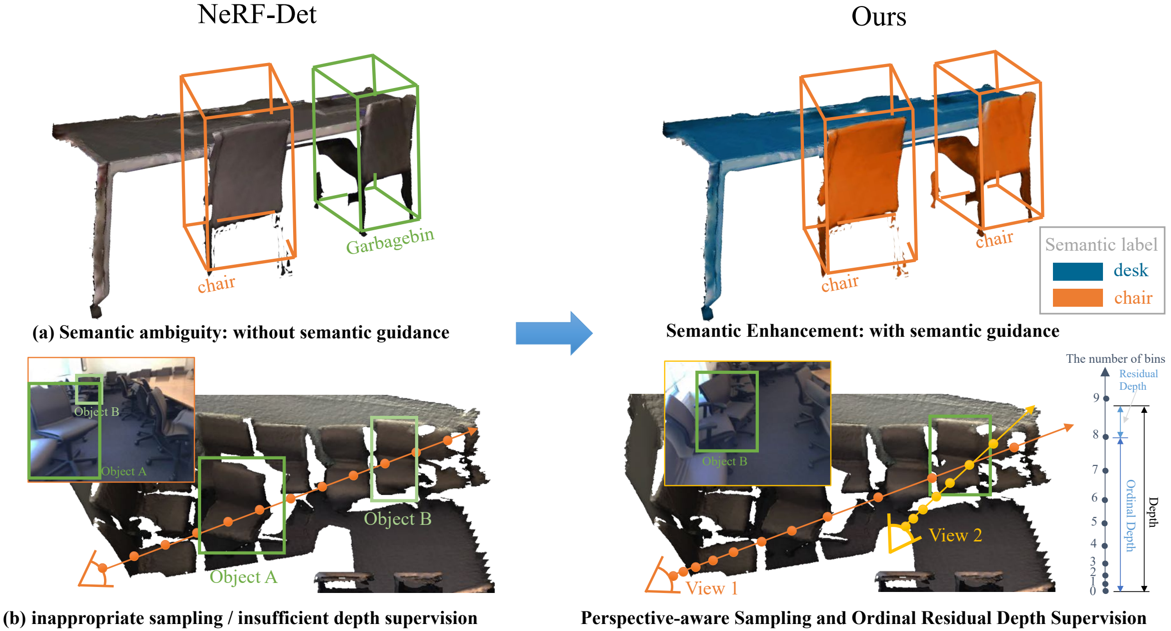
NeRF-Det++: Incorporating Semantic Cues and Perspective-aware Depth Supervision for Indoor Multi-View 3D Detection

Author Name
Affiliation
email@example.com



(a) Semantic ambiguity: without semantic guidance

Semantic Enhancement: with semantic guidance

(b) inappropriate sampling / insufficient depth supervision

Perspective-aware Sampling and Ordinal Residual Depth Supervision

Figure 1: **The shortcomings of NeRF-Det and our corresponding solutions.** Fig. (a) shows the problem of semantic ambiguity. We introduce the Semantic Enhancement module that leverages semantic supervision to enhance the categorical awareness of the detector. Fig. (b) illustrates the limitations of the inappropriate sampling strategy and the insufficient depth supervision. We propose the novel Perspective-aware Sampling, which focuses more on the near, deviating from the conventional uniform sampling approach. Allowing different perspectives to focus on the objects that deserve more attention makes indoor multi-view 3D detection more effective. For example, we can improve the learning of object B from View 2 while allowing View 1 to allocate more attention to its nearby objects. Furthermore, instead of directly regressing the original depth values that are hard to optimize, we propose the Ordinal Residual Depth Supervision. It comprises the classification of the ordinal depth bins and the regression of the residual depth values, which is conducive to more stable depth learning.

Abstract

NeRF-Det has achieved impressive performance in indoor multi-view 3D detection by innovatively utilizing NeRF to enhance representation learning. Despite its notable performance, we uncover three decisive shortcomings in its current design, including **semantic ambiguity**, **inappropriate sampling**, and **insufficient utilization of depth supervision**. To combat the aforementioned problems, we present three corresponding solutions: 1) **Semantic Enhancement**. We project the freely available 3D segmentation annotations onto the 2D plane and leverage the corresponding 2D semantic maps as the supervision signal, significantly enhancing the semantic awareness of multi-view detectors. 2) **Perspective-aware Sampling**. Instead of employing the uniform sampling strategy, we put forward the perspective-aware sampling policy that samples densely near the camera while sparsely in the distance, more effectively collecting the valuable geometric clues. 3) **Ordinal Residual Depth Supervision**. As opposed to directly regressing the depth values that are difficult to optimize, we divide the depth range of each scene into a fixed number of ordinal bins and reformulate the depth prediction as the combination of the classification of depth bins as well as the regression of the residual depth values, thereby benefiting the depth learning process. The resulting algorithm, NeRF-Det++, has exhibited appealing performance in the ScanNetV2 and ARKITScenes datasets. Notably, in ScanNetV2, NeRF-Det++ outperforms the competitive NeRF-Det by **+1.9%** in mAP@0.25 and **+3.5%** in mAP@0.50. The code will be publicly available upon publication.

1 Introduction

3D object detection, which localizes and classifies the objects in the 3D space, serves as a cornerstone technique for various applications, such as augmented reality, robotics, and autonomous driving [Guo *et al.*, 2020]. Multi-view 3D detection draws increasing attention from academic and industrial communities due to the wide accessibility and cheap cost of cameras and reconstruction techniques [Huang *et al.*, 2021; Wang *et al.*, 2023b]. Although image-based outdoor 3D detection has been extensively studied in the past decades, the exploration of indoor multi-view 3D detection is still in its infancy. Indoor 3D detection is challenging due to high object density, severe occlusion, large morphological diversity, irregular spatial distributions of objects, etc.

Previous efforts in indoor multi-view 3D detection have primarily focused on leveraging geometric priors, as images lack precise geometric measurement, which poses significant challenges to obtaining a comprehensive and accurate 3D representation of the surrounding environment. One line of methods (*e.g.*, [Chen *et al.*, 2023], [Rukhovich *et al.*, 2022b], [Rukhovich *et al.*, 2022a]) constructs geometric constraints based on the relationship among different views, and utilizes depth maps obtained through Multi-View Stereo (MVS) algorithms as supervision signals. Fusing these estimated depth maps effectively constructs relatively accurate representations of the 3D world. Another line of methods (*e.g.*, [Xie *et al.*, 2023]) utilizes attention [Vaswani *et al.*, 2017] to incorporate 3D position as a geometric signal and resort to the attention mechanism to establish connections between 2D images and the 3D world, enabling the network to grasp global geometric information.

Recently, the Neural Radiance Field (NeRF) [Mildenhall *et al.*, 2021] has emerged as a powerful tool for geometry modeling. Among the NeRF-based pipelines, NeRF-RPN [Hu *et al.*, 2023] and NeRF-Det [Xu *et al.*, 2023] are two representatives that leverage NeRF to enhance geometry modeling and detect objects more precisely with the reinforced 3D representation. NeRF-RPN utilizes NeRF to extract RGB and density from sampled points and feeds the volumetric features to the perception backbone to yield the detection outputs, which causes additional costs and heavily hinges on the learned neural representation. By leveraging the opacity field estimated by NeRF, NeRF-Det introduces no extra cost during inference and enhances the geometric awareness of the detector.

Despite the impressive performance of NeRF-Det [Xu *et al.*, 2023], we identify three critical flaws in the current framework through comprehensive experiments, *i.e.*, **semantic ambiguity**, **inappropriate sampling**, and **insufficient utilization of depth supervision**. As shown in Fig. 1 (a), although NeRF-Det can roughly estimate the spatial location of an object, the category is usually misclassified. Besides, as shown in Fig. 1 (b), since NeRF-Det employs uniform depth supervision, the depth loss is typically overwhelmed by the distant regions, which have few visual cues and more significant errors, making the depth learning unbalanced and ineffective. In addition, uniform sampling adopted by NeRF-Det fails to fully use the valuable visual clues in the multiple views, weakening the benefit of the neural rendering process.

To tackle the above-mentioned flaws, we systematically investigate the reasons and present the following solutions: 1) **Semantic Enhancement**. We project the freely available 3D segmentation annotations onto the 2D plane and leverage the corresponding 2D semantic maps as the supervision signal. The introduced semantic enhancement module leverages semantic cues to increase the categorical awareness of the detector. 2) **Perspective-aware Sampling**. The multi-view complementation motivates us to design perspective-aware sampling. As shown in Fig. 1 (b), an object in a distant region in one view can be located in a nearby region in another view. By employing perspective-aware sampling, the view with richer visual cues is assigned higher learning weights, making the learning more effective. 3) **Ordinal Residual Depth Supervision**. Directly regressing the depth values causes considerable training difficulty. We divide the depth range of each scene into a constant number of depth bins and the depth regression is reformulated as the combination of the classification of depth bins and the regression of the residual depth values, contributing to the depth learning procedure.

The resulting algorithm, termed NeRF-Det++, is extensively evaluated on the widely used ScanNetV2 [Dai *et al.*, 2017] and ARKITScenes [Baruch *et al.*, 2021] datasets, which are prominent benchmarks for indoor 3D detection. Specifically, our approach achieves 53.9 mAP@0.25 and 29.6 mAP@0.5 in ScanNetV2, surpassing the performance of previous detectors by a large margin. Encouraging results are also observed on the ARKITScenes dataset, further demonstrating the superiority of our method.

Our key contributions are summarized as follows:

1. We point out three critical flaws in NeRF-Det, *i.e.*, semantic ambiguity, inappropriate sampling, and insufficient utilization of depth supervision.
2. We propose systematic solutions to address the aforementioned shortcomings, including designing the Semantic Enhancement module, introducing the Perspective-aware Sampling policy, and putting forward the Ordinal Residual Depth Supervision.
3. Our NeRF-Det++ consistently outperforms previous indoor multi-view 3D detectors on the ScanNetV2 and ARKITScenes datasets.

2 Related work

To compare the distinctions between our method and the existing methods, the related works are introduced from three aspects: neural radiance field (NeRF), multi-view 3D detection, and indoor multi-view 3D object detection with NeRF.

Neural Radiance Field (NeRF). NeRF [Mildenhall *et al.*, 2021] is the pioneering work for novel view synthesis and reconstruction. The main idea of NeRF is that the geometry and appearance of a scene can be modeled using a continuous and implicit radiance field parameterized by a Multi-Layer Perceptron (MLP). Based on NeRF, NeuS [Wang *et al.*, 2021a] and VolSDF [Yariv *et al.*, 2021] utilize the Signed Distance Function (SDF) to replace the density and achieve better reconstruction quality. Due to the spectral bias of MLP, it is more inclined to learn low-frequency signals, causing unsatisfactory 3D reconstruction results with volume rendering. NeuralWarp [Grabocka and Schmidt-Thieme, 2018] utilizes a warping-based loss function to extract the color information from alternate viewpoints, improving reconstruction outcomes. RegSDF [Zhang *et al.*, 2022] proposes several practical regularization terms, including using MVS to obtain additional point clouds and Principal Component Analysis (PCA) to obtain normals as geometric constraints. Mip-NeRF [Barron *et al.*, 2021] introduces a multi-scale representation for Anti-Aliasing NeRF and replaces Positional Encoding (PE) with Integrated Positional Encoding (IPE). It employs a visual frustum-based sampling strategy, prioritizing sampling points closer to the viewer and reducing the sampling density for distant regions. This approach leads to substantial improvements in overall performance. Although our method and Mip-NeRF concentrate on sampling strategies, our approach explicitly modifies the sampling strategy, whereas Mip-NeRF capitalizes on the frustum characteristics. NeuS2 [Wang *et al.*, 2023a] and Neuralangelo [Li *et al.*, 2023] use a hash grid to speed up the training of NeuS, remarkably shortening the per-scene optimization duration of NeRF. Since NeRF is notorious for the tedious per-scene optimization, IBRNet [Wang *et al.*, 2021b] resolves this problem by integrating IBR, NeRF, and the designed ray Transformer to predict and render the results, where relative perspective is used instead of the absolute perspective of the original adjacent image. NeRF and its variants can capture intricate structural details of 3D scenes and facilitate obtaining 3D scene understanding using only RGB images with known poses. Therefore, the representation of NeRF is well-suited for multi-view 3D object detection tasks.

Multi-view 3D Detection. Multi-view 3D object detection aims to predict the category and 3D position of objects by taking multi-view images as input. One commonly used approach is to incorporate geometric consistency as a constraint. For example, VEDet [Chen *et al.*, 2023] leverages 3D multi-view consistency to improve object localization by integrating viewpoint awareness and equal variance, enhancing 3D scene understanding and geometry learning. It employs a query-based transformer architecture to encode 3D scenes, enriching image features by incorporating positional encoding of 3D perspective geometry. Furthermore, methods such as DETR3D [Wang *et al.*, 2022b] and PETR [Liu *et al.*, 2022] build upon DETR [Carion *et al.*, 2020] and combine 2D features from multi-views with 3D position information. Though the preceding approaches utilize geometric priors to some extent to construct the 3D world, the knowledge of the 3D world is inherently encoded within the network. Therefore, another category of methods focuses on constructing alternative representations. For example, ImVoxelNet [Rukhovich *et al.*, 2022b] adopts the voxel-based representation and projects 2D features back onto a 3D grid. However, it lacks explicit geometric information. Frustum3D [Qi *et al.*, 2018] constructs a 3D volume using RGB-D data and performs 3D detection on the point cloud. While these construction methods offer more explicit representations of the 3D world, they often consume significant memory resources.

Indoor Multi-view 3D Detection with NeRF. Recent trends favor the incorporation of NeRF into the detector. NeRF-RPN [Hu *et al.*, 2023] breaks new ground by incorporating NeRF into indoor multi-view 3D object detection. It introduces a versatile pre-trained NeRF model for 3D object detection that does not rely on class labels. It utilizes a novel voxel representation, integrating multi-scale 3D neural volumetric features, allowing for direct utilization of NeRF without the need for viewpoint rendering. However, NeRF-RPN does not fully exploit the potential of NeRF. The early-stage grid sampling results in the loss of substantial RGB, density, and geometric information, leading to unsatisfactory performance. NeRF-Det [Xu *et al.*, 2023] is a concurrent work to NeRF-RPN. It combines a NeRF branch with a 3D detection branch, using a shared MLP to exchange the geometric information. However, NeRF-Det is limited by semantic ambiguity, an inappropriate sampling strategy, and inadequate depth supervision, which hinders its effectiveness. In contrast, our NeRF-Det++ overcomes these limitations and substantially improves the performance of indoor 3D detection.

3 Method

In this section, we first have a brief retrospection of the NeRF-Det framework, which serves as the footstone of our work. Then, we point out the shortcomings of NeRF-Det and present corresponding solutions to tackle these limitations.

3.1 Framework Overview

Given T multi-view images and their pose information, the objective of multi-view 3D detection is to estimate the spatial location and category of the objects, where T represents the total number of views. NeRF-Det introduces a novel approach for indoor multi-view 3D detection by leveraging NeRF to enhance the geometric awareness of the detector. In concrete, it first extracts the 2D features of each posed image from the 2D image backbone. Then, NeRF-Det fuses multi-stage features to generate high-resolution features $\mathbf{F}_i \in \mathbb{R}^{C\times H/4\times W/4}$ ($i = 1, 2, \ldots, T$), where H and W denote the height and width of the input image, respectively, and C denotes the number of channels of the feature map. These features are multiplied with the opacity field, estimated by the shared geometry MLP of NeRF to eliminate the feature ambiguity, yielding geometry-aware features. Despite the excellent performance exhibited by NeRF-Det, it still suffers from three decisive flaws, including semantic ambiguity, inappropriate sampling, and insufficient utilization of depth supervision. To cope with the preceding problems, we proposed the following solutions: *i.e.*, semantic enhancement, perspective-aware sampling, and ordinal residual depth supervision. The schematic overview of our method is shown in Fig. 2.

3.2 Semantic Enhancement

Although NeRF-Det achieves impressive performance in image-based indoor detection, it is usually confronted with inaccurate semantic predictions [Ye *et al.*, 2023] and low confidence scores. The semantic ambiguity problem of NeRF-Det adversely affects the performance of 3D object detection. Therefore, it is necessary to incorporate semantic information into the current framework. To this end, we introduce the

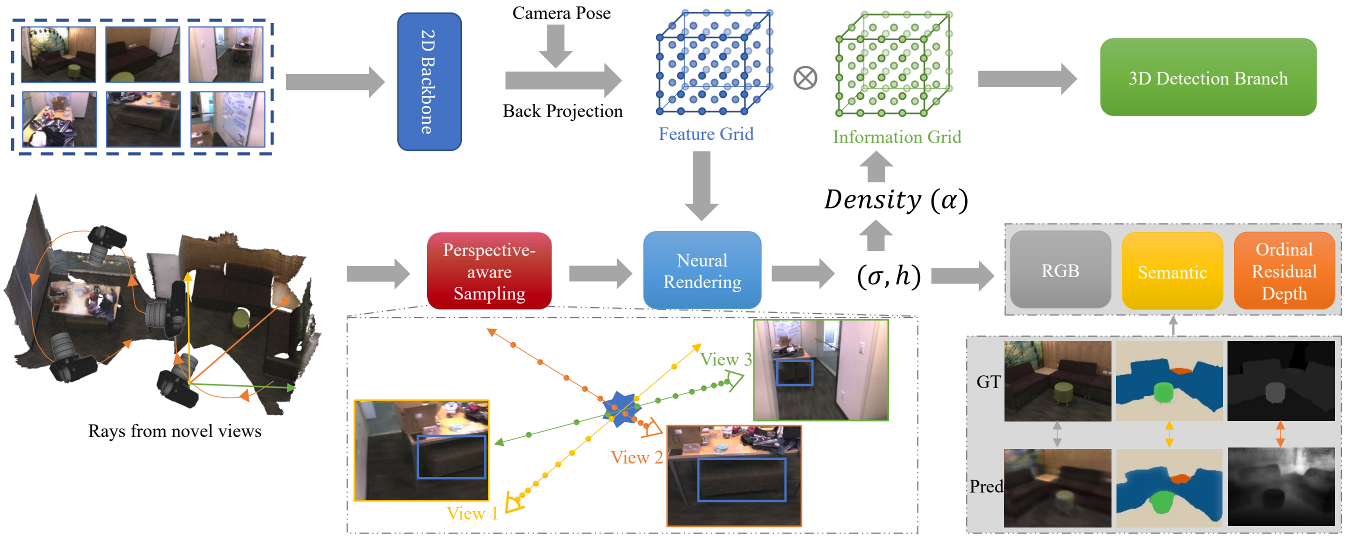


Figure 2: **Schematic overview of our NeRF-Det++.** The framework comprises two branches, i.e., detection and neural rendering. For the detection side, given the multi-view images, we utilize the 2D backbone to extract discriminative features. The camera pose of each image is used to back-project these 2D features to the 3D space, producing the 3D feature grid. As for the rendering branch, we design the Perspective-aware Sampling policy to concentrate on sampling points of more prominent regions. Two MLPs, *i.e.*, the Φ_{G} and Φ_{C}, are employed to estimate the volumetric density and color of the sampled points, respectively. To enhance the semantic awareness of the detection features, we introduce the Semantic Enhancement module Φ_{S}, which applies semantic supervision to the 2D views. The Φ_{G} is shared and used to produce the opacity field multiplied by the 3D feature grid to generate the geometry-enhanced features. Ultimately, the reinforced features are fed to the detection head to yield the detection outputs.

semantic enhancement module to incorporate high-level semantic information while constructing the 3D volume explicitly. Concretely, we integrate the semantic branch Φ_{S} after the geometric module Φ_{G}. As depicted in Fig. 2, the hidden features $\mathbf{h}(\mathbf{x})$, generated by Φ_{G}, are then fed into the Φ_{S} module, producing the semantic predictions:

$$\mathbf{s} = \mathbf{\Phi}_{\text{S}}(\mathbf{x}, \mathbf{d}, \mathbf{h}(\mathbf{x})), \tag{1}$$

where $\Phi_{\text{G}}(.)$ denotes the geometric multi-layer perceptron (MLP), $\mathbf{d}$ is the view direction, $\mathbf{x}$ is the coordinate of the sampled point and $\mathbf{h}(\mathbf{x}) = \mathbf{\Phi}_{\text{G}}(\mathbf{F}, \mathbf{x})$. We use the cross entropy loss to supervise the learning of semantic maps:

$$\mathcal{L}_{\text{Seg}}(\mathbf{s}, \hat{\mathbf{s}}) = \text{CE}(\mathbf{s}, \hat{\mathbf{s}}). \tag{2}$$

Here, $\hat{\mathbf{s}}$ is the ground-truth semantic label. By incorporating a semantic branch and leveraging the Φ_{G} and Φ_{S} modules, our method enables the 3D feature volume to capture and effectively utilize valuable semantic information, thus increasing the semantic awareness of the multi-view detector. Note that the semantic branch is exclusively utilized during training and imposes no additional burden during testing. Therefore, our approach enhances the detection performance without incurring extra computational costs.

3.3 Perspective-aware Depth Supervision

Depth learning is an important component of the multi-view detector as it serves as the bridge to connect the 2D perspective view and the 3D space. Recall that NeRF-Det assigns the same penalty to all regions for the depth estimation. However, since distant objects with fewer visual cues often lead to higher losses, the depth loss of distant areas will dominate the training process, making the depth learning of nearby regions ineffective. Moreover, objects far away in one view can be close in another. The multi-view complementation indicates that these distant objects can benefit from more accurate depth values from alternative perspectives. Consequently, we propose prioritizing the depth learning on close objects and reducing the weight of distant objects.

To tackle the preceding problems, we propose enhancing the depth learning of the multi-view detector from two perspectives, *i.e.*, the loss function, and the sampling policy. Remember that NeRF-Det directly regresses the ground-truth depth values. However, the direct regression of the continuous depth values is extremely difficult. Instead, we transform the original regression prediction of continuous depth values into the classification prediction of discrete depth bins and the regression of the continuous residual depth values. This practice draws inspiration from the depth sampling policy in monocular 3D detection [Fu *et al.*, 2018; Tang *et al.*, 2020], and we redesign it to suit the indoor 3D detection task better, as presented in Eqn. 4 and 5. Compared to monocular 3D detection, the depth range of indoor scenes is relatively limited. Hence, we divide the depth subtly, and N is the number of divided bins.

Perspective-aware Sampling. Considering that distant information is challenging to learn and may also contain inaccuracies, we replace the Uniformly Sampling (US) of NeRF-Det with the perspective-aware sampling strategy. The original uniform sampling strategy is given as follows:

$$l_{\text{US}} = (N-1)\frac{z - z_{\text{min}}}{z_{\text{max}} - z_{\text{min}}}, \tag{3}$$

where z denotes the depth, $z_{\max}$ and $z_{\min}$ represents the maximum and minimum depth of each scene.

To fully harness the precious visual clues in the nearby areas, we propose Logarithmic Increment Sampling (LgIS), which applies logarithmic increments based on the US. The mathematical formulation of LgIS is shown as follows:

$$l_{\text{LgIS}} = (N-1)\frac{\log z - \log z_{min}}{\log z_{\max} - \log z_{\min}}. \quad (4)$$

This sampling policy allows the network to allocate its learning resources more reasonably and prioritize the parts conducive to initial learning. Additionally, we propose Linear Increment Sampling(LnIS), introducing linear increments to guide the attention of the network toward more easily learned parts. The specific formulation is given below:

$$\begin{aligned} l_{\text{LnIS}} &= -0.5 + 0.5\sqrt{1+4\delta}, \\ \text{where} \quad \delta &= N(N-1)\frac{z - z_{\min}}{z_{\max} - z_{\min}}. \end{aligned} \quad (5)$$

The LgIS and LnIS sampling policies constitute our perspective-aware sampling that prioritizes nearby areas while de-emphasizing distant areas.

Ordinal Residual Depth Supervision. We can encode an instance depth z in $l_{\text{int}} = \lfloor l \rfloor$ ordinal bins and the estimated depth is decomposed as $z = z_{l_{\text{int}}} + z_{\text{res}}$, where $\lfloor . \rfloor$ is the floor operation. Furthermore, considering the online learning aspect of the rendering part in NeRF, we have made corresponding adjustments to the depth loss, considering the characteristics of LnIS and LgIS. The proposed Ordinal Residual Depth loss comprises the classification of ordinal depth bins and the regression of the residual depth value, as presented in Eqn. 6.

$$\mathcal{L}_{\text{Depth}}(z, \hat{z}) = \text{CE}(l_{\text{int}}, \hat{l}_{\text{int}}) + \gamma \text{L1}(z_{\text{res}}, \hat{z}_{\text{res}}), \quad (6)$$

where γ is the weight of residual depth loss.

Besides, following the multi-level geometry reasoning framework in NeRF, we further introduce a fine sub-network that aligns with the structure of the coarse sub-network, allowing for more precise sampling.

3.4 Training Objective

Regarding the detection branch, we follow NeRF-Det [Xu *et al.*, 2023] and employ the conventional detection loss $\mathcal{L}_{\text{Det}}$, comprised of the center regression loss, the box size regression loss, and the categorical classification loss. As for the rendering branch, we use a photo-metric loss $\mathcal{L}_{\text{RGB}}$ to learn the low-level appearance information, a geometric loss $\mathcal{L}_{\text{Geo}}$ to learn geometry, and a semantic segmentation loss $\mathcal{L}_{\text{Seg}}$ to learn the high-level semantic information. When the ground-truth depth is used, the depth estimation loss $\mathcal{L}_{\text{Depth}}$, computed by Eqn. 6, can be further utilized. The overall training objective is shown in Eqn. 7.

$$\mathcal{L} = \mathcal{L}_{\text{Det}} + \mathcal{L}_{\text{RGB}} + \mathcal{L}_{\text{Geo}} + \mathcal{L}_{\text{Seg}} + \mathcal{L}_{\text{Depth}}. \quad (7)$$

We empirically find that setting all the loss coefficients to one yields the best performance.

4 Experiments

Dataset. We perform experiments in two popular indoor 3D detection datasets, *i.e.*, ScanNetV2 [Dai *et al.*, 2017] and ARKITScenes [Baruch *et al.*, 2021]. ScanNetV2 contains $1,513$ complex indoor scenes with approximately $2.5M$ RGB-D frames and is annotated with semantic and instance segmentation for 18 object categories. Since ScanNetV2 does not provide amodal or oriented bounding box annotation, we follow NeRF-Det [Xu *et al.*, 2023] and predict axis-aligned bounding boxes instead. The input image resolution is 320×240. We mainly evaluate the methods by mAP with 0.25 IoU and 0.5 IoU threshold, denoted by mAP@.25 and mAP@.50. ARKITScenes contains around 1.6 K rooms with more than $5,000$ scans. Each scan includes a series of RGB-D posed images. Since some labels about the sky direction of each video are inaccurate [Xie *et al.*, 2023], we rectify the images according to the metadata and use the videos with all sky directions except "Left". On ARKITScenes, the input image resolution is 256×192, and the image feature map size is 64×48. We follow the official dataset partition and adopt mAP@.25 and mAP@.50 as the evaluation metric.

4.1 Quantitative results

Table 1 summarizes the comparison between our NeRF-Det++ and state-of-the-art methods on the validation set of ScanNetV2. Without bells and whistles, our method outperforms all multi-view 3D detectors regarding mAP@.25 and mAP@.50. For instance, our NeRF-Det++ significantly outperforms NeRF-Det with ResNet-50 by $+1.9\%$ and $+3.5\%$ in terms of mAP@.25 and mAP@.50, respectively. We also provide experimental results on the ARKITScenes dataset, as presented in Table 2. The performance gap between our NeRF-Det++ and NeRF-Det is $+0.6\%$ in mAP@.25 and $+0.9\%$ in mAP@.50. The consistent improvements observed in two popular indoor 3D detection benchmarks demonstrate the robustness and versatility of our approach.

4.2 Ablation study

We perform comprehensive ablation studies on the ScanNetV2 dataset to evaluate the effectiveness of each component in our method. Our chosen baseline is NeRF-Det with ResNet-50. We assess the efficacy of each component, both with and without depth supervision. To ensure a fair and consistent comparison, we maintain the experimental settings of NeRF-Det throughout the evaluation process.

Effect of Semantic Enhancement Module. The performances are presented in Table 3, indicating that including semantic cues improves the performance of 3D object detection. The semantic enhanced module enriches the 3D feature volume with high-level semantic information. This increased availability of feature information facilitates more accurate 3D detection and mitigates the occurrence of erroneous examples caused by inaccurate categorization in NeRF-Det.

Effect of Sampling Strategy and Depth Loss. We investigate different sampling strategies and depth losses, as shown in Tables 4 and 5. It shows that LgIS and LnIS, which have larger bin sizes in more significant depths, perform better. Additionally, we test inverse depth using uniform sampling

Table 1: Quantitative comparison on ScanNetV2. The first block shows point-cloud-based and RGBD-based methods; the others are multi-view RGB-only methods. ∗ indicates the method with depth supervision. † means our retest follows the official project. The evaluation of each category is based on mAP@0.25. We use **bold** to indicate our method outperforms other approaches under the same configuration.

Methods	cab	bed	chair	sofa	table	door	wind	bkshf	pic	cntr	desk	curt	fridge	shower	toil	sink	bath	ofurn	**mAP**@.25	**mAP**@.50
Seg-Cluster [Wang *et al.*, 2018]	11.8	13.5	18.9	14.6	13.8	11.1	11.5	11.7	0.0	13.7	12.2	12.4	11.2	18.0	19.5	18.9	16.4	12.2	13.4	-
Mask R-CNN [He *et al.*, 2017]	15.7	15.4	16.4	16.2	14.9	12.5	11.6	11.8	19.5	13.7	14.4	14.7	21.6	18.5	25.0	24.5	24.5	16.9	17.1	10.5
SGPN [Wang *et al.*, 2018]	20.7	31.5	31.6	40.6	31.9	16.6	15.3	13.6	0.0	17.4	14.1	22.2	0.0	0.0	72.9	52.4	0.0	18.6	22.2	-
3D-SIS [Hou *et al.*, 2019]	19.8	69.7	66.2	71.8	36.1	30.6	10.9	27.3	0.0	10.0	46.9	14.1	53.8	36.0	87.6	43.0	84.3	16.2	40.2	22.5
VoteNet [Qi *et al.*, 2019]	36.3	87.9	88.7	89.6	58.8	47.3	38.1	44.6	7.8	56.1	71.7	47.2	45.4	57.1	94.9	54.7	92.1	37.2	58.7	33.5
FCAF3D [Rukhovich *et al.*, 2022a]	57.2	87.0	95.0	92.3	70.3	61.1	60.2	64.5	29.9	64.3	71.5	60.1	52.4	83.9	99.9	84.7	86.6	65.4	71.5	57.3
CAGroup3D [Wang *et al.*, 2022a]	60.4	93.0	95.3	92.3	69.9	67.9	63.6	67.3	40.7	77.0	83.9	69.4	65.7	73.0	100	79.7	87.0	66.1	75.12	61.3
ImVoxelNet-R50	34.5	83.6	72.6	71.6	54.2	30.3	14.8	42.6	0.8	40.8	65.3	18.3	52.2	40.9	90.4	53.3	74.9	33.1	48.4	23.7
NeRF-Det-R50	37.2	84.8	75.0	75.6	51.4	31.8	20.0	40.3	0.1	51.4	69.1	29.2	58.1	61.4	91.5	47.8	75.1	33.6	52.0	26.1
NeRF-Det++-R50(Ours)	34.8	**86.3**	**75.7**	**79.9**	**56.9**	**34.3**	**25.4**	**53.3**	**2.3**	**51.8**	**70.5**	**32.6**	**54.3**	51.3	**92**	**53.3**	**80.7**	**34.1**	**53.9(+1.9)**	**29.6(+3.5)**
NeRF-Det-R50∗	37.7	84.1	74.5	71.8	54.2	34.2	17.4	51.6	0.1	54.2	71.3	16.7	54.5	55.0	92.1	50.7	73.8	34.1	51.8	27.4
NeRF-Det++-R50∗(Ours)	**38.9**	83.9	73.9	**77.6**	**57.2**	33.3	**23.5**	47.9	**1.51**	**56.9**	**77.7**	**21.1**	**61.5**	46.8	**92.8**	49.2	**80.2**	**34.5**	**53.2(+1.4)**	**29.6(+2.2)**
ImVoxelNet-R101	30.9	84.0	77.5	73.3	56.7	35.1	18.6	47.5	0.0	44.4	65.5	19.6	58.2	32.8	92.3	40.1	77.6	28.0	49.0	-
NeRF-Det-R101	36.8	85.0	77.0	73.5	56.9	36.7	14.3	48.1	0.8	49.7	68.3	23.5	54.0	60.0	96.5	49.3	78.4	38.4	52.9	-
NeRF-Det++-R101(Ours)	36.1	82.9	74.9	**79.1**	**57.0**	**37.3**	**24.9**	**54.6**	**2.4**	**51.7**	**72.2**	**25.5**	**58.7**	51.5	92.7	**50.8**	**82.2**	35.1	**53.9(+1.0)**	**30.0**
NeRF-Det-R101∗	37.6	84.9	76.2	76.7	57.5	36.4	17.8	47	2.5	49.2	52	29.2	68.2	49.3	97.1	57.6	83.6	35.9	53.3	-
NeRF-Det-R101∗†	38.7	81.3	75.7	76.8	58.1	33.5	23.7	42.1	4.5	56.6	68.2	26.6	51.5	47.4	85.3	49.3	73.1	30.8	51.3	27.4
NeRF-Det++-R101∗(Ours)	38.7	**85.0**	73.2	**78.1**	56.3	**35.1**	22.6	**45.5**	1.9	50.7	**72.6**	26.5	**59.4**	**55.0**	**93.1**	**49.7**	**81.6**	**34.1**	**53.3(+2.0)**	**30.0(+2.6)**

Table 2: Comparisons on ARKITScenes val set.

Methods	**scene**	**mAP**@.25	**mAP**@.50
ImVoxelNet-R50	whole-scene	23.6	-
NeRF-Det-R50	whole-scene	26.7	-
	except “Left”	42.7	26.2
NeRF-Det++-R50	except “Left”	**43.3(+0.6)**	**27.1+(0.9)**

Table 3: Ablation study of Semantic Enhancement module.

Methods	w/o depth		w/ depth	
	mAP@.25	**mAP**@.50	**mAP**@.25	**mAP**@.50
NeRF-Det-R50	52.0	26.1	51.8	27.4
+ Semantic Enhancement	**53.4(+1.4)**	**27.9(+1.8)**	**52.3(+0.5)**	**28.3(+0.9)**

Table 5: Ablation study of depth loss.

Methods	**mAP**@.25	**mAP**@.50
L1 loss	51.8	27.4
Huber loss	**52.9(+1.1)**	27.1(-0.3)
Ordinal Residual Depth loss	**52.8(+1.0)**	**27.5(+0.1)**

Table 6: Ablation study of the fine sub-network.

Methods	w/o depth		w/ depth	
	mAP@.25	**mAP**@.50	**mAP**@.25	**mAP**@.50
NeRF-Det-R50	52.0	26.1	51.8	27.4
+ fine sub-network	**52.6(+0.6)**	**27.4(+1.3)**	**52.0(+0.2)**	**28.7(+1.3)**

(UIS). Although its bin size is increased, the performance is reduced. Due to the squared relationship between increasing bin size and depth, UIS ignores distance information.

For depth loss, we investigate three losses: L1 loss, Huber loss, and the proposed Ordinal Residual Depth loss. Huber loss combines the benefits of L1 and L2 Loss; the incrementally scaled L1 region reduces sensitivity to outliers, while the L2 region provides smoothness. Consequently, we utilize L1 loss to minimize the impact of depth loss in long-distance while employing L2 loss for short-distance to enhance smoothness and capture finer details. The Ordinal Residual Depth loss we design discretizes depth and divides it into a bin classification problem and a residual regression problem, simplifying depth learning. Additionally, based on perceptive-aware sampling, Ordinal Residual Depth loss prioritizes nearby objects. The experimental results demonstrate that our proposed Ordinal Residual Depth loss enhances depth learning, improving 3D detection performance.

Effect of Fine Sub-network. Building upon NeRF and its variants, we incorporate a fine sub-network into our approach. The structure of the fine sub-network mirrors that of the coarse sub-network. It performs a finer sampling within the attention area identified by the coarse sub-network. Our experimental results demonstrate that including the fine sub-network yields significant performance improvements.

Effect of Relative Depth Map. We study whether relative depth will affect performance, as shown in Table 7. The experiment shows that relative depth can improve performance significantly, over 1.1% in mAP@.25 and 0.2% in mAP@.50 compared to absolute depth. As depth prediction is a complex problem, we convert a prediction of absolute depth into relative depth, simplifying the problem.

Table 4: Ablation study of sampling strategies.

Methods	w/o depth		w/ depth	
	mAP@.25	**mAP**@.50	**mAP**@.25	**mAP**@.50
US	52.0	26.1	51.8	27.4
UIS	51.8(-0.2)	25.8(-0.3)	51.7(-0.1)	26.4(-1.0)
LgIS	**53.2(+1.2)**	**26.3(+0.2)**	**52.4(+0.6)**	**28.8(+1.4)**
LnIS	**53.0(+1.0)**	**27.1(+1.0)**	**53.2(+1.4)**	**29.4(+2.0)**

Table 7: Ablation study of depth normalization.

Methods	**mAP**@.25	**mAP**@.50
NeRF-Det-R50*	51.8	27.4
+ depth normalization	**52.9(+1.1)**	**27.6(+0.2)**

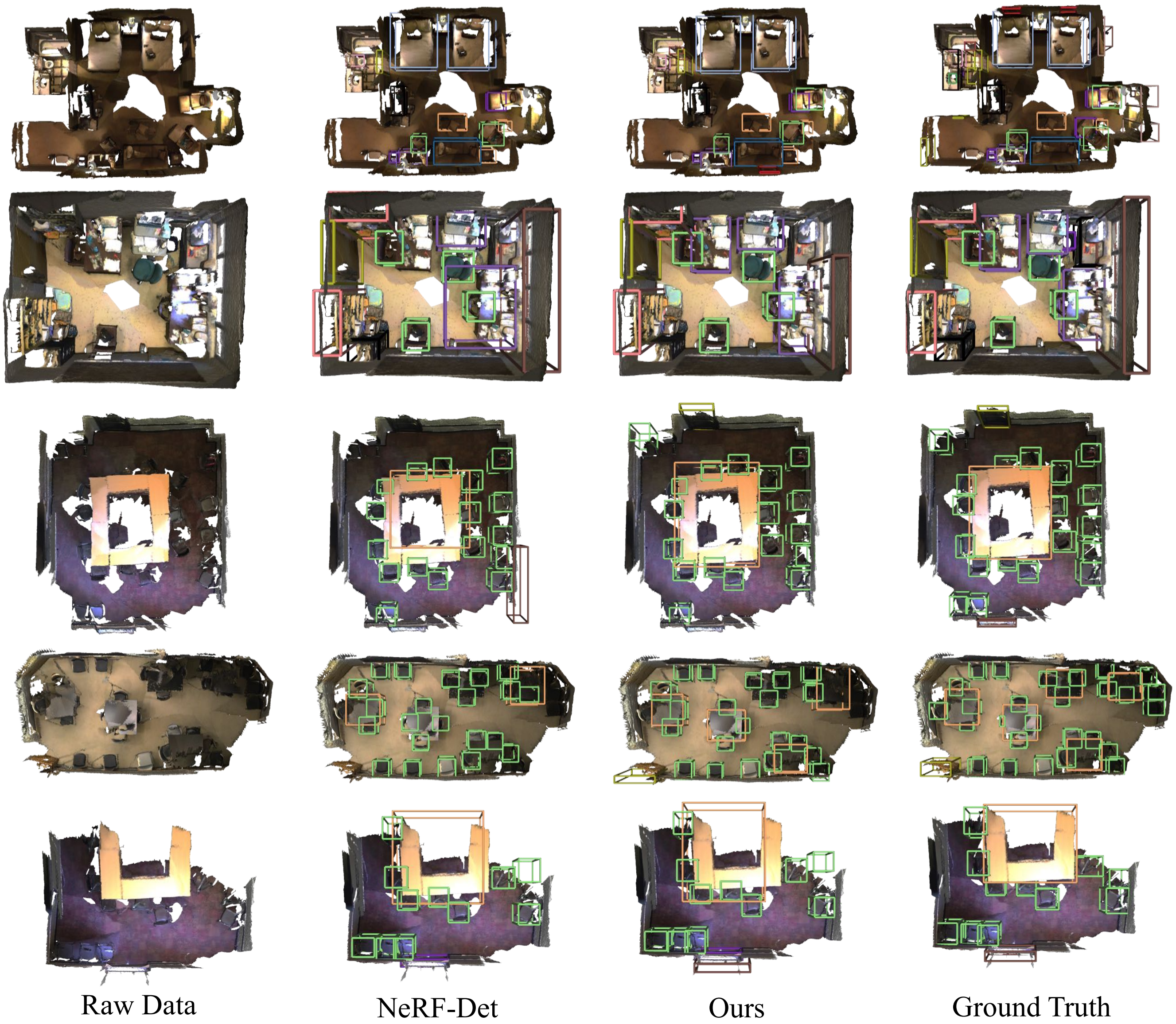


Figure 3: **Visual comparison between NeRF-Det and NeRF-Det++.** Note that our approach only takes posed RGB images as input. The reconstructed mesh is only used for visualization.

4.3 Qualitative results

We present visualizations of the predictions made by NeRF-Det and our enhanced NeRF-Det++, as illustrated in Fig. 3. Our method effectively handles diverse challenges, including scenes with high object density, severe occlusion, and significant morphological diversity. Notably, we rectify specific incorrect predictions made by NeRF-Det. For instance, we successfully resolved the issue of missing bounding boxes in the first, second, and fourth scenes. Furthermore, we overcome misclassification in the fifth scene, where the "curtain" is mistakenly identified as the "garbage bin". Additionally, we favorably alleviate incorrect object dimensions and positional deviations observed in the second and third scenes, respectively. Lastly, in the third scene, our enhanced NeRF-Det++ avoids the false positive of including the "curtain" object, which NeRF-Det mistakenly predicts.

5 Conclusion

In conclusion, this paper presents NeRF-Det++, a novel approach for indoor 3D detection from multi-view images. We identify and address three critical flaws in NeRF-Det. Firstly, to tackle semantic ambiguity, we introduce the Semantic Enhancement module that utilizes semantic supervision for improved classification. Secondly, to address inappropriate sampling, we prioritize nearby objects and leverage the characteristics of multi-views through the design of Perspective-aware Sampling. Lastly, we tackle the issue of insufficient utilization of depth supervision by proposing Ordinal Residual Depth Supervision, which incorporates classification of ordinal depth bins and regression of residual depth values. Extensive experiments conducted on the ScanNetV2 and ARKITScenes validate the superiority of our NeRF-Det++.

References

[Barron *et al.*, 2021] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.

[Baruch *et al.*, 2021] 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.

[Carion *et al.*, 2020] 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*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2023] Dian Chen, Jie Li, Vitor Guizilini, Rares Andrei Ambrus, and Adrien Gaidon. Viewpoint equivariance for multi-view 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9213–9222, 2023.

[Dai *et al.*, 2017] 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.

[Fu *et al.*, 2018] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.

[Grabocka and Schmidt-Thieme, 2018] Josif Grabocka and Lars Schmidt-Thieme. Neuralwarp: Time-series similarity with warping networks. *arXiv preprint arXiv:1812.08306*, 2018.

[Guo *et al.*, 2020] Yulan Guo, Hanyun Wang, Qingyong Hu, Hao Liu, Li Liu, and Mohammed Bennamoun. Deep learning for 3d point clouds: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 43(12):4338–4364, 2020.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Hou *et al.*, 2019] Ji Hou, Angela Dai, and Matthias Nießner. 3d-sis: 3d semantic instance segmentation of rgb-d scans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4421–4430, 2019.

[Hu *et al.*, 2023] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, and Chi-Keung Tang. Nerf-rpn: A general framework for object detection in nerfs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23528–23538, 2023.

[Huang *et al.*, 2021] Xiaoshui Huang, Guofeng Mei, Jian Zhang, and Rana Abbas. A comprehensive survey on point cloud registration. *arXiv preprint arXiv:2103.02690*, 2021.

[Li *et al.*, 2023] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023.

[Liu *et al.*, 2022] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[Mildenhall *et al.*, 2021] 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.

[Qi *et al.*, 2018] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 918–927, 2018.

[Qi *et al.*, 2019] Charles R Qi, Or Litany, Kaiming He, and Leonidas J Guibas. Deep hough voting for 3d object detection in point clouds. In *proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9277–9286, 2019.

[Rukhovich *et al.*, 2022a] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Fcaf3d: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pages 477–493. Springer, 2022.

[Rukhovich *et al.*, 2022b] 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.

[Tang *et al.*, 2020] Yunlei Tang, Sebastian Dorn, and Chiragkumar Savani. Center3d: Center-based monocular 3d object detection with joint depth understanding. In *DAGM German Conference on Pattern Recognition*, pages 289–302. Springer, 2020.

[Vaswani *et al.*, 2017] 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*, 30, 2017.

[Wang *et al.*, 2018] Weiyue Wang, Ronald Yu, Qiangui Huang, and Ulrich Neumann. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2569–2578, 2018.

[Wang *et al.*, 2021a] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021.

[Wang *et al.*, 2021b] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021.

[Wang *et al.*, 2022a] Haiyang Wang, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, Liwei Wang, et al. Cagroup3d: Class-aware grouping for 3d object detection on point clouds. *Advances in Neural Information Processing Systems*, 35:29975–29988, 2022.

[Wang *et al.*, 2022b] 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.

[Wang *et al.*, 2023a] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3295–3306, 2023.

[Wang *et al.*, 2023b] Yingjie Wang, Qiuyu Mao, Hanqi Zhu, Jiajun Deng, Yu Zhang, Jianmin Ji, Houqiang Li, and Yanyong Zhang. Multi-modal 3d object detection in autonomous driving: a survey. *International Journal of Computer Vision*, pages 1–31, 2023.

[Xie *et al.*, 2023] Yiming Xie, Huaizu Jiang, Georgia Gkioxari, and Julian Straub. Pixel-aligned recurrent queries for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18370–18380, 2023.

[Xu *et al.*, 2023] 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.

[Yariv *et al.*, 2021] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.

[Ye *et al.*, 2023] Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

[Zhang *et al.*, 2022] Jingyang Zhang, Yao Yao, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Critical regularizations for neural surface reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6270–6279, 2022.