Holistic Scene Understanding for 3D Object Detection with RGB-D cameras

Dahua Lin, Sanja Fidler, Raquel Urtasun

TTI Chicago

3D object detection

- Goal: Category-level 3D object detection

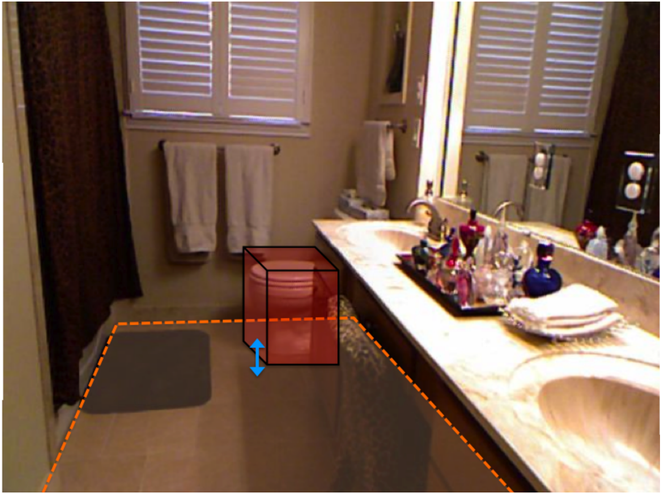
3D object detection

- Goal: Category-level 3D object detection

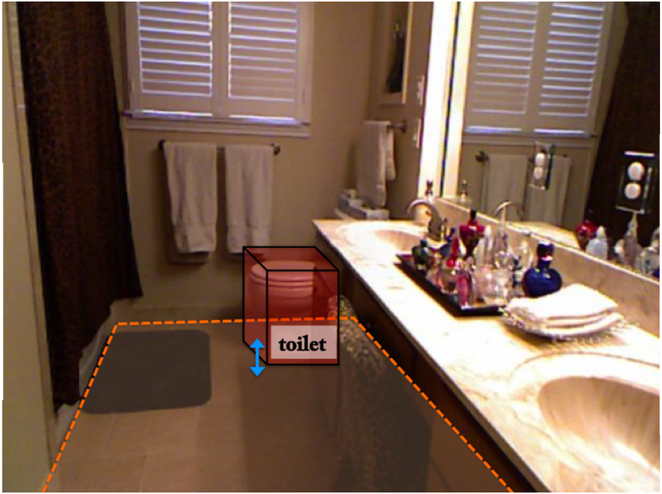
maybe bathroom, maybe kitchen

3D object detection

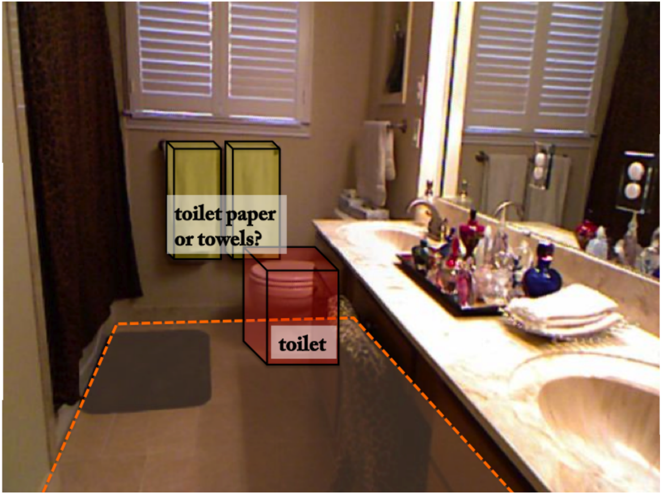
- Goal: Category-level 3D object detection



- Goal: Category-level 3D object detection

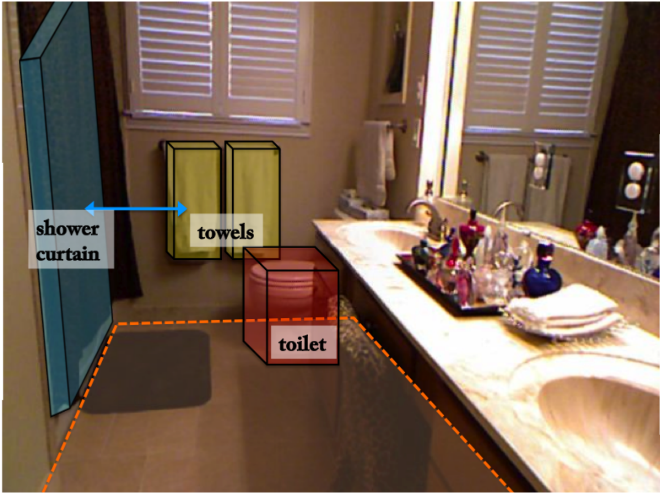


- Goal: Category-level 3D object detection



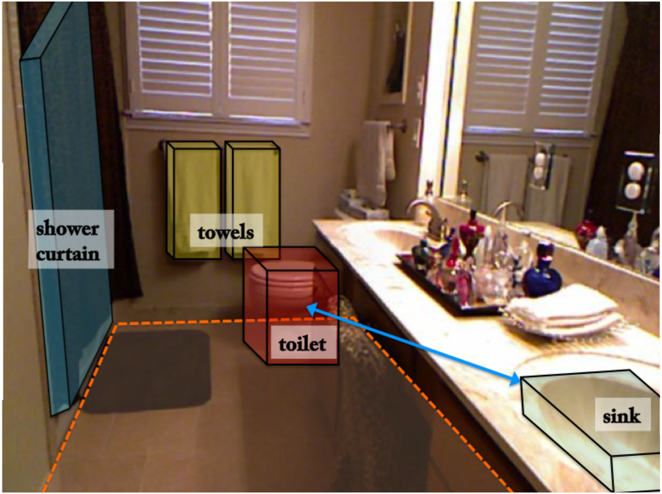
3D object detection

- Goal: Category-level 3D object detection



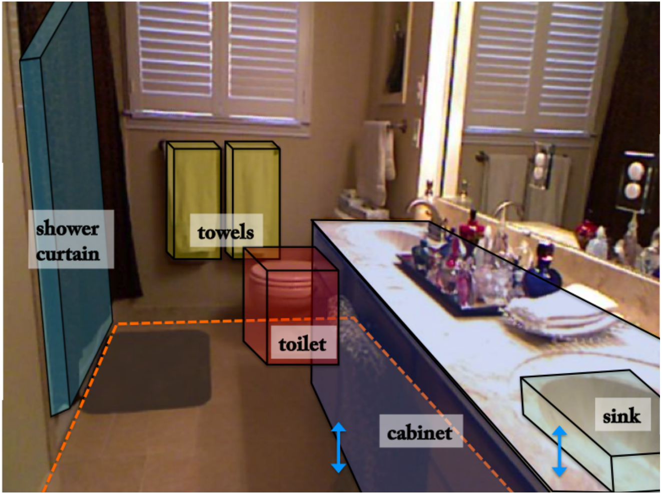
3D object detection

- Goal: Category-level 3D object detection



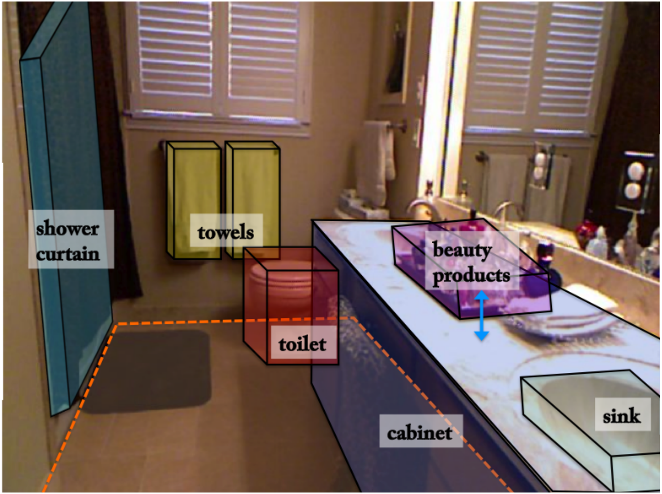
3D object detection

- Goal: Category-level 3D object detection



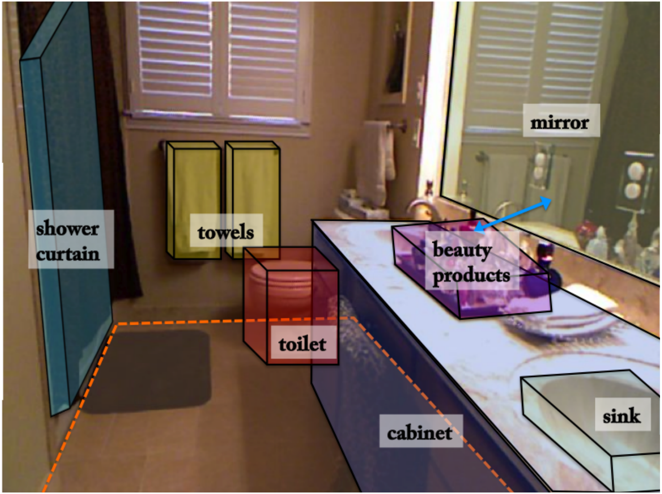
3D object detection

- Goal: Category-level 3D object detection



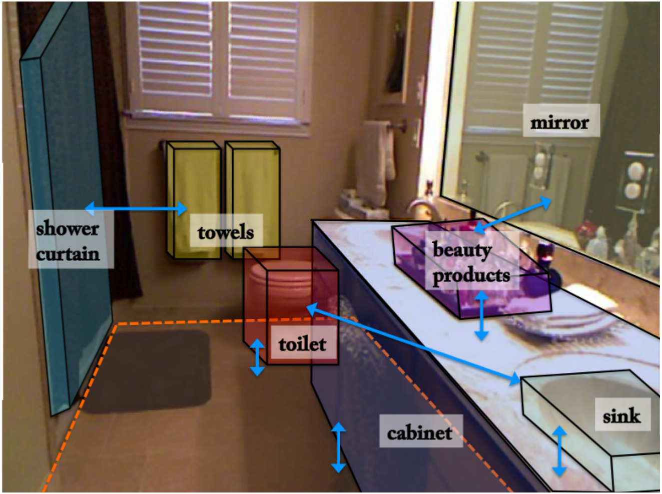
3D object detection

- Goal: Category-level 3D object detection



3D object detection

- Goal: Category-level 3D object detection

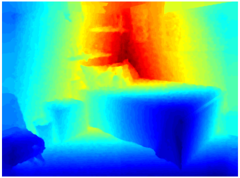


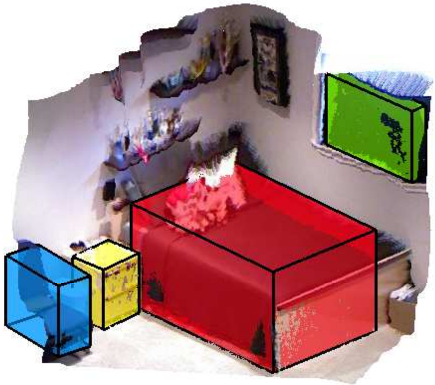
3D object detection in RGB-D images

- Exploit **RGBD imagery** for **category-level 3D object detection**
- **Holistic approach**: jointly reason about **scene**, **objects**, and **contextual relations**

image

depth

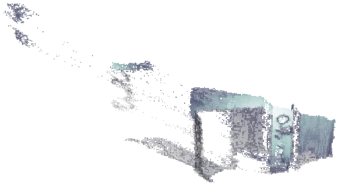




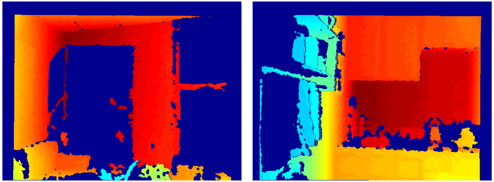
point cloud with **cuboids around objects**

Difficult problem?

Noisy depth

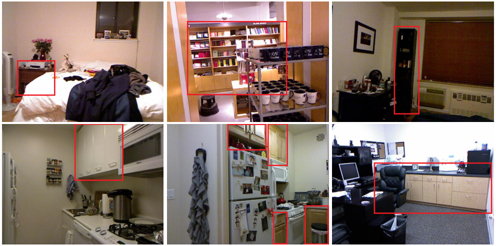


Missing depth



Occlusion

Viewpoint, aspect-ratio variation



Related Work

Holistic models

- Objects, layout: Lee'10 [16], Hedau'10 & '12 [10, 11], Schwing'13 [22]
- Blocks: Gupta'10 [7]

Monocular 3D detection

- Viewpoint: Pepik'12 [19], Sun'10 [25], Liebelt'10 [17]
- Cuboids/polyhedra: Brooks'83 [1], Hedau'10 [10], Lee'10 [16], Fidler'12 [5], Xiang'12 [27]

RGB-D segmentation

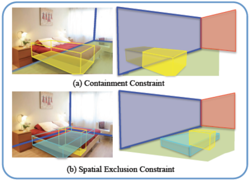
- Koppula'11 [14], Silberman'12 [24], Gupta'13 [8]

RGB-D detection

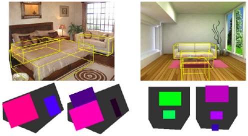
- 2D detector + depth: Gould'08 [6], Walk'10 [26], Saenko'11 [21], Lai'11 [15]

Cuboid generation (no class)

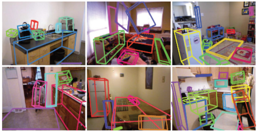
- Jiang'13 [13], Jia'13 [12]



Lee et al., 2010



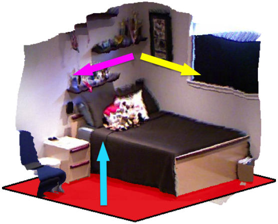
Hedau et al., 2010



Jiang & Xiao, 2013

Overview

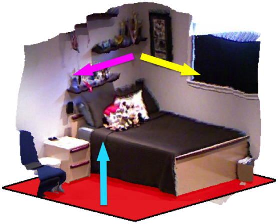
- Rotate the point-cloud to canonical orientation
- Estimate the floor and wall planes



canonical orientation

Overview

- Rotate the point-cloud to canonical orientation
- Estimate the floor and wall planes
- Generate candidate cuboids
- A holistic CRF reasoning about scene and objects, their geometric properties and spatial/semantic relations

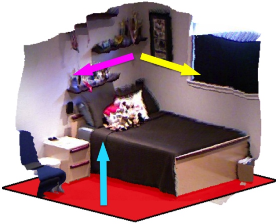


canonical orientation

estimated walls

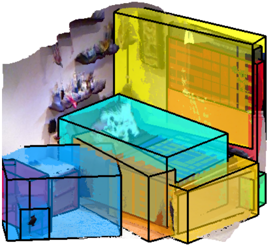
Overview

- Rotate the point-cloud to canonical orientation
- Estimate the floor and wall planes
- Generate candidate cuboids
- A holistic CRF reasoning about scene and objects, their geometric properties and spatial/semantic relations



canonical orientation

estimated walls



top 15 candidates

Cuboid Candidates

- Get candidate "objectness" regions with CPMC [Carreira et al., PAMI 2012 [3]] which we extend to 3D
- Take top K candidates ranked by objectness score
- Project each region to 3D

example regions

Cuboid Candidates

- Get candidate "objectness" regions with CPMC [Carreira et al., PAMI 2012 [3]] which we extend to 3D
- Take top K candidates ranked by objectness score
- Project each region to 3D
- Fit a minimal cube that contains 95% of the 3D points
- Enforce the gravity vector of each cube to be orthogonal to the floor

example regions

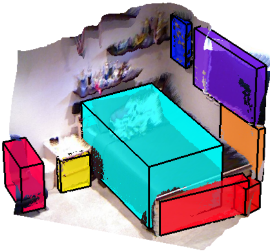
regions in 3D

Cuboid Candidates

- Get candidate "objectness" regions with CPMC [Carreira et al., PAMI 2012 [3]] which we extend to 3D
- Take top K candidates ranked by objectness score
- Project each region to 3D
- Fit a minimal cube that contains 95% of the 3D points
- Enforce the gravity vector of each cube to be orthogonal to the floor

example regions

regions in 3D



fit cuboids

Holistic 3D Scene Model

$$p(\mathbf{y}, s) \propto \exp\left(\mathbf{w}_s^T \phi_s(s) + \mathbf{w}_y^T \sum_{i=1}^{K} \phi_y(y_i) + \mathbf{w}_{yy}^T \sum_{(i,j)} \phi_{yy}(y_i, y_j) + \mathbf{w}_{sy}^T \sum_{i=1}^{K} \phi_{sy}(s, y_i) \right)$$

cuboid class:
$y_i \in \{0, \ldots, C\}$

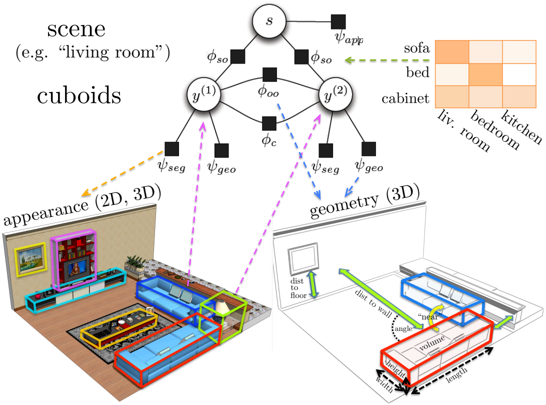
scene class:
$s \in \{1, \ldots, S\}$

Unary:

- appearance
- geometry

Pairwise:

- spatial relations
- semantic relations



Unary potentials

- **Scene appearance:** Classifier on RGB-D features
- **Ranking potential:** Predicts amount of overlap of object candidate with ground-truth [CPMC-o2p, Carreira et al., 2012 [2]]

RGB-D features:

- RGB: gradient, color, LBP, self-similarity, SIFT
- Depth: depth gradient, spin/surface normal

Unary potentials

- **Scene appearance:** Classifier on RGB-D features
- **Ranking potential:** Predicts amount of overlap of object candidate with ground-truth [CPMC-o2p, Carreira et al., 2012 [2]]
- **Segmentation potential:** Classifier on superpixels using RGB-D kernel descriptors [Ren et al., 2012 [20]]

RGB-D features:

- RGB: gradient, color, LBP, self-similarity, SIFT
- Depth: depth gradient, spin/surface normal

Unary potentials

- **Scene appearance:** Classifier on RGB-D features
- **Ranking potential:** Predicts amount of overlap of object candidate with ground-truth [CPMC-o2p, Carreira et al., 2012 [2]]
- **Segmentation potential:** Classifier on superpixels using RGB-D kernel descriptors [Ren et al., 2012 [20]]
- **Object geometry:** Classifier on geometric features

RGB-D features:

- RGB: gradient, color, LBP, self-similarity, SIFT
- Depth: depth gradient, spin/surface normal

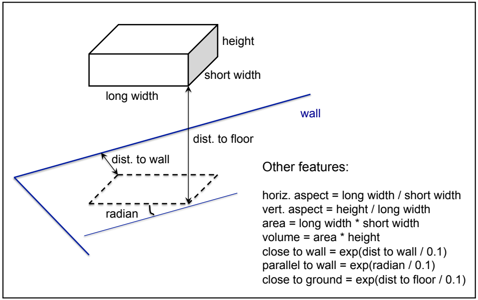
Unary potentials

- **Scene appearance:** Classifier on RGB-D features
- **Ranking potential:** Predicts amount of overlap of object candidate with ground-truth [CPMC-o2p, Carreira et al., 2012 [2]]
- **Segmentation potential:** Classifier on superpixels using RGB-D kernel descriptors [Ren et al., 2012 [20]]
- **Object geometry:** Classifier on geometric features

RGB-D features:

- RGB: gradient, color, LBP, self-similarity, SIFT
- Depth: depth gradient, spin/surface normal

Geometry features:



Pairwise potentials

Semantic context:

- **scene-object potential**:

$$\phi_{sy}(s = k, y = l) = \text{scene-object co-occurrence stats}$$

- **object-object potential**

$$\phi_{yy}(y = l, y' = l') = \text{object-object co-occurrence stats}$$

Geometric relations:

- **close-to**: Two objects are *close to* each other if their distance is less than 0.5 meters.
- **on-top-of**: Object A is *on top of* B if A is higher than B and (at least) 80% of A's bottom face is contained within the top face of B.

Learning and Inference

- **Loss:** how far from GT is each hypothesis
  - Object: 0/1 loss based on IOU with GT
  - Scene: 0/1 loss
- **Learning:** Primal dual method blending learning and inference [Hazan and Urtasun, NIPS 2010 [9]]
- **Inference:** Distributed message passing [Schwing et al., CVPR 2011 [23]]
- **Timings**:
  - **learning** takes **2 minutes** ($\sim$ 800 images)
  - **inference** takes **15 ms per image** (15 cuboids per image)

On Intel i7 quad-core CPU (4 threads)

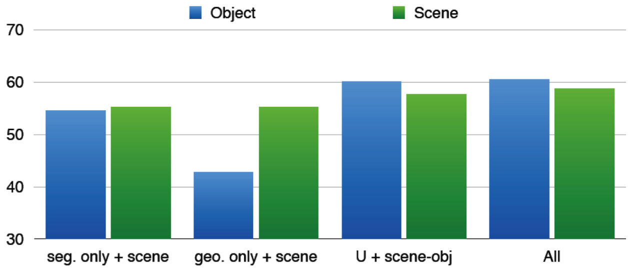
Experimental Results

- NYUv2 [Silberman et al, 2012]: 1449 scenes, 6680 objects, 21 object classes + background
- Ground truth: Fit 3D cuboids around GT regions and correct bad fits
- Standard split: 60% of images used for training and 40% for test

Results on GT cuboids

- Performance of scene measured in classification accuracy
- Performance evaluated on GT cuboids, measured as classification accuracy

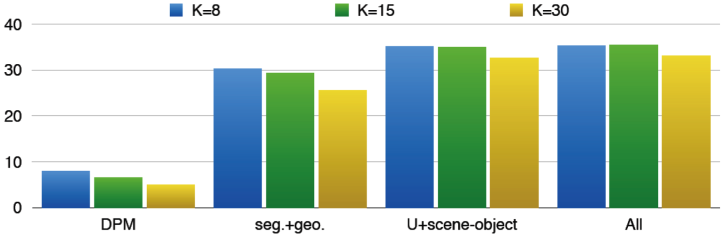
configuration	object	scene
scene appearance only	-	55.20
segmentation only	54.46	-
geometry only	42.85	-
all unaries	59.02	55.20
unaries + scene-obj	60.00	57.65
full model	**60.49**	**58.72**



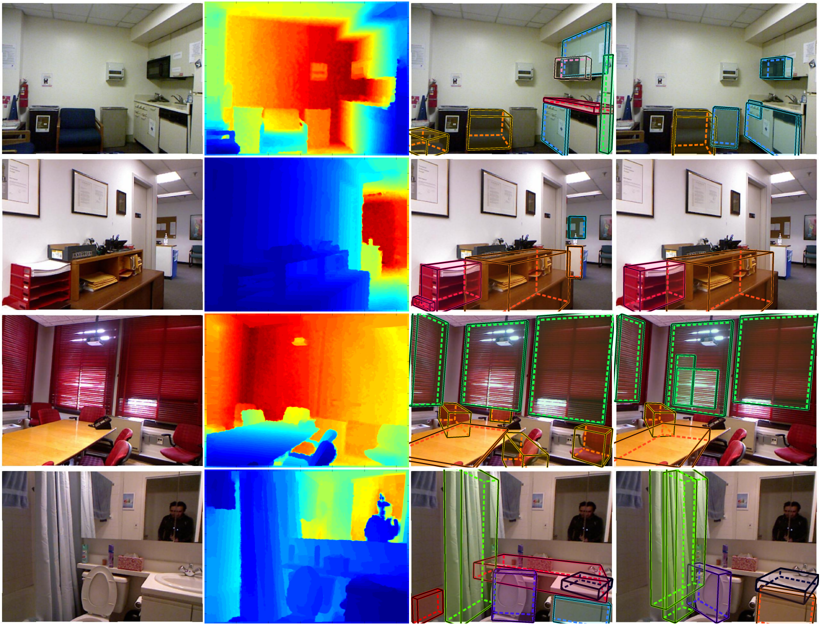
Our Full Detection Pipeline

- Performance measured as average of per-class F-measures
- DPM: [Felzenswalb et al., TPAMI, 2010 [4]]
- Jiang'13: Cuboids from [H. Jiang and J. Xiao, CVPR, 2013 [13]]

	DPM	seg.	seg.+geo.	all unaries	+scene-object	full model
[Jiang'13]	-	11.11	21.13	21.90	22.19	**22.3**
K = 8	8.01	28.98	30.22	35.17	35.18	**35.23**
K = 15	6.54	28.33	29.44	34.92	34.95	**35.56**
K = 30	4.96	24.81	25.58	32.54	32.57	**33.10**



Example detections



Summary and Conclusion

Summary and Conclusion:

- A new 3D holistic model that reasons about the scene and objects of multiple classes in indoor RGB-D scenes
- Experiments demonstrated that our approach significantly outperforms state-of-the-art detectors

Future work:

- Segmentation, 3D detection, support
- Apartment model: large 3D space & video, lots of objects & classes

Code and data available here:

`http://www.cs.utoronto.ca/~fidler/projects/scenes3D.html`

Full paper [18]:

`http://www.cs.utoronto.ca/~fidler/papers/lin_et_al_iccv13.pdf`

Bibliography I

[1] Rodney A. Brooks. Model-based three-dimensional interpretations of two-dimensional images. *PAMI*, 5:140–150, 1983.

[2] J. Carreira, R. Caseiroa, J. Batista, and C. Sminchisescu. Semantic segmentation with second-order pooling. In *ECCV*, 2012.

[3] J. Carreira and C. Sminchisescu. Cpmc: Automatic object segmentation using constrained parametric min-cuts. *TPAMI*, 2012.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9), 2010.

[5] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS*, 2012.

[6] Stephen Gould, Paul Baumstarck, Morgan Quigley, Andrew Y. Ng, and Daphne Koller. Integrating visual and range data for robotic object detection. In *ECCV w. on S. Fusion Alg. & Appl.*, 2008.

[7] Abhinav Gupta, Alexei A. Efros, and Martial Hebert. Blocks world revisited: Image understanding using qualitative geometry and mechanics. In *ECCV*, 2010.

[8] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from rgbd images. In *CVPR*, 2013.

[9] T. Hazan and R. Urtasun. A primal-dual message-passing algorithm for approximated large scale structured prediction. In *NIPS*, 2010.

Bibliography II

[10] V. Hedau, D. Hoiem, and D.A. Forsyth. Thinking inside the box: Using appearance models and context based on room geometry. In *ECCV*, 2010.

[11] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering free space of indoor scenes from a single image. In *CVPR*, 2012.

[12] Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 3d-based reasoning with blocks, support, and stability. In *CVPR*, 2013.

[13] H. Jiang and J. Xiao. A linear approach to matching cuboids in rgbd images. In *CVPR*, 2013.

[14] Hema Koppula, Abhishek Anand, Thorsten Joachims, and Ashutosh Saxena. Semantic labeling of 3d point clouds for indoor scenes. In *NIPS*, 2011.

[15] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *ICRA*, 2011.

[16] David C. Lee, Abhinav Gupta, Martial Hebert, and Takeo Kanade. Estimating spatial layout of rooms using volumetric reasoning about objects and surfaces. In *NIPS*, 2010.

[17] Jorg Liebelt and Cordelia Schmid. Multi-view object class detection with a 3d geometric model. In *CVPR*, pages 1688–1695, 2010.

[18] Dahua Lin, Sanja Fidler, and Raquel Urtasun. Holistic scene understanding for 3d object detection with rgbd cameras. In *ICCV*, 2013.

Bibliography III

[19] Bojan Pepik, Michael Stark, Peter Gehler, and Bernt Schiele. Teaching 3d geometry to deformable part models. In Serge Belongie, Andrew Blake, Jiebo Luo, and Alan Yuille, editors, *CVPR*, 2012.

[20] Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. In *CVPR*, 2012.

[21] K. Saenko, Y. Jia, M. Fritz, J. Long, A. Janoch, A. Shyr, S. Karayev, and T. Darrell. Practical 3-d object detection using category and instance-level appearance models. In *IROS*, 2011.

[22] A. Schwing, S. Fidler, M. Pollefeys, and R. Urtasun. Box in the box: Joint 3d layout and object reasoning from single images. In *ICCV*, 2013.

[23] A. Schwing, T. Hazan, M. Pollefeys, and R. Urtasun. Distributed message passing for large scale graphical models. In *CVPR*, 2011.

[24] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

[25] Min Sun, Hao Su, Silvio Savarese, and Li Fei-Fei. A multi-view probabilistic model for 3d object classes. In *CVPR*, 2009.

[26] S. Walk, K. Schindler, and B. Schiele. Disparity statistics for pedestrian detection: Combining appearance, motion and stereo. In *ECCV*, 2010.

[27] Yu Xiang and Silvio Savarese. Estimating the aspect layout of object categories. In *CVPR*, 2012.