Localizing 3D Cuboids in Single-view Images
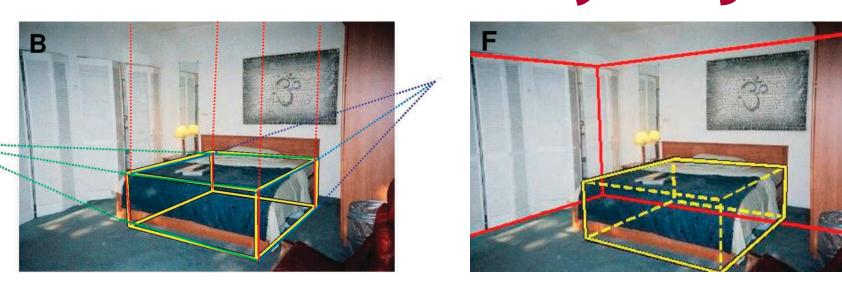
Jianxiong Xiao (MIT) Bryan C. Russell (Intel Labs) Antonio Torralba (MIT)
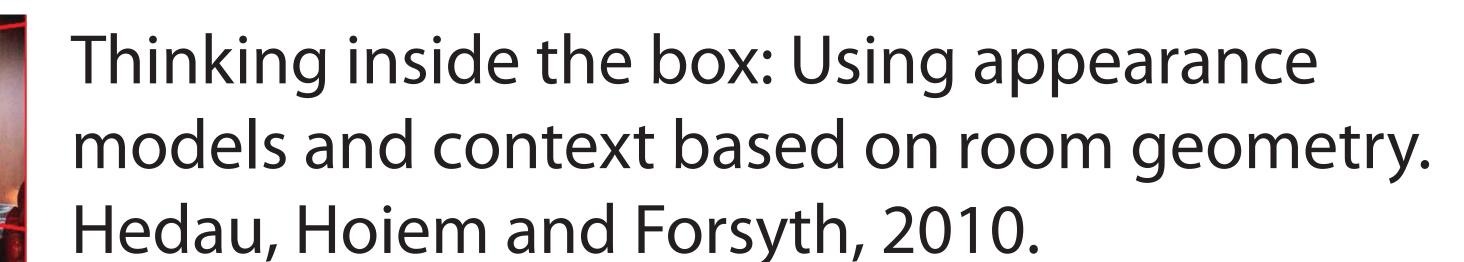
SUNprimitive

Dataset + Code: http://SUNprimitive.csail.mit.edu
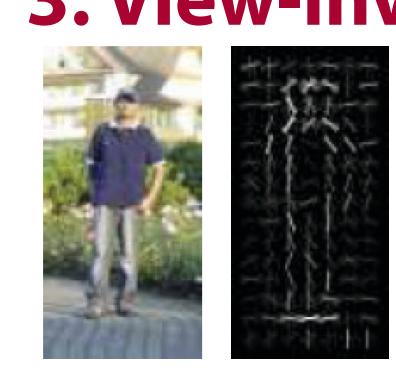
Motivations

1. Standalone boxy object detector

Thinking inside the box: Using appearance models and context based on room geometry. Hedau, Hoiem and Forsyth, 2010.

2. Cuboidness: not category

3. View-invariant 3D detector
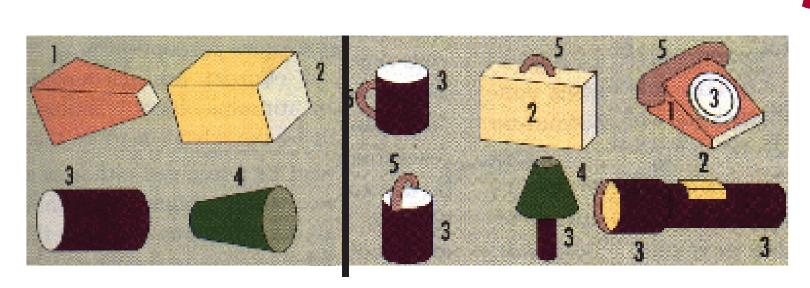
View-based model?
3D view-invariant model?
Mental rotation?

Dalal and Triggs, 2005.

4. Geon and RBC: recognition by geometric primitive components

Recognition by components: a theory of human interpretation. Biederman, 1987.

5. 3D reconstruction is not just low-level

Given a depth map, we still cannot manipulate an object. Because there is no high level representation of the 3D world. Obtaining a good depth map == Inventing a digital camera.
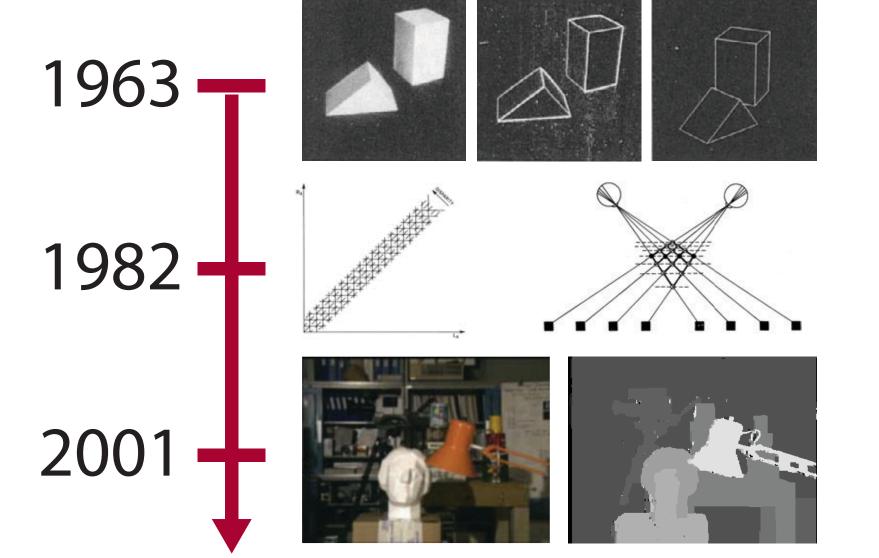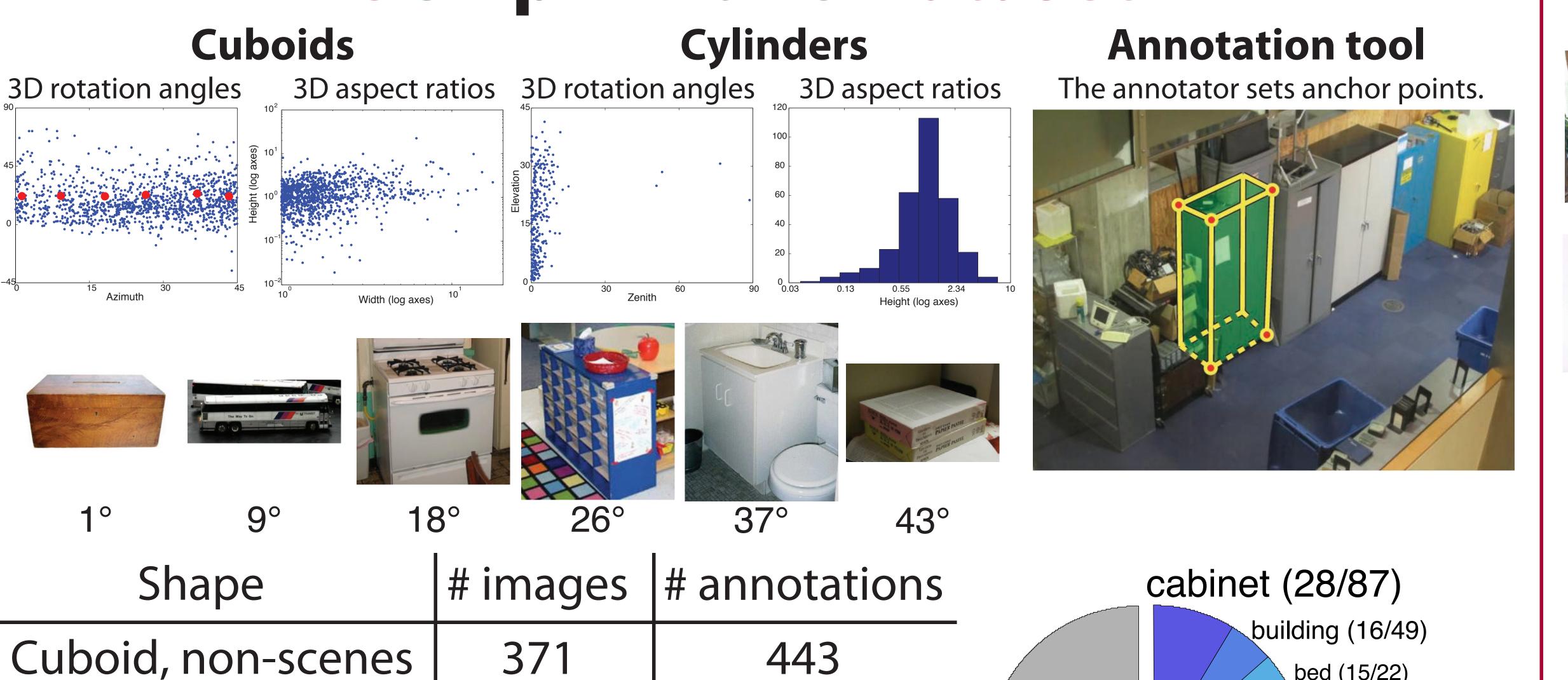
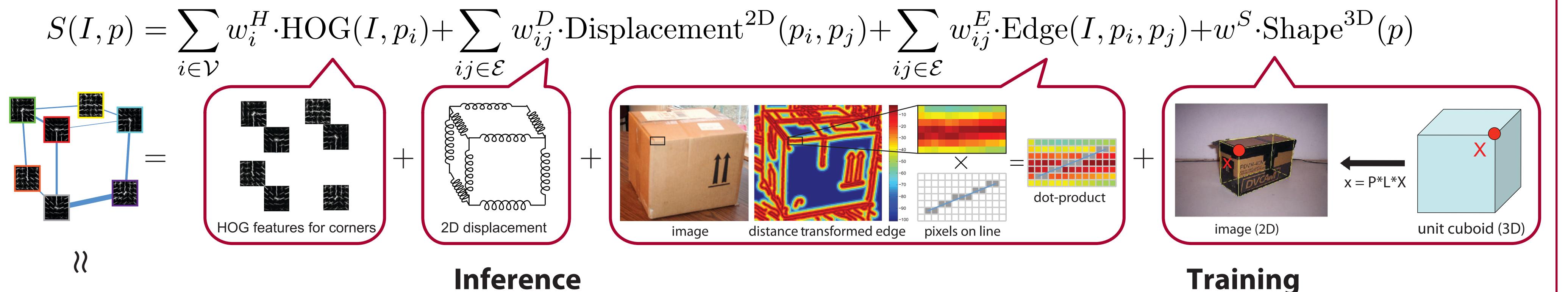|	
1963	Machine perception of three-dimensional solids. Roberts, 1963.
1982	Stereo correspondence. Dev 1974; Marr and Poggio 1976; Marr, 1982.
2001	Dense stereo matching. Boykov, Veksler and Zabih, 2001.

Task

Given a single-view input image, our goal is to detect the 2D corner locations of the cuboids depicted in the image. With the output part locations we can subsequently recover information about the camera and 3D shape via camera resectioning.

Input Image — 3D Cuboid Detector — Output Detection Result — detect — Synthesized New Views — Corner Localization

Model

Scoring function for corner locations in image:

$$S(I,p) = \sum_{i \in \mathcal{V}} w_i^H \cdot \text{HOG}(I, p_i) + \sum_{ij \in \mathcal{E}} w_{ij}^D \cdot \text{Displacement}^{2D}(p_i, p_j) + \sum_{ij \in \mathcal{E}} w_{ij}^E \cdot \text{Edge}(I, p_i, p_j) + w^S \cdot \text{Shape}^{3D}(p)$$

HOG features for corners + 2D displacement + image distance transformed edge × pixels on line = dot-product + image (2D) unit cuboid (3D) x = P*L*X

Inference
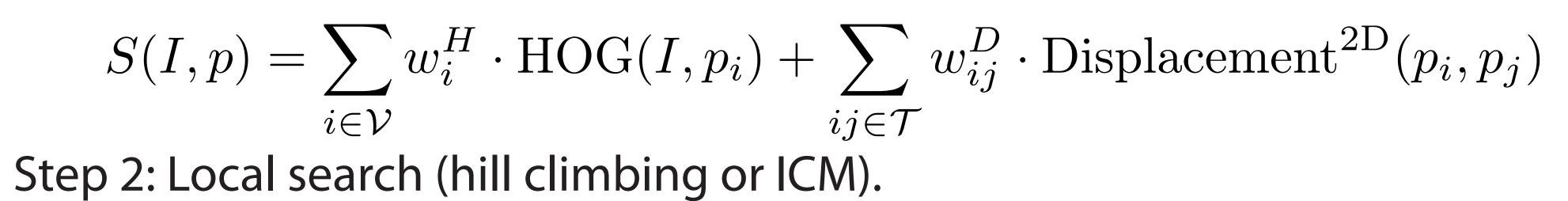
Step 1: Approximation by a tree (dynamic programming + distance transform).

$$S(I,p) = \sum_{i \in \mathcal{V}} w_i^H \cdot \text{HOG}(I, p_i) + \sum_{ij \in \mathcal{T}} w_{ij}^D \cdot \text{Displacement}^{2D}(p_i, p_j)$$

Step 2: Local search (hill climbing or ICM).

Training

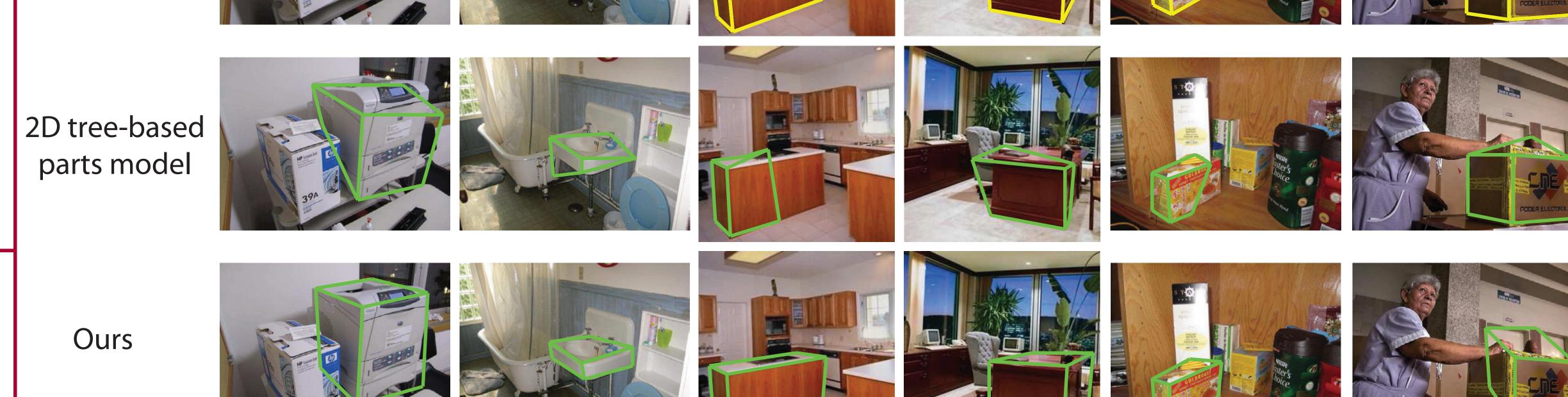Supervised corner-location training with Structural SVM.

$$\min_{\beta, \xi \geq 0} \frac{1}{2} \beta \cdot \beta + C \sum_n \xi_n$$
$$\forall n \in \text{pos} \quad \beta \cdot \Phi(I_n, p_n) \geq 1 - \xi_n$$
$$\forall n \in \text{neg}, \forall p \in P \quad \beta \cdot \Phi(I_n, p) \leq -1 + \xi_n$$

SUNprimitive Dataset

Cuboids
3D rotation angles 3D aspect ratios

Cylinders
3D rotation angles 3D aspect ratios

Annotation tool
The annotator sets anchor points.

1° 9° 18° 26° 37° 43°

Shape	# images	# annotations
Cuboid, non-scenes	371	443
Cuboid, scenes	414	826
Cylinder	156	303
Pyramid	407	407

cabinet (28/87)
building (16/49)
bed (15/22)
night table (15/29)
chest of drawers (10/10)
box (9/18)
desk (8/22)
table (7/8)
island (7/11)
shelves (5/22)
cabinets (5/25)
CPU (7/9)
bench (5/5)
kitchen island (5/9)
wardrobe (5/12)
night table occluded (5/12)
refrigerator (5/9)
stove (5/6)
screen (5/6)
others 97 categories (168/883)

Results

Cuboid

Cylinder and Pyramid

Evaluation

Comparison with 2D parts model

Ground truth

2D tree-based parts model

Ours

Synthesized new views from detections

Detect on Image

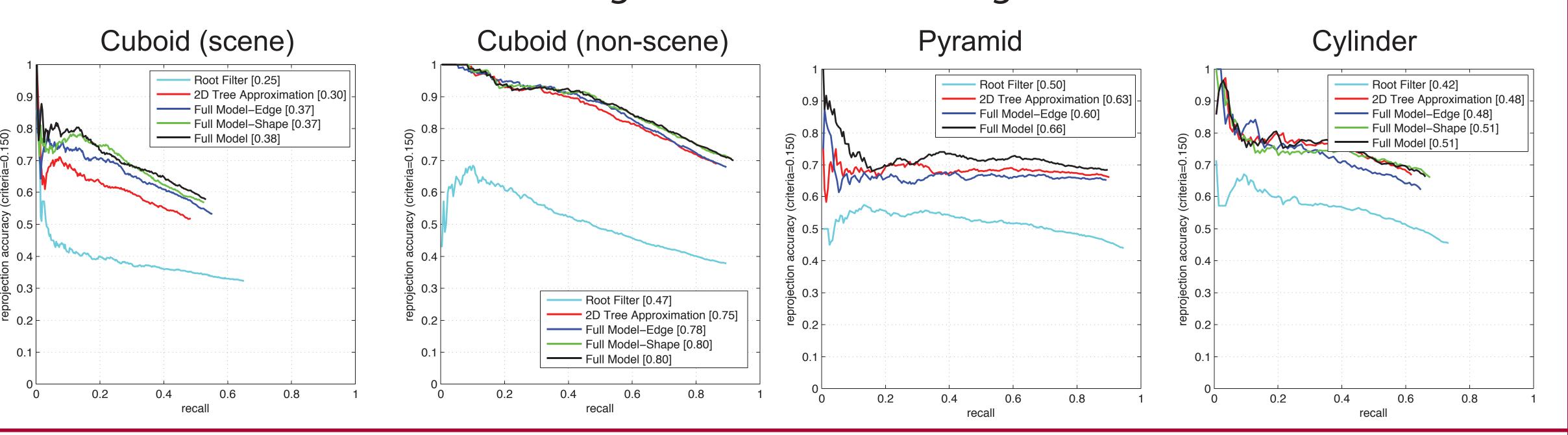Synthesize New Views

Quantitative evaluation: corner localization

- Corner location is correct if it lies within threshold.
- Threshold is set to a fraction of the ground truth bounding box.

Cuboid (scene) Cuboid (non-scene) Pyramid Cylinder

References

[1] Articulated pose estimation using flexible mixtures of parts. Yang and Ramanan. In CVPR, 2011.
[2] Recognizing scene viewpoint using panoramic place representation. Xiao, Ehinger, Oliva and Torralba. In CVPR, 2012.
[3] Reconstructing the world's museums. Xiao and Furukawa. In ECCV, 2012.
[4] Basic level scene understanding: From labels to structure and beyond. Xiao, Russell, Hays, Ehinger, Oliva, and Torralba. In SIGGRAPH Asia, 2012.