

NeRF-RPN: A general framework for object detection in NeRFs

Benran Hu^{1*} Junkai Huang^{1*} Yichen Liu^{1*} Yu-Wing Tai^{1,2} Chi-Keung Tang¹

¹The Hong Kong University of Science and Technology ²Kuaishou Technology

Abstract

This paper presents the first significant object detection framework, NeRF-RPN, which directly operates on NeRF. Given a pre-trained NeRF model, NeRF-RPN aims to detect all bounding boxes of objects in a scene. By exploiting a novel voxel representation that incorporates multi-scale 3D neural volumetric features, we demonstrate it is possible to regress the 3D bounding boxes of objects in NeRF directly without rendering the NeRF at any viewpoint. NeRF-RPN is a general framework and can be applied to detect objects without class labels. We experimented the NeRF-RPN with various backbone architectures, RPN head designs and loss functions. All of them can be trained in an end-to-end manner to estimate high quality 3D bounding boxes. To facilitate future research in object detection for NeRF, we built a new benchmark dataset which consists of both synthetic and real-world data with careful labeling and clean up. Please watch the video for visualizing the 3D region proposals by our NeRF-RPN. Code and dataset will be made available.

1. Introduction

3D object detection is fundamental to important applications such as robotics and autonomous driving which require scene understanding in 3D. Most existing relevant methods require input 3D point clouds or at least RGB-D images acquired from 3D sensors. On the other hand, recent advances in Neural Radiance Fields (NeRF) [34] provide an effective alternative approach to extract highly semantic features of the underlying 3D scenes from 2D multi-view images. Inspired by Region Proposal Network (RPN) for 2D object detection, in this paper, we present the first 3D NeRF-RPN, which directly operates on the NeRF representation of a given 3D scene learned entirely from RGB images and camera poses. Specifically, given the radiance field and the density extracted from a NeRF model, our method produces bounding box proposals, which can be deployed in downstream tasks predicted on aligned features.

Recently, NeRF has provided very impressive results in novel view synthesis. On the other hand, 3D object



Figure 1. Region proposal results on a NeRF. Top 12 proposals in eight orientations with highest confidence are visualized. The NeRF is trained from the *Living Room* scene from INRIA [38].

detection has become increasingly important in many real-world applications such as autonomous driving and augmented reality. Compared to 2D object detection, detection in 3D is more challenging due to the increasing difficulty in data collection where various noises in 3D can be captured as well. Despite some good works, there is a lot of room for exploration in the field of 3D object detection. Imaged-based 3D object detectors either use a single image (e.g., [1, 4, 61]) or utilize multi-view consensus of multiple images (e.g., [29, 50, 62]). Although the latter use multi-view projective geometry to combine information in the 3D space, they still use 2D features to guide the pertinent 3D prediction. Some other 3D detectors based on point cloud representation (e.g., [31, 33, 41, 72]) heavily rely on accurate data captured by sensors. To our knowledge, there is still no representative work on direct 3D object detection in NeRF.

Thus, we propose NeRF-RPN to propose 3D ROIs on a given NeRF representation. Specifically, the network takes as input the 3D volumetric information extracted from NeRF, and directly outputs 3D bounding boxes of ROIs. This RPN will thus be a powerful tool for 3D object detection for NeRF by adopting the “3D-to-3D learning” paradigm, taking full advantages of 3D information inherent in NeRF and predicting 3D region proposals directly in 3D space.

As the first significant attempt to NeRF-RPN for 3D object detection directly from multi-view images, this paper’s focus contributions consist of:

- First significant attempt on introducing RPN to NeRF for 3D objection detection and related tasks.

*Equal contribution. The order of authorship was determined alphabetically.

- A large-scale public indoor NeRF dataset for 3D object detection, based on the existing synthetic indoor dataset Hypersim [46] and 3D-FRONT [11], and real indoor dataset ScanNet [5] and SceneNN [19], carefully curated for NeRF training.
- Implementation and comparisons of NeRF-RPNs on various backbone networks, detection heads and loss functions. Our model can be trained in 4 hrs using 2 NVIDIA RTX3090 GPUs. At runtime, it can process a given NeRF scene in 115 ms (excluding postprocessing) while achieving a 99% recall on our 3D-FRONT NeRF dataset.
- Demonstration of 3D object detection over NeRF and related applications based on our NeRF-RPN.

2. Related Work

2.1. NeRF

Neural radiance field (NeRF) [34] has become the mainstream approach for novel view reconstruction, which models the geometry and appearance of a given scene into a continuous and implicit radiance field parameterized by an MLP. Following this work, instant neural graphics primitive [36] applies hash encoding to reduce the training time dramatically. PlenOctrees [66] uses an octree-based radiance field and a grid of spherical basis functions to accelerate rendering speed and decoding appearance. TensoRF [3] projects a 3D point onto three 2D planes to encode the positional information. Although these works use different approaches to model structures, they achieve the same goal of taking as input xyz coordinates and 3D camera view directions to generate view-dependent RGB colors and volume density at each position to render the images from a given view point. NeRF not only provides structure details of a 3D scene but also is conducive to 3D training, where only RGB images with camera parameters are required, thus making this alternative originally targeted at novel view synthesis also suitable for 3D object detection.

2.2. Object Detection and Region Proposal Network

Subsequent to [20] and recent GPU advances, deep convolutional neural network (CNN) has become the mainstream approach for object detection given single images. Object detection based on deep learning can be divided into anchor-based algorithms and anchor-free algorithms. Anchor-based algorithms include two-stage methods [12, 13, 16, 17, 45] and one-stage methods [10, 23, 25–27, 44, 67]. Anchor-based methods first generate a large number of preset anchors with different sizes and aspect ratios on a given image, then predict the labels and box regression offsets of these anchors. For two-stage methods, relatively coarse region proposals are first generated from anchors, followed by refining such coarse proposals to obtain the final bounding boxes and corresponding labels. One-stage methods directly predict bounding boxes and

labels from anchors. In contrast to anchor-based algorithms, anchor-free algorithms [8, 9, 22, 28, 56, 70, 71] directly use feature maps to predict bounding boxes and labels.

Region Proposal Network (RPN) was first introduced in [45] to use CNN to propose regions that may contain objects in an image for subsequent refinement. RPN uses shared convolutional layers to slide through local regions on the feature maps from feature extraction layers and feeds the transformed features into a box-regression layer and a box-classification layer in parallel. In [45], RPN is applied on the feature map from the last shared convolution layer only, while more recent works such as Feature Pyramid Networks (FPN) [24] extend RPN by utilizing multi-scale feature maps. Our proposed RPN adapts the idea of sliding window of 2D RPN and also utilizes FPN in a 3D fashion.

2.3. 3D Object Detection

Based on the input form, current methods for 3D object detection can be categorized into point cloud-based and RGB-based methods. Point cloud-based 3D object detectors rely on point clouds acquired from LiDARs or depth sensors as its input. Many of them first transform point clouds or RGB-D images into voxel representation and subsequently operate on the 3D feature volume through convolution [15, 33, 49, 55, 72] or Transformers [33, 58]. However, the large memory footprint of voxel representation constrains the output resolution. While sparse convolution [14] and 2D projection have been adopted to alleviate the issue, more recently, works directly operating on raw point clouds have been proposed [21, 31, 35, 40–42, 52, 59]. Most of these approaches involve partitioning points into groups and applying classification and bounding box proposal to each group. Criteria used for grouping include 3D frustums extruded from 2D detection [41], 3D region proposals [51, 52], and voting [39, 40, 59]. GroupFree3D [31] and Pointformer [37], on the other hand, do not need grouping but instead use Transformers to attend over all points.

3D objection detection on single images or multi-view RGB images with camera poses is more challenging and relatively less explored, especially for the latter. Early attempts in monocular 3D objection detection first estimate the per-pixel depths [4, 64], pseudo-LiDAR signals [43, 61, 65], or voxel information [47] from an RGB image, then perform detection on the reconstructed depths or 3D features. Later works have extended 2D object detection methods to operate in 3D. For instance, M3D-RPN [1] and MonoDIS [53] use 2D anchors for detection and predict a 2D to 3D transformation. FCOS3D [60] predicts a series of 3D parameters based on the FCOS architecture [57]. Recently more research has been focused on the multi-view case. ImVoxelNet [50] projects features from multiple images back to a 3D grid and applies a normal voxel-based detector on it. DETR3D [62] and PETR [29] adopt a similar design as DETR [2]. DETR3D predicts 3D

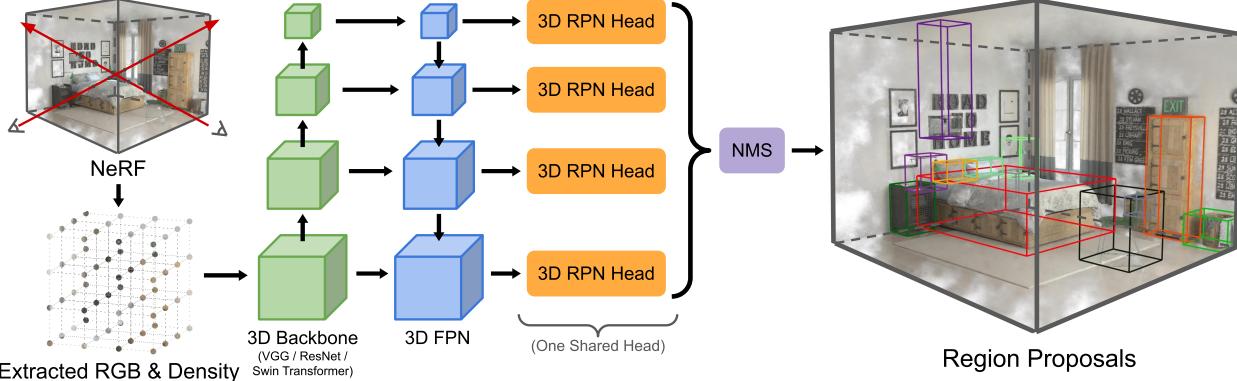


Figure 2. **NeRF-RPN.** Our NeRF-RPN first samples a grid of points and extract their RGB and density in NeRF. The extracted volumetric features are then passed through a 3D backbone architecture to extract deep 3D features at multi-scale. The deep 3D features are fused using a 3D FPN module with a 3D RPN head to regress the potential region proposals.

reference points, which are projected onto 2D images to aggregate image features, whereas PETR directly integrates 3D position embedding into the 2D feature maps. These image-based methods can assist the region proposal task in NeRF, but they do not utilize the inherent or reconstructed 3D information from NeRF and are thus limited in their accuracy.

While it is possible to sample from the radiance and density field of a NeRF model and produce a voxel or point cloud representation on which corresponding 3D object detection methods can be applied for region proposal, such conversion can be ad-hoc depending on both the NeRF model structure as well as the reconstruction quality. Noise and poor geometry approximation at fine details in these converted representations also pose challenges to further applying existing 3D object detectors. It is worth noting that unlike point cloud samples which cover only the surface of an object (crust), the predicted density in NeRF distributes over the interior of an object as well. Clearly, existing methods fail to utilize this important solid object information, which is adequately taken into account by our NeRF-RPN. Besides, there are still no existing 3D object detection datasets that are tailored for the NeRF representation, which also limits the development of 3D object detection in NeRF.

3. Method

Similar to the original RPN, our method has two major components, see Figure 2. The first consists of a feature extractor that takes as input raw radiance and density voxel grid sampled from a NeRF model, and produces a feature pyramid as output. The second is the RPN itself which operates on the feature pyramid and generates object proposals. The volumes on the feature pyramid corresponding to the proposals can subsequently be extracted and further processed for any downstream tasks. Our method is flexible in the form of NeRF input features as well as the network architectures of the feature extractor and the RPN module, which can be adapted to multiple downstream tasks.

3.1. Input Sampling from NeRF

Our method assumes a fully-trained NeRF model with reasonable quality model is provided. The first step is to uniformly sample its radiance and density information to construct a feature volume. Despite the existence of a large number of variants since the original NeRF which adopt different radiance field representations or structures, they share the same property that the reconstructed radiance and density can be queried with view directions and spatial locations. As essentially the radiance and density are used in a similar volumetric rendering process, our method uses the radiance field and density queried from the NeRF as the input so that our NeRF-RPN is agnostic to existing NeRF representation variants.

We uniformly sample the radiance and density on a grid that covers the full traceable volume of the NeRF model. The traceable range is determined by slightly enlarging the bounding box which encloses all the cameras and objects in the scene. The resolution of the grid in each dimension is proportional to the length of the traceable volume in that dimension so that the aspect ratio of the objects are maintained. For NeRF models that use plain RGB for radiance representation, we sample from the same viewing directions used in the camera poses provided to train the NeRF and average the results. If such camera poses are unknown, we uniformly sample directions from a sphere. Generally, the sample at each voxel is in the form of (r, g, b, α) , where (r, g, b) is the averaged radiance and α is converted from the density σ :

$$\alpha = \text{clip}(1 - \exp(-\sigma\delta), 0, 1), \quad (1)$$

where $\delta = 0.01$ is a chosen distance. For NeRF models adopting spherical harmonics or other basis functions for radiance representation, we either supply view directions and compute the corresponding RGB values, or alternatively use the function coefficients as radiance information, depending on the downstream task.

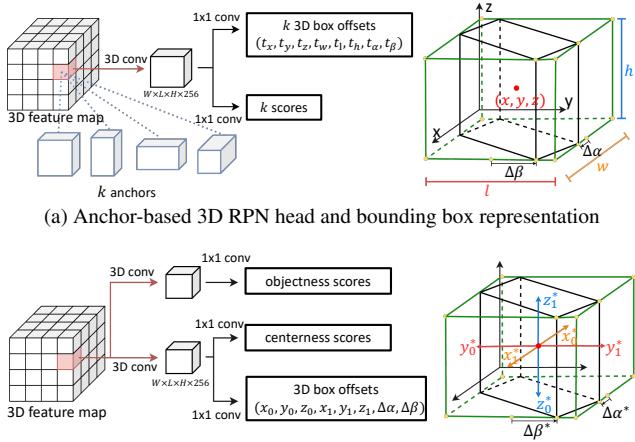


Figure 3. **3D RPN Head.** These two figures illustrate the architectures of anchor-based and anchor-free 3D RPN heads along with their 3D midpoint offset bounding box representations.

3.2. Feature Extractors

Given the grid of raw data, the feature extractor will generate a feature pyramid. We adopt three backbones: VGG [54], ResNet [18] and Swin Transformer [30] in our experiments, but other backbone networks may also be applicable. Considering the large variation in object sizes for indoor NeRF scenes as well as the scale differences between different NeRF scenes, we incorporate an FPN [24] structure to generate multi-scale feature volumes and to enhance high-level semantics information in higher resolution feature volumes. For VGG, ResNet, and the FPN layers, we replace all the 2D convolutions, poolings, and normalization layers with their 3D counterparts. For Swin Transformer, we correspondingly employ 3D position embedding and shifted windows.

3.3. 3D Region Proposal Networks

Our 3D Region Proposal Network takes as input the feature pyramid of the feature extractor and output a set of oriented bounding boxes (OBB) and their corresponding objectness scores. As in most 3D object detection works, we constrain the rotation of the bounding boxes to the world-space z -axis only, which is aligned with the world-space gravity vector and perpendicular to the ground. We experiment two types of region proposal methods for our RPNs: anchor-based and anchor-free methods, see Figure 3.

Anchor-Based RPNs Conventional RPNs as originally proposed in Faster R-CNN [45] place anchors of different sizes and aspect-ratios at each pixel location and predict objectness scores and regression results of the bounding boxes for each anchor. We extend this approach to 3D by placing 3D anchors of different aspect-ratios and scales in voxels on different levels of the feature pyramid. We add k levels of 3D convolutional layers after the feature pyramids (typically $k = 2$ or 4), on top of which two separate 1×1 3D convolutional layers are used to predict the probability

p that an object exists, and the bounding box offset \mathbf{t} for each anchor, see Figure 3(a). The k layers and 1×1 convolutional layers are shared between different levels of the feature pyramid to reduce the number of parameters and improve the robustness to NeRF scenes of different scales. The bounding box offset $\mathbf{t} = (t_x, t_y, t_z, t_w, t_l, t_h, t_\alpha, t_\beta)$ is parametrized similar to [63] but extended with a new dimension:

$$\begin{aligned} t_x &= (x - x_a)/w_a, & t_y &= (y - y_a)/l_a, \\ t_z &= (z - z_a)/h_a, & t_w &= \log(w/w_a), \\ t_l &= \log(l/l_a), & t_h &= \log(h/h_a), \\ t_\alpha &= \Delta\alpha/w, & t_\beta &= \Delta\beta/l, \end{aligned} \quad (2)$$

where $x, y, w, l, \Delta\alpha, \Delta\beta$ are used to describe the projection of the OBB on the xy -plane, while z, h are the additional dimension in height. $x_a, y_a, z_a, w_a, l_a, h_a$ are the position and the size of the reference anchor, see Figure 3(a). Note that this representation does not guarantee the encoded OBBs are cuboids. We follow [63] to transform the projections into rectangles before using them as proposals.

To determine the label of each anchor, we follow the process in Faster R-CNN but with parameters adapted to better fit the 3D setting: we assign a positive label to an anchor if it has an Intersection-over-Union (IoU) overlap greater than 0.35 with any of the ground-truth boxes, or if it has the highest IoU overlap among all anchors with one of the ground-truth box. An anchor that is not labeled positive with IoU values below 0.2 for all ground-truth boxes is regarded negative. Anchors that are neither positive nor negative are ignored in loss computation. The loss we use is similar to that in Faster R-CNN:

$$\begin{aligned} L(\{p_i\}, \{\mathbf{t}_i\}) &= \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) \\ &\quad + \frac{\lambda}{N_{reg}} \sum_i p_i^* L_{reg}(\mathbf{t}_i, \mathbf{t}_i^*), \end{aligned} \quad (3)$$

where p_i, \mathbf{t}_i are predicted objectness and box offsets, p_i^*, \mathbf{t}_i^* are ground-truth targets, N_{cls}, N_{reg} are the number of anchors involved in loss computation, and λ is a balancing factor between the two losses. L_{cls} is the binary cross entropy loss and L_{reg} is the smooth L_1 loss in [12]. The regression loss is only computed for positive anchors.

Anchor-Free RPNs Anchor-free object detectors discard the expensive IoU computation between anchors and ground-truth boxes and can be used for region proposal in specific problem scopes (e.g., figure-ground segmentation). We choose FCOS which is a representative anchor-free method and extend it to 3D.

Unlike anchor-based methods, our FCOS-based RPN predicts a single objectness p , a set of bounding box offsets $\mathbf{t} = (x_0, y_0, z_0, x_1, y_1, z_1, \Delta\alpha, \Delta\beta)$, and a centerness score c for each voxel, see Figure 3(b). We extend the encoding of box offsets in FCOS and define the regression target

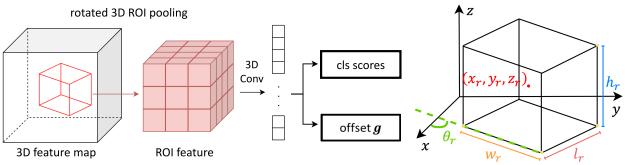


Figure 4. **Binary Classification Network.** The binary classification network architecture with rotated 3D ROI pooling along with the bounding box representation used in this network.

$\mathbf{t}_i^* = (x_0^*, y_0^*, z_0^*, x_1^*, y_1^*, z_1^*, \Delta\alpha^*, \Delta\beta^*)$ as following:

$$\begin{aligned} x_0^* &= x - x_0^{(i)}, & x_1^* &= x_1^{(i)} - x, \\ y_0^* &= y - y_0^{(i)}, & y_1^* &= y_1^{(i)} - y, \\ z_0^* &= z - z_0^{(i)}, & z_1^* &= z_1^{(i)} - z, \\ \Delta\alpha^* &= v_x^{(i)} - x, & \Delta\beta^* &= v_y^{(i)} - y, \end{aligned} \quad (4)$$

where x, y, z are the voxel position, $x_0^{(i)} < x_1^{(i)}$ are the left and right boundary of the **axis-aligned bounding box (AABB)** of i -th ground-truth OBB, and likewise for $y_0^{(i)}, y_1^{(i)}, z_0^{(i)}, z_1^{(i)}$. $v_x^{(i)}$ denotes the x coordinate of the upmost vertex in the xy -plane projection of the OBB, and $v_y^{(i)}$ is the y coordinate of the rightmost vertex, see Figure 3(b). The ground-truth centerness is modified by:

$$c^* = \sqrt{\frac{\min(x_0^*, x_1^*)}{\max(x_0^*, x_1^*)} \times \frac{\min(y_0^*, y_1^*)}{\max(y_0^*, y_1^*)} \times \frac{\min(z_0^*, z_1^*)}{\max(z_0^*, z_1^*)}}. \quad (5)$$

The overall loss is then given by

$$\begin{aligned} L(\{p_i\}, \{\mathbf{t}_i\}, \{c_i\}) &= \frac{1}{N_{pos}} L_{cls}(p_i, p_i^*) \\ &+ \frac{\lambda}{N_{pos}} p_i^* L_{reg}(\mathbf{t}_i, \mathbf{t}_i^*) + \frac{1}{N_{pos}} p_i^* L_{ctr}(c_i, c_i^*), \end{aligned} \quad (6)$$

where L_{cls} is the focal loss in [25] and L_{reg} is the sum of the IoU loss for rotated boxes in [69] and an extra smooth L_1 loss on $\Delta\alpha$ and $\Delta\beta$. L_{ctr} is the binary cross entropy loss. $p_i^* \in \{0, 1\}$ is the ground-truth label of each voxel in the feature pyramid, which is determined using the same center sampling and multi-level prediction process as in [57]; λ is the balancing factor and N_{pos} is the number of voxels with $p_i^* = 1$. The regression and centerness loss only account for positive voxels.

To learn p, \mathbf{t}, c , we adopt the same network for FCOS by adding $k = 2$ or 4 3D convolutional layers independently for the classification and regression branch after the feature pyramid. We append a convolutional layer on top of the classification and regression branch, respectively, to output p and \mathbf{t} , and a parallel convolutional layer on the regression branch for predicting c . As our anchor-based method, we transform the possibly skewed box predictions into cuboids for further post-processing.

3.4. Additional Loss Functions

Objectness Classification Although our NeRF RPN mainly targets on high recalls, some downstream tasks may

prefer a low false-positive rate as well. To improve the precision of ROIs, we add a binary classification network as a sub-component to achieve foreground/background classification. More specifically, the network takes 1) the ROIs from RPN, and 2) the feature pyramid from the feature extractor as input, and outputs an objectness score and bounding box refinement offsets for each ROI, see Figure 4. We extract rotation invariant features for each proposal via rotated ROI pooling. Each proposal is parameterized by $(x_r, y_r, z_r, w_r, l_r, h_r, \theta_r)$, where (x_r, y_r, z_r) describes the center coordinate, w_r, l_r, h_r are the three dimensions, and $\theta_r \in [-\frac{\pi}{2}, \frac{\pi}{2}]$ is the rotated angle along the z -axis. Referring to [63], we first enlarge the box and locate it in the corresponding feature volume. We then apply trilinear interpolation to calculate the value on each feature point, and pad the ROI feature volume with zero before forwarding it through a pooling layer. After pooling, it eventually becomes an $N \times 3 \times 3 \times 3$ feature, which will be used for further regression and classification. Referring to [7], the bounding box offset $\mathbf{g} = (g_x, g_y, g_z, g_w, g_l, g_h, g_\theta)$ is defined as

$$\begin{aligned} g_x &= ((x - x_r) \cos \theta_r + (y - y_r) \sin \theta_r) / w_r \\ g_y &= ((y - y_r) \cos \theta_r - (x - x_r) \sin \theta_r) / l_r, \\ g_z &= (z - z_r) / h_r, & g_w &= \log(w/w_r), \\ g_l &= \log(l/l_r), & g_h &= \log(h/h_r), \\ g_\theta &= (\theta - \theta_r) / 2\pi \end{aligned} \quad (7)$$

The classification layer estimates the probability over 2 classes (namely, *non-object* and *object* class). ROIs with IoU overlap greater than 0.25 with any of the ground-truth boxes are labeled as *object*, while all the others are labeled *non-object*. The loss function is similar to Equation 3, where the box offsets are replaced by \mathbf{g}, \mathbf{g}^* .

2D Projection Loss We project 3D bounding box coordinates $b_i = (x_i, y_i, z_i)$ into 2D $b'_i = (x'_i, y'_i)$ and construct a 2D projection loss as following:

$$L_{2d\ proj}(\{b'_i\}) = \frac{1}{N_{cam} N_{box}} L_{reg}(b'_i, b'^*_i), \quad (8)$$

where N_{cam}, N_{box} are the number of cameras and the number of proposals. We set 4 cameras at 4 top corners of the room, pointing to the room center. In our experiments, adding the 2D projection loss to our existing loss function does not further improve the model performance, as shown in Table 4. We believe in our case the 2D projection loss does not provide new information in the presence of the 3D regression losses prescribed in Eq. (3) and (6). The 2D projection loss may however still be helpful when 3D supervision is unavailable.

4. NeRF Dataset for 3D Object Detection

There has been no representative NeRF dataset constructed for 3D object detection. Thus, we build the



Figure 5. **3D-FRONT NeRF Dataset Samples.** Rows 1–2 show the NeRF reconstruction quality and ground-truth bounding box quality of our 3D-FRONT NeRF dataset. Rows 3–4 show groundtruth boxes with diverse object appearance in the dataset.

first NeRF dataset for 3D object detection utilizing Hypersim [46] and 3D-FRONT [11] datasets. In addition to these synthetic datasets, we incorporate a subset of the real-world datasets from SceneNN [19] and ScanNet [5] to demonstrate that our method is robust to real-world data. Figure 5 shows some selected examples of the 3D groundtruth boxes we carefully labeled from 3D-FRONT. Table 1 summarizes our dataset.

Datasets	# Scenes	# Boxes					
		Total #	Average # (per scene)	# Boxes in size (# voxels)			
				< 16 ³	16 ³ ~32 ³	32 ³ ~64 ³	
Hypersim	250	4798	19.2	3836	770	184	8
3D-FRONT	159	1191	7.5	129	703	324	35
ScanNet	90	1086	12.1	508	488	88	2
SceneNN	16	367	22.9	182	112	54	19

Table 1. **Statistics of our NeRF dataset for 3D Object Detection.**

Hypersim Hypersim is a very realistic synthetic dataset for indoor scene understanding containing a wide variety of rendered objects with 3D semantics. However, the dataset is not specifically designed for NeRF training, where the object annotations provided are noisy for direct use in region proposal tasks. Thus, we perform extensive cleaning based on both the NeRF reconstruction quality and the usability of object annotations (supp mtr). Finally we keep around 250 scenes after cleanup. The original 3D object bounding boxes in Hypersim are not carefully pruned, as some objects are invisible in all images. Furthermore, many instances are too fine in scale, while some are of less or little interest, e.g., floors and windows. We remove ambiguous objects that may interfere our training. Then,

we filter out tiny or thin objects by checking if the smallest dimension of their AABB is below a certain threshold. After these automatic pre-processing, we manually examine each remaining object. Objects that are visible in less than three images, or with over half of their AABBs invisible in all images, are removed.

3D-FRONT 3D-FRONT [11] is a large-scale synthetic indoor scene dataset with room layouts and textured furniture models. Due to its size, effort has been spent on splitting complex scenes into individual rooms and cleaning up bounding boxes (supp mtr). A total of 159 usable rooms are manually selected, cleaned, and rendered in our dataset. More rooms can be generated for NeRF training using our code and 3D-FRONT dataset, which will be released when the paper is accepted for publication. We perform extensive manual cleaning on the bounding boxes in each room. Similar to Hypersim, bounding boxes for construction objects such as ceilings and floors are removed automatically based on their labels. Moreover, we manually merge the relevant parts bounding boxes to label the entire semantic object (e.g., seat, back panel and legs are merged into a chair box). Refer to Figure 5 for examples.

Real-World Dataset We construct our real-world NeRF dataset leveraging ScanNet [5], SceneNN [19], and a dataset from INRIA [38]. ScanNet is a commonly used real-world dataset for indoor 3D object detection which contains over 1,500 scans. We randomly select 90 scenes and for each scene, we uniformly divide the video frames into 100 bins and select the sharpest frame in each bin based on the variance of Laplacian. We use the provided depth and a depth-guided NeRF [48] to train the models. For object annotations, we compute the minimum bounding boxes based on the annotated meshes and discard objects of certain classes and sizes as similarly done for Hypersim.

SceneNN is a real-world indoor dataset with around 100 scenes, where RGB-D images with predicted poses, bounding boxes of objects and reconstructed meshes are provided for each scene. We first filter the images by choosing the image with highest sharpness (variance of Laplacian) among every 20 consecutive frames. Then, we project bounding boxes onto chosen images using camera poses to determine camera pose correctness and eliminate incorrect camera poses manually. A total of 16 scenes survive the above, and we use [36] to reconstruct them.

5. Experiments

5.1. Training & Testing

Training Due to limited memory, we use a maximum resolution of 200 for the longest dimension of NeRF sampling grids for Hypersim, and 160 for all other datasets. During training, input scenes are randomly flipped along x, y axes and rotated along z -axis by $\frac{\pi}{2}$ with probability 0.5 for each augmentation operation. Additionally, the scenes

Methods	Backbones	Hypersim				3D-FRONT				ScanNet			
		Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀	Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀	Recall ₂₅	Recall ₅₀	AP ₂₅	AP ₅₀
Anchor-based	VGG19	57.1	14.9	11.2	1.3	97.8	76.5	65.9	43.2	88.7	42.4	40.7	14.4
	ResNet-50	49.8	13.0	9.7	1.3	96.3	70.6	65.7	45.1	86.2	32.0	34.4	9.0
	Swin-S	69.8	28.3	24.6	6.2	98.5	63.2	51.8	26.6	93.6	44.3	38.7	12.9
Anchor-free	VGG19	66.7	27.3	30.9	11.5	96.3	69.9	85.2	59.9	89.2	42.9	55.5	18.4
	ResNet-50	63.2	17.5	23.2	6.0	95.6	67.7	83.9	55.6	91.6	35.5	55.7	16.1
	Swin-S	70.8	21.0	27.7	7.7	96.3	62.5	78.7	41.0	90.6	39.9	57.5	20.5

Table 2. Ablation on different backbones and heads for NeRF-RPN. Recall₂₅ and Recall₅₀ have IoU threshold of 0.25 and 0.5, respectively.

are slightly rotated along z -axis by $\alpha \in [-\frac{\pi}{18}, \frac{\pi}{18}]$ with a probability of 0.5, which we find can significantly improve the average precision (AP) in RPN outputs. We optimize our network with AdamW [32] with an initial learning rate of 0.0003 and a weight decay of 0.001. In our training, we set $\lambda = 5.0$ for Eq. (3) and $\lambda = 1.0$ for Eq. (6). For the anchor-based approach, we adopt a 4-level FPN and anchors of 13 different aspect ratios, which are 1:1:1, 1:1:2, 1:1:3, 2:2:1, 3:3:1, and their permutations. All anchors on the same level of feature volume share the same size for their shortest side, which is in $\{8, 16, 32, 64\}$, from fine to coarse scale. Following the RPN training strategy in [45], we randomly sample 256 anchors from each scene in each iteration to compute the loss, where the ratio of positive and negative anchors is 1:1. For anchor-free approach, all output proposals are used to compute the loss.

Testing After obtaining the ROIs with objectness scores, we first discard the boxes whose geometry centers are beyond the scene boundary. Then, we select the top 2,500 proposals on each level of the feature volumes independently. To remove redundant proposals, we apply Non-Maximum Suppression (NMS) to the aggregated boxes based on rotated-IoU with threshold 0.1, after which we select the 2,500 boxes with the highest objectness scores.

5.2. Ablation Study

Backbones and Heads Table 2 tabulates the recall and average precision of different combinations of feature extraction backbones and RPN heads. When fixing the backbones and comparing the RPN heads only, we observe that anchor-free models achieve a higher AP on all three datasets. The two RPN methods attain similar recalls on 3D-FRONT and ScanNet, while on Hypersim anchor-free models are generally higher in recalls. We believe the better performance of anchor-free models results are twofold: 1) The centerness prediction of anchor-free models helps suppress proposals that are off from the centers, which is particularly helpful when the bounding box center is misaligned with the mass center, or when the NeRF input is noisy; 2) The limited number of aspect ratios and scales for anchors limits the performance of anchor-based models as 3D objects vary greatly in sizes.

Furthermore, when comparing the performance between different backbones, we notice that models with VGG19 generally achieve better recall and AP compared to others. The major exception concerns the performance of anchor-

Methods	Loss	Recall		AP	
		0.25	0.50	0.25	0.50
Anchor-based	Smooth L_1	98.5	63.2	51.8	26.6
	IoU	98.5	71.3	61.6	36.7
	DIoU	97.1	71.3	59.5	32.8
Anchor-free	Smooth L_1	96.3	56.6	76.5	39.9
	IoU	96.3	62.5	78.7	41.0
	DIoU	97.1	64.0	77.4	40.2

Table 3. Ablation results of the bounding box regression loss.

Methods		Recall		AP	
		0.25	0.50	0.25	0.50
Anchor-based	98.5	63.2	51.8	26.6	
	+2D proj. loss	97.1	65.4	58.4	22.2
Anchor-free	96.3	62.5	78.7	41.0	
	96.3	57.4	78.2	41.3	

Table 4. Ablation of the 2D projection loss run on 3D-FRONT, using Swin-S as the backbone.

based models on Hypersim, where Swin-S demonstrates superior recall and AP. Given that the NeRF results on Hypersim are significantly noisier and the scenes are more complex, we suspect that the larger receptive fields and the richer semantics enabled by the shifted windows and attention of Swin Transformers are crucial for our anchor-based method in this case.

NeRF Sampling Strategies While the density field from NeRF is view-independent, the radiance depends on the viewing direction and can be encoded with different schemes. In the appendix we investigate the importance of these view-dependent radiance information to our method, and conclude that using density alone is the best strategy.

Regression Loss

We test three common loss functions for bounding box regression on the 3D-FRONT dataset using Swin-S as the backbone in Table 3. IoU loss directly optimizes the IoU between the predicted and ground-truth bounding boxes while DIoU loss [68] penalizes the normalized distance between the two for faster convergence. We use the variants for oriented boxes of these two losses as proposed in [69]. Our results illustrate that IoU loss consistently outperforms the other two for the anchor-based approach, while for anchor-free models IoU and DIoU loss produce similar performance.

2D Projection Loss We project the 3D bounding boxes as aforementioned. However, we believe 3D features from NeRF already contain sufficient information for precise 3D bounding box regression, which renders the 2D projection loss redundant. This is corroborated by the results in

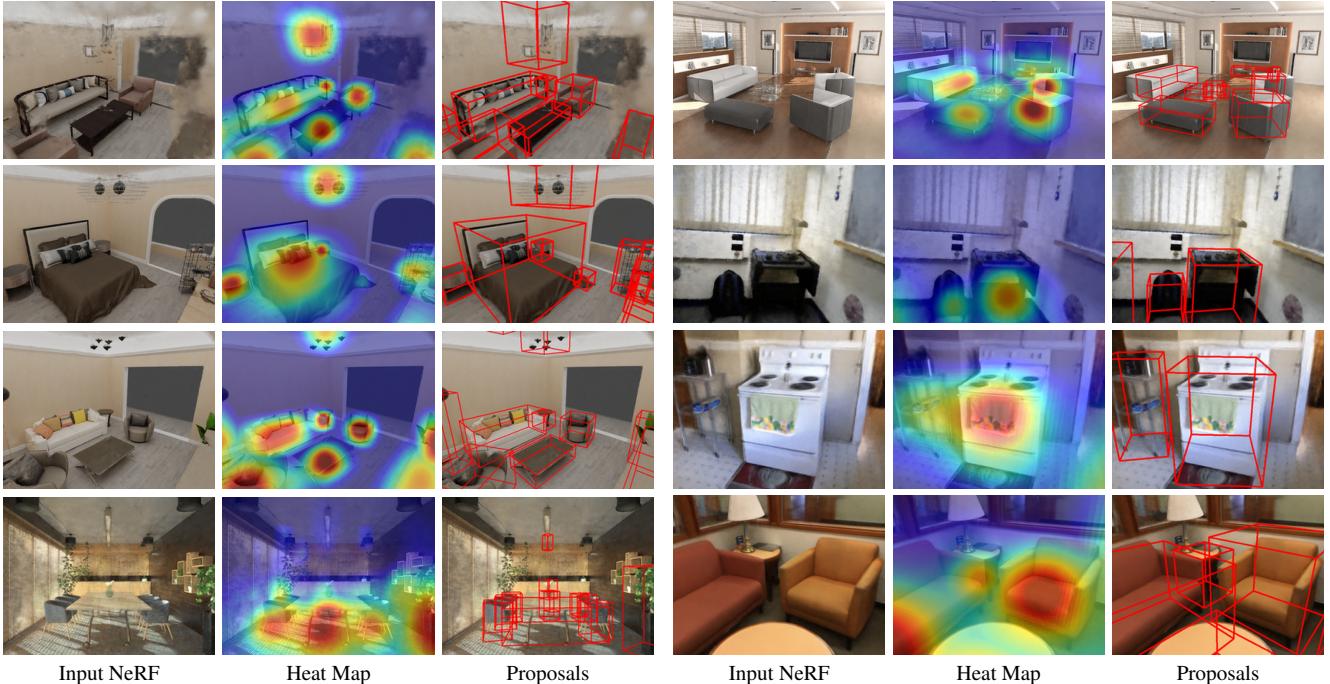


Figure 6. **Qualitative Results.** The “Heat Map” columns show the distribution of proposal confidence scores where red means higher confidence. The “Proposals” columns show a few top bounding boxes after NMS.

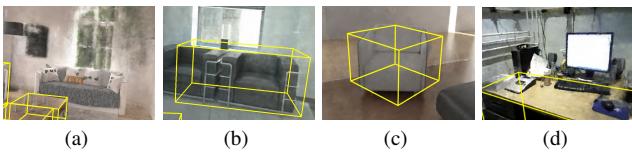


Figure 7. **Failure Cases.** (a)(b) Missing and merging proposals, (c) wrong rotation, (d) no proposal for tiny/second-level objects.

Table 4, where introducing the extra loss does not help with performance. Therefore, we do not use 2D projection loss for other results presented in this paper.

5.3. Results

We performed experiments with different model configurations on various NeRF datasets constructed from Hypersim [46], 3D-FRONT [11], ScanNet [5], SceneNN [19] and INRIA [38], where [19] and [38] are only used in test time due to their relatively small numbers of usable scenes. Detailed quantitative results are shown in Table 2. Figure 6 shows the qualitative results produced by the model with VGG19 backbone and anchor-free RPN head, which shows that the model overall produces reasonably good bounding box proposals. The heat maps demonstrate that our RPN framework can understand 3D scenes with the NeRF inputs. Figure 7 shows typical failure cases. During our experiments, we found that bad NeRF reconstruction can severely hamper the region proposal results. As aforementioned, the region proposal task largely depends on 3D geometry in NeRF. Similar to 2D RPN for images, our method also has missing/merging proposals or wrong rotation after NMS. Presently, our dataset handles first-level objects; tiny or second-level objects are future work.



Figure 8. **Application: Scene Editing.** Removing an object in a bounding box proposed by our NeRF-RPN.

Scene Editing We can edit the scene in NeRF given the proposals produced by our NeRF-RPN. See Figure 8 for a demonstration which sets the volume density inside the proposal bounding box to zero before rendering the image.

6. Conclusion

We propose the first significant 3D object detection framework for NeRF, NeRF-RPN, which operates on the voxel representation extracted from NeRF. By performing comprehensive experiments with different backbone networks, namely, VGG, ResNet, Swin Transformer along with anchor-based, anchor-free RPN heads, and multiple loss functions, we validate our NeRF-RPN can regress high-quality boxes directly from NeRF, without rendering images from NeRF in any view. To facilitate future work on 3D object detection in NeRF, we built a new benchmark dataset consisting of both synthetic and real-world data, with high NeRF reconstruction quality and careful bounding box labeling and cleaning. We hope NeRF-RPN will become a good baseline that can inspire and enable future work on 3D object detection in NeRFs.

References

- [1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9287–9296, 2019.
- [2] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer, 2020.
- [3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022.
- [4] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, Andrew G Berneshawi, Huimin Ma, Sanja Fidler, and Raquel Urtasun. 3d object proposals for accurate object class detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28, 2015.
- [5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.
- [6] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019.
- [7] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2849–2858, 2019.
- [8] Zhiwei Dong, Guoxuan Li, Yue Liao, Fei Wang, Pengju Ren, and Chen Qian. Centripetalnet: Pursuing high-quality keypoint pairs for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [9] Kaiwen Duan, Song Bai, Lingxi Xie, Honggang Qi, Qingming Huang, and Qi Tian. Centernet: Keypoint triplets for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [10] Cheng-Yang Fu, Wei Liu, Ananth Ranga, Ambrish Tyagi, and Alexander C. Berg. Dssd : Deconvolutional single shot detector. *arXiv preprint arXiv:1701.06659*, 2017.
- [11] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10933–10942, 2021.
- [12] Ross Girshick. Fast r-cnn. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, December 2015.
- [13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.
- [14] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3d semantic segmentation with submanifold sparse convolutional networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9224–9232, 2018.
- [15] JunYoung Gwak, Christopher Choy, and Silvio Savarese. Generative sparse detection networks for 3d single-shot object detection. In *European Conference on Computer Vision (ECCV)*, pages 297–313. Springer, 2020.
- [16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.
- [17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1904–1916, 2015.
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [19] Binh-Son Hua, Quang-Hieu Pham, Duc Thanh Nguyen, Minh-Khoi Tran, Lap-Fai Yu, and Sai-Kit Yeung. Scenenn: A scene meshes dataset with annotations. In *International Conference on 3D Vision (3DV)*, pages 92–101. Ieee, 2016.
- [20] Alex Krizhevsky. One weird trick for parallelizing convolutional neural networks. *CoRR*, abs/1404.5997, 2014.
- [21] Jean Lahoud and Bernard Ghanem. 2d-driven 3d object detection in rgb-d images. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4622–4630, 2017.
- [22] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *European Conference on Computer Vision (ECCV)*, September 2018.
- [23] Zuoxin Li and Fuqiang Zhou. Fssd: Feature fusion single shot multibox detector. *arXiv preprint arXiv:1712.00960*, 2017.
- [24] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2017.
- [26] Songtao Liu, Di Huang, and Yunhong Wang. Receptive field block net for accurate and fast object detection. In *European Conference on Computer Vision (ECCV)*, September 2018.
- [27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In *European Conference on Computer Vision (ECCV)*, pages 21–37. Springer International Publishing, 2016.
- [28] Wei Liu, Irtiza Hasan, and Shengcui Liao. Center and scale prediction: Anchor-free approach for pedestrian and face detection. *arXiv preprint arXiv:1904.02948*, 2019.
- [29] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022.
- [30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

- Hierarchical vision transformer using shifted windows. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [31] Ze Liu, Zheng Zhang, Yue Cao, Han Hu, and Xin Tong. Group-free 3d object detection via transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2949–2958, 2021.
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*. OpenReview.net, 2019.
- [33] Jiageng Mao, Yujing Xue, Minzhe Niu, Haoyue Bai, Jia Shi Feng, Xiaodan Liang, Hang Xu, and Chunjing Xu. Voxel transformer for 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3164–3173, 2021.
- [34] 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. In *European Conference on Computer Vision (ECCV)*, 2020.
- [35] Ishan Misra, Rohit Girdhar, and Armand Joulin. An end-to-end transformer model for 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2906–2917, 2021.
- [36] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):102:1–102:15, July 2022.
- [37] Xuran Pan, Zhuofan Xia, Shiji Song, Li Erran Li, and Gao Huang. 3d object detection with pointformer. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7463–7472, June 2021.
- [38] Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. Free-viewpoint indoor neural relighting from multi-view stereo. *CorR*, abs/2106.13299, 2021.
- [39] Charles R Qi, Xinlei Chen, Or Litany, and Leonidas J Guibas. Imvotenet: Boosting 3d object detection in point clouds with image votes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4404–4413, 2020.
- [40] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9277–9286, 2019.
- [41] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgbd data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–927, 2018.
- [42] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.
- [43] Rui Qian, Divyansh Garg, Yan Wang, Yurong You, Serge Belongie, Bharath Hariharan, Mark Campbell, Kilian Q Weinberger, and Wei-Lun Chao. End-to-end pseudo-lidar for image-based 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5881–5890, 2020.
- [44] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [45] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.
- [46] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [47] Thomas Roddick, Alex Kendall, and Roberto Cipolla. Orthographic feature transform for monocular 3d object detection. *arXiv preprint arXiv:1811.08188*, 2018.
- [48] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
- [49] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. *arXiv preprint arXiv:2112.00322*, 2021.
- [50] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2397–2406, 2022.
- [51] Shaoshuai Shi, Li Jiang, Jiajun Deng, Zhe Wang, Chaoxu Guo, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rnn++: Point-voxel feature set abstraction with local vector representation for 3d object detection. *arXiv preprint arXiv:2102.00463*, 2021.
- [52] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [53] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kotschieder. Disentangling monocular 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1991–1999, 2019.
- [54] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- [55] Shuran Song and Jianxiong Xiao. Deep sliding shapes for amodal 3d object detection in rgbd images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 808–816, 2016.
- [56] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.
- [57] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. Fcos: Fully convolutional one-stage object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9627–9636, 2019.
- [58] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and

- Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017.
- [59] Haiyang Wang, Shaoshuai Shi, Ze Yang, Rongyao Fang, Qi Qian, Hongsheng Li, Bernt Schiele, and Liwei Wang. Rbgnet: Ray-based grouping for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1110–1119, 2022.
- [60] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 913–922, 2021.
- [61] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8453, 2019.
- [62] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022.
- [63] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented r-cnn for object detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3520–3529, 2021.
- [64] Bin Xu and Zhenzhong Chen. Multi-level fusion based 3d object detection from monocular images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2345–2353, 2018.
- [65] Yurong You, Yan Wang, Wei-Lun Chao, Divyansh Garg, Geoff Pleiss, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar++: Accurate depth for 3d object detection in autonomous driving. *arXiv preprint arXiv:1906.06310*, 2019.
- [66] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.
- [67] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z. Li. Single-shot refinement neural network for object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.
- [68] Zhaohui Zheng, Ping Wang, Wei Liu, Jinze Li, Rongguang Ye, and Dongwei Ren. Distance-iou loss: Faster and better learning for bounding box regression. In *AAAI*, volume 34, pages 12993–13000, 2020.
- [69] Dingfu Zhou, Jin Fang, Xibin Song, Chenye Guan, Junbo Yin, Yuchao Dai, and Ruigang Yang. Iou loss for 2d/3d object detection. In *International Conference on 3D Vision (3DV)*, pages 85–94. IEEE, 2019.
- [70] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.
- [71] Xingyi Zhou, Jiacheng Zhuo, and Philipp Krähenbühl. Bottom-up object detection by grouping extreme and center points. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [72] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4490–4499, 2018.

A. Additional Details on Dataset Construction

Hypersim As mentioned, we perform extensive cleaning based on the NeRF reconstruction quality. The number of camera poses on each trajectory in the Hypersim dataset is limited to 100, which is too sparse for NeRF training for many larger scenes, and usually produces fuzzy NeRF results strewn with a lot of dangling reconstruction errors or “floaters” to be removed. To remove these unsatisfactory scenes, we train NeRF models on all the scenes, and use a subset of training poses together with randomly interpolated poses as validation camera poses to examine the NeRF quality. We use the NeRF implementation from instant-NGP [36] and run at least 10k training iterations for each scene. By manually checking the NeRF rendering results, we filter out the following types of scenes: 1) scenes containing no objects; 2) scenes where a significant number of object bounding boxes are missing; 3) scenes that are too blurry, or the objects which cannot be clearly separated from floaters. After cleaning the scenes, we further clean the object bounding boxes based on the criteria aforementioned in the paper.

3D-FRONT We spent much effort to split complex scenes into individual rooms and cleaning up bounding boxes. In order to obtain data with suitable size for NeRF training, we first manually partition each scene into individual rooms according to the given layout of the scene. For each selected room, we generate 200~300 camera poses, including 100~150 general views randomly distributed in the room, and 15~20 close-up views for each object within the given room. With these poses, we use [6] to render 2D images for NeRF training.

B. Additional Experiments

B.1. Ablation on NeRF Sampling Strategies

To investigate how view-dependent radiance information from NeRF affects the performance of our method, we experiment the following sampling patterns:

1. use density only;
2. in addition to density, use the average radiance sampled from 18 fixed viewing directions in the form of $(\cos(\phi)\cos(\theta), \cos(\phi)\sin(\theta), \sin(\phi))$, where $\phi \in \{\frac{\pi}{3}, 0, -\frac{\pi}{3}\}$, $\theta \in \{\frac{k\pi}{3} \mid k \in \mathbb{N}, 0 \leq k \leq 5\}$;
3. in addition to density, use the average radiance sampled from all training camera viewing directions;
4. similar as 3) above, but only average from training camera views of which the viewing frustum contains the sample point. If a sample point is invisible in all frustums, we use the same scheme as 3) above;

Methods	Recall		AP	
	0.25	0.50	0.25	0.50
Density only	95.6	82.4	87.9	71.7
Fixed directions	96.3	77.9	84.1	66.3
All cameras	95.6	75.7	86.4	64.0
Filtered cameras	96.3	71.3	86.5	62.1
SH coefficients	95.6	69.9	83.2	57.3

Table B.1. Ablation results of NeRF sampling methods. Reported metrics are calculated on the top 2500 proposals after NMS. Filtered cameras refer to removing training camera views where the sample is outside of the viewing frustums.

5. in addition to density, use the coefficients of the Spherical Harmonics (SH) at the sample point up to degree $l = 3$. The SH function is fitted similarly as in [66] by uniformly sampling radiance from 300 directions on a sphere.

Table B.1 shows the results of different sampling methods above on the 3D-FRONT test set, using VGG19 as the backbone and the anchor-free RPN head. The results might be counter-intuitive as finely-curated radiance information impairs the performance. However, we speculate that density alone is sufficient for the region proposal task as it involves only a binary classification between objects and background which relies less on object semantics. Additionally, in this case, such extra radiance information may lead to more severe over-fitting and thus lower performance on a relatively small dataset such as 3D-FRONT. Nevertheless, the semantics carried in radiance information may be helpful for downstream classification tasks or the detection of secondary object structures.

B.2. Objectness Classification

As mentioned in Section 3.4, we implement a binary objectness classification model. We choose Swin-S [30] as the backbone in the experiments and use the top 2,500 proposals from RPN after NMS with an IoU threshold of 0.3. We fine-tune the feature extractor trained on RPN with AdamW [32], an initial learning rate of 0.0001, and a weight decay of 0.0001. We set $\lambda = 5.0$ in the loss function and also adopt the same augmentation strategy in RPN training. During testing, we use ROIs with objectness scores larger than 0.5 to calculate the average precision (AP). Table B.2 illustrates our results. We find that the APs do not increase as expected and we speculate that this results from the limited resolution of the feature volumes. The ROIs projected onto the coarser-level feature volumes can have similar or smaller sizes compared to our ROI pooling output, while a rotated interpolation over these low-resolution feature volumes can lead to high resampling errors and cannot produce precise features for each rotated ROI. Moreover, the quality of the NeRF models can also affect the ROI quality and the objectness classification performance. However, our objectness classification architecture might still be useful in

Methods	Hypersim		3D-FRONT	
	AP ₂₅	AP ₅₀	AP ₂₅	AP ₅₀
Anchor-based	24.6	6.2	51.8	26.6
+Objectness cls.	12.1	1.2	36.0	7.4
Anchor-free	27.7	7.7	78.7	41.0
+Objectness cls.	14.7	2.5	44.7	16.8

Table B.2. Ablation of the objectness classification component on Hypersim and 3D-FRONT, using Swin-S as the backbone.

many downstream tasks like object detection where the ROI features are required, especially when a higher-resolution feature pyramid is supplied.