Geometry Driven Semantic Labeling of Indoor Scenes

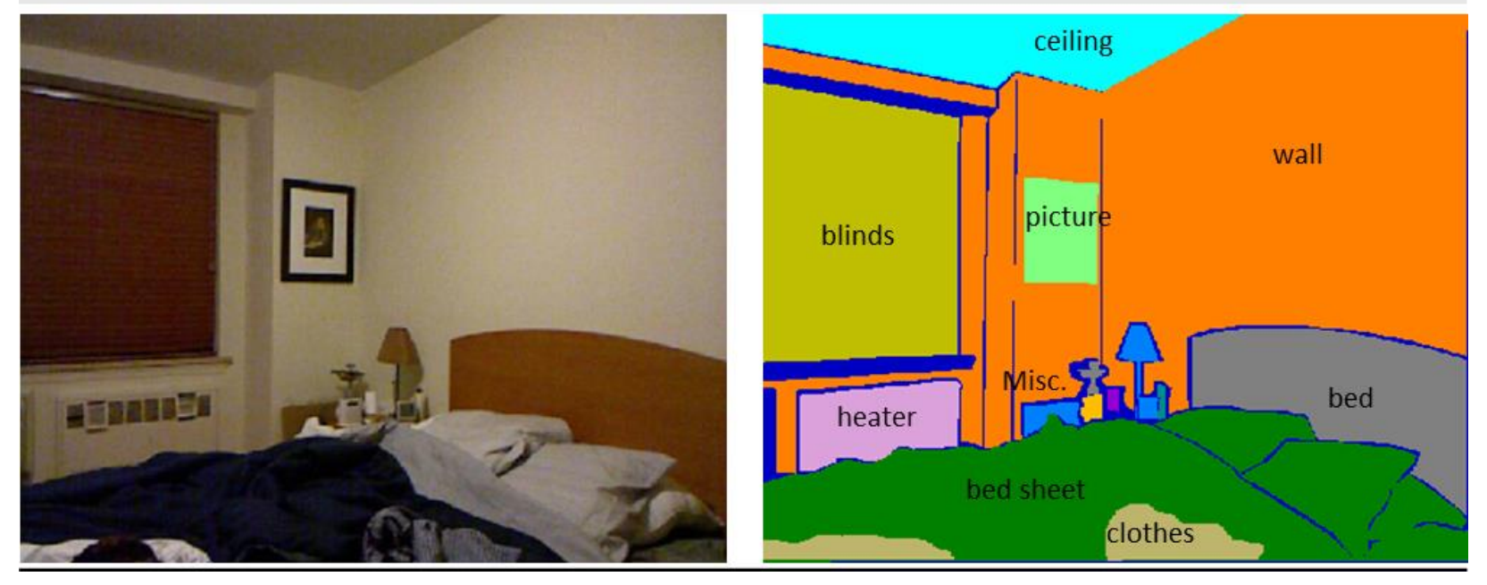
Salman H. Khan[1], Mohammed Bennamoun[1], Ferdous Sohel[1], Roberto Togneri[2]
[1] School of CSSE, [2] School of EECE, The University of Western Australia

ECCV '14
Zürich, September 6-12th, 2014

Introduction

Problem Definition:

Input	Output
RGB-D Image	Segmentation + Labelling



Our Contributions:

- A depth-based geometrical CRF model to more effectively utilize the depth information along side the RGB data
- A novel smoothness constraint based region growing algorithm for plane detection
- A hierarchical fusion scheme to combine appearance and geometry based unary potentials
- A new Spatial Discontinuity Potential which combines various edge strengths
- Our proposed location potential models the spatial location of semantic classes
- The proposed HOP increases the expressivity of the random field model by assimilating the geometric context

Our Approach

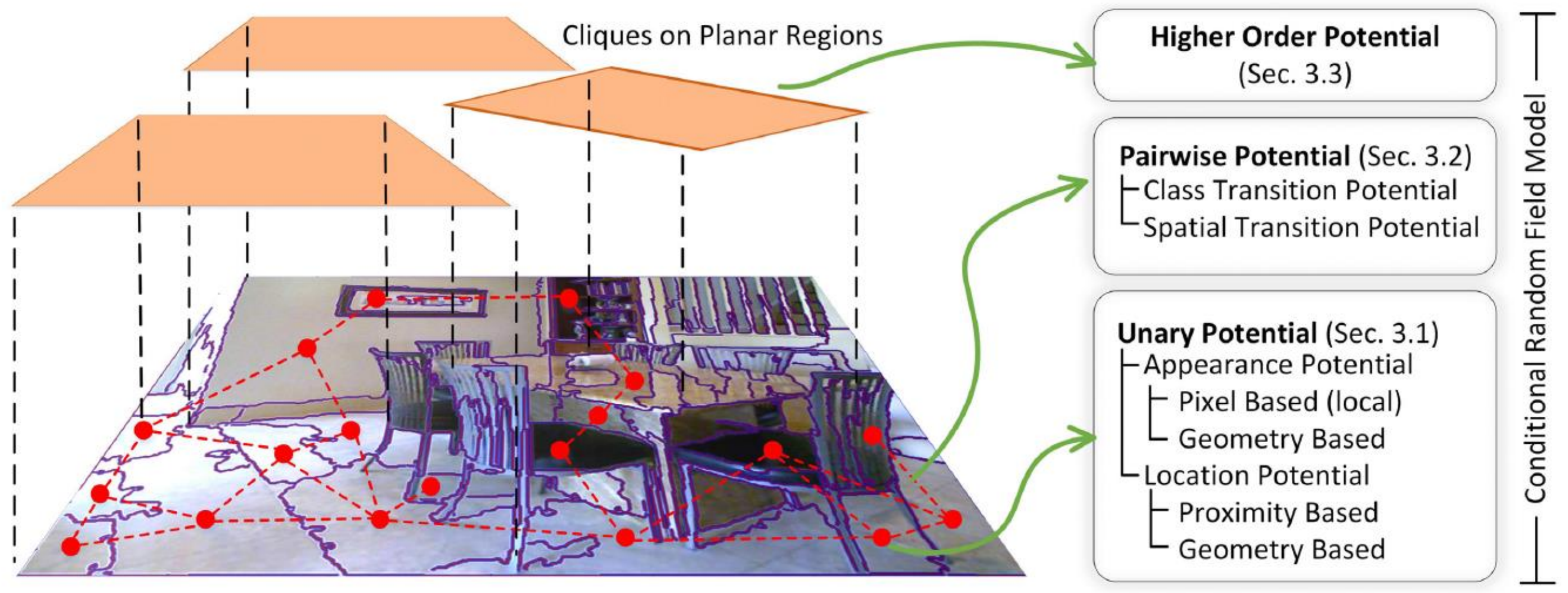

Cliques on Planar Regions
Higher Order Potential (Sec. 3.3)
Pairwise Potential (Sec. 3.2)
Class Transition Potential
Spatial Transition Potential
Unary Potential (Sec. 3.1)
Appearance Potential
Pixel Based (local)
Geometry Based
Location Potential
Proximity Based
Geometry Based
Conditional Random Field Model

Figure: Our approach combines geometrical information with low-level cues with in a CRF model. Only limited graph nodes are shown for the purpose of clear illustration.

- HOPs incorporate long range interactions and enhance the representational power of the CRF model.
- To apply the graph cuts algorithm for efficient approximate inference, we disintegration the HOP into sub-modular energies.
- Local appearance potential is modelled in a discriminative fashion using a Random Forest classifier.
- We use a genetic search algorithm to choose the most useful set of local appearance features on the validation set.

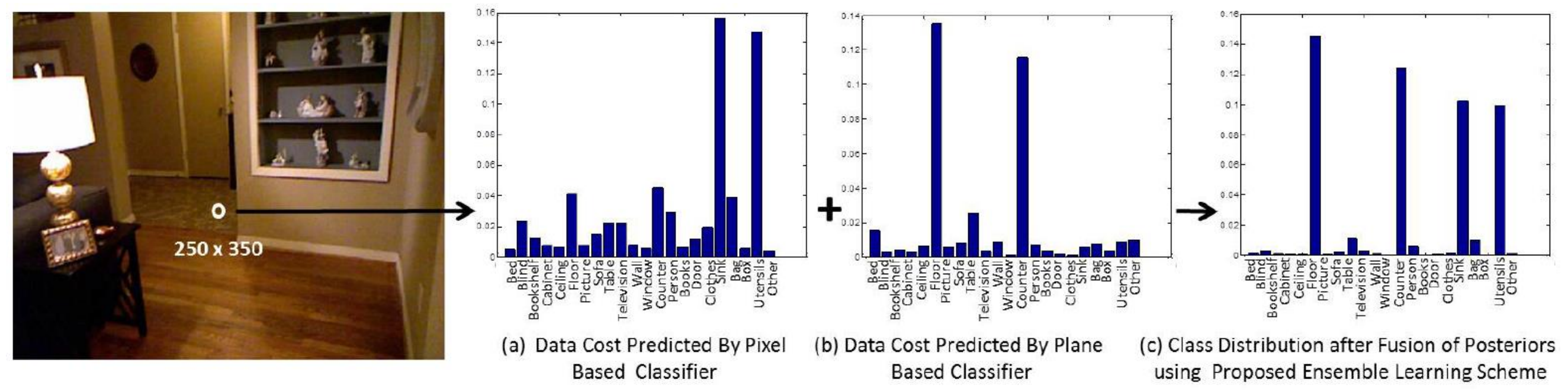


Figure: Effect of Ensemble Learning Scheme: At the pixel location shown in left most image, the pixel based appearance model favors class Sink. On the other hand, planar regions based appearance model takes care of geometrical properties of region and favors class Floor. The right most bar plot shows how our proposed ensemble learning scheme picks the correct class decision.

Learning CRF Model

1. Learning Potentials

Different characteristics of a semantic class (e.g., texture, shape, context, geometry and spatial location) are modelled using a rich set of features:

(a) Features for local appearance potential (LBP, texton, SIFT, color SIFT, depth SIFT, SPIN, HOG)

(b) Features for appearance model on planes (color histograms in the HSV and CIE LAB color spaces, textons, normalized area and height, normal orientation)

(c) Encoding the location potential by incorporating rough geometry as well as the location of semantic classes

2. Learning Parameters

- A structured large-margin learning method (S-SVM [36]) is used to efficiently adjust the probabilistic model parameters.
- We use a single slack formulation which results in a more efficient learning without any performance degradation
- A quadratic program is used to learn parameters of the boundary potentials to get a balanced representation of each edge in the Spatial Discontinuity Potential.

Results

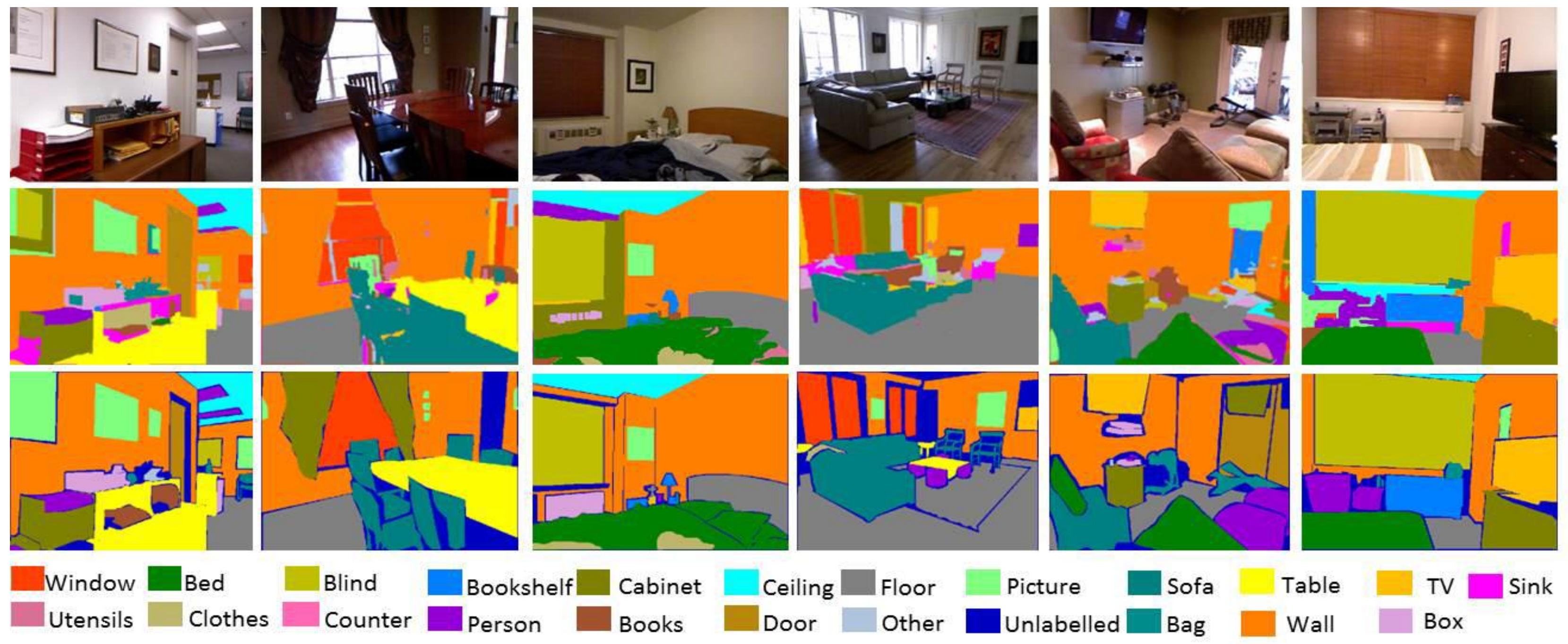


Figure: Examples of semantic labelling results on the NYU-Depth v2 dataset. Figure shows intensity images (top row), ground truths (bottom row) and our results (middle row). Our framework performs well in many cases including some unlabelled regions.

Method	Semantic Classes				Pixel	Class
	Floor	Structure	Furniture	Props	Accuracy	Accuracy
Supp. Inf. [35]	68	59	**70**	**42**	58.6	59.6
ConvNet [5]	68.1	87.8	51.1	29.9	63	59.2
ConvNet + D [3]	87.3	86.1	45.3	35.5	64.5	63.5
Im ∪ 3D [1]	**87.9**	79.7	63.8	27.1	67.0	64.3
This paper	87.1	**88.2**	54.7	32.6	**69.2**	**65.6**

Table: Comparison of results on the NYU-Depth v2 (4-class labeling task): Our method achieved best performance in terms of average pixel and class accuracies

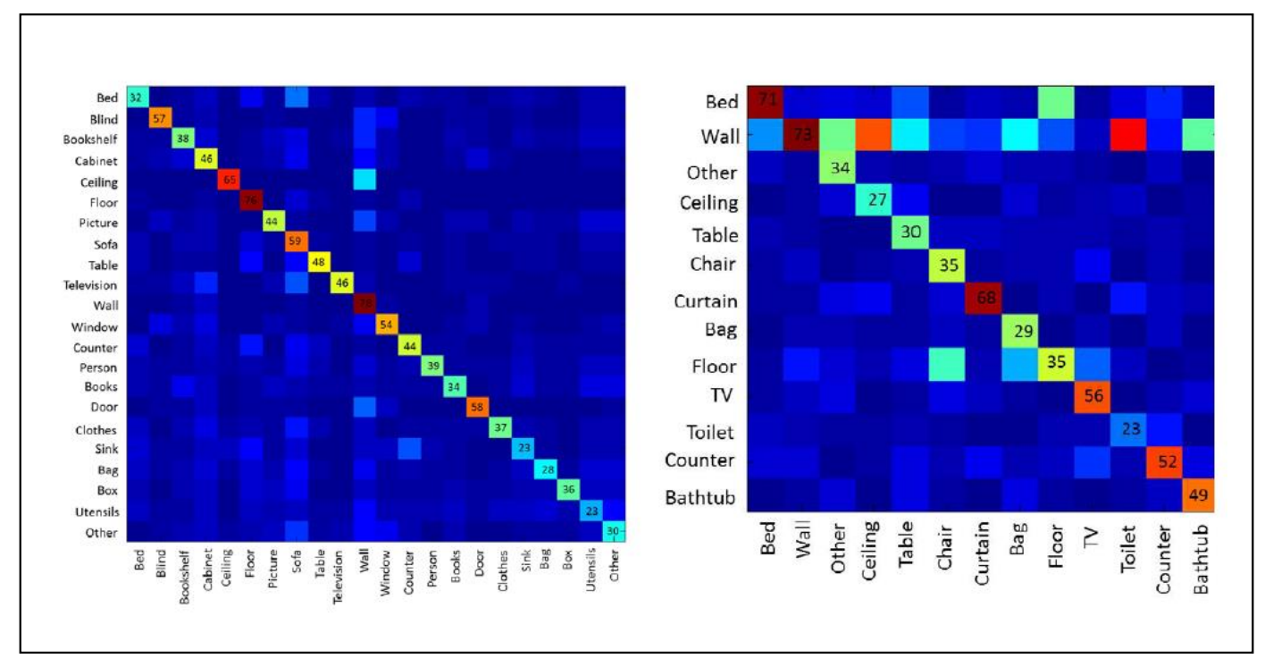
Method	Bed	Blind	Bookshelf	Cabinet	Ceiling	Floor	Picture	Sofa	Table	Television	Wall	Window	Counter	Person	Books	Door	Clothes	Sink	Bag	Box	Utensils	Other	Unlabeled	Mean Class Accuracy	Mean Pixel Accuracy	Classes
Class Freq.	4.7	2.0	4.2	10.7	1.4	10.8	2.2	6.2	2.6	0.5	22.8	2.3	2.7	1.7	0.9	2.3	1.7	0.3	1.7	0.8	0.2	0.1	17.4	-	-	-
ConvNet [5]	30.3	-	31.7	28.5	33.2	68.0	-	35.1	18.0	18.8	**89.4**	37.8	-	-	-	-	-	-	-	-	-	-	-	35.8	51.0	13
CNN+D [3]	**38.1**	-	13.7	42.4	62.6	**87.3**	-	29.8	10.2	6.0	86.1	15.9	-	-	-	-	-	-	-	-	-	-	-	36.2	52.4	13
This paper	32.3	**56.9**	**38.3**	**45.6**	**64.7**	75.8	**43.6**	**58.6**	**47.9**	**45.7**	77.5	**54.0**	**43.8**	**38.8**	**34.0**	**58.3**	**37.2**	**23.1**	**28.4**	**35.7**	**22.6**	**29.9**	-	**45.1**	**58.3**	22

Table: Class wise Accuracies on NYU-Depth v2: Our proposed framework achieves the highest accuracy on 19/22 classes. With nearly double number of classes used in [3, 5], we get ~ 6% and ~ 9% improvement in class and pixel accuracies respectively.

Table: Semantic Labeling Performance: We report the results of our proposed framework when only variants of unary potentials were used (top 3 rows), a CRF with regular Potts model was used (second last row) and the improvements observed when more sophisticated priors and HOPs (last row) were added. Accuracies are reported for 22 and 13 class semantic labeling for NYU v2 and SUN3D datasets respectively.

Variants of Our Method	NYU-Depth v2		SUN3D	
	Pixel Accuracy	Class Acc.	Pixel Accuracy	Class Acc.
Feature Ensemble (FE)	$44.4 \pm 15.8\%$	39.2%	$41.9 \pm 11.1\%$	40.0%
FE + Planar Appearance Model (PAM)	$52.5 \pm 15.5\%$	42.4%	$48.3 \pm 11.5\%$	42.6%
FE + PAM + Planar Location Prior (PLP)	$55.3 \pm 15.8\%$	43.1%	$51.5 \pm 11.9\%$	43.3%
FE + PAM + PLP + CRF (Regular Potts Model)	$55.5 \pm 15.8\%$	43.2%	$51.8 \pm 12.0\%$	43.5%
FE + PAM + PLP + CRF (SDP + HOP)	$\mathbf{58.3 \pm 15.9}\%$	**45.1%**	$\mathbf{54.2 \pm 12.2}\%$	**44.7%**

Table: Comparison of results on the NYU-Depth v2 (4-class labeling task): Our method achieved best performance in terms of average pixel and class accuracies



Plane Detection

Performance Evaluation		
Method	EPC Acc.	E+NPC Acc.
Silberman et al. [35]	0.69 ± 0.09	0.67 ± 0.10
Rabbani et al. [29]	0.60 ± 0.12	0.57 ± 0.14
This paper	$\mathbf{0.76 \pm 0.09}$	$\mathbf{0.81 \pm 0.07}$

Timing Comparison (averaged for NYU v2) (for Matlab prog. running on single core, thread)		
Silberman [35]	Rabbani [29]	This paper
41 sec	73 sec	3.1 sec

Table: Comparison of plane detection results on the NYU-Depth v2 dataset. We report detection accuracies for 'exactly planar classes' (EPC) and 'exact and nearly planar classes' (E+NPC).

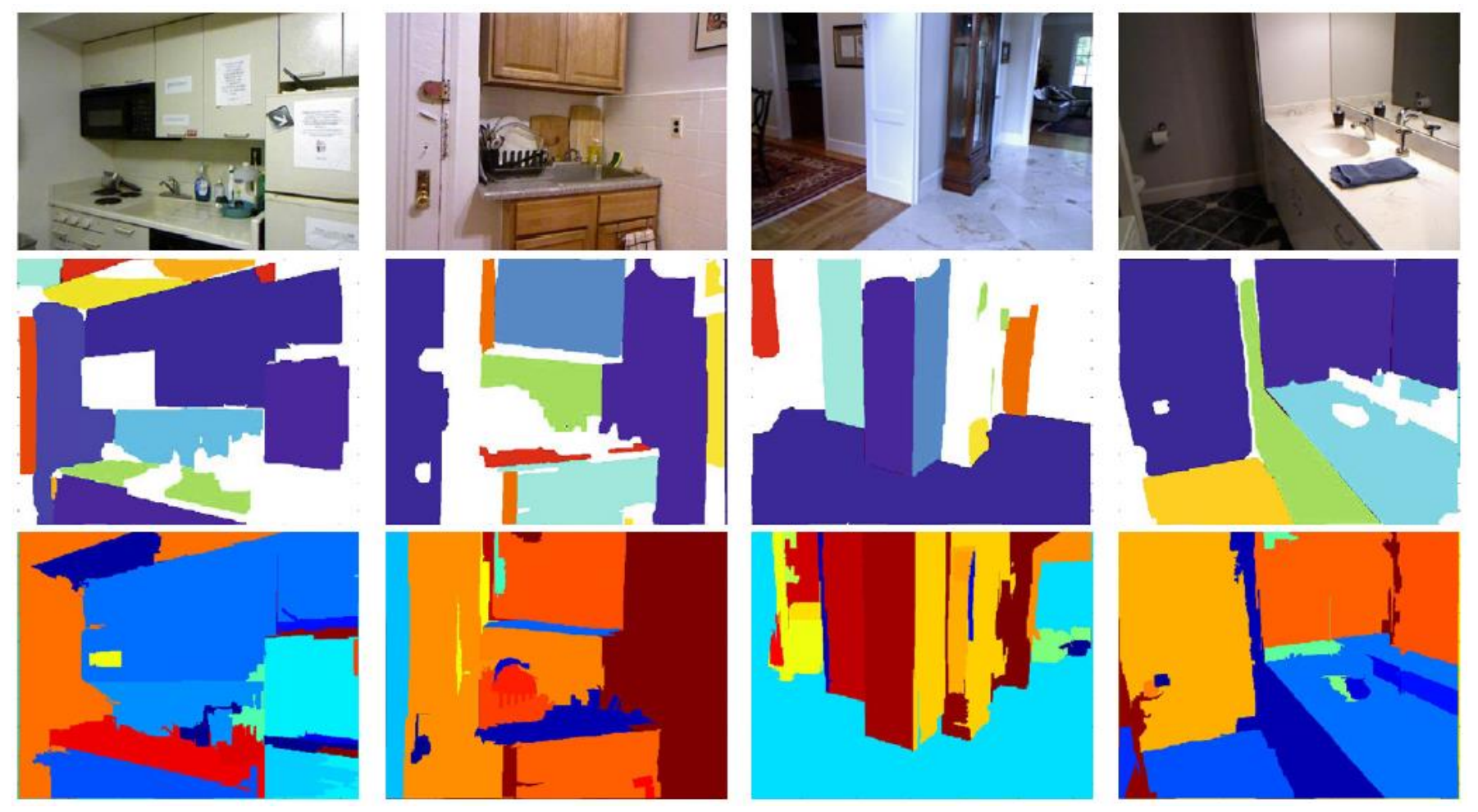


Figure: Comparison of our algorithm (last row) with Silberman et al. [35] (middle row) is shown. Note that the white color in middle row shows no detected planes.

The University of Western Australia
✉ salman.khan@research.uwa.edu.au