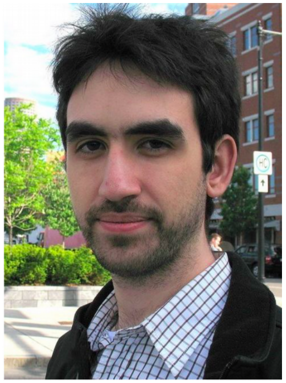
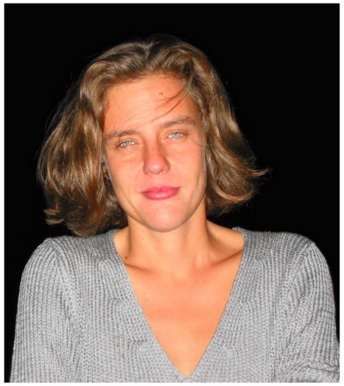
Object Detection with Incomplete Supervision

Jakob Verbeek

LEAR team, INRIA, Grenoble, France

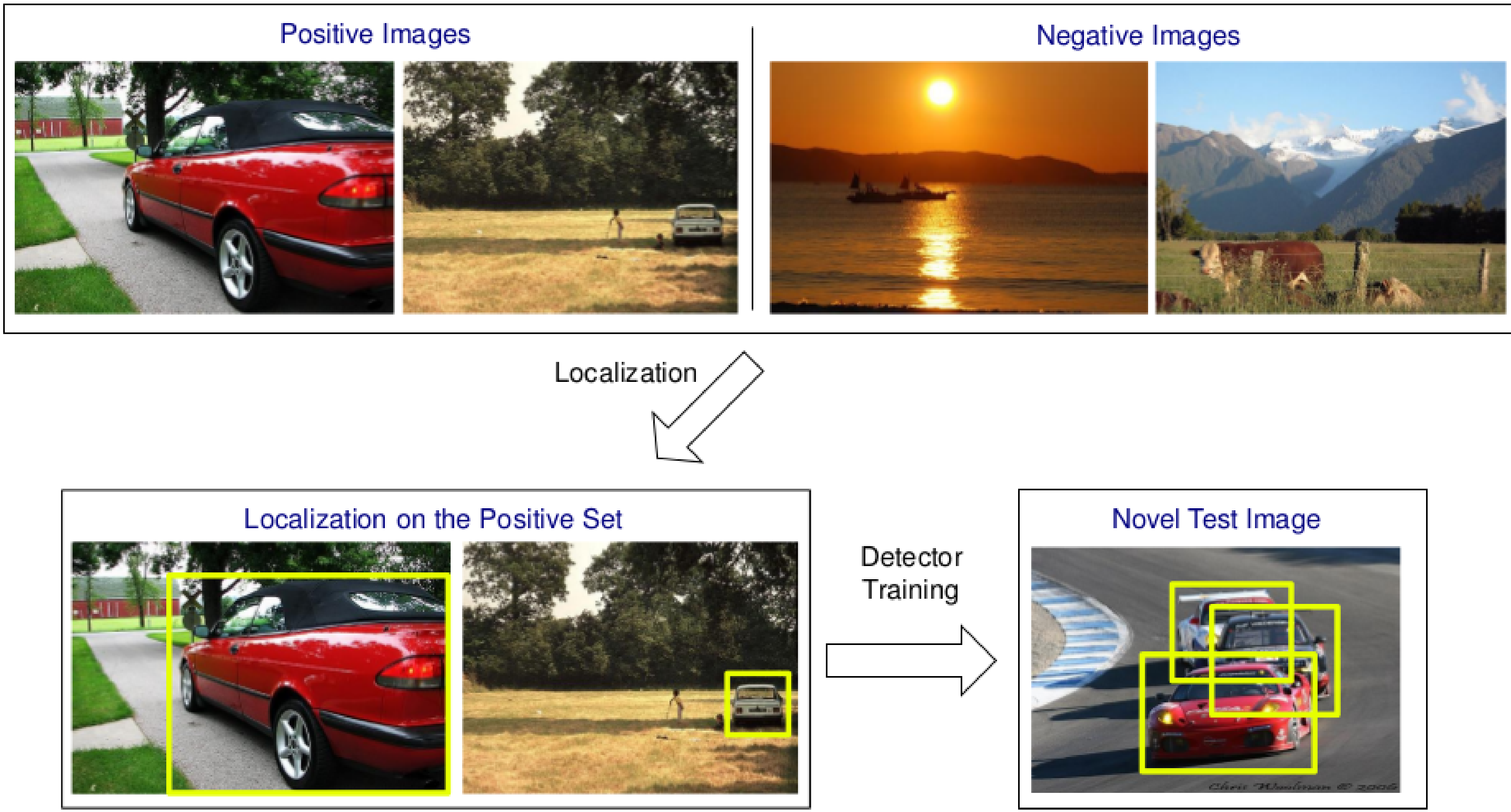
Joint work with: Gokberk Cinbis and Cordelia Schmid





Why learning from incomplete supervision?

- Fully supervised training requires costly bounding box annotations
- Weakly supervised learning only uses image-wide labels



Overview of this presentation

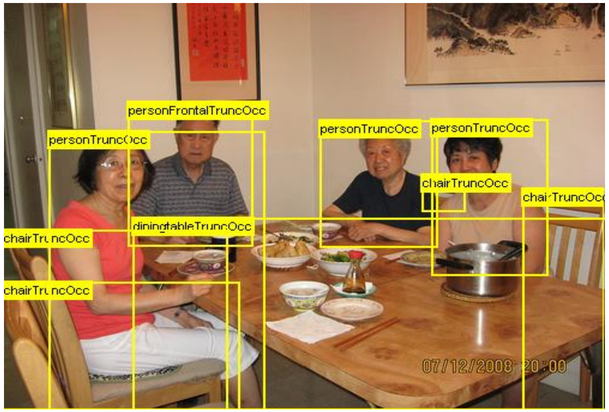
- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning

- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement

- Experimental evaluation and analysis

Challenging factors in object detection

- Intra-class appearance variation
  - Deformable objects: e.g. animals
  - Transparency: e.g. bottles
  - Sub-categories: e.g. ferry vs yacht

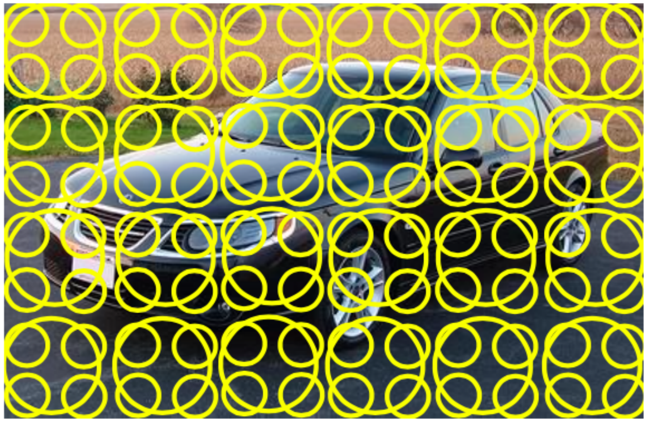


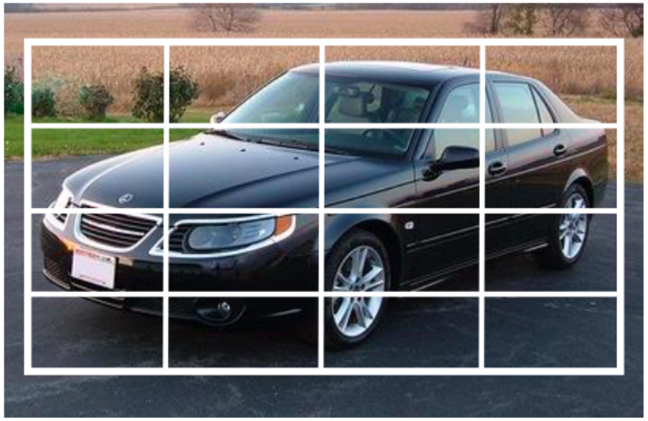
- Scene composition
  - Heavy occlusions: e.g. tables and chairs
  - Clutter: coincidental image content present in bounding box

- Imaging conditions
  - viewpoint, scale, lighting conditions

State-of-the-art visual representations

- Need for strong appearance features to separate classes despite strong intra-class variability and subtle inter-class variations
  - Consider deformability of cats and dogs
  - Similarity between furry cats and dogs in the similar poses





- Fischer vector representation

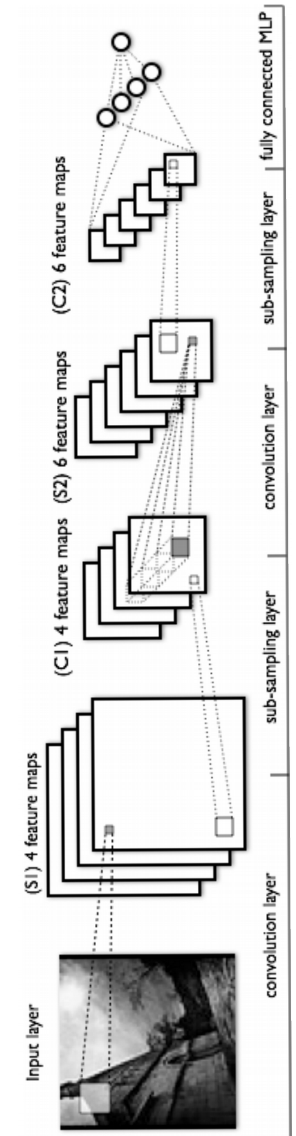
  [Sanchez et al., IJCV, 2013]
  - Local SIFT descriptors, PCA to 64 dim.
  - 64 component GMM for soft quantization
  - Record first and second order moments of features assigned to each Gaussian
  - 4x4 SPM grid, power and L2 normalization
  - 140K dimensional descriptor
  - PQ compression to reduce storage cost

State-of-the-art visual representations

- Need for strong appearance features to separate classes despite strong intra-class variability and subtle inter-class variations
  - Consider deformability of cats and dogs
  - Similarity between furry cats and dogs in the similar poses

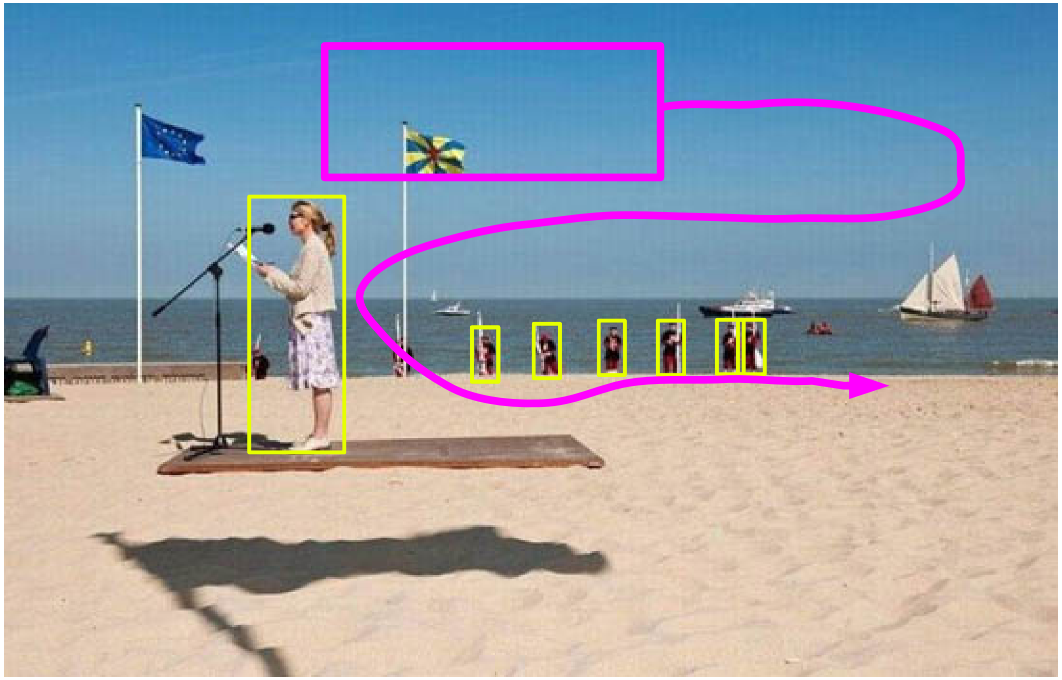
- Global Convolutional Neural Network feature

  [Jia et al., caffe.berkeleyvision.org]
  - Trained on 1000 ImageNet 2012 categories
  - Caffe framework
  - Use last shared layer for representation
  - Resize detection windows to 224x224 pixels
  - L2 normalization
  - 4K dimensional descriptor



A typical object detection system

- Training a binary classifier that will score object windows
  - Positives given by manual annotation (hundreds to thousands)
  - Potential pool of negatives outside positive boxes (zillions)
    - Repetitive access to find useful/hardest negative samples
    - Store or re-extract feature vectors of these examples
- At test image, classify windows of different shapes and sizes
  - Detection speed proportional to number of considered windows

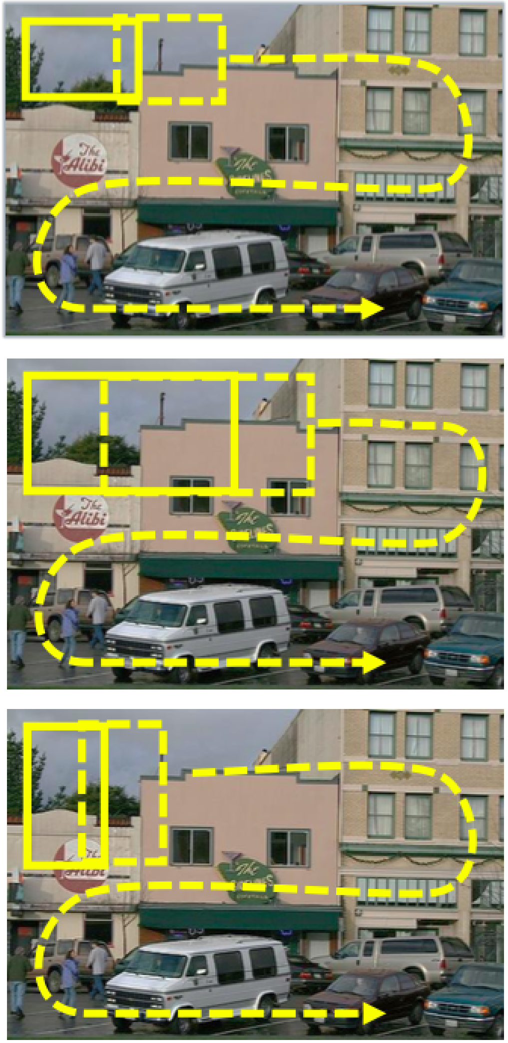


Issues with classic scanning windows

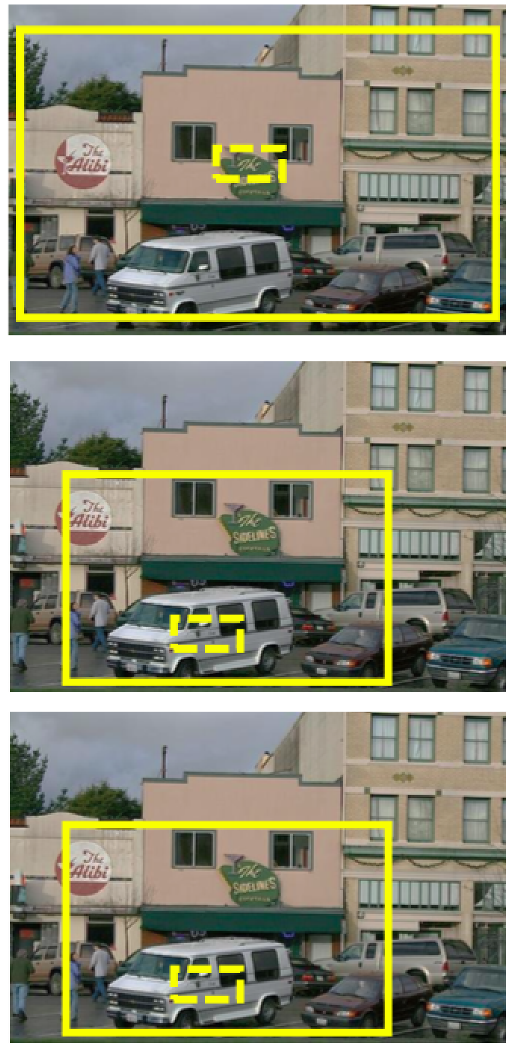
- Number of detection windows in an image is huge
  - Quadratic in image size

- Features are expensive to evaluate
- Features are expensive to store
- Alternatives to dense exhaustive search are needed

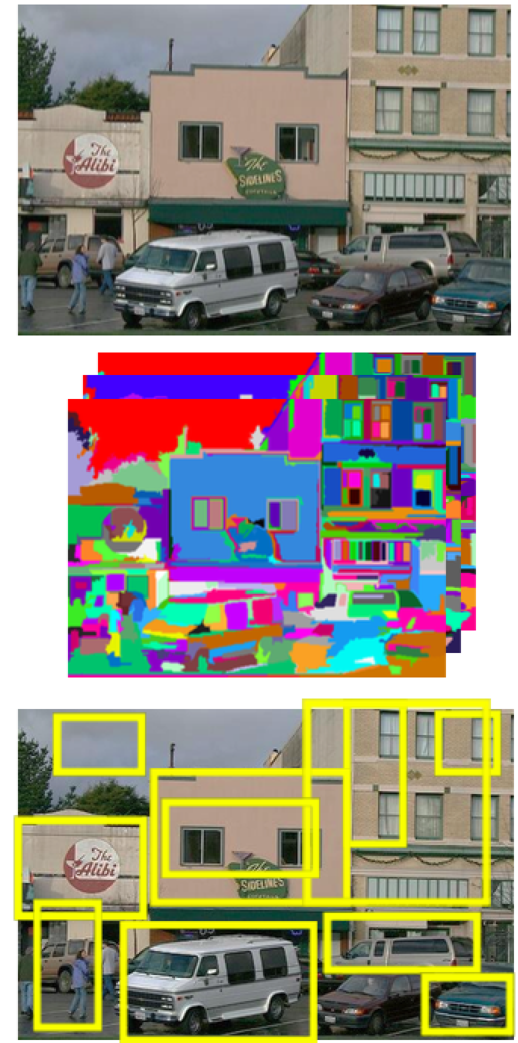
Alternatives to exhaustive sliding window search



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



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



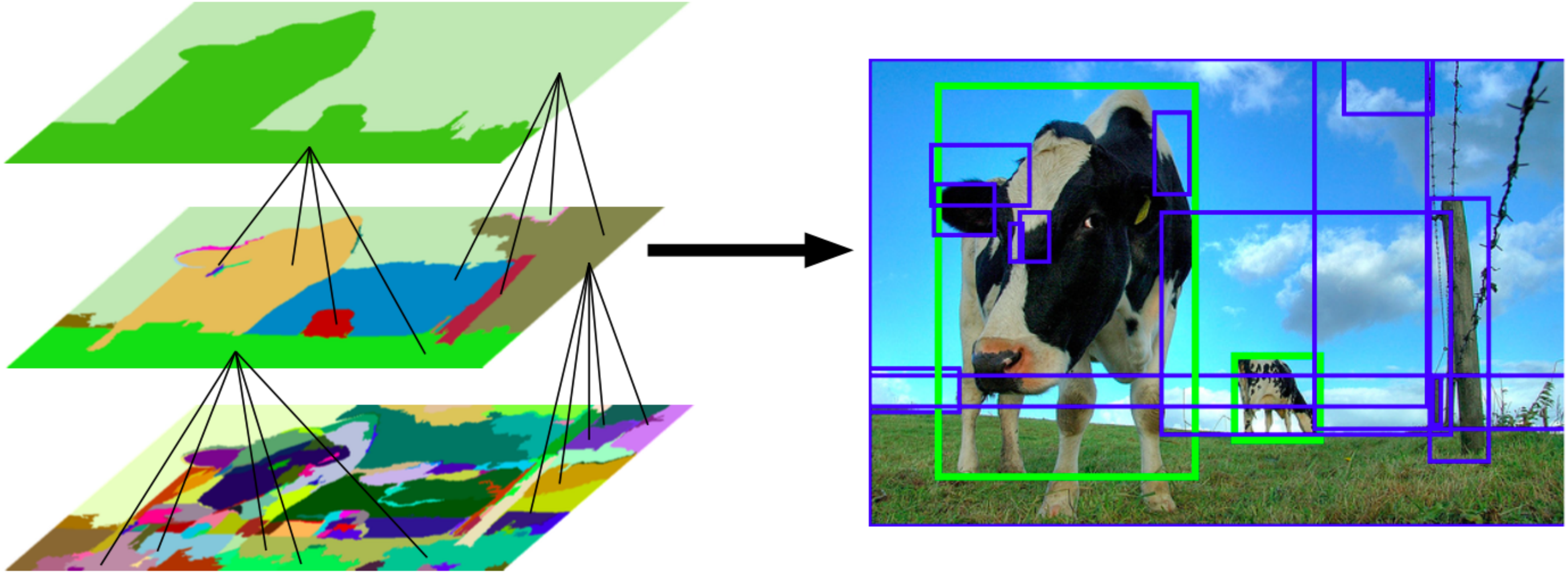
Selective Search
(Alexe *et al*. 2010;
Sande *et al*. 2011)

Alternatives to exhaustive sliding window search

- Branch-and-bound techniques
  - Imposes requirements on type of classifiers / features
    [Lampert, Blaschko, Hofmann, PAMI 2009]
- Feature cascades
  - Requires set of fast features in early stages
    [Viola & Jones, IJCV 2004]
- Coarse-to-fine search
  - Requires compositionality of classifier score
    [Felzenszwalb, Girshick, McAllester, CVPR 2010]
- Data driven generic object hypotheses
  - Consider boxes aligned with low-level image contours
  - Does not impose constraints on classifiers / features
    [Alexe, Deselaers, Ferrari, CVPR 2010]

Search: restricted scanning of bounding box space

- Selective search method [Uijlings et al., IJCV, 2013]
  - 1000 - 2000 windows per image
  - Covers over 95% of true objects with sufficient accuracy
  - Unsupervised multi-resolution hierarchical segmentation
  - Candidate detections generated as bounding box of segments
- Candidate windows used for hard negative mining and testing
- Feature compression using PQ codes and lossless compression

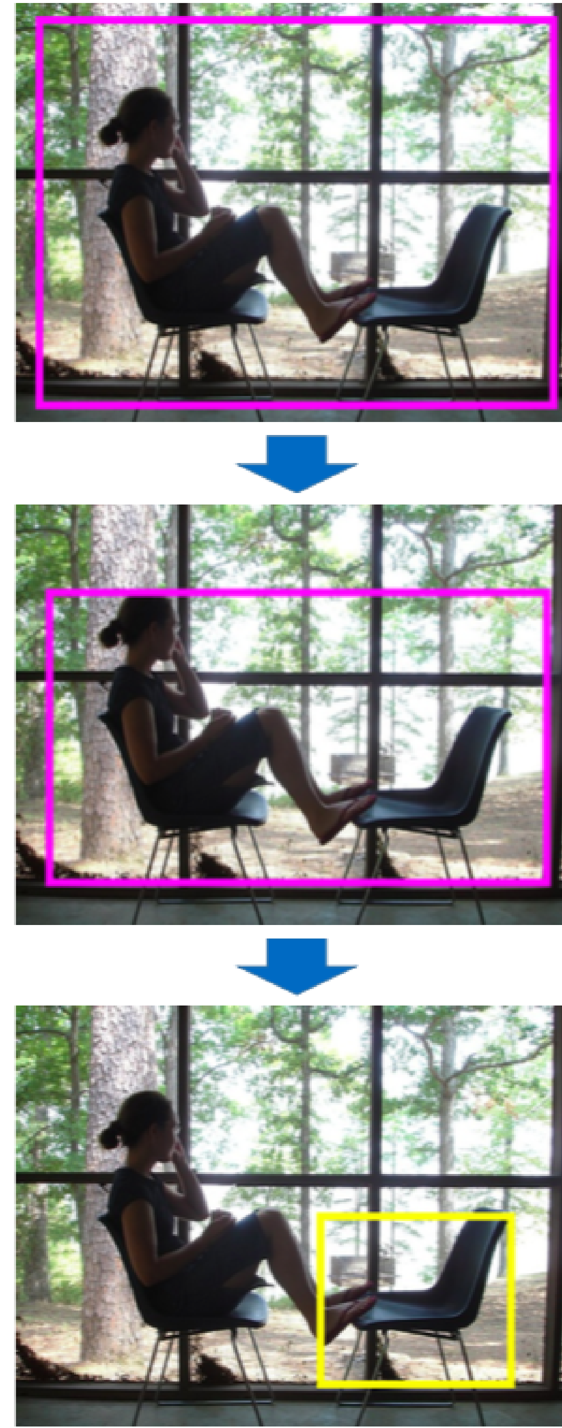


Overview of this presentation

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning

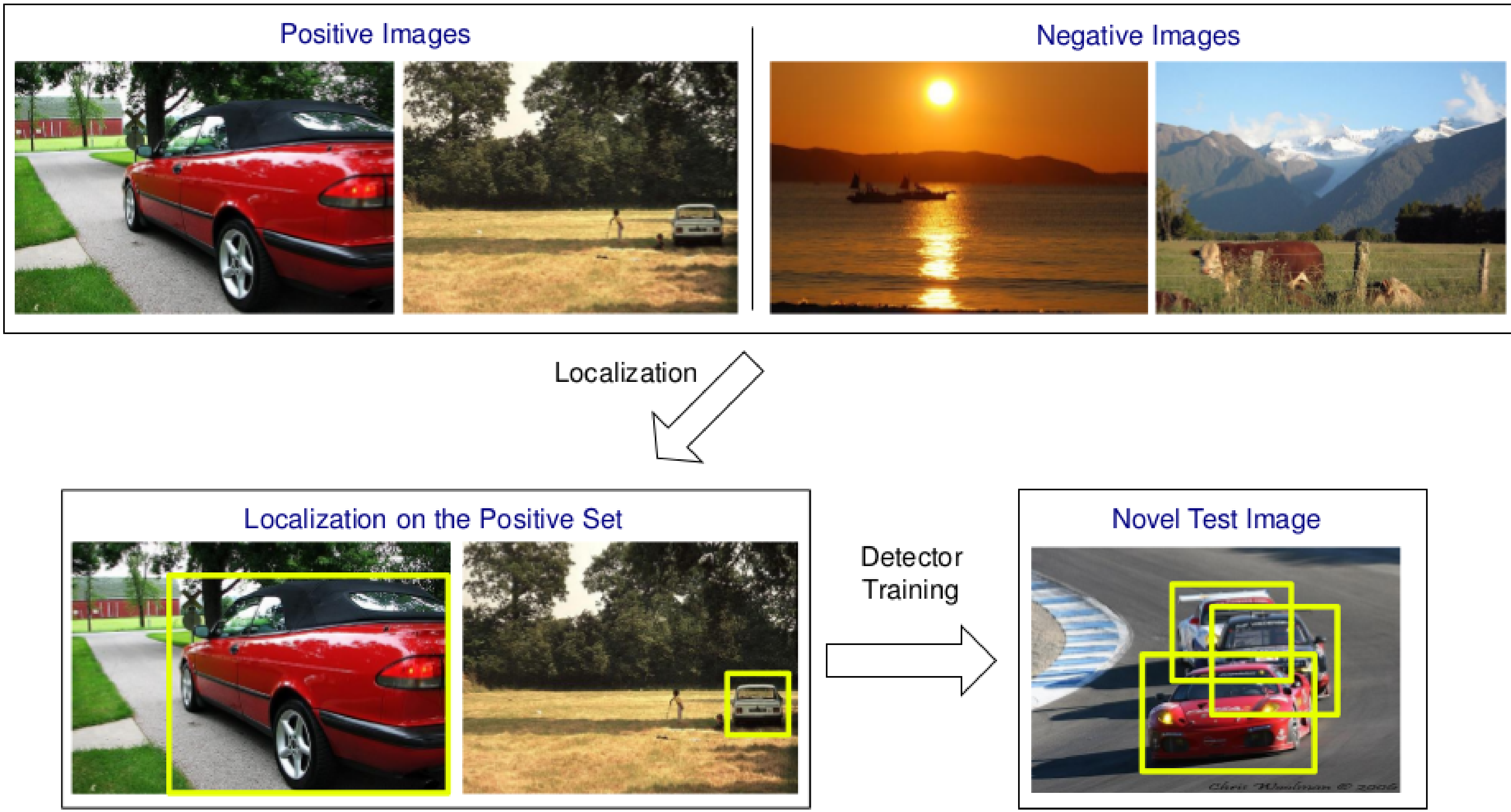
- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement

- Experimental evaluation and analysis



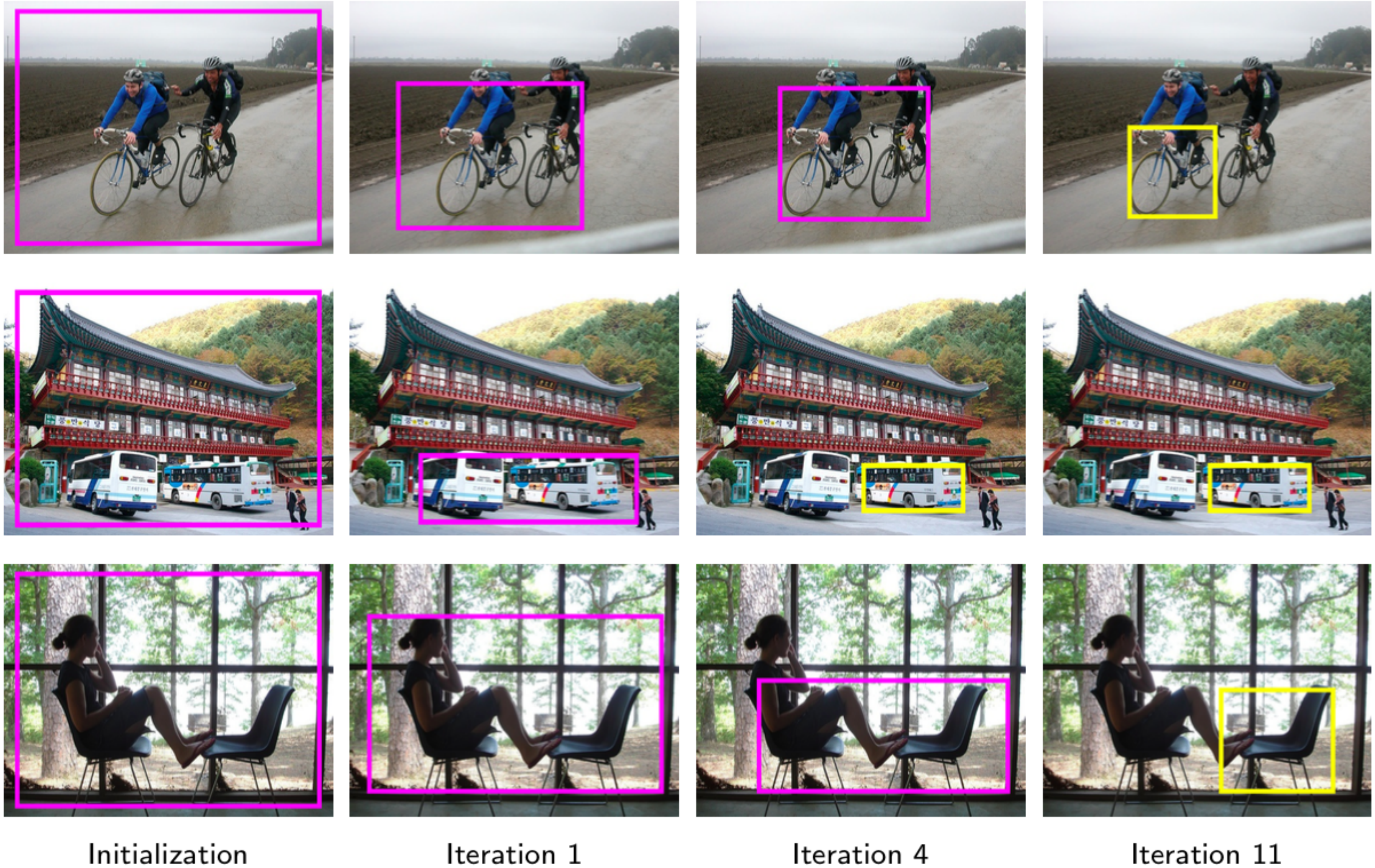
Why learning from incomplete supervision?

- Fully supervised training requires costly bounding box annotations
- Weakly supervised learning only uses image-wide labels



Learning from incomplete supervision

- Joint identification problem: recognition model and training instances
- Alternating optimization: fix one, optimize the other



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

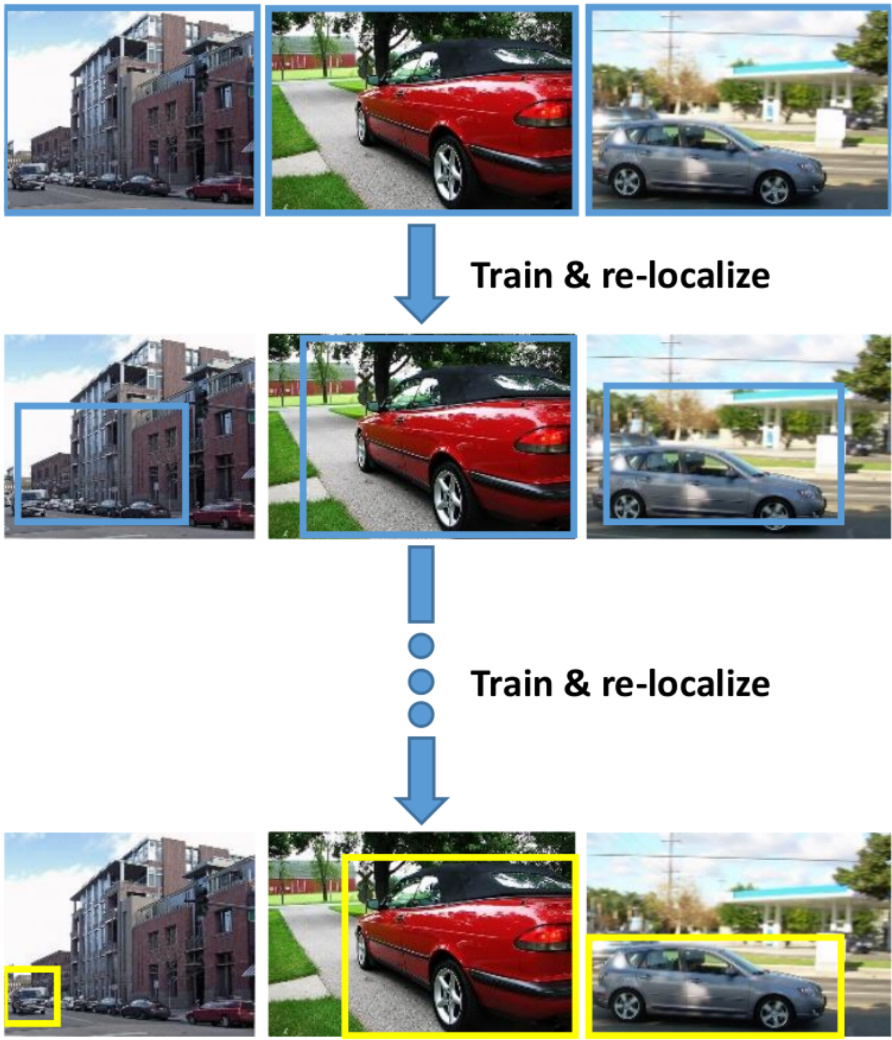
- Vast majority of work relies on multiple-instance learning

  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

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



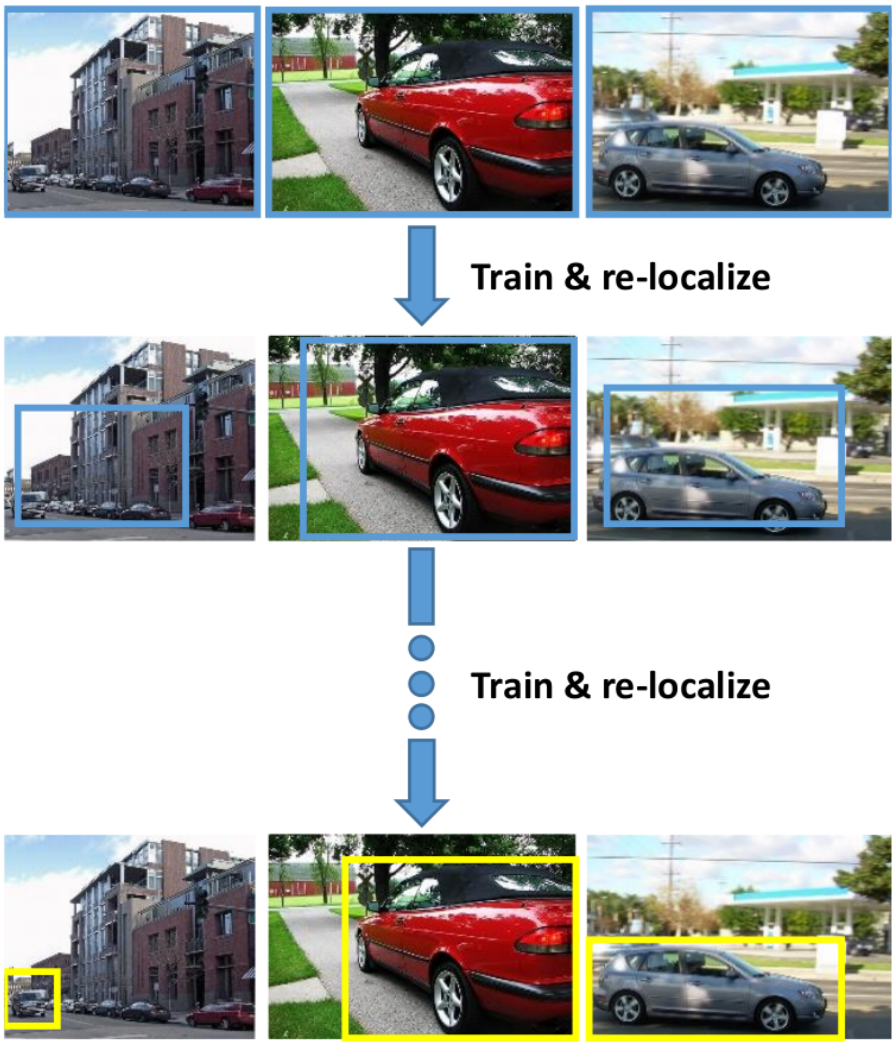
The multiple instance learning (MIL) approach

- Examples come in labeled "bags"

  Dietterich et al., *Artif. Intell.*, 1997
  - Selective search gives ~1500 windows per image = bag
  - Positive images contain at least one positive window
  - Negative images only have negative windows in the bag

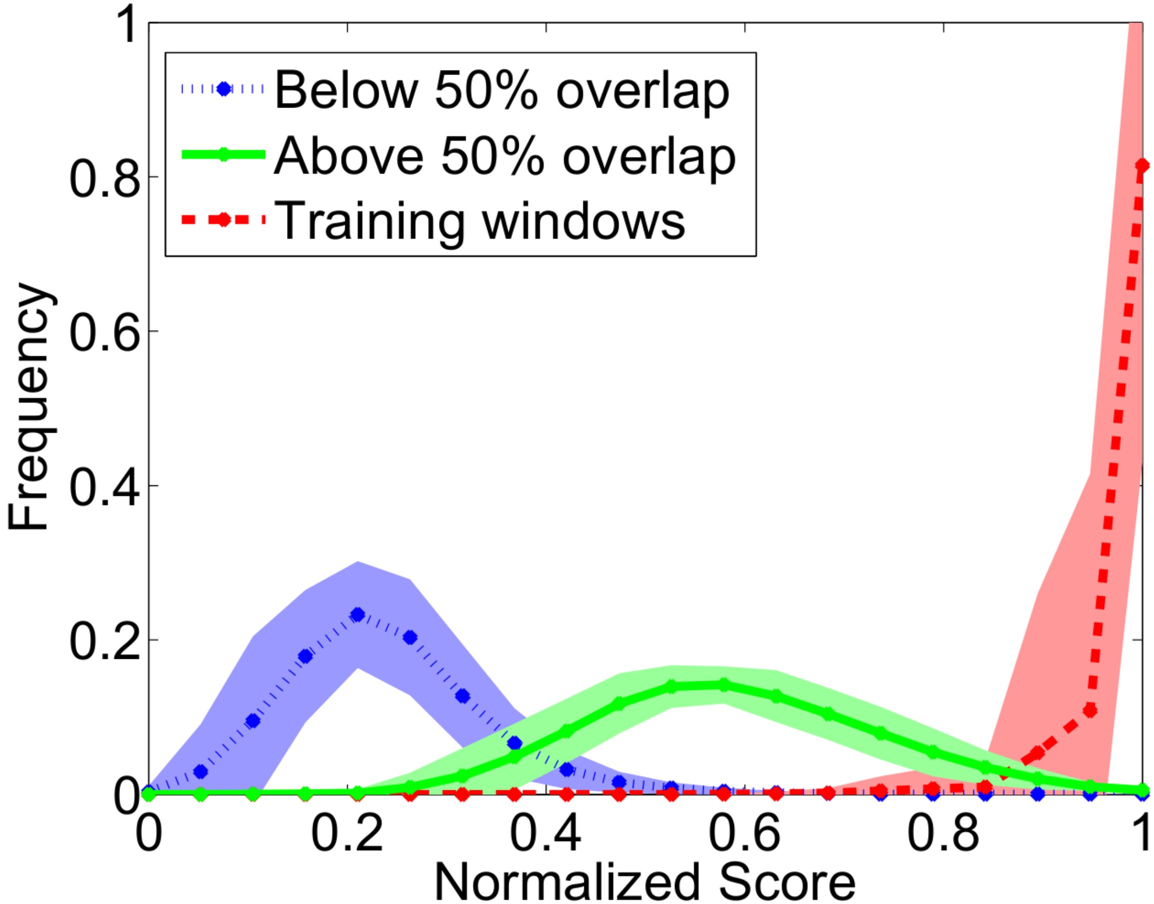
- Multiple Instance SVM

  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



Problems in standard multiple instance learning

- MIL gets stuck at poor local optima
  - Non-convex optimization problem
- Windows used in training get higher score than other windows
  - Biased towards re-localizing on the training windows

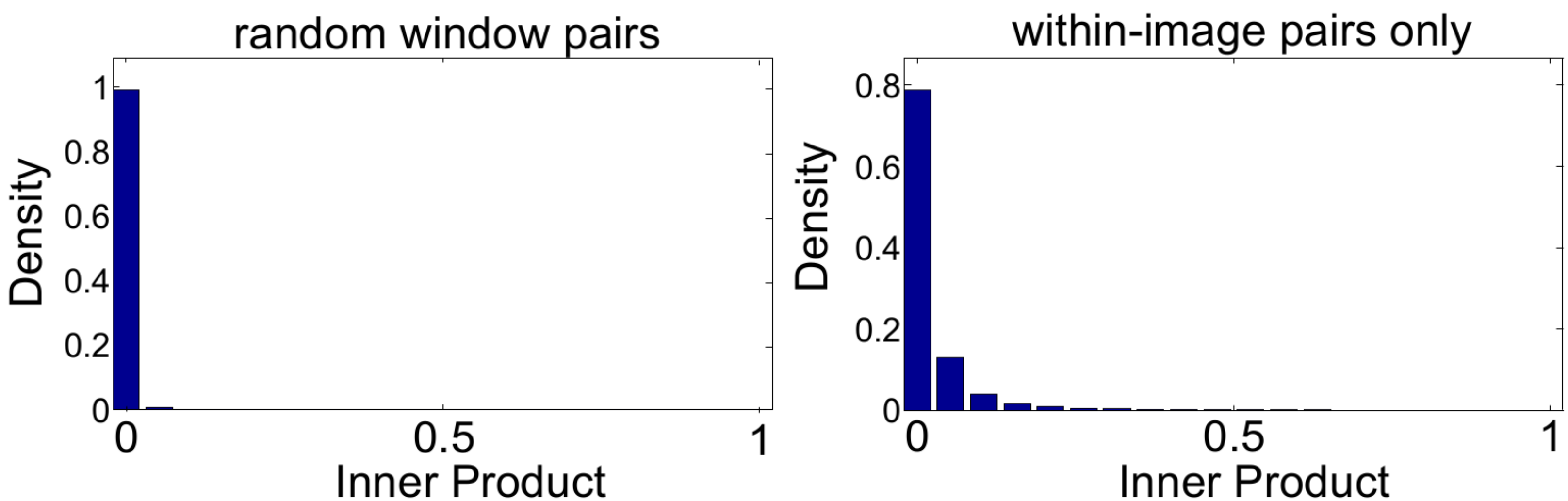


Problems in standard multiple instance learning

- Linear SVM classifier score is weighted sum of dot products

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

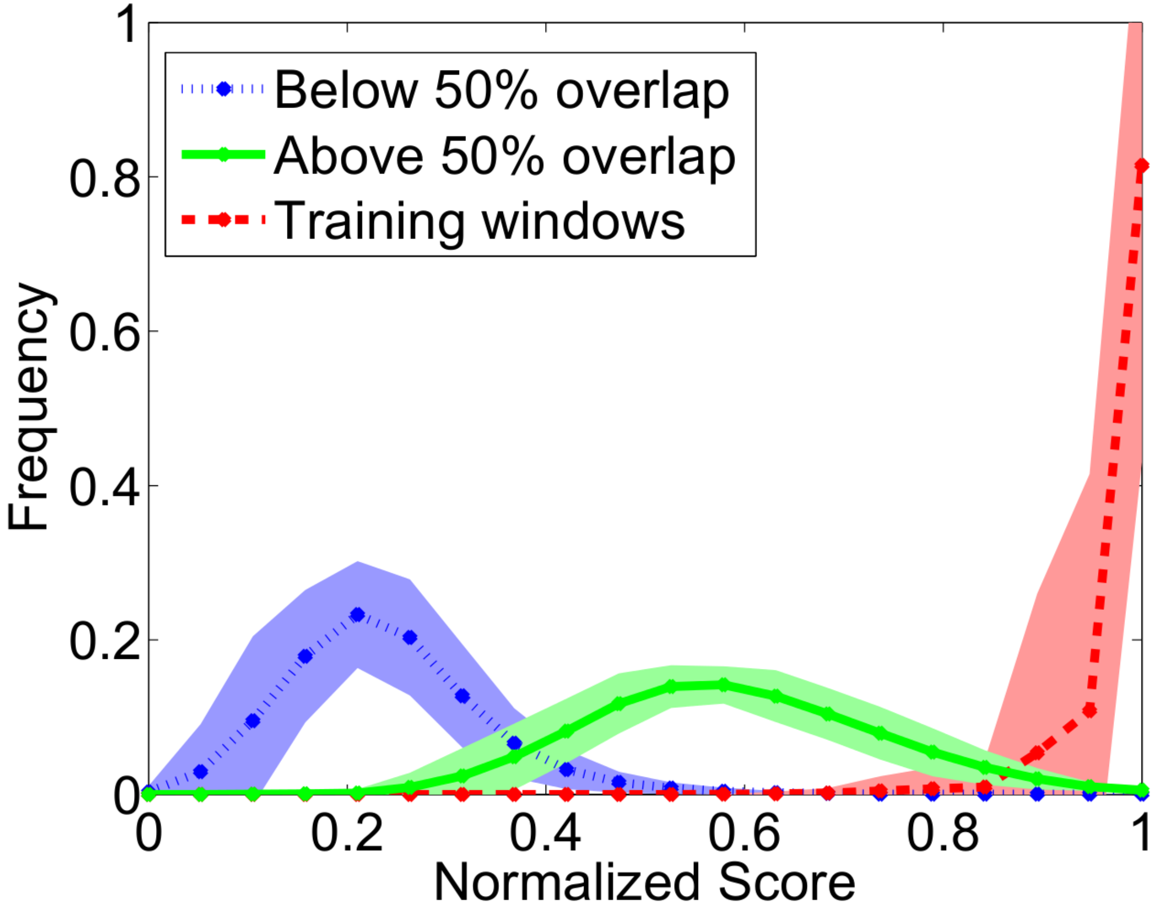
- Fisher Vector descriptors are near-orthogonal = near zero dot product
  - But recall that descriptors are unit normalized !



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

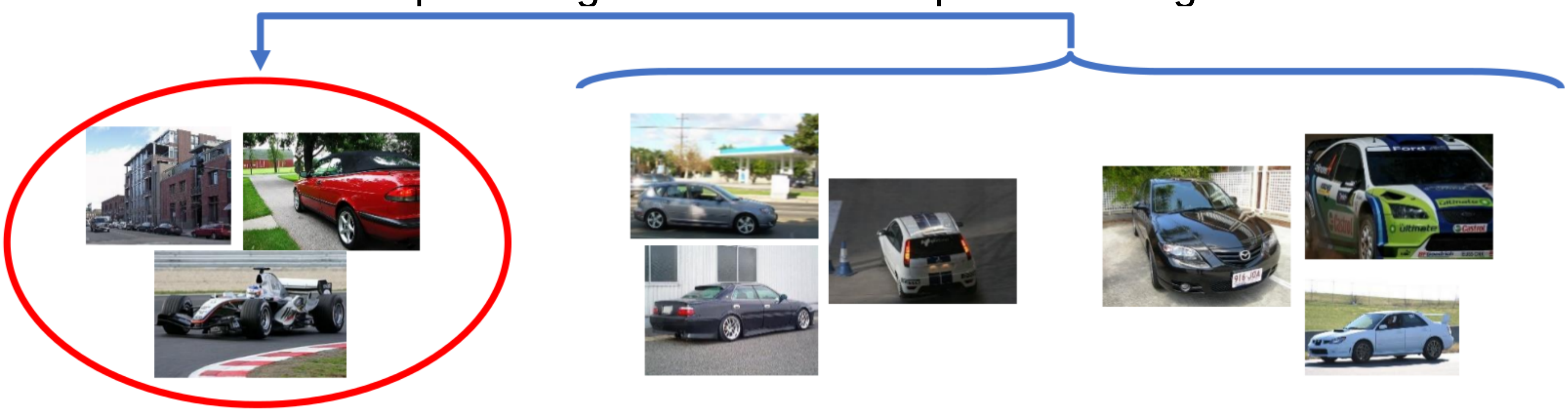
Problems in standard multiple instance learning

- MIL gets stuck at poor local optima
  - Non-convex optimization problem
- Windows used in training get higher score than other windows
  - Biased towards re-localizing on the training windows



Solution: 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
  - Divide positive training images randomly 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 windows used for training and evaluation are always different

Solution: Multi-fold training for multiple instance learning

Algorithm 1 — Multi-fold weakly supervised training

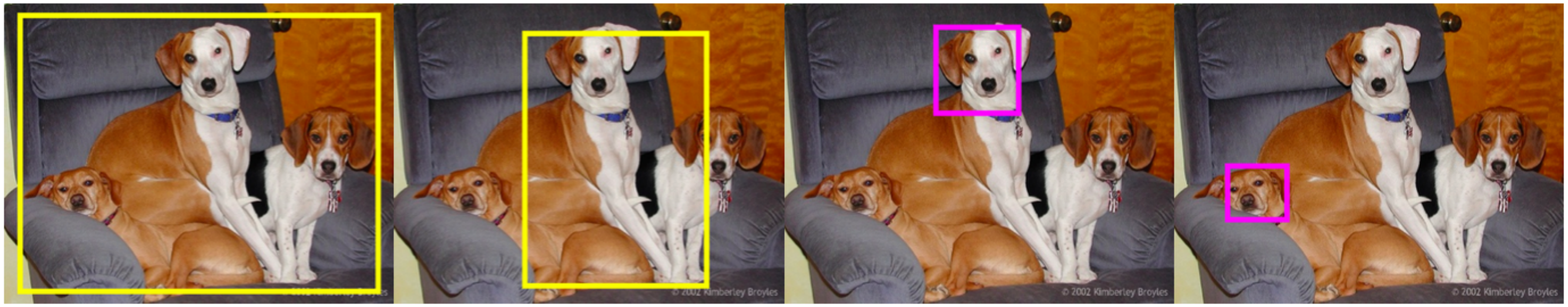
1) Initialization: positive and negative examples are set to entire images up to a 4% border.
2) For iteration $t = 1$ to T
   a) Divide positive images randomly into K folds.
   b) For $k = 1$ to K
      i) Train using positive examples in all folds but k, and all negative examples.
      ii) Re-localize 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.

A quick look at standard and multi-fold training

The trouble with cats and dogs ...

- By construction, weakly supervised learning can only learn the most repetitive and discriminative patterns between the pos. and neg. images
- These patterns sometimes correspond to parts instead of full object
- Exploited before in the context of fully supervised training

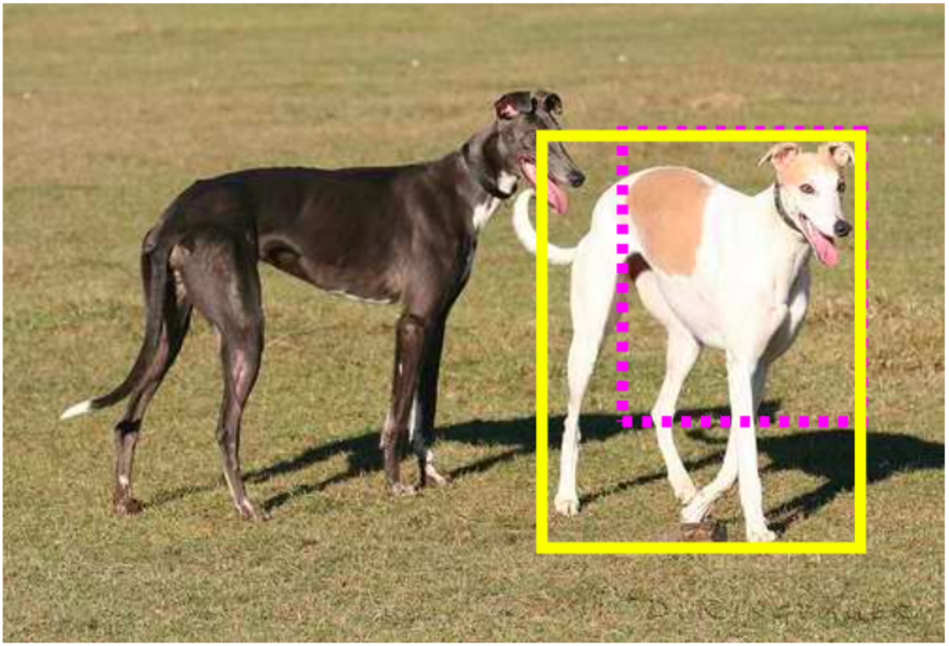
  “The Truth About Cats and Dogs”, Parkhi et al., ICCV 2011.

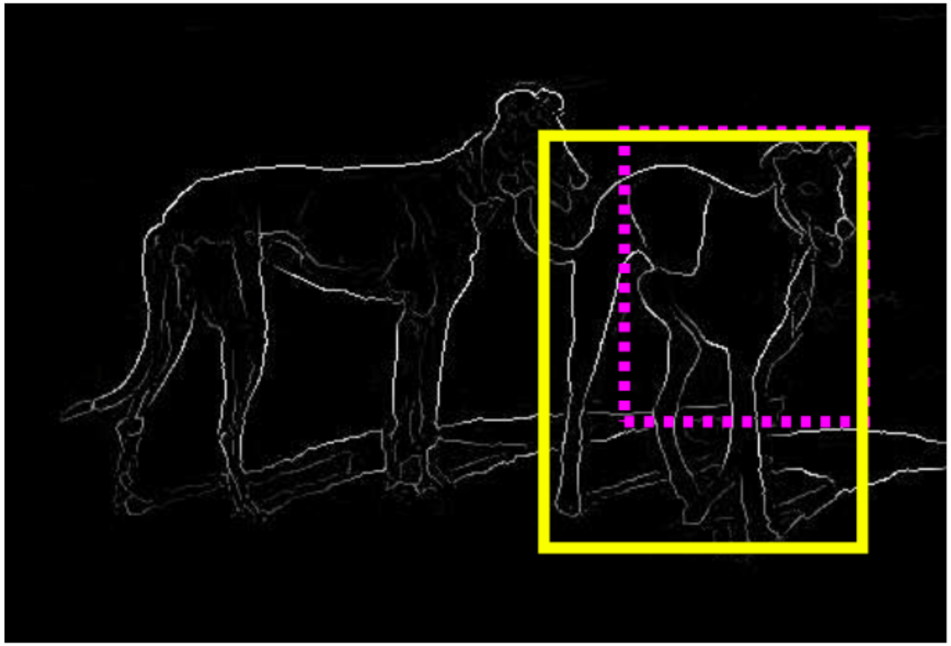


… and our solution to cats and dogs

- Refinement of the output of the multi-fold training procedure
- Final detector trained using these refined hypotheses

- Exploit low-level (non-category) contour detection to promote windows aligning with contours



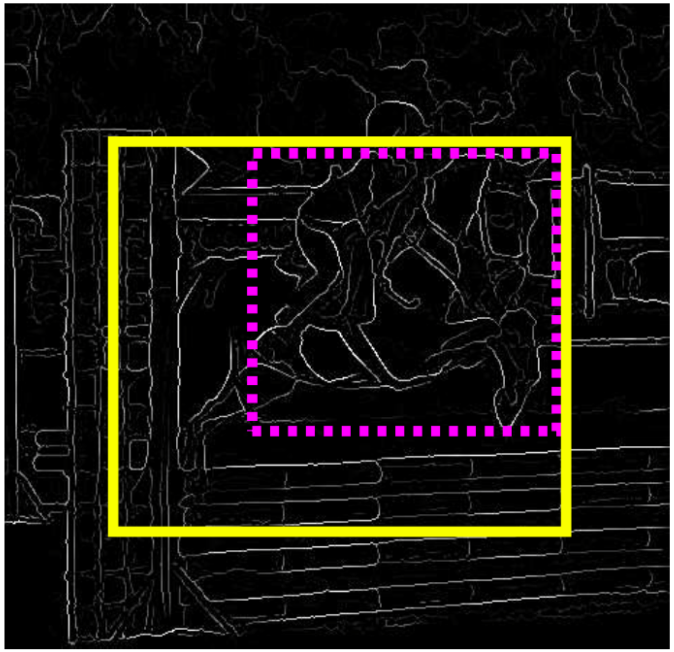


Object hypothesis refinement

- Edge-driven method to generate object hypotheses

  “Edge Boxes”, Zitnick & Dollar, ECCV'14

- Promotes windows that
  - align with long contours,
  - few contours stradlle the window boundary
- Here used to re-assess windows using average of detection and objectness score, only considering top-10 detection windows



Overview of this presentation

- Preliminaries on object localization
  - Challenges
  - Representations
  - Search and learning

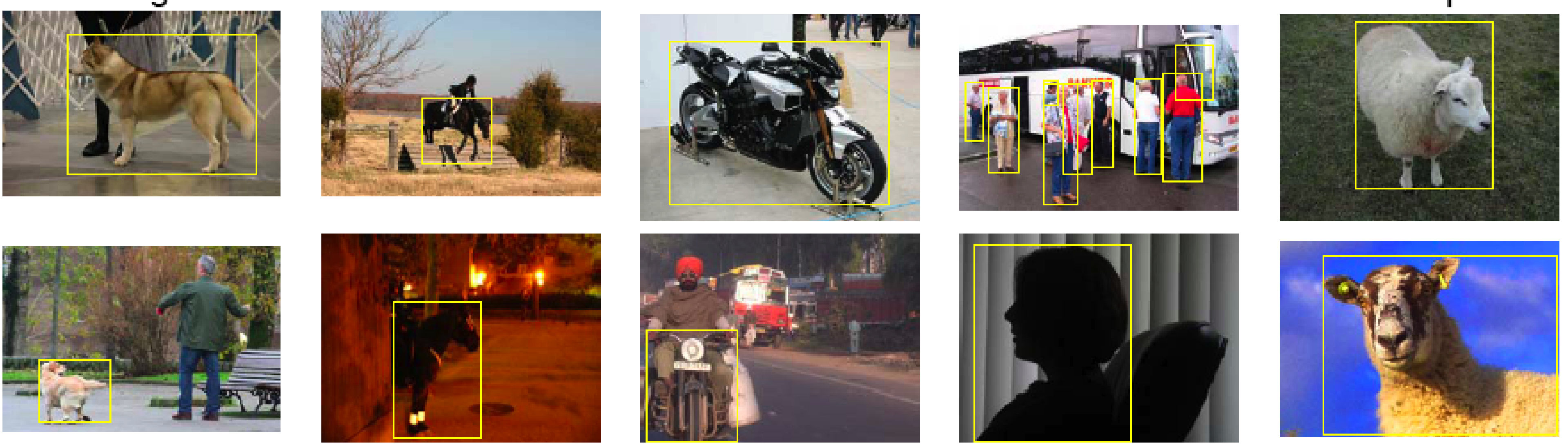
- Learning with incomplete supervision
  - Multiple instance learning approach
  - Multi-fold training to improve performance
  - Object instance hypothesis refinement

- Experimental evaluation and analysis

Evaluations based on PASCAL VOC'07 benchmark

Bicycle | Bus | Car | Cat | Cow

Dog | Horse | Motorbike | Person | Sheep



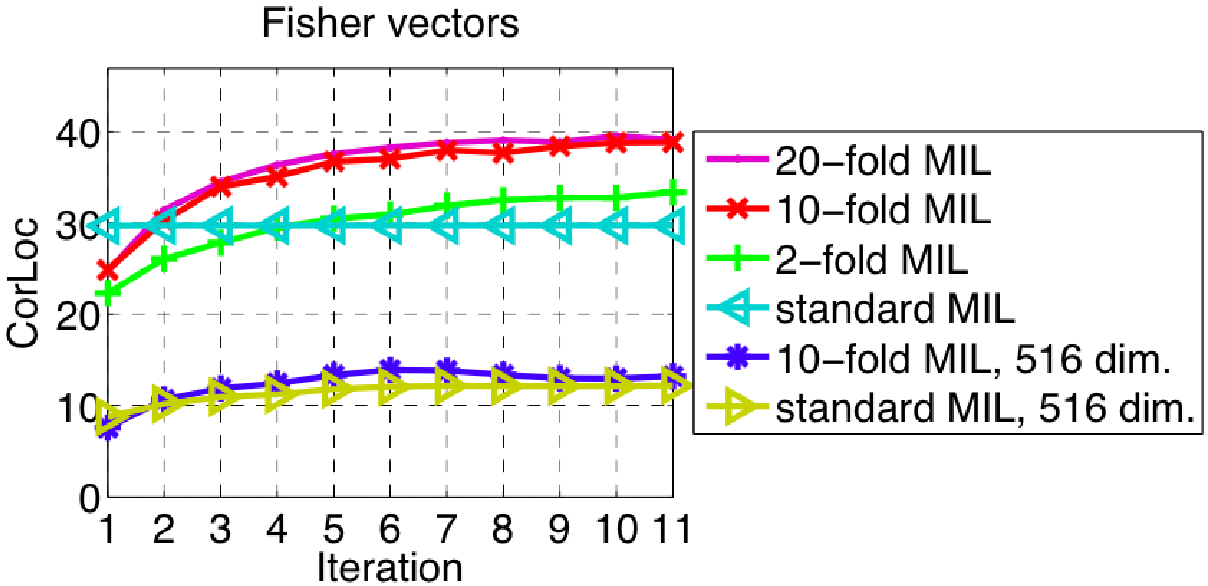
Evaluation of multi-fold training

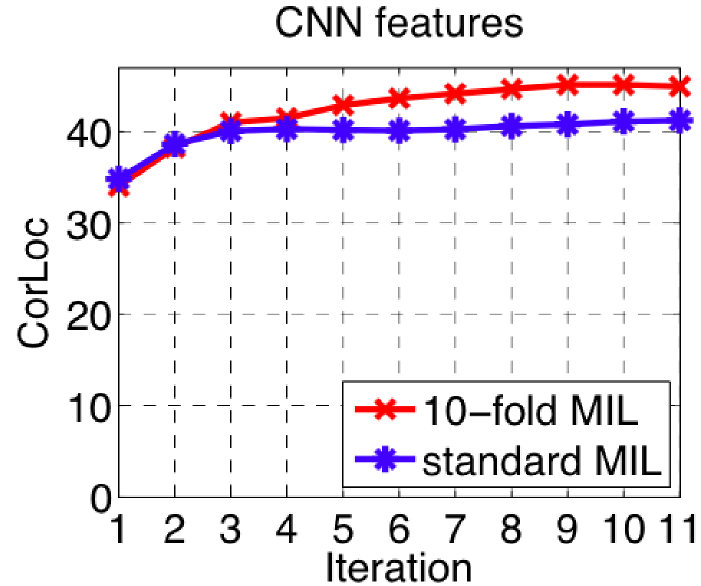
- Standard detection AP on test set
- Localization performance on positive training images
  - Fraction of images with correct localization (CorLoc)
    Deselaers et al., PAMI 2012
- Both averaged over all 20 classes
- Improvements for both features and both performance measures

	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
- For both features: averaged over all 20 classes





- Multi-fold training improves learning from both features
    - 10 folds suffice
    - 5 to 10 iterations suffice

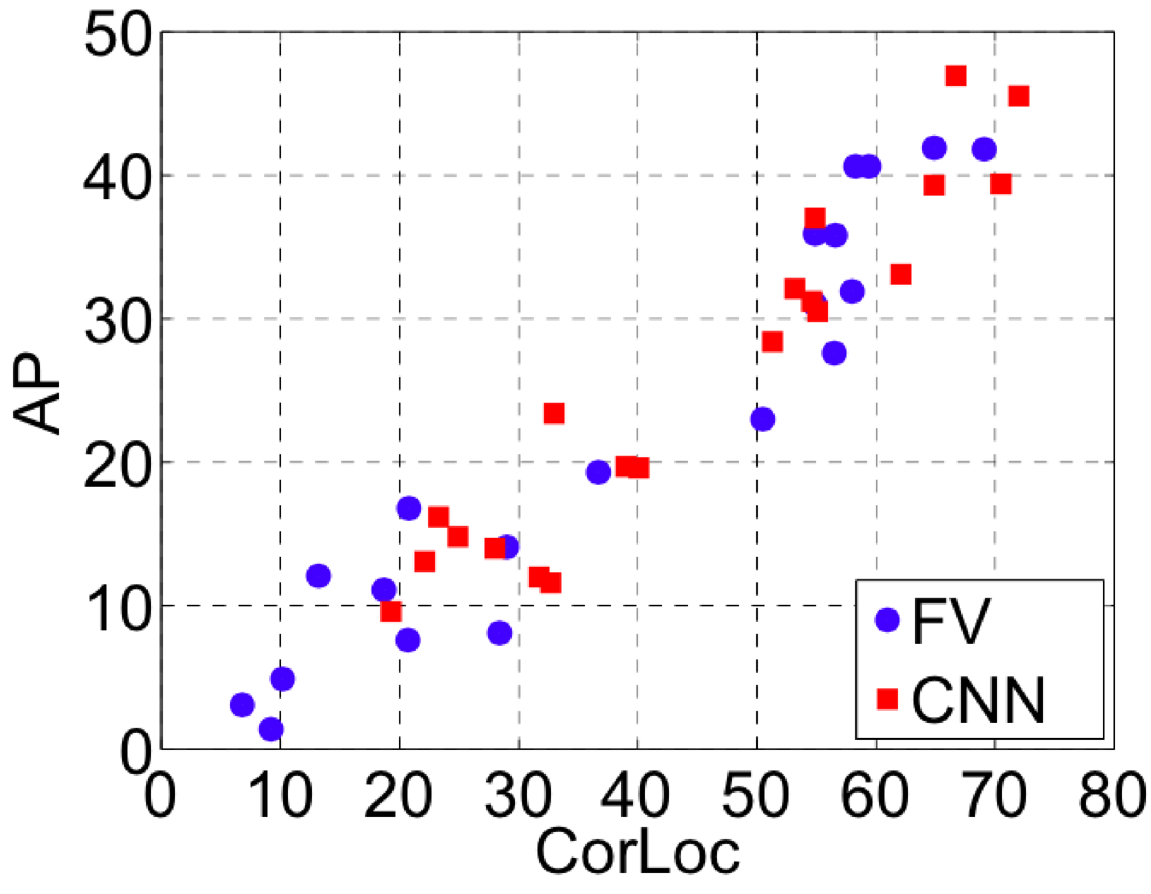
Window refinement and combining features

- Refinement helps improves performance
- Combining features boosts performance

Refinement	No	Yes
	CorLoc	
FV	38.8	46.1 (+7.3)
CNN	45.0	54.2 (+9.2)
FV+CNN	47.3	52.0 (+4.7)
	Detection AP	
FV	22.4	23.3 (+0.9)
CNN	25.9	28.6 (+2.7)
FV+CNN	27.4	30.2 (+2.8)

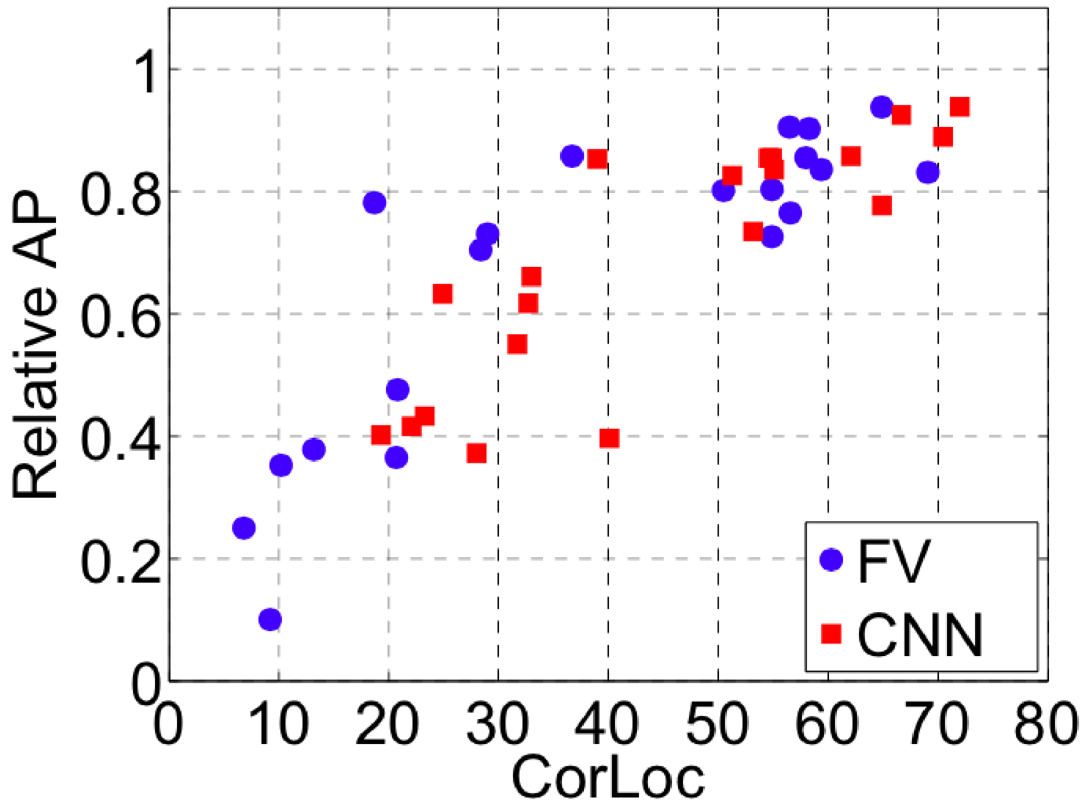
Analysis: The relation between CorLoc and detection AP

- Relation between localization during training and final test performance
  - Each of the 20 classes gives a point on the graph
  - Very highly correlated, similar coefficient for both features



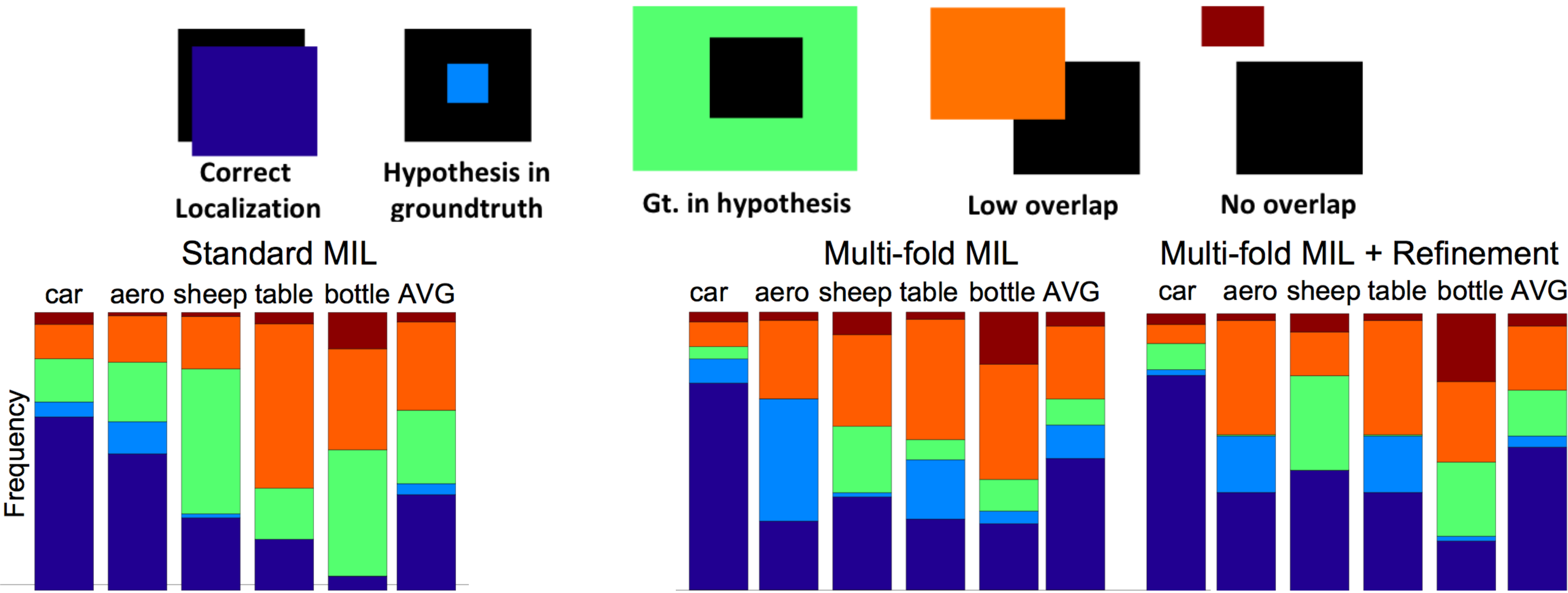
Analysis: The relation between CorLoc and detection AP

- Relative performance of weakly supervised learning with respect to performance with full supervision
  - Ratio of AP with weak vs full supervision
  - Stable performance when CorLoc is > 40%, around 80% relative
  - Smaller CorLoc results in rapid deterioration



Analysis: What type of errors are made?

- More correct localization with multi-fold training
- Less overshoot of true object for multi-fold training, more undershoot
- Refinement fixes “undershoot” cases
- Complete failure (<10%) relatively rare: explains robustness



Analysis: what makes weakly supervised learning hard ?

- Performance for the shades of grey between fully and weakly supervised learning scenario

Supervision	Neg on Pos	Positive Set	mAP(FV)	mAP(CNN)
Image labels only	No	Non-diff/trunc	22.4	25.9
Cand box for one obj	No	Non-diff/trunc	30.8	36.5
Cand box for all obj	No	Non-diff/trunc	30.7	35.7
Cand box for all obj	Yes	Non-diff/trunc	32.0	41.2
Exact box for all obj	Yes	Non-diff/trunc	32.8	40.5
Exact box for all obj	Yes	All	35.4	42.8

- The two most critical factors for performance
  - Getting one example right per positive image
  - Hard-negative mining on positive images

Comparison the recent state of the art

- Separation between methods based on whether they leverage external training data to learn CNN features

	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

- Improvements over the state of the art without external training data
- With external training data: comparable to best methods [Wang et al.,'14]

Summary and outlook

- State-of-the-art weakly supervised object detection performance
  - Strong appearance cues for recognition: FV and CNN descriptor
  - Re-localization bias suppression: Multi-fold MIL training
  - Recognition and localization decoupling: hypothesis refinement

- From here on forward:
  - Dealing with noise on the image labels (eg google-image download)
  - Concurrent training of categories: leverage explaining away
  - Richer interactions between recognition and segmentation

- Relevant publications
  - “Multi-fold MIL training for weakly supervised object localization”, CVPR'14
  - Journal paper under review: CNN features and refinement
  - PhD thesis Gokberk Cinbis, 2014: “Fisher kernel based models for image classification and object localization”

Object Detection with Incomplete Supervision

Jakob Verbeek

LEAR team, INRIA, Grenoble, France

Joint work with: Gokberk Cinbis and Cordelia Schmid