

# 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

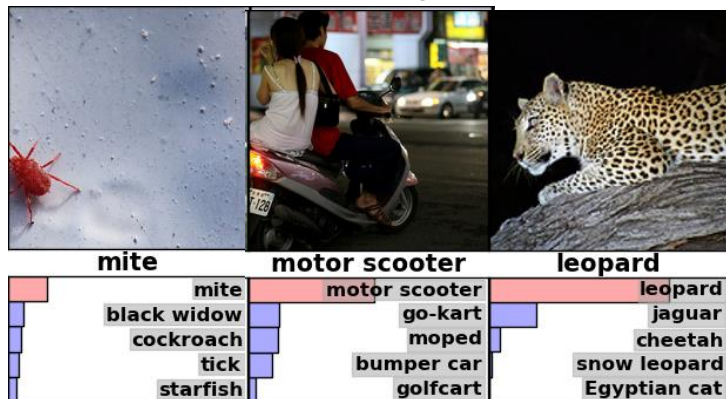
08122635252121314133210507102023222326333642453122202526231614141320343524131413131328231714091819212527212744423326363323191924333640424044352  
08133242231010244837260708101827272228343748543631272928231817161440615352181515151417261108071121242930253250473431383218192631263139394940322  
09112942230507164739281118182026262630364853544234293027211816161533595448171415151414191409070815252529253754473536434024222531243135324540372  
09092641394046405039292129282727272834405055573016202627221917161517272920141515151514152013080912252628223451463436403217202532223335334436311  
09102343545752454839301412121528282835374752552915182527273226212020181818181918171817171814091013232628223654442735453622232532283332385341401  
070918394518312238403214040813282828313339465235161825292633302625232323232628293131282413131208222727223648402735392916222432334043434748441  
06081538421427223037311706091228303130343945487117142225243033302927252627262830323533292212171517263024203647382642453120252631243234454538280  
0812143338171411263632210509102730312938404240311012192425171319262425293025262728322923161316162331292318324033223740271623253025331465448340  
22251628314170112035312005060926282824434036301608101719171204172724223524100813112835140612171823303739303337251843442619262729243331394142270  
23261923312528091531292205050922191318272929181413172221191503182425162517050912153538150506192321215771272936342939402217232828213131535042180  
18191819281603071123201506051021110318232319161314161618191504111823093114000615122831100404233129222628232933293546412120252927172927424747190  
05182414181606091223161411172022150315131513121209121212161207061122142915040210072333100404211935252422212223243343402617172733342933473833100  
05132513161614171918161507131619180406131607091515161721201711071031132617020410092734110508111024272927252122203450393741261719364142504837100  
09102615151211141515161506111413140710181918182318101728181921111332081410040708093429070508030112231907251826263148403638303228181736363630070  
11082417141309121313141412151719161109101010121508041322201917101635110505070805083313040722230804131506211830303338303535252326313340453821050  
57222117091417221622211625232020192014132019100500000307140602051313040304060605153614091114171904020205181215042824212632312624233635364022050  
48241320091319351719191524272425262922182212030201030308090505081308020303040405121412141004050802090804071111081711142330282926263739363216301  
08041022091312352328210907101410060509121109040708090907060805060707090806050504070703060607050604101610040407685103072134332722213237383617363  
0504041306070834221908030100020205050813107131914110810071209020408131210111110080407080825000508041519050211856703032546443311162631343116362  
020508030304071009110400020100000605081722381611040201010203040506091111111011080201020710100107090571330403201208043129233312122425272118372  
08050503040505030405040000000404000203071119030503010202030305060708090911101111030101021315030913041411020403072005033322131712142725241616262  
05020103041514030608131208061114040101000104050804020203040506060707080913101117140500021507011220041410020203061804042912081511202522241416232  
04040304041616132019131317251214150302030306050504040404050606060707101918121219261101010901011623070802030203071303052207050508162308151021242  
0506081010151705060201030918131319070303040606060606050405050506071016263020141215382606030200001517020203051417121103051014120706202204101124262  
14171714091406021401020305060503121201020406060606060506060605071223363823111219373316090503060905020406051819150702020304090205231603061424333  
1014122409150304100404050503020104030202030606060606060605060718102034472917172335352010060607070905050507121706030001010100006210305121335623  
13090707081006040303020203080400010303030305050606050812050606070202233730181725475025090605061215070606091107000302010202010002060308160870996  
15140907040302010101010101080400020304050404050607050504030606020311291020191725444319090606071214120807161708010101020202010000000108131162976  
21090302010101010000000000010102030405060606060606060911103040709030402030717192227231613110809081012110825351606000002020100000000005061525331  
080212070000000000000010203030405050506070708080807081732230408290800040608111825354134251711100906070909294219110400020201010202020100000610050  
0102050301010101020302030406060607080808090907040708180104131100021534242146444437272015110907050707284424160802030100010102020100010101071  
0206020403030508120705030405060807080909100909090604030202010516010517130910444945363124191311080506051843232614074010000000010101000101010  
0002091107040711130904061006060707080809101010100906030302020206161513021210044757494135282216151306050513392126241009040302000000000000101000  
00000305050604040203020516070707070808080910101008060202010103180425350204001954535144372923181216090405092933262612090702020101000000010100000  
00000104050603000002021614050608060607070809100906060302020101270903040101012753494947393123181210150705041533362208130704020101000000010201000  
00000001020401000001001409020303030404040405060704040202010101130601000000010936373535292116130906080703040916301604130504030101000000000501000



# Recent Progress: Convolutional Neural Networks

## Object classification

ILSVRC'12: 1.2M images, 1K classes



Top 5 error:

2012:

<i>SIFT + FVs [7]</i>	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%
<i>Human</i>	5.1%
ResNet	3.6%

## Face Recognition

LFW



--2013:

LBP	87.3%
FVF	93.0%

2014-2016:

DeepFace	97.3%
VGG	99.1%
<i>Human</i>	99.2%
VisionLabs	99.3%
FaceNet	99.6%
BAIDU	99.7%

Accuracy:



# How does it work?



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

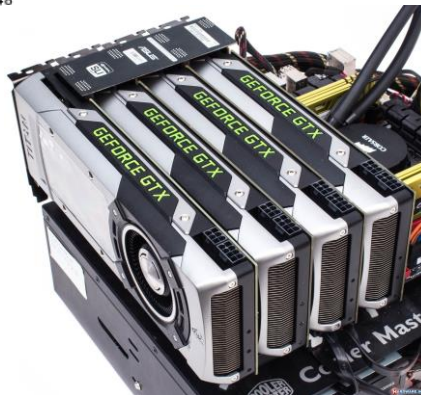
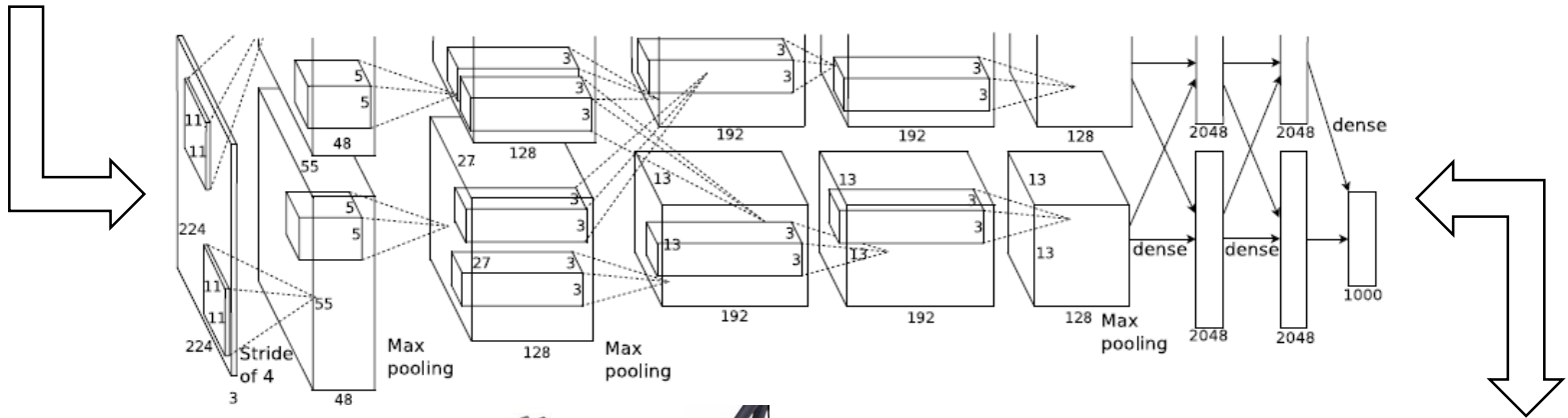
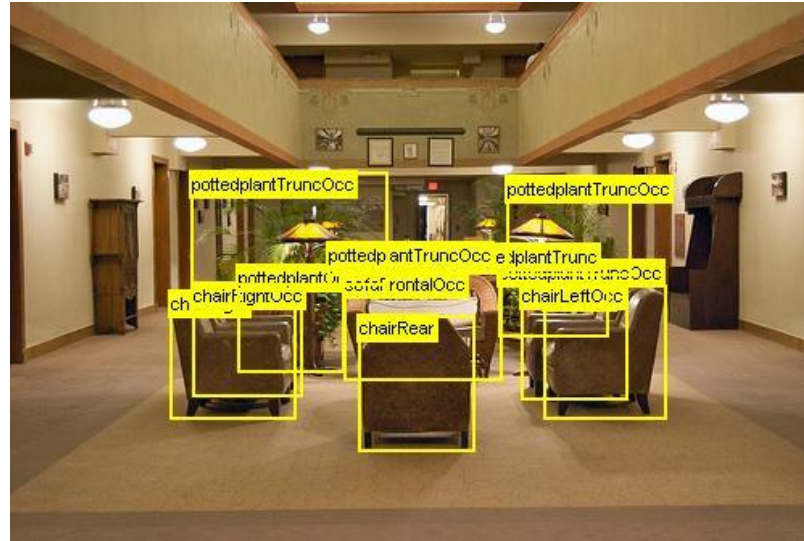


Image  
annotation

# Problems with annotation

- Expensive



- Ambiguous



Table? Dining table? Desk? ...

# Problems with annotation

## What action class?





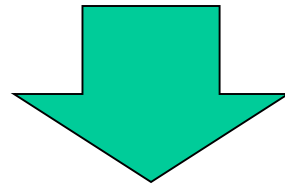
# Problems with annotation

## What action class?



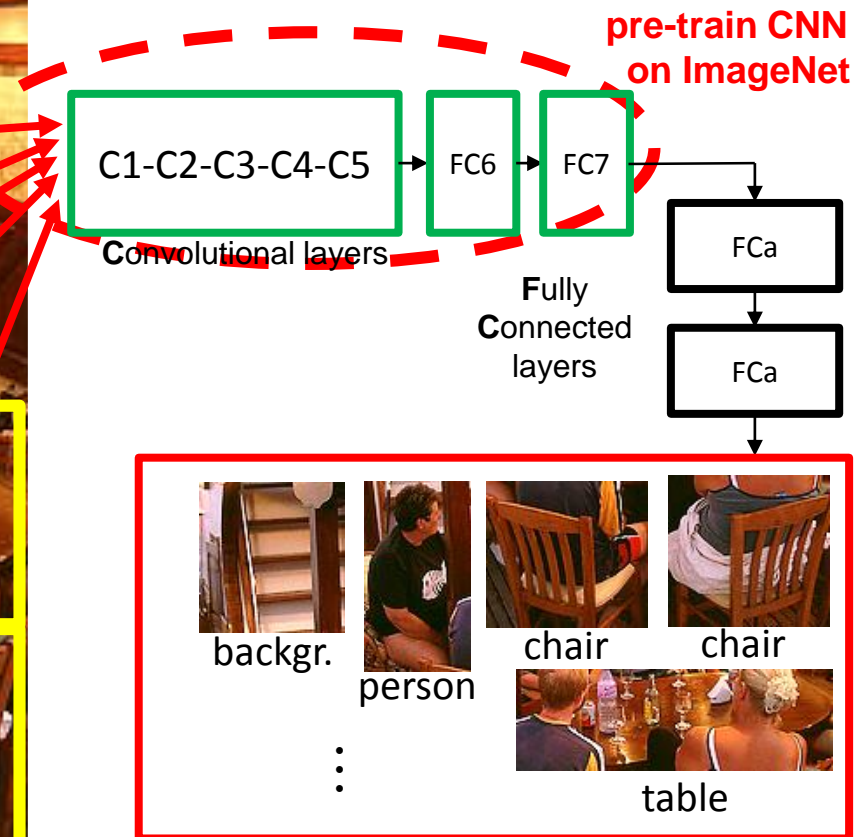
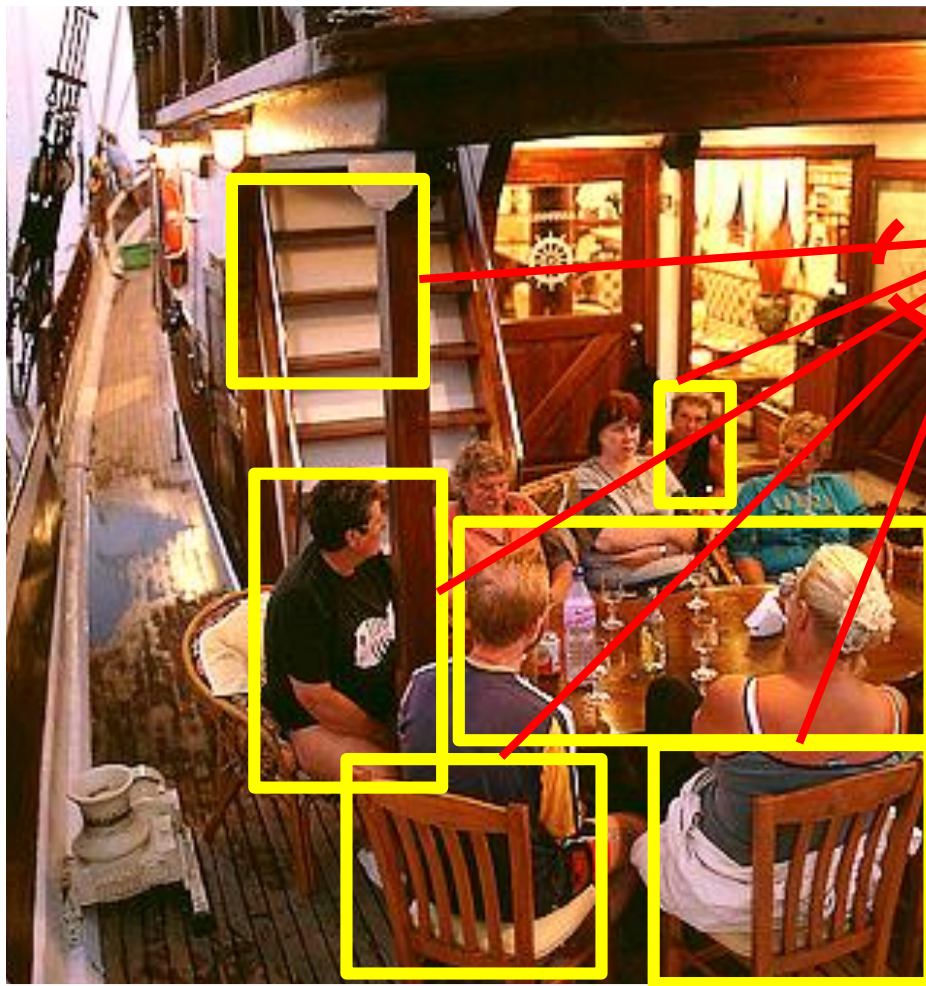
**This talk:**

**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] ...



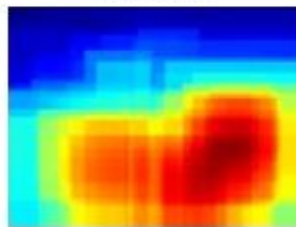
# Results

## Pascal VOC

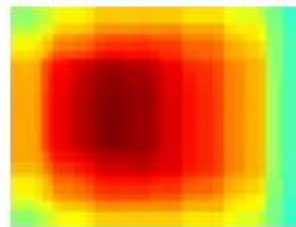
Oquab, Bottou, Laptev and Sivic  
CVPR 2014



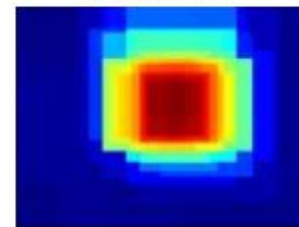
chair



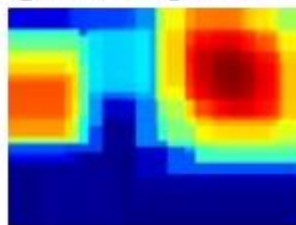
diningtable



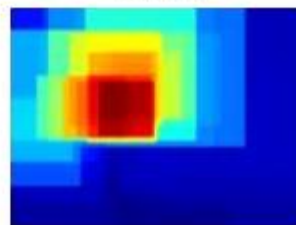
person



pottedplant



sofa



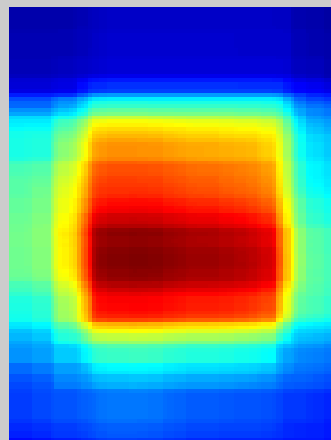
tvmonitor



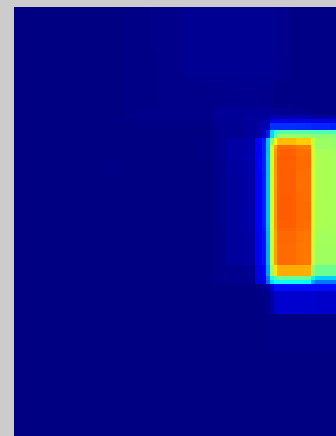
# Results



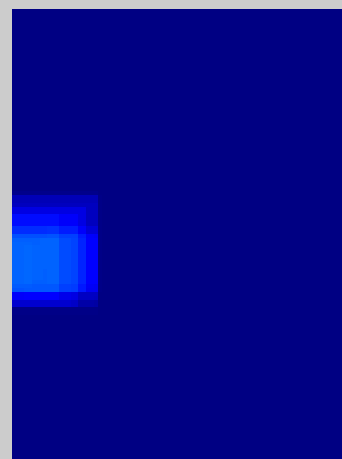
bus 203.2477



person 7.8236



car 2.2312



# 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



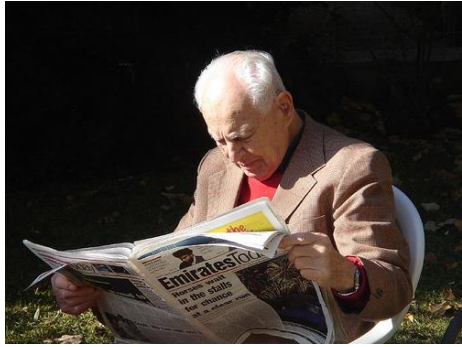
“Public bikes in Warsaw during night”

[https://www.flickr.com/photos/jacek\\_kadaj/8776008002/in/photostream/](https://www.flickr.com/photos/jacek_kadaj/8776008002/in/photostream/)



# Goal

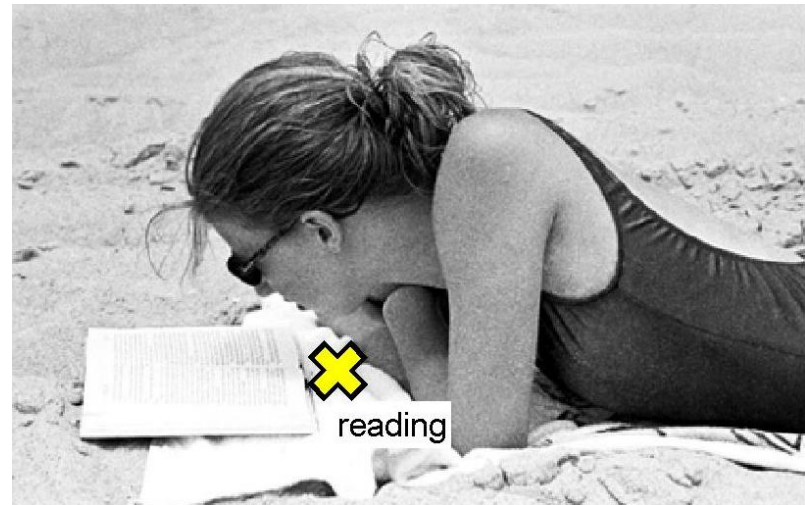
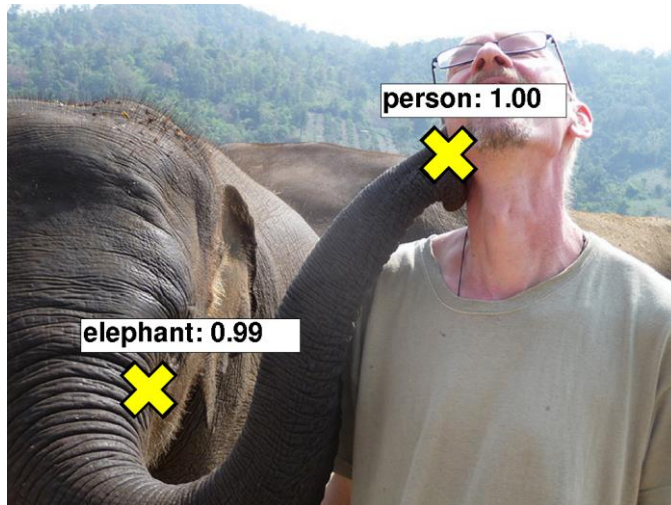
## Training input



+

✓ Person	✓ Reading
✓ Chair	✗ Riding bike
✗ Airplane	✗ Running
...	...

## Test output

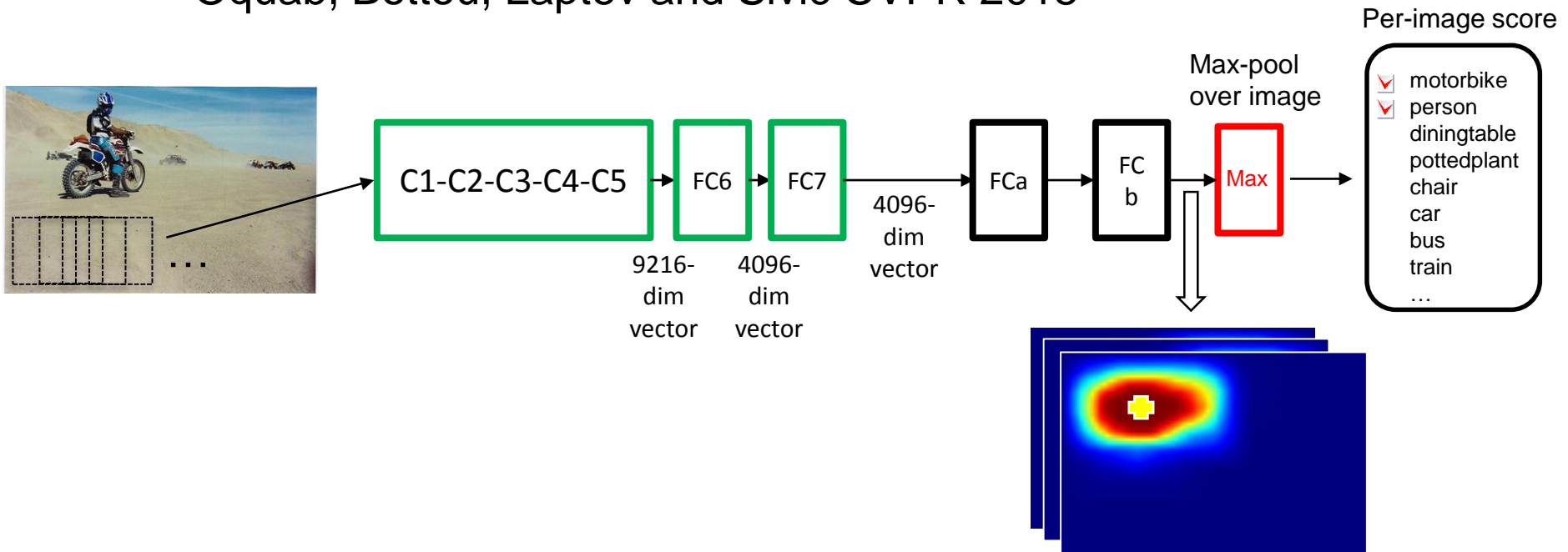


More details in <http://www.di.ens.fr/willow/research/weakcnn/>



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

Oquab, Bottou, Laptev and Sivic CVPR 2015

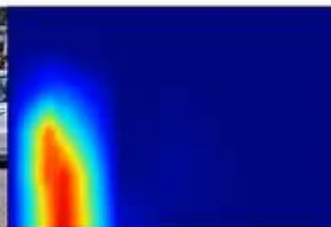
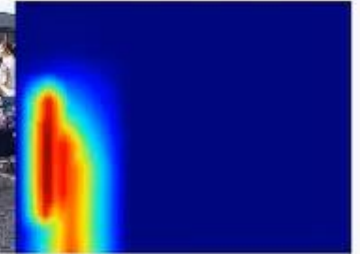


1. Fully convolutional network
2. Image-level aggregation (max-pool)
3. Multi-label loss function (allow multiple objects in image)

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

# Training Motorbikes

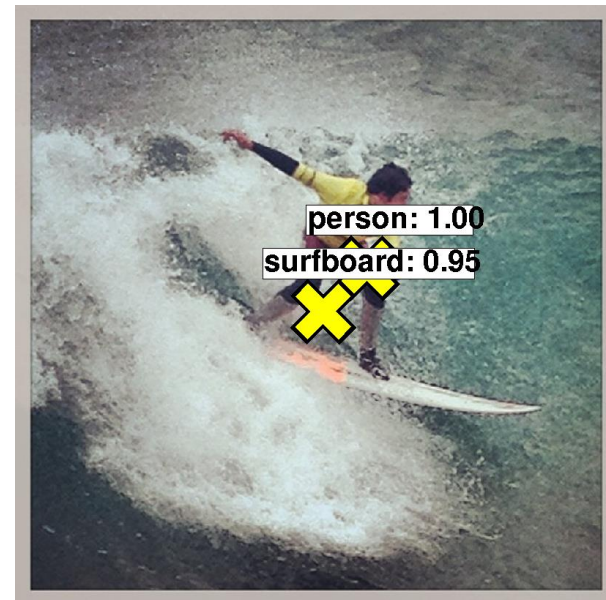
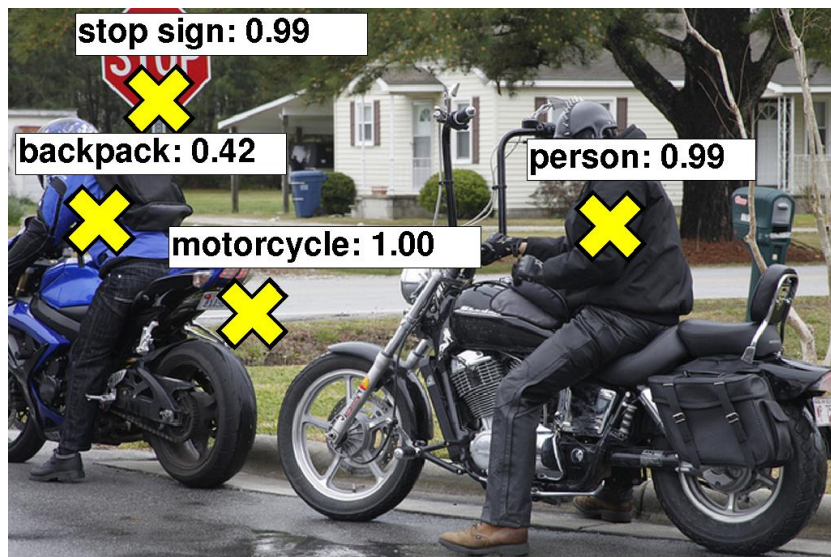
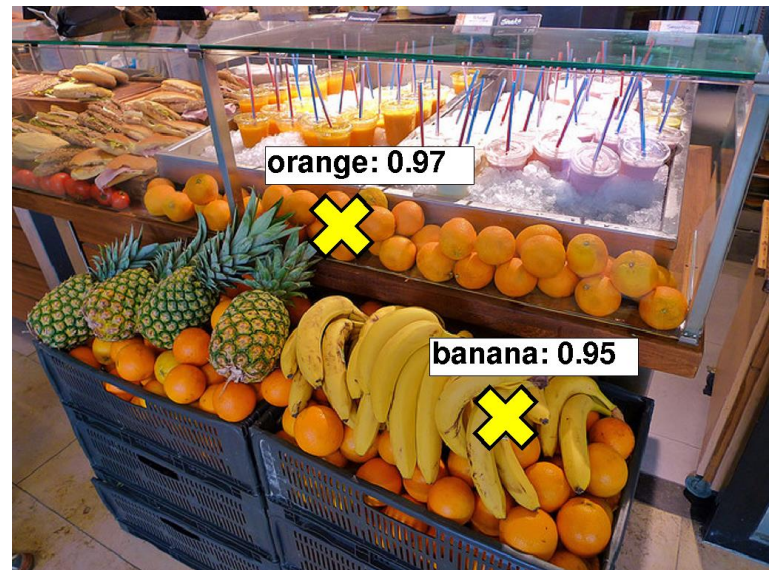
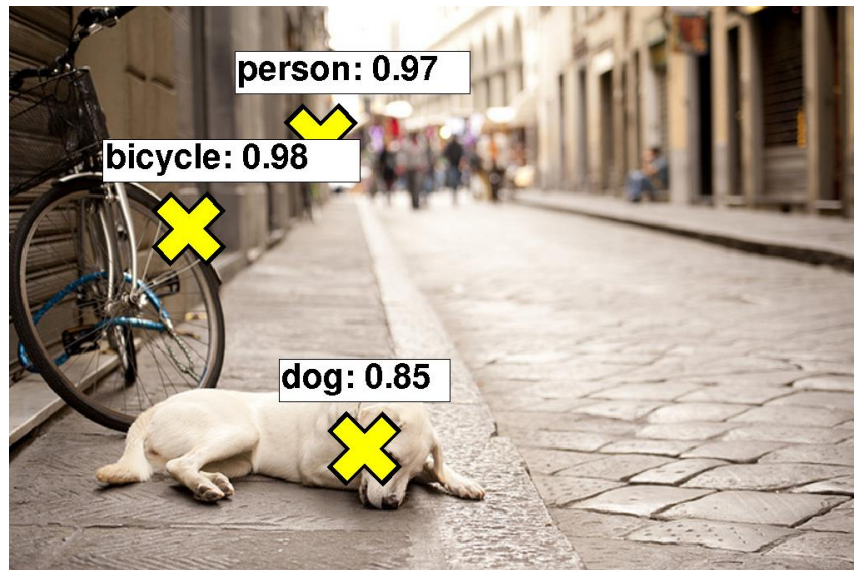
motorbike - training iteration 0030



Evolution of  
localization  
score maps  
over training  
epochs

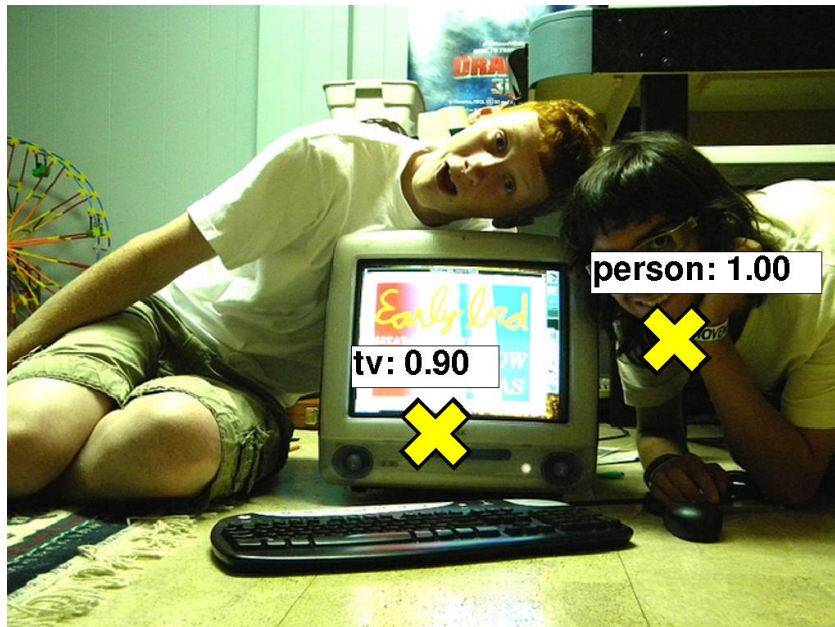


# Test results on 80 classes in Microsoft COCO dataset

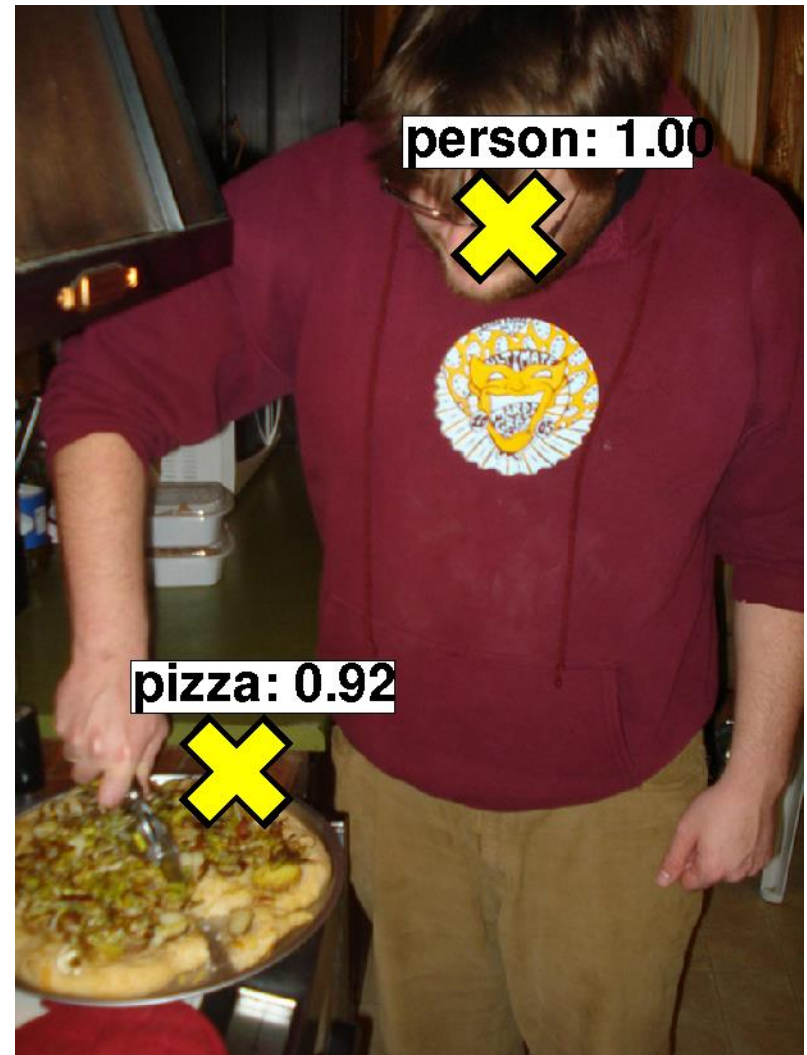
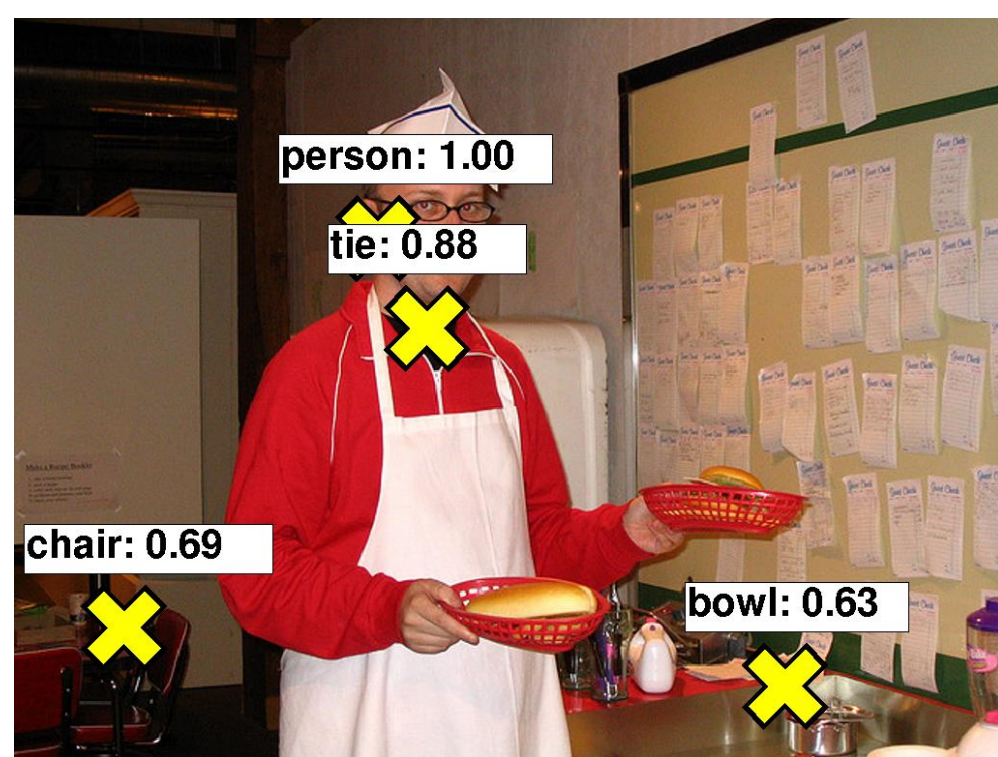




# Test results on 80 classes in Microsoft COCO dataset



# Test results on 80 classes in Microsoft COCO dataset





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

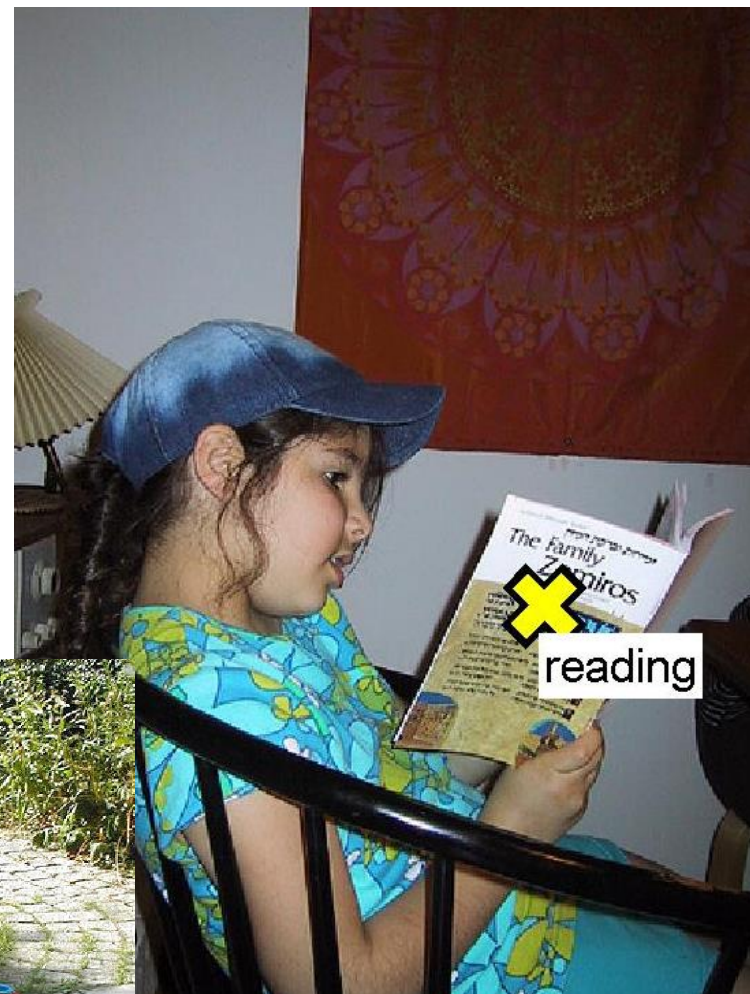
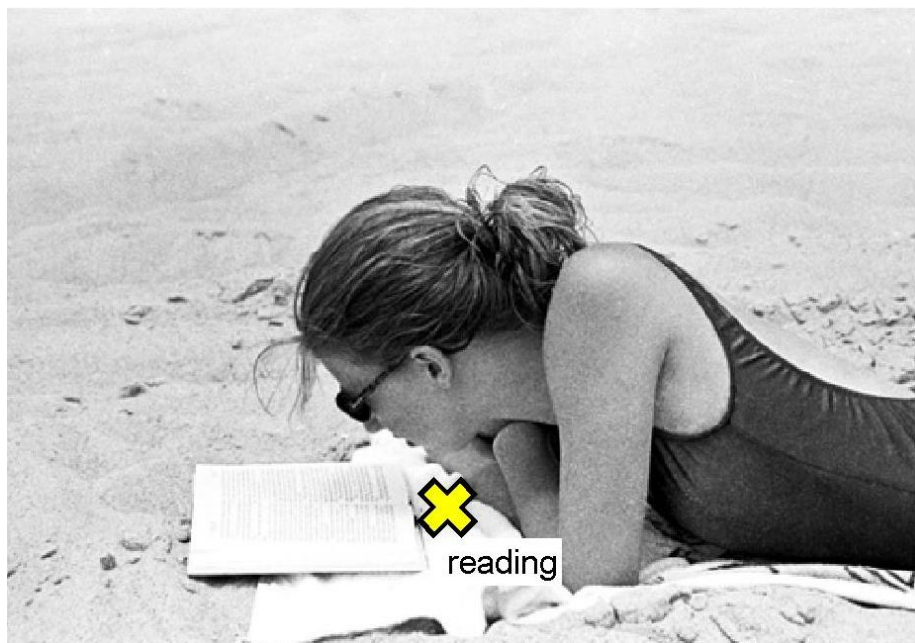


# Test results for **10 action classes** in Pascal VOC12





# 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

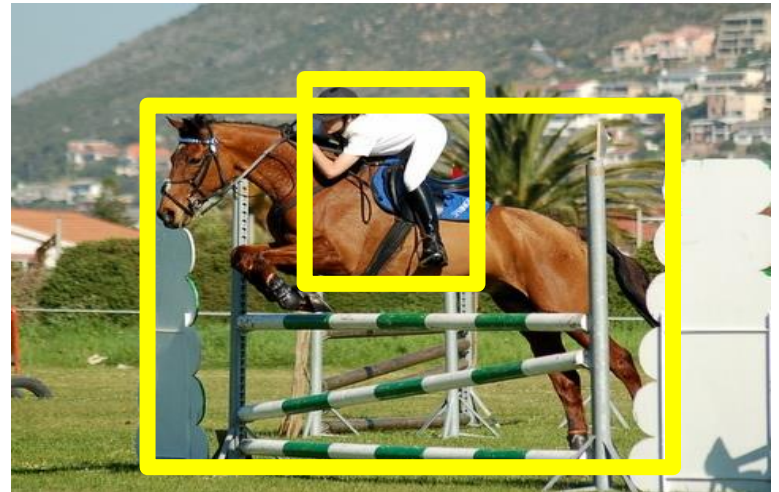
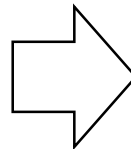
## Failure cases



# Context-aware deep network models for weakly supervised localization

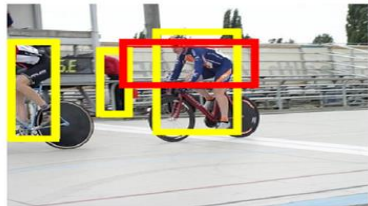
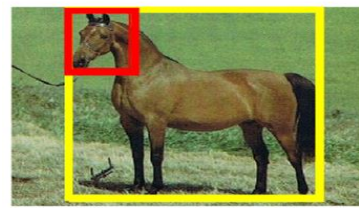
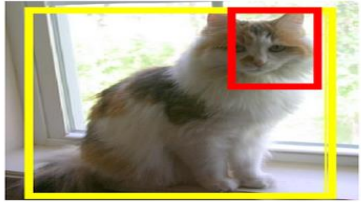
Vadim Kantorov, Maxime Oquab, Minsu Cho, Ivan Laptev

(In submission)





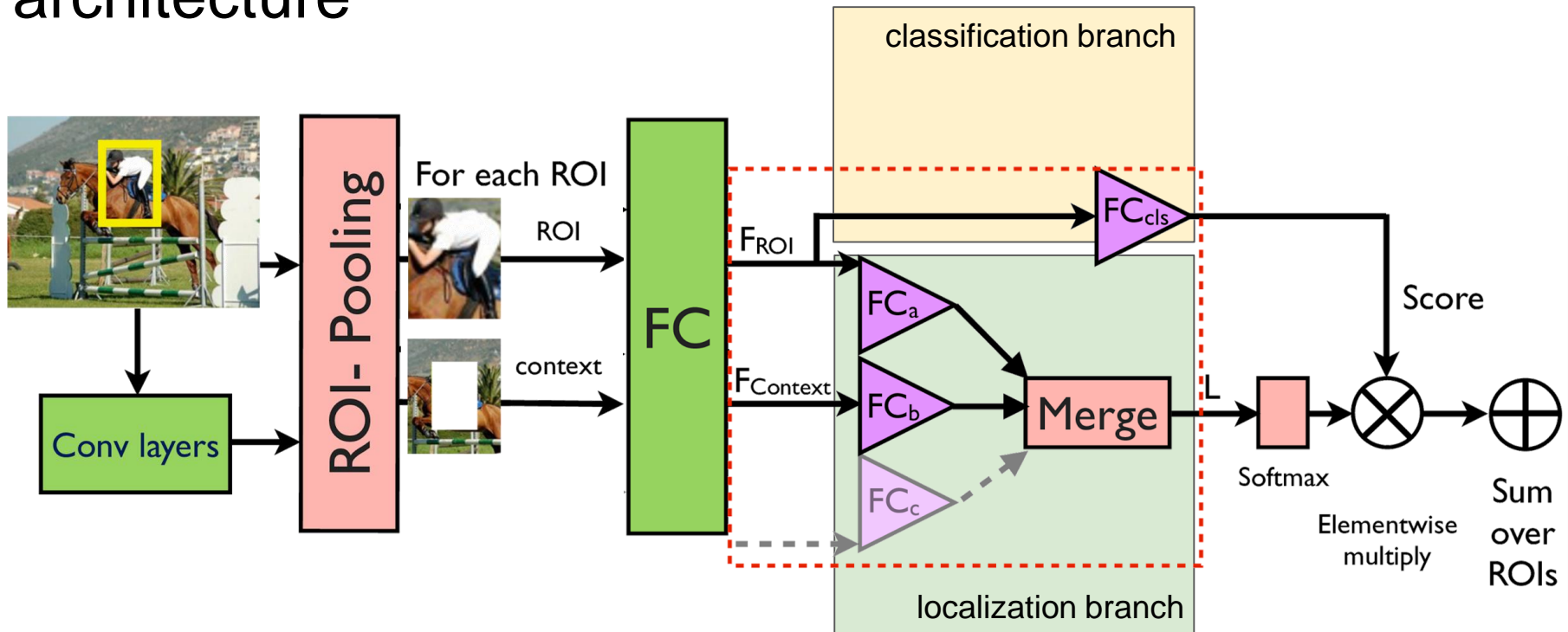
# Problems: shrinking / expansion



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

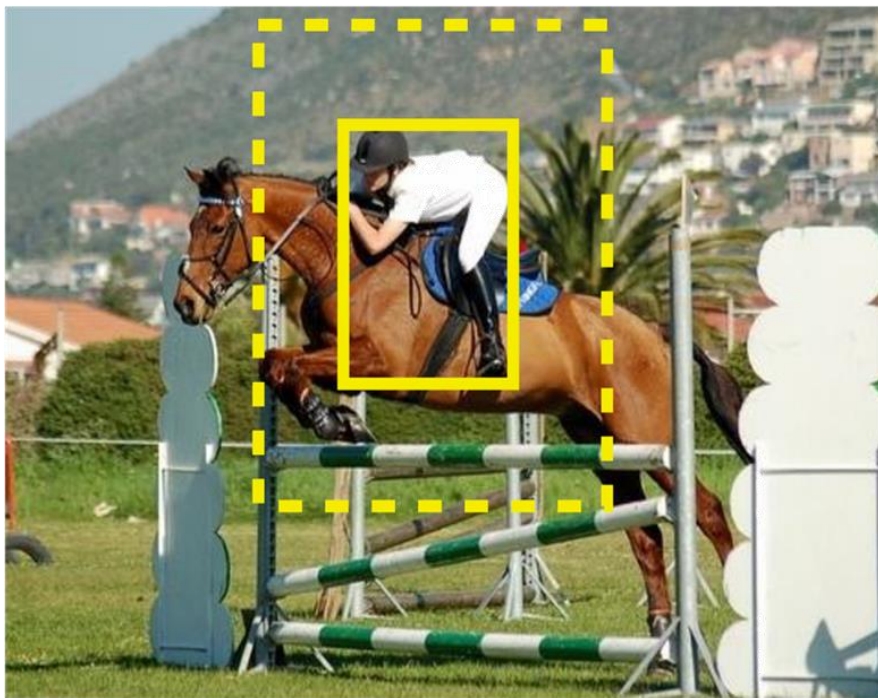


# Context-aware architecture

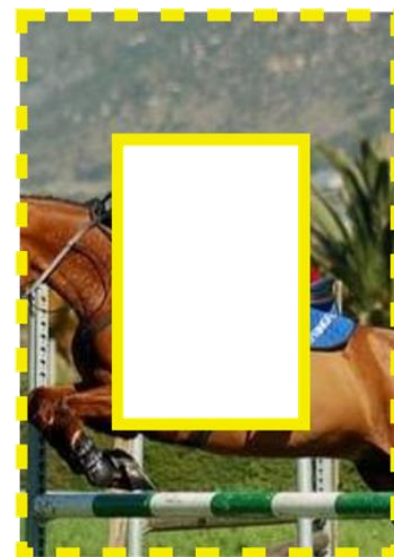


- We process context around a proposal separately
- Everything else in the architecture is pretty much like in [Bilen et al]

## ROI transforms

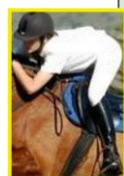


ROI



Context

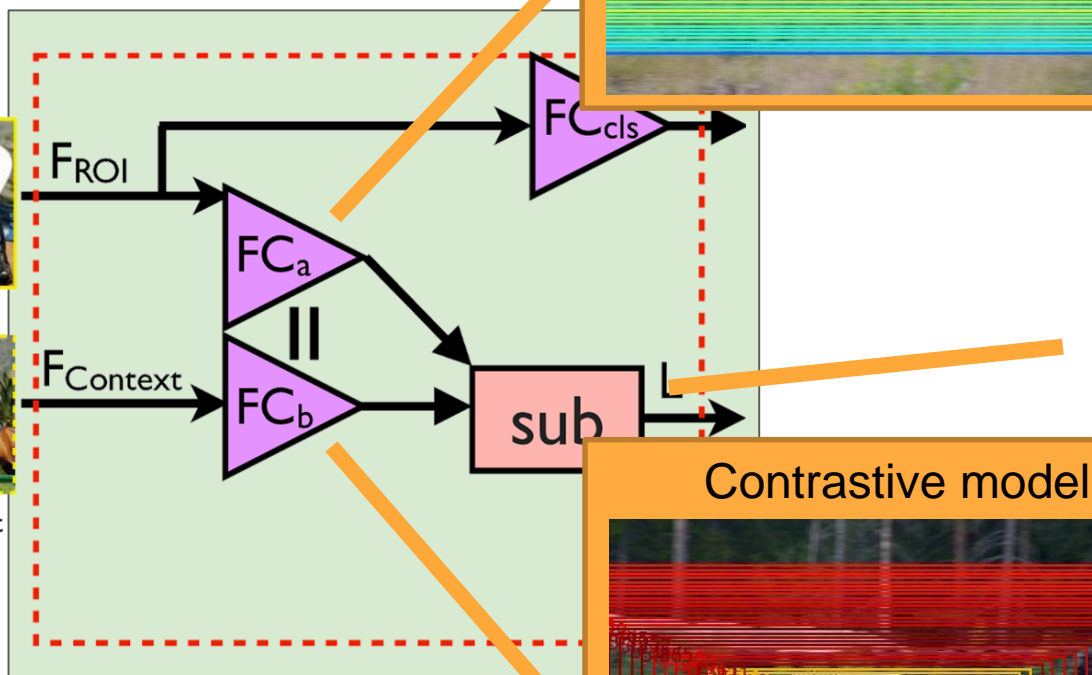
# Contrastive model



ROI



Context



Contrastive model,  $FC_a$



Contrastive model,  $FC_a - FC_b$



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	<b>52.15</b>	<b>33.66</b>
(d)	our WSDDN-8 [6]	48.09	28.19
(e)	ensemble	<b>54.14</b>	<b>34.76</b>
(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]

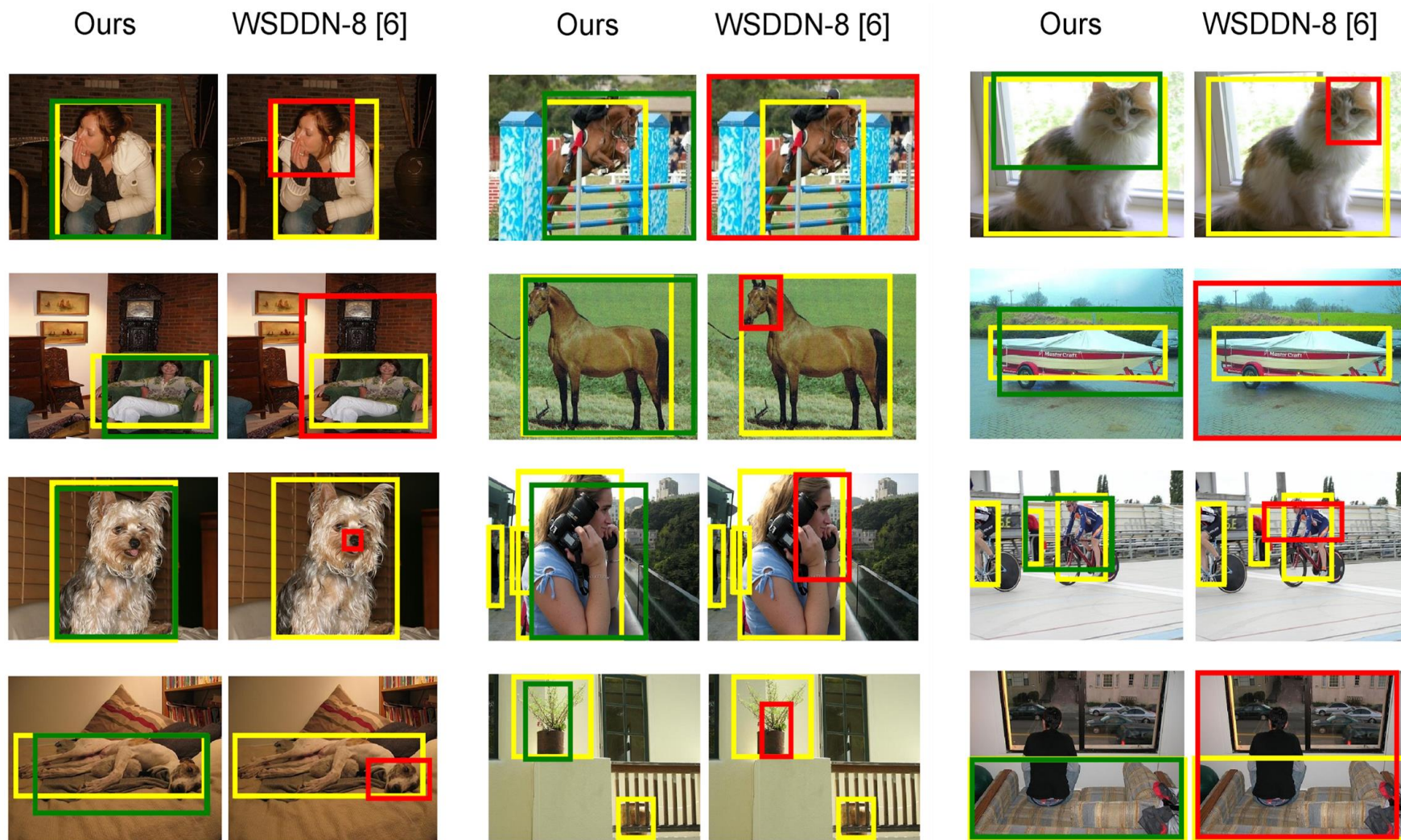
[6] Bilen et al., Weakly Supervised Deep Detection Networks, CVPR 2016]

## PASCAL VOC 2012

Model	Avg. mAP
additive	58.7 32.6
contrastive S	60.6 <b>35.4</b>
ensemble <sup>†</sup>	<b>60.7 35.4</b>
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 conjunction with:*



**VisionLabs**  
visual recognition company

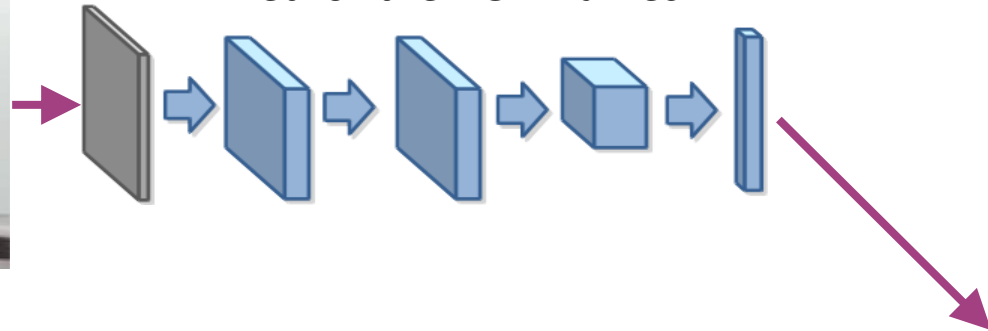
**Skoltech**  
Skolkovo Institute of Science and Technology

# Two-stream neural network

RGB frames



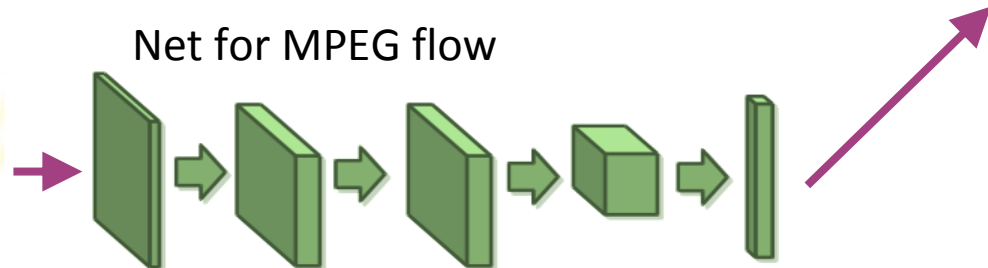
Net for the RGB frames



Decoded motion vectors



Net for MPEG flow



«playing guitar»

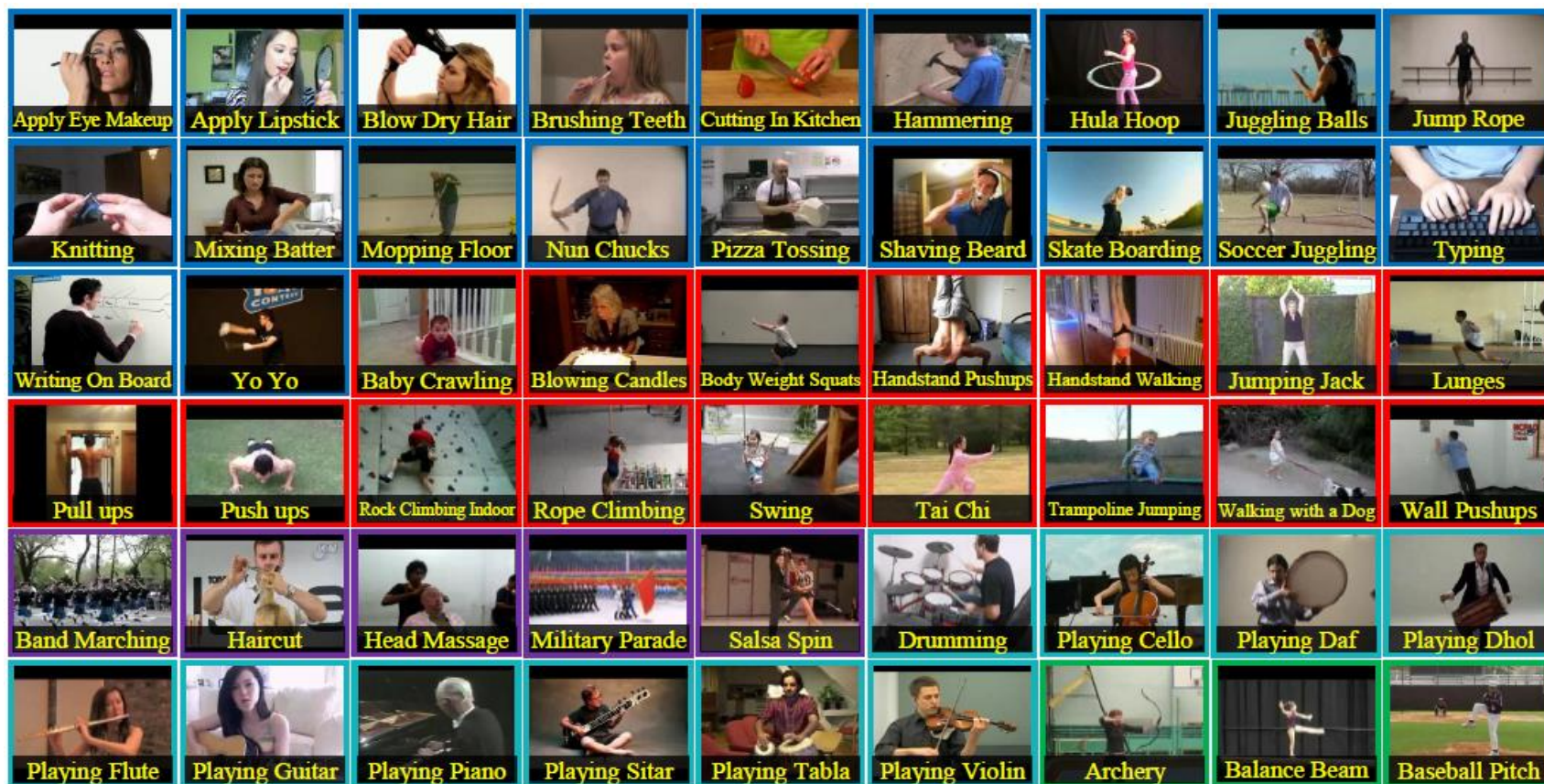


# Results

UCF101 benchmark

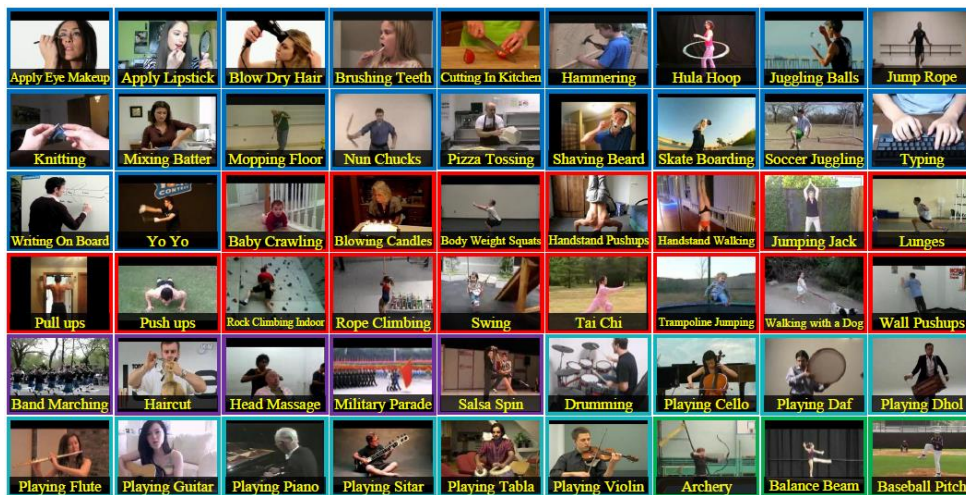
101 action classes

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?



CAM-EC Front Left  
2020-08-19 19:54:14



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]

Learning from Internet for robotics



[Microsoft HoloLens]

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

Sample 1



Start by **loosening** each **bolt**. Then locate the jack and **lift** the **car**. Now you can **remove** the bolts and then the **wheel**.

Sample 2



First **undo** the **nuts**. Once that done, you can **jack** the **car**. Then withdraw the nuts completely so that you can **remove** the flat **tire**.

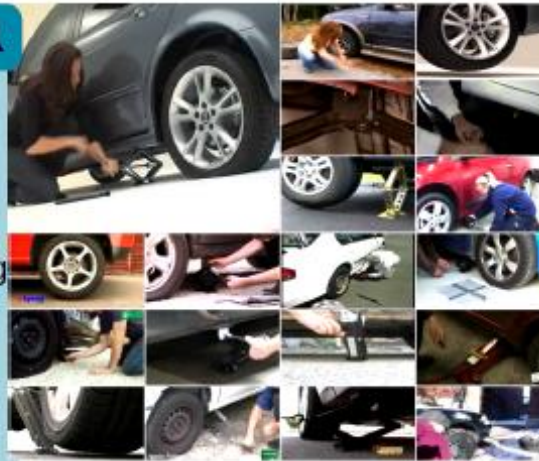


# New dataset of Internet instruction videos

## Changing a car tire

### JACK CAR

"all the way around before we jack the car. So now what we're gonna..." "so that it lifts the car by the frame..."  
"Make sure the jack is lifting nice and square..." "At this stage the car is being raised to a safe height..." "we can jack the car up..."



"placing the replacement wheel on..." "The wheel is then removed and replaced with the spare wheel..." "your spare tire is now ready to install..." "Line up the spare and lift it to place..."  
"make sure you put the new tire on the..."

"Then tighten down each nut with the tire iron..." "Tight one nut a few turns and then do one opposite until each lug nut..."  
"Just tighten these up now..." "finally screw the wheel tight..." "give them that final tight..."



### TIGHT THE WHEEL NUTS



### PLACE SPARE TIRE



# New dataset of Internet instruction videos

## JACK CAR

"all the way around before we jack the car. So now what we're gonna..." "so that it lifts the car by the frame ..."

"Make sure the jack is lifting nice and square ..." "At this stage the car is being raised to a safe height..." "we can jack the car up..."



# New dataset of Internet instruction videos

“placing the replacement wheel on ...” “The wheel is then removed and replaced with the spare wheel...” “your spare tire is now ready to install...” “Line up the spare and lift it to place...” “make sure you put the new tire on the...”





# New dataset of Internet instruction videos

"Then tighten down  
each nut with the tire  
iron..." "Tight one nut  
a few turns and then do  
one opposite until each lug  
nut..."

"Just tighten these  
up now..." "finally SCREW  
the wheel tight..." "  
give them that final  
tight..."

**TIGHT THE WHEEL NUTS**





# New dataset of Internet instruction videos

Make  
coffee

FILL WATER

"you put in water..."  
"you wouldn't put water more  
than above this valve..."  
"They say to fill it up to that line..."  
"fill the bottom  
part up with water up"



"we have the moka pot  
reassembled..."  
"just grab it and you twist the  
top on..."  
"just screw it together"  
"Make it tight..."  
"screw the top on and set it on  
the stove"

SCREW FILTER



SEE COFFEE

"we're gonna see that the  
coffee is  
beginning to come out..."  
"now here's a sneak  
peek..."  
"know that it is sucking the  
coffee up..."  
"when you see this you want to  
turn off the heat..."  
"after the coffee comes out  
it starts..."

# New dataset of Internet instruction videos

## Repotting a plant

"give the plant a drink of water..."  
"the next step is going to be watering..."  
"we are gonna put water in here..."  
"if you don't water the plant, you will get air pockets in the soil..."

WATER PLANT

REMOVE PLANT

"the next step is to take the plant out of the pot..."  
"I'm going to remove my plant from its old pot..."  
"remove it from its beginning container..."



"make sure there aren't any roots..."  
"you need to tease them out..."

"sometimes you just want to cut the roots..."  
"loosen the old soil away from the roots..."

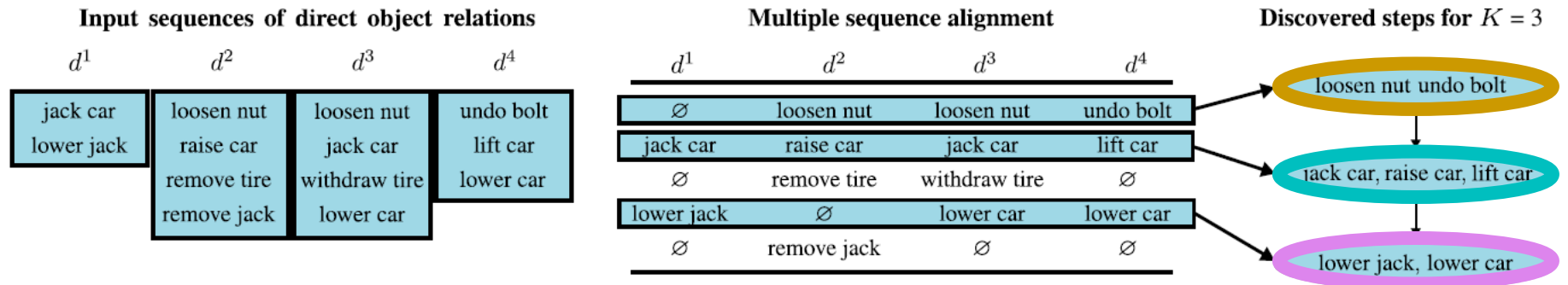


LOOSEN ROOTS



# Approach: two linked clustering problems

## 1. Text clustering into a **sequence** of common steps



## 2. Video clustering to **localize the actions** with text constraints

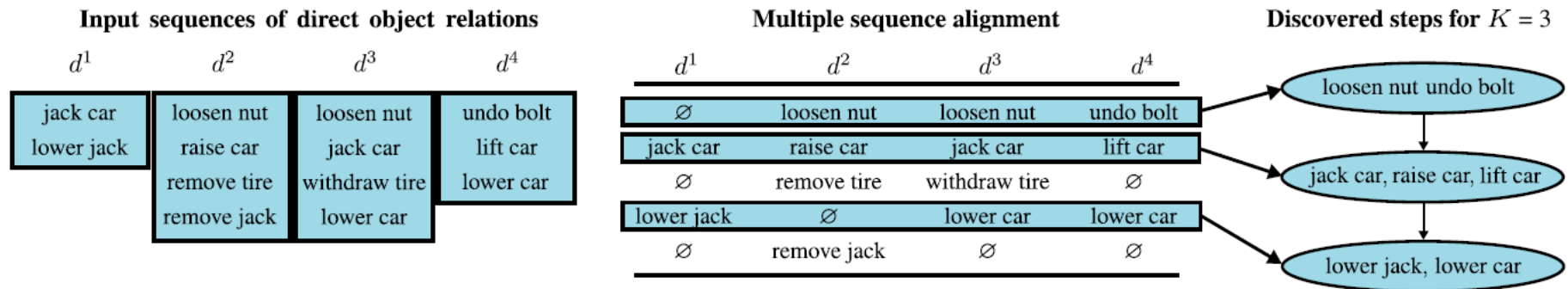
Start by **loosening** each **bolt**. Then locate the jack and **lift** the **car**. Now you can **remove** the bolts and then the **wheel**.

First **undo** the **nuts**. Once that done, you can **jack** the **car**. Then withdraw the nuts completely so that you can **remove** the flat **tire**.



# Approach: two linked clustering problems

## 1. Text clustering into a **sequence** of common steps



## 2. Video clustering to **localize the actions** with text constraints

$$h(Z) = \min_{W \in \mathbb{R}^{K \times d}} \underbrace{\frac{1}{2T} \|Z - XW\|_F^2}_{\text{Discriminative loss on data}} + \underbrace{\frac{\lambda}{2} \|W\|_F^2}_{\text{Regularizer}}$$

Discovered temporal localization [TxK matrix]
Representation of video chunks (IDTF, CNN) [Txd] matrix
Linear action classifier [dxK] matrix

s.t.  $\underbrace{Z \in \mathcal{Z}}_{\text{ordered script}}, \underbrace{AZ \geq R}_{\text{weak textual constraints}}$   
Temporal constraints from text



# Future challenges



# Future challenges

*What is unusual in this*



*What is unusual in this scene?*



*attention of this person?*

