Boosting Multi-View Indoor 3D Object Detection via Adaptive 3D Volume Construction Supplementary Materials

Runmin Zhang[1] Zhu Yu[1*] Si-Yuan Cao[2,3*] Lingyu Zhu[4]
Guangyi Zhang[1] Xiaokai Bai[1] Hui-Liang Shen[1]
[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]Ningbo Global Innovation Center, Zhejiang University [3]NingboTech University [4]City University of Hong Kong
{runmin_zhang, yu_zhu, cao_siyuan}@zju.edu.cn, lingyzhu-c@my.cityu.edu.hk,
{zhangguangyi, shawnnnkb, shenhl}@zju.edu.cn

1. More Quantitative Results

Discussion about the DepthNet. We note that the depth quality has a significant impact on the detection performance, and our current DepthNet represents a trade-off between accuracy and efficiency. The generalization ability can be enhanced by incorporating advanced depth estimation methods. We use the relative depth from Depth Anything v2 [6] as an additional input to the monocular branch of DepthNet, with results reported in Table 1. Stronger depth networks can further improve detection performance, and our framework is flexible to support such upgrades.

Performance under novel light conditions. To explore the applicability of SGCDet under more challenging conditions, we conduct experiments under novel lighting variations. We simulate such conditions by applying random brightness and contrast adjustments, and random Gaussian light spot overlays, as illustrated in Fig. 1. As shown in Table 2, SGCDet remains robust under these conditions.

Per-category performance. We report the per-category AP@0.25 and AP@0.50 scores on the ScanNet [2] and ARKitScenes [1] datasets in Tables 3, 4, 5, and 6. These results demonstrate that SGCDet consistently outperforms existing methods across most categories.

2. More Qualitative Results

We provide additional qualitative results on the ScanNet and ARKitScenes datasets in Fig. 2 and Fig. 3, respectively. Compared to ImGeoNet [3], which relies on ground-truth scene geometry for supervision, our SGCDet detects more target objects and achieves more accurate object classification results.

*Corresponding authors.

Table 1. Comparison of DepthNet on the ScanNet dataset.

Performance	SGCDet	+ Depth Any. v2-Small	+ Depth Any. v2-Base
mAP@0.25	61.2	62.3 (↑ +1.1)	62.6 (↑ +1.4)
mAP@0.50	35.2	37.1 (↑ +1.9)	37.4 (↑ +2.2)

Figure 1. Visualization of the synthetic novel light condition.

Table 2. Comparison of light conditions on the ScanNet dataset.

Light condition	Normal light condition		Synthetic novel light condition	
	mAP@0.25	mAP@0.50	mAP@0.25	mAP@0.50
SGCDet	61.2	35.2	61.0	34.9

3. Visualization of the Sparse Volume Construction

We present the visualization of two refinement stages of our sparse volume construction in Fig. 4. The voxel resolutions at the two stages are $20 \times 20 \times 8$ and $40 \times 40 \times 16$, respectively. As observed, the pseudo labels roughly indicate the occupancy of 3D scenes, providing flexible supervision for the occupancy estimation module. Benefiting from this design, our network effectively identifies and selects regions likely to contain objects for volume feature refinement, thereby avoiding redundant computations in free space.

4. Limitations

While SGCDet achieves superior performance compared to previous approaches, its accuracy on certain categories

(*e.g.*, TV monitor, picture) remains sub-optimal. Additionally, the perception range of 3D volumes is constrained by the predefined voxel size and resolution, which may not be suitable for all 3D scenes. As illustrated in Fig. 4 (Scene 3), the fixed 3D volume fails to cover objects located at both sides. Future work could explore methods to dynamically adjust the perception range of 3D volumes based on the input images. Despite these limitations, we believe SGCDet contributes to advancing indoor 3D perception.

References

[1] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. ARKitScenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021. 1

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 1

[3] Tao Tu, Shun-Po Chuang, Yu-Lun Liu, Cheng Sun, Ke Zhang, Donna Roy, Cheng-Hao Kuo, and Min Sun. ImGeoNet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6996–7007, 2023. 1

[4] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. NeRF-Det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23320–23330, 2023. 3

[5] Yating Xu, Chen Li, and Gim Hee Lee. MVSDet: Multi-view indoor 3d object detection via efficient plane sweeps. In *Advances in Neural Information Processing Systems*, 2024. 3

[6] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, 2024. 1

Table 3. Per-category AP@0.25 scores on the ScanNet dataset. The best results are in **bold**.

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn
NeRF-Det [4]	42.3	**84.6**	75.9	78.5	56.3	33.4	21.4	49.9	2.4	50.6	73.9	21.3	64.3	62.5	90.9	57.7	75.5	32.3
MVSDet [5]	40.5	82.4	79.2	80.2	55.6	40.3	25.4	60.9	3.5	47.3	73.4	28.9	**64.6**	64.1	94.8	52.1	76.7	41.8
SGCDet (Ours)	**43.6**	83.1	**82.3**	**84.7**	**61.7**	**42.1**	**33.6**	**70.4**	**4.0**	**57.3**	**75.4**	**47.1**	61.1	**69.8**	**95.6**	**59.9**	**83.8**	**46.2**

Table 4. Per-category AP@0.50 scores on the ScanNet dataset. The best results are in **bold**.

Method	cab	bed	chair	sofa	tabl	door	wind	bkshf	pic	cntr	desk	curt	fridg	showr	toil	sink	bath	ofurn
NeRF-Det [4]	15.8	**73.1**	45.3	40.6	39.5	8.1	2.0	20.3	0.2	**13.8**	42.5	5.3	25.3	10.0	63.9	26.0	49.1	12.7
MVSDet [5]	14.9	71.4	48.9	54.4	38.8	9.5	3.1	29.6	0.8	9.8	48.5	5.6	40.2	10.2	77.3	**29.0**	52.9	17.7
SGCDet (Ours)	**18.0**	72.5	**57.3**	**61.8**	**45.6**	**12.1**	**3.6**	**44.1**	**1.1**	8.5	**52.1**	**14.1**	**39.1**	**11.8**	**83.0**	28.0	**58.7**	**21.7**

Table 5. Per-category AP@0.25 scores on the ARKitScenes dataset. The best results are in **bold**.

Method	cab	fridg	shlf	stove	bed	sink	wshr	tolt	bthtb	oven	dshwshr	frplce	stool	chr	tble	TV	sofa
NeRF-Det [4]	57.1	81.8	43.0	**20.1**	89.0	38.0	**81.2**	92.2	94.9	65.5	**52.7**	**59.2**	29.8	74.6	67.9	1.3	78.6
MVSDet [5]	58.9	**84.1**	50.9	15.7	86.1	46.4	78.3	**93.1**	94.9	67.6	39.1	51.6	**35.2**	77.2	70.1	**3.5**	**80.4**
SGCDet (Ours)	**61.6**	**84.1**	**53.5**	16.2	**92.7**	**48.6**	79.8	92.8	**95.0**	**70.0**	46.1	48.2	34.5	**79.3**	**73.4**	0.4	79.4

Table 6. Per-category AP@0.50 scores on the ARKitScenes dataset. The best results are in **bold**.

Method	cab	fridg	shlf	stove	bed	sink	wshr	tolt	bthtb	oven	dshwshr	frplce	stool	chr	tble	TV	sofa
NeRF-Det [4]	25.5	70.8	15.3	4.2	69.4	8.5	67.6	74.2	81.7	37.7	**43.0**	12.6	10.4	44.8	34.6	0.0	51.2
MVSDet [5]	30.2	**76.9**	10.8	3.1	72.8	**16.2**	61.7	81.9	85.9	39.5	21.4	**15.5**	13.5	51.3	40.5	**0.5**	60.6
SGCDet (Ours)	**38.0**	68.1	**19.0**	**4.3**	**87.8**	13.9	**71.6**	**87.5**	**93.9**	**48.5**	35.2	7.7	**17.6**	**55.5**	**49.8**	0.0	**62.0**

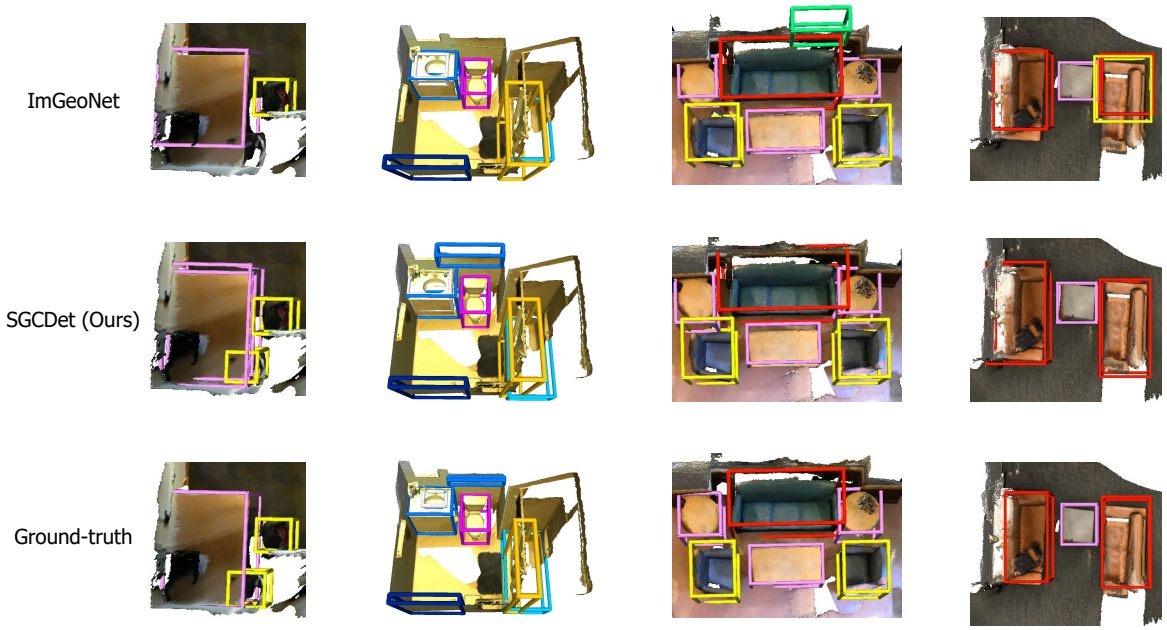

Figure 2. Qualitative results on the ScanNet dataset.

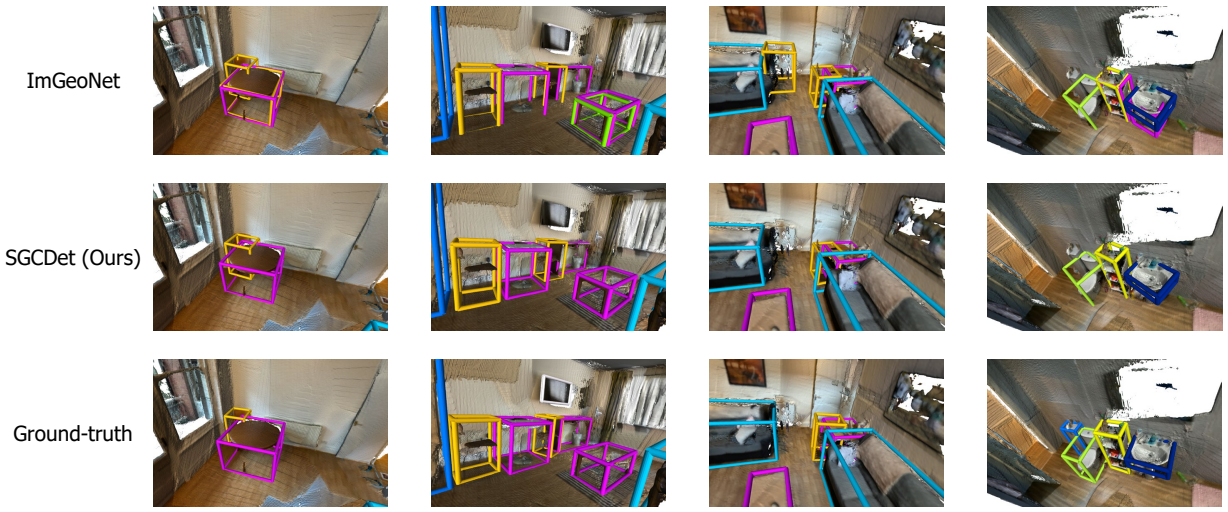

Figure 3. Qualitative results on the ARKitScenes dataset.

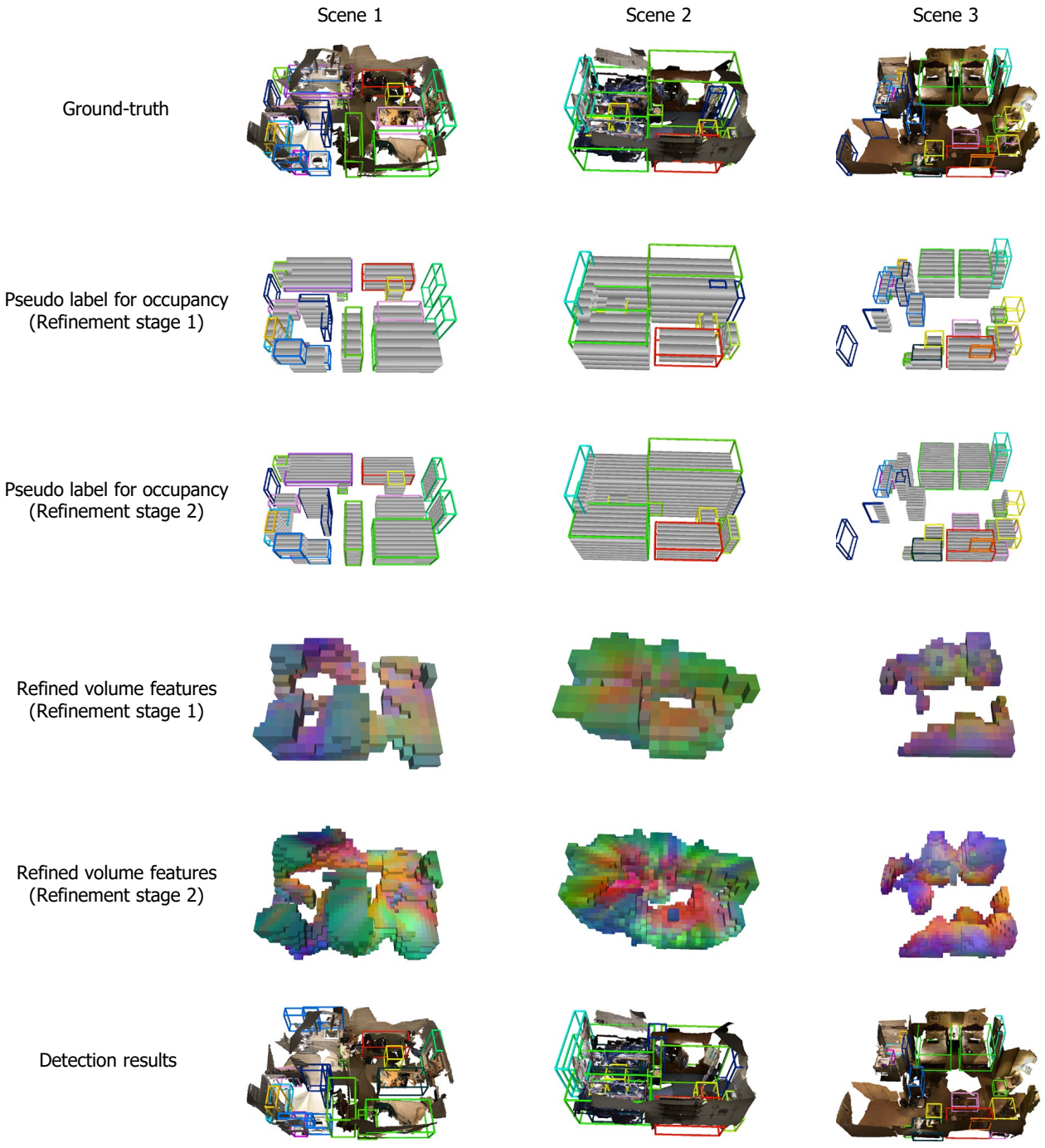

Figure 4. Visualization of sparse volume construction. The voxel resolutions of the two refinement stages are $20 \times 20 \times 8$ and $40 \times 40 \times 16$, respectively. Our sparse volume construction adaptively identifies and selects regions likely to contain objects. This supports efficient feature refinement while avoiding redundancy.