arXiv:2302.06291v1 [cs.CV] 13 Feb 2023

表面偏置的多层上下文三维目标检测

Sultan Abu Ghazal[1] a,Jean Lahoud[1] b and Rao Anwer[1] c
穆罕默德·本·扎耶德人工智能大学，阿联酋阿布扎比{
sultan.abughazal, jean.lahoud, rao.anwer}@mbzuai.ac.ae

关键词: 检测、分割、多层上下文、三维。

摘要: 三维点云中的目标检测是计算机视觉多个应用领域中的一项关键任务，包括机器人技术、自动驾驶汽车和增强现实。本研究采用一种高效的表面偏置特征提取方法（Wang 等，2022），处理三维点云中的目标检测任务，该方法还能够捕获多个层面的上下文线索。我们提出了一种三维目标检测器，可提取候选对象的精确特征表示，并在三维场景中的点块、候选对象和全局场景上运用自注意力。Qian 等，2020 证明，自注意力能够有效编码三维点云中的相关性信息。其他三维检测器则致力于通过选择性地获取更有意义的局部特征来增强点云特征提取（Wang 等，2022），但忽略了上下文信息。为此，本文提出的架构采用基于射线的表面偏置特征提取和多层上下文编码，性能超越了最先进的三维目标检测器。本文在 ScanNet数据集的场景上进行了三维检测实验，并逐一引入自注意力模块，以分离分析各层级自注意力的作用。代码见
https://github.com/SultanAbuGhazal/SurfaceBaisedMLevelContext

1 引言

三维点云中的目标检测与图像中的目标检测类似，其任务是通过三维边界框在三维场景中定位一个或多个物体，并识别或分类这些物体。点云除了分布不规则之外，还是一种无序数据集；以不同方式排列点云中的点列表，并不会改变点云所表示的形状。此外，点云的意义体现在点邻域中；单个或孤立的三维点不包含任何信息。因此，必须考虑相邻点之间的交互。最后，形状的点云表示会因平移和旋转等变换而发生剧烈变化，因此，所学习的几何特征表示必须对这类变换保持不变。

此前有多项研究提出了不同的方法，以实现可靠的三维目标检测。值得一提的是，Lin 等，2013 提出了一种利用几何特征以及三维场景的二维图像分割的方法，用于实现整体场景理解，并预测物体性分数和物体边界框。这类方法容易受到遮挡和扫描角度剧烈变化的影响。

其他检测器采用基于模板的方法，通过三维场景中的关键点搜索预构建的常见三维对象数据库，以寻找与场景中对象的匹配项。在一些情况下（Li 等人，2015年；Nan 等人，2012年；Litany 等人，2017年），会对预构建数据库中匹配的手工制作物体模型进行缩放和定向，使其与扫描对象相匹配，并用它替换场景中的对象。（Song 和 Xiao，2014）使用单个物体类别的数百张二维渲染图生成物体的合成深度图，并据此生成用于拟合支持向量机的点云。测试时，三维滑动窗口在三维场景中移动，每一步都使用先前训练的支持向量机对窗口重叠区域内的三维点进行分类。与滑动窗口类似，后续工作（Song 和 Xiao，2016）提出使用区域提议网络（RPN）检测点云中的对象，其计算开销远高于用于二维图像目标检测的区域提议网络。较新的方法（Hou 等人，2019年）结合点云和二维图像，在三维目标检测任务上训练模型。

a https://orcid.org/0000-0003-0416-5964 b
https://orcid.org/0000-0003-0315-6484 c
https://orcid.org/0000-0002-9041-2214

对于仅经过最少预处理便处理点云的基于点的三维检测器，（Qian 等，2020）提出一种方法，利用自注意力和多尺度特征融合，从点云中提取多层级上下文信息，以完成三维目标检测任务。一个三维检测网络基于（Qi 等人，2019）的成功构建，能够识别三维对象与全局场景之间的相关性。该网络在（Qi 等人，2019）的不同阶段引入了三个上下文模块：分别作用于种子点层级和投票簇层级，并添加一个模块来编码全局场景特征。当前性能最佳的三维检测网络（Wang 等，2022）采用了一种有效的采样与分组方案，该方案倾向于选取物体表面上的点。它用基于射线的分组阶段替换（Qi 等人，2019）的采样与分组阶段，并引入前景偏置采样（FBS）方案，以替换骨干网络中的最远点采样（FPS），从而选择种子点。该网络采用（Qi 等人，2017）作为骨干网络来提取点特征并执行点集抽象，与（Qi 等人，2019）类似。此外，该检测器借鉴（Qi 等人，2019）的投票模块，从每个种子点生成一个投票。

在以下各节中，我们将讨论相关工作，并依次详细介绍所提出架构的不同组成部分。随后，我们将介绍所进行的实验，并讨论所得结果。

2 相关工作

点云通常是一种高度不规则的数据格式。数据采集过程无法保证所捕获的数据点均匀分布。这一特性使得难以直接应用常见的卷积操作，从三维点云中提取特征。传统上，人们设计了多种不同的方法来应对点云的不规则性。一种直观的方法是使用体素卷积神经网络（Wu 等，2014；Qi 等，2016b；Maturana 和 Scherer，2015；Wang 和 Posner，2015；Li 等，2016；Yin 等，2021；Shi 等，2019；Deng 等，2020；Song 和 Xiao，2015），或进行体素化，将输入点云插值到三维体素网格上。在这类方法中，最终体素表示的分辨率由体素的尺寸决定。体素类似于三维像素，构成密度均匀的规则数据格式。这种表示方式使得可以通过卷积操作从三维空间中提取特征。这些三维操作的计算量很大，尤其是在分辨率较高时。

另一种方法是使用多视图卷积神经网络（Su 等人，2015年；Savva 等人，2016年），将三维点云渲染或投影到二维平面上，再通过卷积神经网络进行分类。这类方法表现出色，但在很大程度上限制了上下文理解，也无法扩展到语义级的逐点分类。其他方法则使用三维数据的网格，并应用谱卷积神经网络，对具有流形网格的有机物体进行分类（Bronstein 和 Kokkinos，2010年；Masci 等人，2015年）。

本研究旨在拓展和进一步发展基于点的三维检测器，因为这类检测器简单，且计算需求相对较低。基于点的方法还可以减少体素方法和基于投影的方法中可能发生的信息损失。

2.1 多层级上下文 VoteNet

Qian 等，2020提出了 MLCVNet，这是一种利用自注意力和多尺度特征融合，从多个层级捕获点云上下文信息、用于三维目标检测任务的方法。现有的三维目标检测器会独立识别物体，而不考虑场景中物体之间上下文信息的利用。与此同时，MLCVNet 能够识别三维物体与全局场景之间的相关性。该方法在 Qi 等人，2019的不同阶段引入了三个上下文模块，以嵌入多个层级的上下文线索。在种子点层级，引入块到块上下文（PPC）模块，以捕获点块之间的关系；随后进入投票阶段，种子点会投票给相应的物体质心点。在投票簇层级，聚类阶段之后、预测阶段之前会加入物体到物体上下文（OOC）模块，以捕获物体候选框之间的相关性。最后，还加入了第三个模块来编码全局场景特征，称为全局场景上下文（GSC）模块。

PPC 模块旨在捕捉输入点云中各点块之间的关系。在此实现中，点块对应于通过 PointNet ++骨干网络对输入点集进行抽象后生成的种子点。该模块有两项优势：首先，汇集跨补丁信息并从其他相似图块中补充信息，可以减轻部分补丁中数据缺失的影响。此外，该模块还能捕捉跨补丁关系，将每个图块的投票信息与其他图块的信息相结合，从而增强生成的投票结果。该

过程通过紧凑型广义非局部（Compact Generalized Non-Local，简称 CGNL）自注意力模块实现（Yue 等人，2018）；该模块在投票阶段之前按通道融合所有图块的信息。

$$\mathbf{A}' = f(\theta(\mathbf{A}), \phi(\mathbf{A}))g(\mathbf{A}), \quad (1)$$

这里，θ(·)、ϕ(·)、g(·) 是变换函数，而 f (·, ·) 是用于编码输入特征中不同位置之间相似性的函数。

在 Qi 等人，2019 中，投票簇会被逐个独立处理，以生成边界框和类别预测。这使网络无法捕捉彼此靠近的不同物体之间常见的关系。因此，OOC 模块采用了图像二维目标检测中使用的一种技术（Chen 等人，2018），使相关物体能够共享加权信息。与 Qi 等人，2019 类似，在投票结果围绕选定的 K 个投票簇中心聚类后，每个簇都会通过多层感知机和最大池化层抽象为一个特征向量，其中包含一个位置分量和一个特征分量，如公式 2 所示；v_i 表示簇中的每个投票结果，n 表示该簇中的投票结果总数。随后，生成的特征会经过自注意力步骤，以在提议与分类阶段之前捕捉簇间或物体间的关系。CGNL自注意力模块会生成新的簇特征，用于编码所有簇之间的相关性。

$$C_{OOC} = \mathrm{CGNL}\left(\max_{i=1,\cdots,n}\{MLP(v_i)\}\right), \quad (2)$$

最后，第三个上下文模块旨在捕获整个场景的上下文。该模块通过在提议与分类阶段之前，将场景中的全局上下文信息编码到生成的特征中来发挥作用。受 Liu 等人，2018 年针对二维目标检测提出的结构推断方法启发，GSC 模块旨在利用全局场景上下文线索来增强 K 个候选对象的特征表示。具体做法是在 PPC 和 OOC 模块之前添加一个分支，将种子特征或点块与物体提议连接起来。

GSC 模块首先分别对 M 个图块和 K 个簇进行逐通道最大池化，得到两个特征向量：一个表示点的图块，另一个表示投票簇。随后将这两个向量拼接，并通过多层感知机进一步聚合，将特征维度缩减至与提议的维度一致。接着，将这个全局特征向量扩展至与物体提议 K 的数量一致。最后，将扩展后的特征与物体提议进行组合，并使用逐元素乘法。这一操作表示为公式 3。

$$C_{new} = MLP([max(C); max(\mathbf{P})]) + C_{OOC}, \quad (3)$$

在 ScanNet数据集上训练三维对象检测模型时，依次添加 PPC、OOC 和 GSC 模块，分别使基线平均精度均值（平均精度均值）提高 2.6、1.2 和 1.1 个百分点。总体而言，MLCVNet 在 ScanNet数据集上以 0.25 的 IoU阈值检测三维对象时，平均精度均值达到 64.5。

2.2 基于射线的分组

Wang等人，2022提出了RBGNet，这是一种有效的采样与分组方案，倾向于选择物体表面上的点。简而言之，该方法用基于射线的分组阶段替换了 Qi 等人，2019中的采样与分组阶段，并引入前景偏置采样（FBS）方案来选择种子点，以替代骨干网络中的最远点采样（FPS）。尽管 Qi 等人，2019采用的投票方案在生成物体提议方面发挥着重要作用，但对投票结果及其对应特征向量进行聚类与分组的方式却相当通用且简单，仍有改进空间。物体提议的质量直接受到分组与聚类过程质量的影响。因此，RBGNet旨在通过提出一种基于射线的分组方法来改进投票结果的分组与聚类，该方法有选择地将物体形状信息编码到物体提议中，以提高准确性。该方法不再仅仅依据投票结果与投票簇中心的距离对其进行聚类，并直接通过集合抽象模块聚合特征，而是沿着从每个投票簇中心发出的射线对投票结果进行聚类，然后以选择性方式聚合，同时考虑每个投票结果与其所属物体的相关性以及物体表面几何。基于射线的分组模块以 M 个簇中心 $\{\mathrm{c}_i\}_{i=1}^{M}$ 开始，其中 $\mathrm{c}_i = [\mathrm{v}_i;\ \mathrm{f}_i]$ 表示投票中心位置 $\mathrm{v}_i \in \mathbb{R}^3$ 及投票中心特征向量 $\mathrm{f}_i \in \mathbb{R}^3$。随后，该模块从簇中心发出 N 条射线，并依据球坐标系下的三条规则对射线的角度和长度进行标准化。首先，使用公式 4 确定与 z轴之间的 P 个极角，其中 θ_p 是第 p^{th} 个极角。

$$\theta_p = \frac{\pi p}{(P-1)},\ p \in \{0, \cdots, P-1\} \quad (4)$$

其次，根据公式 5 中的情况，待生成的 N 条射线分布在 P 个极角上。这始终会使恰好有一条射线指向

正 z轴方向，恰好有一条射线指向负 z轴方向，并且在 p^{th} 极角上有 A_p 条射线。公式 5 中的数字 4 是一个超参数，可与射线总数 N 和极角数 P 一起调整。

$$A_p = \begin{cases} 1, & \text{if } p = 0 \text{ or } P-1 \\ 4 \times p, & \text{if } 0 < p \leq \frac{P-1}{2} \\ 4 \times (P-p-1), & \text{if } P-1 > p > \frac{P-1}{2} \end{cases} \quad (5)$$

第三，每个极角上的射线的方位角按如下方式确定：同一极角上的 A_p 条射线围绕 2π 的方位角均匀分布，形成垂直于 z 轴的环。每条射线的极角 $\theta_{p,\ a} \in [0,\pi]$ 和方位角 $\psi_{p,\ a} \in [0,2\pi]$ 均使用式 6 确定。

$$\psi_{p,a} = \frac{2\pi a}{A_p},\ \theta_{p,a} = \theta_p,\ a \in \{0,\cdots,A_p-1\} \quad (6)$$

最后，为了确定射线从投票簇中心向外延伸的距离，该模块使用簇特征向量 f_i学习预测每个物体的尺度 l_i，该尺度表示物体边界框对角线长度的一半。如式 7 所示，该子网络使用平滑-ℓ_1 范数进行监督，其中 l_i^*是真实边界框对角线长度的一半，I 为正簇中心的数量。指示函数$\mathbb{I}[i_{th}$ 为正] 表示簇中心 c_i 距离真实物体中心不超过 0.3 米。

$$L_{scale-reg} = \frac{1}{I}\sum_i \|l_i - l_i^*\|_\eta \mathbb{I}[i_{th} \text{ is positive}] \quad (7)$$

确定要生成的射线只是基于射线的分组模块的第一步。为了实现聚焦于物体表面几何和细粒度形状特征的优势，按照（Mildenhall等，2020）的方法，以系统化、由粗到细的方式选择每条射线上的点进行特征提取。本质上，每条射线的长度被划分为若干区段，并根据每个区段中种子点的位置及其密度，在每个区段中以粗粒度和细粒度两种层次进行采样。在每个层次中生成锚点，并围绕这些锚点系统地采样种子点。在粗粒度和细粒度特征提取过程之前，利用骨干网络中第一个集合抽象层的信息，将种子点从1024个上采样回2048个。

从粗采样过程开始，沿每条射线以等间距生成 K_c 个锚点，将射线划分为 K_c 个区间，即区间。该锚点集合可表示为公式 8，其中 n 表示第 n^{th} 条射线。随后，根据与各锚点 $q^{(c)}{}_{n,k}$ 的接近程度，在其周围采样种子点，并使用集合抽象层进行处理，以生成一个表示射线对应区段的局部特征 $\rho^{(c)}{}_{n,k}$。

$$Q_n^{(c)} = \{q_{n,k}^{(c)} = (x_{n,k}^{(c)}, y_{n,k}^{(c)}, z_{n,k}^{(c)})\},\ k \in \{1,\cdots,K_c\} \quad (8)$$

为了有选择地对锚点进行加权，训练一个分类子网络：如果锚点位于物体表面，则为其赋予正标签，否则赋予负标签。这样，便可根据锚点的局部特征向量 $\rho^{(c)}{}_{n,k}$ 和作为输入的投票簇特征向量 f_i，预测哪些锚点及其周围的种子点属于该物体。正锚点的预测表面掩码 $m^{(c)}{}_{n,k}$ 由公式9表示。该子网络通过监督训练：对于在其邻近范围内存在真实物体表面点的锚点赋予正标签，否则赋予负标签。

生成细锚点时会重复这一过程。最终得到粗局部特征集 $P(c) = \{\rho_{k=1}^{(c)} n,k\}^{k=1}$,$n=1$ Kc,N、细特征集 $P(f) = \{\rho(f)\ n,k\}^{k=1}$,$n=1$ Kf ,N、粗点的表面种子点掩码集 $M^{(c)} = \{m^{(c)}\ n,k\}^{k=1}$,$n=1$ Kc,N、细点的表面种子点掩码集 $M(f) = \{m(f)\ n,k\}^{k=1}$,$n=1$ Kf ,N，以及对应生成的锚点的位置。

$$m_{n,k}^{(c)} = \mathcal{F}_{mask}^{(c)}(\rho_{n,k}^{(c)}, f_i) \quad (9)$$

最后，在融合阶段将提取的特征合并。首先，将粗、细表面点掩码分别应用于粗、细特征，生成掩码特征集 $\hat{\mathcal{P}}_{(c)}$ 和 $\hat{\mathcal{P}}_{(f)}$，其中掩码中对应负锚点的特征设为零。然后，将粗、细特征分别输入两个独立的聚合分支。给定射线 n，其掩码粗点特征 $\{\hat{\rho}_{n,k}^{c}\}^{K_c}{}_{k=1}$ 按顺序拼接为单个特征 $r_n^{(c)}$，用于表示该射线，随后投影为大小为 32 的特征向量。式 10 对此进行了表示，其中拼接记为 $\odot$。

$$r_n^{(c)} = \mathcal{F}_{\text{point}}^{(c)}(\{\hat{\rho}_{n,k}^{(c)}\}_{k=1}^{K_c}, \odot) \quad (10)$$

同样，粗粒度射线特征 $\mathcal{R}_{(c)} = \{r_n^{(c)}\}_{n=1}^{N}$按顺序拼接，并输入一个具有两个隐藏层、输出维度为 128 的多层感知机进行处理，如公式 11 所示。在这两个拼接步骤中，物体提议中的射线顺序均保持一致。

$$\mu^{(c)} = \mathcal{F}_{\text{ray}}^{(c)}(\{r_n^{(c)}\}_{n=1}^{N}, \odot) \tag{11}$$

对于细粒度点特征聚合分支，采用相同的策略来获得一个 128 维特征向量 $\mu^{(f)}$。在融合阶段结束时，粗粒度和细粒度聚合向量按照公式 12 进行融合，生成融合特征 g；随后，融合特征 g 与投票簇特征 f 结合，最终提升网络的性能。

$$g = \mathcal{F}_{\text{fuse}}(\mu^{(c)}, \mu^{(f)}) \tag{12}$$

RBGNet 还提出了一种更有效的种子点采样方法，用于替代骨干网络中的 FPS（Qi 等人，2017）。FPS 算法通常会大量采样不属于感兴趣物体的背景点，尤其是在第一个将点下采样至 2048 个点的集合抽象层中。因此，FBS 分别对前景点和背景点应用 FPS。为了区分前景点和背景点，骨干网络中增加了一个额外的分割头，通过真实边界框提供的掩码进行监督，以预测前景点。如式 13 所示，分割头会根据每个点的$_j$位置和 ν_j特征向量，预测该点是否属于前景。

$$\varepsilon_j = \mathcal{F}^{\text{fore}}(\nu_j, \rho_j) \in [0,1] \tag{13}$$

随后，根据得分对点进行排序，得分最高的 κ 个点被标记为前景点 $\mathcal{D}^{(f)}=\{d^{(f)}{}_j\}_{j=1}^{\kappa}$，其余点则被标记为背景点 $\mathcal{D}^{(b)} = \{d^{(b)}{}_j\}_{j=1}^{2048-\kappa}$。最后，分别对这两组点应用 FPS，并将结果合并为一个样本集，如式 14 所示，其中 $\mathcal{S}$ 是最终合并后的点集。

$$\mathcal{D}^{(\hat{f})} = \text{FPS}(\mathcal{D}^{(f)}),$$
$$\mathcal{D}^{(\hat{b})} = \text{FPS}(\mathcal{D}^{(b)}),\ \mathcal{S} = \mathcal{D}^{(\hat{f})} \oplus \mathcal{D}^{(\hat{b})} \tag{14}$$

引入基于射线的分组模块和 FBS，使三维对象检测网络的准确率超越了当前模型（Liu 等人，2021；Zhang 等人，2020）。RBGNet 在 ScanNet数据集上进行三维对象检测时，在 IoU阈值为 0.25 的条件下，25 次运行的平均结果为 69.9，最大平均精度均值（平均精度均值）达到 70.6。

3 方法

所提出的架构整合了补丁级自注意力模块、对象级自注意力模块，以及全局上下文特征，并结合基于射线的由粗到细特征分组来生成对象候选框。如图 1 所示，输入点云首先经过 PointNet++ 骨干网络进行特征提取和点集抽象，将 N 个输入点缩减为 M 个种子点。PointNet++ 中的点集抽象模块通常以 N $\times$(d $+$C) 输入矩阵作为起点，其中 N 是点的数量，d 是位置向量的维度，C 是特征向量的维度。首先使用 FPS 对输入进行采样，以尽可能全面地覆盖 M 个采样种子点。随后，以 M 个种子点为中心，将输入点云划分为多个组或局部区域，使输出为 M $\times$ n $\times$(d $+$C)，其中 n 是组大小。各组的大小可能有所不同，但幸运的是，PointNet（Qi 等人，2016a）模块能够将任意数量的点映射为固定长度的特征向量。最后，使用 PointNet 层对该输出进行抽象，生成输出 M $\times$(d $+C'$)。

随后是一个紧凑型广义非局部（CGNL）自注意力模块，用于编码补丁间信息，如图 1a 所示。该模块通过融合相似图块之间的信息来增强补丁特征。CGNL模块利用点特征中的通道相关性，以及每一对点图块之间的关系。

随后，将编码了自注意力信息的种子点特征传递至投票模块。对于每个种子点 $s_i = [x_i$，投票多层感知机预测一个指向物体中心的投票，并利用 $x_i \in \mathbb{R}^3$ 和 $f_i \in \mathbb{R}^C$（Qi 等人，2019）。该模块预测欧几里得空间中的偏移量 $\Delta x_i \in \mathbb{R}^3$，以及特征空间中的偏移量 $\Delta f_i \in \mathbb{R}^C$，使每个种子点 s_i 的投票可表示为 $v_i = [y_i; g_i]$，其中 $y_i = x_i + \Delta x_i$，$g_i = f_i + \Delta f_i$。训练时使用真实边界框中心，通过公式 15 对物体中心偏移学习进行正则化。

$$L_{vote-reg} = \frac{1}{M_{pos}} \sum_i \| \Delta x_i - \Delta x_i^* \| \, \mathbb{I}[s_i \text{ on object}], \tag{15}$$

其中，$\mathbb{I}[s_i \text{ on object}]$ 是一个指示函数，用于表示种子点是否位于物体表面上；M_{pos} 是物体表面上的种子点总数，Δx_i^* 是种子点相对于对应物体真实边界框中心的偏移量。

整体学习目标可由公式 16 表示，其中 $\mathcal{L}_{\text{fbs}}$ 是前景偏置采样损失，$\mathcal{L}^{\text{rbfg}}$ 是基于射线的分组损失，$\mathcal{L}_{\text{obj-cls}}$是物体性得分损失，$\mathcal{L}_{\text{box}}$ 是边界框预测损失，$\mathcal{L}_{\text{sem-cls}}$ 是语义分类损失。

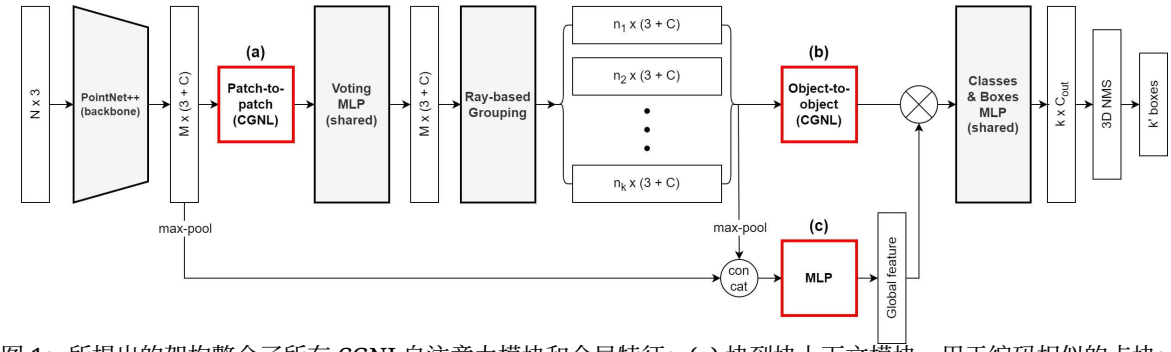


图 1：所提出的架构整合了所有 CGNL自注意力模块和全局特征；(a) 块到块上下文模块，用于编码相似的点块；(b) 对象到对象上下文模块，用于编码相关的候选对象；(c) 全局场景上下文模块，用于聚合低层和高层特征。

$$L = \lambda_{\text{vote-reg}}\mathcal{L}_{\text{vote-reg}} + \lambda_{\text{fbs}}\mathcal{L}_{\text{fbs}} + \lambda_{\text{rbfg}}\mathcal{L}_{\text{rbfg}} + \lambda_{\text{obj-cls}}\mathcal{L}_{\text{obj-cls}} + \lambda_{\text{box}}\mathcal{L}_{\text{box}} + \lambda_{\text{sem-cls}}\mathcal{L}_{\text{sem-cls}} \quad (16)$$

生成的投票结果由基于射线的分组模块（ Wang 等，2022）进行聚类和分组，该模块会生成 K 个簇，即候选对象。该模块对投票中心进行由粗到细的特征提取，并捕获对象的表面几何结构，以增强将投票中心分组为候选对象的效果。它首先将点上采样回 2048 个点，然后沿着 18 条生成的射线，以粗、细两个层级对这些点进行采样，并将其传入集合抽象层。提取出的特征经过系统融合，为每个候选对象生成一个特征向量。

为正则化基于射线的特征分组，$\mathcal{L}_{\text{rbfg}}$ 损失定义为粗粒度和细粒度交叉熵损失$\mathcal{L}_{\text{c-cls}}$ 和$\mathcal{L}_{\text{f-cls}}$之和，以及物体尺度平滑损失$\mathcal{L}_{\text{scale-reg}}$，如公式 17 所示。

$$\mathcal{L}_{\text{rbfg}} = \lambda_{\text{scale-reg}}\mathcal{L}_{\text{scale-reg}} + \lambda_{\text{c-cls}}\mathcal{L}_{\text{c-cls}} + \lambda_{\text{f-cls}}\mathcal{L}_{\text{f-cls}} \quad (17)$$

随后，使用第二个 CGNL自注意力模块处理候选对象，以编码对象级相关性（图 1b）。此时，将该模块的输出与通过聚合块和簇特征得到的全局特征相结合（图 1c）。最后，将组合后的特征向量输入全连接层，以进行分类和边界框预测。

3.1 数据集与评估指标

所提出的架构在 ScanNet（Dai 等人，2017）上进行了评估。ScanNet 是一个大型室内三维场景数据集，评估遵循标准的训练集和评估集划分。

表1：RBGNet 的基线结果，取自于在 ScanNet数据集上进行的 5 次训练。评估时，IoU阈值为 0.25。最大和平均 mAP@0.25 均低于 Wang 等（ 2022 ）报告的数值；他们在 ScanNet 上进行了 25 次训练，报告的最大平均精度均值为 70.6，平均值为 69.9。

	R1	R2	R3	R4	R5	max	avg
mAP	68.4	69.2	69.6	69.5	69.1	**69.6**	69.2
cbnt	52.2	54.9	53.4	53.8	51.5	54.9	53.1
bed	89.7	91.1	87.1	91.9	88.8	91.9	89.7
椅子	94.0	93.9	93.8	93.0	94.1	94.1	93.8
sofa	91.3	91.0	90.6	91.6	92.5	92.5	91.4
桌子	73.9	73.7	74.2	75.6	74.5	75.6	74.4
door	61.3	62.5	60.0	63.0	60.9	63.0	61.5
wdw	56.4	53.0	55.1	57.2	57.9	57.9	55.9
书架	55.4	62.5	61.8	58.8	61.0	62.5	59.9
pic	22.8	23.2	21.3	20.8	23.6	23.6	22.3
cntr	67.0	65.8	75.3	74.9	63.3	75.3	69.3
desk	76.7	76.2	76.7	78.3	79.5	79.5	77.5
crtn	53.0	49.1	51.7	49.3	52.2	53.0	51.0
冰箱	50.6	50.4	58.9	54.6	51.0	58.9	53.1
淋浴帘	73.9	72.6	74.0	70.2	75.4	75.4	73.2
马桶	95.3	99.8	98.8	99.6	99.1	99.8	98.5
sink	67.4	73.7	68.6	68.3	65.8	73.7	68.7
tub	92.7	94.3	92.1	93.6	95.0	95.0	93.5
gbin	58.1	57.1	60.1	56.0	57.5	60.1	57.8

该数据集默认不包含真实边界框，因此，边界框是根据提供的语义点标签生成的。ScanNet 包含 1513 个样本，标注了 18 种不同的对象类别。评估指标采用平均精度均值（mAP），IoU 阈值为 0.25，遵循 Qi 等人，2019 的做法。

4 结果

作为实验基线，按照作者的建议，使用4个GPU以相同的配置在ScanNet数据集上训练RBG Net（Wang 等，2022）网络。表1列出了获得的基线结果。

4.1 图块注意力

在第一个实验中，仅将块到块上下文模块添加到网络中。具体做法是对特征提取骨干网络的输出进行重塑，并将其输入CGNL模块。表2列出了在ScanNet数据集上训练该网络所获得的结果。可以观察到，这一简单的改动对整体性能的提升贡献最为显著。

表 2：在补丁级（图块到图块上下文）上加入 CGNL 自注意力模块后的结果，取自 ScanNet数据集上的 5 次训练。评估时，IoU阈值为 0.25。

	R1	R2	R3	R4	R5	max	avg
mAP	70.4	69.2	70.5	69.3	69.3	**70.5**	69.7
cbnt	54.3	55.4	53.5	56.8	51.2	56.8	54.2
bed	90.2	90.3	92.0	91.1	91.6	92.0	91.0
椅子	94.0	95.0	94.2	94.9	93.8	95.0	94.4
sofa	92.8	91.4	91.7	90.8	91.0	92.8	91.5
桌子	75.1	75.6	76.2	74.0	74.2	76.2	75.0
door	64.5	62.3	61.6	61.2	62.1	64.5	62.3
wdw	58.6	55.3	58.0	53.9	56.7	58.6	56.5
书架	58.5	56.4	65.5	49.7	54.2	65.5	56.9
pic	25.8	22.3	21.8	21.4	21.6	25.8	22.6
cntr	68.9	62.3	64.0	66.6	70.5	70.5	66.5
desk	77.6	78.0	79.1	77.3	76.3	79.1	77.7
crtn	56.9	53.0	50.9	55.2	54.3	56.9	54.1
冰箱	51.5	54.9	61.1	59.0	58.5	61.1	57.0
淋浴帘	73.8	71.9	78.1	78.0	74.6	78.1	75.3
马桶	99.1	97.3	97.4	98.6	97.3	99.1	97.9
sink	71.0	71.9	71.4	66.9	67.2	71.9	69.7
tub	95.6	93.3	94.2	94.0	93.8	95.6	94.2
gbin	58.7	58.9	57.9	57.2	58.0	58.9	58.1

4.2 块到块注意力

在第二个实验中，块到块上下文模块与物体到物体上下文模块结合使用，两者均附加到网络中。为适配新模块，我们对聚类特征进行重塑，并依次通过两个 CGNL 模块。表3列出了在ScanNet数据集上训练网络所得的结果。在所呈现的实验中，这种组合的表现不如块到块实验，与基线持平。

4.3 结合全局特征的块与物体注意力

在最后一组实验中，保留 patch-to-patch 和物体到物体上下文模块的原样，并向网络引入全局特征。为了生成全局特征，对种子特征和簇特征分别进行最大池化并拼接，然后通

表3：在补丁级（块到块上下文）和对象候选级别（物体到物体上下文）引入 CGNL 自注意力模块后，通过在 ScanNet数据集上进行 5 次训练得到的结果。评估时，IoU阈值为 0.25。

	R1	R2	R3	R4	R5	max	avg
mAP	69.7	69.8	69.4	69.6	66.8	**69.8**	69.1
cbnt	55.7	52.1	53.8	55.3	49.5	55.7	53.3
bed	89.7	89.4	88.7	89.6	89.5	89.7	89.4
椅子	94.0	94.2	93.4	94.0	93.5	94.2	93.8
sofa	91.6	88.9	91.7	90.0	87.2	91.7	89.9
桌子	74.3	75.1	76.9	74.2	72.6	76.9	74.6
door	61.6	62.2	60.4	62.4	57.6	62.4	60.8
wdw	57.8	57.4	56.3	54.5	58.0	58.0	56.8
书架	60.7	60.7	62.1	60.0	56.5	62.1	60.0
pic	22.5	22.2	23.3	23.6	19.1	23.6	22.1
cntr	65.2	67.1	66.8	68.3	48.0	68.3	63.1
desk	79.0	78.5	77.0	82.2	74.1	82.2	78.1
crtn	54.1	56.9	54.8	52.3	60.0	60.0	55.6
冰箱	55.1	60.8	55.5	52.6	48.3	60.8	54.5
淋浴帘	74.5	70.3	71.3	76.6	72.5	76.6	73.0
马桶	99.4	99.6	99.9	98.8	97.3	99.9	99.0
sink	72.2	72.2	66.1	67.6	71.5	72.2	69.9
tub	91.7	91.7	92.5	94.9	89.8	94.9	92.1
gbin	56.1	57.5	58.2	56.7	56.5	58.2	57.0

过多层感知机进行聚合，将特征维度缩减至与簇特征相匹配。表4列出了在 ScanNet数据集上训练网络所得的结果。此组合优于基线，并取得了更高的检测性能。在5次训练中，最高平均精度均值为70.2，平均值为69.6。相比之下，基线的最高平均精度均值为69.6，平均值为69.2。

表4：在补丁级（补丁到补丁上下文）和对象候选级别（物体到物体上下文）引入 CGNL 自注意力模块，并结合全局上下文特征后，在 ScanNet 数据集上进行 5 次训练所得的结果。评估时，IoU 阈值为 0.25。

	R1	R2	R3	R4	R5	max	avg
mAP	69.2	70.2	69.8	69.8	69.2	**70.2**	69.6
cbnt	54.4	57.6	55.3	55.8	52.1	57.6	55.0
bed	90.4	91.4	88.7	90.9	90.4	91.4	90.3
椅子	94.3	93.8	93.7	93.4	93.7	94.3	93.8
sofa	90.4	88.8	91.9	92.2	90.7	92.2	90.8
桌子	75.1	74.6	73.6	73.2	75.8	75.8	74.4
door	61.5	59.9	61.5	62.5	62.4	62.5	61.5
wdw	53.4	57.0	57.1	57.9	57.7	57.9	56.6
书架	58.6	56.4	57.0	55.6	61.2	61.2	57.8
pic	23.6	24.7	22.6	22.9	24.1	24.7	23.6
cntr	65.1	67.3	62.9	69.3	68.7	69.3	66.7
desk	78.9	80.3	81.4	82.2	79.9	82.2	80.5
crtn	53.9	58.9	49.6	54.7	52.0	58.9	53.8
冰箱	57.1	59.3	65.3	59.2	60.2	65.3	60.2
淋浴帘	71.1	73.6	75.2	73.8	69.6	75.2	72.7
马桶	99.2	98.9	97.5	99.3	99.3	99.3	98.9
sink	67.1	70.8	71.5	67.4	68.9	71.5	69.1
tub	92.9	92.4	93.7	89.8	84.5	93.7	90.6
gbin	58.9	57.6	57.3	56.4	55.1	58.9	57.0

5 结论

通过在三维点的多个层级应用自注意力来融入上下文线索，是一种强大的方法，能够显著增强特征表示，并最终提升三维场景理解能力。本文展示的实验在采用精确的、基于射线且偏向表面的特征分组的网络中，于不同层级引入了自注意力模块：首先是补丁级，然后是对象级，最后是全局场景级。我们通过多次运行报告了各模块的效果，并发现补丁级自注意力能够显著提升三维目标检测器的性能。通过更精细的集成和进一步实验，基线能够更好地利用附加的自注意力模块，尤其是更好地配合从投票簇中提取的细粒度和粗粒度特征的融合阶段。

参考文献

Bronstein, M. 和 Kokkinos, I.（2010年）。用于非刚性形状识别的尺度不变热核签名。第 1704–1711 页。Chen, Z.、Huang, S. 和 Tao, D.（2018）。目标检测的上下文细化。发表于 ECCV。Dai, A.、Chang, A. X.、Savva, M.、Halber, M.、Funkhouser, T. 和 Nießner, M.（2017）。ScanNet：室内场景的丰富标注三维重建。发表于计算机视觉与模式识别会议（CVPR）论文集，电气电子工程师学会。Deng, J.、Shi, S.、Li, P.、Zhou, W.、Zhang, Y. 和 Li, H.（2020年）。Voxel R-CNN：迈向高性能的基于体素的三维目标检测。CoRR，abs/2012.15712。Hou, J.、Dai, A. 和 Nießner, M.（2019年）。3D-SIS：RGB-D 扫描的三维语义实例分割。发表于计算机视觉与模式识别会议（CVPR）论文集，电气电子工程师学会。Li, Y.、Dai, A.、Guibas, L. 和 Nießner, M.（2015年）。用于实时三维重建的数据库辅助目标检索。计算机图形学论坛，34。Li, Y.、Pirk, S.、Su, H.、Qi, C. R. 和 Guibas, L. J.（2016年）。场探测神经网络：用于三维数据的场探测神经网络。CoRR，abs/1605.06240。Lin, D.、Fidler, S. 和 Urtasun, R.（2013）。利用 RGB-D 摄像机进行三维目标检测的整体场景理解。发表于2013年 IEEE 国际计算机视觉会议，第 1417–1424 页。Litany, O.、Remez, T.、Freedman, D.、Shapira, L.、Bronstein, A. 和 Gal, R.（2017）。ASIST：自动语义不变场景变换。计算机视觉与图像理解，157：284–299。基于图像和距离扫描的城市室内或室外场景大规模三维建模。

Liu, Y., Wang, R., Shan, S., and Chen, X. (2018). 结构推理网络：利用场景级上下文和实例级关系进行目标检测。2018 IEEE/CVF计算机视觉与模式识别会议，页码 6985–6994。Liu, Z., Zhang, Z., Cao, Y., Hu, H., and Tong, X. (2021). 通过 Transformer 实现无分组的三维目标检测。2021 IEEE/CVF国际计算机视觉会议（ICCV），页码 2929–2938。Masci, J., Boscaini, D., Bronstein, M. M., and Van-dergheynst, P. (2015). ShapeNet：非欧几里得流形上的卷积神经网络。CoRR，abs/1501.06297。Maturana, D. and Scherer, S. (2015). Voxnet：用于实时目标识别的三维卷积神经网络。页码 922–928。Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). NeRF：将场景表示为神经辐射场以进行视图合成。载于 Vedaldi, A., Bischof, H., Brox, T., and Frahm, J.-M. 编，《计算机视觉——ECCV 2020》，页码 405–421，Cham。Springer International Publishing。Nan, L., Xie, K., and Sharf, A. (2012). 一种用于杂乱室内场景理解的搜索-分类方法。ACM图形学汇刊，31。Qi, C. R., Litany, O., He, K., and Guibas, L. J. (2019). 用于点云三维目标检测的深度霍夫投票。载于 IEEE国际计算机视觉会议论文集。Qi, C. R., Su, H., Mo, K., and Guibas, L. J. (2016a). PointNet：面向点集三维分类与分割的深度学习。CoRR，abs/1612.00593。Qi, C. R., Su, H., Nießner, M., Dai, A., Yan, M., and Guibas, L. J. (2016b). 用于三维数据目标分类的体积与多视图卷积神经网络。CoRR，abs/1604.03265。Qi, C. R., Yi, L., Su, H., and Guibas, L. J. (2017). PointNet++：度量空间中点集的深度层次特征学习。CoRR，abs/1706.02413。Qian, X., Yu-kun, L., Jing, W., Zhoutao, W., Yiming, Z., Kai, X., and Jun, W. (2020). MLCVNet：用于三维目标检测的多层上下文 VoteNet。载于 IEEE计算机视觉与模式识别会议（CVPR）。Savva, M., Yu, F., Su, H., Aono, M., Chen, B., Cohen-Or, D., Deng, W., Su, H., Bai, S., Bai, X., Fish, N., Han, J., Kalogerakis, E., Learned-Miller, E., Li, Y., Liao, M., Maji, S., Tatsuma, A., Wang, Y., and Zhou, Z. (2016). SHREC'16 赛道：从 ShapeNet Core55 进行大规模三维形状检索。Shi, S., Wang, Z., Wang, X., and Li, H. (2019). Part-a^2 net：面向点云目标检测的三维部件感知与聚合神经网络。CoRR，abs/1907.03670。Song, S. and Xiao, J. (2014). 用于深度图像中三维目标检测的滑动形状。载于 Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T. 编，《计算机视觉——ECCV 2014》，页码 634–651，Cham。Springer International Publishing。

Song, S. and Xiao, J. (2015). RGB-D图像中非模态三维目标检测的深度滑动形状。CoRR，abs/1511.02300。Song, S. and Xiao, J. (2016). RGB-D图像中非模态三维目标检测的深度滑动形状。载于 IEEE计算机视觉与模式识别会议论文集（CVPR）。Su, H., Maji, S., Kalogerakis, E., and Learned-Miller, E. G. (2015). 用于三维形状识别的多视图卷积神经网络。CoRR, abs/1505.00880。Wang, D. and Posner, I. (2015). 在线点云目标检测中的投票用于投票。Wang, H., Shi, S., Yang, Z., Fang, R., Qian, Q., Li, H., Schiele, B., and Wang, L. (2022). RBGNet：用于三维目标检测的基于射线的分组。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 1110–1119。Wu, Z., Song, S., Khosla, A., Tang, X., and Xiao, J. (2014). 用于2.5D目标识别和下一最佳视点预测的三维形状网络。CoRR, abs/1406.5670。Yin, T., Zhou, X., and Krahenbuhl, P. (2021). 基于中心的三维目标检测与跟踪。页码 11779–11788。Yue, K., Sun, M., Yuan, Y., Zhou, F., Ding, E., and Xu, F. (2018). 紧凑型广义非局部网络。NIPS。Zhang, Z., Sun, B., Yang, H., and Huang, Q.-X. (2020). H3dnet：利用混合几何基元进行三维目标检测。载于 ECCV。

此图“orcid.png”提供“PNG”格式，获取地址：

http://arxiv.org/ps/2302.06291v1

