A. Dataset and Implementation Details

Dataset. Our experiments are conducted on ScanNetV2 [7] and ARKITScenes dataset [2]. ScanNetV2 dataset is a challenging dataset containing 1513 complex scenes with around 2.5 million RGB-D frames and annotated with semantic and instance segmentation for 18 object categories. Since ScanNetV2 does not provide amodal or oriented bounding box annotation, we predict axis-aligned bounding boxes instead, as in [13, 30, 34]. We mainly evaluate the methods by mAP with 0.25 IoU and 0.5 IoU threshold, denoted by mAP@.25 and mAP@.50.

ARKITScenes dataset contains around 1.6K rooms with more than 5000 scans. Each scan includes a series of RGB-D posed images. In our experiments, we utilize the subset of the dataset with low-resolution images. The subset contains 2,257 scans of 841 unique scenes, and each image in the scan is of size 256 × 192. We follow the dataset setting provided by the official repository [1]. We mainly evaluate the methods by mAP with 0.25 IoU as follow [2].

Detection Branch. We follow ImVoxelNet, mainly use ResNet50 with FPN as our backbone and the detection head consists of three 3D convolution layers for classification, location, and centerness, respectively. For the experiment on the ARKITScenes, we additionally predict the rotation. We use the same size $40 \times 40 \times 16$ of the voxels, with each voxel represents a cube of $0.16m, 0.16m, 0.2m$. Besides, we also keep the training recipe as same as ImVoxelNet. During training, we use 20 images on the ScanNet datatset and 50 images on the ARKITScenes dataset by default. During test we use 50 images and 100 images on the ScanNet dataset and ARKITScenes dataset, respectively. The network is optimized by Adam optimizer with an initial learning rate set to 0.0002 and weight decay of 0.0001, and it is trained for 12 epochs, and the learning rate is reduced by ten times after the $8th$ and $11th$ epoch.

NeRF Branch. In our NeRF branch, 2048 rays are randomly sampled at each iteration from 10 novel views for supervision. Note that the 10 novel views are ensured to be different with the views input to detection branch for both training and inference. We set the near-far range as (0.2 meter - 8 meter), and uniformly sample 64 points along each ray. During volumetric rendering, if more than eight points on the ray are projected to empty space, then we would throw it and do not calculate the loss of the ray. The geometry-MLP (G-MLP) is a 4-layer MLP with 256 hidden units and skip connections. The color-MLP (C-MLP) is a one-layer MLP with 256 hidden units. Our experiments are conducted on eight V100 GPUs with 16G memory per GPU. We batched the data in a way such that each GPU carries a single scene during training. During training, the two branches are end-to-end jointly trained. During inference, we can keep either one of the two branches for desired tasks. The whole Our implementation is based on MMDetection3D [5].

Table 10: Ablation on number of views. Due to the GPU memory limitation, we downsample the image resolution 2x when conduct experiments on 100 views (denoted as ImVoxelNet-R50-2x' and NeRF-Det-R50-2x'.). Experiments on each setting run three times. We report the mean and standard deviations of our experiments.

Methods	mAP@.25	mAP@.50
ImVoxelNet-R50-2x (10 views)	37.8±1.2	17.5±1.0
ImVoxelNet-R50-2x (20 views)	46.5±0.5	21.1±0.5
ImVoxelNet-R50-2x (50 views)	48.4±0.3	23.7±0.2
ImVoxelNet-R50-2x'(100 views)	48.1±0.1	24.7±0.1
NeRF-Det-R50-2x (10 views)	41.4 ±1.0 (+3.6)	19.2±0.9 (+1.7)
NeRF-Det-R50-2x (20 views)	50.2 ±0.5 (+3.7)	23.6±0.4 (+2.5)
NeRF-Det-R50-2x (50 views)	51.8 ±0.2 (+3.4)	26.0±0.1 (+2.3)
NeRF-Det-R50-2x'(100 views)	52.2±0.1 (+4.1)	27.4±0.1 (+2.7)

B. Evaluation Protocol of Novel View Synthesis and Depth Estimation.

To evaluate the novel view synthesis and depth estimation performance, we random select 10 views of each scene as the novel view (indicated as target view in IBRNet [44]), and choose the nearby 50 views as the support views. To render the RGB and depth for the 10 novel views, each points shooting from the pixels of novel views would be projected to the all support views to sample features, and then pass into the NeRF MLP as illustrated in Method section. We keep the same novel view and support view for both setting in Table. 6 of the main text. Note that the evaluation is conducted on the test set of ScanNet dataset, which are never seen during training. The non-trivial results also demonstrate the generazability of the proposed geometry-aware volumetric representation.

C. Additional Results

Ablation studies on number of views. We conducted an analysis of how the number of views affects the performance of 3D detection, as shown in Table 10. Specifically, we used the same number of training images (20 images) and tested with different numbers of images. Our proposed NeRF-Det-R50-2x showed a significant improvement in performance as the number of views increased. In contrast, the performance of ImVoxelNet-R50-2x had limited improvement, and even worse, the performance decreased when the number of views increased to 100. We attribute the performance improvements of NeRF-Det to its effective scene modeling. NeRF performs better as the number of views increases, typically requiring over 100 views for

[1] https://github.com/apple/ARKitScenes/tree/main/threedod

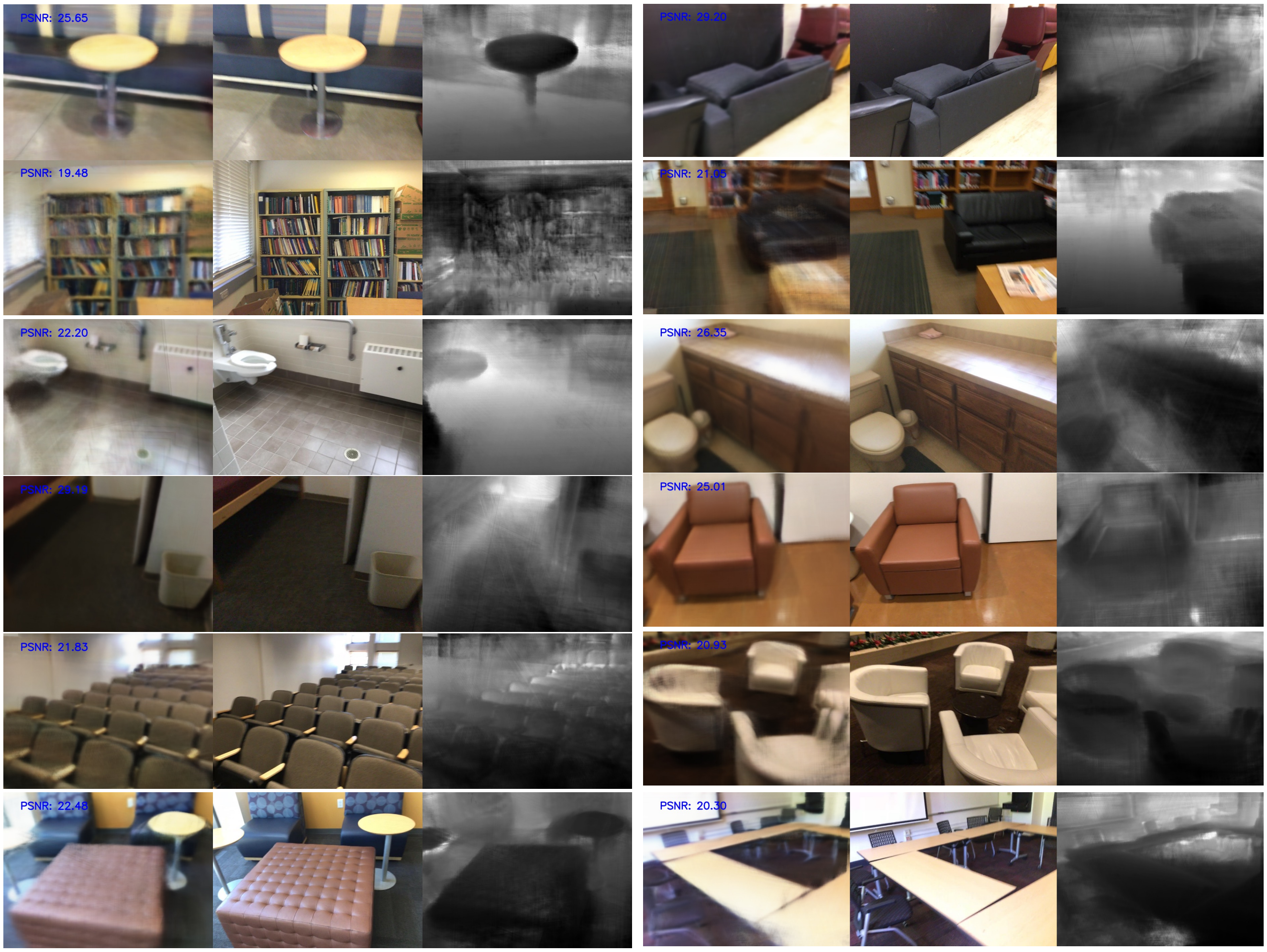

Figure 6: Novel-view synthesis results on top of NeRF-Det-R50-2x*. For each triplet group, the left figure is the synthesized results, the middle one is the ground truth RGB image, and the right part is the estimated depth map. Note that the visualization is from test set, which is never seen during training.

an object [24]. Our proposed NeRF-Det inherits this advantage, leading to a drastic performance gain of 4.1 mAP@.25 and 2.7 mAP@.50 on 100 views.

Overall, our analysis demonstrates the effectiveness of our proposed NeRF-Det in leveraging multi-view observations for 3D detection and the importance of utilizing a method that can effectively model the scene geometry.

More Qualitative Results We provide more visualization results of novel-view synthesis and depth estimation, as shown in Fig. 6. The results come from the test set of ScanNet. We can observe that the proposed method generalizes well on the test scenes. Remarkably, it achieves non-trivial results on the relatively hard cases. For example, the left of the second row presents a bookshelf with full of colorful books, our method can give reasonable novel-view synthesis results. On the other hand, for the left of fifth row, the extremely dense chairs are arranged in the scenes and we can observe the method can predict accurate geometry.

D. Discussion about outdoor 3D detection

We emphasize the differences of NeRF-Det and the other 3D detection works in outdoor scenes. Our proposed NeRF-Det shares the similar intuition with many of outdoor 3D detection works, such as [26, 45, 21], which try to learn geometric-aware representations. However, the proposed NeRF-Det and the other works differ intrinsically. The outdoor 3D detection works [26, 45, 21] propose to use cost volume or explicitly predicted depth to model the scene geometry. Instead, NeRF-Det leverage the discrepancy of multi-view observations, i.e., the augmented variance features in our method section, as the priors of NeRF-MLP input. Beyond the cost volume, we step forward to leverage the photo-realistic principle to predict the density fields, and then transform it into the opacity field. Such a geometry representation is novel to the 3D detection task. The analysis in our experiment part also demonstrates the advantages of the proposed opacity field. In addition to the different method of modeling scene geometry, our design of combining NeRF and 3D detection in an end-to-end manner allows the gradient of NeRF to back-propagate and benefit the 3D detection branch. This is also different from previous NeRF-then-perception works [16, 40].

Our NeRF-Det is specifically designed for 3D detection in indoor scenes, where objects are mostly static. Outdoor scenes present unique challenges, including difficulties in ensuring multi-view consistency due to moving objects, unbounded scene volume, and rapidly changing light conditions that may affect the accuracy of the RGB value used to guide NeRF learning. We plan to address these issues and apply NeRF-Det to outdoor 3D detection in future work.