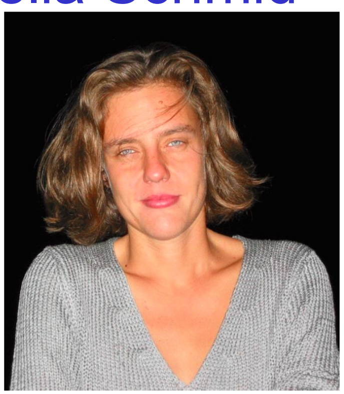
Object category localization with incomplete supervision

Jakob Verbeek

LEAR team, INRIA, Grenoble, France

Joint work with: Gokberk Cinbis and Cordelia Schmid



Object category localization

- Locate object category instances by means of bounding box

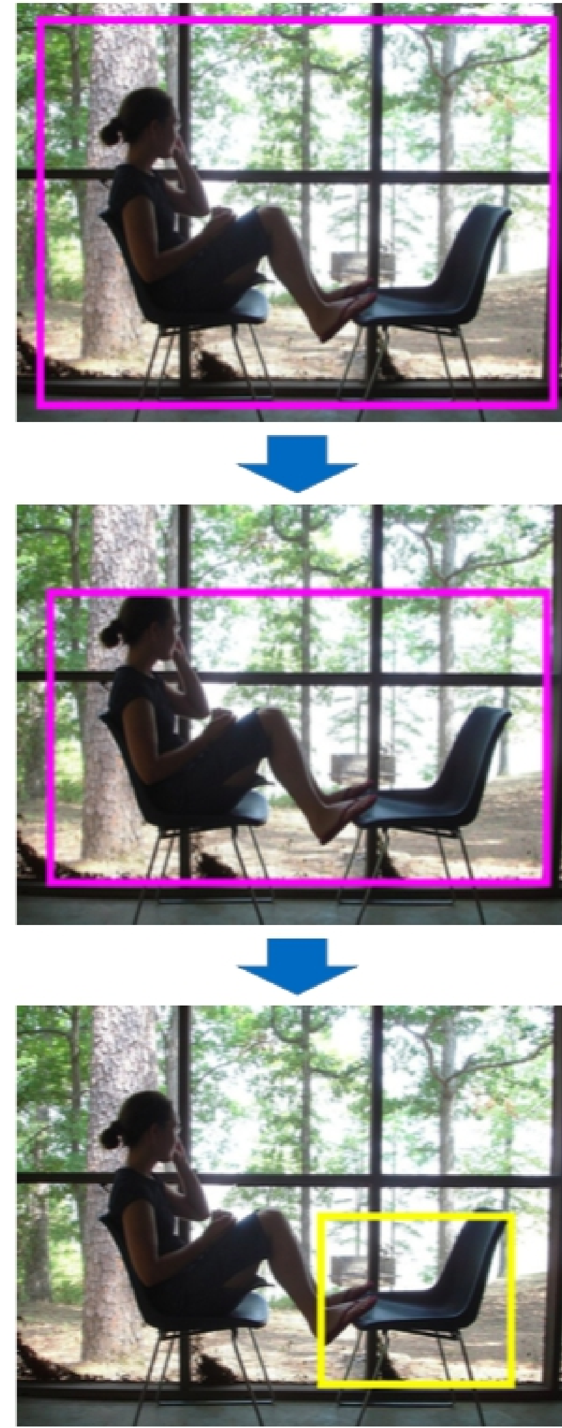
- Supervised learning setup:
  - Training images with bounding-box annotations of object instances
  - Learn a binary classifier: windows are a category instance or not
- Numerous applications
  - Surveillance
  - Traffic safety: autonomous or assisted driving systems
  - ...

Why learning from incomplete supervision?

- Bounding boxes are much more expensive to get than image labels
- Weakly supervised learning only uses image-wide labels
  - For positive images we only know there's at least one instance, but we don't know how many and where they are
  - Less detail in supervision than in target outputs

Presentation outline

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning
- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement
- Experimental evaluation results



Challenging factors in object detection

- Intra-class appearance variation
  - Objects deformation due to pose
  - Transparency: e.g. bottles
  - Sub-categories: e.g. ferry vs yacht

- Scene composition
  - Heavy occlusions: e.g. tables and chairs
  - Clutter: coincidental image content present in bounding box

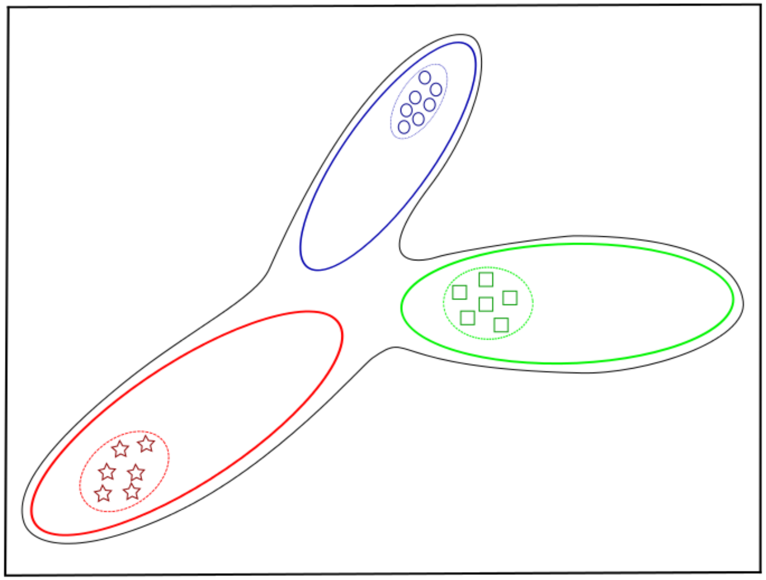
- Imaging conditions
  - viewpoint, scale, lighting conditions

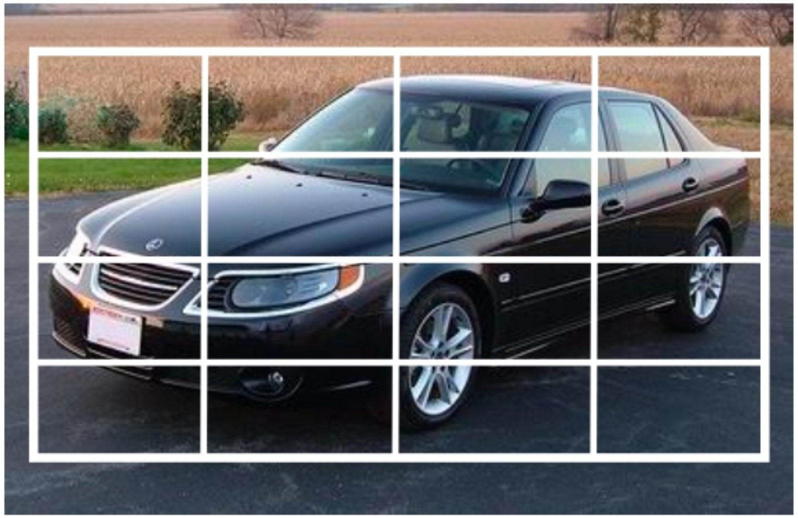
State-of-the-art visual representations (1/2)

- Fischer vector image representation [Sanchez et al., IJCV, 2013]
  - Represent data with gradient of log-likelihood of generative model
- Densely sampled SIFT descriptors modeled with Gaussian mixture
- Encode an image by gradient w.r.t. means and variances: 2KD vector
  - Results in a 140K dimensional signature

$$\nabla_{\mu_k} \ln p(x_{1:N}) = \frac{1}{\sigma_k \sqrt{\pi_k}} \sum_{n=1}^{N} p(k|x_n)(x_n - \mu_k)$$

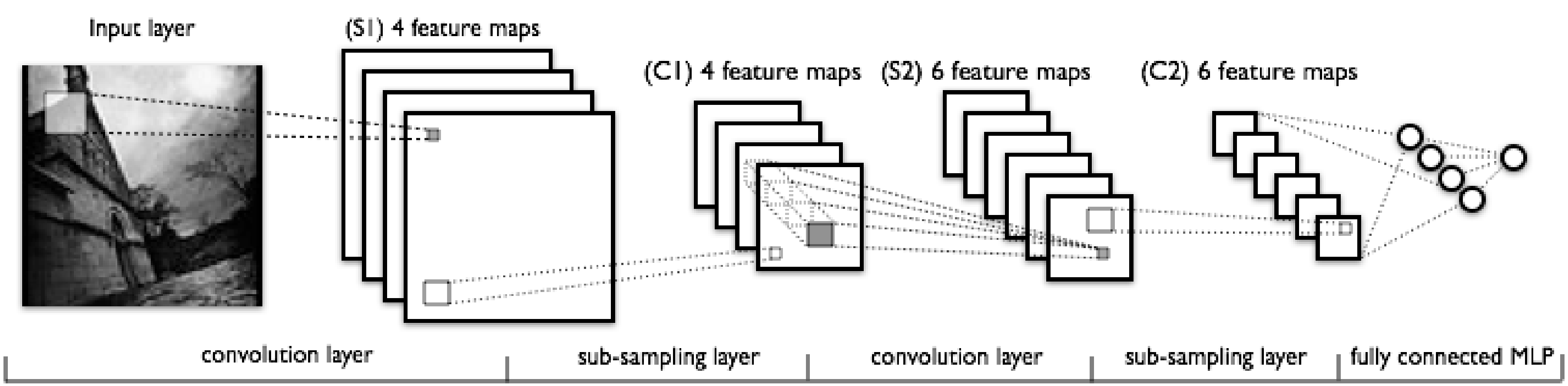
$$\nabla_{\sigma_k} \ln p(x_{1:N}) = \frac{1}{\sigma_k \sqrt{2\pi_k}} \sum_{n=1}^{N} p(k|x_n)\left((x_n - \mu_k)^2 - \sigma_k^2\right)$$



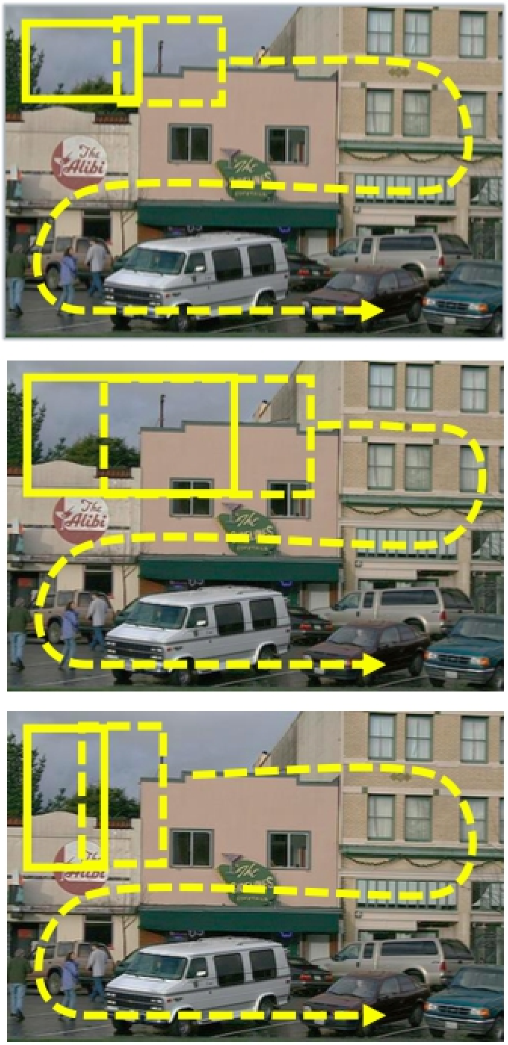


State-of-the-art visual representations (2/2)

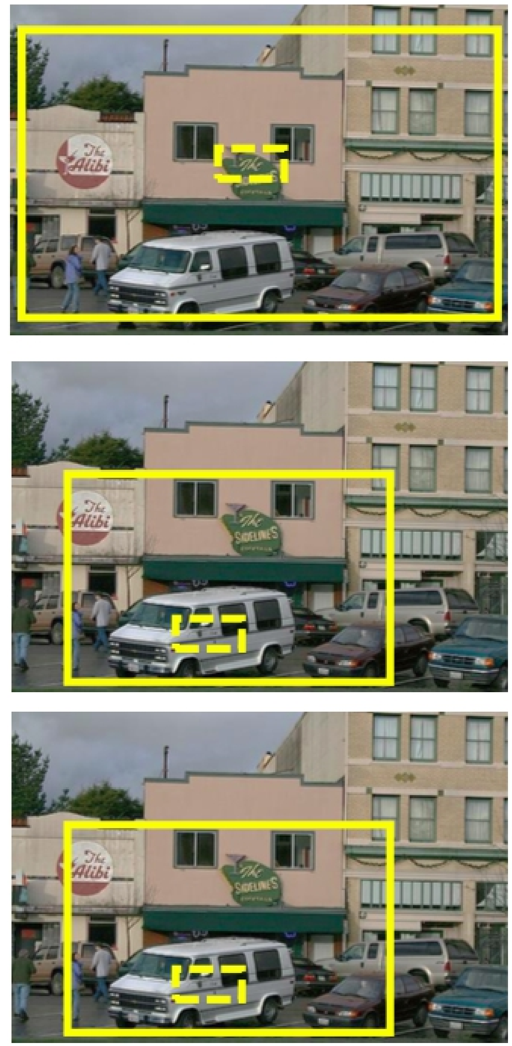
- Use a Convolutional Neural Network as a feature extraction method for object detection [R-CNN, Girschik et al., 2013]
- Trained on 1 million images of 1000 categories (ImageNet 2012)
- Caffe framework [Jia et al., caffe.berkeleyvision.org]
- Use last shared layer as a 4K dimensional representation



How to avoid exhaustive sliding window search



Sliding window
(Viola and Jones 2002;
Felzenszwalb *et al*. 2008, ...)

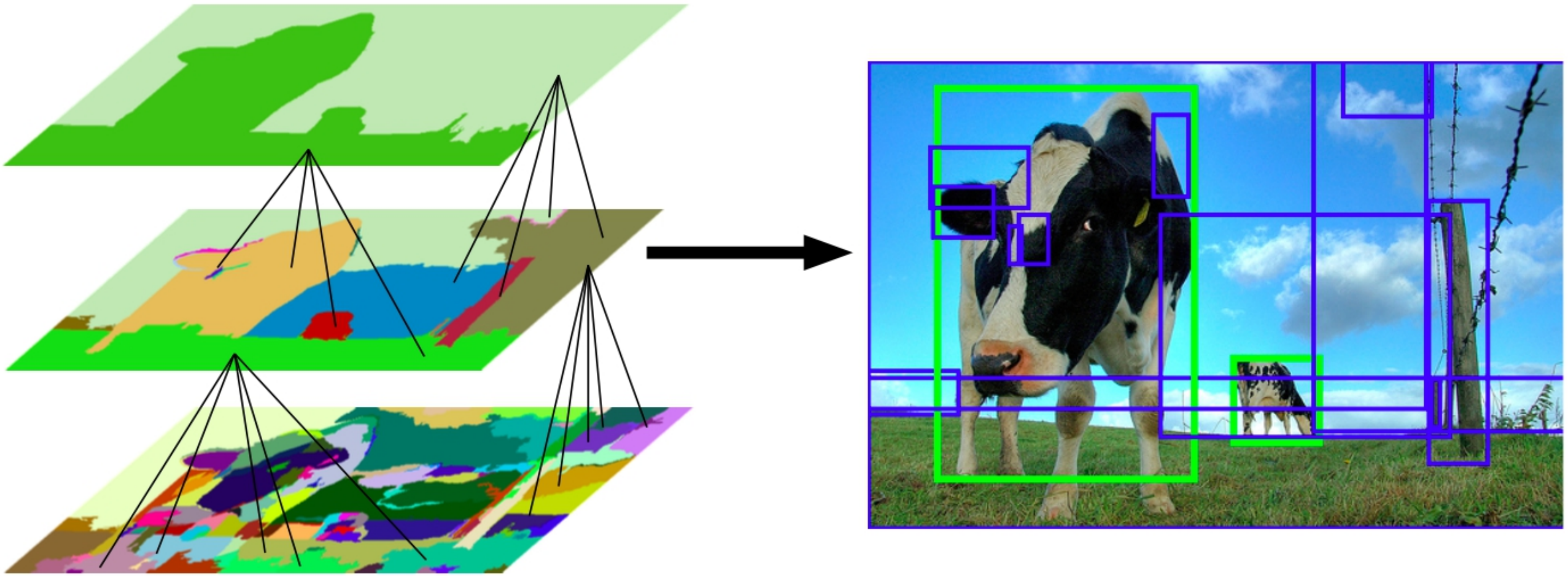


Branch & bound
(Lampert *et al*. 2008;
Lehmann *et al*. 2013)

Selective Search
(Alexe *et al*. 2010;
Sande *et al*. 2011)

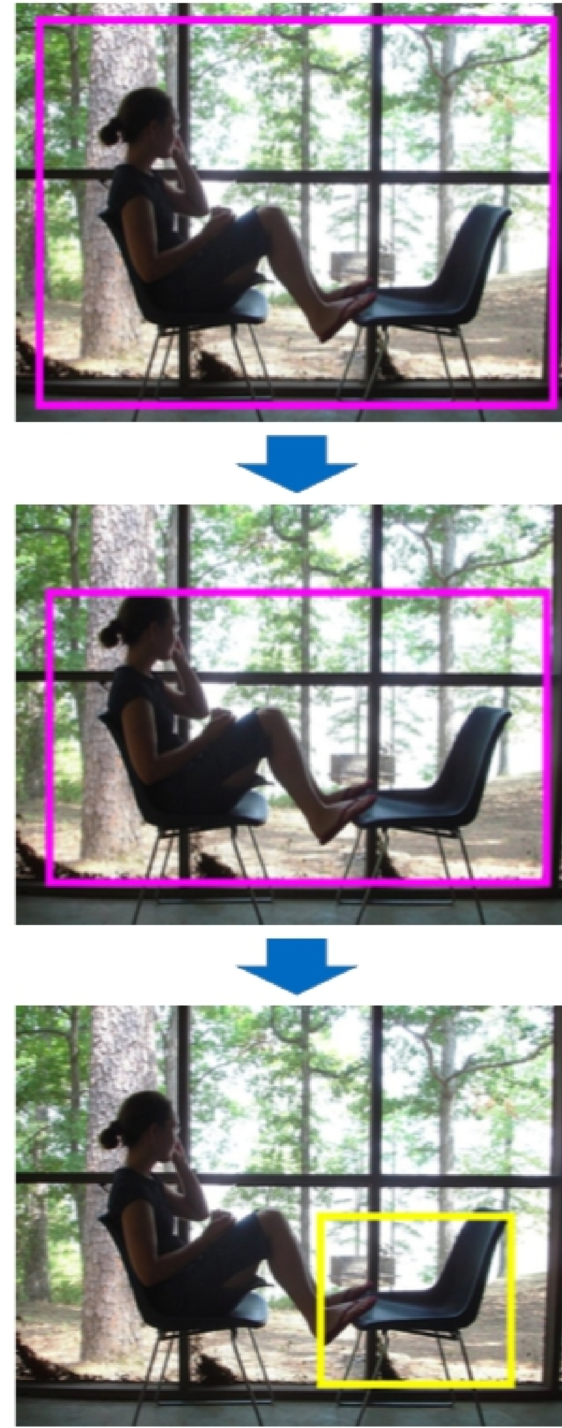
Search: restricted scanning of bounding box space

- Selective search method [Uijlings et al., IJCV, 2013]
  - Unsupervised multi-resolution hierarchical segmentation
  - Detections proposals generated as bounding box of segments
  - 1500 windows per image suffice to cover over 95% of true objects with sufficient accuracy



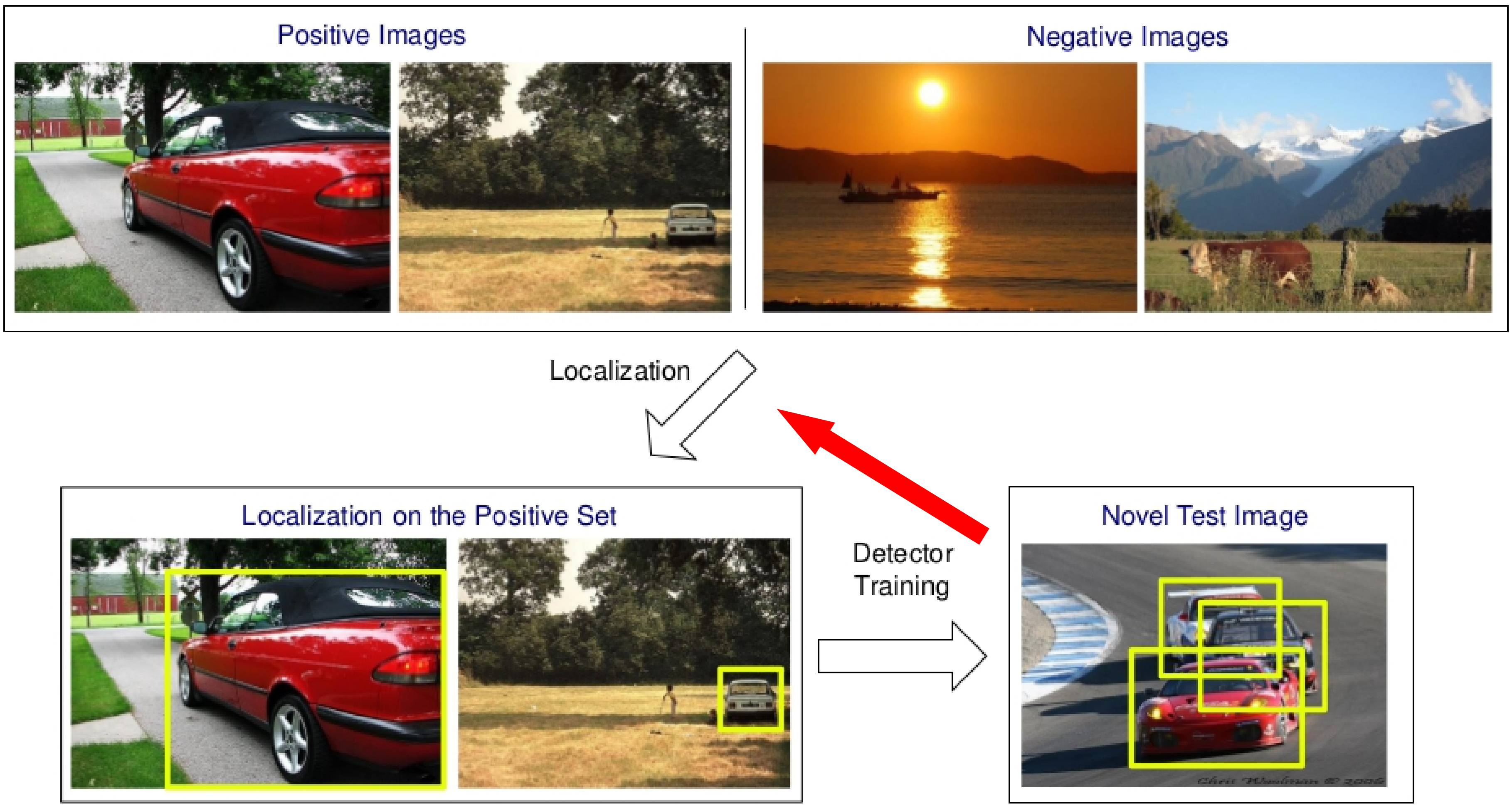
Presentation outline

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning
- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement
- Experimental evaluation results



Learning from incomplete supervision

- A joint identification problem:
  - Locating object instances in positive images
  - Learning detector from positive and negative examples



State-of-the-art weakly-supervised detector training

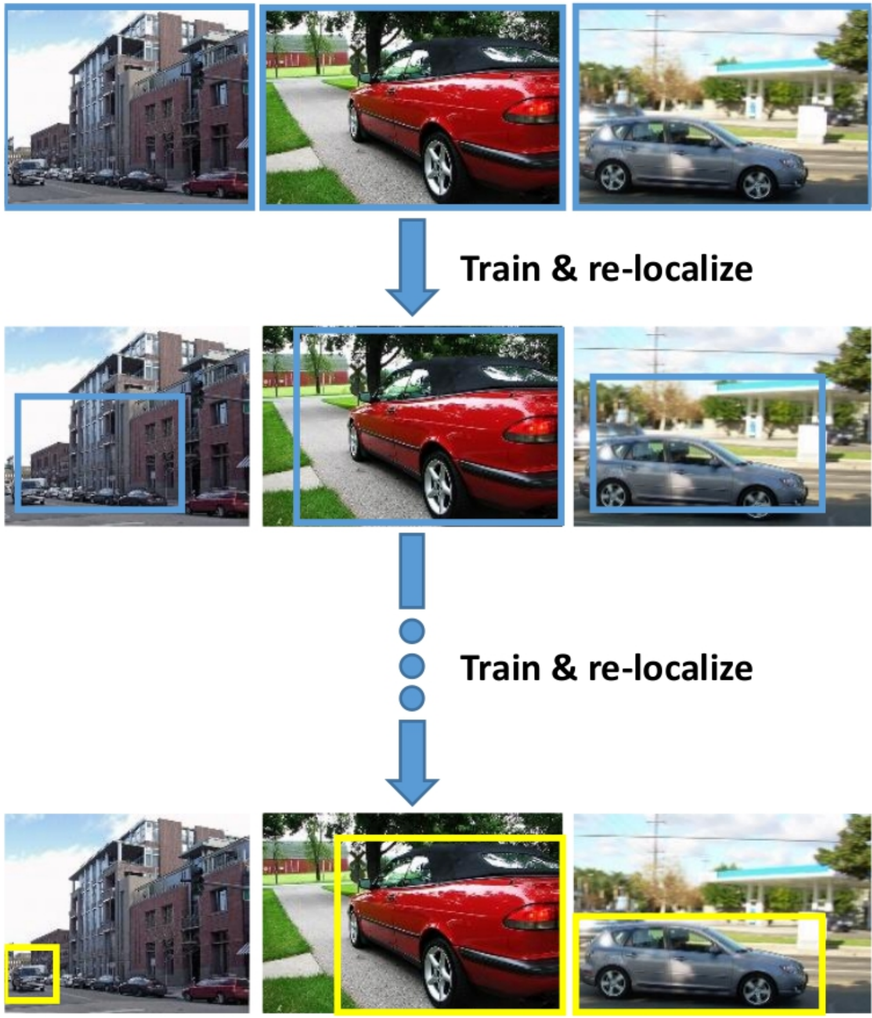
- Vast majority of work relies on multiple-instance learning

  Pandey & Lazebnik 2011, Siva et al. 2011, 2012, 2013, Russakovsky et al. 2012, Shi et al. 2013, ...

- Approaches vary in terms of
  - Initialization strategy
  - Object descriptors and detector
  - Utilization of pair-wise window similarities

- Some alternative recent approaches are based on topic models, e.g. LDA

  Shi, Hospedales, Xiang, ICCV 2013.
  Wang, Ren, Huang, Tan, ECCV 2014.



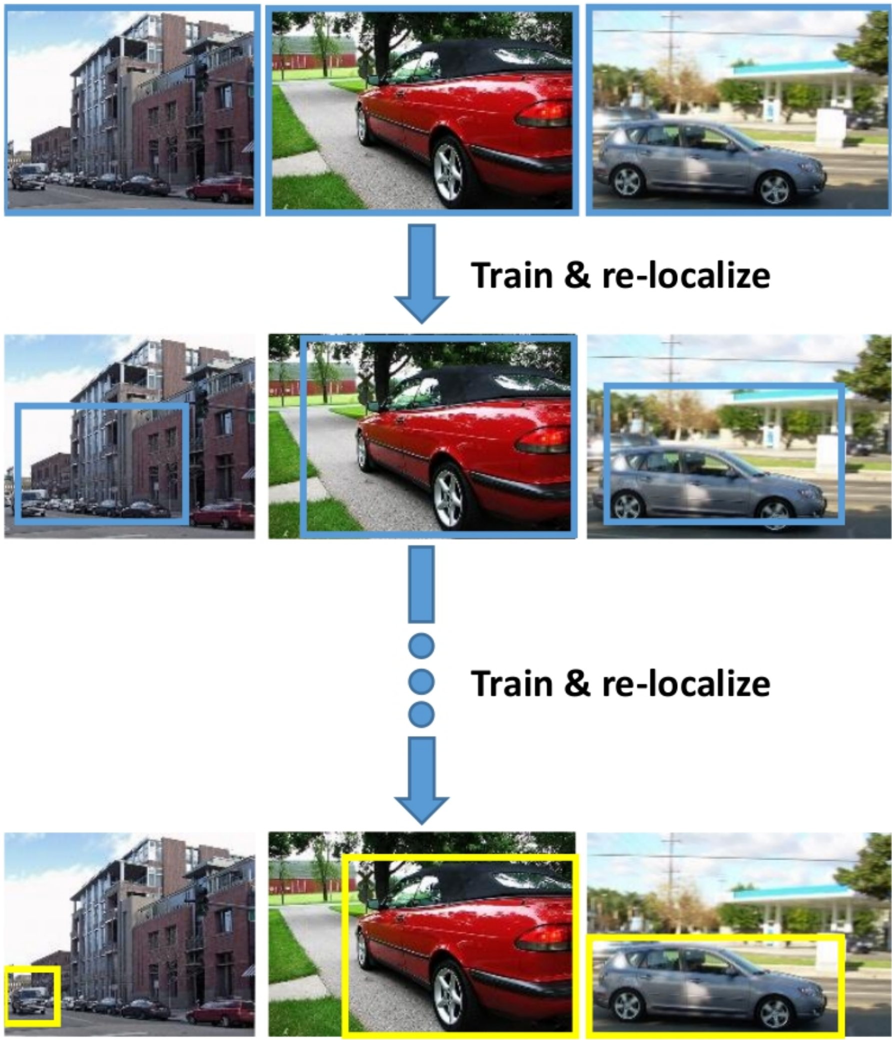
The multiple instance learning (MIL) approach

- Examples come in labeled "bags"

  [Dietterich et al., *Artif. Intell.,* 1997]
  - Positive bags contain at least one positive sample
  - Negative bags only contain negative samples

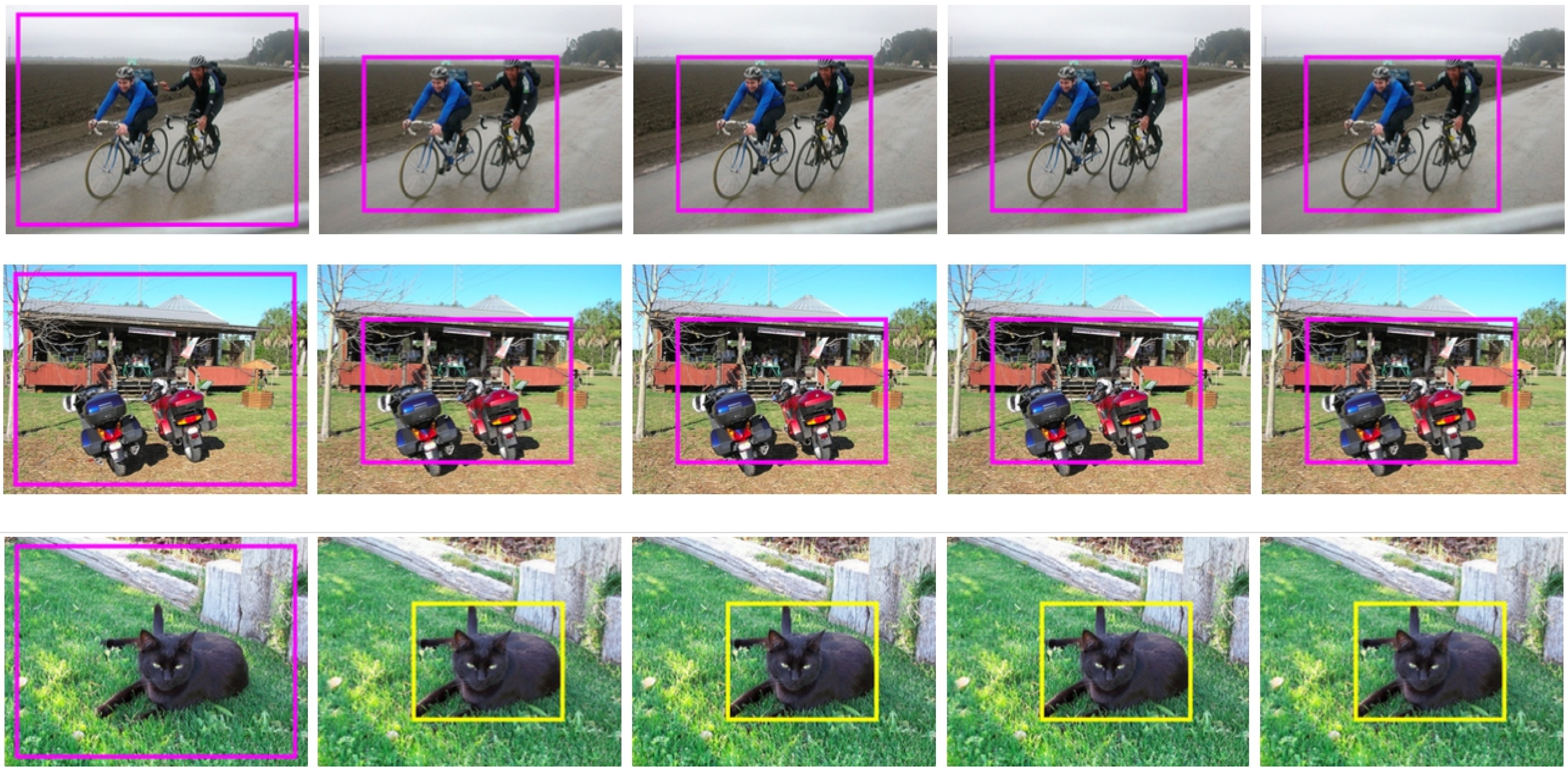
- Multiple Instance SVM

  [Andrews et al., NIPS 2002]
  - Initialize initial selection of samples from positive bags
  - Train SVM with selection
  - Select top scoring sample in each positive bag
  - Repeat until convergence



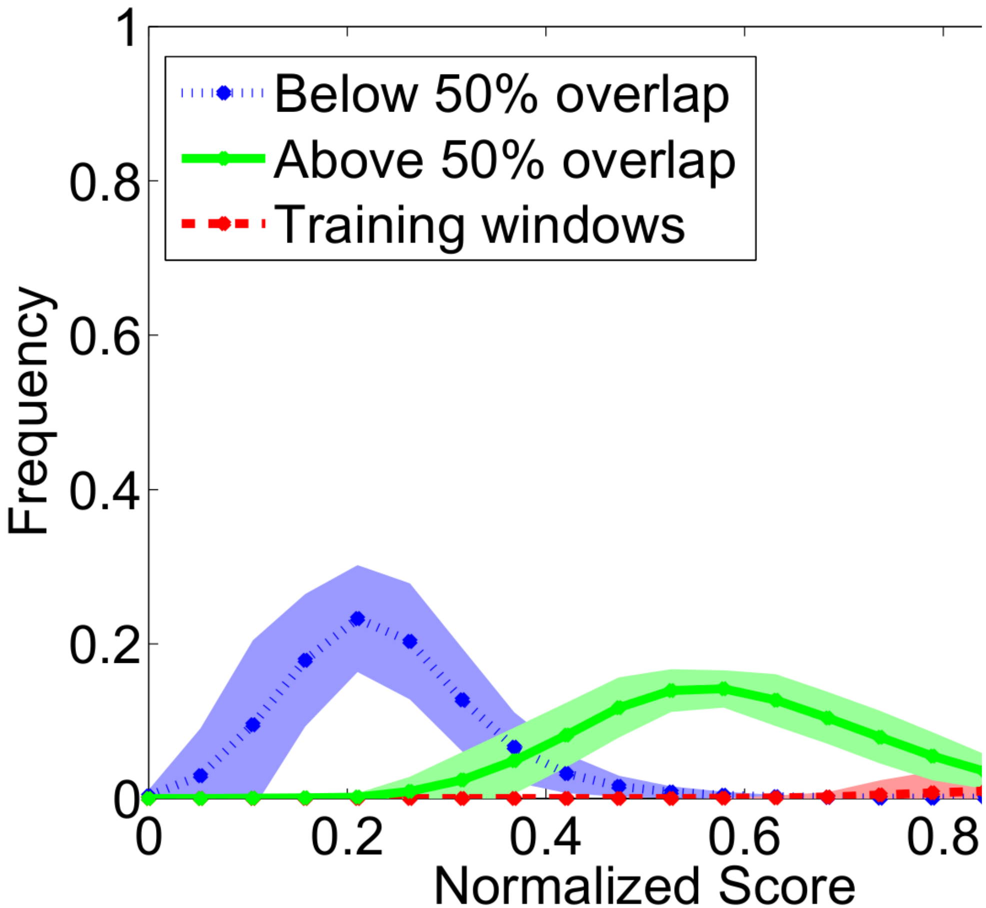
Multiple instance learning in practice...

- Converges rapidly to poor local optima



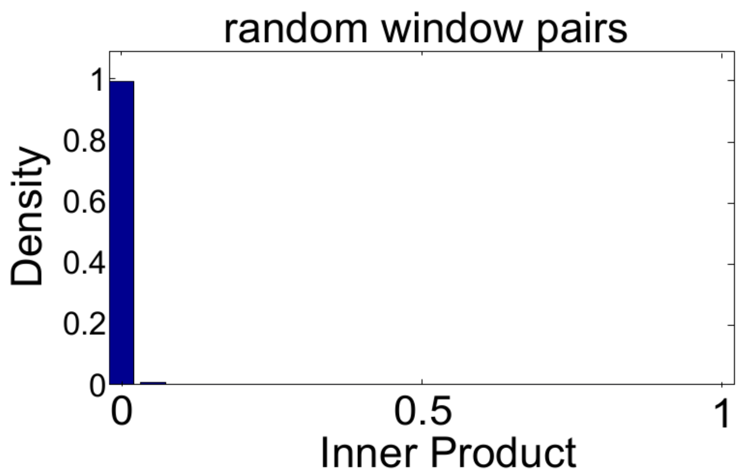
The problems in multiple instance learning

- Given a trained detector, consider score of windows that
  - do not match true objects
  - do match true objects
  - were used as positive samples to train the detector (might be wrong)



Problems in standard multiple instance learning

- Our window descriptors are high dimensional
  - Descriptors are L2 normalized
  - Most pairs are near orthogonal, i.e. near-zero dot products



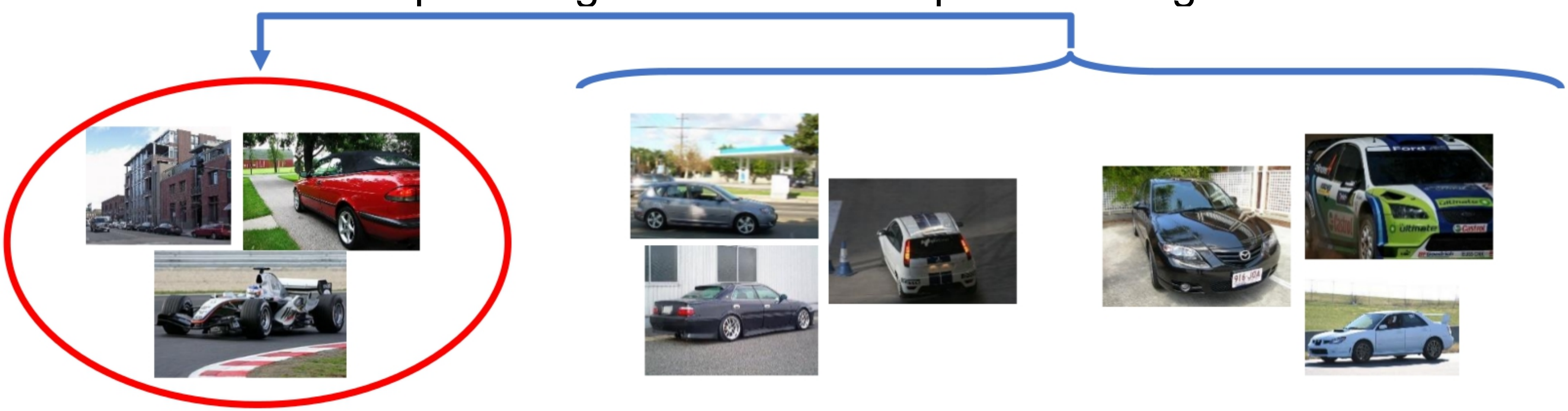
- Linear classifier score is weighted sum of dot products

$$w^T x = \sum_i \alpha_i (x_i^T x)$$

- Classifier scores much higher for positive windows used in training
  - This causes the degenerate re-localization behavior

Multi-fold training for multiple instance learning

- Separate sets of positive images for training and re-localization
  - Negative images do not need to be split, since no relocalization there
- Repeat two steps
  - Partition positive training images into K folds
  - For fold k = 1,...,K
    - Train detector from all training images, except those in fold k
    - Select top-scoring window in each positive image in fold k



- Avoids the re-localization bias since images used for training and re-localization are always different

Multi-fold training for multiple instance learning algorithm

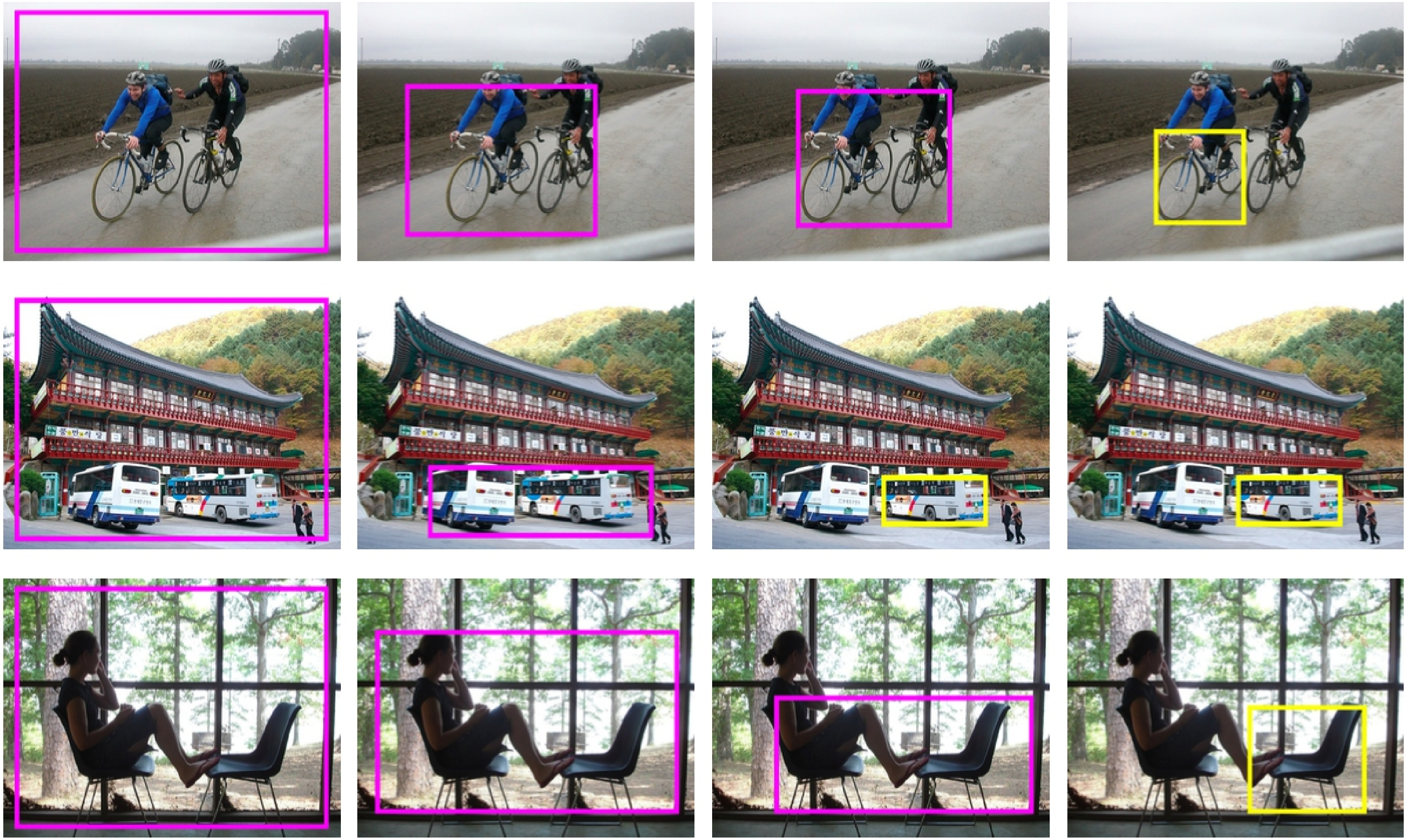
Algorithm 1 — Multi-fold weakly supervised training

---

1) Initialization: positive and negative examples are set to entire images
2) For iteration $t = 1$ to T
   a) Divide positive images randomly into K folds.
   b) For $k = 1$ to K
      i) Train using positive examples in all folds but k, and all negative examples.
      ii) Re-localize positives by selecting the top scoring window in each image of fold k using this detector.
   c) Train detector using re-localized positives and all negative examples.
   d) Add new negative windows by hard-negative mining.
3) Return final detector and object windows in train data.

Multi-fold training for multiple instance learning

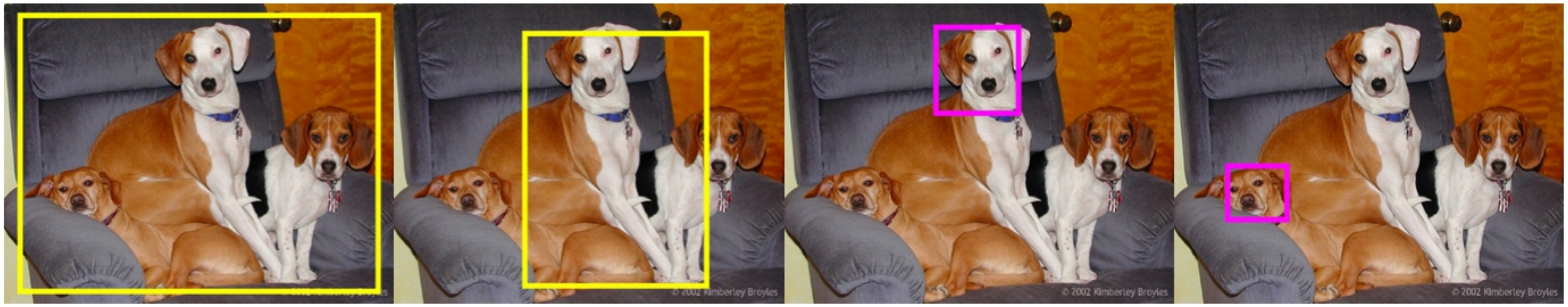
- Resolves the degenerate re-localization of standard MIL training



Initialization | Iteration 1 | Iteration 4 | Iteration 11

Limitation of weakly supervised learning

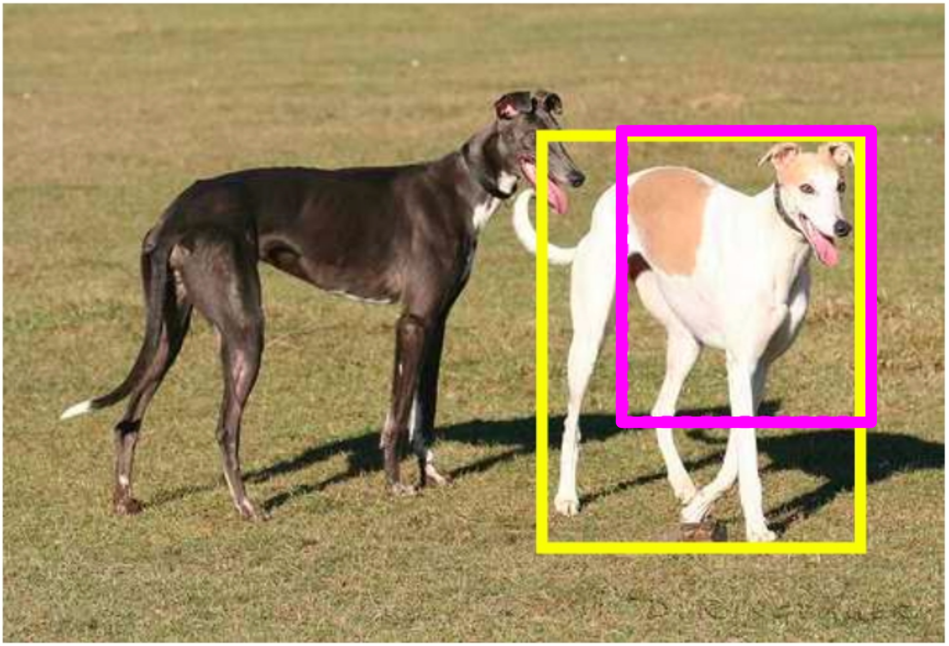
- Weakly supervised learning learns the most discriminative pattern between the positive and negative images
- These patterns may correspond to parts instead of full objects
  - For example the faces of cats and dogs due to body poses

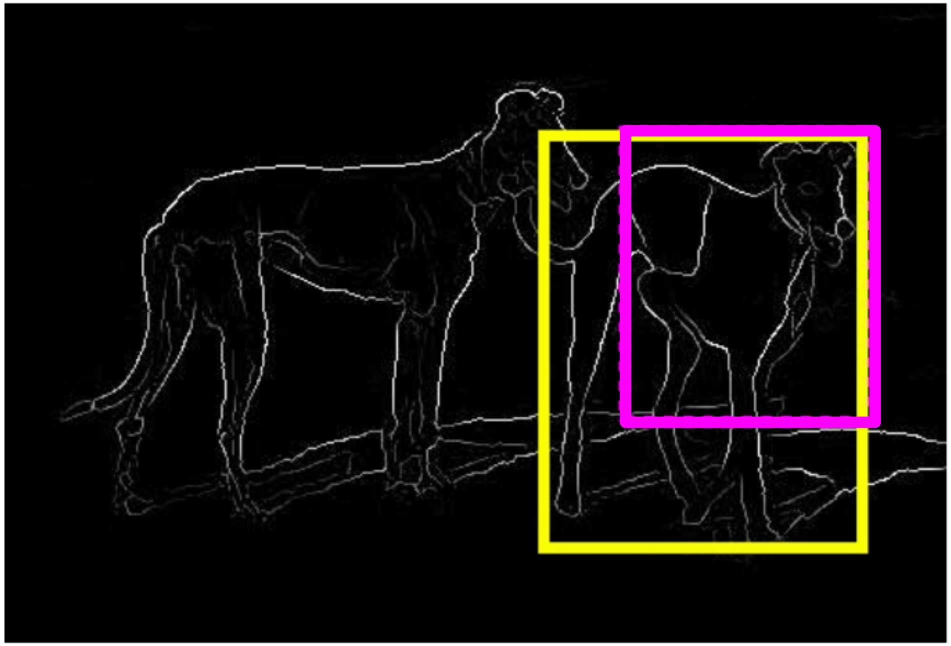


Hypothesis refinement using low-level contours

- Encourage object hypotheses to align with long image contours
  - Using efficient contour alignment score [Zitnick & Dollar, ECCV'14]

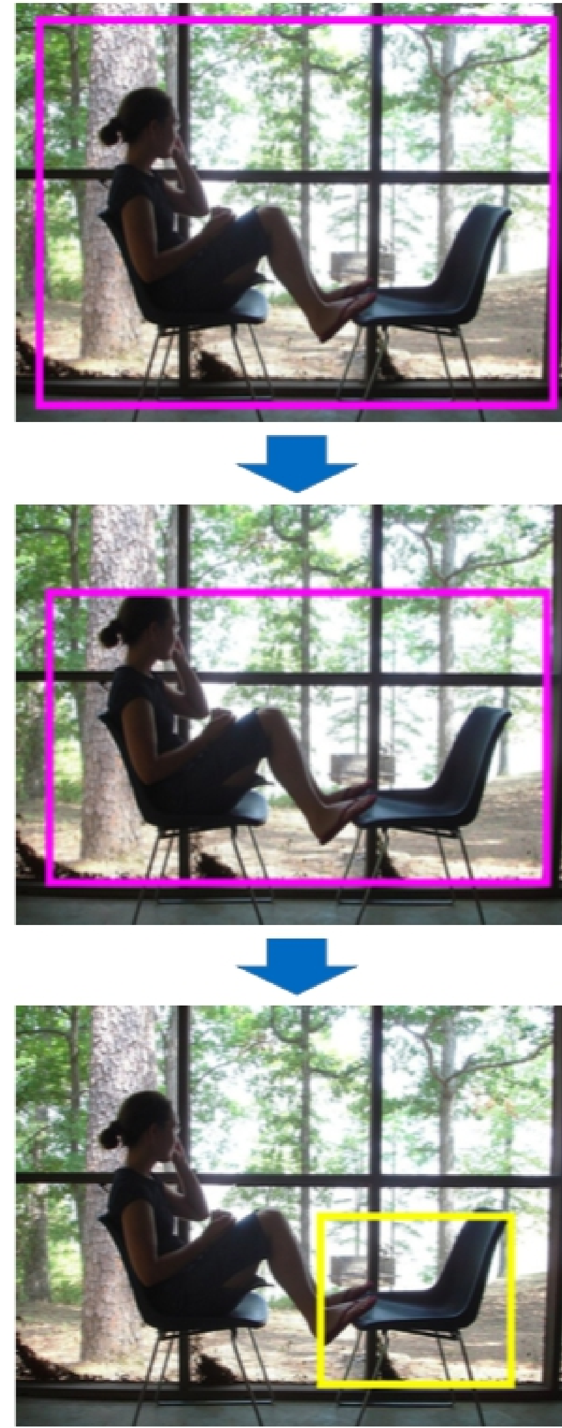
- After multi-fold training iterations: use weighted combination of detection and contour alignment score
- Final detector trained using the refined hypotheses





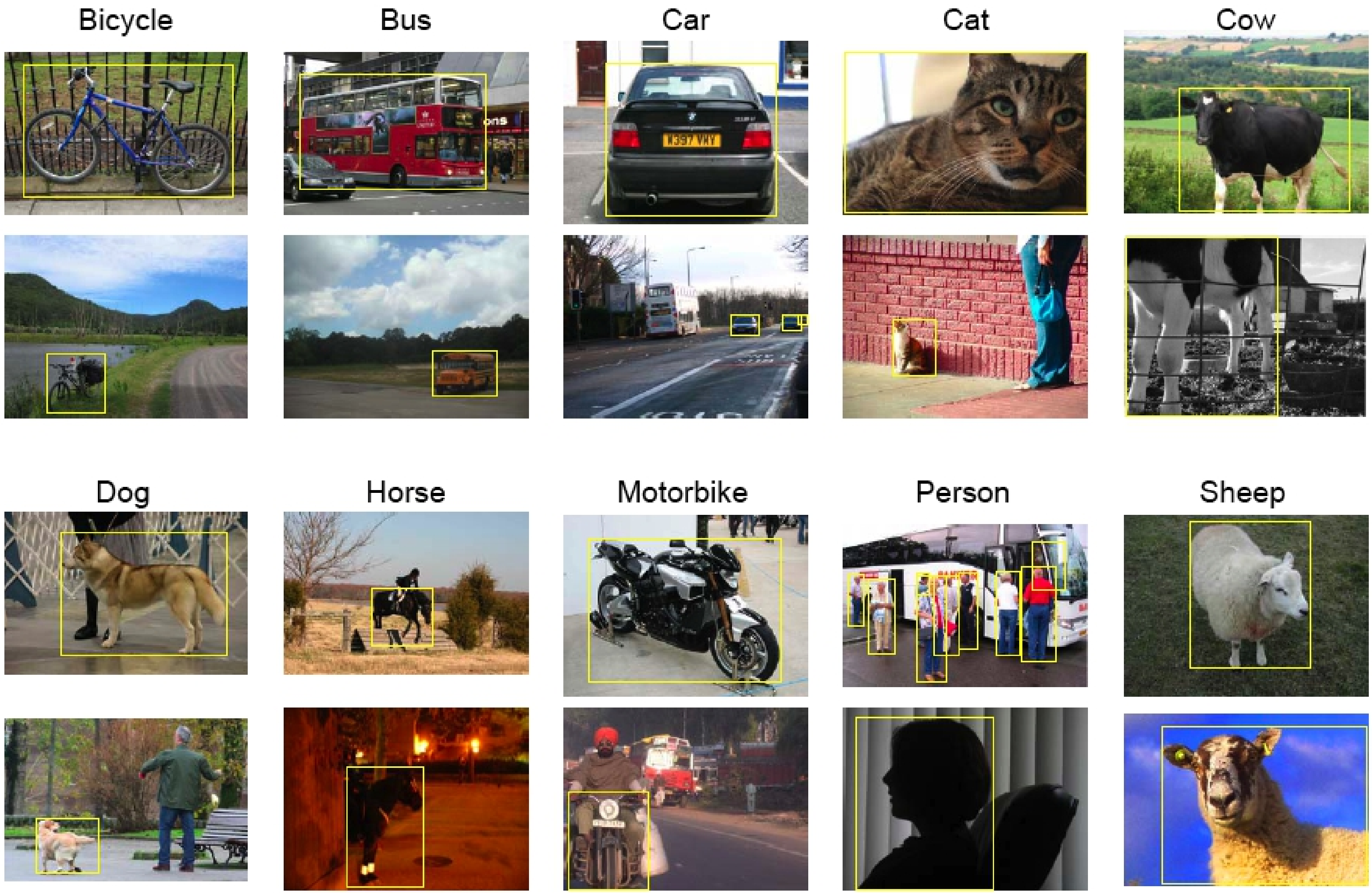
Presentation outline

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning
- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement
- Experimental evaluation results



Evaluations based on PASCAL VOC'07 benchmark

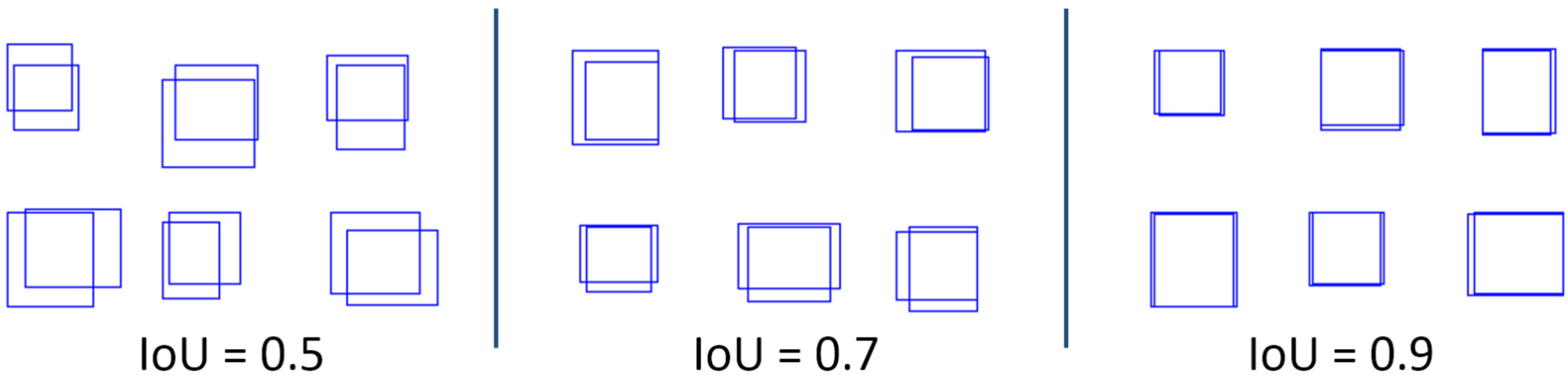
- Most challenging dataset for weakly supervised detection



Evaluation protocols

- Localization success on positive training images
  - Fraction of images with correct localization (**CorLoc**) [Deselaers et al., PAMI 2012]
- Standard PASCAL-VOC detection Average Precision (**AP**) on test set
- Both measures averaged over 20 different object categories

- Detection declared a success if highly overlapping with ground-truth
  - Intersection-over-union of window areas larger than 50%



Evaluation of multi-fold training

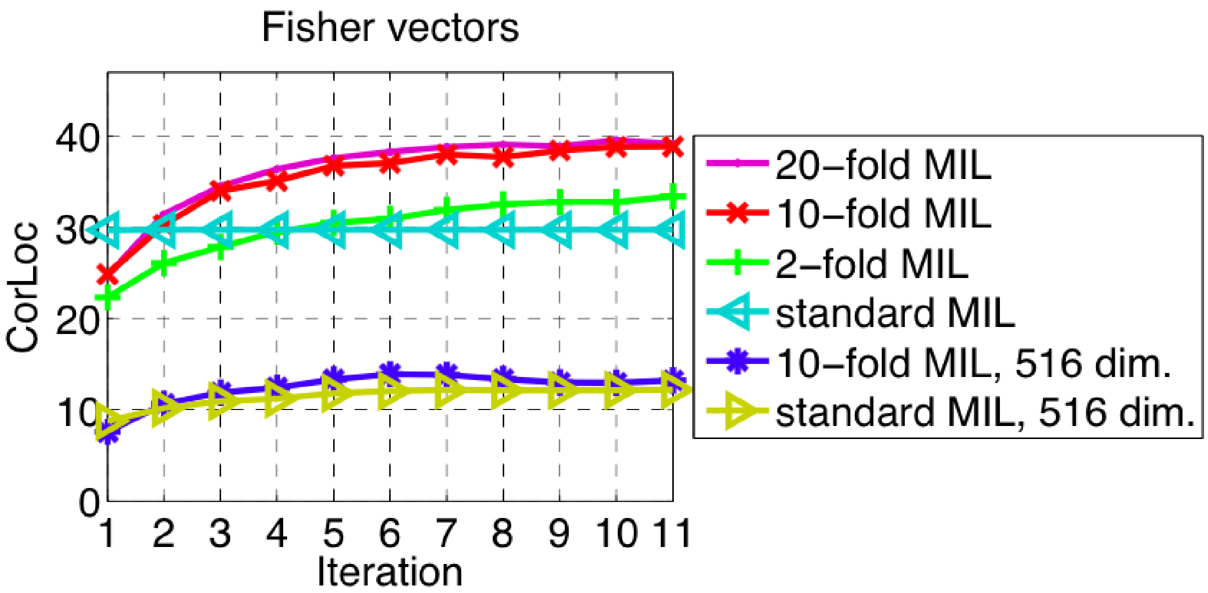
- Comparison of standard MIL training and multi-fold strategy

- Multi-fold training improves both performance measures using either Fisher vector or CNN features

	Standard	Multi-fold
	CorLoc	
FV	29.7	38.8 (+9.1)
CNN	41.2	45.0 (+3.8)
	Detection AP	
FV	15.5	22.4 (+6.9)
CNN	24.3	25.9 (+1.6)

Evaluation of multi-fold training

- CorLoc over the re-training / re-localization iterations
- Iteration n: n-th iteration after initialization from full image



Window refinement and combining features

- Contour alignment score improves performance
- Combining features boosts performance

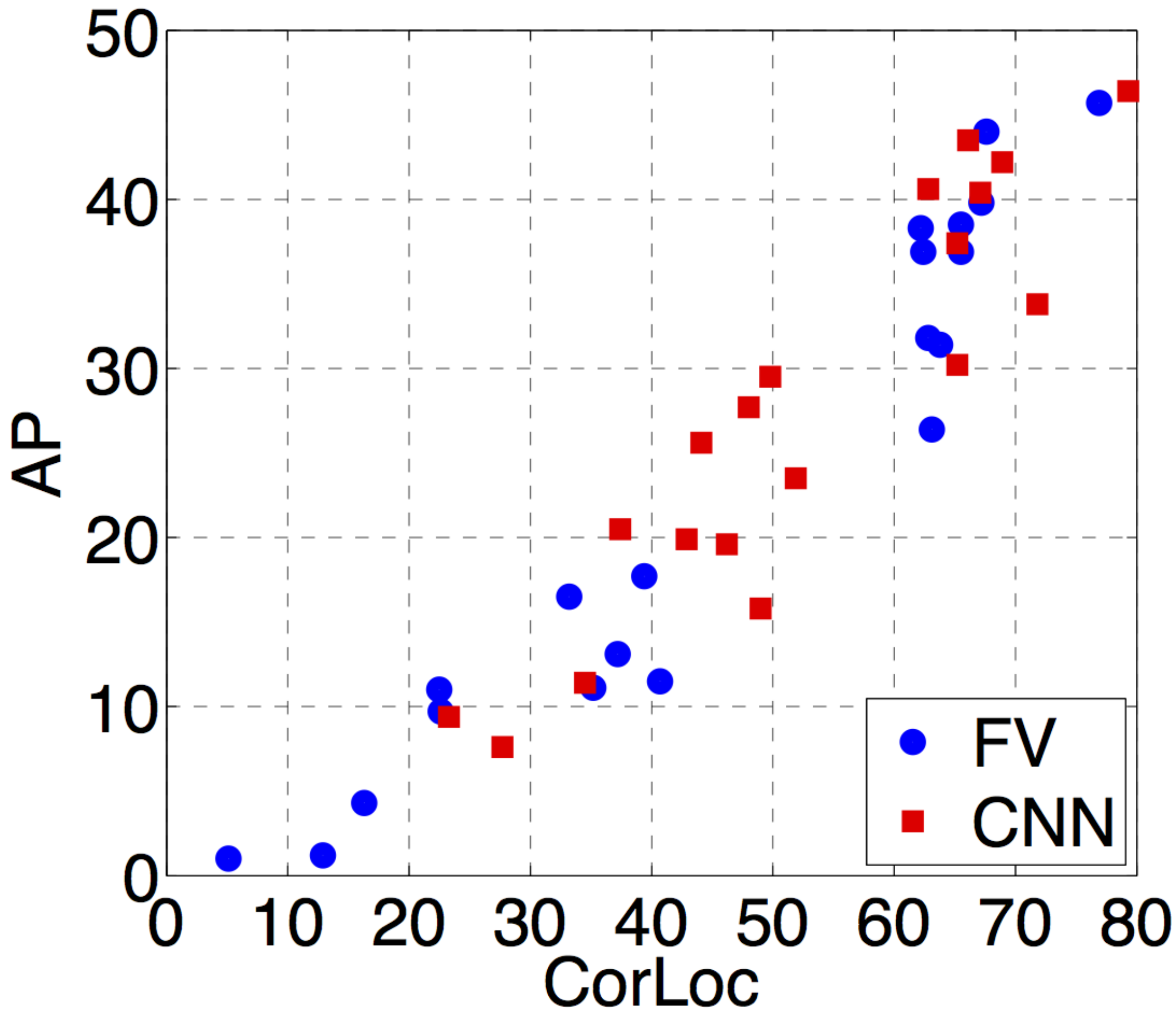
Refinement	No	Yes
	Detection AP	
FV	22.4	23.3 (+0.9)
CNN	25.9	28.6 (+2.7)
FV+CNN	27.4	30.2 (+2.8)

- Classes with largest improvements due to contour alignment

Refinement	No	Yes
	CorLoc for FV+CNN	
Horse	55.6	70.5 (+14.9)
Dog	37.3	48.4 (+11.4)
Cat	24.8	35.6 (+10.8)

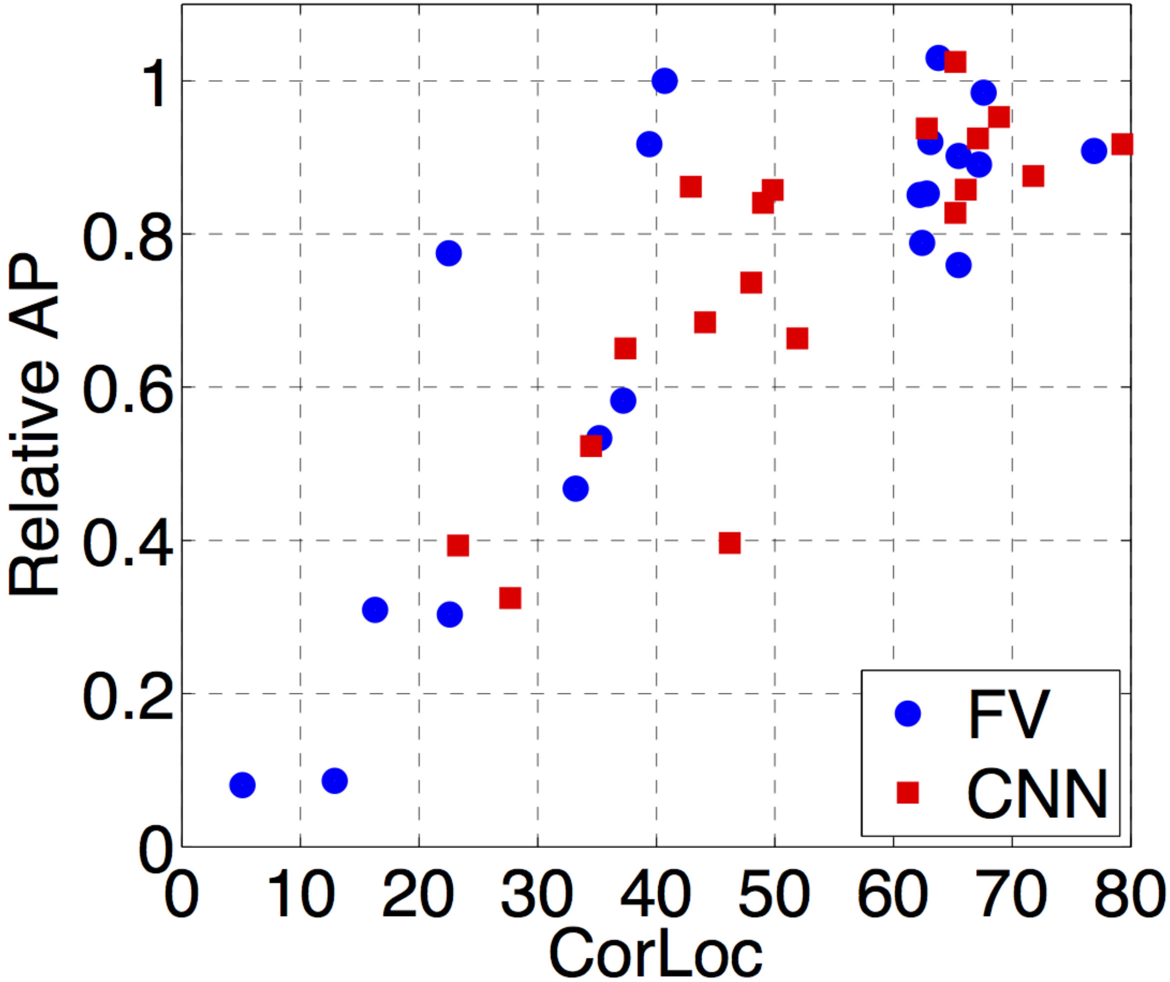
The relation between CorLoc and detection AP

- Relation between localization during training and final test performance
  - Each of the 20 classes gives a point on the graph



Relative performance of weakly supervised learning

- Ratio of detection AP with weakly supervised training (image-labels) and AP with same detector trained from bounding box annotations
  - Each point represents one object category



Overview of the state of the art

- Methods divided into those that use external training data to learn CNN features and those that do not

	aero	bicy	bird	boa	bot	bus	car	cat	cha	cow	dtab	dog	hors	mbik	pers	plnt	she	sofa	trai	tv	Av.
Pandey and Lazebnik'11 [32]	11.5	—	—	3.0	—	—	—	—	—	—	—	—	20.3	9.1	—	—	—	—	13.2	—	—
Siva and Xiang'11 [42]	13.4	44.0	3.1	3.1	0.0	31.2	43.9	7.1	0.1	9.3	9.9	1.5	29.4	38.3	4.6	0.1	0.4	3.8	34.2	0.0	13.9
Russakovsky *et al*.'12 [35]	30.8	25.0	—	3.6	—	26.0	—	—	—	—	—	—	21.3	29.9	—	—	—	—	—	—	15.0
Ours (FV-only)	36.9	38.3	11.5	11.1	1.0	39.8	45.7	16.5	1.2	26.4	4.3	17.7	31.8	44.0	13.1	11.0	31.4	9.7	38.5	36.9	23.3
	methods using additional training data																				
Song *et al*.'14 [43]	27.6	41.9	19.7	9.1	10.4	35.8	39.1	33.6	0.6	20.9	10.0	27.7	29.4	39.2	9.1	19.3	20.5	17.1	35.6	7.1	22.7
Song *et al*.'14 [44]	36.3	**47.6**	23.3	12.3	11.1	36.0	46.6	25.4	0.7	23.5	12.5	23.5	27.9	40.9	14.8	19.2	24.2	17.1	37.7	11.6	24.6
Bilen *et al*.'14 [6]	42.2	43.9	23.1	9.2	**12.5**	44.9	45.1	24.9	8.3	24.0	13.9	18.6	31.6	43.6	7.6	**20.9**	26.6	20.6	35.9	29.6	26.4
Wang *et al*.'14 [50]	48.8	41.0	23.6	12.1	11.1	42.7	40.9	**35.5**	**11.1**	**36.6**	18.4	**35.3**	34.8	51.3	17.2	17.4	26.8	32.8	35.1	45.6	30.9
Wang *et al*.'14 [50] +context	**48.9**	42.3	26.1	11.3	11.9	41.3	40.9	34.7	10.8	34.7	18.8	34.4	35.4	**52.7**	19.1	17.4	**35.9**	**33.3**	34.8	**46.5**	**31.6**
Ours	39.3	43.0	**28.8**	**20.4**	8.0	**45.5**	**47.9**	22.1	8.4	33.5	**23.6**	29.2	**38.5**	47.9	**20.3**	20.0	35.8	30.8	**41.0**	20.1	30.2

- Results comparable with the state of the art (with CNN features), or better when no external training data is used

- A lot of improvement in performance of weakly supervised detection in recent years: AP values have doubled !

Conclusion

- Presented a state-of-the-art weakly supervised object detection method
  - Strong appearance cues for recognition: FV and CNN descriptors
  - Re-localization bias suppression: Multi-fold MIL training
  - Localization refinement: alignment with long contours
- Future directions:
  - Dealing with noise on the image labels
  - Concurrent training of categories: leverage explaining away
  - Richer interactions between recognition and segmentation

Object category localization with incomplete supervision

Jakob Verbeek

LEAR team, INRIA, Grenoble, France

Joint work with: Gokberk Cinbis and Cordelia Schmid

informatics mathematics