3D Scene Understanding from a single RGB-D Image

Andreas Geiger
Chaohui Wang

MPI for Intelligent Systems
Tübingen, Germany

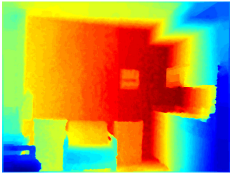
October 09, 2015

Introduction

Task: Infer scene layout and objects from a single RGB-D Image

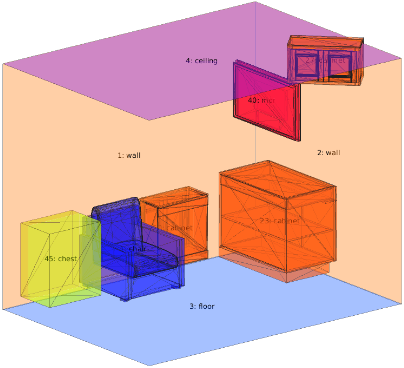
Input

RGB



Depth

Output

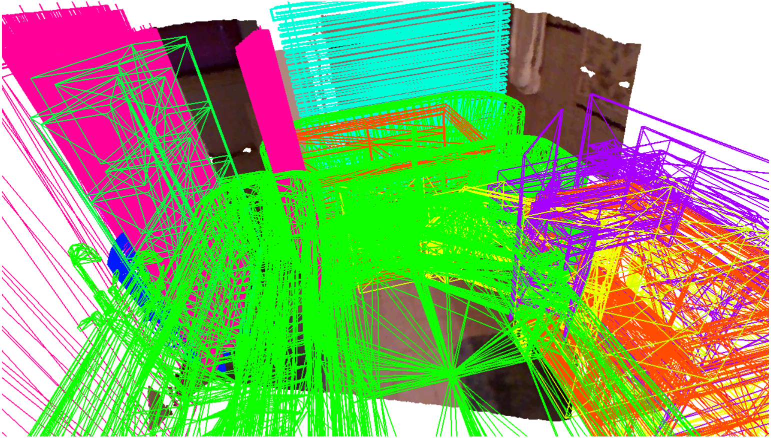


System Overview

Input: RGB-D image

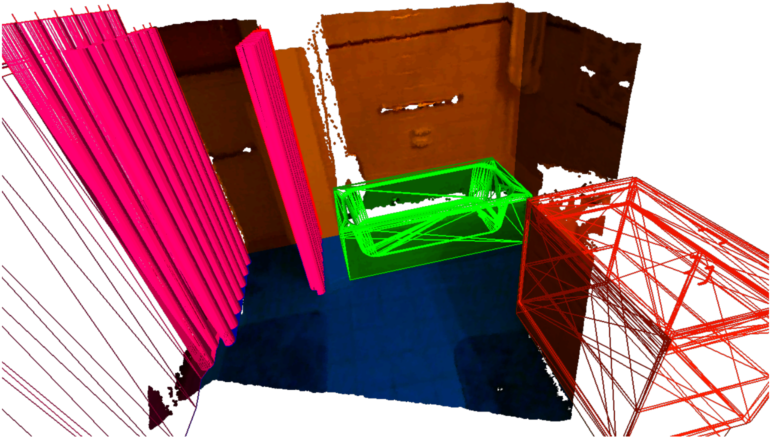
System Overview

Proposals: Layout elements and CAD objects



System Overview

Model: High-order CRF



Model

$$
\begin{aligned}
E(\mathbf{x}|\mathbf{I}) \quad = \quad & \sum_{i \in \mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i \in \mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k \in \mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j \in \mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}} \\
& + \sum_{i \in \mathcal{L}, j \in \mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i \in \mathcal{L} \cup \mathcal{O}, k \in \mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c \in \mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}
\end{aligned}
$$

Notation:

- $\mathcal{L}$: layout elements $\quad$ $\mathcal{O}$: objects $\quad$ $\mathcal{S}$: superpixels$\mathcal{L}$: layout elements $\quad$ $\mathcal{O}$: objects $\quad$ $\mathcal{S}$: superpixels
- $x_i \in \{0, 1\}$: object/layout present?
- $x_k \in \{0, 1\}$: superpixel explained?

Goal:

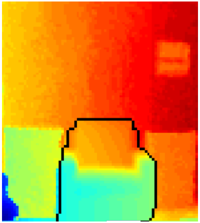
- Given an RGB-D image $\mathbf{I}$, infer presence of each object/layout $x_i \in \{0, 1\}$ while explaining as many superpixels x_k as possible!

Model

$$
\begin{aligned}
E(\mathbf{x}|\mathbf{I}) \;=\; & \sum_{i \in \mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i \in \mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k \in \mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j \in \mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}} \\
& + \sum_{i \in \mathcal{L}, j \in \mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i \in \mathcal{L} \cup \mathcal{O}, k \in \mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c \in \mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}
\end{aligned}
$$

Layout and Object Unary Potentials

Overlap with 2D proposal
[Carreira2012,Gupta2013]



3D geometry
depth/normals

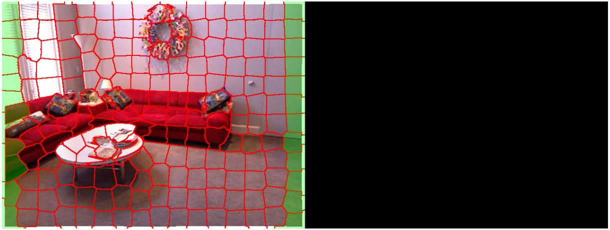
3D object scale
estimated from data

Model

$$E(\mathbf{x}|\mathbf{I}) = \sum_{i\in\mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i\in\mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k\in\mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}}$$

$$+ \sum_{i\in\mathcal{L}, j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i\in\mathcal{L}\cup\mathcal{O}, k\in\mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c\in\mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}$$

Superpixel Unary Potential

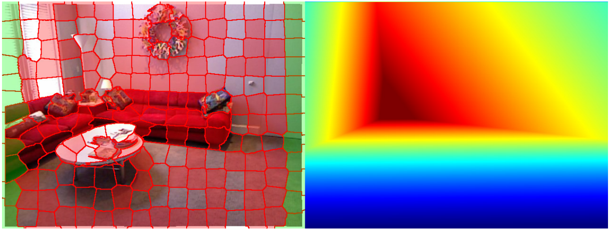


Penalty for unexplained superpixels

Model

$$E(\mathbf{x}|\mathbf{I}) = \sum_{i\in\mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i\in\mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k\in\mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}}$$
$$+ \sum_{i\in\mathcal{L}, j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i\in\mathcal{L}\cup\mathcal{O}, k\in\mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c\in\mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}$$

Superpixel Unary Potential

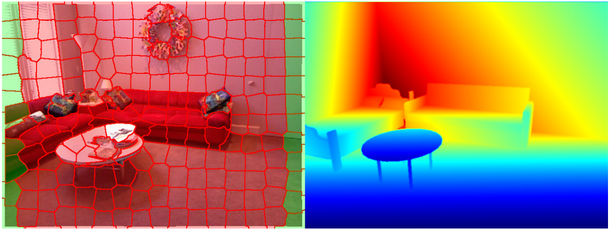


Penalty for unexplained superpixels

Model

$$E(\mathbf{x}|\mathbf{I}) = \sum_{i \in \mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i \in \mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k \in \mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j \in \mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}}$$
$$+ \sum_{i \in \mathcal{L}, j \in \mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i \in \mathcal{L} \cup \mathcal{O}, k \in \mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c \in \mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}$$

Superpixel Unary Potential



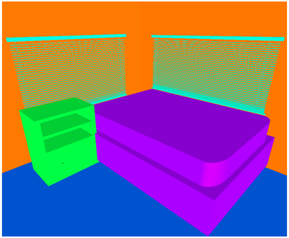
Penalty for unexplained superpixels

Model

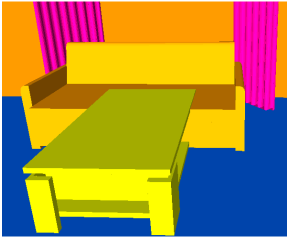
$$E(\mathbf{x}|\mathbf{I}) = \sum_{i\in\mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i\in\mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k\in\mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}}$$

$$+ \sum_{i\in\mathcal{L},j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i\in\mathcal{L}\cup\mathcal{O},k\in\mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c\in\mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}$$

Object-Object and Layout-Object Potentials



Geometric constellation



Volumetric exclusion

Model

$$
\begin{aligned}
E(\mathbf{x}|\mathbf{I}) \quad = \quad & \sum_{i\in\mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i\in\mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k\in\mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}} \\
& + \sum_{i\in\mathcal{L},j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i\in\mathcal{L}\cup\mathcal{O},k\in\mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c\in\mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}
\end{aligned}
$$

Occlusion Potential

$$
\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I}) = \begin{cases} \infty & \text{if } x_i = 1 \wedge x_k = 1 \wedge \text{“}i \text{ occludes } k\text{”} \\ 0 & \text{otherwise} \end{cases}
$$

- Only non-occluded superpixels can be explained
- Penalizes objects “occluding” the actual 3D scene

Model

$$
\begin{aligned}
E(\mathbf{x}|\mathbf{I}) \quad = \quad & \sum_{i\in\mathcal{L}} \underbrace{\phi_i^{\mathcal{L}}(x_i|\mathbf{I})}_{\text{layout}} + \sum_{i\in\mathcal{O}} \underbrace{\phi_i^{\mathcal{O}}(x_i|\mathbf{I})}_{\text{object}} + \sum_{k\in\mathcal{S}} \underbrace{\phi_k^{\mathcal{S}}(x_k)}_{\text{superpixel}} + \sum_{i,j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{O}}(x_i, x_j)}_{\text{object-object}} \\
& + \sum_{i\in\mathcal{L}, j\in\mathcal{O}} \underbrace{\psi_{ij}^{\mathcal{L},\mathcal{O}}(x_i, x_j)}_{\text{layout-object}} + \sum_{i\in\mathcal{L}\cup\mathcal{O}, k\in\mathcal{S}} \underbrace{\psi_{ik}^{\mathcal{S}}(x_i, x_k|\mathbf{I})}_{\text{occlusion}} + \sum_{c\in\mathcal{C}_{\mathbf{I}}} \underbrace{\kappa_c(\mathbf{x}_c)}_{\text{consistency}}
\end{aligned}
$$

Consistency Potential

$$
\kappa_c(\mathbf{x}_c) = \begin{cases} \infty & \text{if } x_k = 1 \wedge \sum_{i\in c\setminus\{k\}} x_i = 0 \\ 0 & \text{otherwise} \end{cases}
$$

- Only superpixels supported by objects can be explained
- Ensures consistency between superpixels and objects/layout

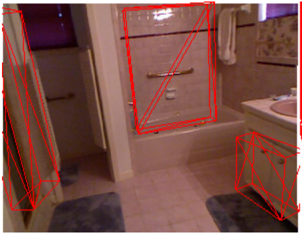
Proposal Generation

- [Lin2013,Jiang2013] fit 3D cuboids to the RGB-D point cloud

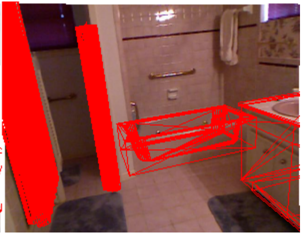
[Lin2013] Our Results

Proposal Generation

- [Lin2013,Jiang2013] fit 3D cuboids to the RGB-D point cloud
- **Problems:** Shrinking bias, rotation errors



[Lin2013]

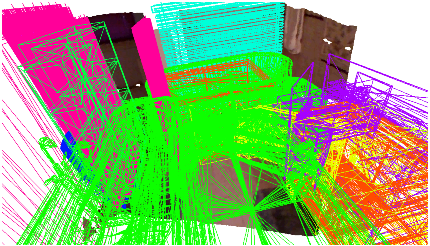


Our Results

Proposal Generation

- [Lin2013,Jiang2013] fit 3D cuboids to the RGB-D point cloud
- **Problems:** Shrinking bias, rotation errors
- **Here:** Directly sample proposals from unary distribution

$$p(\theta_i) \propto \exp\left(-\phi_i^{\mathcal{L}/\mathcal{O}}(\theta_i|\mathbf{I})\right)$$



Inference

Min-Sum Belief Propagation:

$$\underbrace{m^t_{c\to i}(x_i)}_{\text{factor}\to\text{variable}} = \min_{\mathbf{x}_c \setminus \{x_i\}} \left(f_c(\mathbf{x}_c) + \sum_{j \in c \setminus \{i\}} m^{t-1}_{j\to c}(x_j) \right)$$

$$\underbrace{m^t_{i\to c}(x_i)}_{\text{variable}\to\text{factor}} = \sum_{c' \in \mathcal{N}_i \setminus \{c\}} m^{t-1}_{c'\to i}(x_i)$$

$f_c(\mathbf{x}_c)$: factor of clique c $\mathbf{x}_c$: variables in clique c $\mathcal{N}_i$: factors involving variable i

Sparse Factor:

$$f(\mathbf{x}) = \begin{cases} \xi(\mathbf{x}) & \text{if } \mathbf{x} \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

$\mathcal{S}$: set of special states $\xi(\mathbf{x})$: arbitrary function $(-\infty \leq \xi(\mathbf{x}) \leq \infty)$

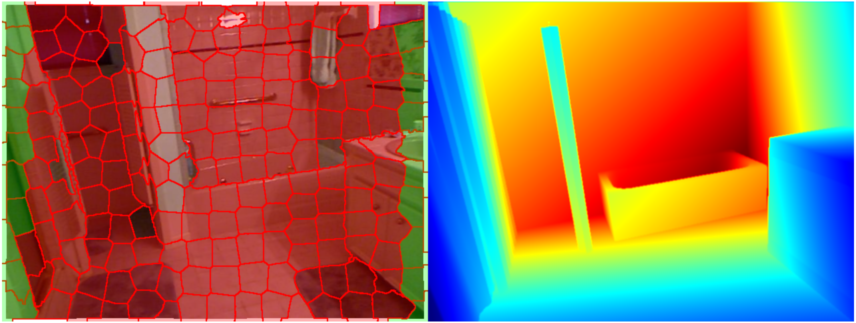
Inference

Partition: $\mathcal{T}_k(\mathbf{y}) = \{\mathbf{x} | x_i = y_i, i \leq k\}$

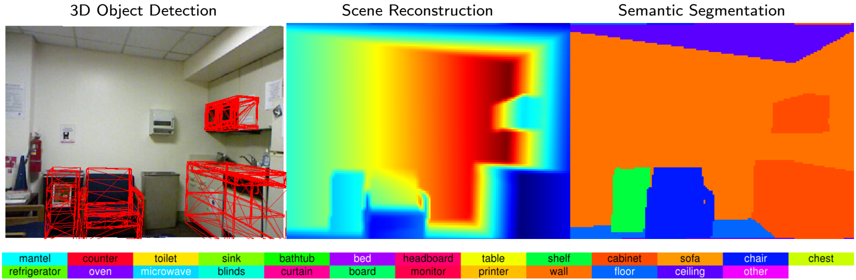
x_1	x_2	x_3	$f(\mathbf{x})$	
0	0	0	0	$\mathcal{T}_1(0)$
0	0	1	0	
0	1	0	0	
0	1	1	0	
1	0	0	0	$\mathcal{T}_2(1,0)$
1	0	1	0	
1	1	0	0	$\mathcal{T}_3(1,1,0)$
1	1	1	1	$\mathcal{T}_3(1,1,1)$

Complexity: $O(2^N) \Rightarrow O(N^2)$

Inference



Experiments



- Evaluation on NYUv2 dataset [Silberman2012]
- 795 training and 654 test images, 3D annotations [Guo2013]
- 3D object detection (F1 score, 3D BBox IOU; $\geq 30\%$)

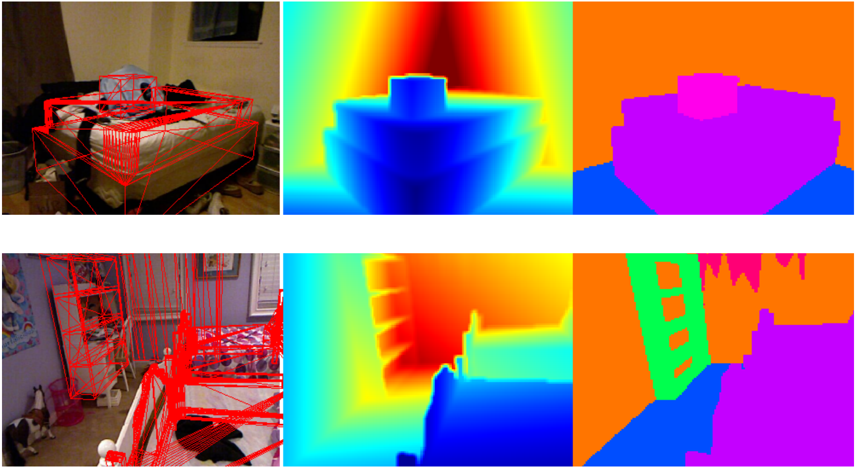
Quantitative Results

3D Object Detection (F1 score in %)

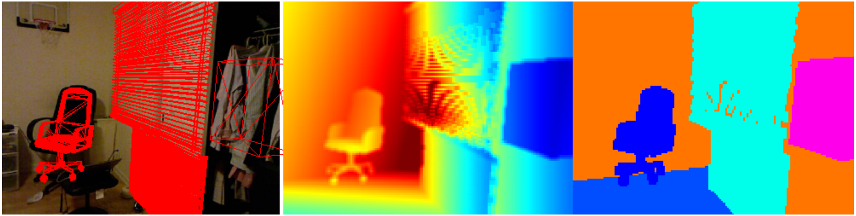
	mantel	counter	toilet	sink	bathtub	bed	headboard	table	shelf	cabinet	sofa	chair	chest	refrigerator	oven	microwave	blinds	curtain	board	monitor	printer	overall
#obj	10	126	30	36	25	169	23	455	242	534	228	703	137	42	29	40	111	91	50	81	25	3187
[Lin2013] - 8 Proposals	0	4	27	**12**	0	13	0	8	**13**	3	16	8	5	0	0	0	**13**	5	3	**8**	0	7.90
[Lin2013] - 15 Proposals	0	3	27	10	0	11	0	7	11	3	19	8	4	0	0	11	11	6	3	6	0	7.71
[Lin2013] - 30 Proposals	0	3	24	11	12	10	0	7	10	3	18	9	5	0	0	**11**	11	5	3	6	0	7.61
Base-Det-Cuboid	0	8	3	2	13	12	5	8	3	6	6	4	14	14	**7**	3	3	2	1	4	2	5.80
Base-NMS-Cuboid	0	3	16	0	0	51	6	11	8	14	12	7	24	10	6	0	10	**7**	2	7	4	11.93
NoOcclusion-Cuboid	0	5	8	3	22	51	7	15	9	17	17	10	21	17	0	0	6	6	2	1	5	13.68
NoContext-Cuboid	0	**9**	7	2	27	51	6	17	7	18	16	6	21	23	5	0	4	2	1	5	**6**	13.38
FullModel-Cuboid	0	6	8	3	23	51	7	15	8	18	17	7	24	21	0	0	6	6	2	6	5	13.45
Base-Det-CAD	0	8	13	2	11	10	5	10	4	6	8	9	14	14	**7**	4	5	3	4	4	1	7.66
Base-NMS-CAD	0	2	43	3	0	48	6	16	9	14	21	15	23	14	5	6	6	5	2	5	4	15.05
NoOcclusion-CAD	0	4	52	4	25	49	0	21	9	17	30	18	24	24	0	0	0	6	4	3	0	17.57
NoContext-CAD	0	8	47	4	28	45	7	23	8	**20**	28	**20**	25	22	0	4	2	4	**5**	4	0	18.61
FullModel-CAD	0	4	**61**	4	**31**	**55**	**7**	**24**	10	19	**33**	18	**27**	**24**	0	0	1	6	3	5	0	**19.22**

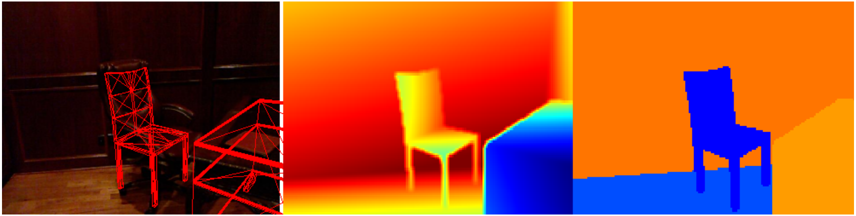
[Lin2013] Dahua Lin, Sanja Fidler and Raquel Urtasun:
Holistic scene understanding for 3D object detection with RGB-D cameras.
IEEE International Conference on Computer Vision (ICCV), 2013.

Qualitative Results

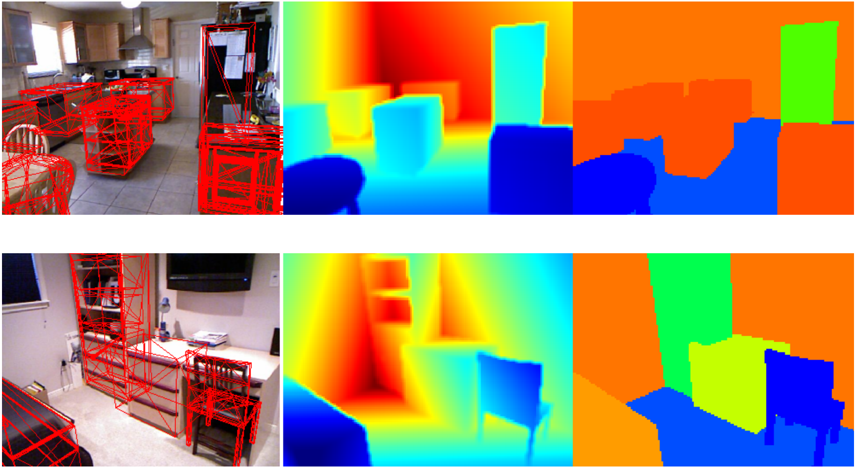


Qualitative Results

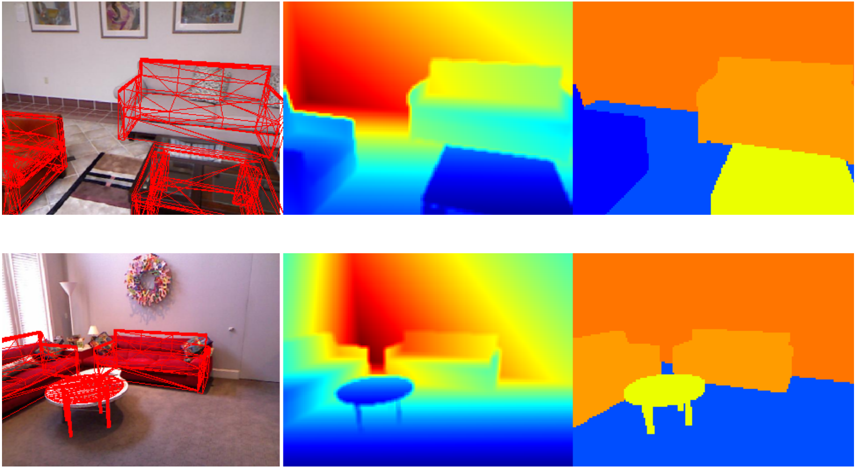




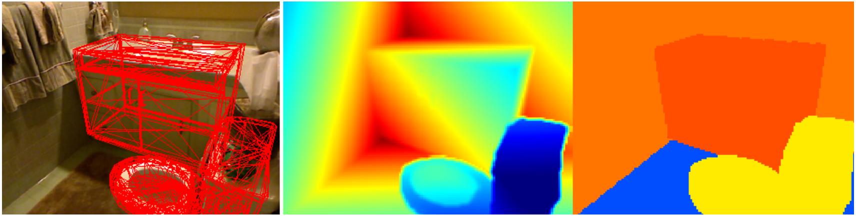
Qualitative Results

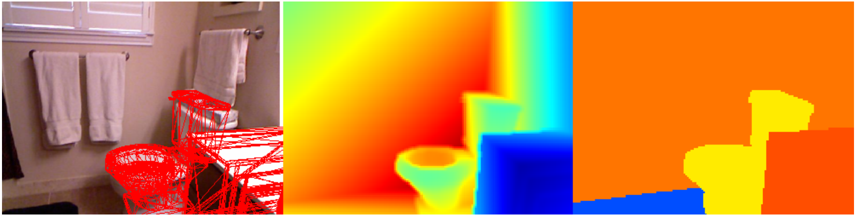


Qualitative Results

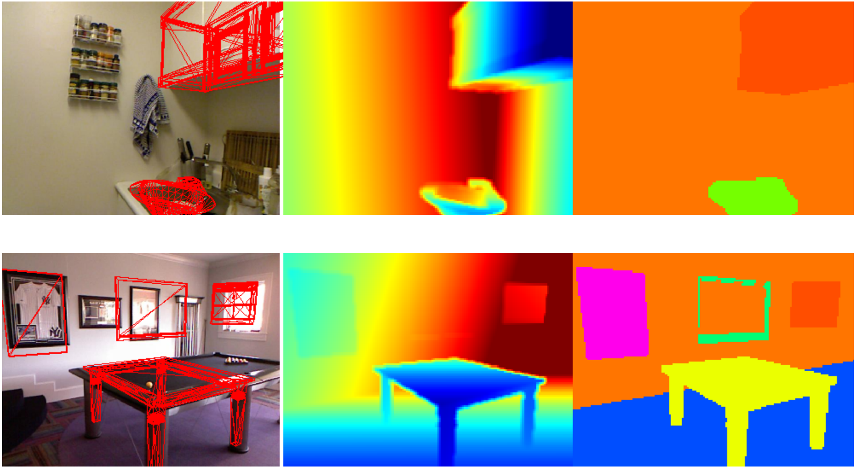


Qualitative Results

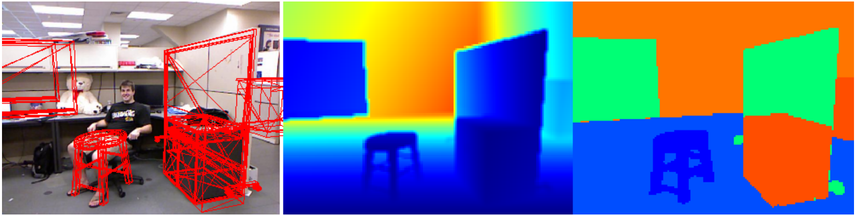


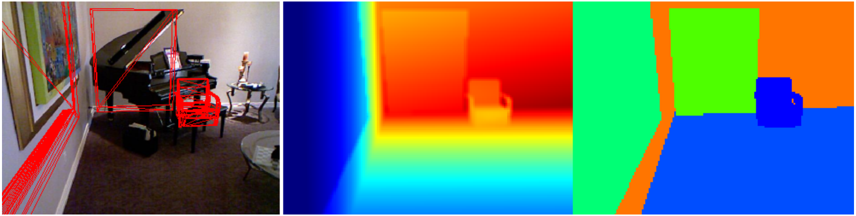


Failure Cases



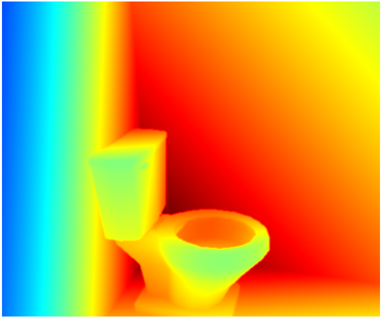
Failure Cases





Conclusion

- Novel model for indoor 3D scene understanding
- Accurate modeling of visibility constraints
- 3D CAD models improve recognition



Project page:
www.cvlibs.net/projects/indoor_scenes