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

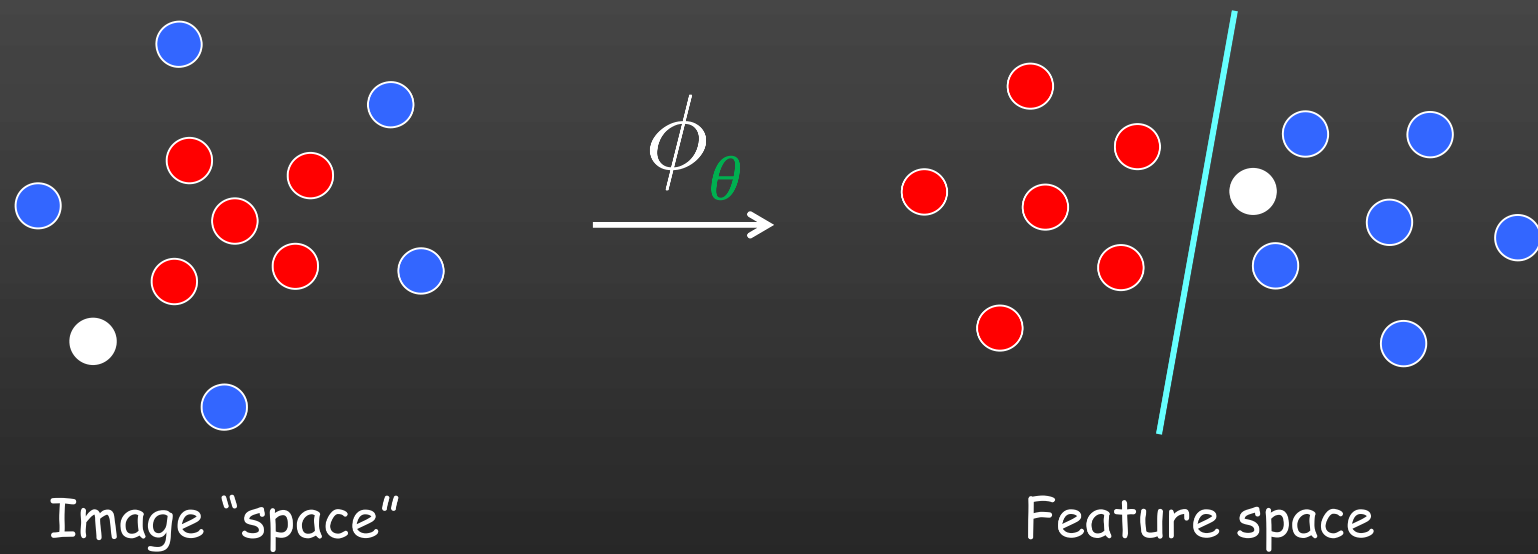
Trees

Labelled training examples

??

Test image

Image categorization as supervised classification



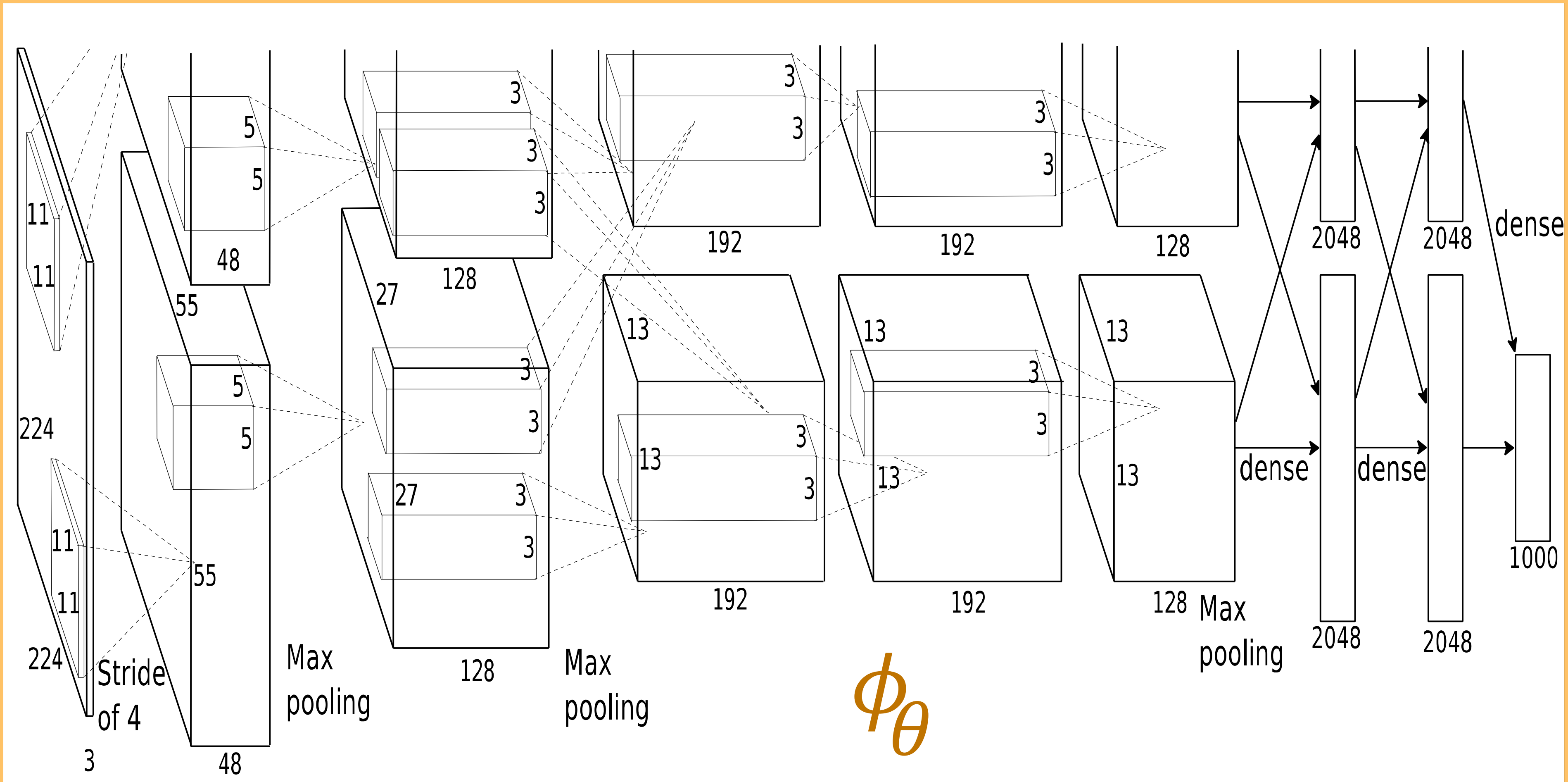
$$\min_{f \in \mathcal{F}} \frac{1}{N} \sum_n \ell(z_n, f(\phi_\theta(x_n))) + \Omega(f)$$

Training datum

Label

Prediction function

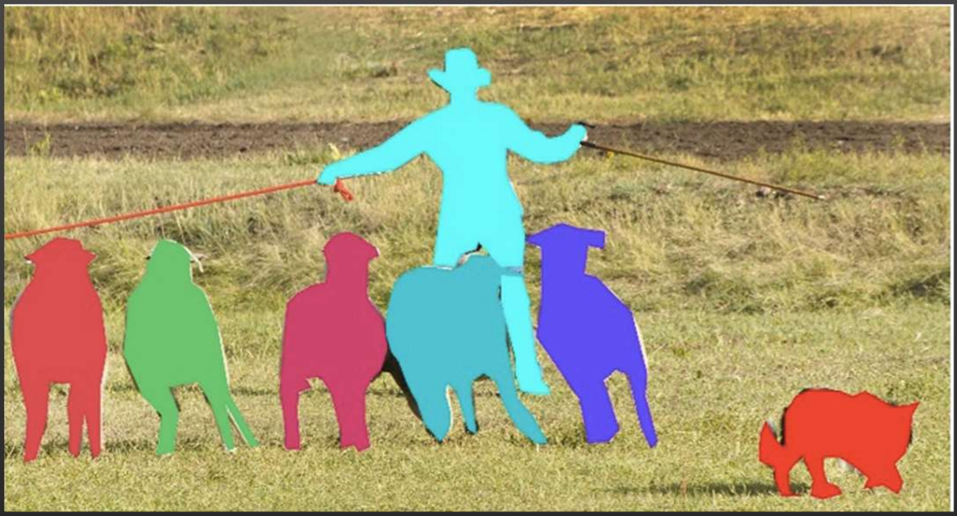
Convolutional neural networks (LeCun et al., 1998)



(Krizhevsky et al.'12)

Supervision: Where do the labels come from?

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



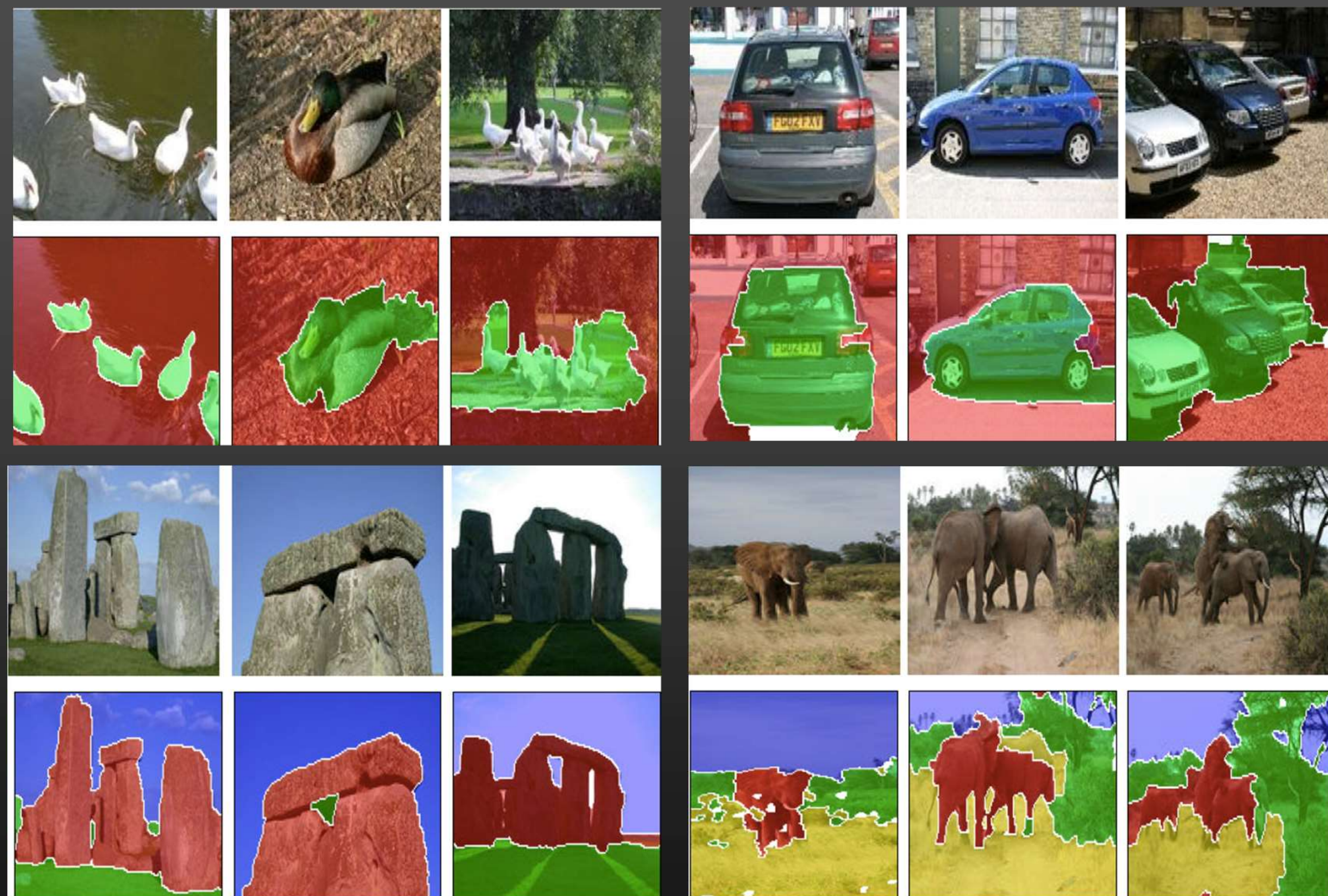
- Example: MS COCO (Lin et al., 2015)
    - 328K images of 91 object categories
    - 2.5M labelled instances

(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

Square loss

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

$$\min_{Z} \; Tr(ZZ^T A(X, \lambda \;))$$

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

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

$$\min_{\substack{y\in\{0,1\}^{N\times K},\\ y1_K=1_N}} \left[\min_{\substack{A\in\mathbb{R}^{d\times K},\\ b\in\mathbb{R}^K}} E_U(y,A,b) \right] + E_B(y) - H(y)$$

Softmax

Spectral clustering term

Entropy term

Discriminative clustering term

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

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

$$\min_{Z} \; Tr(ZZ^T A(X,\lambda \;))$$

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

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

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

Optimization:

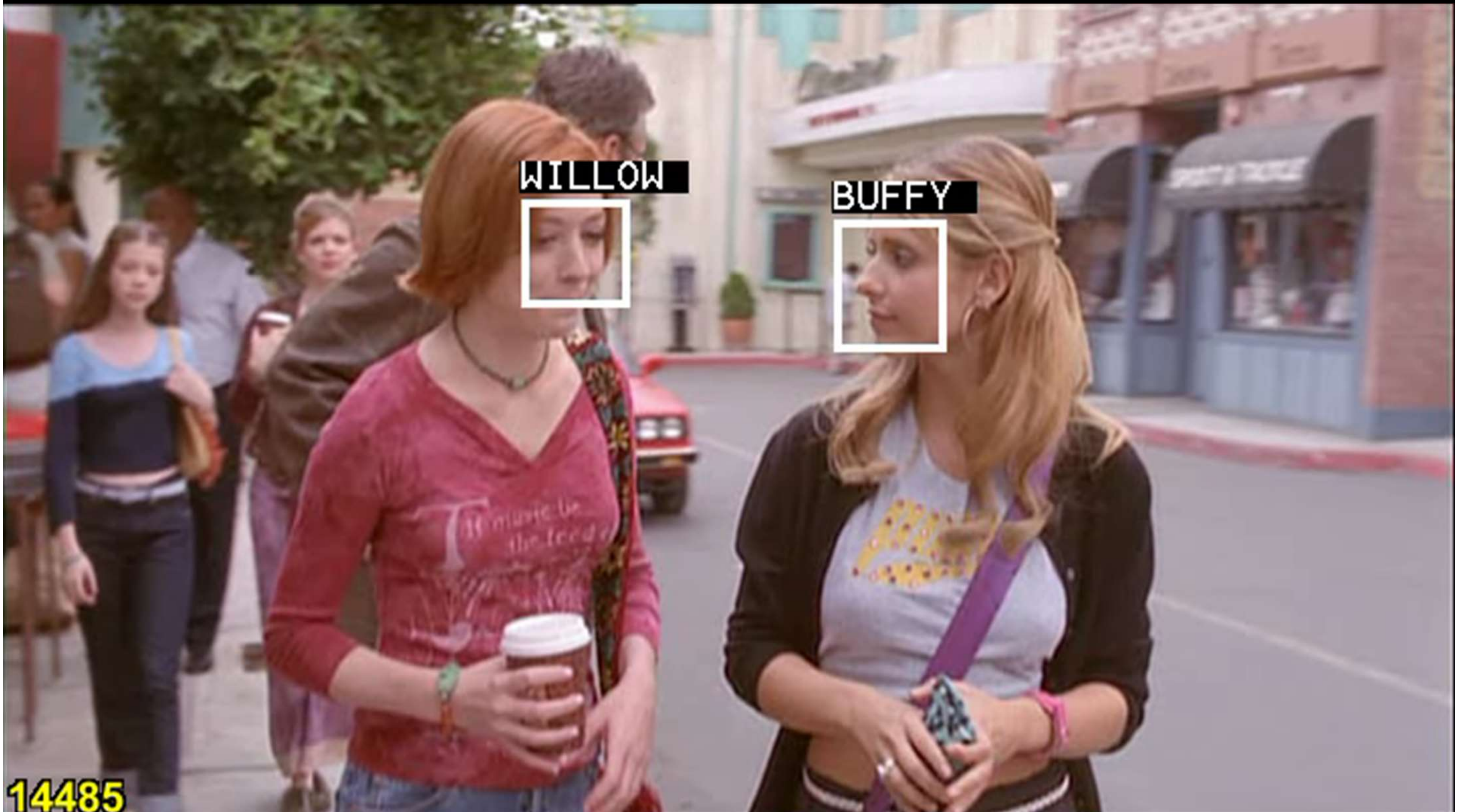
- Relax to continuous problem
- EM/block-coordinate descent procedure with quasi-Newton and projected gradient descent for the two steps, initialized with quadratic approximation
- Round up the solution

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

Multi-class cosegmentation results

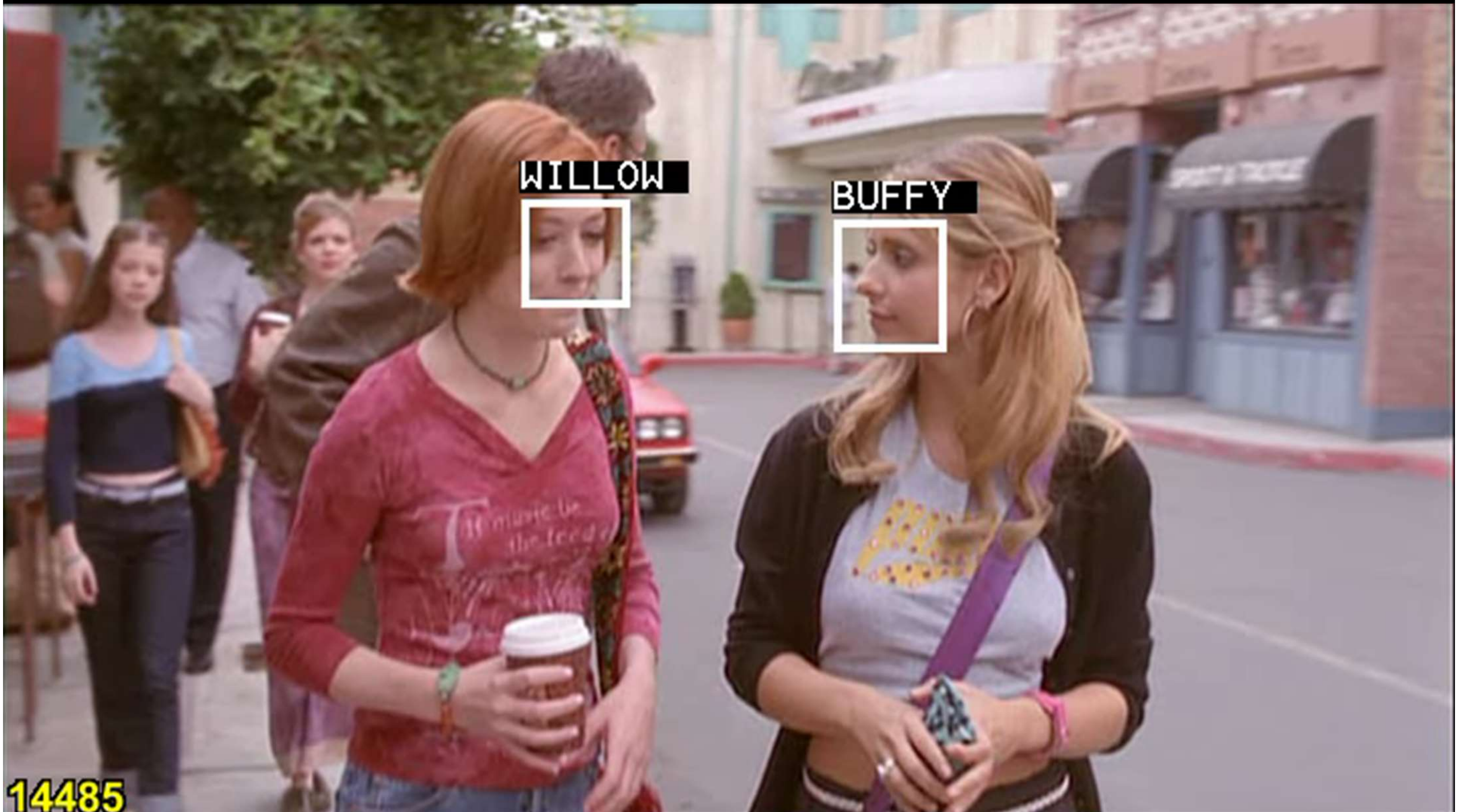
(Joulin et al., CVPR'12)

Naming the characters of TV series



(Sivic, Everingham, Zisserman, 2009)

TV series come with their own metadata

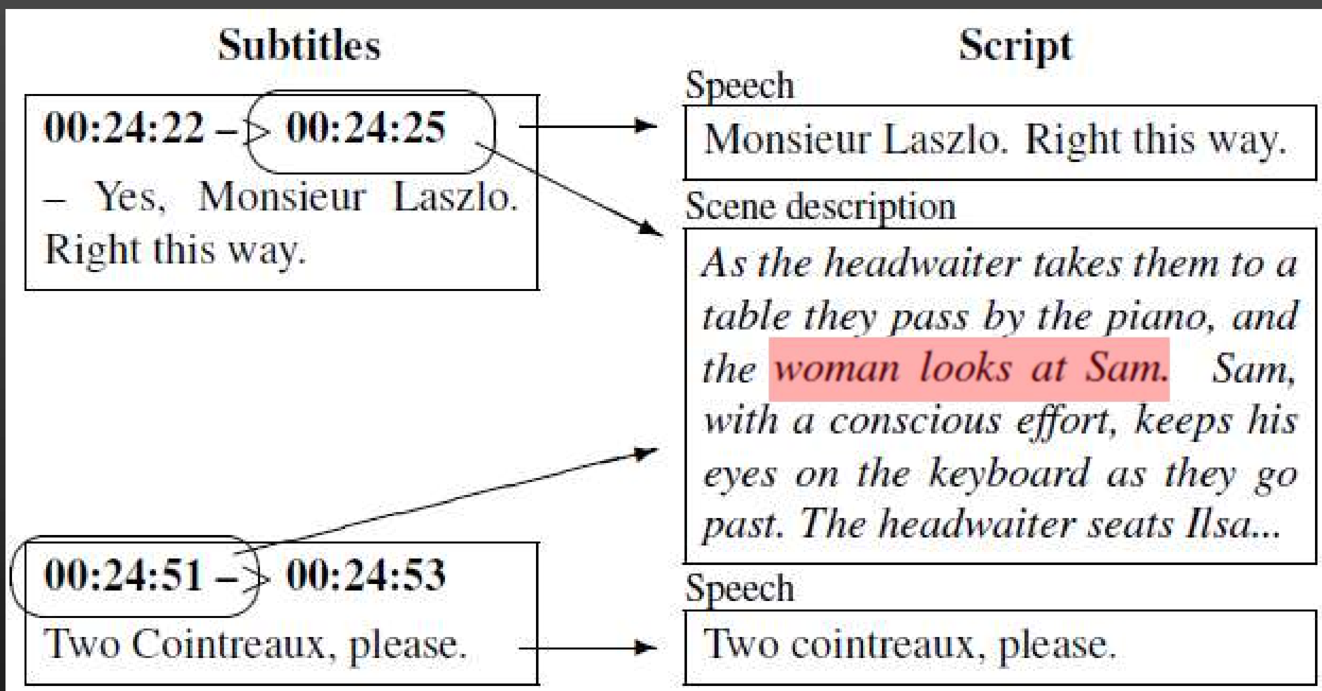


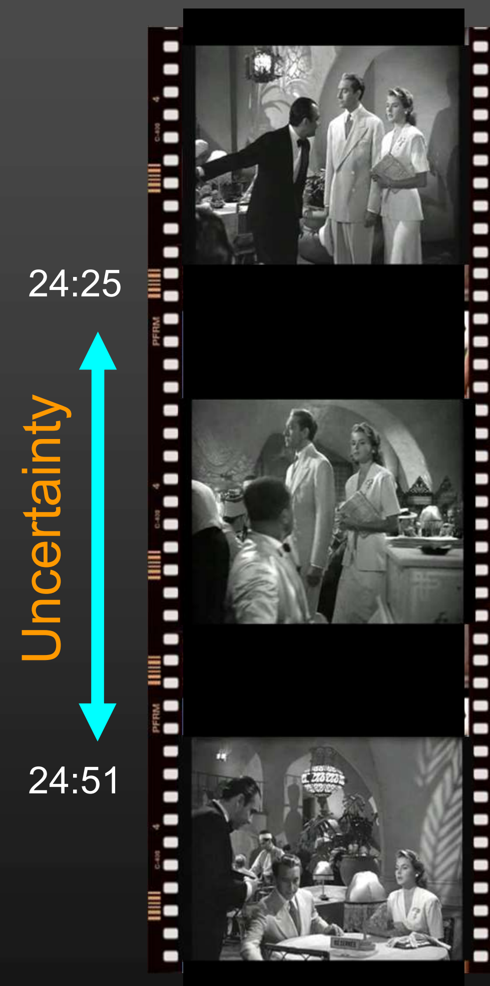
(Sivic, Everingham, Zisserman, 2009)

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...

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

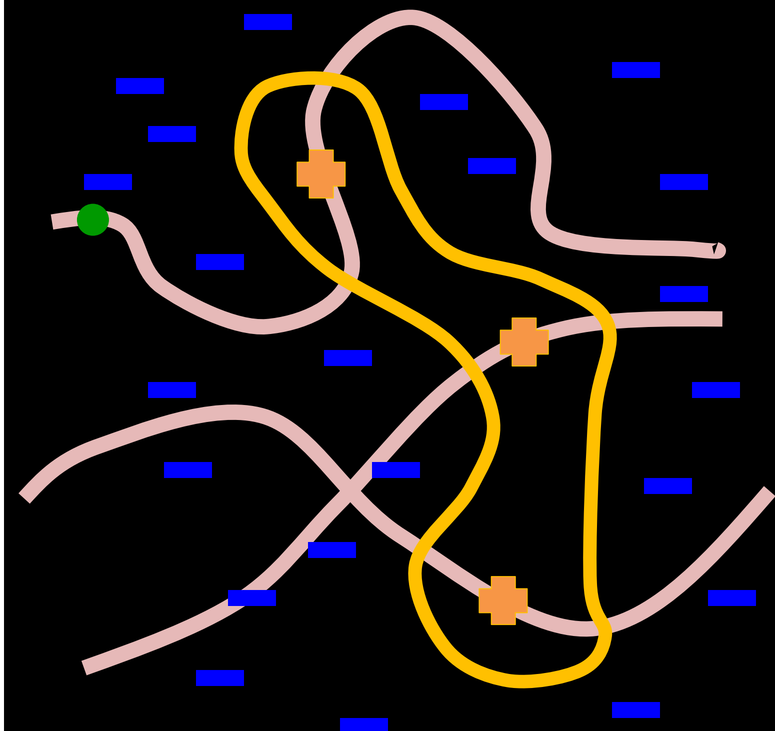
Output: temporal action clusters

... **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

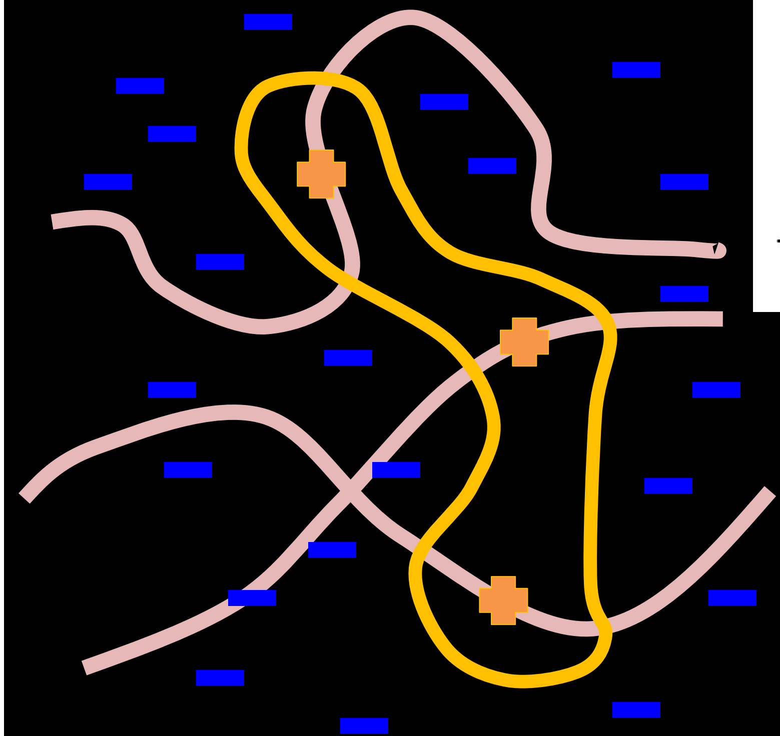
Negative samples

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



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

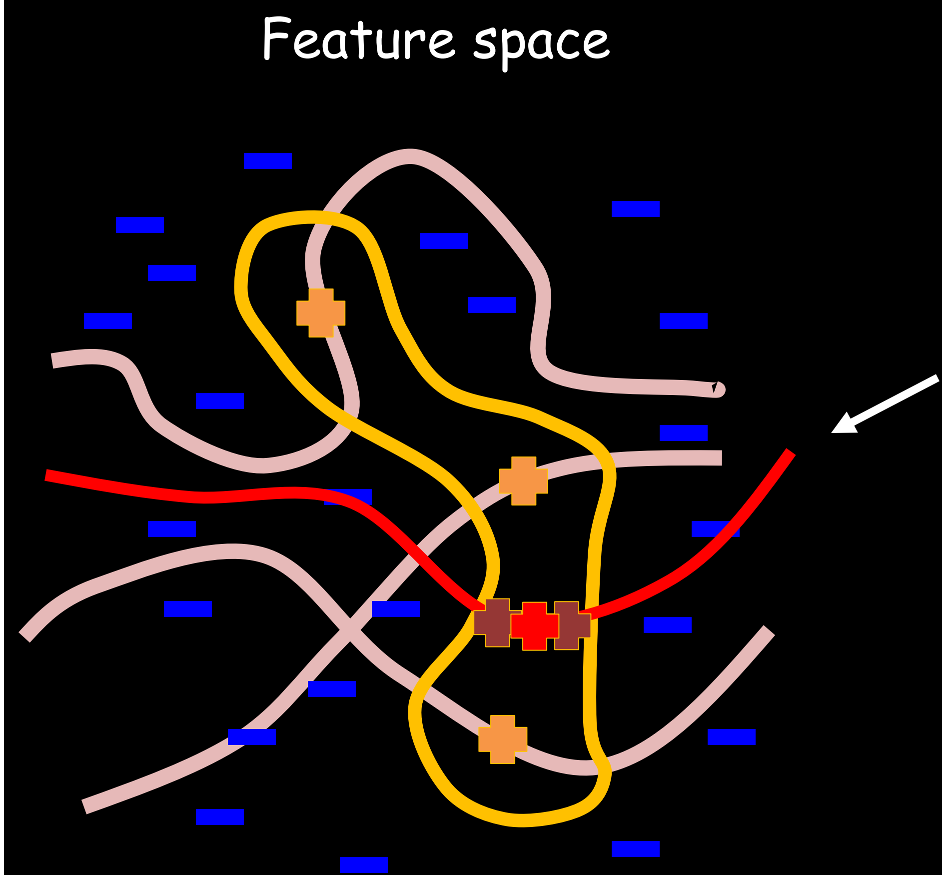
Optimization: Block-coordinate descent

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



New video

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

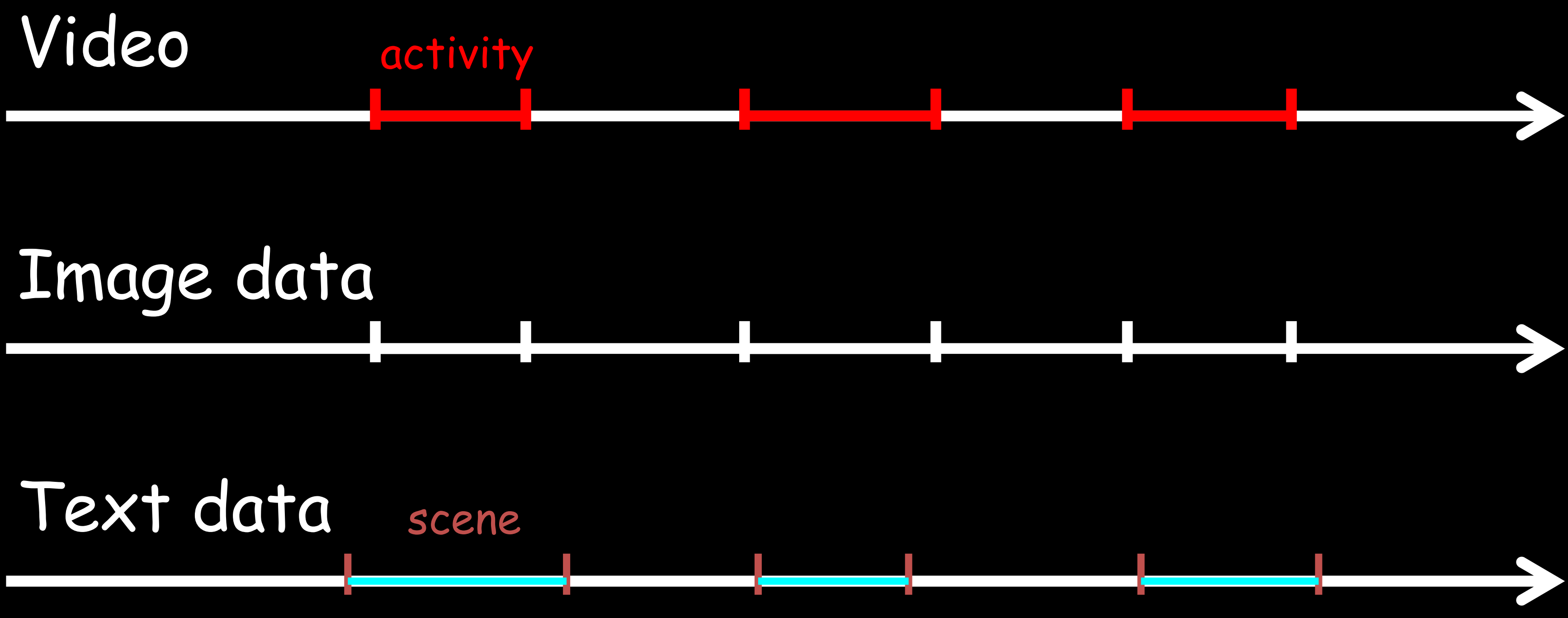
Automatic Annotation of Human Actions in Video

ICCV 2009 DEMO

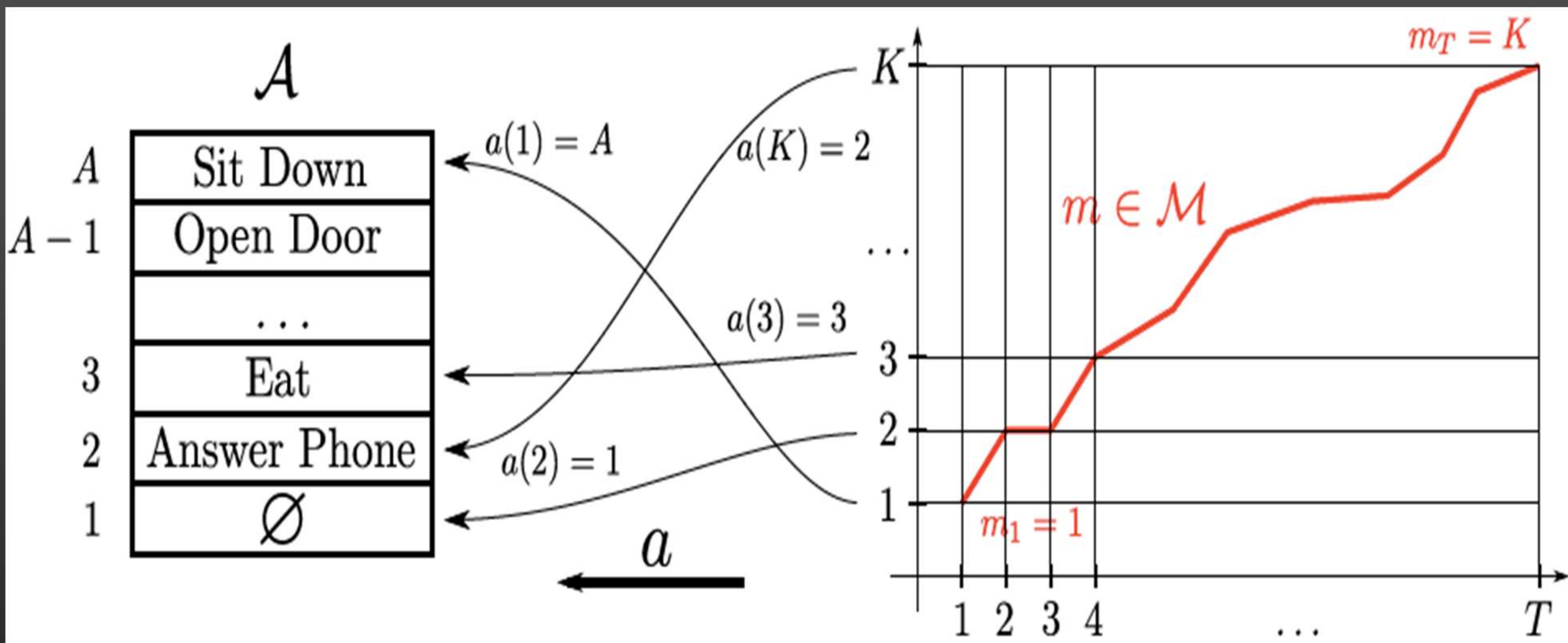
O.Duchenne, I.Laptev, J.Sivic, F.Bach and J.Ponce

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)

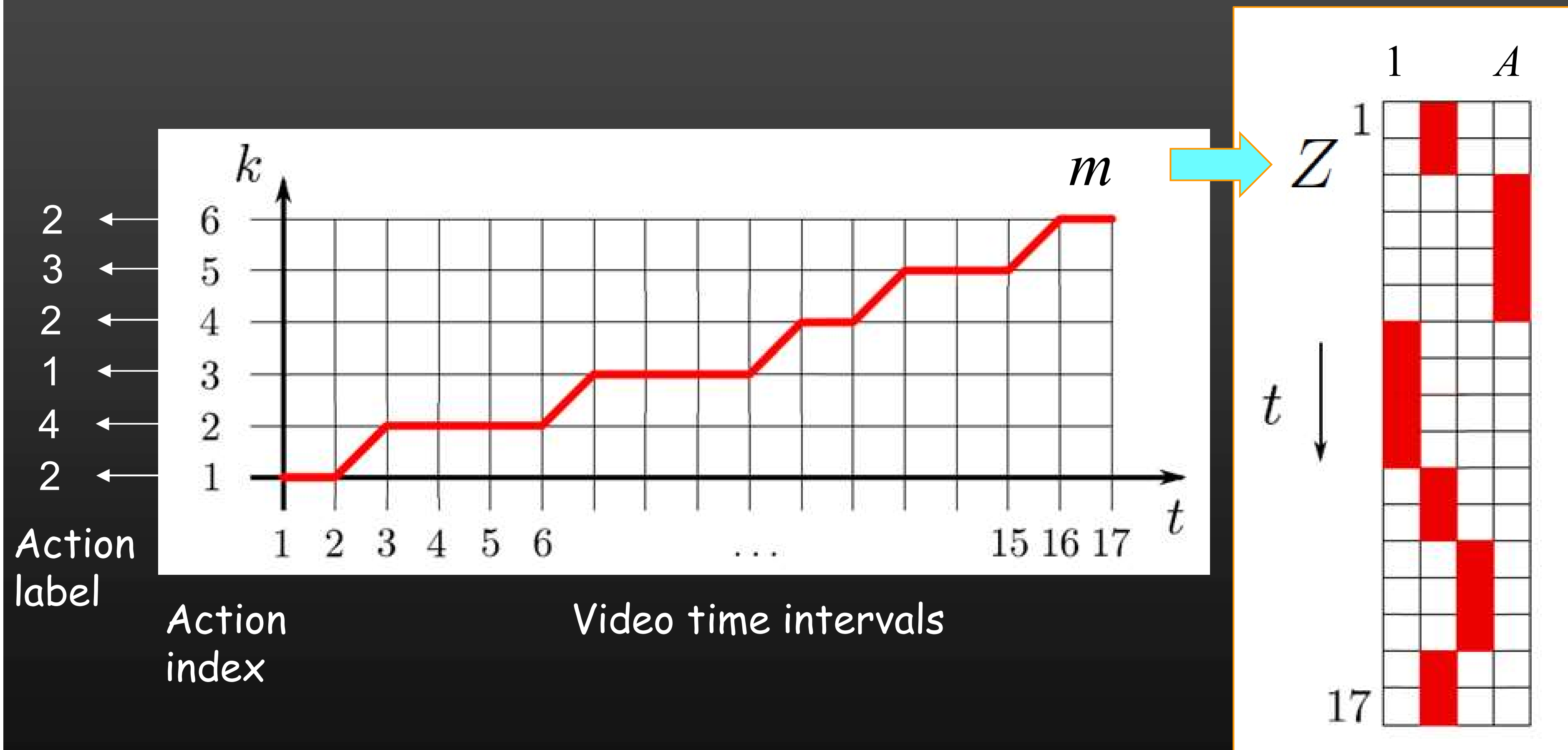


Dictionary | Script metadata a | Alignment m

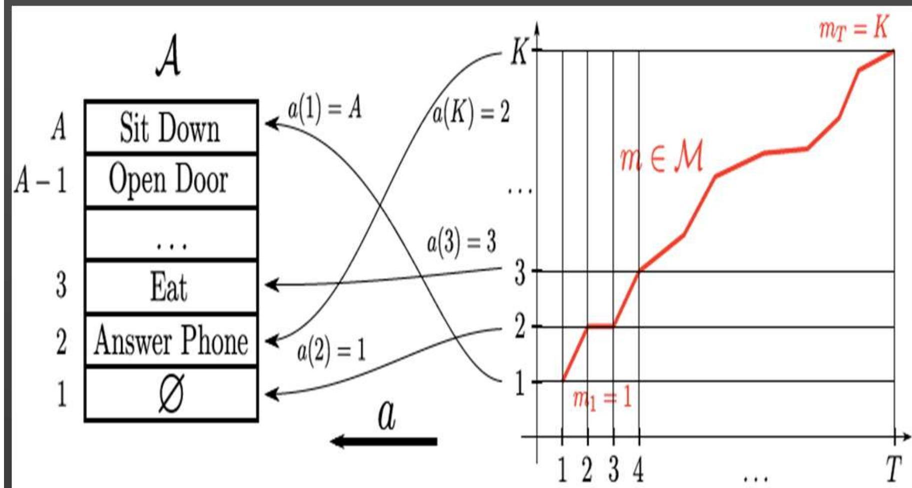
$$\min_{f \in \mathcal{F}} \left[\sum_{n=1}^{N} \min_{m \in \mathcal{M}} \frac{1}{T} \sum_{t=1}^{T} \ell\left(a_n(m_t), f(x_n(t))\right) \right] + \lambda \Omega(f)$$

Changing the representation

$a, m \rightarrow Z$



Action labeling under ordering constraints



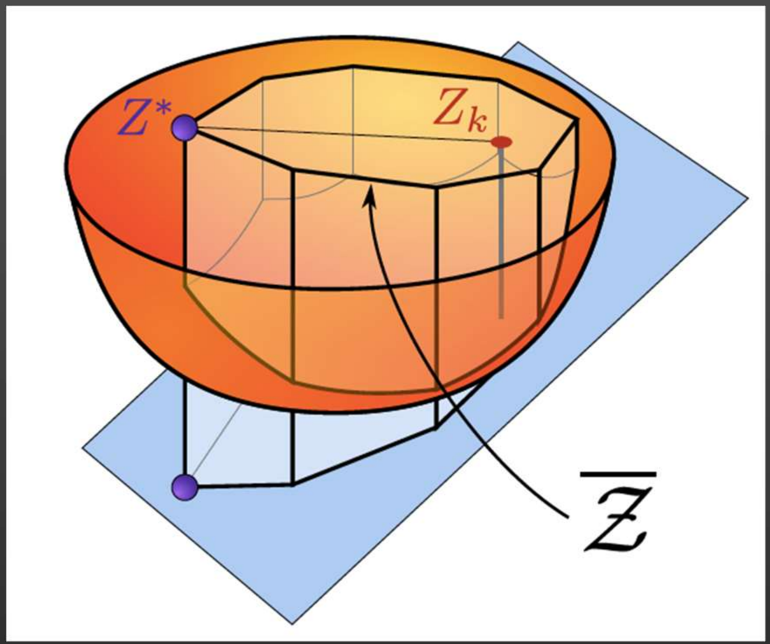
$$\min_{f\in\mathcal{F}} \left[\sum_{n=1}^{N} \min_{m\in\mathcal{M}} \frac{1}{T} \sum_{t=1}^{T} \ell\left(a_n(m_t), f(x_n(t))\right) \right] + \lambda \Omega(f)$$

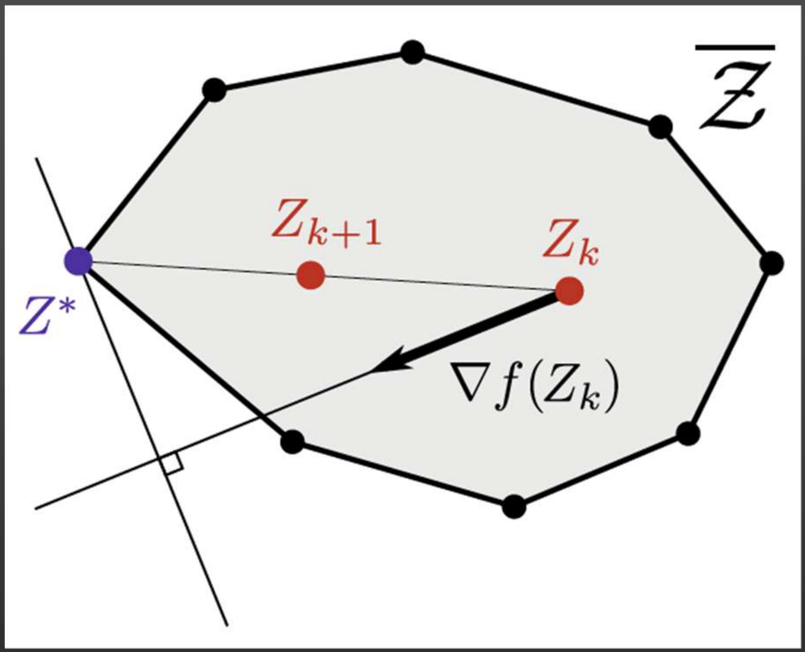
$$\min_{f\in\mathcal{F}, Z\in\mathcal{Z}} \frac{1}{T} \sum_{t=1}^{T} \ell(Z_t, f(x_t)) + \lambda\Omega(f) = \frac{1}{T}\|Z - XW - b\|_F^2 + \frac{\lambda}{2}\|W\|_F^2$$

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

- Minimize a convex quadratic function over a large discrete domain $\mathbb{Z}$
- Relaxed problem: minimize instead over $\overline{\mathbb{Z}}$=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 $\overline{\mathbb{Z}}$
- $Z_{k+1} = (1-\gamma)\, Z_k + Z^*$

- No need for a projection step, converges to global minimum
- DP can be used to minimize linear functions over $\mathbb{Z}$ and thus $\overline{\mathbb{Z}}$
- DP can also be used for rounding

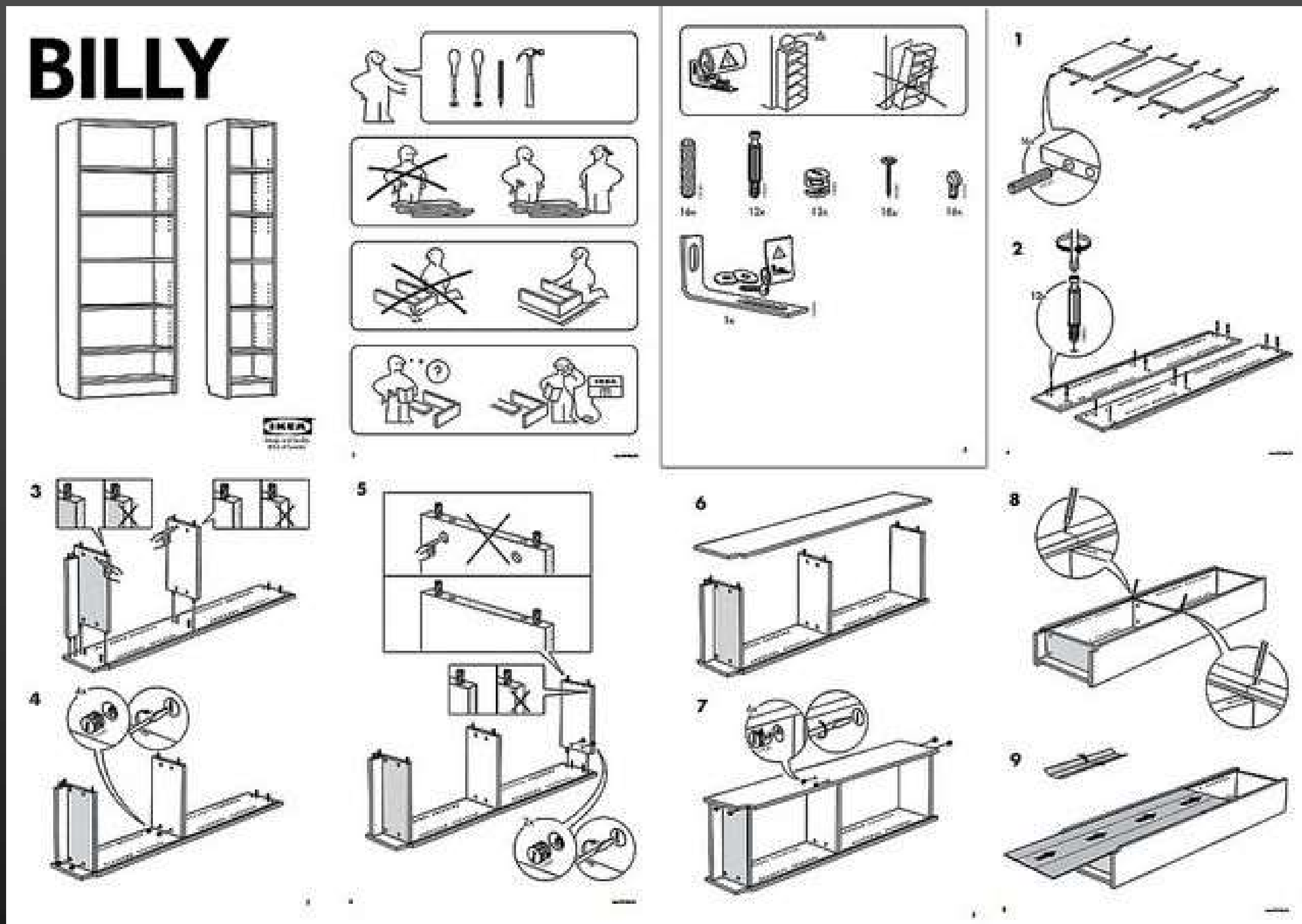
Temporal action localization

Clip number 0101

(Bojanowski et al., ECCV'14)

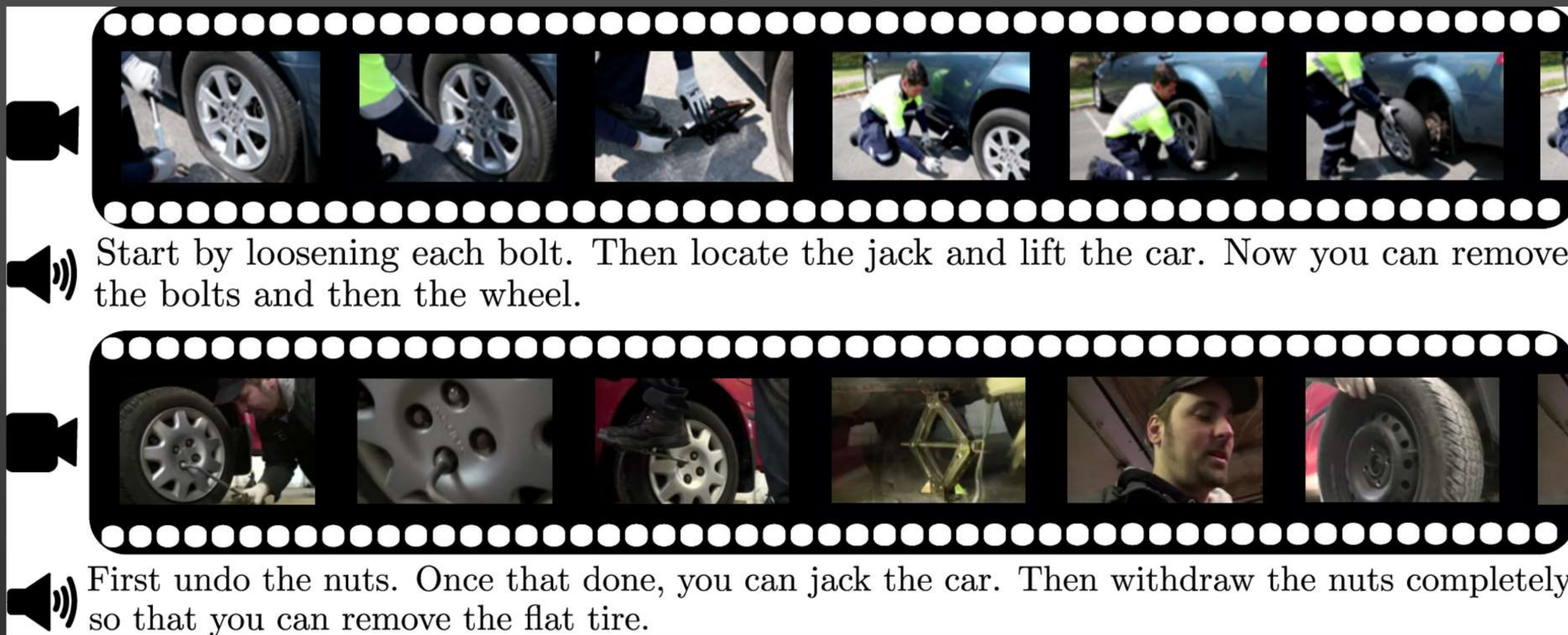
Learning from narrated instructional videos (Alayrac et al., CVPR'16, PAMI'17)

Making assembly plans

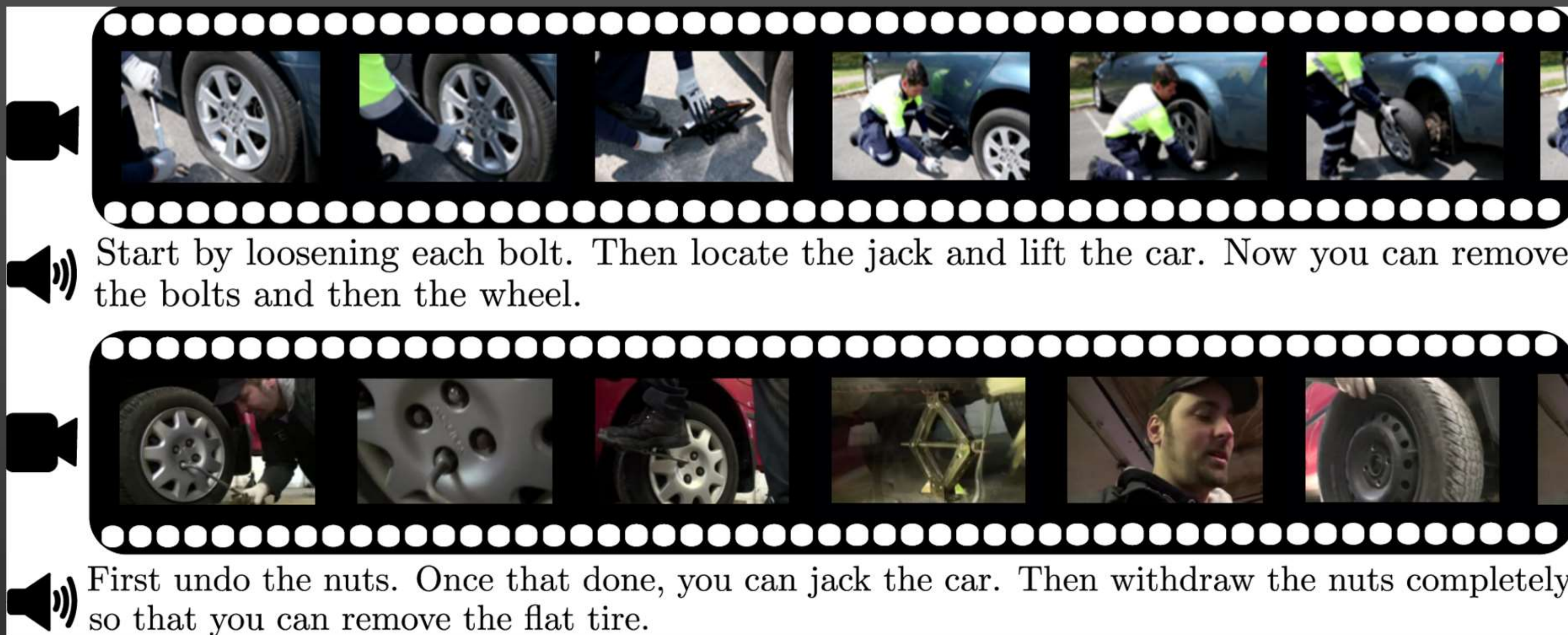


Automatically produce a sequence of instructions from narrated videos

Input: a set of narrated videos and their text transcriptions



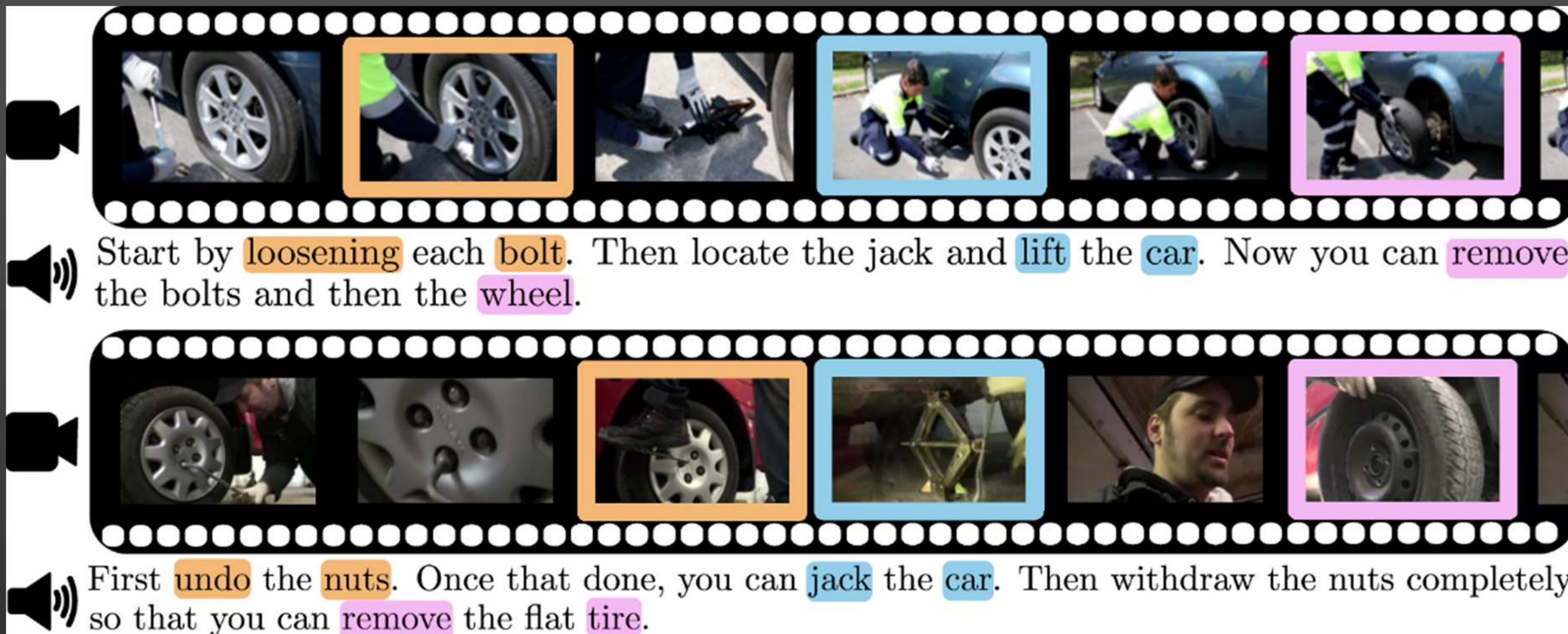
Input: a set of narrated videos and their text transcriptions



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

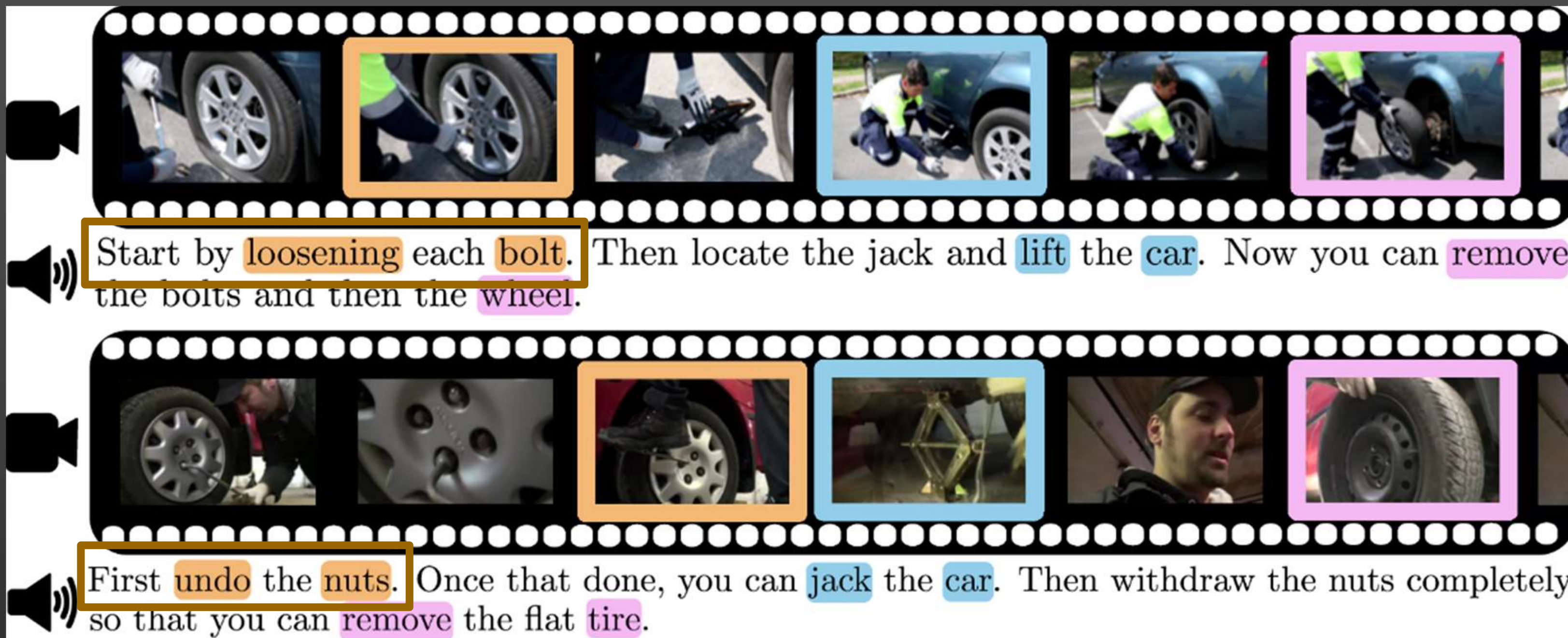


Output:

- 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	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
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
∅	loosen nut	loosen nut	undo bolt
jack car	raise car	jack car	lift car
∅	∅	unscrew nut	∅
remove wheel	remove tire	withdraw tire	lower ∅ car
∅	lower jack	∅	

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
∅	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

$$C = c(\varnothing, \text{`loosen nut'}) + \cdots + c(\varnothing, \text{`undo bolt'})$$
$$+ \cdots + c(\text{`jack car'}, \text{`raise car'})$$
$$+ \cdots + c(\varnothing, \text{`unscrew nut'}) + \cdots$$

Text alignment in multiple sequences

Alignment cost

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

Sum over all pairs

Sum over template lines

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

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

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

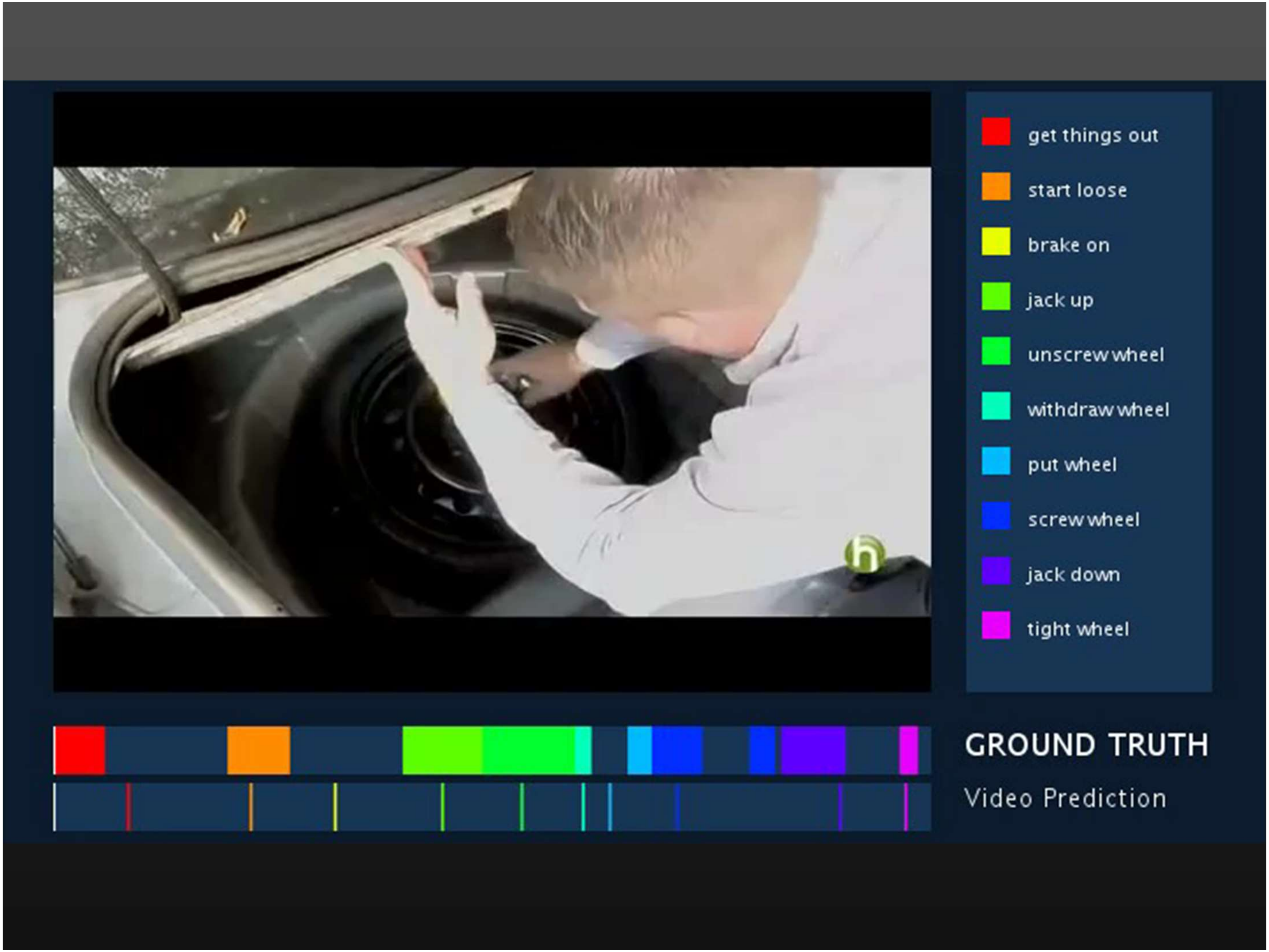
Text alignment in multiple sequences

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

Discovered list of steps

1) Loosen nut
2) Jack car
3) Remove wheel

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

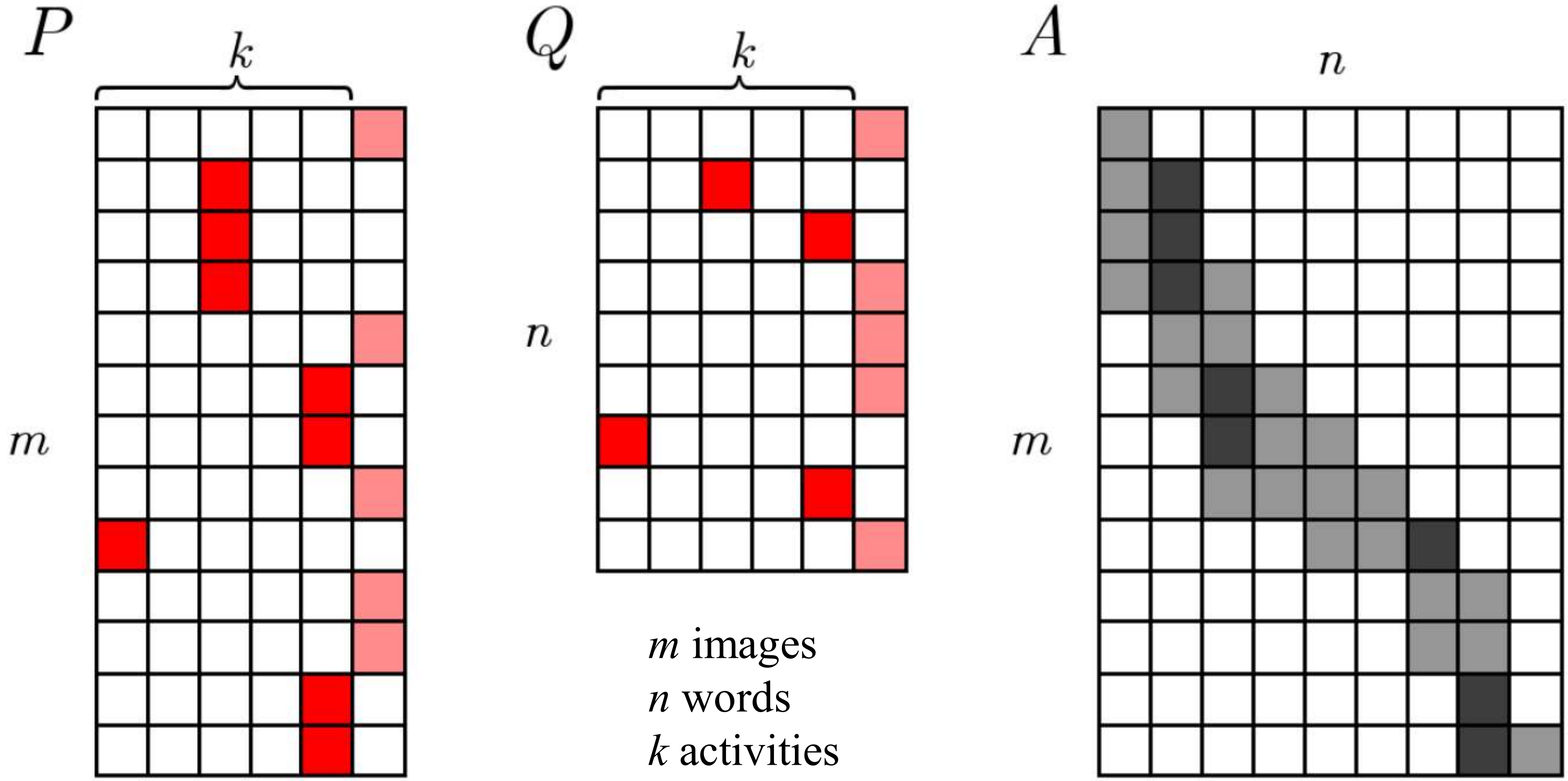


get things out
start loose
brake on
jack up
unscrew wheel
withdraw wheel
put wheel
screw wheel
jack down
tight wheel
GROUND TRUTH
Video Prediction

Activity discovery from images and words

(Alakuijala, Mairal, Ponce, Schmid, 2019)

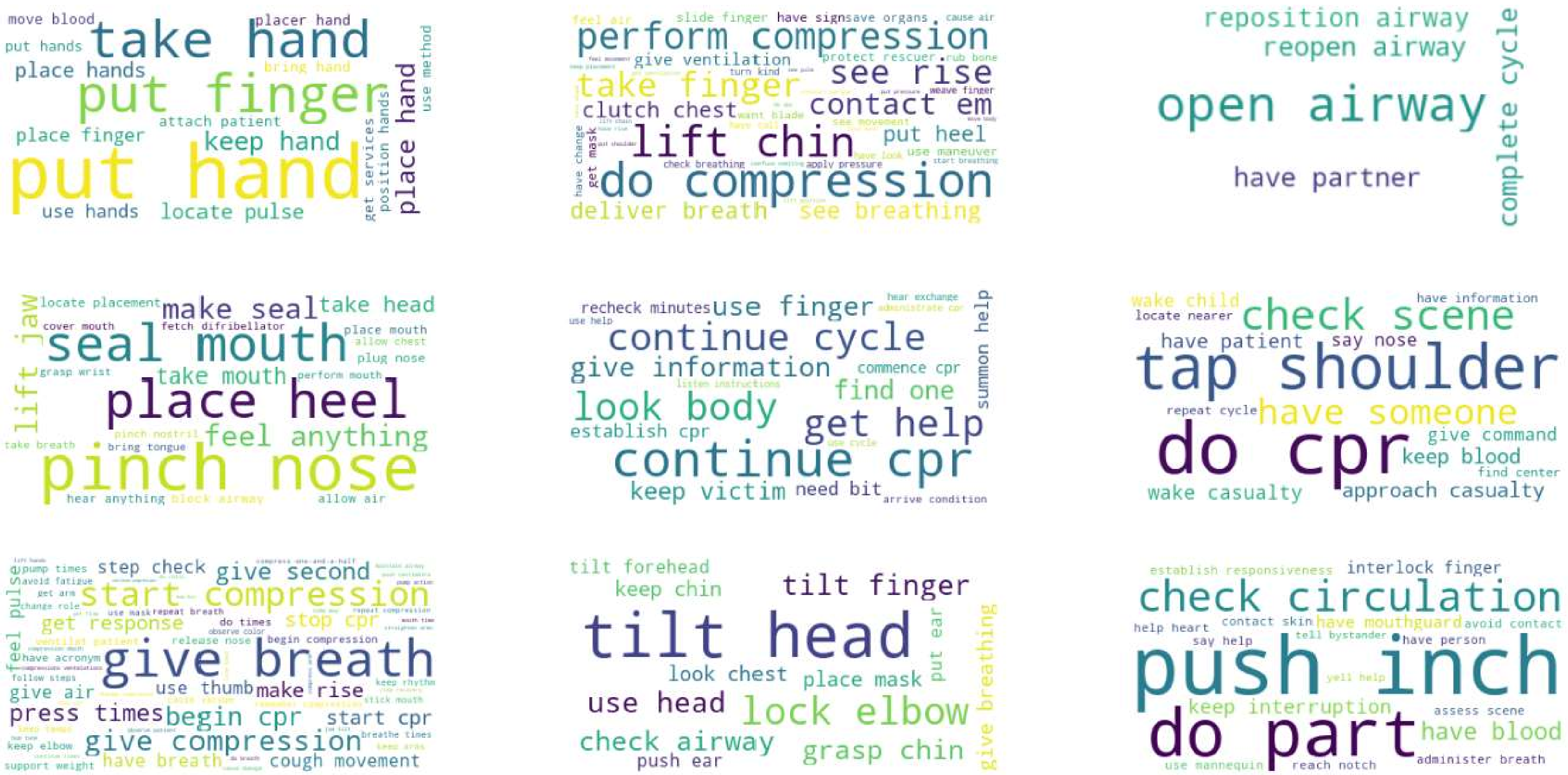
Activity discovery from images and words



m images
n words
k activities

(Alakuijala, Mairal, Ponce, Schmid, 2019)

Discovered ingredients of "performing CPR"



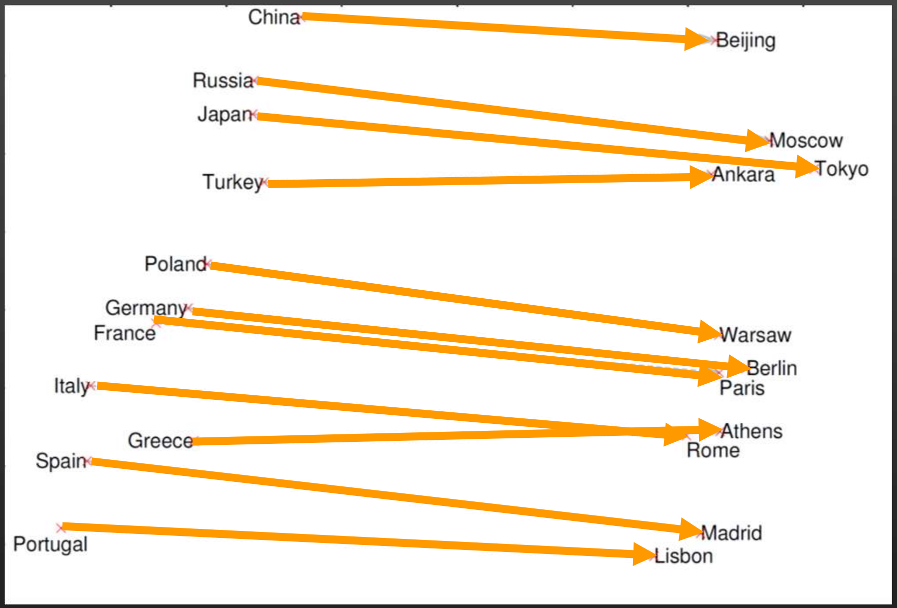
How much supervision do we really need?
(Cho et al., CVPR'15)

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): $w \to u(w) \in R^d$

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

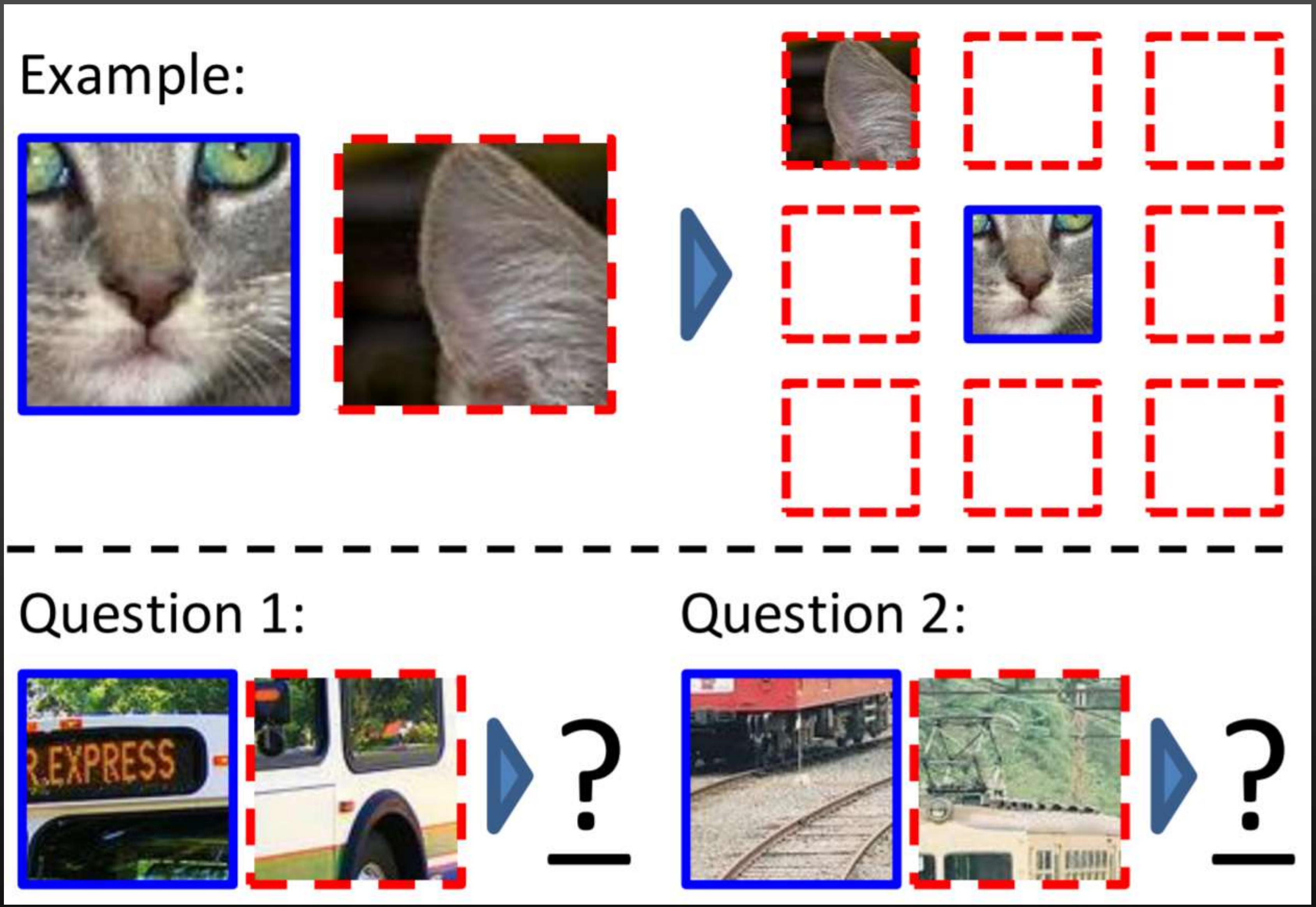


Note: Visualization in 2D but $d \approx 300$

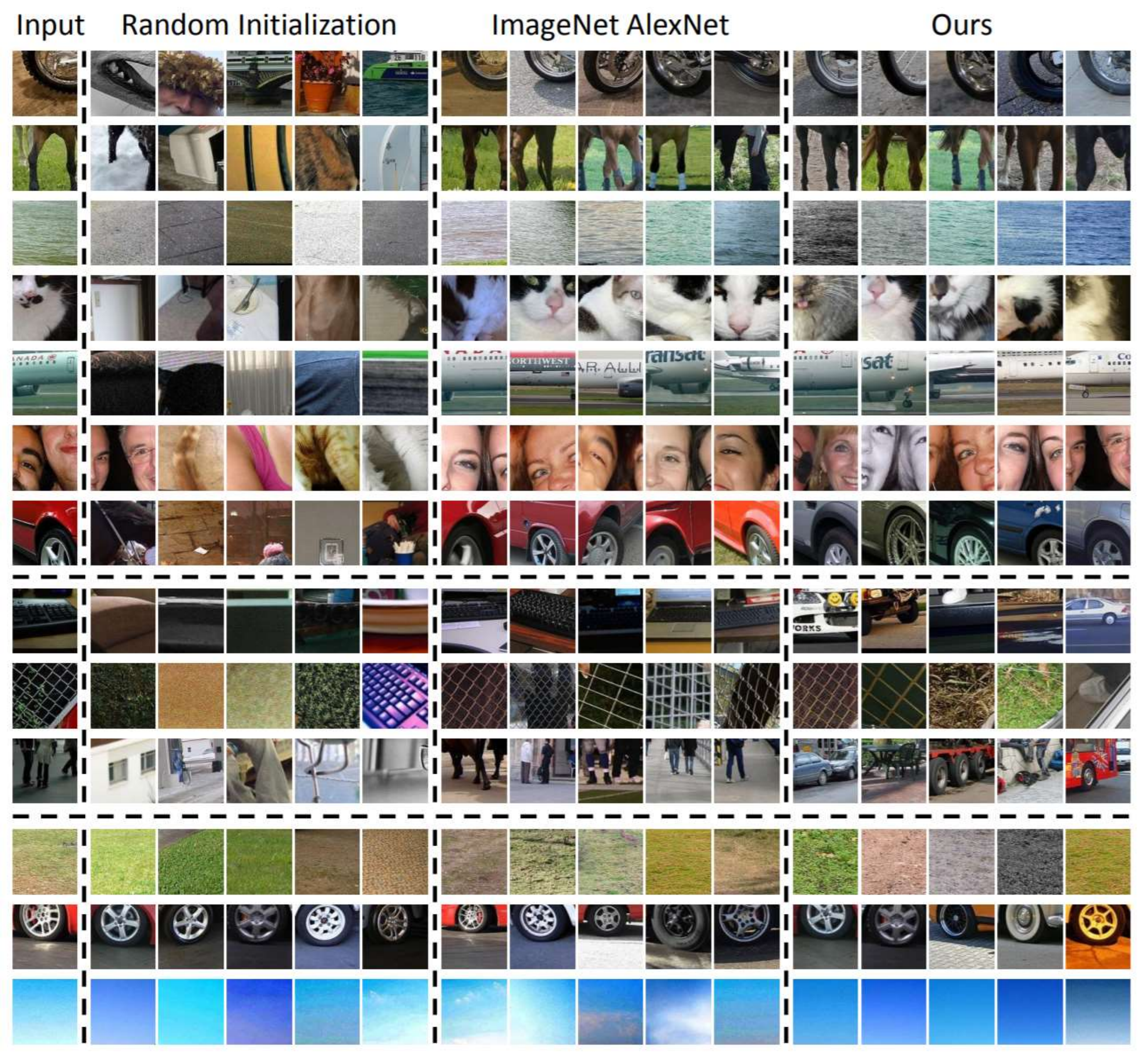
Modeling contextual info with co-occurrence statistics

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

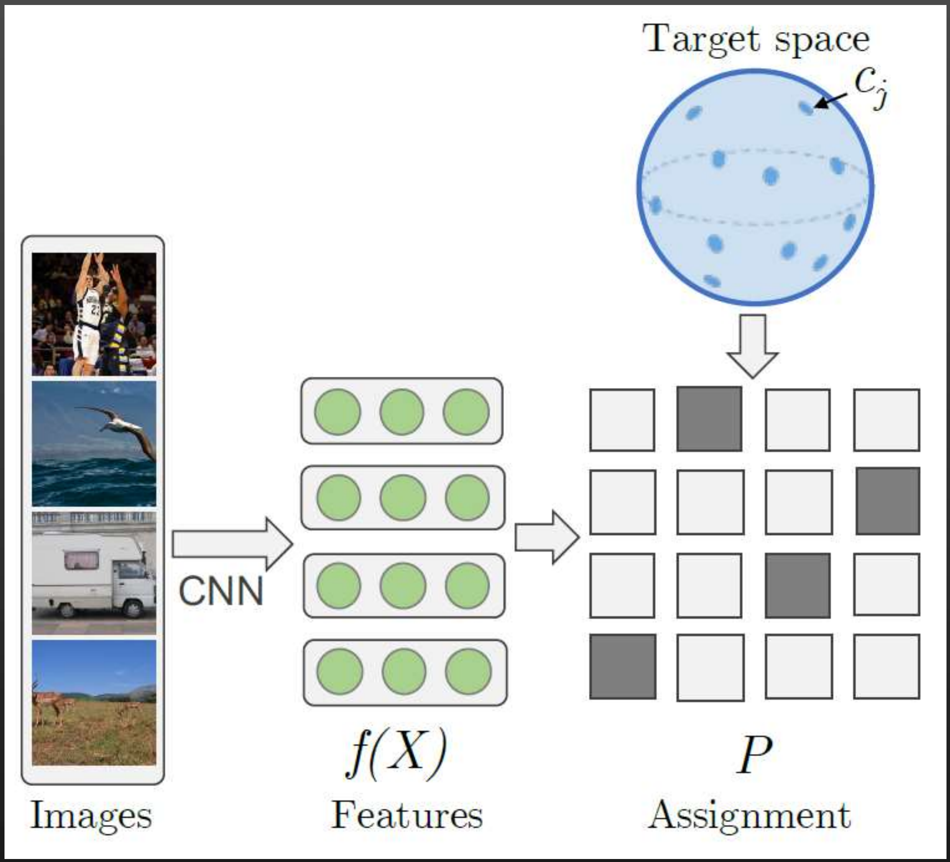
Example: Unsupervised Visual Representation Learning by Context Prediction (Doersch, Gupta, Efros, 2016)



Retrieved nearest neighbors



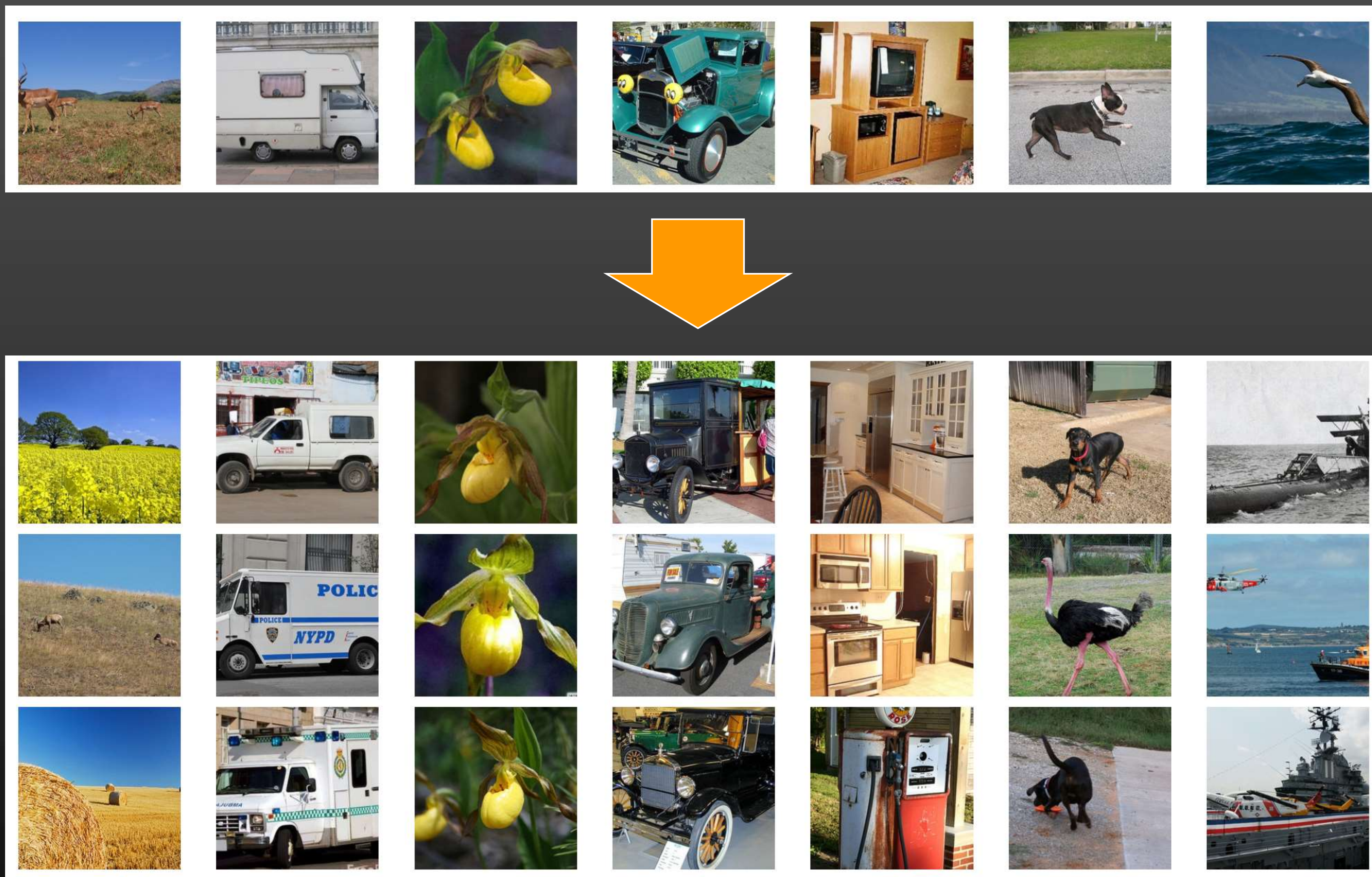
Unsupervised feature learning (Bojanowski & Joulin, ICML'17)

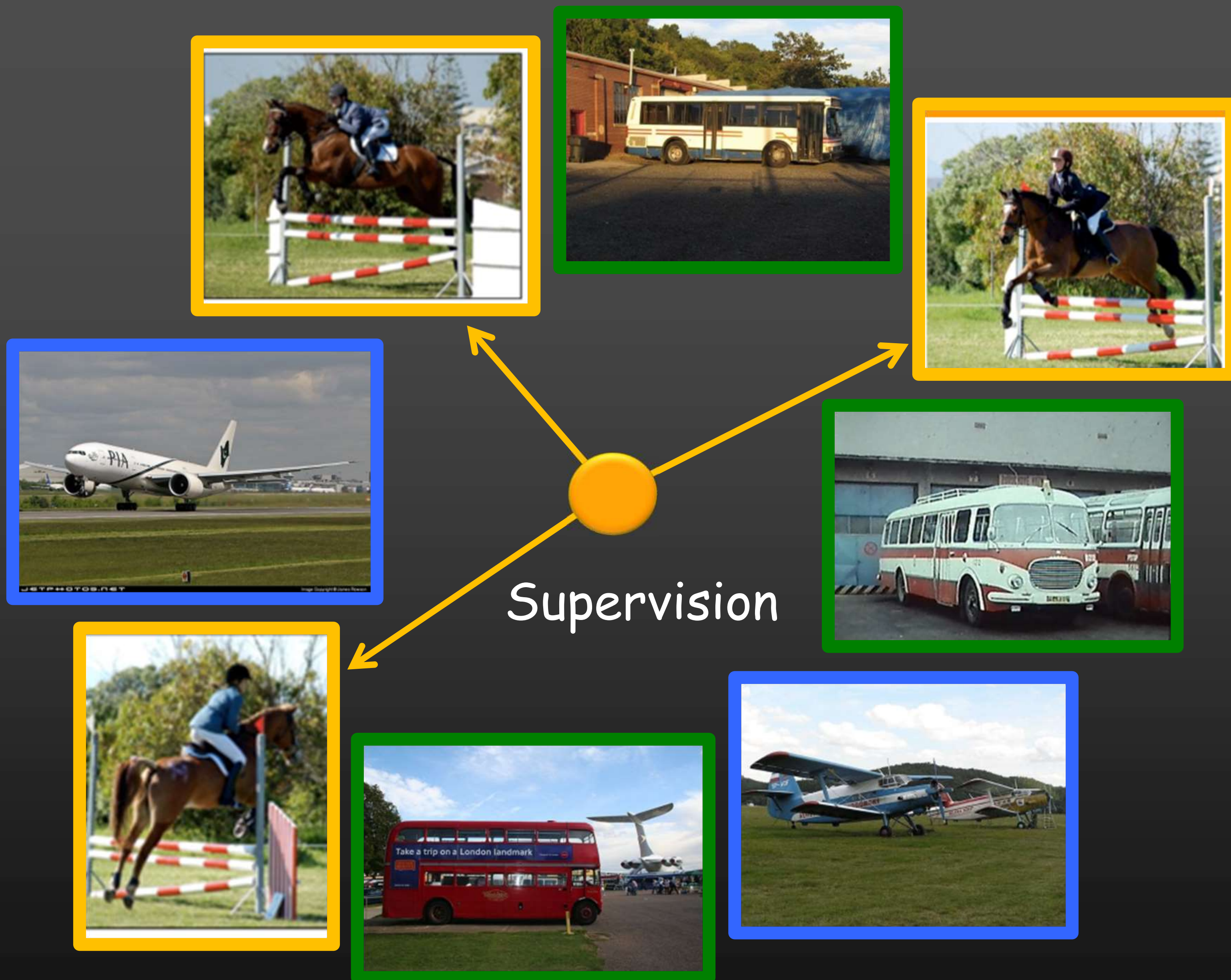


$$\min_{\theta} \min_{Y \in \mathbb{R}^{n \times d}} \frac{1}{2n} \| f_\theta(X) - Y \|_F^2$$

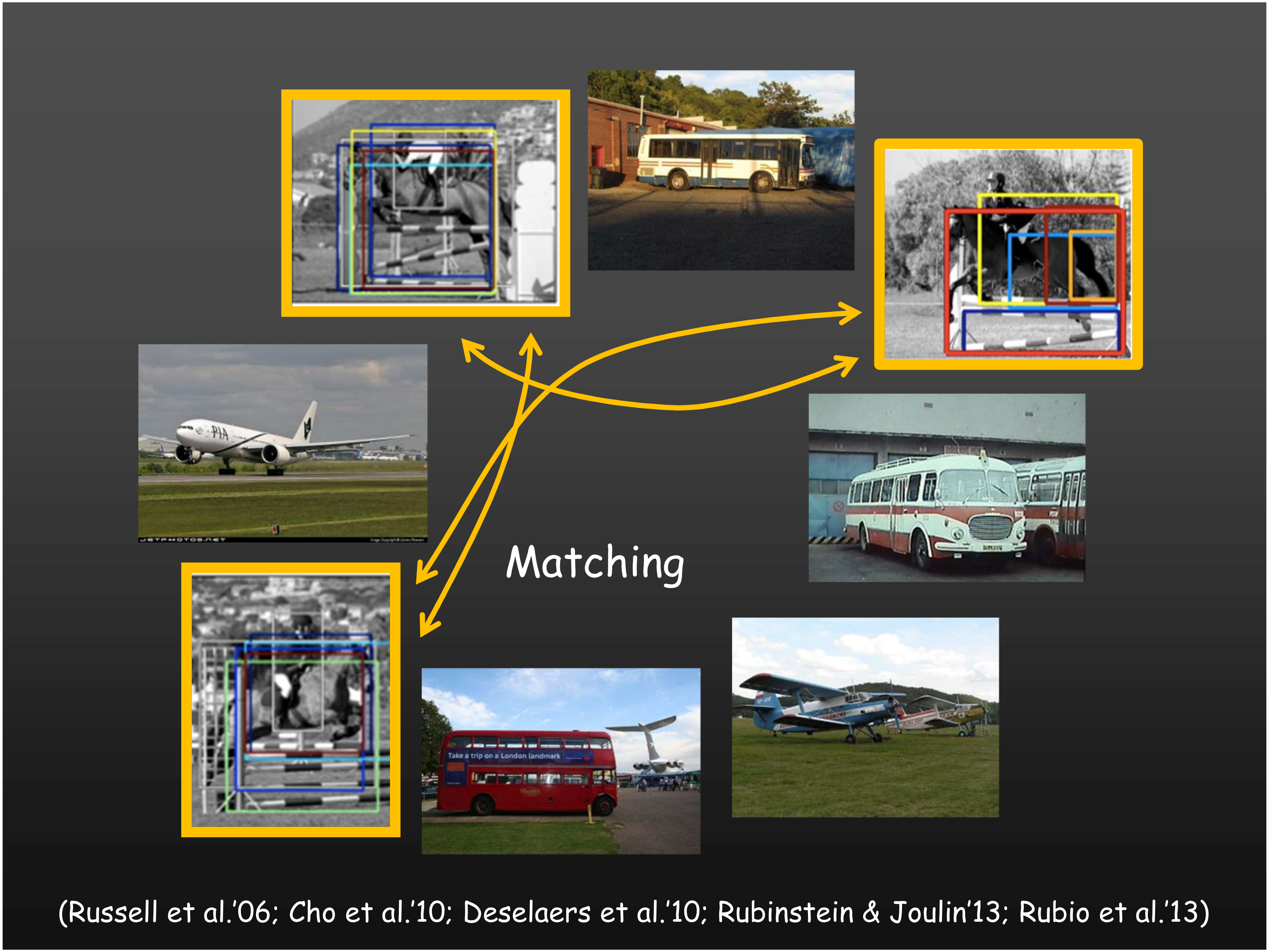
$$\max_{\theta} \max_{P \in \mathcal{P}} \mathrm{Tr}\left(P C f_\theta(X)^\top \right)$$

Retrieved nearest neighbors

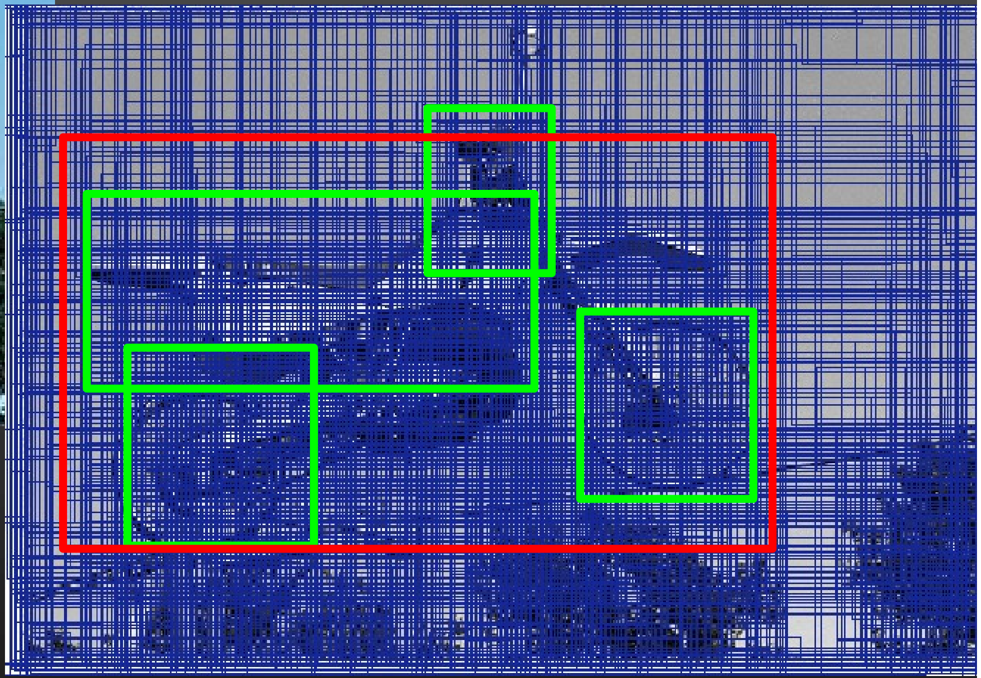




(Cho, Kwak, Schmid, Ponce, 2015)


PIA
Take a trip on a London landmark
Matching
(Russell et al.'06; Cho et al.'10; Deselaers et al.'10; Rubinstein & Joulin'13; Rubio et al.'13)

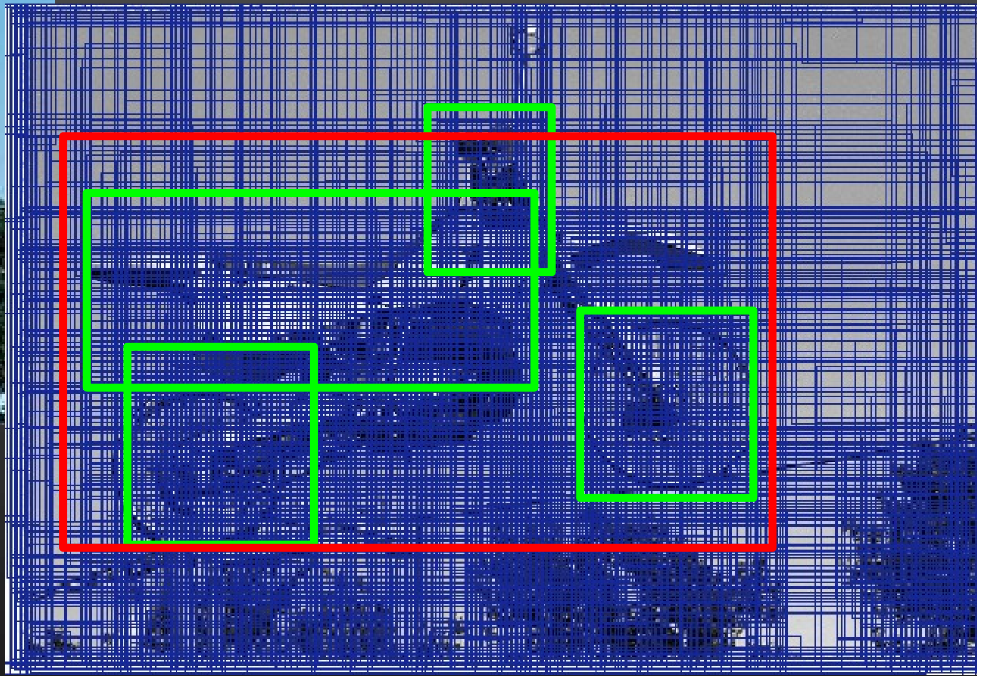
Finding parts and objects among region candidates



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

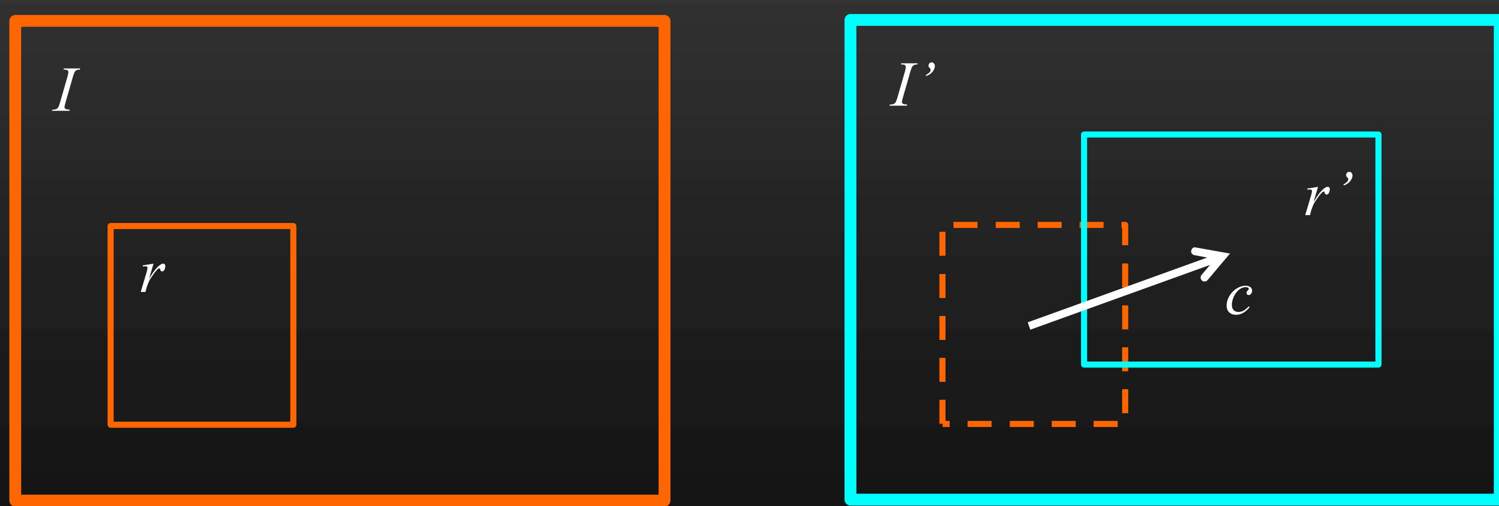
Matching model - Probabilistic Hough matching

match data configuration

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

two regions region proposals position+scale

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



Matching model – Probabilistic Hough matching

match   data   configuration

$$P(m \mid d) = \sum_c P(m \mid c)\; P(c \mid d)$$

$$= P(m_a) \sum_c P(m_g \mid c)\; P(c \mid d)$$

appearance   geometry

Matching model – Probabilistic Hough matching

- Bayesian model

$$P(m \mid d) = \sum_c P(m \mid c)\; P(c \mid d)$$

$$= P(m_a) \sum_c P(m_g \mid c)\; P(c \mid d)$$

Matching model – Probabilistic Hough matching

- Bayesian model

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

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

- Probabilistic Hough transform

$$P(c \mid d) \approx H(c \mid d) = \sum_{m \in d} P(m \mid c)$$

$$= \sum_{m \in d} P(m_a)\, P(m_g \mid c)$$

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

Matching model – Probabilistic Hough matching

- Bayesian model

$$P(m \mid d) = \sum_c P(m \mid c)\ P(c \mid d)$$
$$= P(m_a) \sum_c P(m_g \mid c)\ P(c \mid d)$$

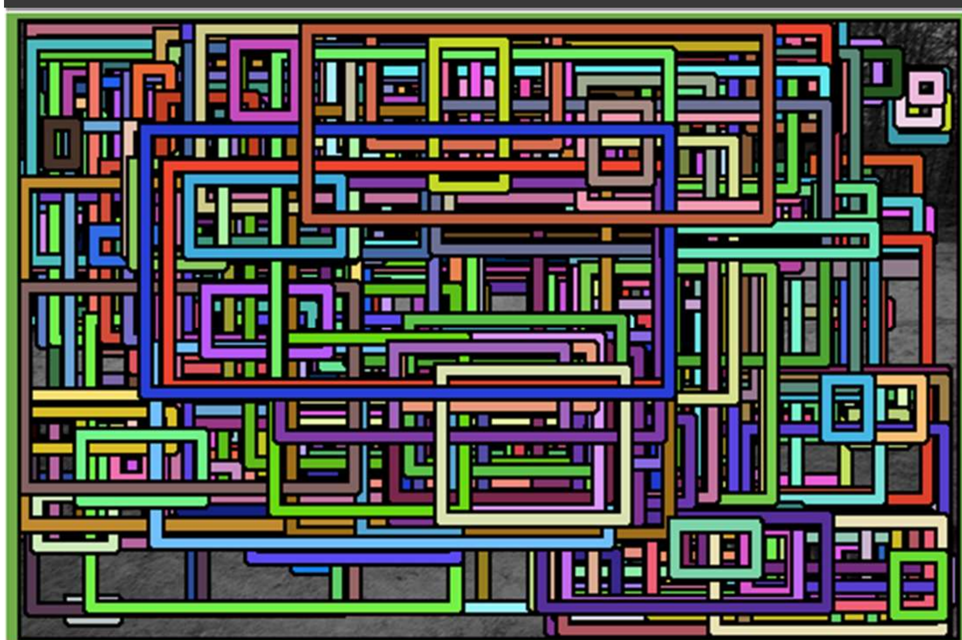
- Probabilistic Hough transform

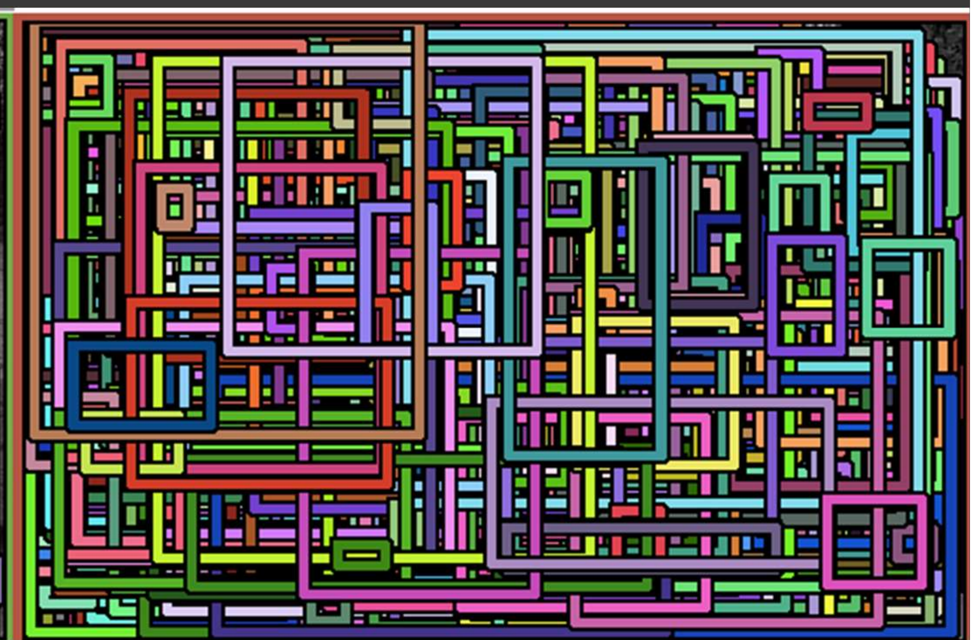
$$P(c \mid d) \approx H(c \mid d) = \sum_{m \in d} P(m \mid c)$$
$$= \sum_{m \in d} P(m_a)\, P(m_g \mid c)$$

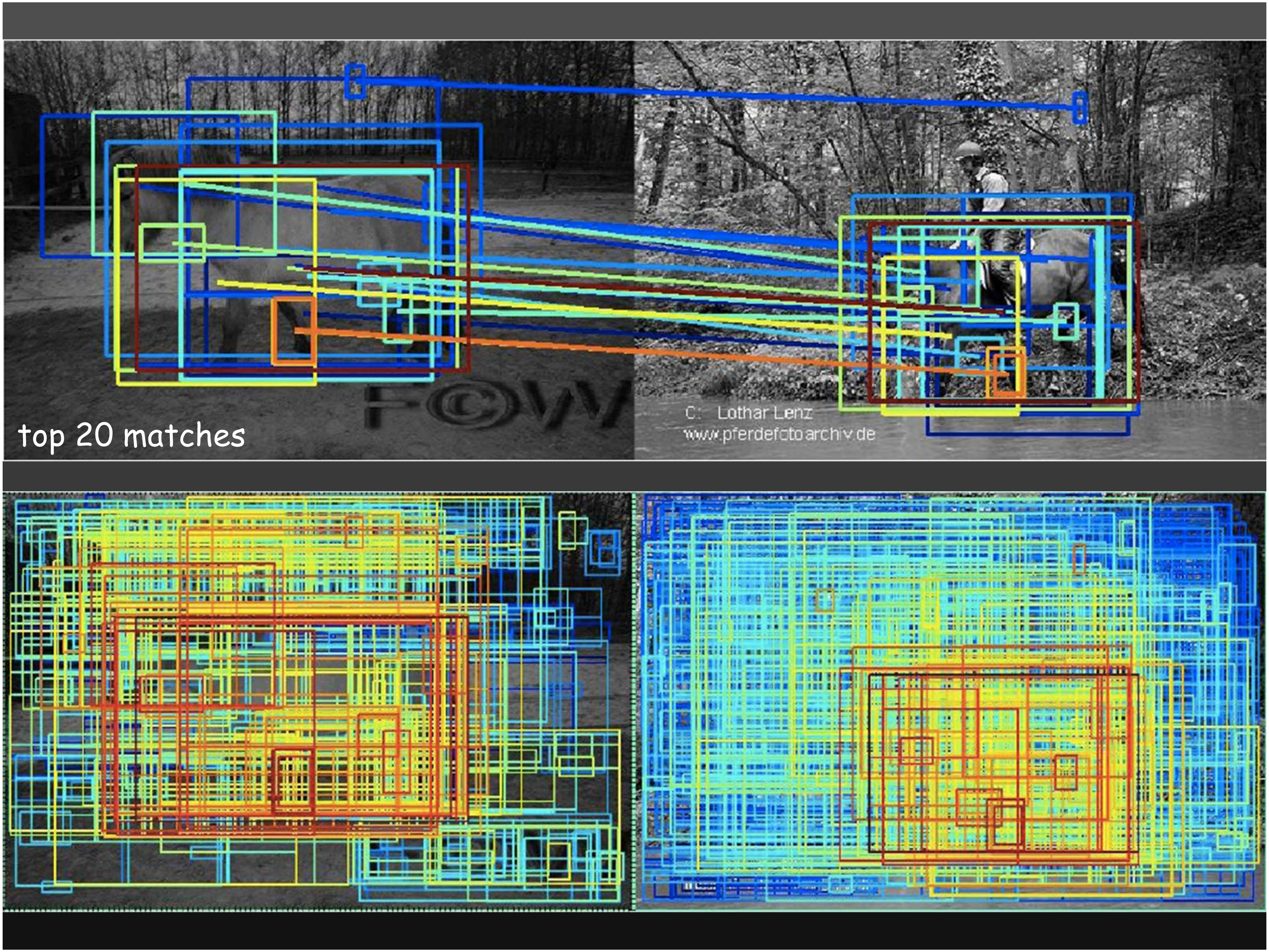
- Region confidence

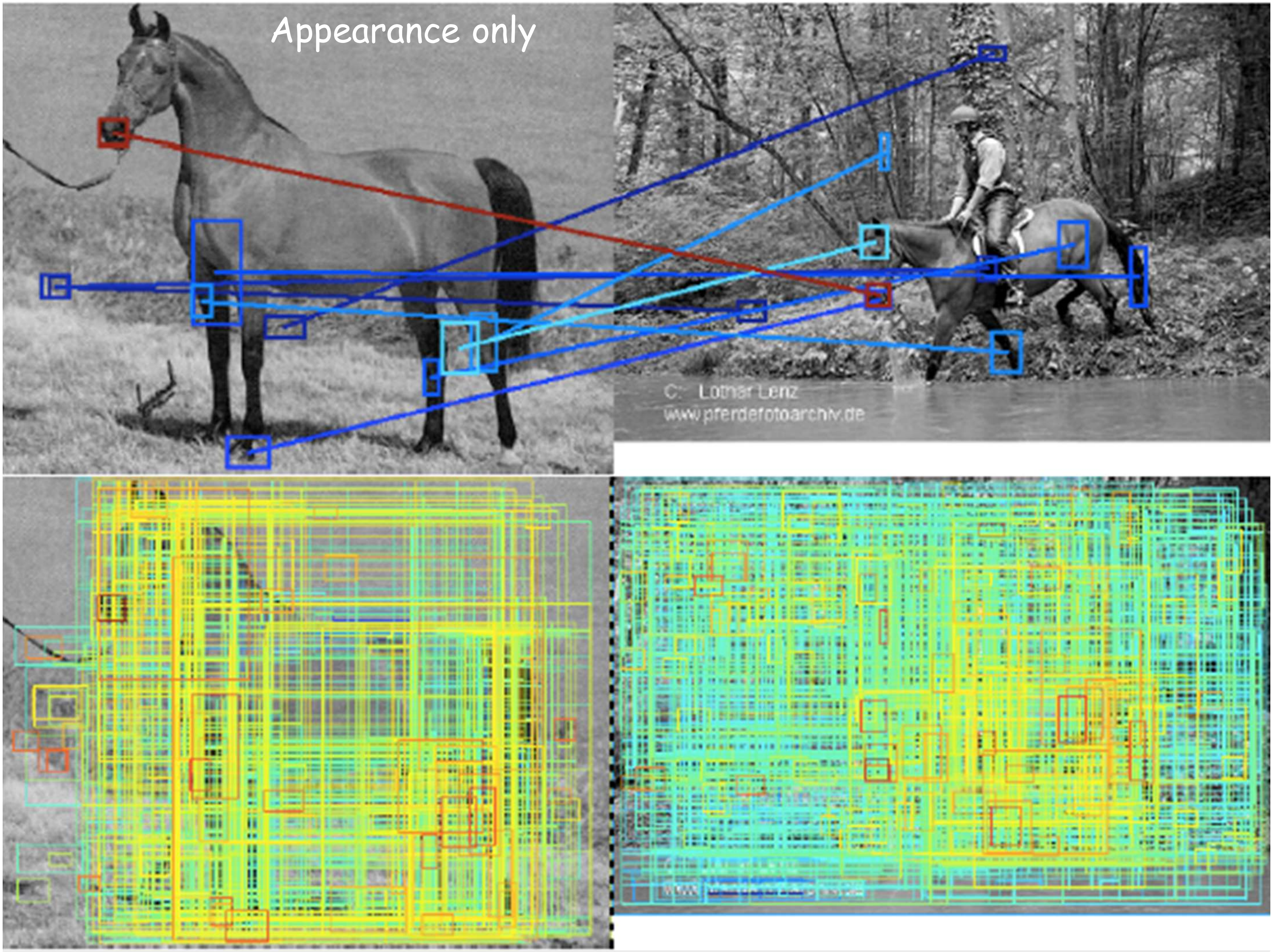
$$C(r' \mid d) = \max_{r''} P(r' \longleftrightarrow r'' \mid d)$$

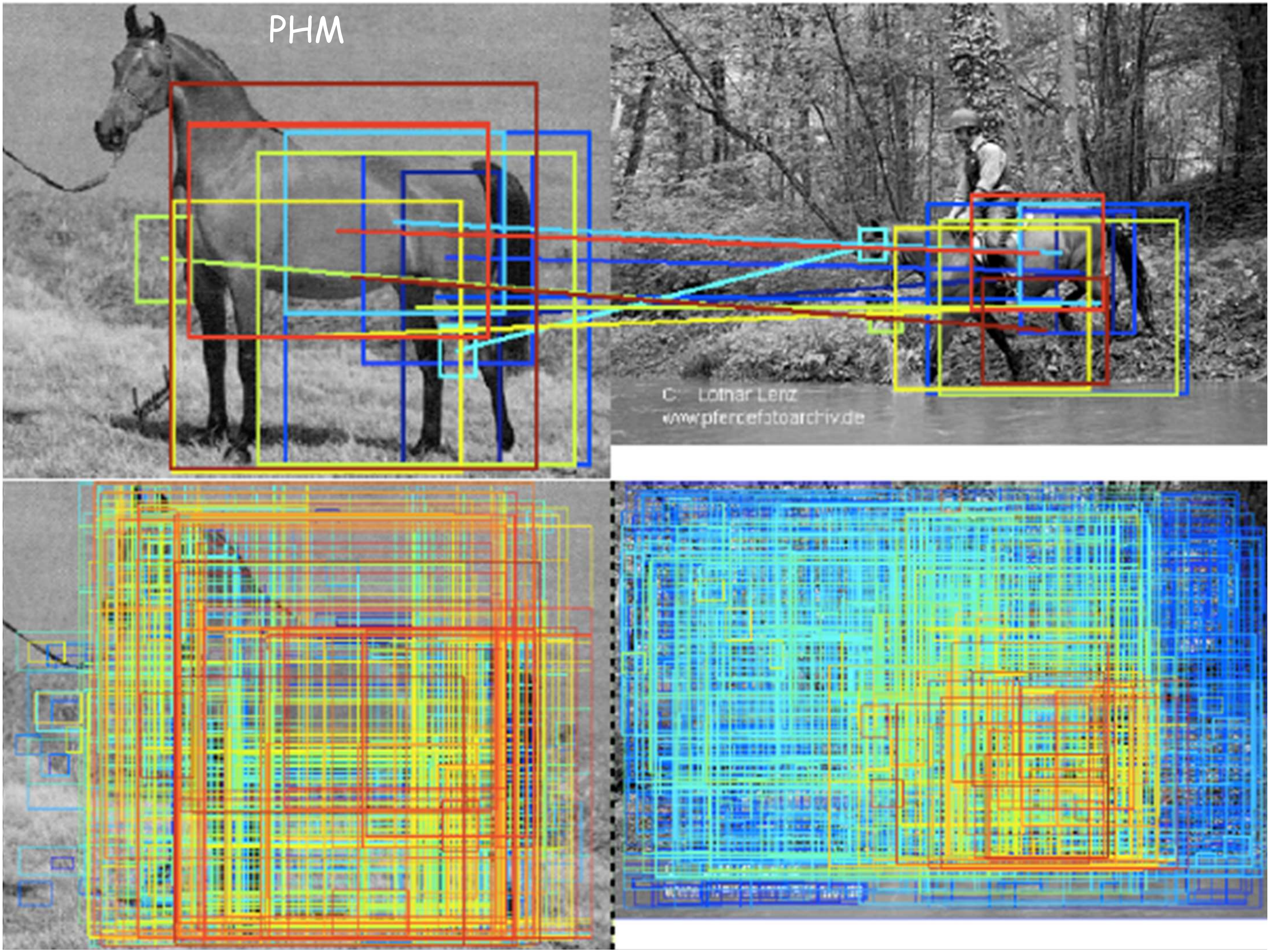

C: Lothar Lenz
www.pferdefotoarchiv.de






top 20 matches
C: Lothar Lenz
www.pferdefotoarchiv.de


Appearance only
C: Lothar Lenz
www.pferdefotoarchiv.de


PHM
C Lothar Lenz
www.pferdefotoarchiv.de

Matching model – Probabilistic Hough matching

- Bayesian model

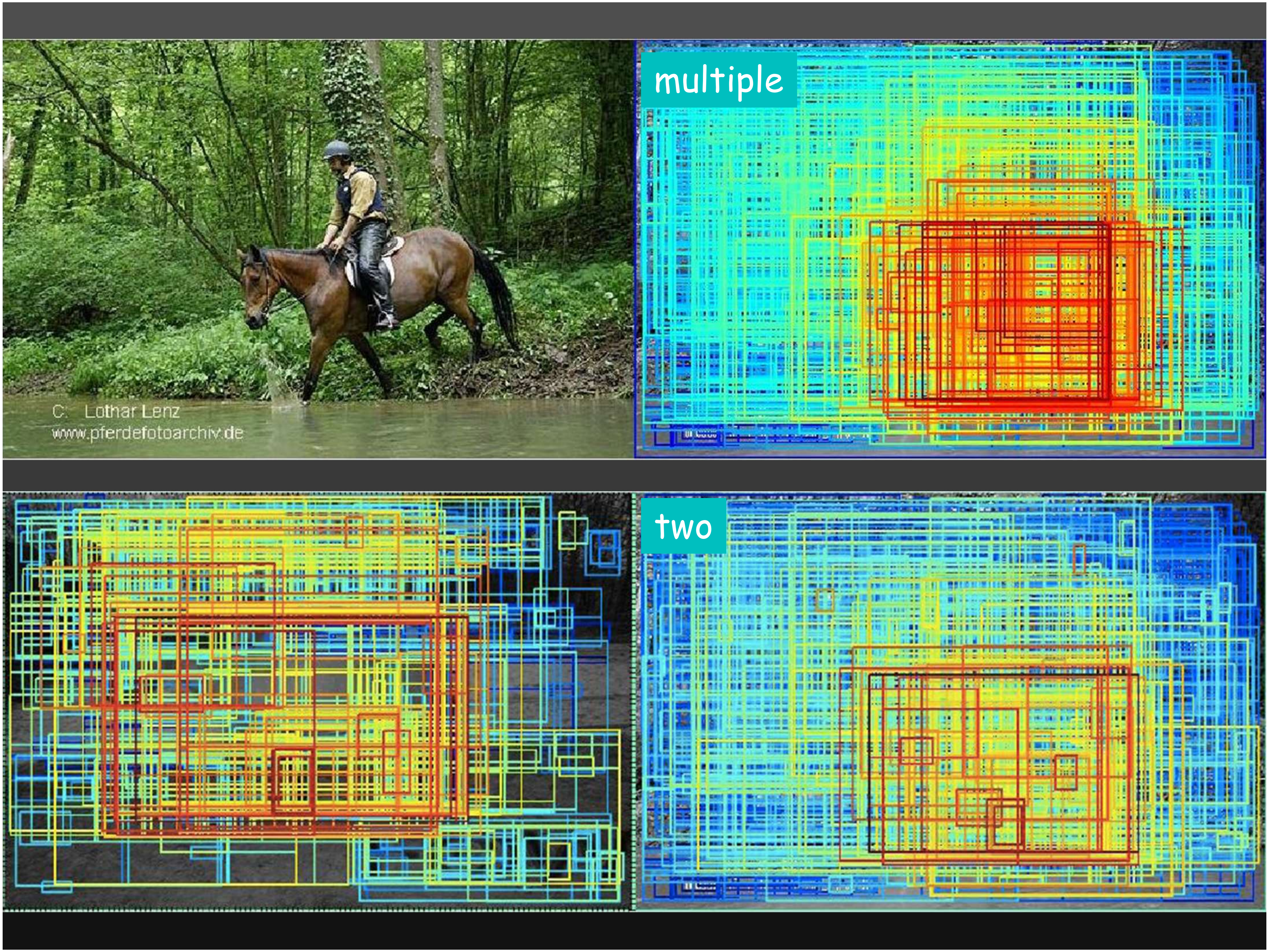
$$P(m \mid d) = \sum_c P(m \mid c)\ P(c \mid d)$$
$$= P(m_a) \sum_c P(m_g \mid c)\ P(c \mid d)$$

- Probabilistic Hough transform

$$P(c \mid d) \approx H(c \mid d) = \sum_{m \in d} P(m \mid c)$$
$$= \sum_{m \in d} P(m_a)\, P(m_g \mid c)$$

- Two images -> multiple images

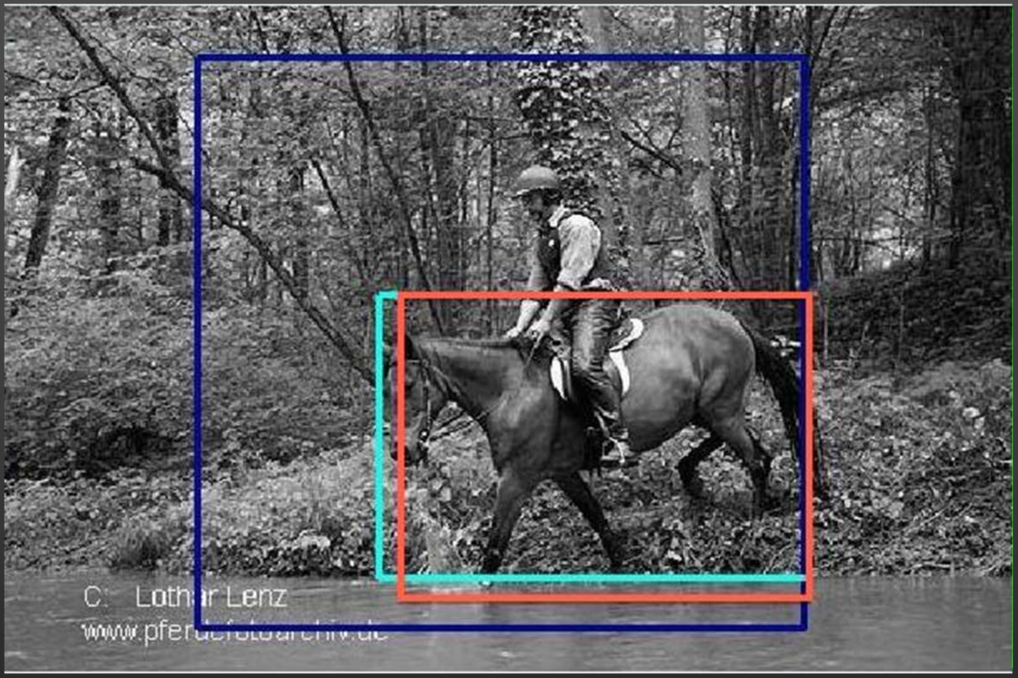
$$C_{d'}(r') = \sum_{d''} C(r' \mid [d', d''])$$


C. Lothar Lenz
www.pferdefotoarchiv.de
multiple
two

Stand-out scoring of part hierarchies

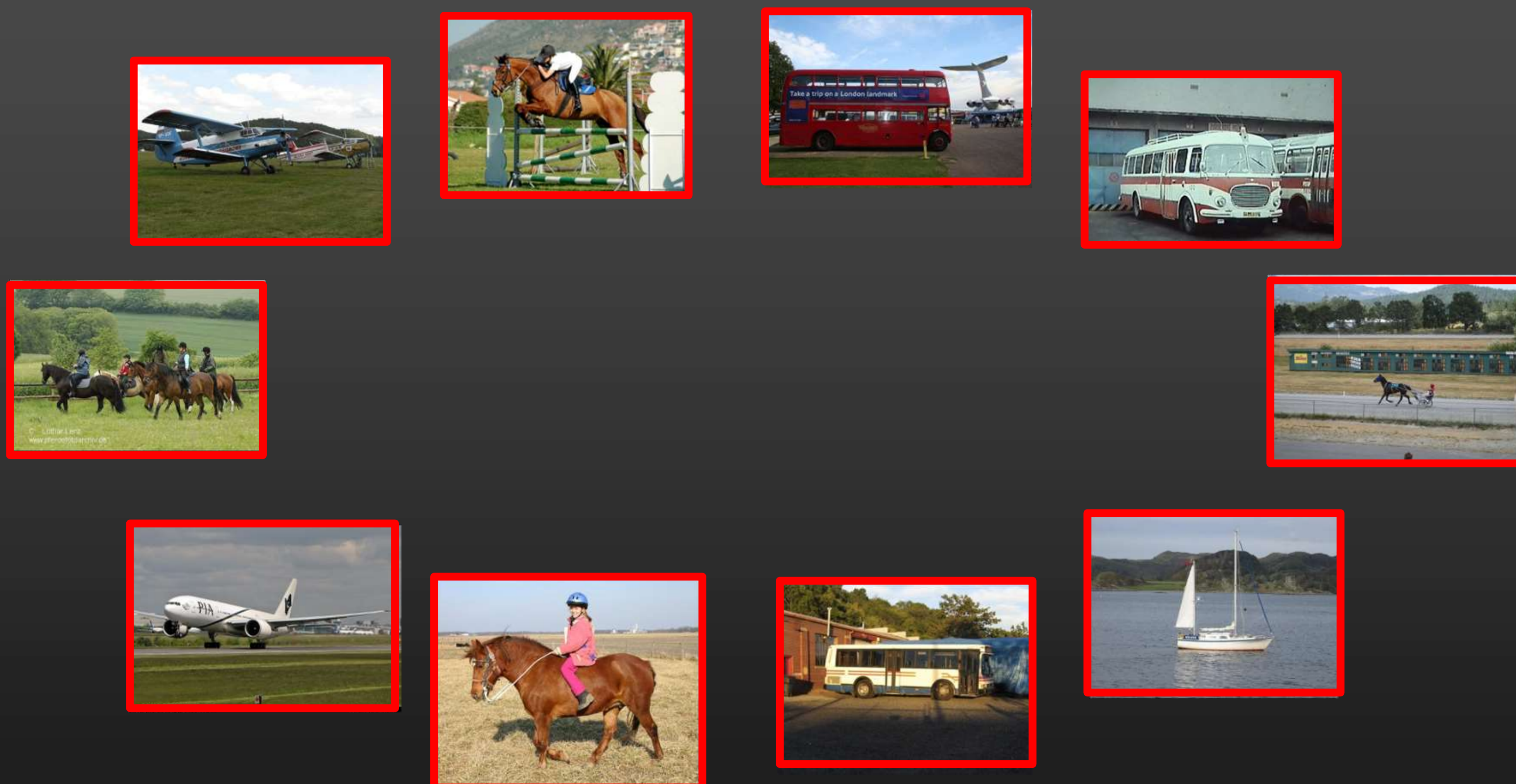
- Object regions should contain
  - more foreground than part regions
  - less background than larger regions

Stand-out scoring of part hierarchies



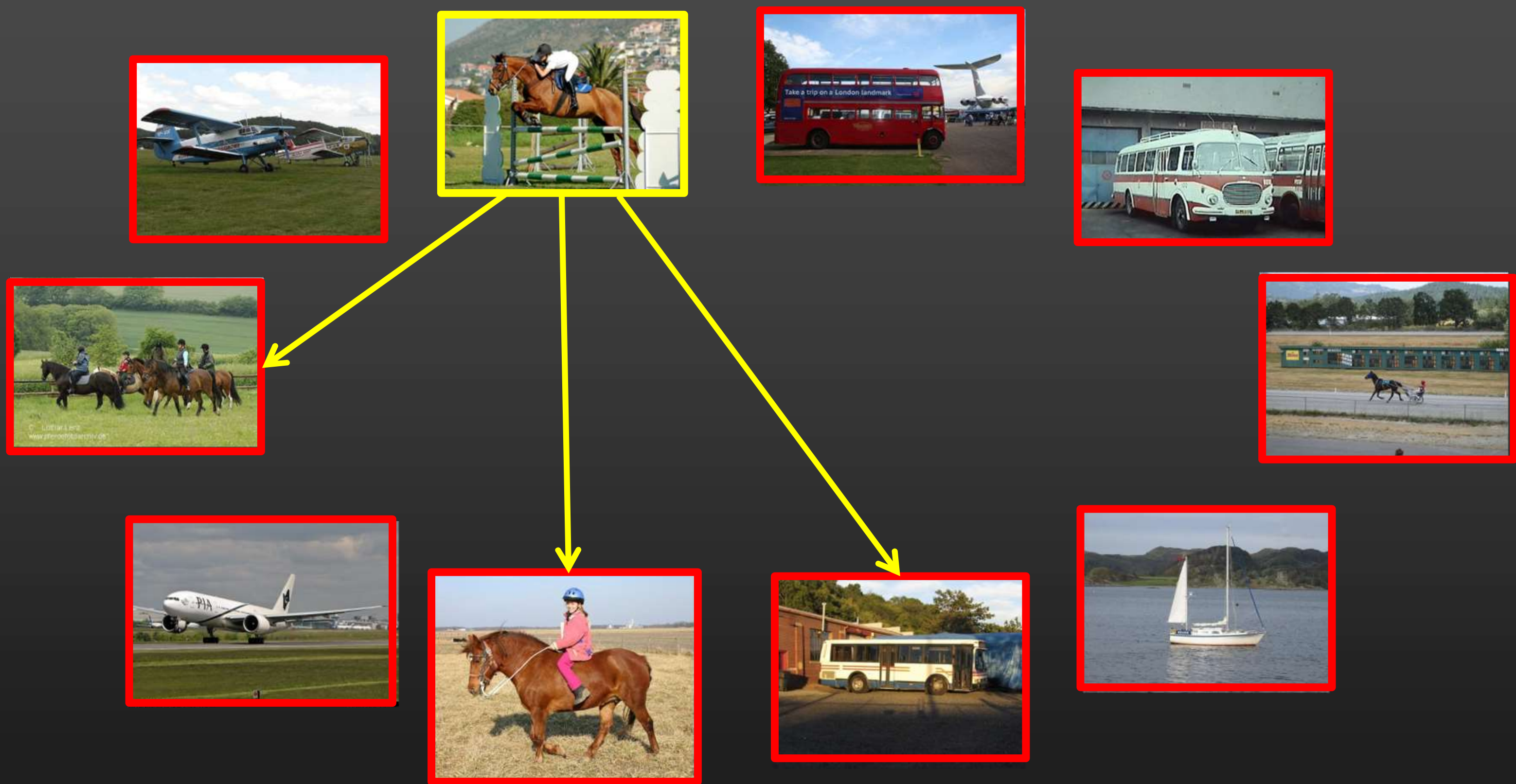
- Object regions should contain
  - more foreground than part regions
  - less background than larger regions
- $S_d(r) = C_d(r) - \max_{r' \supset r} C_d(r')$

A simple iterative algorithm



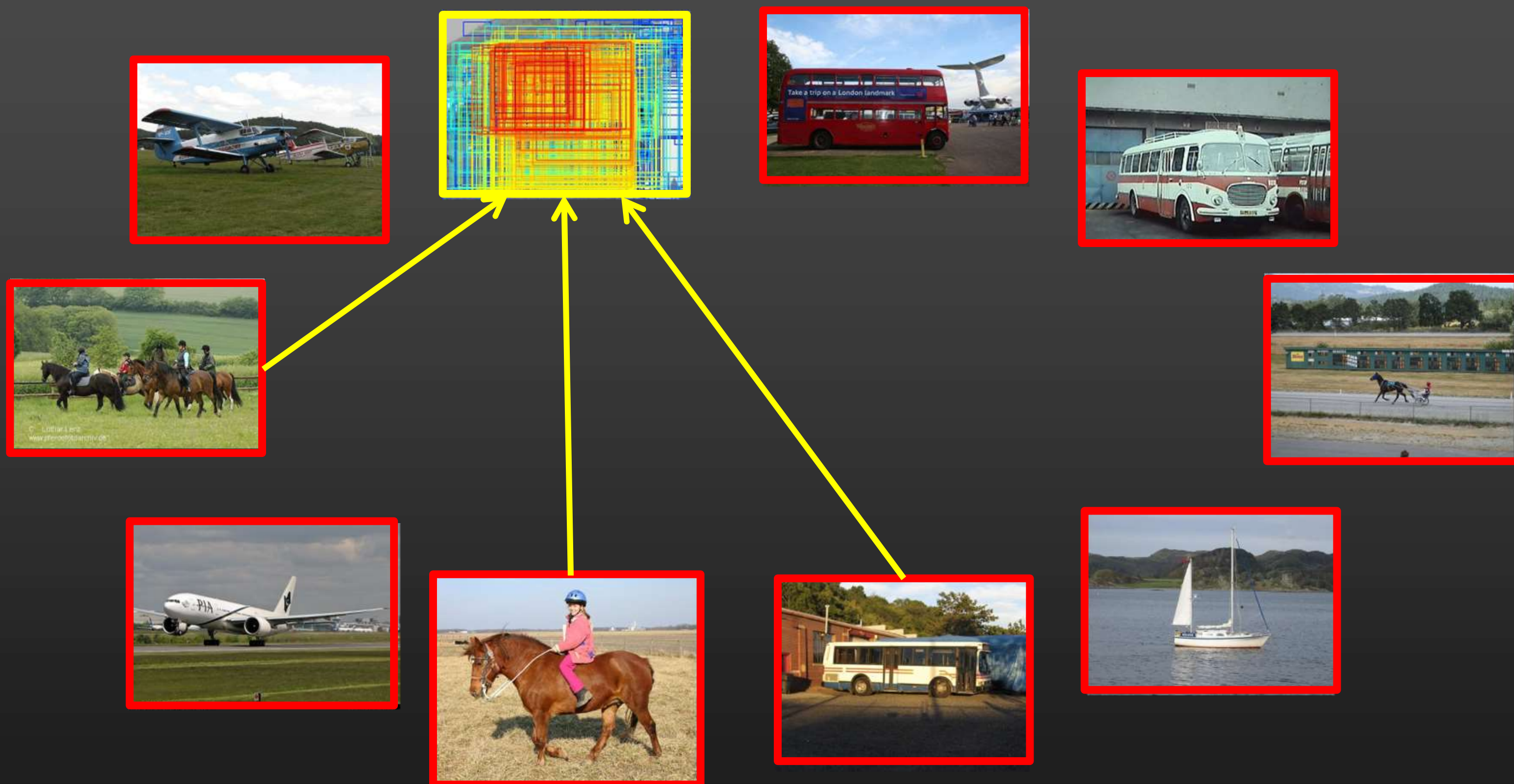
Initialize

A simple iterative algorithm



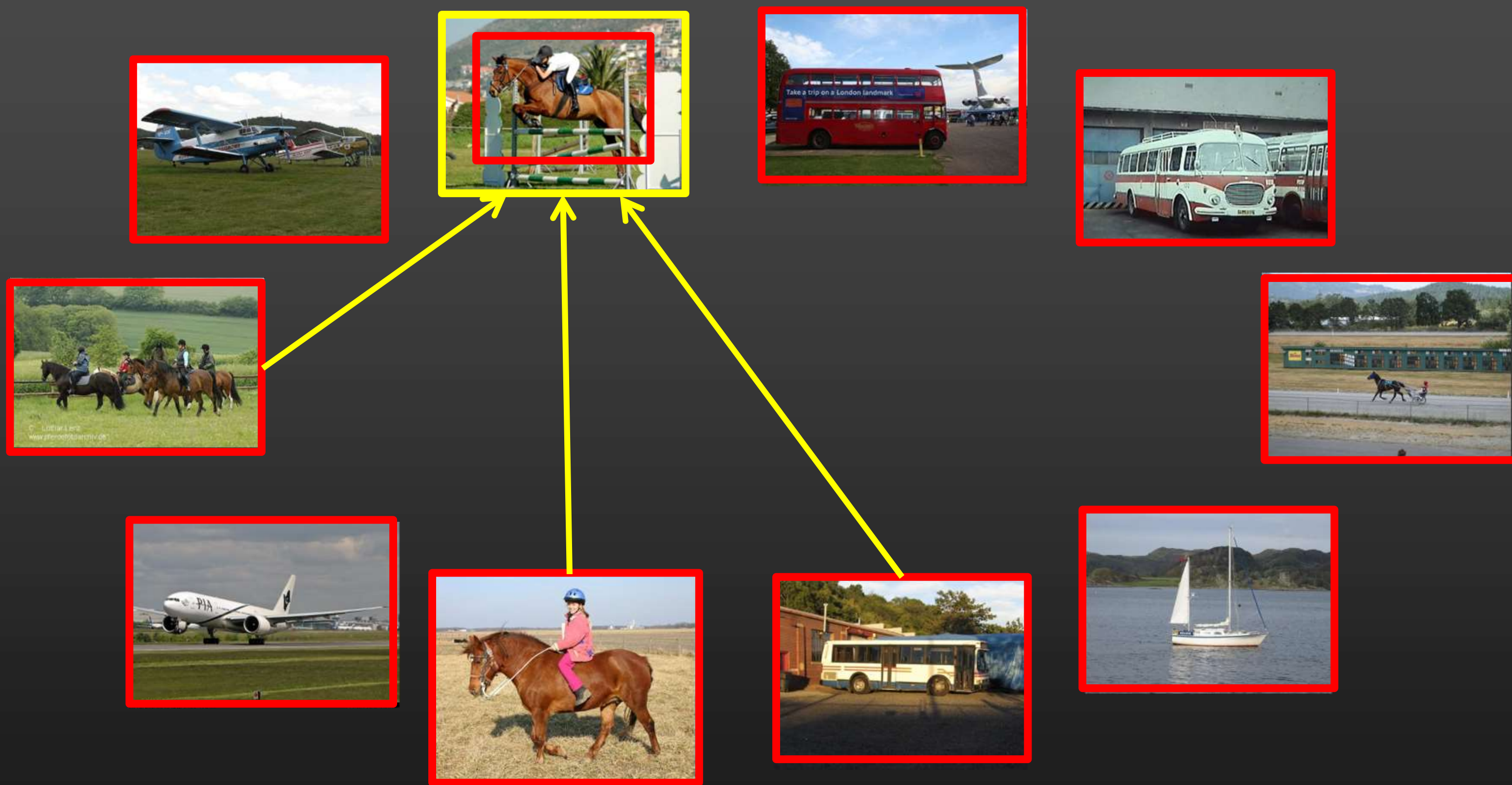
Retrieve 10 nearest neighbors (Oliva & Torralba'06)

A simple iterative algorithm



Match

A simple iterative algorithm

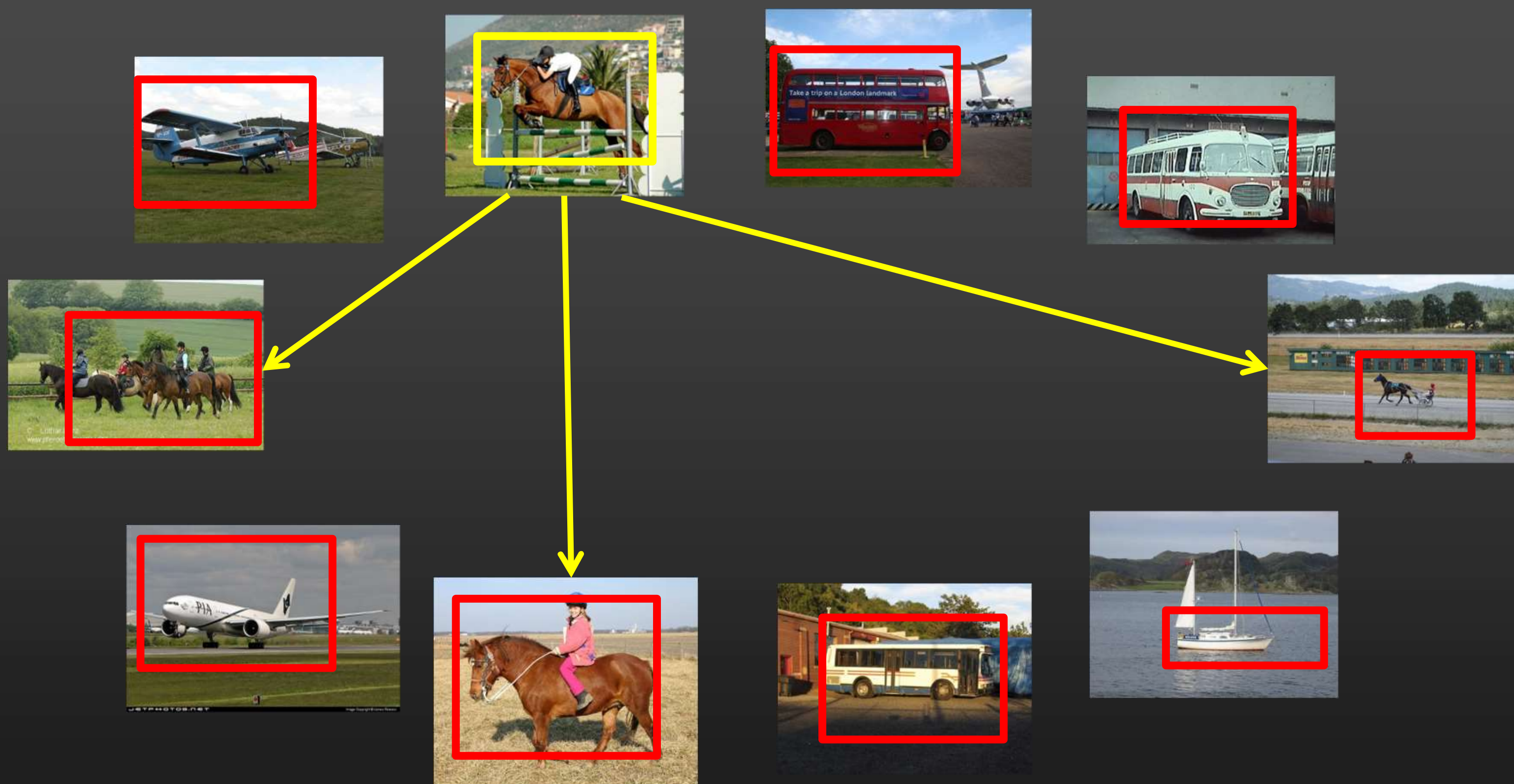


Localize

A simple iterative algorithm

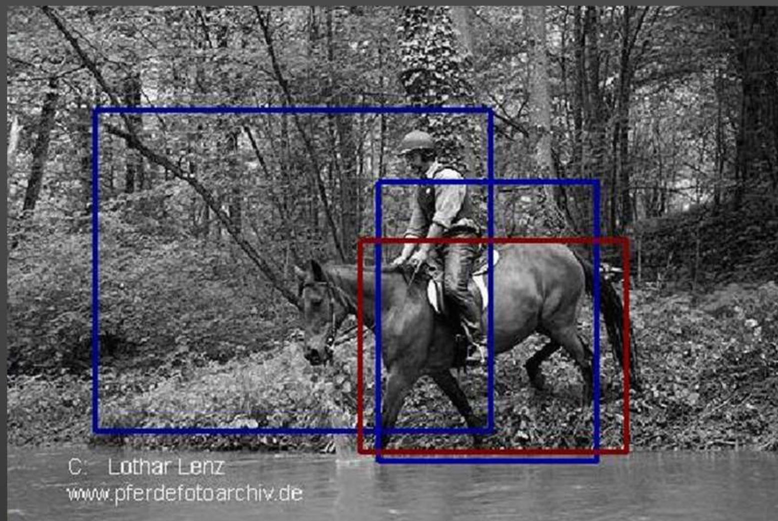
Localize

A simple iterative algorithm



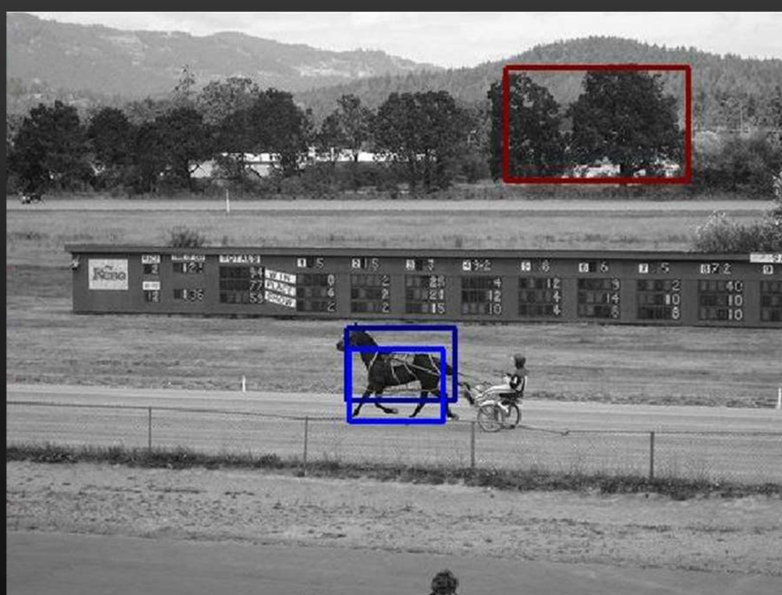
Retrieve using top 20 confidence scores, etc.

Localization improvement over iterations



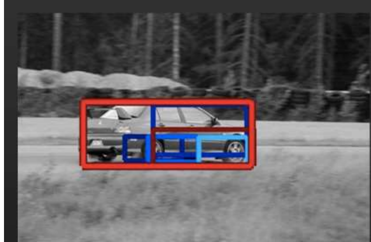
After 1 iteration

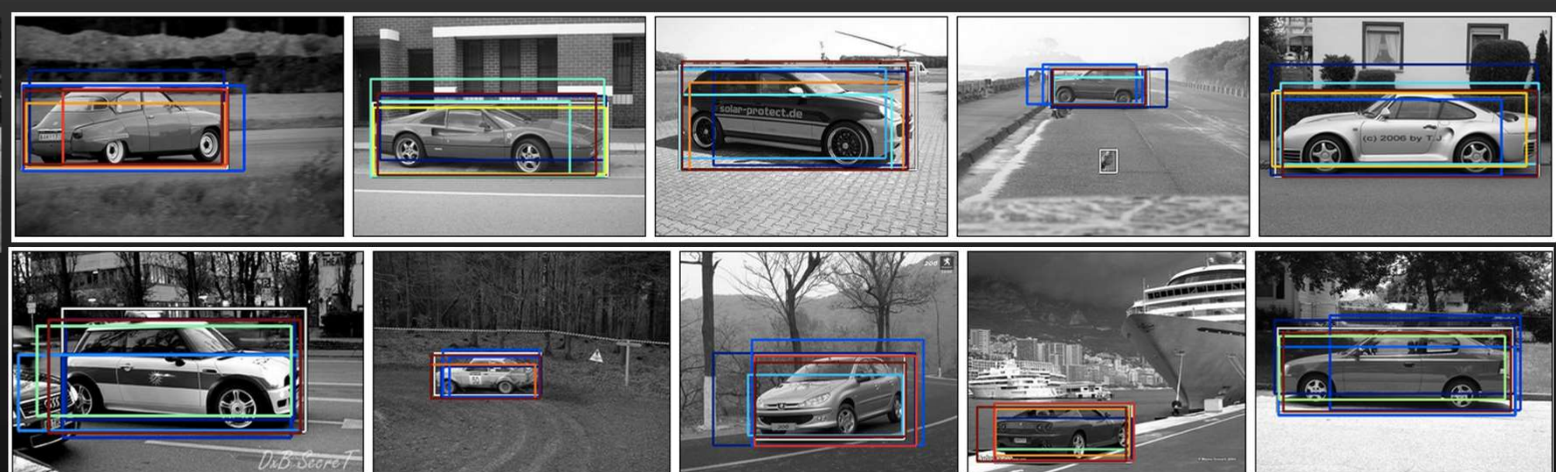
After 3 iterations



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 *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

Uses pre-trained CNN features

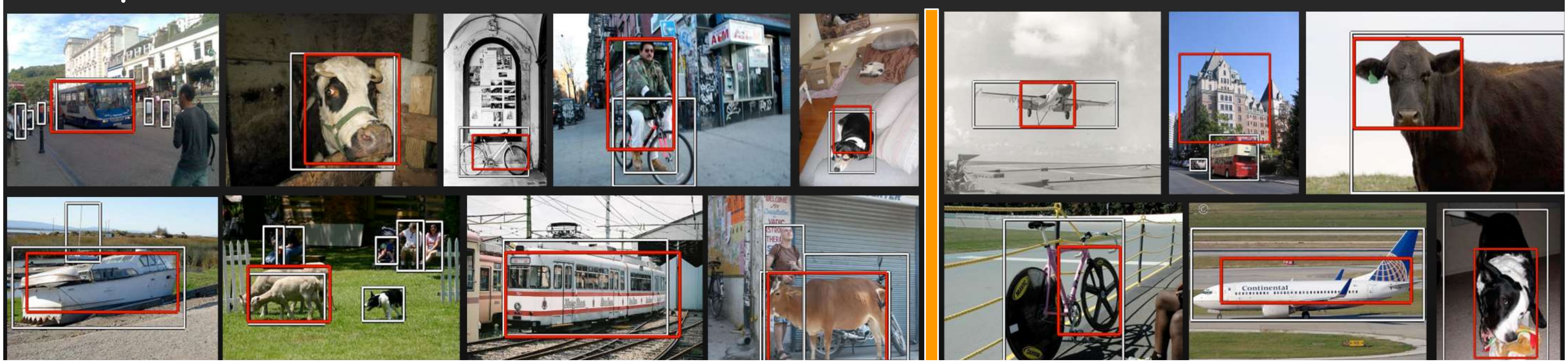
CorLoc and CorRet – mixed classes

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

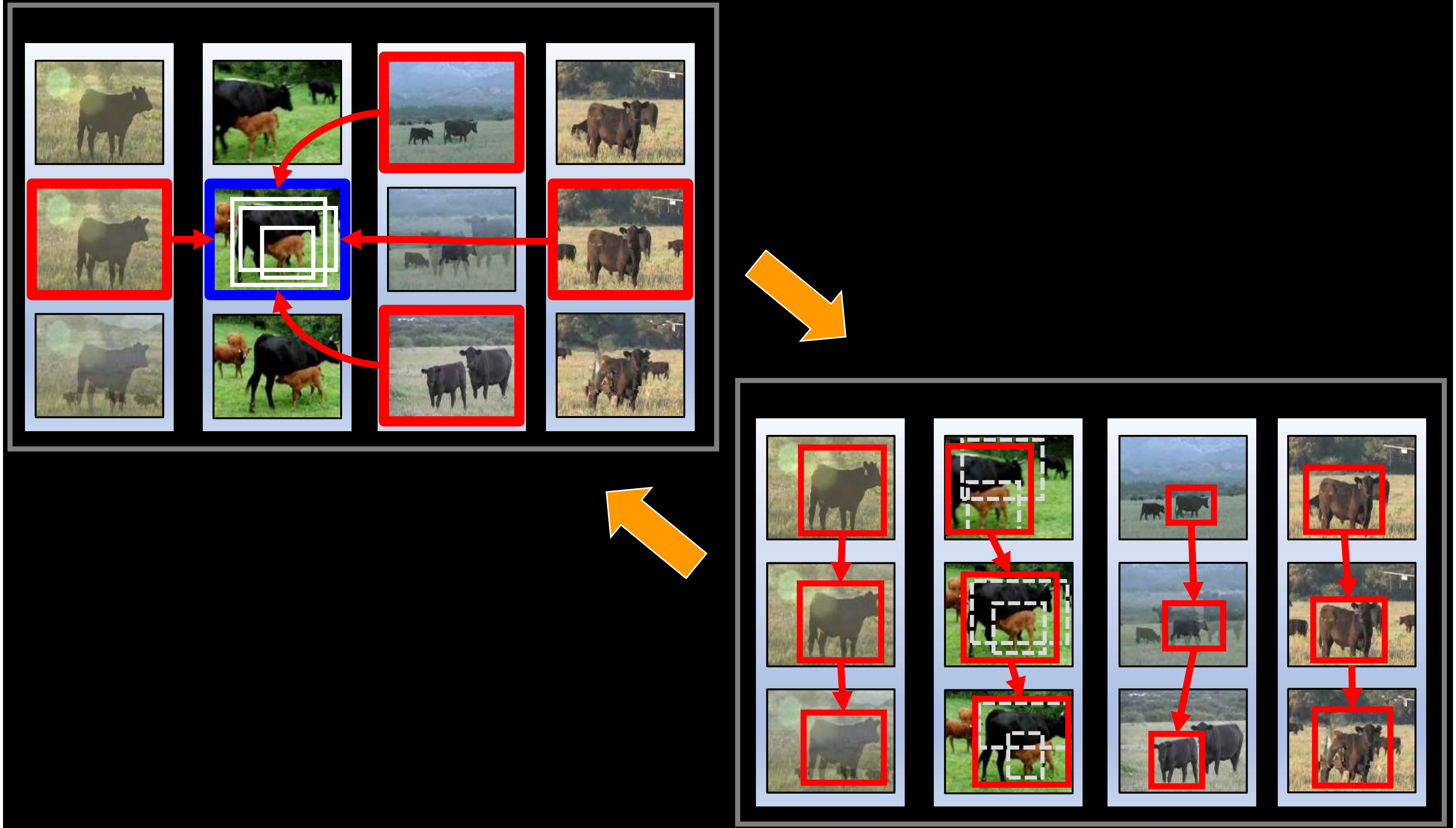
Examples – mixed classes

Successes

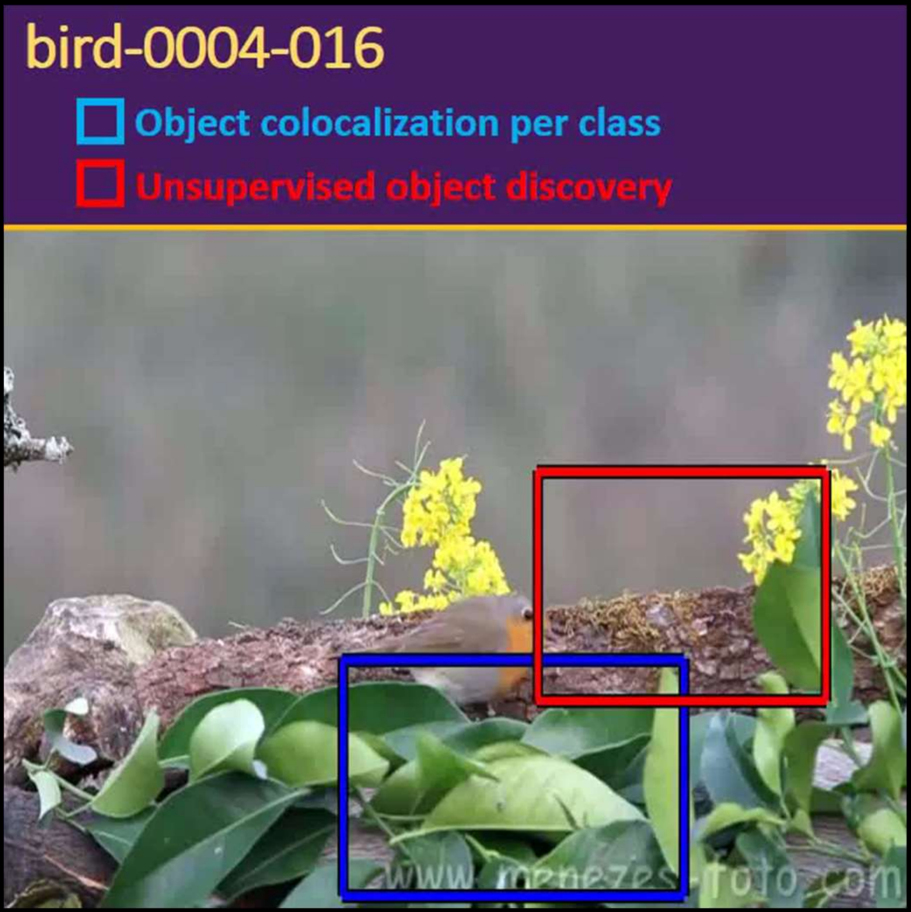
Failures



Unsupervised object discovery in multiple videos



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



(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 *Animal*s 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

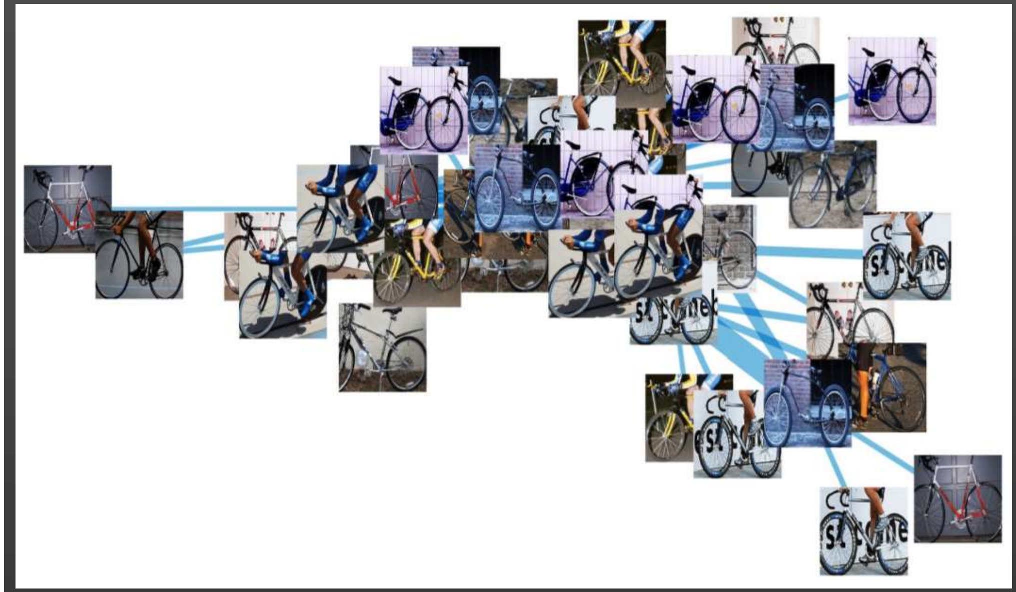
$$\text{Maximize} \quad \sum_{1 \leq i < j \leq n} e_{ij} \sum_{k,l=1}^{p} S_{ij}^{kl} x_i^k x_j^l = \sum_{1 \leq i < j \leq n} e_{ij} x_i^T S_{ij} x_j$$

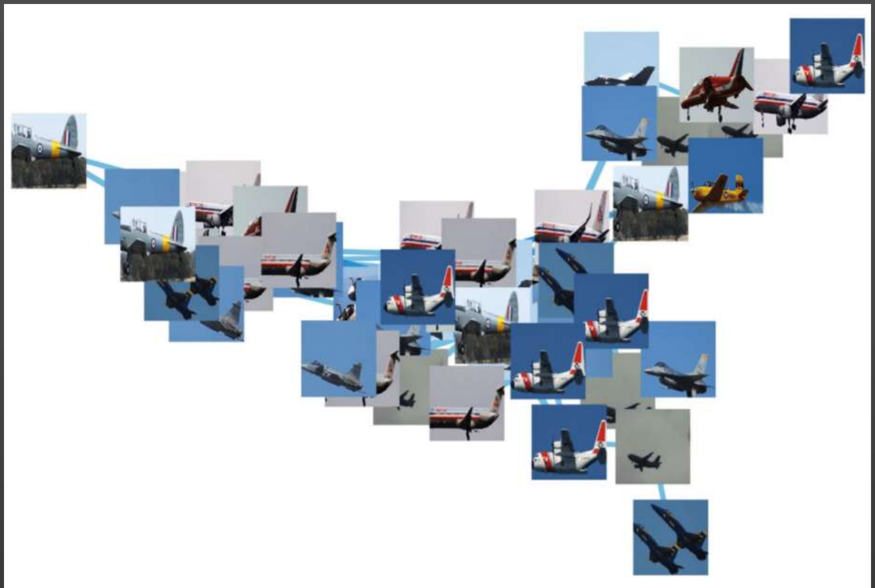
similarity

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

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

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

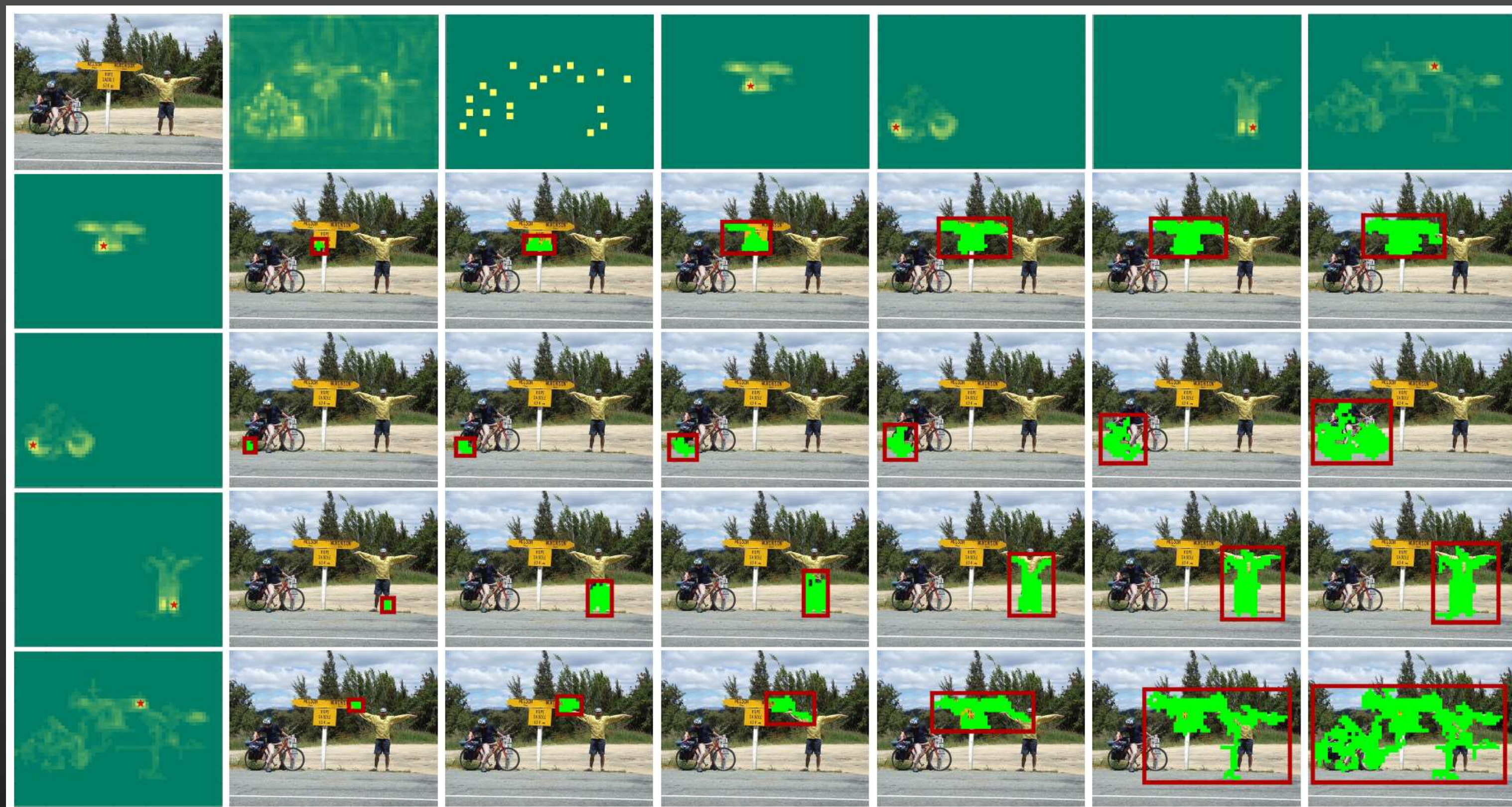




w/o CO: greedy combinatorial search
w CO: use gradient ascent

Method	OD	VOC_6x2	VOC_all
Cho *et al.*	-	-	37.6
Cho *et al.*, our execution	82.2	55.9	37.5
w/o CO	**83.0 $\pm$ 0.4**	**60.2 $\pm$ 0.4**	**39.8 $\pm$ 0.2**
w CO	80.8 $\pm$ 0.5	59.3 $\pm$ 0.4	38.5 $\pm$ 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 ν=50 to find promising proposals
- True structure. Run the algorithm with the selected proposals and ν=5 on the whole image collection

Small-scale CorLoc results

Method	Features	OD	VOC_6x2	VOC_all
Cho *et al.*	WHO	82.2	55.9	37.6
Vo *et al.* RP	WHO	82.3 ± 0.3	62.5 ± 0.6	40.7 ± 0.2
Wei *et al.* [33]	VGG16	75.8	57.9	39.8
Wei *et al.* [34]	VGG16	73.5	66.2	41.9
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]	82.5 ± 0.1	70.6 ± 0.4	44.5 ± 0.1
Ours	**87.5 ± 0.3**	**70.9 ± 0.3**	**48.6 ± 0.1**

Large-scale CorLoc results

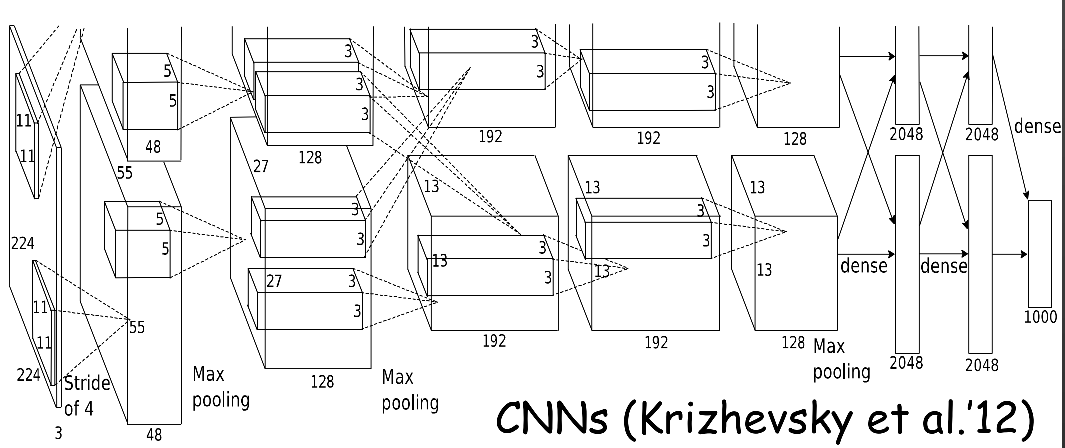
Method	VOC_all	VOC12	COCO_20k
Wei *et al.* [33]	43.4	46.2	38.6
Wei *et al.* [34]	43.4	46.3	40.5
Baseline 1	43.3 ± 0.2	40.1 ± 0.1	45.0 ± 0.1
Baseline 2	**48.6 ± 0.1**	**49.3 ± 0.1**	-
Ours	46.5 ± 0.1	46.2 ± 0.1	**47.3 ± 0.1**

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**

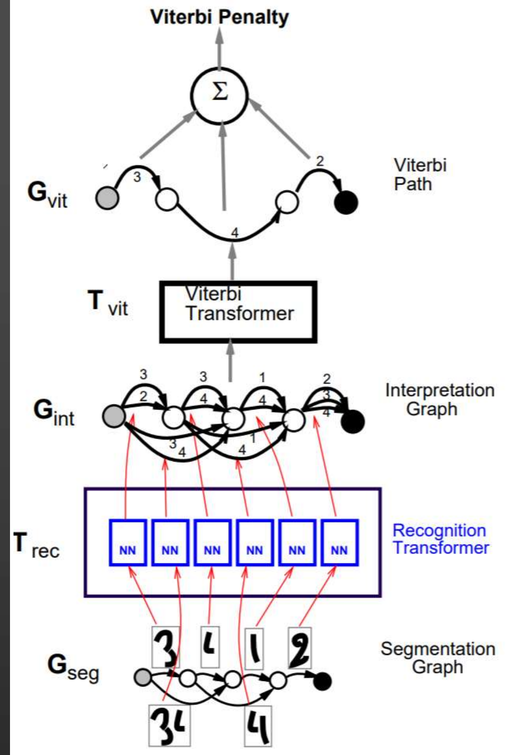
Beyond block diagrams

Deep learning
(LeCun et al.'98)



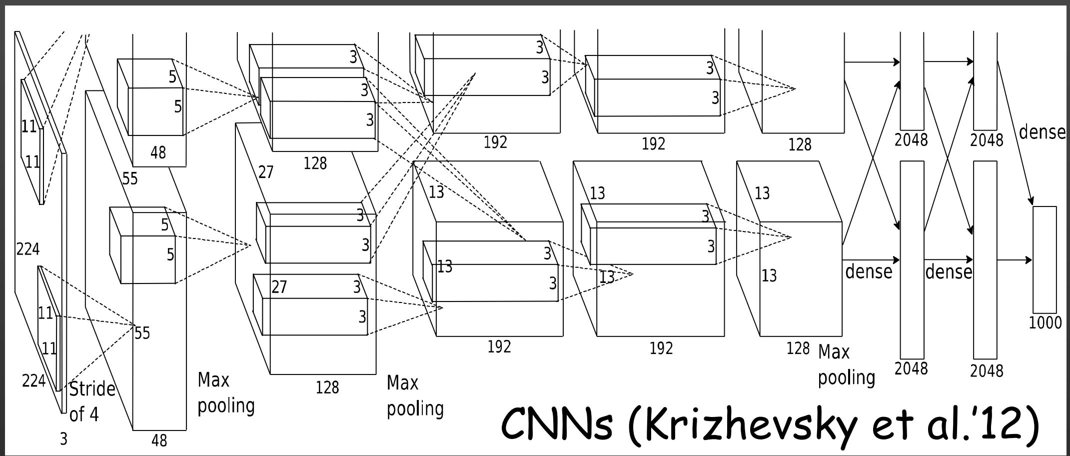
CNNs (Krizhevsky et al.'12)

Graph transformer networks



Beyond block diagrams and pattern recognition

Deep learning
(LeCun et al.'98)

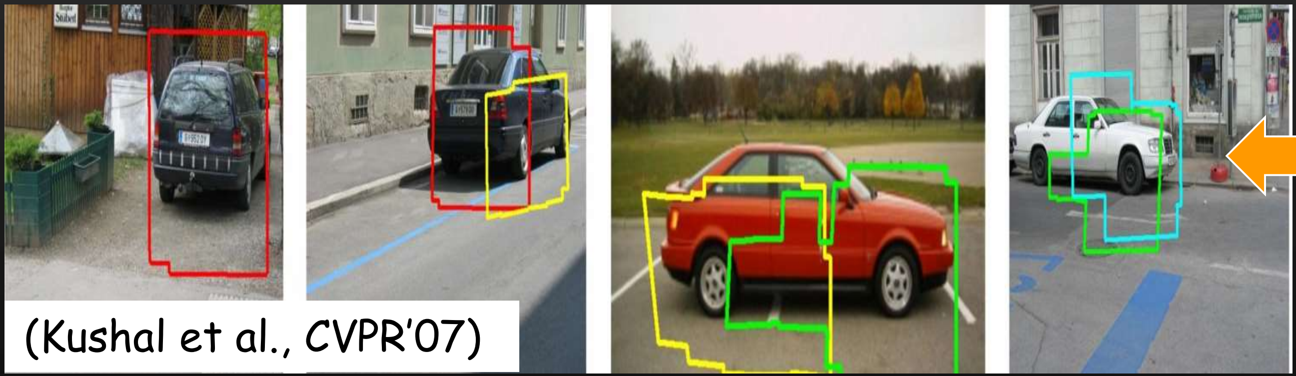


CNNs (Krizhevsky et al.'12)

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