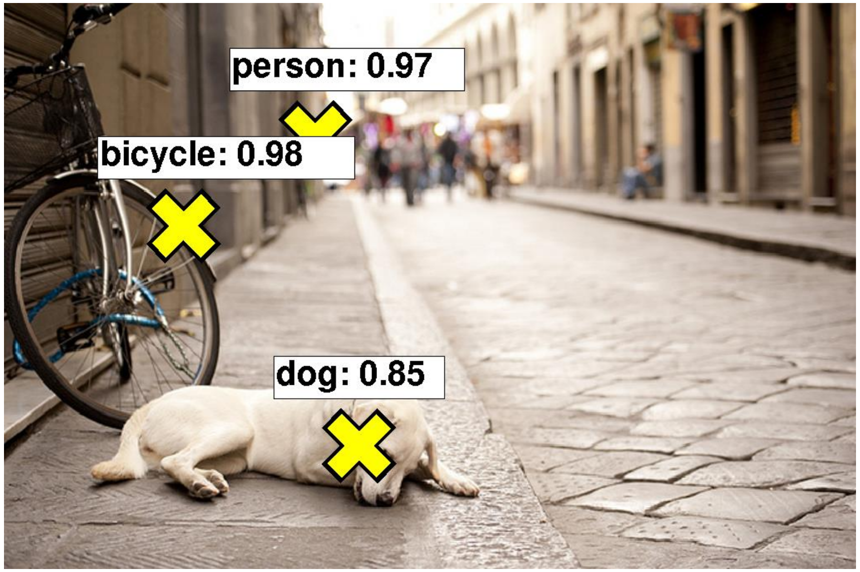
See also [Papandreou et al. '15, Sermanet et al. '14, Chaftield et al.'14]

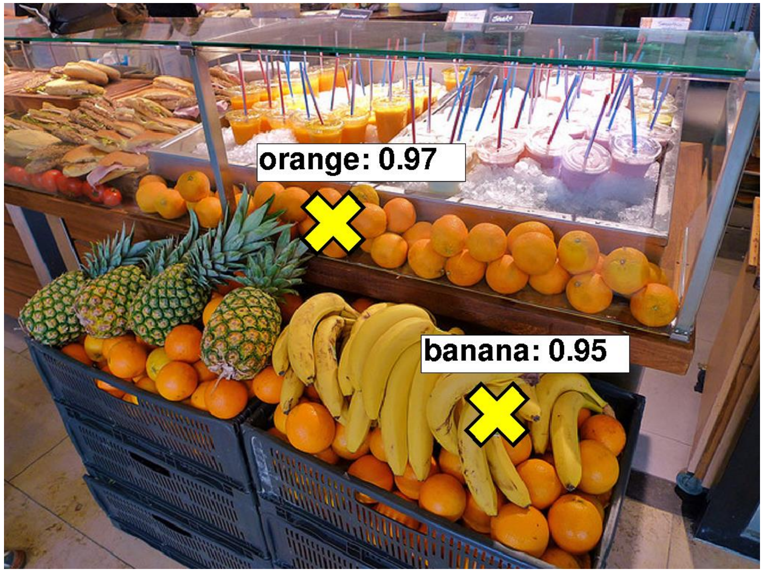
Training Motorbikes

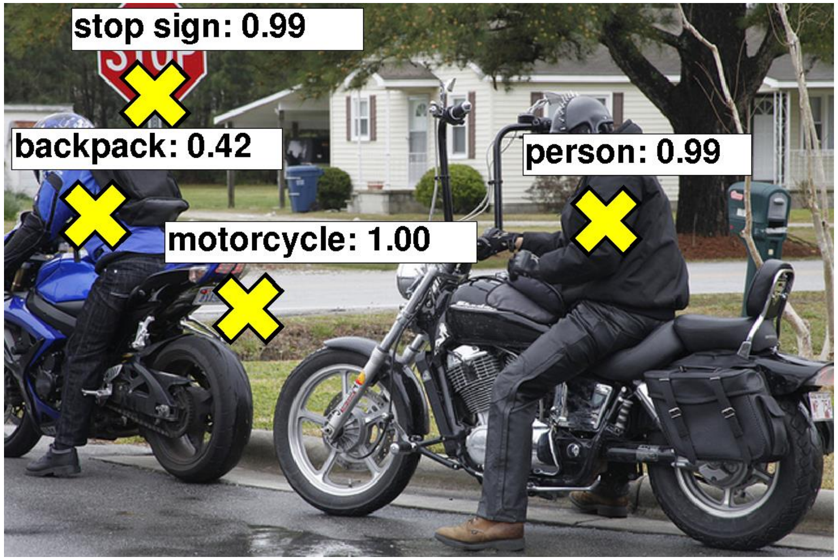
Evolution of localization score maps over training epochs



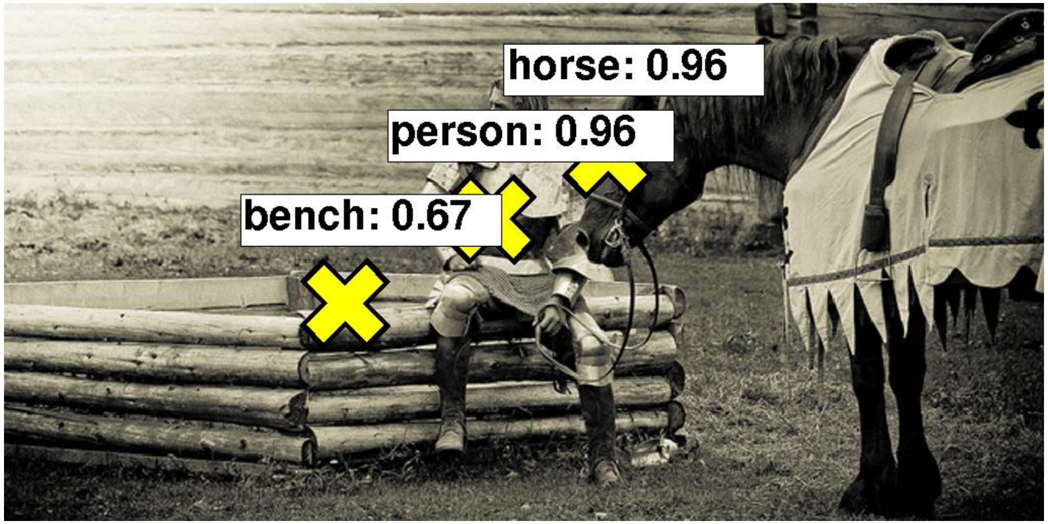
Test results on 80 classes in Microsoft COCO dataset

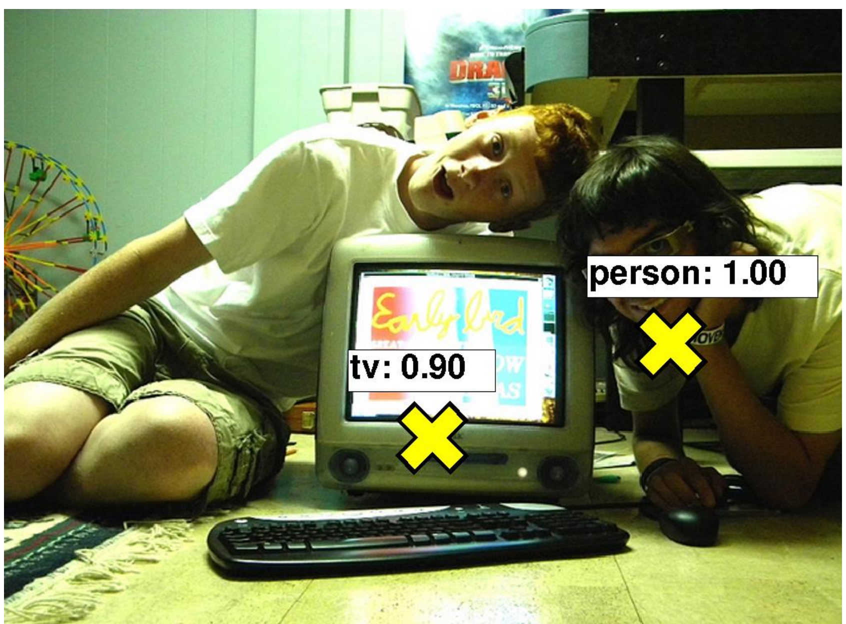




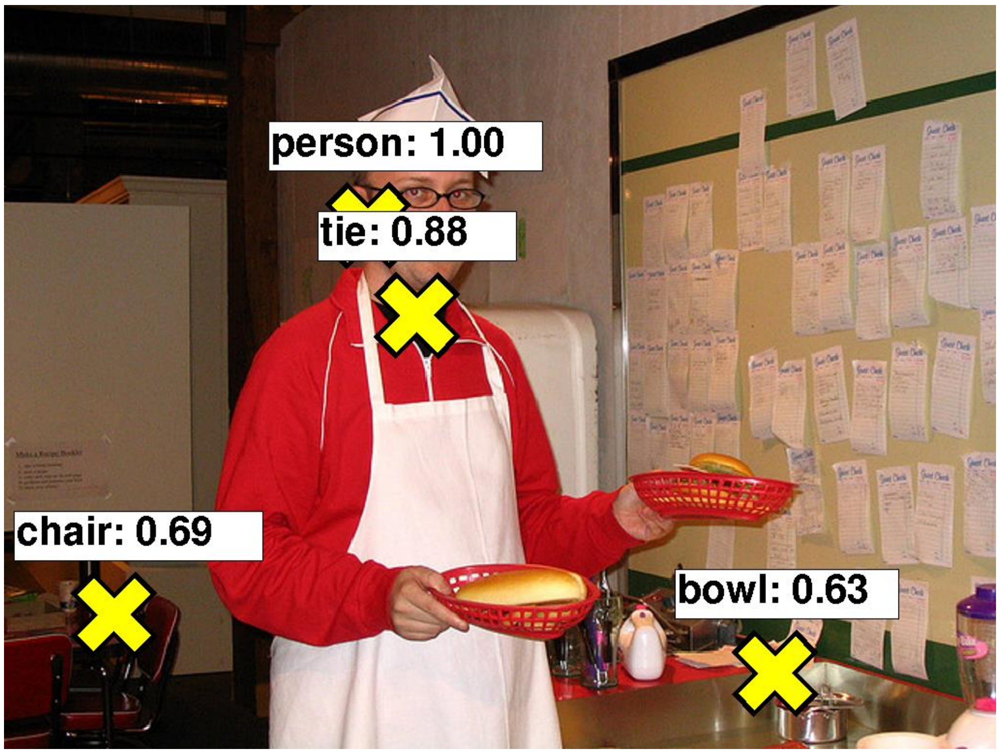


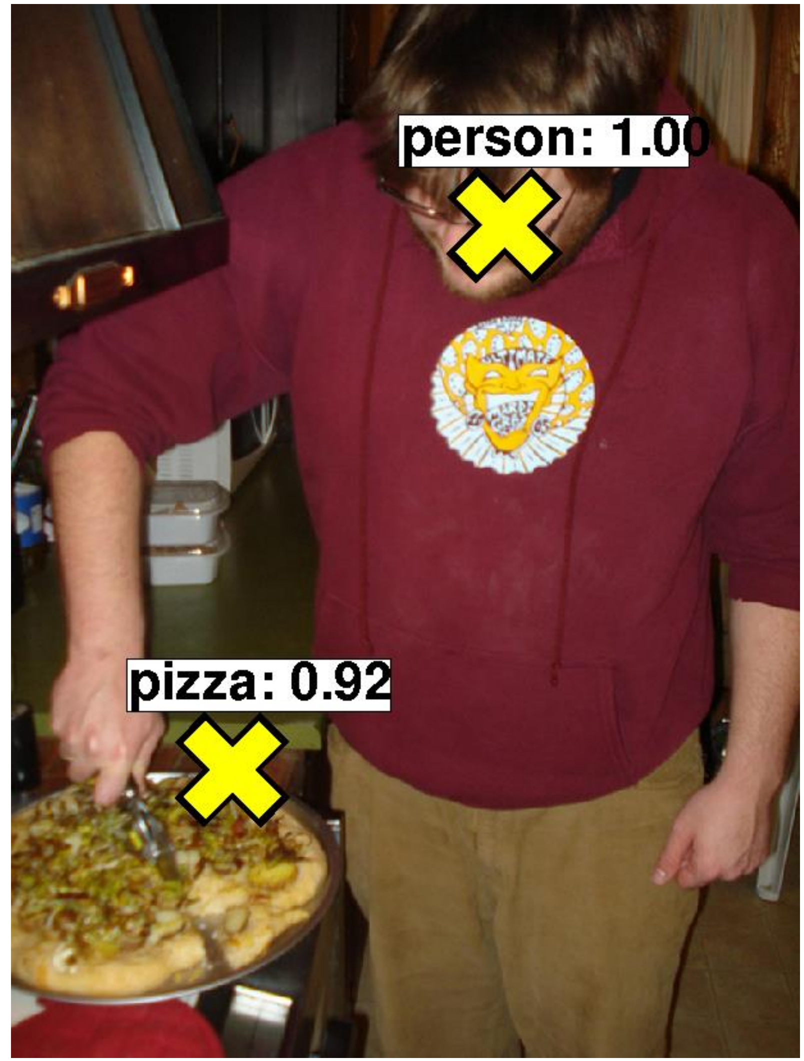
Test results on 80 classes in Microsoft COCO dataset





Test results on 80 classes in Microsoft COCO dataset

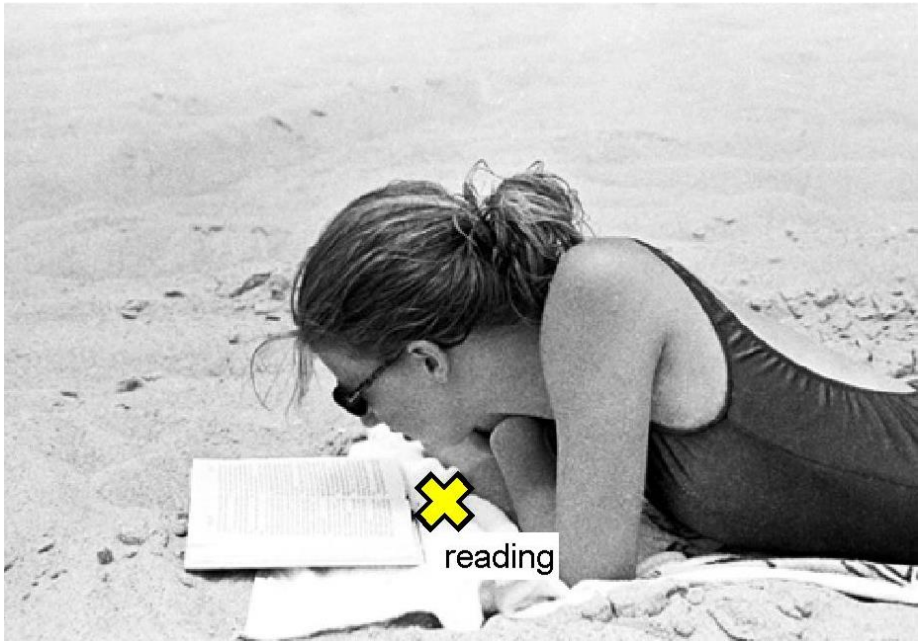


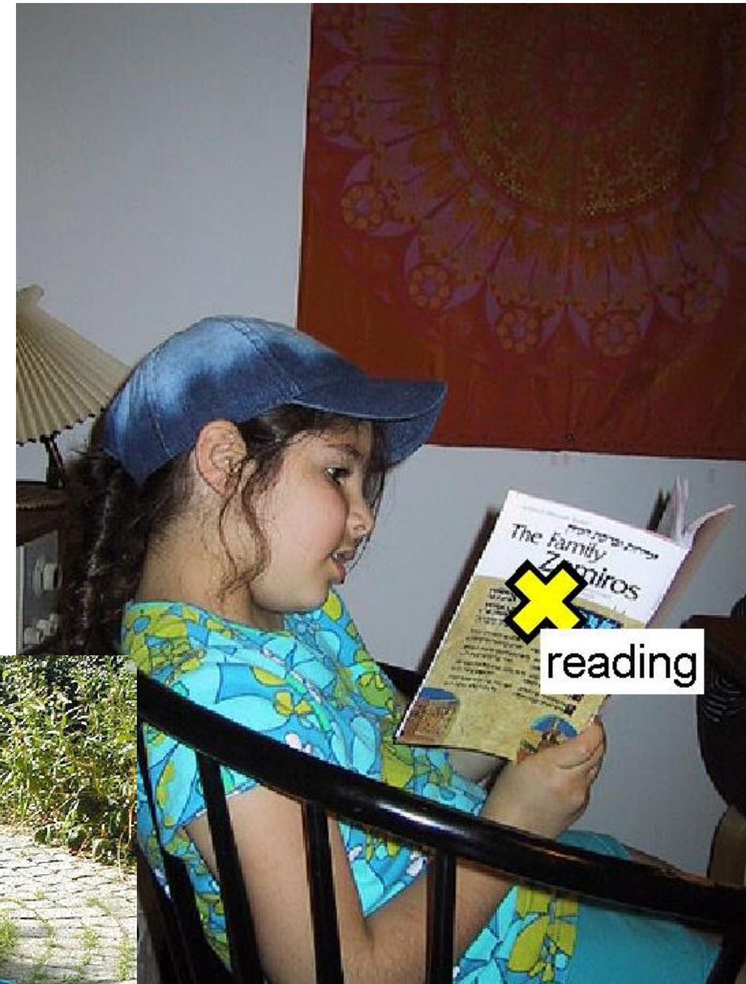


Results for weakly-supervised *action* recognition in Pascal VOC'12 dataset

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

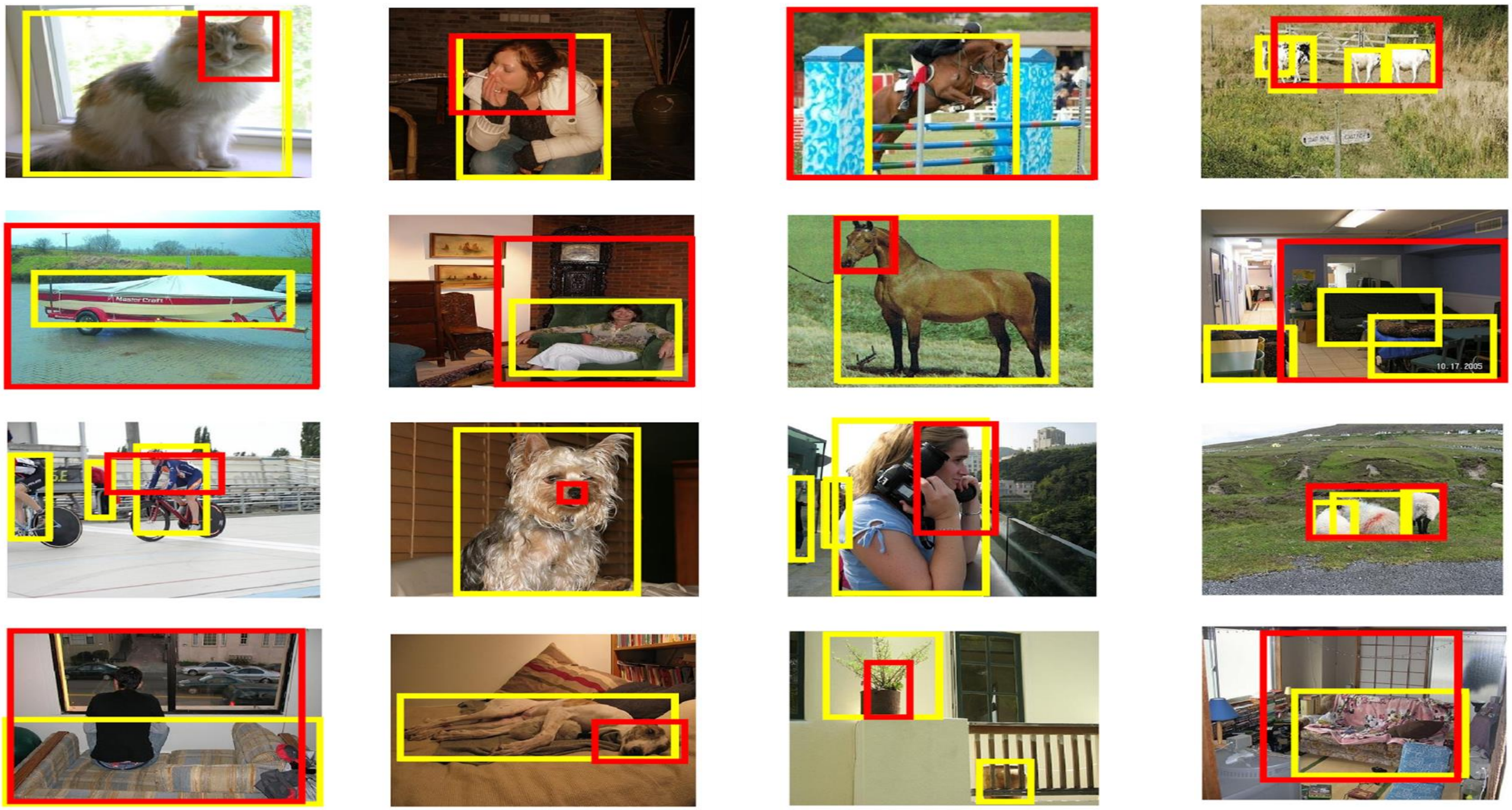
Failure cases

Context-aware deep network models for weakly supervised localization

Vadim Kantorov, Maxime Oquab, Minsu Cho, Ivan Laptev

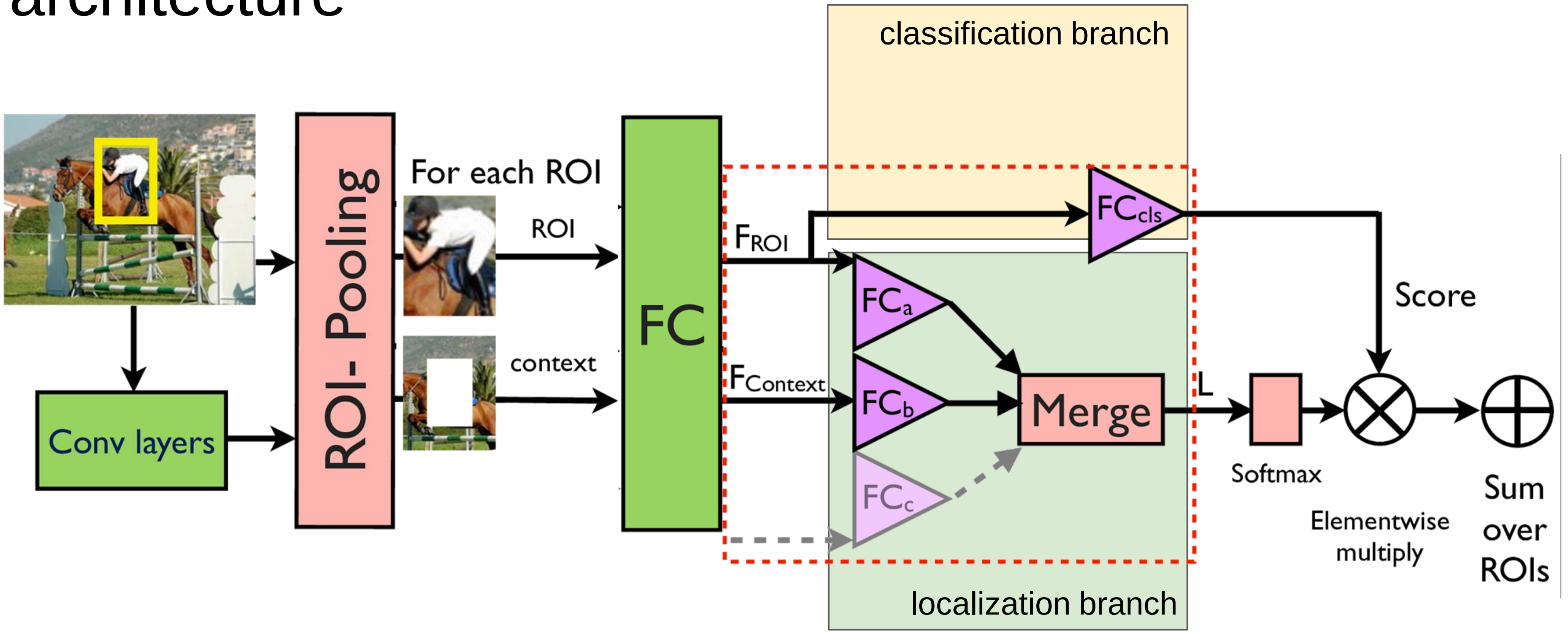
(In submission)

Problems: shrinking / expansion



[Bilen et al., Weakly Supervised Deep Detection Networks, CVPR 2016]

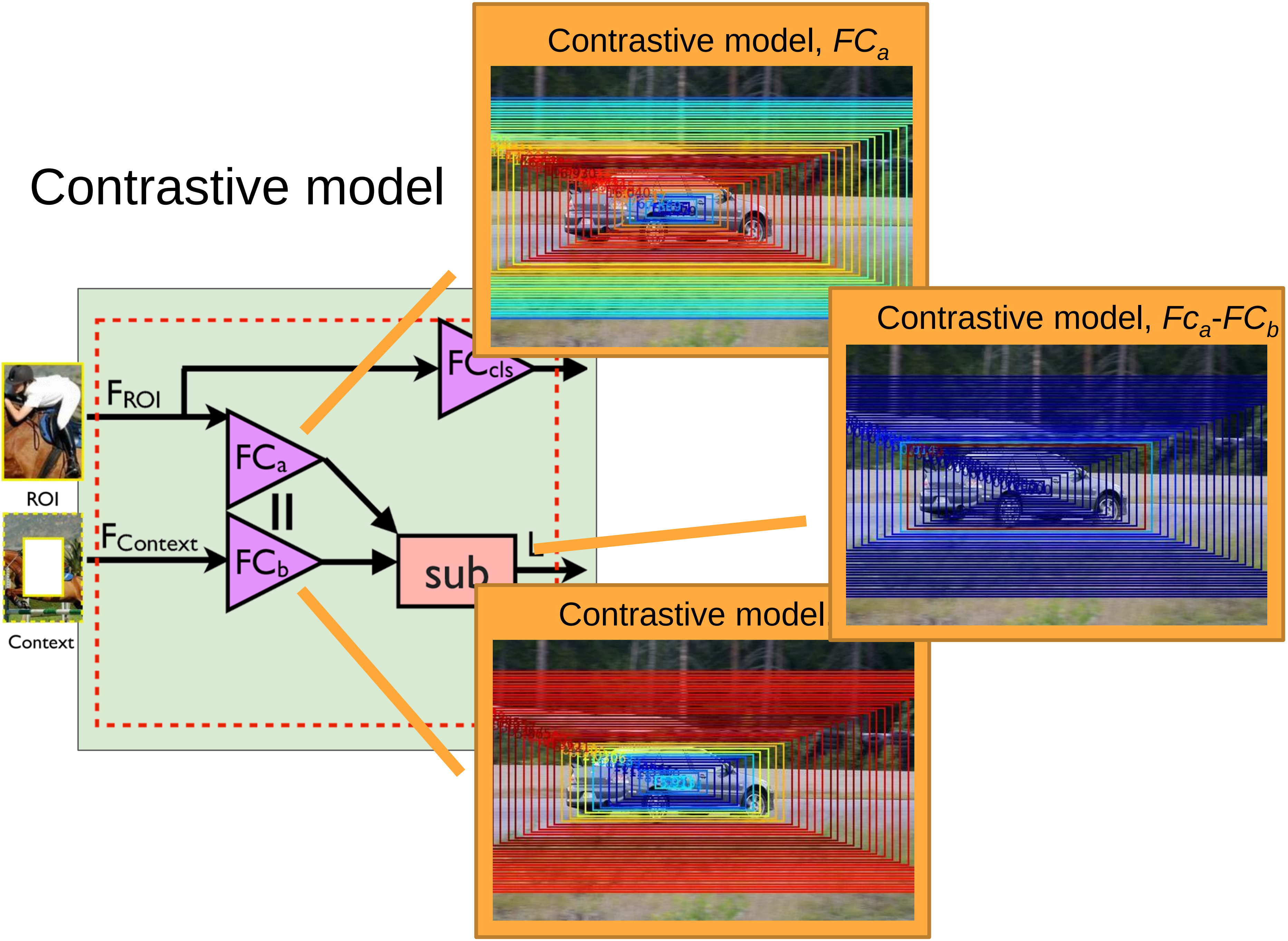
Context-aware architecture



➔ We process context around a proposal separately
➔ Everything else in the architecture is pretty much like in [Bilen et al]

ROI transforms

Contrastive model



Results

PASCAL VOC 2007

	Model	CorLoc	mAP
(a)	additive	49.42	31.13
(b)	contrastive A	50.07	31.94
(c)	contrastive S	**52.15**	**33.66**
(d)	our WSDDN-8 [6]	48.09	28.19
(e)	ensemble	**54.14**	**34.76**
(x)	WSDDN-ens [6]	51.0	30.6
(y)	WSDDN-8 [6]	∼49	28.7
(z)	Wang et al. [1]	48.5	30.9

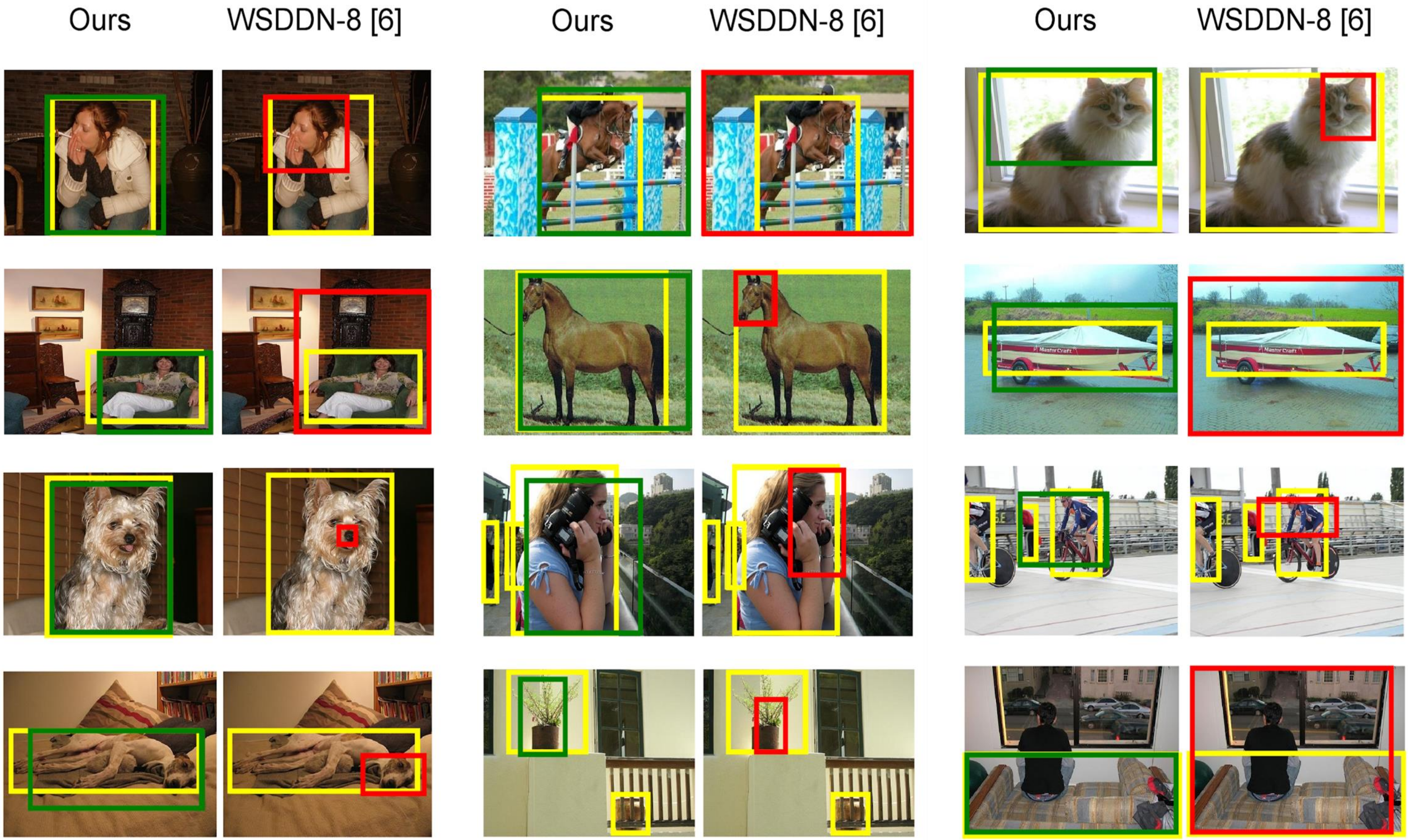
[1] Wang et al., Weakly supervised object localization with latent category learning, ECCV 2014]
[6] Bilen et al., Weakly Supervised Deep Detection Networks, CVPR 2016]

PASCAL VOC 2012

Model	Avg.	mAP
additive	58.7	32.6
contrastive S	60.6	**35.4**
ensemble†	**60.7**	**35.4**
WSDDN-8* [1]	56.4	31.2

[1] Bilen et al., Weakly Supervised Deep Detection Networks, CVPR 2016]

Results



[6] Bilen et al., Weakly Supervised Deep Detection Networks, CVPR 2016

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

Intelligent analytics for a large video surveillance systems

The goal of the project is the R&D for creating a software which provides:

- Video stream recognition and indexing
- Information retrieval in large-scale video surveillance systems and image/video storages

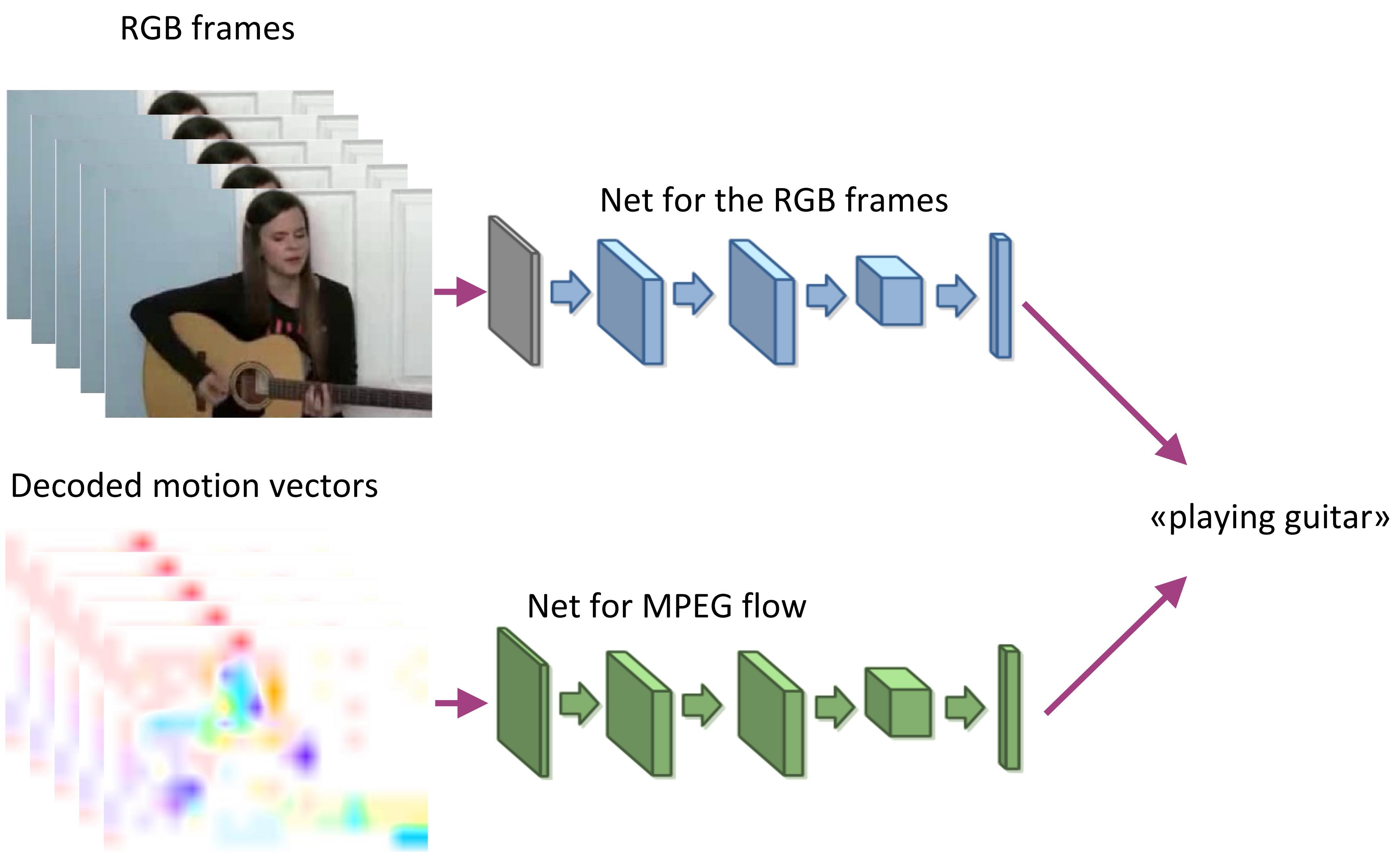
This research is supported by the Russian Ministry of Science and Education grant RFMEFI57914X0071

In conjuction with:

VisionLabs
visual recognition company

Skoltech
Skolkovo Institute of Science and Technology

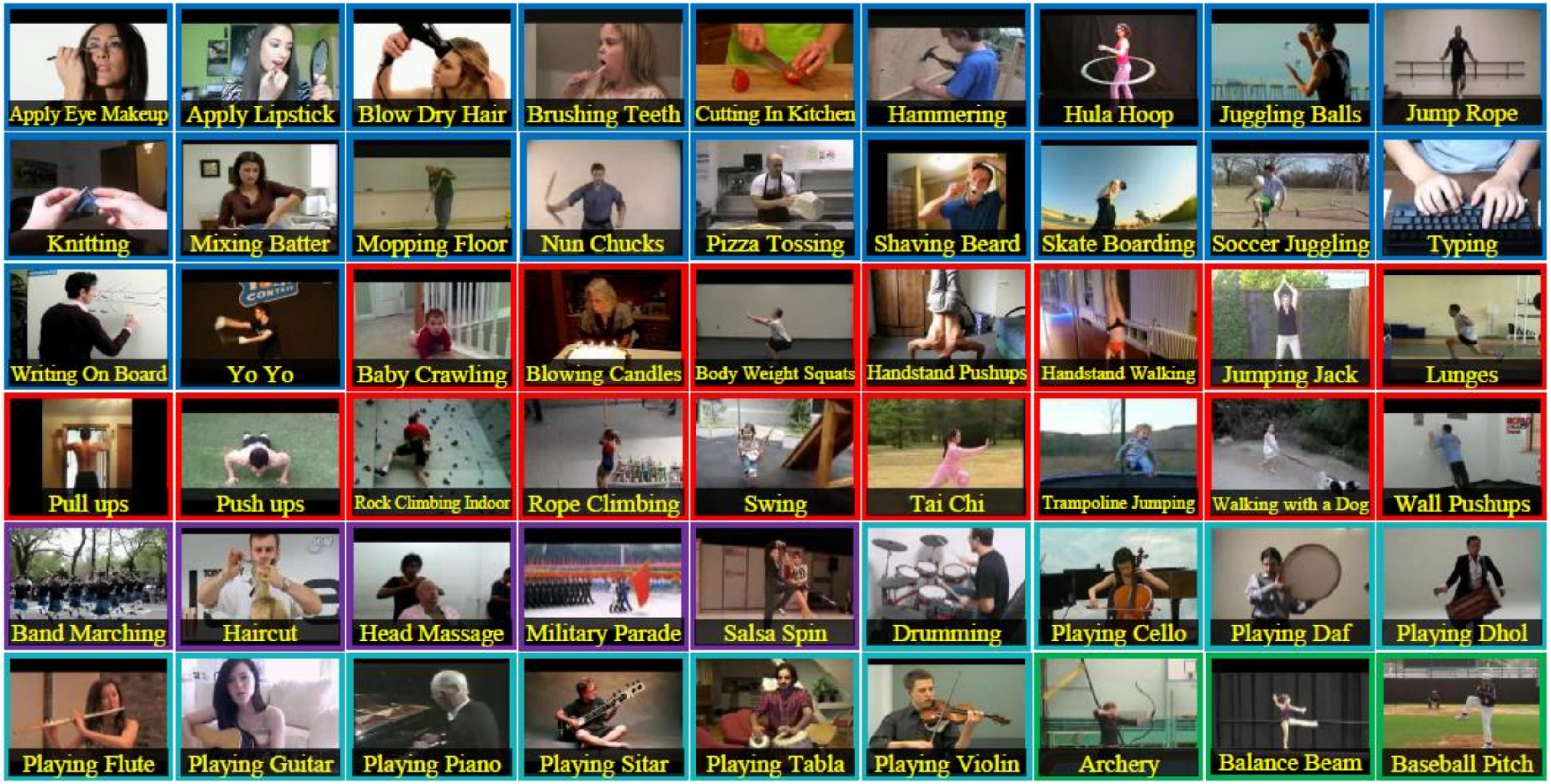
Two-stream neural network



Results

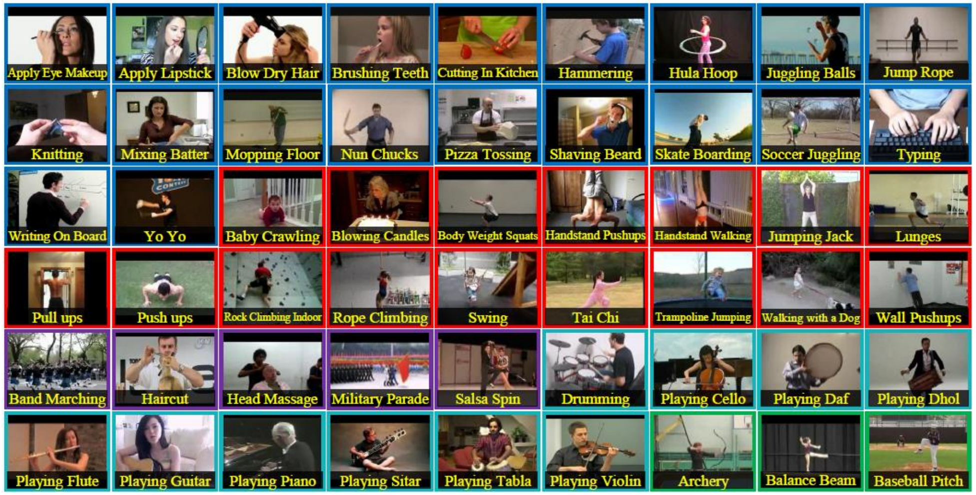
UCF101 benchmark

101 action classes
13K videos



Results

Method	UCF 101 accuracy	Speed
CNN RGB, Simonyan et al. [1]	72.8%	real-time
CNN RGB + opt. flow, Simonyan et al. [1]	87.0%	non real-time
C3D RGB, Tran et al. [2]	85.2%	real-time
Ours CNN RGB + MPEG flow	82.5%	real-time
Ours C3D RGB + MPEG flow	86.8%	real-time



How to define actions?

- Is action vocabulary well-defined?

Examples of “Open” action:






- What granularity of action vocabulary shall we consider?


Front Left

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

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

What are action classes?





open

What is the right action granularity?

person-throws-cat-into-trash-bin

Less ambiguity if *actions are defined in relation to concrete tasks*

Learning from narrated instruction videos

J.-B. Alayrac, P. Bojanowski, N. Agrawal, J. Sivic, I. Laptev and S. Lacoste-Julien

CVPR 2016

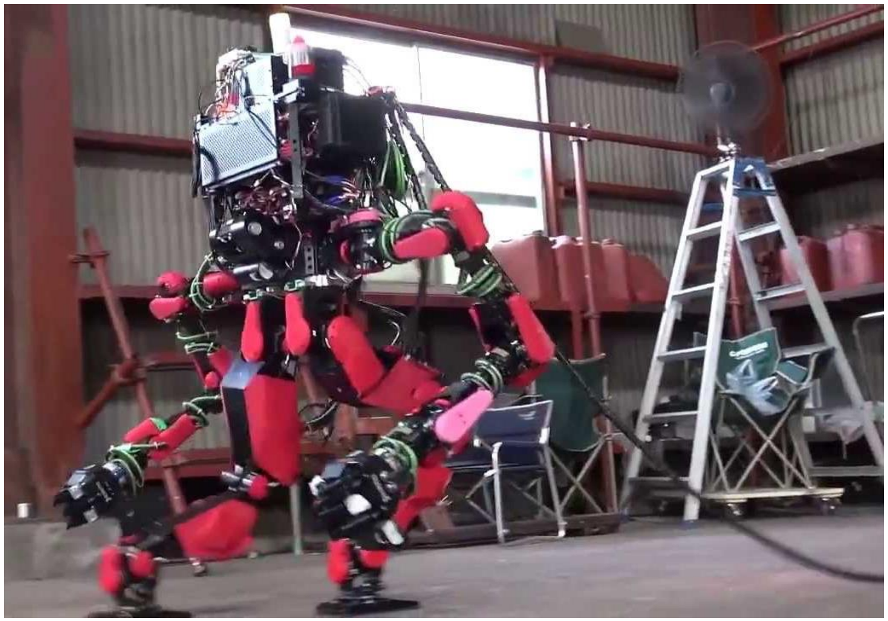
Goals

Given a set of narrated instruction videos of a task

- Discover main steps
- Learn their visual and linguistic representation
- Temporally localize each step in input videos

“How to” instruction videos: changing tire

Motivation



[Darpa robot challenge]

[Microsoft HoloLens]

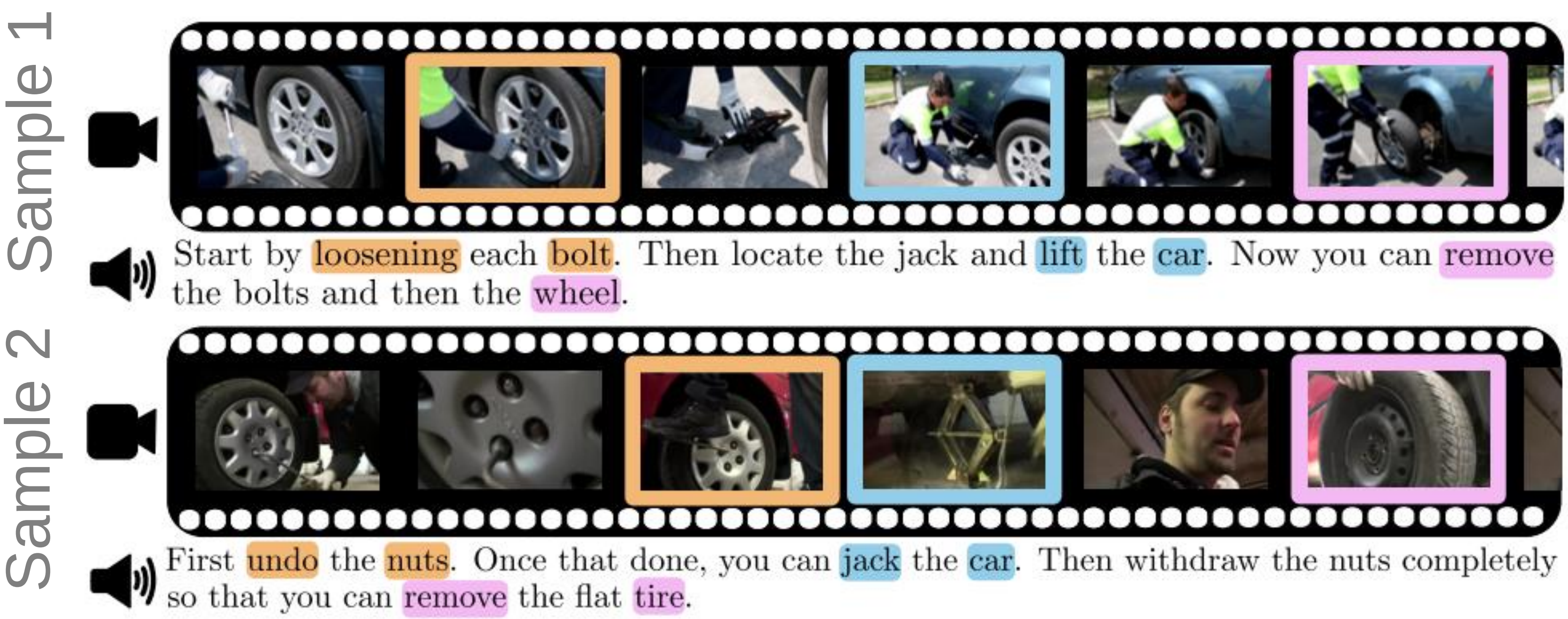
Learning from Internet for robotics

Personal assistant

Why is this difficult?

1. Variation in appearance (viewpoint, tools, actions, …)
2. Variation in natural language narration
3. Variability in temporal structure of videos

Example Task: Changing car tire



New dataset of Internet instruction videos

New dataset of Internet instruction videos

New dataset of Internet instruction videos

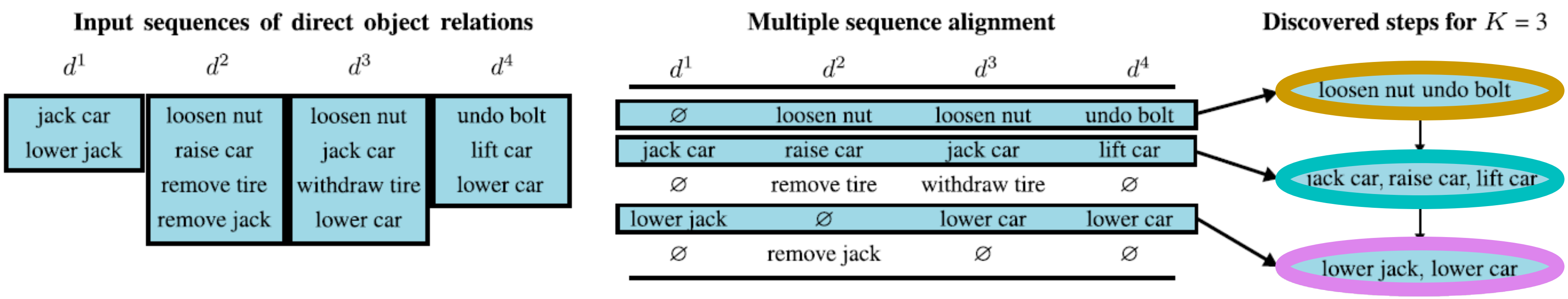
New dataset of Internet instruction videos

New dataset of Internet instruction videos

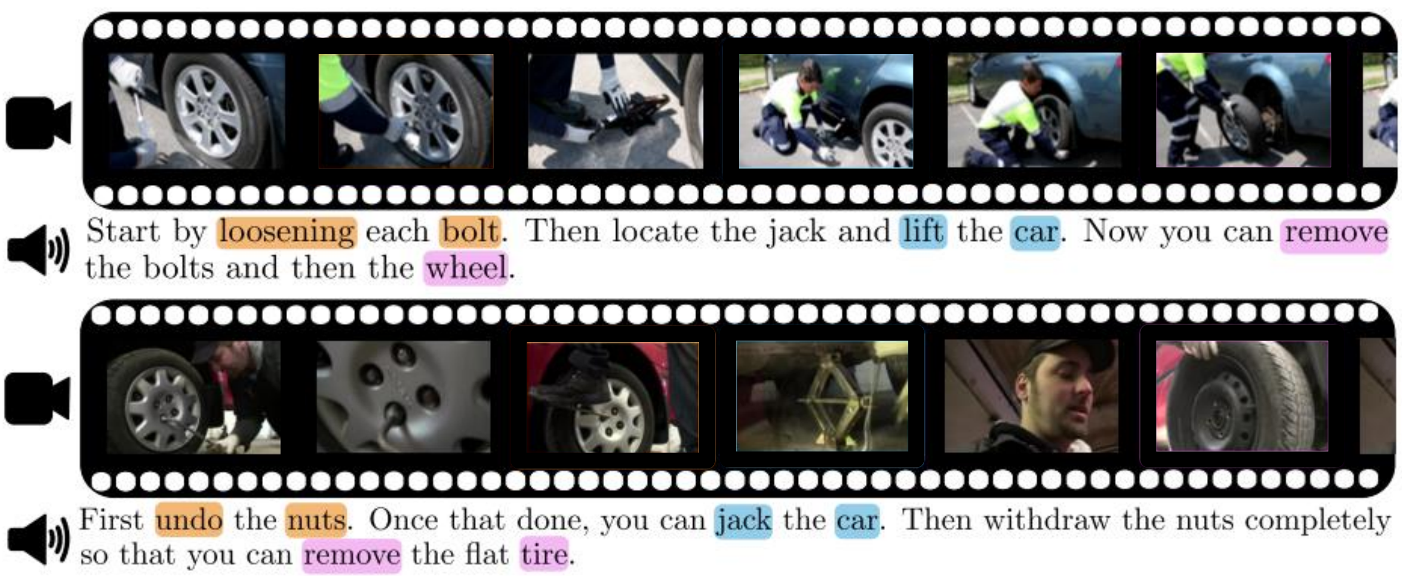
New dataset of Internet instruction videos

Approach: two linked clustering problems

1. Text clustering into a sequence of common steps

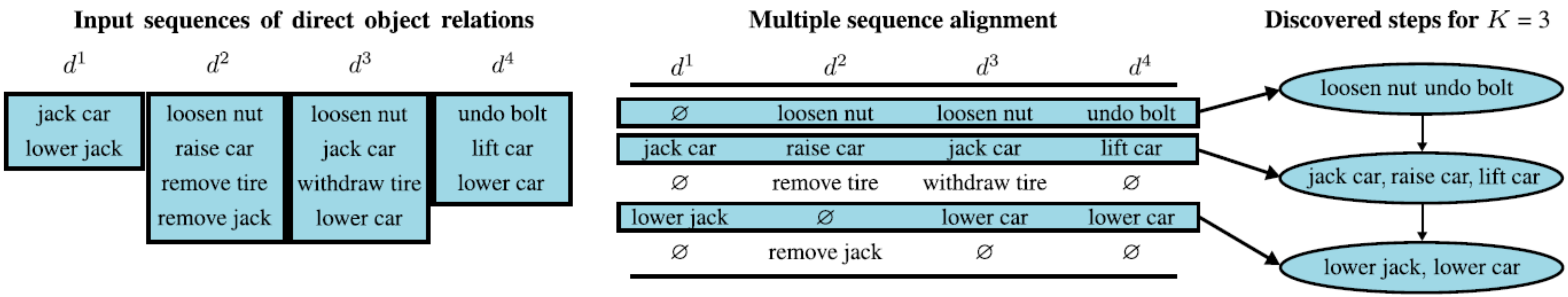


2. Video clustering to localize the actions with text constraints



Approach: two linked clustering problems

1. Text clustering into a sequence of common steps



2. Video clustering to localize the actions with text constraints

$$h(Z) = \min_{W\in\mathbb{R}^{K\times d}} \underbrace{\frac{1}{2T}\|Z - XW\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\lambda}{2}\|W\|_F^2}_{\text{Regularizer}} \quad \text{s.t.} \quad \underbrace{Z\in\mathcal{Z}}_{\text{ordered script}}, \quad \underbrace{AZ \geq R}_{\text{weak textual constraints}}.$$

Discovered temporal localization [TxK matrix]

Representation of video chunks (IDTF,CNN) [Txd] matrix

Linear action classifier [dxK] matrix

Temporal constraints from text

[Bach and Harchaoui'08, Xu et al.'04, Bojanowski et al.'13,'14,'15]

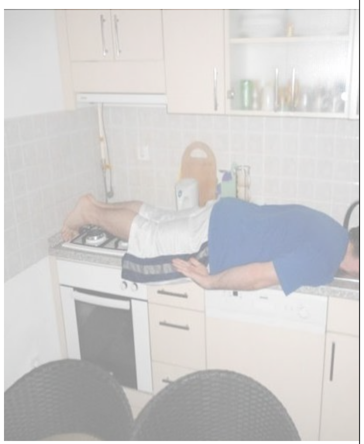
Changing a tire

K = 5

Future challenges

Future challenges

What is unusual in this



What is unusual in this scene?

tention of this person?