OPUS: 使用稀疏集的占用预测

王家宝 [1]*; **刘钊江** [2]*, **孟强** [3], **颜柳江** [3], **王珂** [3], **杨杰** [3],
刘伟 [2], **侯淇彬** [1,4]†, **程明明** [1,4]
[1] 南开大学媒体计算实验室
[2] 上海交通大学 [3] 卡尔动力 [4] 南开国际先进研究院，深圳福田
https://github.com/jbwang1997/OPUS

Abstract

占用预测旨在预测三维环境中的占用状态，在自动驾驶领域的发展势头迅猛。主流的占用预测工作首先将三维环境离散为体素，然后在这种密集网格上进行分类。然而，对样本数据的检查发现，绝大多数体素是没被占用的。在这些空体素上进行分类需要次优的计算资源分配，而减少这些空体素则需要复杂的算法设计。为此，我们从一个新的角度提出了占位预测任务：将其表述为一个精简的集合预测范例，而无需明确的空间建模或复杂的稀疏化程序。我们提出的框架被称为 OPUS，它利用 Transformer 的编码器-解码器架构，使用一组可学习的查询同时预测占用位置和类别。首先，我们利用倒置距离损失（Chamfer distance loss）将集与集之间的比较问题扩展到前所未有的程度，从而使端到端训练这种模型成为现实。随后，根据学习到的位置，使用近邻搜索自适应地分配语义类别。此外，OPUS 还采用了包括从粗到细的学习、一致的点采样和自适应重加权的一系列非平凡策略来提高模型性能。最后，与当前最先进的方法相比，我们最轻的模型在 Occ3D-nuScenes 数据集上以接近 2× 的 FPS 实现了更优秀的 RayIoU，而我们最大的模型 RayIoU 比之前的最佳结果高出 6.1。

1 引言

与已有的成熟目标级别边界框表示方法 [7, 21, 18, 35, 27, 44**?**] 相比，基于体素的占用预测方法 [14, 33, 9, 34, 30] 能够为周围场景提供更精细的几何结构与语义信息。例如，使用边界框来描述车门开启的车辆或部署了支架的吊车是十分困难的，而占用表示可以自然地描述这类非常规形状。因此，占用预测正在自动驾驶领域快速获得关注。

目前的方法 [3, 42, 8, 25, 14**?**] 大多依赖于密集的数据表示，其中每个特征点与物理空间中的一个体素直接一一对应。然而我们观察到，绝大多数物理体素其实是空的。例如，

*Equal contribution. †Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

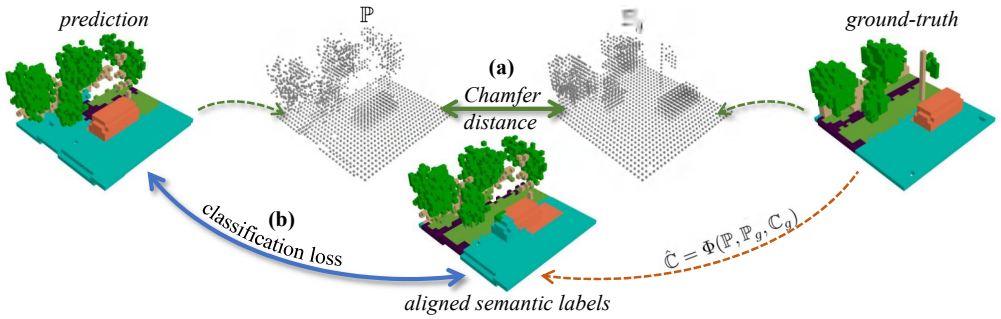

图 1: 我们将占用预测建模为集合预测问题。对于每一个场景，我们预测一组点的位置 $\mathbb{P}$ 和对应的语义类别 $\mathbb{C}$。给定标注的体素位置集合 $\mathbb{P}_g$ 和类别 $\mathbb{C}_g$，我们将集合到集合的匹配任务分解为两个子任务：(a) 通过 Chamfer 距离约束 $\mathbb{P}$ 与 $\mathbb{P}_g$ 的点分布相似性；(b) 基于与 $\mathbb{P}$ 最近的真实点集生成 $\mathbb{P}$ 的真实类别集合 $\hat{\mathbb{C}} = \Phi(\mathbb{P}, \mathbb{P}_g, \mathbb{C}_g)$，并将预测的类别集合 $\mathbb{C}$ 和之对齐。

在 SemanticKITTI 数据集 [1] 中，大约 67% 的体素是空的；而在 Occ3D-nuScenes 数据集 [34] 中，这一比例甚至超过 90%。这种稀疏的占用数据特性使得直接的密集表示极其低效，因为大部分计算资源都消耗在空体素上。为了解决这一低效问题，已有工作探索了稀疏的潜在表示，例如三视图表示 [33, 8] 或缩小解空间的方案 [19, 9]，显著降低了计算开销。然而，这些方法仍然将占用预测视为一个在特定位置进行分类的问题，依然依赖于复杂的中间表示设计和对三维空间的显式建模。

本工作中，我们提出将占用预测任务建模为一个直接的集合预测问题，直接回归被占用的位置并同时预测其语义类别。我们提出的名为 OPUS 的框架，利用了 Transformer 的编码器-解码器架构，主要包括：(1) 一个从多视角图像中提取 2 维特征的图像编码器；(2) 一组可学习的用于预测占用位置与语义类别的查询（queries）；(3) 一个用于利用相关图像特征来更新 query 特征的稀疏解码器。我们的 OPUS 无需显式建模三维空间或复杂的稀疏化设计，提供了一种简洁优雅的端到端占用预测方案。然而，一个关键挑战在于如何将预测结果与标注集合进行匹配，尤其是当标注结果没有顺序时。我们指出，虽然 Hungarian 算法被广泛应用于 DETR 系列方法 [4, 43, 26, 20, 11, 38] 中，但它并不适用于本任务。该算法的时间复杂度为 $O(n^3)$，空间复杂度为 $O(n^2)$，无法处理大规模体素集合。在我们的实验中，当匹配两个包含 1 万个点的集合时，Hungarian 算法在一张 80G 的 A100 显卡上耗时约 24 秒，占用约 2304MB 显存。而在 Occ3D-nuScenes [34] 中，单个场景的体素数可能高达 7 万个。因此直接应用 Hungarian 算法在占用预测语境下的集合匹配任务中并不可行。

但在占用预测中，是否真的需要精确的一一匹配呢？我们认为，进行一一匹配的目标是为了获得监督信号，即完整且准确的**点位置**与**语义类别**。如果我们能通过其他方式获得这些监督信号，那么复杂的一一匹配过程是可以完全省略的。因此，我们提出将占用预测任务拆分为两个并行子任务，如 Fig. 1所示：第一个子任务通过 Chamfer 距离，将预测点分布与标注点分布对齐，从而获得点位置的监督。这是点云领域广泛使用的技术 [5, 28]；第二个子任务通过将每个预测点分配至其最近的标注点，并继承其类别，实

现点分类监督。值得注意的是，这些操作均可并行执行，并且在 GPU 上运行效率极高。在 Occ3D-nuScenes 中，我们的匹配过程可在毫秒级时间内完成，几乎不占用显存。该方法的时间复杂度为 $O(n^2)$，空间复杂度为 $O(n)$，为大规模占用预测模型的训练铺平了道路。

此外，我们还提出了多个策略，以进一步提升我们的端到端稀疏框架下的占用预测性能，包括**粗到细的学习方式**、**一致性点采样策略**以及**自适应损失重加权机制**。在 Occ3D-nuScenes 上，我们所有模型及其变体的性能均明显超过现有所有方法，验证了我们方法的有效性与先进性。其中，我们最轻量的模型在速度提升超过 2× 的同时，比 SparseOcc [19] 提升了 3.3 的绝对 RayIoU 分数；而我们最强配置则取得了 41.2 的 RayIoU，相较现有最优方法提升了 14%，建立了新的性能上限。

我们的贡献总结如下：

- 据我们所知，本文首次将占用预测任务视为直接的集合预测问题，促进了稀疏框架的端到端训练。
- 我们引入了多个非平凡策略，包括粗到细学习、一致性采样、自适应重加权，有效提升了 OPUS 框架的性能。
- 在 Occ3D-nuScenes 数据集上的大量实验表明，OPUS 在保持实时推理速度的同时，在 RayIoU 指标上显著超越现有方法。

2 相关工作

2.1 占用预测

三维占据预测旨在推断三维空间中各体素的占据状态及语义类别，现已成为自动驾驶领域重要的基础感知任务并在学术和业界引发广泛关注。传统方法 [3, 42, 8, 25, 14, 37, 34**?**] 多采用连续密集特征表示，但由于占据数据固有的稀疏性易产生计算冗余。针对此问题，Tang et al. [33]通过三视图特征压缩密集特征实现高效建模。近期，基于 Transformer 的稀疏查询方法 [19, 9, 12] 逐渐兴起：OccupancyDETR [9] 通过目标检测引导的查询机制实现占据补全；VoxFormer [12] 从一组稀疏查询中生成三维体素，对应于借助深度估计的占用位置识别；SparseOcc [19] 则采用多级稀疏解码器逐级滤除空体素并预测剩余体素的占用状态。这些方法虽有效降低了计算成本，但仍需多阶段处理流程和复杂的空间建模。相较之下，本文方法直接通过稀疏查询来回归占用位置，无需预设位置先验，实现了优雅的的端到端占用预测。

2.2 基于 Transformer 的集合预测

DETR [4] 开创性地将 Transformer 应用于集合预测，通过稀疏查询机制生成带有特征和检测交互的无序检测结果。通过将目标检测任务视为一个直接的集合预测问题，该方法摒弃了复杂的后处理，实现了端到端的目标检测。后续研究 [43, 26, 20, 11, 38, 41, 32] 在性能提升与高效训练方面取得显著进展。稀疏查询范式在三维目标检测 [39, 21, 16–18, 36] 中亦获验证：DETR3D [39] 通过稀疏的三维目标查询来索引二维特征，将三维

的位置和多视角图像通过相机变换矩阵链接在一起；PETR [21] 将三维位置嵌入编码为二维图像特征，使查询能够直接聚合特征而不进行三维到二维的投影，以生成三维位置感知特征。Sparse4D [16] 将检测结果与时空特征融合进一步提升了三维目标检测的精度。尽管取得了巨大的成功，基于 Transformer 的集合预测主要面向目标检测任务，其查询数量受场景目标数限制。将集合预测扩展至占位预测面临巨大挑战，因所需查询数量呈数量级增长。

3 方法

本章中，我们首先在 Sec. 3.1 回顾当前基于查询的稀疏化方法在占位预测中的应用。随后在 Sec. 3.2 中，我们提出将该任务视为一个直接的集合预测问题的全新建模方式。最后，我们在 Sec. 3.3 中详细介绍所提出的 OPUS 框架。

3.1 回顾基于查询的占用稀疏化方法

利用稀疏查询的 Transformer 架构，为建模占用表示（occupancy representation）的稀疏性提供了新的可能性。其中一种典型的减少查询数量的方法是，如 PETRv2 [22] 所提出的，将每个查询分配给一块体素区域，而非单个体素。然而，这种方式仍会对整个三维空间进行密集预测，未能有效缓解冗余问题。另一种方式如 VoxFormer [12] 和 SparseOcc [19]，仅将查询分配给可能被占用的体素。VoxFormer 利用深度估计模块预测潜在的被占用体素，而 SparseOcc 则采用多阶段策略逐步剔除空区域。尽管如此，它们对体素占用状态的依赖导致了预测过程中的误差累积。此外，这些方法需要中间表示对三维空间进行复杂建模，难以实现端到端训练。

当前方法的根本问题在于它们将任务建模为一个分类问题，每个查询被限制在固定的空间区域内进行语义分类。这种限制极大地束缚了查询的灵活性，使其难以自适应地聚焦于关键区域。为解决这一问题，我们提出取消该空间限制，使每个查询能够自主决定其关注的区域。最终，我们将占用预测建模为一个集合预测任务：每个查询直接预测一组点的位置及其语义类别。

3.2 集合预测视角下的建模方式

本工作核心在于将占用预测视为一个集合预测问题。我们将 V_g 个真实占用体素表示为集合 $\{\mathbb{P}_g, \mathbb{C}_g\}$，其中 $|\mathbb{P}_g| = |\mathbb{C}_g| = V_g$。每个元素 $\{\mathbf{p}_g, \mathbf{c}_g\} \in \{\mathbb{P}_g, \mathbb{C}_g\}$ 表示一个体素中心的三维坐标 $\mathbf{p}_g$ 及其语义类别 $\mathbf{c}_g$。对于模型预测 V 个点得到的 $\{\mathbb{P}, \mathbb{C}\}$，核心挑战在于如何进行集合间的监督匹配，即：如何将无序的预测结果与有监督的目标对齐。一种可选方式是采用 Hungarian 匹配算法，但我们在附录中已讨论并实验证明其在大规模场景下的可扩展性受限。我们观察到，无须追求预测和真实值之间的一一对应，本质上匹配的目标是预测的点在位置和语义上的准确性，因此我们将任务拆解为两个并行目标：1. 使预测位置尽可能精确且覆盖全面；2. 保证每个预测点能够被赋予准确的语义类别。

第一个目标可通过 Chamfer 距离实现，广泛用于点云相关任务中 [5, 28, 10, 40]，具体定义如下：

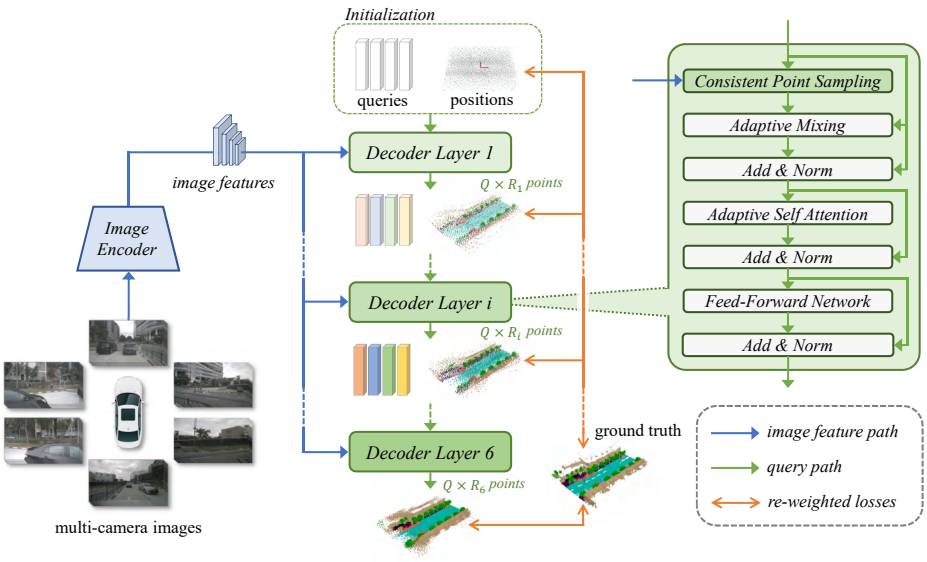

图 2: OPUS 采用了一个 Transformer 编码器-解码器架构，包括：(1) 一个图像编码器，用于从多视角图像中提取二维特征；(2) 一系列解码器，通过图像特征逐步细化查询点，这些特征通过 一致性点采样模块建立关联；(3) 一组可学习的查询点，用于预测占据点的位置和类别。每个查询点遵循一个 由粗到细的策略，逐步增加预测点的数量。最终，整个模型通过我们提出的 自适应重加权集合到集合损失进行端到端训练。

$$\mathrm{CD}(\mathbb{P}, \mathbb{P}_g) = \frac{1}{|\mathbb{P}|} \sum_{\mathbf{p} \in \mathbb{P}} D(\mathbf{p}, \mathbb{P}_g) + \frac{1}{|\mathbb{P}_g|} \sum_{\mathbf{p}_g \in \mathbb{P}_g} D(\mathbf{p}_g, \mathbb{P}), \text{ where } D(\mathbf{x}, \mathbb{Y}) = \min_{\mathbf{y} \in \mathbb{Y}} ||\mathbf{x} - \mathbf{y}||_1. \tag{1}$$

最小化 Chamfer 距离可以有效对齐预测点与真实点的分布，无需考虑点的排列顺序，从而实现对占用体素的直接学习。

第二个目标中，尽管 $\mathbb{C}$ 与 $\mathbb{C}_g$ 所对应的点位置不同，无法直接对比，但我们可以利用体素的局部空间一致性：同一物体中相邻点通常具有相同语义标签。据此，我们为每个预测点分配其在真实点集中最近邻体素的语义类别，具体为：

$$\{\hat{\mathbb{C}}, \hat{\mathbb{P}}\} = \left\{ \arg\min_{\{\mathbf{c}_g, \mathbf{p}_g\} \in \{\mathbb{C}_g, \mathbb{P}_g\}} \|\mathbf{p}_g - \mathbf{p}\|_2, \quad \mathbf{p} \in \mathbb{P} \right\}. \tag{2}$$

这里，$\hat{\mathbb{C}}$ 是用以监督预测类别 $\mathbb{C}$ 的标签集合。

值得一提的是，Eq. (1) 与 Eq. (2) 中的计算均可在 GPU 上高效并行执行，单次匹配在毫秒级内完成，使得将占用预测作为集合预测问题成为可能，从而支持大规模端到端训练。接下来我们深入研究 OPUS 的架构。

3.3 OPUS 框架详解

本部分介绍 *OPUS* 框架，如 Fig. 2 所示。首先，从多视角图像中提取图像特征，并初始化一组可学习的查询 $\mathbb{Q}$、点位置 $\mathbb{P}$ 和置信得分 $\mathbb{C}$。随后，这些查询特征与预测结果被输入到一系列解码器中，通过与图像特征的关联进行迭代式细化。在每个阶段，预测的位置和得分都会接受来自真实标签的监督，从而实现整个框架的端到端训练。可以看

出，我们的核心结构是由多个解码器构成的序列。因此，接下来我们将详细介绍解码器的输入/输出形式，以及特征在解码器内部的聚合与更新方式。

基本符号

令 $\{\mathbb{Q}_0, \mathbb{P}_0, \mathbb{C}_0\}$ 为初始可学习查询、点位置信息和类别预测分数；经过第 i 个解码器后的输出为 $\{\mathbb{Q}_i, \mathbb{P}_i, \mathbb{C}_i\}$。这些集合的长度都是 Q，对应着查询的数量。每个查询特征 $\mathbf{q}_i \in \mathbb{Q}_i, i \in \{0, 1, \cdots, 6\}$ 的通道数为 $C = 256$。为了减少查询的数量来提升效率，我们设计每个查询一次预测 R_i 个点，因此 $\mathbf{p}_i \in \mathbb{P}_i$ 和 $\mathbf{c}_i \in \mathbb{C}_i$ 的尺寸分别为 $Q \times R_i \times 3$ 和 $Q \times R_i \times N$，其中 N 为语义类别数。

粗到细的预测策略

高层语义信息通常难以直接从低层特征中准确预测。因此，我们不试图去预测整个三维环境的占用，而是在早期阶段允许模型仅预测稀疏的占用点，如 Fig. 2 所示。为了实现这个目标，我们采用逐步细化策略，使得预测点数量随阶段递增，即 $R_{i-1} \leq R_i$ 对 $i \in \{1, 2, \cdots, 6\}$。

Chamfer 距离相比 Hungarian 算法具有天然优势：即使预测数量少于真实点，也不会陷入局部匹配，因为 Hungarian 算法可能由于缺少分布的限制，将预测指派给任意的真实标注的子集。Chamfer 更关注点集的整体分布，而非一对一的严格匹配。这使得预测点即使数量有限，也能较均匀地分布，并且能代表真实的三维空间。

解码器结构

我们的解码器与 SparseBEV [18] 中的解码器类似，后者是一种高效且稀疏的目标检测器。对于给定的查询 $\mathbf{q}_{i-1} \in \mathbb{Q}_{i-1}$ 及其对应的点位置 $\mathbf{p}_{i-1} \in \mathbb{P}_{i-1}$，第 i 个解码器首先通过一致性点采样模块进行图像特征聚合，该新方案将在后文中详细讨论。随后，通过自适应地混合图像特征与查询特征，并结合所有查询之间的自注意力机制（与 SparseBEV 中的操作相同），将查询特征更新为 $\mathbf{q}_i$。最后，一个仅由 Linear、LayerNorm 和 ReLU 层组成的预测模块生成语义类别 $\mathbf{c}_i$（尺寸 $R_i \times N$）和位置偏移量 $\Delta\mathbf{p}_i$（尺寸 $R_i \times 3$）。由于 $\Delta\mathbf{p}_i$ 与 $\mathbf{p}_{i-1}$ 在维度上不匹配，无法直接相加，我们首先沿第一维对 $\mathbf{p}_{i-1}$ 求均值，然后将结果复制 R_i 次，得到 $\mathbf{p}_{i-1}$。最终位置为 $\mathbf{p}_i = \bar{\mathbf{p}}_{i-1} + \Delta\mathbf{p}_i$。

一致性点采样 SparseBEV 中所用的特征采样方法并不适用于我们的框架，因为它是针对检测输入设计的。因此，我们提出了一种新的“一致性点采样”（Consistent Point Sampling, CPS）流程，用于从 M 张图像特征中采样 3D 点并聚合相应特征。给定输入 $\{\mathbf{q}, \mathbf{p}\} \in \{\mathbb{Q}, \mathbb{P}\}$，我们先采样 S 个点，并通过下式在第 m 张图像特征中计算它们的坐标：

$$\mathbf{c}_m = \mathbf{T}_m \mathbf{r}, \quad \text{其中 } \mathbf{r} = \mathbf{m}_p + \phi(\mathbf{q}) \cdot \sigma_p, \tag{3}$$

这里，$\mathbf{T}_m$ 表示从当前 3D 空间到第 m 张图像坐标系的投影矩阵；$\phi(\mathbf{q})$ 是一个线性映射层，用于从查询特征 $\mathbf{q}$ 生成 S 个 3D 点；$\mathbf{m}_p$ 和 σ_p 分别表示 $\mathbf{p}$ 中 R 个点的均值和标准差。值得注意的是，我们用标准差 σ_p 对预测偏移量 $\phi(\mathbf{q})$ 进行重加权，以继承先

前预测的离散程度。本质上，当输入点集 $\mathbf{p}$ 分布较广时，我们会进行更大范围的采样；反之，则在更窄范围内采样。这一操作在我们的实验中显著提升了预测性能。

并非所有 $\mathbf{c}_m$ 中的坐标都是可行的，因为采样点可能在对应相机中不可见。因此，我们生成一个掩码集合 $\mathbb{V}_m$，对于每个 $s \in \{1, 2, \ldots, S\}$ 和 $m \in \{1, 2, \ldots, M\}$，若 $\mathbf{c}_{s,m}$ 有效则掩码值 $v_{s,m} = 1$，否则为 0。接着，我们从图像特征集合 $\{F_m\}_{m=1}^{M}$ 中聚合信息，用于后续的自适应混合阶段。具体地，我们有：

$$f_s = \frac{1}{\sum_{m=1}^{M} |\mathbb{V}_m|} \sum_{s=1}^{S} \sum_{m=1}^{M} w_{s,m}\, v_{s,m}\, \mathcal{B}(F_m, \mathbf{c}_{s,m}), \tag{4}$$

其中，$v_{s,m}$ 表示掩码集合 $\mathbb{V}_m$ 中的第 s 个元素，$\mathbf{c}_{s,m}$ 是第 s 个点 $\mathbf{r}_s$ 在第 m 张图像特征中的映射坐标。操作 $\mathcal{B}$ 表示双线性插值，$w_{s,m}$ 是从查询特征 $\mathbf{q}$ 通过线性变换生成的第 m 张图像特征对点 $\mathbf{r}_s$ 的权重。

自适应重加权的训练损失我们的框架训练目标是用真实标注 $\{\mathbb{P}_g, \mathbb{C}_g\}$ 来监督 $\{\mathbb{P}_i, \mathbb{C}_i\}_{i=1}^{6}$ 的学习。点的位置可以通过 Eq. (1) 进行训练。然而，原始的 Chamfer 距离损失侧重于点分布的整体相似性，忽略了每个点是否足够准确。这一点在我们的实验中导致了不理想的性能表现。为了解决该问题，我们采用了一种简单而有效的重加权策略，以强调误差较大的点，并将 Chamfer 距离损失修改如下：

$$\begin{aligned} \mathrm{CD}_R(\mathbb{P}, \mathbb{P}_g) &= \frac{1}{|\mathbb{P}|} \sum_{\mathbf{p} \in \mathbb{P}} D_R(\mathbf{p}, \mathbb{P}_g) + \frac{1}{|\mathbb{P}_g|} \sum_{\mathbf{p}_g \in \mathbb{P}_g} D_R(\mathbf{p}_g, \mathbb{P}), \\ \text{其中} D_R(\mathbf{x}, \mathbb{Y}) &= W(d) \cdot d, \quad d = \min_{\mathbf{y} \in \mathbb{Y}} \|\mathbf{x} - \mathbf{y}\|_1. \end{aligned} \tag{5}$$

这里，$W(d)$ 是对距离最近真实点较远的点进行惩罚的重加权函数。在我们的实现中，当 $d \geq 0.2$ 时取 $W(d) = 5$，否则取 $W(d) = 1$。

对于分类任务，我们首先使用 Eq. (2) 为 $\mathbb{C}_i$ 生成目标类别 $\hat{\mathbb{C}}_i$。随后，可用常规的分类损失对语义类别进行训练。在我们的实现中，采用了带有手动搜索权重的焦点损失（focal loss）[15]，并将修改后的损失记为 $\mathrm{FocalLoss}_R$。最后，所提方法 OPUS 的训练目标为

$$L_{\mathrm{OPUS}} = \mathrm{CD}_R(\mathbb{P}_0, \mathbb{P}_g) + \sum_{i=1}^{6} \big(\mathrm{CD}_R(\mathbb{P}_i, \mathbb{P}_g) + \mathrm{FocalLoss}_R(\mathbb{C}_i, \hat{\mathbb{C}}_i)\big), \tag{6}$$

其中 $\mathrm{CD}_R(\mathbb{P}_0, \mathbb{P}_g)$ 显式地鼓励初始点 $\mathbb{P}_0$ 捕捉数据集的一般模式。

4 实验

4.1 实验设置

数据集和评估指标。所有模型均在 Occ3D-nuScenes [34] 数据集上进行评估，该数据集为大规模 nuScenes [2] 基准测试中的 18 个类别（1 个自由类别和 17 个语义类别）提供占用标签。在 1,000 个带标签的驾驶场景中，分别使用 750、150 和 150 个场景进行训练、验证和测试。评估中使用常用的 mIoU 指标。最近，SparseOcc [19] 指出过估计很容易破坏 mIoU 指标，并提出了 RayIoU 作为解决方案。因此，按照他们的工作，我们还在 1 米、2 米和 4 米的不同距离阈值下报告 RayIoU 结果，分别表示为 $\mathrm{RayIoU}_{1\mathrm{m}}$、$\mathrm{RayIoU}_{2\mathrm{m}}$ 和 $\mathrm{RayIoU}_{4\mathrm{m}}$。最终的 RayIoU 得分为这三个值的平均值。

表 1: Occ3D-nuScenes [34] 上的占用预测性能。“8f” 和 “16f” 分别表示从 8 或 16 帧融合时间信息的模型。基线结果直接从相应论文或 SparseOcc [19] 复制而来。FPS 结果在 A100 GPU 上测得。

方法	骨干网络	图像大小	mIoU	RayIoU_{1m}	RayIoU_{2m}	RayIoU_{4m}	RayIoU	FPS
RenderOcc [29]	Swin-B	1408×512	24.5	13.4	19.6	25.5	19.5	-
BEVFormer [13]	R101	1600×900	39.3	26.1	32.9	38.0	32.4	3.0
BEVDet-Occ [7]	R50	704×256	36.1	23.6	30.0	35.1	29.6	2.6
BEVDet-Occ (8f) [7]	R50	704×384	39.3	26.6	33.1	38.2	32.6	0.8
FB-Occ (16f) [14]	R50	704×256	39.1	26.7	34.1	39.7	33.5	10.3
SparseOcc (8f) [19]	R50	704×256	-	28.0	34.7	39.4	34.0	17.3
SparseOcc (16f) [19]	R50	704×256	30.6	29.1	35.8	40.3	35.1	12.5
OPUS-T (8f)	R50	704×256	33.2	31.7	39.2	44.3	38.4	22.4
OPUS-S (8f)	R50	704×256	34.2	32.6	39.9	44.7	39.1	20.7
OPUS-M (8f)	R50	704×256	35.6	33.7	41.1	46.0	40.3	13.4
OPUS-L (8f)	R50	704×256	36.2	34.7	42.1	46.7	41.2	7.2

实现细节。按照以往工作 [19, 14, 7]，我们将图像大小调整为 704×256，并使用 ResNet50 [6] 骨干网络提取特征。我们将一系列模型分别标记为 OPUS-T、OPUS-S、OPUS-M 和 OPUS-L，查询数量分别为 0.6K、1.2K、2.4K 和 4.8K。在每个模型中，所有查询预测的点数相同，最终阶段总计为 76.8K 点。在我们的 CPS 中，OPUS-T 的采样数量为 4，其他模型为 2。更多模型细节请参见 Appendix D.2。所有模型均在 8 个 nvidia 4090 GPU 上使用 AdamW [24] 优化器进行训练，批量大小为 8。学习率在前 500 次迭代中逐步升至 $2e^{-4}$，然后按照余弦退火 [23] 方案衰减。除非另有说明，主结果中的模型训练 100 个周期，消融研究中的模型训练 12 个周期。

4.2 主要结果

定量性能。在此部分中，我们将 OPUS 与 Occ3D-nuScenes 数据集上的先前最先进方法进行了比较。我们的方法不仅在 RayIoU 方面实现了卓越的性能，在 mIoU 方面也取得了具有竞争力的结果，并且展示了值得称道的实时性能。如 Tab. 1 所示，OPUS-T (8f) 达到了 22.4 FPS，显著快于密集对应方法，几乎是稀疏对应方法 SparseOcc (8f) 速度的 1.3 倍。尽管仅使用了 7 个历史帧，其 38.4 RayIoU 结果轻松超越了其他模型，包括 RayIoU 为 33.5(-4.9) 的 FB-Occ (16f) 和 RayIoU 为 35.1(-3.3) 的 SparseOcc (16f)。同样，OPUS-S (8f) 和 OPUS-M (8f) 在性能和效率之间取得了良好的平衡。最大的 OPUS 版本最终实现了 41.2 的 RayIoU，比之前的最佳结果好 6.1RayIoU，展现了显著优势。

在预测点总数相同的情况下，我们通过改变查询数量并相应地改变每个查询的点数，得到了不同版本的 OPUS。可以观察到，随着查询数量的增加，FPS 值从 22.4 降至 7.2，同时在 mIoU 和 RayIoU 方面提升了模型性能。OPUS-M (8f) 使用 2.4K 查询，在保持竞争力的 FPS 的同时实现了相当的 RayIoU，从而达到了平衡。

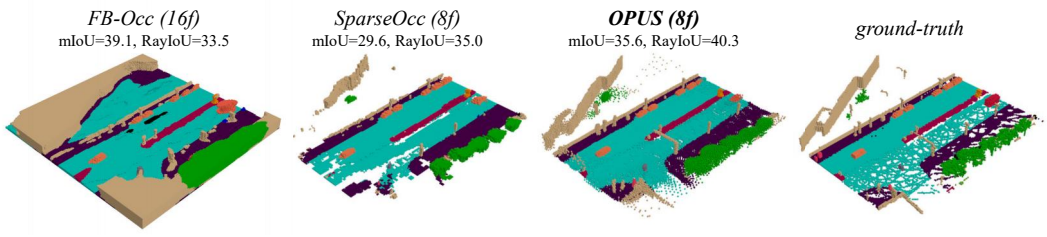

图 3: 占用预测的可视化。**最好以彩色查看**。

表 2: 不同策略组合下的模型性能。

CD_R	$FocalLoss_R$	CPS	粗到细	mIoU	$RayIoU_{1m}$	$RayIoU_{2m}$	$RayIoU_{4m}$	RayIoU
				17.4	23.6	29.7	34.3	29.2
✓				23.7 (6.3↑)	23.9	30.7	35.6	30.1 (0.9↑)
✓	✓			25.1 (1.4↑)	25.2	32.3	37.0	31.5 (1.4↑)
✓	✓	✓		25.5 (0.4↑)	26.0	33.1	37.9	32.3 (0.8↑)
✓	✓	✓	✓	27.2 (1.7↑)	26.1	33.3	38.4	32.6 (0.3↑)

尽管 mIoU 指标容易受到高估操作的影响 [19]，我们的 OPUS 仍实现了 36.2 mIoU，显著缩小了密集和稀疏模型在该指标上的差距。这些在不同指标下的结果共同证明了我们 OPUS 的优越性。

可视化我们在 Fig. 3 中可视化了预测的占用情况。可以观察到，与稀疏方法相比，FB-Occ 倾向于产生更密集的结果。尽管在 3D 环境中看起来完整，但其预测的占用结果是严重高估的，特别是对于远距离区域。高估可能会破坏 mIoU 指标 [19]，而主要考虑沿光线的第一个占用体素的 RayIoU 则会对其进行严厉惩罚。因此，FB-Occ 实现了 39.1 的最佳 mIoU 但 RayIoU 值最差。另一方面，SparseOcc 偶尔会在长距离出现不连续的假阴性预测。这归因于 SparseOcc 逐渐移除空体素，导致早期阶段的错误过滤累积并导致最终的错误预测。相比之下，我们的 OPUS 由于其端到端方法，保持了更连续的预测，从而实现了更合理的可视化。

4.3 消融研究和可视化

此部分详细介绍了使用 OPUS-M (8f) 模型进行的消融研究和可视化。

OPUS 中提出策略的影响。在我们的工作中，我们引入了 Chamfer 距离损失和 focal 损失的自适应重加权，以及一致点采样和粗到细预测策略。我们在 Tab. 2 中检查了这些策略的影响。在没有任何装饰的情况下，OPUS 实现了基线 17.4 mIoU 和 29.2 RayIoU。将原始 CD 损失替换为我们的修订版 CD_R 显著提升 mIoU 和 RayIoU 的结果分别为 6.4 和 0.9，证明了在此任务中关注错误预测位置的重要性。$FocalLoss_R$ 进一步将两个指标分别提高了 1.4。在 Eq. (3) 中引入 σ_p 项进一步分别提升了 mIoU 和 RayIoU 0.4 和 0.8，证明了在当前采样过程中考虑先前采样分布的有效性。我们提出的粗到细查询预测逐渐增加了阶段中的点数。该方案不仅减少了早期阶段的计算量，还显著提升了模型性能，特别是在 mIoU 方面提高了 1.7。这些结果突出了每个组件的累积优势，展示了它们的集成如何带来显著的性能提升。

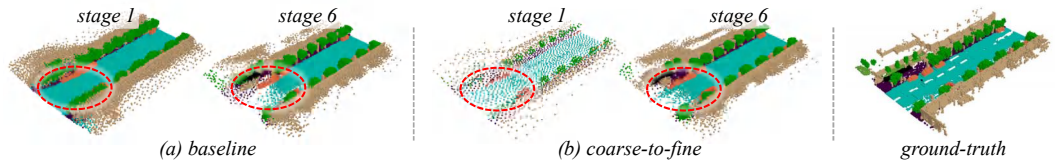

图 4: 粗到细预测的可视化。

表 3: 初始点 $\mathbb{P}^0$ 的不同处理方法的比较。“网格”和“随机”分别表示点在 BEV 空间中均匀采样和在 3D 空间中随机采样“优化”表示点随机初始化但通过 CD_R 损失对真实值进行监督。

类型	mIoU	RayIoU_{1m}	RayIoU_{2m}	RayIoU_{4m}	RayIoU
网格	22.8	22.2	28.9	33.9	28.3
随机	23.1	23.6	30.5	35.6	29.9
优化	23.7	23.9	30.7	35.6	30.1

粗到细预测的可视化。我们在 Fig. 4 中可视化了不同阶段的预测结果。在基线场景中，如 Fig. 4(a) 所示，所有解码器回归相同数量的点，我们观察到跨阶段的点分布不一致以及远距离中的许多假阴性预测，如圆圈所示。这可能归因于在早期阶段学习细粒度占用表示的困难，阻碍了整个框架的有效训练。相比之下，我们的粗到细策略显著缓解了早期阶段的学习难度，从而提升了模型性能。因此，不同阶段的点分布更加一致，最终预测显示出更少的假阴性，如 Fig. 4(b) 所示。

预测点的可视化。在 Fig. 6 中，我们选择了一些查询并可视化了它们的预测点。值得注意的是，大多数查询表现出倾向于预测具有相同类别甚至来自同一实例的点，如 Fig. 6(a)-(g) 所示。一个有趣的观察是，对于具有大体积的类别（如可行驶表面和人行道），预测点倾向于表现出多样的分布。相反，对于尺寸有限的物体（如交通锥、摩托车和汽车），点的分布相对于实例尺寸更加紧密。这些模式可以通过 Fig. 5 进一步验证，其中我们展示了来自三个选定类别的查询的点的标准差。这些结果突出了我们的模型在适应不同物体类别独特空间特征方面的有效性。

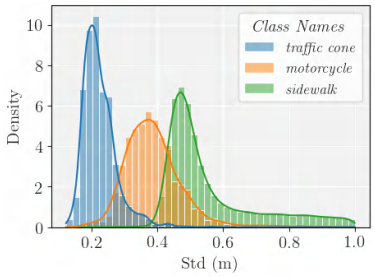

图 5: 来自一个查询的点的标准差分布。

由于我们没有明确约束单个查询中的点具有相同的类别，因此一个查询可能会产生不同类别的点。我们发现这种现象通常发生在物体边界处。然而，即使类别不同，这些点仍然紧密分布，如 Fig. 6(h)-(j) 所示。

初始点处理的影响 Tab. 3 比较了对初始点 $\mathbb{P}^0$ 的三种不同处理方法。网格初始化将 BEV 空间划分为均匀分布的柱状体，并有序地将柱状体中心指定为初始位置，这种方法在 BEVFormer [13] 中使用。随机初始化在 3D 空间中均匀分布地为每个位置分配值。初始化后，$\mathbb{P}^0$ 在训练期间仍可学习。在随机初始化的基础上，我们的 OPUS 进一步对

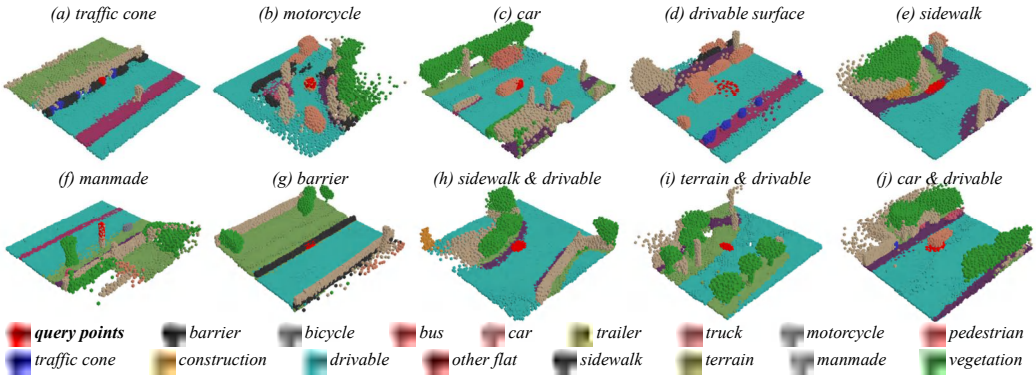

图 6: 不同查询生成的点的可视化。**最佳以彩色查看**。

$\mathbb{P}^0$ 添加了真实值分布的监督（即，Eq. (6) 中的 $\text{CD}_R(\mathbb{P}^0, \mathbb{P}_g)$。Tab. 4 中的结果表明，随机初始化优于网格初始化，实现了 23.1 的 mIoU 对比 22.8，以及 29.9 的 RayIoU 对比 28.3。这一改进可能归因于随机初始化提供了更多样化的 3D 分布。此外，引入的监督在 mIoU 上额外提高了 0.6，在 RayIoU 上提高了 0.2。这些结果揭示了随机初始化的效率以及对初始位置的额外监督的有效性。

自注意力的可视化为了更好地理解查询点关注的内容，我们可视化了每个查询中具有最高 10 个自注意力权重的查询点。我们将它们投影到 2D 图像上以便更好地可视化。以下是来自 Fig. 7 的一些有趣发现。一般来说，查询倾向于关注邻近的查询点。例如，第一幅图中的路边查询和第二幅图中的汽车查询允许从邻近查询点流入本地信息，使查询能够捕获详细的本地信息。此外，即使查询点不非常接近，查询仍能关注语义相关的位置。例如，Fig. 7 第一幅图中的植被查询不仅关注树干，还关注草地，表明查询能够捕获语义相关的信息以进行更准确的预测。另一个值得注意的观察是，查询可以感知地形相关的信息。例如，在第三幅图中，路边查询点沿着道路边缘的直线关注其他点，突出了模型理解环境场景相关结构的能力。

不同稀疏化策略的比较 在 Tab. 4 中，我们将 OPUS 与两种具有不同稀疏化策略的模型进行了比较。第一个基线是 SparseOcc，它通过在各种级联阶段过滤空体素来实现稀疏化。按照 PETRv2 [22] 的方法，第二个基线是一种基于柱状块的方法，将 3D 空间划分为少量的柱状块。我们使用 50×50 查询，每个查询对应对邻近 $4 \times 4 \times 16$ 个体素进行分类。为了公平比较，所有这些模型都训练了 100 个周期。相比之下，我们的模型在充分训练后以 38.0 的 RayIoU 得分取得了最佳结果，显著超越了 SparseOcc 的

图 7: 解码器中的自注意力。对于每个支点（标记为 ×），具有最高 10 个自注意力权重的查询点以圆圈显示，圆圈大小与权重成正比。**最佳以彩色查看**。

表 4: 不同稀疏化策略的对比。

模型	Q	R	RayIoU_{1m}	RayIoU_{2m}	RayIoU_{4m}	RayIoU	FPS
SparseOcc	(4000/16000/64000)		28.4	34.9	39.6	34.3	17.3
PETR v2	2500	256	24.4	31.0	36.3	30.6	13.8
OPUS	2400	32	31.7	38.8	43.4	38.0	13.4

表 5: 在 Waymo-Occ3D 数据集上的性能表现。

模型	General	Vehicle	Bicyclist	Ped.	Sign	Tfc. light	Pole	Cons. cone	Bicycle	Motorcycle	Building	Vegetaion	Treetrunk	Road	Sidewalk	mIoU	RayIoU	FPS
BEVDet	0.13	13.06	2.17	10.15	7.80	5.85	4.62	0.94	1.49	0.0	7.27	10.06	2.35	48.15	34.12	9.88	-	-
TPVFormer	3.89	17.86	12.03	5.67	13.64	8.49	8.90	9.95	14.79	0.32	13.82	11.44	5.8	73.3	51.49	16.76	-	-
BEVFormer	3.48	17.18	13.87	5.9	13.84	2.7	9.82	12.2	13.99	0.0	13.38	11.66	6.73	74.97	51.61	16.76	-	4.6
CTF-Occ	6.26	28.09	14.66	8.22	15.44	10.53	11.78	13.62	16.45	0.65	18.63	17.3	8.29	67.99	42.98	18.73	-	2.6
OPUS-L	4.66	27.07	19.39	6.53	18.66	6.41	11.44	10.40	12.90	0.0	18.73	18.11	7.46	72.86	50.31	19.00	24.7	8.5

34.3。另一方面，我们的模型也可以实现实时速度。这些结果证明了我们稀疏化程序的优越性。

在 Waymo-Occ3D 数据集上的比较我们进一步在 Waymo-Occ3D [31] 数据集上简单实现了 OPUS，以探索 OPUS 的泛化和鲁棒性。由于 Waymo-Occ3D 通常不作为以视觉为中心的方法的标准基准测试，我们在该数据集上唯一找到的具有报告结果的基于视觉的方法是 Occ3D 论文，该论文评估了 BEVDet、TPVFormer、BEVFormer 和新提出的 CTF-Occ [34]。为了与这些基线公平比较，我们在数据集的 20 如 Tab. 5 所示，尽管没有微调训练配置，OPUS-L 仍然实现了 19.0 的 mIoU，超越了所有先前方法。此外，OPUS-L 在 Waymo-Occ3D 数据集上还达到了 8.5 FPS，这大约是 CTF-Occ 速度的 3 倍，BEVFormer 的 2 倍。

5 结论和限制

本文提出了一种全新的占用预测建模视角，将其表述为一个集合预测（set prediction）问题。基于 Transformer 编码器-解码器架构，所提出的 OPUS 从一组可学习查询中并行预测被占用的位置与语义类别。预测结果与真实值之间的匹配通过两个高效的并行子任务完成，从而支持大规模点集上的端到端训练。此外，查询特征通过一系列非平凡的设计得到增强（*i.e.*，粗到细学习、一致性点采样以及损失重加权），从而显著提升了预测性能。在 Occ3D-nuScenes 基准上的实验表明，OPUS 在精度与效率方面均超越现有方法，得益于其稀疏化设计。

然而，OPUS 也带来了新的挑战，特别是在收敛速度方面。该问题或可借鉴 DETR 后续工作中的策略来缓解，这些方法在很大程度上解决了原始 DETR 的收敛问题。另一个挑战是，尽管稀疏方法通常在 RayIoU 指标上优于密集方法，但在 mIoU 上则相对较弱。如何在保持 RayIoU 优势的同时提升 mIoU 表现，是未来值得探索的方向。最后，

尽管我们当前实验仅基于视觉模态，所提出的核心建模方式同样适用于多模态任务。我们将多模态占用预测作为未来工作的一部分。

致谢

本研究得到国家自然科学基金（编号 62225604，编号 62276145）、中央高校基本科研业务费（南开大学，070-63223049）的支持，并部分得到国家自然科学基金（编号 62376153，62402318，24Z990200676）的资助。计算由南开大学超算中心（NKSC）提供支持。

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[3] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022.

[4] 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.

[5] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 605–613, 2017.

[6] 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.

[7] 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.

[8] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023.

[9] Yupeng Jia, Jie He, Runze Chen, Fang Zhao, and Haiyong Luo. Occupancydetr: Making semantic scene completion as straightforward as object detection. *arXiv preprint arXiv:2309.08504*, 2023.

[10] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1116–1124, 2023.

[11] Feng Li, Hao Zhang, Shilong Liu, Jian Guo, Lionel M Ni, and Lei Zhang. Dn-detr: Accelerate detr training by introducing query denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13619–13627, 2022.

[12] 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.

[13] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird' s-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022.

[14] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[16] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

[17] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d v2: Recurrent temporal fusion with sparse model. *arXiv preprint arXiv:2305.14018*, 2023.

[18] 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.

[19] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023.

[20] Shilong Liu, Feng Li, Hao Zhang, Xiao Yang, Xianbiao Qi, Hang Su, Jun Zhu, and Lei Zhang. Dab-detr: Dynamic anchor boxes are better queries for detr. *arXiv preprint arXiv:2201.12329*, 2022.

[21] 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.

[22] Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Aqi Gao, Tiancai Wang, and Xiangyu Zhang. Petrv2: A unified framework for 3d perception from multi-camera images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3262–3272, 2023.

[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. In *International Conference on Learning Representations*, 2018.

[25] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. *arXiv preprint arXiv:2312.01919*, 2023.

[26] Depu Meng, Xiaokang Chen, Zejia Fan, Gang Zeng, Houqiang Li, Yuhui Yuan, Lei Sun, and Jingdong Wang. Conditional detr for fast training convergence. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3651–3660, 2021.

[27] Qiang Meng, Xiao Wang, JiaBao Wang, Liujiang Yan, and Ke Wang. Small, versatile and mighty: A range-view perception framework. *arXiv preprint arXiv:2403.00325*, 2024.

[28] Benedikt Mersch, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In *Conference on Robot Learning*, pages 1444–1454. PMLR, 2022.

[29] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. *arXiv preprint arXiv:2309.09502*, 2023.

[30] Yining Shi, Kun Jiang, Ke Wang, Kangan Qian, Yunlong Wang, Jiusi Li, Tuopu Wen, Mengmeng Yang, Yiliang Xu, and Diange Yang. Effocc: A minimal baseline for efficient fusion-based 3d occupancy network. *arXiv preprint arXiv:2406.07042*, 2024.

[31] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[32] Zhiqing Sun, Shengcao Cao, Yiming Yang, and Kris M Kitani. Rethinking transformer-based set prediction for object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3611–3620, 2021.

[33] Pin Tang, Zhongdao Wang, Guoqing Wang, Jilai Zheng, Xiangxuan Ren, Bailan Feng, and Chao Ma. Sparseocc: Rethinking sparse latent representation for vision-based semantic occupancy prediction. *arXiv preprint arXiv:2404.09502*, 2024.

[34] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024.

[35] Jiabao Wang, Qiang Meng, Guochao Liu, Liujiang Yan, Ke Wang, Ming-Ming Cheng, and Qibin Hou. Towards stable 3d object detection. In *European conference on computer vision*. Springer, 2024.

[36] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023.

[37] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023.

[38] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2567–2575, 2022.

[39] 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.

[40] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14673–14684, 2024.

[41] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2022.

[42] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023.

[43] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2020.

[44] Ziyue Zhu, Qiang Meng, Xiao Wang, Ke Wang, Liujiang Yan, and Jian Yang. Curricular object manipulation in lidar-based object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1125–1135, 2023.

附录

A 广泛影响

我们的工作提出了一种用于占用预测的端到端范式，以快速推理速度实现了最先进的 RayIoU 性能。这一进步可以带来实时且精确的占用结果，这对于自动驾驶（AD）的实际应用至关重要。因此，我们工作的最大积极社会影响是提高了 AD 系统的安全性和响应速度。

然而，与任何 AD 系统组件一样，这项工作的最大负面社会影响是安全性问题。自动驾驶系统与人类生命直接相关，错误的预测或决策可能导致危险的结果。因此，在提高占用结果的准确性以及开发解决错误预测的补充方法方面，需要大量的后续工作。

B 所涉及资产的许可证

我们的代码建立在 SparseBEV 提供的代码库[1] 之上，该代码库受 MIT 许可证约束。我们的实验是在 Occ3D-nuScenes 上进行的，该数据集为 nuScenes 数据集提供占用标签。Occ3D-nuScenes 也采用 MIT 许可证，而 nuScenes 则采用 CC BY-NC-SA 4.0 许可证。

C 复杂度分析

在本部分中，我们提供了将 m 个预测与 n 个真实值进行匹配所涉及的时间和空间复杂度的详细分析。

匈牙利算法。匈牙利算法的核心是寻找 $\min(m,n)$ 次迭代中的增广路径。每次迭代可以视为试图通过在残差图中找到最短增广路径来改进当前匹配，而残差图使用 Dijkstra 算法的复杂度为 $O(\max(m,n)^2)$。因此，匈牙利算法的时间复杂度为 $O(\min(m,n)\cdot \max(m,n)^2)$。

与此同时，匈牙利算法需要计算一个大小为 $m \times n$ 的成本矩阵来存储与每个潜在分配相关的成本。在匹配过程中，跟踪的标签和匹配对每个需要 $O(\min(m,n))$ 空间。因此，最终的空间复杂度为 $O(m \times n)$。

我们的方法。我们的方法采用了查姆弗距离损失，这涉及到计算成对距离并确定每个点的最小距离。第一步需要 $O(m \times n)$ 的时间复杂度，下一步也需要 $O(m \times n)$。语义标签的分配可以重用先前最近搜索的结果，因此不需要额外的计算。最终，时间复杂度为 $O(m \times n)$。

对于每个集合中的一个点，算法需要跟踪到另一个集合中任何点的最小距离。这可以通过为每个点使用一个变量来实现，分别在两个方向上实现 $O(m)$ 和 $O(n)$。同时，语义标签分配的空间复杂度为 $O(m)$。总体而言，这总计为 $O(2m+n)$。

两种方法的比较。总之，当 m 和 n 在规模上相当且可比较时，匈牙利算法展示了 $O(n^3)$ 的时间复杂度和 $O(n^2)$ 的空间复杂度，而我们的方法分别以 $O(n^2)$ 和 $O(n)$ 的复杂度

[1] https://github.com/MCG-NJU/SparseBEV

显著提高了效率。这代表了在时间和空间需求上的显著减少，使其成为大规模应用的更高效解决方案。

D 附加实验。

D.1 匈牙利匹配与我们方法的比较

表 6: 匈牙利算法和我们标签分配方案的比较。

点数	时间 (ms)		GPU (Mb)	
	匈牙利算法	我们的方法	匈牙利算法	我们的方法
100	0.52	0.12	39	14
1,000	78.34	0.13	81	14
10,000	24,216.35	1.25	2,304	15
100,000	-	28.85	-	39

Tab. 6 展示了在匹配点数相同的两个点云时的耗时和 GPU 使用情况。很明显，匈牙利算法存在可扩展性问题。例如，当点数为 10K 时，它消耗大约 24 秒和 2,304Mb 的 GPU 内存用于单次匹配。当扩展到 100K 点时，由于 CUDA 内存限制，即使在 80G A100 GPU 上也无法进行匹配。

相比之下，我们的标签分配方法实现了显著的效率，对于 10K 和 100K 点分别仅需要大约 1.25ms 和 28.85ms。此外，训练期间的 GPU 内存消耗可以忽略不计。这些发现揭示了我们标签分配方法的实用性和有效性，特别是在占用预测中，点数很容易超过 10K。

表 7: 不同模型的配置。

模型	Q	S	点数					
			s1	s2	s3	s4	s5	s6
OPUS-T	600	4	1	4	16	32	64	128
OPUS-S	1200	2	1	4	8	16	32	64
OPUS-M	2400	2	1	2	4	8	16	32
OPUS-L	4800	2	1	2	4	8	16	16

D.2 不同版本的详细配置。

在本节中，我们详细介绍了我们模型各种版本的设置，如 Tab. 7 所示，每个版本都针对性能和速度的不同方面进行了优化。我们最快的模型 OPUS-T 仅使用 0.6K 查询，每个查询在图像中采样 4 个点。在 6 个阶段中，预测的点数分别为 1、4、16、32、64 和 128。这种配置确保了快速的处理时间，同时保持了竞争力的性能。我们的其他模型版本，如 OPUS-S、OPUS-M、OPUS-L 在 CPS 模块中每个查询采样 2 个点，逐步加倍查询数量并相应调整预测点数以平衡速度和准确性。所有这些模型最终预测的点数相同。

表 8: 不同预测点数的性能。

模型	点数	mIoU	RayIoU_{1m}	RayIoU_{2m}	RayIoU_{4m}	RayIoU
OPUS-M	64	28.4	22.2	29.5	34.8	28.8
	32	27.2	26.1	33.3	38.4	32.6
	16	22.8	28.1	35.3	40.2	34.5
	8	16.4	27.4	34.6	39.6	33.9

D.3 最后一层不同精炼点数的影响。

Tab. 8 评估了在最后一层中改变预测点数的影响。我们使用 OPUS-M 作为该实验的模型。如表所示，随着点数从 8 增加到 64，mIoU 稳定上升，从 16.4 增加到 28.4。这种趋势是可以预期的，因为增加点数通常可以通过覆盖更多的体素来提高 mIoU，因为 mIoU 严重惩罚假阴性（FN）。然而，当模型预测 16 个点时，RayIoU 结果达到峰值，随着点数的进一步增加而下降。这种下降部分是由于在一定程度上增加更多点会引入噪声，从而对强调沿射线首次被占用的体素的 RayIoU 产生负面影响。

表 9: 不同距离范围的性能。

模型	总体	$0m \sim 20m$	$20m \sim 40m$	$> 40m$
FB-Occ	33.5	41.3	24.2	12.1
OPUS-L	41.2	49.10	31.15	13.73

D.4 不同距离范围的预测

我们在 Tab. 9 中报告了 FB-Occ 和 OPUS 在不同范围内的 RayIoU。很明显，OPUS 在近距离区域的优势更为明显，而在远距离区域的优势较小。这可能归因于 SparseOcc 指出的现象：密集方法倾向于高估表面，特别是在近距离区域。

E 附加定性分析

E.1 SparseOcc 与 OPUS 的差异

占用预测的视角。占用预测的基本差异在于视角。如主稿所述，所有先前的方法，包括 SparseOcc [19]，都将占用预测视为一个标准的分类任务。而 OPUS 率先提出了一种集合预测的视角，提供了一种新颖、优雅且端到端的稀疏化方法。

多阶段与端到端稀疏化过程。 SparseOcc 通过多个阶段逐渐丢弃体素来生成稀疏占用。在早期阶段丢弃空体素是不可逆的，导致明显的累积错误，如 Fig. 3 所示。相比之下，OPUS 通过直接预测一个稀疏集合来规避复杂的过滤机制，从而产生更连贯的结果。

详细模型设计。在结构的更详细视角下，还有许多差异，例如：

- **查询数量**。在 NuScene-Occ3D 中，SparseOcc 在其最终阶段需要 32K 查询。OPUS 则仅使用 0.6K-4.8K 查询进行占用预测，利用其灵活的特性并有助于快速推理。

- **粗到细过程**。SparseOcc 的粗到细策略涉及逐步过滤空体素并将占用体素细分。相比之下，OPUS 将粗到细解释为跨阶段预测点数的增加。

- **学习目标**。我们的学习目标同时预测语义类别和占用位置。后者是 OPUS 引入的新目标，通过修改后的查姆弗距离损失实现。

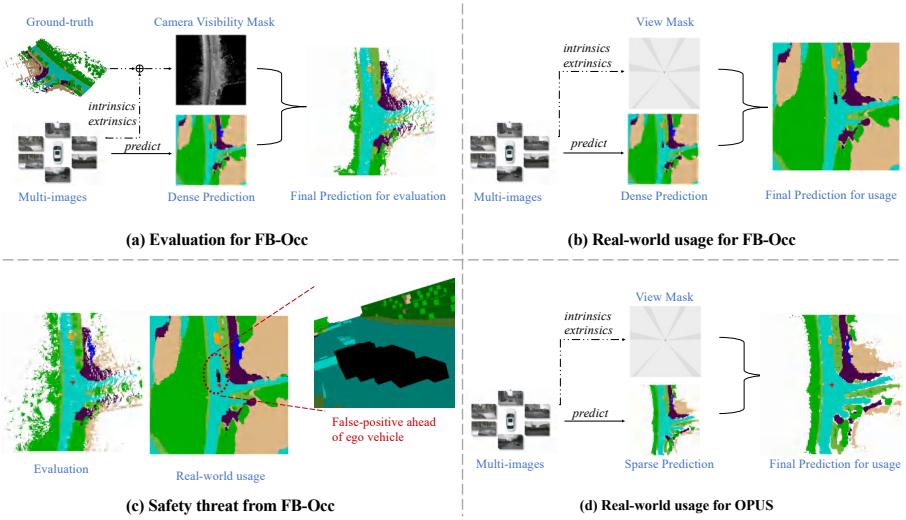

图 8: 由于评估指标与实际场景之间的差异所导致的安全威胁示例。(a) 在评估之前，首先根据相机内参和外参生成相机可见性掩膜。然后，将密集预测掩膜化以获得最终评估预测。(b) 在实际应用中，由于不知道真实占用情况，无法进行相机可见性推理。我们只能根据相机内参和外参生成视图掩膜，无法过滤密集模型中的过度估计体素。(c) 在自车附近出现了大量误报，用红星标记。这些错误预测的体素在评估 mIoU 时会被过滤，但在实际应用中可能会导致严重的安全问题。(d) OPUS 生成稀疏占用预测，受到过度估计的影响较小。因此，在这种情况下不会出现此类安全威胁。**最佳以彩色查看**。

E.2 mIoU、RayIoU 与驾驶安全之间的关系分析

我们的 OPUS-L (8f) 实现了 41.17 的最先进的 RayIoU，超过了之前的稀疏模型 SparseOcc 6.07 和密集模型 FB-Occ 7.7。稀疏和密集方法之间的 mIoU 差距也从 SparseOcc 的 8.5 缩小到 OPUS 的 3.0。然而，这一差距对安全的影响仍不明确。在自动驾驶的背景下，这个问题尤为重要，我们希望澄清如下：

密集预测的风险。密集预测的最大问题是评估指标与实际场景之间的差异。如 Fig. 8 所示，评估指标仅考虑相机掩膜内的体素，该掩膜来自相机参数和真实数据。然而，在实际应用中，我们只能基于相机内参和外参生成视图掩膜，无法过滤过度估计的体素。从 Fig. 8 和 Fig. 3 可以看出，密集方法甚至可能在自车附近误识别占用体素。这些错误在评估中被忽略，但在实际场景中可能构成重大安全隐患。相比之下，OPUS 受此问题的影响较小，因为它不会过度估计占用情况。

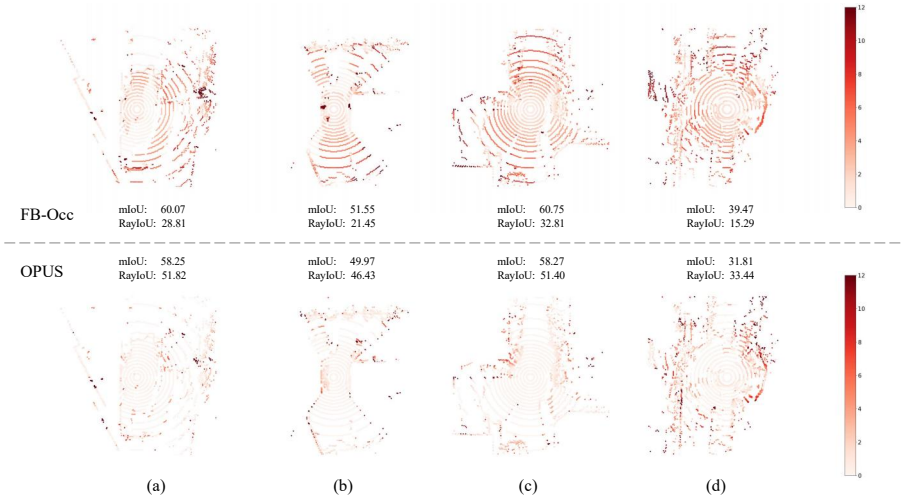

图 9: FB-Occ 和 OPUS 的预测误差图。与 FB-Occ 相比，OPUS 的 mIoU 较低，但 RayIoU 结果较高，误差明显更小。**最佳以彩色查看**。

OPUS 的深度误差远小于 FB-Occ。在 Fig. 9 中，我们比较了沿相机光线的 FB-Occ 和 OPUS 的深度误差。尽管 OPUS 的 mIoU 性能相对较低，但它在所有场景中显示出较低的深度误差。鉴于首次占用体素对安全的重要性，OPUS 在此方面的精确性提高了安全性，而不是降低安全性。

总之，尽管有必要尽量缩小稀疏和密集方法之间的 mIoU 差距，但我们的分析表明，mIoU 可能无法完全代表潜在的危险情况。因此，在占用任务中考虑 mIoU 和 RayIoU 是更合理的。

NeurIPS Paper Checklist

1. **Claims**

   Question: 摘要和引言中提出的主要观点是否准确反映了论文的贡献和范围?

   Answer: [Yes]

   Justification: 我们在摘要和引言中明确描述了我们的贡献和范围。

   Guidelines:

   - 回答 NA 表示摘要和引言中不包括论文中的观点。
   - 摘要和/或引言应清楚地阐述所提出的观点，包括论文的贡献以及重要的假设和限制。对这个问题回答 No 或 NA 将不会被审稿人看好。
   - 所提出的观点应与理论和实验结果相匹配，并反映结果在其他设置中的可推广性。
   - 涵盖论文未达到的目标是可以的，只要明确这些目标是作为动机而存在的。

2. **Limitations**

   Question: 论文是否讨论了作者完成的工作的局限性?

   Answer: [Yes]

   Justification: 我们在主文本的“结论和局限性”部分描述了我们的局限性。

   Guidelines:

   - 回答 NA 表示论文没有局限性，而回答 No 表示论文存在局限性，但未在论文中讨论。
   - 鼓励作者在论文中创建一个单独的“局限性”部分。
   - 论文应指出任何强假设以及结果对这些假设的违反有多鲁棒（例如，独立性假设、无噪声设置、模型的良好规范性、仅在局部成立的渐近近似）。作者应反思这些假设在实践中可能被违反的方式及其影响。
   - 作者应反思所提出观点的范围，例如，如果该方法仅在少数数据集或少数运行中进行了测试。一般来说，实验结果通常依赖于隐含的假设，这些假设应被明确说明。
   - 作者应反思影响该方法性能的因素。例如，面部识别算法在图像分辨率低或在低光照条件下拍摄的图像中可能表现不佳。或者，语音转文本系统可能无法可靠地用于在线讲座的字幕，因为它无法处理技术术语。

- 作者应讨论所提出算法的计算效率以及它们如何随数据集规模扩展。
- 如果适用，作者应讨论其方法在解决隐私和公平性问题方面的可能局限性。
- 虽然作者可能担心完全诚实地说明局限性可能会被审稿人用作拒绝的理由，但更糟糕的结果可能是审稿人发现论文未承认的局限性。作者应根据自己的最佳判断行事，并认识到个别行动在促进透明度方面的重要作用，以维护社区的诚信。审稿人将被特别指示不要因诚实地说明局限性而进行处罚。

3. **Theory Assumptions and Proofs**

   Question: 对于每个理论结果，论文是否提供了完整的假设集和完整（且正确）的证明?

   Answer: [NA]

   Justification: 本论文不包含理论结果。

   Guidelines:

   - 回答 NA 表示论文不包含理论结果。
   - 论文中的所有定理、公式和证明都应编号并交叉引用。
   - 所有假设都应在任何定理的陈述中明确说明或引用。
   - 证明可以出现在主论文或补充材料中，但如果它们出现在补充材料中，作者被鼓励提供一个简短的证明草图以提供直觉。
   - 反之，如果论文的核心提供了非正式证明，则应在附录或补充材料中提供正式证明。
   - 证明所依赖的定理和引理应被正确引用。

4. **Experimental Result Reproducibility**

   Question: 论文是否完全披露了所有需要重现论文主要实验结果的信息（无论是否提供了代码和数据），以便影响论文的主要观点和/或结论?

   Answer: [Yes]

   Justification: 我们在“实验”部分提供了实现细节。此外，我们的代码和配置直接附加在补充材料中，以便直接重现我们的结果。

   Guidelines:

   - 回答 NA 表示论文不包含实验。

- 如果论文包含实验，对该问题的回答为 No 将不会被审稿人看好：无论是否提供了代码和数据，使论文可重现都是重要的。
- 如果贡献是数据集和/或模型，作者应描述他们采取了哪些步骤来使他们的结果可重现或可验证。
- 根据贡献的不同，可重现性可以通过多种方式实现。例如，如果贡献是一个新的架构，充分描述架构可能就足够了，或者如果贡献是一个特定的模型和实证评估，可能需要使其他人能够使用相同的数据集复制模型，或者提供访问模型的方式。一般来说。发布代码和数据通常是实现这一目标的一种好方法，但可重现性也可以通过详细说明如何复制结果、访问托管模型（例如，在大型语言模型的情况下）、发布模型检查点或其他适当的方式提供。
- 虽然 NeurIPS 不要求发布代码，但会议确实要求所有提交都提供某种合理的途径来实现可重现性，这可能取决于贡献的性质。例如
  (a) 如果贡献主要是新的算法，论文应清楚说明如何重现该算法。
  (b) 如果贡献主要是新的模型架构，论文应清晰且完整地描述架构。
  (c) 如果贡献是一个新的模型（例如，大型语言模型），则应有一种方式可以访问该模型以重现结果，或者有一种方式可以重现该模型（例如，使用开源数据集或如何构建数据集的说明）。
  (d) 我们认识到在某些情况下可重现性可能很棘手，在这种情况下，作者被欢迎描述他们提供的特定的可重现性方式。在封闭源模型的情况下，可能模型的访问受到某种方式的限制（例如，仅限注册用户），但其他研究人员应该有一些途径来重现或验证结果。

5. **Open access to data and code**

   Question: 论文是否提供对数据和代码的开放访问，并提供足够的说明以忠实地重现论文中描述的主要实验结果，如补充材料所述?

   Answer: [Yes]

   Justification: 是的，我们在补充材料中提供了代码和配置。在准备所需的 Occ3D 数据集后，可以轻松重现我们的结果。

   Guidelines:

   - 回答 NA 表示论文不包含需要代码的实验。
   - 请参阅 NeurIPS 代码和数据提交指南（https://nips.cc/public/guides/CodeSubmissionPolicy）以获取更多详细信息。

- 虽然我们鼓励发布代码和数据，但我们理解这可能不可能，因此 “否” 是一个可接受的回答。论文不能仅因未包含代码而被拒绝，除非这是贡献的核心（例如，对于新的开源基准）。
- 说明应包含运行以重现结果的确切命令和环境。请参阅 NeurIPS 代码和数据提交指南（https://nips.cc/public/guides/CodeSubmissionPolicy）以获取更多详细信息。
- 作者应提供数据访问和准备的说明，包括如何访问原始数据、预处理数据、中间数据和生成的数据等。
- 作者应提供脚本以重现新提出方法和基线的所有实验结果。如果只有部分实验可以重现，他们应说明哪些实验未包含在脚本中以及原因。
- 在提交时，为了保持匿名性，作者应发布匿名版本（如果适用）。
- 建议在补充材料（附加到论文中）中提供尽可能多的信息，但也可以包含指向数据和代码的 URL。

6. **Experimental Setting/Details**

   Question: 论文是否指定了所有必要的训练和测试细节（例如，数据分割、超参数、如何选择它们、优化器类型等），以便理解结果?

   Answer: [Yes]

   Justification: 是的，我们的训练/测试细节在 “实验设置” 部分中进行了详细说明。

   Guidelines:

   - 回答 NA 表示论文不包含实验。
   - 实验设置应在论文的核心部分呈现，详细程度足以理解结果并使其有意义。
   - 可以在代码、附录或补充材料中提供完整的详细信息。

7. **Experiment Statistical Significance**

   Question: 论文是否适当地报告了误差条并正确定义了其他适当的信息以说明实验的统计显著性?

   Answer: [No]

   Justification: 在本论文中，未报告误差条，因为这将过于计算昂贵。

   Guidelines:

   - 回答 NA 表示论文不包含实验。

- 如果结果附有误差条、置信区间或统计显著性检验，则作者应回答“Yes”，至少对于支持论文主要观点的实验。
- 应清楚说明误差条所捕捉的变异性因素（例如，训练/测试分割、初始化、某些参数的随机抽取，或在给定实验条件下进行的整体运行）。
- 应解释计算误差条的方法（闭式公式、调用库函数、自助法等）
- 应给出所做的假设（例如，误差呈正态分布）。
- 应明确误差条是标准差还是均值的标准误差。
- 报告 1误差条是可以的，但作者应说明这一点。如果未验证误差的正态分布假设，作者应报告 2误差条并说明其具有 96
- 对于非对称分布，作者应小心不要在表格或图表中显示对称的误差条，这可能导致结果超出范围（例如，负误差率）。
- 如果在表格或图表中报告了误差条，作者应在文本中解释它们是如何计算的，并在文本中引用相应的图表或表格。

8. **Experiments Compute Resources**

   Question: 对于每个实验，论文是否提供了足够的信息来说明重现实验所需的计算资源（计算工作类型、内存、执行时间）?

   Answer: [Yes]

   Justification: 计算资源在“实验设置”部分中进行了详细说明。

   Guidelines:

   - 回答 NA 表示论文不包含实验。
   - 论文应指出使用的计算工作类型（CPU 或 GPU、内部集群或云提供商，包括相关的内存和存储）。
   - 论文应提供每个单独实验运行所需的计算量以及总的计算量估计。
   - 论文应披露整个研究项目是否需要比论文中报告的实验更多的计算资源（例如，初步或失败的实验未被纳入论文）。

9. **Code Of Ethics**

   Question: 论文中的研究是否在各个方面都符合 NeurIPS 伦理准则 https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

Justification: 我们的工作遵守 NeurIPS 伦理准则。

Guidelines:

- 回答 NA 表示作者未审查 NeurIPS 伦理准则。
- 如果作者回答 No，他们应解释需要偏离伦理准则的特殊情况。
- 作者应确保保持匿名性（例如，如果由于其司法管辖区的法律或法规有特殊考虑）。

10. **Broader Impacts**

Question: 论文是否讨论了所完成工作的潜在积极社会影响和负面社会影响?

Answer: [Yes]

Justification: 我们在附录的“广泛影响”部分讨论了我们工作的社会影响。

Guidelines:

- 回答 NA 表示所完成的工作没有任何社会影响。
- 如果作者回答 NA 或 No，他们应解释为什么他们的工作没有社会影响或为什么论文未涉及社会影响。
- 负面社会影响的例子包括潜在的恶意或意外使用（例如，虚假信息、生成假档案、监控）、公平性考虑（例如，部署可能对特定群体做出不公平决策的技术）、隐私考虑和安全考虑。
- 会议期望许多论文将是基础研究，并未与特定应用或部署相关联。然而，如果存在直接的负面应用路径，作者应指出这一点。例如，指出生成模型质量的提高可能被用于生成虚假信息的深度伪造是合理的。另一方面，指出用于优化神经网络的通用算法可能使人们更快地训练生成深度伪造的模型是不需要的。
- 作者应考虑当技术按预期使用且正常运行时可能出现的危害，当技术按预期使用但给出错误结果时可能出现的危害，以及由于技术的（故意或无意）误用而产生的危害。
- 如果存在负面社会影响，作者还可以讨论可能的缓解策略（例如，模型的门控发布、除了攻击之外还提供防御、监控误用的机制、监控系统如何随时间从反馈中学习、提高机器学习的效率和可访问性）。

11. **Safeguards**

Question: 论文是否描述了为负责任地发布具有高误用风险的数据或模型（例如预训练语言模型、图像生成器或抓取的数据集）所采取的保障措施?

Answer: [NA]

Justification: 所涉及的数据/模型不存在高误用风险。

Guidelines:

- 回答 NA 表示论文不存在此类风险。
- 具有高误用风险或双重用途的模型应与必要的保障措施一起发布，以允许模型的受控使用，例如要求用户遵守使用指南或限制以访问模型或实施安全过滤器。
- 从互联网抓取的数据集可能存在安全风险。作者应描述他们如何避免发布不安全的图像。
- 我们认识到提供有效的保障措施是具有挑战性的，许多论文不需要此部分，但我们鼓励作者考虑这一点并尽最大努力。

12. **Licenses for existing assets**

Question: 论文中使用资产（例如代码、数据、模型）的创作者或原始所有者是否得到了适当的署名，并且明确提及了许可证和使用条款并正确遵守?

Answer: [Yes]

Justification: 我们在附录的 "所涉及资产的许可证" 部分详细说明了许可证，并在论文中引用了相关论文。

Guidelines:

- 回答 NA 表示论文未使用现有资产。
- 作者应引用生成代码包或数据集的原始论文。
- 作者应说明所使用的资产版本，并尽可能包含 URL。
- 每个资产的许可证名称（例如 CC-BY 4.0）应包含在内。
- 对于从特定来源（例如网站）抓取的数据，应提供该来源的版权和服务条款。
- 如果发布资产，包中应包含许可证、版权声明和使用条款。对于流行数据集，paperswithcode.com/datasets 已经整理了一些数据集的许可证。他们的许可证指南可以帮助确定数据集的许可证。
- 对于重新打包的现有数据集，应同时提供原始许可证和衍生资产的许可证（如果已更改）。
- 如果此信息无法在线获得，作者被鼓励联系资产的创作者。

13. **New Assets**

    Question: 论文中引入的新资产是否得到了充分的记录，并且文档与资产一起提供?

    Answer: [Yes]

    Justification: 我们的代码中的 readme 文件提供了关于训练、许可证和限制的详细信息。

    Guidelines:

    - 回答 NA 表示论文未发布新资产。
    - 研究人员应通过结构化模板作为提交的一部分传达数据集/代码/模型的详细信息。这包括关于训练、许可证、限制等的详细信息。
    - 论文应讨论是否以及如何从使用资产的人那里获得同意。
    - 在提交时，记得匿名化你的资产（如果适用）。你可以创建一个匿名化 URL 或包含一个匿名化 zip 文件。

14. **Crowdsourcing and Research with Human Subjects**

    Question: 对于众包实验和人类受试者研究，论文是否包括参与者收到的完整指令文本以及适用的截图，以及有关补偿（如果有）的详细信息?

    Answer: [NA]

    Justification: 该论文不涉及众包或人类受试者研究。

    Guidelines:

    - 回答 NA 表示论文不涉及众包或人类受试者研究。
    - 将此信息包含在补充材料中是可以的，但如果论文的主要贡献涉及人类受试者，则应在主论文中包含尽可能多的详细信息。
    - 根据 NeurIPS 伦理准则，参与数据收集、策划或其他劳动的工人应至少获得数据收集地的最低工资。

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: 论文是否描述了研究参与者可能面临的风险，这些风险是否已向受试者披露，以及是否获得了机构审查委员会 (IRB) 批准（或基于你所在国家或机构的要求的等效批准/审查)?

    Answer: [NA]

Justification: 本论文不涉及众包或人类受试者研究。

Guidelines:

- 回答 NA 表示论文不涉及众包或人类受试者研究。
- 根据研究进行的国家，机构审查委员会批准（或等效）可能对任何人类受试者研究是必需的。如果你获得了机构审查委员会批准，你应清楚地在论文中说明。
- 我们认识到不同机构和地点的程序可能差异很大，我们期望作者遵守 NeurIPS 伦理准则和其机构的指导方针。
- 对于初始提交，不要包含任何可能破坏匿名性的信息（如果适用），例如进行审查的机构。