

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

Runmin Zhang¹ Zhu Yu^{1*} Si-Yuan Cao^{2,3*} Lingyu Zhu⁴
 Guangyi Zhang¹ Xiaokai Bai¹ Hui-Liang Shen¹

¹College of Information Science and Electronic Engineering, Zhejiang University

²Ningbo Global Innovation Center, Zhejiang University ³NingboTech University ⁴City University of Hong Kong

{runmin_zhang, yu_zhu, cao_siyuan}@zju.edu.cn, lingyuzhu-c@my.cityu.edu.hk,

{zhangguangyi, shawnnkb, shenhl}@zju.edu.cn

Abstract

This work presents *SGCDet*, a novel multi-view indoor 3D object detection framework based on adaptive 3D volume construction. Unlike previous approaches that restrict the receptive field of voxels to fixed locations on images, we introduce a geometry and context aware aggregation module to integrate geometric and contextual information within adaptive regions in each image and dynamically adjust the contributions from different views, enhancing the representation capability of voxel features. Furthermore, we propose a sparse volume construction strategy that adaptively identifies and selects voxels with high occupancy probabilities for feature refinement, minimizing redundant computation in free space. Benefiting from the above designs, our framework achieves effective and efficient volume construction in an adaptive way. Better still, our network can be supervised using only 3D bounding boxes, eliminating the dependence on ground-truth scene geometry. Experimental results demonstrate that *SGCDet* achieves state-of-the-art performance on the *ScanNet*, *ScanNet200* and *ARKitScenes* datasets. The source code is available at <https://github.com/RM-Zhang/SGCDet>.

1. Introduction

Indoor 3D object detection is a fundamental 3D perception task, with broad applications in embodied AI, AR/VR, and robotics. Leveraging precise scene geometry as input, point cloud-based 3D object detectors [11, 18, 26, 28, 32, 43] have achieved impressive performance. However, capturing accurate scene geometry typically requires high-cost 3D sensors. Recently, there has been a shift towards using multi-view posed images for 3D object detection.

*Corresponding authors.

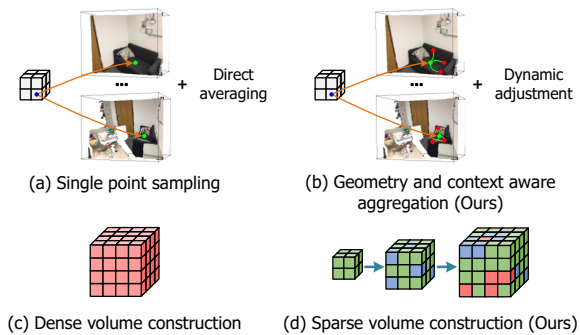


Figure 1. Comparison of feature lifting and volume construction strategies between previous approaches and our *SGCDet*. (a) The single point sampling strategy used in previous approaches restricts the receptive field of voxels to a limited region, neglecting contextual information across multiple views. (b) Our geometry and context aware aggregation adaptively integrates geometric and contextual features within deformable regions across different views, enhancing the representational capability of voxel features. (c) Previous approaches construct high-resolution, dense 3D volumes without considering the inherent sparsity of 3D scenes, leading to unnecessary computational overhead. (d) Our sparse volume construction adaptively refines voxels that are likely to contain objects, reducing redundant computational cost in free space.

To bridge the gap between 2D images and 3D representations, the pioneering work *ImVoxelNet* [29] lifts 2D features along the overall ray. Each voxel then adopts the averaged results across multiple views as its features. However, this approach uses the same weights for features derived from different images, leading to a coarse and error-prone 3D voxel representation. Although the following works [31, 38] introduce an opacity probability to suppress voxel features in free space through post-processing, they still fail to address the occlusion issue during the 2D-to-3D projection process. More recent approaches [30, 40] introduce explicit geometry constraints to assist the feature lifting. Nevertheless, the final performance of these

approaches heavily depends on the accuracy of the estimated geometric information, either complicating the training pipeline or significantly increasing computational cost.

As analyzed above, previous approaches primarily enhance the quality of 3D voxel representations from a geometric perspective, overlooking the valuable contextual information of images. The sampling locations on 2D feature maps are constrained to fixed positions determined by the predefined voxel centers and camera poses. This single point sampling strategy limits the receptive field of voxels to a small region, restricting their ability to perceive visual information. In addition, this strategy further amplifies the dependency on accurate geometric information [30, 40], as illustrated in Fig. 1(a).

To address these issues, we propose a **geometry and context aware aggregation** module to adaptively lift the 2D features. Instead of simply performing a weighted average of the sampled features across multi-view images, we take the sampled features as queries to aggregate relevant geometric and contextual features within a deformable region. Furthermore, we introduce a multi-view attention mechanism to dynamically adjust the contributions from different views, enhancing the representation capabilities of the transformed 3D volumes, as illustrated in Fig. 1(b).

On the other hand, previous approaches generally construct high-resolution, dense 3D volumes, as shown in Fig. 1(c). This dense representation fails to account for the inherent sparsity of 3D scenes, leading to unnecessary computational overhead. To address this issue, we propose a **sparse volume construction** strategy that constructs 3D volumes in an adaptive manner, as illustrated in Fig. 1(d). Specifically, we employ an occupancy prediction module to identify voxels likely to contain objects for refinement, thereby reducing redundant computations in free space. A critical aspect of this strategy is the supervision of occupancy prediction. While a straightforward solution is to directly use ground-truth geometry for supervision [30, 31], it is infeasible when such data is unavailable [1, 40]. To eliminate reliance on ground-truth geometry, we leverage 3D bounding boxes to generate pseudo labels for occupancy, achieving flexible network supervision.

By combining the Sparse volume construction and the Geometry and Context aware aggregation, we propose a novel framework for multi-view indoor 3D object Detection, named **SGCDet**. Thanks to above designs, SGCDet performs effective and efficient 3D volume construction. We evaluate our SGCDet on the ScanNet [6], ScanNet200 [27], and ARKitScenes [2] datasets. SGCDet achieves state-of-the-art performance among approaches that do not rely on ground-truth geometry for supervision. Compared to the previous state-of-the-art approach MVSDet [40], SGCDet significantly improves mAP@0.5 by 3.9 on ScanNet, while reducing training memory, training time,

inference memory, and inference time by 42.9%, 47.2%, 50%, and 40.8%, respectively. Remarkably, SGCDet also surpasses some approaches that use ground-truth geometry during training.

Our contributions are summarized as follows:

- We propose the geometry and context aware aggregation module to enhance feature lifting. It enables each voxel to adaptively aggregate geometric and contextual features within a deformable region, and dynamically adjusts feature contributions across different views.
- We introduce the sparse volume construction strategy, which adaptively refines voxels likely to contain objects, reducing computations in free space. Notably, the overall network can be supervised using only 3D bounding boxes, eliminating the need for ground-truth geometry.
- Extensive experiments demonstrate that SGCDet outperforms the previous state-of-the-art approach by a large margin, while significantly reducing computational overhead. These results validate both the effectiveness and efficiency of SGCDet.

2. Related Works

Image-based 3D Object Detection. Image-based 3D object detection has gained significant attention due to its cost-effectiveness and fine-grained visual perception capabilities. Current approaches primarily focus on constructing 3D representations from input images, including bird’s-eye-view (BEV) [10, 12, 14, 16] and voxel-based approaches [29–31, 34, 38, 40]. Given the variability in camera viewpoints and object distributions, voxel-based representations are better suited for indoor scenes. ImVoxelNet [29] is the pioneer that introduces an end-to-end pipeline for multi-view indoor 3D object detection. It directly lifts 2D features along 3D rays without incorporating scene geometry, leading to ambiguities in volume features. Building on ImVoxelNet, ImGeoNet [31] and NeRF-Det [38] compute opacity probabilities for the 3D volume to suppress features in free space. NeRF-Det++ [9] and GoN3RDet [17] further enhance NeRF-Det through semantic and geometric constraints. However, these approaches treat opacity solely as a post-processing step, and fail to address occlusion issues during the feature lifting process. Alternatively, recent approaches [30, 40] explicitly estimate scene geometry to achieve occlusion-aware projection. CN-RMA [30] combines a 3D reconstruction network with a point cloud-based 3D object detector, and uses a reconstructed TSDF to guide the feature lifting process. Nevertheless, it requires a time-consuming multi-stage training pipeline, and relies on ground-truth geometry for supervision. In contrast, MVSDet [40] leverages multi-view stereo to compute depth probabilities from input images, and applies 3D Gaussian Splatting [4] for self-supervision. However, it still suffers from high computational costs.

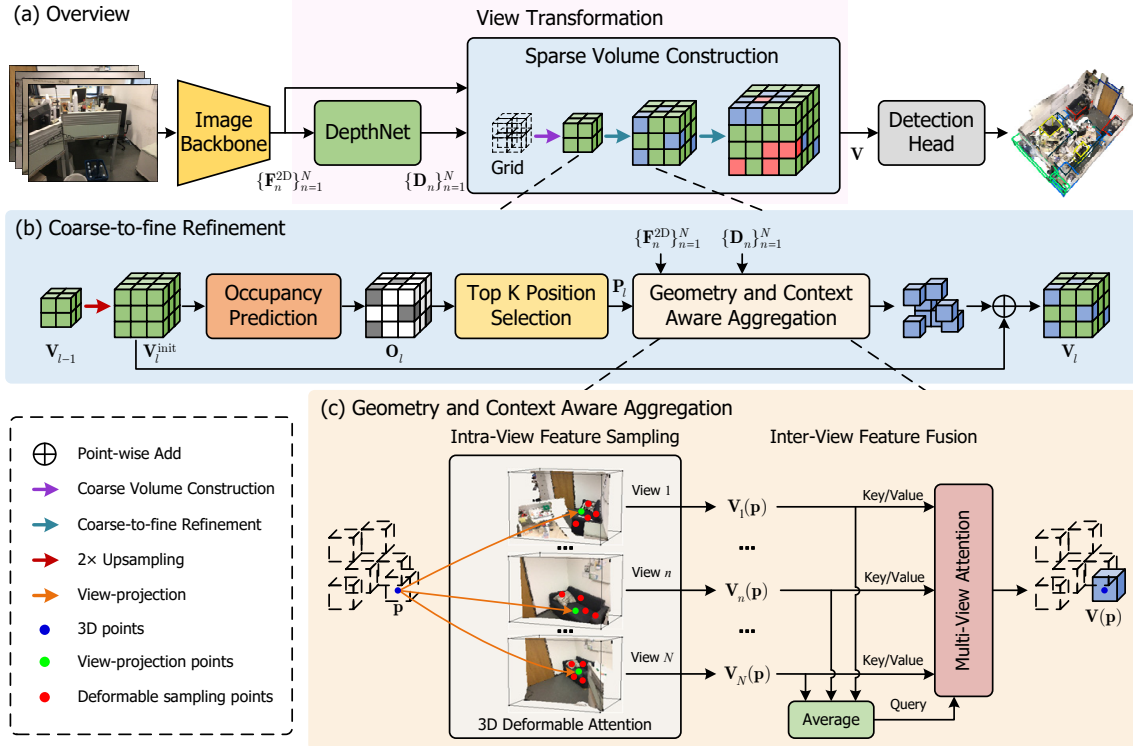


Figure 2. Schematics and detailed architectures of SGCDet. (a) Overview of SGCDet, which consists of an image backbone to extract image features, a view transformation module to lift image features to 3D volumes, and a detection head to predict 3D bounding boxes. (b) Details of the coarse-to-fine refinement in our sparse volume construction strategy. (c) Details of our geometry and context aware aggregation module.

Sparse Design in 3D Vision. Inspired by DETR [3], several 3D detection methods [20, 22, 35, 37] employ a sparse set of object queries to enable 3D-to-2D interaction. However, due to the absence of explicit 3D representations, these methods typically suffer from slow convergence. Other occupancy prediction approaches reduce the number of voxel queries through depth-based query proposal initialization [13, 15, 42], multi-scale sparse reconstruction [21, 25], or by reformulating the problem as a sparse set prediction [33]. While these methods have made notable progress, they still rely on precise geometry for supervision, limiting their applicability in scenarios where ground-truth geometric information is unavailable.

3. Method

3.1. Overview

Given N posed images $\{\mathbf{I}_n\}_{n=1}^N$ as input, SGCDet aims to predict 3D bounding boxes of the scene. As illustrated in Fig. 2(a), the overall framework of SGCDet consists of three main components: an image backbone that extracts 2D features $\{\mathbf{F}_n^{2D} \in \mathbb{R}^{H \times W \times C}\}_{n=1}^N$, a view transformation module that lifts these 2D features to 3D volumes $\mathbf{V} \in \mathbb{R}^{X \times Y \times Z \times C}$, and a detection head that predicts 3D bounding boxes. Here, (H, W) and (X, Y, Z) represent the

spatial resolution of 2D features and 3D volumes, respectively. C denotes the number of channels.

Our core design focuses on the view transformation module, which achieves adaptive 3D volume construction. Specifically, we adopt a simple yet effective **DepthNet** (Sec. 3.4) to estimate the depth distributions $\{\mathbf{D}_n \in \mathbb{R}^{H \times W \times D}\}_{n=1}^N$ for the input images, where D is the number of depth bins. The depth distributions provide geometric information for the view transformation process. To address the inefficiency of dense volume construction, we propose the **sparse volume construction** (Sec. 3.2) that adaptively builds the 3D volume in a coarse-to-fine manner. Within this process, we introduce the **geometry and context aware aggregation** (Sec. 3.3), which ensures adaptive feature lifting by integrating geometric and contextual information within a flexible region.

3.2. Sparse Volume Construction

Given the dense 3D grid $\mathbf{G} \in \mathbb{R}^{X \times Y \times Z \times 3}$, previous approaches [29, 31, 38, 40] typically project each 3D voxel to 2D features for volume construction. However, since most voxels in a 3D scene are free space, such dense construction is inefficient for object detection and incurs significant computational overhead. To address this issue, we introduce a sparse volume construction strategy that constructs

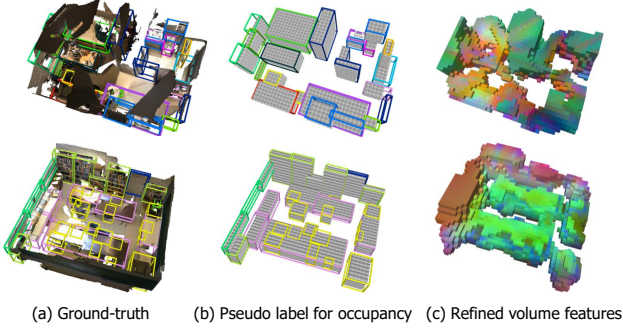


Figure 3. Visualization of our sparse volume construction. (a) Ground-truth 3D bounding boxes. (b) Pseudo-labels for occupancy supervision, generated from 3D bounding boxes. (c) Refined volume features. Our occupancy prediction network effectively filters free space, focusing the feature refinement on voxels that are likely to contain objects.

the 3D volume in a coarse-to-fine manner. As shown in Fig. 2(b), the key idea is to progressively upsample a coarse 3D volume, adaptively refining only the voxels that likely contain objects. Specifically, we first construct a coarse 3D volume $\mathbf{V}_0 \in \mathbb{R}^{\frac{X}{2^L} \times \frac{Y}{2^L} \times \frac{Z}{2^L} \times C}$ with a spatial resolution of $(\frac{X}{2^L}, \frac{Y}{2^L}, \frac{Z}{2^L})$. This coarse volume captures the overall scene geometry, and can be used to identify regions that may contain objects for further refinement.

Coarse-to-fine Refinement. The overall refinement process composes of L stages. As illustrated in Fig. 2(b), at the l -th stage, we first upsample the output volume of the $(l-1)$ -th stage by a factor of 2, obtaining $\mathbf{V}_l^{\text{init}} \in \mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}} \times C}$. Then, we estimate the occupancy probability of each voxel by

$$\mathbf{O}_l = \mathcal{F}(\mathbf{V}_l^{\text{init}}), \quad (1)$$

where $\mathbf{O}_l \in \mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}}}$ denotes the occupancy probability, and \mathcal{F} is a lightweight occupancy prediction head. Next, we select the positions with top- k occupancy probability for feature refinement, formulated as

$$\mathbf{V}_l = \mathbf{V}_l^{\text{init}} + \mathcal{P}(\mathbf{P}_l, \{\mathbf{F}_n^{2D}\}_{n=1}^N, \{\mathbf{D}_n\}_{n=1}^N), \quad (2)$$

where \mathbf{P}_l is the set of coordinates of the top $k\%$ points, and \mathcal{P} denotes the geometry and context aware aggregation described in Sec. 3.3. This strategy avoids redundant computation in free space, while effectively capturing fine structures in regions likely containing objects.

Supervision on Occupancy Probability. A straightforward way is to supervise the occupancy probability via ground-truth scene geometry. However, it is not feasible when the precise geometry is unavailable [1, 40]. To address this issue, we use the ground-truth 3D bounding boxes to generate pseudo labels for occupancy, providing a flexible supervision strategy. Given the 3D grid $\mathbf{G}_l \in$

$\mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}} \times 3}$ at the l -th stage, the ground truth occupancy probability $\mathbf{O}_{l,\text{gt}} \in \mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}}}$ is defined as:

$$\mathbf{O}_{l,\text{gt}}(x, y, z) = \begin{cases} 1, & \mathbf{G}_l(x, y, z) \text{ is inside any bounding box,} \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

The network is supervised by the binary cross entropy loss between \mathbf{O}_l and $\mathbf{O}_{l,\text{gt}}$. Fig. 3 displays two examples of the pseudo-labels generated from 3D bounding boxes and refined volume features in the last refinement stage. It can be observed that our sparse volume construction effectively captures scene geometry and adaptively focuses on regions containing objects.

3.3. Geometry and Context Aware Aggregation

A detailed diagram of our geometry and context aware aggregation is shown in Fig. 2(c). To obtain features for a voxel with center $\mathbf{p} = (x, y, z)^\top$, we first perform intra-view feature sampling to independently sample features from each view for initial information aggregation. Subsequently, we apply inter-view feature fusion to fuse the features from multiple views for further refinement.

Intra-view Feature Sampling. Previous methods [29–31, 38, 40] simply sample image features at the locations that derived from voxel centers and camera poses as the voxel features, limiting the receptive field of voxels. To address this problem, we introduce a 3D deformable attention mechanism [12] to incorporate geometric and contextual information within an adaptive region. Specifically, for each view n , we lift the 2D image features to a 3D pixel space, formulated as

$$\mathbf{F}_n^{3D} = \mathbf{F}_n^{2D} \otimes \mathbf{D}_n, \quad (4)$$

where $\mathbf{F}_n^{3D} \in \mathbb{R}^{H \times W \times D \times C}$ denotes the lifted 3D features, and \otimes refers to the outer product conducted at the last dimension. Next, we project \mathbf{p} to view n as

$$\mathbf{p}_n = (u_n, v_n, d_n)^\top = \mathbf{K}_n \mathbf{E}_n(\mathbf{p}, 1)^\top, \quad (5)$$

where \mathbf{p}_n is the coordinate in the 3D pixel space, \mathbf{K}_n and \mathbf{E}_n are the intrinsic and extrinsic matrices of view n , respectively. Instead of directly sampling \mathbf{F}_n^{3D} at \mathbf{p}_n as the voxel features $\mathbf{V}_n(\mathbf{p})$, we take the sampled features as queries to aggregate information from neighboring regions, formulated as

$$\begin{aligned} \mathbf{V}_n(\mathbf{p}) &= \text{DeformAttn}(\mathbf{p}_n, \phi(\mathbf{F}_n^{3D}, \mathbf{p}_n), \mathbf{F}_n^{3D}) \\ &= \sum_{m=1}^M A_{n,m} W \phi(\mathbf{F}_n^{3D}, \mathbf{p}_n + \Delta \mathbf{p}_{n,m}), \end{aligned} \quad (6)$$

where M is the number of sampled points, W is the matrix for value projection, and ϕ denotes the trilinear interpolation used to sample features from \mathbf{F}_n^{3D} . $\Delta \mathbf{p}_{n,m}$ and $A_{n,m}$

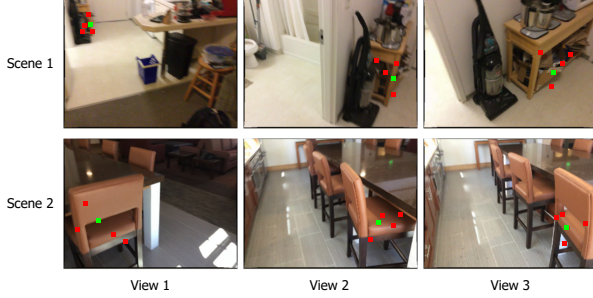


Figure 4. Visualization of sampling locations in our intra-view feature sampling. Green points represent positions derived from voxel centers and camera poses, while red points indicate deformable sampling locations. We note that the deformable attention is performed in the 3D pixel space, For clarity, the depth dimension is omitted in this visualization.

are the 3D offset and attention weight of the m -th sampled point, respectively, and they are generated from the query $\phi(\mathbf{F}_n^{3D}, \mathbf{p}_n)$ via a linear layer. For simplicity, we exclude the multi-head operation in Eq. 6. We show the locations of deformable sampling points across different views in Fig. 4. Compared to the single point sampling, our geometry and context aware aggregation effectively integrates geometric and contextual information within a flexible region, thus enhancing the representation capabilities of voxel features.

Inter-view Feature Fusion. Due to variations in object appearance and size across different views, the features $\{\mathbf{V}_n(\mathbf{p})\}_{n=1}^N$ sampled from different views may differ significantly. To adaptively adjust the contribution of each view, we propose a multi-view attention mechanism. Specifically, we use the average pooling of features from all views $\mathbf{V}_{\text{avg}}(\mathbf{p})$ as the query, while $\{\mathbf{V}_n(\mathbf{p})\}_{n=1}^N$ serves as both the key and value. This process can be formulated as

$$\mathbf{V}(\mathbf{p}) = \text{Attn}(\mathbf{V}_{\text{avg}}(\mathbf{p}), \{\mathbf{V}_n(\mathbf{p})\}_{n=1}^N, \{\mathbf{V}_n(\mathbf{p})\}_{n=1}^N), \quad (7)$$

where $\mathbf{V}(\mathbf{p})$ denotes the final features of the 3D point \mathbf{p} , and $\text{Attn}(\cdot)$ refers to the standard attention operation [7, 36]. Here, we assume that \mathbf{p} can be projected to all views for notation simplicity. In practice, we discard any views where \mathbf{p} is projected outside the image boundaries.

Discussions. Our geometry and context aware aggregation is highly inspired by DFA3D [12], upon which we introduce substantial modifications to better accommodate indoor scenes. DFA3D employs view-agnostic 3D queries to predict sampling offsets and weights across all views, which performs well in autonomous driving scenarios with fixed camera layouts. However, its effectiveness is limited in indoor environments, where camera poses vary significantly and objects exhibit large shape and scale differences across views. In contrast, our intra-view feature sampling leverages view-specific features as queries, enabling adaptive aggregation tailored to each individual view. Furthermore, our inter-view fusion module assigns learnable atten-

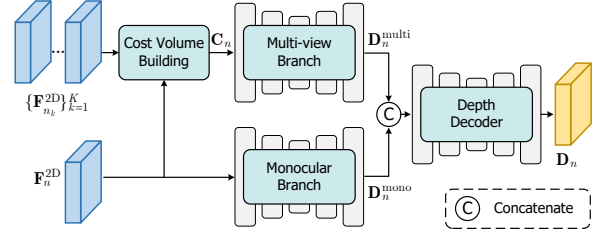


Figure 5. Detailed architecture of the DepthNet.

tion weights to each view’s contribution, resulting in more consistent and robust scene-level voxel representations.

3.4. DepthNet

The depth distributions provide geometric information for the 2D-to-3D projection process, whose accuracy significantly influences the final detection performance. To fully leverage multi-view images for accurate depth estimation, we introduce a simple yet effective DepthNet. As illustrated in Fig. 5, it fuses both multi-view and monocular depth features for depth estimation, where the former provides geometric properties through feature matching, and the latter contributes detailed structures of input images.

For any view n with 2D features \mathbf{F}_n^{2D} , we select the nearest K views with 2D features $\{\mathbf{F}_{n_k}^{2D}\}_{k=1}^K$, and use plane sweeping [5] to construct the cost volume. Specifically, we discretize the depth range $[d_{\min}, d_{\max}]$ into D depth bins as $[d_1, \dots, d_i, \dots, d_D]$. For each depth plane d_i , we warp the 2D features of nearby views to view n using camera matrices:

$$\mathbf{F}_{n_k, d_i}^{2D} = \mathcal{W}(\mathbf{F}_{n_k}^{2D}, d_i, \mathbf{K}_n, \mathbf{E}_n, \mathbf{K}_{n_k}, \mathbf{E}_{n_k}), \quad (8)$$

where \mathcal{W} is warping operation in [39, 41]. Then, we build the cost volume $\mathbf{C}_n \in \mathbb{R}^{H \times W \times D}$ as:

$$\mathbf{C}_n(h, w, i) = \frac{1}{K} \sum_{k=1}^K \frac{\mathbf{F}_n^{2D}(h, w) \cdot \mathbf{F}_{n_k, d_i}^{2D}(h, w)^\top}{\sqrt{C}}. \quad (9)$$

We then process the cost volume and image features through two parallel branches, producing the multi-view depth features $\mathbf{D}_n^{\text{multi}} \in \mathbb{R}^{H \times W \times D}$ and monocular depth features $\mathbf{D}_n^{\text{mono}} \in \mathbb{R}^{H \times W \times C}$, respectively. Finally, these two features are concatenated and passed through a depth decoder to output the depth distributions \mathbf{D} .

3.5. Overall Training Objective

The loss function of SGCDet comprises two components: detection loss \mathcal{L}_{det} , and occupancy loss \mathcal{L}_{occ} .

Detection Loss. Following [29, 31, 38, 40], we use an anchor-free detection head. The detection loss \mathcal{L}_{det} consists of cross-entropy loss $\mathcal{L}_{\text{center}}$ for centerness, IoU loss \mathcal{L}_{iou} for location, and focal loss \mathcal{L}_{cls} for classification, which can be formulated as $\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{center}} + \mathcal{L}_{\text{iou}} + \mathcal{L}_{\text{cls}}$.

Occupancy Loss. We supervise \mathbf{O}_l of each coarse-to-fine layer in the sparse volume construction as $\mathcal{L}_{\text{occ}} =$

Table 1. Quantitative results and computational cost on the ScanNet dataset. * denotes the results are directly cited from [30, 40].

Method	Voxel Resolution	Performance		Training Cost		Inference Cost	
		mAP@0.25	mAP@0.50	Memory (GB)	Time (Hours)	Memory (GB)	FPS
<i>With ground-truth geometry supervision.</i>							
ImGeoNet* [31]	40×40×16	54.8	28.4	13	16	11	2.50
CN-RMA* [30]	256×256×96	58.6	36.8	43	242	12	0.26
<i>Without ground-truth geometry supervision.</i>							
ImVoxelNet* [29]	40×40×16	46.7	23.4	11	13	9	2.60
NeRF-Det* [38]	40×40×16	53.5	27.4	13	14	12	1.30
MVSDet* [40]	40×40×16	56.2	31.3	35	36	28	0.87
SGCDet (Ours)	40×40×16	61.2	35.2	20	19	14	1.46

$\sum_{l=1}^L \mathcal{L}_{\text{bce}}(\mathbf{O}_l, \mathbf{O}_{l,gt})$, where \mathcal{L}_{bce} denotes the binary cross entropy loss.

The total loss is represented as

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{occ}}, \quad (10)$$

where we set the weight of occupancy loss λ to 0.5.

4. Experiments

4.1. Datasets and Metrics

We evaluate SGCDet on ScanNet [6], ScanNet200 [27], and ARKitScenes [2] datasets. ScanNet contains 1,201 scenes for training and 312 for testing, covering 18 categories. ScanNet200 extends ScanNet to 200 object categories with a broader range of object sizes. For both ScanNet and ScanNet200, we predict axis-aligned bounding boxes. ARKitScenes contains 4,498 scans for training and 549 for testing, with annotations for 17 classes. In this case, we detect oriented bounding boxes. We employ mean average precision (mAP) with thresholds of 0.25 and 0.5 for evaluation.

4.2. Implementation Details

Network Details. In alignment with [40], we use 40 images for training and 100 images for testing. We employ ResNet-50 [8] with a feature pyramid network (FPN) [19] as the image backbone. The spatial resolutions of input images and 2D feature maps are 320×240 and 80×60 , respectively. For DepthNet, we set the depth range and depth bins to $[0.2m, 5m]$ and 12, respectively, and use 2 nearest views to construct the cost volume. The sparse volume construction has 2 refinement stages, and we select the voxels with top 25% occupancy probability for refinement. The number of sampling points in the deformable attention is set to 4.

We present two variants of our network, SGCDet and SGCDet-L, with channel dimensions of 256 and 128, respectively. SGCDet computes a 3D volume with a spatial resolution of $40 \times 40 \times 16$ and a voxel size of $0.2m \times 0.2m \times 0.16m$, while SGCDet-L produces a 3D volume with a higher spatial resolution of $80 \times 80 \times 32$ and a finer voxel size of $0.1m \times 0.1m \times 0.08m$.

Table 2. Quantitative results on the ScanNet200 dataset. * denotes the results are directly cited from [31]. The voxel resolution of all approaches is $80 \times 80 \times 32$.

Method	Performance (mAP@0.25)			
	Total	Head	Common	Tail
ImVoxelNet* [29]	19.0	34.1	14.0	7.7
ImGeoNet* [31]	22.3	38.1	17.3	9.7
SGCDet-L (Ours)	28.9	46.0	24.0	14.9

Table 3. Quantitative results on the ARKitScenes dataset. * denotes the results are directly cited from [30, 40].

Method	Voxel Resolution	mAP@0.25	mAP@0.50
<i>With ground-truth geometry supervision.</i>			
ImGeoNet* [31]	40×40×16	60.2	43.4
CN-RMA* [30]	192×192×80	67.6	56.5
<i>Without ground-truth geometry supervision.</i>			
ImVoxelNet* [29]	40×40×16	27.3	4.3
NeRF-Det* [38]	40×40×16	39.5	21.9
MVSDet* [40]	40×40×16	42.9	27.0
ImVoxelNet [29]	40×40×16	58.0	33.2
NeRF-Det [38]	40×40×16	60.4	38.3
MVSDet [40]	40×40×16	60.7	40.1
SGCDet (Ours)	40×40×16	62.3	44.7
SGCDet-L (Ours)	80×80×32	70.4	57.0

Table 4. Ablation on the geometry and context aware aggregation. ‘2D Deform.’ and ‘3D Deform.’ denote the deformable attention is performed on 2D features \mathbf{F}_n^{2D} and lifted 3D features \mathbf{F}_n^{3D} , respectively. ‘MV Attn.’ denotes the multi-view attention.

Setting	2D Deform.	3D Deform.	MV Attn.	mAP@0.25	mAP@0.50
(a)				56.0	29.8
(b)	✓			56.2	30.5
(c)		✓		59.5	34.1
(d)		✓	✓	61.2	35.2

Training Setup. We adopt the AdamW [24] optimizer, and set the maximum learning rate to 0.0002. The cosine decay strategy [23] is used to decrease the learning rate. The models are trained on NVIDIA A6000 GPUs. We train for 12 epochs on the ScanNet and ARKitScenes datasets, and for 30 epochs on the ScanNet200 datasets.

4.3. Quantitative Results

We compare our method with the previous state-of-the-art approaches, including ImVoxelNet [29], ImGeoNet [31],

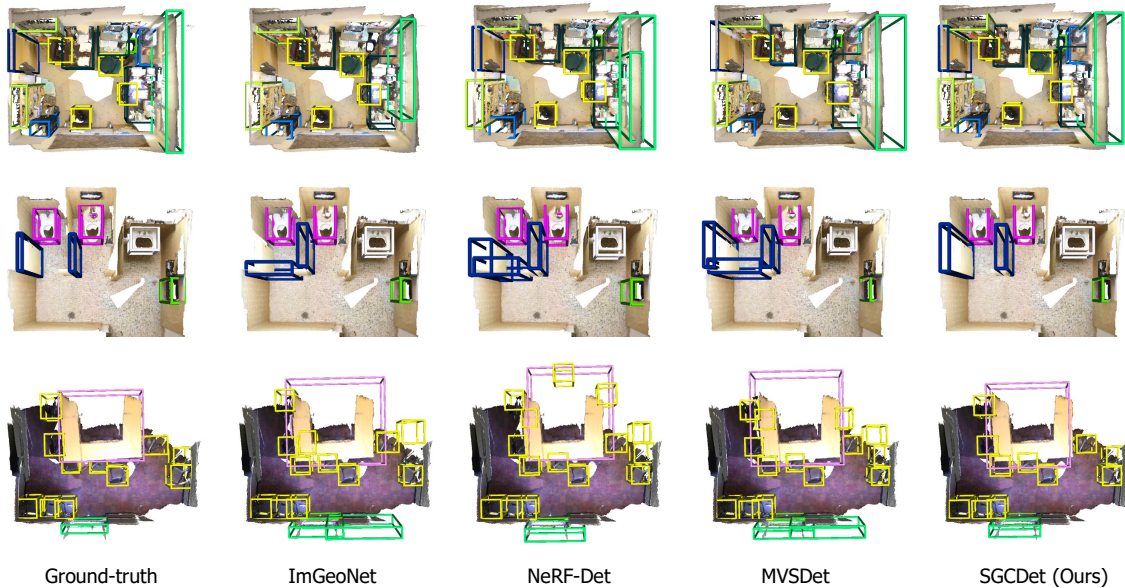


Figure 6. Qualitative comparison of different methods on the ScanNet dataset.

Table 5. Ablation on the sparse volume reconstruction, including the number of refinement stages and the selection ratio for refinement. The setting (e) is used in our SGCDet.

Setting	Voxel Resolution (Selection Ratio)	Performance		Training Cost		Inference Cost	
		mAP@0.25	mAP@0.50	Memory (GB)	Time (Hours)	Memory (GB)	FPS
(a)	40×40×16 (100%)	61.0	36.0	31	24	22	1.33
(b)	20×20×8 (100%) + 40×40×16 (25%)	60.6	35.6	21	21	14	1.40
(c)	10×10×4 (100%) + 20×20×8 (100%) + 40×40×16 (100%)	61.3	36.2	34	26	26	1.28
(d)	10×10×4 (100%) + 20×20×8 (50%) + 40×40×16 (50%)	60.9	35.4	22	22	15	1.40
(e)	10×10×4 (100%) + 20×20×8 (25%) + 40×40×16 (25%)	61.2	35.2	20	19	13	1.46
(f)	10×10×4 (100%) + 20×20×8 (10%) + 40×40×16 (10%)	57.0	31.7	19	19	13	1.53

NeRF-Det [38], CN-RMA [30], and MVSDet [40]. It is noted that ImGeoNet and CN-RMA require ground-truth geometry for training. Additionally, CN-RMA relies on a time-consuming multi-stage training pipeline, and uses a higher voxel resolution compared to other approaches.

Table 1 lists the performance and computational cost on the ScanNet dataset. The computational cost is measured on a single NVIDIA A6000 GPU. SGCDet achieves an mAP@0.25 of 61.2 and an mAP@0.50 of 35.2, surpassing all comparison approaches without using ground-truth geometry for supervision. Compared to the previous state-of-the-art approach MVSDet, SGCDet attains gains of 5.0 and 3.9 in terms of mAP@0.25 and mAP@0.50, respectively. Furthermore, SGCDet even achieves better or comparable performance than those approaches requiring ground-truth geometry during training. In terms of computational cost, SGCDet substantially reduces both training and inference costs compared to CN-RMA and MVSDet, which explicitly estimate geometry for feature lifting. Although ImVoxelNet, ImGeoNet, and NeRF-Det are efficient, their detection performance is notably lower than ours. Overall, SGCDet achieves a remarkable balance between accuracy and computational cost, while eliminating reliance on ground-truth

geometry. We further evaluate SGCDet-L on the ScanNet200 dataset, with the results shown in Table 2. The object sizes decrease from the head to the tail group. SGCDet-L consistently outperforms other approaches, demonstrating strong robustness to small objects and complex scenes with dense object distributions.

Table 3 presents the results on the ARKitScenes dataset. It is observed that some approaches exhibit a substantial performance drop compared to their results on ScanNet. This discrepancy arises because the coordinate origin of 3D scenes in ARKitScenes is positioned far from the scene center, causing the perception region of the constructed 3D volume to fail to cover the 3D scenes. To ensure a fair comparison, we follow ImGeoNet [31] to relocate the coordinate origin to the center of the input camera poses and reproduce these approaches. As shown in Table 3, SGCDet consistently provides the best performance compared to all approaches with the same 3D voxel resolution. Moreover, our SGCDet-L, with a voxel resolution of $80 \times 80 \times 32$, outperforms CN-RMA, which uses a higher voxel resolution of $192 \times 192 \times 80$ and ground-truth geometry supervision. These results further demonstrate the effectiveness of our proposed SGCDet.

Table 6. Ablation on the occupancy loss.

Setting	mAP@0.25	mAP@0.50
w/o occupancy loss	54.5	29.0
w/ occupancy loss	61.2	35.2

4.4. Qualitative Results

Fig. 6 presents visualizations of predicted 3D bounding boxes obtained from ImGeoNet[29], NeRF-Det [38], MVS-Det [40], and our proposed SGCDet. It is observed that the comparison approaches often miss some objects or predict incorrect bounding boxes in free space. In contrast, SGCDet produces more accurate detection results.

4.5. Ablation Studies

We conduct ablation studies on the ScanNet dataset.

Ablation on the geometry and context aware aggregation. Table 4 shows the ablation study of the geometry and context aware aggregation. Setting (a) serves as our baseline, employing a single-point sampling strategy for feature lifting. Settings (b) and (c) examine the impact of aggregating image features within a deformable region. Although 2D deformable attention enlarges the receptive field of voxels, it suffers from depth ambiguity, resulting in limited performance gains. In contrast, our 3D deformable attention simultaneously incorporates geometric and contextual information within an adaptive region, leading to notable improvements of 3.3 and 3.6 in mAP@0.25 and mAP@0.50, respectively. The performance is further enhanced by integrating multi-view attention, which dynamically adjusts contributions from different views (setting (d)).

Ablation on the sparse volume reconstruction. Table 5 presents a detailed analysis of the sparse volume reconstruction. Setting (a) is the baseline that directly builds the 3D volume with a fixed resolution of $40 \times 40 \times 16$. Comparing setting (a), (b), and (e), we observe that the coarse-to-fine strategy significantly reduces computational cost, while maintaining performance. We further vary the selection ratio for refinement in settings (c)-(f). Although reducing the selection ratio improves efficiency, an overly small selection ratio (*e.g.*, 10%) may miss object regions, degrading detection accuracy. To balance both accuracy and computational overhead, we set the selection ratio to 25%.

Ablation on the occupancy loss. As shown in Table 6, removing occupancy loss leads to a performance drop of 6.7 mAP@0.25 and 6.2 mAP@0.50, demonstrating the importance of explicit occupancy supervision. Thanks to our pseudo-labeling strategy based on 3D bounding boxes, we eliminate the reliance on ground-truth scene geometry.

A deeper look at the pseudo-labeling strategy. The 3D bounding boxes may produce noisy occupancy labels, particularly at the box boundaries or in cluttered scenes. However, these noisy occupancy labels are only used in the

Table 7. Ablation on the 3D bounding boxes label quality.

Label quality	Ground-truth labels		Noisy and incomplete labels	
	mAP@0.25	mAP@0.50	mAP@0.25	mAP@0.50
ImGeoNet [31]	54.8	28.4	54.0 (\downarrow 0.8)	26.2 (\downarrow 2.2)
SGCDet (Ours)	61.2	35.2	60.7 (\downarrow 0.5)	33.6 (\downarrow 1.6)

Table 8. Ablation on modules in DepthNet and depth quality.

Setting	mAP@0.25	mAP@0.50
(a) w/o monocular branch	59.6	33.8
(b) w/o multi-view branch	57.7	32.1
(c) full model	61.2	35.2
(d) w/ depth supervision	62.2	37.1
(e) ground-truth depth	64.3	42.3

training stage, serving as an explicit supervision for occupancy prediction. For inference, the top 25% selection for refinement ensures sufficient coverage of occupied regions, including areas not annotated by pseudo-labels (Fig. 3). To further assess the impact of noisy bounding box annotations, we randomly drop 15% of the ground-truth boxes, and apply 15% random scaling to the remaining boxes during training. These imperfect annotations affect both occupancy prediction and the learning of 3D detection. Nevertheless, Table 7 demonstrates that our SGCDet exhibits higher robustness compared to ImGeoNet [31], which uses ground-truth geometry supervision.

Ablation on the DepthNet. We present the ablation analysis of the DepthNet in Table 8. As shown in setting (a)-(c), removing any component of the DepthNet leads to a decrease of the detection accuracy. We then evaluate the influence of the depth quality. Adding depth supervision (setting (d)) achieves an mAP@0.25 of 62.2 and an mAP@0.50 of 37.1. Notably, these results even surpass CN-RMA [30] that needs a multi-stage training pipeline and ground-truth scene geometry for supervision. Setting (e) refers to directly using the ground-truth depth as input, indicating the upper bound of our model. It reveals a big improvement space by further study on more accurate depth estimation.

5. Conclusions

We have proposed SGCDet, a novel multi-view indoor 3D object detection framework. To enhance the representation capability of voxel features, we introduce a geometry and context aware aggregation module that adaptively integrates image features across multiple views. Additionally, we develop a sparse volume construction strategy that selectively refines voxels with high occupancy probability, significantly reducing redundant computation in free space. Our framework is trained using only 3D bounding boxes for supervision, eliminating the need for ground-truth scene geometry. Experimental results demonstrate that SGCDet achieves state-of-the-art performance on the ScanNet, ScanNet200, and ARKitScenes datasets.

Acknowledgments

This work was supported in part by the National Key Research and Development Program of China under grant 2023YFB3209800, in part by the National Natural Science Foundation of China under grant 62301484, in part by the Ningbo Natural Science Foundation of China under grant 2024J454, and in part by the Aeronautical Science Foundation of China under grant 2024M071076001. We also thank the generous help from Sijin Li, Zhejiang University.

References

- [1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7822–7831, 2021. 2, 4
- [2] 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. 2, 6
- [3] 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, 2020. 3
- [4] Yuedong Chen, Haofei Xu, Chuanxia Zheng, Bohan Zhuang, Marc Pollefeys, Andreas Geiger, Tat-Jen Cham, and Jianfei Cai. MVSplat: Efficient 3d gaussian splatting from sparse multi-view images. In *European Conference on Computer Vision*, pages 370–386, 2024. 2
- [5] Robert T Collins. A space-sweep approach to true multi-image matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 358–363, 1996. 5
- [6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017. 2, 6
- [7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6
- [9] Chenxi Huang, Yuenan Hou, Weicai Ye, Di Huang, Xiaoshui Huang, Binbin Lin, and Deng Cai. NeRF-Det++: Incorporating semantic cues and perspective-aware depth supervision for indoor multi-view 3d detection. *IEEE Transactions on Image Processing*, 2025. 2
- [10] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. BEVDet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2
- [11] Maksim Kolodiazhnyi, Anna Vorontsova, Matvey Skripkin, Danila Rukhovich, and Anton Konushin. UniDet3D: Multi-dataset indoor 3d object detection. *arXiv preprint arXiv:2409.04234*, 2024. 1
- [12] Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. DFA3D: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6684–6693, 2023. 2, 4, 5
- [13] Wuyang Li, Zhu Yu, and Alexandre Alahi. VoxDet: Rethinking 3d semantic occupancy prediction as dense object detection. *arXiv preprint arXiv:2506.04623*, 2025. 3
- [14] Yin hao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. BEVDepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2
- [15] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. VoxFormer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 3
- [16] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. BEVFormer: Learning bird’s-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [17] Zechuan Li, Hongshan Yu, Yihao Ding, Jinhao Qiao, Basim Azam, and Naveed Akhtar. GO-N3RDet: Geometry optimized nerf-enhanced 3d object detector. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27211–27221, 2025. 2
- [18] Yingping Liang and Ying Fu. CascadeV-Det: Cascade point voting for 3d object detection. *arXiv preprint arXiv:2401.07477*, 2024. 1
- [19] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2117–2125, 2017. 6
- [20] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. SparseBEV: High-performance sparse 3d object detection from multi-camera videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18580–18590, 2023. 3
- [21] Haisong Liu, Yang Chen, Haiguang Wang, Zetong Yang, Tianyu Li, Jia Zeng, Li Chen, Hongyang Li, and Limin Wang. Fully sparse 3d occupancy prediction. In *European Conference on Computer Vision*, pages 54–71, 2024. 3
- [22] 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, 2022. 3
- [23] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6
- [24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 6
- [25] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. OcctreeOcc: Efficient and multi-granularity occupancy prediction using octree queries. In *Advances in Neural Information Processing Systems*, 2024. 3
- [26] 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. 1
- [27] David Rozenberszki, Or Litany, and Angela Dai. Language-grounded indoor 3d semantic segmentation in the wild. In *European Conference on Computer Vision*, pages 125–141, 2022. 2, 6
- [28] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. FCAF3D: Fully convolutional anchor-free 3d object detection. In *European Conference on Computer Vision*, pages 477–493, 2022. 1
- [29] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. ImVoxelNet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022. 1, 2, 3, 4, 5, 6, 8
- [30] Guanlin Shen, Jingwei Huang, Zhihua Hu, and Bin Wang. CN-RMA: Combined network with ray marching aggregation for 3d indoor object detection from multi-view images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21326–21335, 2024. 1, 2, 6, 7, 8
- [31] Tao Tu, Shun-Po Chuang, Yu-Lun Liu, Cheng Sun, Ke Zhang, Donna Roy, Cheng-Hao Kuo, and Min Sun. ImGeoNet: Image-induced geometry-aware voxel representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6996–7007, 2023. 1, 2, 3, 4, 5, 6, 7, 8
- [32] Haiyang Wang, Lihe Ding, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, and Liwei Wang. CA-Group3D: Class-aware grouping for 3d object detection on point clouds. In *Advances in Neural Information Processing Systems*, pages 29975–29988, 2022. 1
- [33] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Ming-Ming Cheng. OPUS: Occupancy prediction using a sparse set. In *Advances in Neural Information Processing Systems*, 2024. 3
- [34] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. EmbodiedScan: A holistic multi-modal 3d perception suite towards embodied AI. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 2
- [35] 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, 2022. 3
- [36] A Waswani, N Shazeer, N Parmar, J Uszkoreit, L Jones, A Gomez, L Kaiser, and I Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, 2017. 5
- [37] 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. 3
- [38] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. NeRF-Det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23320–23330, 2023. 1, 2, 3, 4, 5, 6, 7, 8
- [39] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezaatofghi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 5
- [40] Yating Xu, Chen Li, and Gim Hee Lee. MVSDet: Multi-view indoor 3d object detection via efficient plane sweeps. In *Advances in Neural Information Processing Systems*, 2024. 1, 2, 3, 4, 5, 6, 7, 8
- [41] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision*, pages 767–783, 2018. 5
- [42] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *Advances in Neural Information Processing Systems*, 2024. 3
- [43] Zaiwei Zhang, Bo Sun, Haitao Yang, and Qixing Huang. H3DNet: 3d object detection using hybrid geometric primitives. In *European Conference on Computer Vision*, pages 311–329, 2020. 1