Deep Sliding Shapes for Amodal 3D Object Detection in RGB-D Images

PRINCETON VISION GROUP

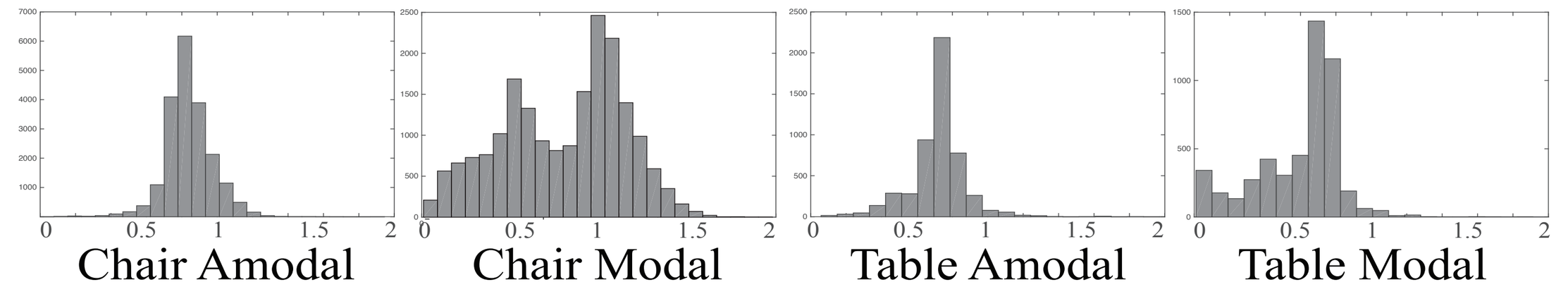
Shuran Song   Jianxiong Xiao

Introduction

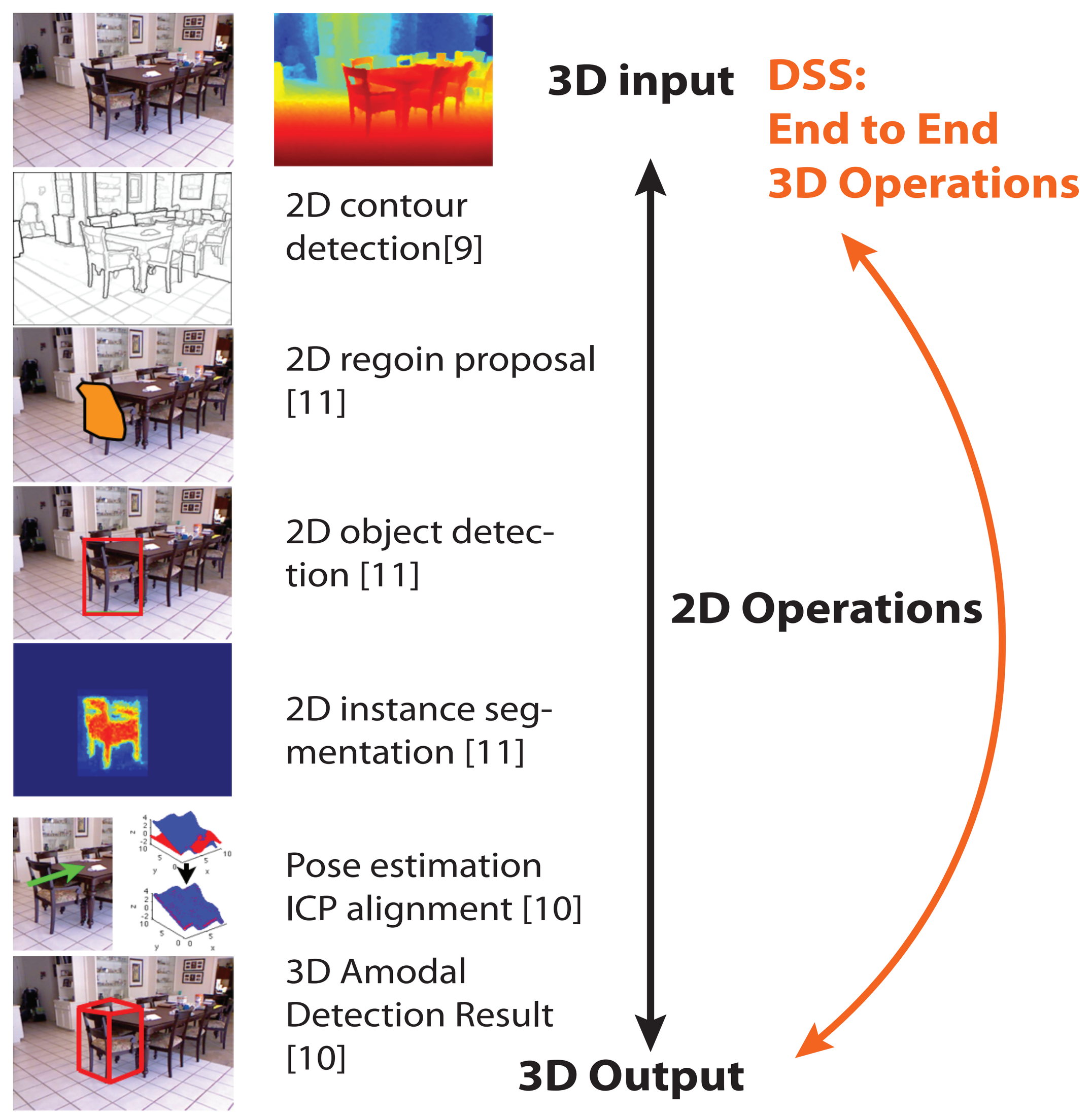
In this paper, we focus on the task of amodal 3D object detection in RGB-D images, which aims to produce an object's **3D bounding box** that gives **real-world dimensions** at the object's **full extent**, regardless of truncation or occlusion.

This kind of recognition is useful for many applications, for instance, in the perception-manipulation loop for robotics. But adding a new dimension for prediction significantly enlarges the search space, and makes the task much more challenging.

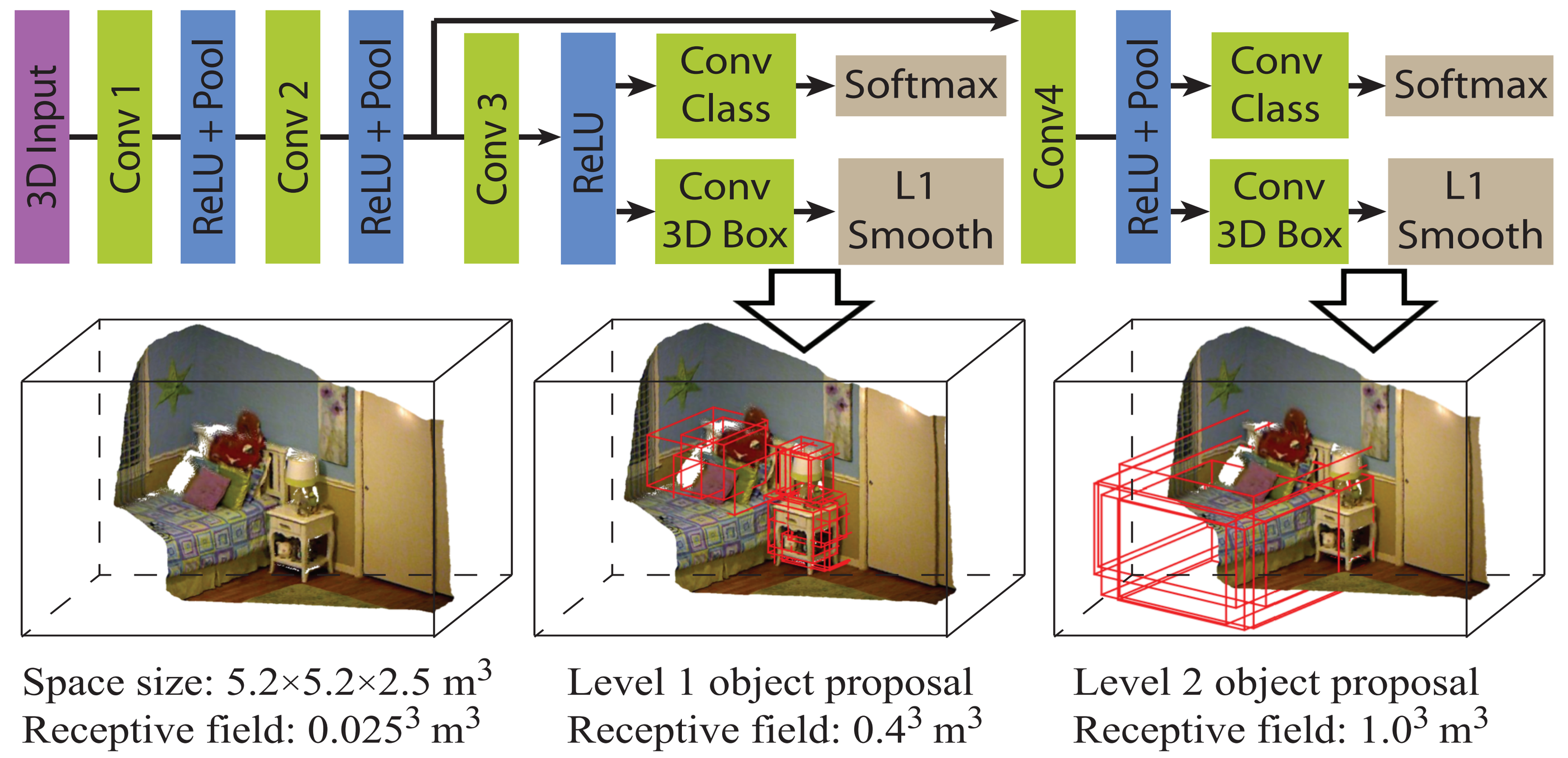
In this paper, we introduce Deep Sliding Shapes, a complete 3D formulation to learn object proposals and classifiers using 3D convolutional neural networks (ConvNets).



Previous Works

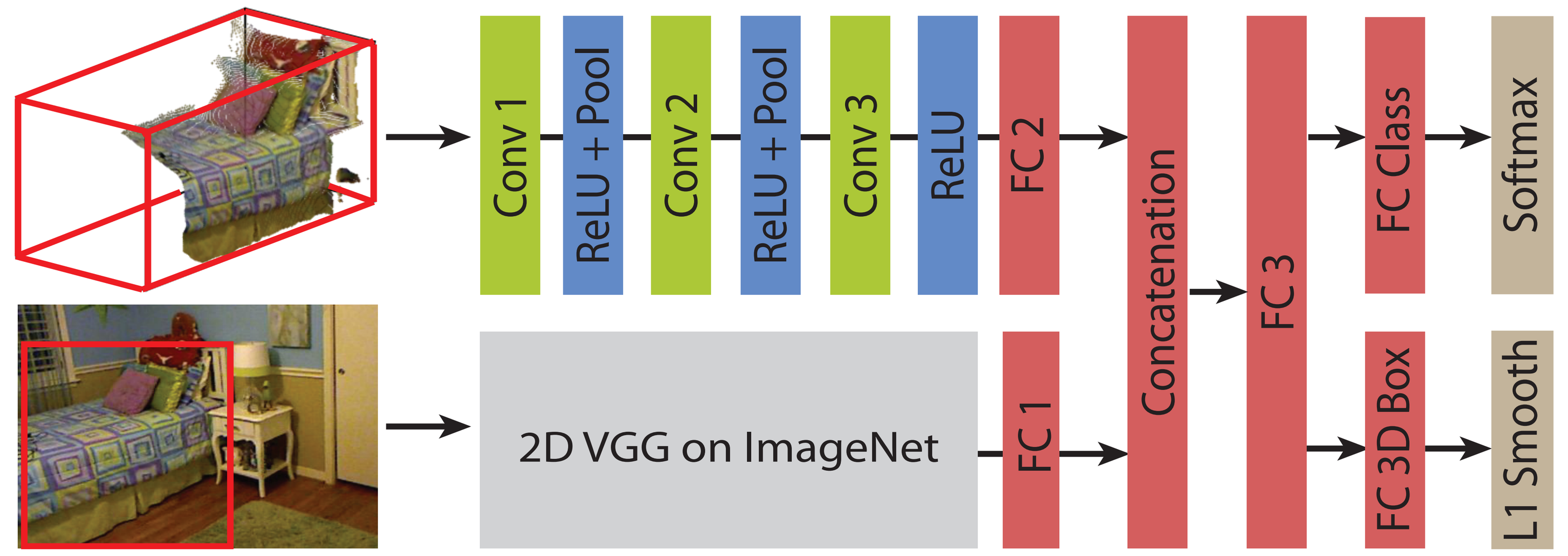


3D Amodal Region Proposal Network



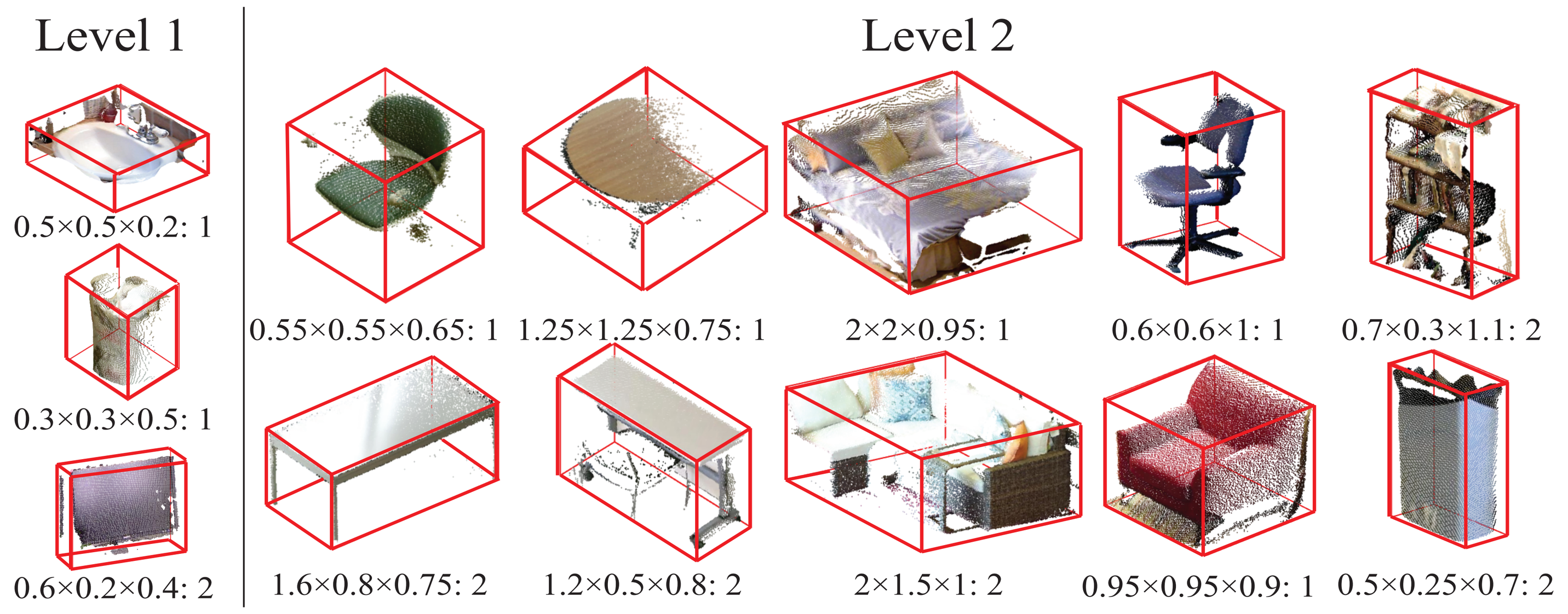
Taking a 3D volume from depth as input, our fully convolutional 3D network extracts 3D proposals at two scales with different receptive fields.

Joint Object Recognition Network



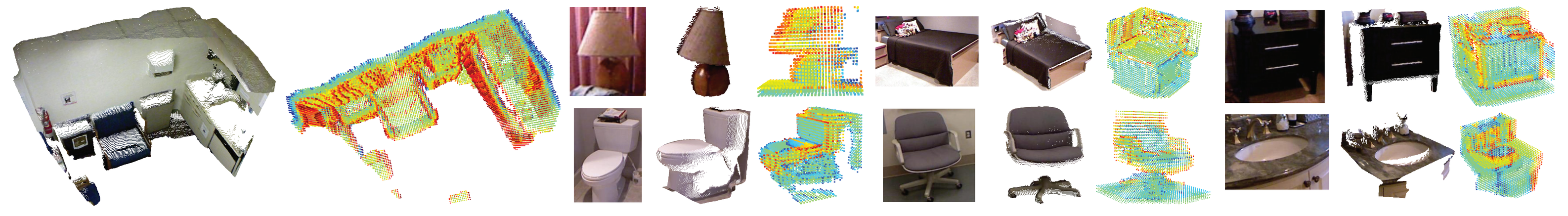
For each 3D proposal, we feed the 3D volume from depth to a 3D ConvNet, and feed the 2D color patch to a 2D ConvNet, to jointly learn object categories and 3D box regression.

List of All Anchors Types



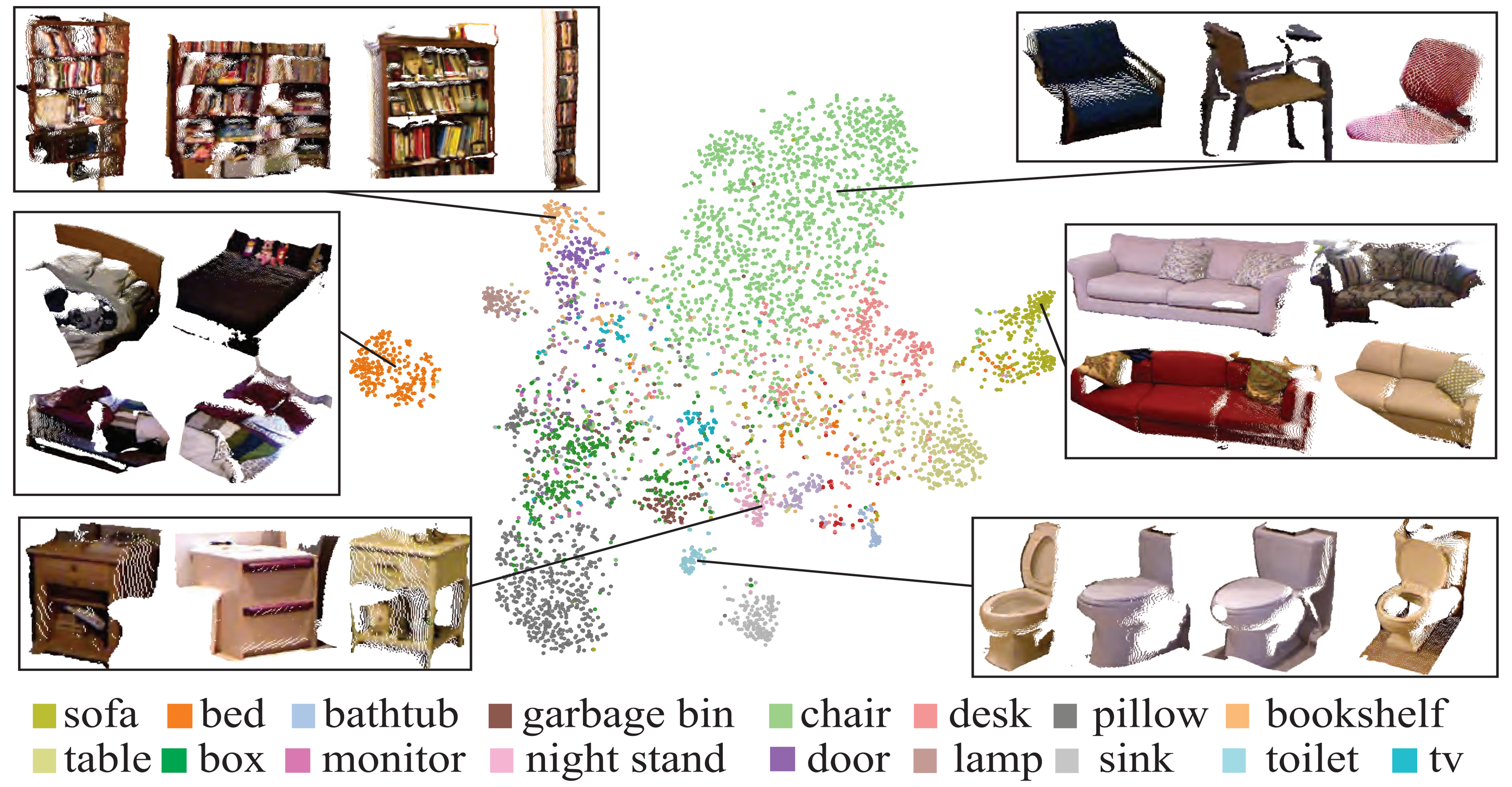
The subscripts show the width × depth × height in meters: number of orientations

Visualization of TSDF Encoding



We only visualize the TSDF values when close to the surface. Red indicates the voxel is in front of surfaces; and blue indicates the voxel is behind the surface. The resolution is 208×208×100 for the Region Proposal Network, and 30×30×30 for the Object Recognition Network

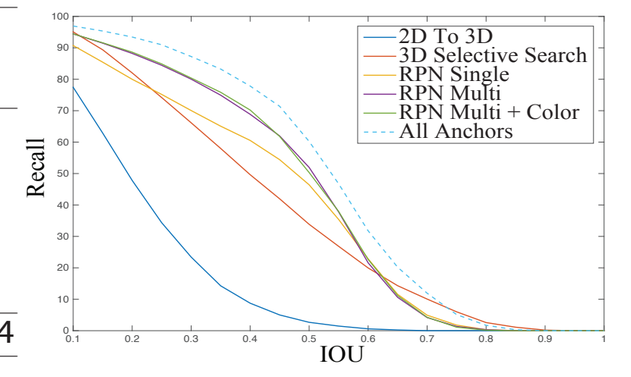
t-SNE Embedding



2D t-SNE embedding of the last layer features from the 3D ConvNet. Color encodes object category.

Evaluation

	bathtub	bed	bookshelf	box	chair	counter	desk	door	dresser	garbage bin	lamp	monitor	night stand	pillow	sink	sofa	table	tv	toilet	Recall	ABO	#Box
2D To 3D	41.7	53.5	37.9	22.0	26.9	46.2	42.2	11.8	47.3	33.9	41.8	12.5	45.8	20.7	49.4	55.8	54.1	15.2	50.0	34.4	0.210	2000
3D Selective Search	79.2	80.6	74.7	66.0	66.5	92.3	80.9	53.9	89.1	89.8	83.6	45.8	85.4	75.9	83.1	85.5	80.9	69.7	83.3	74.2	0.409	2000
RPN Single	87.5	98.7	70.1	15.6	95.0	100.0	93.0	20.6	94.5	49.2	49.1	12.5	100.0	34.2	81.8	94.9	93.3	57.6	96.7	75.2	0.425	2000
RPN Multi	100.0	98.7	73.6	42.6	94.7	100.0	92.5	21.6	96.4	78.0	69.1	37.5	100.0	75.2	97.4	97.1	96.4	66.7	100.0	84.4	0.460	2000
RPN Multi Color	100.0	98.1	72.4	42.6	95.0	100.0	93.0	19.6	96.4	79.7	76.4	37.5	100.0	79.0	97.4	97.1	95.4	57.6	100.0	84.9	0.461	2000
All Anchors	100.0	98.7	75.9	50.4	97.2	100.0	97.0	45.1	100.0	94.9	96.4	83.3	100.0	91.2	100.0	97.8	96.9	84.8	100.0	91.0	0.511	107674

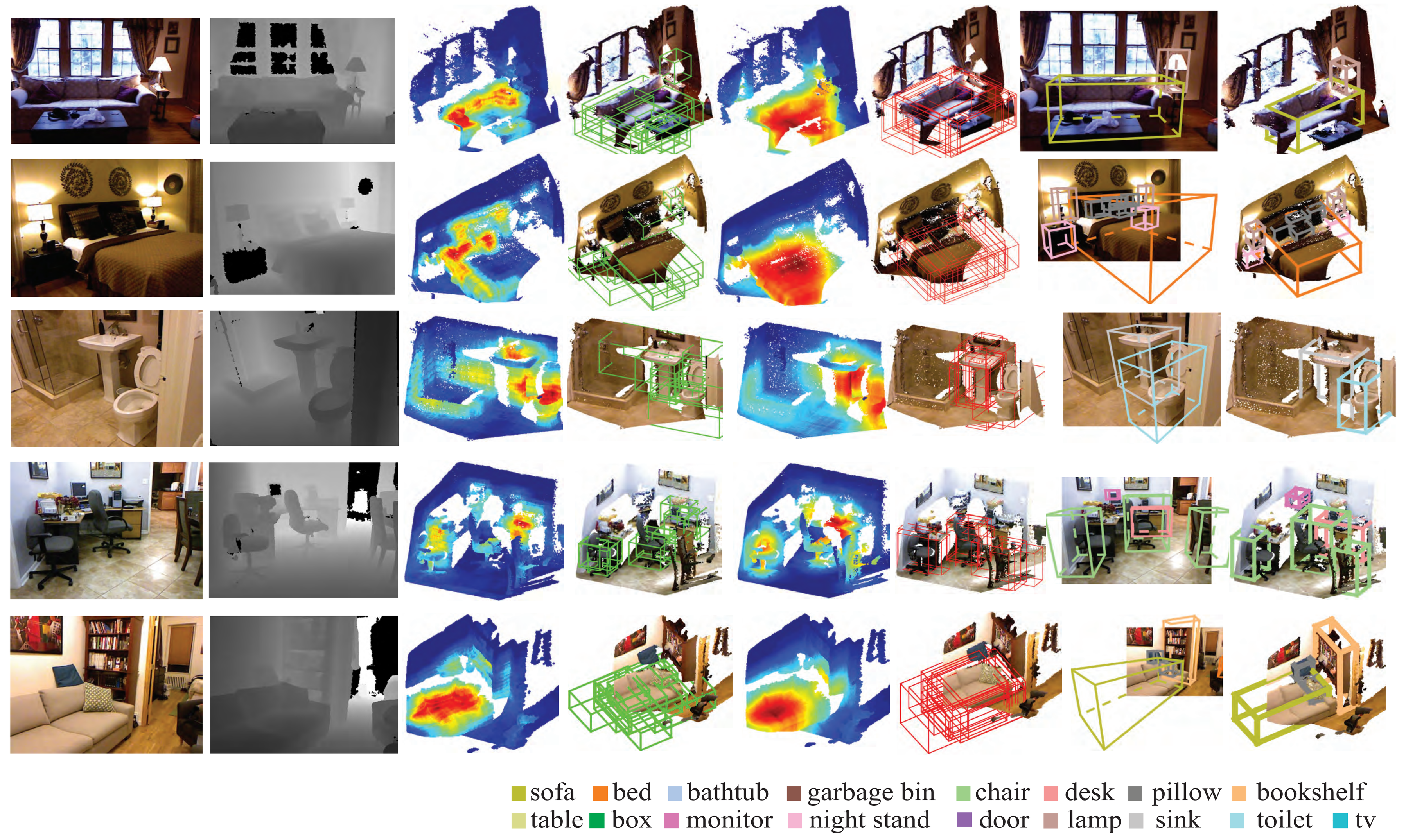


Evaluation for Amodal 3D Object Proposal on NYUv2

poposal	algorithm	bathtub	bed	bookshelf	chair	counter	desk	dresser	garbage bin	lamp	night stand	pillow	sink	sofa	table	toilet	box	door	monitor	tv	mAP
3D SS	dxdydz no bbreg	43.3	55.0	16.2	23.1	3.4	10.4	17.1	30.7	10.9	35.4	20.3	41.2	47.2	25.2	43.9	1.9	1.6	0.1	9.9	23.0
	dxdydz	52.1	60.5	19.0	30.9	2.2	15.4	23.1	36.4	19.7	36.2	18.9	52.5	53.7	32.7	56.9	1.9	0.5	0.3	8.1	27.4
RPN	dxdydz no bbreg	51.4	74.8	7.1	51.5	15.5	22.8	24.9	11.4	12.5	39.6	15.4	43.4	58.0	40.7	61.6	0.2	0.0	1.5	2.8	28.2
	dxdydz no size	59.9	78.9	12.0	51.5	15.6	24.6	27.7	12.5	18.6	42.3	15.1	59.4	59.6	44.7	62.5	0.3	0.0	1.1	12.9	31.5
	dxdydz	59.0	80.7	12.0	59.3	15.7	25.5	28.6	12.6	18.6	42.5	15.3	59.5	59.9	45.3	64.8	0.3	0.0	1.4	13.0	32.3
	tsdf dis	61.2	78.6	10.3	61.1	2.7	23.8	21.1	25.9	12.1	34.8	13.9	49.5	61.2	45.6	70.8	0.3	0.0	0.1	1.7	30.2
	dxdydz+rgb	58.3	79.3	9.9	57.2	8.3	27.0	22.7	4.8	18.8	46.5	14.4	51.6	56.7	45.3	65.1	0.2	0.0	4.2	0.9	30.1
	proj dxdydz+img	58.4	81.4	20.6	53.4	1.3	32.2	36.5	18.3	17.5	40.8	19.2	51.0	58.7	47.9	71.4	0.5	0.2	0.3	1.8	32.2
	dxdydz+img+hha	55.9	83.0	18.8	63.0	17.0	33.4	43.0	33.8	16.5	54.7	22.6	53.5	58.0	49.7	75.0	2.6	0.0	1.6	6.2	36.2
	dxdydz+img	62.8	82.5	20.1	60.1	11.9	29.2	38.6	31.4	23.7	49.6	21.9	58.5	60.3	49.7	76.1	4.2	0.0	0.5	9.7	36.4

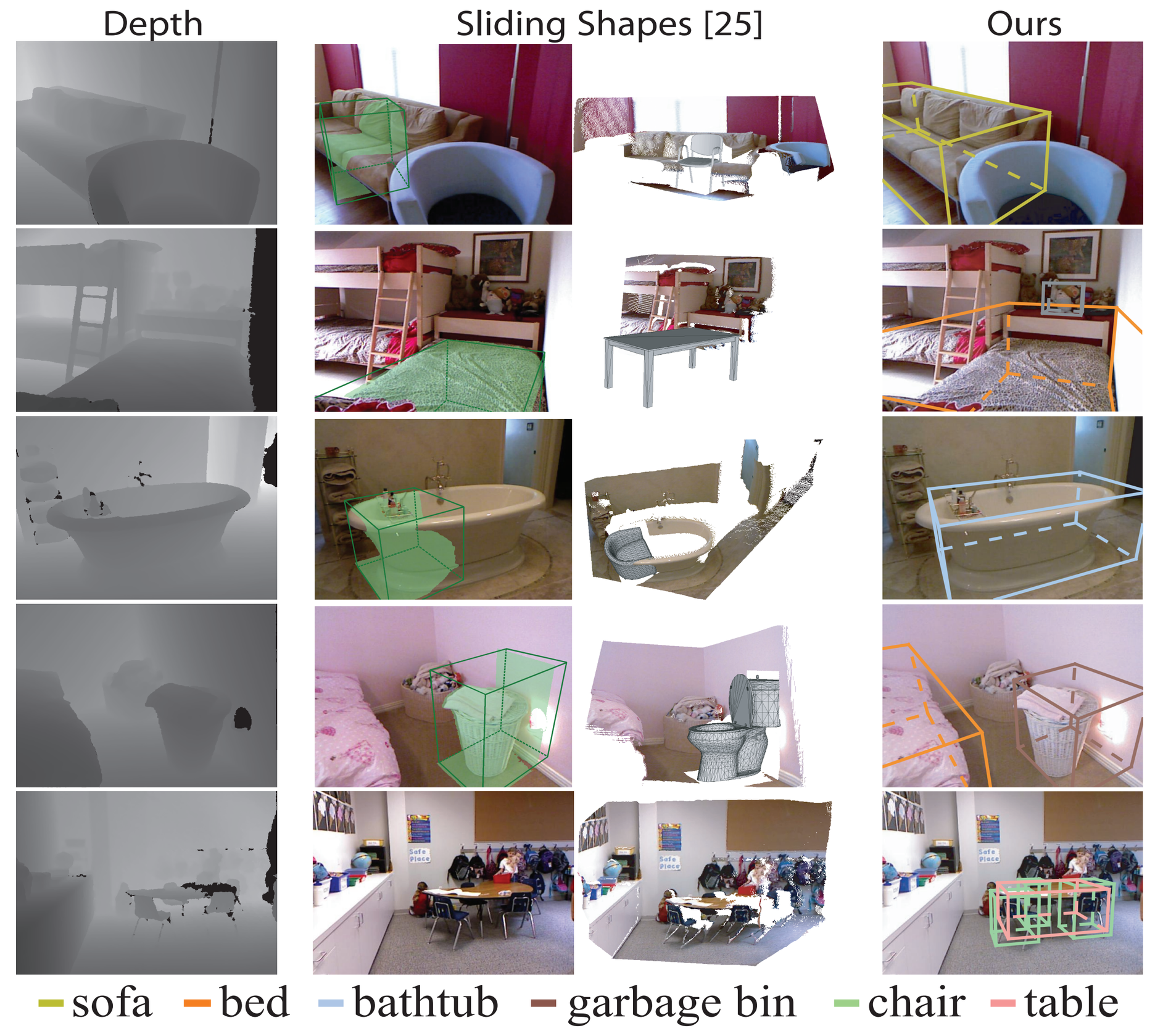
3D Amodal Object Detection on NYUv2

Results



Examples for Detection Results. For the proposal results, we show the heat map for the distribution of the top proposals (red is the area with more concentration), and a few top boxes after NMS. For the recognition results, our amodal 3D detection can estimate the full extent of 3D both vertically and horizontally.

Comparison



Comparison with Sliding Shapes [25]. Deep Sliding Shapes is able to better use shape, color, and contextual information to handle more object categories, resolve ambiguous cases, and detect objects with atypical sizes.

Algorithm	input	bed	chair	table	sofa	toilet	mAP
Sliding Shapes [25]	d	33.5	29	34.5	33.8	67.3	39.6
[10] on instance seg	d	71.0	18.2	30.4	49.6	63.4	46.5
[10] on instance seg	rgbd	74.7	18.6	28.6	50.3	69.7	48.4
[10] on estimated model	d	72.7	47.5	40.6	54.6	72.7	57.6
[10] on estimated model	rgbd	73.4	44.2	33.4	57.2	84.5	58.5
ours [depth only]	d	83.0	58.8	68.6	49.5	79.2	67.8
ours [depth + img]	rgbd	84.7	61.1	70.5	55.4	89.9	72.3

Reference

[25] S. Song and J. Xiao. Sliding Shapes for 3D object detection in depth images. In ECCV, 2014

[10] S.Gupta,P.A.Arbelaez, R.B.Girshick, andJ. Malik. Aligning 3D models to RGB-D images of cluttered scenes. In CVPR, 2015.

[9] S. Gupta, P. Arbelaez, and J. Malik. Perceptual organization and recognition of indoor scenes from RGB-D images. In CVPR, 2013.

[11] S. Gupta, R. Girshick, P. Arbelaez, and J. Malik. Learning rich features from RGB-D images for object detection and segmentation. In ECCV,