

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

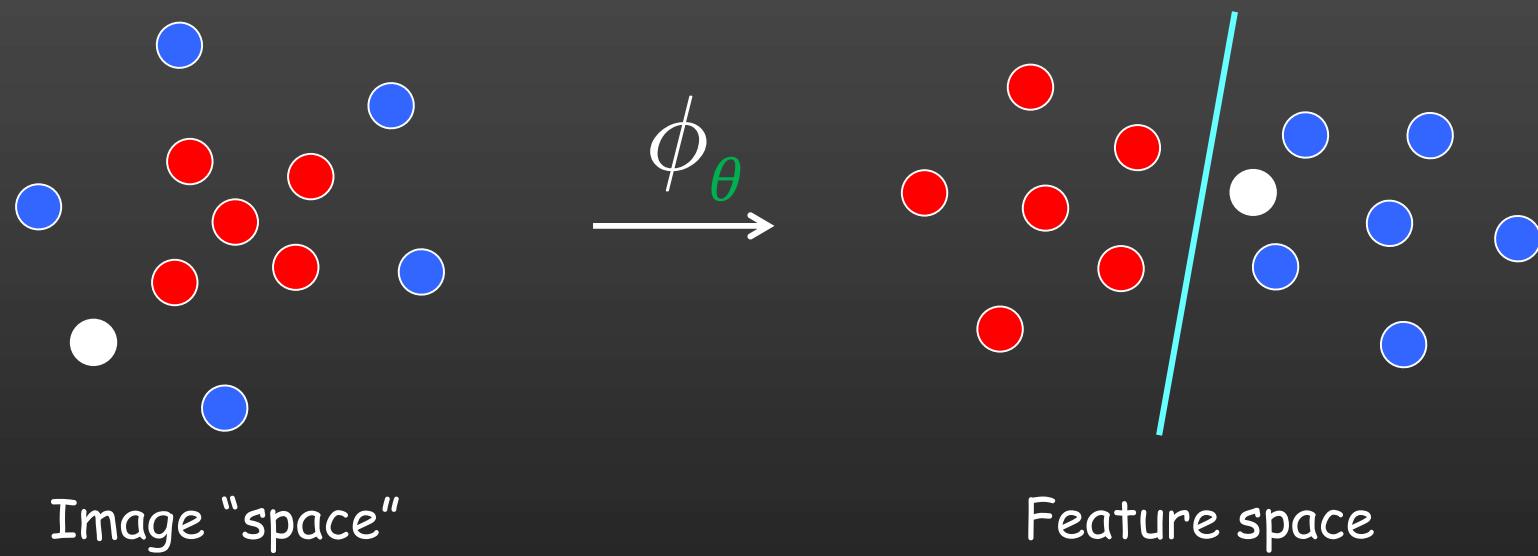


Labelled training examples



Test image

Image categorization as supervised classification



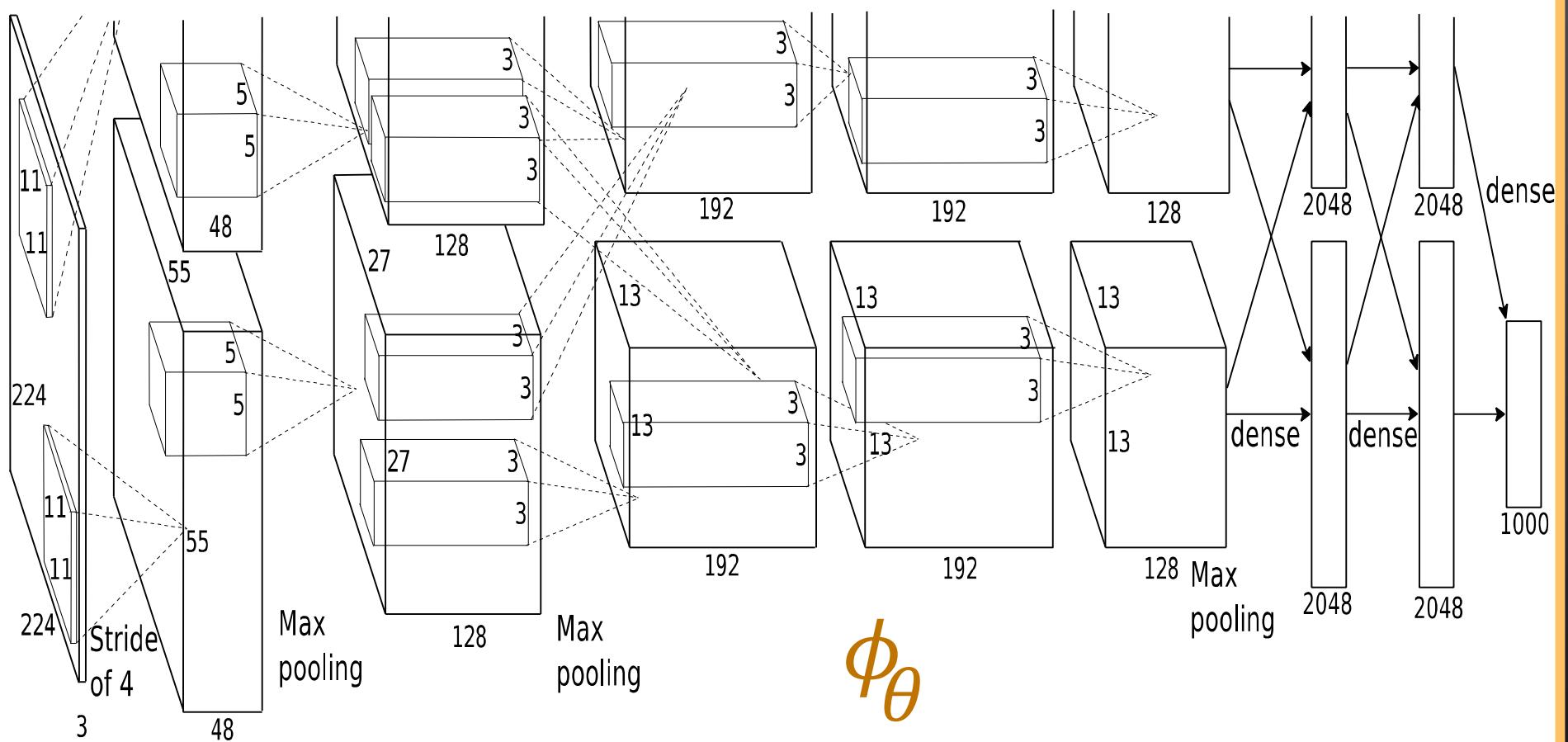
$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_n \ell(z_n, f(\phi_{\theta}(x_n))) + \Omega(f)$$

Training datum

Label

Prediction function

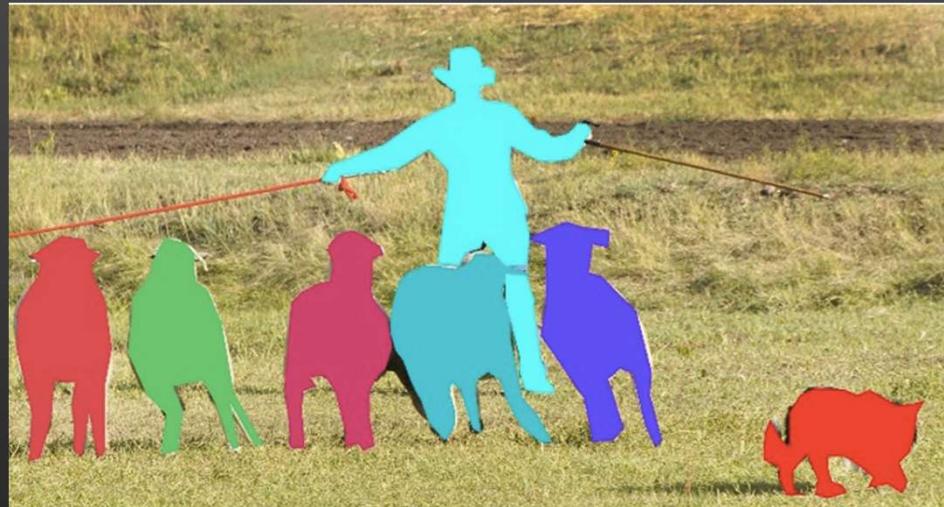
Convolutional neural networks (LeCun et al., 1998)



(Krizhevsky et al.'12)

Supervision: Where do the labels come from?

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



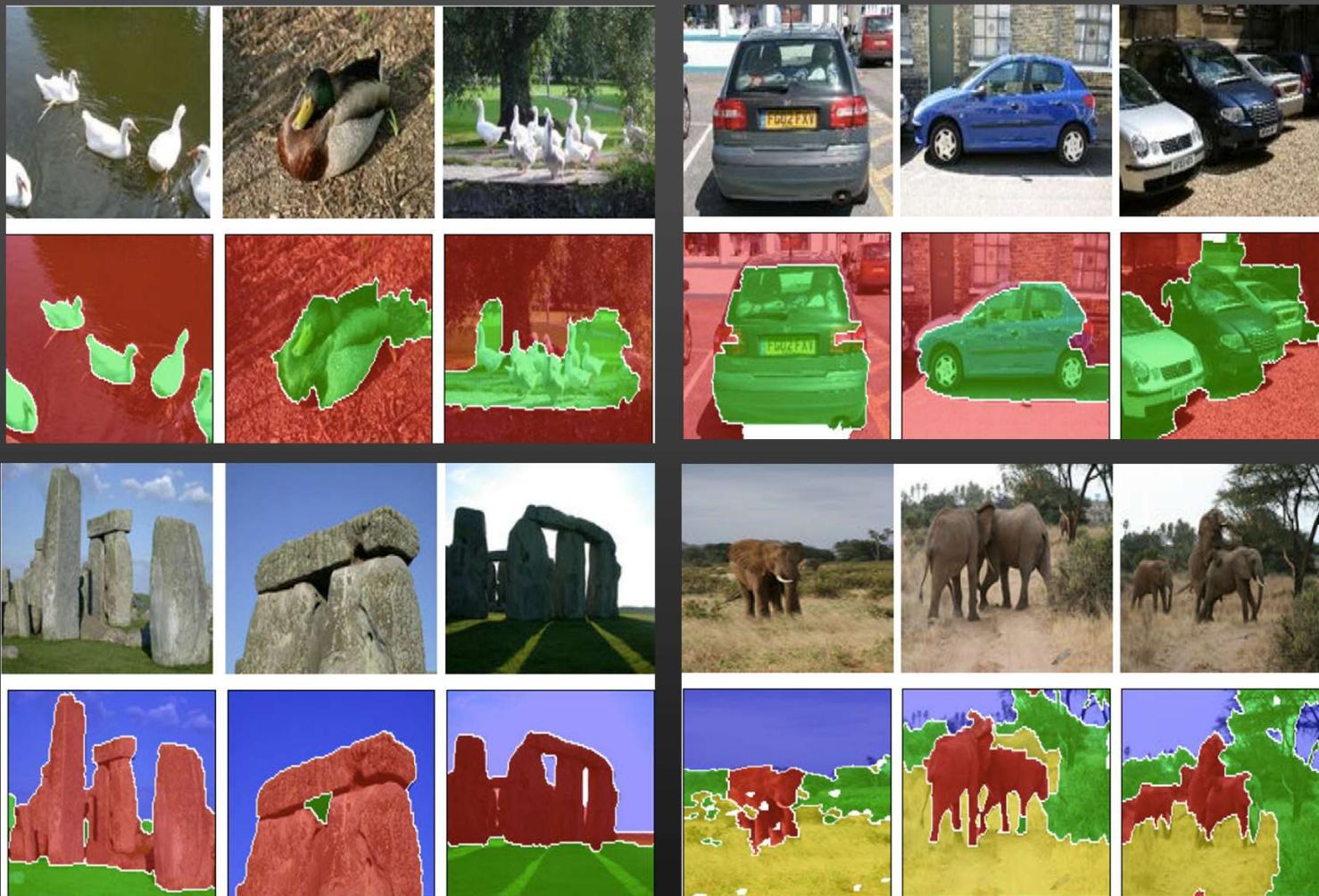
- Example: MS COCO (Lin et al., 2015)
 - 328K images of 91 object categories
 - 2.5M labelled instances

(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 ≈ “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

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

Optimize over labels too

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

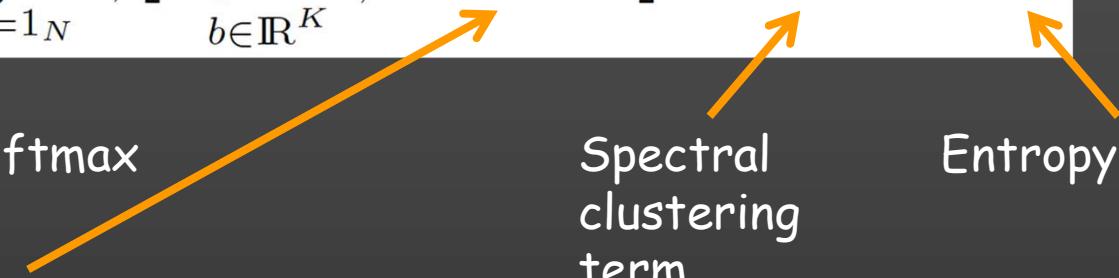
Square loss

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

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

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

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



 Softmax → Discriminative clustering term

 Spectral clustering term → Spectral clustering term

 Entropy term → Entropy 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)

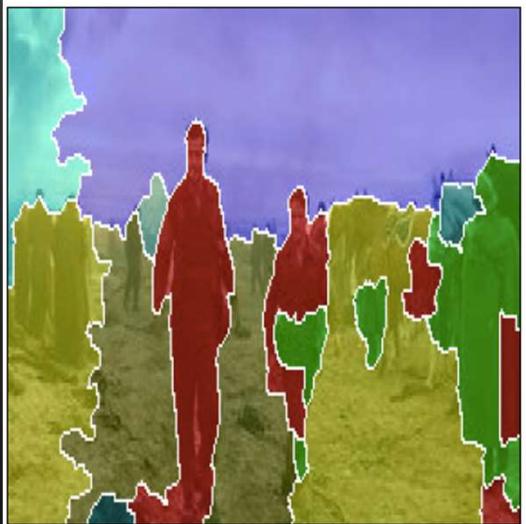
$$\min_{\substack{y \in \{0,1\}^{N \times K}, \\ y \mathbf{1}_K = \mathbf{1}_N}} \left[\min_{\substack{A \in \mathbb{R}^{d \times K}, \\ b \in \mathbb{R}^K}} E_U(y, A, b) \right] + E_B(y) - H(y)$$

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



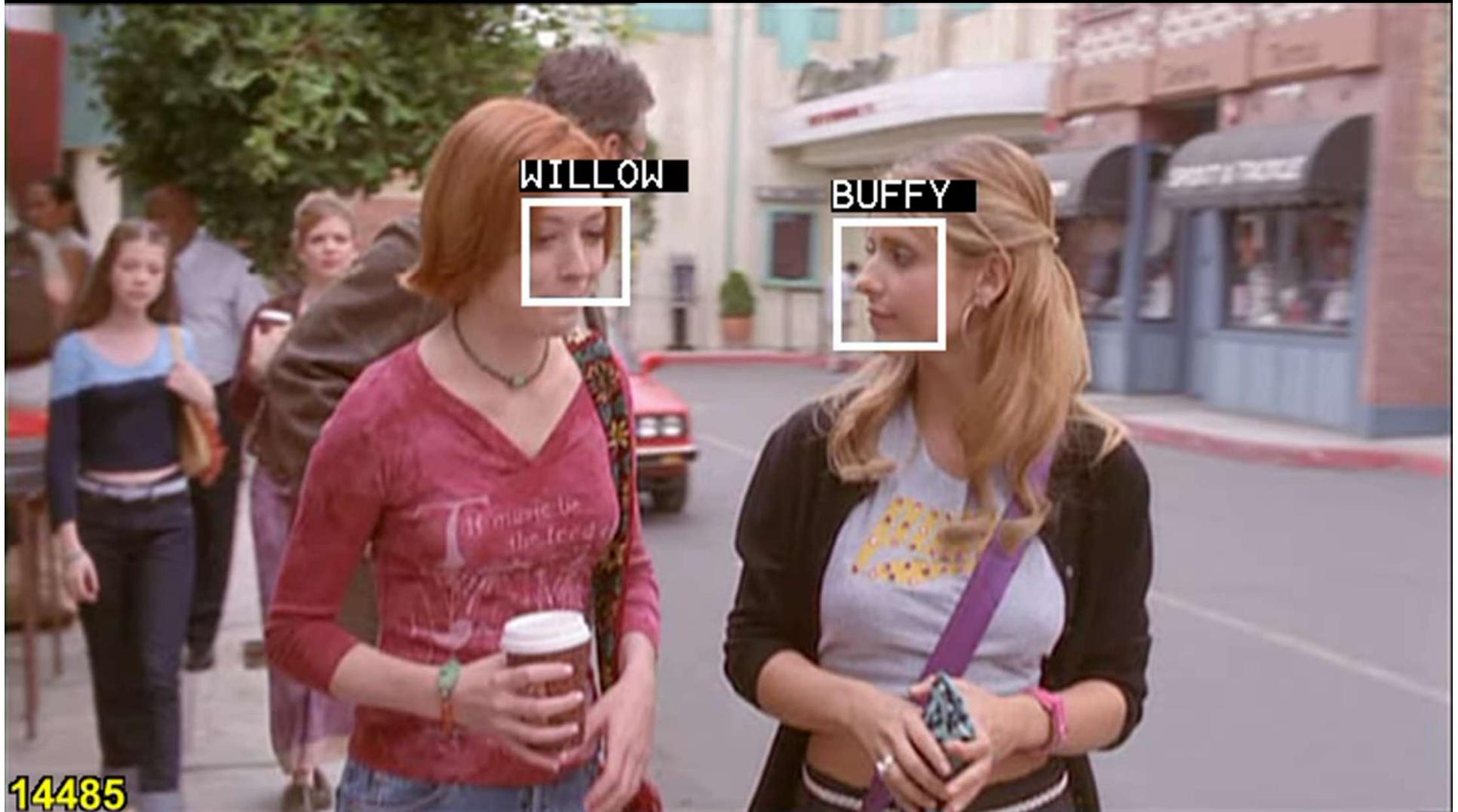
(Joulin et al., CVPR'12)

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...

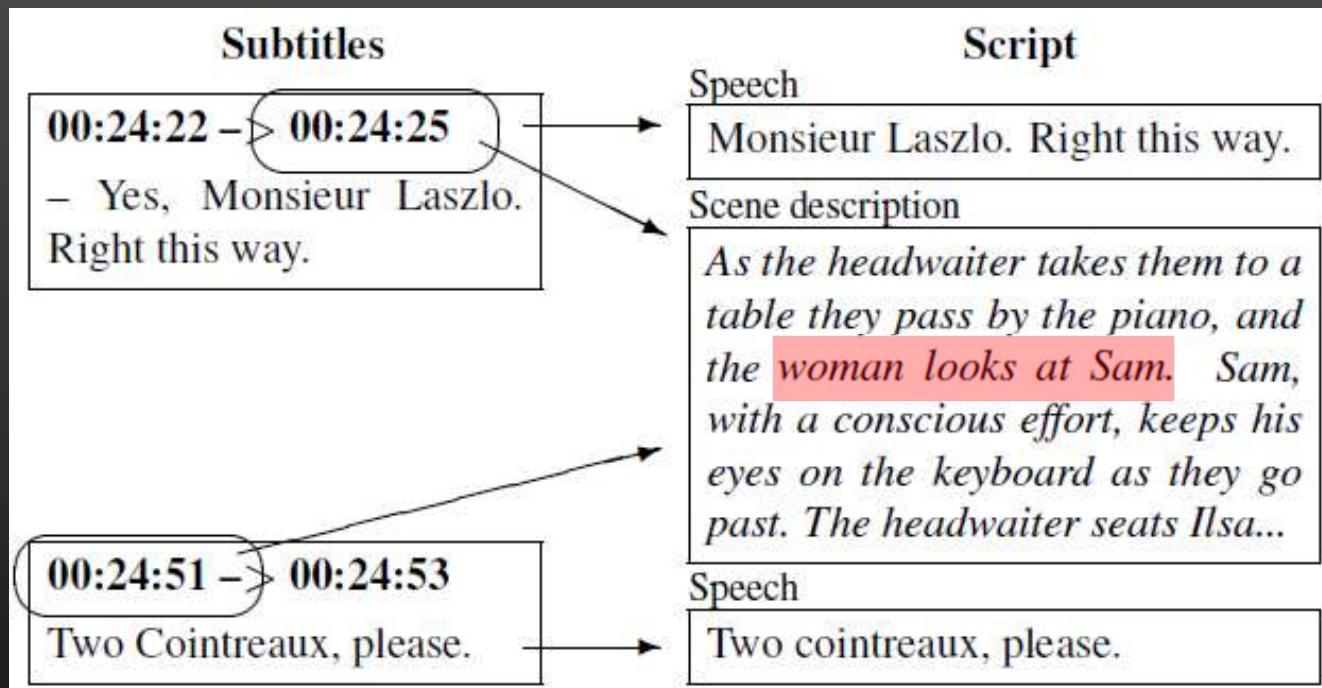


Videos (often) come with their own metadata!

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



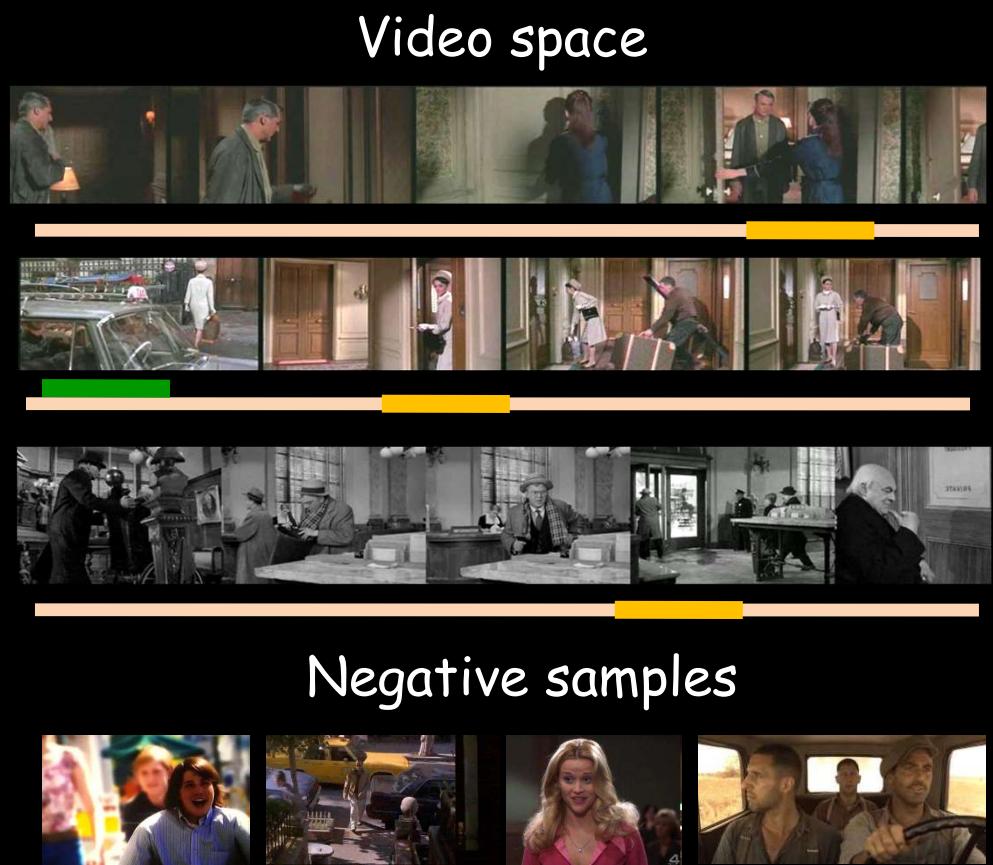
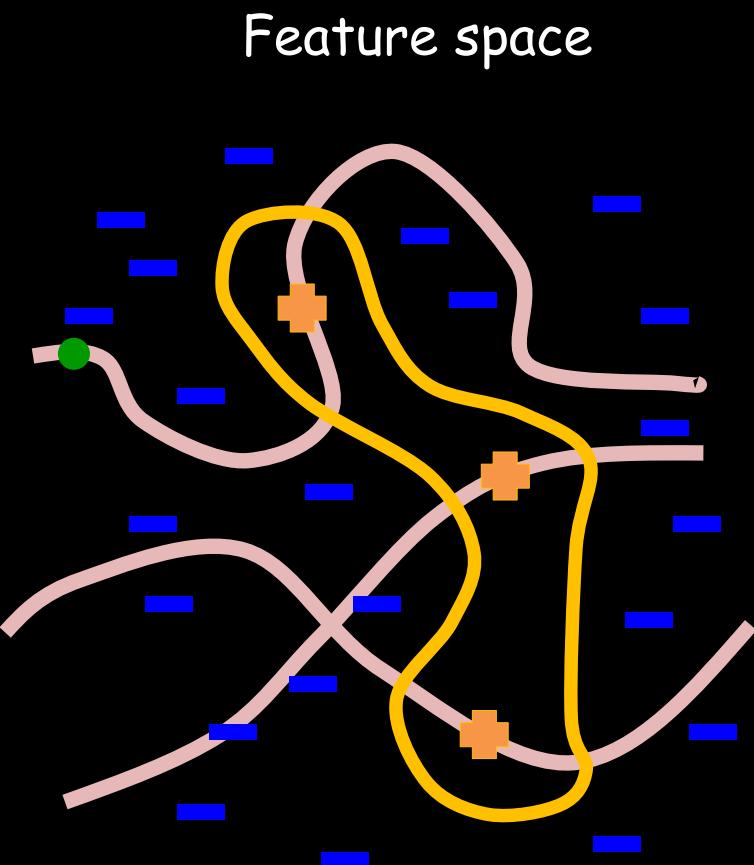
Output: temporal action clusters

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



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

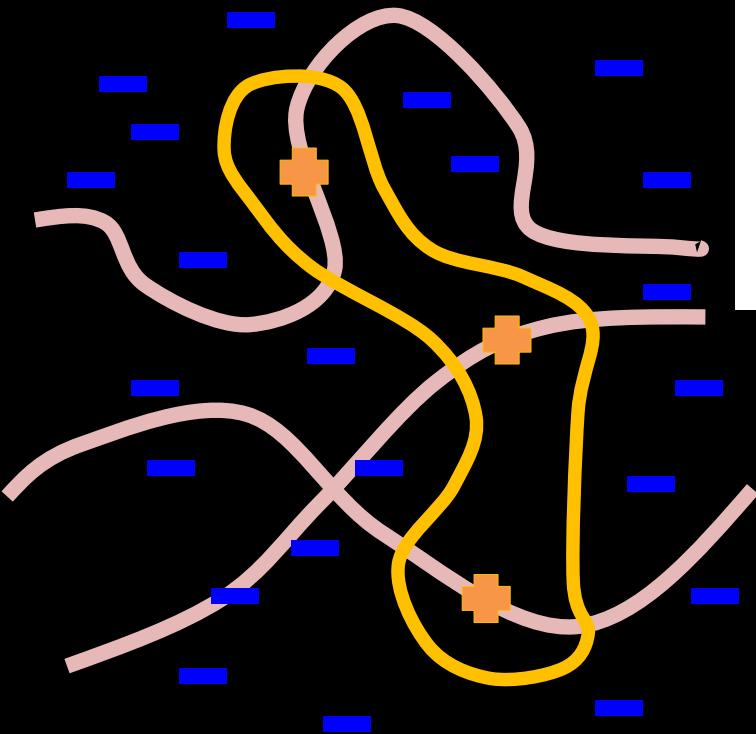
Temporal localization as classification



A latent SVM model for temporal localization

(Felzenszwalb, McAllester, Ramanan, 2008)

Feature space



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

Optimization: Block-coordinate descent

1. Exhaustive search for f

2. SVM training for w, b

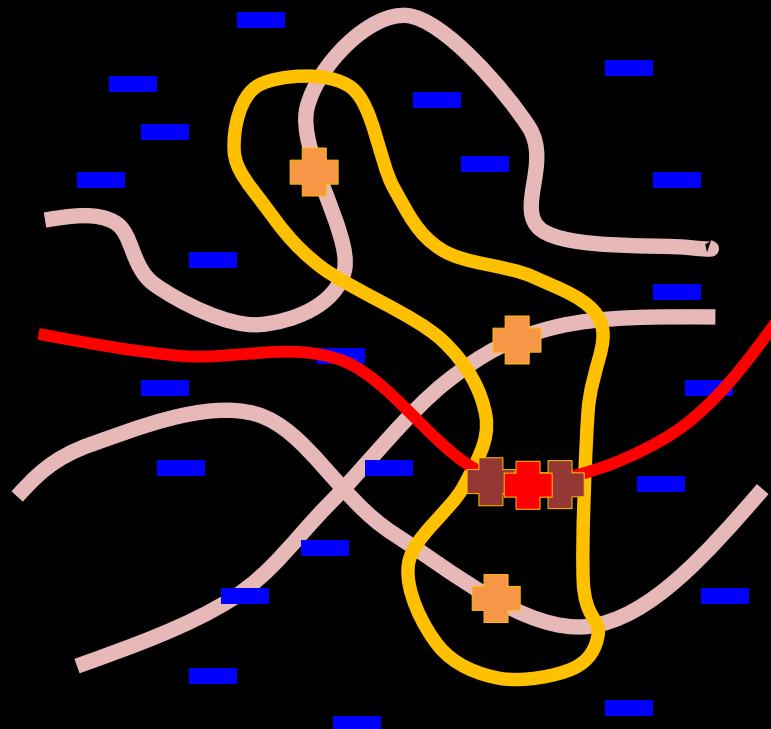
This is an instance of discriminative clustering

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

Video

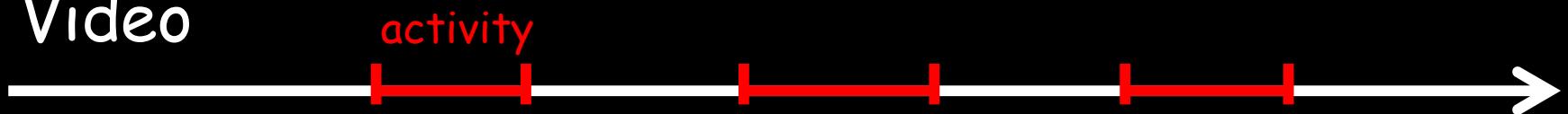
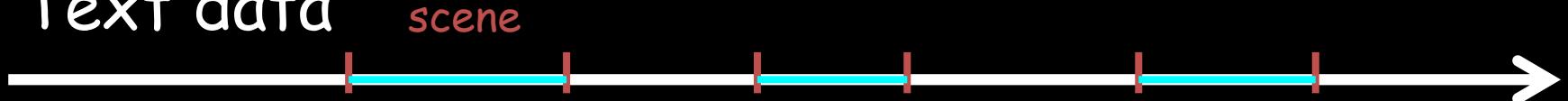


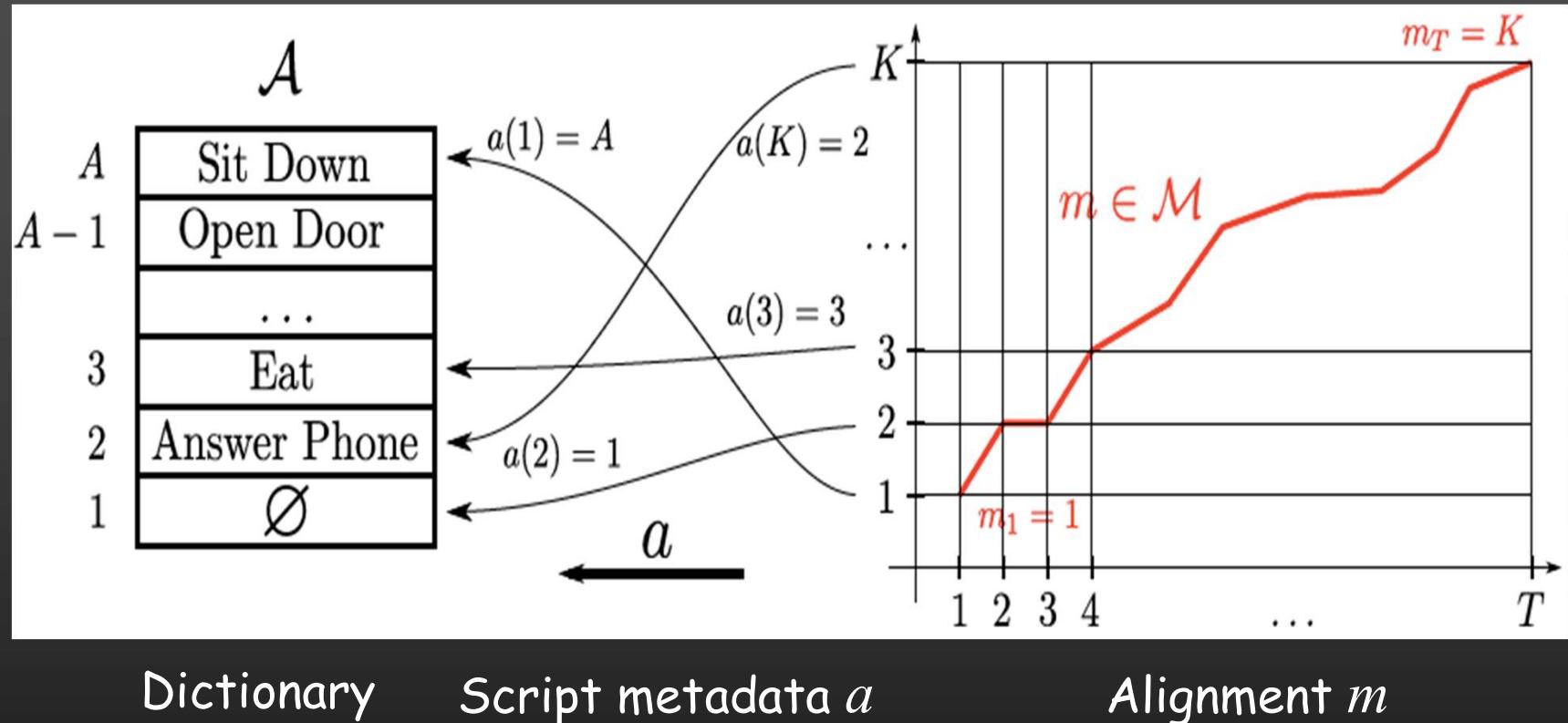
Image data



Text data



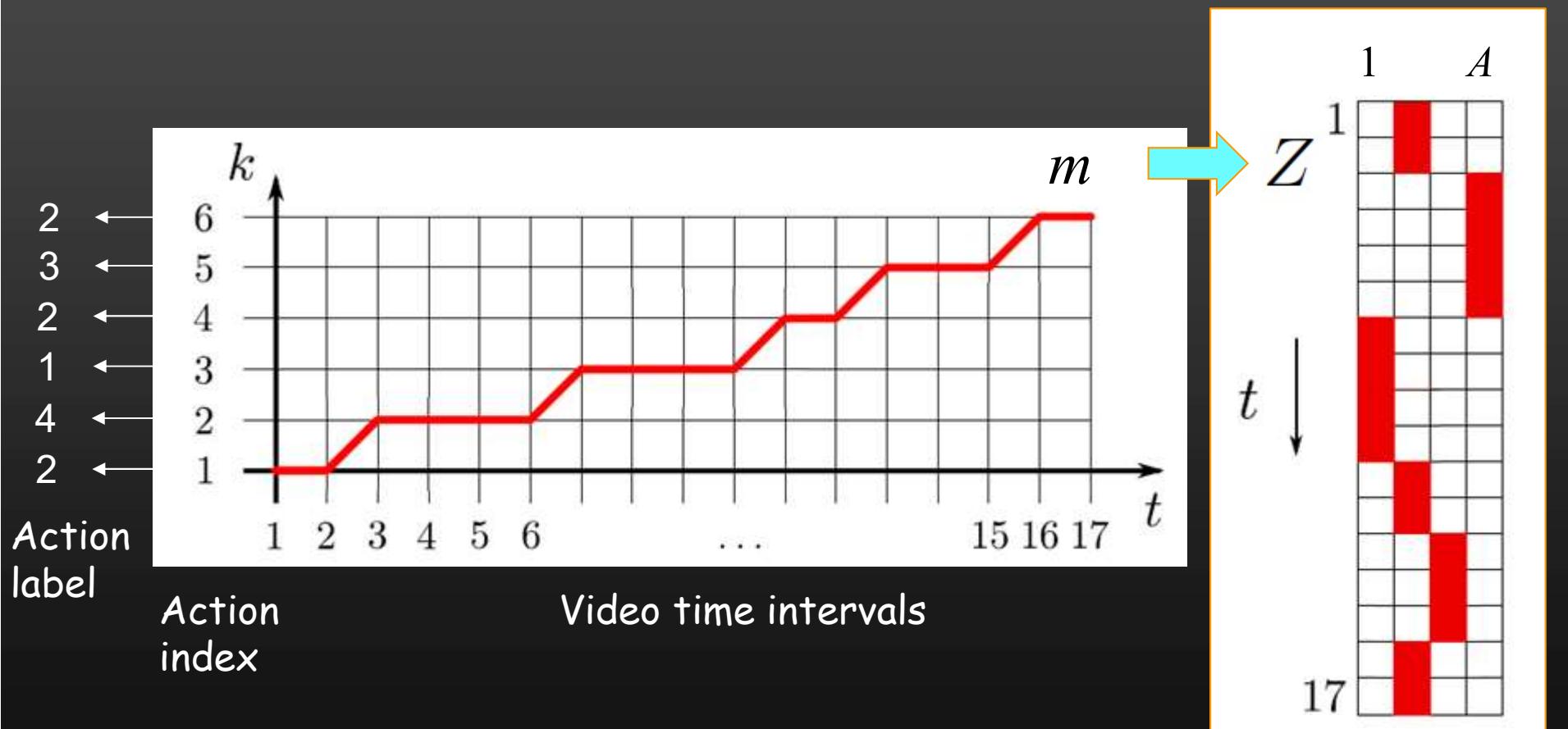
Action labeling under ordering constraints (Bojanowski et al., ECCV'14)



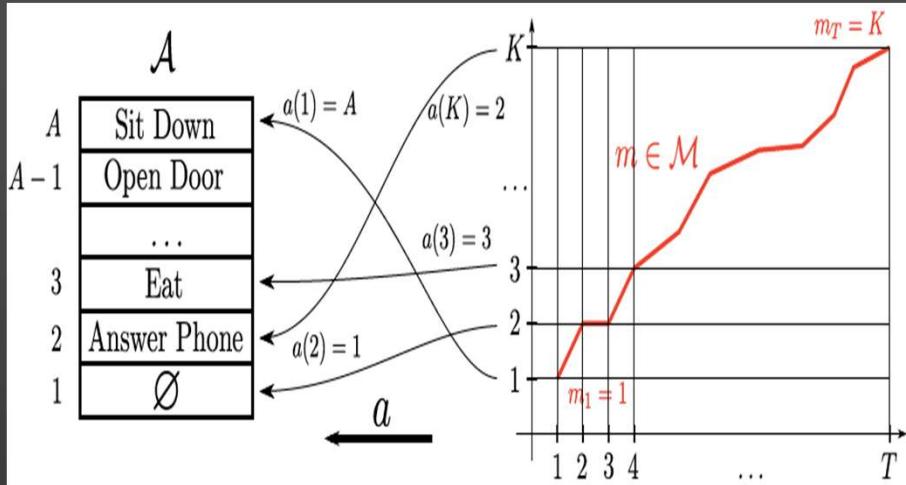
$$\min_{f \in \mathcal{F}} \left[\sum_{n=1}^N \min_{m \in \mathcal{M}} \frac{1}{T} \sum_{t=1}^T \ell(a_n(m_t), f(x_n(t))) \right] + \lambda \Omega(f)$$

Changing the representation

a, m → Z



Action labeling under ordering constraints



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

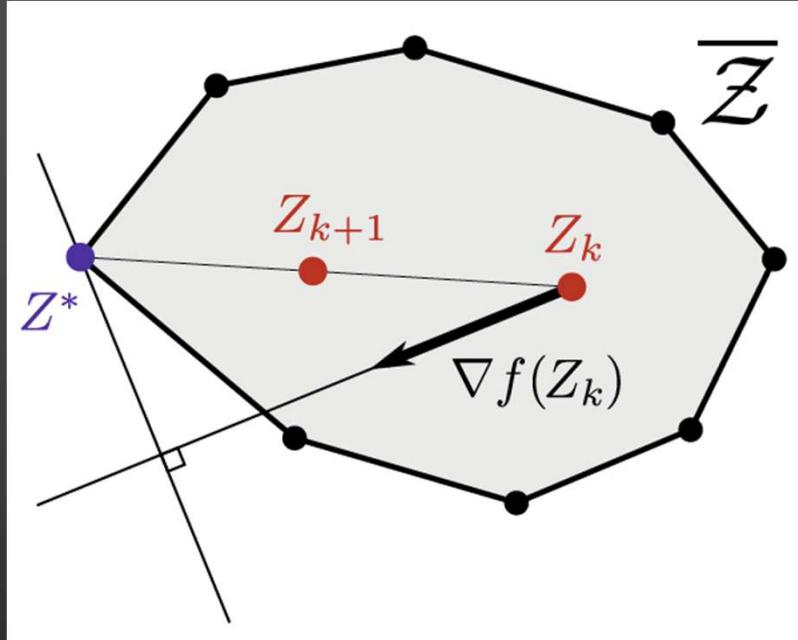
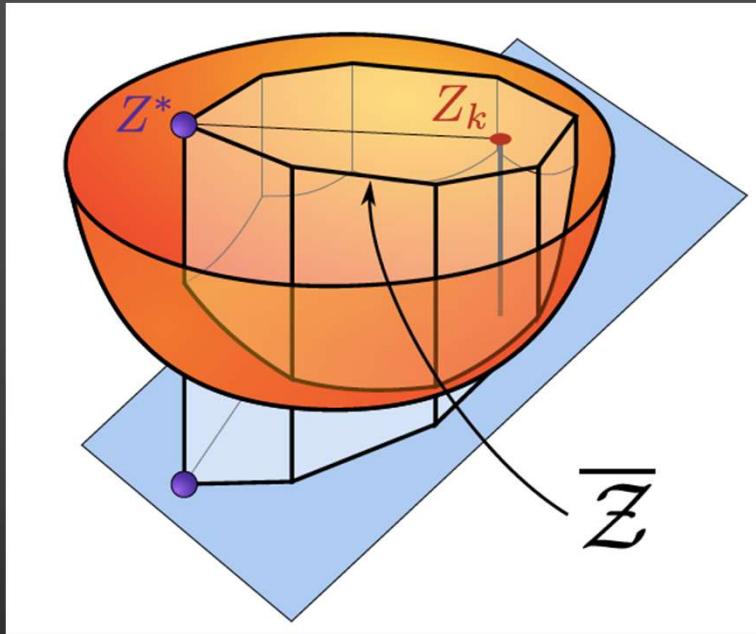


$$\min_{f \in \mathcal{F}, Z \in \mathcal{Z}} \frac{1}{T} \sum_{t=1}^T \ell(Z_t, f(x_t)) + \lambda \Omega(f) = \frac{1}{T} \|Z - XW - b\|_F^2 + \frac{\lambda}{2} \|W\|_F^2$$

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

- Minimize a convex quadratic function over a large discrete domain \mathbb{Z}
- Relaxed problem: minimize instead over $\overline{\mathbb{Z}} = \text{conv}(\mathbb{Z})$, then round up
- Difficulty: \mathbb{Z} (and thus $\overline{\mathbb{Z}}$) 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 \bar{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 Z and thus \bar{Z}
- DP can also be used for rounding

Temporal action localization

Clip number 0101

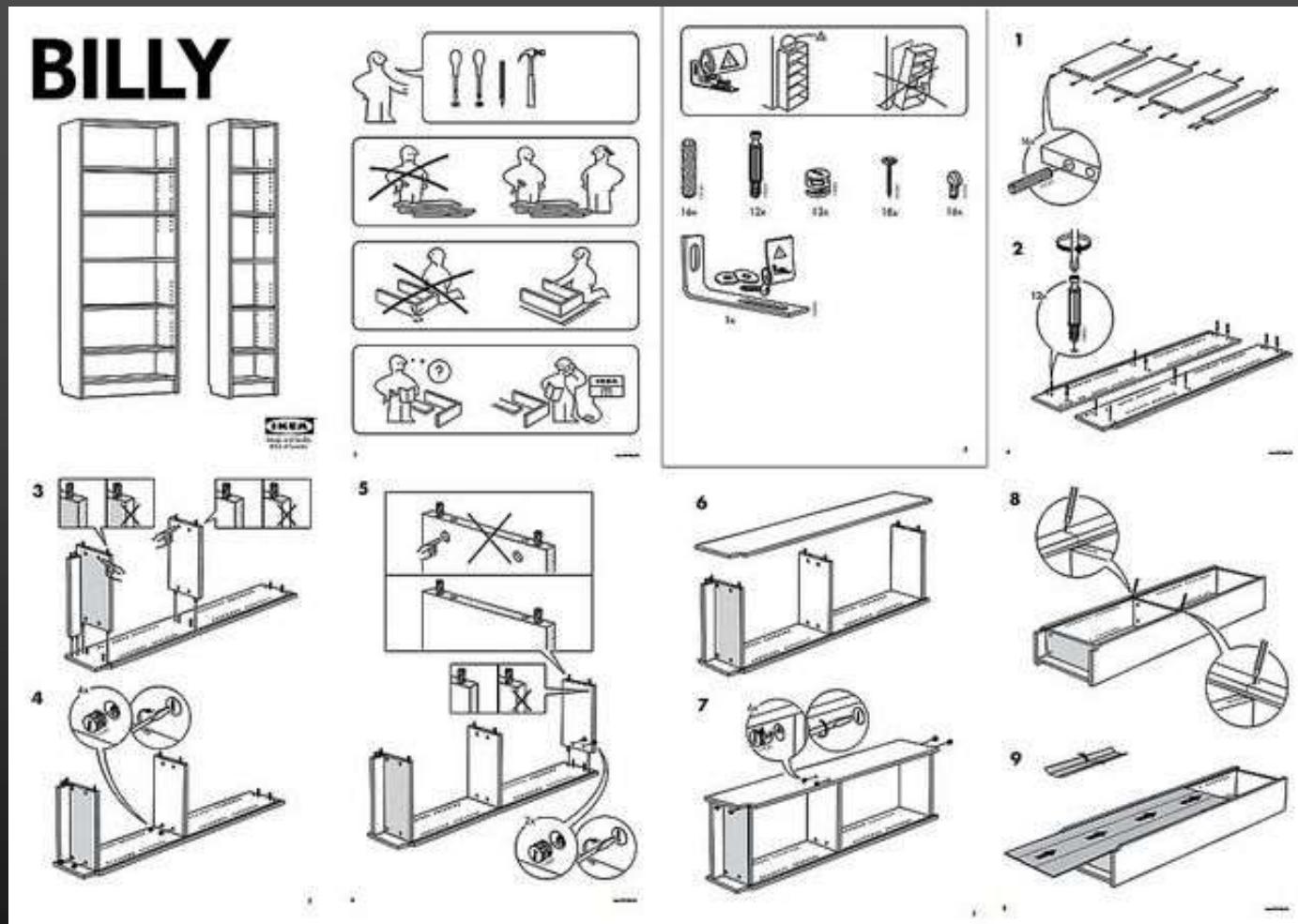
(Bojanowski et al., ECCV'14)

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**.



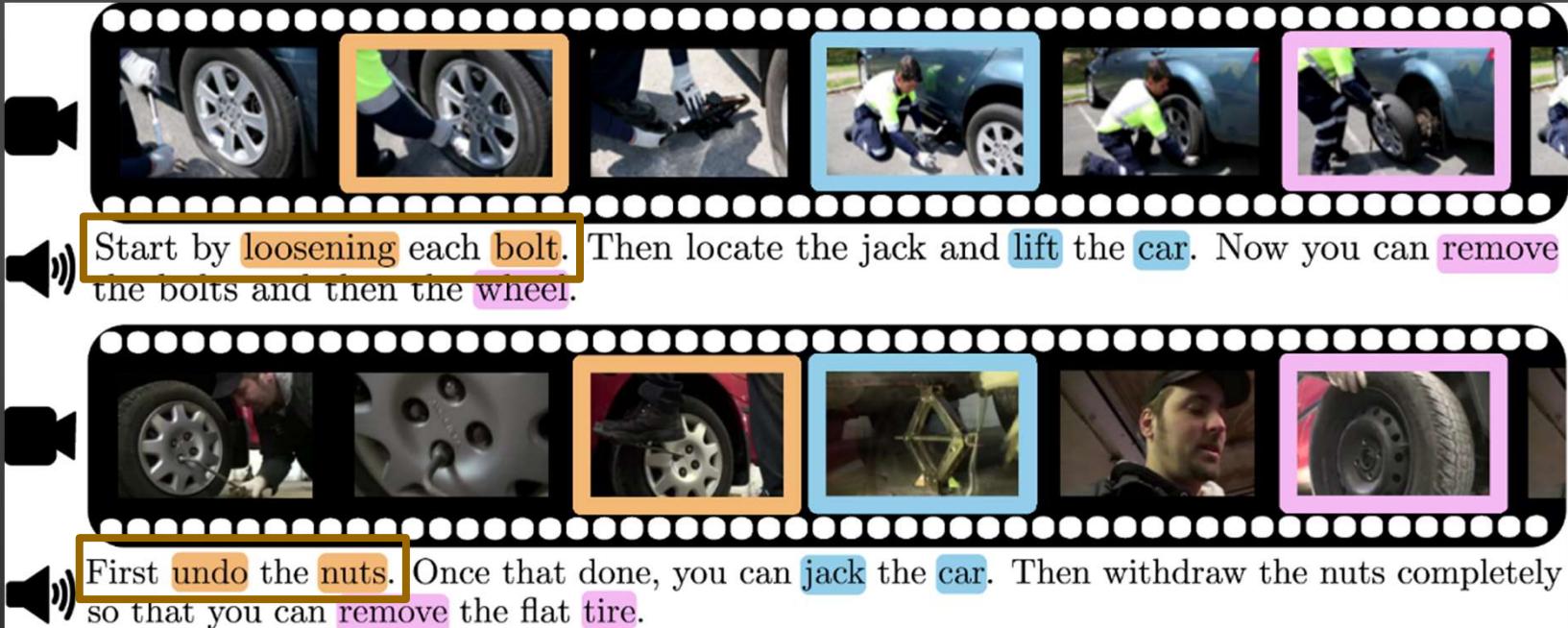
First **undo** the **nuts**. Once that done, you can **jack** the **car**. Then withdraw the nuts completely so that you can **remove** the flat **tire**.

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



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

Video 1

jack car
remove wheel

Video 2

loosen nut
raise car
remove tire
lower jack

Video 3

loosen nut
jack car
unscrew nut
withdraw tire

Video 4

undo bolt
lift car
lower car

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

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
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	∅
∅	lower jack	∅	lower car

Text alignment in multiple sequences

We seek to minimize the sum of pairwise costs:

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

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

Sum over all pairs Sum over template lines Alignment cost
Mapping of sequence n to a common template

The diagram illustrates the formula for text alignment in multiple sequences. The formula is:

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

Annotations explain the components:

- "Sum over all pairs" points to the first summation term $\sum_{(n,m)}$.
- "Sum over template lines" points to the second summation term $\sum_{l=1}^L$.
- "Alignment cost" points to the function call $c(\phi(d^n)_l, \phi(d^m)_l)$.
- "Mapping of sequence n to a common template" is a descriptive label for the argument $\phi(d^n)_l$.

[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

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

Text alignment in multiple sequences

Video 1	Video 2	Video 3	Agreement	Video 4
∅	loosen nut	loosen nut	3	ndo bolt
jack car	raise car	jack car	4	lift car
∅	∅	unscrew nut	1	∅
remove wheel	remove tire	withdraw tire	3	∅
∅	lower jack	∅	2	ower 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	
jack car	raise car	jack car	lift car	4	
∅	∅	unscrew nut	∅	4	
remove wheel	remove tire	withdraw tire	∅	3	
∅	lower jack	∅	lower car	2	<ol style="list-style-type: none">1) Loosen nut2) Jack car3) Remove wheel

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



- get things out
- start loose
- brake on
- jack up
- unscrewwheel
- withdrawwheel
- put wheel
- screwwheel
- jack down
- tight wheel

GROUND TRUTH

Video Prediction



Activity discovery from images and words

Make kimchi fried rice



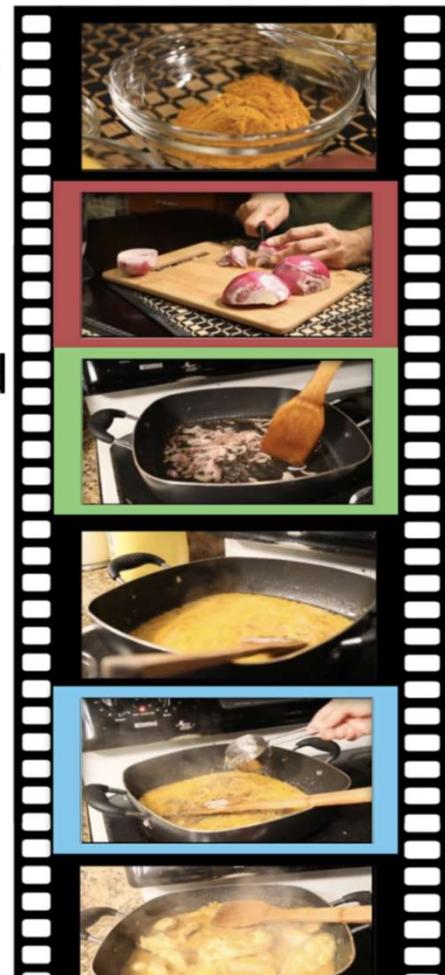
"I'm going to start off by chopping up an onion.

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



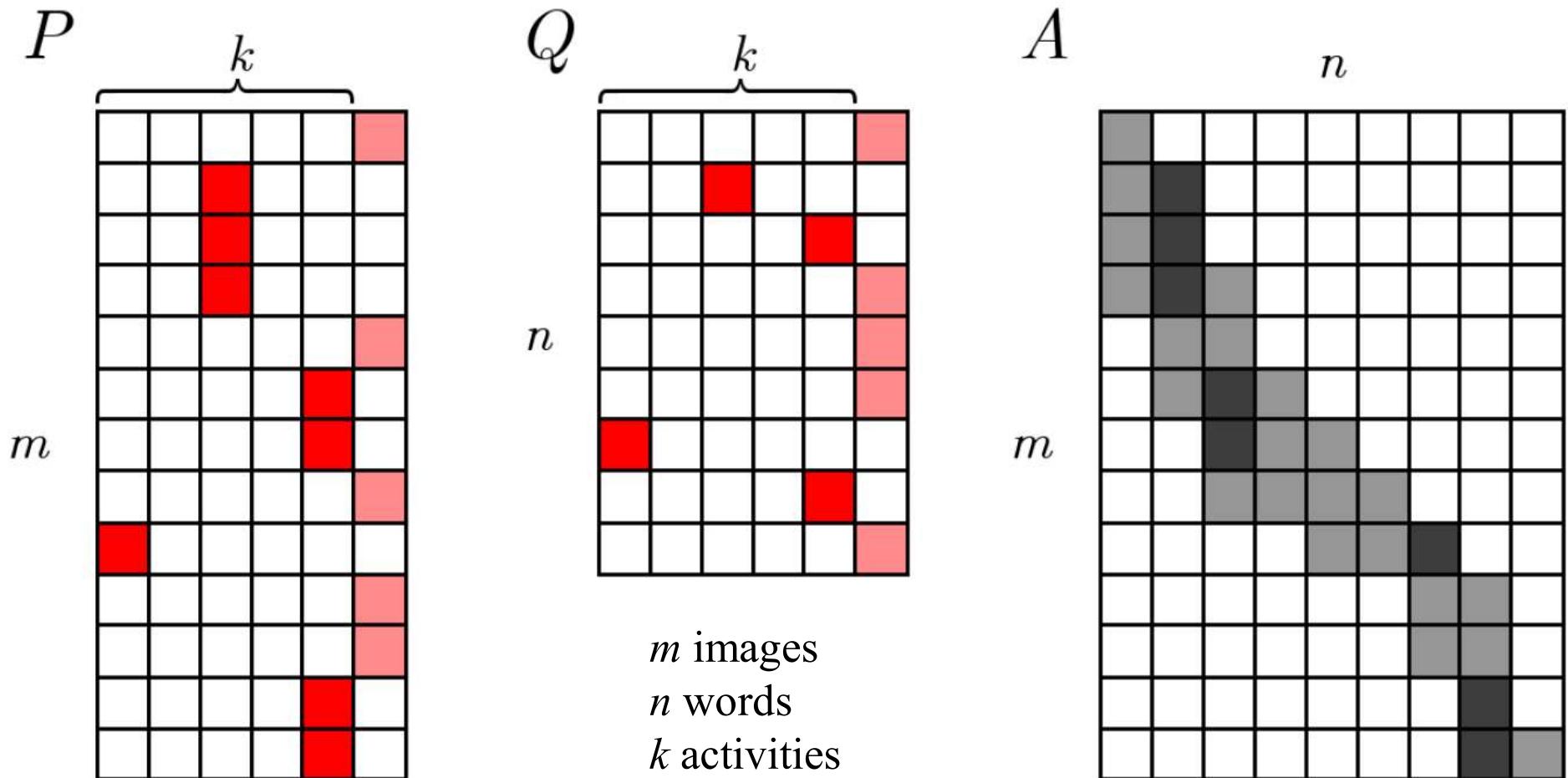
"...next you're going to take a full onion...

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

locate placement make seal take head
cover mouth fetch defibrillator place mouth allow chest
grasp wrist take mouth perform mouth
seal mouth
take breath bring tongue
pinch nostril feel anything
hear anything block airway allow air
place heel
pinch nose

left hand pump times step check give second
right hand avoid fatigue get arm change role
start compression do times stop cpr
get response observe color
ventilate patient release nose begin compression
have acronym follow steps
give air use thumb make rise keep rhythm
press times begin cpr start cpr
keep elbow support weight have breath do team cough movement
give breath
give compression

feel air slide finger have signs save organs cause air
give ventilation protect rescuer rub bone
turn kind see rise
clutch chest want blade leave finger
perform compression
see movement contact em
lift chin put heel
do compression
deliver breath see breathing

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

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

reposition airway
reopen airway
open airway
have partner
complete cycle

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

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

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)

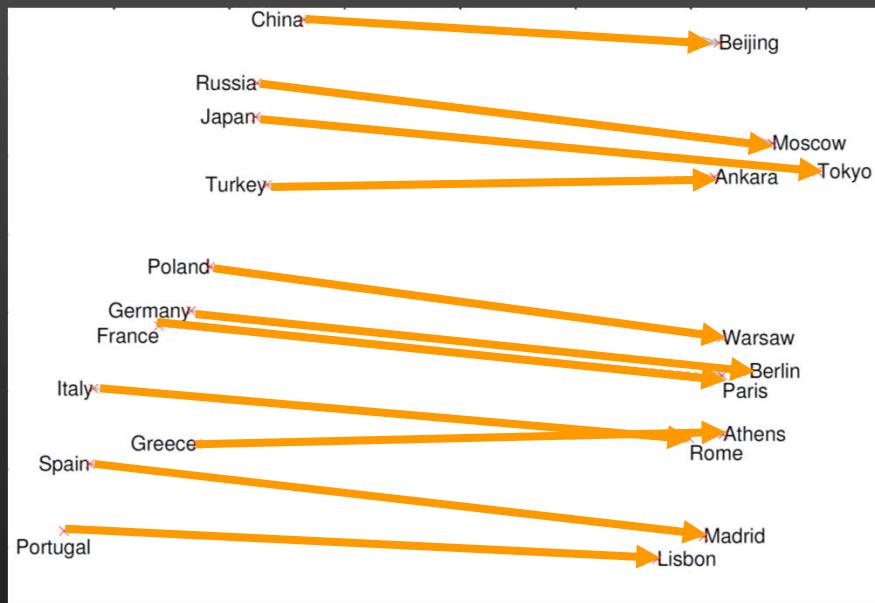
Co-segmentation/localization (Rother'06; Russell'06; Joulin'10; Kim'11; Vicente'11; Joulin'14; Tang'14)

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”



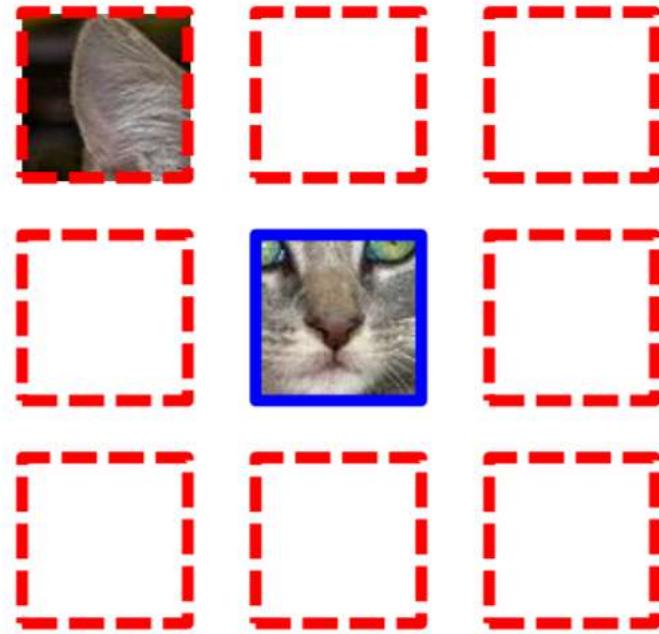
Note: Visualization in 2D
but $d \approx 300$

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)

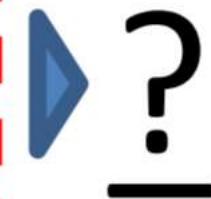
Example:



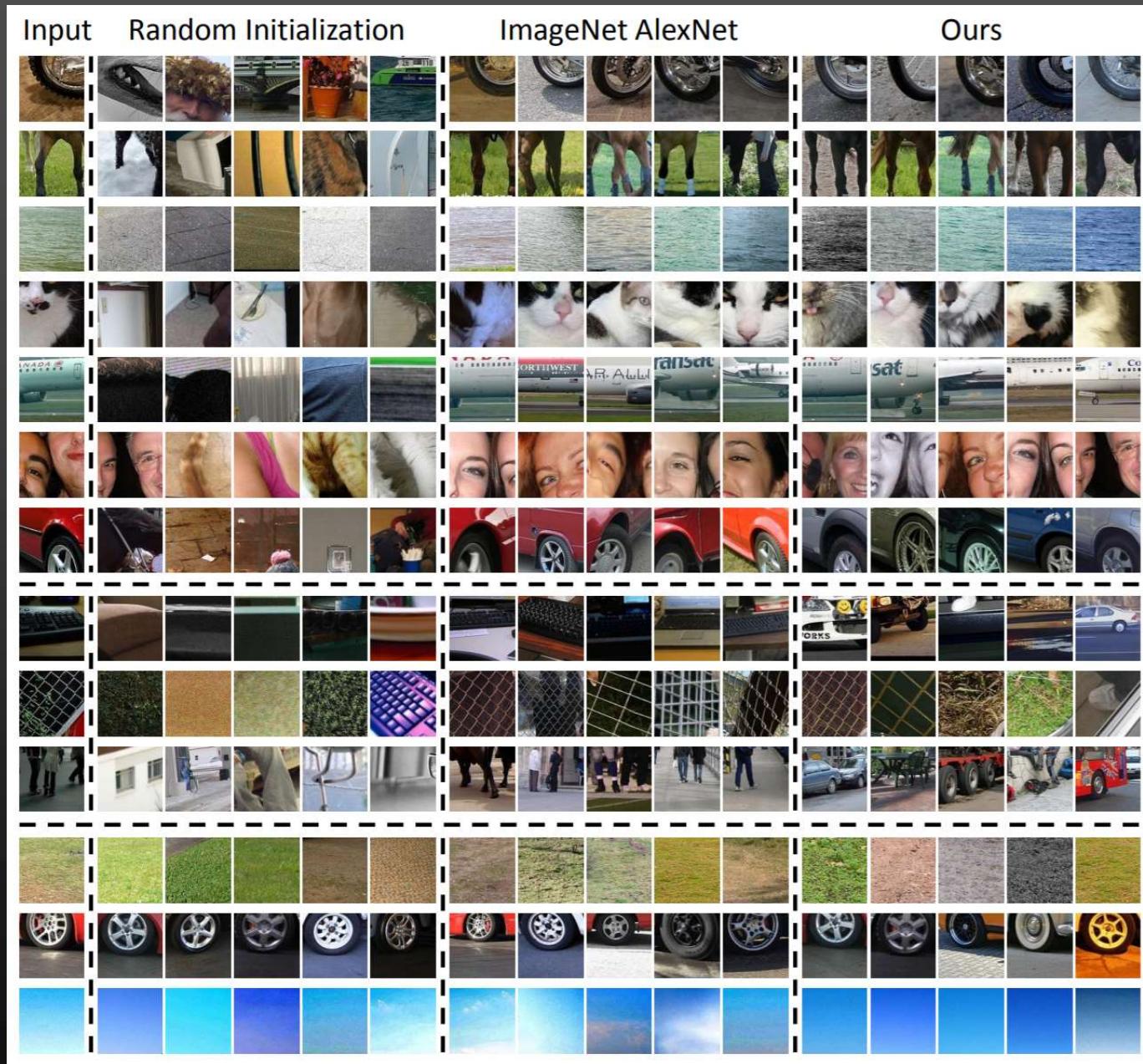
Question 1:



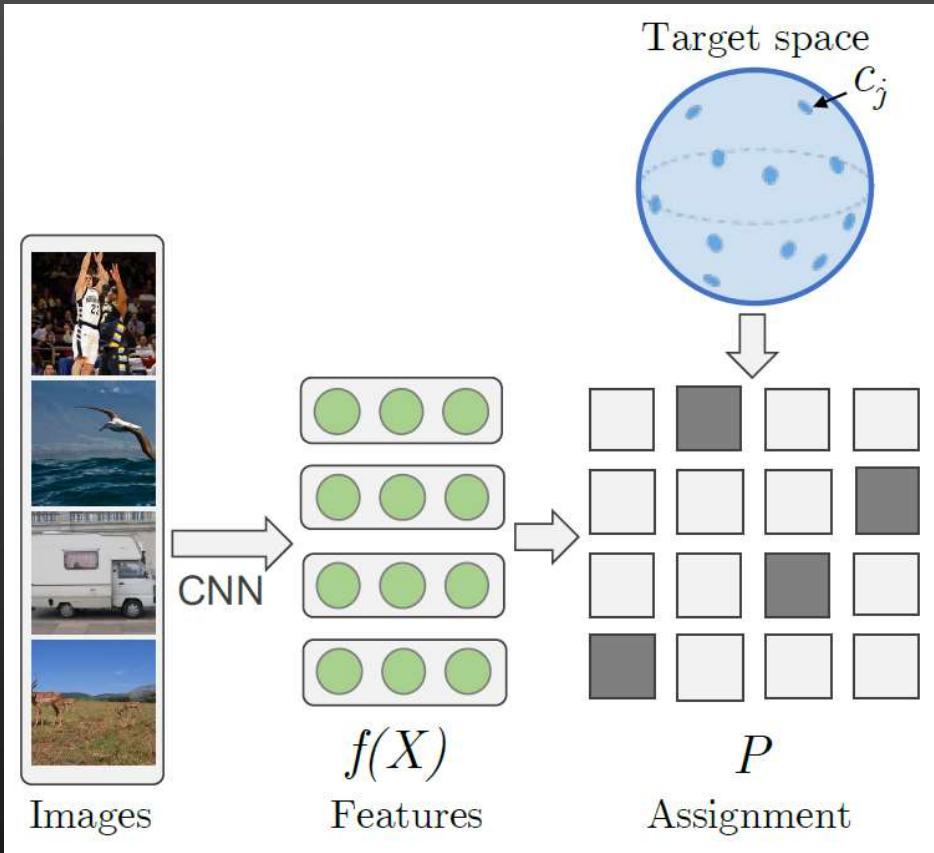
Question 2:



Retrieved nearest neighbors



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



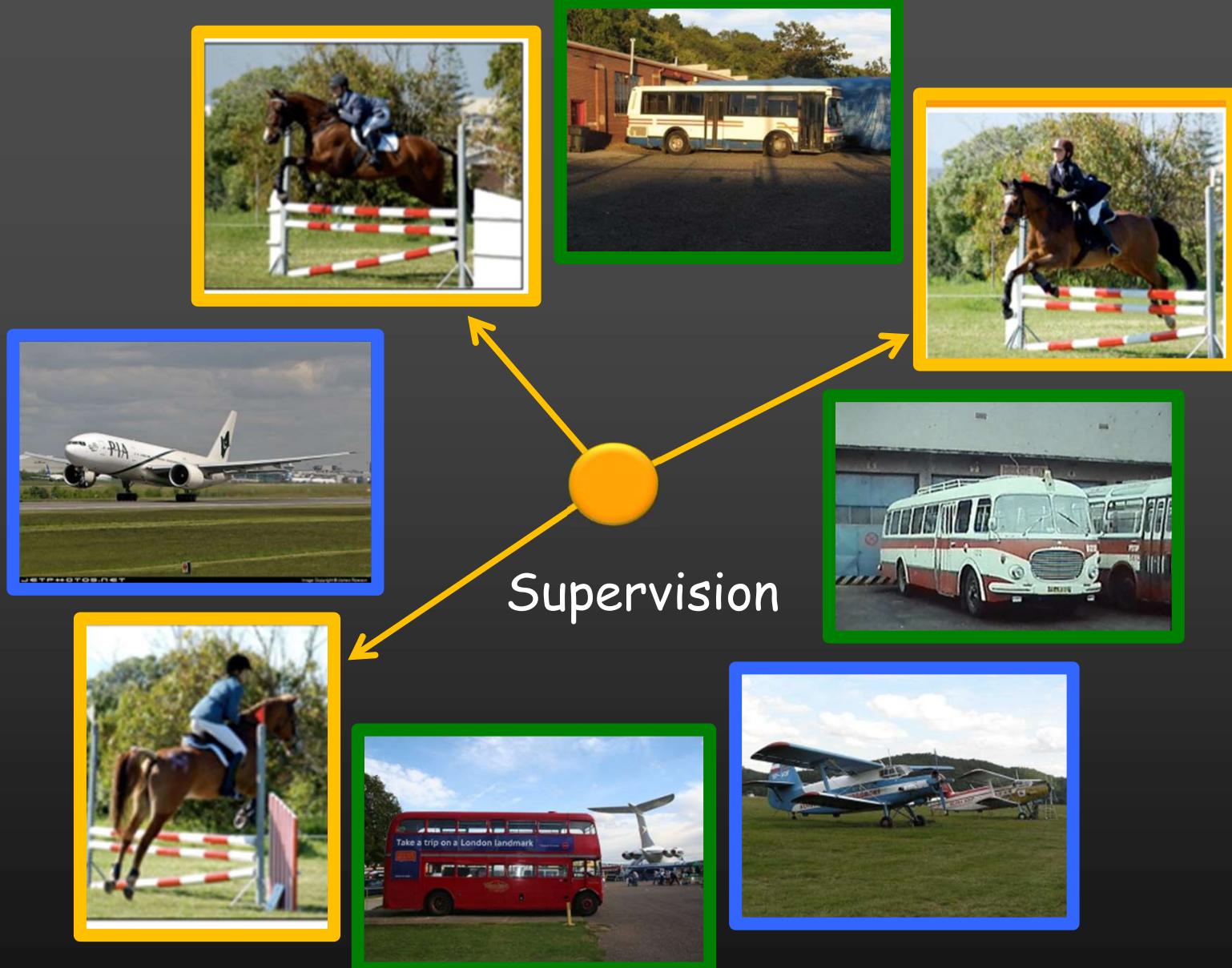
$$\min_{\theta} \min_{Y \in \mathbb{R}^{n \times d}} \frac{1}{2n} \|f_{\theta}(X) - Y\|_F^2$$



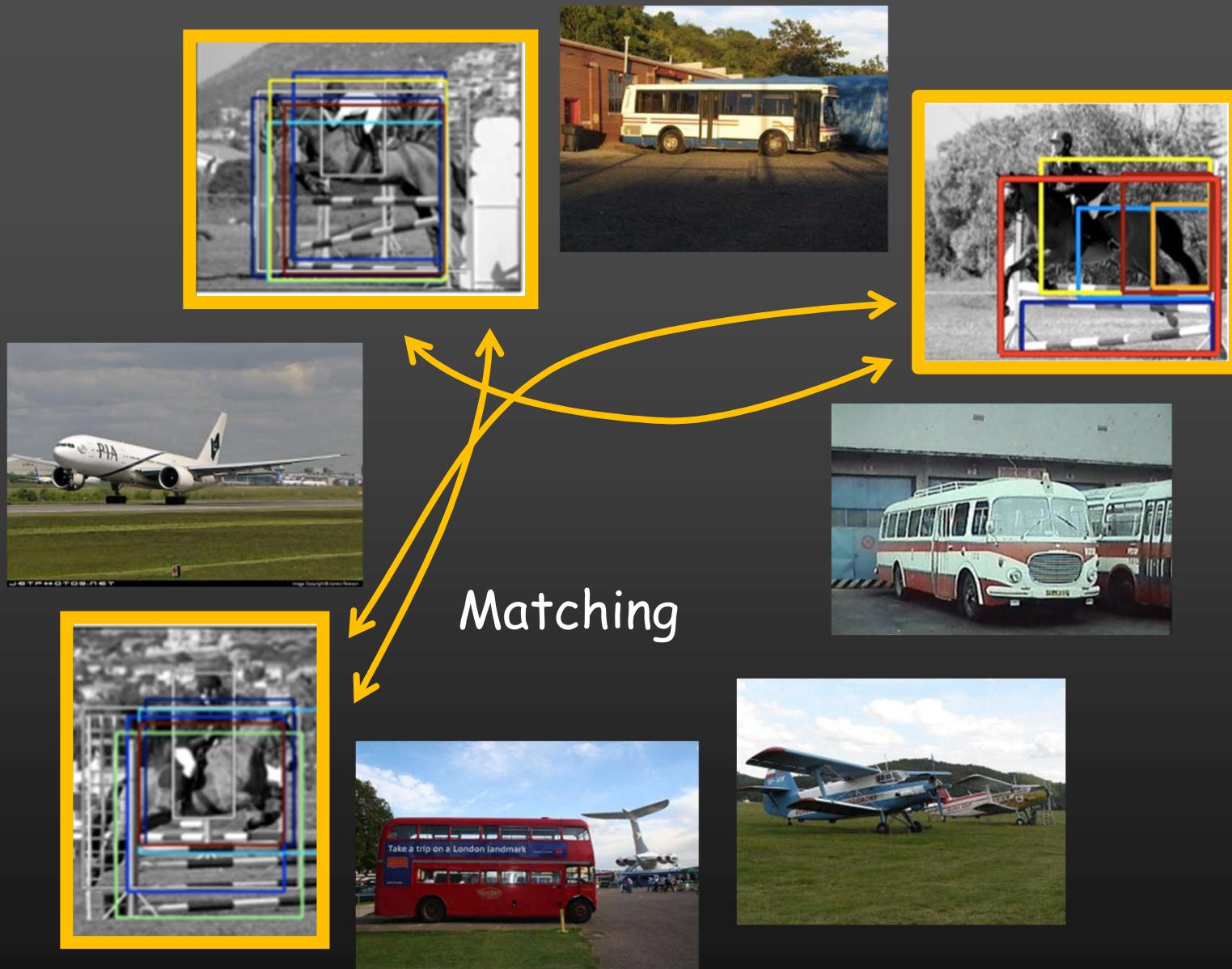
$$\max_{\theta} \max_{P \in \mathcal{P}} \text{Tr} (P C f_{\theta}(X)^{\top})$$

Retrieved nearest neighbors



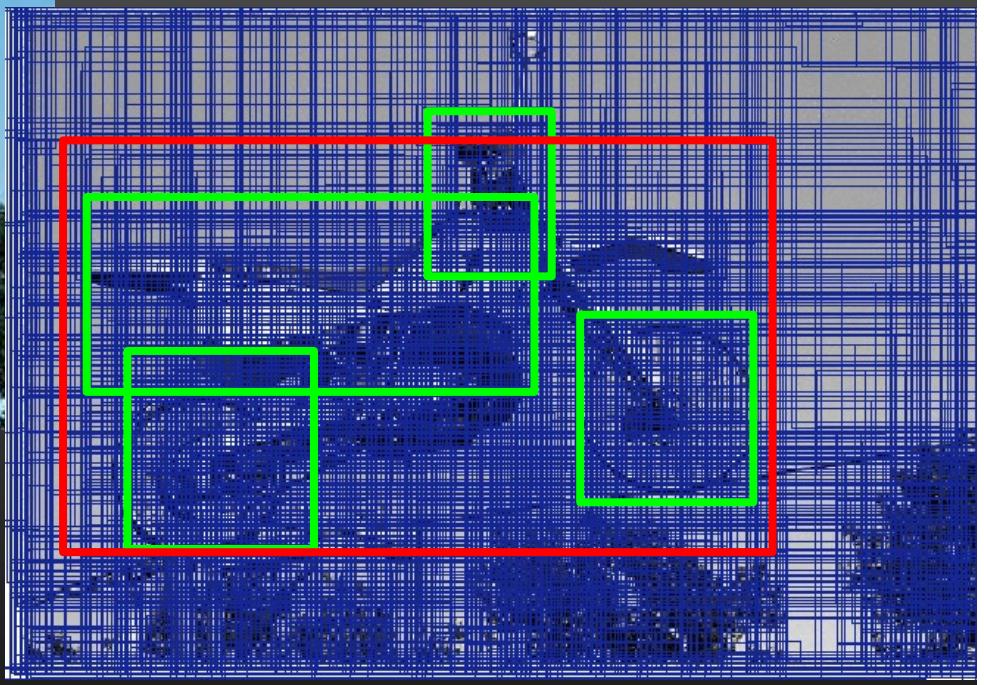


(Cho, Kwak, Schmid, Ponce, 2015)



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

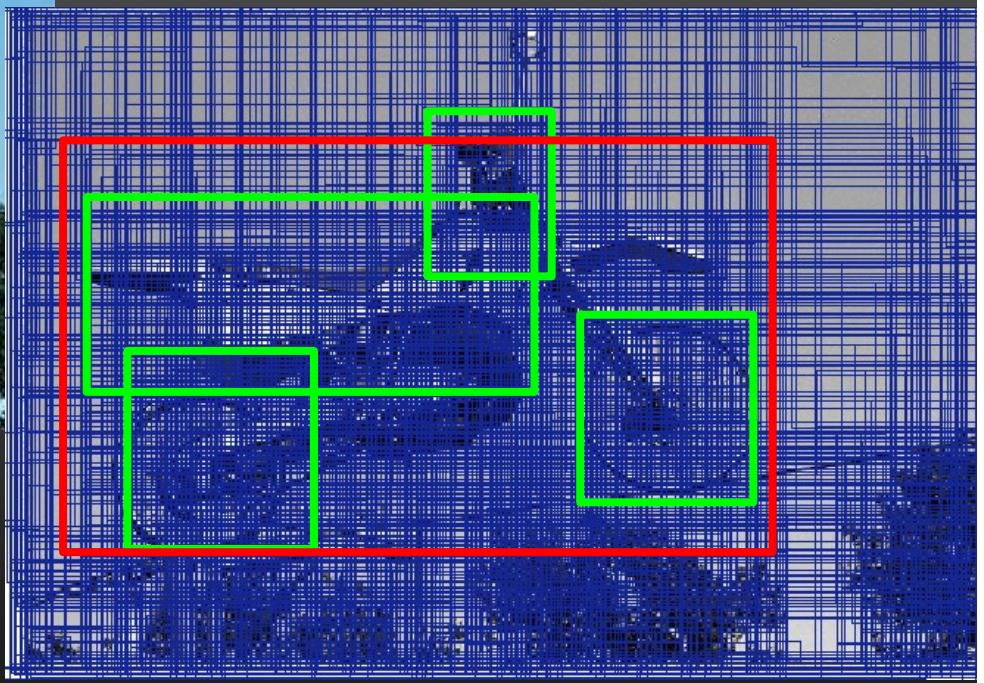
Finding **parts** and **objects** among region candidates



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

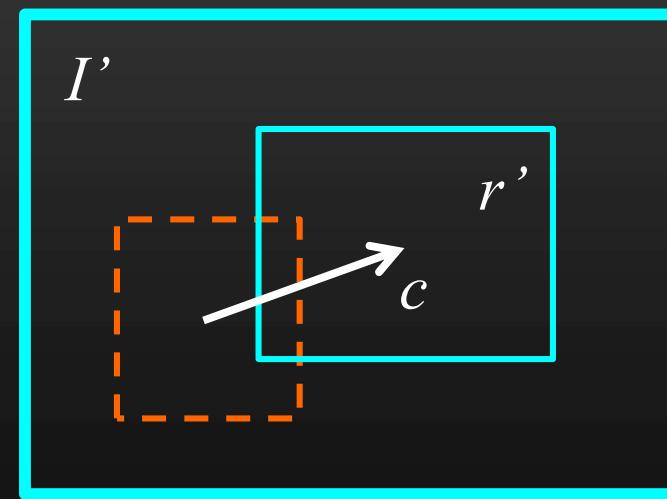
Caveat: These region proposals are supervised

Matching model - Probabilistic Hough matching

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

match data configuration
two regions region proposals position+scale

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



Matching model - Probabilistic Hough matching

$$\begin{aligned} \text{match} & \quad \text{data} & \text{configuration} \\ P(m | d) &= \sum_c P(m | c) P(c | d) \\ &= P(m_a) \sum_c P(m_g | c) P(c | d) \\ \text{appearance} & & \text{geometry} \end{aligned}$$

Matching model - Probabilistic Hough matching

- Bayesian model

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

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

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

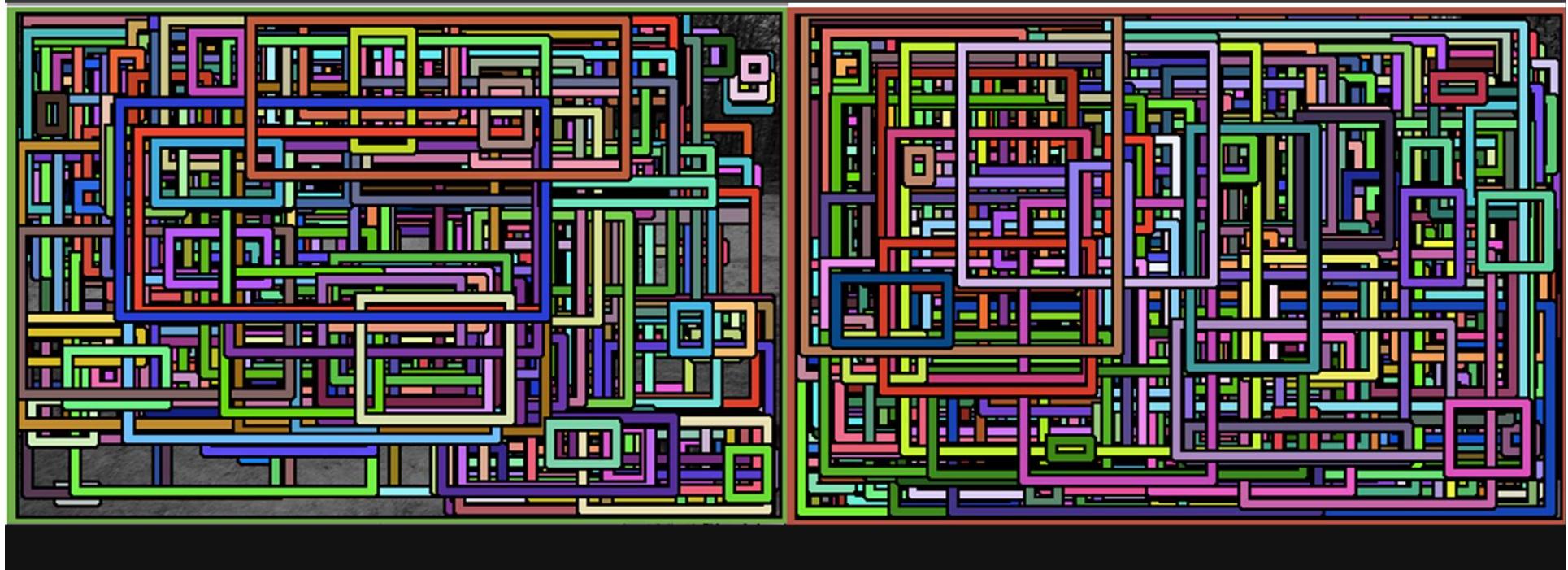
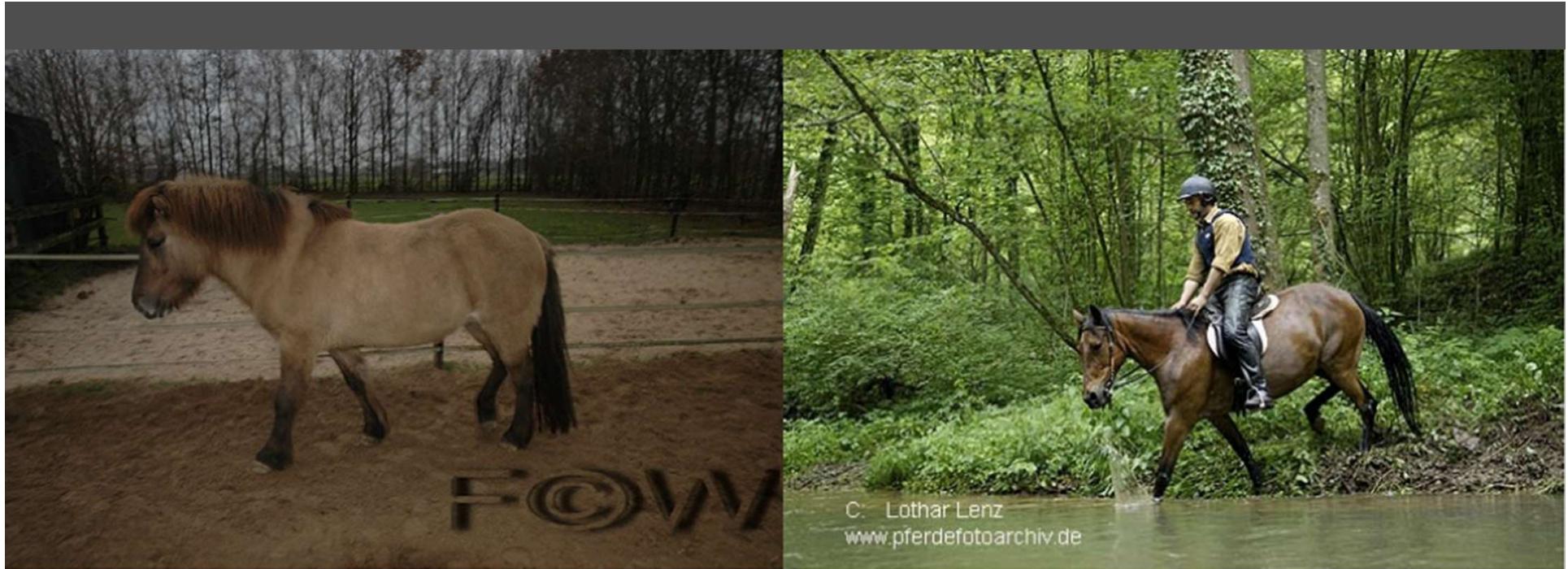
$$\begin{aligned} P(m | d) &= \sum_c P(m | c) P(c | d) \\ &= P(m_a) \sum_c P(m_g | c) P(c | d) \end{aligned}$$

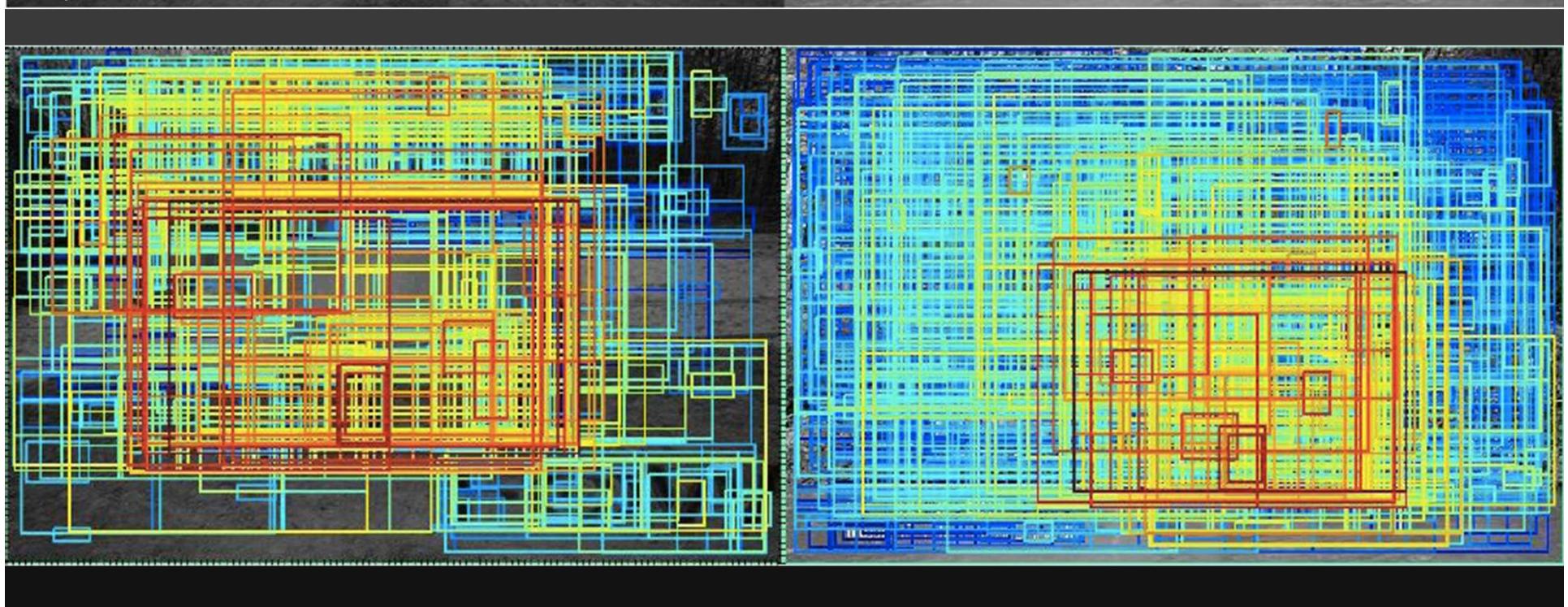
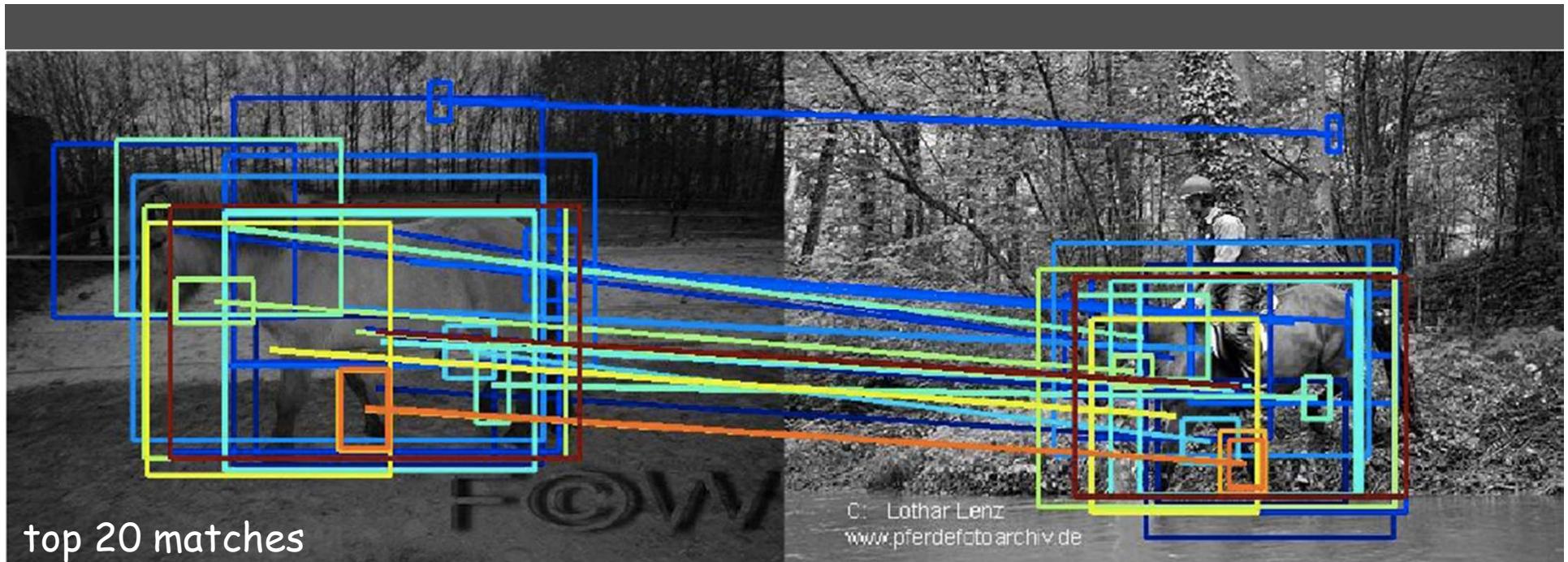
- Probabilistic Hough transform

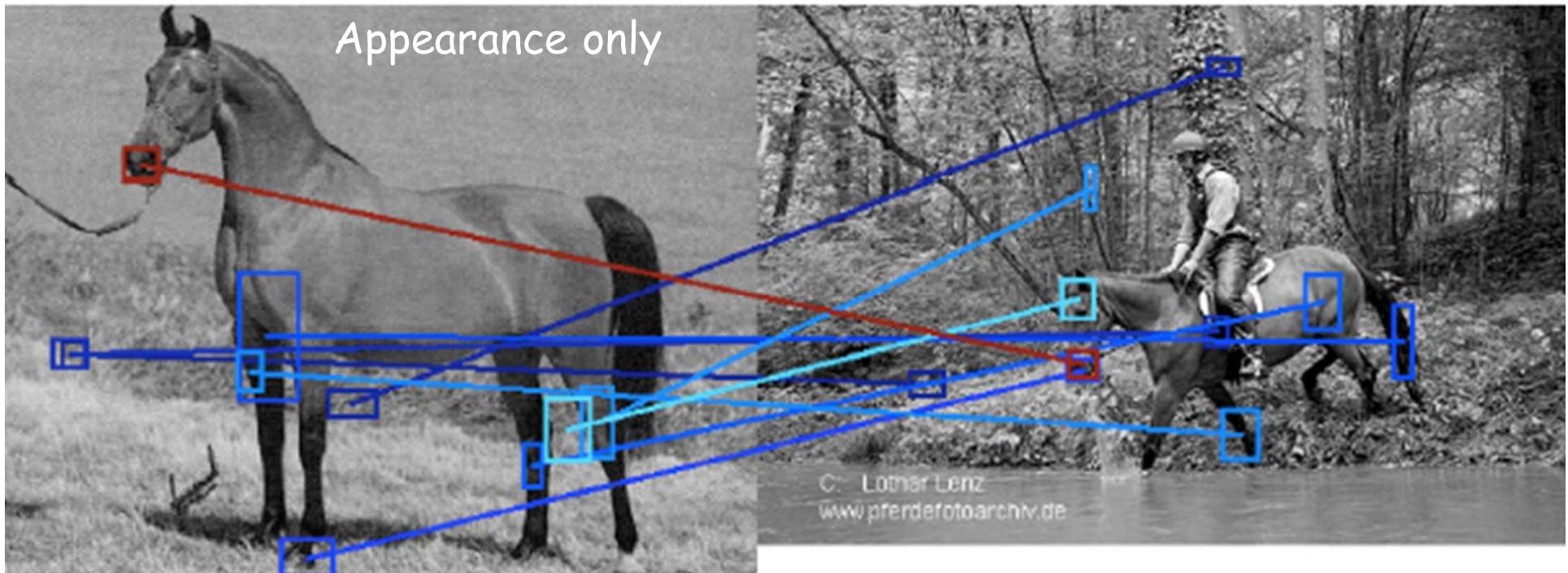
$$\begin{aligned} P(c | d) &\approx H(c | d) = \sum_{m \in d} P(m | c) \\ &= \sum_{m \in d} P(m_a) P(m_g | c) \end{aligned}$$

- Region confidence

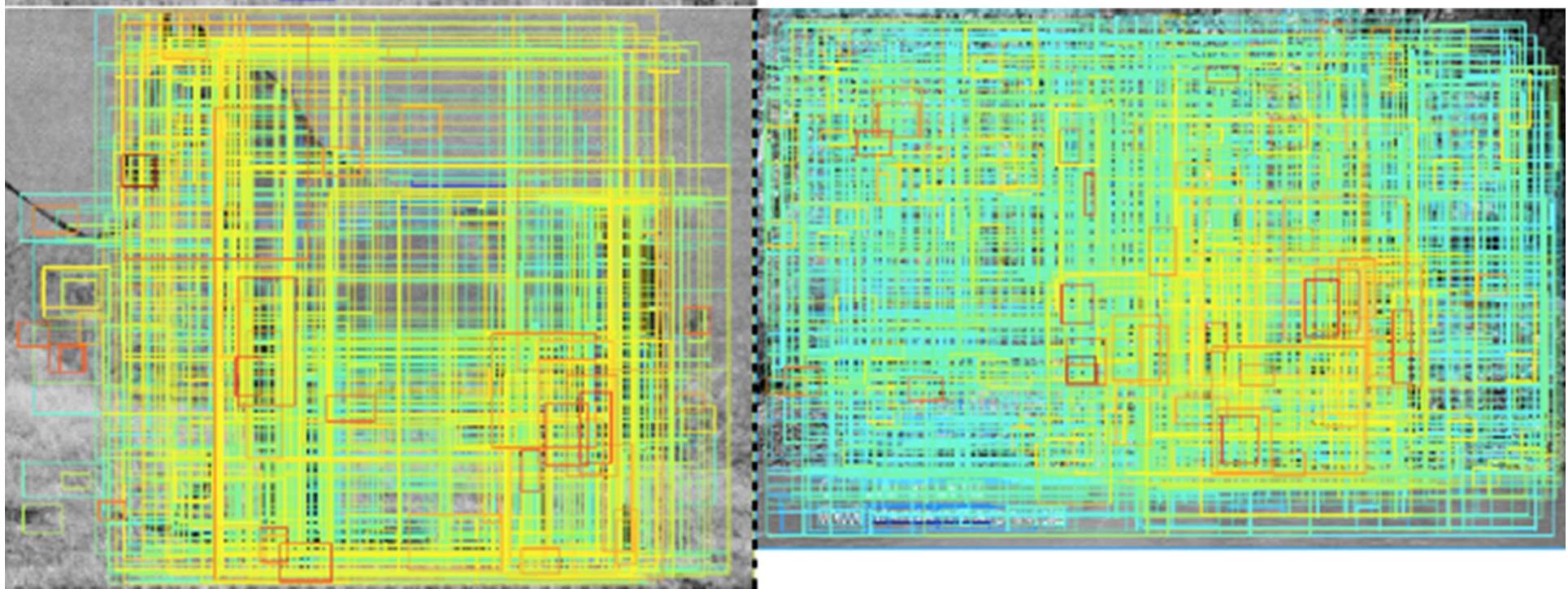
$$C(r' | d) = \max_{r''} P(r' \leftrightarrow r'' | d)$$

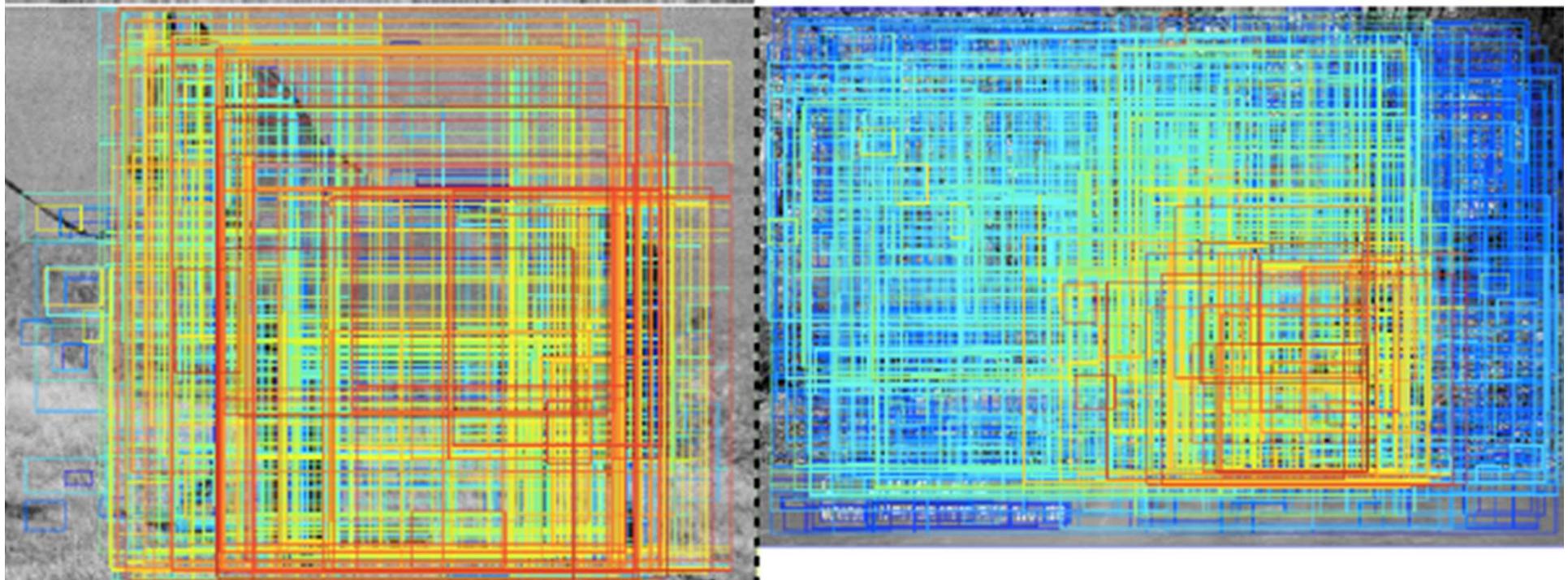
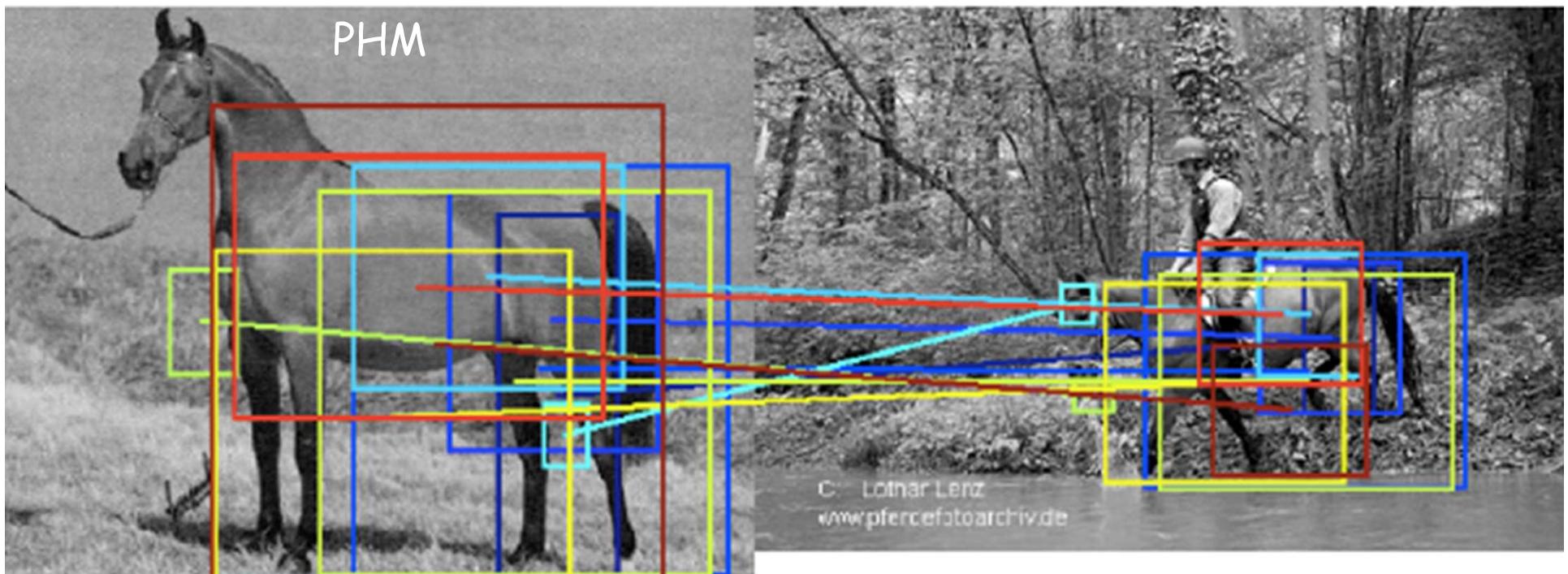






C: Lothar Lenz
www.pferdefotoarchiv.de





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

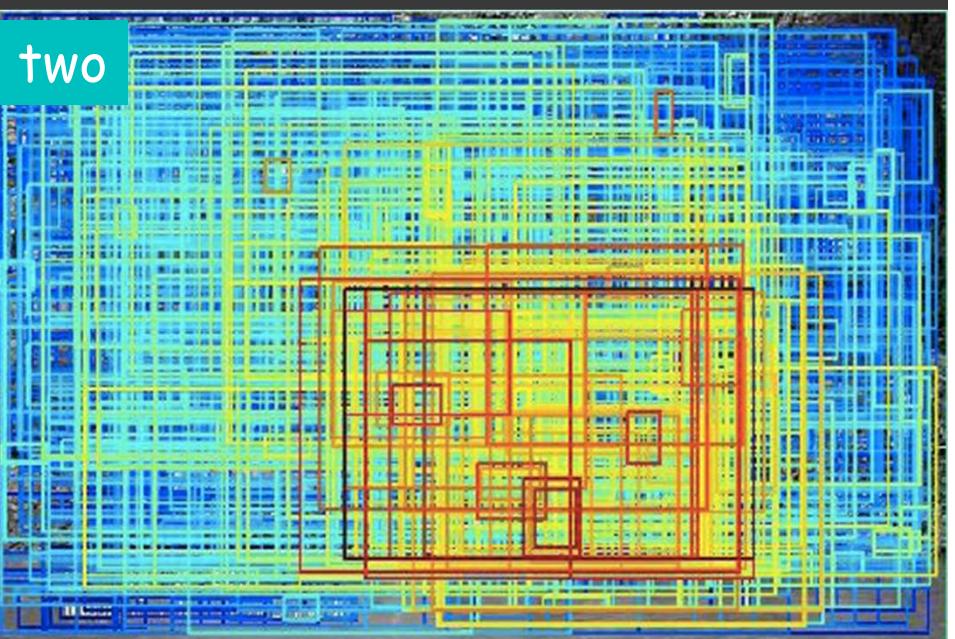
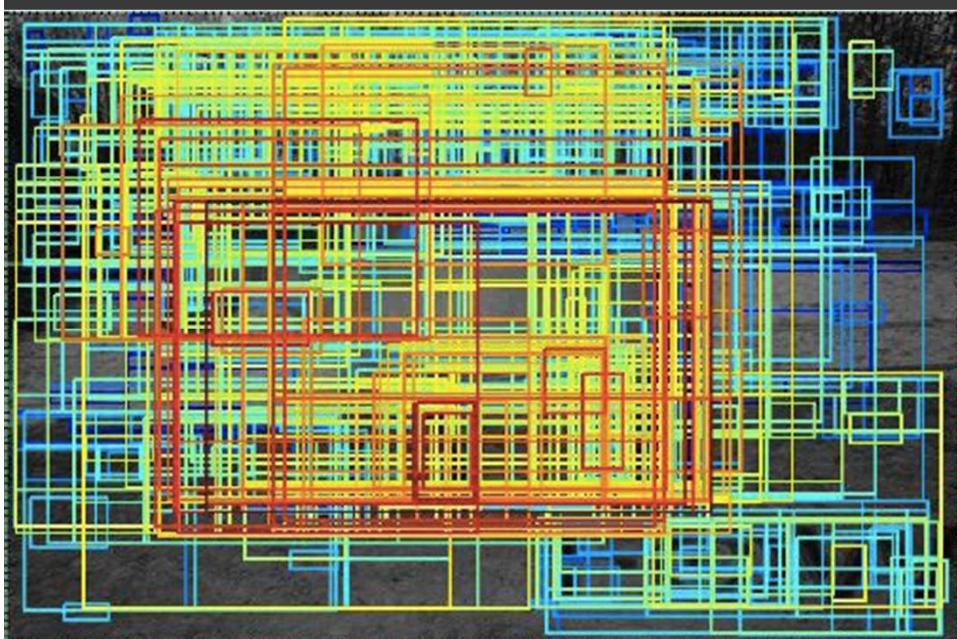
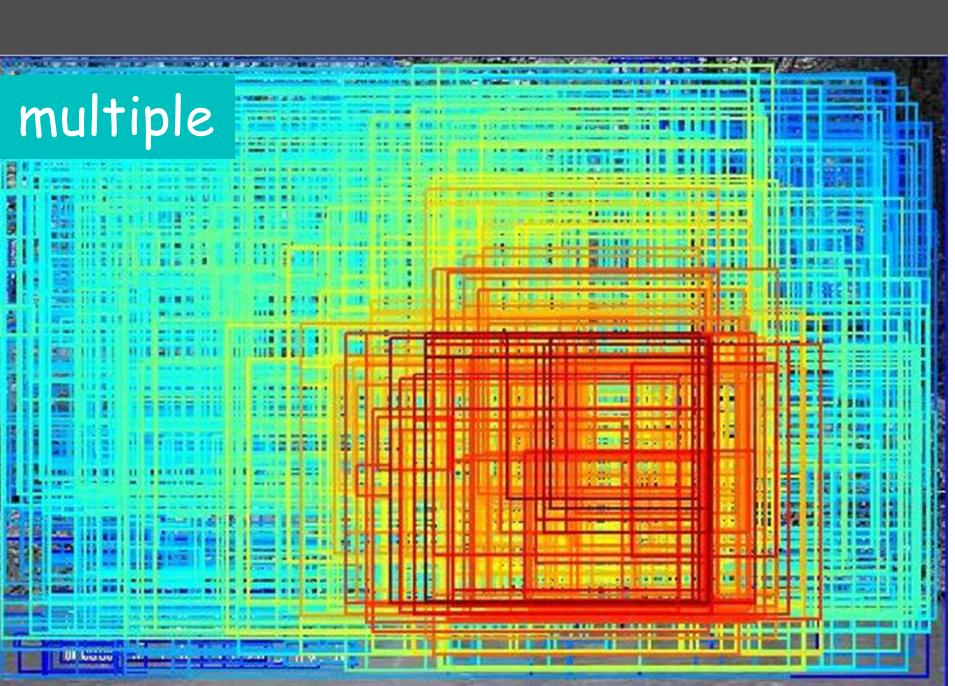
$$\begin{aligned} P(c | d) &\approx H(c | d) = \sum_{m \in d} P(m | c) \\ &= \sum_{m \in d} P(m_a) P(m_g | c) \end{aligned}$$

- Two images \rightarrow multiple images

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



C. Lothar Lenz
www.pferdefotoarchiv.de



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



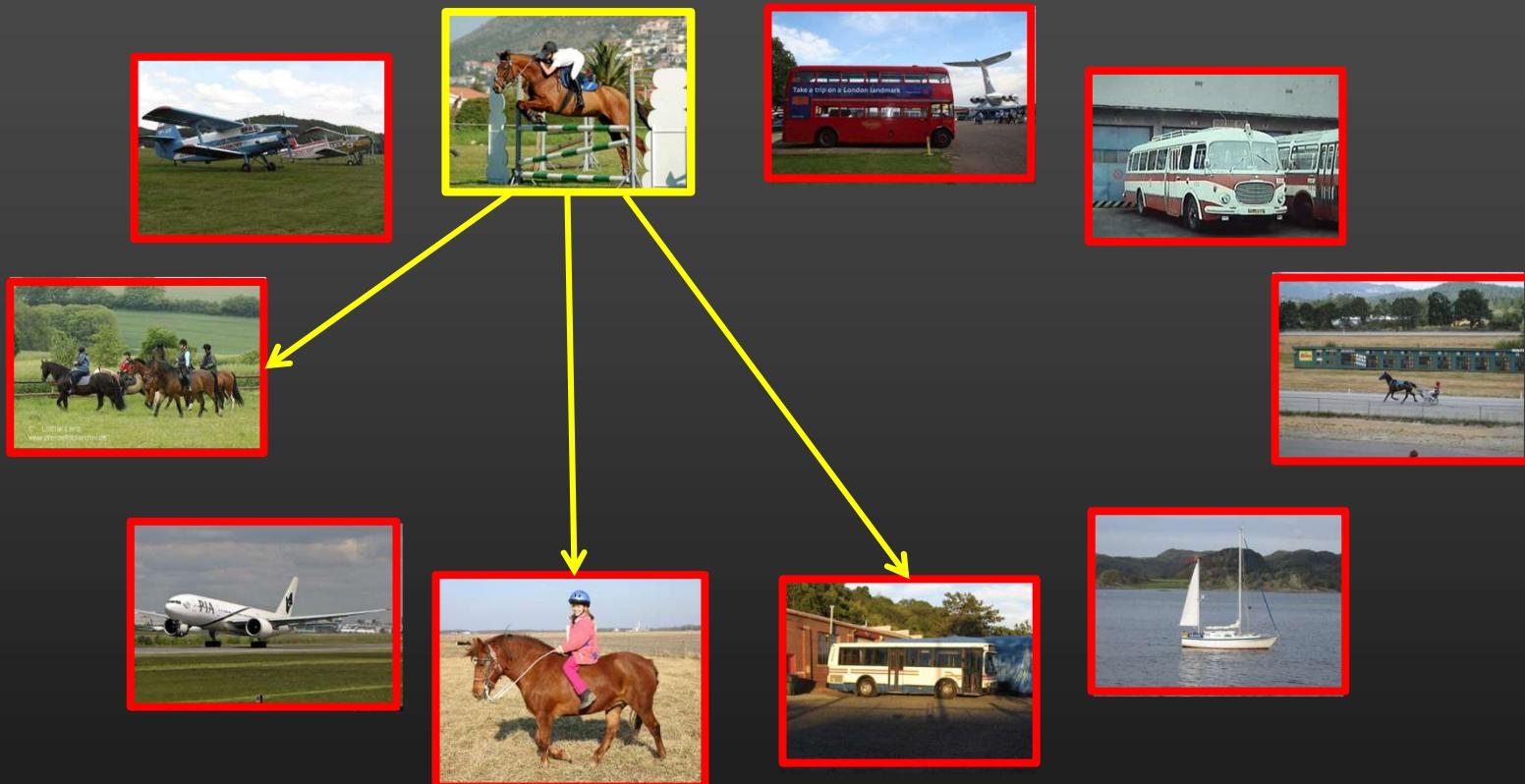
- Object regions should contain
 - more foreground than part regions
 - less background than larger regions
- $S_d(r) = C_d(r) - \max_{r' \supset r} C_d(r')$

A simple iterative algorithm



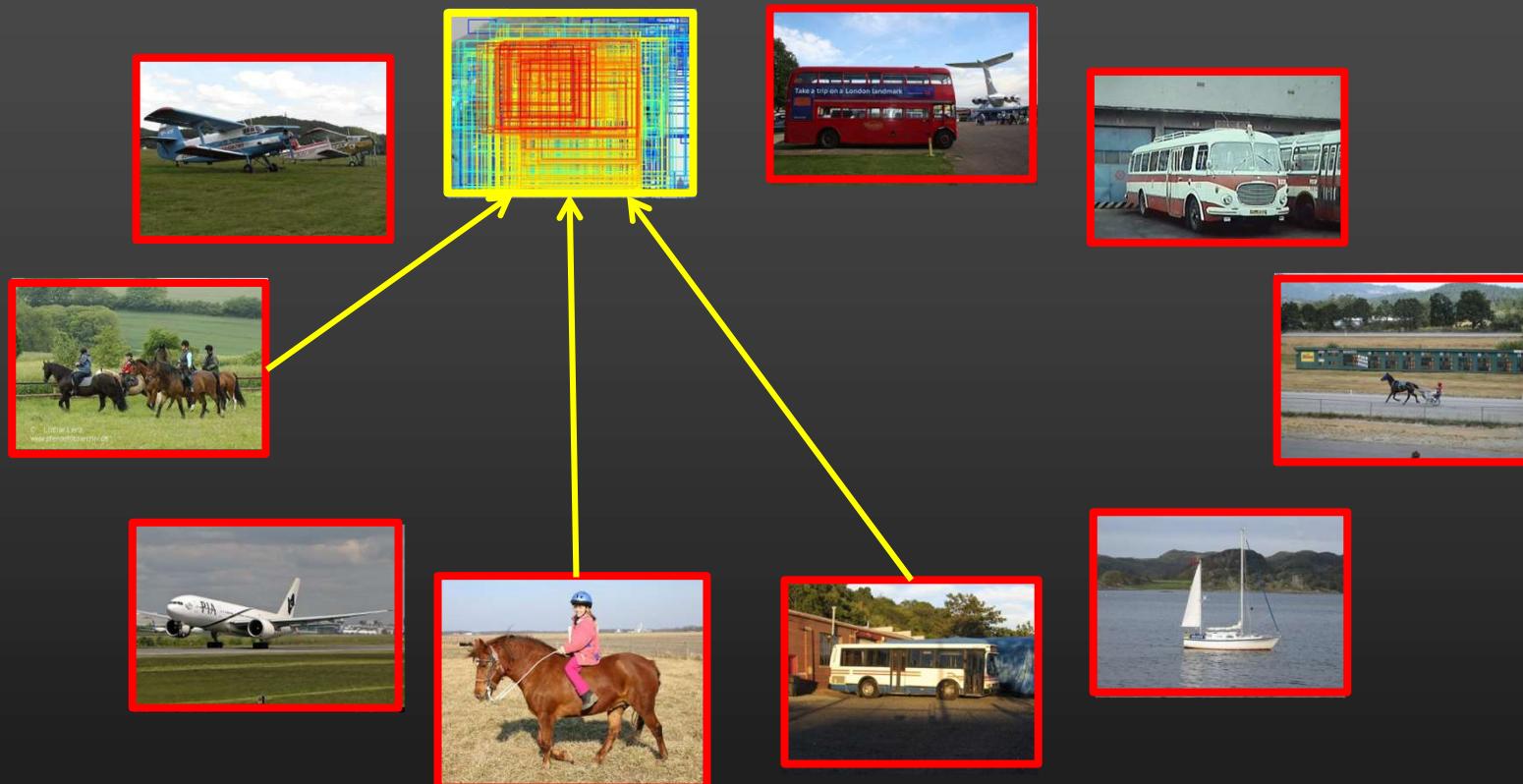
Initialize

A simple iterative algorithm



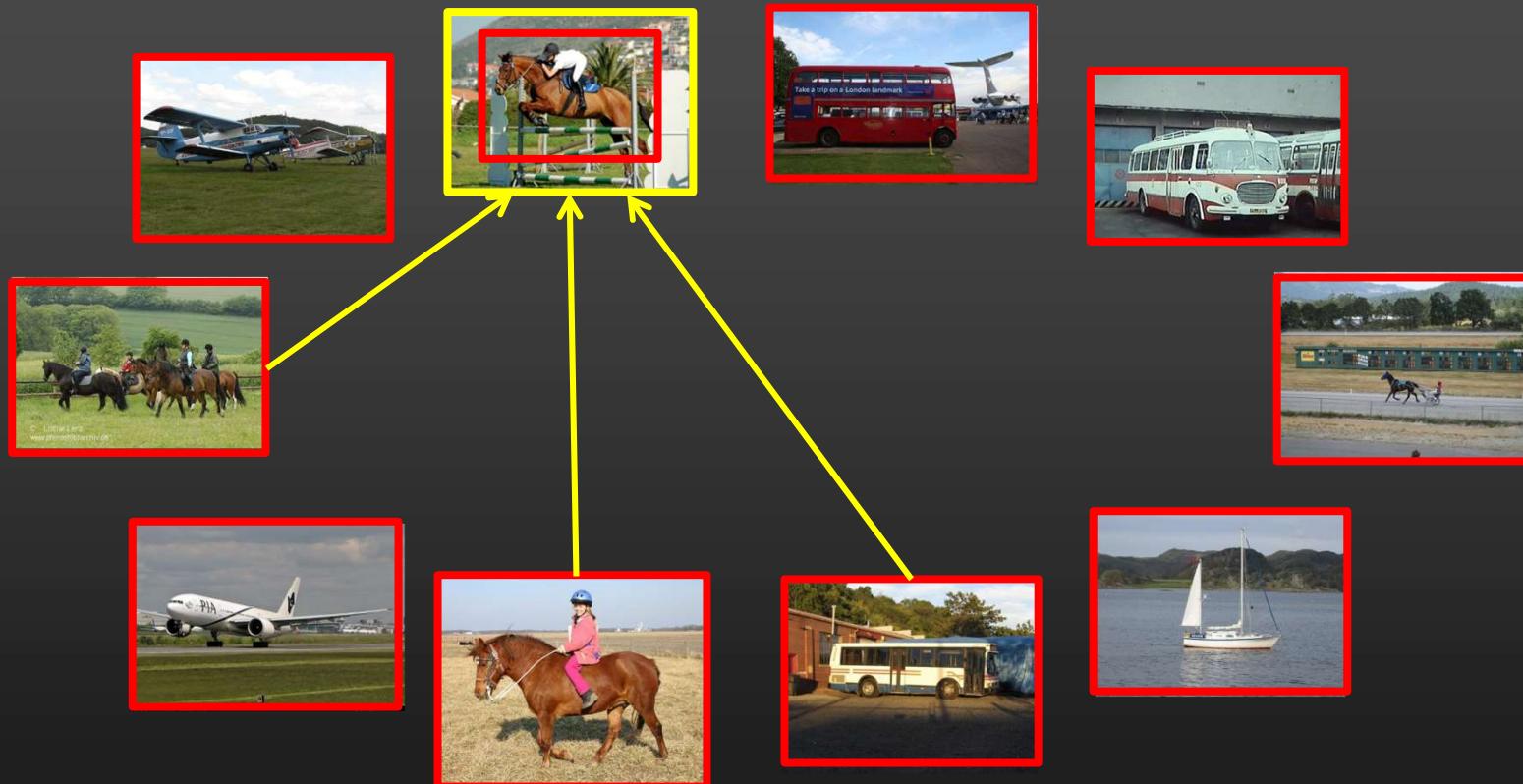
Retrieve 10 nearest neighbors (Oliva & Torralba'06)

A simple iterative algorithm



Match

A simple iterative algorithm



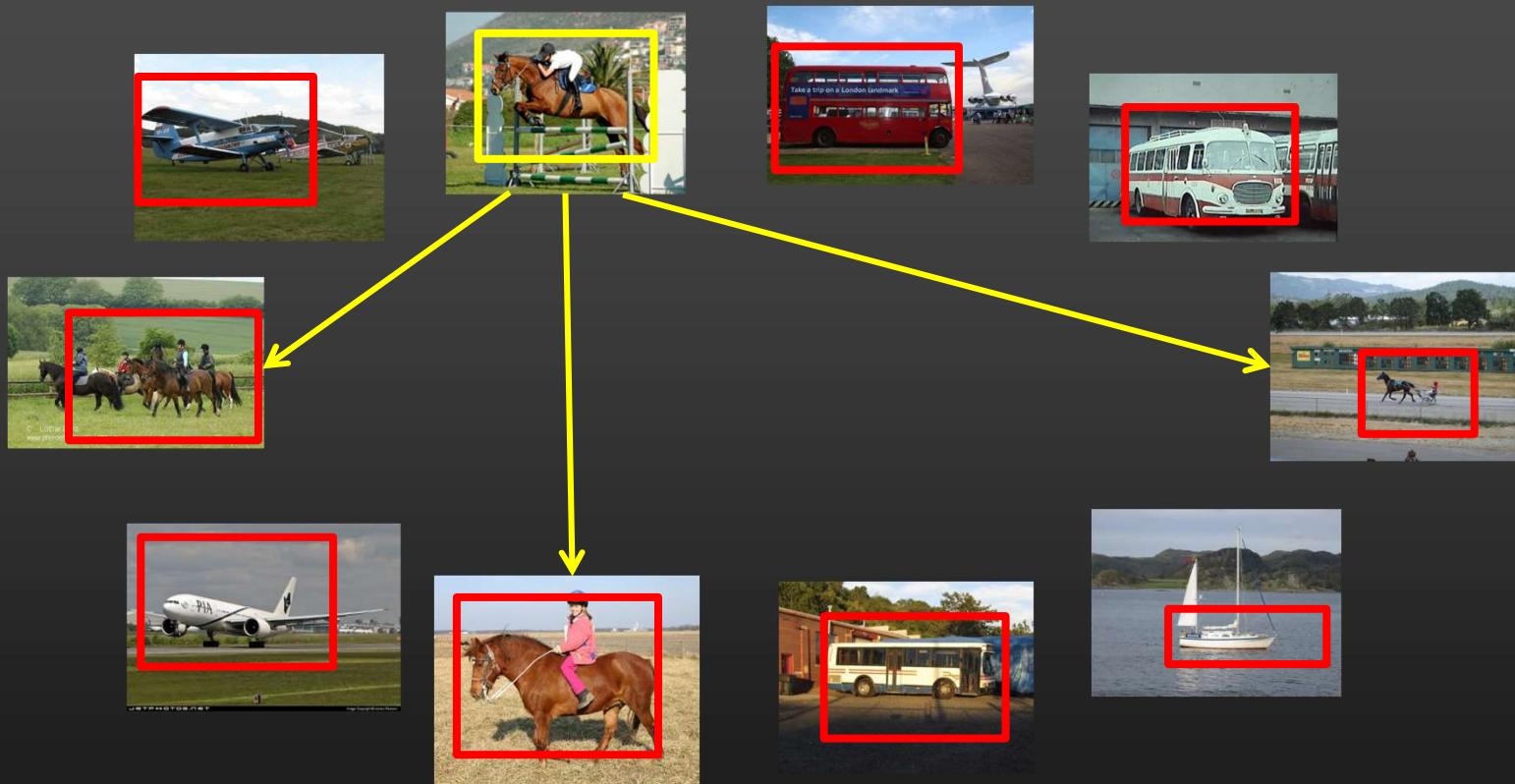
Localize

A simple iterative algorithm



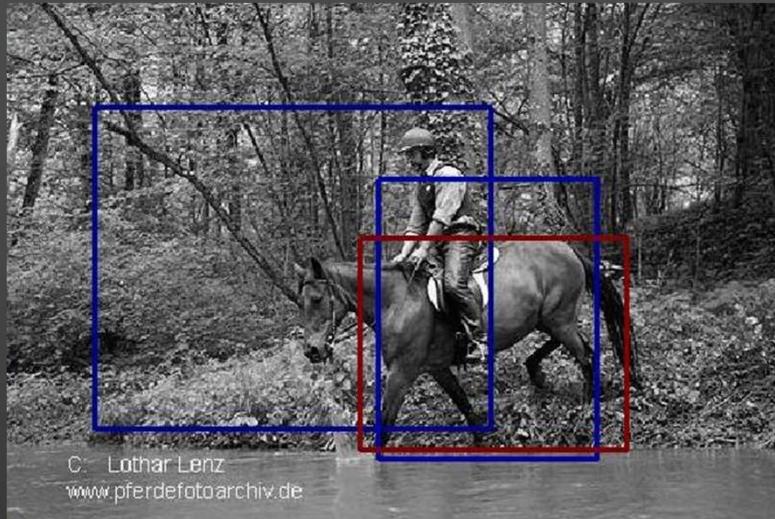
Localize

A simple iterative algorithm



Retrieve using top 20 confidence scores, etc.

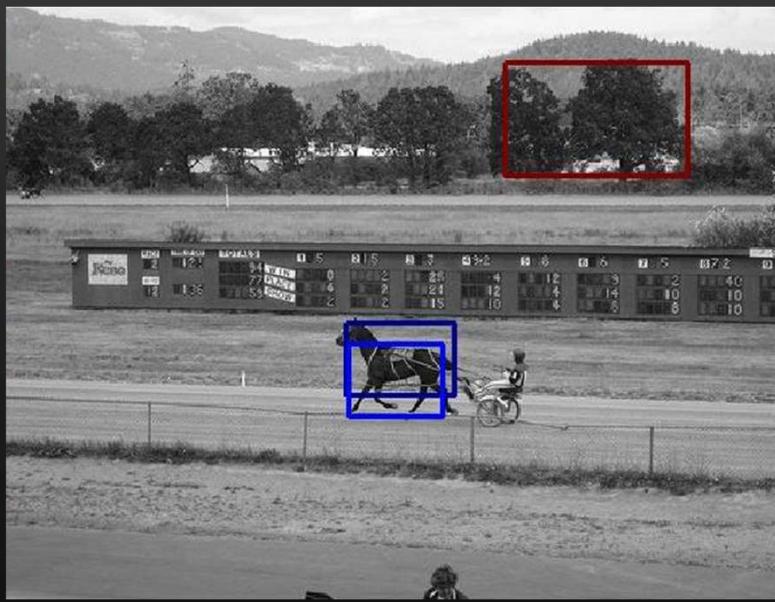
Localization improvement over iterations



After 1 iteration



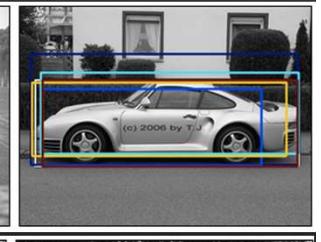
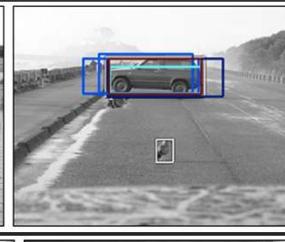
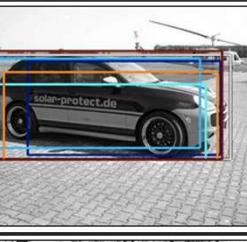
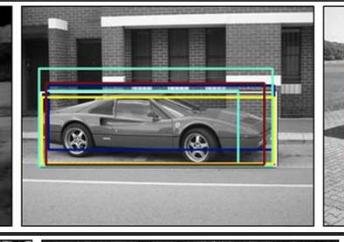
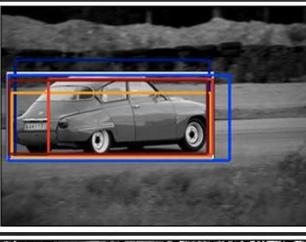
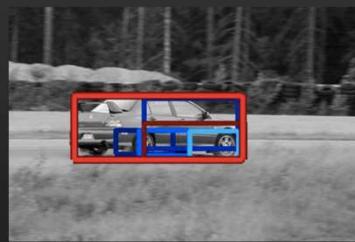
After 3 iterations



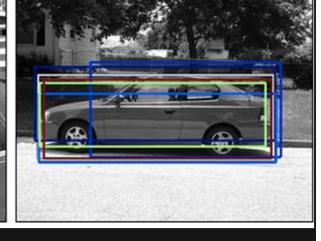
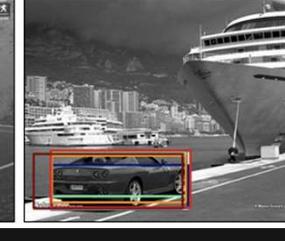
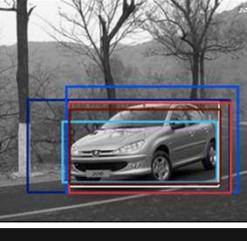
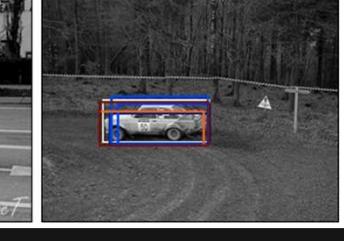
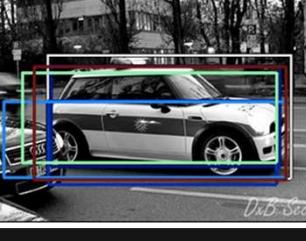
Retrieval improvement over iterations



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	trai	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 <i>et al.</i> [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 <i>et al.</i> [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 <i>et al.</i> [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 <i>et al.</i> [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 <i>et al.</i> [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

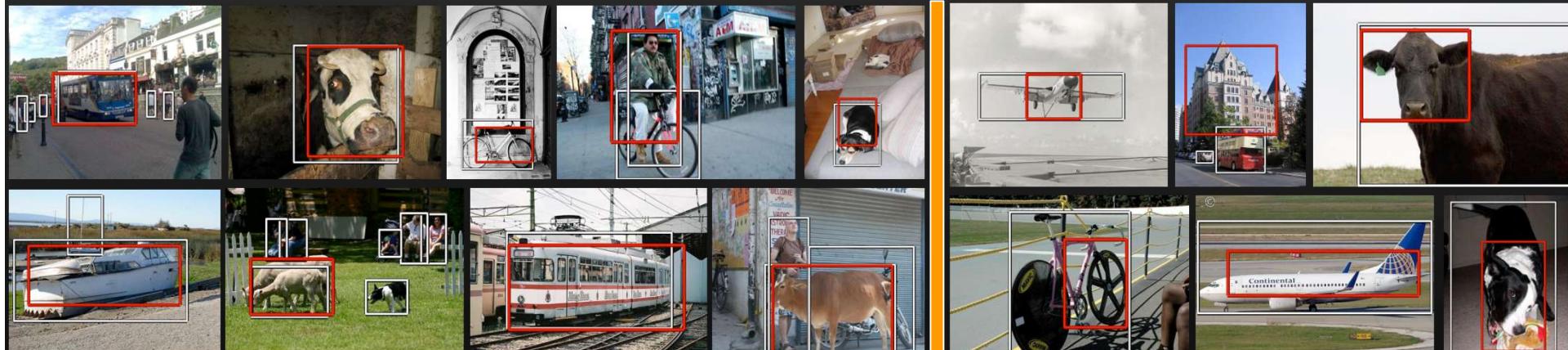
Uses pre-trained CNN features

Evaluation metric	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	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

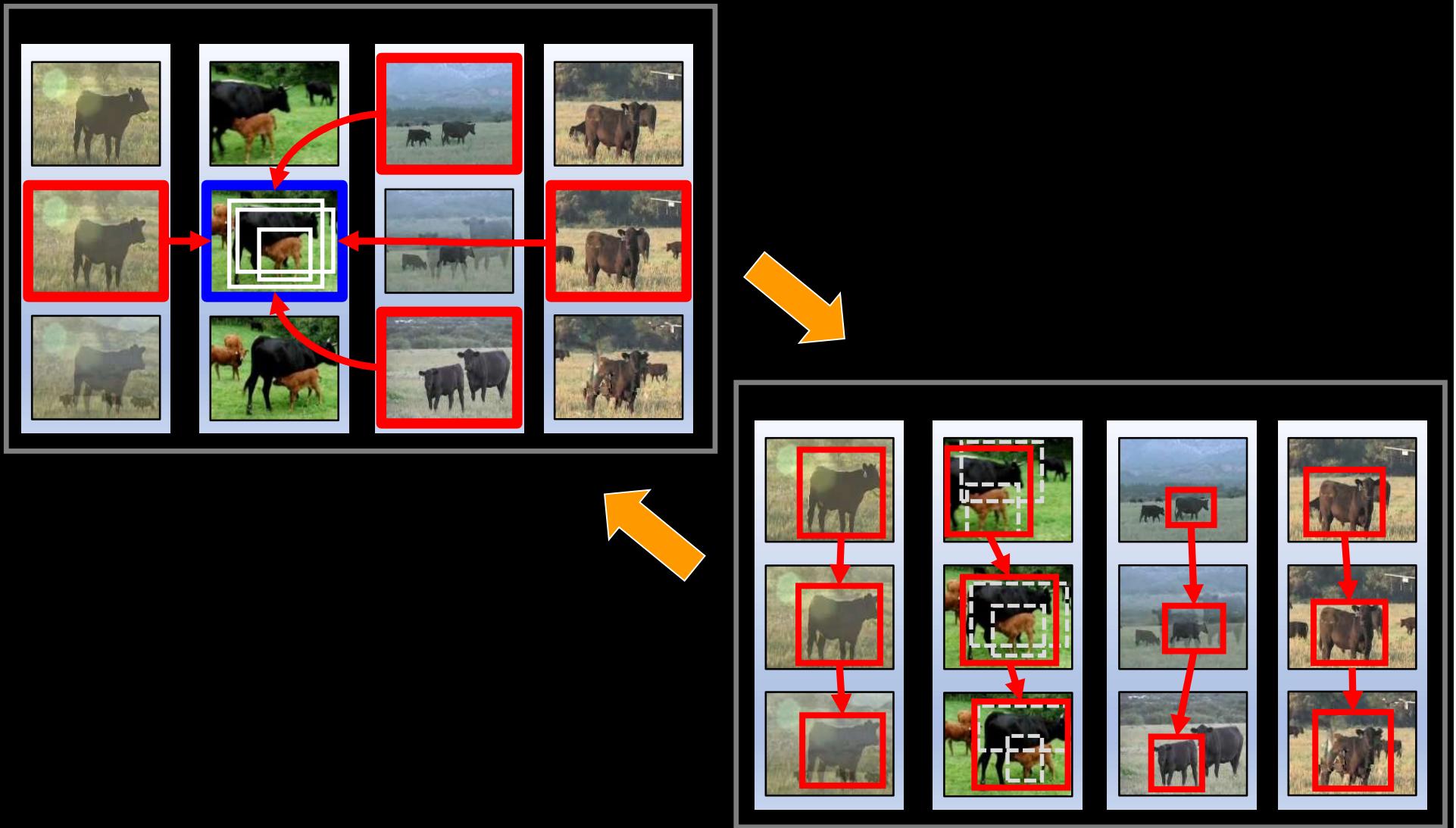
Examples - mixed classes

Successes

Failures



Unsupervised object discovery in multiple videos



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

aeroplane-0004-029

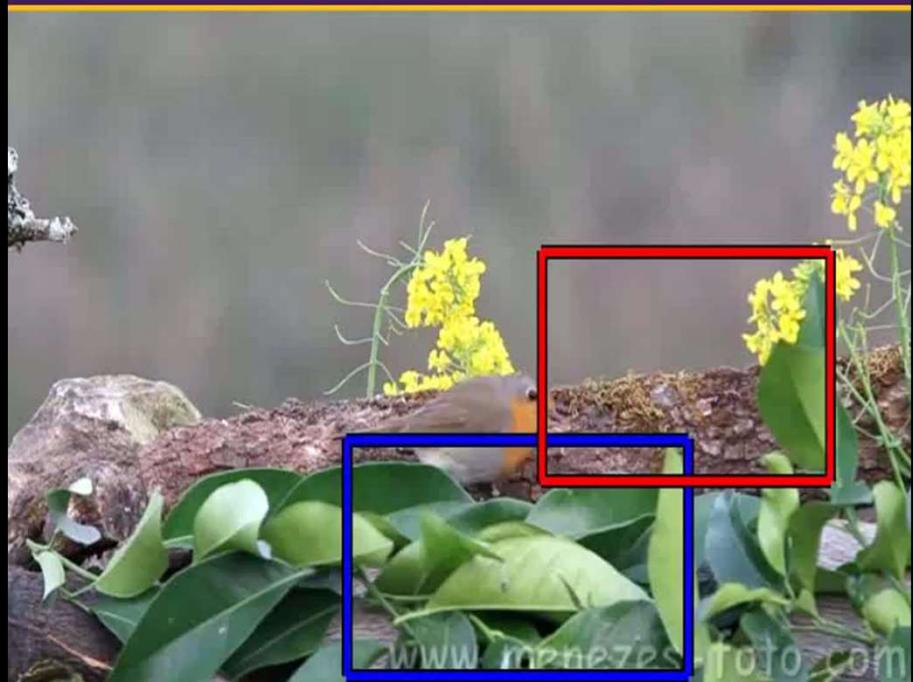
- Object colocalization per class
- Unsupervised object discovery



Copyright © Simon Lowe

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

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

subject to $\forall i \in 1 \dots n, \sum_{k=1}^p x_i^k \leq \nu$

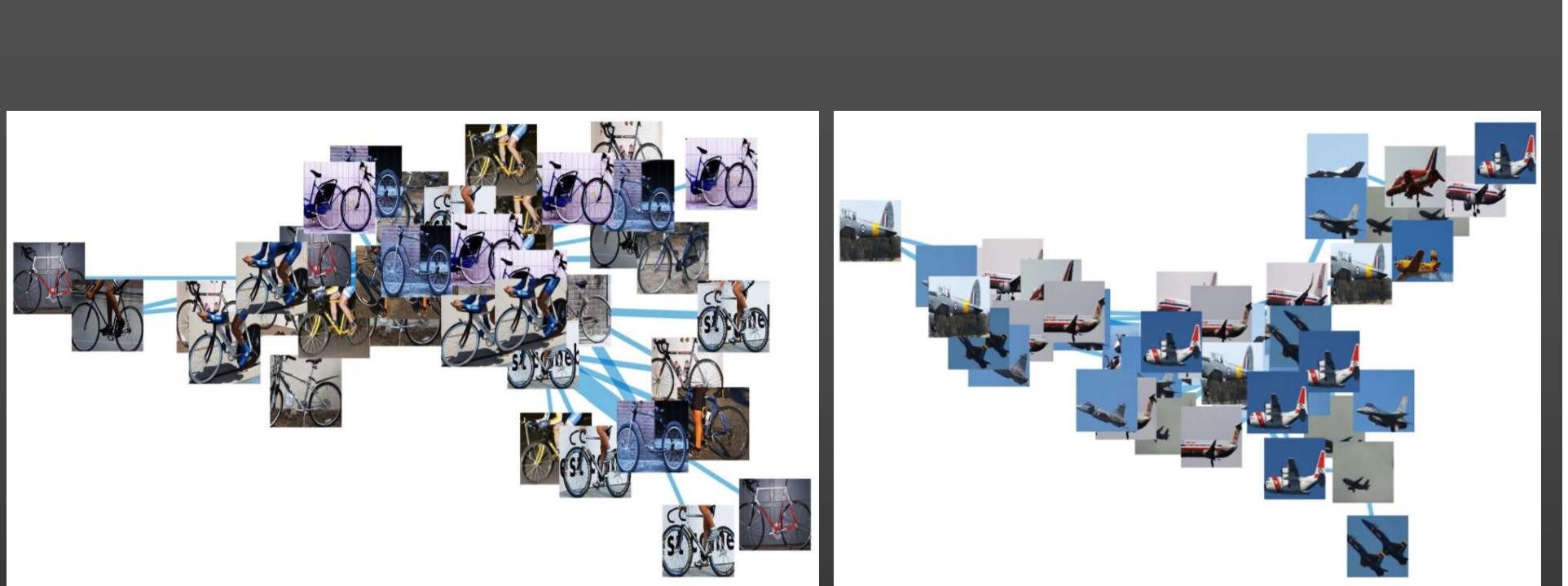
$\forall i \in 1 \dots n, \sum_{j=1}^n e_{ij} \leq \tau$

images i, j are linked e_{ij}

boxes k, l in images i, j are active S_{ij}^{kl}

similarity

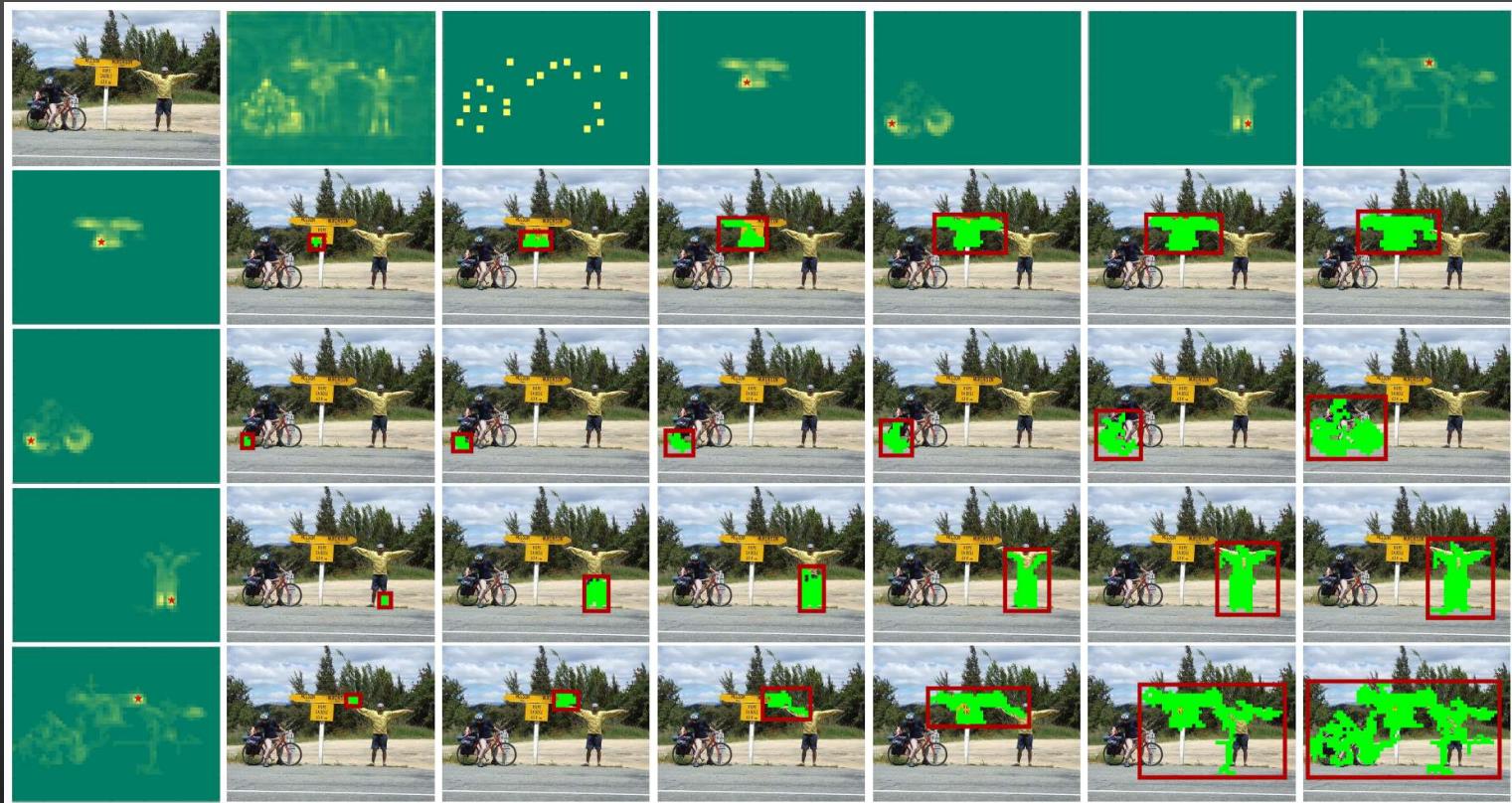
(Vo, Bach, Cho, Han, LeCun, Perez, Ponce, CVPR'19)



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

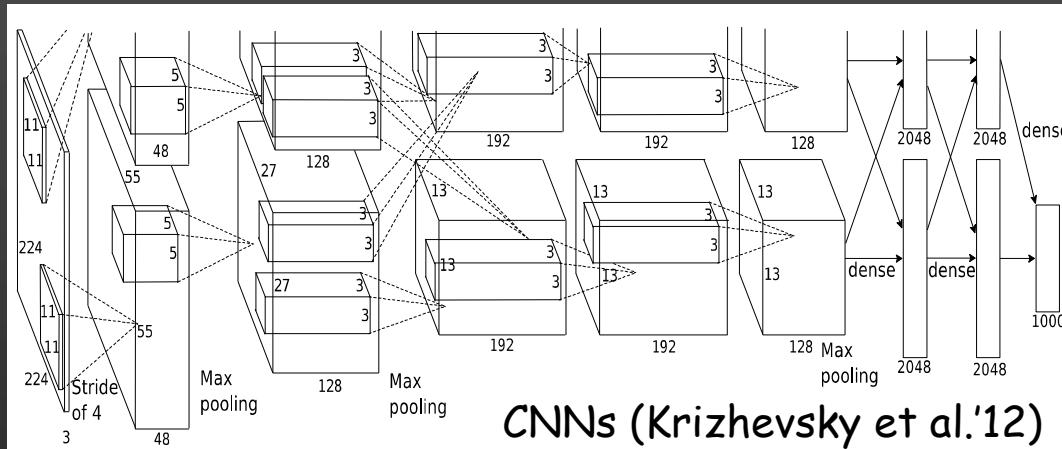
Method	VOC_all	VOC12	COCO_20k
Wei <i>et al.</i> [33]	43.4	46.2	38.6
Wei <i>et al.</i> [34]	43.4	46.3	40.5
Baseline 1	43.3 ± 0.2	40.1 ± 0.1	45.0 ± 0.1
Baseline 2	48.6 ± 0.1	49.3 ± 0.1	-
Ours	46.5 ± 0.1	46.2 ± 0.1	47.3 ± 0.1

Large-scale CorRet results

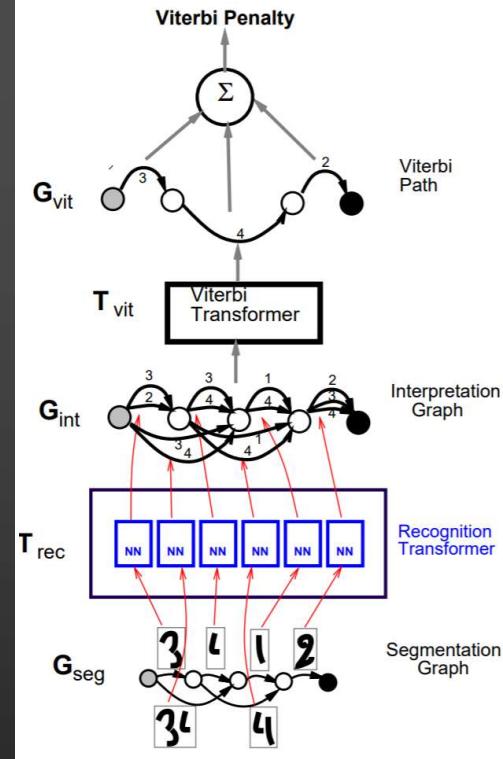
Dataset	VOC_all	VOC12	COCO_20k
Baseline	50.7	57.5	36.8
Ours	61.3 ± 0.0	64.7 ± 0.0	40.1 ± 0.0

Beyond block diagrams

Deep learning
(LeCun et al.'98)

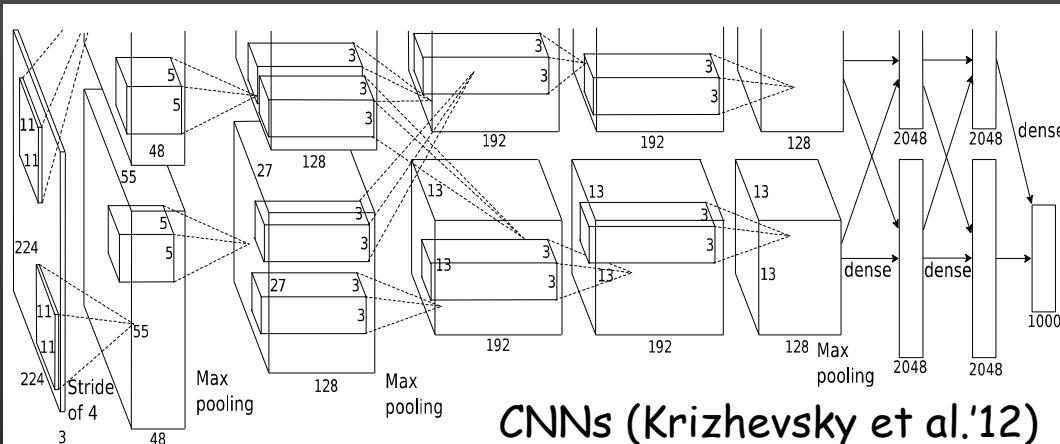


Graph transformer networks



Beyond block diagrams and pattern recognition

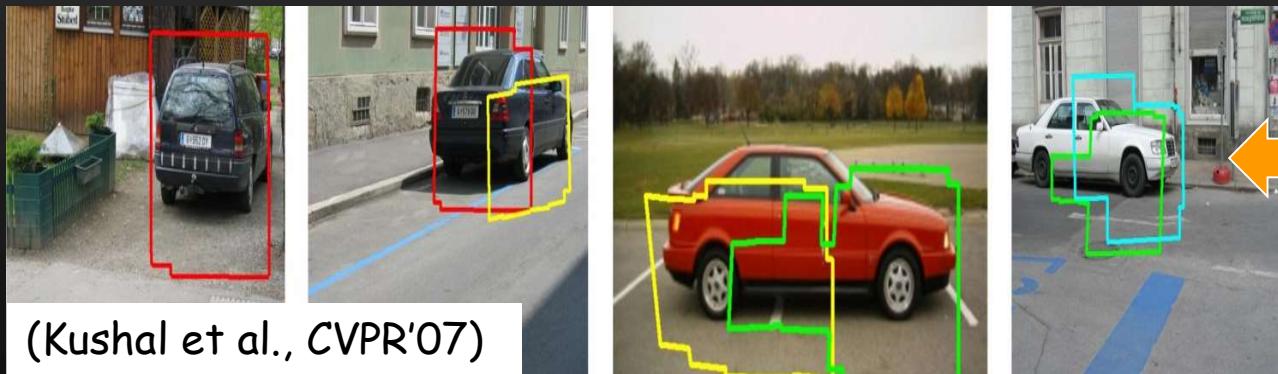
Deep learning
(LeCun et al.'98)



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
     $\hat{z} = [y; \sqrt{\mu} \varphi_{\lambda/\mu}^\theta(K \star x)];$ 
     $\hat{K} = [k_0; \sqrt{\mu} K];$ 
     $C = \operatorname{argmin}_C \|\delta - C \star \hat{K}\|_F^2 + \rho \sum_{i=0}^n \|c_i\|_F^2;$ 
     $x = \text{CPCR}(\hat{K}, \hat{z}, \psi^\nu(C), x);$ 
     $\mu = \mu + \delta_t;$ 
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!