arXiv:2301.02650v2 [cs.CV] 8 May 2024

用于室内三维目标检测的分层点注意力

Manli Shu[1, 2, *] Le Xue[2] Ning Yu[2] Roberto Martín-Martín[2,3]
Caiming Xiong[2] Tom Goldstein[1] Juan CarlosNiebles[2,4] 和 Ran Xu[2]

摘要—三维目标检测是多种机器人系统（如增强现实和家用机器人）所必需的视觉技术。作为通用网络架构，Transformer 近年来在 3D 点云目标检测领域取得了巨大成功。然而，普通 Transformer 缺乏层级结构，限制了其学习不同尺度特征的能力。这一局限使 Transformer 检测器在小物体上的表现较差，也影响了它们在小物体占多数的室内环境中的可靠性。本文为基于点云的 Transformer 检测器提出了两种新颖的注意力操作，作为通用的层次化设计。首先，我们提出聚合多尺度注意力（MS-A），从单尺度输入特征构建多尺度词元，以实现更细粒度的特征学习。其次，我们提出尺寸自适应局部注意力（Local-A），通过自适应注意力区域，在边界框提议内进行局部特征聚合。这两种注意力操作都是与模型无关的网络模块，可接入现有点云 Transformer，实现端到端训练。我们在两个广泛使用的室内检测基准上评估了所提出的方法。将我们提出的模块接入最先进的基于 Transformer的三维检测器后，我们在两个基准上都刷新了此前的最佳结果，并且在小物体上的提升更为显著。

一、引言

三维计算机视觉模型（例如目标检测器）帮助机器人和控制系统从三维数据（例如点云）中感知并理解环境；三维数据能够提供更准确的几何和空间信息，并且对光照变化和域偏移具有鲁棒性。由于点云不像图像那样具有网格结构，先前的研究提出了各种用于点云理解的神经网络架构 [1]–[13]。随着基于注意力的架构（即 Transformer）在其他学习领域 [14]–[16]，取得成功，近年来，Transformer 已被应用于点云 [17]–[23]。Transformer 的一些特性使其非常适合对点云进行建模。例如，其置换不变性是对点云这类无序集合进行建模所必需的，而其注意力机制有助于学习长程关系并捕获全局上下文。

尽管 Transformer 在处理点云方面具有优势，我们发现，最先进的 Transformer 检测器在不同物体尺寸上的表现并不均衡，对较小物体的平均精度较低（见第 IV-B 节）。这种表现失衡会影响下游应用的鲁棒性，尤其是在环境中杂乱分布着许多小物体的室内场景中 [24], [25]。我们推测，对小物体的这种偏差可归因于两个因素。首先，出于效率考虑，现有模型是在经过下采样的点云上训练的，其点数远少于原始数据。大幅下采样会丢失几何细节，对较小物体的影响更为显著。其次，普通Transformer [14], [15] 只能学习全局尺度的特征，而较小物体可能需要更精细的特征提取。

受这些观察结果的启发，我们认为点云Transformer能够从层次化特征学习策略中受益 [26], [27], 例如多尺度和局部化特征学习。然而，考虑到点云Transformer的计算量，使用更高分辨率（即更高点密度）的点云特征可能效率低下。此外，由于点云具有不规则性，将层次化设计融入用于基于点的三维目标检测的Transformer并非易事。

我们的方法。我们提出了两种基于点的注意力模块，用于点云的层次化特征学习。这两种模块均与模型无关，可插入任何现有的基于点的Transformer中进行端到端训练。

我们首先提出聚合多尺度注意力（MS-A），它从单尺度输入构建更高分辨率的特征。随后，它通过多尺度令牌聚合 [28] 利用多尺度特征进行交叉注意力，且几乎不增加参数开销。第二个提出的模块是尺寸自适应局部注意力（Local-A），其注意力区域由检测器针对每个物体候选生成的候选框定义。因此，它能够实现自适应的局部特征学习。

我们在两个广泛使用的室内 3D 检测基准 ScanNetV 2 [24]和 SUN RGB-D [25]上评估了我们的方法。我们将注意力模块接入现有的基于 Transformer 的三维检测器，并进行端到端训练。我们的方法以极少的参数开销刷新了此前的最佳结果，并且对较小物体的提升更为显著。我们的主要贡献总结如下：

- 我们发现，基于点的三维 Transformer 检测器存在性能不均衡问题，对较小物体的检测效果较差。
- 受这一发现启发，我们提出了两种通用的点注意力操作，以实现多尺度和局部化特征学习。
- 我们将该方法应用于多种基于点的 Transformer 检测器，并在两个

[1] 马里兰大学大学公园市校区计算机科学系，美国马里兰州大学公园市 {manlis, tomg}@umd.edu[2] Salesforce Research，美国加利福尼亚州帕洛阿尔托{manli.shu, lxue, ning.yu, cxiong, ran.xu}@salesforce.com[3] 德克萨斯大学奥斯汀分校计算机科学系，美国德克萨斯州robertomm@cs.utexas.edu[4] 斯坦福大学计算机科学系，美国加利福尼亚州jniebles@cs.stanford.edu

* 在马里兰大学工作期间完成。

广泛使用的室内 3D 检测基准。

II. 相关工作

用于 3D 点云的深度神经网络。现有的点云学习网络架构，根据其点云表示方式，大致可分为两类：基于网格的 和 基于点的；此外，还有一些混合架构同时采用这两种表示方式 [29]–[33]。基于网格的 方法将不规则点云投影到网格结构中，例如 3D 体素或柱体 [34]–[38]。另一方面，基于点的 方法直接从原始点云中学习特征。在这一类方法中，基于图的方法 [39]–[42]利用图来建模点与点之间的关系。其他研究则将点云视为一个集合，并通过集合抽象来学习特征 [4], [6], [43],[44]。近期研究探索了将 Transformer 架构用于基于点的学习 [12], [13], [17]–[19], [45] ：将单个点作为词元输入 Transformer，由注意力机制在全局范围内学习点特征。以往大多数方法通过开发新的骨干网络来改进点云学习，而我们的工作旨在提供一种通用的模型无关解决方案。PAConv [46] 提出了一种通用的卷积操作，而我们研究的是可集成到新兴 Transformer 模型中的交叉注意力操作。

面向 3D 视觉 Transformer 的层次化设计。受 2D 视觉 [16], [28], [47], [48] 尤其是卷积神经网络 [49] 领域相关研究的启发，面向 3D Transformer 的层次化设计同样致力于学习不同粒度的特征。一些方法 [19], [50]提出了能够进行多尺度特征学习的注意力机制，但这些方法只能作用于体素，无法应用于基于点的 3D 模型。Lai 等人 [51] 提出了分层自注意力，可直接在点云上学习多尺度特征，但其仅支持下采样，从而限制了更细粒度的特征学习。相比之下，我们的多尺度注意力能够生成任意分辨率的特征。另一类工作 [12], [17], [18] 通过对由 k 最近邻或固定半径/窗口指定的局部区域应用自注意力来学习局部特征。不同于固定区域，我们的局部注意力能够自适应尺寸，使用与近似物体尺寸相匹配的自适应局部区域。

III. 方法

A. 背景

点云目标检测包含 P 个点的点云 $\mathcal{P}_{\text{raw}}$ $\mathcal{P} = \{\mathbf{p_i}\}_{i=1}^{P}$，每个点 $p_i \in \mathrm{R}^3$均由其三维坐标表示。点云上的三维目标检测旨在预测场景中物体的一组边界框，包括其位置（作为边界框的中心）、边界框的尺寸和方向，以及相应物体的语义类别。需要注意的是，由于计算限制，模型在早期阶段会将点云下采样为$\mathcal{P}_{\text{raw}}$的子集，其中包含 N （$N << P$）个点。 $\mathcal{P} = \text{SA}(\mathcal{P}_{\text{raw}}) = \{\mathbf{p_i}\}_{i=1}^{N}$ 包含围绕 N聚合的点组

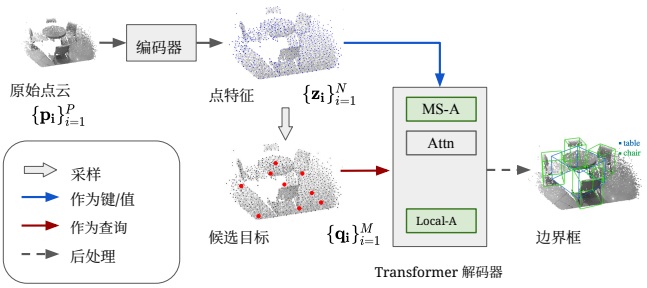


图 1. **采用我们提出的模块（MS-A 和 Local-A）的基于点的三维 Transformer 检测器**。检测器概览：原始点云在编码过程中经过下采样，以获得点特征，这些特征在 Transformer 解码器中充当键和值。物体候选从点特征中采样得到（例如使用最远点采样）。Transformer 解码器通过交替进行自注意力机制和交叉注意力来学习对象特征。所提出的 MS-A 和 Local-A 是可插入 Transformer 的交叉注意力模块。有关各模块的设计细节，请参见图 2 和图 3。

组中心，其中 SA（集合抽象）是聚合函数，而组中心则使用最远点采样（*FPS*） [4], 一种能够良好覆盖整个点云的随机采样算法从原始点云中采样得到。

基于点的三维 Transformer 检测器。 我们的方法建立在基于点的三维目标检测器之上 [7], [13], [52],, 以自底向上的方式在点云中检测三维物体。与其他在鸟瞰图或体素化点云上以自顶向下方式生成候选框的三维检测器相比 [53], [54], , 基于点的方法直接作用于不规则点云，不会产生量化误差。

我们在图 1 中展示了通用的基于点的 3D Transformer 检测器。输入点云的特征$\{\mathbf{z_i}\}_{i=1}^{N}$，$\mathbf{z_i} \in \mathrm{R}^d$ 通过骨干模型（例如，PointNet++ [6]）提取，其中 d 为特征维度。基于点的检测器以 $M(M < N)$ 个初始物体候选 $\{\mathbf{q_i}\}_{i=1}^{M}$，$\mathbf{q_i} \in \mathrm{R}^C$作为起点生成边界框预测，这些候选目标从点云中采样得到，并作为物体中心。一种常用的候选目标采样方法是最远点采样（FPS）。获得初始候选目标后，检测器会为每个物体候选提取特征。基于注意力机制的方法 [13] 通过在物体候选之间进行自注意力，以及在候选目标（即查询）与点特征 $\{\mathbf{z_i}\}_{i=1}^{N}$之间进行交叉注意力来学习特征。

物体候选的学习特征随后会传递给预测头，由其预测每个物体候选的边界框属性。3D 边界框的属性包括其位置（框中心）$\hat{\mathbf{c}} \in \mathrm{R}^3$, 尺寸 $\hat{\mathbf{d}} \in \mathrm{R}^3$, 方向（航向角）$\hat{\mathbf{a}} \in \mathrm{R}$，以及物体的语义标签 $\hat{\mathbf{s}}$。采用这些参数化表示，我们可以将边界框提议表示为 $\hat{\mathbf{b}} = \{\hat{\mathbf{c}}, \hat{\mathbf{d}}, \hat{\mathbf{a}}, \hat{\mathbf{s}}\}$。**注意力机制** 是 Transformer 的基本构建模块。注意力函数以查询（Q）、键（K）和值（V）为输入。注意力函数的输出是值的加权和，其中注意力权重为键与查询之间的缩放点积：

$$\text{Attn}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_h}})V, \tag{1}$$

其中， d_h 是注意力层的隐藏维度。对于自注意力， $Q \in \mathrm{R}^{d_h}$、 $K \in \mathrm{R}^{d_h}$ 和 $V \in \mathrm{R}^{d_v}$ 分别通过参数矩阵 $W_i^Q \in \mathrm{R}^{d \times d_h}$、 $W_i^K \in \mathrm{R}^{d \times d_h}$和$W_i^V \in \mathrm{R}^{d \times d_v}$ 进行线性投影，从输入 $X \in \mathrm{R}^d$变换而来。对于交叉注意力， Q、 K和 V可以来自不同的来源。

实际上，Transformer 采用**多头注意力**设计，即在不同的注意力头上并行应用多个注意力函数。每个注意力头的输入都是该层输入的一部分。具体而言，查询、键和值沿隐藏维度拆分为 $(Q_i, K_i, V_i)_{i=1}^h$，其中 $Q_i \in \mathrm{R}^{d_h/h}, K_i \in \mathrm{R}^{d_h/h}, V_i \in \mathrm{R}^{d_v/h}$， h 为注意力头的数量。多头注意力层的最终输出是所有注意力头输出拼接后的投影：

$$\begin{aligned}\mathrm{MultiHead}(Q,K,V) = \mathrm{Concat}(\{\mathrm{Attn}(Q_0,K_0,V_0);\\ ...;\mathrm{Attn}(Q_{h-1},K_{h-1},V_{h-1})\})W^O,\end{aligned} \tag{2}$$

其中，第一项表示输出的拼接， W^O 是输出投影矩阵。

B. 聚合多尺度注意力

在现有的基于点的 Transformer 检测器中，交叉注意力模块应用于物体候选与点云中的所有其他点之间。然而，由于注意力函数的计算开销，模型实际使用的点云是由 1024 个点组成的下采样集合 [12], [13], ，而原始点云通常包含数万个点 [24], [25]。如此大规模的下采样会导致详细几何信息和细粒度特征的损失，而这些信息和特征对于目标检测等密集预测任务至关重要。

为此，我们提出聚合多尺度注意力（MS-A），通过学习从单尺度特征输入构建任意更高分辨率（即，更高点密度）的特征图。随后，它利用两种特征的特征，作为物体候选与其他点之间交叉注意力中的多尺度键/值。多尺度特征聚合通过多头令牌聚合实现，其中我们在不同的注意力头子集中使用不同尺度的键和值。我们的目标是为物体级特征学习提供细粒度的几何细节。

MS-A 的第一步是从单尺度输入构建更高分辨率的特征图。我们提出了一种可学习上采样过程。给定层的输入点云特征 $\{\mathbf{z_i}\}_{i=1}^N, \mathbf{z_i} \in \mathrm{R}^d$，MS-A 可以创建一个特征图

点的数量（N'）可以任意。为获取 N'个点的位置（即坐标），我们使用最远点采样从原始点云 中采样 N'个点，并获得 $\{\mathbf{p_i}\}_{i=1}^{N'}, \mathbf{p_i} \in \mathrm{R}^3$。接下来，对于 N'点云中的每个采样点$\mathbf{p_i}$，我们通过三近邻加权插值来初始化其特征 [6], ，这是一种许多基于点的方法所采用的通用操作 [17], [46], [52]。与以往工作使用插值在编码器和解码器之间建立跳跃连接不同，我们的 N' 点是从原始点云中任意采样的。我们可以采样一个 N'点特征

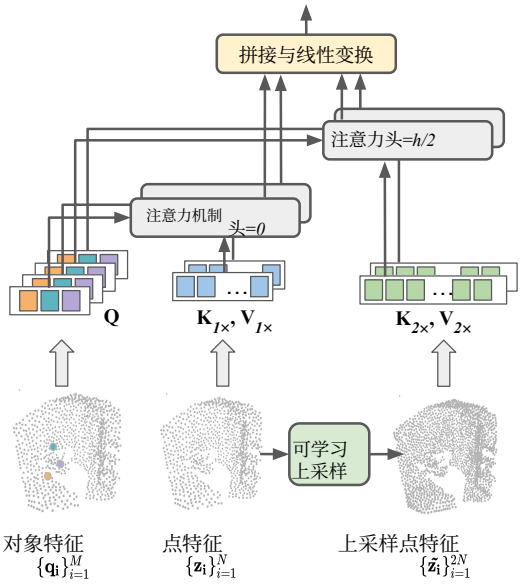


图 2. **聚合多尺度注意力 (MS-A)** 在多头交叉注意力设计中学习不同尺度的特征。它根据单尺度输入点特征构建更高分辨率（即更高点密度）的点特征，并使用两个尺度的键和值。

使其点数多于模型中的任何特征图，从而提供更细粒度的几何细节。

通过插值特征初始化，我们利用可学习模块 Φ_θ，获得 N' 点特征图的最终特征表示。我们使用多层感知机作为可学习投影函数。学习得到的 N' 点特征图为：

$$\{\tilde{\mathbf{z}}_\mathbf{i}\}_{i=1}^{N'}, \tilde{\mathbf{z}}_\mathbf{i} = \Phi_\theta(\mathrm{interpolate}(\{z_i^0, z_i^1, z_i^2\})) \tag{3}$$

上采样后，我们得到两组尺度不同的点特征 $\{\mathbf{z_i}\}_{i=1}^N$，$\{\tilde{\mathbf{z}}_\mathbf{i}\}_{i=1}^{N'}$。为避免增加计算量，我们对两组点特征同时进行多头交叉注意力计算。不同注意力头使用不同尺度的特征。我们将注意力头平均分为两组，第一组使用 $\{\mathbf{z_i}\}_{i=1}^N$ 得到 K 和 V，另一组则使用 $\{\mathbf{z_i}\}_{i=1}^{N'}$。两组共享同一组由物体候选 $\{\mathbf{q_i}\}_{i=1}^M$ 得到的查询。由于该模块的输入和输出与普通注意力模块相同，因此可以将 MS-A 插入任意基于注意力的模型，以实现不同尺度的特征学习。实践中，我们设置 $N' = 2N$，并仅在 Transformer 解码器的第一层应用 MS-A，以尽可能降低额外计算开销。有关设计选择的进一步分析见第IV-D节。

C. 尺寸自适应局部注意力

尽管注意力机制能够有效地建模点之间的长程关系，但无法保证所学习的模块会对特定物体的重要点（例如，属于该物体的点）给予更多关注。另一方面，普通 Transformer缺乏层级结构，因此无法支持显式的局部特征提取。尽管一些工作通过限制

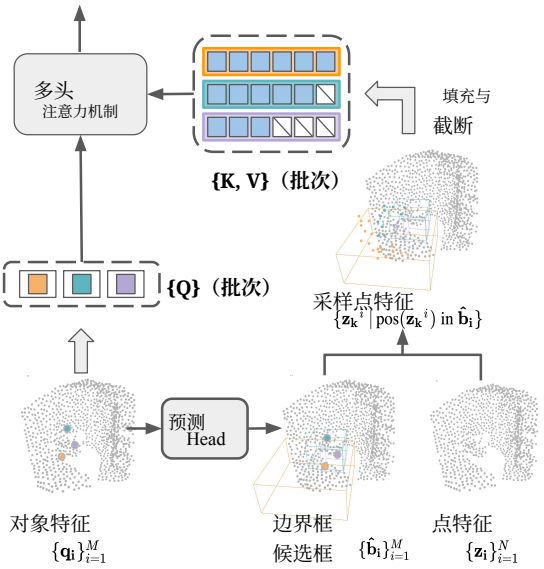


图 3. **尺寸自适应局部注意力（Local-A）**在每个物体候选（查询）与其对应边界框提议内的点之间执行自适应局部注意力。不同物体候选的注意力范围（键和值的 token 长度）各不相同，因此需要进行填充或截断，以支持批处理。

注意力仅作用于固定的局部区域，[12], [19], 我们提出了尺寸自适应局部注意力（Local-A），根据每个边界框提议的尺寸自适应地定义局部区域。

基于点的三维检测器可以根据解码器层的对象特征（$\{\mathbf{q_i}\}_{i=1}^{M}$）生成中间边界框提议 $\{\hat{}\mathbf{b_i}\}_{i=1}^{M}$。给定为每个物体候选 $\mathbf{q_i}$生成的边界框，我们在$\hat{}\mathbf{q_i}$ 与从其对应边界框$\mathbf{b_i}$内采样的点之间执行交叉注意力。因此，每个查询点都拥有定制的尺寸自适应注意力区域。对于解码器层 l 中的每个输入物体候选 $\mathbf{q_i}^l \in \mathrm{R}^d$，通过 Local-A 对其进行更新（$i$ 表示查询点的索引）：

$$\mathbf{q_i}^{l+1} = \mathrm{Attn}(Q_i^l, K_i, V_i), \text{where} \tag{4}$$

$$Q_i^l = \mathbf{q_i}^l W^Q, K_i = Z_i W^K, V_i = Z_i W^V \text{with} \tag{5}$$

$$Z_i = \{\mathbf{z_j}^i \,|\, \mathrm{pos}(\mathbf{z_j}^i) \text{ in } \hat{\mathbf{b}}_\mathbf{i}\}, \hat{\mathbf{b}}_\mathbf{i} = \mathrm{Pred}_{box}^l(\mathbf{q_i}^l). \tag{6}$$

在公式（6）中，pos(·)获取点特征$\mathbf{z}$的三维坐标（x, y, z），而$\mathbf{b_i}$ 是模型对查询点$\mathbf{q_i}^l$ 的边界框预测（由预测头 Pred_{box}^l生成）。$\hat{}Z_i$ 表示边界框$\mathbf{b_i}$内的一组点。 K_i 和 V_i由这组点特征得出。请注意， $\{\mathbf{z_i}\}_{i=1}^{N}$是编码器提取的完整点特征集合，在解码过程中不会更新。

由于每个$\mathbf{q_i}$ 的键和值集合取决于其边界框预测的大小，因此 K_i/V_i的大小会因查询而异。为实现批量计算，我们为采样过程设定最大点数 (N_{local})，并使用 N_{local} 作为每个查询点的固定词元长度。对于包含少于 N_{local} 个点的边界框，我们用未使用的令牌填充点序列至 N_{local}，并在交叉注意力函数中屏蔽未使用的令牌；对于包含多于 N_{local} 个点的边界框，我们会随机丢弃多余的点，并截断序列，使其包含 N_{local} 个点作为键和值。最后，当边界框为空时，我们会在物体候选周围执行球查询 [4]，以采样 N_{local} 个点。

与 MS-A 类似，Local-A 不会对模块的输入提出额外要求。因此，我们可以将其应用于任意 Transformer 层。我们在 Transformer 的末端应用 Local-A，此处的边界框提议通常更准确。结合第IV-D节中的相关消融研究，我们通过实验将 $N_{local} = 16$ 设为该值。

IV. 实验

本节首先在两个广泛使用的室内点云检测数据集 ScanNetV2 和 SUN RGB-D 上评估我们的方法。接着，我们对边界框和注意力权重的可视化结果进行定性分析。我们还使用提出的尺寸感知指标开展评估。最后，我们针对方法的设计选择进行了消融研究。

数据集。 *ScanNetV2* [24] 由 1513 个重建网格组成，涵盖数百个室内场景，并为三维场景理解任务提供丰富的标注，包括分类、语义分割和目标检测。对于目标检测，它提供了包含 18 个目标类别的轴对齐边界框。我们遵循官方数据集划分，使用 1201 个样本进行训练，312 个样本进行测试。*SUN RGB-D* [25] 是一个包含 10335 个样本的单视图 RGB-D 数据集。对于三维目标检测，它提供了包含 37 个目标类别的有向边界框标注；不过，我们遵循标准评估协议 [52]，仅使用十个常见类别。训练集和测试集分别包含 5285 个和 5050 个样本。

评估指标。对于两个数据集，我们遵循标准评估协议 [52]，并使用平均精度均值（mAP）作为评估指标。我们报告了在两个不同交并比（IoU）阈值下的 mAP 分数：mAP@0.25 和 mAP@0.5。此外，在第 IV-B 节中，为了评估模型在不同物体尺寸下的性能，我们沿用 2D 视觉领域的做法 [55]，采用尺寸感知指标，分别衡量小型、中型和大型物体的 mAP。由于点云训练和推理具有随机性，我们对模型进行 5 次训练，并对每个模型测试 5 次。我们报告 25 次试验中的最佳结果和平均结果。

基线方法。 我们将方法应用于三种不同的基于 Transformer 的点云检测器：Group-Free [13],RepSurf [56], 和 3DETR [12]，以验证方法并展示其与模型无关的特性。Group-Free 使用 PointNet++ [6] 骨干网络对点特征进行编码，并通过采用普通注意力的 Transformer 解码器学习对象特征。我们考虑两种 Group-Free 配置：Group-Free^{6,256}和 Group-Free^{12,512}，，其中 Group-Free$^{L,\ O}$ 表示具有 L 个解码器层和 O 个物体候选的变体。RepSurf-U 从一种新颖的多表面表示中学习点特征，该表示能够显式描述局部几何。它采用与 [13]相似的 Transformer 解码器，并有两种配置。

RepSurf-U 用于目标检测的官方实现和平均结果尚未公开，因此我们纳入了自行复现的 RepSurf-U 结果。3DETR 是另一种端到端 3D Transformer 检测器。与其他基线方法不同，3DETR 的编码器和解码器都是 Transformer。3DETR 的官方结果并非基于多次运行，因此我们报告了自行复现的基线结果。

我们还纳入了以往基于点的三维检测器进行比较。VoteNet [52] 通过端到端可优化的霍夫投票聚合物体候选的特征。H3DNet [57] 提出了一组用于目标检测的混合几何基元，并为每种基元训练独立的骨干网络。Pointformer [18] 提出了一种层次化架构作为骨干网络，并采用 VoteNet 的投票算法进行目标检测。

实现细节。对于具有 L 个 Transformer 层的基线模型，我们将第 1 层的交叉注意力替换为 MS-A，以实现多尺度特征学习。在第 L 层（即最后一个解码器层），我们将其交叉注意力替换为 Local-A。我们遵循基线模型 [12], [13], [56] 的原始训练设置。

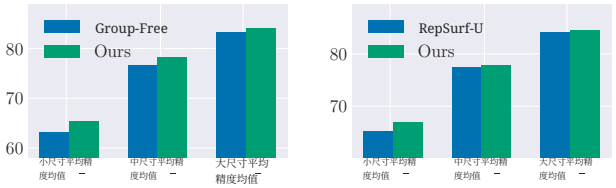
表 I**ScanNetV2 上的目标检测性能**。

我们遵循标准流程 [52]，报告 25 次试验（5 次训练，每次训练进行 5 次测试）中的最佳结果，括号内为平均结果。Backbone 表示编码器模型架构，PN++ 表示 PointNet++。Group-Free$^{L,\ O}$ 表示具有 L 个解码器层和 O 个物体候选的变体。RepSurf-U 也采用相同的标记方式。请注意，RepSurf 的检测代码尚未发布，因此我们实现了自己的 RepSurf-U 版本，并报告复现结果（重复）。

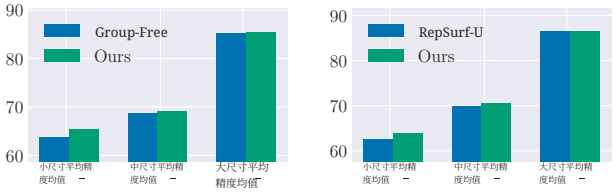
方法	#参数	骨干网络	ScanNet V2	
			mAP@0.25	mAP@0.50
VoteNet [52]	-	PN++	62.9	39.9
H3DNet [57]	-	PN++	64.4	43.4
H3DNet [57]	-	4×PN++	67.2	48.1
Pointformer [18]	-	Transformer	64.1	42.6
3DETR-m [12] (重复)	7.4M	Transformer	64.1 (62.9)	45.8 (44.1)
采用多尺度 + 局部（我们的方法）	7.5M	Transformer	**64.8 (63.7)**	**46.4 (45.2)**
Group-Free^{6,256} [13]	14.5M	PN++	67.3 (66.3)	48.9 (48.5)
搭配 多尺度局部（我们的方法）	14.6M	PN++	67.8 (66.8)	50.8 (49.5)
RepSurf-U^{6,256} [56]	14.5M	PN++	68.8 (-)	50.5 (-)
RepSurf-U^{6,256} (重复)	14.5M	PN++	68.0 (67.4)	50.2 (48.7)
采用多尺度 + 局部（我们的方法）	14.6M	PN++	**69.9 (68.7)**	**53.0 (51.3)**
Group-Free ^{12,512} [13]	29.6M	PN++w2x	69.1 (68.6)	52.8 (51.8)
采用多尺度 + 局部（我们的方法）	29.6M	PN++w2x	70.3 (69.0)	53.5 (52.3)
RepSurf-U^{12,512} [56]	29.7M	PN++w2x	71.2 (-)	54.8 (-)
RepSurf-U^{12,512} （重复）	29.7M	PN++w2x	70.8 (70.2)	54.4 (53.6)
采用多尺度局部（我们的方法）	29.8M	PN++w2x	**71.4 (70.6)**	**55.6 (54.3)**

表 II**SUN RGB-D 上的目标检测性能**。

方法	mAP@0.25	mAP@0.50
VoteNet [52]	59.1	35.8
H3DNet [57]	60.1	39.0
3DETR-m [12]	59.1	32.7
Pointformer [18]	61.1	36.6
Group-Free^{6,256} [13]	63.0 (62.6)	45.2 (44.4)
使用多尺度 +局部（我们的方法）	63.3 (62.8)	45.8 (44.9)
RepSurf-U^{6,256} [56]	64.3(-)	45.9 (-)
RepSurf-U^{6,256} (重复)	64.0 (63.3)	45.7 (45.2)
采用多尺度 +局部（我们的方法）	**64.2 (63.6)**	**47.0 (45.7)**



(a) ScanNetV2 上按尺寸类别（S/M/L）划分的平均精度均值。



（b）SUN-RGBD 上按尺寸类别（S/M/L）划分的平均精度均值。

图 4. **不同大小物体上的性能**。我们根据各数据集的统计信息（体积分布）定义小/中/大阈值。

A. 主要结果

从表 I 可以看出，基线模型配备我们的注意力模块后，性能均有稳定提升。在 ScanNetV2 上，将多尺度-A 和 Local-A 应用于 Group-Free 后，我们的表现与最先进的 RepSurf-U 检测器不相上下；此外，只需增加很少的参数开销，就能在不同模型配置下进一步提升 RepSurf-U 的性能。表 II 显示了 SUN RGB-D 上的类似趋势。我们的方法在 ScanNetV2 上的提升幅度大于在 SUN RGB-D 上的提升幅度，这可归因于两个数据集的复杂度不同：ScanNetV2 中的一个样本包含整个场景的扫描数据，而 SUN RGB-D 中的一个样本只包含单视图。在处理具有精细原始数据的复杂场景时，我们的层次注意力优势更加显著。

B. 不同大小物体上的性能。

除了标准评估指标外，我们还考察了模型在不同大小物体上的性能。我们遵循二维检测领域的惯例，定义用于三维检测的尺寸感知指标 [55]。我们将 mAP_S 的阈值设为所有物体体积的第 $30th$ 百分位数，并将第 $70th$ 百分位数设为 mAP_L 的阈值。

我们在两个基准数据集上评估了 Group-Free 和 RepSurf 模型。从图 4 的结果可以看出，对于这两种模型，较小物体的 mAP（mAP_S）都远低于较大物体。当采用我们提出的注意力模块时，mAP_S 的性能提升最大。

C. 定性结果

在图5中，我们展示了两个数据集上的定性结果。可视化结果是将我们的方法应用于 Group-Free 检测器后得到的。定性结果表明，即使在包含十多个物体的复杂场景中，我们的模型也能检测并分类不同尺度的物体（例如，底行中的示例）。通过查看 Transformer 检测器中的交叉注意力权重，我们发现

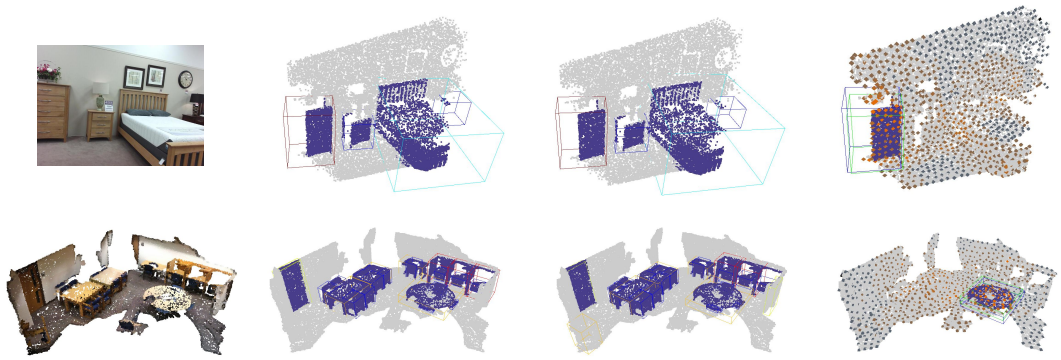


图 5. **SUN RGB-D（上）和 ScanNetV2（下）上的定性结果**。中间两列中边界框的颜色表示物体的语义标签。最后一列中，我们同时绘制了物体的真实标注（绿色）和预测结果（蓝色）。为便于观察，我们突出显示了属于物体的点。最后一列中，我们展示了应用 Local-A 之前最后一个 Transformer 层的注意力权重。我们展示了物体候选与点云之间的交叉注意力权重。

物体候选往往与属于其对应物体的点具有更高的相关性。

D. 消融研究

Local-A 中的最大点数（N_{local}）。在 Local-A 中，对于每个物体候选（即查询），我们在其对应的边界框提议内采样一组点，并在交叉注意力函数中将这些点特征用作该物体候选的键和值。如第 III-C 节所述，我们将采样点数限制为N_{local}，以便进行批量计算。

表 III**LOCAL-A 中 N_{local} 的影响**。

当点数充足时，较大的 N_local 表示在每个边界框提议内采样的点更加密集。

N_{local}	mAP@0.25	mAP@0.50	mAP_S	mAP_M	mAP_L
8	67.8	51.1	64.3	77.2	82.8
16	**68.8**	**52.3**	65.1	**77.9**	83.4
24	68.7	**52.3**	**65.2**	77.7	83.5
32	68.3	52.1	64.7	77.3	**83.8**

从表 III 可以看出，Local-A 中的点数过少（例如 $N_{local} = 8$）会导致性能下降。随着 N_{local} 增加，当其超过 $N_{local} = 16$时，我们没有观察到明显的性能提升。直观而言，较小的 N_{local} 意味着每个边界框内的点采样较为稀疏，可能稀疏到无法为任何物体提供足够的信息。这解释了为什么 $N_{local} = 8$效果不佳。另一方面，较大的 N_{local} 可能只对大型物体有益，对较小物体影响甚微，因为较小物体会填充未使用的令牌。

具有不同特征分辨率的 MS-A。与 MS-A 中的可学习上采样能够生成更高分辨率的特征类似，我们可以使用传统的集合抽象来实现可学习的下采样 [4]，该方法会聚合局部组内的点特征，并生成点数更少的特征图（即分辨率更低的特征图）。直观来说，分辨率更高的特征图能提供更细粒度的几何细节，而分辨率较低的特征图则可能提供更全面的全局上下文。我们对具有不同采样比例的 MS-A 进行了实证分析，以研究不同粒度特征图的影响。我们选择 0.5 和 2.0 作为采样比例，分别代表粗粒度和细粒度特征。

表 IV**采用不同特征尺度的 MS-A**。

特征尺度 $= s$ 表示特征图包含 $s \times N$ 个点。较大的 s 表示特征图具有更高的点密度（即，分辨率）。

尺度 s	mAP@0.25	mAP@0.50	mAP_S	mAP_M	mAP_L
[1]	68.6	51.8	63.1	76.6	83.2
[1, 2]	**68.9**	**52.5**	**65.0**	**77.5**	**83.9**
[0.5, 1, 2]	67.9	51.7	64.6	76.7	**83.9**

表 IV 中的结果表明，粗粒度特征（$s = 0.5$）对 Transformer 检测器没有帮助。这是预期之中的，因为 Transformer 的感受野不受限制，因此不需要粗粒度特征图来学习全局上下文。

V. 结论

本文提出聚合多尺度注意力（MS-A）和尺寸自适应局部注意力（Local-A）这两种通用的基于点的注意力操作，可应用于各种 3D Transformer 检测器，并实现细粒度特征学习。我们在两个具有挑战性的室内 3D 检测基准上提升了基于点的 Transformer 检测器的性能，其中小型物体的提升幅度最大。由于我们的方法促进细粒度 3D 特征学习，而这对许多 3D 视觉系统都很重要，未来的工作将把所提出的注意力模块应用于其他任务，例如分割。另一个未来方向是将我们的方法用于提升自动驾驶车辆等应用中的室外检测器。鉴于主流室外检测器通常不是基于点的，因此可能需要将这些注意力操作调整为适用于其他 3D 表示。

参考文献

[1] B. Graham、M. Engelcke 和 L. van der Maaten，“利用子流形稀疏卷积网络进行三维语义分割”，发表于 CVPR，2018年。[2] L. Landrieu 和 M. Simonovsky，“利用超点图进行大规模点云语义分割”，发表于 CVPR，2018年。[3] C. R. Qi、W. Liu、C. Wu、H. Su 和 L. J. Guibas，“用于从 RGB-D数据进行三维目标检测的视锥 PointNets”，发表于 CVPR，2018年。[4] C. R. Qi、H. Su、K. Mo 和 L. J. Guibas，“PointNet：点集上的深度学习，用于三维分类与分割”，发表于 CVPR，2017年。[5] C. R. Qi、H. Su、M. Nießner、A. Dai、M. Yan 和 L. J. Guibas，“用于三维数据目标分类的体积表示与多视图卷积神经网络”，发表于 CVPR，2016年。[6] C. R. Qi、L. Yi、H. Su 和 L. J. Guibas，“PointNet++：度量空间中点集上的深度层次化特征学习”，发表于神经信息处理系统大会，2017年。[7] S. Shi、X. Wang 和 H. Li，“PointRCNN：从点云生成三维目标提议并进行检测”，发表于 CVPR，2019年。[8] H. Su、V. Jampani、D. Sun、S. Maji、E. Kalogerakis、M. Yang 和 J. Kautz，“SplatNet：用于点云处理的稀疏格网络”，发表于 CVPR，2018年。[9] Y. Xu、T. Fan、M. Xu、L. Zeng 和 Y. Qiao，“SpiderCNN：利用参数化卷积滤波器对点集进行深度学习”，发表于 ECCV，2018年。[10] Y. Yang、C. Feng、Y. Shen 和 D. Tian，“FoldingNet：通过深度网格变形实现的点云自编码器”，发表于 CVPR，2018年。[11] T. Guan、J. Wang、S. Lan、R. Chandra、Z. Wu、L. Davis 和 D. Manocha，“M3DETR：利用 Transformer 进行多表征、多尺度和互关系三维目标检测”，发表于 IEEE应用计算机视觉冬季会议，2022年。[12] I. Misra、R. Girdhar 和 A. Joulin，“用于三维目标检测的端到端Transformer模型”，发表于 ICCV，2021年。[13] Z. Liu、Z. Zhang、Y. Cao、H. Hu 和 X. Tong，“通过 Transformer 实现无分组三维目标检测”，发表于 ICCV，2021年。[14] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，“注意力机制就是一切”，发表于自然语言处理与计算语言学研讨会，2017年。[15] A. Dosovitskiy、L. Beyer、A. Kolesnikov、D. Weissenborn、X. Zhai、T. Unterthiner、M. Dehghani、M. Minderer、G. Heigold、S. Gelly、J. Uszkoreit 和 N. Houlsby，“一张图像胜过 16×16 个词元：用于大规模图像识别的 Transformer”，发表于 ICLR，2021年。[16] Z. Liu、Y. Lin、Y. Cao、H. Hu、Y. Wei、Z. Zhang、S. Lin 和 B. Guo，“Swin Transformer：使用移位窗口的层次视觉 Transformer”，发表于 ICCV，2021年。[17] H. Zhao、L. Jiang、J. Jia、P. H. S. Torr 和 V. Koltun，“Point Transformer”，发表于 ICCV，2021年。[18] X. Pan、Z. Xia、S. Song、L. E. Li 和 G. Huang，“利用 Pointformer 进行三维目标检测”，发表于 CVPR，2021年。[19] C. Zhang、H. Wan、X. Shen 和 Z. Wu，“PatchFormer：一种高效的、具有补丁注意力的点云 Transformer”，发表于 CVPR，2022年。[20] X. Yu、L. Tang、Y. Rao、T. Huang、J. Zhou 和 J. Lu，“Point-BERT：通过掩码点建模对三维点云 Transformer 进行预训练”，发表于 CVPR，2022年。[21] J. Yang、Q. Zhang、B. Ni、L. Li、J. Liu、M. Zhou 和 Q. Tian，“利用自注意力和 Gumbel子集采样对点云进行建模”，发表于 CVPR，2019年。[22] Z. Liu、Z. Chen、S. Xie 和 W. Zheng，“TransGrasp：用于七自由度抓取检测的多尺度层次点云 Transformer”，发表于 ICRA，IEEE，2022年，第 1533–1539 页。[23] J. Kini、A. Mian 和 M. Shah，“3DMoDT：利用注意力引导亲和度在三维点云中联合检测与跟踪”，发表于 ICRA，IEEE，2023年，第 841–848 页。[24] A. Dai、A. X. Chang、M. Savva、M. Halber、T. A. Funkhouser 和 M. Nießner，“ScanNet：带有丰富标注的室内场景三维重建”，发表于 CVPR，2017年。[25] S. Song、S. P. Lichtenberg 和 J. Xiao，“SUN RGB-D：RGB-D 场景理解基准套件”，发表于 CVPR，2015年。[26] S. Feng、Z. Zhou、J. S. Smith、M. Asselmeier、Y. Zhao 和 P. A. Vela，“GPF-BG：用于安全四足导航的层次化视觉规划框架”，发表于 ICRA，IEEE，2023年，第 1968–1975 页。[27] C. Wen、H. Huang、Y. Liu 和 Y. Fang，“用于三维激光雷达地点识别的金字塔可学习令牌”，发表于 ICRA，IEEE，2023 年，第 4143–4149 页。[28] S. Ren 、 D. Zhou 、 S. He 、 J. Feng 和 X. Wang，“通过多尺度令牌聚合实现分流自注意力”，发表于 CVPR，2022年。[29] Y. Zhou 和 O. Tuzel，“VoxelNet：用于基于点云的三维目标检测的端到端学习”，发表于 CVPR，2018年。[30] A. H. Lang、S. Vora、H. Caesar、L. Zhou、J. Yang 和 O. Beijbom，“PointPillars：用于点云目标检测的快速编码器”，发表于 CVPR，2019年。[31] S. Shi、C. Guo、L. Jiang、Z. Wang、J. Shi、X. Wang 和 H. Li，“PV-RCNN：用于三维目标检测的点-体素特征集抽象”，发表于 CVPR，2020年。[32] M. Ye、S. Xu 和 T. Cao，“HVNet：用于基于激光雷达的三维目标检测的混合体素网络”，发表于 CVPR，2020年。[33] H. Wang、L. Ding、S. Dong、S. Shi、A. Li、J. Li、Z. Li 和 L. Wang，“CAGroup3D：用于点云三维目标检测的类别感知分组”，发表于 NeurIPS，2022年。[34] D. Maturana 和 S. Scherer，“VoxNet：用于实时目标识别的三维卷积神经网络”，发表于2015年 IEEE/RSJ智能机器人与系统国际会议（IROS），2015年。[35] G. Riegler、A. O. Ulusoy 和 A. Geiger，“OctNet：以高分辨率学习深度三维表示”，发表于 CVPR，2017年。[36] P. Wang、Y. Liu、Y. Guo、C. Sun 和 X. Tong，“O-CNN：用于三维形状分析的基于八叉树的卷积神经网络”，《ACM图形学汇刊》，第36卷，第4期，第72:1–72:11页，2017年。[37] C. He、R. Li、S. Li 和 L. Zhang，“体素集变换器：一种基于点云的三维目标检测集到集方法”，发表于 CVPR，2022年。[38] X. Wang、J. Lei、H. Lan、A. Al-Jawari 和 X. Wei，“DueqNet：用于自动驾驶室外三维目标检测的双等变网络 ”，发表于 ICRA。IEEE，2023年，第6951–6957页。[39] M. Simonovsky 和 N. Komodakis，“图卷积神经网络中的动态边条件滤波器”，发表于 CVPR，2017年。[40] Y. Wang、Y. Sun、Z. Liu、S. E. Sarma、M. M. Bronstein 和 J. M. Solomon，“用于点云学习的动态图卷积神经网络”，《ACM图形学汇刊》，第38卷，第5期，第146:1–146:12页，2019年。[41] Q. Xu、X. Sun、C. Wu、P. Wang 和 U. Neumann，“用于快速、可扩展点云学习的 Grid-GCN”，发表于 CVPR，2020年。[42] H. Zhou、Y. Feng、M. Fang、M. Wei、J. Qin 和 T. Lu，“用于点云分析的自适应图卷积”，发表于 ICCV，2021年。[43] X. Ma、C. Qin、H. You、H. Ran 和 Y. Fu，“重新思考点云中的网络设计与局部几何：一种简单的残差多层感知机框架”，发表于 ICLR，2022年。[44] H. Wang、S. Shi、Z. Yang、R. Fang、Q. Qian、H. Li、B. Schiele 和 L. Wang，“RBGNet：用于三维目标检测的基于射线的分组”，发表于 CVPR，2022年。[45] X. Wu、Y. Lao、L. Jiang、X. Liu 和 H. Zhao，“点变换器V2：分组向量注意力与基于分区的池化”，发表于 NeurIPS，2022年。[46] M. Xu、R. Ding、H. Zhao 和 X. Qi，“PAConv：在点云上采用动态核组装的位置自适应卷积”，发表于 CVPR，2021年。[47] Y. Li、C. Wu、H. Fan、K. Mangalam、B. Xiong、J. Malik 和 C. Feichtenhofer，“MViTv2：用于分类和检测的改进型多尺度视觉变换器”，发表于 CVPR，2022年。[48] X. Zhu、W. Su、L. Lu、B. Li、X. Wang 和 J. Dai，“可变形DETR：用于端到端目标检测的可变形变换器”，发表于 ICLR，2021年。[49] Y. LeCun、B. E. Boser、J. S. Denker、D. Henderson、R. E. Howard、W. E. Hubbard 和 L. D. Jackel，“反向传播在手写邮政编码识别中的应用”，《神经计算》，1989年。[50] Z. Chen、Z. Li、S. Zhang、L. Fang、Q. Jiang 和 F. Zhao，“用于动态多模态三维目标检测的可变形特征聚合”，发表于 ECCV，2022年。[51] X. Lai、J. Liu、L. Jiang、L. Wang、H. Zhao、S. Liu、X. Qi 和 J. Jia，“用于三维点云分割的分层变换器”，发表于 CVPR，2022年。[52] C. R. Qi、O. Litany、K. He 和 L. J. Guibas，“点云三维目标检测的深度霍夫投票”，发表于 ICCV，2019年。[53] X. Chen、H. Ma、J. Wan、B. Li 和 T. Xia，“用于自动驾驶的多视角三维目标检测网络”，发表于 CVPR，2017年。[54] K. Tian、Y. Ye、Z. Zhu、P. Li 和 G. Huang，“用于鸟瞰视图局部地图构建的高效混合解码器”，发表于 ICRA。IEEE，2023年，第8378–8385页。[55] T. Lin、M. Maire、S. J. Belongie、J. Hays、P. Perona、D. Ramanan、P. Dollár 和 C. L. Zitnick，“Microsoft COCO：上下文中的常见物体”，发表于 ECCV，2014年。[56] H. Ran、J. Liu 和 C. Wang，“点云的表面表示”，发表于 CVPR，2022年。[57] Z. Zhang、B. Sun、H. Yang 和 Q. Huang，“H3DNet：使用混合几何基元进行三维目标检测”，发表于 ECCV，2020年。