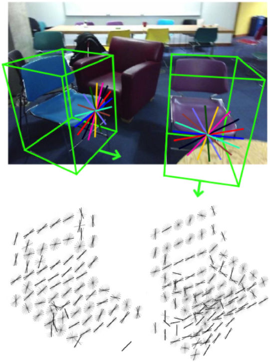
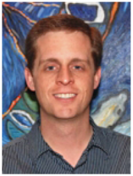
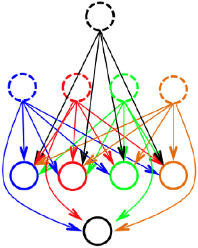
3D Object Detection and Layout Prediction using Clouds of Oriented Gradients







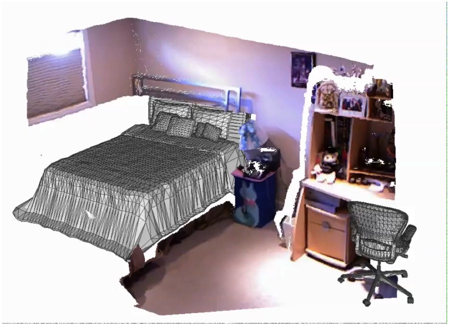
Zhile Ren Erik Sudderth

Brown University

June 28, 2016

CVPR 2016

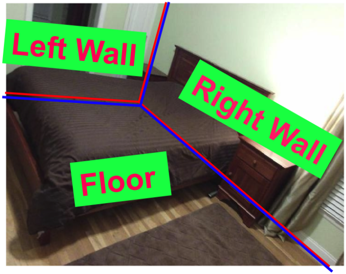
Modeling the 3D World

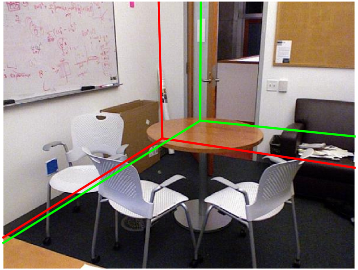


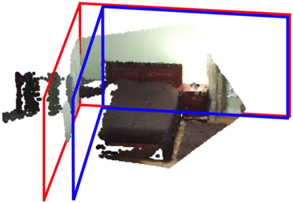
SUN RGB-D Dataset (Song et al., 2015)

- 10335 indoor images with (noisy) depths and camera parameters.
- Annotations: 3D **cuboids with orientations** for indoor objects (bed, sofa, chair, etc.) and 3D **room layout** (walls, floor, & ceiling).
- Baseline: Object detection using CAD models (Song et al., 2014).

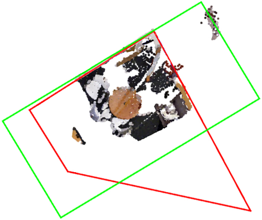
Room Layout Prediction







Small change in 2D lead to huge error in 3D (Schwing et al., 2012)



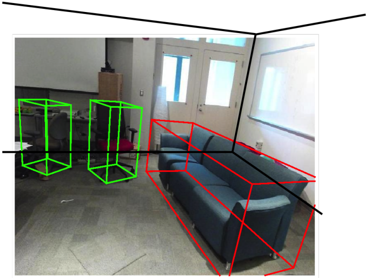
Sensor errors can mislead simple layout prediction heuristics (Song et al., 2015)

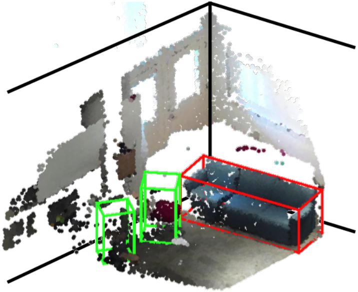
Our Goals & Contributions

Goal: RGB-D $\rightarrow$ 3D cuboid object detection + layout prediction.

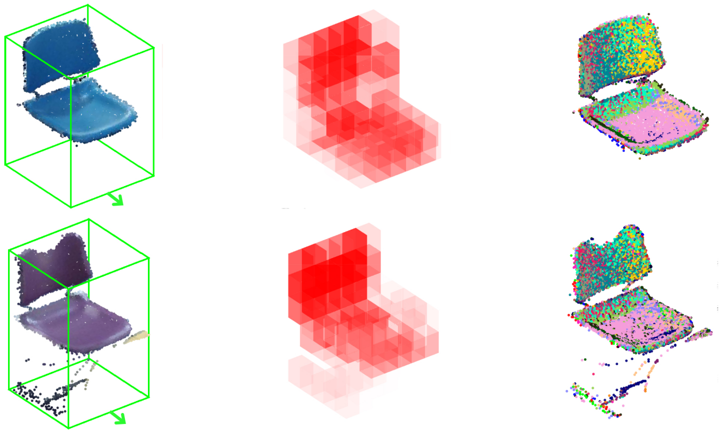
Our contributions:

- **Cloud of Oriented Gradient (COG)** descriptor to model object categories from general 3D viewpoints.
- **Manhattan Voxel** to model 3D room layout
- **Cascaded Detection** to model contextual relationships between objects and scenes.



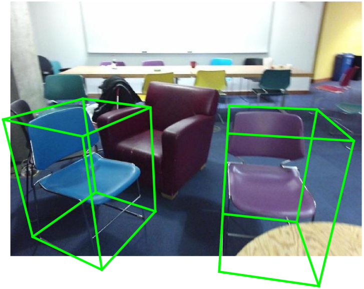


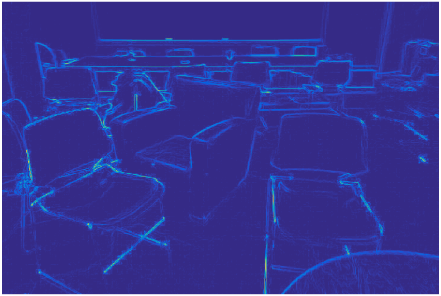
Geometric Features for 3D Cuboids



- Discretize into $6 \times 6 \times 6$ grid of (large) voxels
- **Point Cloud Density**: Fraction of points in 2D area.
- **3D Normal Distribution**: Estimated from local planar fit, 25 bins.

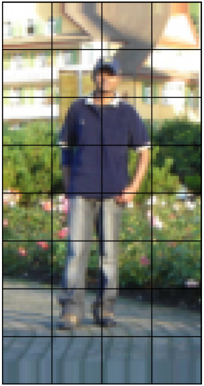
Modeling Object Appearance





Goal: Model 3D object appearance using image gradients.

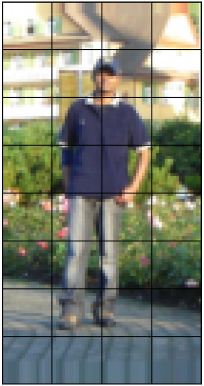
Histograms of Oriented Gradients (HOG) descriptor

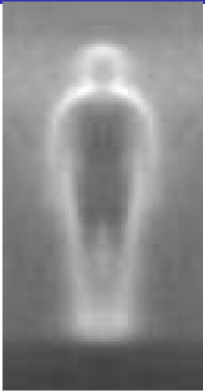


Input Image

[Dalal and Triggs, 2005]

Histograms of Oriented Gradients (HOG) descriptor

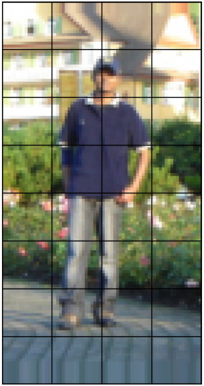


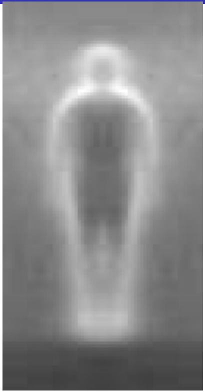


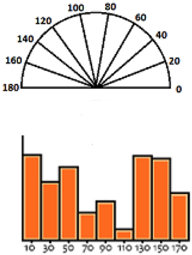
Input Image 2D Gradients

[Dalal and Triggs, 2005]

Histograms of Oriented Gradients (HOG) descriptor



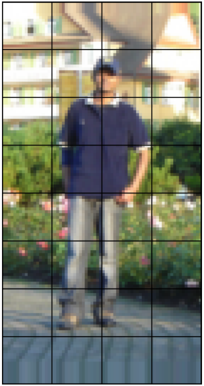


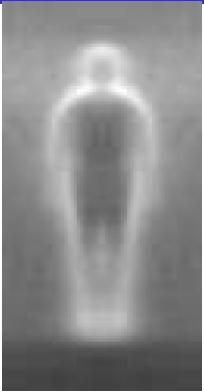


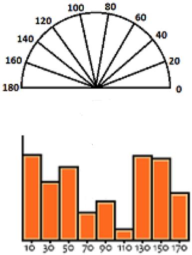
Input Image 2D Gradients Gradient Orientation Bins

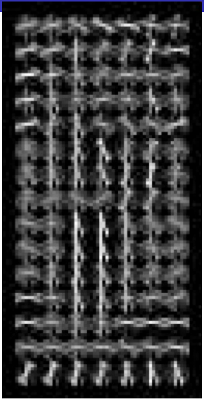
[Dalal and Triggs, 2005]

Histograms of Oriented Gradients (HOG) descriptor





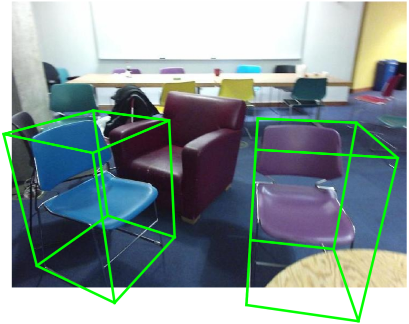




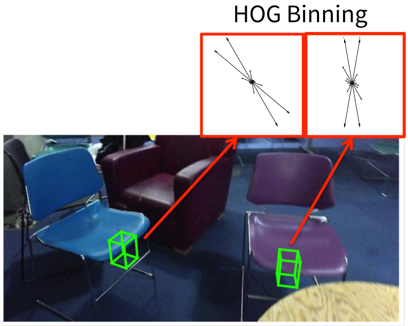
Input Image   2D Gradients   Gradient Orientation Bins   HOG

[Dalal and Triggs, 2005]

Cloud of Oriented Gradient (COG) Feature

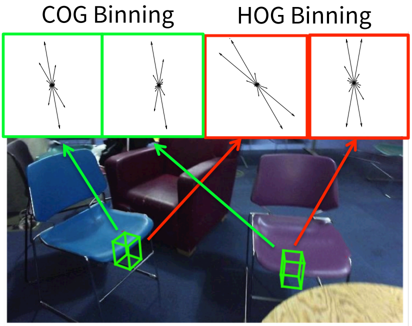


Cloud of Oriented Gradient (COG) Feature



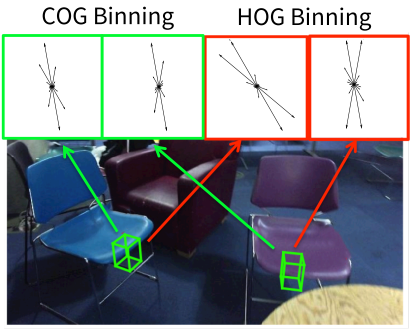
Inconsistent binning of HOG

Cloud of Oriented Gradient (COG) Feature

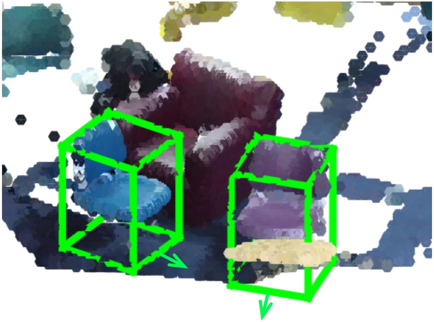


Inconsistent binning of HOG
Consistent binning of COG

Cloud of Oriented Gradient (COG) Feature

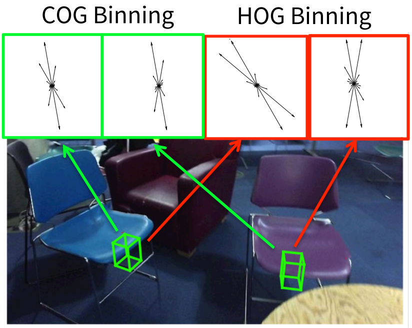


Inconsistent binning of HOG
Consistent binning of COG

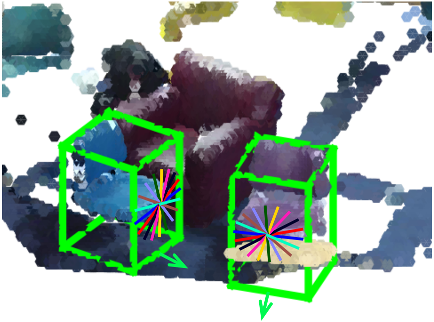


Proposed cuboids in 3D

Cloud of Oriented Gradient (COG) Feature

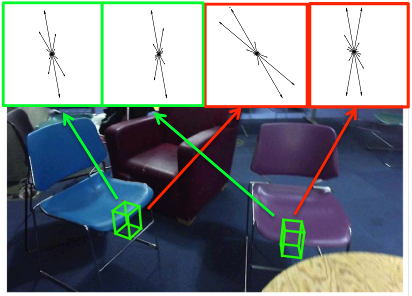


Inconsistent binning of HOG
Consistent binning of COG

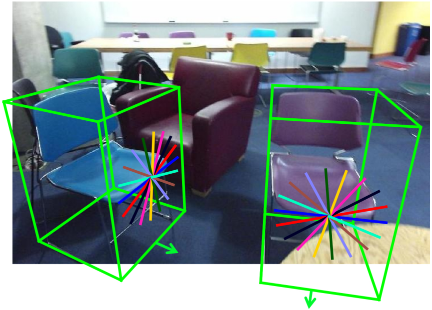


Orientation bins in 3D.
(Same color: Same orientation bin)

Cloud of Oriented Gradient (COG) Feature

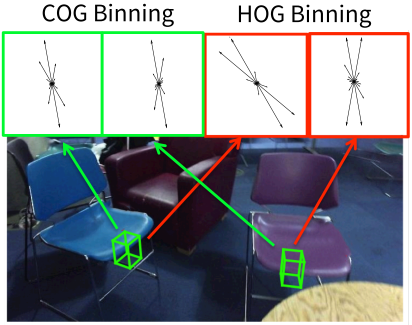


Inconsistent binning of HOG
Consistent binning of COG



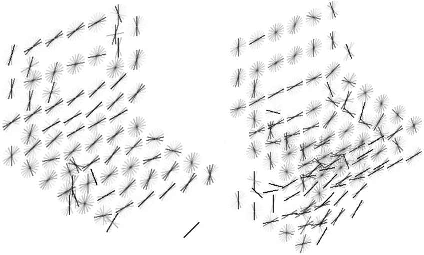
Perspective projection to 2D image.
(Same color: Same orientation bin)

Cloud of Oriented Gradient (COG) Feature



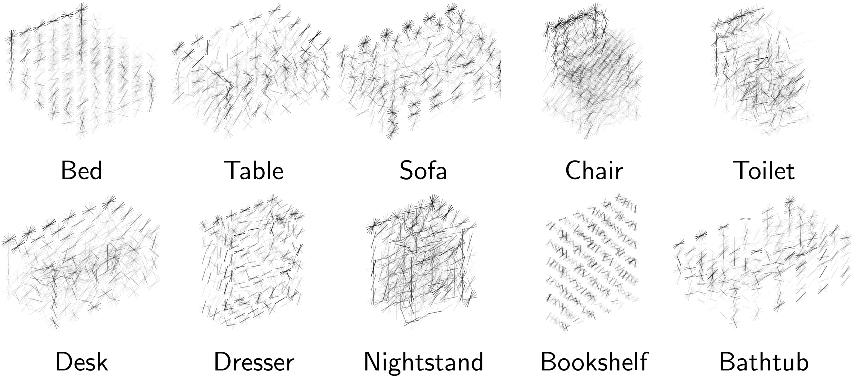
Inconsistent binning of HOG

Consistent binning of COG



Edge binning is stable across viewpoints!

Learned 3D COG Features



Learning 3D Object Detectors

Structural SVM training For each object category

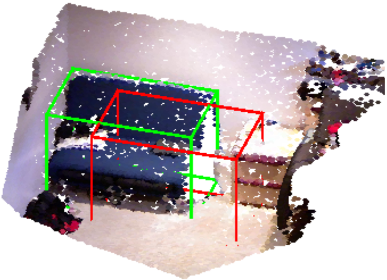
$$I \to B$$

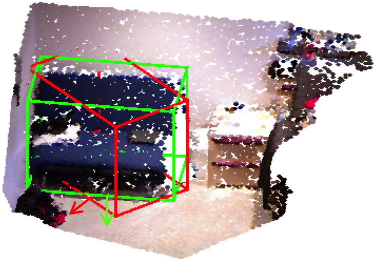
I: RGB-D image. $B = (L, \theta, S)$: 3D bounding box.

- L: location. θ: orientation. S: size.
- Confidence score for each cuboid j of image i:

$$F(I_i, B_j) = w^T \Psi(I_i, B_j)$$

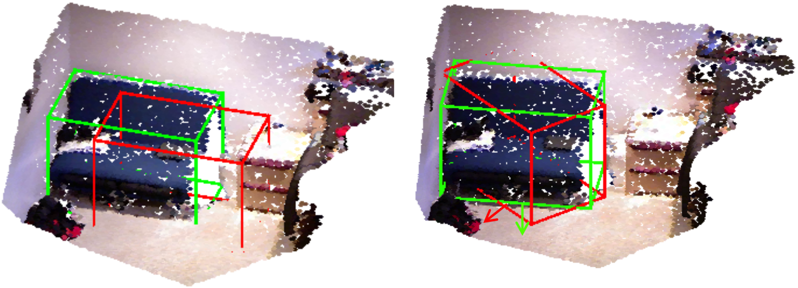
- Testing: Sliding-windows in 3D.





The loss function for each training example i:

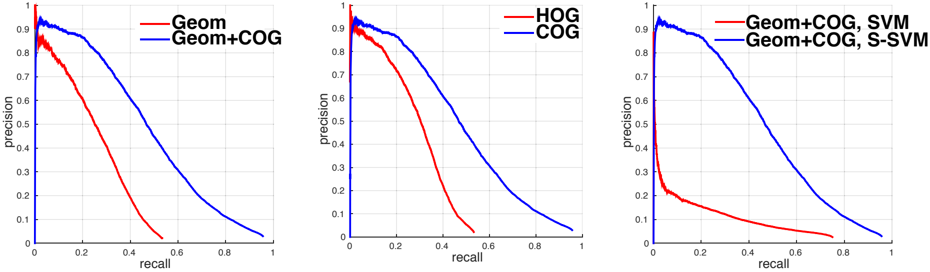
$$\Delta(B_i, \bar{B}_i) = 1 - \underbrace{IOU(B_i, \bar{B}_i)}_{\text{location/size overlap}} \cdot \underbrace{\left(\frac{1 + cos(\bar{\theta}_i - \theta_i)}{2}\right)}_{\text{pose alignment}} \quad (1)$$



Visualize loss score of detection (red) with ground truth (green)

All of our modeling pieces are important

- COG is an informative object appearance feature.
- Perspective-corrected COG bins are more accurate than standard HOG bins.
- Loss-sensitive S-SVM training outperforms binary SVMs with hard-negative mining.



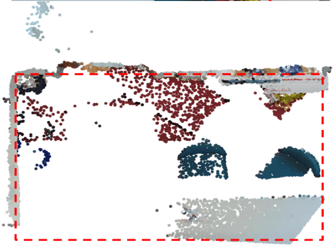
Precision-Recall curve for chair detector.

Room layout: Manhattan Structure.

Goal: RGB-D $\rightarrow$ 3D cuboid object detection + **layout prediction**.

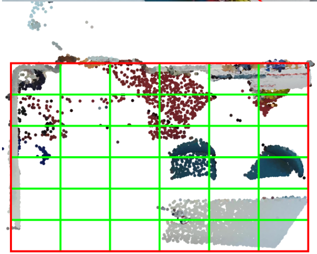
http://www.aliexpress.com/store/product/Series-assembled-model-diy-birthday-gift-for-boys-girls-handmade-gift/1702169_32320724705.html

Traditional Voxel Discretization



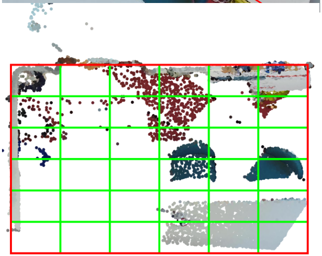
Features for 3D Manhattan Room Layout

Traditional Voxel Discretization



Features for 3D Manhattan Room Layout

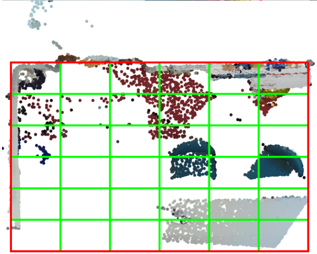
Traditional Voxel Discretization



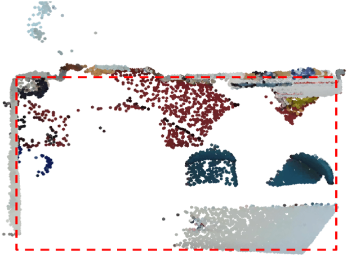
Manhattan Voxel Discretization

Features for 3D Manhattan Room Layout

Traditional Voxel Discretization

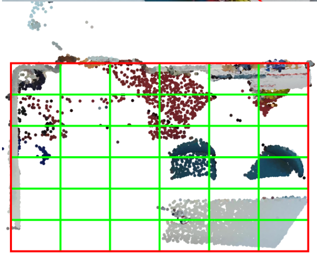


Manhattan Voxel Discretization

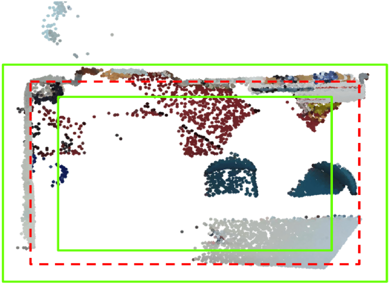


Features for 3D Manhattan Room Layout

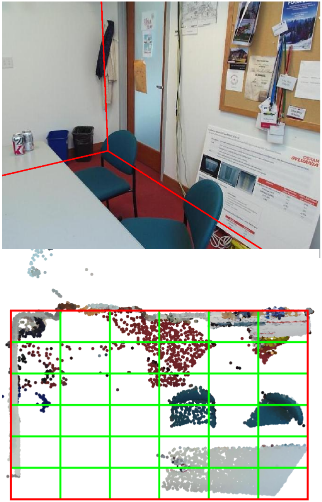
Traditional Voxel Discretization



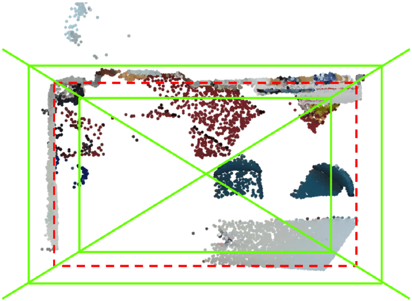
Manhattan Voxel Discretization



Traditional Voxel Discretization

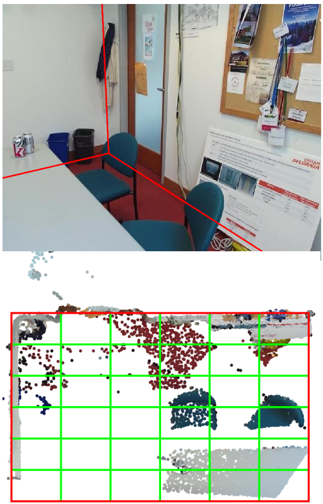


Manhattan Voxel Discretization

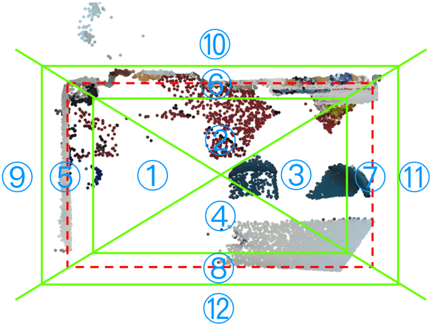


Features for 3D Manhattan Room Layout

Traditional Voxel Discretization



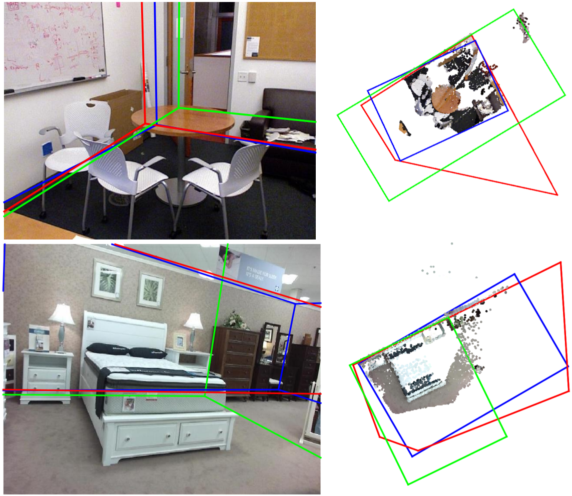
Manhattan Voxel Discretization



Advantages:

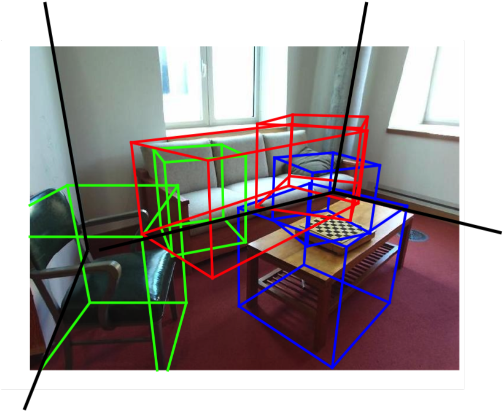
- Models statistics along, inside, and outside the wall.
- Handles the intersection of walls.

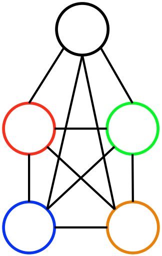
Experiment: Layout Prediction



- **Manhattan voxel** 3D layout predictions (meanIOU=**78.96**)
- **Baseline** in SUN RGB-D (meanIOU=**73.40**)
- **Ground truth** annotations.

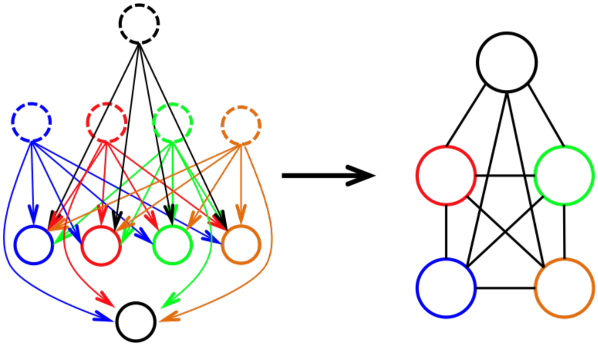
Handling False Positives





- Use simple heuristics: won't work in general.
- MRF with fully connected graph: inference is extremely challenging.
- Colored nodes: Object categories. Black nodes: Room layout.

Context via Cascaded Classifier



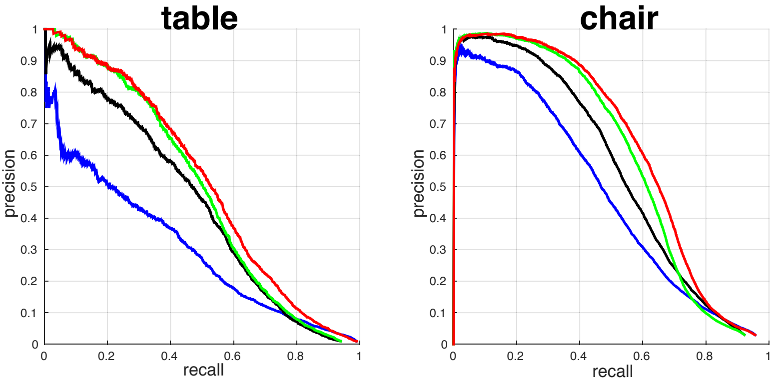
Cascaded classifier (Heitz et al., 2008)

- Model first-stage detector as latent variables in directed graph.
- Marginalizing first-stage variables recovers fully-connected graph.

Context via Cascaded Classifier

- Object-object and object-layout features: 3D relative overlap, confidence difference, and distance & angle to wall.
- Use SVM to train, add new confidence scores to initial scores.

Experiment: Contextual Detection



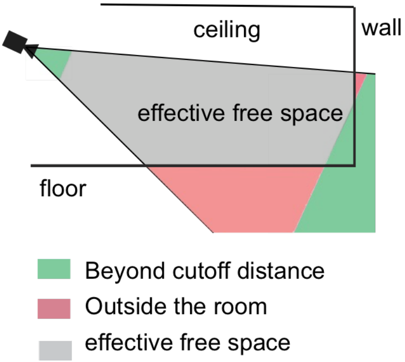
- **Blue**: Geometric & COG features
- **Black**: Add context from 5 categories
- **Green**: Add context from 5 more categories (10 total)
- **Red**: Add context from inferred 3D room layout

Experiment: Average Precision on SUN RGB-D

					
Sliding-Shape (Song et al., 2014)	42.95	19.66	20.60	28.21	60.89
Geom+COG	52.98	28.64	42.16	45.14	43.00
Geom+COG+Context-10	61.29	48.68	49.80	59.03	66.31
Geom+COG+Context-10+Layout	**63.67**	**51.29**	**51.02**	**62.17**	**70.07**

					
Sliding-Shape (Song et al., 2014)	No CAD models				
Geom+COG	28.17	7.93	14.25	12.83	47.69
Geom+COG+Context-10	44.58	12.97	25.14	30.05	56.78
Geom+COG + Context-10 + Layout	**45.19**	**15.47**	**27.36**	**31.80**	**58.26**

Experiment: Total Scene Understanding

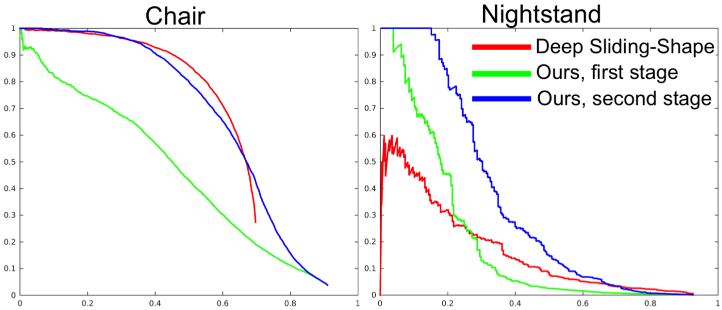


	P_g	R_g	R_r	IoU
Sliding-Shape + Plane-Fitting	37.8	32.3	23.7	66.0
COG + Manhattan Voxel + Context	**47.3**	**36.8**	**35.8**	**72.0**

Geometric Precision/Recall, Recognition Recall, and free-space IOU.

(Song et al., 2015)

Independent Work at CVPR 2016: Deep Sliding-Shape

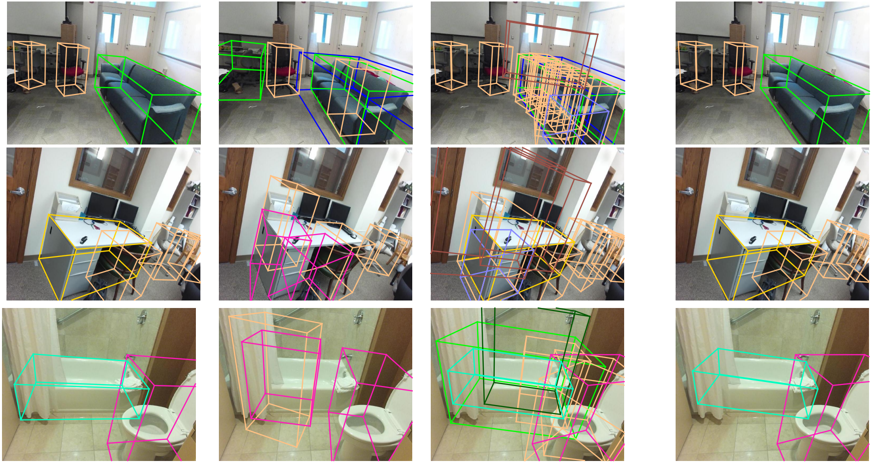


Deep Sliding-Shape (Song et al., 2016)	**0.80**	0.54	0.55	0.61	**0.78**
COG + Context-20 + Extra Cuboid Features	0.75	**0.63**	**0.67**	**0.64**	0.73

Deep Sliding-Shape (Song et al., 2016)	0.23	0.08	0.16	0.15	0.44
COG + Context-20 + Extra Cuboid Features	**0.47**	**0.37**	**0.34**	**0.34**	**0.76**

Results and Conclusions

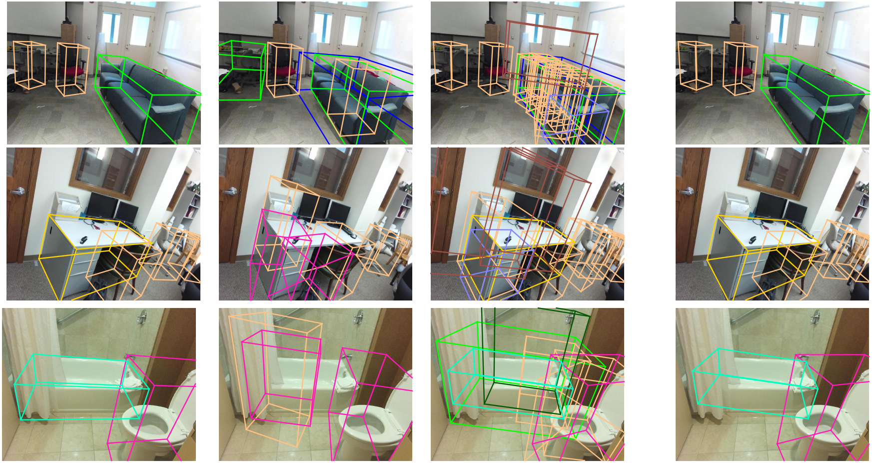
Bed | Table | Sofa | Chair | Toilet | Desk | Dresser | Nightstand | Bookshelf | Bathtub



Ground Truth | Sliding Shape | Geom+COG | Geom+COG+Context-10

Results and Conclusions

Bed | Table | Sofa | Chair | Toilet | Desk | Dresser | Nightstand | Bookshelf | Bathtub



Ground Truth | Sliding Shape | Geom+COG | Geom+COG+Context-10

Thanks!