Appendix

A. Implementation Details

In this section, we provide a comprehensive overview of the implementation details for the GO-N3RDet model architecture and its training process. Specifically, we detail the two main branches of our model: the detection branch, which focuses on 3D object detection, and the NeRF branch, which enhances scene representation and geometry understanding. Our elaboration of the training details also includes hyper-parameter settings and dataset-specific setups, to offer a thorough understanding of our implementation.

A.1. Detection Branch

In the Detection Branch, 3D voxel features are constructed by aggregating multi-view image features. These voxel features are then processed through a 3D convolutional neural network [6] to extract spatial and semantic information. Finally, the refined voxel features are passed into a voxel-based detection head, which predicts 3D bounding boxes and object categories.

Feature Volume. To construct the feature volume, multi-scale features are first extracted from input images using a ResNet [4] backbone. These features are then merged across levels through a Feature Pyramid Network (FPN), resulting in a 256-channel output, which is subsequently used as input for voxel generation. Our detection network supports 10 to 200 input views, as demonstrated in Tab. 1, where experiments show the effect of varying view counts on detection accuracy.

The multi-view features are organized into a 3D voxel grid with dimensions $[40, 40, 16]$, voxel size $[0.16\,\text{m}, 0.16\,\text{m}, 0.2\,\text{m}]$, and spatial bounds ranging from $[-2.7\,\text{m}, -2.7\,\text{m}, -0.78\,\text{m}]$ to $[3.7\,\text{m}, 3.7\,\text{m}, 1.78\,\text{m}]$. Input images are resized to $(480, 360)$ before feature extraction. For each voxel, features are aggregated by projecting its center onto the multi-view images and applying average pooling across views, ensuring consistency in voxel representation. This voxel construction method, which originates from NeRF-Det [8] and ImVoxelNet [6], serves as the foundation for our approach. Building upon this baseline, our GO-N3RDet introduces the PEOM to further enhance feature representation and improve detection performance.

PEOM Architecture. The PEOM module improves voxel feature alignment and spatial representation by offsetting voxel centers and integrating image features, which enhances 3D geometric scene perception and helps mitigate the impact of camera calibration errors on voxel field quality.

Initially, each voxel center is projected into 2D image space, where its position is refined through learned offsets (constrained within $[-10, 10]$ pixels). The adjusted position retrieves the corresponding image feature, which is then backprojected into the voxel grid to replace the original voxel center with a refined point. A depth filter ensures consistency with the voxel grid resolution, resulting in a spatially calibrated 3D feature volume.

In the feature fusion phase, multi-layer perceptrons (MLPs) transform these backprojected features and 3D positional data into enriched, position-sensitive representations. Max pooling is then applied to aggregate these features, creating voxel representations that are both highly salient and optimized for downstream tasks.

Through the combined use of positional adjustments during backprojection and spatially-aware feature fusion, PEOM produces precise, feature-rich voxel representations that significantly enhance the performance of 3D object detection.

A.2. The NeRF branch

In the NeRF branch, we leverage NeRF [5] to predict the opacity of each voxel, thereby enriching the voxel's geometric information with opacity-based enhancement.

For Double Importance Sampling (DIS), we sample 2,048 rays per scene, with 128 points sampled along each ray. This sampling strategy integrates information from both offset points and the density predicted through uniform sampling, balancing contributions from these two sources. Specifically, uniform sampling of 128 points is performed within the voxel grid, and the density of these points is computed within the offset voxel field produced by PEOM. Concurrently, the NeRF model predicts densities for these same points. These two sets of densities are then combined using equal weights for balanced integration, as described in Eq. 15 of the main paper. The weights $\alpha = 0.5$ and $\beta = 0.5$ ensure an even contribution from the offset geometry (via PEOM) and the NeRF density predictions.

A.3. Training details

The model is implemented on the MMDetection3D [2] framework to ensure modularity and scalability. Training is conducted with a batch size of 1 on four NVIDIA A100 GPUs. An AdamW optimizer is used with an initial learning rate of 0.0002 and a weight decay of 0.01. A custom parameter configuration applies a learning rate multiplier of 0.1 specifically for the backbone parameters, while the weight decay remains at 1.0. Gradient clipping is applied with a maximum norm of 35 to stabilize training.

The learning rate schedule follows a MultiStepLR policy over 14 epochs, with milestones set at epochs 13. At each milestone, the learning rate is reduced by a factor of 0.1, allowing for gradual decay to enhance convergence in later stages of training.

Views	Method	mAP@.25	mAP@.50
10	NeRF-Det	39.4	17.2
10	GO-N3RDet	38.6	18.3
20	NeRF-Det	45.1	22.5
20	GO-N3RDet	47.9	24.2
50	NeRF-Det	48.8	25.9
50	GO-N3RDet	55.3	30.0
100	NeRF-Det	53.3	26.9
100	GO-N3RDet	57.9	32.5
150	NeRF-Det	52.9	27.6
150	GO-N3RDet	58.0	32.0
200	NeRF-Det	53.4	27.5
200	GO-N3RDet	58.6	33.7

Table 1. Ablation study on the number of views for NeRF-Det [8] and GO-N3RDet on the ScanNet [3] dataset.

B. More on Ablation Study and Discussion

B.1. Supplementary Notes on Ablation Studies

The following provides supplementary explanations for the experiments presented in Tables 4, 5, and 6 of the main paper, offering additional insights into the effectiveness of various modules and parameters in our model.

In the experiments shown in Table 4, we evaluated the performance impact of several modules, including the Positional Information Embedded Voxel Optimization Module (POEM) and Double Importance Sampling (DIS). For these evaluations, POEM was applied to shift voxel centers when assessing the impact of DIS. However, to isolate the effect of DIS independently, we using only DIS to obtain offset points for the NeRF branch, while maintaining the original voxel features in the detection branch, without POEM adjustments. For the ablation studies on DIS and Offset Optimization Module (OOM), the NeRF branch was utilized to enhance feature representation. Conversely, in the experiments specifically examining POEM's effects, the NeRF branch was excluded to isolate POEM's impact on voxel optimization.

Table 5 examines the influence of opacity adjustments on model performance. In this experiment, POEM was consistently applied both with and without opacity adjustments. Meanwhile, Table 6 evaluates the effect of varying the number of sampling points; in this setup, POEM was used only to provide offset coordinates to DIS, without incorporating POEM-enhanced voxel features into the detection branch.

B.2. Number of Views

Based on the experimental results in Tab. 1, our method, GO-N3RDet, consistently outperforms the baseline method, NeRF-Det. As the number of views increases, GO-N3RDet achieves progressively higher detection accuracy in terms of both mAP@0.25 and mAP@0.50, with a particularly notable advantage in high-view configurations. For instance, with 50 views, GO-N3RDet attains an mAP@0.25 of 55.3%, exceeding NeRF-Det by 6.5%, and an mAP@0.50 of 30.0%, surpassing NeRF-Det by 4.1%. When the number of views increases to 200, GO-N3RDet achieves mAP@0.25 and mAP@0.50 scores of 58.6% and 33.7%, respectively, significantly outperforming NeRF-Det's scores of 53.4% and 27.5%. This improvement can be attributed to GO-N3RDet's optimized approach to multi-view information processing, which enables effective integration of details from different perspectives. As the number of views grows, our method more accurately captures spatial relationships and depth information within the scene, enhancing object localization and recognition accuracy. This advantage is particularly evident in high-view configurations, where GO-N3RDet demonstrates increased robustness by leveraging refined opacity calculations, allowing for more precise extraction of geometric and textural details across multiple views.

Point Num	64	128	192	256
GO-N3RDet	57.9	58.6	57.7	57.5

Table 2. Ablation study on the number of sampling points per ray in GO-N3RDet (ScanNet [3] dataset).

Values of (α, β)	(0, 1)	(1, 0)	(0.5, 0.5)
mAP@.50	57.9	58.2	58.6

Table 3. Effect of α and β values on mAP for GO-N3RDet (ScanNet [3] dataset).

B.3. Number of Sample Points

In the main paper, we presented an ablation study on different sampling methods. Here, we extend this analysis to examine the impact of varying the number of sampling points per ray, as shown in Tab. 2. Initially, increasing the sampling points per ray enhances detection performance by enabling the model to capture finer scene details, thereby improving geometric and depth information. This improvement is particularly evident when the point count increases from 64 to 128, with mAP rising from 57.9 to 58.6. However, as the point count is further increased to 192 and 256, detection performance begins to decline slightly. This decline can be attributed to the redundancy introduced by excessive sampling, which may obscure critical global features and hinder the model's ability to focus on the most relevant information. Thus, an optimal number of sampling points strikes a balance between detailed scene capture and efficient feature extraction, while excessive sampling may ultimately reduce detection accuracy.

Disturbance	NeRF@0.25	NeRF@0.5	Ours@0.25	Ours@0.5
None	53.3	27.4	58.6	33.7
$\pm 1^\circ / \pm$ 1 cm	53.1	27.6	58.5	34.1
$\pm 3^\circ / \pm$ 3 cm	50.6	22.5	58.7	32.9
$\pm 5^\circ / \pm$ 5 cm	46.2	18.4	55.4	31.5
$\pm 10^\circ / \pm$ 10 cm	32.1	8.7	39.5	19.3

Table 4. Performance comparison under different disturbances.

Method	Image Size	mAP@.25	mAP@.50
NeRF-Det	(320, 240)	53.1	27.8
GO-N3RDet	(320, 240)	58.2	32.6
NeRF-Det	(480, 360)	53.4	27.5
GO-N3RDet	(480, 360)	58.6	33.7

Table 5. Impact of image resolution on detection performance for NeRF-Det and GO-N3RDet (ScanNet [3] dataset).

B.4. Perturbation experiment

We also introduce random perturbations to camera poses during testing to simulate inaccurate AR/VR conditions. The experimental results are presented in Table 4, demonstrating the robust performance of our method under such disturbances.

B.5. Double Importance Sample Weight

In Double Importance Sampling (DIS), α and β represent the weights assigned to two density-based sampling strategies, determining their relative contributions to the sampling process. The results in Tab. 3 demonstrate the impact of different (α, β) combinations on the mAP of GO-N3RDet.

When $\alpha = 0$ and $\beta = 1$, the model achieves an mAP of 57.9, indicating that relying solely on the β parameter provides a reasonable performance baseline. Conversely, when $\alpha = 1$ and $\beta = 0$, the mAP improves to 58.2, suggesting that α contributes more effectively to model performance when used independently. The highest mAP, 58.6, is achieved with a balanced combination of $\alpha = 0.5$ and $\beta = 0.5$, demonstrating that equal weighting leverages the complementary strengths of both sampling strategies. This balance allows the model to capture richer information, thereby enhancing detection accuracy.

B.6. Image Size

The results in Tab. 5 demonstrate the impact of image resolution on detection performance for NeRF-Det and GO-N3RDet. Across both metrics (mAP@0.25 and mAP@0.5), GO-N3RDet consistently outperforms NeRF-Det at both image resolutions, highlighting the effectiveness of our approach. At the lower resolution (320, 240), GO-N3RDet achieves an mAP@0.25 of 58.2, significantly surpassing NeRF-Det's 53.1. Similarly, for mAP@0.5, GO-N3RDet records 32.6 compared to NeRF-Det's 27.8. When the resolution increases to (480, 360), both methods see slight performance improvements. GO-N3RDet reaches an mAP@0.25 of 58.6 and an mAP@0.5 of 33.7, again outperforming NeRF-Det, which scores 53.4 and 27.5, respectively. These results suggest that while higher resolution offers minor gains for both methods, GO-N3RDet benefits more substantially, demonstrating its robustness and superior ability to leverage additional image details for improved 3D detection accuracy.

B.7. Extended Comparison with Recent Advances

Recent methods such as CN-RMA [7] and MVSDet [9] have demonstrated strong performance in multi-view indoor object detection. Therefore, we further compare our GO-N3RDet with these approaches to highlight the advantages of our method in this context.

Specifically, CN-RMA [7] leverages a pre-trained 3D reconstruction network to reconstruct point cloud scenes from multiple views, followed by sparse convolution for 3D feature extraction and subsequent detection. However, this approach requires real point cloud data as a supervision signal to guide the reconstruction network, and due to its staged training process, it is not end-to-end, which limits its practicality in real-world applications. Despite these constraints, our GO-N3RDet achieves comparable performance on the ScanNet [3] dataset, with an accuracy of 58.6%, matching that of CN-RMA. Our method eliminates the need for real point cloud supervision, supports end-to-end training, and demonstrates superior training efficiency by requiring significantly fewer training epochs—14 compared to 192—while maintaining competitive performance. Furthermore, our end-to-end method only requires $\sim$12 hr on 4 A100 GPUs, whereas CN-RMA necessitates $\sim$10 hr for MVS pre-training, $\sim$2 hr for the detection module, and $\sim$20 hr for fine-tuning on the same hardware, highlighting the efficiency and practicality of our approach.

The latest MVSDet [9] incorporates Gaussian Splatting into indoor 3D object detection, introducing an end-to-end modeling approach. On the ScanNet [3] dataset, MVSDet achieves significant improvements over NeRF-Det, demonstrating an mAP of 56.2 compared to NeRF-Det's performance. However, our GO-N3RDet outperforms MVSDet with an mAP of 58.6, highlighting its superior ability to leverage NeRF and multi-view features for enhanced multi-view 3D object detection.

C. More Results

C.1. Per-category results

We evaluate per-category performance on the ARKITScenes dataset at an IoU threshold of 0.25. Table 6 reports results across 18 classes. Compared to existing methods, our approach demonstrates a significant improvement,

Methods	cab	fridg	shlf	stove	bed	sink	wshr	tolt	bthtb	oven	dshwshr	frplce	stool	chr	tble	TV	sofa	**mAP@.25**
ImVoxelNet-R50 [6]	32.2	34.3	4.2	0.0	64.7	20.5	15.8	68.9	80.4	9.9	4.1	10.2	0.4	5.2	11.6	3.1	35.6	23.6
NeRF-Det-R50 [8]	36.1	40.7	4.9	0.0	69.3	24.4	17.3	75.1	84.6	14.0	7.4	10.9	0.2	4.0	14.2	5.3	44.0	26.7
Ours-R50	33.5	73.8	30.4	15.5	78.8	47.3	72.0	80.1	66.3	51.9	16.4	10.0	21.9	55.8	44.1	2.2	59.9	44.7

Table 6. Results on ARKITScenes [1] validation set with mAP@0.25. R50 refer to the ResNet50 backbone networks.

achieving a mAP@0.25 of 44.7%. Specifically, our method outperforms ImVoxelNet-R50 [6] and NeRF-Det-R50 [8] , which achieve mAP@0.25 scores of 23.6% and 26.7%, respectively, with a margin of 21.1% and 18.0%. These improvements highlight the effectiveness of our method in enhancing voxel representation and improving 3D detection accuracy across diverse object categories in ARKITScenes.

C.2. More Qualitative Results

In Fig.1 , we present visualization results of novel view synthesis using the NeRF branch on the ScanNet [3] dataset, including RGB and depth images. The synthesized novel views effectively capture objects within the scene. For example, the left side of the first row shows a refrigerator filled with beverages of various colors, for which our method generates realistic novel view synthesis results.

As illustrated in Fig.2 and Fig.3, We provide more qualitative comparisons between our method and the baseline methods on ScanNet [3] and ARKITScenes [1] datasets.

D. Limitations

While our GO-N3RDet model demonstrates substantial accuracy improvements over prior NeRF-based methods, integrating multi-view techniques with NeRF inevitably incurs higher computational costs, especially as the number of views and sampling points increases. This reflects the inherent trade-off between capturing richer geometric information and computational efficiency, a challenge common to most multi-view approaches. In our work, we observed that constructing voxel features using a single pixel feature per voxel, obtained through max pooling across views, is sufficient to achieve promising results. This approach avoids the computational cost of aggregating features from all views, as required in methods like average pooling. This observation suggests that future work could explore more effective view selection strategies or develop methods to utilize image features more efficiently, further improving the detection network's overall performance and computational efficiency.

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. 4, 5

[2] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020. 1

[3] 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. 2, 3, 4, 6

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1

[5] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[6] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022. 1, 4

[7] Guanlin Shen, Jingwei Huang, Zhihua Hu, and Bin Wang. Cnrma: Combined network with ray marching aggregation for 3d indoor object detection from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21326–21335, June 2024. 3

[8] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, et al. 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. 1, 2, 4, 5, 6

[9] Yating Xu, Chen Li, and Gim Hee Lee. Mvsdet: Multi-view indoor 3d object detection via efficient plane sweeps. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024. 3

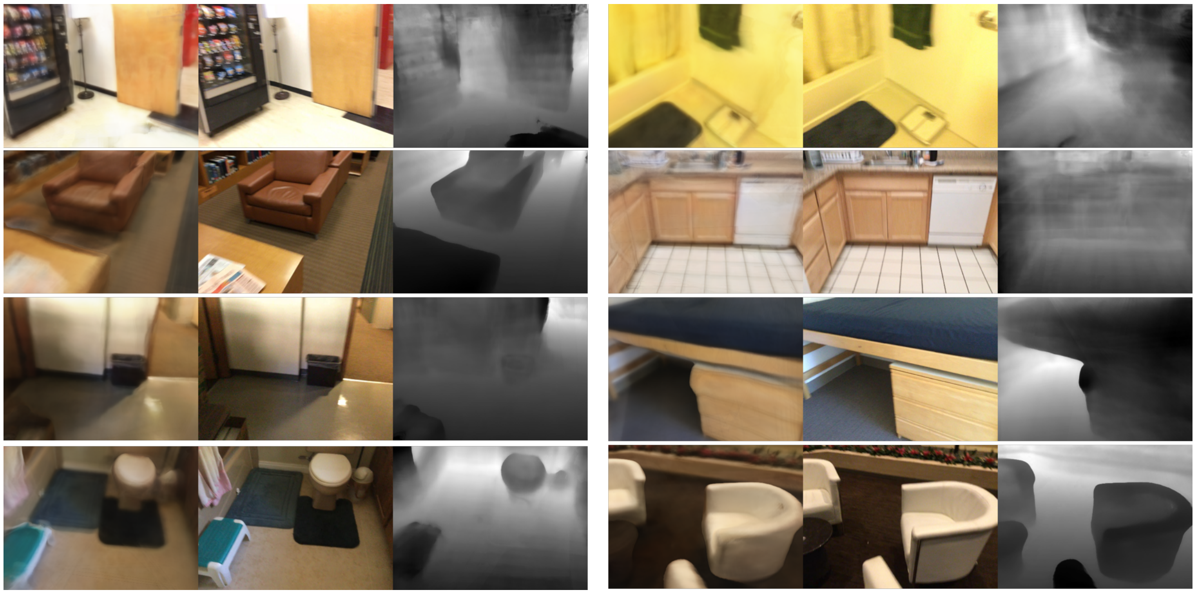

Figure 1. Novel view synthesis results for GO-N3RDet. For each triplet, the left image shows the synthesized result, the middle image presents the real RGB image, and the right image displays the estimated depth map. Note that these visualizations are from the test set and were never seen during training.

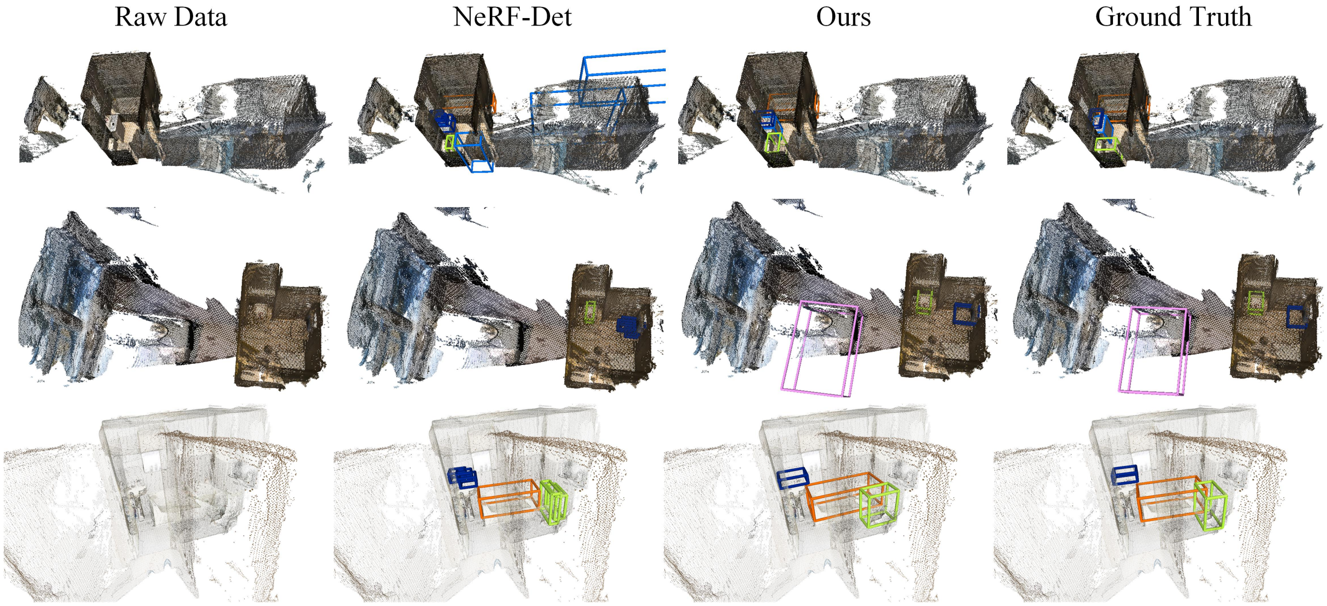

Figure 2. Representative qualitative results on ARKITScenes dataset [1]. As compared to the baseline, i.e., NeRF-DET [8], GO-N3RDet enhancement not only enables detection of more challenging objects, but also reduces false positive detections. Best viewed on screen.

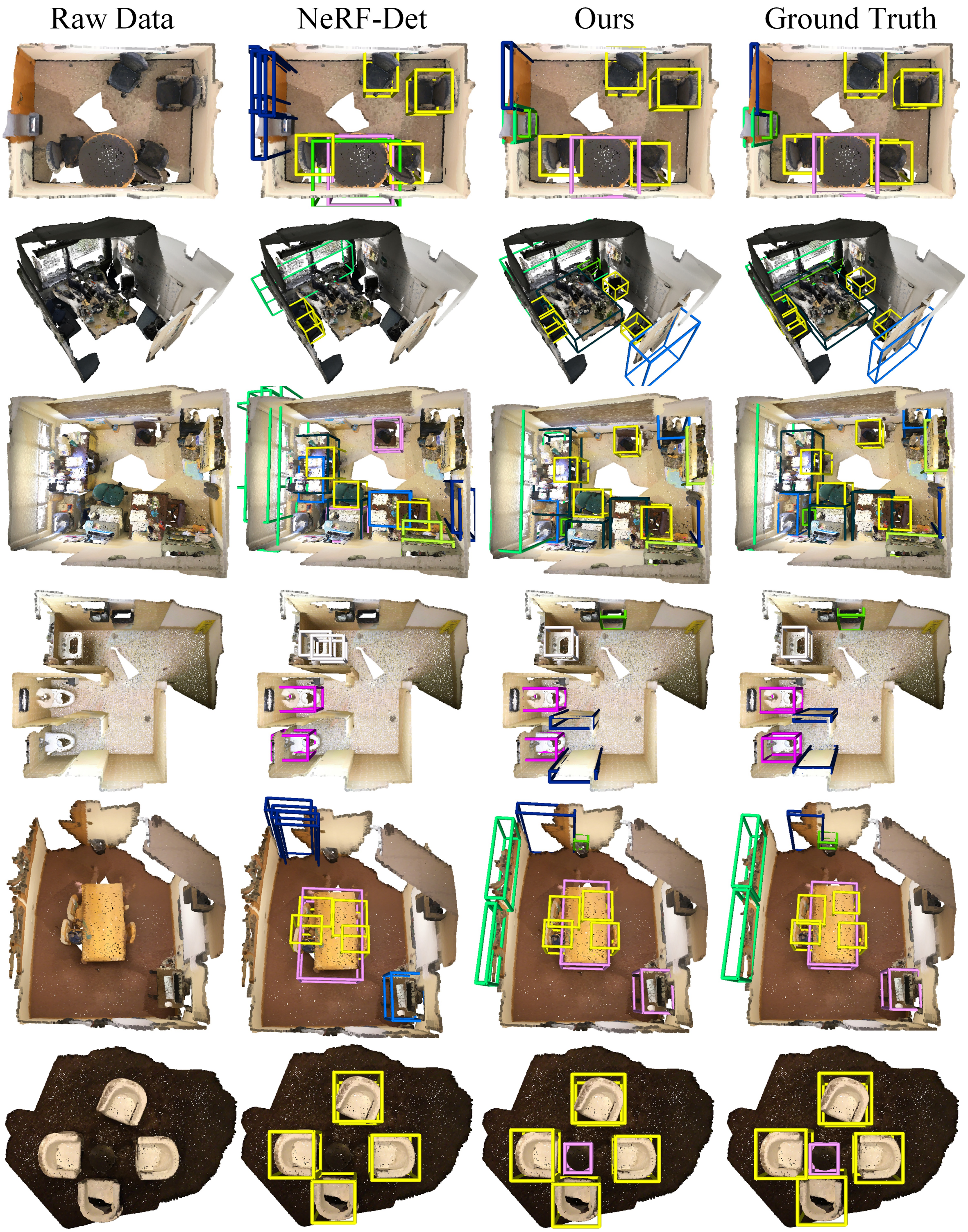

Figure 3. Representative qualitative results on ScanNet dataset [3]. As compared to the baseline, i.e., NeRF-DET [8], GO-N3RDet enhancement not only enables detection of more challenging objects, but also reduces false positive detections. Best viewed on screen.