Appendix of ImVoxelNet

Danila Rukhovich[1,2], Anna Vorontsova[1], Anton Konushin[1,2]
[1]Samsung AI Center Moscow; [2]Lomonosov Moscow State University
{d.rukhovich, a.vorontsova, a.konushin}@samsung.com

Abstract

In A, we report the results of the extensive evaluation of the proposed method on the SUN RGB-D dataset. In Section B we visualize predicted bounding boxes for several samples taken from all four datasets we use in our experiments.

A. More results on SUN RGB-D

For a comprehensive comparison, we also mention PerspectiveNet [4], which is evaluated following a different protocol. In that protocol, the annotations are mapped into 30 object categories. Accordingly, we train ImVoxelNet using the same object categories. The results are reported in Tab. 3. Among these 30 categories, 10 object categories are consistent with 10 categories used in [3, 2, 5]. So, we can merge these benchmarks and report metrics for [3, 2, 5, 4] that are obtained on the same subset of 10 object categories 2. Following [4], we assume camera poses are known, so we optimize only L_{indoor} and do not use any additional camera pose loss.

Another SUN RGB-D benchmark has been proposed in [7] for point cloud-based methods evaluation. This benchmark implies detecting objects of 10 categories with mAP@0.25 chosen as the main metric. In Tab. 1, we report the results of our method against point cloud-based methods. This comparison is unfair, favoring point cloud-based methods since they have access to more complete data. Nevertheless, we report the metrics to establish a baseline for monocular 3D object detection on SUN RGB-D.

Comparison with Total3DUnderstanding [5] on all NYU-37 object categories is present in Tab. 4. In this experiment, we optimize $L_{\mathsf{indoor}} + L_{\mathsf{extra}}$ since camera pose is assumed unknown.

B. Visualization

All visualized images belong to validation subsets of the corresponding datasets. Different colors of the depicted bounding boxes mark different object categories; the color encoding is consistent within each dataset.

Method	RGB	PC	bath	bed	bkshf	chair	desk	dresser	nstand	sofa	table	toilet	mAP
F-PointNet[6]	✓	✓	43.3	81.1	33.3	64.2	24.7	32.0	58.1	61.1	51.1	**90.9**	54.0
VoteNet[7]	✗	✓	74.4	83.0	28.8	75.3	22.0	29.8	62.2	64.0	47.3	90.1	57.7
H3DNet[9]	✗	✓	73.8	85.6	31.0	**76.7**	**29.6**	33.4	65.5	66.5	50.8	88.2	60.1
ImVoteNet[8]	✓	✓	**75.9**	**87.6**	**41.3**	**76.7**	28.7	**41.4**	**69.9**	**70.7**	**51.1**	90.5	**63.4**
ImVoxelNet	✓	✗	71.7	69.6	5.7	53.7	21.9	21.2	34.6	51.5	39.1	76.8	40.7

Table 1. AP@0.25 scores for 10 object categories [7] from the SUN RGB-D dataset. All methods but ImVoxelNet use point cloud (PC) as an input.

Method	bed	chair	sofa	table	desk	toilet	bin	sink	shelf	lamp	mAP
3DGP[1]	5.62	2.31	3.24	1.23	–	–	–	–	–	–	–
HoPR[3]	58.29	13.56	28.37	12.12	4.79	16.50	0.63	2.18	1.29	2.41	14.01
CooP[2]	63.58	17.12	41.22	26.21	9.55	58.55	10.19	5.34	3.01	1.75	23.65
PerspectiveNet[4]	**79.69**	40.42	62.35	44.12	20.19	81.22	22.42	**41.35**	8.29	13.14	39.09
ImVoxelNet	77.87	**65.94**	**63.89**	**51.17**	**31.91**	**84.53**	**33.35**	39.91	**21.65**	**17.19**	**48.74**

Table 2. AP@0.15 scores for 10 out of 30 object categories [4] from the SUN RGB-D dataset.

Method	toilet	recycle bin	night stand	end table	drawer	computer	key board	table	chair	monitor	stool
PerspectiveNet[4]	81.22	37.68	35.16	19.77	1.28	1.24	**2.86**	44.12	40.42	1.14	**22.65**
ImVoxelNet	**84.53**	**52.20**	**46.29**	**25.31**	**6.05**	**2.71**	0.01	**51.17**	**65.94**	**19.82**	10.37

Method	lamp	dresser	picture	garbage bin	shelf	sofa chair	cabinet	sink	desk	book shelf	coffee table
PerspectiveNet[4]	13.14	**27.38**	0.00	22.42	0.97	51.86	1.70	**41.35**	20.19	8.29	28.80
ImVoxelNet	**17.19**	22.32	**0.82**	**33.35**	**4.00**	**54.61**	**7.90**	39.91	**31.91**	**21.65**	**36.48**

Method	box	sofa	white board	bed	pillow	paper	painting	cpu
PerspectiveNet[4]	1.64	62.35	0.02	**79.69**	11.36	0.00	0.17	**21.60**
ImVoxelNet	**3.29**	**63.89**	**0.95**	77.87	**14.65**	0.00	**0.53**	5.30

Table 3. AP@0.15 scores for 30 object categories [4] from the SUN RGB-D dataset.

Method	cabinet	bed	chair	sofa	table	door	window	book shelf	picture	counter	blinds
CooP[2]	10.47	57.71	15.21	36.67	31.16	0.14	0.00	3.81	0.00	27.67	**2.27**
T3DU[5]	11.39	59.03	15.98	43.95	35.28	0.36	0.16	5.26	**0.24**	**33.51**	0.00
ImVoxelNet	**19.24**	**79.17**	**63.07**	**60.59**	**51.14**	**0.74**	**0.18**	**16.37**	0.14	14.89	0.26

Method	desk	shelves	curtain	dresser	pillow	mirror	floor mat	clothes	books	fridge	tv
CooP[2]	19.90	2.96	1.35	15.98	2.53	**0.47**	–	0.00	**3.19**	21.50	5.20
T3DU[5]	23.65	4.96	2.68	19.20	2.99	0.19	–	0.00	1.30	20.68	4.44
ImVoxelNet	**31.20**	**5.47**	**3.34**	**35.45**	**11.01**	0.22	–	**1.40**	0.13	**23.28**	**12.41**

Method	paper	towel	shower curtain	box	white board	person	night stand	toilet	sink	lamp	bathtub
CooP[2]	0.20	2.14	**20.00**	2.59	0.16	20.96	11.36	42.53	15.95	3.28	24.71
T3DU[5]	**0.41**	**2.20**	**20.00**	2.25	0.43	23.36	6.87	48.37	14.40	3.46	27.85
ImVoxelNet	0.00	1.92	0.00	**2.71**	**1.17**	**42.02**	**38.38**	**77.28**	**45.12**	**13.27**	**43.59**

Method	bag	wall	floor	ceiling
CooP[2]	1.53	–	–	–
T3DU[5]	**2.27**	–	–	–
ImVoxelNet	0.53	–	–	–

Table 4. AP@0.15 scores for 37 object categories [5] from the SUN RGB-D dataset.

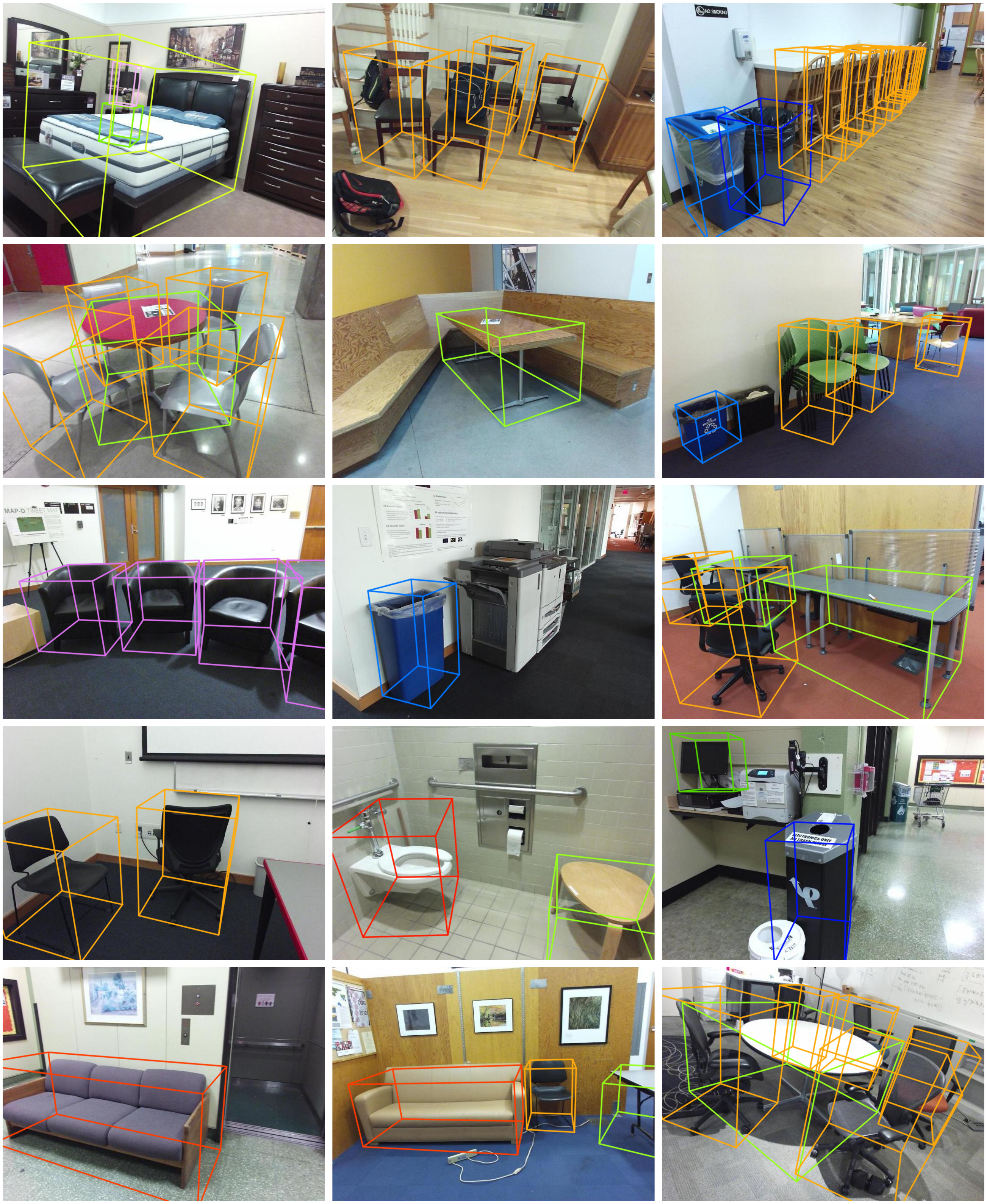
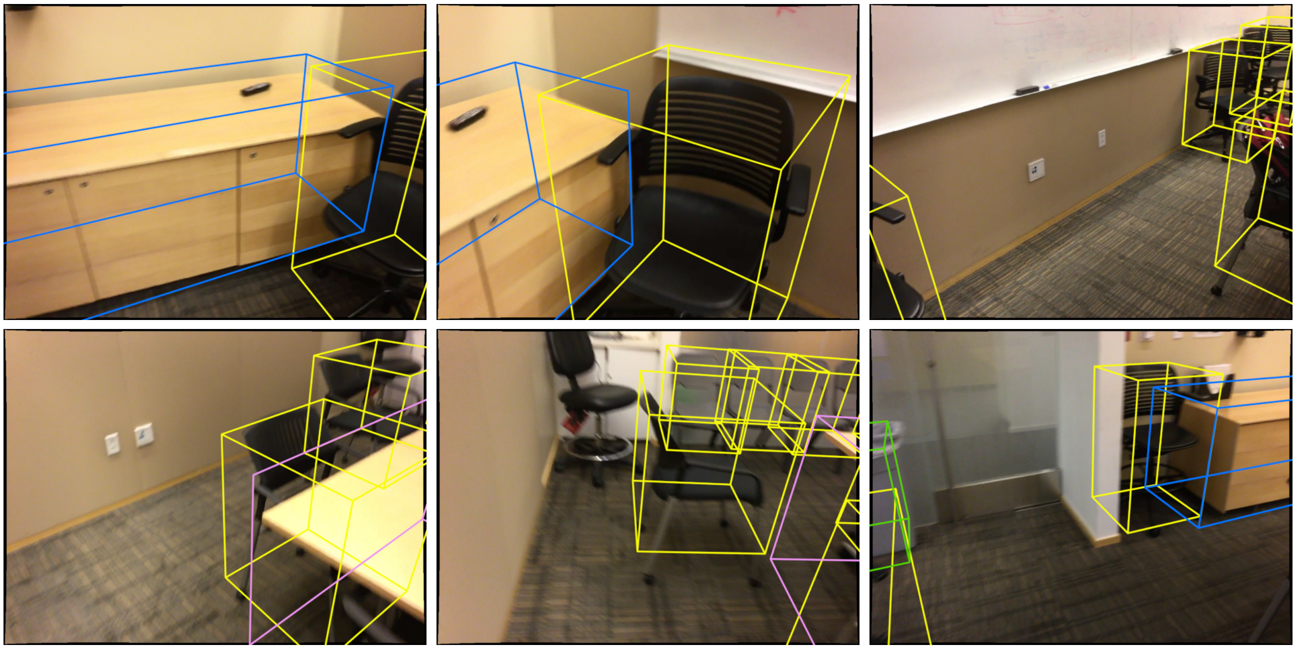
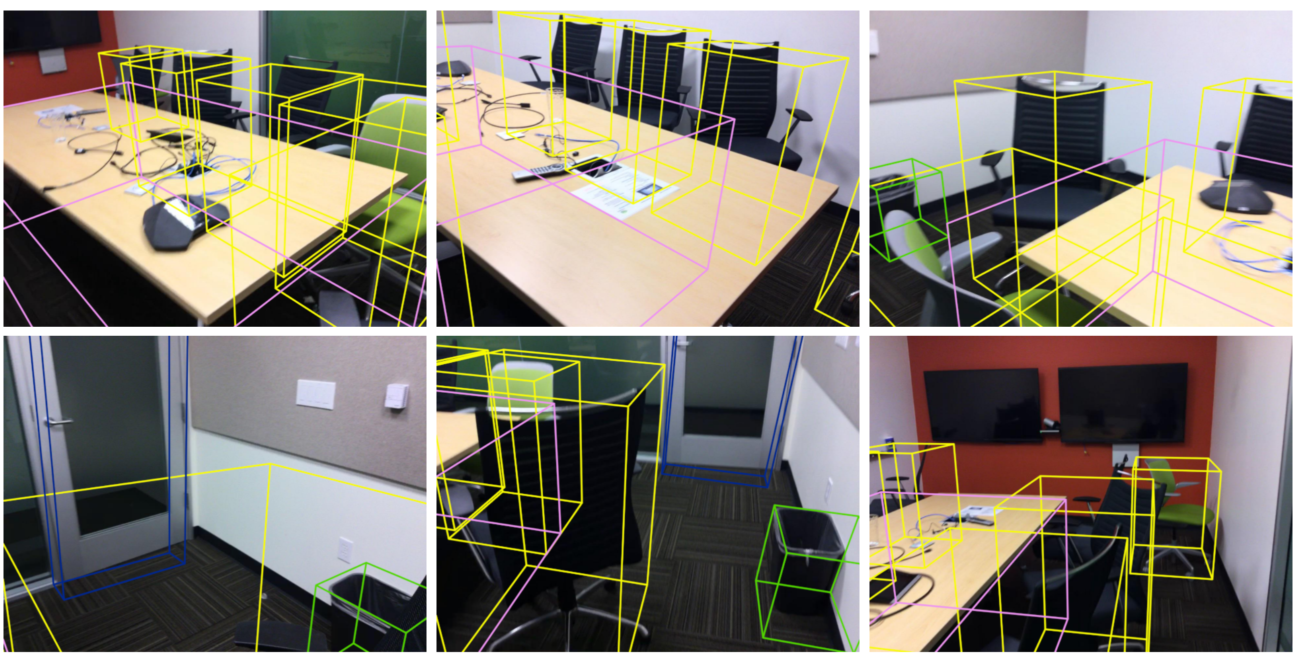


Figure 1. Objects detected on the monocular images from the validation subset of the SUN RGB-D dataset.



a) Scene *0169_00*.

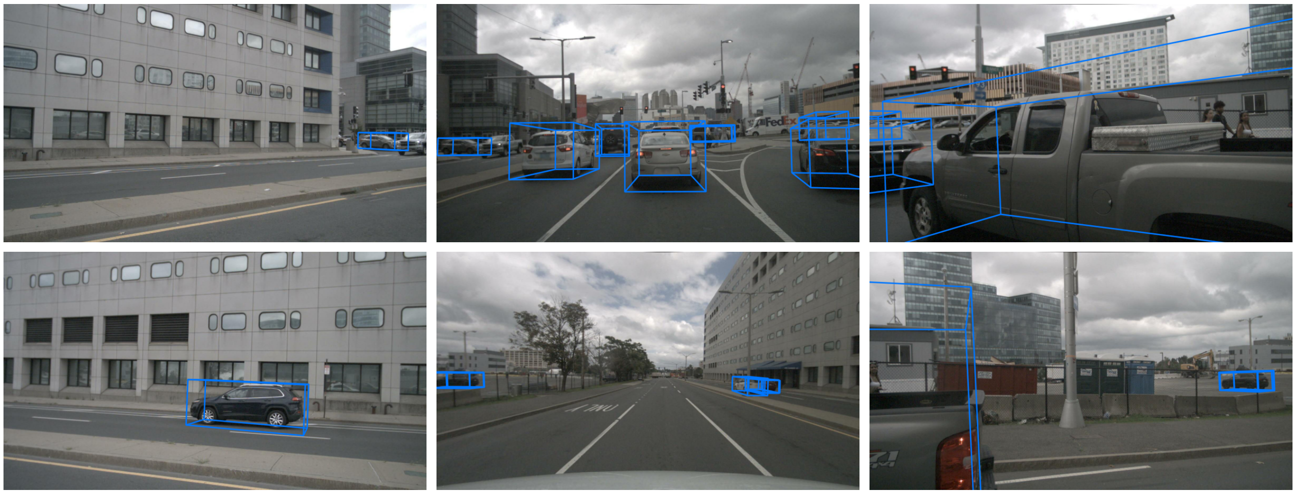


b) Scene *0575_00*.

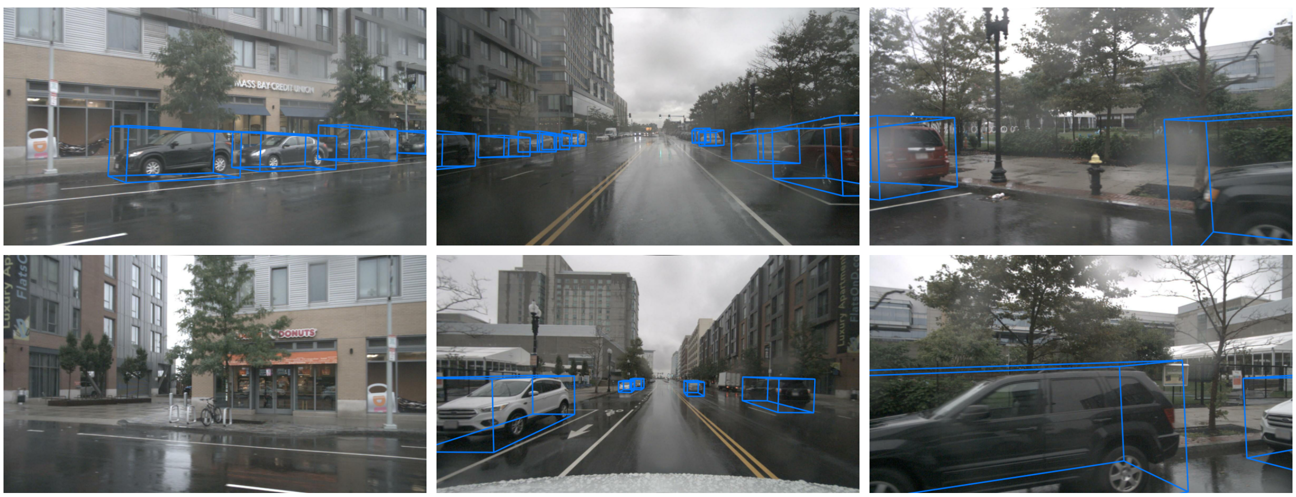
Figure 2. Objects detected on the multi-view inputs from the validation subset of the ScanNet dataset.

Figure 3. Cars detected on the monocular images from the validation subset of the KITTI dataset.

.



a) Scene *n008-2018-08-01-15-16-36-0400__15331512526*.



b) Scene *n008-2018-09-18-15-12-01-0400__15372981046*.

Figure 4. Cars detected in the images of two scenes from the validation subset of the nuScenes dataset. The predictions were obtained in multi-view settings. The first two rows correspond to the first scene, and the last two rows correspond to another one. For each scene, the upper row consists of images taken with a front-left, front, and front-right camera (from left to right). The second row contains images taken with a back-left, back, and back-right camera, respectively.

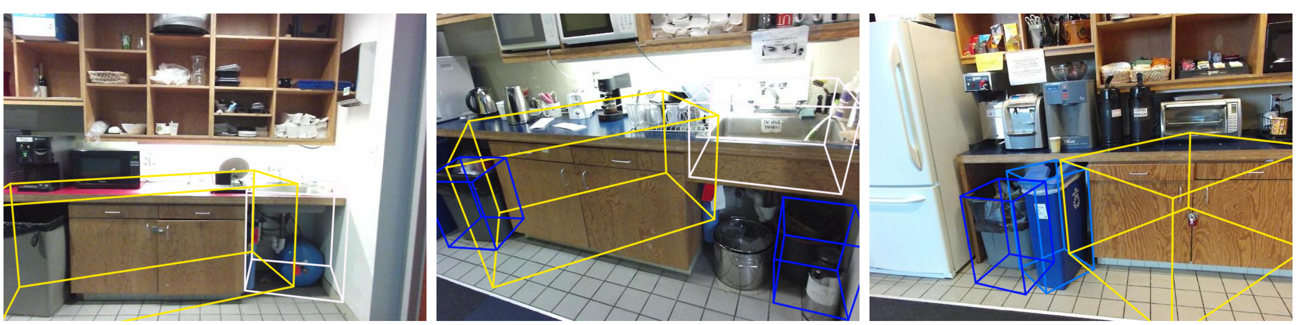


Figure 5. Examples of the detection failures for images from the validation subset of the SUN RGB-D dataset. These examples depict typical error cases: small objects of *sink*, *garbage bin* and *recycle bin* categories are detected quite precisely, but rotation angles for large object such as *cabinet* are estimated poorly.

References

[1] W. Choi, Y.-W. Chao, C. Pantofaru, and S. Savarese. Understanding indoor scenes using 3d geometric phrases. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 33–40, 2013.

[2] S. Huang, S. Qi, Y. Xiao, Y. Zhu, Y. N. Wu, and S.-C. Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. *arXiv preprint arXiv:1810.13049*, 2018.

[3] S. Huang, S. Qi, Y. Zhu, Y. Xiao, Y. Xu, and S.-C. Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 187–203, 2018.

[4] S. Huang, Y. Chen, T. Yuan, S. Qi, Y. Zhu, and S.-C. Zhu. Perspectivenet: 3d object detection from a single rgb image via perspective points. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] Y. Nie, X. Han, S. Guo, Y. Zheng, J. Chang, and J. J. Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020.

[6] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[7] C. R. Qi, O. Litany, K. He, and L. J. Guibas. Deep hough voting for 3d object detection in point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019.

[8] C. R. Qi, X. Chen, O. Litany, and L. J. Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4404–4413, 2020.

[9] Z. Zhang, B. Sun, H. Yang, and Q. Huang. H3dnet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329. Springer, 2020.