Separating Objects and Clutter in Indoor Scenes

Salman H. Khan

School of Computer Science & Software Engineering,
The University of Western Australia

Co-authors: Xuming He, Mohammed Bennamoun,
Ferdous Sohel, Roberto Togenri

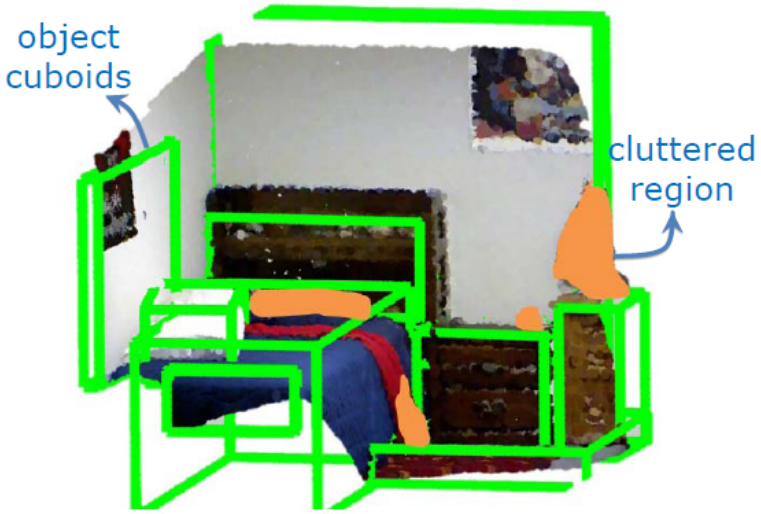
CVPR, June 2015, Boston

Outline

Outline

Problem Description

Fine-grained Structure Categorization in RGB-D Images

Input	Output
RGB-D Image	Object cuboids + Clutter
	

Outline

Motivation

- Volumetric reasoning about generic 3D objects and their 3D spatial layout is a fundamental problem in scene understanding.

Motivation

- Volumetric reasoning about generic 3D objects and their 3D spatial layout is a fundamental problem in scene understanding.
- Real-world scenes are composed of not only large regular-shaped structures and objects (such as walls, door, furniture), but also irregular shaped objects and cluttered regions which cannot be represented well by object-level primitives.

Outline

Related Works

3D Spatial Layout Estimation

- Representing objects as 3D geometric primitives (such as cuboids) [Lin'13, Jiang'13, Schwing'13].

Clutter Identification

- Clutter reasoning in the scene layout estimation setting [Hedau'09, Wang'13, Zhang'13].

[Lin'13] D. Lin, et al., Holistic scene under- standing for 3d object detection with rgbd cameras. ICCV, 2013.
[Jiang'13] H. Jiang and J. Xiao. A linear approach to matching cuboids in rgbd images. In CVPR. IEEE, 2013.
[Schwing'13] A. G. Schwing, et al., Box in the box: Joint 3d layout and object reasoning from single images. In ICCV. IEEE, 2013.
[Hedau'09] V. Hedau, et al., Recovering the spatial layout of cluttered rooms. In ICCV. IEEE, 2009.
[Wang'13] H. Wang, et al., Discriminative learning with latent variables for cluttered indoor scene understanding. Communications of the ACM, 2013.
[Zhang'13] J. Zhang, et al. Estimating the 3d layout of indoor scenes and its clutter from depth sensors. 2013.

Outline

Our Approach

Contributions:

Goal: "3D object cuboid detection in a cluttered scene"

So, what's new?

- We jointly localize generic 3D objects (represented by cuboids) and label cluttered regions from an RGBD image.

Our Approach

Contributions:

Goal: "3D object cuboid detection in a cluttered scene"

So, what's new?

- We jointly localize generic 3D objects (represented by cuboids) and label cluttered regions from an RGBD image.
- Our definition of clutter is more refined.
  'Any unordered region other than the main structures and major regular shaped objects is a cluttered region'.

Our Approach

Contributions:

Goal: "3D object cuboid detection in a cluttered scene"

So, what's new?

- We jointly localize generic 3D objects (represented by cuboids) and label cluttered regions from an RGBD image.
- Our definition of clutter is more refined.
  'Any unordered region other than the main structures and major regular shaped objects is a cluttered region'.
- We propose a higher order CRF (defined on cuboid hypotheses and superpixels) whose parameters are learned using structural learning.

CRF Formulation

The CRF model describes an RGBD image with an optimal set of cuboids and pixel-level labeling of cluttered regions.

The CRF model encodes:

- interactions b/w cuboids
- interactions b/w superpixels
- interactions between superpixels and cuboids

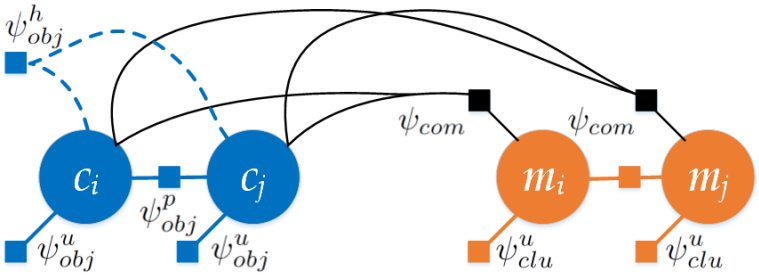


Figure: Graph structure representation for the potentials defined on the object cuboids and the cluttered/non-cluttered regions.

The Gibbs energy is given by,

$$E(\mathbf{m}, \mathbf{c}|\mathcal{I}) = E_{obj}(\mathbf{c}) + E_{sp}(\mathbf{m}) + E_{com}(\mathbf{m}, \mathbf{c}), \tag{1}$$

Outline

Cuboid Hypothesis Generation

- We define two types of cuboids in the hypothesis set:
  - Scene bounding cuboids ($\mathcal{O}_{sbc}$),
  - Object cuboids ($\mathcal{O}_{oc}$)
- Hypothesis generation process is based on bottom-up clustering and cuboid fitting procedure.
- Cuboid fitting procedure is different for scene bounding cuboids and object cuboids.
- Cuboid hypothesis set significantly reduces the search space of 3D cuboids.
  - Top 60 cuboids are selected based on cuboid unary potential ranking.

Outline

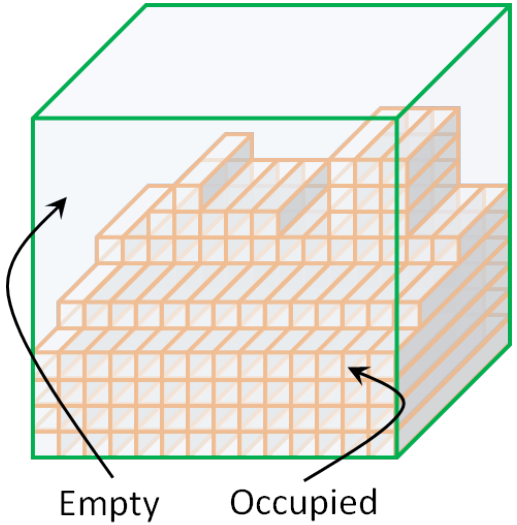
Potentials on Cuboids

$E_{obj}(\mathbf{c})$, is defined as a combination of three potential functions:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K} \left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k) \right] + \sum_{i<j} \psi^{p}_{obj}(c_i, c_j), \tag{2}$$

cuboid likelihood MDL prior pairwise potential

- The unary potential encodes appearance, physical and geometrical based properties.
  - Volumetric occupancy feature



We use a learning approach to predict this potential.

Potentials on Cuboids

$E_{obj}(\mathbf{c})$, is defined as a combination of three potential functions:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K} \left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k) \right] + \sum_{i<j} \psi^{p}_{obj}(c_i, c_j), \tag{2}$$

cuboid likelihood MDL prior pairwise potential

- The unary potential encodes appearance, physical and geometrical based properties.
  - Volumetric occupancy feature
  - Color consistency feature

We use a learning approach to predict this potential.

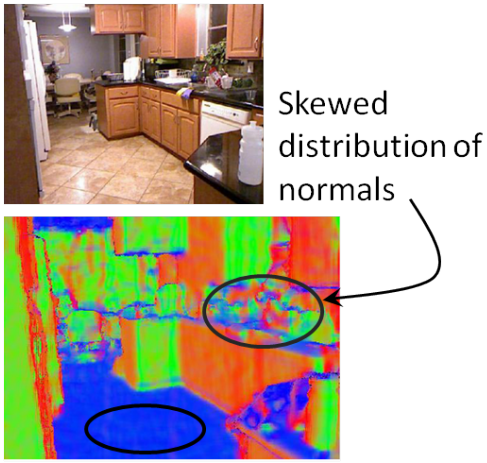
Potentials on Cuboids

$E_{obj}(\mathbf{c})$, is defined as a combination of three potential functions:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K} \left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k) \right] + \sum_{i<j} \psi^{p}_{obj}(c_i, c_j), \tag{2}$$

cuboid likelihood MDL prior pairwise potential

- The unary potential encodes appearance, physical and geometrical based properties.
  - Volumetric occupancy feature
  - Color consistency feature
  - Normal consistency feature



We use a learning approach to predict this potential.

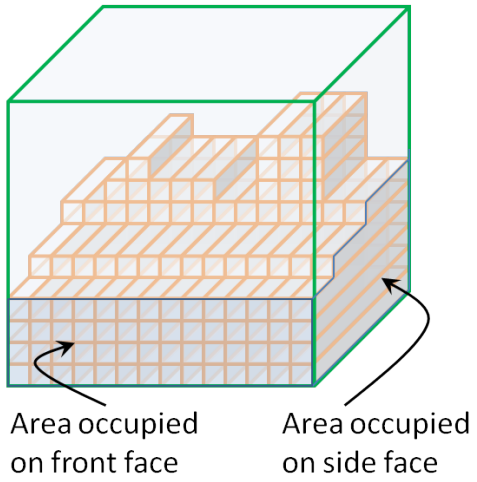
Potentials on Cuboids

$E_{obj}(\mathbf{c})$, is defined as a combination of three potential functions:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K} \left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k) \right] + \sum_{i<j} \psi^{p}_{obj}(c_i, c_j), \tag{2}$$

cuboid likelihood MDL prior pairwise potential

- The unary potential encodes appearance, physical and geometrical based properties.
  - Volumetric occupancy feature
  - Color consistency feature
  - Normal consistency feature
  - Tightness feature



We use a learning approach to predict this potential.

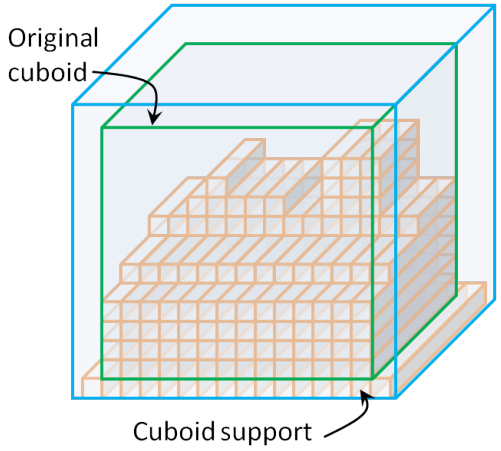
Potentials on Cuboids

$E_{obj}(\mathbf{c})$, is defined as a combination of three potential functions:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K} \left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k) \right] + \sum_{i<j} \psi^{p}_{obj}(c_i, c_j), \tag{2}$$

cuboid likelihood MDL prior pairwise potential

- The unary potential encodes appearance, physical and geometrical based properties.
  - Volumetric occupancy feature
  - Color consistency feature
  - Normal consistency feature
  - Tightness feature
  - Support feature



We use a learning approach to predict this potential.

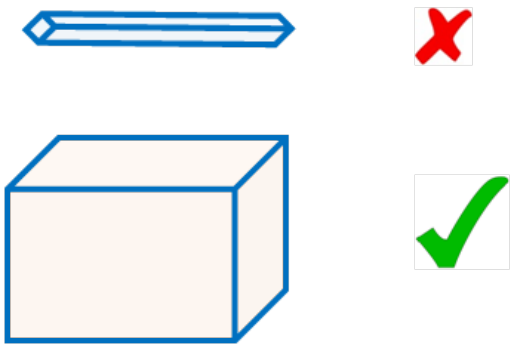
Potentials on Cuboids

$E_{obj}(\mathbf{c})$, is defined as a combination of three potential functions:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K} \left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k) \right] + \sum_{i<j} \psi^{p}_{obj}(c_i, c_j), \tag{2}$$

cuboid likelihood MDL prior pairwise potential

- The unary potential encodes appearance, physical and geometrical based properties.
  - Volumetric occupancy feature
  - Color consistency feature
  - Normal consistency feature
  - Tightness feature
  - Support feature
  - Geometric plausibility feature



We use a learning approach to predict this potential.

Potentials on Cuboids

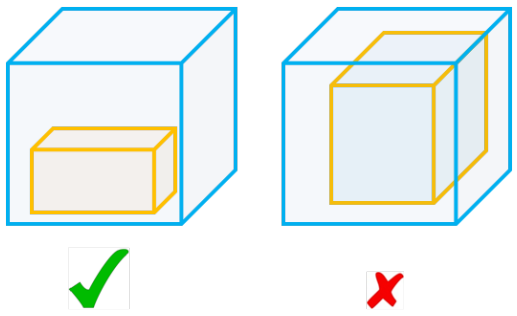
$E_{obj}(\mathbf{c})$, is defined as a combination of three potential functions:

$$E_{obj}(\mathbf{c}) = \sum_{k=1}^{K} \left[\psi^{u}_{obj}(c_k) + \psi^{h}_{obj}(c_k) \right] + \sum_{i<j} \psi^{p}_{obj}(c_i, c_j), \tag{2}$$

cuboid likelihood MDL prior pairwise potential

- The unary potential encodes appearance, physical and geometrical based properties.
  - Volumetric occupancy feature
  - Color consistency feature
  - Normal consistency feature
  - Tightness feature
  - Support feature
  - Geometric plausibility feature
  - Cuboid size feature

  We use a learning approach to predict this potential.



Potentials on Cuboids

Continued ..

- The MDL potential prefers to explain a given image in terms of a small number of cuboids.
- The pairwise energy encodes interactions between cuboids. It has two components:

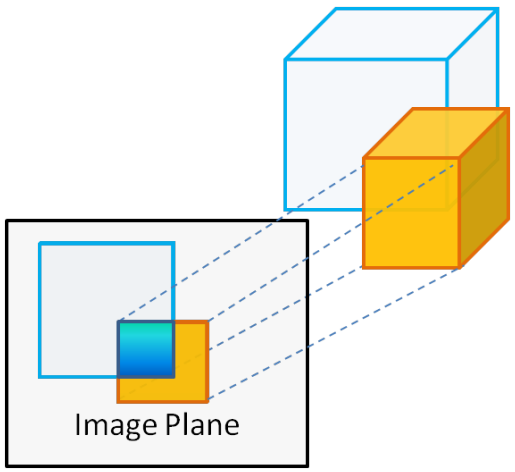
- View obstruction potential

$$\psi^{p}_{obs}(c_i, c_j) = \lambda_{obs}\tilde{\mu}^{obs}_{i,j} c_i c_j \quad (3)$$

$$\mu^{obs}_{i,j} = (A_{c_i} \cap A_{c_j})/A_{c_i} \quad (4)$$

- Box intersection potential

$$\psi^{p}_{int}(c_i, c_j) = \lambda_{int}\tilde{\mu}^{int}_{i,j} c_i c_j = \lambda_{int}\tilde{\mu}^{int}_{i,j} x_{i,j} \quad (5)$$



Potentials on Cuboids

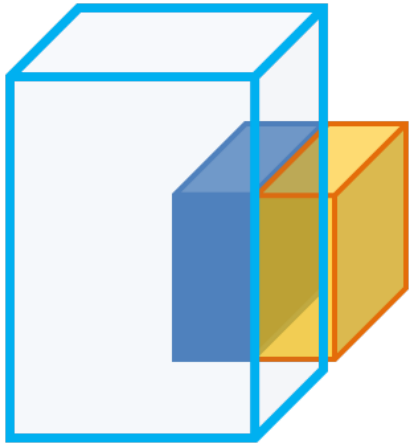
Continued ..

- The MDL potential prefers to explain a given image in terms of a small number of cuboids.
- The pairwise energy encodes interactions between cuboids. It has two components:

- View obstruction potential

$$\psi^{p}_{obs}(c_i, c_j) = \lambda_{obs}\tilde{\mu}^{obs}_{i,j} c_i c_j \quad (3)$$

$$\mu^{obs}_{i,j} = (A_{c_i} \cap A_{c_j})/A_{c_i} \quad (4)$$



- Box intersection potential

$$\psi^{p}_{int}(c_i, c_j) = \lambda_{int}\tilde{\mu}^{int}_{i,j} c_i c_j = \lambda_{int}\tilde{\mu}^{int}_{i,j} x_{i,j} \quad (5)$$

Outline

Potentials on Superpixels

E_{sp}, consists of unary and pairwise smoothness terms:

$$E_{sp}(\mathbf{m}) = \underbrace{\sum_{j=1}^{J} \psi_{sp}^{u}(m_j)}_{\text{data likelihood}} + \underbrace{\sum_{(i,j)\in N_s} \psi_{sp}^{p}(m_i, m_j)}_{\text{pairwise potential}}, \tag{6}$$

- The unary potential captures appearance and texture properties of cluttered and non-cluttered regions. A learning based approach is used on top of kernel descriptors proposed by [Bo et al., 2011].
- A contrast sensitive Potts model is used as pairwise potential to encourage smoothness:

$$\psi_{sp}^{p}(m_i, m_j) = \lambda_{smo}\mu_{i,j}^{smo}(m_i + m_j - m_i \cdot m_j), \tag{7}$$

where, $\mu_{i,j}^{smo} = \exp(-\|\bar{v}_i - \bar{v}_j\|^2/\sigma_c^2)$, $\bar{v}_i, \bar{v}_j$ are the mean color of superpixel s_i and s_j.

Outline

Cuboid-Superpixel Compatibility

This term enforces the consistency of the cuboid activations and the superpixel labeling:

$$E_{com}(\mathbf{m}, \mathbf{c}) = \sum_{j=1}^{J} \psi_{com}(m_j, \mathbf{c}). \tag{8}$$

The compatibility potential consists of two terms:

$$\psi_{com}(m_j, \mathbf{c}) = \psi_{mem}(m_j, \mathbf{c}) + \sum_{k} \psi_{occ}(m_j, c_k), \tag{9}$$

- Superpixel membership term,

$$\psi_{mem}(m_j, \mathbf{c}) = \lambda_{\infty} [\![m_j \neq \max_{k: s_j \in o_k} c_k]\!], \tag{10}$$

- Occlusion relationship term,

$$\psi_{occ}(m_j, c_k) = \lambda_{occ} \mu_{jk}^{occ} \overline{m}_j c_k = \lambda_{occ} \mu_{jk}^{occ} z_{jk} \tag{11}$$

where, $\overline{m}_j = 1 - m_j$, and z_{jk} is the auxiliary variable for linearization.

Outline

Model Inference and Learning

Inference as MILP

For model inference, we minimize the CRF energy:

$$\{\mathbf{m}^*, \mathbf{c}^*\} = \underset{\mathbf{m},\mathbf{c}}{\operatorname{argmin}}\ E(\mathbf{m}, \mathbf{c}|\mathcal{I}). \tag{12}$$

The relaxation method in [Jiang et al., 2013] is used to transform Eq. 12 into a MILP with linear constraints.

- For pairwise view obstruction cost in Eq. (3), we introduce $y_{i,j}$ for $c_i \cdot c_j$ with constraints: $c_i \geq y_{i,j}, c_j \geq y_{i,j}, y_{i,j} \geq c_i + c_j - 1$.
- Similarly, auxiliary boolean variables $x_{i,j}$, $w_{i,j}$ and $z_{j,k}$ are introduced in subsequent potentials.
- MILP formulation can be solved efficiently compared to the original ILP, using the branch and bound method.

Model Inference and Learning

Inference as MILP - continued

- The complete set of linear inequality constraints is:

$$c_i \geq y_{i,j},\ c_j \geq y_{i,j},\ y_{i,j} \geq c_i + c_j - 1, \tag{13}$$

$$\forall i,j:\ o_i \text{ and } o_j \in \mathcal{O}_{oc},\ 0 \leq \mu^{obs}_{i,j} < \alpha_{obs},$$

$$\forall i,j:\ o_i \text{ or } o_j \in \mathcal{O}_{sbc},\ 0 \leq \mu^{obs}_{i,j} < \alpha'_{obs}.$$

$$c_i \geq x_{i,j},\ c_j \geq x_{i,j},\ x_{i,j} \geq c_i + c_j - 1, \tag{14}$$

$$\forall i,j:\ o_i \text{ and } o_j \in \mathcal{O}_{oc},\ 0 \leq \mu^{int}_{i,j} < \alpha_{int},$$

$$\forall i,j:\ o_i \text{ or } o_j \in \mathcal{O}_{sbc},\ 0 \leq \mu^{int}_{i,j} < \alpha'_{int}.$$

$$c_i + c_j \leq 1, \tag{15}$$

$$\forall i,j:\ o_i \text{ and } o_j \in \mathcal{O}_{oc},\ \mu^{int}_{i,j} \geq \alpha_{int} \vee \mu^{obs}_{i,j} \geq \alpha_{obs},$$

$$\forall i,j:\ o_i \text{ or } o_j \in \mathcal{O}_{sbc},\ \mu^{int}_{i,j} \geq \alpha'_{int} \vee \mu^{obs}_{i,j} \geq \alpha'_{obs}.$$

$$m_i \geq w_{i,j},\ m_j \geq w_{i,j},\ w_{i,j} \geq m_i + m_j - 1,\ \forall i,j \tag{16}$$

$$m_j \leq \sum_{k:s_j \in o_k} c_k . \forall j \quad c_k \geq z_{j,k}, m_j \leq 1 - z_{j,k}, z_{j,k} \geq c_k - m_j, \tag{17}$$

$$\forall k:\ o_k \in \mathcal{O}_{oc},\ 0 \leq \mu^{occ}_{j,k} < \alpha_{int}, \qquad \forall k:\ o_k \in \mathcal{O}_{sbc},\ 0 \leq \mu^{occ}_{j,k} < \alpha'_{int}.$$

Model Inference and Learning

Inference as MILP - continued

We empirically evaluate the efficiency of Branch and Bound algorithm for inference.

- On average, $819 \pm 48\%$ variables were involved in each inference.
- Final MILP gap was zero for 98.5% of the cases on the whole test set.

	Small gap	Large gap	Cuts	LP relax.
Time (sec)	$1.84 \pm 31\%$	$1.31 \pm 24\%$	$0.45 \pm 13\%$	$0.001 \pm 0.4\%$
Det. Rate	26.8%	26.1%	24.4%	19.9%

Table: Inference running time comparisons for variants of MILP formulation.

Outline

Model Inference and Learning

Parameter Learning

- Our model has a large number of parameters ($\boldsymbol{\lambda} = \{\lambda_{bbu}, \lambda_{mdl}, \lambda_{obs}, \lambda_{int}, \lambda_{app}, \lambda_{smo}, \lambda_{occ}\}$).
- We adopt a systematic approach to learn all the model parameters using cutting plane algorithm of [Joachims et al., 2009].
- The algorithm efficiently adds low energy labelings to the active constraints set and updates the parameters such that the ground-truth has the lowest energy.
- We use the IOU loss function on cuboid matching as our loss function in learning:

$$\Delta(\mathbf{t}^{(n)}, \mathbf{t}) = \sum_i \left(1 - \frac{|o_i^{(n)} \cap o_i|}{|o_i^{(n)} \cup o_i|}\right). \tag{18}$$

Experimental Results

Dataset and Setup

Evaluation is performed on 3D cuboid detection dataset released with RMRC, 2013.

- It contains 1074 RGBD Images taken from NYU Depth v2 dataset.
- There are 7701 annotated 3D bounded boxes in total ($\sim$7 cuboids/image).
- Experiments are performed on complete dataset using 3-fold cross validation.
- For each fold, 716/358 train/test split is used.

Experimental Results

Cuboid Detection Task

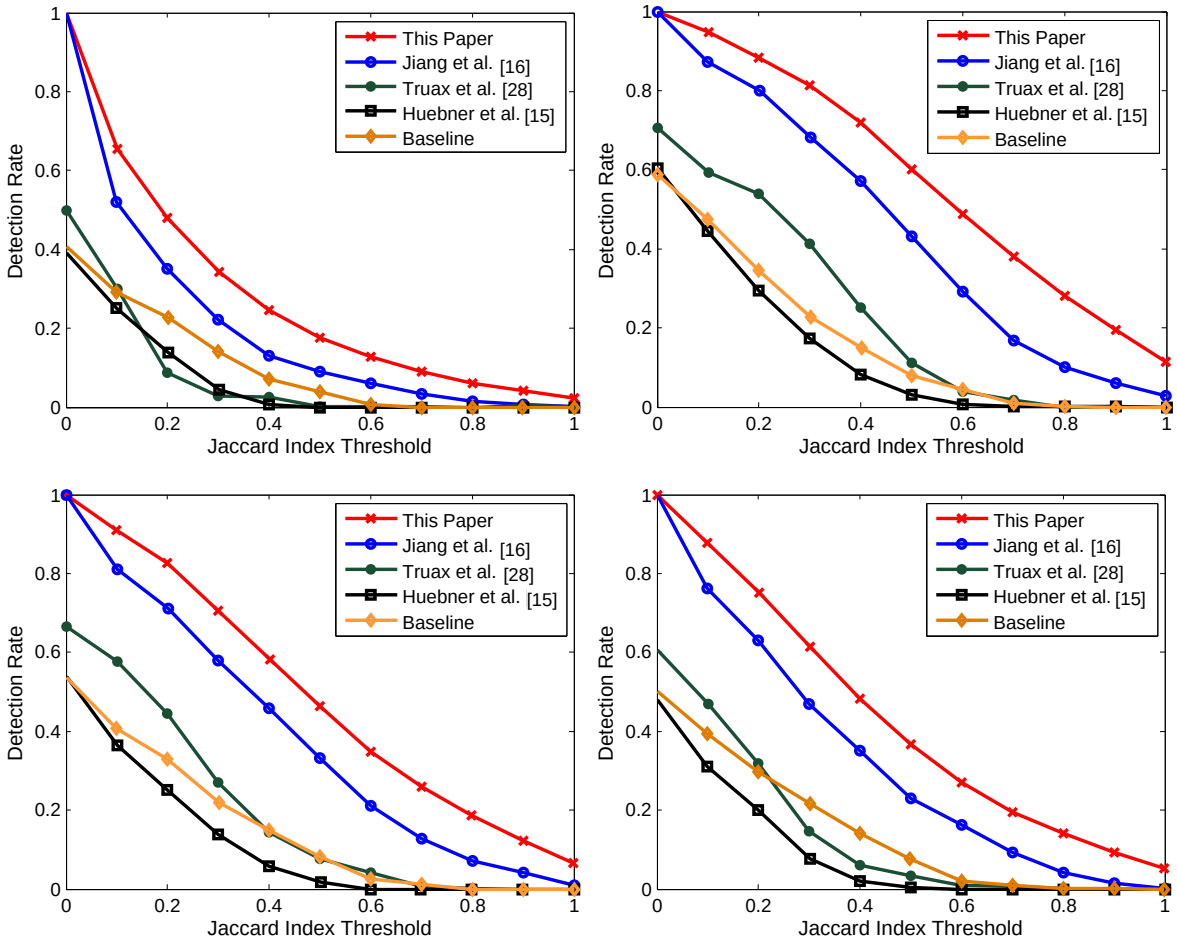


Figure: Jaccard Index comparisons for all annotated cuboids (*top left*), for the most salient cuboid (*top right*), for top two salient cuboids (*bottom left*) and top three salient cuboids (*bottom right*).

Experimental Results

Cuboid Detection Task - Continued

Ablative Analysis

Both the newly introduced features and the joint modeling contribute to the overall improvement in detection accuracy.

Method	Accuracy
Unary cuboid cost of [Jiang et al., 2013]	6.5%
Our unary cuboid cost only	8.8%
Our unary + pairwise cuboid cost only	19.4%
Our full model	26.1%

Table: An ablation study on the model potentials/features for the cuboid detection task at the 40% JI threshold.

Qualitative Results - I

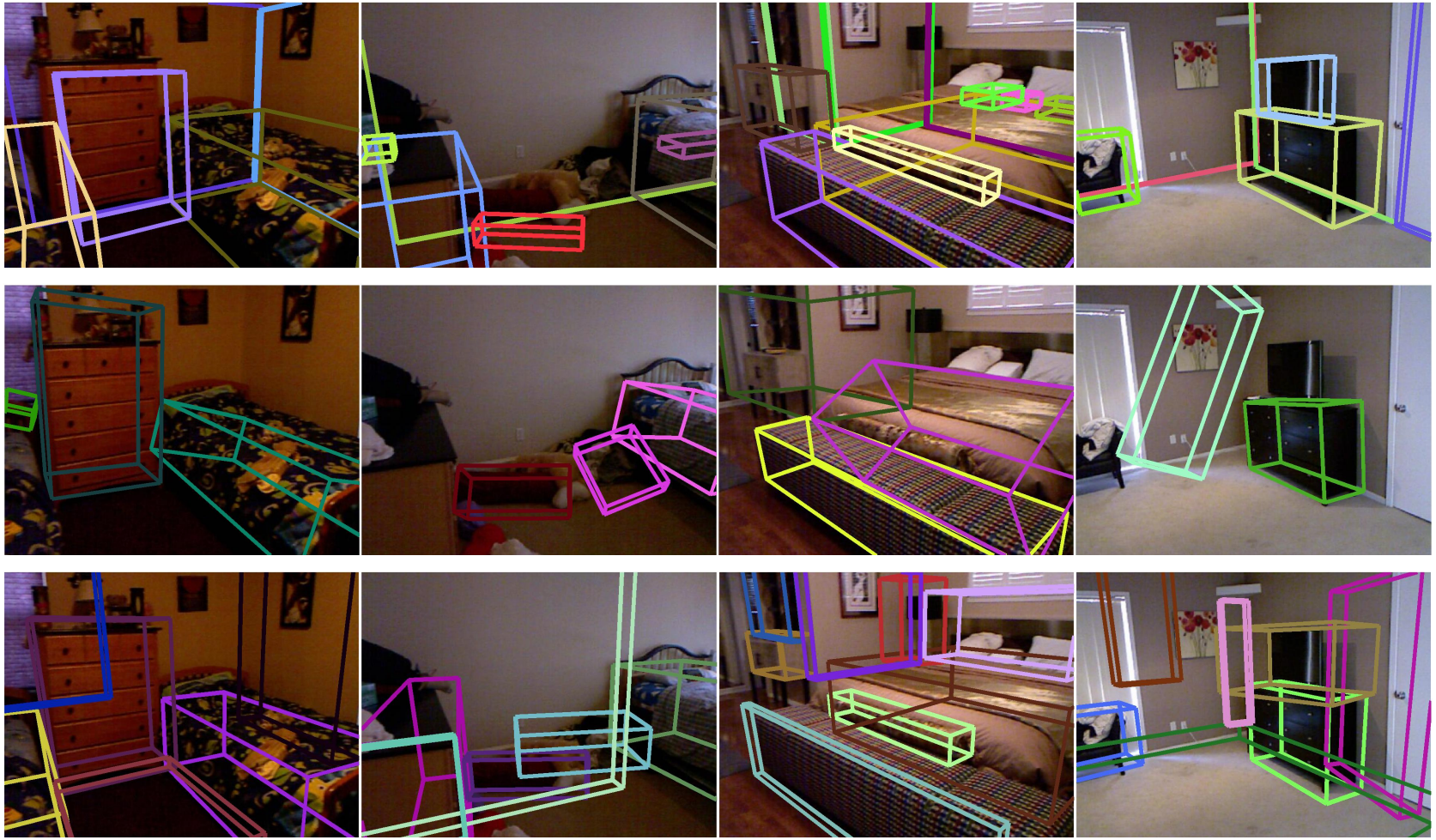


Figure: Comparison of our results (*3rd row*) with the state of the art technique [Jiang et al., 2013] (*2nd row*) and Ground Truth (*1st row*).

Qualitative Results - II

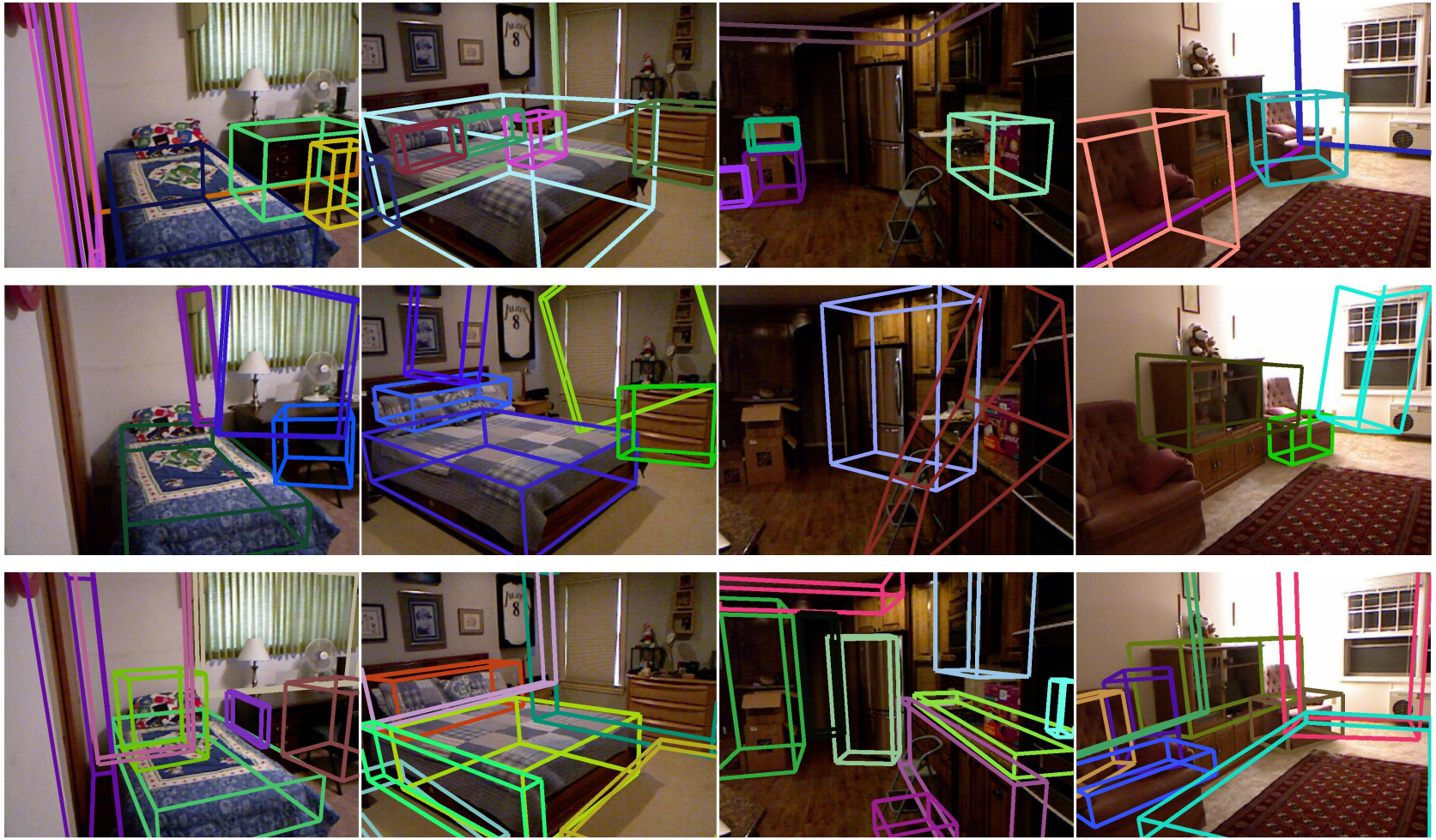


Figure: Comparison of our results (*3rd row*) with the state of the art technique [Jiang et al., 2013] (*2nd row*) and Ground Truth (*1st row*).

Experimental Results

Cluter/Non-Clutter Segmentation Task

Ablative Analysis

Method	Precision	Recall	F-Score
Superpixel unary only	$0.43 \pm 13\%$	$0.45 \pm 11\%$	$0.44 \pm 16\%$
Unary + pairwise	$0.46 \pm 12\%$	$0.48 \pm 10\%$	$0.47 \pm 16\%$
Full model (all classes)	$0.65 \pm 9\%$	$0.68 \pm 8\%$	$0.66 \pm 12\%$
Full model (only object classes)	$0.75 \pm 6\%$	$0.71 \pm 8\%$	$0.73 \pm 10\%$

Table: Evaluation on Clutter/Non-Clutter Segmentation Task. Precision signifies the accuracy of clutter classification.

Qualitative Results - I

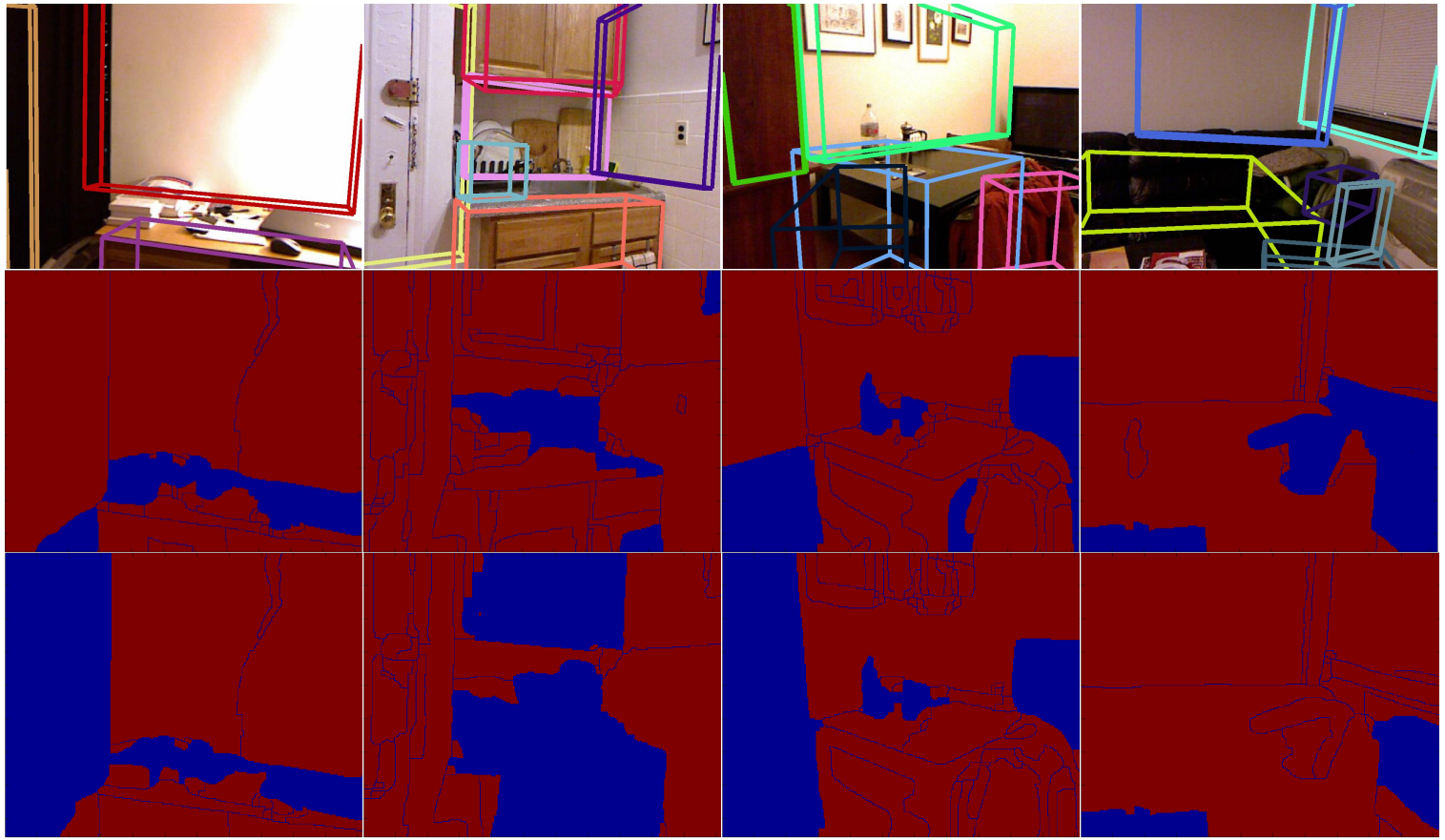


Figure: Our method is able to accurately detect cuboids in the case of cluttered indoor scenes (1st row). The 2nd and 3rd rows show our clutter labelling and the ground-truth labelling on superpixels, respectively. In the *bottom* two rows, *red* color represents *non-clutter* while *blue* color represents *clutter*.

Qualitative Results - II

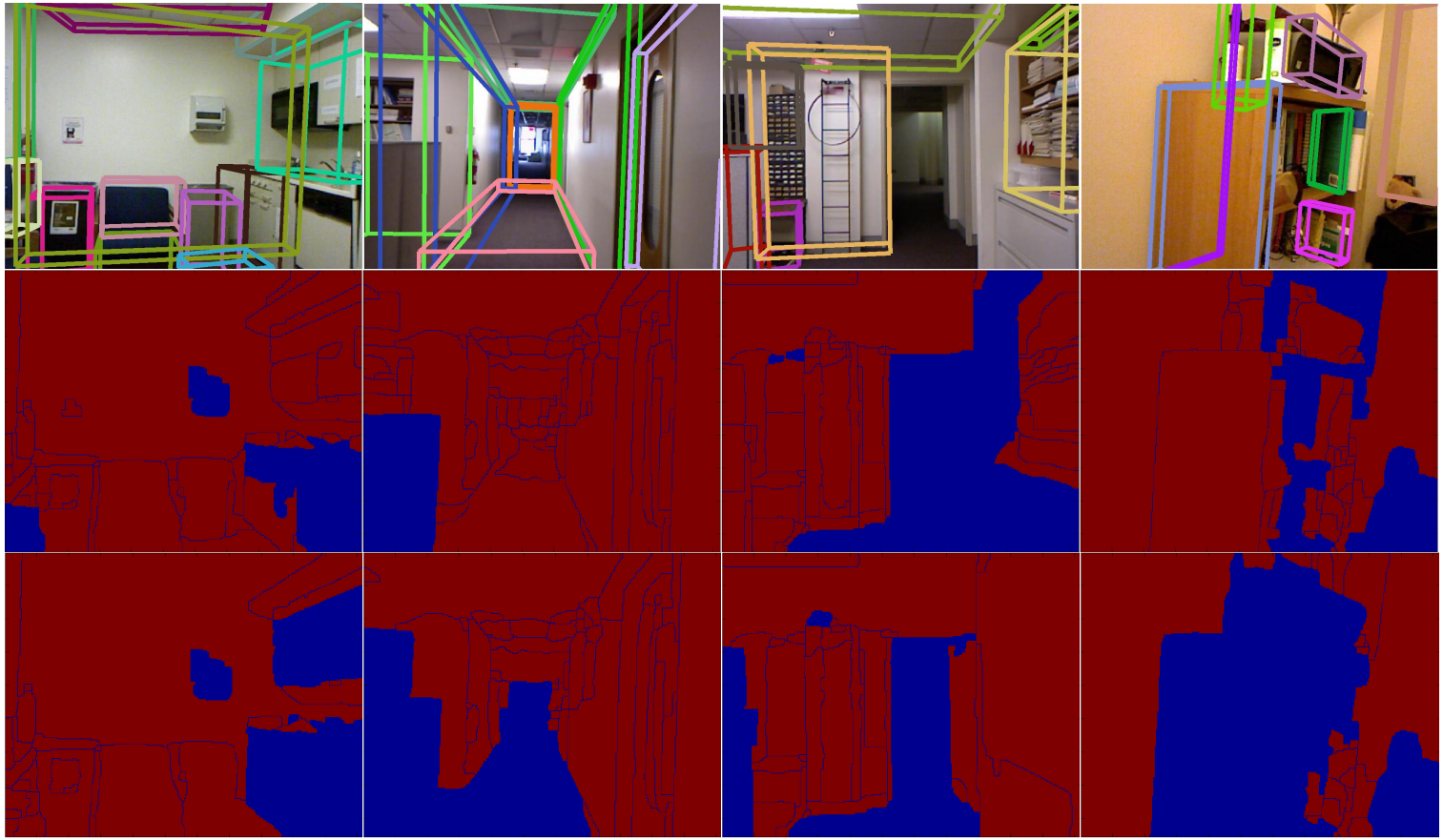


Figure: Our method is able to accurately detect cuboids in the case of cluttered indoor scenes (1st row). The 2nd and 3rd rows show our clutter labelling and the ground-truth labelling on superpixels, respectively. In the *bottom* two rows, *red* color represents *non-clutter* while *blue* color represents *clutter*.

Experimental Results

Foreground Segmentation Task

Eval. Criterion	CPMC [Carreira et al., 2012]		This Paper	
	Pre.	Rec.	Pre.	Rec.
Most salient obj.	$0.83 \pm 11\%$	0.79 ± 12	$0.85 \pm 15\%$	$0.82 \pm 15\%$
Top 2 salient obj.	$0.77 \pm 12\%$	0.73 ± 14	$0.81 \pm 16\%$	$0.79 \pm 16\%$
Top 3 salient obj.	$0.69 \pm 15\%$	0.66 ± 17	$0.79 \pm 21\%$	$0.76 \pm 19\%$
All objects	$0.54 \pm 17\%$	0.51 ± 20	$0.73 \pm 23\%$	$0.69 \pm 21\%$

Table: Evaluation on Foreground/Background Segmentation Task. Precision signifies the accuracy of foreground detection.

Experimental Results

Limitations/Failure Cases

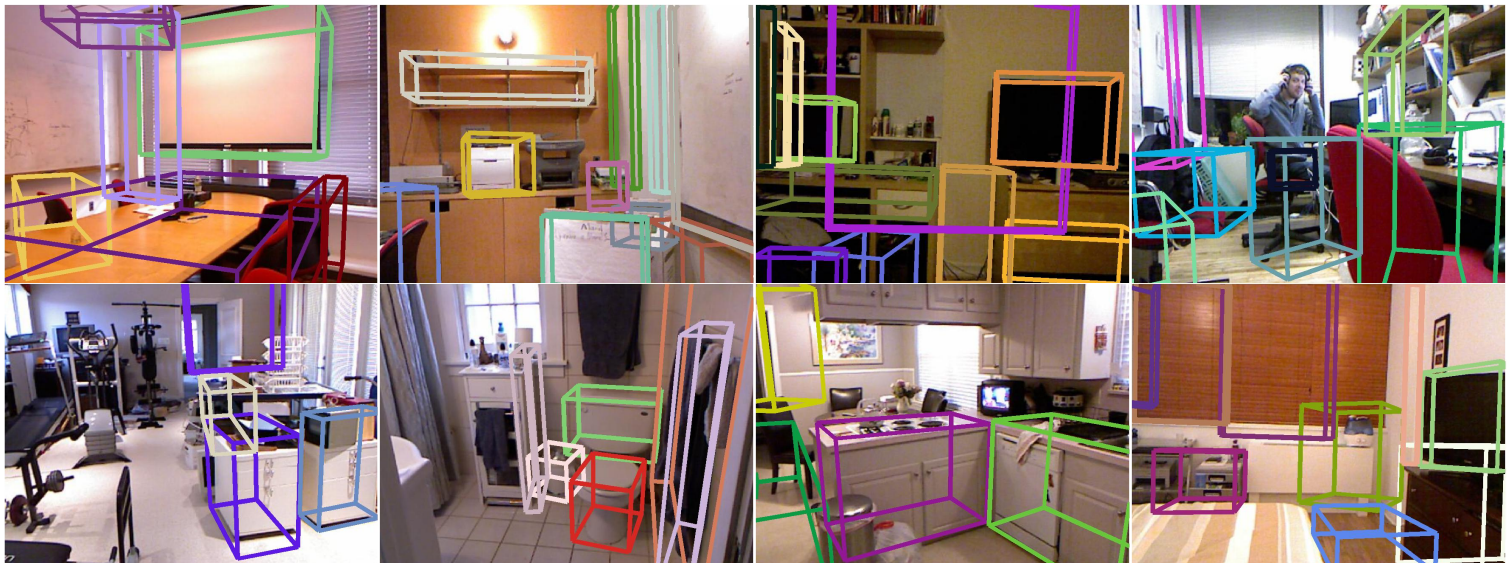


Figure: Ambiguous Cases: Examples of detection errors.

- Object detection relies on initial cuboid hypothesis set.
- No object proposals when only one side was visible.
- Due to missing depth values, specular surfaces were confused with cluttered regions.

Summary

- Cuboid detection and clutter estimation to develop better holistic understanding of indoor scenes.
- Our approach jointly models 3D generic objects as cuboids and cluttered regions as local surfaces.
- We build a CRF model and incorporate rich set of appearance and geometric features.
- We learn its parameters in a systematic manner using structural learning.
- Model inference is formulated as a MILP, which can be solved efficiently.

Thank You!

Questions?