通过
自适应 3D 体积构建提升多视角室内 3D 对象检测

Runmin Zhang[1] Zhu Yu[1*] Si-Yuan Cao[2,3*] Lingyu Zhu[4]
Guangyi Zhang[1] Xiaokai Bai[1] Hui-Liang Shen[1]
[1]College of Information Science and Electronic Engineering, Zhejiang University
[2]Ningbo Global Innovation Center, Zhejiang University [3]NingboTech University [4]City University of Hong Kong
{ runmin_zhang, yu_zhu, cao_siyuan } @zju.edu.cn, lingyzhu-c@my.cityu.edu.hk,
{ zhangguangyi, shawnnnkb, shenhl } @zju.edu.cn

Abstract

这项工作提出了 SGCDet，这是一种基于自适应 3D 体积构建的新型多视角室内 3D 物体检测框架。与先前的方法将体素的接收场限制在图像的固定位置不同，我们引入了几何和上下文感知聚合模块，以便在每个图像的自适应区域内整合几何和上下文信息，并动态调整来自不同视图的贡献，增强了体素特征的表示能力。此外，我们提出了一种稀疏体积构建策略，该策略自适应地识别和选择具有高占据概率的体素进行特征细化，从而最小化自由空间中的冗余计算。得益于上述设计，我们的框架实现了以自适应方式进行的有效体积构建。更好的是，我们的网络可以仅使用 3D 边界框进行监督，消除了对真实场景几何的依赖。实验结果表明，SGCDet 在 ScanNet、ScanNet200 和 ARKitScenes 数据集上达到了最新的性能。源代码可在 https://github.com/RM-Zhang/SGCDet 获得。

室内三维物体检测是一项基础的三维感知任务，广泛应用于具身人工智能、增强/虚拟现实和机器人领域。利用精确的场景几何作为输入，基于点云的三维物体检测器 [11, 18, 26, 28, 32, 43] 已经取得了令人印象深刻的性能。然而，捕获准确的场景几何通常需要高成本的三维传感器。最近，已经转向使用多视图姿态图像进行三维物体检测。

为了弥合二维图像与三维表示间的差距，开创性工作 ImVoxelNet [29] 沿着整体射线提升二维特征。然后，每个体素采用多个视图的平均结果作为其特征。然而，这种方法对来自不同图像的特征使用相同权重，导致三维体素表示粗糙且容易出错。尽管后续的工作 [31, 38] 通过后处理引入不透明概率以抑制自由空间中的体素特征，它们仍未能在二维到三维投影过程中解决遮挡问题。更近期的方法 [30, 40] 引入显式几何约束以协助特征提升。然而，这些方法的最终性能严重依赖于估计几何信息的准确性，这要么使得训练流程复杂化，要么显著增加计算成本。

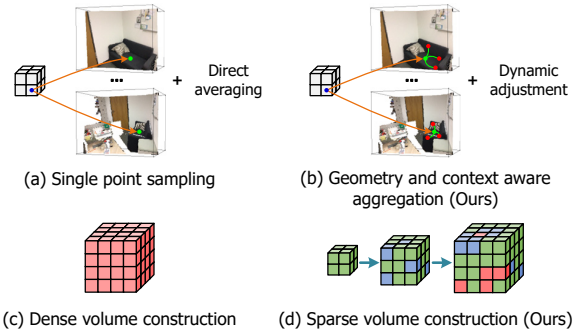


Figure 1. 对比我们提出的 SGCDet 与之前方法在特征提升和体素构建策略上的差异。(a) 之前方法使用的单点采样策略将体素的感受野限制在一个有限的区域，忽略了跨多个视图的上下文信息。(b) 我们的几何和上下文感知聚合方法自适应地整合了来自不同视图的可变形区域内的几何和上下文特征，增强了体素特征的表示能力。(c) 之前的方法构建高分辨率的密集 3D 体素，未考虑 3D 场景的固有稀疏性，导致不必要的计算开销。(d) 我们的稀疏体素构建方法自适应地优化可能包含物体的体素，并在空闲空间中减少冗余计算成本。

如上所述，先前的方法主要从几何角度增强 3D 体素表示的质量，而忽视了图像中有价值的上下文信息。二维特征图上的采样位置被限制在由预定义的体素中心和相机姿态确定的固定位置。这种单点采样策略将体素的感受野限制在一个小区域，限制了它们感知视觉信息的能力。此外，这种策略进一步加剧了对精确几何信息的依赖 [30, 40] ，如图 1 (a) 所示。

为了解决这些问题，我们提出了一种几何和上下文感知的聚合模块，用于自适应提升 2D 特征。我们不是简单地对多视图图像中的采样特征进行加权平均，而是将采样特征视为查询，在可变形区域内聚合相关几何和上下文特征。此外，我们引入了一种多视图注意机制，以动态调整来自不同视图的贡献，增强变换后 3D 体积的表示能力，如图 1 (b) 所示。

另一方面，之前的方法通常构建高分辨率的稠密三

*Corresponding authors.

维体积，如图 1 (c) 所示。这种稠密表示未能考虑到三维场景的固有稀疏性，导致不必要的计算开销。为了解决这个问题，我们提出了一种稀疏体积构建策略，以自适应的方式构建三维体积，如图 1 (d) 所示。具体而言，我们采用一个占用预测模块来识别可能含有待细化目标的体素，从而减少自由空间中的冗余计算。该策略的一个关键方面是占用预测的监督。虽然一个简单的解决方案是直接使用真实几何进行监督 [30, 31]，但在没有这些数据时这是不可行的 [1, 40]。为了消除对真实几何的依赖，我们利用三维边界框生成占用的伪标签，实现灵活的网络监督。

通过结合稀疏体积构建和几何与上下文感知聚合，我们提出了一种新的多视图室内 3D 物体检测框架，称为 SGCDet。得益于上述设计，SGCDet 能够有效地构建 3D 体积。我们在 ScanNet [6]、ScanNet200 [27] 和 ARKitScenes [2] 数据集上评估了 SGCDet。在不依赖于地面实况几何监督的方法中，SGCDet 实现了最先进的性能。与之前的最先进方法 MVSDet [40] 相比，SGCDet 在 ScanNet 上的 mAP@0.5 指标显著提高了 3.9，同时减少了训练内存、训练时间、推理内存和推理时间，分别降低了 42.9 %、47.2 %、50 % 和 40.8 %。值得注意的是，SGCDet 也超越了一些在训练中使用地面实况几何的数据的方法。

我们的贡献总结如下：

- 我们提出了几何和上下文感知聚合模块以增强特征提升。它使每个体素能够在可变形区域中自适应地聚合几何和上下文特征，并在不同视角中动态调整特征的贡献。
- 我们介绍了一种稀疏体素构建策略，该策略自适应地细化可能包含物体的体素，从而减少在空闲空间中的计算量。值得注意的是，该网络整体上可以仅使用 3D 边界框进行监督，无需真实几何信息。
- 大量实验表明，SGCDet 在性能上大幅超越了以往的领先方法，同时显著减少了计算开销。这些结果验证了 SGCDet 的有效性和高效性。

1. 相关工作

基于图像的 3D 目标检测。基于图像的 3D 目标检测由于其成本效益和细粒度视觉感知能力而受到广泛关注。目前的方法主要集中在从输入图像构建 3D 表示，包括鸟瞰图（BEV）[10, 12, 14, 16] 和基于体素的方法 [29–31, 34, 38, 40]。鉴于相机视点和目标分布的多样性，体素表示更适合室内场景。ImVoxelNet [29] 是引入用于多视角室内 3D 目标检测的端到端流程的先驱。它直接沿 3D 光线提升 2D 特征，而未结合场景几何，导致体积特征的模糊性。在 ImVoxelNet 的基础上，ImGeoNet [31] 和 NeRF-Det [38] 计算 3D 体积的不透明度概率，以抑制自由空间中的特征。NeRF-Det++ [9] 和 Go-N3RDet [17] 通过语义和几何约束进一步增强了 NeRF-Det。然而，这些方法仅将不透明度视为后处理步骤，未能解决特征提升过程中的遮挡问题。或者，最近的方法 [30, 40] 明确估计场景几何，以实现遮挡感知投影。CN-RMA [30] 将 3D 重建网络与基于点云的 3D 目标检测器结合使用，利用重建的 TSDF 引导特征提升过程。然而，它需要耗时的多阶段训练流程，并依赖于用于监督的真实几何。相反，MVSDet [40] 利用多视图立体从输入图像计算深度概率，并应用 3D 高斯散射 [4] 进行自我监督。然而，它仍然面临高计算成本问题。

三维视觉中的稀疏设计。受到 DETR [3] 的启发，一些三维检测方法 [20, 22, 35, 37] 使用稀疏的目标查询集来实现三维到二维的交互。然而，由于缺乏明确的三维表示，这些方法通常存在收敛缓慢的问题。其他占用预测方法通过基于深度的查询提议初始化 [13, 15, 42]，多尺度稀疏重建 [21, 25]，或者通过将问题重新表述为稀疏集预测 [33] 来减少体素查询的数量。虽然这些方法已经取得了显著进展，但它们仍然依赖于精确的几何形态进行监督，限制了它们在缺乏真实几何信息的场景中的适用性。

2. 方法

给定 N 姿态图像 $\{\boldsymbol{I}_n\}_{n=1}^N$ 作为输入，SGCDet 的目标是预测场景的 3D 包围框。如图 2 (a) 所示，SGCDet 的整体框架由三个主要组件组成：用于提取 2D 特征 $\{\boldsymbol{F}_n^{\text{2D}} \in \mathbb{R}^{H\times W\times C}\}_{n=1}^N$ 的图像骨干，一个将这些 2D 特征提升为 3D 体积 $\boldsymbol{V} \in \mathbb{R}^{X\times Y\times Z\times C}$ 的视图变换模块，以及用于预测 3D 包围框的检测头。这里，(H, W) 和 (X, Y, Z) 分别代表 2D 特征和 3D 体积的空间分辨率，C 表示通道数量。

我们的核心设计集中在视图转换模块，该模块实现自适应 3D 体积构建。具体而言，我们采用一个简单但有效的 DepthNet（第 2.3 节）来估计输入图像的深度分布 $\{\boldsymbol{D}_n \in \mathbb{R}^{H\times W\times D}\}_{n=1}^N$，其中 D 是深度分层的数量。深度分布为视图转换过程提供了几何信息。为了解决稠密体积构建的低效问题，我们提出了稀疏体积构建（第 2.1 节），该方法以由粗到精的方式自适应地构建 3D 体积。在这个过程中，我们引入了几何和上下文感知的聚合（第 2.2 节），通过在灵活区域内整合几何和上下文信息来确保自适应特征提升。

2.1. 稀疏体积构建

给定密集的三维网格 $\boldsymbol{G} \in \mathbb{R}^{X\times Y\times Z\times 3}$，之前的方法 [29, 31, 38, 40] 通常将每个三维体素投影到二维特征以构建体积。然而，由于三维场景中的大多数体素是自由空间，这种密集构建对于目标检测来说效率低下，并导致显著的计算开销。为了解决这个问题，我们引入了一种稀疏体积构建策略，以由粗到细的方式构建三维体积。如图 2 (b) 所示，其关键思想是逐渐对粗略的三维体积进行上采样，只自适应地细化可能包含物体的体素。具体来说，我们首先构建一个空间分辨率为 $(\frac{X}{2^L}, \frac{Y}{2^L}, \frac{Z}{2^L})$ 的粗略三维体积 $\boldsymbol{V}_0 \in \mathbb{R}^{\frac{X}{2^L}\times\frac{Y}{2^L}\times\frac{Z}{2^L}\times C}$。这个粗略体积捕捉了整个场景的几何形状，可以用于识别可能包含物体的区域以进行进一步细化。

粗到细的细化。整体细化过程由 L 个阶段组成。如图 2 (b) 所示，在 l 阶段，我们首先将第 $(l-1)$ 阶段的输

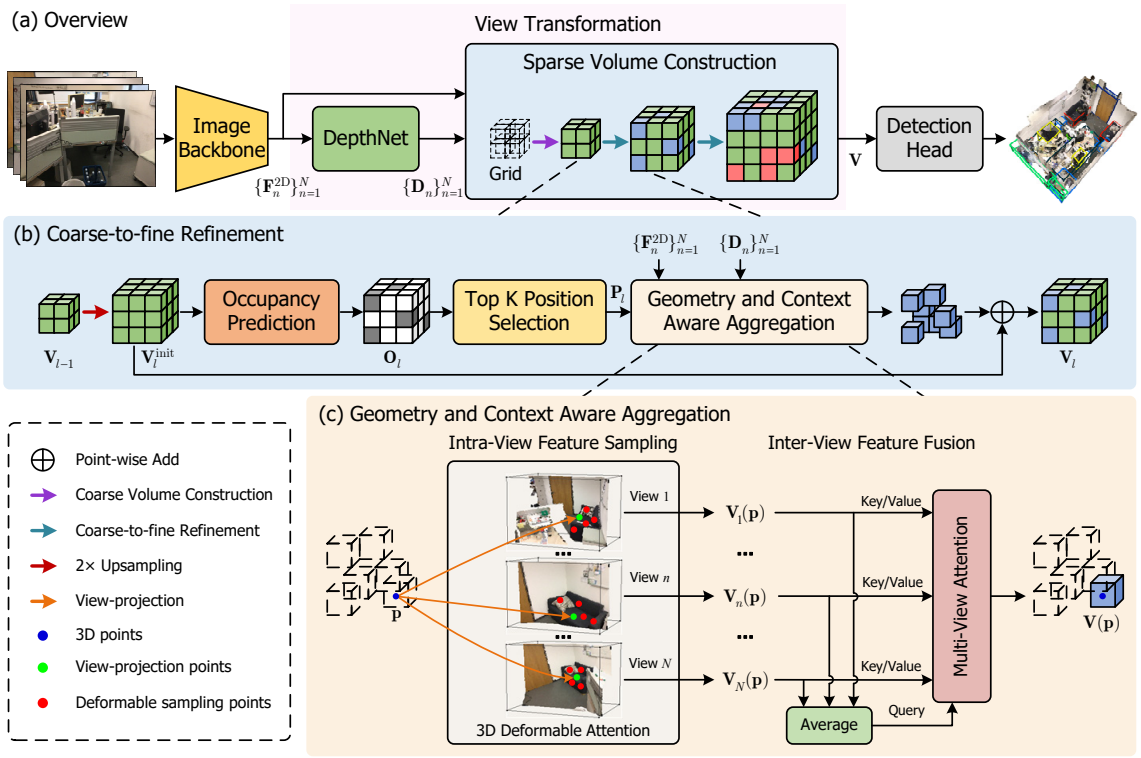


Figure 2. SGCDet 的示意图和详细架构。(a) SGCDet 的概览，其中包括一个用于提取图像特征的图像骨干，一个用于将图像特征提升到 3D 体积的视图转换模块，以及一个用于预测 3D 边界框的检测头。(b) 我们稀疏体积构建策略中粗到细细化的细节。(c) 我们的几何和上下文感知聚合模块的细节。

出体积上采样 2 倍，得到 $\boldsymbol{V}_l^{\text{init}} \in \mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}} \times C}$。然后，我们通过

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

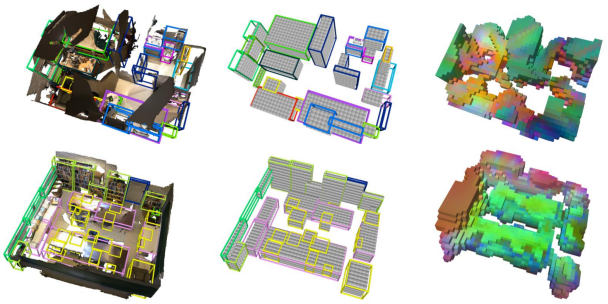
估计每个体素的占用概率，其中 $\boldsymbol{O}_l \in \mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}}}$ 表示占用概率，$\mathcal{F}$ 是一个轻量级占用预测头。接下来，我们选择占用概率最高的 k 个位置进行特征细化，公式为

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

，其中 $\boldsymbol{P}_l$ 是前 $k\%$ 个点的坐标集合，$\mathcal{P}$ 表示在第 Sec. 2.2 节中描述的几何和上下文感知聚合。这一策略避免了在自由空间的冗余计算，同时有效地捕捉到可能包含物体的区域中的细小结构。

占用概率监督。一种简单的方法是通过真实场景几何来监督占用概率。然而，当精确几何不可用时，这是不可行的 [1, 40]。为了解决这个问题，我们使用真实的 3D 边界框来生成占用的伪标签，提供了一种灵活的监督策略。给定在 l 阶段的 3D 网格 $\boldsymbol{G}_l \in \mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}} \times 3}$，真实的占用概率 $\boldsymbol{O}_{l,\text{gt}} \in \mathbb{R}^{\frac{X}{2^{L-l}} \times \frac{Y}{2^{L-l}} \times \frac{Z}{2^{L-l}}}$ 定义为：

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



(a) Ground-truth (b) Pseudo label for occupancy (c) Refined volume features

Figure 3. 我们的稀疏体积构建的可视化。(a) 真实的 3D 边界框。(b) 用于占用监督的伪标签，源自 3D 边界框。(c) 精炼后的体积特征。我们的占用预测网络有效地过滤了空闲空间，将特征精炼集中在可能包含物体的体素上。

网络由 $\boldsymbol{O}_l$ 和 $\boldsymbol{O}_{l,\text{gt}}$ 之间的二元交叉熵损失进行监督。图 3 展示了在最后调整阶段，从 3D 边界框和细化体积特征生成的两个伪标签示例。可以观察到，我们的稀疏体积构建有效捕捉了场景几何，并自适应地关注包含物体的区域。

2.2. 几何和上下文感知聚合

我们几何和上下文感知聚合的详细图示如图 2 (c) 所示。为了获取以中心 $\boldsymbol{p} = (x,y,z)^\top$ 为中心的体素的特征，我们首先执行视图内特征采样，以独立地从每个视图中采样特征，以进行初始的信息聚合。随后，我们应

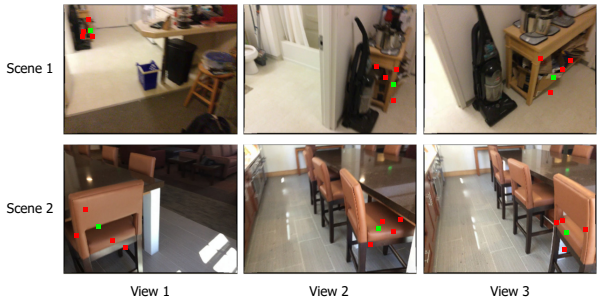


Figure 4. 在我们的视图内特征采样中，采样位置的可视化。绿色点表示从体素中心和相机姿态得出的位置，而红色点表示可变形采样位置。我们注意到，可变形注意力是在 3D 像素空间中执行的，为了清晰起见，此可视化中省略了深度维度。

用视图间特征融合，以融合多个视图的特征进行进一步的细化。

视图内特征采样。之前的方法 [29–31, 38, 40] 只是简单地在从体素中心和相机姿态导出的特定位置信息中采样图像特征作为体素特征，这限制了体素的感受野。为了解决这个问题，我们引入了一种 3D 可变形注意机制 [12]，以在自适应区域内结合几何和上下文信息。具体而言，对于每个视图 n，我们将 2D 图像特征提升到 3D 像素空间，公式为

$$\boldsymbol{F}_n^{\mathrm{3D}} = \boldsymbol{F}_n^{\mathrm{2D}} \otimes \boldsymbol{D}_n, \tag{4}$$

，其中 $\boldsymbol{F}_n^{\mathrm{3D}} \in \mathbb{R}^{H\times W\times D\times C}$ 表示提升的 3D 特征，$\otimes$ 指的是在最后一个维度上进行的外积。接下来，我们将 $\boldsymbol{p}$ 投影到视图 n，公式为

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

，其中 $\boldsymbol{p}_n$ 是 3D 像素空间中的坐标，$\boldsymbol{K}_n$ 和 $\boldsymbol{E}_n$ 分别是视图 n 的内在和外在矩阵。我们并不直接在 $\boldsymbol{p}_n$ 处采样 $\boldsymbol{F}_n^{\mathrm{3D}}$ 作为体素特征 $\boldsymbol{V}_n(\boldsymbol{p})$，而是将采样的特征作为查询，从邻近区域聚合信息，公式为

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

，其中 M 是采样点的数量，W 是值投影的矩阵，ϕ 表示用于从 $\boldsymbol{F}_n^{\mathrm{3D}}$ 中采样特征的三线性插值。$\Delta \boldsymbol{p}_{n,m}$ 和 $A_{n,m}$ 分别是第 m 个采样点的 3D 偏移和注意力权重，它们通过一个线性层从查询 $\phi(\boldsymbol{F}_n^{\mathrm{3D}}, \boldsymbol{p}_n)$ 中生成。为简单起见，我们在方程 6 中省略了多头操作。我们在图 4 中展示了不同视图中可变形采样点的位置。与单点采样相比，我们具有几何和上下文感知的聚合能够有效地在灵活的区域内整合几何和上下文信息，从而增强了体素特征的表示能力。

视图间特征融合。由于不同视图中的物体外观和大小的变化，从不同视图采样的特征 $\{\boldsymbol{V}_n(\boldsymbol{p})\}_{n=1}^N$ 可能存在显著差异。为了自适应地调整每个视图的贡献，我们提出了一种多视图注意力机制。具体来说，我们使用从所有视图 $\boldsymbol{V}_{\mathrm{avg}}(\boldsymbol{p})$ 的特征平均池化作为查询，而 $\{\boldsymbol{V}_n(\boldsymbol{p})\}_{n=1}^N$ 同时作为键和值。这个过程可以公式化为

$$\boldsymbol{V}(\boldsymbol{p}) = \mathrm{Attn}(\boldsymbol{V}_{\mathrm{avg}}(\boldsymbol{p}), \{\boldsymbol{V}_n(\boldsymbol{p})\}_{n=1}^N, \{\boldsymbol{V}_n(\boldsymbol{p})\}_{n=1}^N), \tag{7}$$

，其中 $\boldsymbol{V}(\boldsymbol{p})$ 表示 3D 点 $\boldsymbol{p}$ 的最终特征，$\mathrm{Attn}(\cdot)$ 是标准注意力操作 [7, 36]。这里，我们假设 $\boldsymbol{p}$ 可以投影到所有视图上以简化符号表示。在实际操作中，我们会丢弃任何 $\boldsymbol{p}$ 投影在图像边界之外的视图。

讨论。我们的几何和上下文感知聚合受到 DFA3D [12] 的高度启发，在此基础上，我们进行了大量修改以更好地适应室内场景。DFA3D 使用与视图无关的 3D 查询来预测所有视图中的采样偏移和权重，这在具有固定相机布局的自动驾驶场景中性能良好。然而，在相机姿态变化显著且物体在各个视图中表现出较大形状和尺度差异的室内环境中，其效果有限。相反，我们的视图内特征采样利用特定视图的特征作为查询，能够实现针对每个视图的自适应聚合。此外，我们的视图间融合模块为每个视图的贡献分配可学习的注意力权重，从而产生更一致和更鲁棒的场景级体素表示。

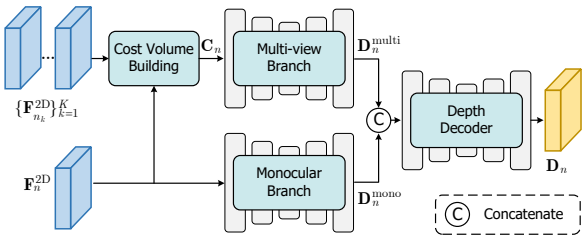


Figure 5. DepthNet 的详细架构。

2.3. 深度网络

深度分布为 2D 到 3D 投影过程提供几何信息，其准确性显著影响最终的检测性能。为了充分利用多视图图像进行准确的深度估计，我们引入了一个简单但有效的 DepthNet。如图 5 所示，它融合了多视图和单目深度特征用于深度估计，其中前者通过特征匹配提供几何属性，而后者提供输入图像的详细结构。

对于任何具有二维特征 $\boldsymbol{F}_n^{\mathrm{2D}}$ 的视图 n，我们选择最近的具有二维特征 $\{\boldsymbol{F}_{n_k}^{\mathrm{2D}}\}_{k=1}^K$ 的 K 视图，并使用平面扫描 [5] 构建代价体积。具体而言，我们将深度范围 $[d_{\min}, d_{\max}]$ 离散为 $[d_1, ..., d_i, ..., d_D]$ 的 D 个深度区间。对于每个深度平面 d_i，我们利用相机矩阵将相邻视图的二维特征变换到视图 n：

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

其中 $\mathcal{W}$ 是在 [39, 41] 中的扭曲操作。然后，我们构建代价体积 $\boldsymbol{C}_n \in \mathbb{R}^{H\times W\times D}$ 为：

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

Table 1. 在 ScanNet 数据集上的定量结果和计算成本。* 表示结果直接引用自 [30, 40] 。

Method	Voxel Resolution	Performance		Training Cost		Inference Cost	
		mAP@0.25	mAP@0.50	Memory (GB)	Time (Hours)	Memory (GB)	FPS
With ground-truth geometry supervision.							
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
Without ground-truth geometry supervision.							
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

然后，我们通过两个平行分支处理代价体积和图像特征，分别生成多视角深度特征 $\boldsymbol{D}_n^{\text{multi}} \in \mathbb{R}^{H\times W\times D}$ 和单目深度特征 $\boldsymbol{D}_n^{\text{mono}} \in \mathbb{R}^{H\times W\times C}$ 。最后，这两个特征被拼接并通过一个深度解码器来输出深度分布 $\boldsymbol{D}$ 。

SGCDet 的损失函数包括两个组成部分：检测损失 $\mathcal{L}_{\text{det}}$ 和占用损失 $\mathcal{L}_{\text{occ}}$ 。

检测损失。根据 [29, 31, 38, 40] ，我们使用一个无锚点的检测头。检测损失 $\mathcal{L}_{\text{det}}$ 包括用于中心度的交叉熵损失 $\mathcal{L}_{\text{center}}$ ，用于定位的 IoU 损失 $\mathcal{L}_{\text{iou}}$ ，和用于分类的焦点损失 $\mathcal{L}_{\text{cls}}$ ，可以表示为 $\mathcal{L}_{\text{det}} = \mathcal{L}_{\text{center}} + \mathcal{L}_{\text{iou}} + \mathcal{L}_{\text{cls}}$. 。

占用损失。我们监督稀疏体积构造中每个粗到细层的 $\boldsymbol{O}_l$ 作为 $\mathcal{L}_{\text{occ}} = \sum_{l=1}^{L} \mathcal{L}_{\text{bce}}(\boldsymbol{O}_l, \boldsymbol{O}_{l,gt})$ ，其中 $\mathcal{L}_{\text{bce}}$ 表示二元交叉熵损失。

总损失表示为

$$\mathcal{L} = \mathcal{L}_{\text{det}} + \lambda \mathcal{L}_{\text{occ}}, \tag{10}$$

，其中我们将占用损失 λ 的权重设置为 0.5 。

3. 实验

我们在 ScanNet [6] ，ScanNet200 [27] 和 ARKitScenes [2] 数据集上评估 SGCDet。ScanNet 包含 1,201 个用于训练的场景和 312 个用于测试的场景，涵盖 18 个类别。ScanNet200 将 ScanNet 扩展到 200 个物体类别，具有更广泛的物体尺寸范围。对于 ScanNet 和 ScanNet200，我们预测轴对齐的边界框。ARKitScenes 包含 4,498 个用于训练的扫描和 549 个用于测试的扫描，具有 17 个类别的标注。在这种情况下，我们检测定向边界框。我们采用平均精度 (mAP) 作为评估标准，阈值为 0.25 和 0.5。

3.1. 实现细节

网络详情。根据 [40] 的描述，我们使用 40 张图像进行训练，100 张图像进行测试。我们采用 ResNet-50 [8] ，并结合特征金字塔网络（FPN） [19] 作为图像主干。输入图像和二维特征图的空间分辨率分别为 320 × 240 和 80 × 60 。对于 DepthNet，我们将深度范围和深度分箱分别设置为 $[0.2m, 5m]$ 和 12 ，使用 2 个最近视图来构建代价体积。稀疏体积构建有 2 个精细化阶段，我们选择占用概率最高的 25% 个体素进行精细化。在可变形注意力中，采样点的数量设置为 4 。

Table 2. ScanNet200 数据集上的定量结果。* 表示结果直接引用自 [31] 。所有方法的体素分辨率为 80 × 80 × 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. ARKitScenes 数据集上的定量结果。* 表示结果直接引用自 [30, 40] 。

Method	Voxel Resolution	mAP@0.25	mAP@0.50
With ground-truth geometry supervision.			
ImGeoNet* [31]	40 × 40 × 16	60.2	43.4
CN-RMA* [30]	192 × 192 × 80	67.6	56.5
Without ground-truth geometry supervision.			
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

我们提出了我们的网络的两个变体，SGCDet 和 SGCDet-L，它们的通道维数分别为 256 和 128。SGCDet 计算空间分辨率为 40 × 40 × 16 和体素大小为 0.2 m × 0.2 m × 0.16 m 的 3D 体积，而 SGCDet-L 则生成一个具有更高空间分辨率 80 × 80 × 32 和更精细体素大小 0.1 m × 0.1 m × 0.08 m 的 3D 体积。

训练设置。我们采用 AdamW [24] 优化器，最大学习率设置为 0.0002 。使用余弦衰减策略 [23] 来降低学习率。模型在 NVIDIA A6000 GPU 上进行训练。我们在 ScanNet 和 ARKitScenes 数据集上训练 12 个周期，在 ScanNet200 数据集上训练 30 个周期。

3.2. 定量结果

我们将我们的方法与之前的最新方法进行了比较，包括 ImVoxelNet [29] 、ImGeoNet [31] 、NeRF-Det [38] 、CN-RMA [30] 和 MVSDet [40] 。需要注意的是，ImGeoNet 和 CN-RMA 需要真实几何数据进行训练。此外，CN-RMA 依赖于耗时的多阶段训练流程，并且使用了比其他方法更高的体素分辨率。

表格 1 列出了在 ScanNet 数据集上的性能和计算成

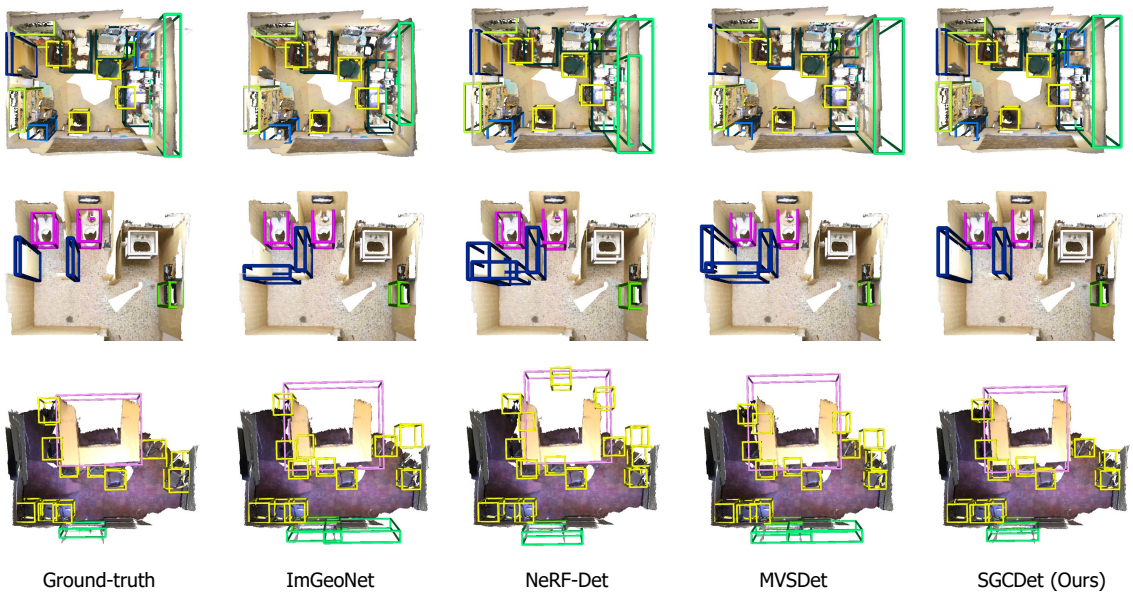


Figure 6. 在 ScanNet 数据集上不同方法的定性比较。

Table 4. 几何与上下文感知聚合的消融实验。“2D Deform.”和“3D Deform.”分别表示可变形注意力在二维特征 $\boldsymbol{F}_n^{\text{2D}}$ 和提升后的三维特征 $\boldsymbol{F}_n^{\text{3D}}$ 上执行。“MV Attn.”表示多视图注意力。

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

本。计算成本是在单个 NVIDIA A6000 GPU 上测量的。SGCDet 达到了 mAP@0.25 为 61.2 和 mAP@0.50 为 35.2，超过了所有没有使用真实几何作为监督的比较方法。与之前的最新方法 MVSDet 相比，SGCDet 分别在 mAP@0.25 和 mAP@0.50 上取得了 5.0 和 3.9 的提升。此外，SGCDet 甚至在性能上优于或可与需要在训练过程中使用真实几何的那些方法相比。在计算成本方面，SGCDet 相比 CN-RMA 和 MVSDet 大幅降低了训练和推理成本，后者明确估计几何特征提升。虽然 ImVoxelNet、ImGeoNet 和 NeRF-Det 效率较高，但它们的检测性能明显低于我们的。总体而言，SGCDet 在准确性和计算成本之间实现了显著的平衡，同时消除了对真实几何的依赖。我们进一步在 ScanNet200 数据集上评估了 SGCDet-L，其结果如表格 2 所示。对象尺寸从头组到尾组逐渐减小。SGCDet-L 始终优于其他方法，展示了对于小物体和复杂场景中密集物体分布的强鲁棒性。

表格 3 展示了在 ARKitScenes 数据集上的结果。观察到与其在 ScanNet 上的结果相比，某些方法性能有显著下降。这种差异的产生是因为 ARKitScenes 的 3D 场景的坐标原点距离场景中心非常远，导致构建的 3D 体积的感知区域未能覆盖 3D 场景。为了确保公平比较，我们遵循 ImGeoNet [31]，将坐标原点重新定位到输入相机姿态的中心并重新实现这些方法。如表格 3 所示，SGCDet 在与所有具有相同 3D 体素分辨率的方法相比时，始终提供最佳性能。此外，我们的 SGCDet-L 以 $80 \times 80 \times 32$ 的体素分辨率，超越了使用更高体素分辨率 $192 \times 192 \times 80$ 和真实几何监督的 CN-RMA。这些结果进一步证明了我们提出的 SGCDet 的有效性。

图 6 展示了从 ImGeoNet [29]、NeRF-Det [38]、MVSDet [40] 和我们提出的 SGCDet 获得的预测 3D 边界框的可视化。观察发现，比较方法经常遗漏一些物体，或在空旷区域预测不正确的边界框。相比之下，SGCDet 能够产生更准确的检测结果。

我们在 ScanNet 数据集上进行消融研究。

关于几何和上下文感知聚合的消融实验。表格 4 显示了几何和上下文感知聚合的消融研究。设置 (a) 是我们的基线，采用单点采样策略进行特征提升。设置 (b) 和 (c) 探讨了在可变形区域内聚合图像特征的影响。尽管二维可变形注意力扩大了体素的感受野，但它受到深度模糊的影响，导致性能提升有限。相比之下，我们的三维可变形注意同时在自适应区域内融入几何和上下文信息，在 mAP@0.25 和 mAP@0.50 上分别显著提升了 3.3 和 3.6。通过融合多视图注意力，性能进一步提升，该方法可动态调整不同视图的贡献（设置 (d)）。

关于稀疏体积重建的消融研究。表 5 详细分析了稀疏体积重建。设置 (a) 是基线方法，直接以固定分辨率 $40 \times 40 \times 16$ 构建 3D 体积。比较设置 (a)、(b) 和 (e)，我们观察到由粗到细的策略在保持性能的同时显著降低了计算成本。我们进一步在设置 (c)-(f) 中改变了精细化的选择比例。尽管减少选择比例提高了效率，但过小的选择比例（e.g., 10%）可能错过目标区域，降低检测精度。为了平衡准确性和计算负担，我们将选择比例设为 25% 。

关于占用损失的消融实验。如表 6 所示，去除占用

Table 5. 对稀疏体积重建的消融研究，包括精细化阶段的数量和精细化的选择比例。设置 (e) 被用于我们的 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)	$40 \times 40 \times 16$ (100 %)	61.0	36.0	31	24	22	1.33
(b)	$20 \times 20 \times 8$ (100 %) + $40 \times 40 \times 16$ (25 %)	60.6	35.6	21	21	14	1.40
(c)	$10 \times 10 \times 4$ (100 %) + $20 \times 20 \times 8$ (100 %) + $40 \times 40 \times 16$ (100 %)	61.3	36.2	34	26	26	1.28
(d)	$10 \times 10 \times 4$ (100 %) + $20 \times 20 \times 8$ (50 %) + $40 \times 40 \times 16$ (50 %)	60.9	35.4	22	22	15	1.40
(e)	$10 \times 10 \times 4$ (100 %) + $20 \times 20 \times 8$ (25 %) + $40 \times 40 \times 16$ (25 %)	61.2	35.2	20	19	13	1.46
(f)	$10 \times 10 \times 4$ (100 %) + $20 \times 20 \times 8$ (10 %) + $40 \times 40 \times 16$ (10 %)	57.0	31.7	19	19	13	1.53

Table 6. 占用损失的消融研究。

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

Table 7. 关于三维边界框标签质量的消融研究。

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. 深度网络中的模块消融及深度质量。

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

损失导致 mAP@0.25 下降 6.7 和 mAP@0.50 下降 6.2，说明显式占用监督的重要性。得益于我们基于 3D 边界框的伪标记策略，我们消除了对真实场景几何的依赖。

深入分析伪标签策略。3D 边界框可能会产生噪声的占用标签，特别是在盒子边界或杂乱的场景中。然而，这些噪声占用标签仅在训练阶段使用，作为占用预测的显式监督。在推理时，精选的前 25 % 用于细化，确保对已占用区域的充分覆盖，包括那些未被伪标签标注的区域（图 3）。为了进一步评估噪声边界框标注的影响，我们随机删除 15 % 的真实框，并在训练期间对剩余框进行 15 % 的随机缩放。这些不完美的标注影响了占用预测和 3D 检测的学习。然而，表 7 显示，与使用真实几何监督的 ImGeoNet [31] 相比，我们的 SGCDet 表现出更高的鲁棒性。

关于 DepthNet 的消融分析。我们在表 8 中展示了 DepthNet 的消融分析。如设置 (a)-(c) 所示，去除 DepthNet 的任何组件都会导致检测准确率的下降。接下来我们评估深度质量的影响。添加深度监督（设置 (d)）实现了 62.2 的 mAP@0.25 和 37.1 的 mAP@0.50。值得注意的是，这些结果甚至超过了需要多阶段训练管线和监督用场景几何真值的 CN-RMA [30]。设置 (e) 指的是直接使用真实深度作为输入，这表明我们的模型的上限。它揭示了通过对更精确深度估计的进一步研究可以有很大的提升空间。

4. 结论

我们提出了 SGCDet，这是一种新颖的多视角室内 3D 物体检测框架。为了增强体素特征的表示能力，我们引入了一个几何和上下文感知的聚合模块，该模块自适应地整合了多视角的图像特征。此外，我们开发了一种稀疏体积构建策略，该策略选择性地细化具有高占用概率的体素，显著减少了自由空间中的冗余计算。我们的框架仅使用 3D 边界框进行监督训练，消除了对真实场景几何信息的需求。实验结果表明，SGCDet 在 ScanNet、ScanNet200 和 ARKitScenes 数据集上实现了最先进的性能。

5.

致谢 部分工作得到了中国国家重点研发计划 2023YFB3209800 号项目的支持，部分得到了中国国家自然科学基金 62301484 号项目的支持，部分得到了宁波市自然科学基金 2024J454 号项目的支持，以及中国航空科学基金 2024M071076001 号项目的支持。我们也感谢浙江大学李思瑾提供的慷慨帮助。

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] 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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[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.

[13] Wuyang Li, Zhu Yu, and Alexandre Alahi. VoxDet: Rethinking 3d semantic occupancy prediction as dense object detection. arXiv preprint arXiv:2506.04623, 2025.

[14] Yinhao 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.

[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.

[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.

[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.

[18] Yingping Liang and Ying Fu. CascadeV-Det: Cascade point voting for 3d object detection. arXiv preprint arXiv:2401.07477, 2024.

[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.

[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.

[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.

[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.

[23] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983, 2016.

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[25] Yuhang Lu, Xinge Zhu, Tai Wang, and Yuexin Ma. OctreeOcc: Efficient and multi-granularity occupancy prediction using octree queries. In Advances in Neural Information Processing Systems, 2024.

[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.

[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.

[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.

[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.

[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.

[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.

[32] Haiyang Wang, Lihe Ding, Shaocong Dong, Shaoshuai Shi, Aoxue Li, Jianan Li, Zhenguo Li, and Liwei Wang. CAGroup3D: Class-aware grouping for 3d object detection on point clouds. In Advances in Neural Information Processing Systems, pages 29975–29988, 2022.

[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.

[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.

[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.

[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.

[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.

[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.

[39] Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. Unifying flow, stereo and depth estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.

[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.

[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.

[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.

[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.