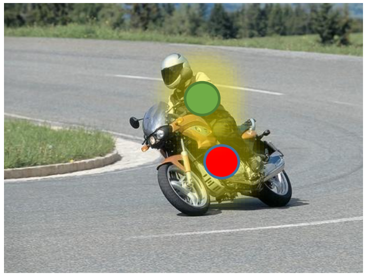
Weakly supervised object detection (WSOD)

CVPR 2018 Tutorial

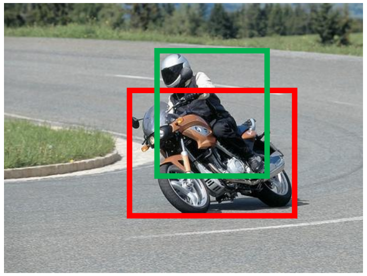
Hakan Bilen
University of Edinburgh

Manual supervision for object recognition

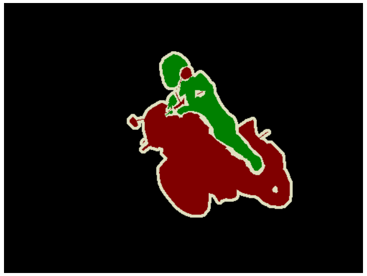
{motorbike,person}



{motorbike (point), person (point)}



{motorbike (b-box), person (b-box)}



{motorbike (pixel labels), person (pixel labels)}

1 sec per class	2.4 sec per instance	10 sec per instance	78 sec per instance

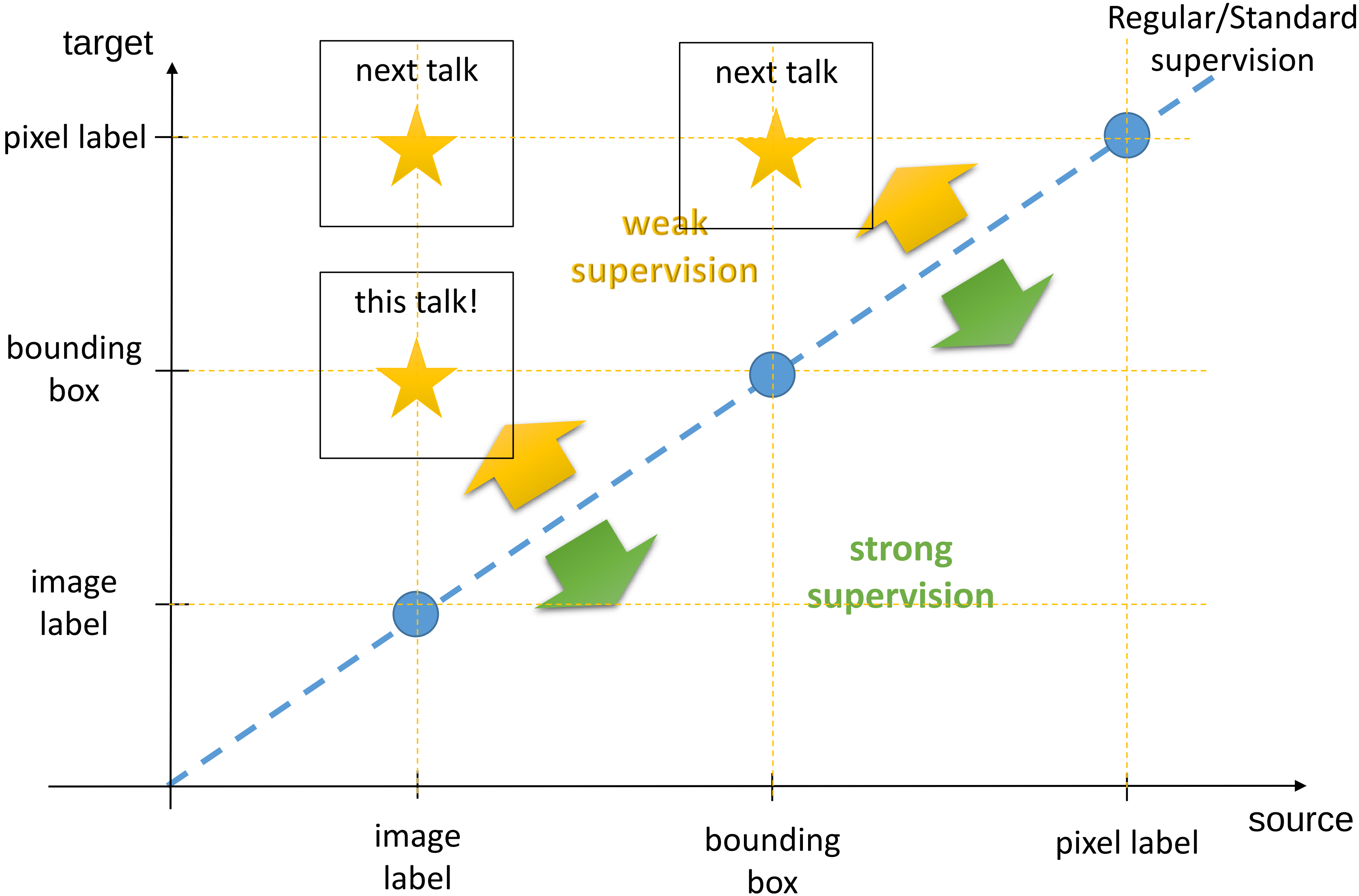
annotation time

Berman et al., What's the Point: Semantic Segmentation with Point Supervision, ECCV 16

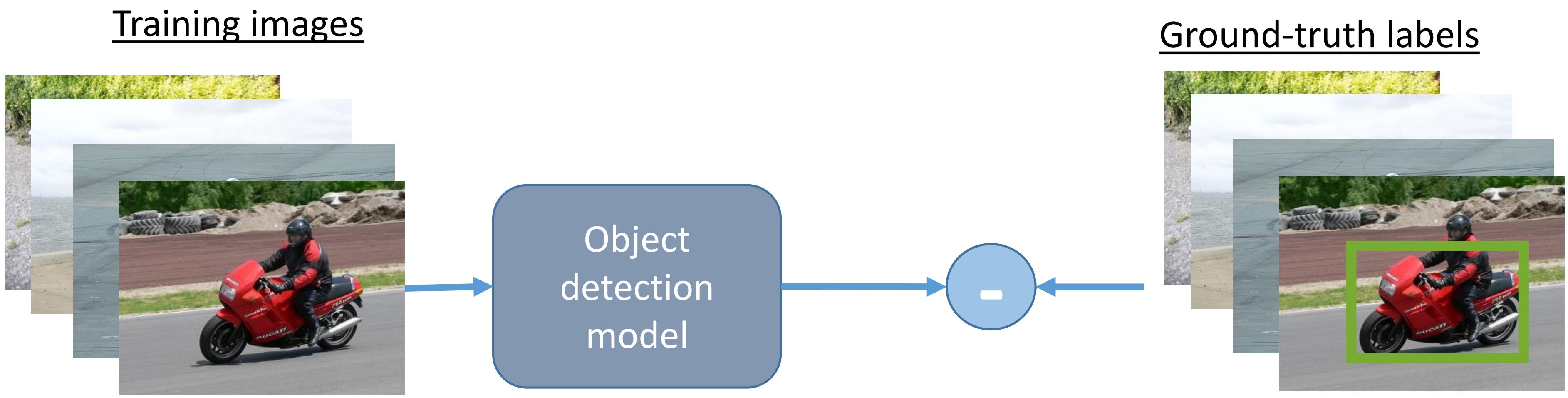
Weak supervision

Lower degree (or cheaper) annotation at train time than the required output at test time

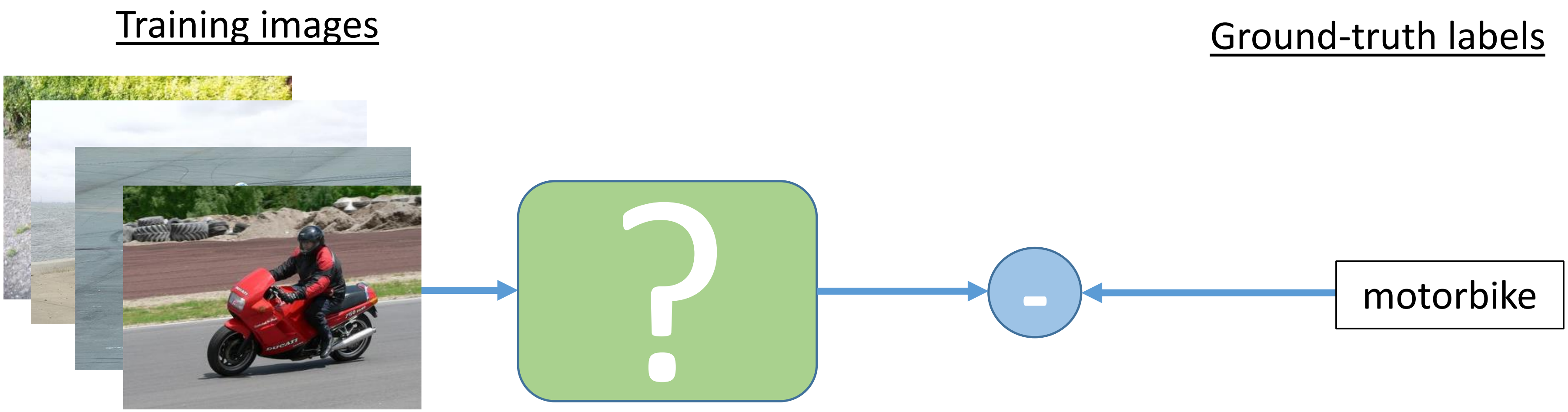
Manual supervision for object recognition



Standard supervised object detection



Weakly supervised object detection (WSOD)



What can we say at minimum?

1- When image is positive, at least one object instance from target category is present

2- When image is negative, no object instance from target category is present

Assumptions

1- There exists a set of features present in positive images and absent in negative images

2- The same features are only present on the target object instances

Challenges

Intra-class variations

- Appearance
- Transformations
- Scale
- Aspect ratio

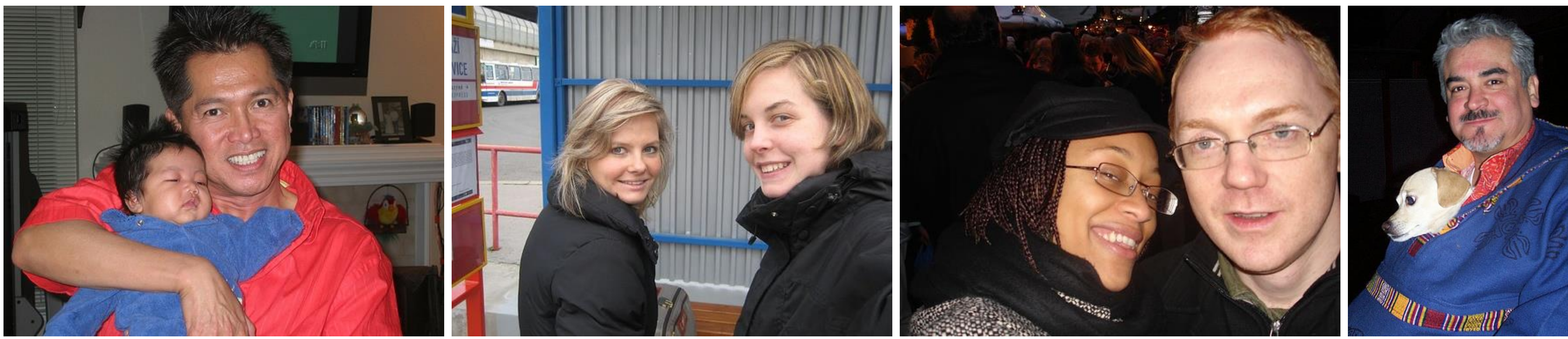
Background clutter

Occlusions

Challenges

Ambiguity in defining commonality

- Parts



Question: What is a person?

a) Face
b) Face + upper body
c) Face + whole body

Ambiguity in defining commonality

- Context

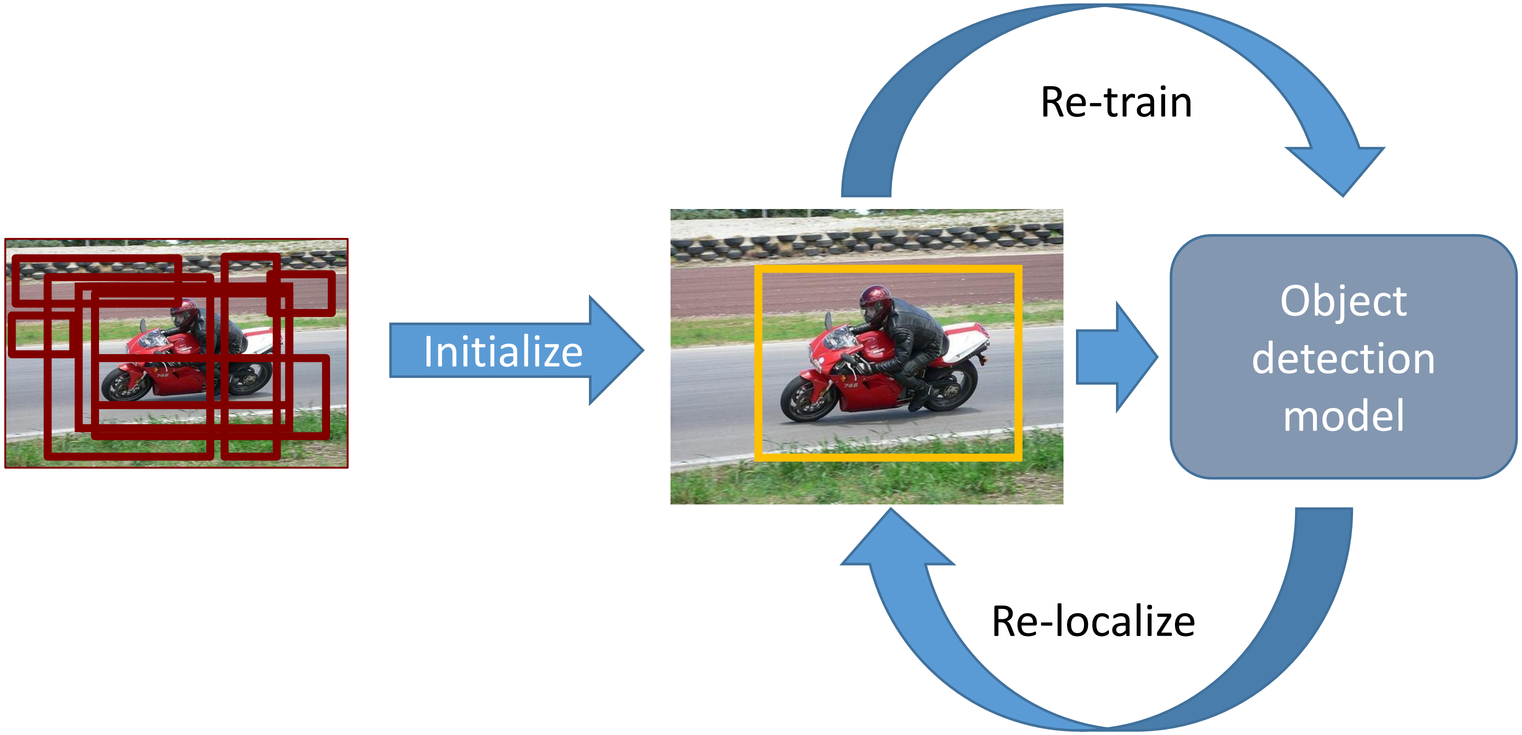
Question: What is a motorbike?

a) Motorbike + Person
b) Person
c) Motorbike + Motorbike
d) Motorbike ☺

Challenges

Alternating optimization (Re-localize + Re-train)

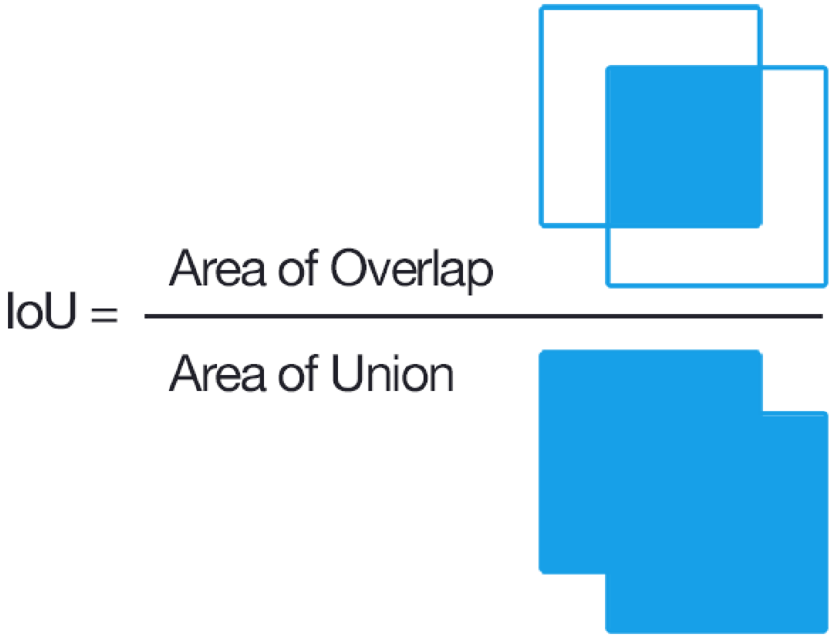
- Sensitive to initialization (local minimum)
- Overfitting (locking) to predicted windows



1. **Standard (PASCAL) object evaluation criterion:** average precision at intersection over union (IoU) 50%

2. **Correct Localization (CorLoc)** [Alex IJCV 12]: the percentage of positive training images is correctly localized at IoU 50%

- Diagnostic measure
- 100% = Supervised training



Diagnosing WSOD

Four modes of failure

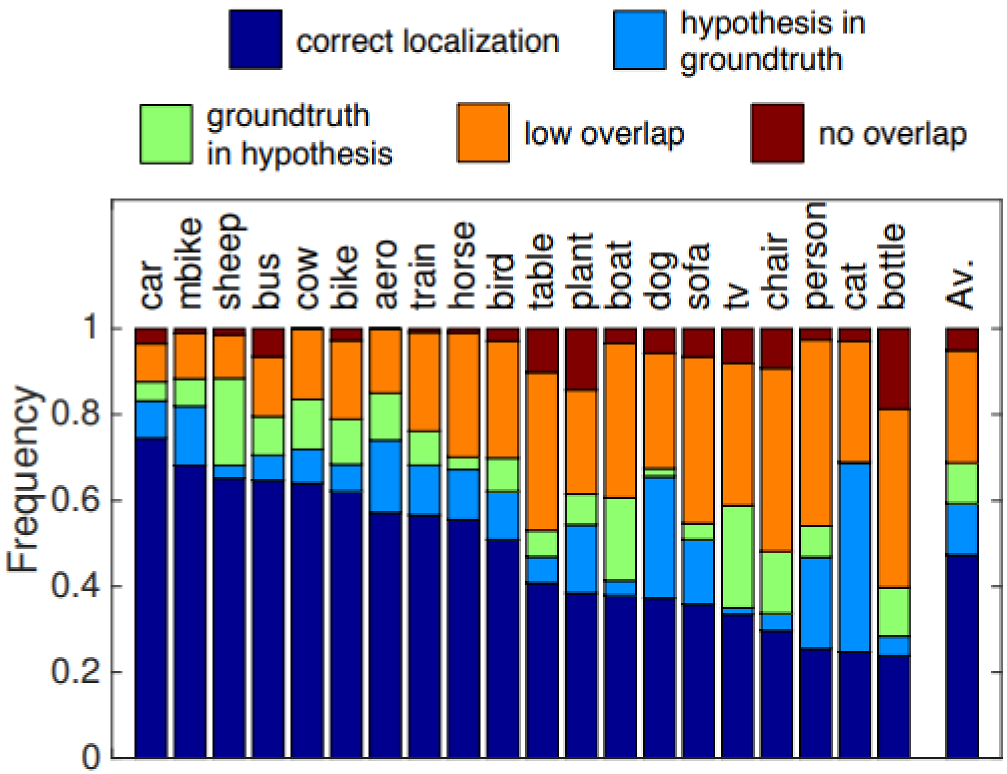


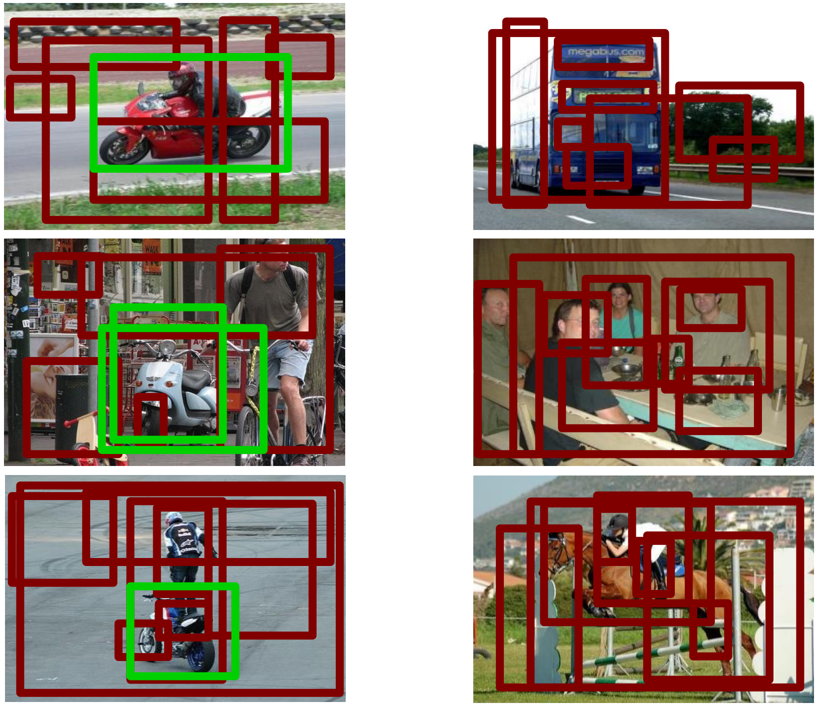
Fig: Cinbis PAMI 17

- 4 modes of failure
- In average most failures are in low overlap
- Person, cat & dog face detection (hypothesis in gt)
- Sheep, boat and tv context detection (gt in hypothesis)

Multiple-instance learning (MIL)

Dietterich et al. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence

Positive bags **Negative bags**



bags = images
instances = windows

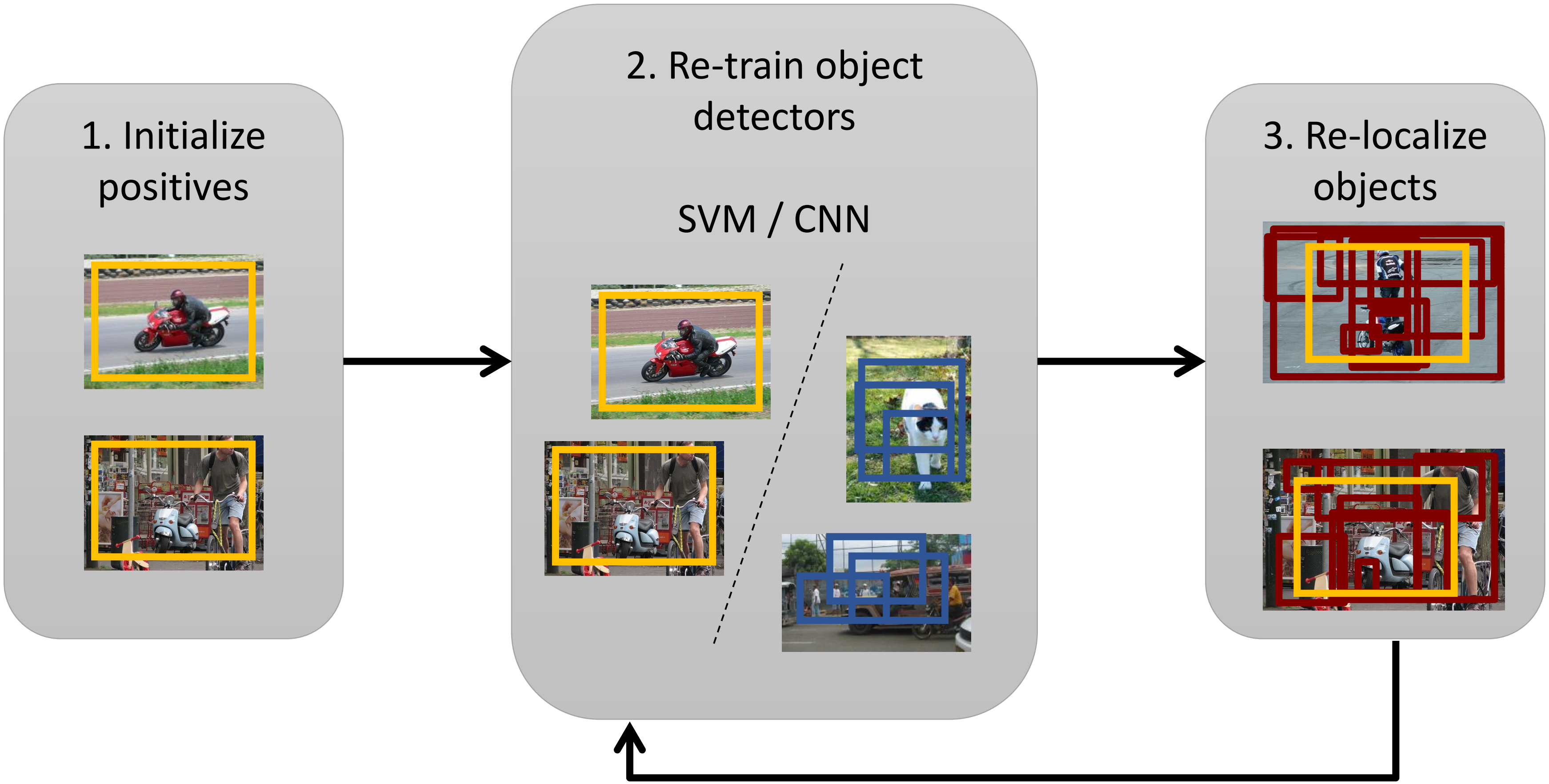
Goals:

- find true positive instances
- train window classifier

[Blaschko NIPS 10, Cinbis CVPR 14, Deselaers ECCV 10, Nguyen ICCV 09, Bilen BMVC 11, Russakovsky ECCV 12, Siva ICCV 11, Siva ECCV 12, Song NIPS 14, Song ICML 14, Bilen BMVC 14]

Slide credit: Vitto Ferrari

Standard MIL pipeline



1. Window space
2. Initialization
3. Re-localization & Re-training

Slide credit: Vitto Ferrari

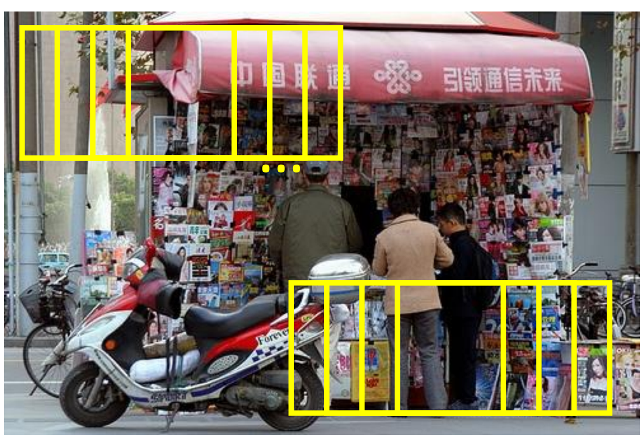
1. Window space

How to generate bags?

Sliding windows

- >100k per image
- dense
- translations, scales and aspect-ratios (4D space)

[Chum CVPR 07, Nguyen ICCV 09, Pandey ICCV 11]

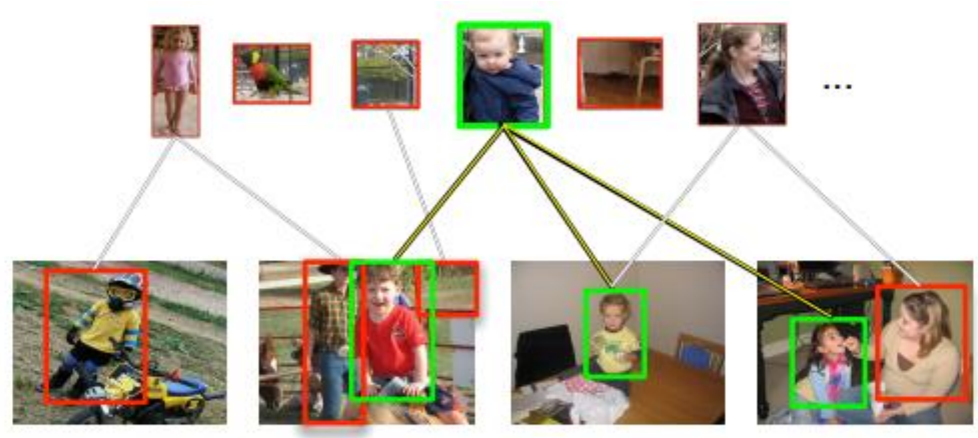


Object proposals

- ~2k per image
- sparse
- [Alexe CVPR 10, van de Sande ICCV 11, Dollar ECCV 14]
- Commonly used in WSOD
  [Deselaers ECCV 10, Siva ICCV 11, Russakovsky ECCV 12, Cinbis CVPR 14, Wang ECCV 14, Bilen CVPR 16]

Slide credit: Vitto Ferrari

2. Initialization



[Song et al ICML 14]

Simple strategies

- Whole image
  [Nguyen ICCV 09, Bilen BMVC 14]
- Whole image minus a margin
  [Pandey ICCV11, Russakovsky ECCV12, Bilen CVPR 14]

Constructs a graph to find initial boxes:

1. relevant (occur in many positive images)
2. discriminative (dissimilar to the boxes in the negative images)
3. complementary (capture multiple modes)

3. Re-localization and Re-training

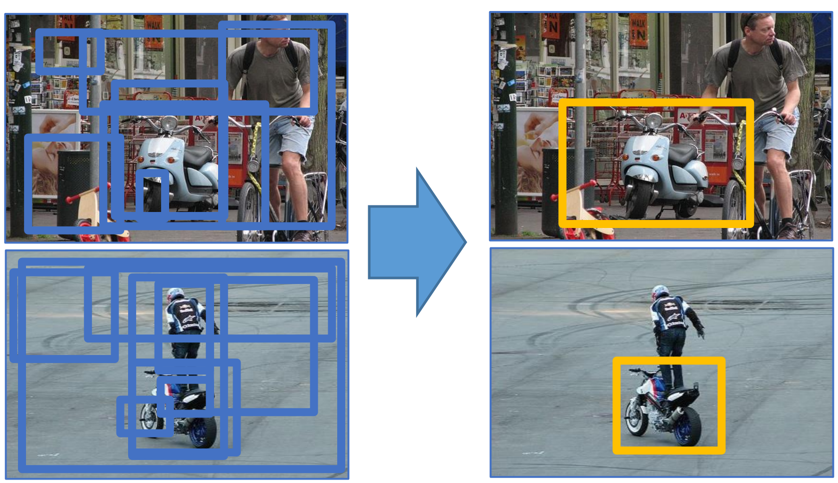
Standard max margin formulation

<u>Re-localizing object instances</u>

$$\arg\max_b A(x_b)$$

Proposal Appearance model Features

- Only one positive instance per image



[Nguyen ICCV 09, Bilen CVPR 14, Cinbis CVPR 14, Papadopoulos CVPR 16]

<u>Re-training object detectors</u>

- For positive images:

$$\max_b A(x_b) > \Delta \ \ (\Delta\text{:margin})$$

- For negative images:

$$\max_b A(x_b) < -\Delta$$

- Different from supervised learning

1. 1 pos instance in each pos image
2. No neg instances from pos image

(Think about Fast(er)-RCNN)

[Nguyen ICCV 09, Bilen BMVC 11, Russakovsky ECCV 12, …]

More robust optimization: Relaxing max operator

- Hedge your bets on multiple proposals:

[Bilen BMVC14, CVPR15-6, Kantorov ECCV16]

$$\max_b A(x_b)$$

$$\log \sum_b \exp A(x_b)$$

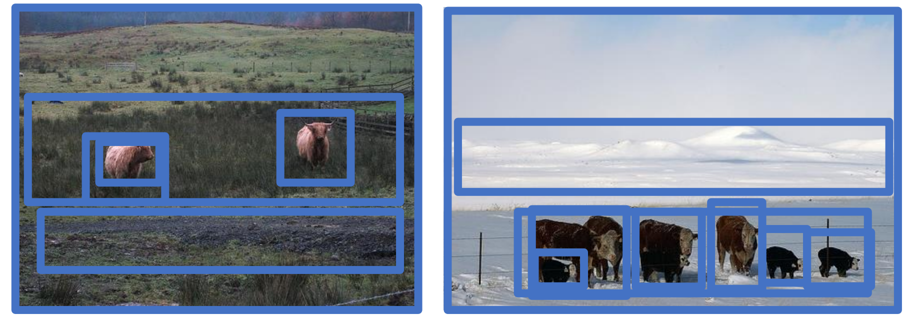
- Re-train object detectors:

For positive images

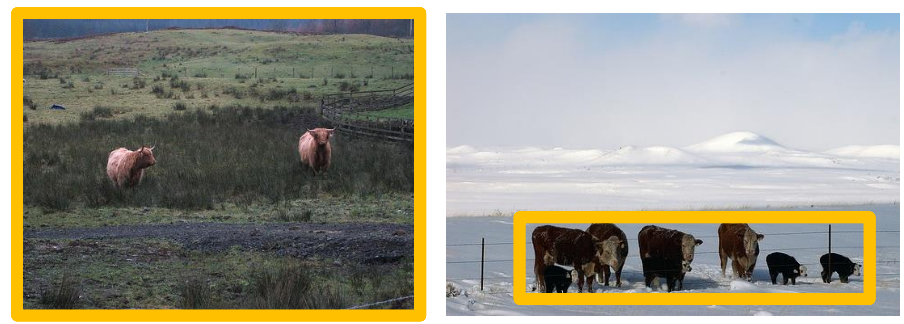
$$\log \sum_b \exp A(x_b^+) > \Delta$$

For negative images

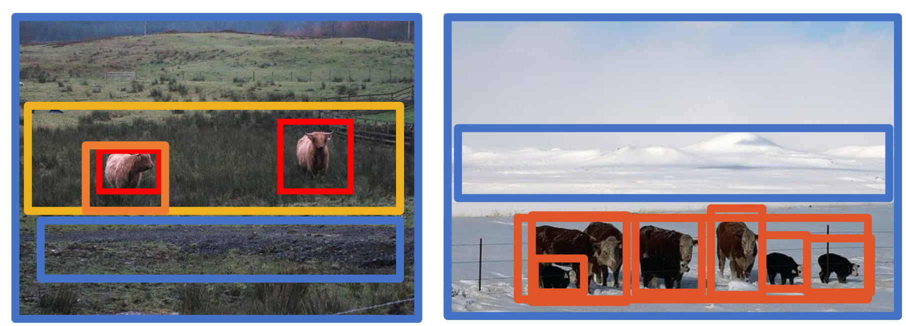
$$\log \sum_b \exp A(x_b^-) < -\Delta$$



Max



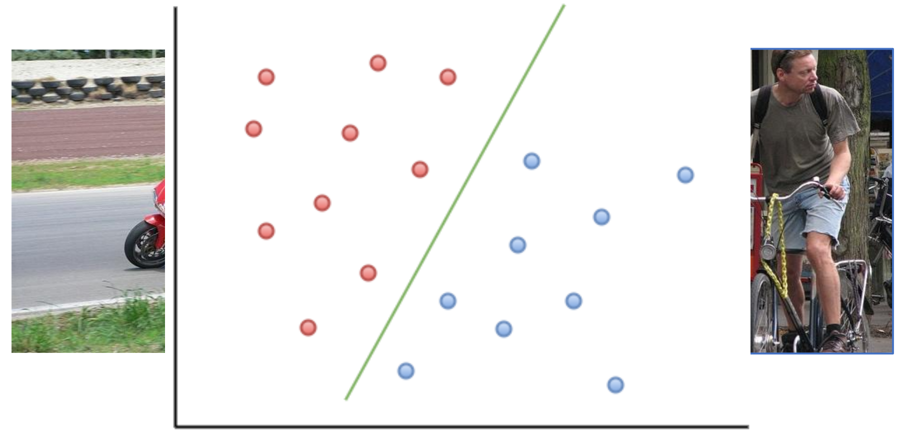
Soft-max



Re-localization and Re-training

More robust optimization: Self-paced learning [Kumar NIPS 10]

- Inspired from Curriculum Learning [Bengio ICML 09]
- Start with easy samples, then consider hard ones in training
- Easiness for human:
  - scale, clutter, occlusion
- Easiness for machine:
  - Selection of samples via confidence of max scoring window [Kumar NIPS 10]
  - Selection of window space by allowing smaller windows [Bilen IJCV 14, Shi ECCV 14]
  - Selection of samples via inter-category competition [Sangineto PAMI 17]



(d) iter 0 (e) iter 1 (f) iter 5

Fig [Bilen IJCV 14]

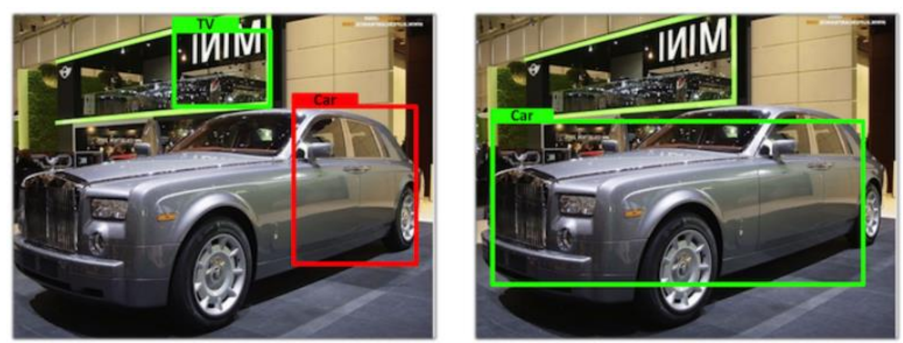
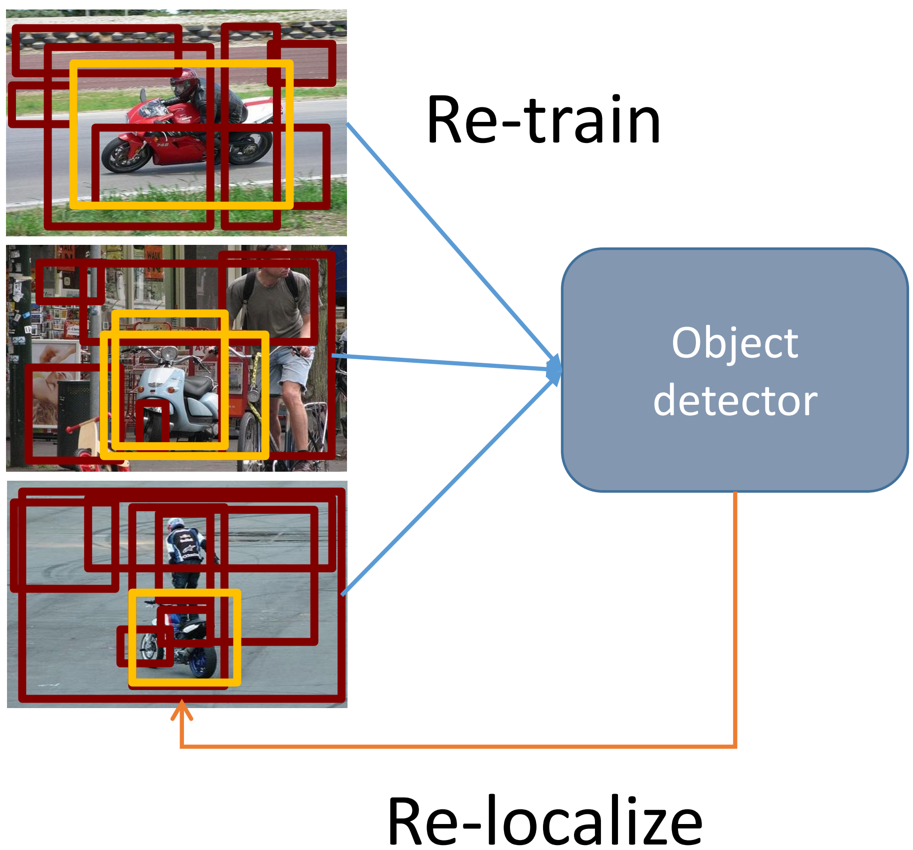


Fig [Sangineto PAMI 17]

More robust re-localization: Multifold MIL [Cinbis CVPR 14]

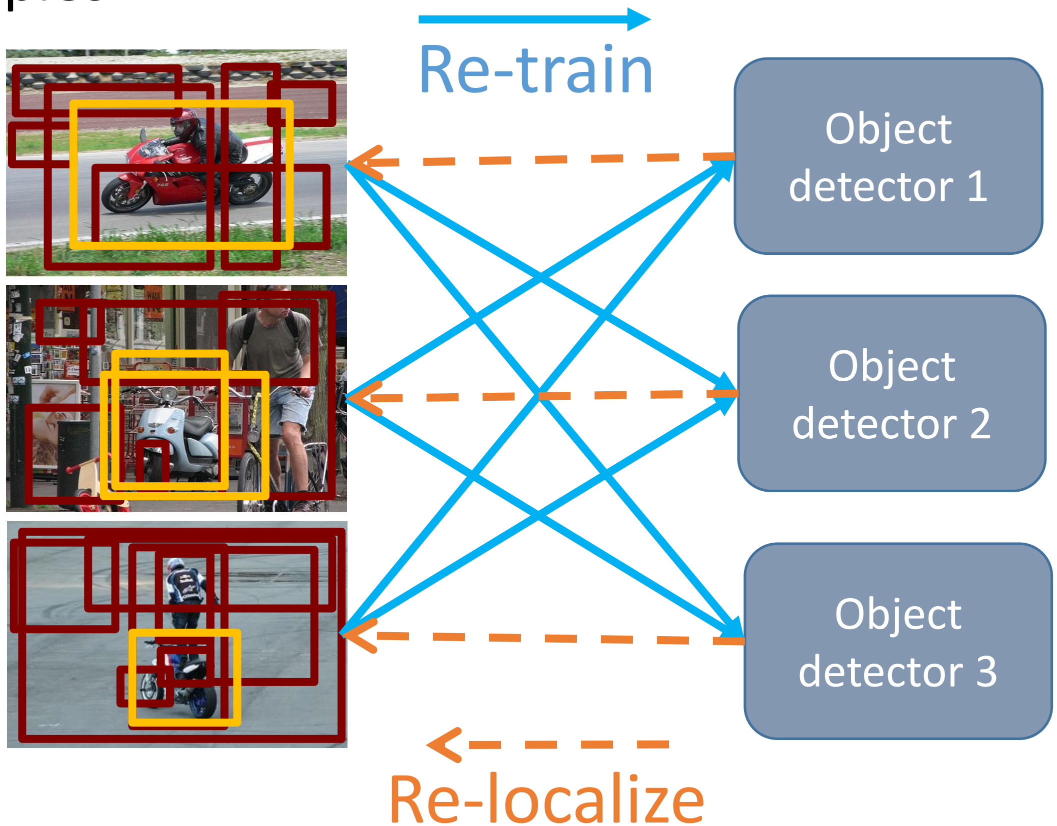
Problem: Detector overfits into the given proposal



More robust re-localization: Multifold MIL [Cinbis CVPR 14]

Problem: Detector overfits into the given proposal

Solution: Train using positive examples in all folds but k, and all negative examples

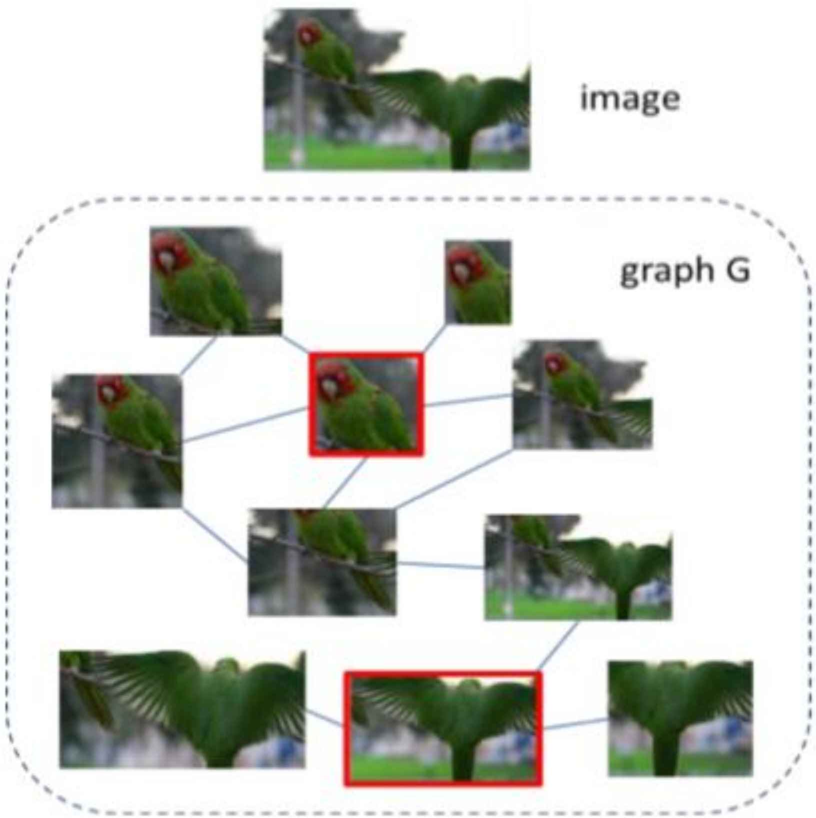


More robust re-localization: Self-taught learning [Jie CVPR 17]

Idea: Replace max with a more sophisticated technique that considers spatial neighborhood

Dense sub-graph discovery

1. Connect if IoU>0.5
2. Select proposal with most connections
3. Remove connected nodes



Priors

Assume that we have N positive images, each with W windows

- W^N possible configurations
- Only 1 of them is correct
- Can we eliminate some of bad ones by using our prior knowledge?

Priors: Pairwise similarity

- Similarity between selected windows across positive images

[Chum CVPR 07, Deselaers ECCV 10, Siva ICCV 11, Bilen CVPR 15]

☺ Less overfitting

☹ Expensive to optimize

☹ Ignores intra-class variation

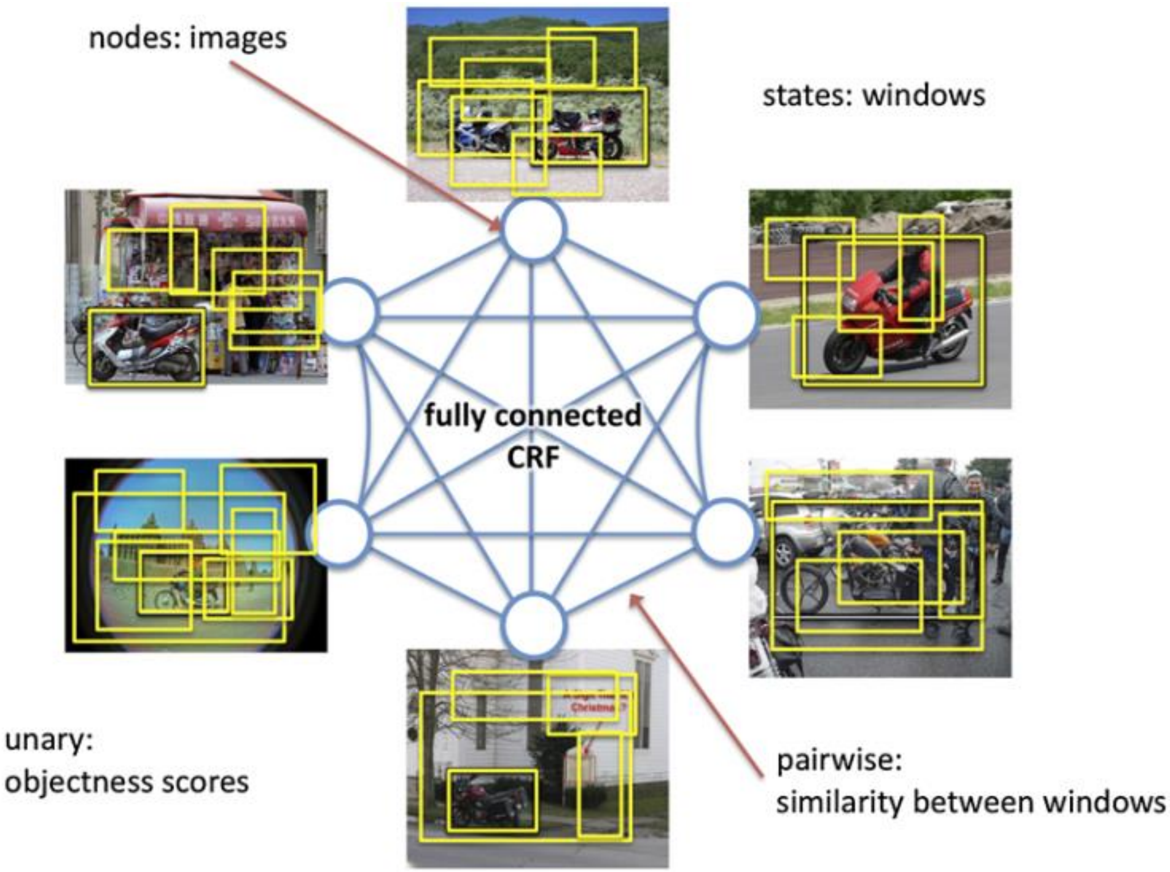


Fig: [Deselaers ECCV 10]

Priors: Pairwise similarity

Sub-categories

- Clustering via probabilistic latent Semantic Analysis (pLSA)
- [Wang ECCV 14]
- ☺ Modeling intra-class variations
- ☹ Sensitive to number of clusters

Exemplars

- [Chum CVPR 07, Bilen CVPR 15]
- ☺ No need to set number of clusters
- ☹ Memory expensive

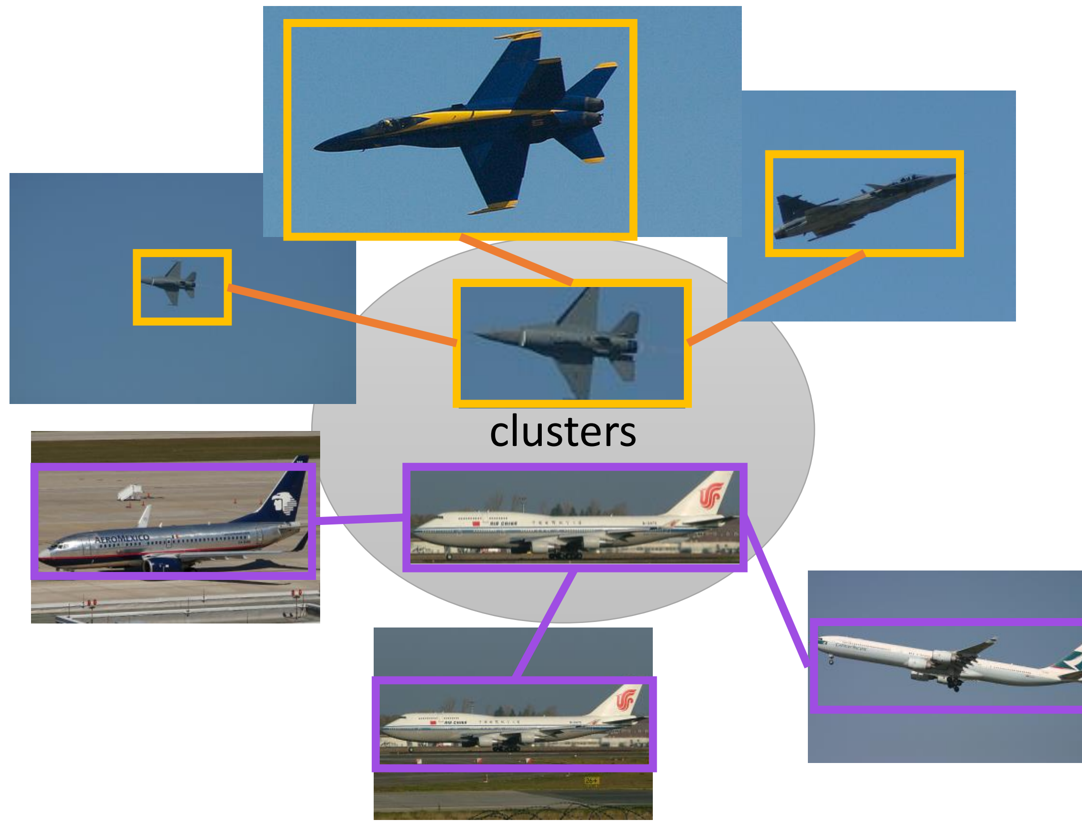


Fig: [Bilen CVPR 15]

Priors: Context

- Background provides contextual cues for recognition [Russakovsky ECCV 12, Bilen CVPR 14, Kantorov ECCV 16]
- Better separation of foreground and background
- Additive: select a ROI that is semantically compatible with its context
- Contrastive: select a ROI that is outstanding from its context

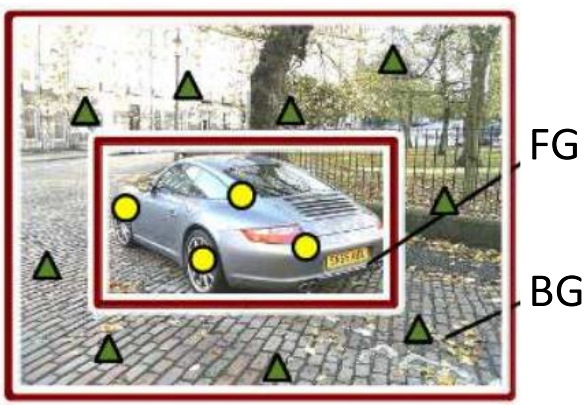


Fig: [Russakovsky ECCV 12]

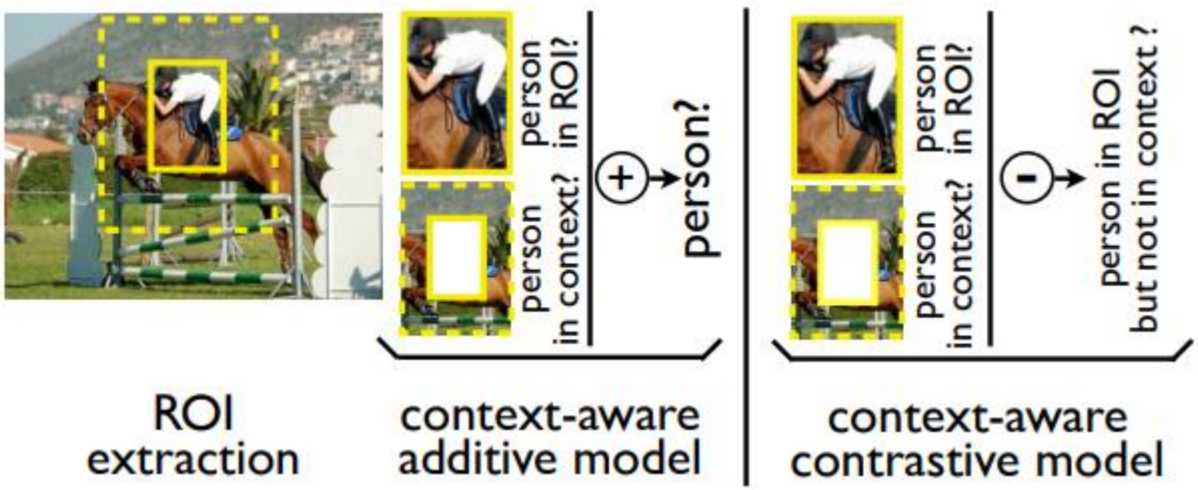


Fig: [Kantorov ECCV 16]

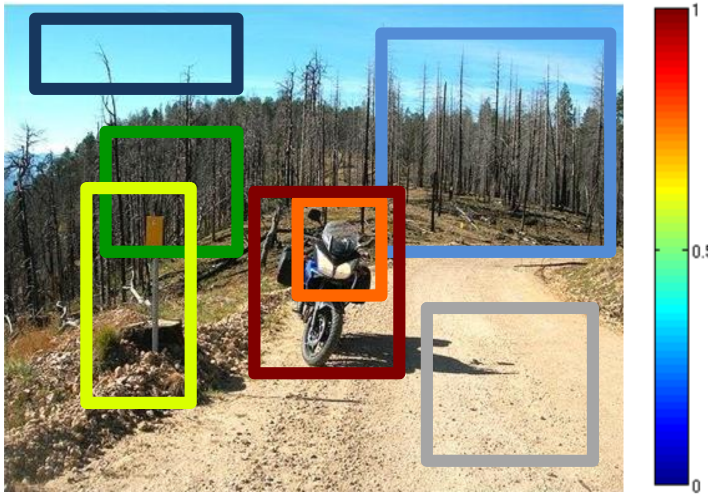
Priors: Objectness

Quantify how likely a window is to contain an object of *any* class
[Alexe CVPR 10, Zitnick&Dollar ECCV 14]

- Steers re-localization towards objects and away from background
- Pushes towards whole objects instead of subregions

$$argmax_b \lambda A(x_b) + (1-\lambda)\mathrm{Obj}(b)$$

Commonly used for weakly supervised object localization
[Deselaers ECCV 10, Khan OAGMW 11, Siva ICCV 11, Guillaumin CVPR 12, Prest CVPR 12, Shapovalova ECCV 12, Shi BMVC 12, Tang CVPR 14, Wang ECCV 14, Jerripothula ECCV 16, Cinbis PAMI 16, Bilen CVPR 16, …]

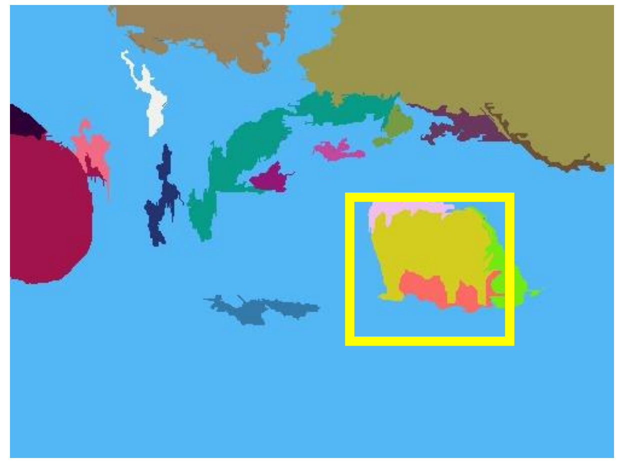


Slide credit: Vittorio Ferrari

Priors: Objectness, example cues

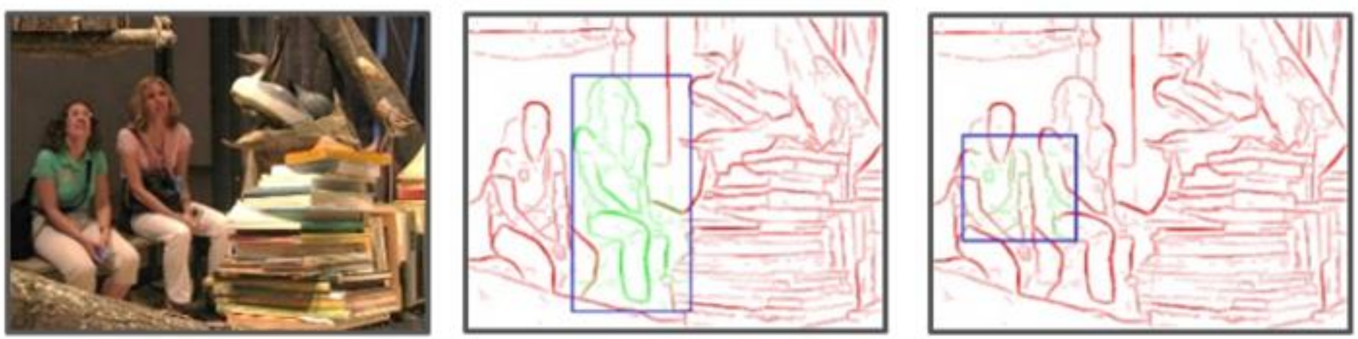
Color contrast

Segments straddling



[Alexe CVPR 10]

Edges straddling

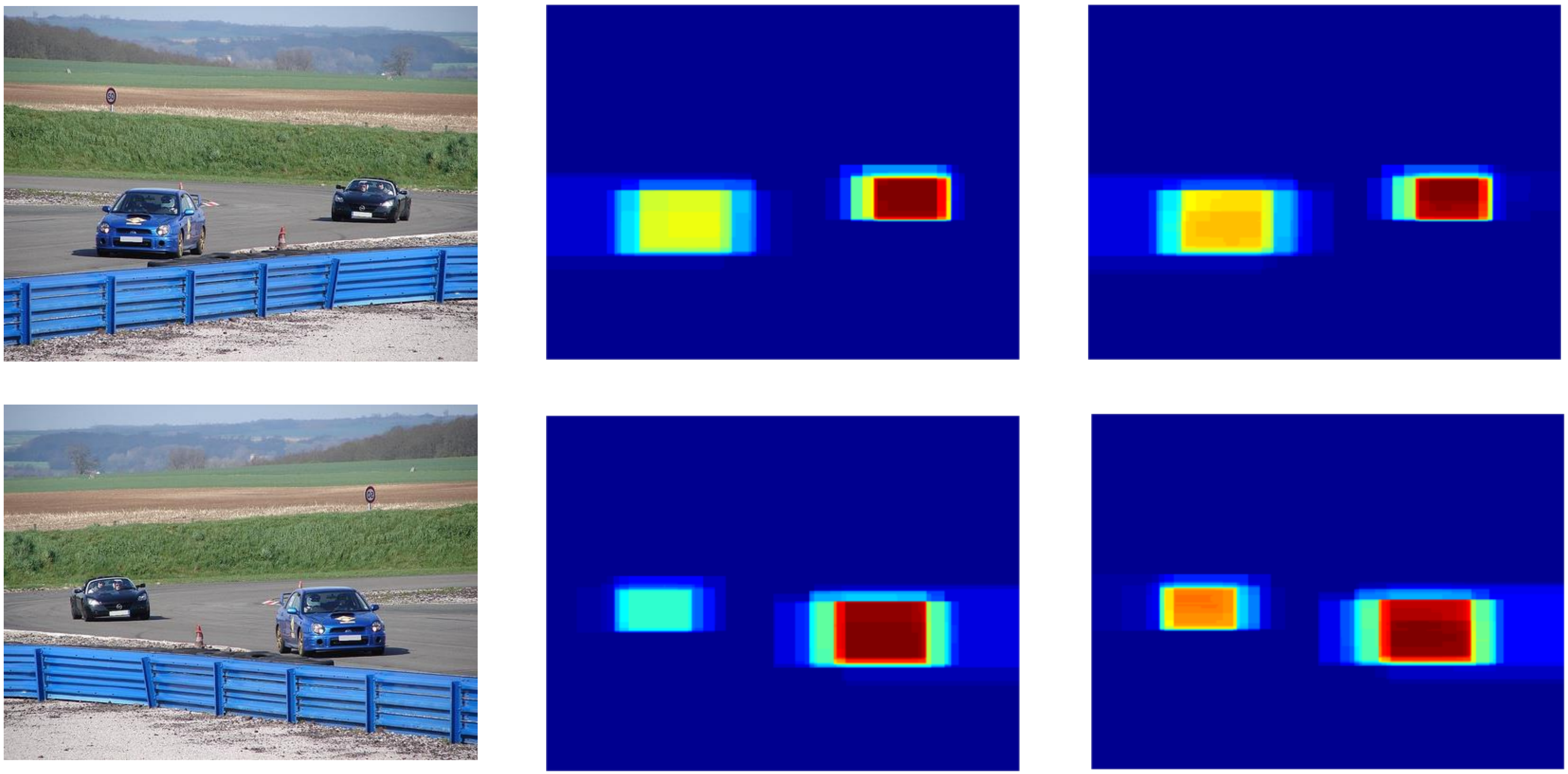


[Zitnick ECCV 14]

Slide credit: Vittorio Ferrari

Priors: Symmetry [Bilen BMVC 14]

What can we say about object locations for these two images?

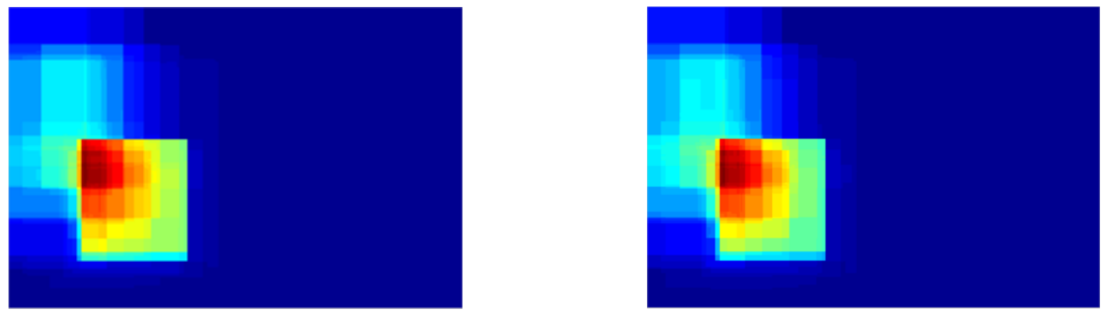


Minimize KL divergence between prediction scores across images

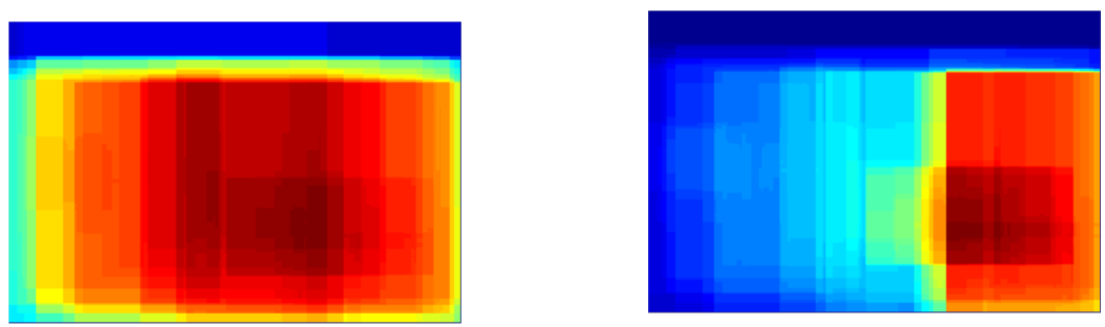
Priors: Mutual exclusion [Bilen BMVC 14]

Assumption: A box can tightly cover only one object instance

Not always true but in most cases!



Tv



Sofa

Minimize KL divergence between box scores across different classes

Priors: Scale [Shi ECCV 16]

- Curriculum learning (bigger objects down to smaller ones)
- Weight object proposals according to estimated size
- Requires training a size estimator from a small set

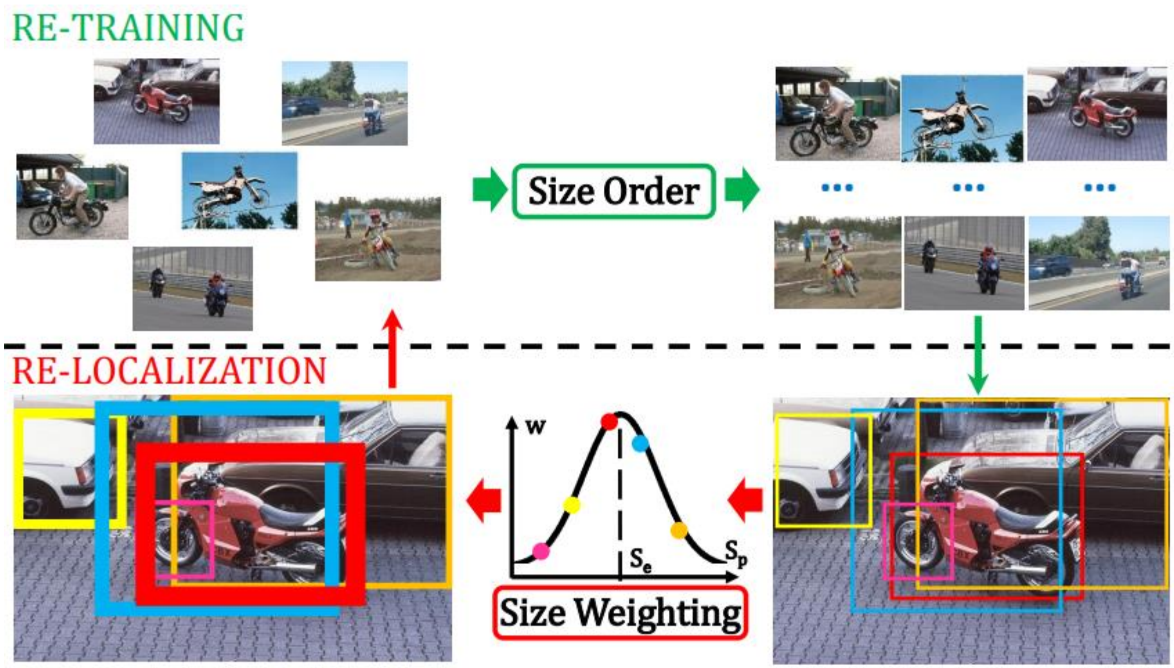


Fig: [Shi ECCV 16]

3. Re-localization and Re-training

Priors: Motion

[Prest CVPR 12, Tang CVPR 13, Joulin ECCV 14, Kuznetsova CVPR 15, Liang ICCV 15, Liang ICCV 15, Kalogeiton PAMI 15]

☺ Motion cues for object boundaries

☹ Noisy data

1. Get spatio-temporal bounding-boxes by using long-term point trajectories [Brox & Malik ECCV 10]
2. Filter tubes with variation over time and objectness
3. Domain adaptation: videos to images

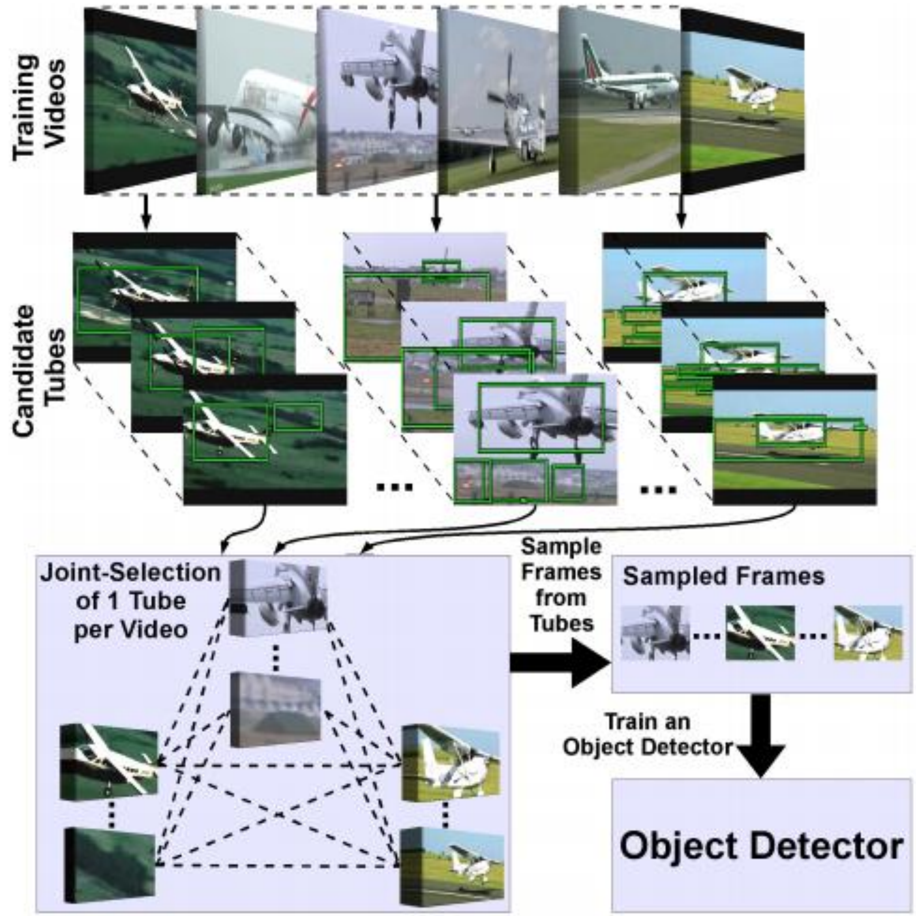
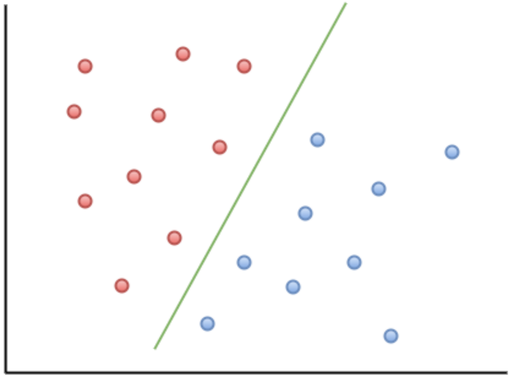


Figure [Prest CVPR 12]

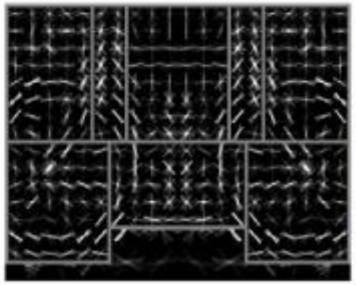
Feature representation

- SVMs on oldies (SIFT + Bag-of-words or Fisher Vectors, HOG templates)

[Chum CVPR 07, Nguyen ICCV 09, Deselaers ECCV 10, Siva ICCV 11, Russakovsky ECCV 12, Cinbis CVPR 14]



- DPM [Pandey ICCV 2011]



- CNNs as black box feature generator

[Song ICML 14, Song NIPS 14, Bilen BMVC 14, Wang ECCV 14, Bilen CVPR 15, Cinbis PAMI 16, Papadopoulos CVPR 16]

Weakly Supervised Deep Detection Nets (WSDNN)

End-to-end training with CNN [Bilen CVPR 16]

Finetuning CNNs

☺ Impressive results for supervised object detection [Fast-RCNN]

☺ CNNs learn objects and object parts in image classification [Zhou ICLR 15]

☹ High capacity leads to overfitting (standard MIL performs worse than CNN as black box feature generator)

Divide object detection into two sub-tasks with a two stream architecture

- Classification stream: assign each region to a class
- Detection stream: picks most promising windows in an image given a class
- This is **not** standard MIL (maybe mini-batch MIL)

Weakly Supervised Deep Neural Nets (WSDNN)

End-to-end training with CNN [Bilen CVPR 16]

Two stream architecture

- Classification
- Detection

CNN

Classification stream

	R_1	R_2	R_3	R_4
horse	0.52	0.47	0.04	0.93
person	0.48	0.53	0.96	0.07

Normalize over classes

Detection stream

	R_1	R_2	R_3	R_4
horse	0.04	0.01	0.07	0.88
person	0.02	0.03	0.91	0.04

Normalize over regions

$\otimes$

	R_1	R_2	R_3	R_4
horse	0.02	0.05	0.00	0.82
person	0.01	0.02	0.87	0.00

$\oplus$

horse	**0.89**
person	**0.90**

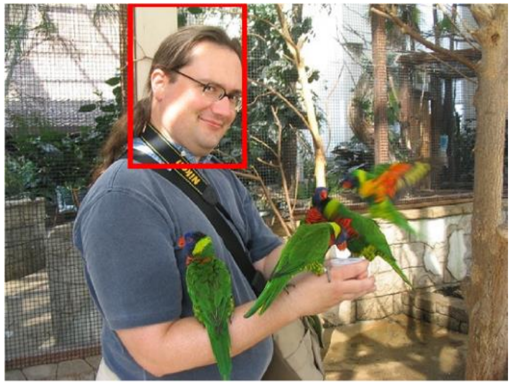
Fig: [Bilen CVPR 16]

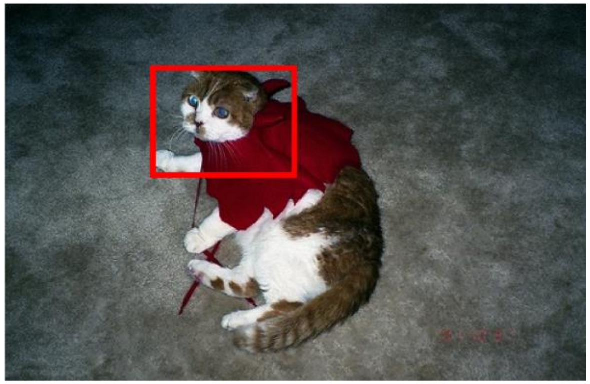
End-to-end training with CNN [Bilen CVPR 16]

☺ End-to-end learning + No custom deep learning layers

☺ State-of-the-art results with AlexNet (62% of supervised)

☹ Does not work so well with deeper networks VGG16 (56% of supervised)





It focuses on smaller regions with deeper networks.

Question: Why?

My answer: Deeper networks can recognize fine-grained differences!

How to improve WSOD for deeper nets?

Cascaded object detection [Diba CVPR 17]

- Stage 1: Better class activation maps, provides a subset of windows
- Stage 2: Selects highest scoring proposal window
- Additional final step: Trains a Fast-RCNN
- Back to 64% of supervised counterpart (Fast-RCNN)

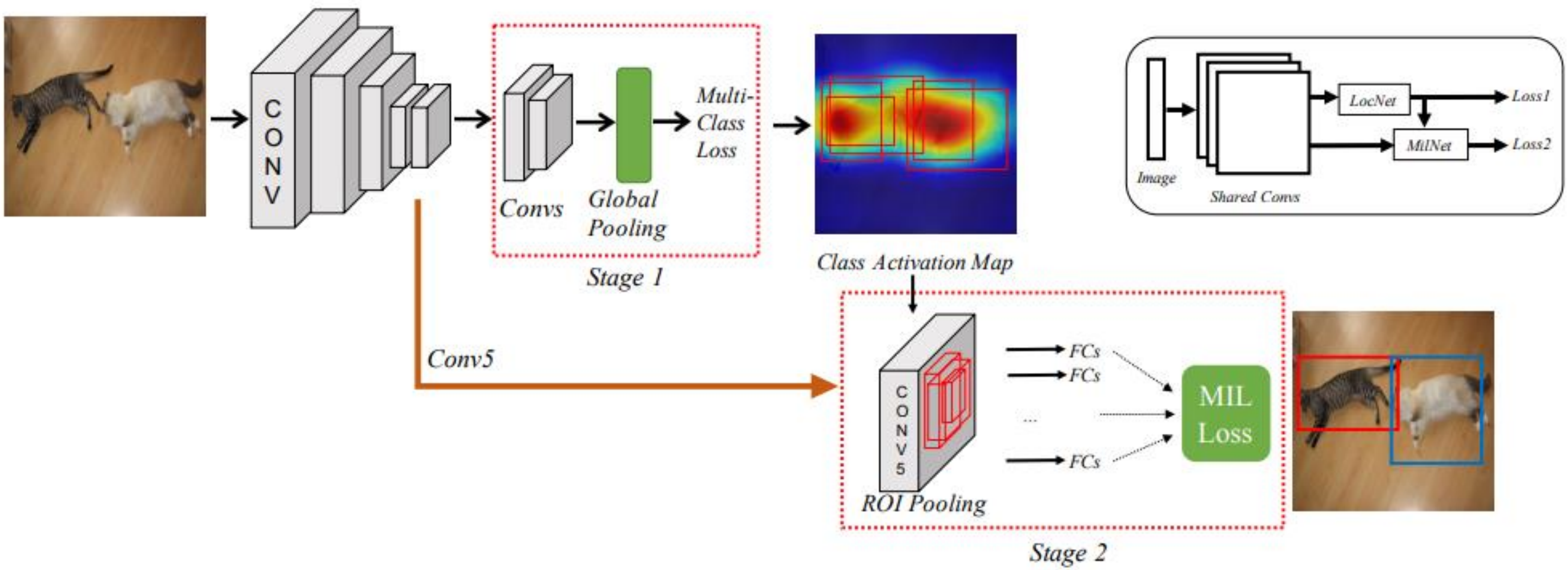


Figure [Diba CVPR 17]

How to improve WSOD for deeper nets?

Refining predictions [Tang CVPR 17]

1. Train a WSDDN
2. Get highest scoring proposal for positives and find overlapping proposals
3. Gradually add them to training as positive instances

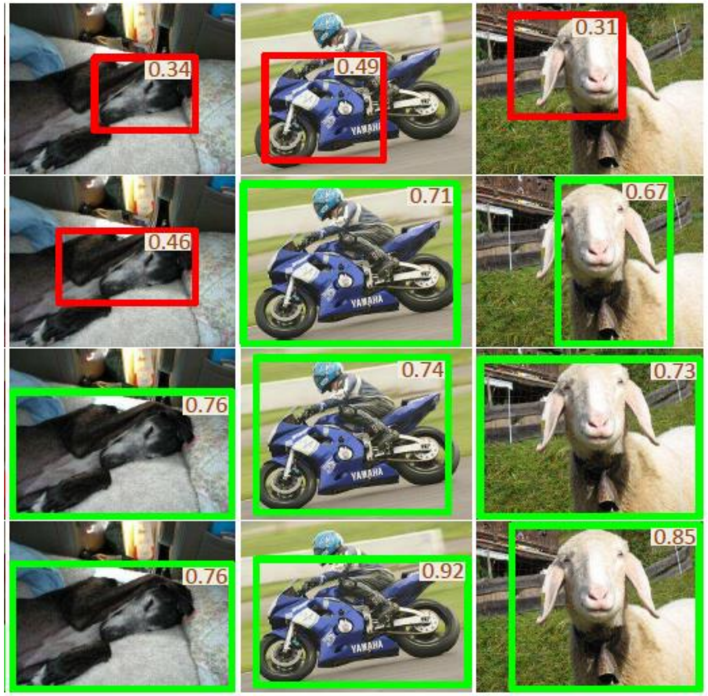
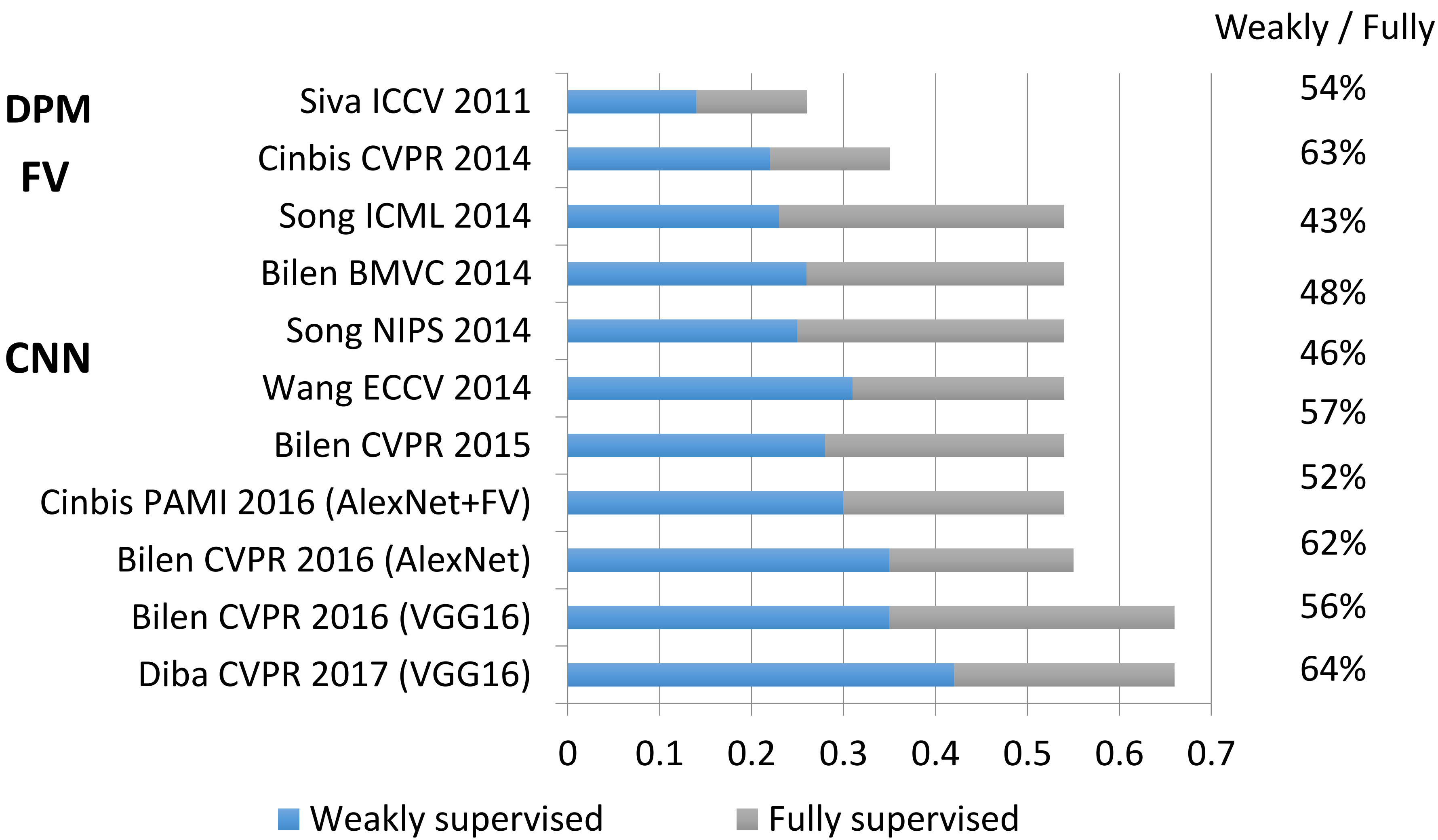


Figure [Tang CVPR 17]

Performance at test time

WSL on PASCAL 07 trainval all views, test on test (mAP)



Performance still far from fully supervised detector

Conclusions on weakly supervised object detection

- WSOD is challenging due to
  - intra-class variations,
  - ambiguity with parts and context,
  - sensitive to initialization,
  - prone to overfitting

- Solutions are
  - Using smart initialization strategies
  - Robust re-localization and re-training methods
  - Incorporating prior knowledge

Looking for a post-doc

University of Edinburgh

Starting date: October 2018 or after

Please contact: hbilen@ed.ac.uk

Questions