

Introduction to computer vision XIV

Instructors: Jean Ponce and Matthew Trager
jean.ponce@inria.fr, matthew.trager@cims.nyu.edu

TAs: Jiachen (Jason) Zhu and Sahar Siddiqui
jiachen.zhu@nyu.edu, ss12414@nyu.edu

Slides will be available after class at:
<https://mtrager.github.io/introCV-fall2019/>

Image categorization as supervised classification

Beavers



Chairs



Trees



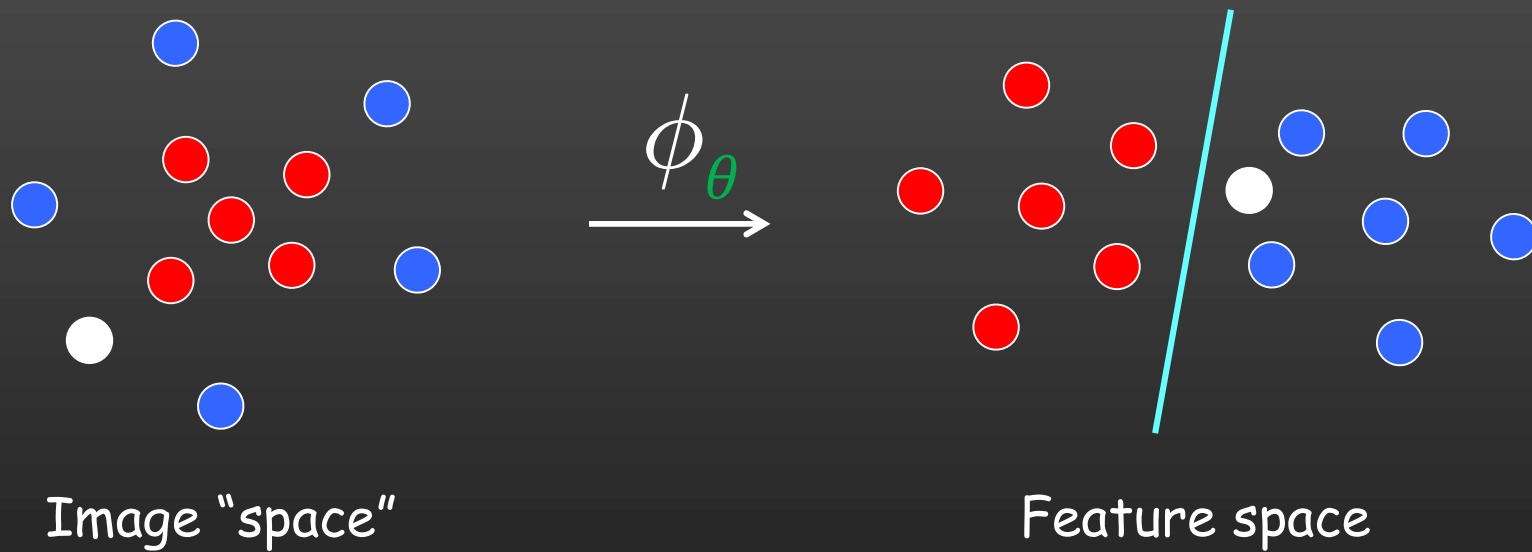
Labelled training examples



??

Test image

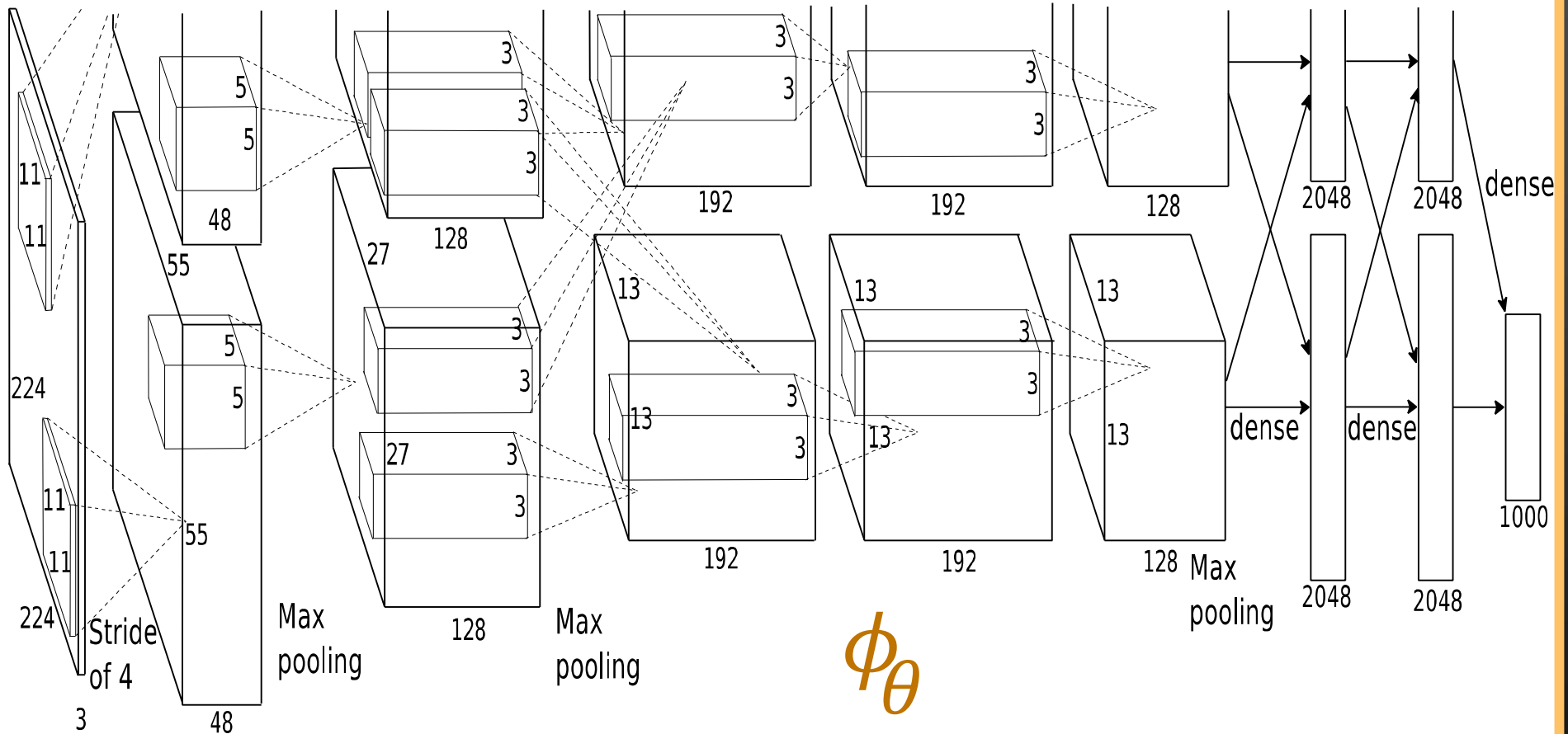
Image categorization as supervised classification



$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_n \ell(z_n, f(\phi_\theta(x_n))) + \Omega(f)$$

Labels: Training datum (pointing to x_n), Label (pointing to z_n), Prediction function (pointing to f)

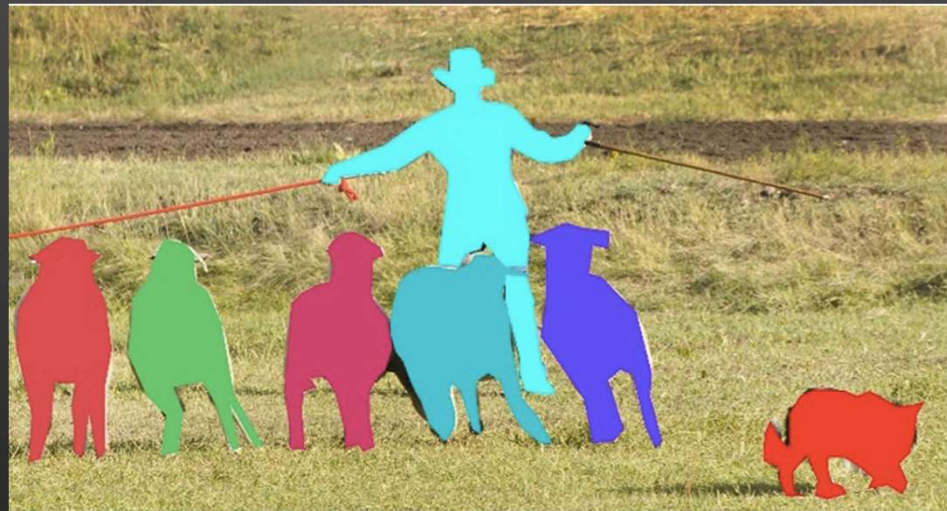
Convolutional neural networks (LeCun et al., 1998)



(Krizhevsky et al.'12)

Supervision: Where do the labels come from?

- A trend toward manually annotating the whole wide world with crowd sourcing



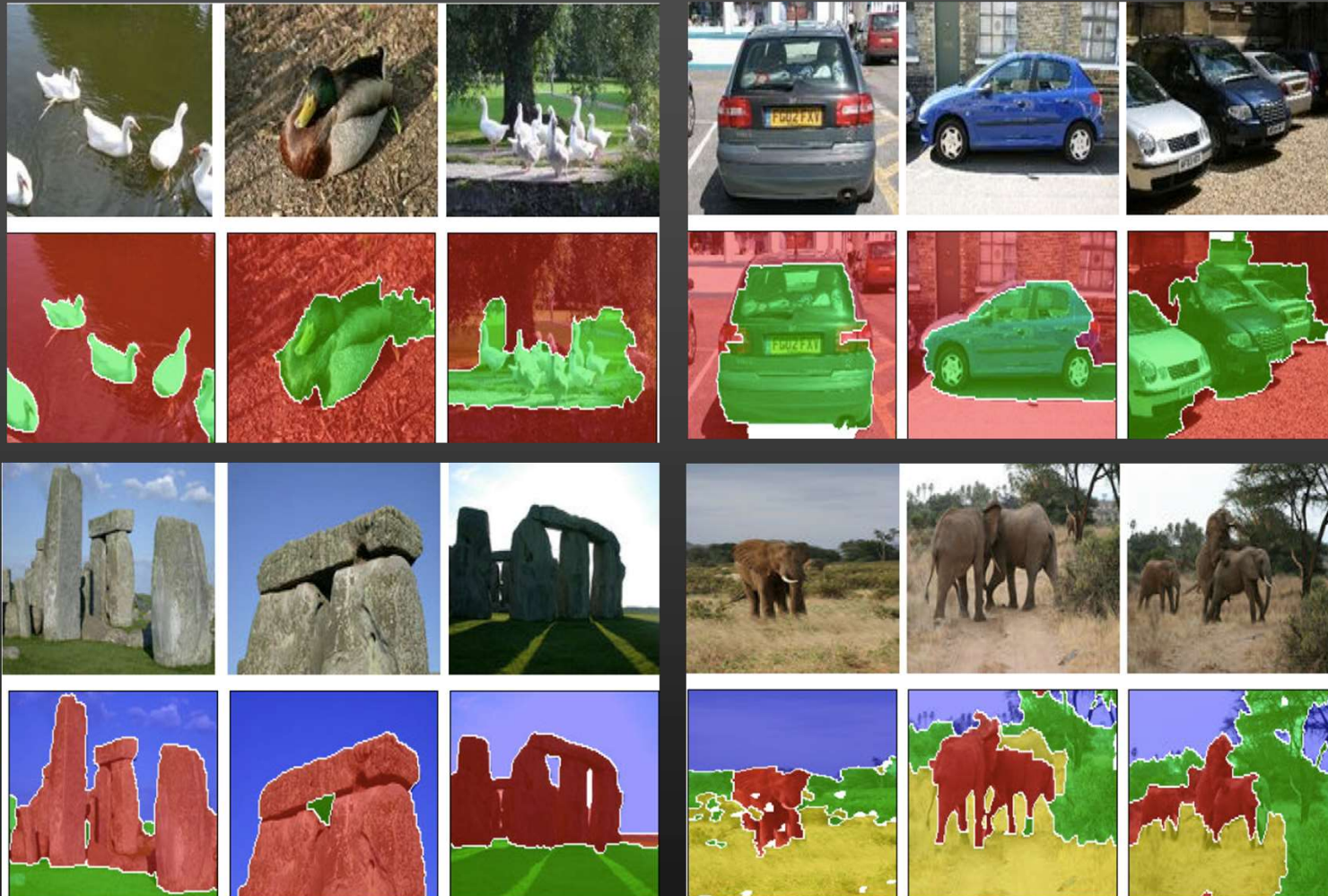
- Example: MS COCO (Lin et al., 2015)
 - 328K images of 91 object categories
 - 2.5M labelled instances

(Russell et al., 2008; Deng et al., 2009; Everingham et al., 2010; Xiao et al., 2010)

Outline

- Weaker forms of supervision, e.g.,
 - image-level labels
 - existing meta data
- Not covered: Semi-supervised methods
 - with some labelled data
- Totally unsupervised methods,
 - self-supervised \approx "free" labels
 - and alternatives
- Musings about parts, semantics, etc.

Using weaker supervision: Cosegmentation



(Lazebnik et al.'04; Rother et al.'06; Hochbaum & Singh'09; Joulin et al.'10)
(Kim & Xing'11; Joulin et al.'12; Rubio et al.'12; Wang et al.'13)

Conventional supervised classification

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_n \ell(z_n, f(\phi(x_n))) + \Omega(f)$$

Conventional supervised classification

$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_n \ell(z_n, f(\phi(x_n))) + \Omega(f)$$

Discriminative clustering

$$\min_{Z, f} \frac{1}{N} \sum_{i \in I} \sum_{n \in \mathcal{N}_i} \ell(z_n, f(\phi(x_n))) + \Omega(f)$$

Square loss

Optimize over labels too

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

$$\min_Z \text{Tr}(ZZ^T A(X, \lambda))$$

(Xu et al., 2004; Bach & Harchaoui, 2007)

Multi-class cosegmentation (Joulin et al., CVPR'12)

$$\min_{\substack{y \in \{0,1\}^{N \times K}, \\ y \mathbf{1}_K = \mathbf{1}_N}} \left[\min_{\substack{A \in \mathbb{R}^{d \times K}, \\ b \in \mathbb{R}^K}} E_U(y, A, b) \right] + E_B(y) - H(y)$$

Softmax

Spectral
clustering
term

Entropy term

Discriminative clustering term

$$\min_{Z, f} \frac{1}{N} \sum_{i \in I} \sum_{n \in \mathcal{N}_i} \ell(z_n, f(\phi(x_n))) + \Omega(f)$$



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



$$\min_Z \operatorname{Tr}(ZZ^T A(X, \lambda))$$

(Shi & Malik, 2000; Ng et al., 2001; Xu et al., 2004; Bach & Harchaoui, 2007)

Multi-class cosegmentation (Joulin et al., CVPR'12)

$$\min_{\substack{y \in \{0,1\}^{N \times K} \\ y \mathbf{1}_K = \mathbf{1}_N}} \left[\min_{\substack{A \in \mathbb{R}^{d \times K} \\ b \in \mathbb{R}^K}} E_U(y, A, b) \right] + E_B(y) - H(y)$$

Optimization:

- Relax to continuous problem
- EM/block-coordinate descent procedure with quasi-Newton and projected gradient descent for the two steps, initialized with quadratic approximation
- Round up the solution

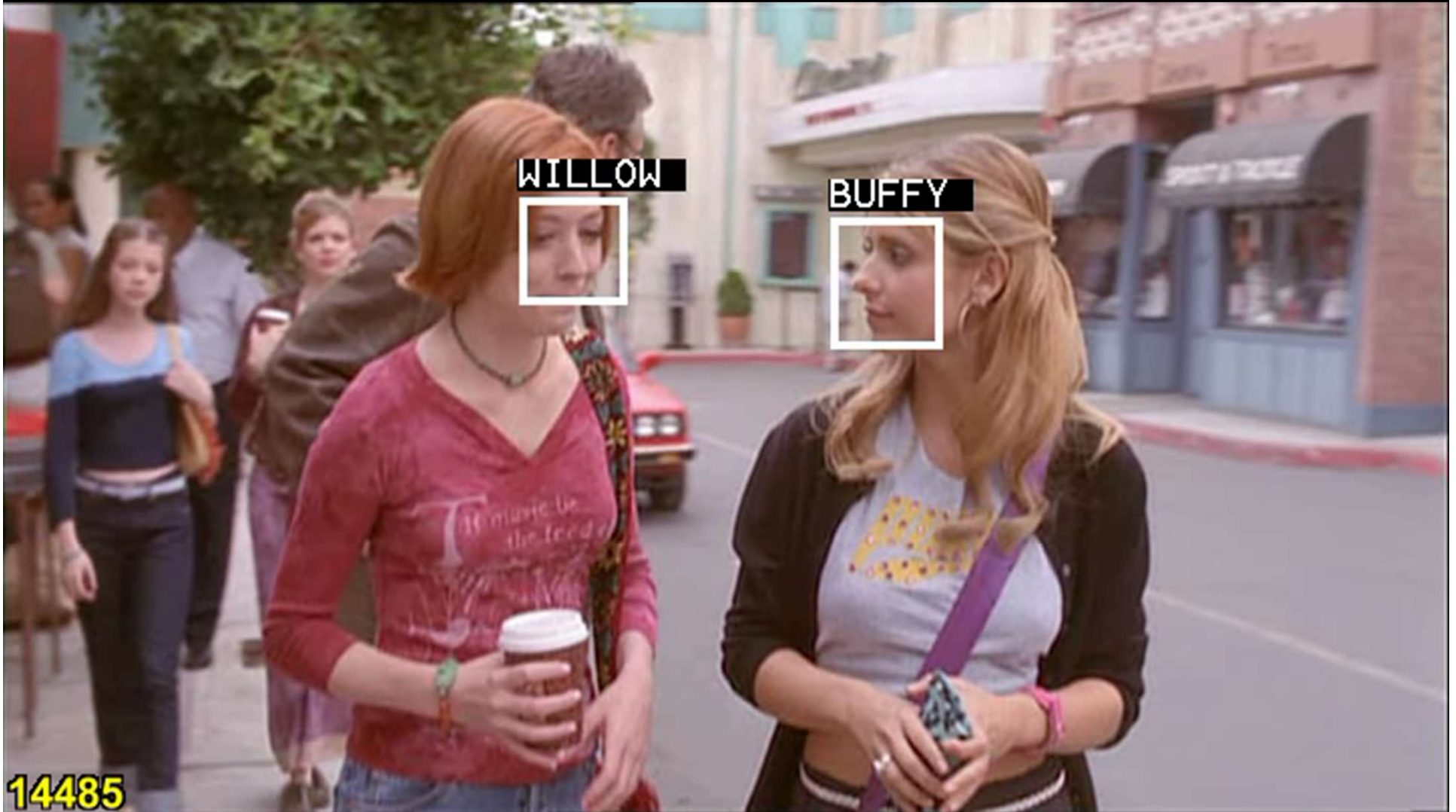
Missing: no foreground model (Rother et al., 2006)

Multi-class cosegmentation results



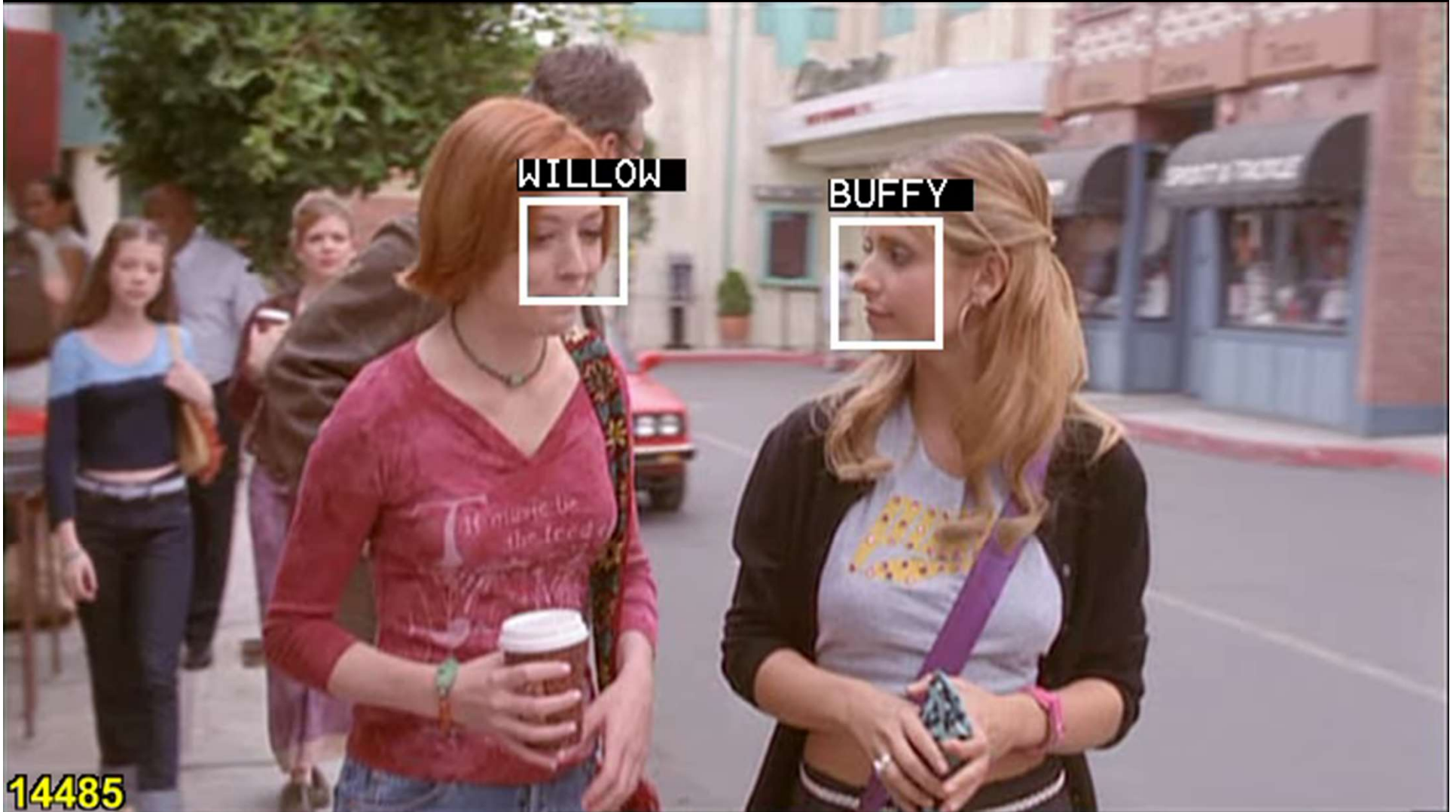
(Joulin et al., CVPR'12)

Naming the characters of TV series



(Sivic, Everingham, Zisserman, 2009)

TV series come with their own metadata



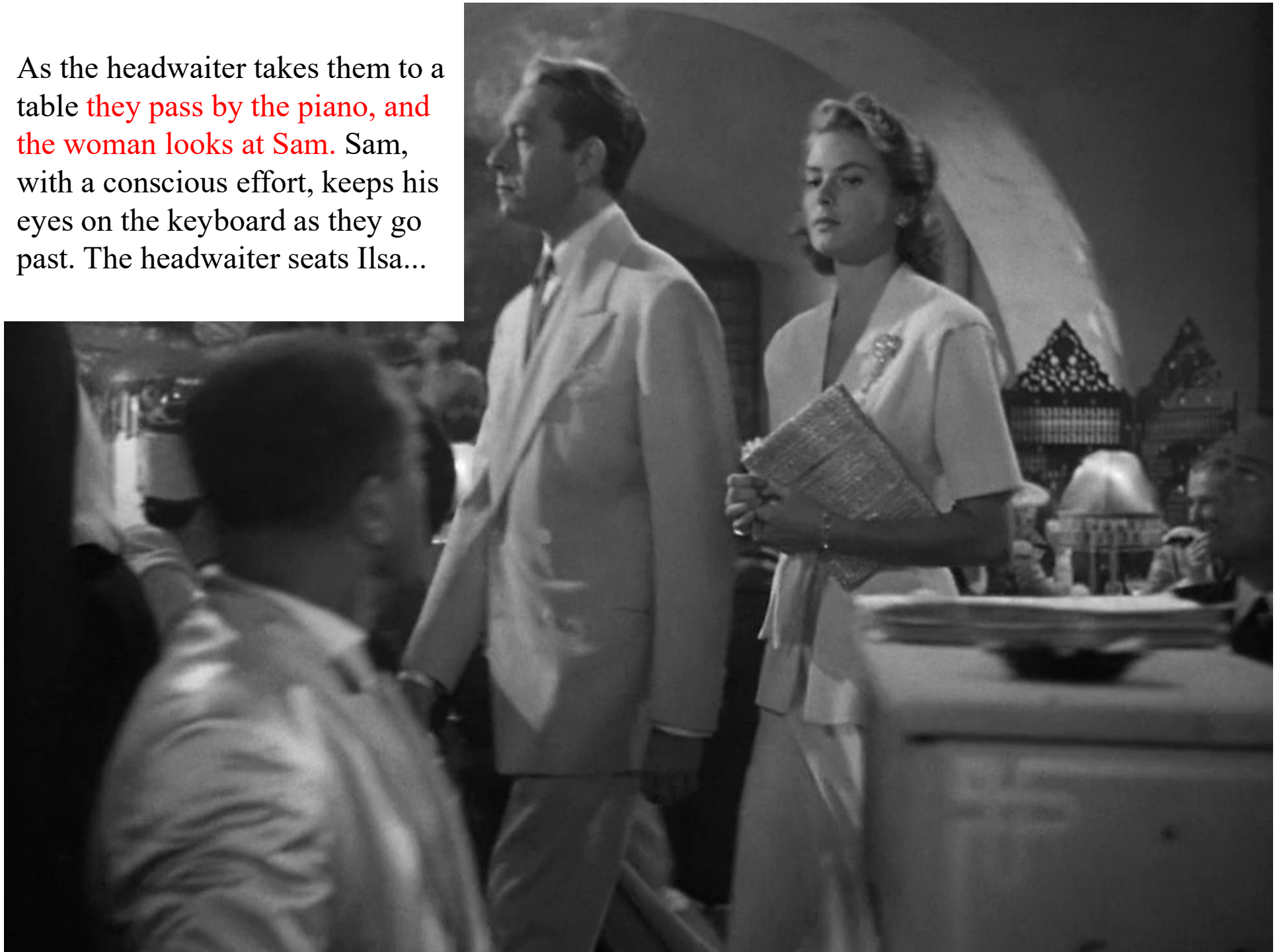
(Sivic, Everingham, Zisserman, 2009)

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...

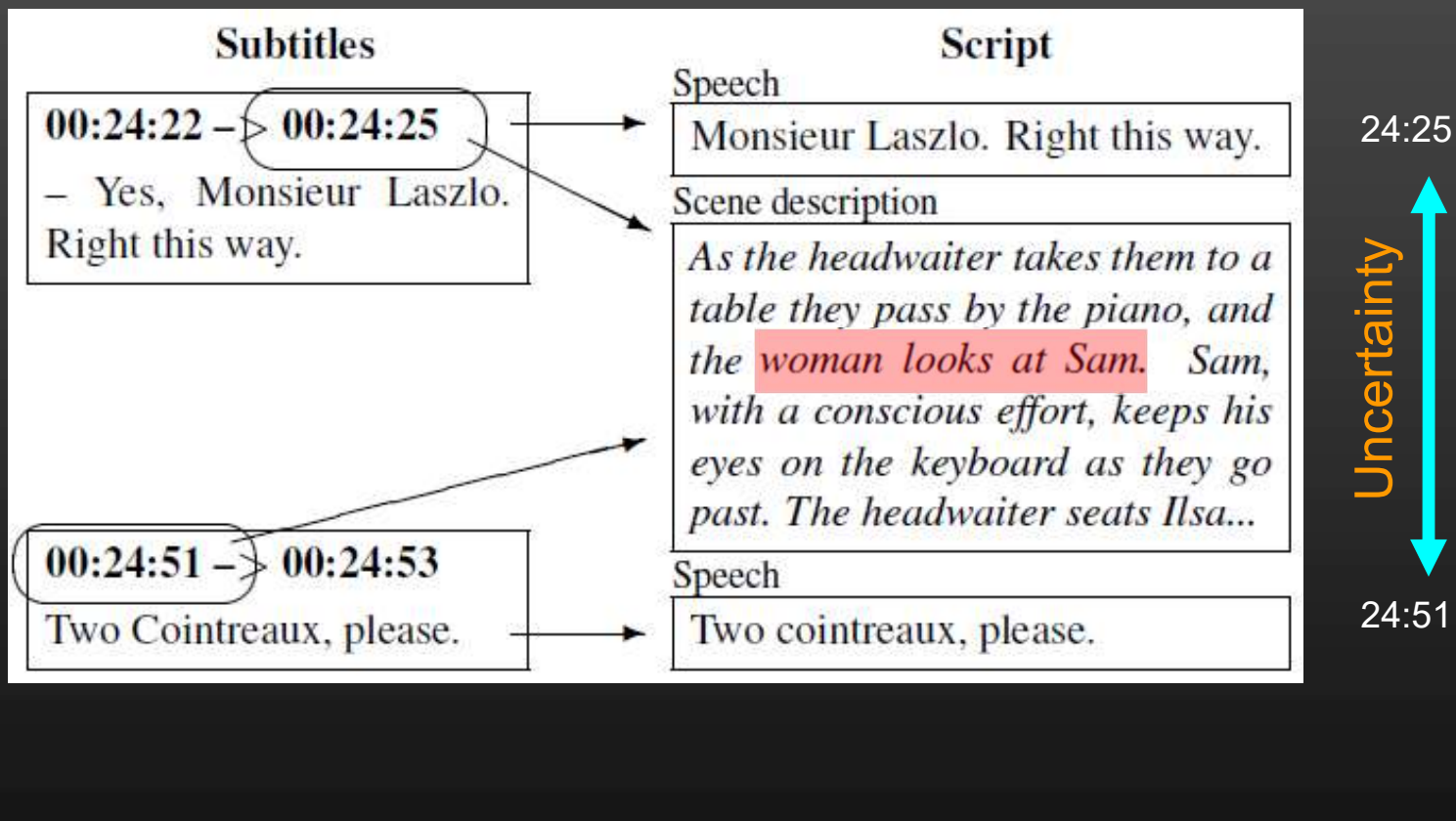


Videos (often) come with their own metadata!

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...



Scripts as a source of supervision



(Laptev et al., 2008; Sivic et al., 2009; Duchenne et al., 2009)

Automated temporal action localization

Input:

- Action type, e.g. "Person opens door"
- Videos + aligned scripts



Output: temporal action clusters

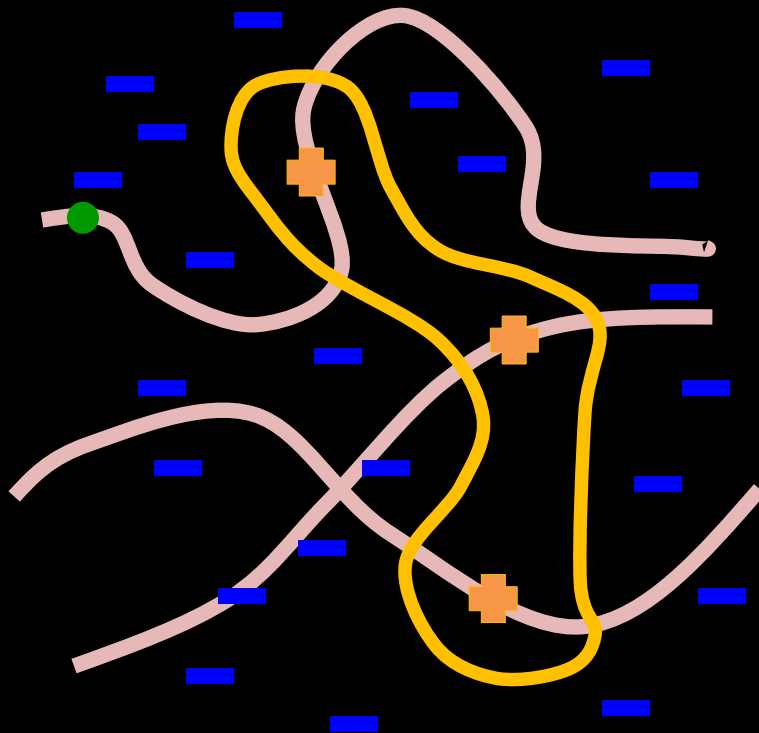
... Jane jumps up and opens the door ...
... Carolyn opens the front door ...
... Jane opens her bedroom door ...



(Duchenne, Laptev, Sivic, Bach, Ponce, 2009)

Temporal localization as classification

Feature space



Video space



Negative samples

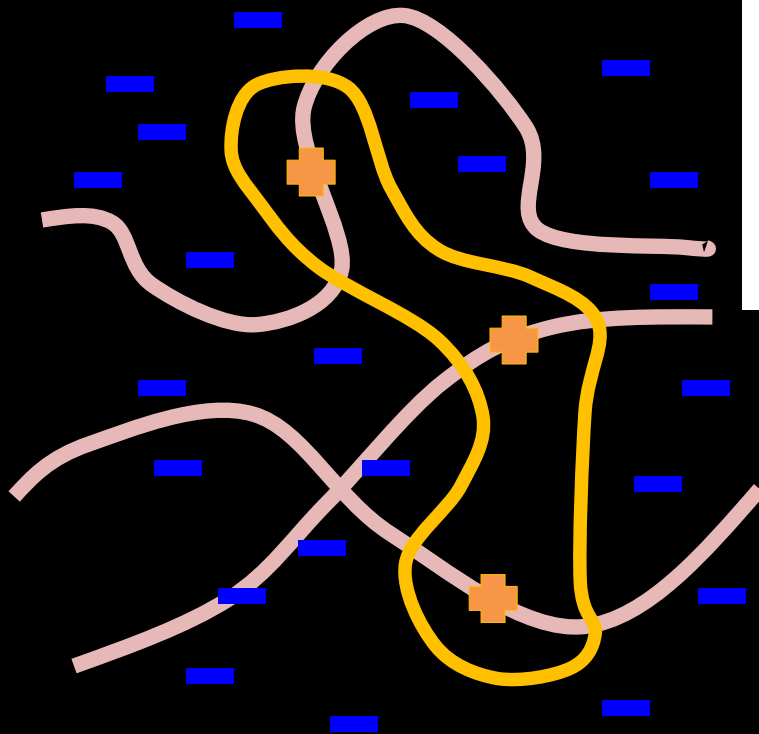


Random video clips: lots of them,
very low chance to be positives

A latent SVM model for temporal localization

(Felzenszwalb, McAllester, Ramanan, 2008)

Feature space



$$\min_{w,b} C_+ \sum_{i=1}^M \max\{0, 1 - \max_f w^\top \Phi(c_i[f]) - b\} \\ + C_- \sum_{i=1}^P \max\{0, 1 + w^\top \Phi(x_i^-) + b\} + \|w\|^2$$

Optimization: Block-coordinate descent

1. Exhaustive search for f
2. SVM training for w, b

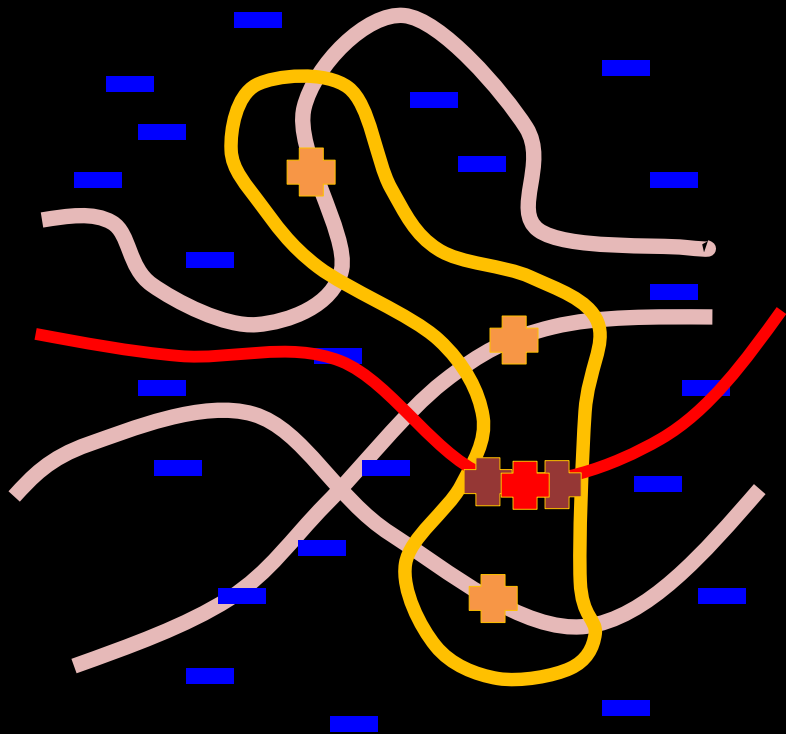
This is an instance of discriminative clustering

Clustering results on "Coffee and cigarettes"



Using the learned models for action detection

Feature space



New video

- Find local maxima
- aka non maximum suppression
- aka sliding window

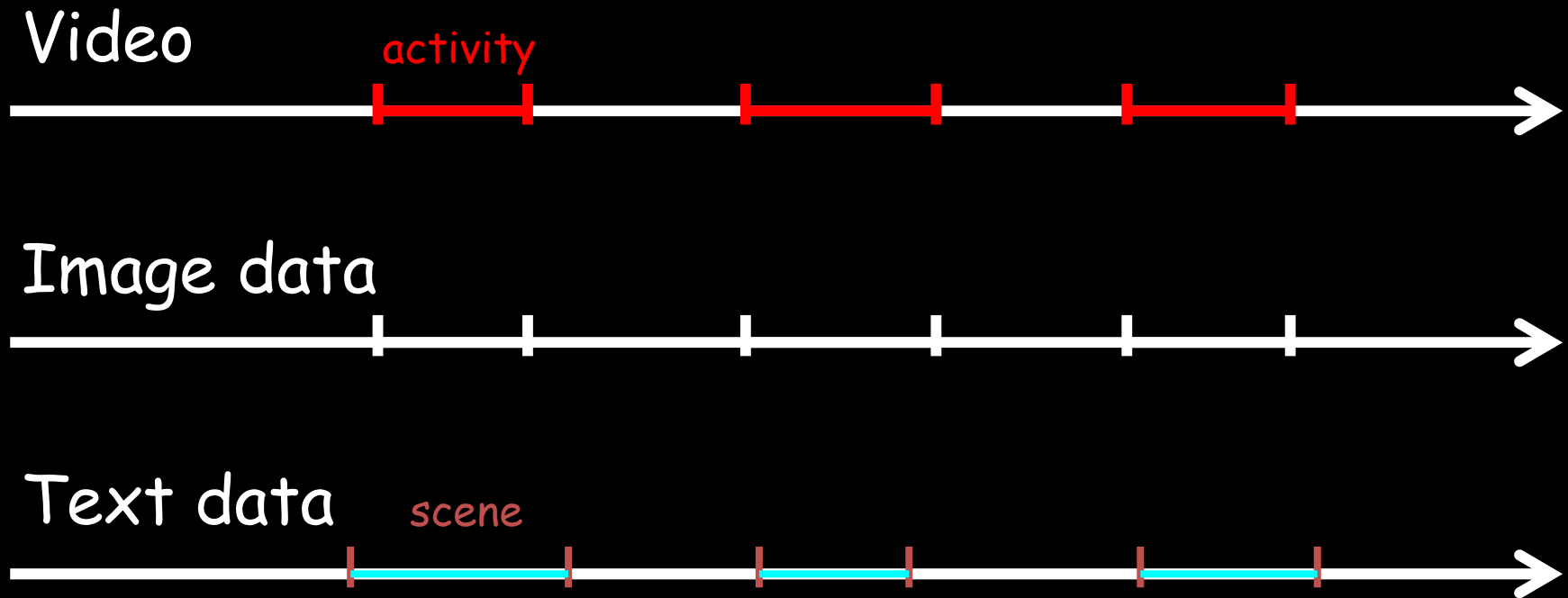
Automatic Annotation of Human Actions in Video

ICCV 2009 DEMO

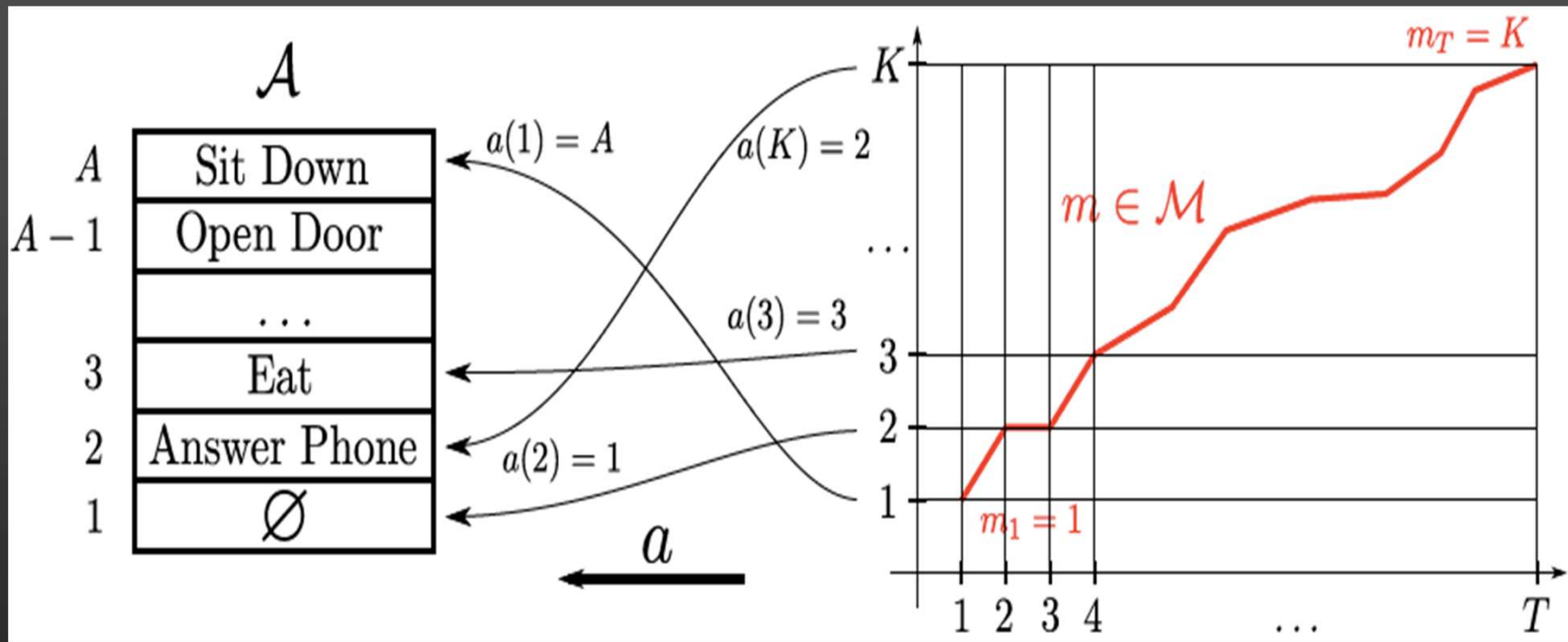
O.Duchenne, I.Laptev, J.Sivic, F.Bach and J.Ponce

**Temporal detection of actions OpenDoor and SitDown in episodes of
The Graduate, The Crying Game, Living in Oblivion**

Exploiting temporal constraints



Action labeling under ordering constraints (Bojanowski et al., ECCV'14)



Dictionary

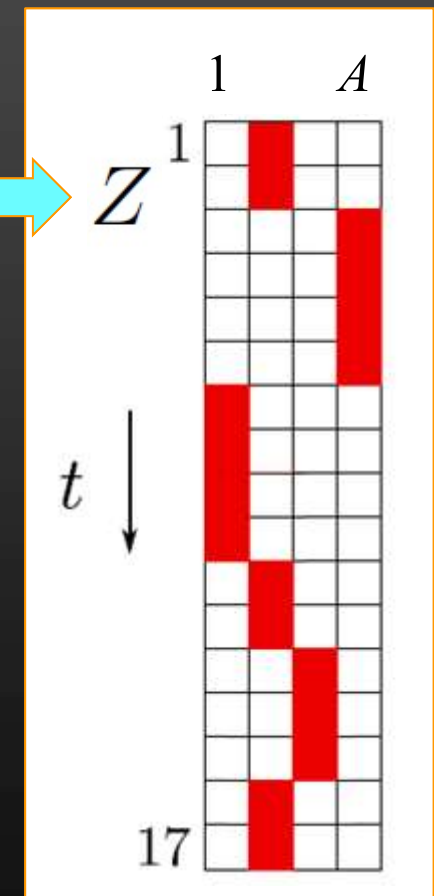
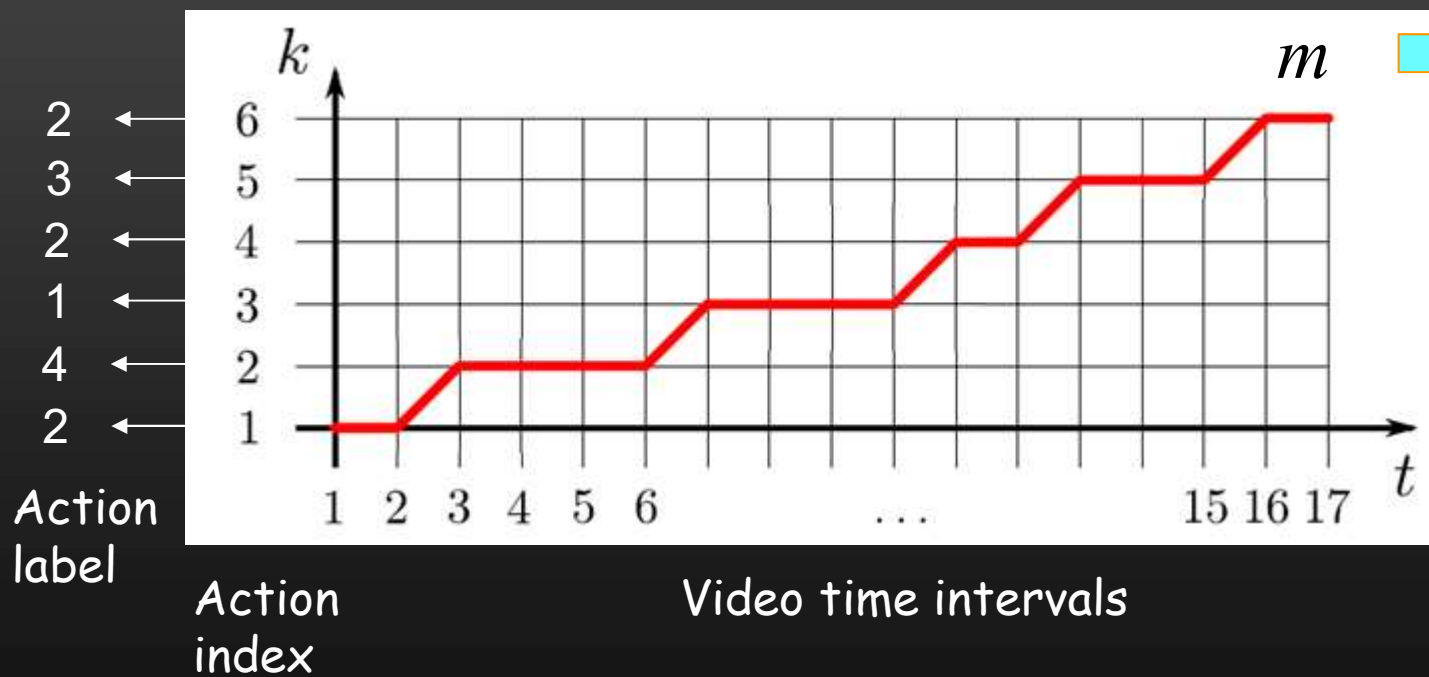
Script metadata a

Alignment m

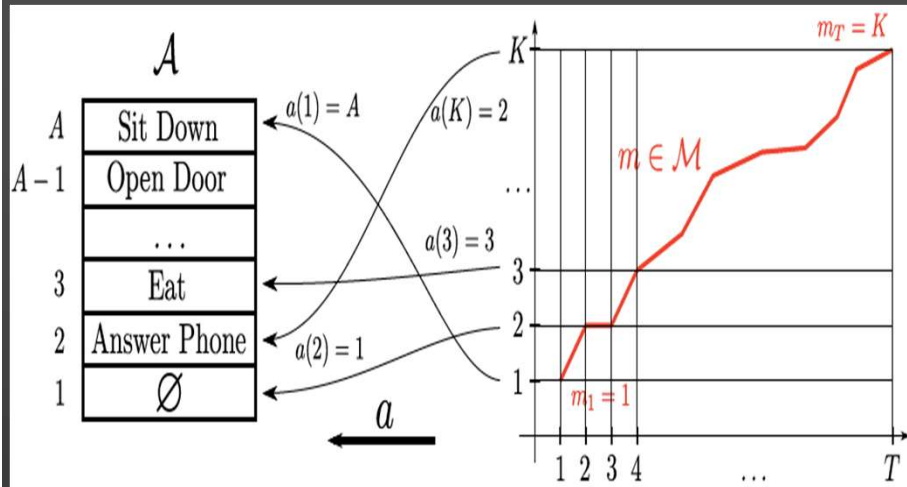
$$\min_{f \in \mathcal{F}} \left[\sum_{n=1}^N \min_{m \in \mathcal{M}} \frac{1}{T} \sum_{t=1}^T \ell(a_n(m_t), f(x_n(t))) \right] + \lambda \Omega(f)$$

Changing the representation

$$a, m \rightarrow Z$$



Action labeling under ordering constraints



$$\min_{f \in \mathcal{F}} \left[\sum_{n=1}^N \min_{m \in \mathcal{M}} \frac{1}{T} \sum_{t=1}^T \ell(a_n(m_t), f(x_n(t))) \right] + \lambda \Omega(f)$$

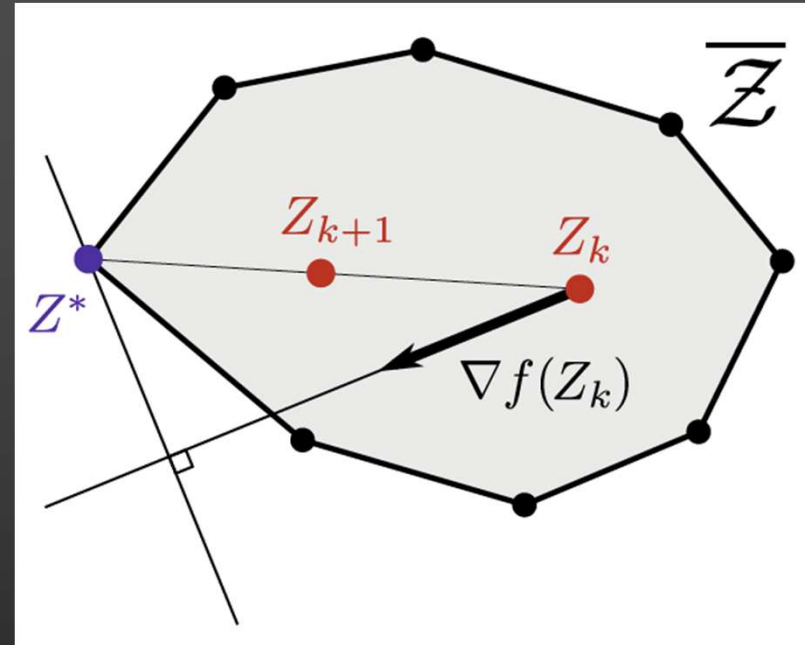
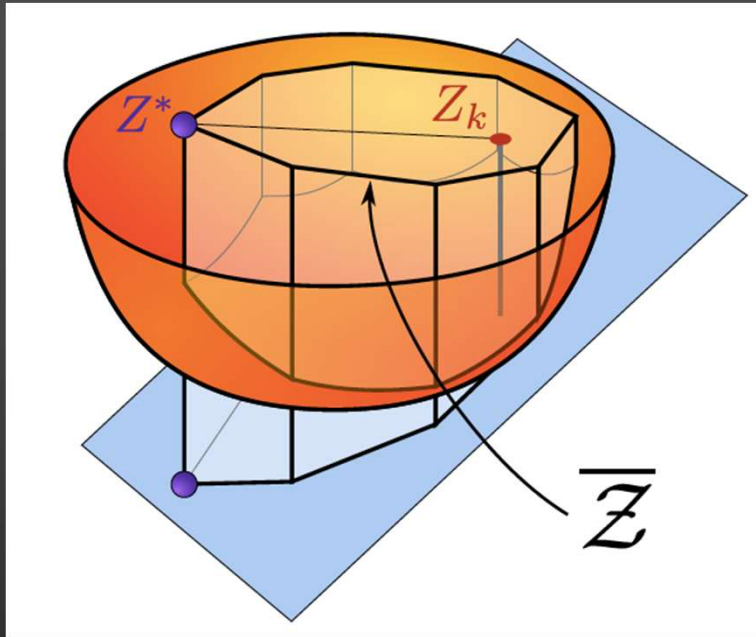


$$\min_{f \in \mathcal{F}, Z \in \mathcal{Z}} \frac{1}{T} \sum_{t=1}^T \ell(Z_t, f(x_t)) + \lambda \Omega(f) = \frac{1}{T} \|Z - XW - b\|_F^2 + \frac{\lambda}{2} \|W\|_F^2$$

$$\min_{Z \in \mathcal{Z}} \text{Tr}(ZZ^T B), \text{ where } B = \frac{1}{T} \Pi_T (I_T - X (X^T \Pi_T X + T \lambda I_d)^{-1} X^T) \Pi_T$$

- Minimize a convex quadratic function over a large discrete domain \mathcal{Z}
- Relaxed problem: minimize instead over $\overline{\mathcal{Z}} = \text{conv}(\mathcal{Z})$, then round up
- Difficulty: \mathcal{Z} (and thus $\overline{\mathcal{Z}}$) are defined by complex implicit constraints
- Frank-Wolfe to the rescue!

The Frank-Wolfe algorithm (1956)



Repeat until convergence :

- Replace the cost surface by its tangent plane, and minimize over $\overline{\mathcal{Z}}$
- $Z_{k+1} = (1-\gamma) Z_k + Z^*$
- No need for a projection step, converges to global minimum
- DP can be used to minimize linear functions over \mathcal{Z} and thus $\overline{\mathcal{Z}}$
- DP can also be used for rounding

Temporal action localization

Clip number 0101

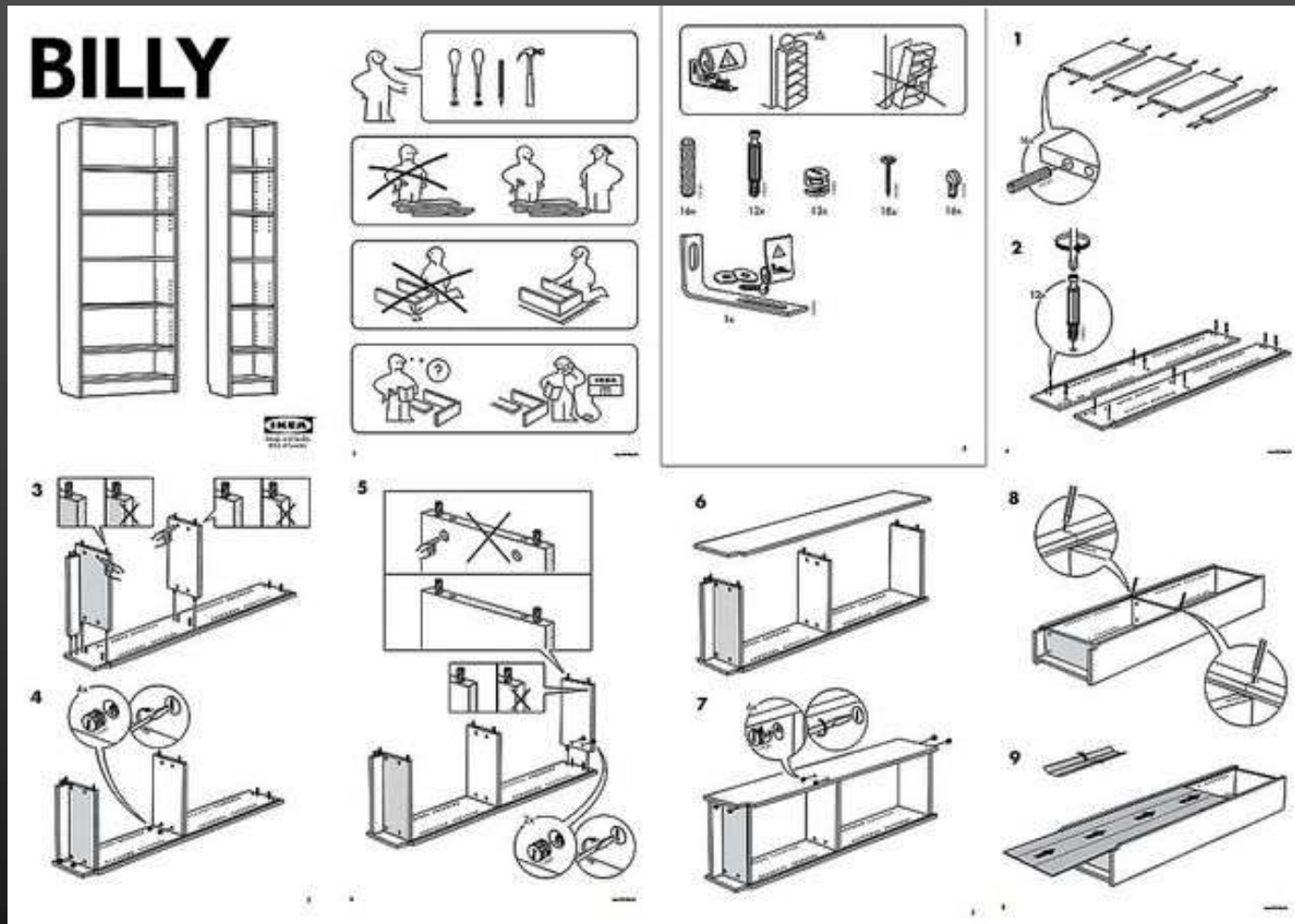
(Bojanowski et al., ECCV'14)

Learning from narrated instructional videos (Alayrac et al., CVPR'16, PAMI'17)



We can get two hands on it and we can exert some real leverage

Making assembly plans



Automatically produce a sequence of instructions from narrated videos

Input: a set of narrated videos and their text transcriptions



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.

Input: a set of narrated videos and their text transcriptions



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.

Output:

- Sequence of main steps

1. Loosen nuts

2. Jack the car

3. Remove the flat tire

Input: a set of narrated videos and their text transcriptions



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.

Output:

- Sequence of main steps
- Visual and textual models of the steps

1. Loosen nuts

2. Jack the car

3. Remove the flat tire

Input: a set of narrated videos and their text transcriptions



The image displays two filmstrips representing video frames. The top filmstrip shows a person working on a car wheel. The first frame is highlighted with an orange box, the second with a blue box, and the fifth with a pink box. Below it, a text transcription reads: "Start by loosening each bolt. Then locate the jack and lift the car. Now you can remove the bolts and then the wheel." The words "loosening", "bolt", "lift", "car", and "remove" are highlighted in orange, blue, blue, blue, and pink respectively, matching the boxes in the filmstrip above. The bottom filmstrip shows a person using a jack to lift a car. The third frame is highlighted with an orange box, the fourth with a blue box, and the sixth with a pink box. Below it, a text transcription reads: "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." The words "undo", "nuts", "jack", "car", "remove", and "tire" are highlighted in orange, blue, blue, blue, pink, and pink respectively, matching the boxes in the filmstrip above.

Output:

- Sequence of main steps
- Visual and textual models of the steps
- Temporal localization of the steps

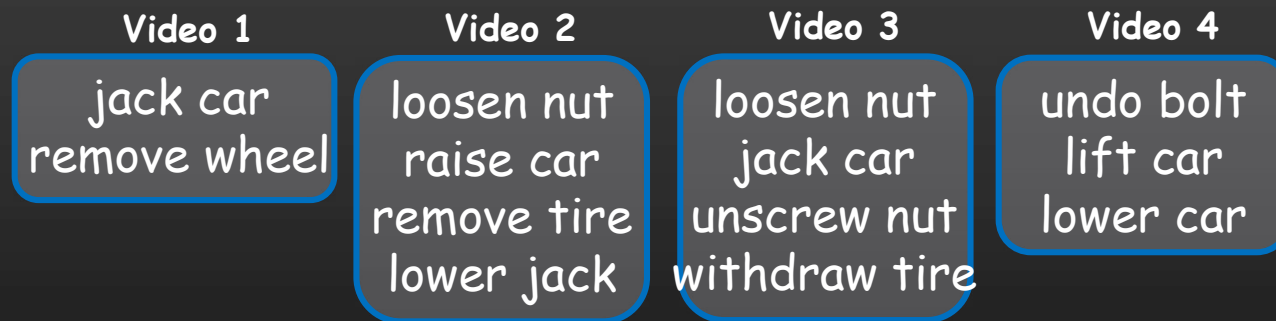
1. Loosen nuts

2. Jack the car

3. Remove the flat tire

Text alignment in multiple sequences


- Narrations are first processed into sequence of direct object relations (dobj)
 - Ex: "Let's now jack the car" → dobj = [jack car]
- Similarity scores from Wordnet
 - Ex: undo bolt ≈ loosen nut, jack car ≠ remove wheel




Text alignment in multiple sequences

Video 1	Video 2	Video 3	Video 4
jack car	loosen nut	loosen nut	undo bolt
remove wheel	raise car	jack car	lift car
	remove tire	unscrew nut	lower car
	lower jack	withdraw tire	

Text alignment in multiple sequences

Video 1	Video 2	Video 3	Video 4
	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
remove wheel	remove tire	unscrew nut	lower car
	lower jack	withdraw tire	

Text alignment in multiple sequences

Video 1	Video 2	Video 3	Video 4
	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
remove wheel	remove tire	unscrew nut	lower car
	lower jack	withdraw tire	

Text alignment in multiple sequences

Video 1	Video 2	Video 3	Video 4
<input type="checkbox"/>	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
<input type="checkbox"/>	<input type="checkbox"/>	unscrew nut	<input type="checkbox"/>
remove wheel	remove tire	withdraw tire	lower car
<input type="checkbox"/>	lower jack	<input type="checkbox"/>	

Text alignment in multiple sequences

Video 1	Video 2	Video 3	Video 4
<input type="checkbox"/>	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
<input type="checkbox"/>	<input type="checkbox"/>	unscrew nut	<input type="checkbox"/>
remove wheel	remove tire	withdraw tire	<input type="checkbox"/>
<input type="checkbox"/>	lower jack	<input type="checkbox"/>	lower car

Text alignment in multiple sequences

We seek to minimize the sum of pairwise costs:

Video 1	Video 2	Video 3	Video 4
<input type="checkbox"/>	loosen nut	loosen nut	<input type="checkbox"/> undo bolt
jack car	raise car	jack car	lift car
<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/> unscrew nut	<input type="checkbox"/>
remove wheel	remove tire	withdraw tire	<input type="checkbox"/>
<input type="checkbox"/>	lower jack	<input type="checkbox"/>	lower car

$$\begin{aligned} C = & c(\emptyset, \text{'loosen nut'}) + \dots + c(\emptyset, \text{'undo bolt'}) \\ & + \dots + c(\text{'jack car'}, \text{'raise car'}) \\ & + \dots + c(\emptyset, \text{'unscrew nut'}) + \dots \end{aligned}$$

Text alignment in multiple sequences

$$\min_{\phi} \sum_{(n,m)} \sum_{l=1}^L c(\phi(d^n)_l, \phi(d^m)_l)$$

Alignment cost

Sum over all pairs

Sum over template lines

Mapping of sequence n to a common template

[Wang and Jiang 1994, Higgins and Sharp, 1988]

Text alignment in multiple sequences

- Rewrite as an integer quadratic program

$$\min_U \text{Tr}(U^T B U), \text{ subject to } U \in \bar{\mathcal{U}}$$

- Solve relaxed problem with Frank-Wolfe
- Round up the solution

Text alignment in multiple sequences

Video 1	Video 2	Video 3	Agreement	Video 4
<input type="checkbox"/>	loosen nut	loosen nut	3	undo bolt
jack car	raise car	jack car	4	lift car
<input type="checkbox"/>	<input type="checkbox"/>	unscrew nut	1	<input type="checkbox"/>
remove wheel	remove tire	withdraw tire	3	<input type="checkbox"/>
<input type="checkbox"/>	lower jack	<input type="checkbox"/>	2	lower car

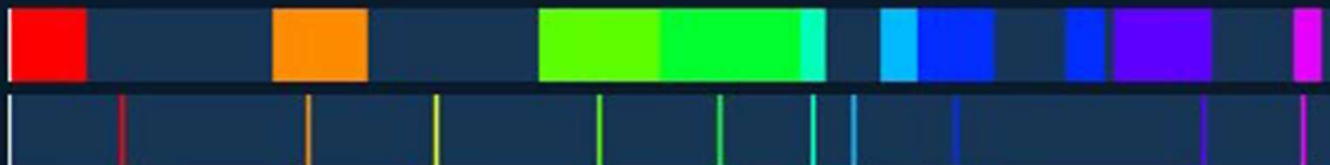
Text alignment in multiple sequences

Video 1	Video 2	Video 3	Video 4	Agreement	Discovered list of steps
∅	loosen nut	loosen nut	undo bolt	3	1) Loosen nut 2) Jack car 3) Remove wheel
jack car	raise car	jack car	lift car	4	
∅	∅	unscrew nut	∅	1	
remove wheel	remove tire	withdraw tire	∅	3	
∅	lower jack	∅	lower car	2	

.. and then use method similar to previous one for temporal localization



- get things out
- start loose
- brake on
- jack up
- unscrew wheel
- withdraw wheel
- put wheel
- screw wheel
- jack down
- tight wheel



GROUND TRUTH

Video Prediction

Activity discovery from images and words

Make kimchi fried rice



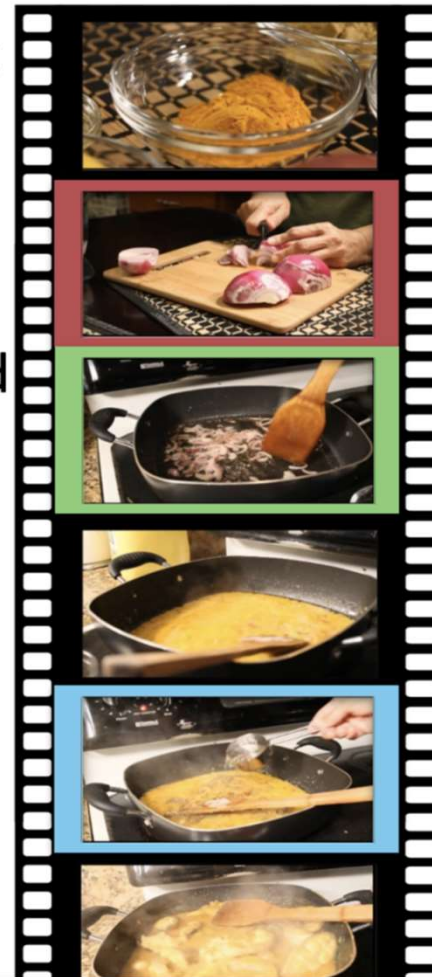
"I'm going to start off by chopping up an onion.

Get your pan on a nice high heat and add some oil...

...and just stir this through...

... You want to fry the rice now mixing every now and then..."

Make Kerala fish curry



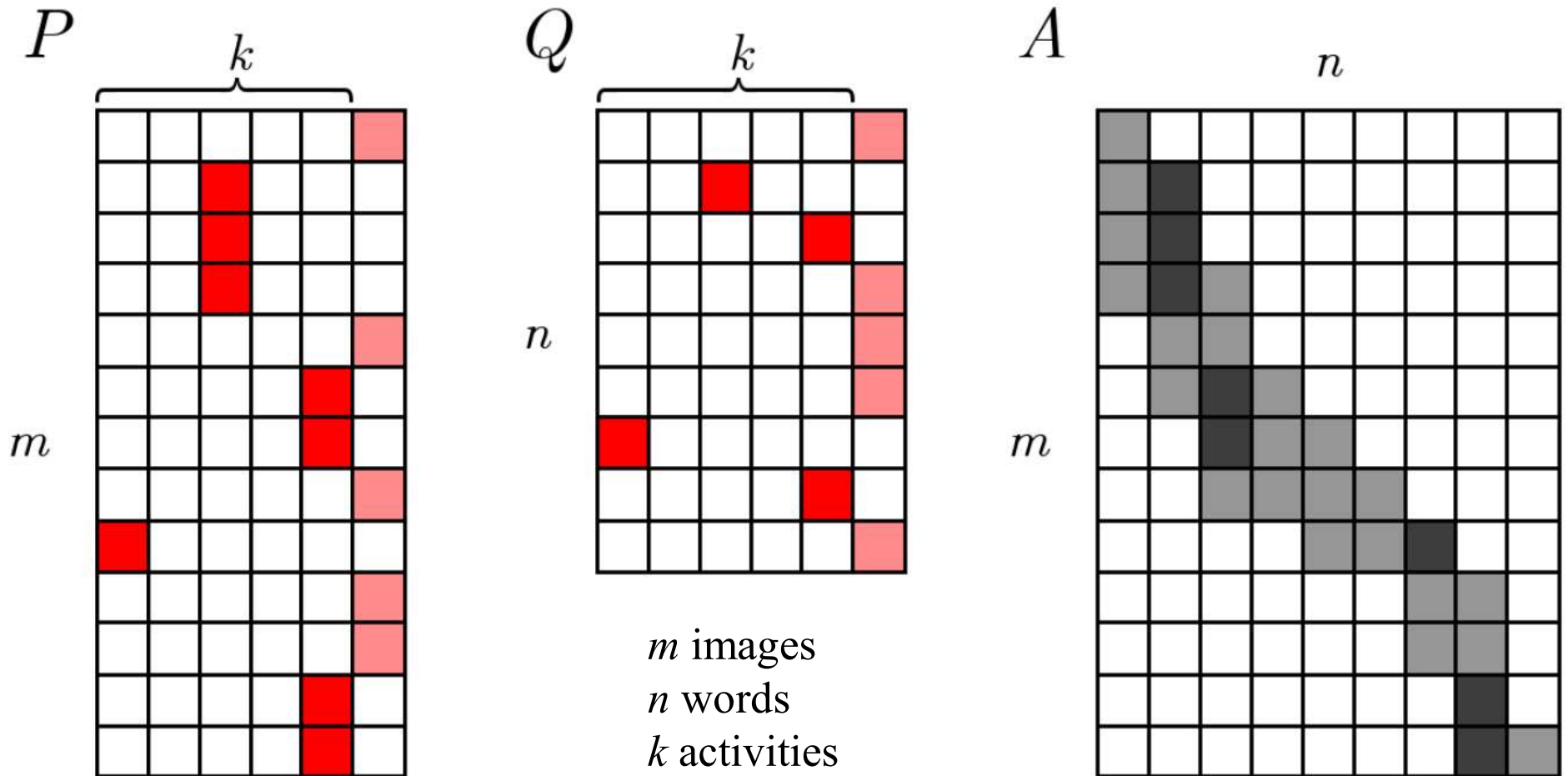
"...next you're going to take a full onion...

...So make sure you stir the onions around...

...you're also going to add one cup of water to the pan..."

(Alakuijala, Mairal, Ponce, Schmid, 2019)

Activity discovery from images and words



(Alakuijala, Mairal, Ponce, Schmid, 2019)

Discovered ingredients of "performing CPR"

move blood
put hands
place hands
place finger
put hand
use hands
locate pulse
take hand
put finger
hand
place hand
use method
attach patient
keep hand
get services
position hands

feel air
perform
take finger
clutch chest
lift chin
do compression
deliver breath
see breathing
slide finger
have signs
save organs
cause air
compression
protect rescuer
rub bone
see rise
contact em
put heel
use maneuver
start breathing
check breathing
apply pressure

reposition airway
reopen airway
open airway
have partner
complete cycle

locate placement
cover mouth
lift jaw
take breath
hear anything
make seal
take head
seal mouth
place mouth
allow chest
plug nose
take mouth
perform mouth
place heel
feel anything
pinch nostril
bring tongue
pinch nose
allow air
block airway

recheck minutes
use help
continue cycle
look body
continue cpr
keep victim
use finger
hear exchange
administrate cpr
summon help
give information
commence cpr
find one
get help
get cycle
need bit
arrive condition

wake child
locate nearer
have patient
repeat cycle
wake casualty
check scene
say nose
tap shoulder
have someone
do cpr
give command
keep blood
find center
approach casualty

step check
pump times
avoid fatigue
get arm
change role
use mask
repeat breath
do times
stop cpr
begin compression
release nose
begin compression
follow steps
give air
press times
begin cpr
start cpr
keep elbow
give compression
support weight
give second
compress one and a half
start cpr
use thumb
make rise
keep rhythm
stick mouth
breathe times
see app
have breath
cough movement

tilt forehead
keep chin
tilt finger
tilt head
look chest
use head
check airway
push ear
tilt finger
place mask
lock elbow
grasp chin
give breathing
put ear

establish responsiveness
interlock finger
help heart
say help
tell bystander
have person
check circulation
have mouthguard
avoid contact
push
keep interruption
do part
use mannequin
reach notch
administer breath
yell help
assess scene
have blood

How much supervision do we really need? (Cho et al., CVPR'15)



Strong

Weak

Very weak



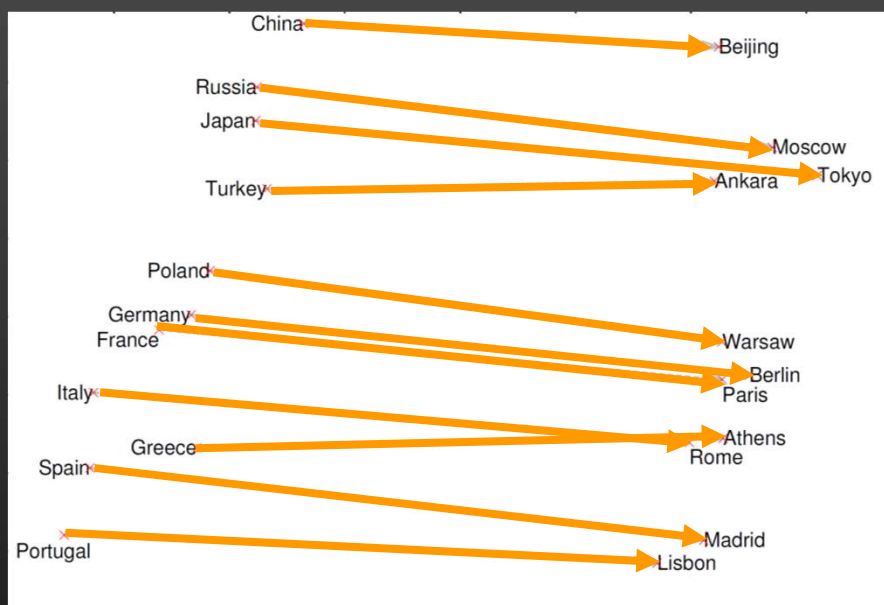
None

Object detection (Leibe et al.'08; Felzenszwalb et al.'10; Girshick et al.'14)
Weakly supervised localization (Chum'07; Pandey'11; Desaelers'12; Siva'12; Shi'13; Cinbis'14; Wang'14)
Co-segmentation/localization (Rother'06; Russell'06; Joulin'10; Kim'11; Vicente'11; Joulin'14; Tang'14)
Unsupervised discovery (Grauman & Darrell'05; Sivic et al.'05,08; Kim et al.'05,09)

Using context for self supervision

Ex: Word2vec (Mikolov et al., 2013): $w \rightarrow u(w) \in R^d$

- $u(\text{Paris}) \approx u(\text{France}) + [u(\text{Berlin}) - u(\text{Germany})]$
- Analogies as "linear algebra"



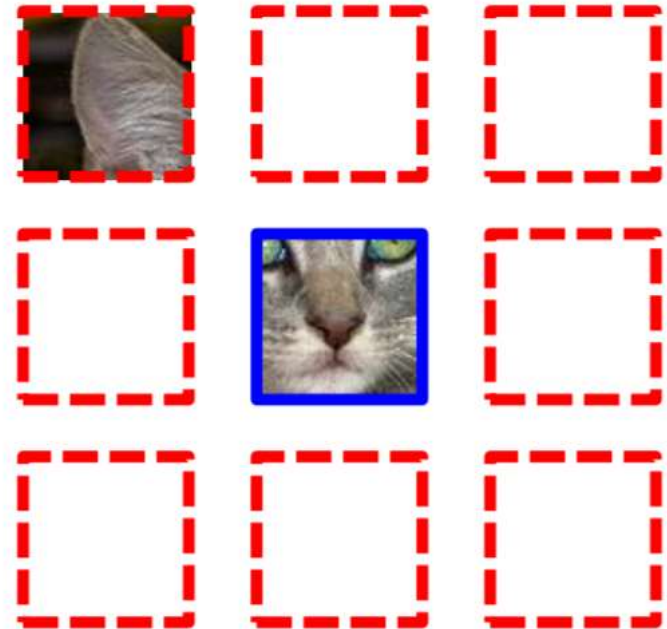
Note: Visualization in 2D
but $d \approx 300$

Modeling contextual info with co-occurrence statistics

$$\max_{u, v} \frac{1}{T} \sum_{t=1}^T \sum_{c \in N_t} \left(\log \sigma[u(w_t) \cdot v(c)] + \sum_{k=1}^K \log \sigma[-u(w_t) \cdot v(w_{\text{random}})] \right).$$

Example: Unsupervised Visual Representation Learning by Context Prediction (Doersch, Gupta, Efros, 2016)

Example:



Question 1:



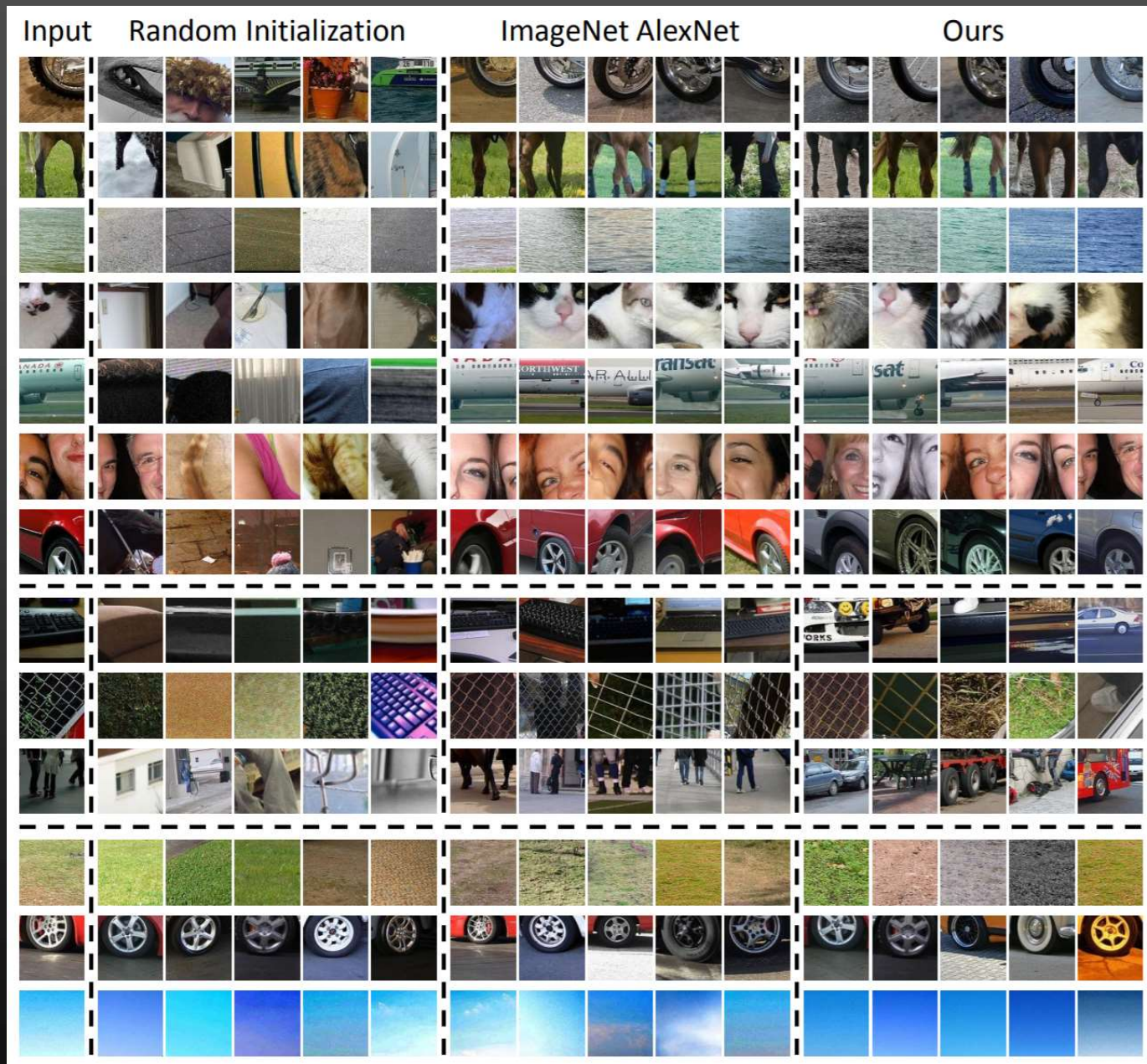
?

Question 2:

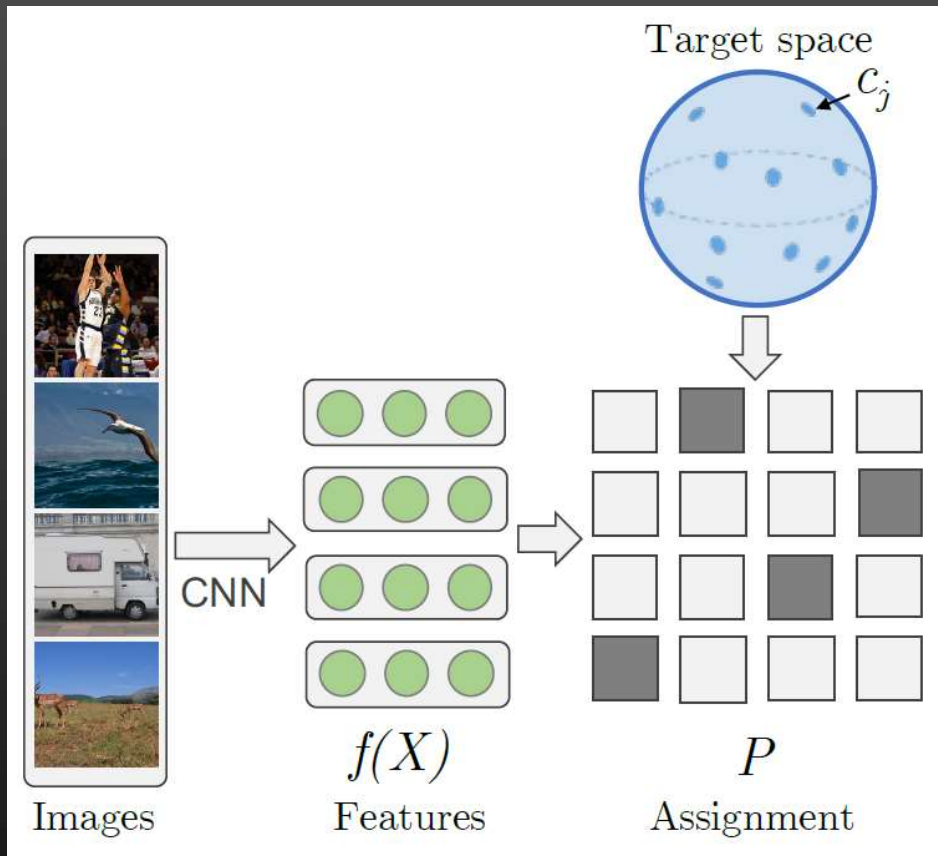


?

Retrieved nearest neighbors



Unsupervised feature learning (Bojanowski & Joulin, ICML'17)

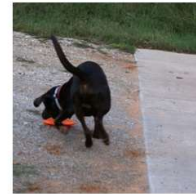


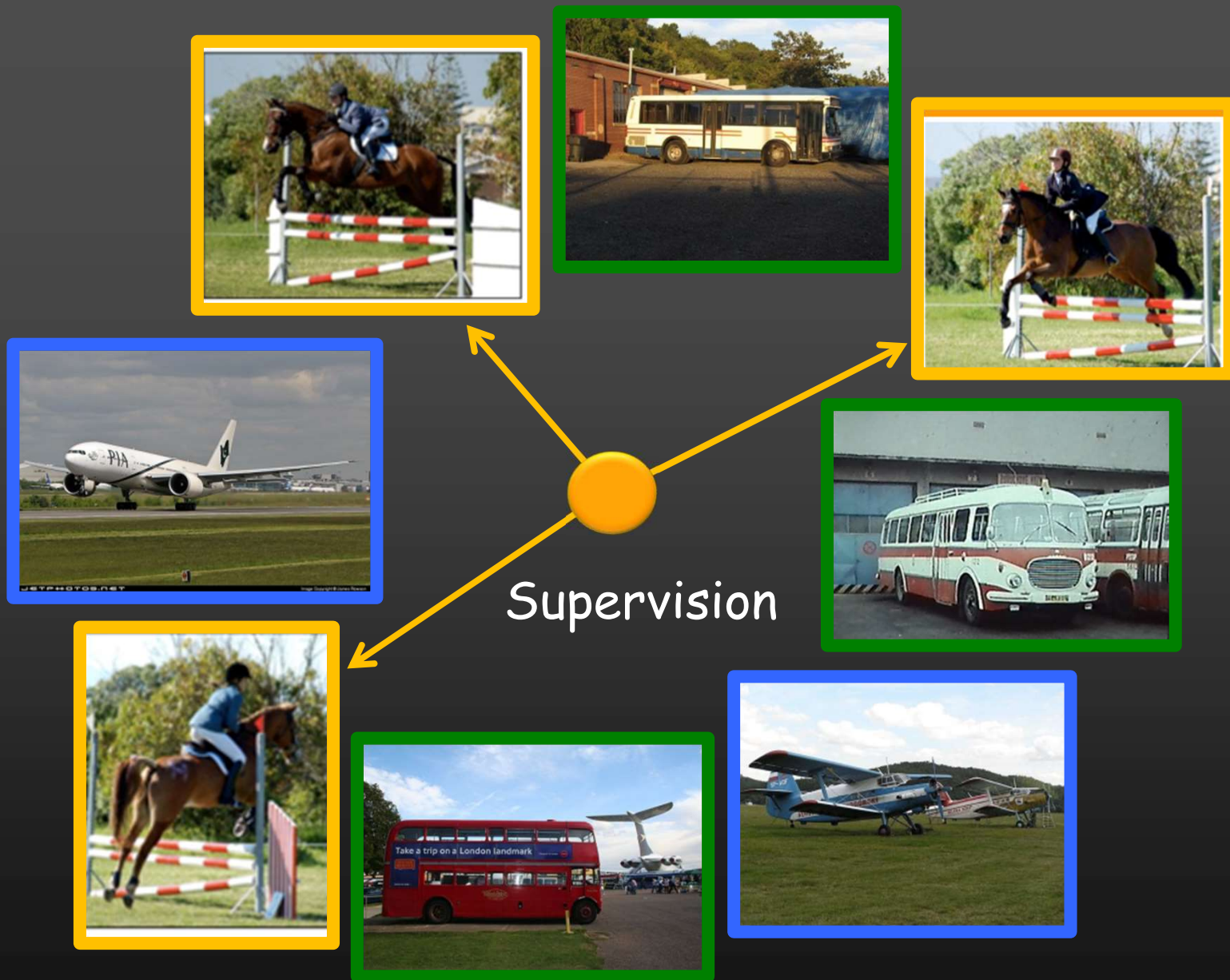
$$\min_{\theta} \min_{Y \in \mathbb{R}^{n \times d}} \frac{1}{2n} \|f_{\theta}(X) - Y\|_F^2$$



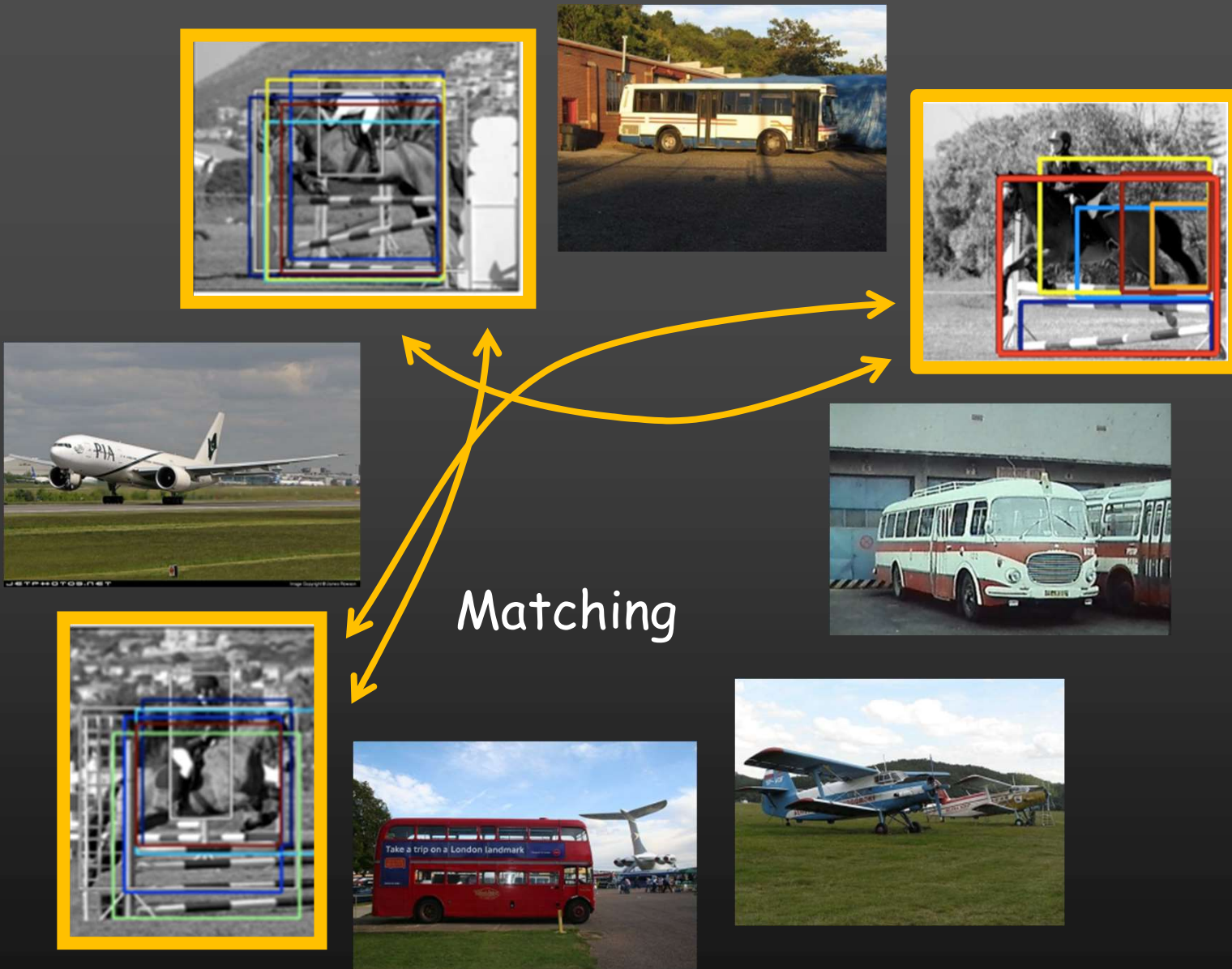
$$\max_{\theta} \max_{P \in \mathcal{P}} \text{Tr}(PC f_{\theta}(X)^{\top})$$

Retrieved nearest neighbors



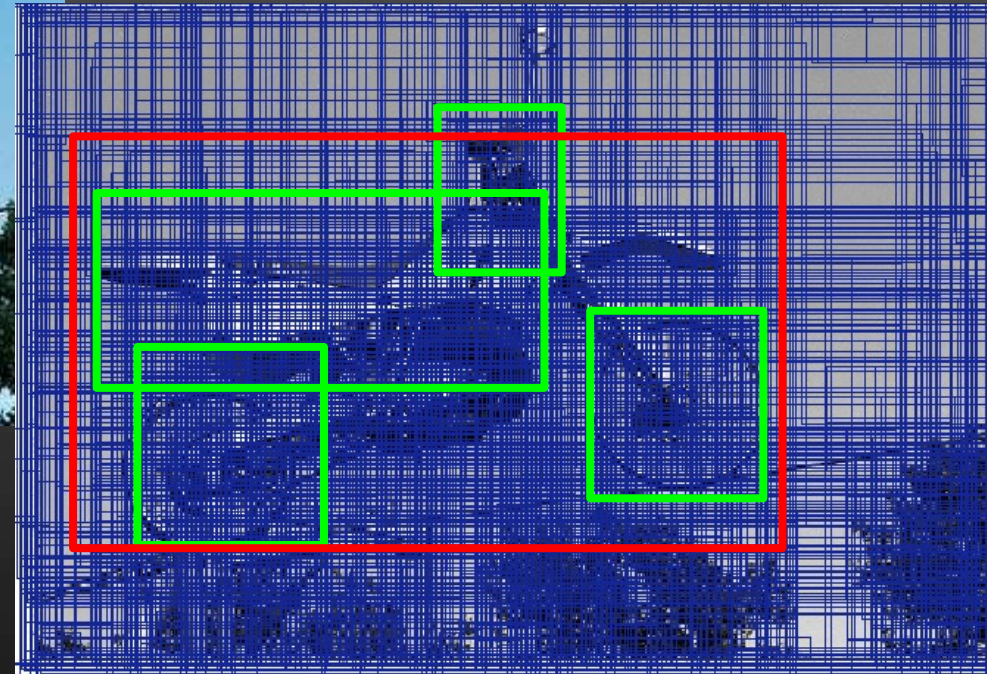


(Cho, Kwak, Schmid, Ponce, 2015)



(Russell et al.'06; Cho et al.'10; Deselaers et al.'10; Rubinstein & Joulin'13; Rubio et al.'13)

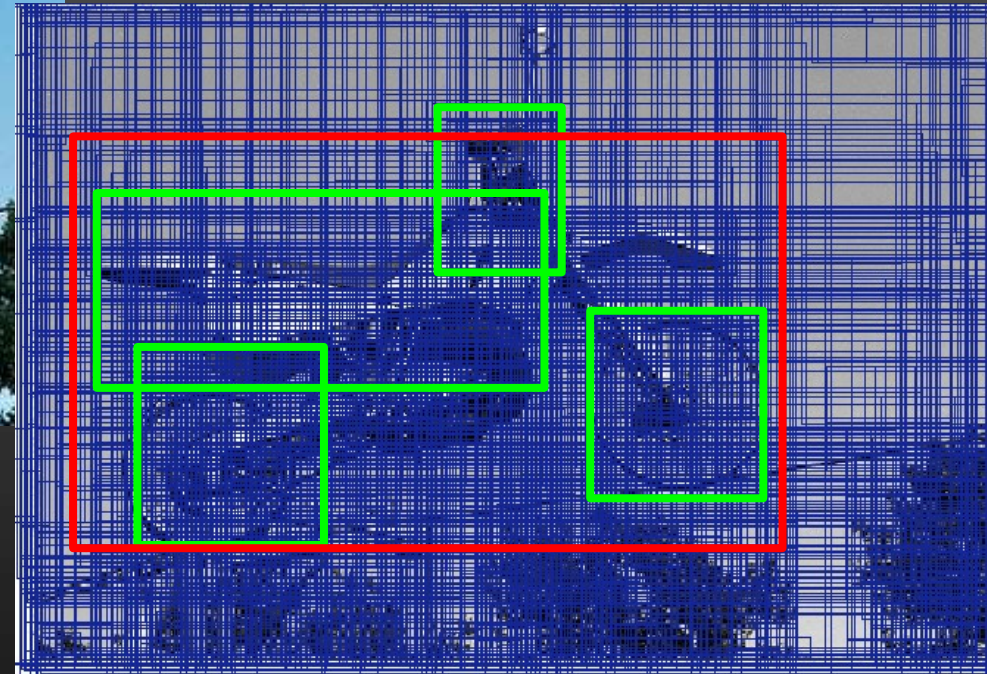
Finding **parts** and **objects** among region candidates



1000 to 4000 candidates per object

Here: Region proposals (Manen et al.'13, Uijlings et al.'13)
and HOG descriptors (Dalal & Triggs'05)

Finding **parts** and **objects** among region candidates



1000 to 4000 candidates per object

Caveat: These region proposals are supervised

Matching model - Probabilistic Hough matching

match

data

configuration

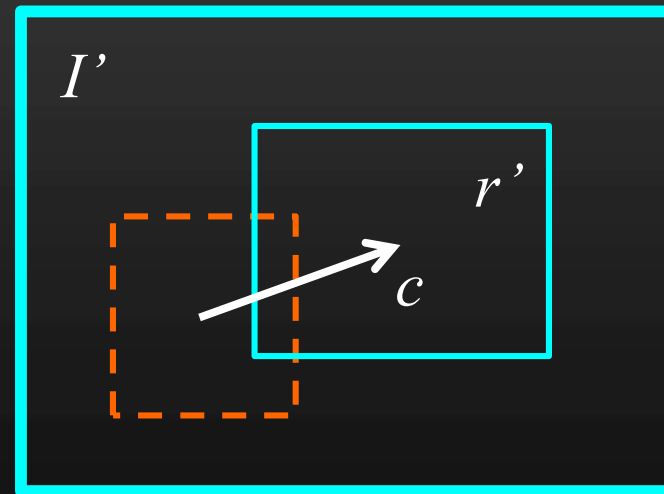
$$P(m | d) = \sum_c P(m | c, d) P(c | d)$$

two regions

region proposals

position+scale

$$m = [r, r']$$



Matching model - Probabilistic Hough matching

match

data

configuration

$$P(m | d) = \sum_c P(m | c) P(c | d)$$

$$= P(m_a) \sum_c P(m_g | c) P(c | d)$$

appearance

geometry

Matching model - Probabilistic Hough matching

- Bayesian model

$$\begin{aligned} P(m | d) &= \sum_c P(m | c) P(c | d) \\ &= P(m_a) \sum_c P(m_g | c) P(c | d) \end{aligned}$$

Matching model - Probabilistic Hough matching

- Bayesian model

$$\begin{aligned} P(m | d) &= \sum_c P(m | c) P(c | d) \\ &= P(m_a) \sum_c P(m_g | c) P(c | d) \end{aligned}$$

- Probabilistic Hough transform

$$\begin{aligned} P(c | d) &\approx H(c | d) = \sum_{m \in d} P(m | c) \\ &= \sum_{m \in d} P(m_a) P(m_g | c) \end{aligned}$$

(Hough'59; Ballard'81; Stephens'91; Leibe et al.'04; Maji & Malik'09; Barinova et al.'12)

Matching model - Probabilistic Hough matching

- Bayesian model

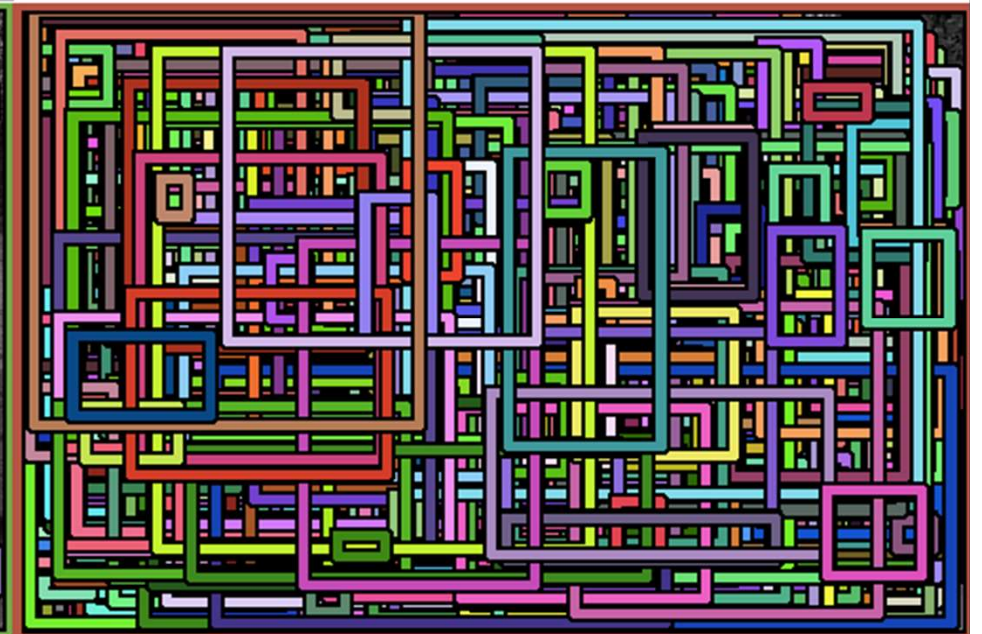
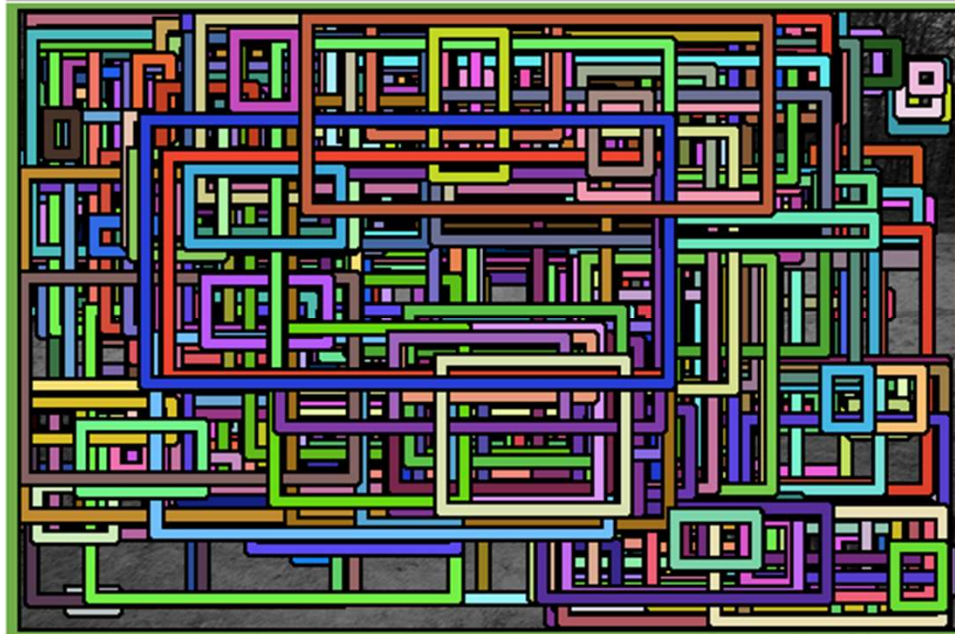
$$\begin{aligned} P(m | d) &= \sum_c P(m | c) P(c | d) \\ &= P(m_a) \sum_c P(m_g | c) P(c | d) \end{aligned}$$

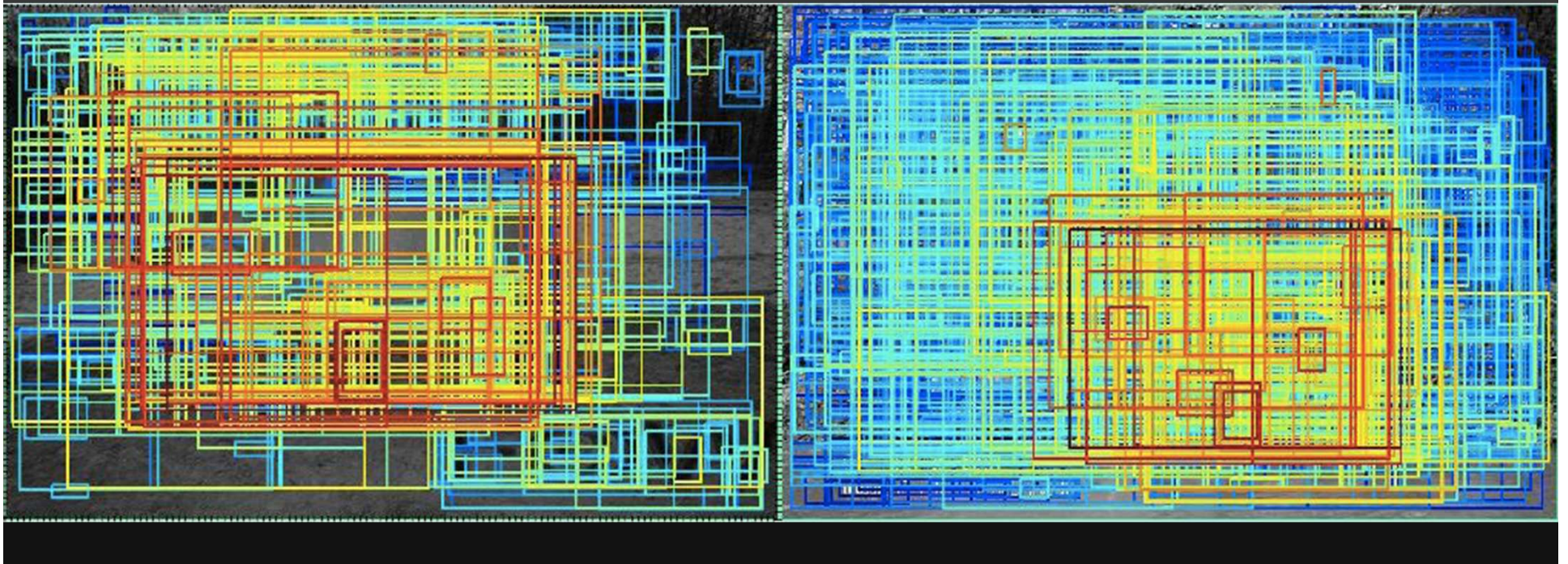
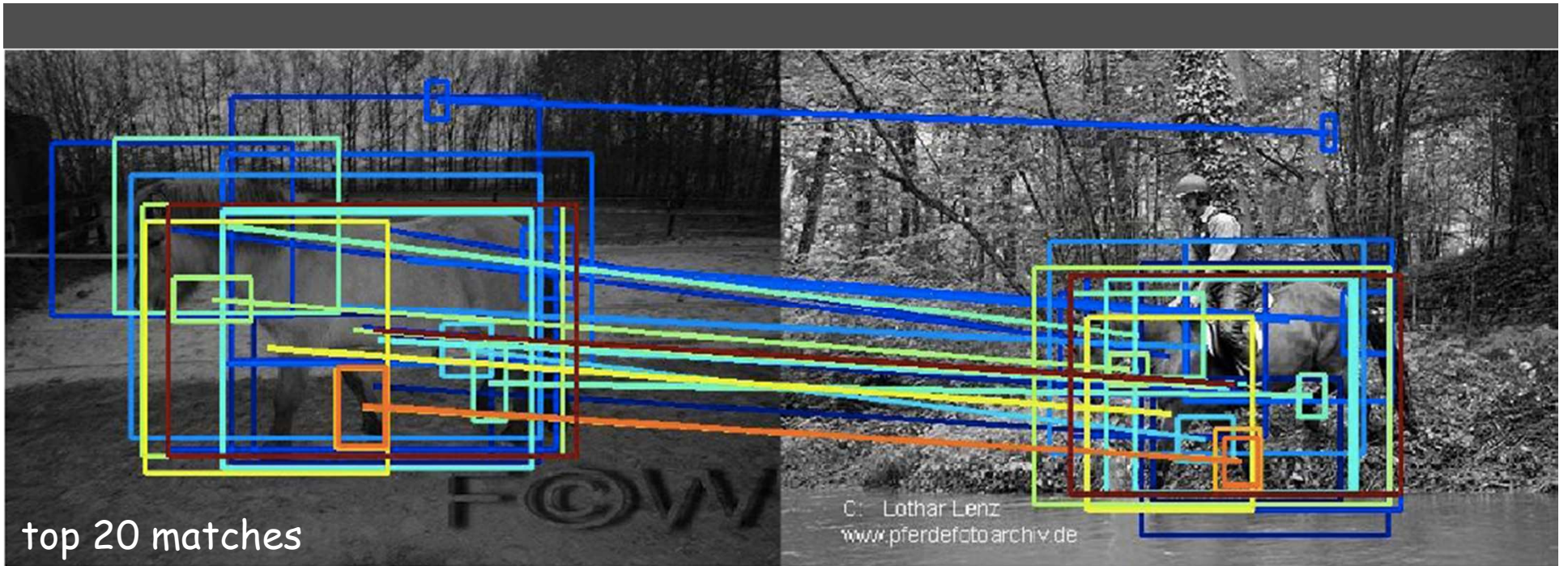
- Probabilistic Hough transform

$$\begin{aligned} P(c | d) &\approx H(c | d) = \sum_{m \in d} P(m | c) \\ &= \sum_{m \in d} P(m_a) P(m_g | c) \end{aligned}$$

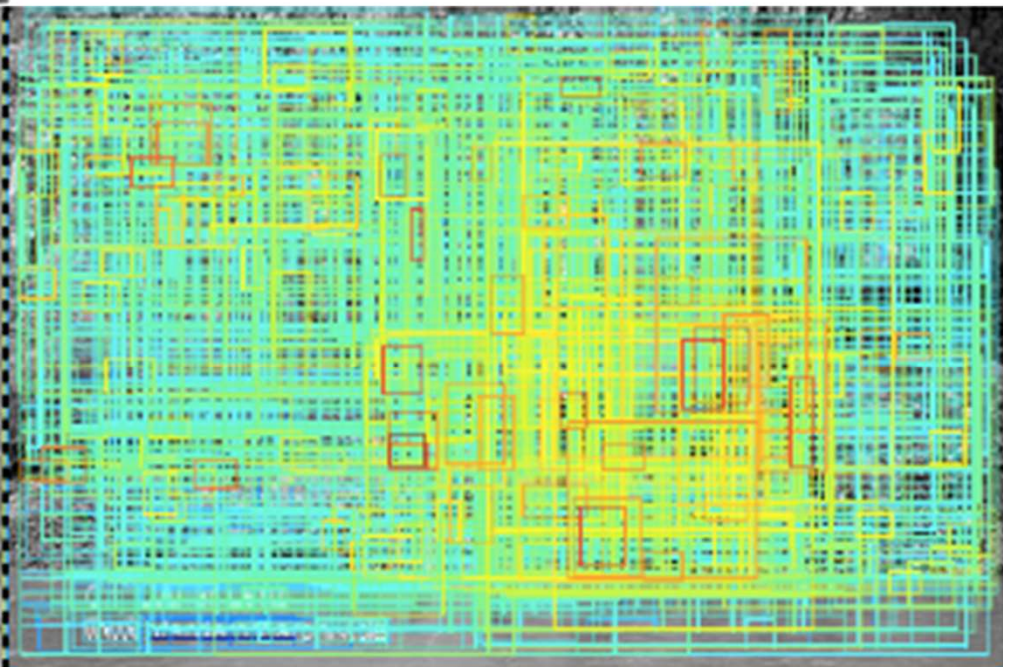
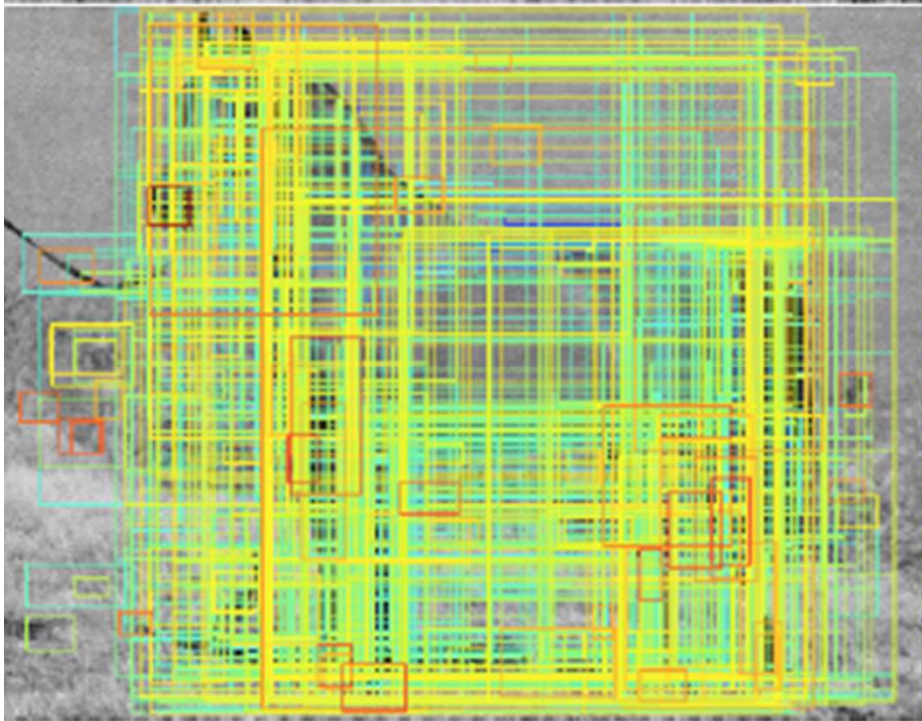
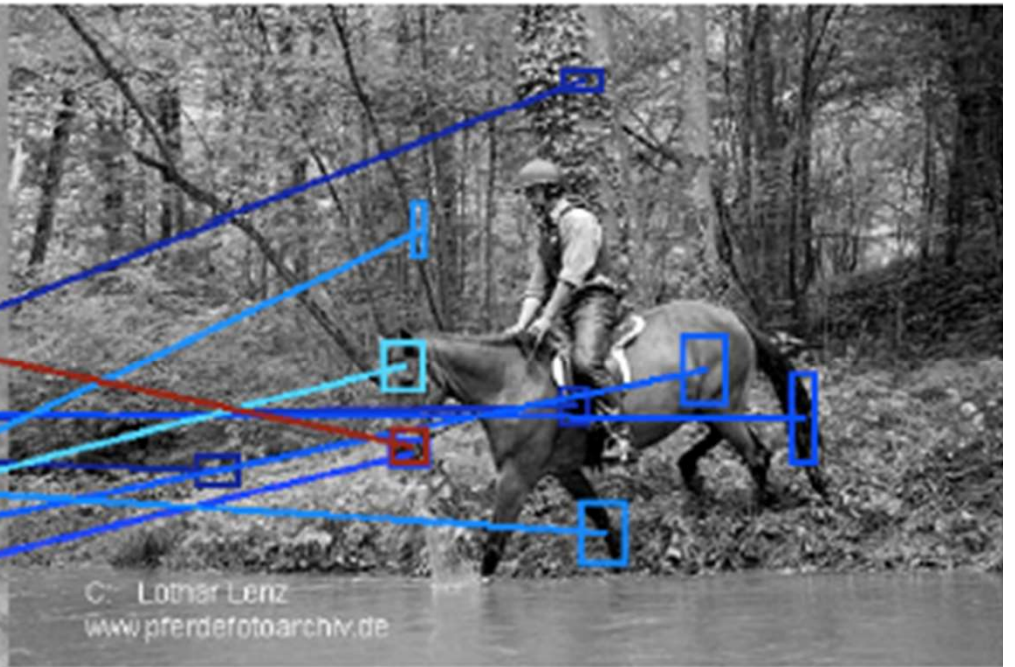
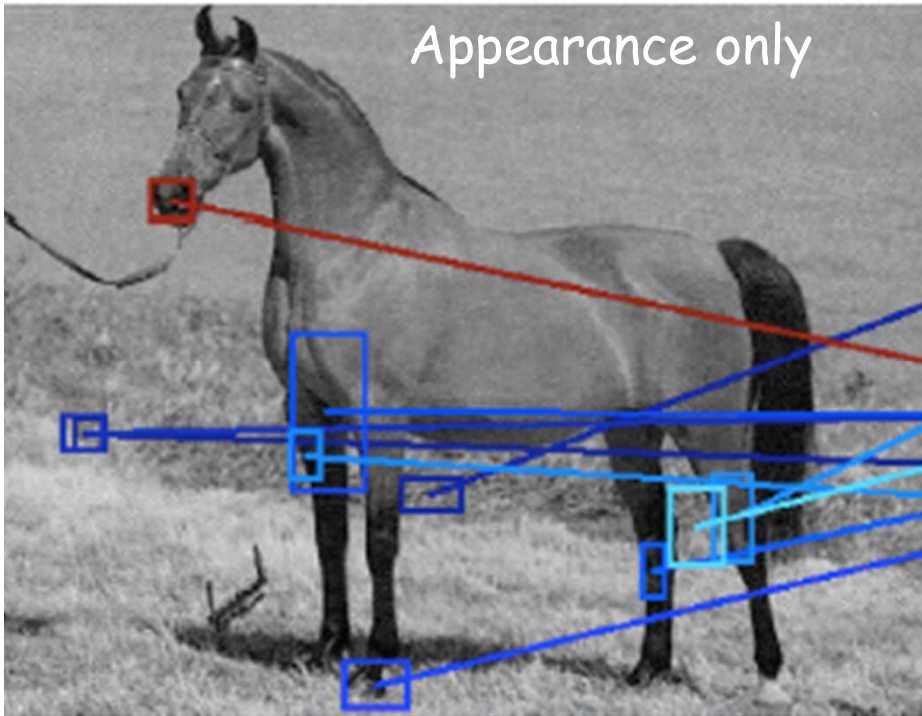
- Region confidence

$$C(r' | d) = \max_{r''} P(r' \leftrightarrow r'' | d)$$

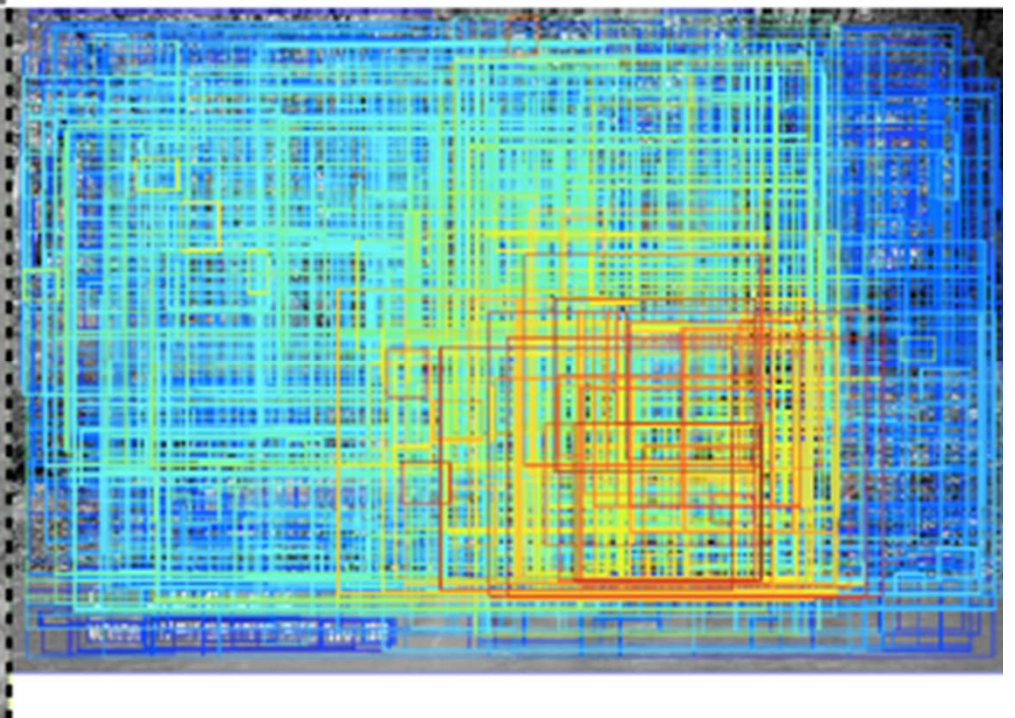
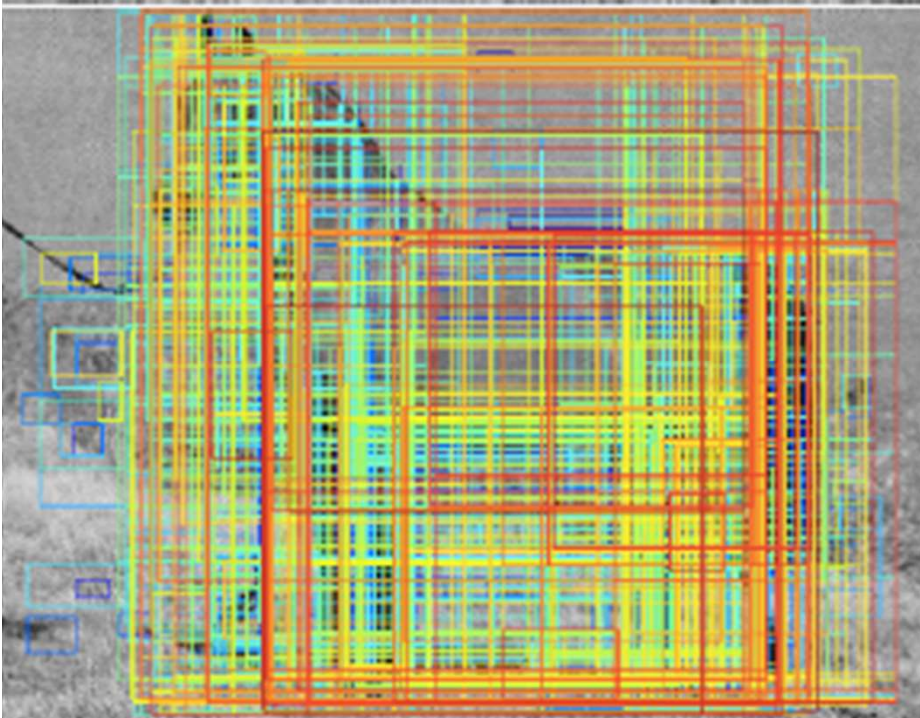
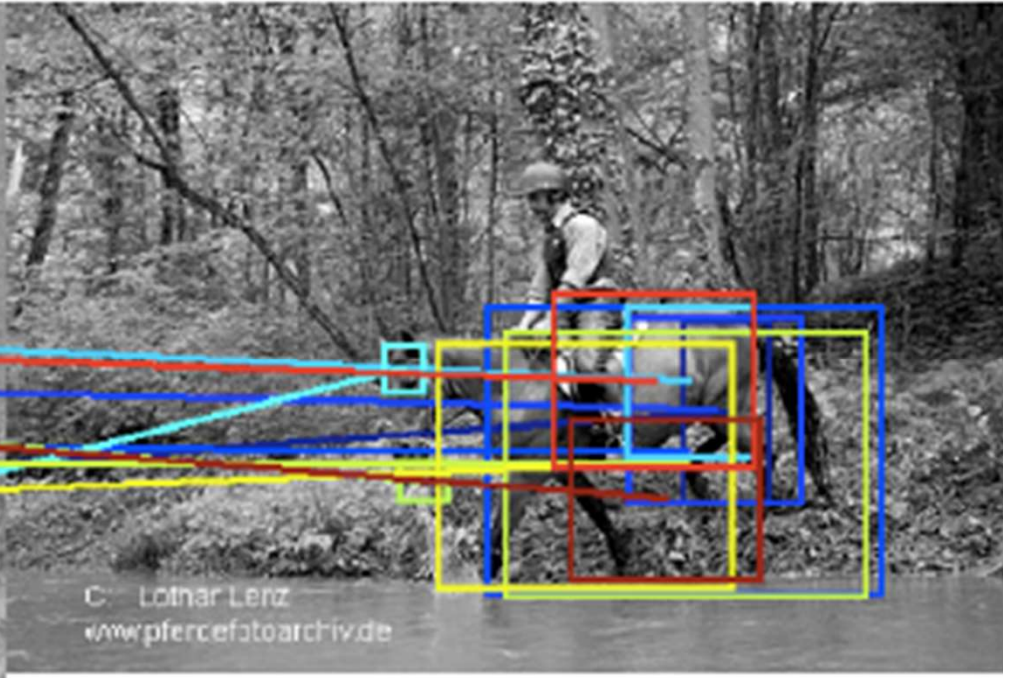
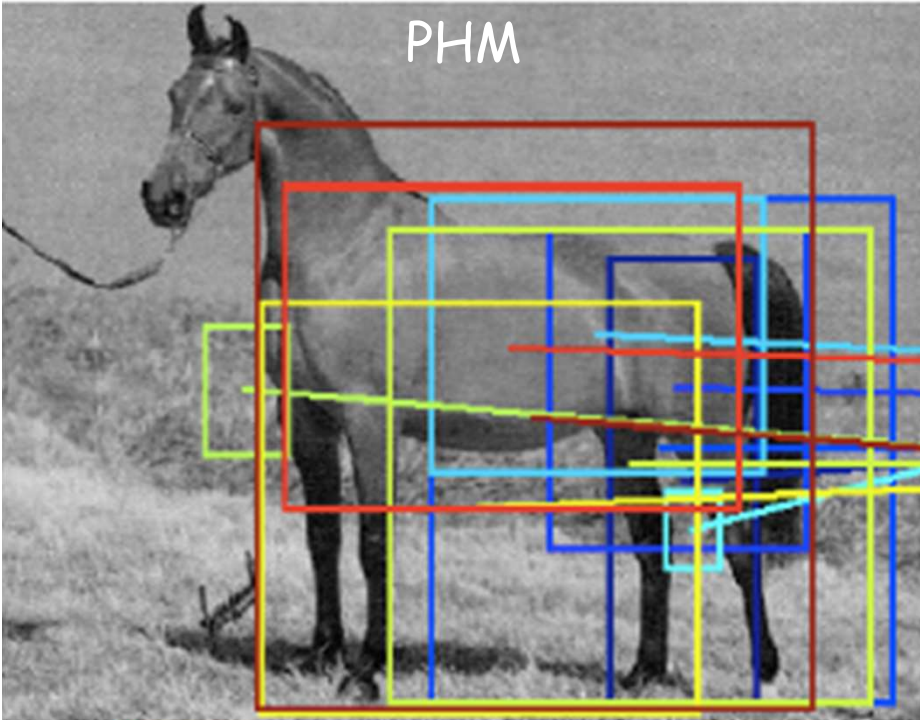




Appearance only



PHM



Matching model - Probabilistic Hough matching

- Bayesian model

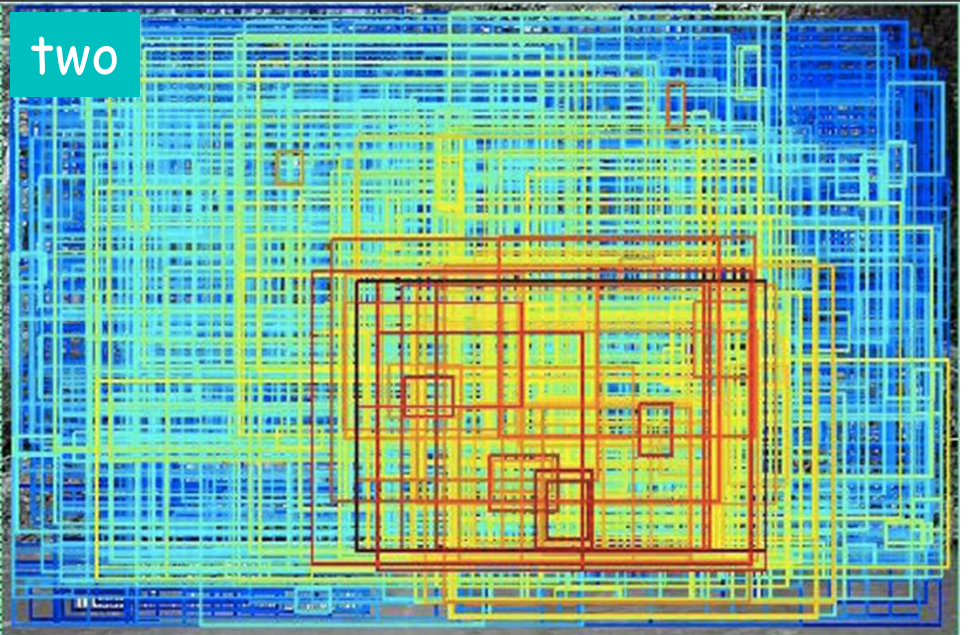
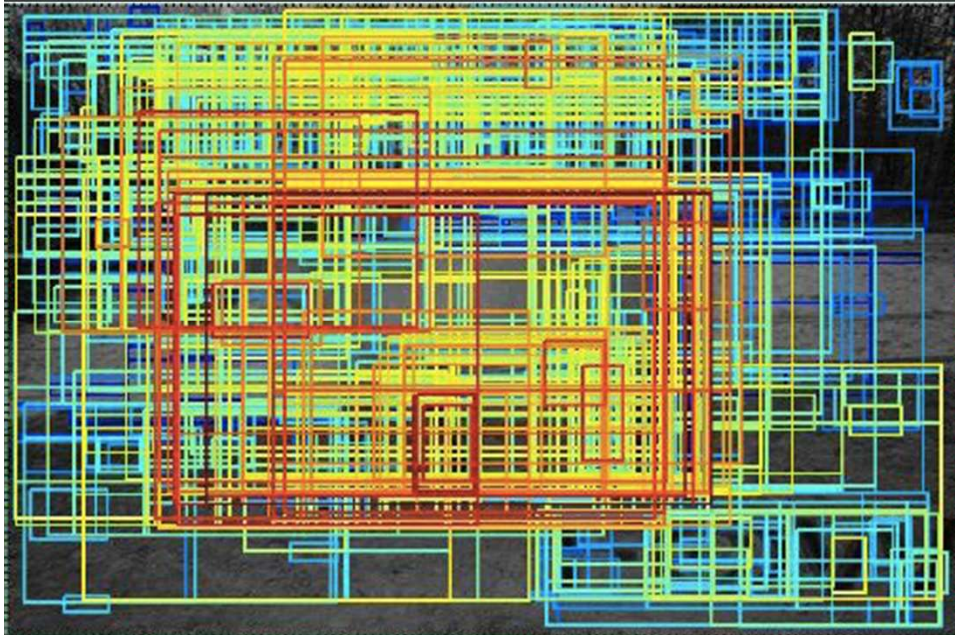
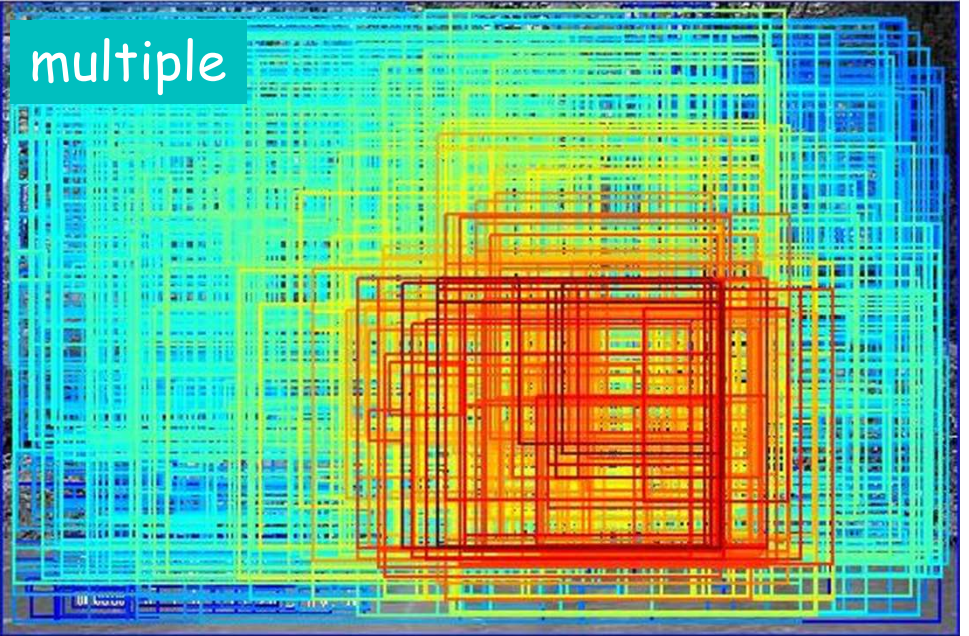
$$\begin{aligned} P(m | d) &= \sum_c P(m | c) P(c | d) \\ &= P(m_a) \sum_c P(m_g | c) P(c | d) \end{aligned}$$

- Probabilistic Hough transform

$$\begin{aligned} P(c | d) &\approx H(c | d) = \sum_{m \in d} P(m | c) \\ &= \sum_{m \in d} P(m_a) P(m_g | c) \end{aligned}$$

- Two images -> multiple images

$$C_{d'}(r') = \sum_{d''} C(r' | [d', d''])$$

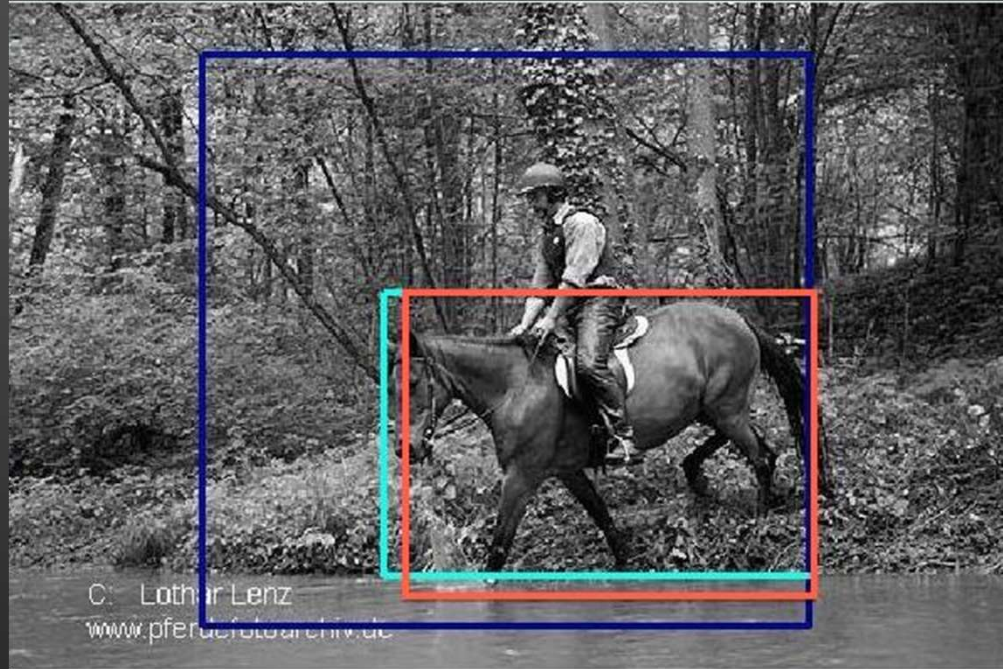


Stand-out scoring of part hierarchies



- Object regions should contain
 - more foreground than part regions
 - less background than larger regions

Stand-out scoring of part hierarchies



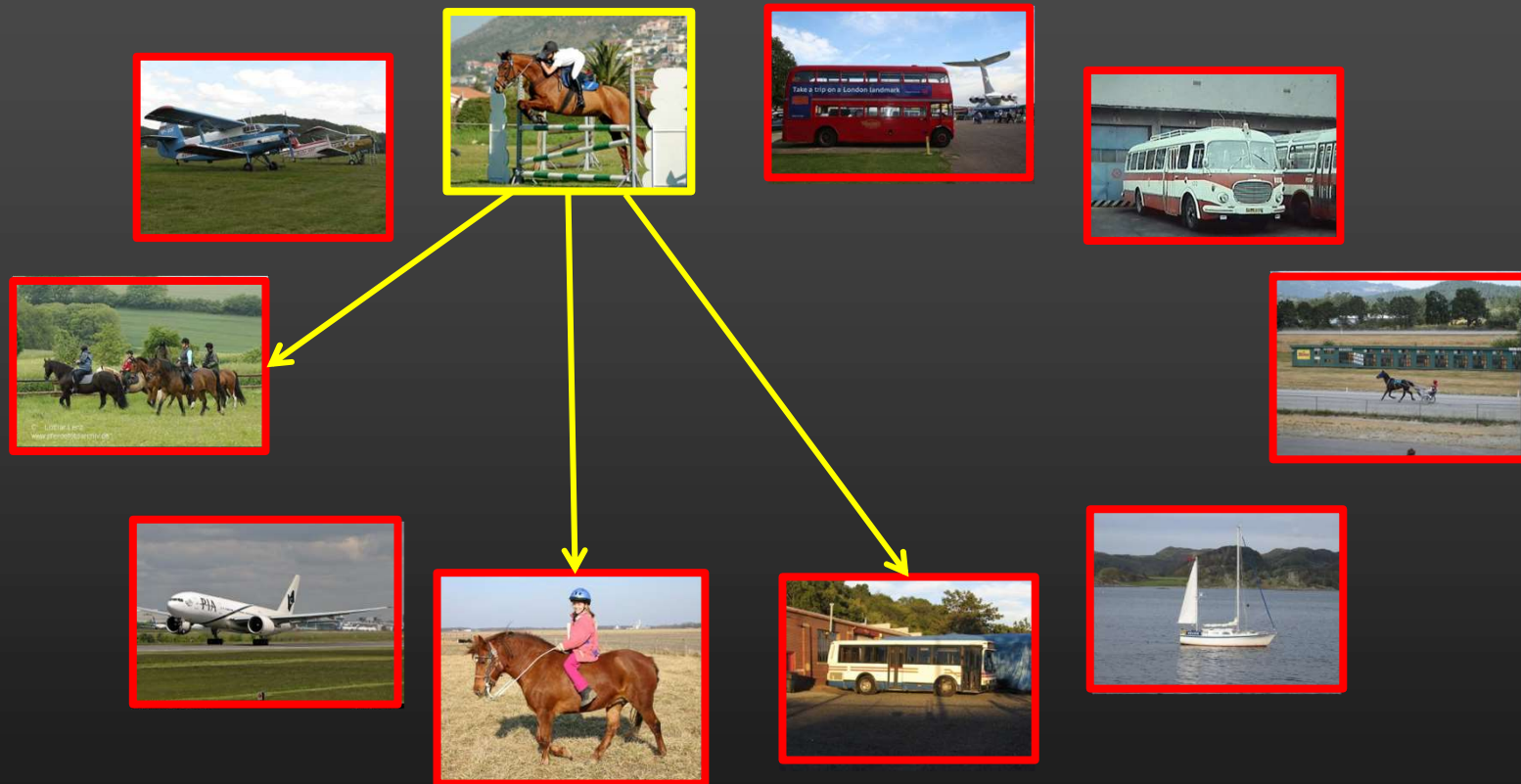
- Object regions should contain
 - more foreground than part regions
 - less background than larger regions
- $S_d(r) = C_d(r) - \max_{r' \supset r} C_d(r')$

A simple iterative algorithm



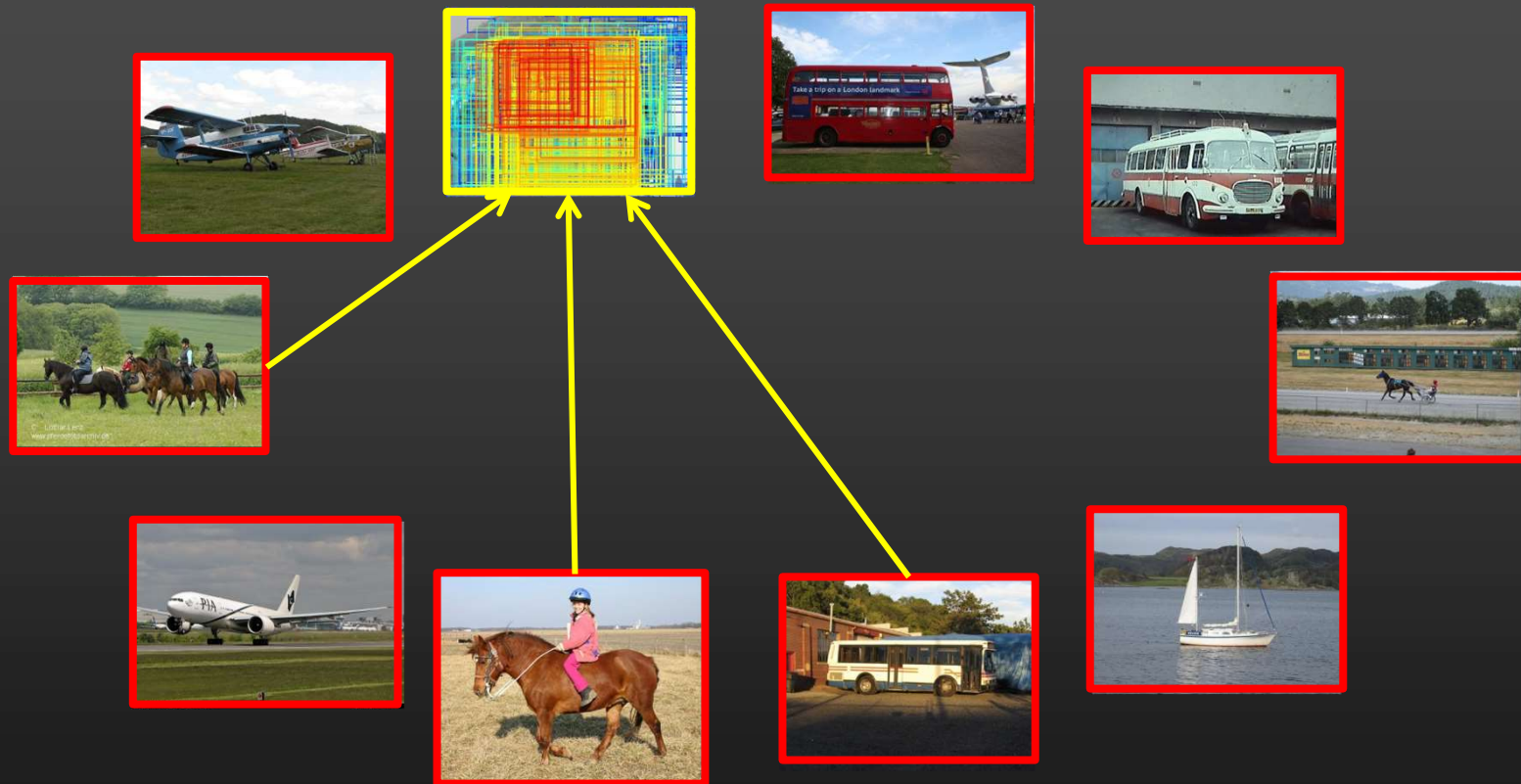
Initialize

A simple iterative algorithm



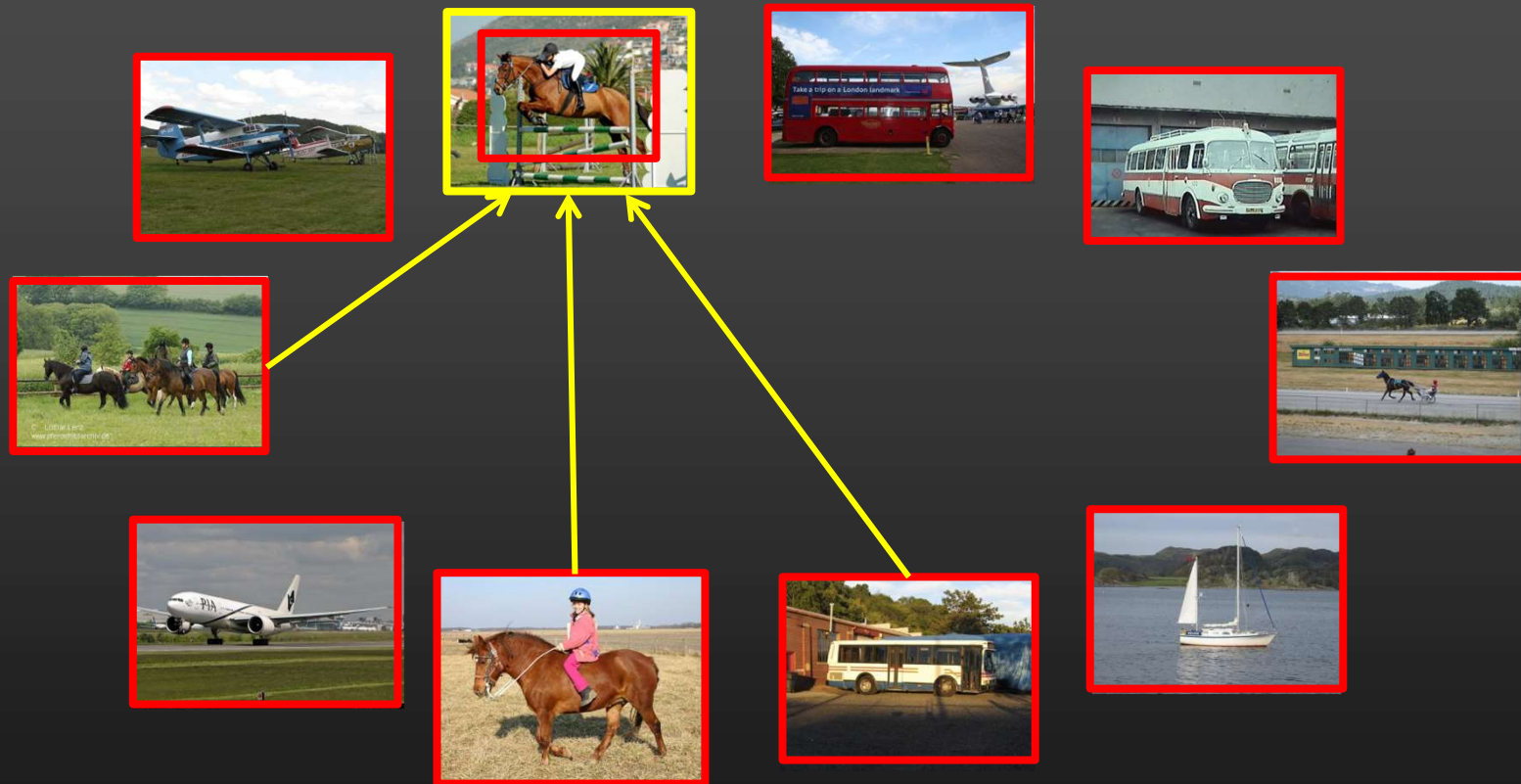
Retrieve 10 nearest neighbors (Oliva & Torralba'06)

A simple iterative algorithm



Match

A simple iterative algorithm



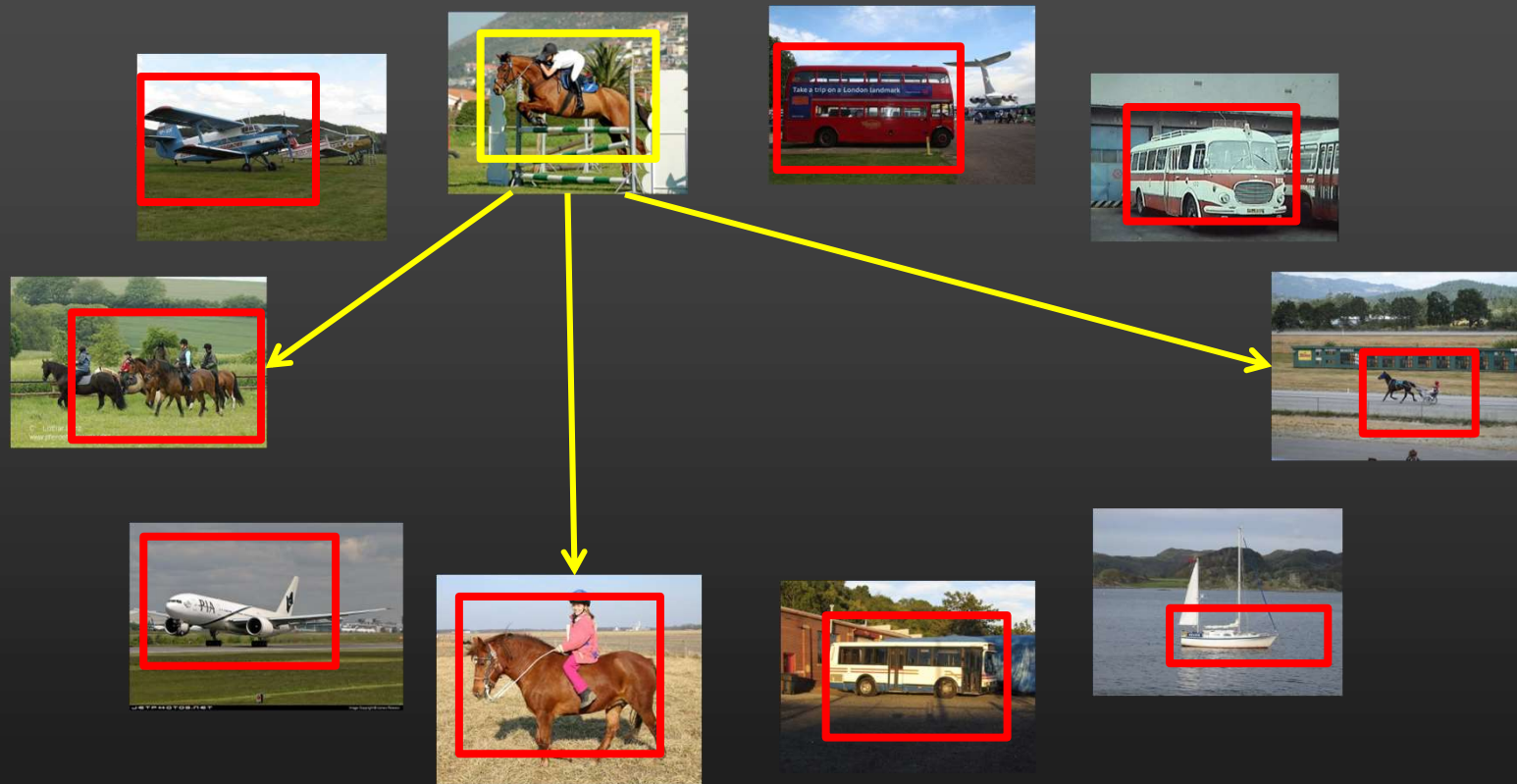
Localize

A simple iterative algorithm



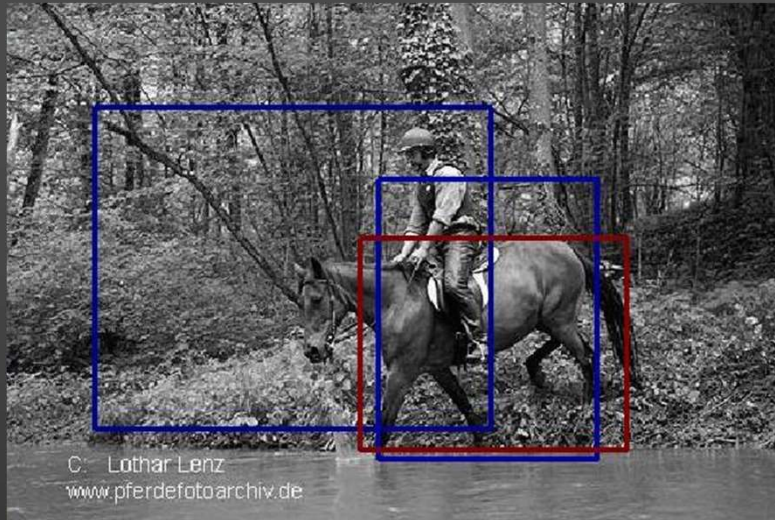
Localize

A simple iterative algorithm



Retrieve using top 20 confidence scores, etc.

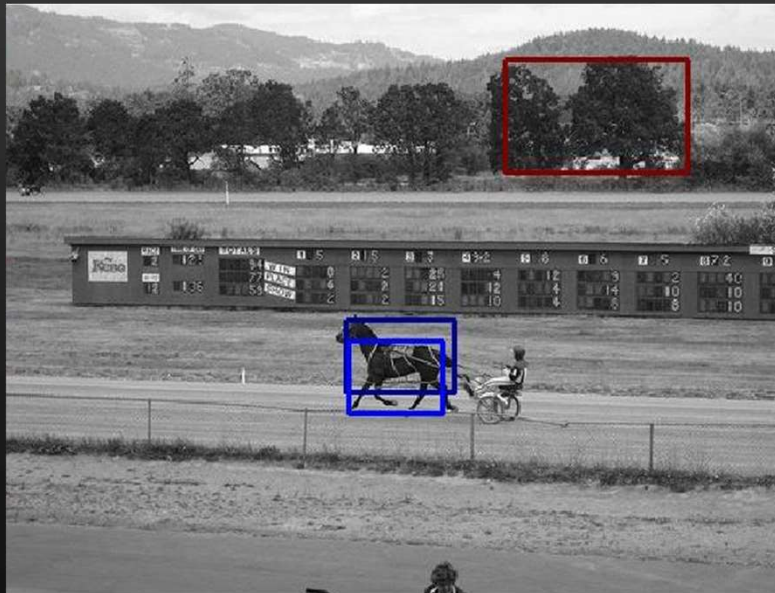
Localization improvement over iterations



After 1 iteration



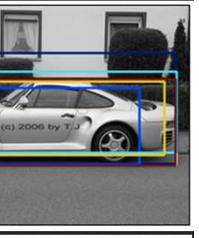
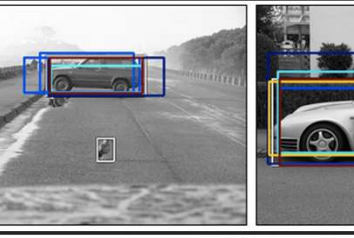
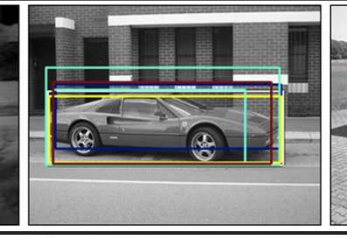
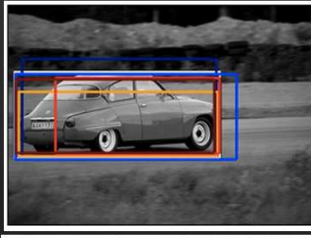
After 3 iterations



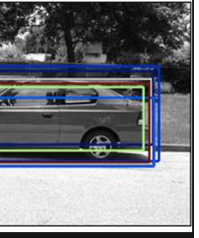
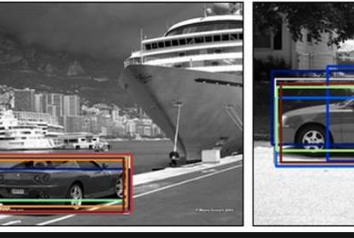
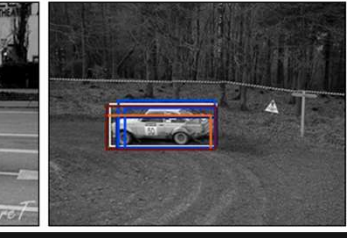
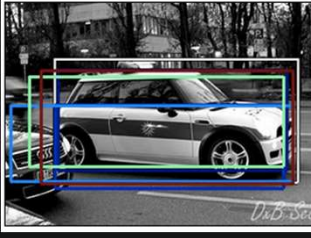
Retrieval improvement over iterations



1st iteration



5th iteration



Pascal'07 results (Cho et al., CVPR'15)

CorLoc - separate classes

Method	Data used	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	tra	tv	Av.
Pandey & Lazebnik [26]	P+N	50.9	56.7	-	10.6	0	56.6	-	-	2.5	-	14.3	-	50.0	53.5	11.2	5.0	-	34.9	33.0	40.6	-
Siva & Xiang [36]	P+A	42.4	46.5	18.2	8.8	2.9	40.9	73.2	44.8	5.4	30.5	19.0	34.0	48.8	65.3	8.2	9.4	16.7	32.3	54.8	5.5	30.4
Siva et al. [34]	P+N	45.8	21.8	30.9	20.4	5.3	37.6	40.8	51.6	7.0	29.8	27.5	41.3	41.8	47.3	24.1	12.2	28.1	32.8	48.7	9.4	30.2
Shi et al. [33]	P+N	67.3	54.4	34.3	17.8	1.3	46.6	60.7	68.9	2.5	32.4	16.2	58.9	51.5	64.6	18.2	3.1	20.9	34.7	63.4	5.9	36.2
Cinbis et al. [6]	P+N	56.6	58.3	28.4	20.7	6.8	54.9	69.1	20.8	9.2	50.5	10.2	29.0	58.0	64.9	36.7	18.7	56.5	13.2	54.9	59.4	38.8
Wang et al. [42]	P+N+A	80.1	63.9	51.5	14.9	21.0	55.7	74.2	43.5	26.2	53.4	16.3	56.7	58.3	69.5	14.1	38.3	58.8	47.2	49.1	60.9	48.5
Joulin et al. [18]	P	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	24.6
Ours	P	50.3	42.8	30.0	18.5	4.0	62.3	64.5	42.5	8.6	49.0	12.2	44.0	64.1	57.2	15.3	9.4	30.9	34.0	61.6	31.5	36.6

CorLoc and CorRet - mixed classes

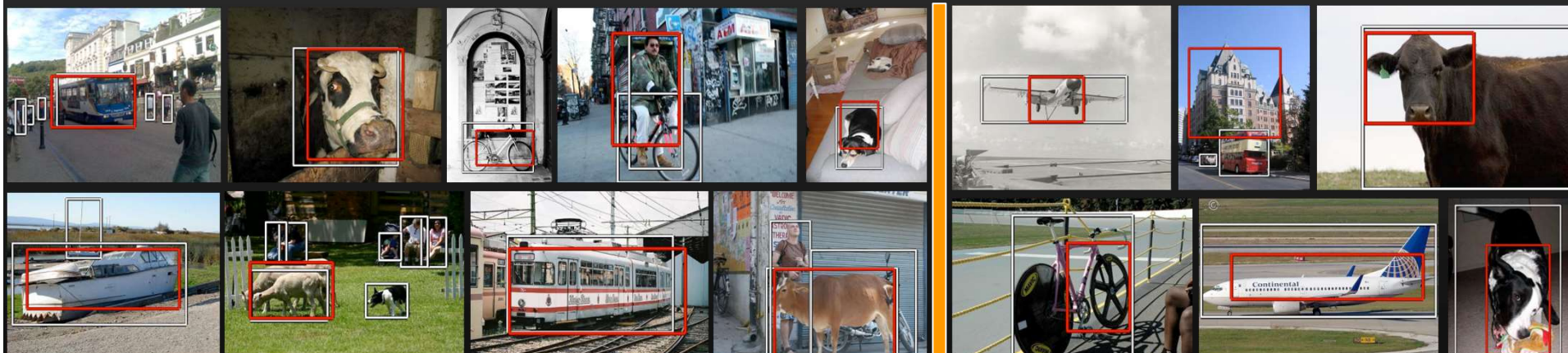
Uses pre-trained CNN features

Evaluation metric	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	tra	tv	Av.	any
CorLoc	40.4	32.8	28.8	22.7	2.8	48.4	58.7	41.0	9.8	32.0	10.2	41.9	51.9	43.3	13.0	10.6	32.4	30.2	52.7	21.8	31.3	37.6
CorRet	51.1	45.3	12.7	12.1	11.4	21.2	61.9	11.6	19.2	9.70	3.9	17.2	29.6	34.0	43.7	10.2	8.1	9.9	23.7	27.3	23.2	36.6

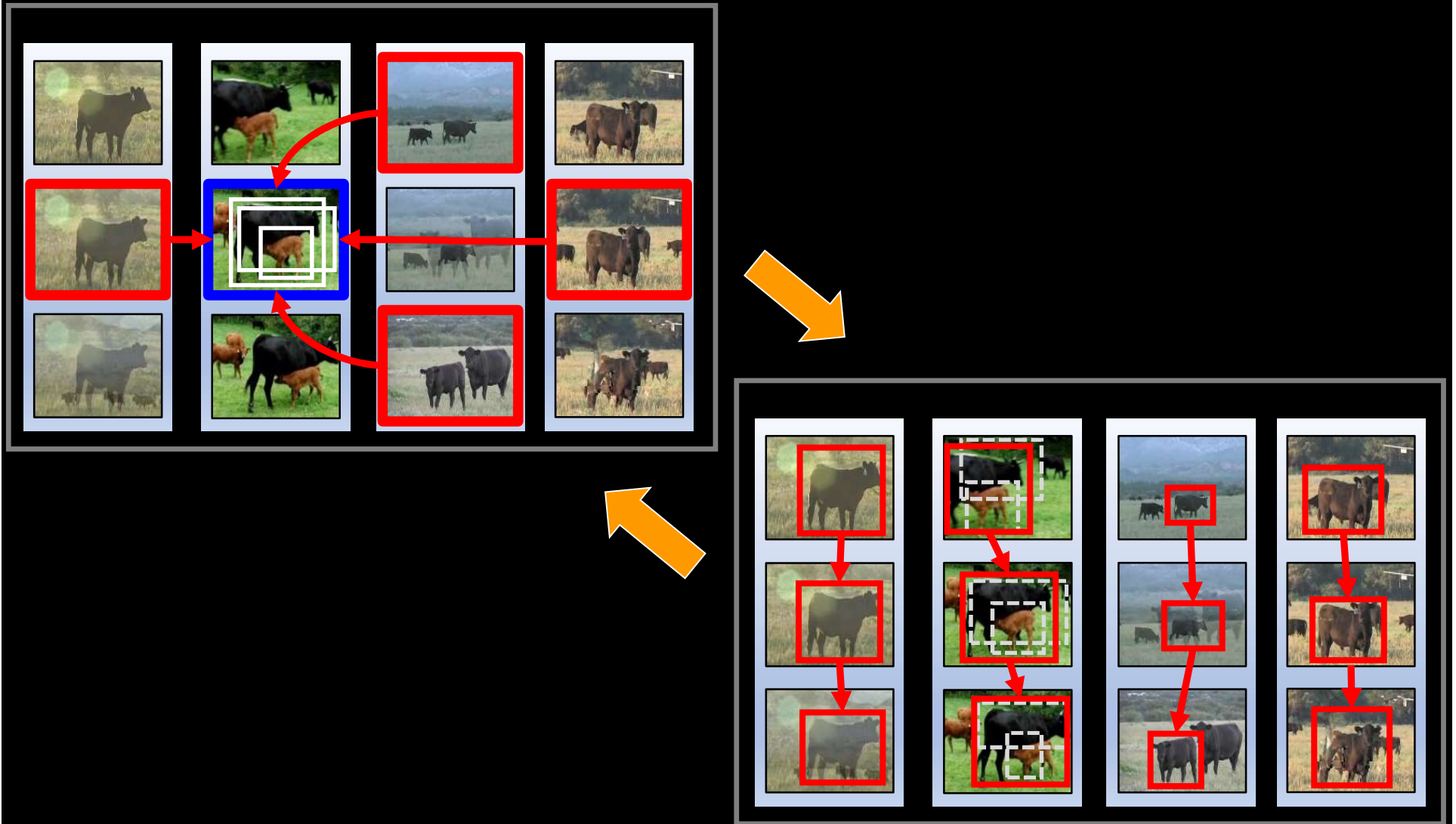
Examples - mixed classes

Successes

Failures



Unsupervised object discovery in multiple videos



(Suha, Cho, Laptev, Ponce, Schmid, 2015)

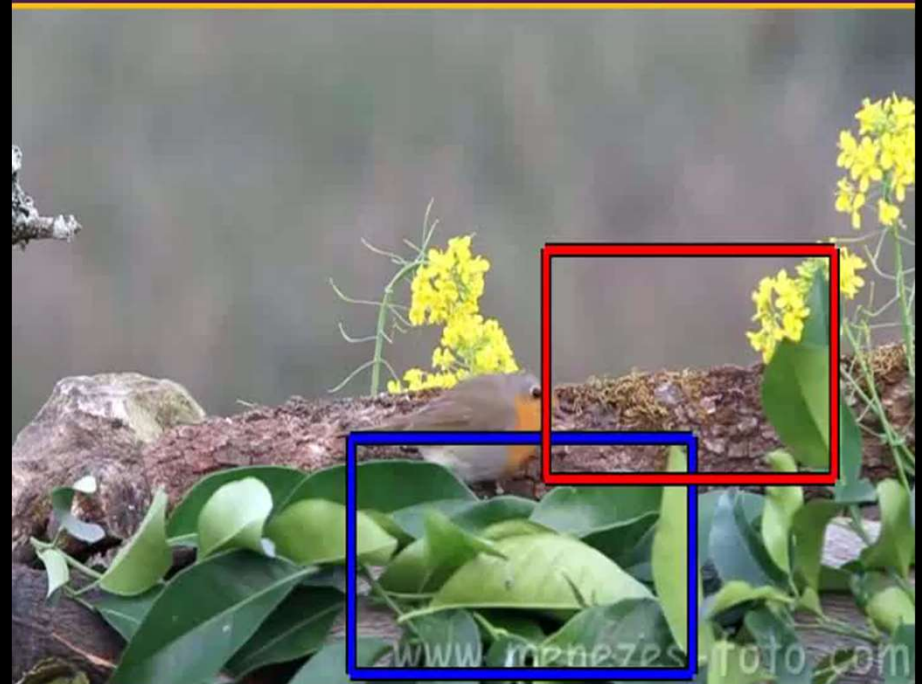
aeroplane-0004-029

- Object colocalization per class
- Unsupervised object discovery



bird-0004-016

- Object colocalization per class
- Unsupervised object discovery



(Suha, Cho, Laptev, Ponce, Schmid, 2015)

45 clips selected manually from the Bourne trilogy

Discovering *Cars* from movie clips



About 90mn (excluding preprocessing) on a 12-core 1.2GHz machine

44 clips selected manually from two movies

Discovering *Animals* from movie clips



About 90mn (excluding preprocessing) on a 12-core 1.2GHz machine

Going further

Unsupervised object discovery as optimization

images i, j are linked

boxes k, l in images i, j are active

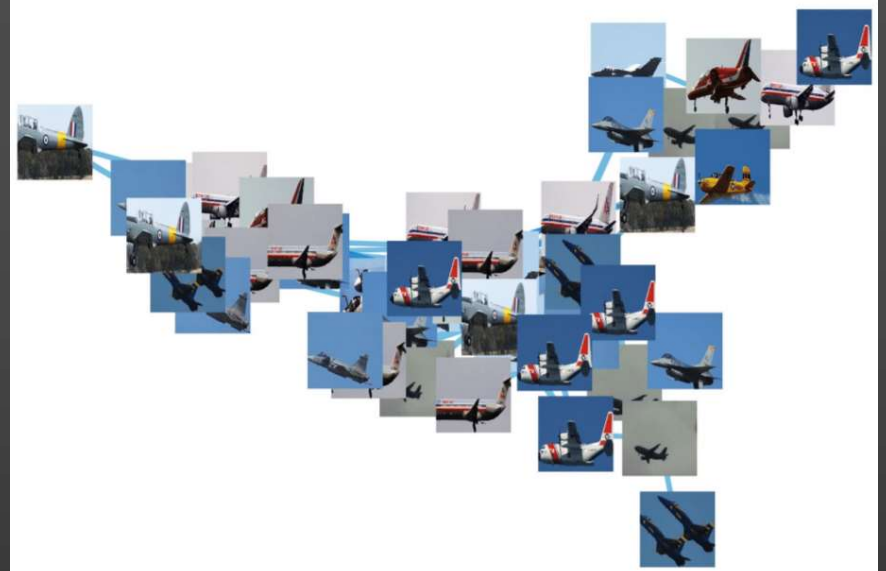
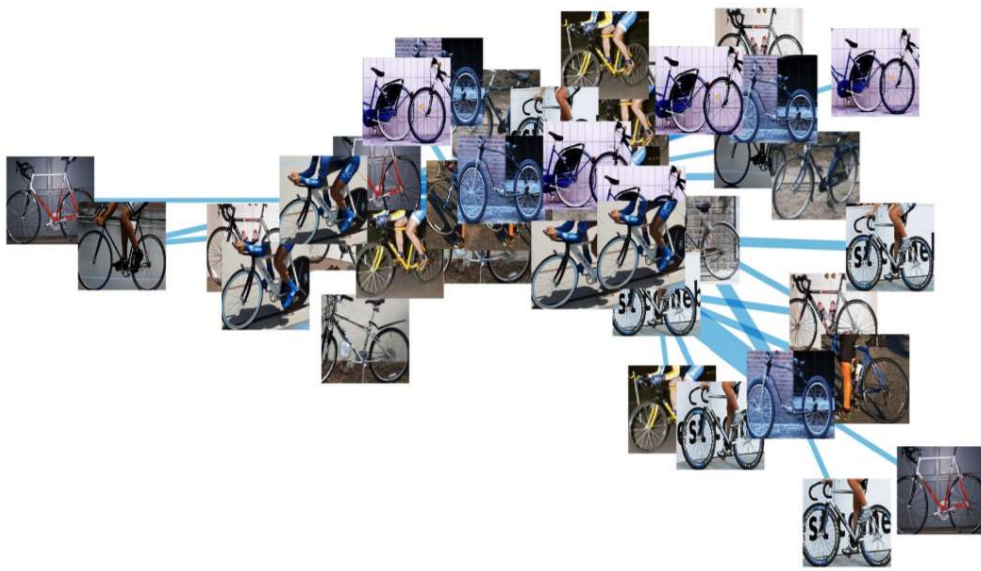
Maximize
$$\sum_{1 \leq i < j \leq n} e_{ij} \sum_{k, l=1}^p S_{ij}^{kl} x_i^k x_j^l = \sum_{1 \leq i < j \leq n} e_{ij} x_i^T S_{ij} x_j$$

similarity

subject to

$$\forall i \in 1 \dots n, \sum_{k=1}^p x_i^k \leq \nu$$
$$\forall i \in 1 \dots n, \sum_{j=1}^n e_{ij} \leq \tau$$

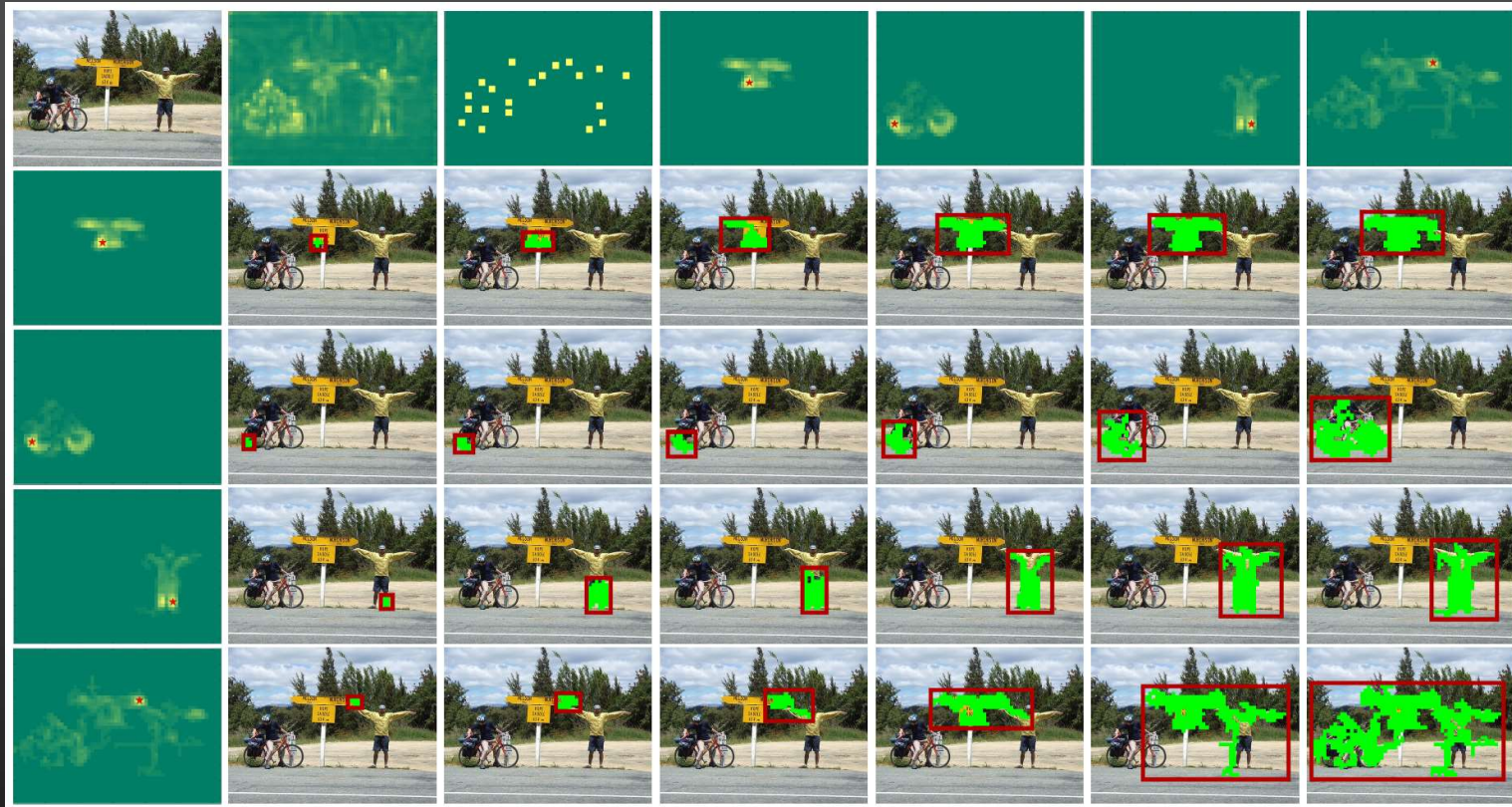
(Vo, Bach, Cho, Han, LeCun, Perez, Ponce, CVPR'19)



w/o CO: greedy combinatorial search
 w CO: use gradient ascent

Method	OD	VOC_6x2	VOC_all
Cho <i>et al.</i>	-	-	37.6
Cho <i>et al.</i> , our execution	82.2	55.9	37.5
w/o CO	83.0 ± 0.4	60.2 ± 0.4	39.8 ± 0.2
w CO	80.8 ± 0.5	59.3 ± 0.4	38.5 ± 0.2

Unsupervised region proposals and large-scale object discovery using features trained on an auxiliary task



(Vo, Perez, Ponce, 2019)

Insight: sum of feature map values provide good saliency maps (Wei et al., 2017), and thus a good basis for region proposals

Insight 2: Use two interpretations of the graph:

- Proxy for the true structure. Run algorithm on small subgroups of images with $v=50$ to find promising proposals
- True structure. Run the algorithm with the selected proposals and $v=5$ on the whole image collection

Small-scale CorLoc results

Method	Features	OD	VOC_6x2	VOC_all
Cho <i>et al.</i>	WHO	82.2	55.9	37.6
Vo <i>et al.</i> RP	WHO	<u>82.3 ± 0.3</u>	62.5 ± 0.6	40.7 ± 0.2
Wei <i>et al.</i> [33]	VGG16	75.8	57.9	39.8
Wei <i>et al.</i> [34]	VGG16	73.5	<u>66.2</u>	<u>41.9</u>
Ours	VGG16	87.5 ± 0.3	70.9 ± 0.3	48.6 ± 0.1

Proposal comparison

Region proposals	OD	VOC_6x2	VOC_all
Edgeboxes [37]	81.4 ± 0.3	55.2 ± 0.3	32.6 ± 0.1
Selective search [30]	81.3 ± 0.3	57.8 ± 0.2	33.0 ± 0.1
Randomized Prim [22]	<u>82.5 ± 0.1</u>	<u>70.6 ± 0.4</u>	<u>44.5 ± 0.1</u>
Ours	87.5 ± 0.3	70.9 ± 0.3	48.6 ± 0.1

Large-scale CorLoc results

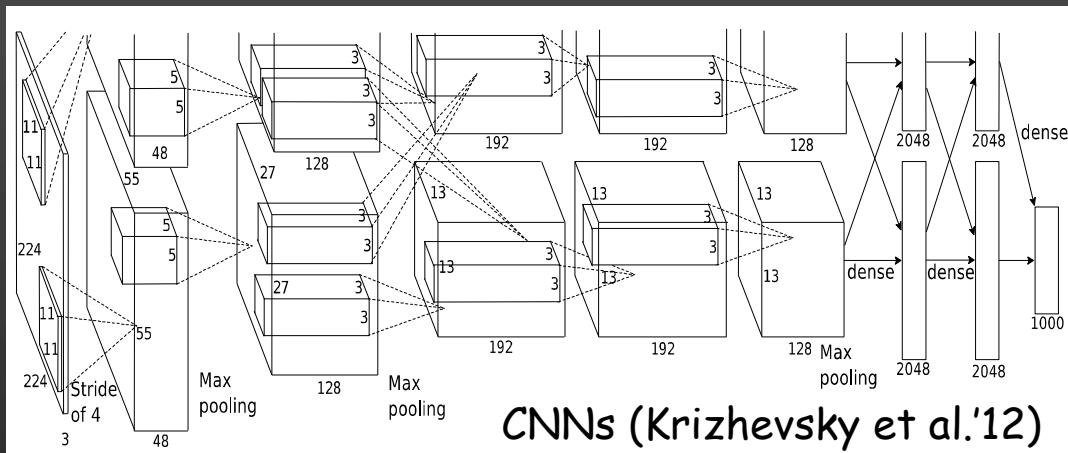
Method	VOC_all	VOC12	COCO_20k
Wei <i>et al.</i> [33]	43.4	46.2	38.6
Wei <i>et al.</i> [34]	43.4	46.3	40.5
Baseline 1	43.3 ± 0.2	40.1 ± 0.1	45.0 ± 0.1
Baseline 2	48.6 ± 0.1	49.3 ± 0.1	-
Ours	46.5 ± 0.1	46.2 ± 0.1	47.3 ± 0.1

Large-scale CorRet results

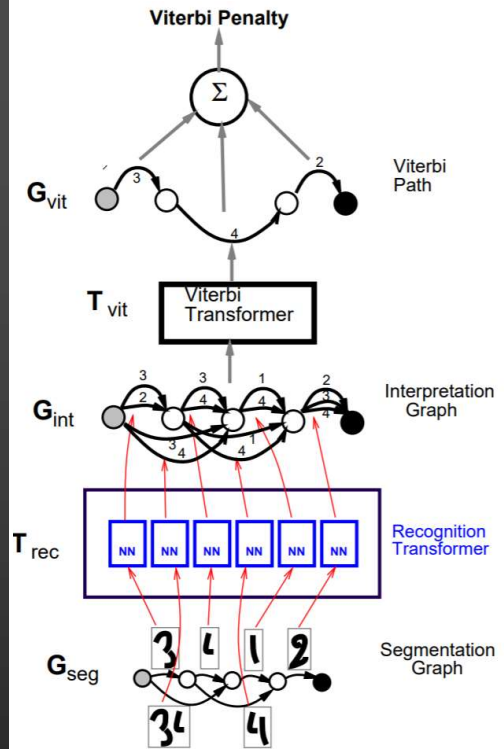
Dataset	VOC_all	VOC12	COCO_20k
Baseline	50.7	57.5	36.8
Ours	61.3 ± 0.0	64.7 ± 0.0	40.1 ± 0.0

Beyond block diagrams

Deep learning
(LeCun et al.'98)

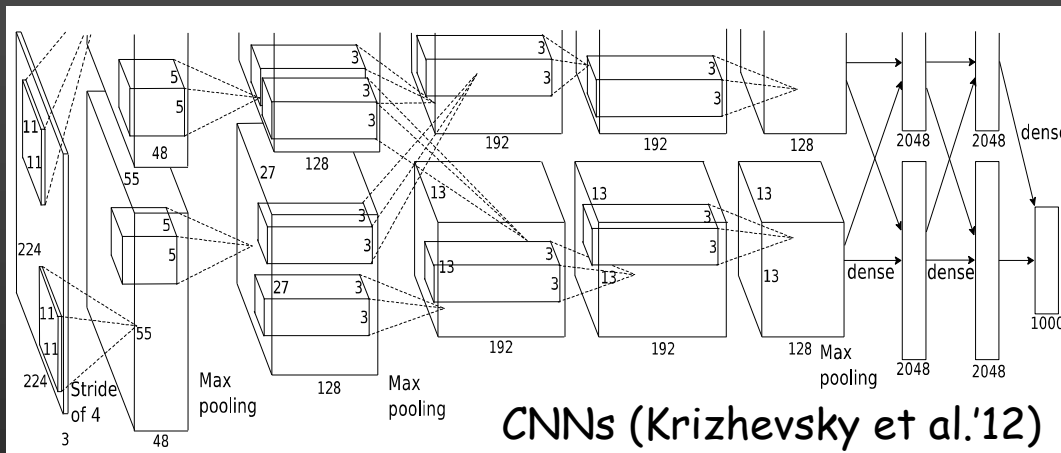


Graph transformer networks



Beyond block diagrams and pattern recognition

Deep learning
(LeCun et al.'98)



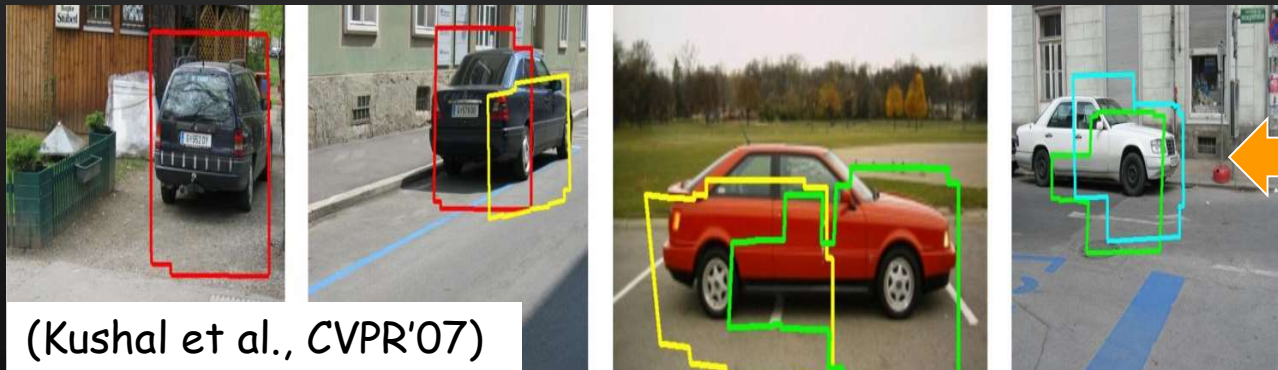
Tunable algorithms
(Eboli et al.'19, Lecouat et al.'19)

```

function  $x = \text{LCHQS}(y, k_0, K, \theta, \nu)$ 
 $x = y; \mu = 0;$ 
for  $t = 0 : T - 1$  do
     $\hat{z} = [y; \sqrt{\mu} \varphi_{\lambda/\mu}^{\theta}(K \star x)];$ 
     $\hat{K} = [k_0; \sqrt{\mu} K];$ 
     $C = \text{argmin}_C \|\delta - C \star \hat{K}\|_F^2 + \rho \sum_{i=0}^n \|c_i\|_F^2;$ 
     $x = \text{CPCR}(\hat{K}, \hat{z}, \psi^{\nu}(C), x);$ 
     $\mu = \mu + \delta_t;$ 
end for
end function
    
```

```

function  $x = \text{CPCR}(A, b, C, x_0)$ 
 $x = x_0;$ 
for  $u = 0 : U - 1$  do
     $x = x - C \star (A \star x - b);$ 
end for
end function
    
```



(Kushal et al., CVPR'07)

← Didn't work so well
but the problem is
important!