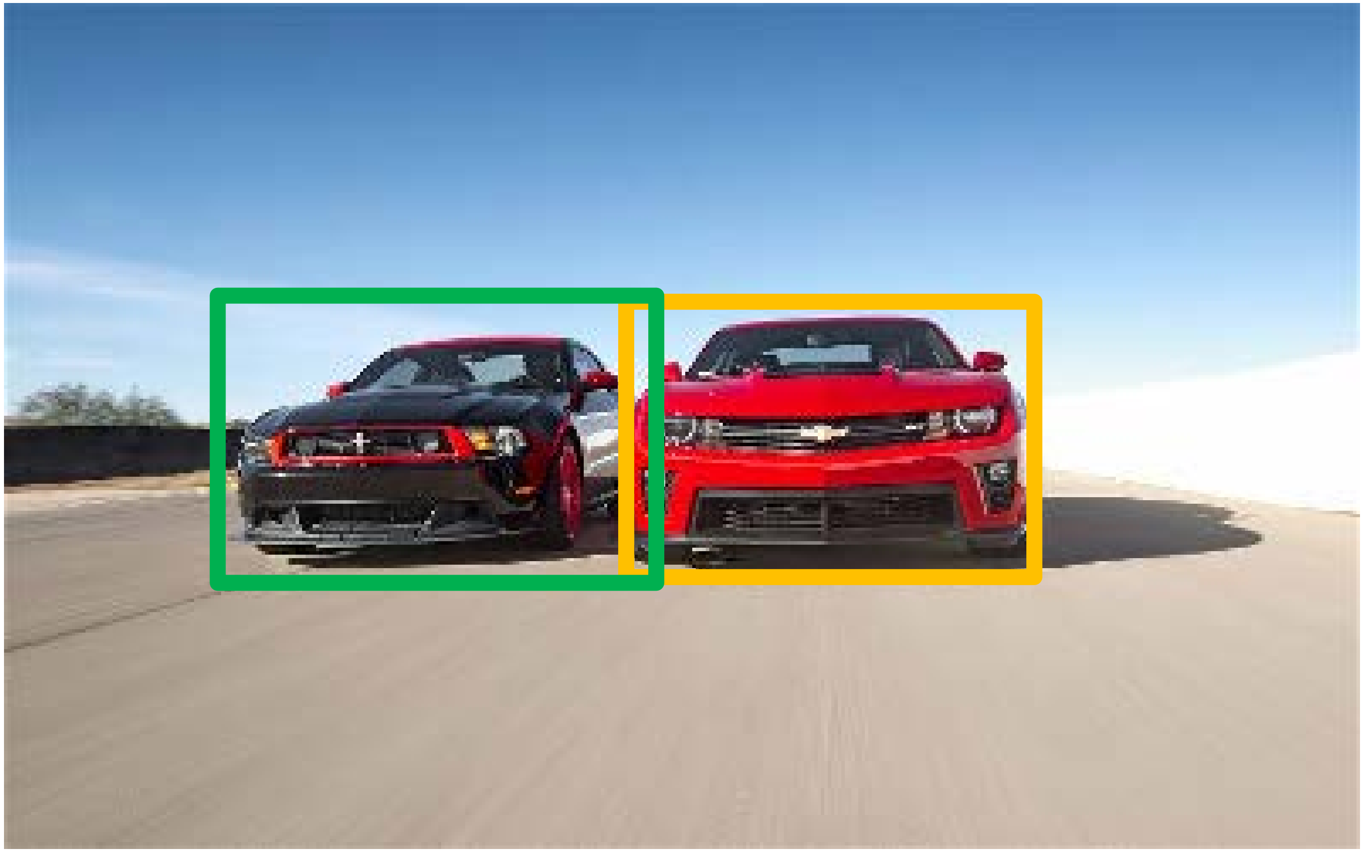
Object Co-detection

Yingze (Sid) Bao, Yu Xiang, Silvio Savarese

Midwest Workshop 2012, ECCV 2012

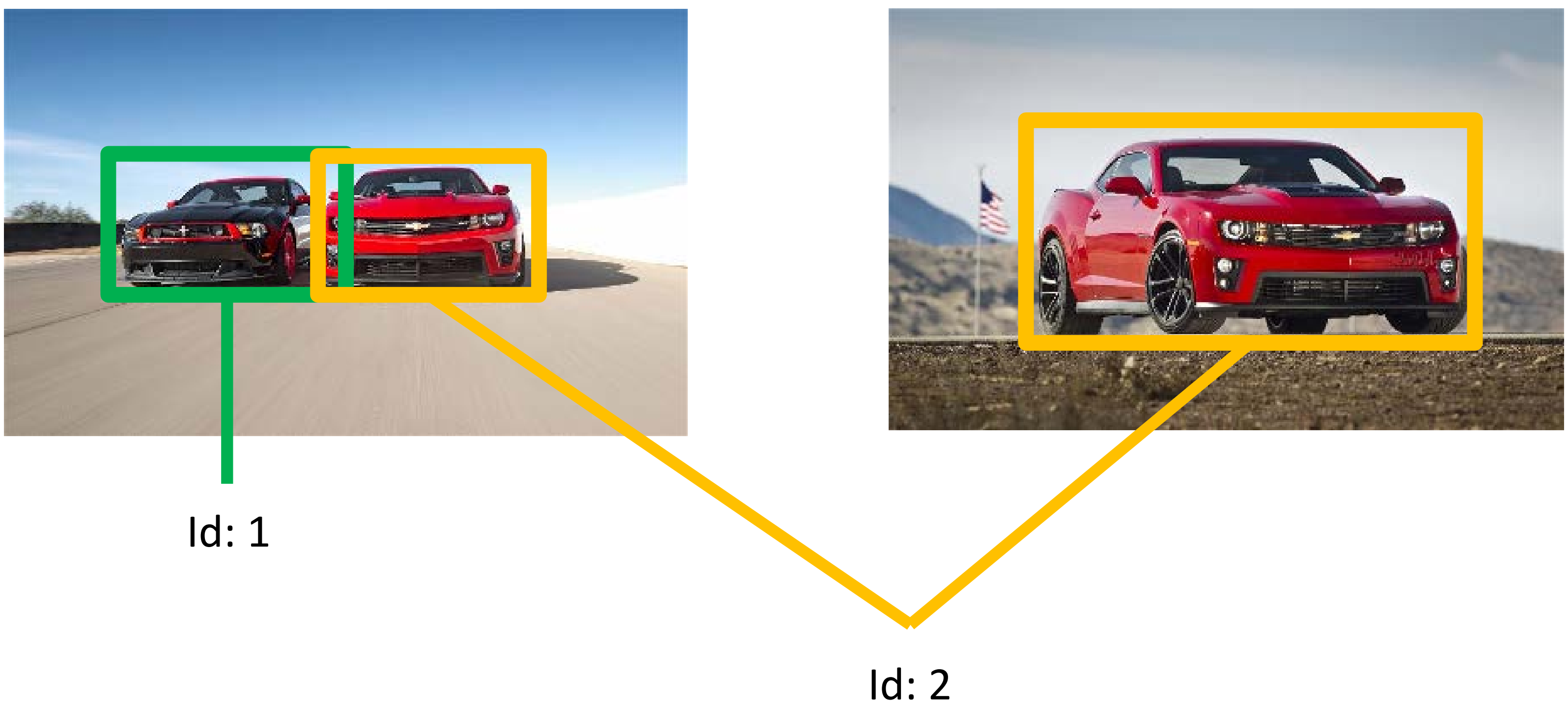
Object Detection: A Review



Object Co-Detection

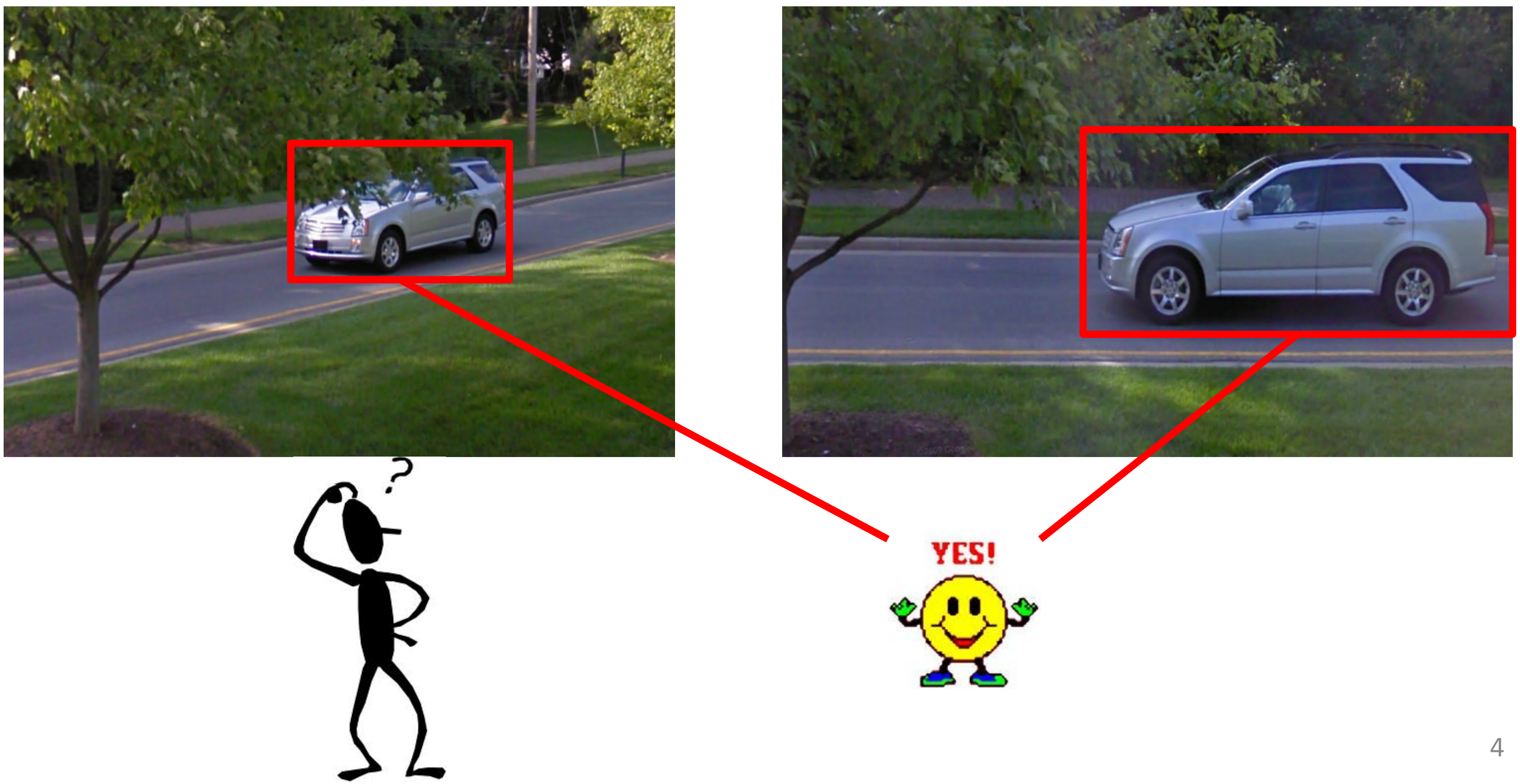
- Detect objects in multiple images
- Establish object correspondences
- Estimate object viewpoint changes

A Coherent Algorithm



Motivations

- Better detection accuracy than single-image methods



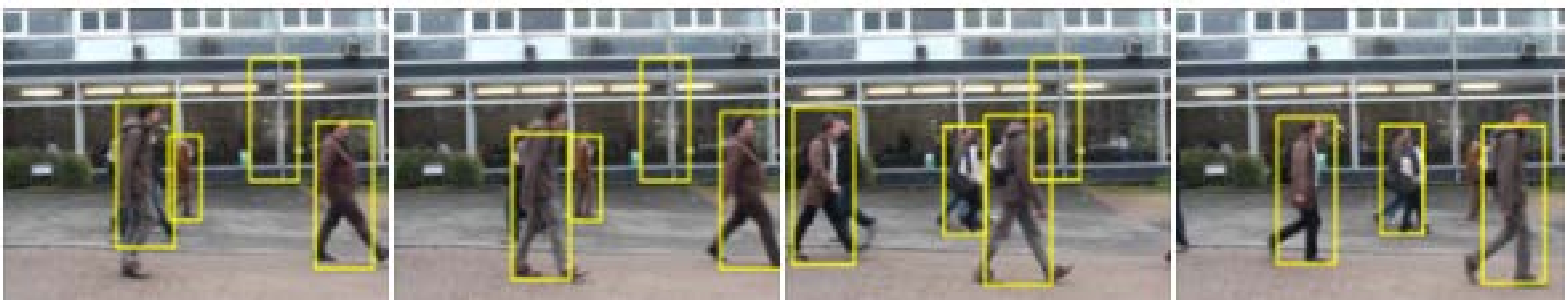
Motivations

- Better detection accuracy than single-image methods
- Better matching accuracy than low-level methods

Where is the car?

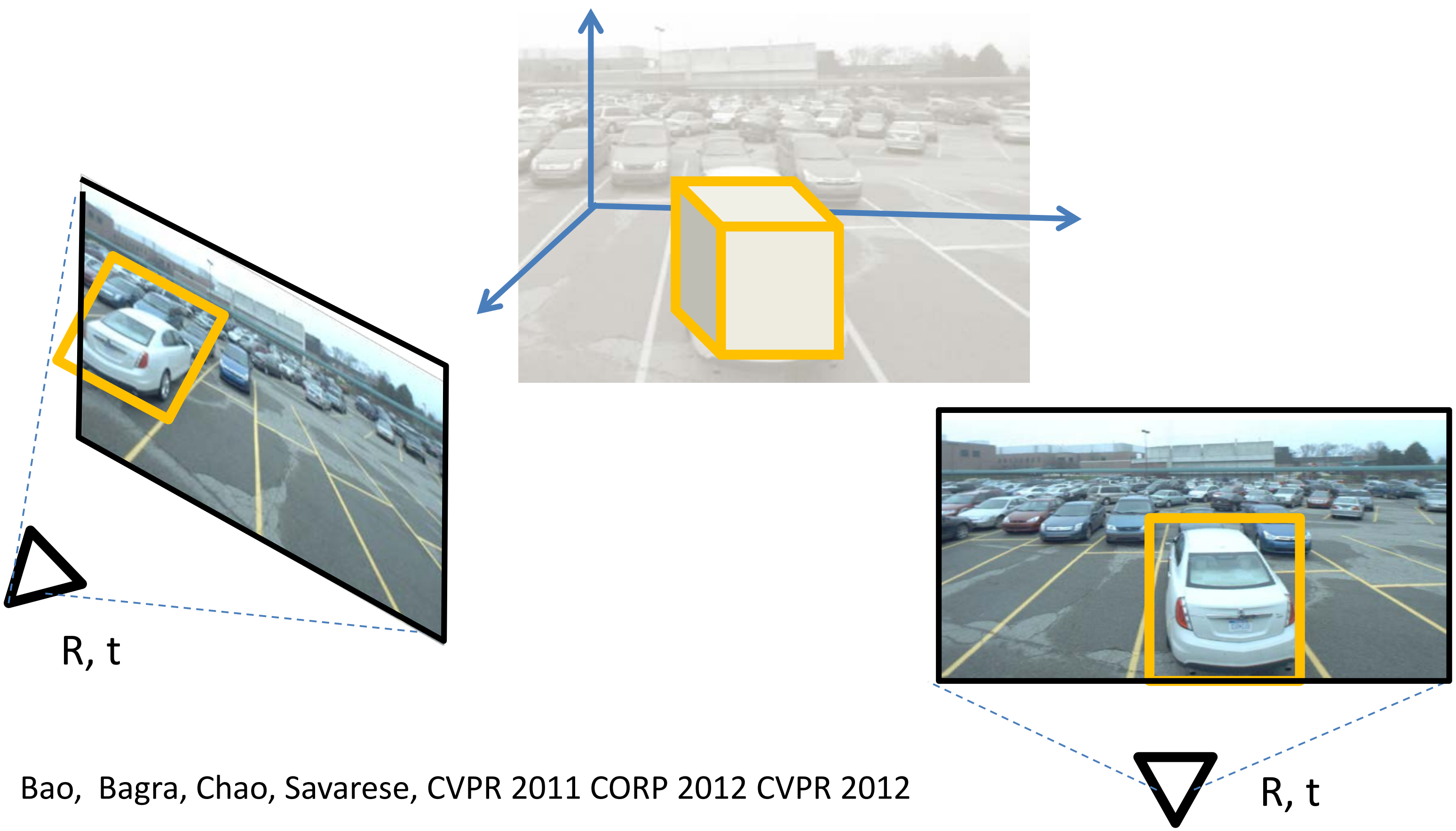
Motivations

- Better detection accuracy than single-image methods
- Better matching accuracy than low-level methods
- Tracking by Detection
  - Co-detection provides consistent detection across frames



Motivations

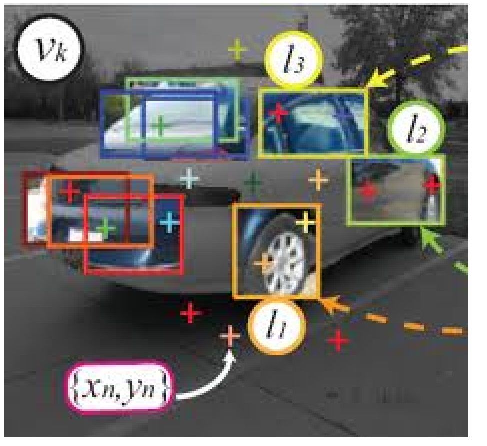
- Better detection accuracy than single-image methods
- Better matching accuracy than low-level methods
- Tracking by Detection
- Semantic Structure From Motion

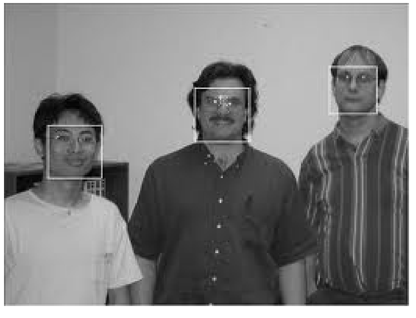


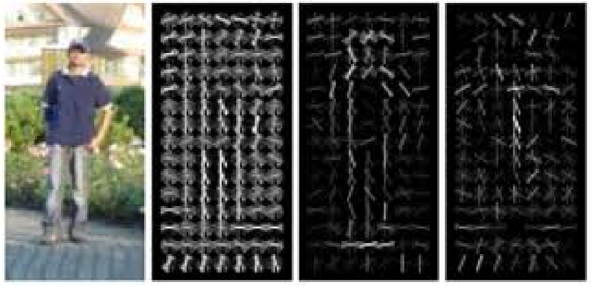
Bao, Bagra, Chao, Savarese, CVPR 2011 CORP 2012 CVPR 2012

Related Problems

- Object detection in a single image
  - Viola et al. 2001
  - Fergus et al. 2003
  - Leibe et al. 2004
  - Dalal et al. 2005
  - Savarese et al. 2006
  - Felzenszwalb et al. 2009
  - etc….

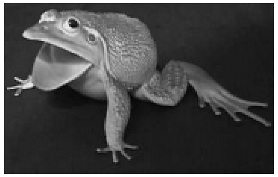






Related Problems

- Object detection in a single image
- Single instance detection
  - Low level image features
  - Small pose variation
  - Rich texture



Lowe 1999

- Lowe 1999
- Berg et al. 2005
- Ferrari et al. 2006
- Nister et al. 2006
- Rothganger et al.2006
- Hsiao et al. 2010
- etc.

Related Problems

- Object detection in a single image
- Single instance detection
- Co-segmentation
  - No semantic information
  - Hard to handle objects with different poses

- Rother et al. 2006
- Batra et al. 2010
- Hochbaum et al. 2009
- etc.

Input Image pair

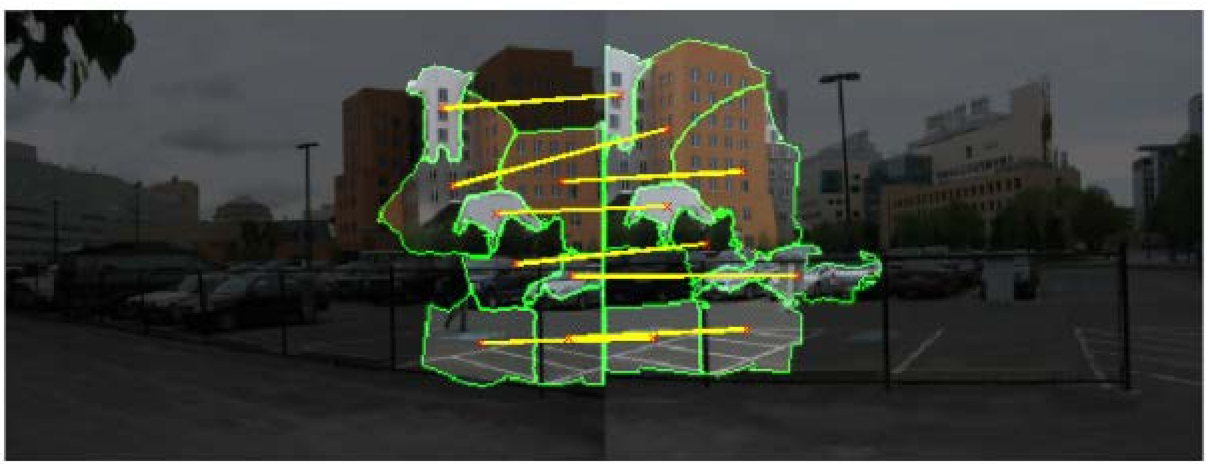
Rother et al. 2006

Cosegmentation

Related Problems

- Object detection in a single image
- Single instance detection
- Co-segmentation
- Region matching
  - No semantic information 
  - May require epipolar geometry validation
  - Sensitive to segmentation noise

- Schaffalitzky et al. 2001
- Tuytelaars et al. 2004
- Matas et al. 2004
- Toshev et al. 2007
- etc.

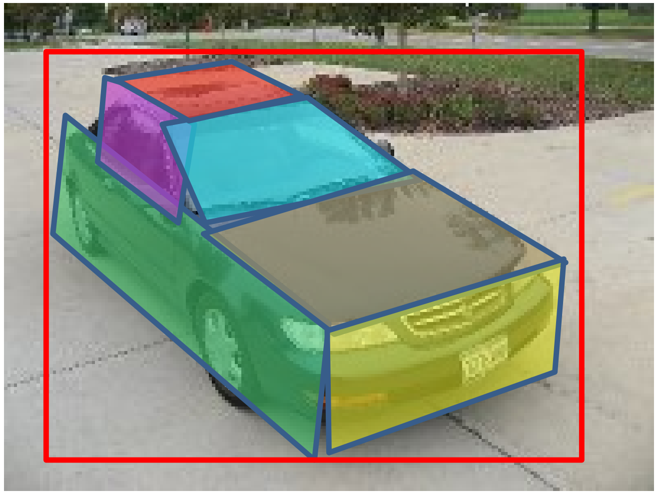


Toshev et al. 2007

Object Co-detection is challenging

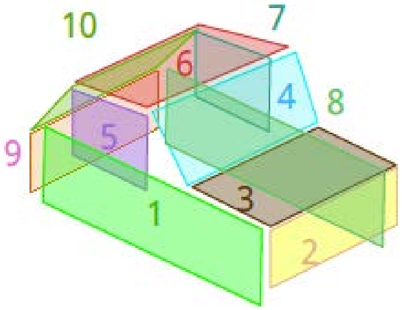
Pose Variation & Self-occlusion

Object Representation

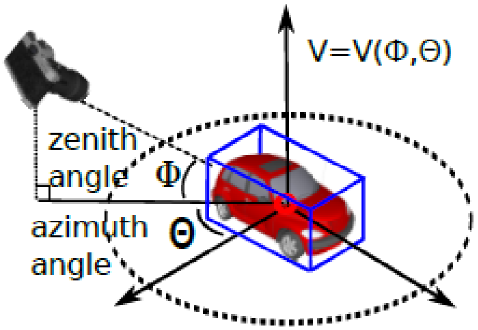


- p : part
- r : root filter
- V : view point

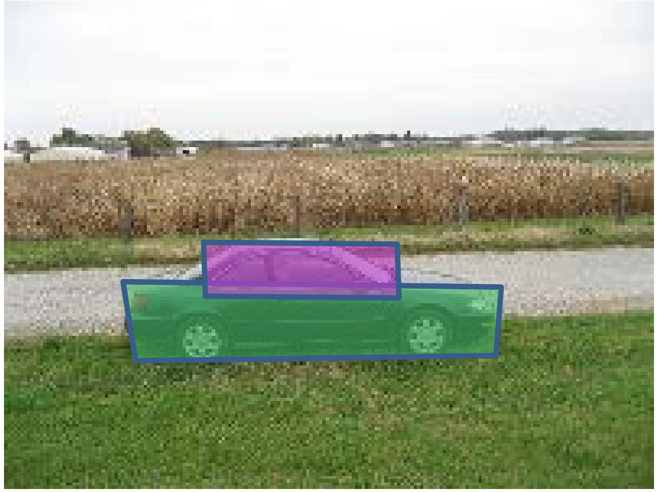
$O = \{r, V, p_1, p_2, \ldots, p_n\}$

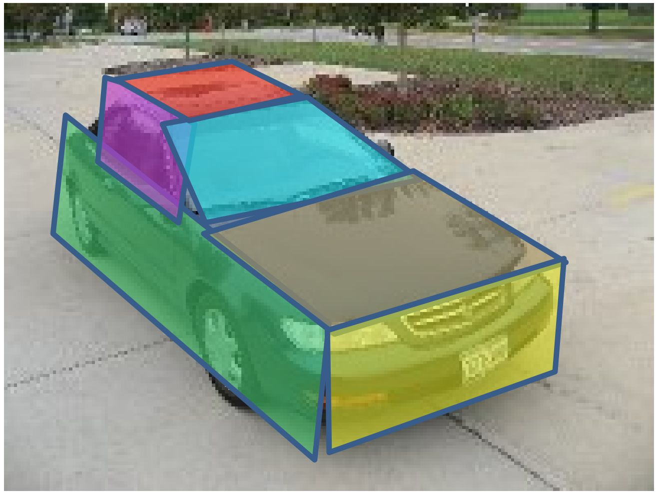


id, location, pose, visibility and scale



Pose Variation & Self-occlusion





Goal of object co-detection

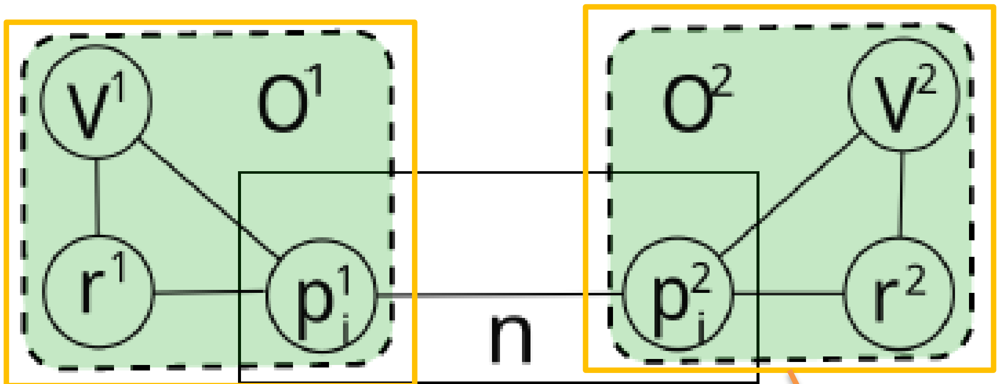
- Identify matching objects in every input images respectively

$$\{O_1, O_2, \dots, O_K\} = arg\max_{\{O_k\}} E(O_1, O_2, \dots, O_K; \boldsymbol{I})$$

- O_k: an object in image I_k
- E: energy based on the co-detection model
  - What is the co-detection model?

Co-detection Model

E_{unit}

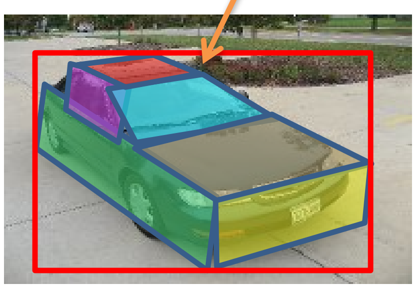


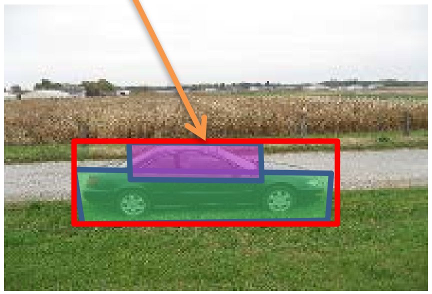
p : part

r : bounding box

V : view point

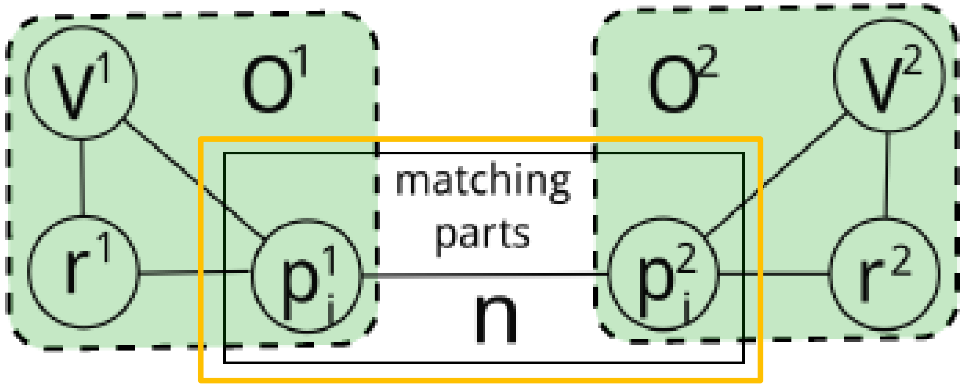
- In the case of 2 input images





Co-detection Model

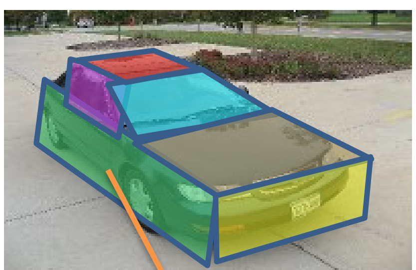
E_{unit}



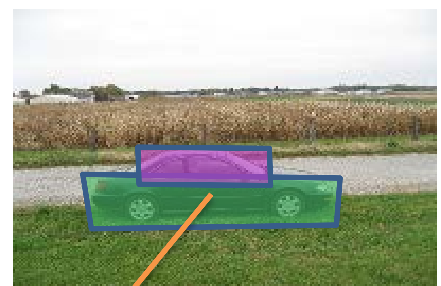
E_{match}

p : part
r : bounding box
V : view point

- In the case of 2 input images



Account for visibility!

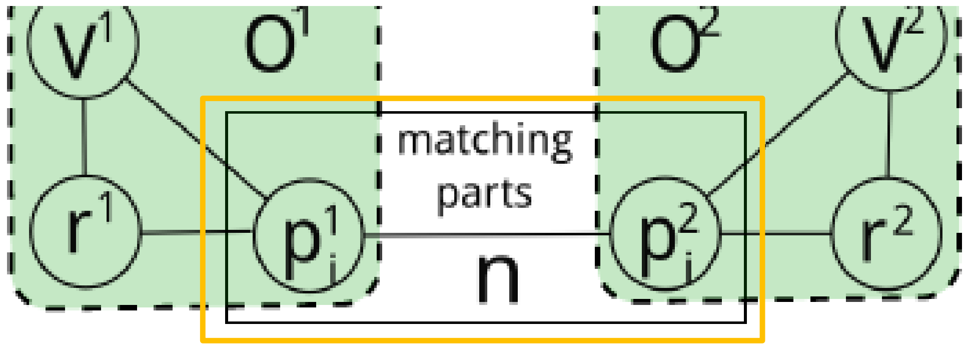


Match

Rectification ⟺ Rectification

$$E(\mathcal{O},\mathcal{I}) = \sum_{k=1}^{K} E_{\text{unit}}(O^k, I^k) + \sum_{i=1}^{n} E_{\text{match}}(\{p_i^k\}_{k=1}^{K}, \{V^k\}_{k=1}^{K}, \mathcal{I})$$

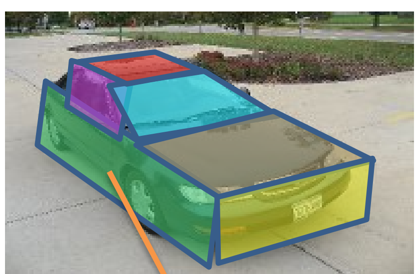
E_{unit}



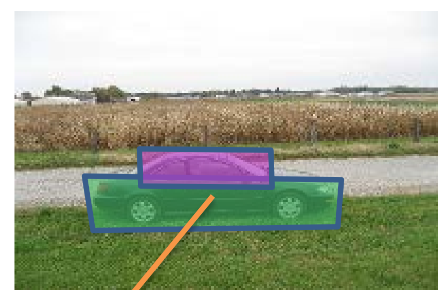
p : part
r : bounding box
V : view point

- In the case of 2 input images

E_{match}



Account for visibility!



Match

Rectification

Rectification

$$E(\mathcal{O},\mathcal{I}) = \sum_{k=1}^{K} E_{\text{unit}}(O^k, I^k) + \sum_{i=1}^{n} E_{\text{match}}(\{p_i^k\}_{k=1}^{K}, \{V^k\}_{k=1}^{K}, \mathcal{I})$$

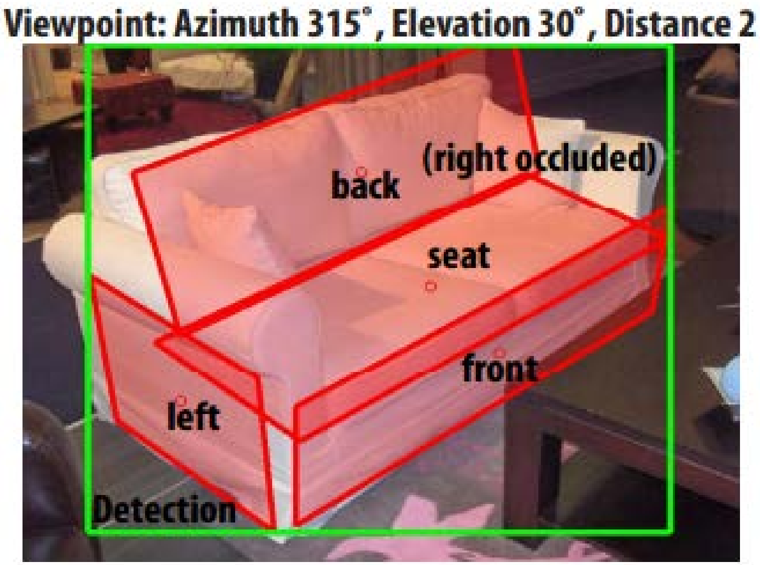
Set of objects

Set of input images

$$E(\mathcal{O},\mathcal{I}) = \sum_{k=1}^{K} \boxed{E_{\text{unit}}(O^k, I^k)} + \sum_{i=1}^{n} E_{\text{match}}(\{p_i^k\}_{k=1}^{K}, \{V^k\}_{k=1}^{K}, \mathcal{I})$$

- Unitary potential
  - Single image object detector

$$E_{\text{unit}}(O^k, I^k) = E_{\text{root}}(r^k, V^k, I^k) + \sum_{i=1}^{n} E_{\text{part}}(p_i^k, V^k, I^k) + \sum_{i=1}^{n} E_{\text{rp}}(r^k, p_i^k, V^k, I^k)$$



Xiang & Savarese CVPR 12

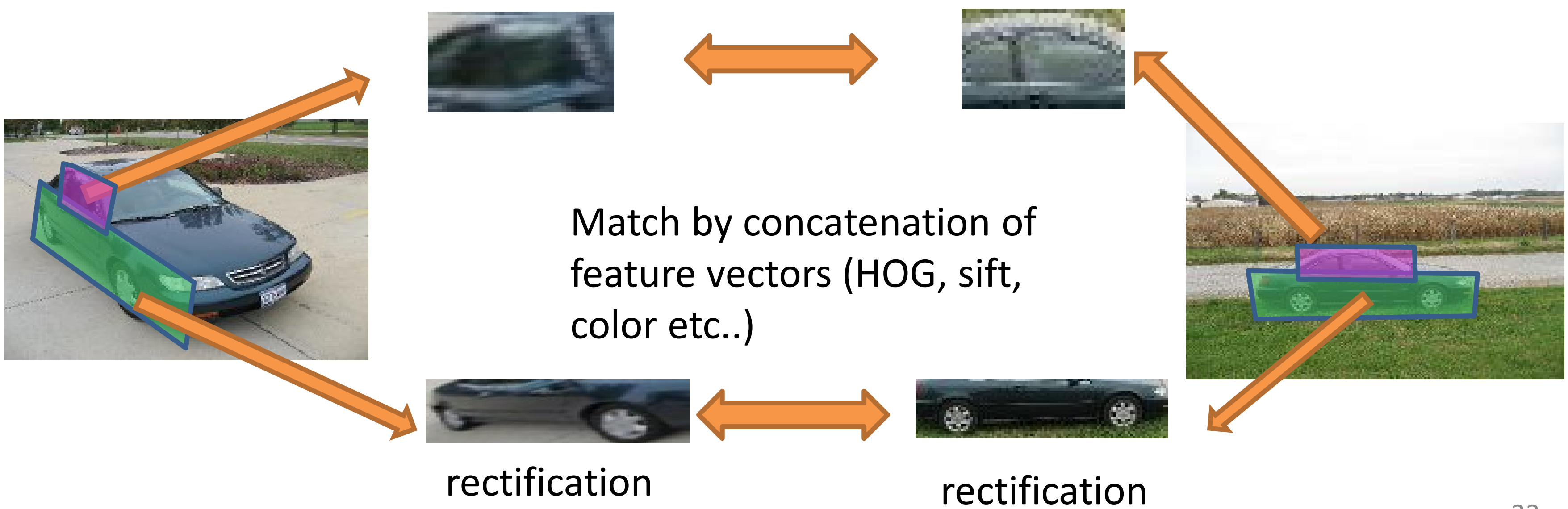
Good with many part-based object detection models!

- Fergus et al. 03
- Leibe et al. 04
- Silvio & Feifei 06
- Kushal et. al. 07
- Chiu et. al. 07
- Sun et al. 09
- Felzenszwalb 09
- Filder 09

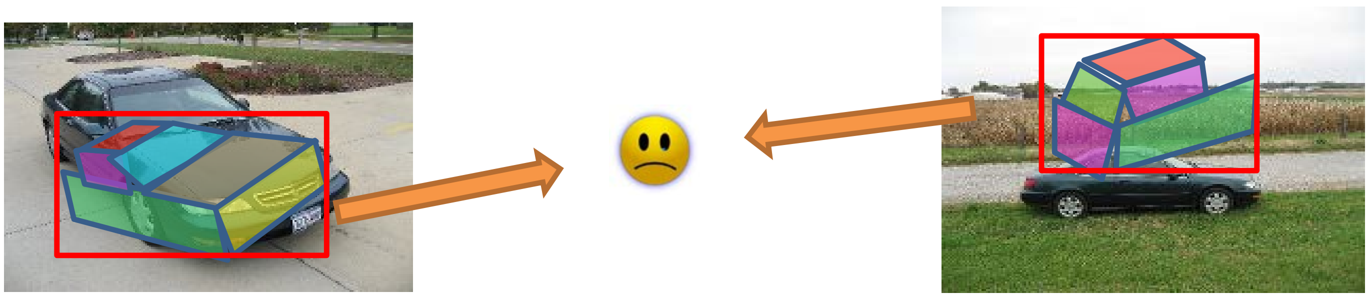
$$E(\mathcal{O},\mathcal{I}) = \sum_{k=1}^{K} E_{\mathrm{unit}}(O^k, I^k) + \sum_{i=1}^{n} \boxed{E_{\mathrm{match}}(\{p_i^k\}_{k=1}^{K}, \{V^k\}_{k=1}^{K}, \mathcal{I})}$$

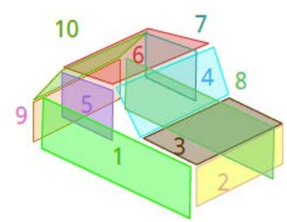
- Matching potential
- Decompose into pair-wise terms

$$E_{\mathrm{match}}(\{p_i^k\}_{k=1}^{K}, \{V^k\}_{k=1}^{K}, \mathcal{I}) = \frac{1}{C_K^2} \sum_{k_1,k_2} \boxed{M(p_i^{k_1}, p_i^{k_2}, V^{k_1}, V^{k_2}, I^{k_1}, I^{k_2})}$$



Single Image Detector False Alarms





Inference

$$\mathcal{O}^* = \arg\max_{\mathcal{O}} E(\mathcal{O}, \mathcal{I})$$

Goal: detect matching objects

All possible configurations of matching objects in different images

- Loopy model
  - approximating inference with 2 steps

$$E(\mathcal{O},\mathcal{I}) = \sum_{k=1}^{K} E_{\text{unit}}(O^k, I^k) + \sum_{i=1}^{n} E_{\text{match}}(\{p_i^k\}_{k=1}^K, \{V^k\}_{k=1}^K, \mathcal{I})$$

- Step 1: $\tilde{O}$ objects with high unitary potential $\sum_{k=1}^{K} E_{\text{unit}}(O^k, I^k)$
- Step 2: $O^* = arg \max_{\tilde{O}} \sum_{k=1}^{K} E_{\text{unit}}(O^k, I^k) + \sum_{i=1}^{n} E_{\text{match}}(\{p_i^k\}_{k=1}^K, \{V^k\}_{k=1}^K, \mathcal{I})$

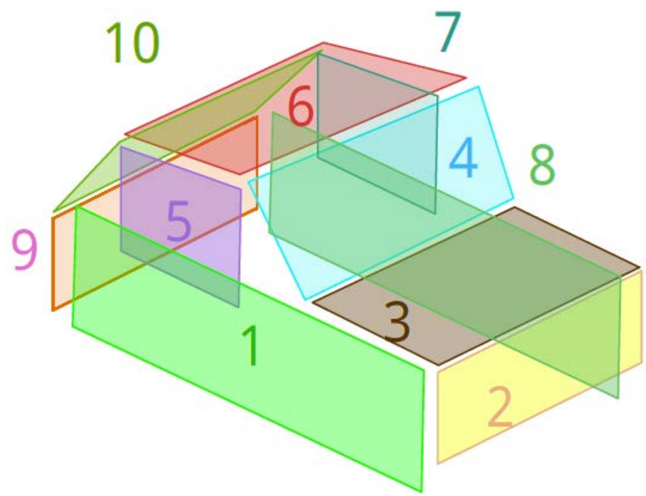
Learn the model

- Learn parameters for E_{unit}
  - Standard learning process in a part-based object detection model

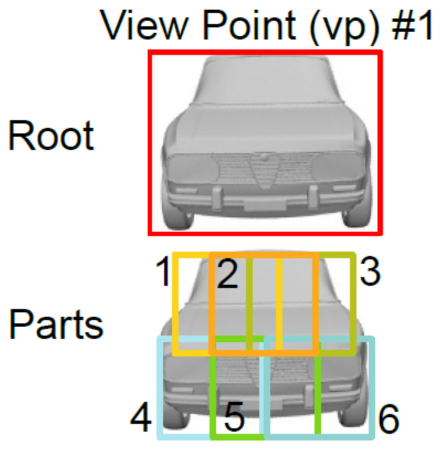
$$\{\beta_{\text{root}}, \beta_{\text{part}}, \beta_{\text{rp}}, \mathbf{w}\}$$
$$= \arg \min_{\beta_{\text{root}}, \beta_{\text{part}}, \beta_{\text{rp}}, \mathbf{w}} \frac{1}{2}(\|\beta_{\text{root}}\|^2 + \|\beta_{\text{part}}\|^2 + \|\beta_{\text{rp}}\|^2 + \|\mathbf{w}\|^2) +$$
$$\lambda \sum_t \max(0, 1 - y_t \max_{\mathcal{P}^t} E(\mathcal{O}^t, \mathcal{I}^t)),$$

- Learn parameters for E_{match}
  - Learning weights **w** of different matching cues (e.g. HOG, sift, color) in a SVM learning framework

$$\mathrm{w} = \arg\min_{\mathrm{w}} \frac{1}{2} \sum_i \|\mathrm{w}_i\|^2 + \lambda \sum_{t=1}^{T} \max(0, 1 - y_t [\sum_{i=1}^{n} E_{match}(\{\bar{p}_i^k\}_{k=1}^K, \{V^k\}_{k=1}^K, \mathcal{I})])$$


10
7
6
4
8
5
9
3
1
2

2D Object Part Representation – A Simplification



- Treat different view points as different object categories
  - Cannot match objects with large pose variations
- More choices to compute E_{unit} 
  - Fergus et al. CVPR'03
  - Leibe et al. 04
  - Felzenszwalb 09
  - Etc.

Experiments

Experiments

- 3 Datasets
  - Cars
    - 300 image pairs
    - Pandey et al. 2009, Bao et al. 2010
  - Pedestrians
    - 200 image pairs
    - Ess et al. 2007
  - 3d objects
    - ~400 image pairs for 8 categories each
    - Savarese & Fei-fei. 2006

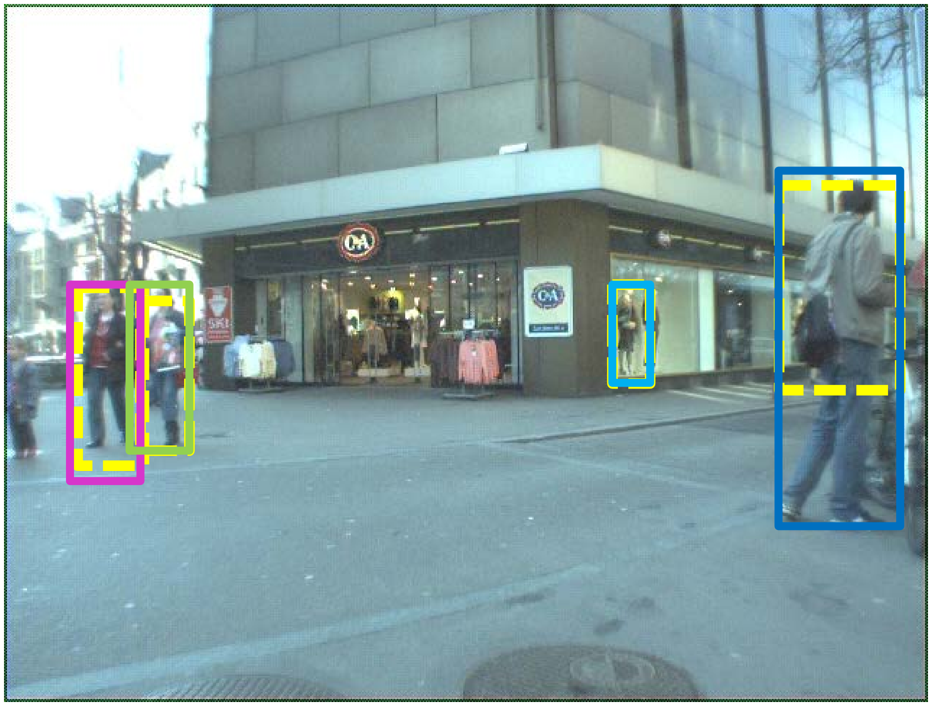
Car dataset

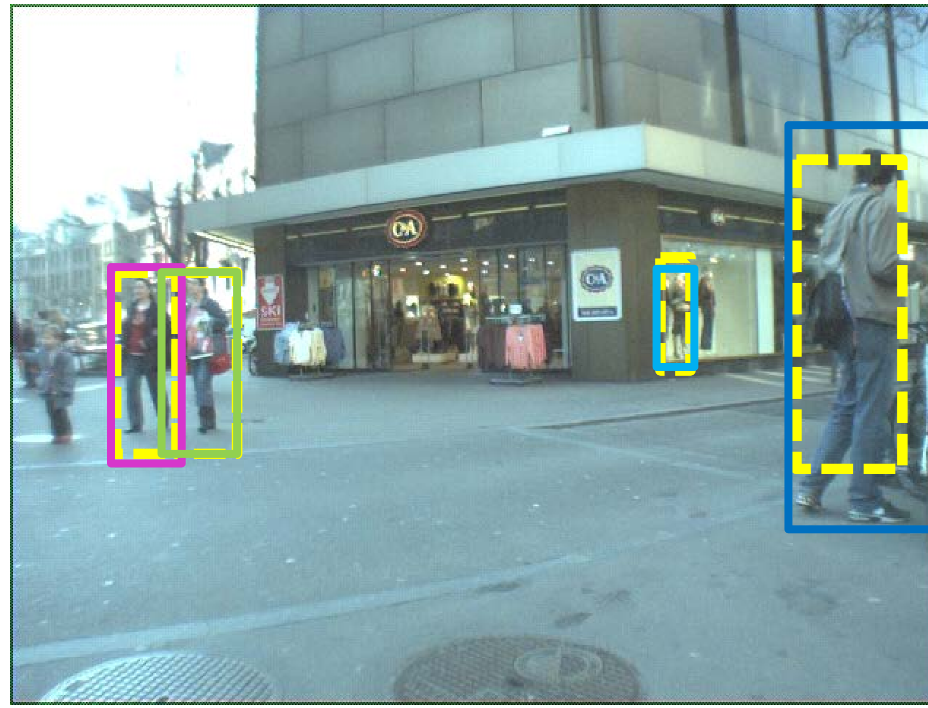
Single Img. Det.

Co-detector

(2d object representation applied)

Pedestrian dataset



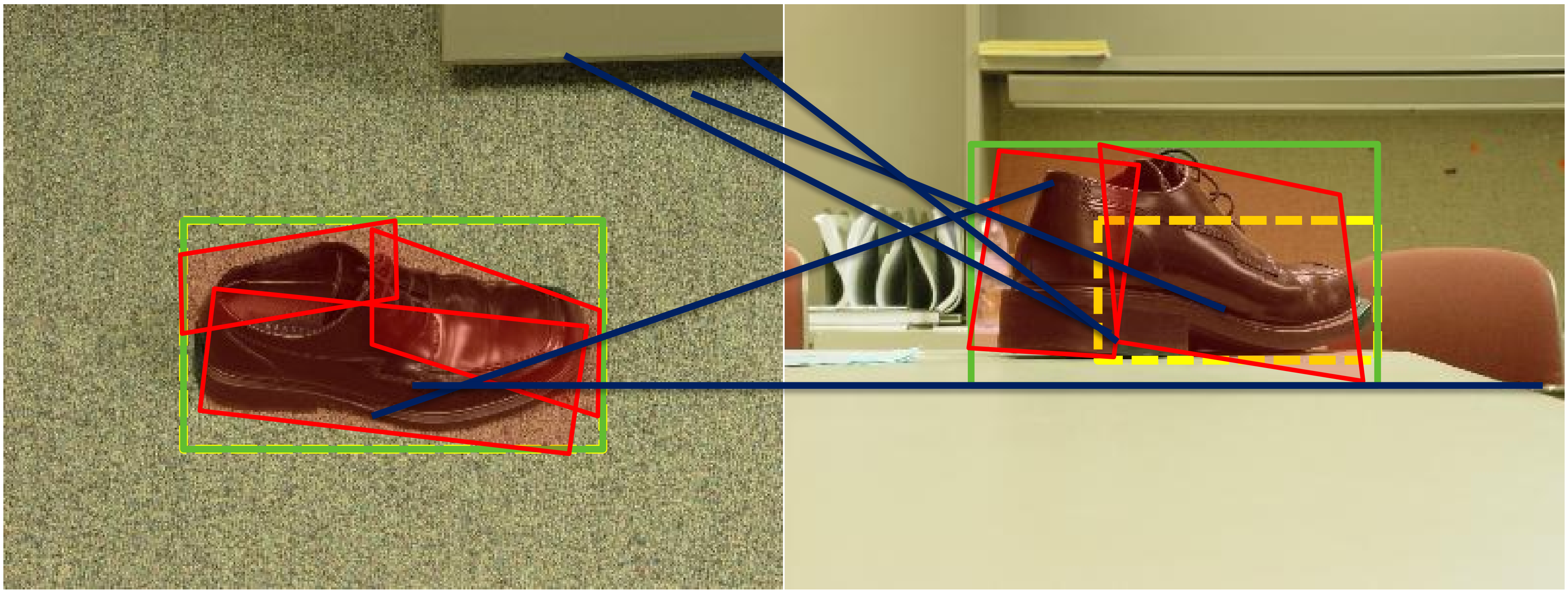


Single Img. Det.

Co-detector

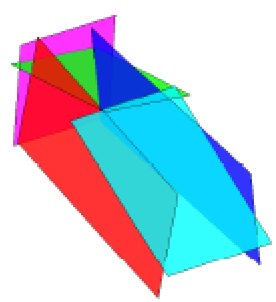
(2d object representation applied)

3d object dataset



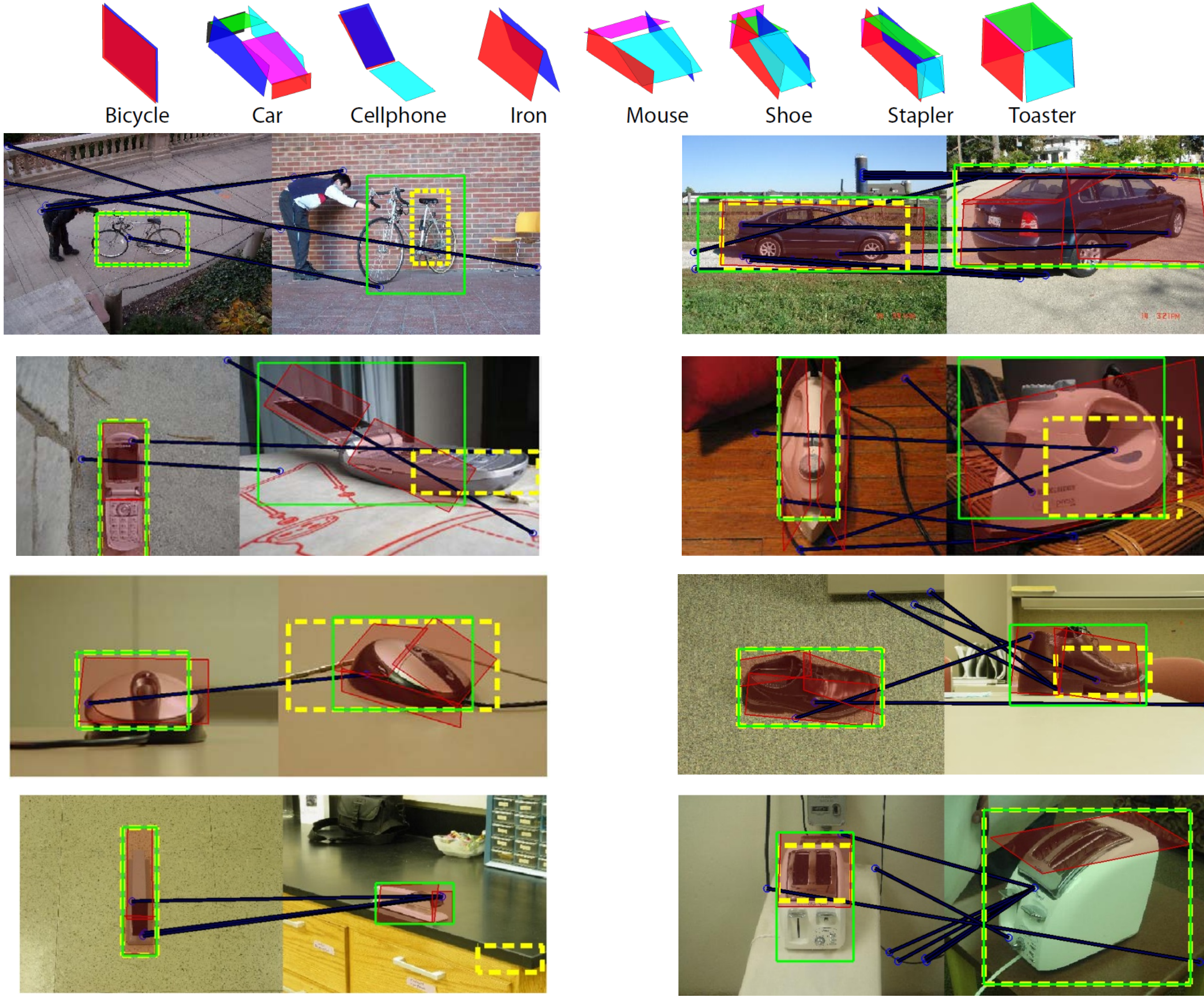
Single Img. Det.

Co-detector



Shoe model

SIFT match

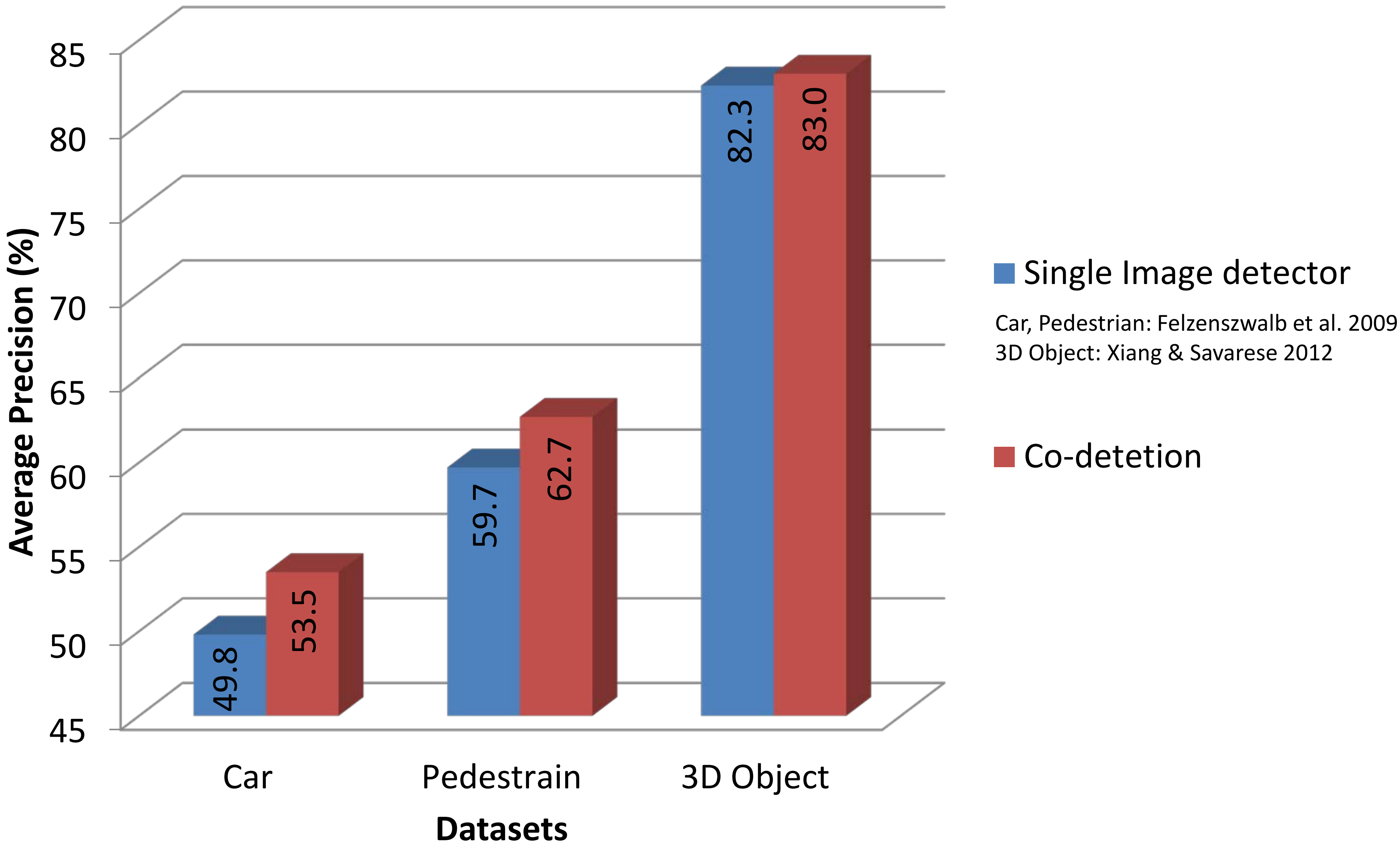

Bicycle
Car
Cellphone
Iron
Mouse
Shoe
Stapler
Toaster

Quantitative Evaluation

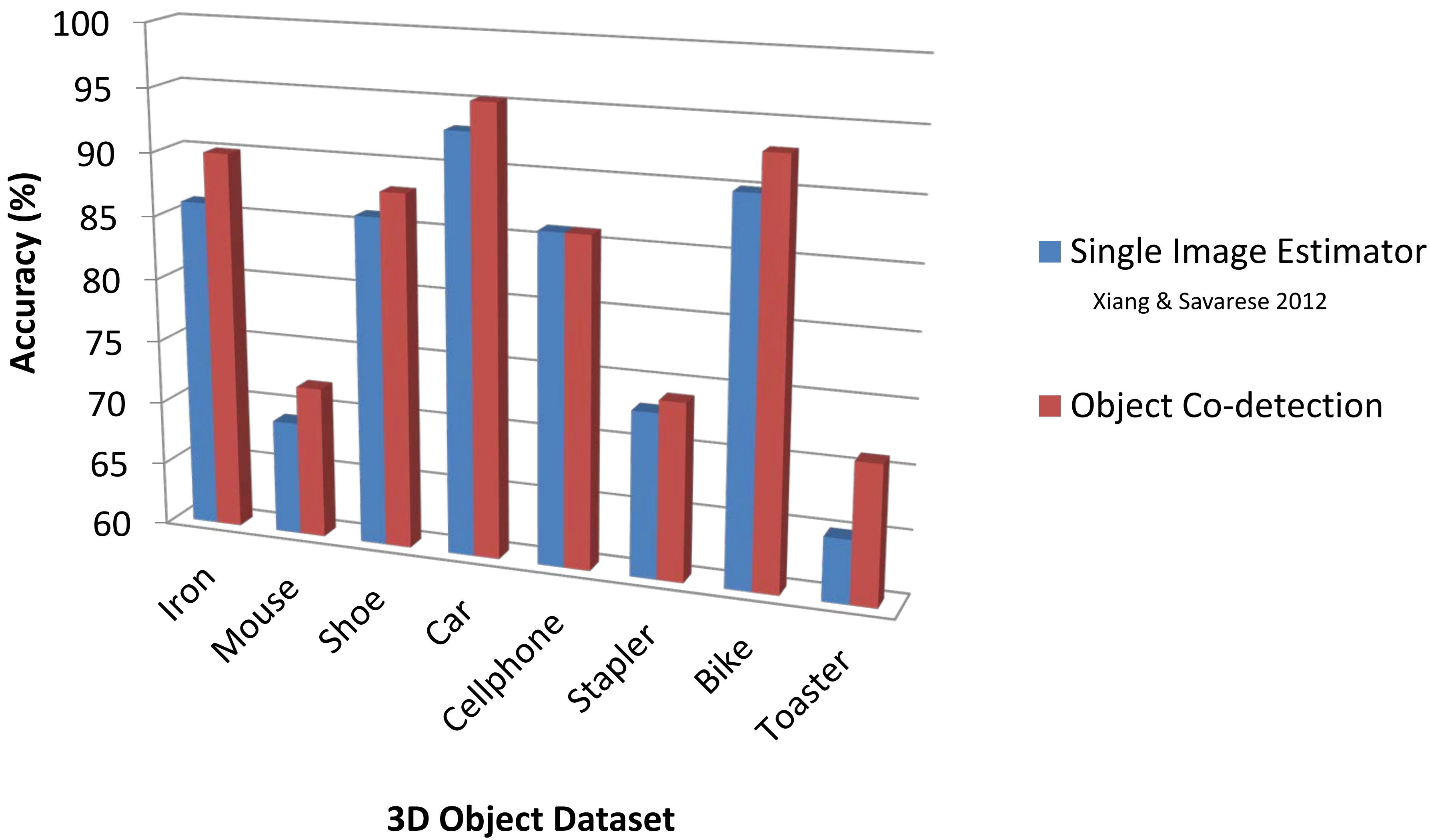
- Object detection
- Pose estimation
- Single instance detection

(More results in the paper)

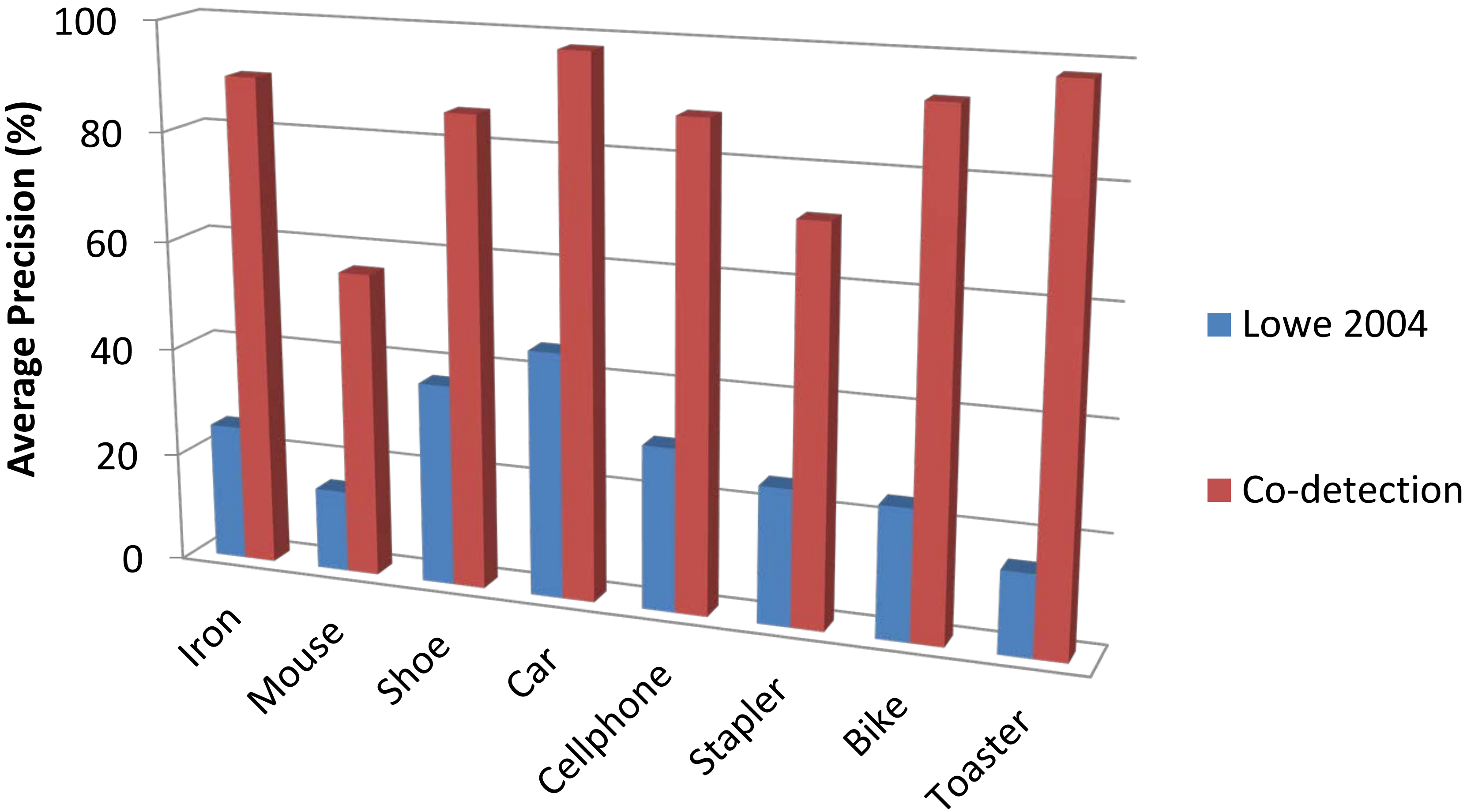
Object Detection Accuracy



Pose Estimation Accuracy

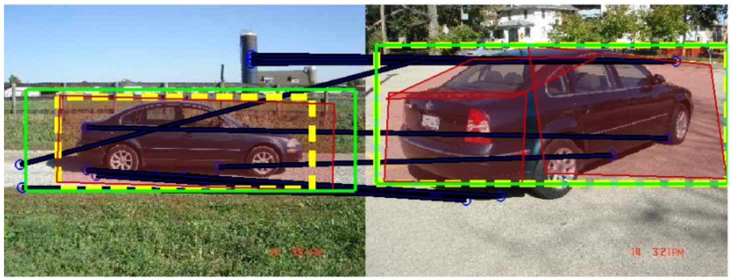


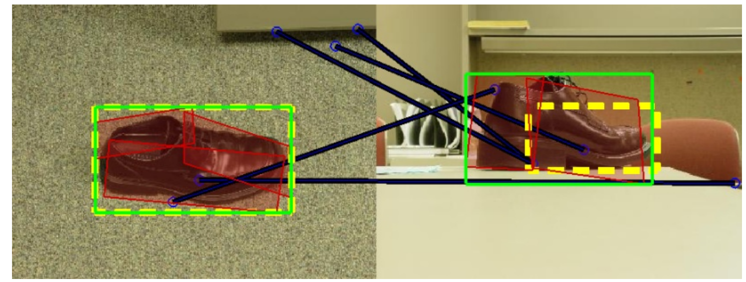
Detecting the same instance

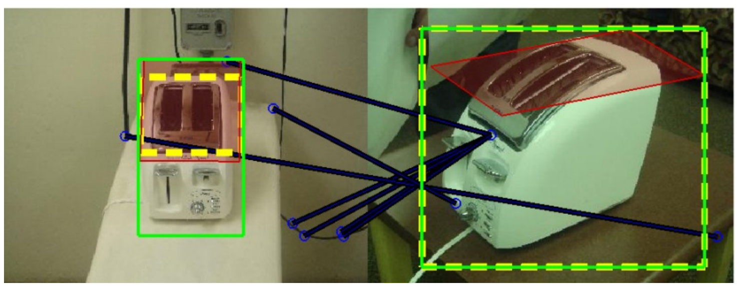


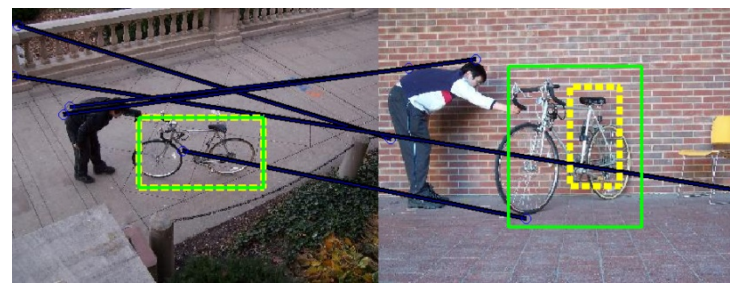
(same poses)

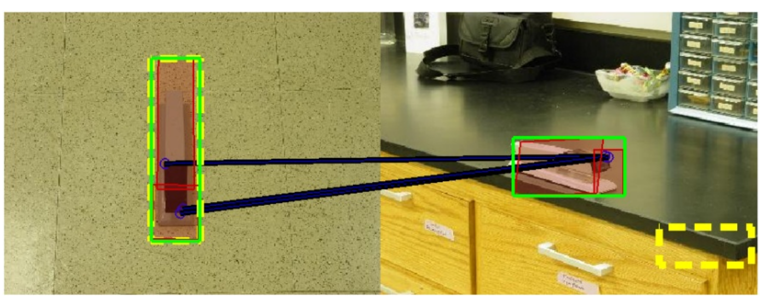
Why the co-detection is better at the instance detection task?

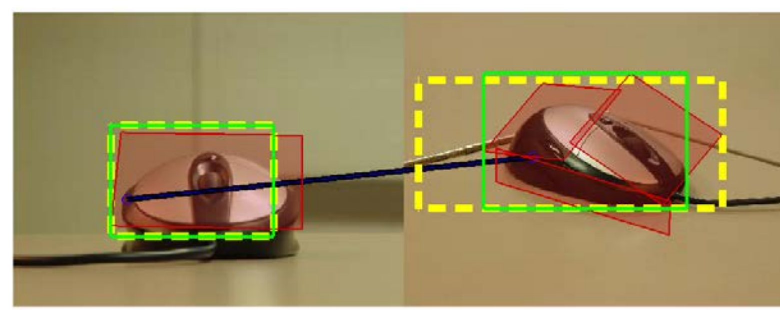












- Object co-detection problem
  - A generalization of the object detection problem
- Our solution
  - Exploit existing object representation models
  - Measure object similarity by parts
- Experiments
  - Superior performance in extensive tests
- Acknowledgement