基于 SAM 分割大模型的三维点云分割

目　录

摘要

在本文中，我们首先提出了一种新的大模型框架来同时对点云中的实例和语义进行分割；其次，为了获取更多的点云局部特征，提出了一种局部特征编码模块来聚合每个点云附近的位置和颜色信息；此外，设计了一种有效的特征融合分割模块，将嵌入高维特征空间的语义特征与实例特征进行融合，得到融合后的实例特征，同时，将融合后的实例特征嵌入语义特征空间，并与语义特征进行融合，用来促进语义分割，以得到优化后的精细语义分割结果。最后，在大型点云数据集 ScanNet V2 和 S3DIS 上对提出的方法进行了测试与评价，并与现有方法进行了比较。我们的方法在实例分割精度指标与语义分割精度指标上都实现了优秀的性能。

关键词：语义分割；实例分割；SAM 分割大模型；特征融合；三维点云。

1.引言

近年来，随着深度学习与大模型技术的快速发展，视觉通用分割大模型在二维图像理解领域取得了突破性进展。其中，Segment Anything Model（SAM）[1]作为代表性工作，通过在大规模、多样化数据上进行预训练，展现出强大的类别无关分割能力与良好的泛化性能。SAM 不依赖于固定类别标签，而是通过提示（Prompt）机制实现对任意目标区域的精准分割，在自然图像、医学影像及遥感影像等多种视觉场景中均表现出优异性能。这种“通用分割”范式为跨模态、跨任务与跨维度的视觉理解研究提供了新的思路。

与此同时，三维点云室内场景理解作为计算机视觉与机器人领域的重要研究方向，在智能机器人、增强现实、数字孪生和自动驾驶等应用中具有广泛需求。与二维图像不同，三维点云以离散点的形式直接描述真实世界的几何结构，通常包含空间坐标（x,y,z）以及颜色、反射强度等属性。室内点云场景具有结构复杂、物体密集、尺度变化大、遮挡严重等特点，同时点云数据还存在无序性、稀疏性和密度不均衡等问题，这些因素都为高精度的语义理解与目标分割带来了显著挑战。

在三维点云室内场景理解任务中，语义分割与实例分割是两类核心问题。三维点云语义分割旨在为场景中每一个点分配一个语义类别标签，如墙面、地面、桌椅等，其关注重点在于对整体场景语义结构的理解；而三维点云实例分割则进一步要求在同一语义类别下区分不同物体个体，实现“同类异体”的分离，其目标是为每个点同时赋予实例或个体的标识。两者在任务目标上存在明显差异，但在实际应用中又具有紧密联系：语义分割为实例分割提供类别先验，而实例分割则在语义理解的基础上实现更精细的目标或个体级建模，共同支撑高层次的三维场景感知。

尽管近年来基于深度学习的三维分割方法取得了显著进展，但当前研究仍面临诸多亟待解决的问题。一方面，三维点云标注成本高昂，导致大规模高质量训练数据难以获取，模型泛化能力受限；另一方面，现有三维分割模型多为任务或数据集专用，难以适应复杂多变的真实室内场景。此外，如何有效引入二维视觉大模型的先验知识，实现二维到三维跨维度、跨模态的信息对齐与协同建模，仍然缺乏成熟且统一的解决方案。因此，探索将视觉通用分割大模型 SAM 微调并迁移至三维点云室内场景语义分割与实例分割任务中，具有重要的研究价值与现实意义。

2. 研究现状

为了解决点云数据非结构化、离散且无序所带来的特征学习困难问题，早期大量研究工作[4–10]通常采用中间表示转换的方式对点云进行预处理，即将原始

点云数据映射为多视角二维图像或规则化的三维体素表示，再利用成熟的二维卷积神经网络（2D CNN）或三维卷积神经网络（3D CNN）进行特征提取与学习。这类方法在一定程度上缓解了点云数据难以直接建模的问题，并能够复用在图像和体素领域中已被广泛验证的深度学习结构。

其中，多视角图像投影方法通过从不同视角对点云物体进行投影，将空间信息转换为一组二维图像特征，从而实现对三维目标的间接建模。然而，该类方法对视角数量、视角分布及投影方向的选择高度敏感，在复杂室内场景中，点云之间的相互遮挡不可避免地导致部分结构信息无法被有效投影，进而造成跨视角特征不一致以及被遮挡区域的空间信息缺失。此外，多视角方法在视角融合过程中往往依赖简单的特征聚合策略，难以充分刻画点云在三维空间中的连续几何关系。

与此同时，体素化点云的方法通过将三维空间划分为规则网格，将点云离散化为体素占据或体素特征表示，使得三维卷积网络得以直接应用。然而，受限于计算资源与显存开销，体素分辨率通常难以设置得足够精细，导致在体素化过程中不可避免地丢失大量点云的原生几何细节与局部结构信息。在低分辨率的体素表示下，物体的边缘、细小结构以及相邻目标之间的空间区分性往往被显著削弱，从而影响分割精度。

基于上述局限性，近年来的研究逐渐转向直接对原始点云数据进行建模的方法[11–30]，以避免中间表示转换所带来的信息损失。这类方法通常将三维点云作为网络的直接输入，通过端到端的深度学习框架完成目标识别与分割任务，在一定程度上保留了点云的真实空间结构。然而，尽管直接点云方法在整体性能上取得了明显进展，其仍然面临着若干关键挑战：一方面，现有方法在局部几何特征与细粒度结构建模能力方面仍显不足，难以准确捕捉复杂室内场景中物体的细节变化；另一方面，由于点云采样不均与空间分布复杂，全局结构信息在特征聚合过程中容易被过度平滑，导致目标物体的边缘轮廓与实例边界出现模糊现象，从而限制了模型在高精度语义分割与实例分割任务中的表现。

受近年来视觉通用分割大模型（以 SAM 为代表）的成功经验启发，相关研究表明，大规模预训练模型通过在海量、多样化二维图像数据上学习，能够同时兼顾局部细粒度特征建模能力与全局结构一致性表达能力，从而有效缓解传统分割方法中局部信息不足与全局语义不稳定的问题。SAM 采用统一的编码–解码框架，并结合提示驱动的分割机制，使模型在不同尺度、不同场景下均能够生成高质量、边界清晰的分割结果，展现出极强的泛化能力与鲁棒性。这一特性为解决复杂视觉场景中的目标分割问题提供了新的技术路径。

基于上述优势，本文尝试将通用视觉分割大模型的设计思想与能力引入至三维点云室内场景分割任务中，以弥补现有点云分割方法在特征表达上的不足。与传统点云网络相比，SAM 在二维图像领域中已隐式学习到丰富的几何结构先验、边缘轮廓信息以及高层语义关系，这些先验知识对于提升三维点云中物体边界感知、实例区分能力具有重要潜在价值。因此，将 SAM 的分割能力与点云数据相

结合，有望在保持点云原生空间结构的同时，引入更强的全局上下文建模能力。

进一步地，本文通过构建二维–三维之间的有效关联机制，将二维图像中的语义与结构先验迁移至三维点云空间中，实现跨模态特征协同建模。一方面，二维图像具有更高的纹理分辨率和清晰的物体轮廓信息，能够为三维点云提供可靠的语义约束；另一方面，三维点云所包含的真实空间几何关系又能够反向补充二维图像在深度与遮挡建模方面的不足。通过融合二维通用分割大模型的先验知识与三维点云的几何特性，可以在语义分割与实例分割任务中同时提升类别判别能力与实例边界精度，从而获得更加一致、精细的三维场景分割结果。

3. 实验内容与方法

3.1 Segment Anything Model（SAM）

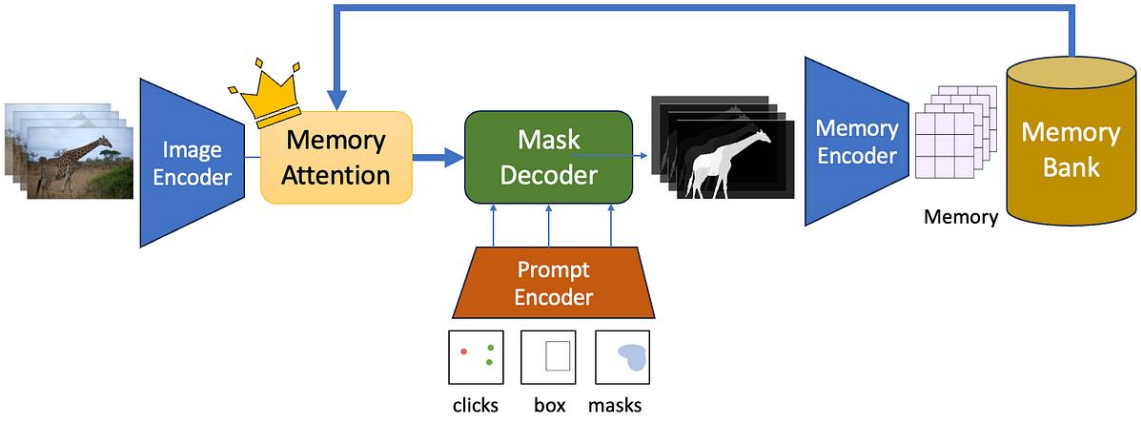

图 1 SAM 大模型的网络结构

SAM [1]采用一种统一的“编码 - 提示 - 解码（Encoder - Prompt - Decoder）”的通用分割架构，其核心思想在于将传统依赖类别监督的分割问题，重新表述为一个条件掩码生成问题。具体而言，SAM 不再直接预测固定类别的分割结果，而是将分割任务统一建模为：在给定输入图像及相应提示（Prompt）的条件下，生成对应的语义或实例掩码。这种建模方式使得模型能够摆脱对具体类别定义的强依赖，从而具备更强的任务泛化能力与场景适应性。

在该框架下，提示信息被视为一种条件约束，用于明确模型“关注什么区域”或“分割哪一个目标”，而模型的学习目标则是刻画输入图像与目标区域之间的映射关系。通过引入提示驱动机制，SAM 能够在同一模型参数下灵活应对不同尺度、不同形态以及不同数量的目标分割需求，从而实现“任意目标、任意场景”的通用分割能力。

从整体结构上看，SAM 由三个功能上相互解耦、但在特征空间中紧密协同的核心模块构成：图像编码器（Image Encoder）、提示编码器（Prompt Encoder）和掩码解码器（Mask Decoder）。其中，图像编码器负责从输入图像中提取具有全局语义一致性与局部结构表达能力的高维特征表示；提示编码器则将不同形式

的提示信息映射为可与图像特征对齐的条件嵌入，用于引导分割过程；掩码解码器在融合图像特征与提示特征的基础上，生成对应的高分辨率分割掩码及其置信度预测。通过上述模块的协同作用，SAM 实现了对复杂视觉场景中目标区域的高精度、可泛化分割建模。

3.2 图像编码器

图像编码器（Image Encoder）负责视觉特征提取，其结构基于 Vision Transformer（ViT）架构，通过多头自注意力机制对输入图像进行全局建模。与传统卷积神经网络依赖局部感受野逐层堆叠不同，ViT 通过自注意力机制在特征计算阶段即可显式建模图像中任意位置之间的长程依赖关系，从而赋予编码器大感受野与强全局上下文建模能力。这一特性使得模型能够在复杂场景中有效捕捉目标与背景之间的全局结构关系。

在具体实现上，输入的 RGB 图像首先被划分为一系列固定大小的图像块（Patch），并通过线性映射转换为序列化的 Patch 嵌入表示。在加入位置编码后，这些 Patch 特征被送入由多层 Transformer Block 组成的编码网络中进行特征建模。每个 Transformer Block 通过多头自注意力（Multi-Head Self-Attention, MHSA）与前馈网络（FFN）协同作用，使模型能够在不同子空间中并行关注图像的多种结构与语义关系，从而形成具有丰富表达能力的高维特征表示。

得益于 ViT 在不同层级对 Patch 的逐步聚合与重编码，图像编码器能够在全局建模的同时保留多尺度语义信息。一方面，较浅层的 Patch 特征更关注局部纹理与边缘结构，有助于刻画目标轮廓与细节变化；另一方面，深层特征则融合了更大范围的上下文信息，具备更强的语义判别能力。这种多尺度 Patch 表达机制使得图像编码器在兼顾局部细节表达的同时，维持全局语义一致性，为高质量语义分割与实例分割提供了关键支撑。

在编码器输出阶段，SAM 不直接生成分割结果，而是输出一组高维、稠密的图像特征嵌入，作为后续提示引导分割的基础表示。设输入图像为$I \in \mathbb{R}^{H\times W\times 3}$，图像编码器输出特征表示为：

$$F_I = \mathrm{ViT}(I), F_I \in \mathbb{R}^{\frac{\mu}{p}\times\frac{\mu}{p}\times C}$$

其中p表示 Patch 的空间尺寸，C为特征通道维度。该稠密特征图在空间上保留了与原始图像一致的几何布局关系，在语义上融合了局部与全局信息，为提示编码器与掩码解码器提供了高质量的视觉表征基础。

3.3 提示编码器

提示编码器（Prompt Encoder）是 SAM 中最具代表性的创新模块，其主要作用在于将用户或系统提供的分割提示信息转化为可学习的条件特征表示，从而显式引导模型关注特定目标区域。与传统分割模型依赖固定类别监督不同，SAM 通过提示驱动机制，将分割任务从“预测类别标签”转化为“在条件约束下生成

目标区域掩码”，使模型具备高度灵活的交互性与泛化能力。

在具体形式上，提示编码器支持多种类型的提示输入，包括点提示、框提示以及已有分割掩码提示。其中，点提示可进一步区分为正样本点与负样本点，用于分别指示目标区域内部与非目标区域；框提示通过提供目标的大致空间范围，引导模型聚焦于指定区域内的潜在实例；而分割掩码提示则可作为更强的先验条件，用于对已有分割结果进行细化与修正。这种多样化的提示设计使得 SAM 能够在不同应用场景下灵活适配多种分割需求。

在特征建模层面，提示编码器的核心任务是将不同形式、不同语义强度的提示信息统一映射到一个共享的嵌入空间中，使其能够与图像编码器输出的视觉特征进行有效融合。设提示集合为P，提示编码器通过可学习的映射函数将其转换为嵌入表示：

$$F_P = \text{Encode}_{prompt}(P)$$

其中F_P表示提示特征嵌入。对于点与框等稀疏提示，编码器通常结合其空间位置信息与类型标识进行编码；而对于掩码提示，则通过卷积或下采样操作将其映射为与图像特征分辨率一致的稠密条件特征。

通过上述方式，提示编码器在特征空间中实现了对“分割意图”的显式建模，使得模型在解码阶段能够根据提示信息有选择性地激活对应的空间区域。这一机制不仅有效降低了分割任务的歧义性，还在一定程度上缓解了复杂场景中目标粘连、实例边界模糊等问题，为生成高质量的语义与实例掩码提供了关键约束

3.4 掩码编码器

掩码解码器（Mask Decoder）是 SAM 中负责生成最终分割结果的关键模块，其主要作用是在融合图像编码器输出的视觉特征与提示编码器生成的条件特征的基础上，预测目标区域的分割掩码。为兼顾计算效率与表达能力，SAM 的掩码解码器通常采用一种轻量级 Transformer 结构结合多层感知机（MLP Head）的设计方案，使模型能够在保持高分割精度的同时，适应大规模推理与交互式应用需求。在结构上，掩码解码器以图像特征F_I与提示特征F_P作为输入，通过 Transformer 解码模块对两类特征进行联合建模。其中，图像特征提供稠密的空间语义信息，而提示特征则作为条件查询，引导模型关注与分割目标相关的区域。通过多头注意力机制，掩码解码器能够在特征空间中动态建立提示与图像局部区域之间的对应关系，从而实现对目标实例的精确定位与轮廓建模。

与传统分割模型仅输出单一预测结果不同，SAM 在解码阶段引入了多候选掩码预测机制。具体而言，对于同一组输入提示，掩码解码器会并行生成多个候选分割掩码 ，并为每个掩码预测一个对应的置信度分数，用于估计该掩码与真实目标之间的重叠程度。该置信度通常通过 IoU 预测分支进行建模，从而为后续的掩码选择提供可靠依据。这种设计有效提升了模型在提示存在歧义或目标边界复杂情况下的鲁棒性。

在数学形式上，掩码解码器的核心计算过程可表示为：

$$M_k = \sigma\big(\mathrm{MLP}_k(\mathrm{Decoder}(F_I, F_P))\big), k = 1, \ldots, K$$

其中，Decoder(·)表示 Transformer 解码模块，$\mathrm{MLP}_k(\cdot)$为第k个掩码预测头，$\sigma(\cdot)$为 Sigmoid 激活函数，用于输出像素级概率掩码。同时，模型还会对每个候选掩码预测其 IoU 置信度：

$$s_k = MLP_{IoU}(\mathrm{Decoder}(F_I, F_P))$$

其中 s_k 表示第k个掩码的质量评分。最终，系统可根据预测的 IoU 分数选择最优掩码作为输出结果。

通过上述设计，掩码解码器不仅能够充分融合全局语义信息与提示约束，还能够在实例边界不清晰或目标重叠的复杂场景中生成结构完整、轮廓清晰的分割结果。

3.5 局部增强特征编码

我们在 ViT 的每一层 transformer 块中增加一个局部增强特征编码单元，用于增强网络提取局部点云特征的能力。在局部增强特征编码单元中，对于第 i 个点，通过 KNN 算法聚合其最近的 k 个点的位置特征，KNN 算法是基于逐点的欧几里得距离得到的。公式如下所示：

$$P_i = p_i \oplus (p_i^k - p_i) \oplus p_i^k \oplus \left\| p_i^k - p_i \right\|$$

式中 p_i 表示第 i 个中心点的位置特征，p_i^k 表示在中心点附近的 k 个点的位置特征。所以 $p_i^k - p_i$ 表示邻域点到中心点的相对位置特征，$\left\| N_i^k - p_i \right\|$ 表示邻域点和中心点间的欧式距离，P_i 表示在每个中心点聚合得到的局部结构特征，$\oplus$ 表示聚合操作，。

接着，通过前面的 KNN 算法聚合 k 个邻域点的 RGB 或者中间特征，公式如下所示：

$$F_i = f_i \oplus (f_i^k - f_i) \oplus f_i^k$$

式中 f_i 表示第 i 个中心点的 RGB 或网络学习得到的中间特征，f_i^k 表示在中心点附近的 RGB 或中间特征，$f_i^k - f_i$ 表示中心点周围的相对特征，公式中没有欧几里得距离那一部分，因为很难用较低维度的距离关系去表示相邻点之间的高维特征关系，F_i 表示在每个中心点聚合得到的局部语义特征，这将会有利于网络

学习低级语义和高级语义的信息。

我们通过多层感知机将中心点的局部结构特征 P_i 嵌入到局部语义特征 F_i 中，我们采用一个有效的注意力机制使得网络关注更重要的局部特征，我们的方法如下：

$$S_i = \lambda \cdot P_i \oplus F_i$$

式中，λ 参数表示共享的多层感知机中可训练的参数，这将会有利于中心点的局部结构特征很好地嵌入到局部语义特征空间中。我们采用 Softmax 操作将局部特征 S_i 回归到 0~1 之间，得到每个特征的注意力分数 s_i，如下式所示，β 表示共享的多层感知机中可训练的参数：

$$s_i = Softmax(\beta \cdot S_i)$$

之后将注意力分数 s_i 与局部特征 S_i 相乘得到的注意力权重进行累加，得到可感知的局部特征，如下式所示。这使得网络在保留平均池化或最大池化的平移不变性的特点的同时，更加关注到了点云局部比较重要的区域：

$$\tilde{S}_i = \sum_{i=1}^{k} (S_i \cdot s_i)$$

如图 2 所示，我们的局部双特征编码单元共同作用，用 $\oplus$ 表示，使得网络学习到更有效的局部特征。

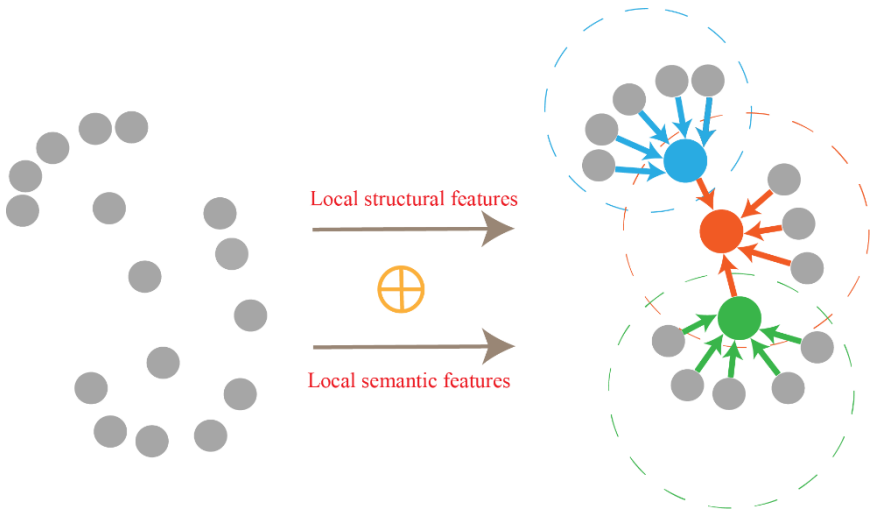

图 2 局部特征聚合图

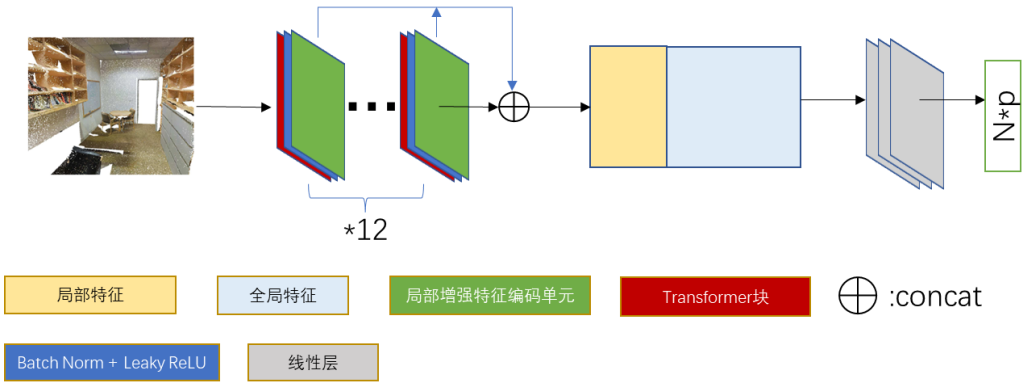

图 3 改进的图像编码器的结构

4. 实验设置

为了进行有效的评估，我们遵循标准的六折交叉验证的方法，即在每个训练中选择五个区域作为训练集，一个区域作为测试集，使用交叉验证的方法建立六个模型来覆盖完成的数据集。对于语义分割，我们采用整体精度(oAcc)、平均精度(mAcc)和平均交并比(mIou)来评估我们的方法，对于实例分割，我们采用平均实例精度指标mAP_{50}、mPrec 和平均召回率 mRec。

对于 S3DIS[40]和 ScanNet V2 数据集，我们采用与 RandLA-Net[15]中相同的预处理设置，首先将每个房间通过网格采样的方式减少输入点云的个数，网格的大小为 0.04 米。然后将整个点云输入到网络中进行训练和测试，实验采用随机采样的方法，因为这将会大大提高点云处理的计算效率。在模型训练过程中，学习率为 1e-5，batch size 为 8，epochs 为 100。

5. 实验结果与分析

5.1 语义分割

表 1 显示了我们的方法在 S3DIS 数据集上语义分割的定量结果。从表 1 可以看出，在 6 折交叉验证实验中，相比于最经典的方法 Pointnet[11]，我们的方法的 mAcc、oAcc 和 mIoU 分别提高了 33.5%、9.8%和 29.5%，方法的三项指标分别提高了 34.1%、10.2%和 30.1%。同时，我们的方法的三项指标值均要优于现有较好的方法 KVGCN[32]、JSNet[36]和 RandLA-Net[15]，但 mAcc 和 mIoU 要略低于 FG-Net[39]。通过方法中的特征融合分割，我们达到了最好的水平。同时，从表 2 可以看出，方法提高了 6 个类别的 IoU 结果。以上定量研究结果表明，网络有利于更有效的局部特征聚合，并且特征融合分割网络促进了语义更精确的分割。

表 1 S3DIS 上的语义分割结果

Method	mAcc	oAcc	mIoU

Pointnet[11]	49.0	78.5	41.1
KVGCN[32]	72.3	87.4	60.9
JSNet[36]	71.7	88.7	61.7
RandLA-Net[15]	82.0	88.0	70.0
FG-Net[39]	82.9	88.2	70.8
Transformer[2]	82.5	88.3	70.6
Ours	**83.1**	**88.7**	**71.2**

表 2 S3DIS 中每个类别的 IoU 结果

Method	Ceiling	Floor	Wall	Beam	Column	Window	Door	Chair	Table	Bookcase	Sofa	Board	Clutter
Pointnet[11]	88.6	97.3	69.8	0.05	3.92	46.3	10.8	52.6	58.9	40.3	5.9	26.4	33.2
RSNet[17]	92.5	92.8	78.6	32.8	34.4	51.6	68.1	60.1	59.7	50.2	16.4	44.9	52.0
KVGCN[32]	**94.5**	94.1	79.5	53.4	36.3	56.8	63.2	67.5	64.3	23.6	54.3	43.1	53.2
RandLA-Net[15]	93.1	96.1	80.6	62.4	48.0	64.4	69.4	76.4	69.4	64.2	60.0	65.9	60.1
KPConv[22]	93.6	92.4	**83.1**	**63.9**	**54.3**	66.1	**76.6**	57.8	64.0	**69.3**	**74.9**	61.3	60.3
Transformer[3]	93.2	96.8	80.0	57.3	46.0	65.5	69.0	83.3	70.6	62.4	64.6	67.4	61.2
Ours	94.0	**97.4**	82.5	57.7	47.1	**66.3**	71.1	**83.6**	**71.2**	63.7	65.5	**68.1**	**61.5**

为了呈现出更直观的结果，我们将 S3DIS 中 area5 区域作为测试集，其他区域作为训练集。我们对比了与原始标签的效果，并选择了三个样本来可视化语义分割的结果，如图 4 所示,我们的方法较好地分割出了点云中的大部分区域(例如第一个示例中的门边缘、后两个示例中的黑板和沙发等区域，白色方框表示 Ours 分割良好的区域)，说明 Ours 有效地增强了局部特征的表达。Ours 通过特征融合分割的方法，将实例特征融合进了语义特征空间，得到具有实例感知的语义特征，实现了更精确的语义分割(例如沙发的角落，凳子的边缘等。黑色方框表示 Ours 良好分割的部分)。

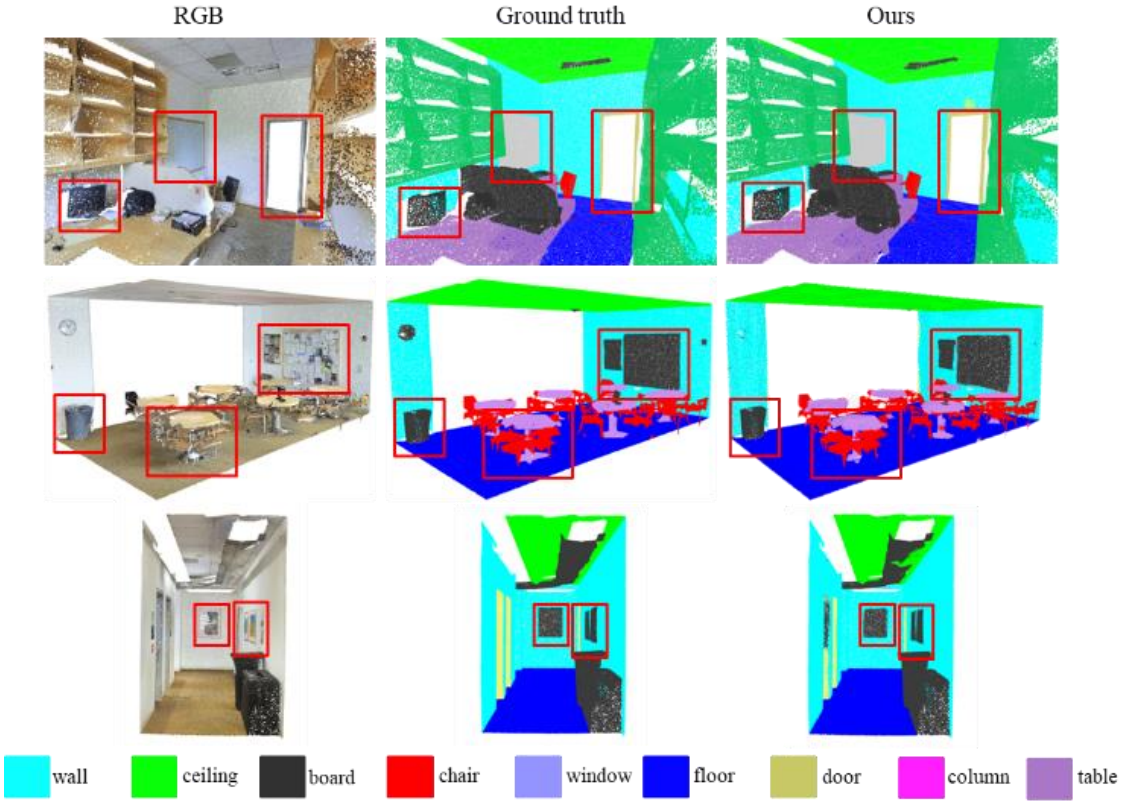

图 4 语义分割效果

5.2 实例分割

点云实例分割的定性结果如图 5 所示。我们的模型能够预测非常接近实际情况的实例分割结果。值得注意的是，我们的模型可以捕获详细的结构信息，并为具有挑战性的场景预测正确的实例语义。例如，在有马桶的场景中，Ours 能够清晰地预测马桶的边缘。

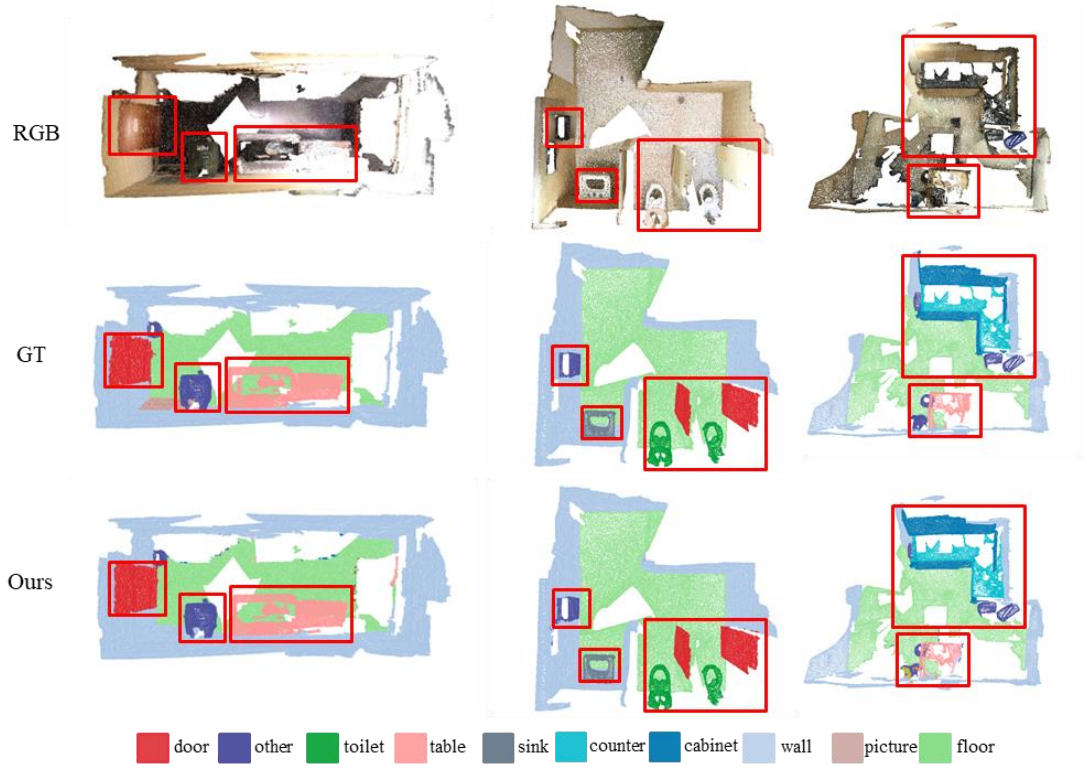

图 5 实例分割效果

此外，Ours 直接将点作为输入也优于那些可以方便地使用 3D 卷积的体素输入方法[34,36,38]。我们的 Ours 在浴室场景中表现良好。同样值得注意的是，Ours 可以很好地处理场景中的许多对象。

表 3 ScanNet V2 上的实例分割结果

Method	mAP_{50}	mPrec	mRec
SGPN[33]	37.9	38.2	31.2
ASIS[35]	51.2	63.6	47.5
3D-BoNet[34]	-	65.6	47.6
GICN[38]	-	68.5	50.8
JSNet[36]	54.1	66.9	53.9
Ours	**58.3**	**68.8**	**55.9**

6. 总结

本文围绕视觉通用分割大模型 SAM 的网络结构与工作机制进行了系统分析，并探讨了其在三维点云室内场景语义分割与实例分割任务中的迁移与应用潜

力。通过对图像编码器、提示编码器与掩码解码器三大核心模块的深入剖析，可以看出，SAM 依托 Vision Transformer 强大的全局建模能力与提示驱动的条件分割范式，有效兼顾了局部细节表达与全局结构一致性，为高质量分割结果提供了坚实基础。进一步地，将 SAM 在二维图像中学习到的丰富几何与语义先验引入三维点云分割任务，有助于缓解传统点云方法在实例边界模糊、全局语义不稳定等方面的不足。总体而言，SAM 为构建跨模态、可泛化的三维场景分割框架提供了一种新的技术路径，也为后续开展基于通用视觉大模型的三维感知研究奠定了重要基础。

7. 参考文献

[1] Kirillov A, Mintun E, Ravi N, et al. Segment anything[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 4015-4026.

[2] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R.R. Martin, S.-M. Hu, PCT: Point cloud transformer, Comput. Vis. Media 7 (2021) 187–199. https://doi.org/10.1007/s41095-021-0229-5.

[3] H. Zhao, L. Jiang, J. Jia, P.H.S. Torr, V. Koltun, Point Transformer, in: 2021: pp. 16259–16268.

[4] Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In the Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 945–953.

[5] Z. Yang and L. Wang, “Learning relationships for multi-view 3D object recognition,” in ICCV, 2019

[6] X. Wei, R. Yu, and J. Sun, “View-gcn: View-based graph convolutional network for 3D shape analysis,” in CVPR, 2020.

[7] Qi, C.R.; Su, H.; Niessner, M.; Dai, A.; Yan, M.; Guibas, L.J. Volumetric and Multi-view CNNs for Object Classification on 3D Data. In the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5648–5656.

[8] Shuran Song, Fisher Y u, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In CVPR, 2017. 1, 2

[9] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In IROS, 2015. 1, 2, 6

[10] Klokov , R.; Lempitsky , V . Escape from cells: Deep kd-networks for the recognition of 3d point cloud models. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), V enice, Italy , 22–29 October 2017; pp. 863–872.

[11] Qi, C.R.; Su, H.; Mo, K.; Guibas, L.J. PointNet: Deep learning on point sets for 3d classification and segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp.

652–660.
[12] Qi, C.R.; Yi, L.; Su, H.; Guibas, L.J. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. arXiv 2017, arXiv:1706.02413.
[13] M. Jiang, Y . Wu, and C. Lu, “PointSIFT: A sift-like network module for 3D point cloud semantic segmentation,” arXiv preprint arXiv:1807.00652, 2018.
[14] F. Engelmann, T. Kontogianni, J. Schult, and B. Leibe, “Know what your neighbors do: 3D semantic segmentation of point clouds,” in ECCVW, 2018.
[15] Qingyong Hu, Bo Yang, Linhai Xie, Stefano Rosa, Yulan Guo, Zhihua Wang, Niki Trigoni, and Andrew Markham. Randla-net: Efficient semantic segmentation of large-scale point clouds. In CVPR, 2020. 2
[16] C. Zhao, W. Zhou, L. Lu, and Q. Zhao, “Pooling scores of neighboring points for improved 3D point cloud segmentation,” in ICIP, 2019.
[17] Q. Huang, W. Wang, and U. Neumann, “Recurrent slice networks for 3D segmentation of point clouds,” in CVPR, 2018.
[18] X. Ye, J. Li, H. Huang, L. Du, and X. Zhang, “3D recurrent neural networks with context fusion for point cloud semantic segmentation,” in ECCV, 2018.
[19] F. Liu, S. Li, L. Zhang, C. Zhou, R. Ye, Y . Wang, and J. Lu,“3DCNN-DQN-RNN: A deep reinforcement learning framework for semantic parsing of large-scale 3D point clouds,” in ICCV,
2017.
[20] B.-S. Hua, M.-K. Tran, and S.-K. Yeung, “Pointwise convolutional neural networks,” in CVPR, 2018.
[21] S. Wang, S. Suo, W.-C. Ma, A. Pokrovsky , and R. Urtasun, “Deep parametric continuous convolutional neural networks,” in CVPR, 2018.
[22] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, “KPConv: Flexible and deformable convolution for point clouds,” in ICCV, 2019.
[23] F. Engelmann, T. Kontogianni, and B. Leibe, “Dilated point convolutions: On the receptive field of point convolutions,” in ICRA, 2020.
[24] L. Landrieu and M. Simonovsky , “Large-scale point cloud semantic segmentation with superpoint graphs,” in CVPR, 2018.
[25] L. Landrieu and M. Boussaha, “Point cloud oversegmentation with graph-structured deep metric learning,” in CVPR, 2019.
[26] Xie, Z.Y .; Chen, J.Z.; Peng, B. Point clouds learning with attention-based graph convolution networks. Neurocomputing 2020, 402, 245–255.
[27] Xu, M.X.; Dai, W.R.; Shen, Y .M.; Xiong, H.K. MSGCNN: Multi-scale Graph Convolutional Neural Network for Point Cloud Segmentation. In Proceedings of the Fifth IEEE International Conference on Multimedia Big Data, Singapore, 11–13 September 2019; pp. 118–127.
[28] FPConv: Learning Local Flattening for Point Convolution
[29] Wang, Y .; Sun, Y .B.; Liu, Z.W.; Sarma S.E; Bronstein M.M; Solomon J.M. Dynamic Graph CNN for Learning on Point Clouds. ACM T rans. Graph (TOG) 2019, 38, 146.
[30] Hu, Z.; Zhen, M.; Bai, X.; Fu, H.; Tai, C.L. JSENet: Joint semantic segmentation and edge detection network for 3d point clouds. In Proceedings of the European

Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.
[31] Zhao, H.; Jiang, L.; Jia, J.; Torr, P .; Koltun, V . Point Transformer. Avaliable online: https://arxiv .org/pdf/2012.09164v1.pdf (accessed on 2 March 2021).
[32] KVGCN: A KNN Searching and VLAD Combined Graph Convolutional Network for Point Cloud Segmentation.
[33] Wang, W.; Y u, R.; Huang, Q.; and Neumann, U. 2018a. Sgpn: Similarity group proposal network for 3d point cloud instance segmentation. In CVPR.
[34] Y ang, B.; Wang, J.; Clark, R.; Hu, Q.; Wang, S.; Markham, A.; and Trigoni, N. 2019. Learning object bounding boxes for 3d instance segmentation on point clouds. arXiv preprint arXiv:1906.01140.
[35] Wang, X.; Liu, S.; Shen, X.; Shen, C.; and Jia, J. 2019b. Associatively segmenting instances and semantics in point clouds. In CVPR.
[36] Zhao and W. Tao, "JSNet: Joint instance and semantic segmentation of 3D point clouds," in AAAI, 2020.
[37] Bo Yang, Sen Wang, Andrew Markham, and Niki Trigoni. Robust attentional aggregation of deep feature sets for multi-view 3D reconstruction. IJCV, 2019.
[38] Liu S H , Yu S Y , Wu S C , et al. Learning Gaussian Instance Segmentation in Point Clouds[J]. 2020.
[39] Liu K , Gao Z , Lin F , et al. FG-Net: Fast Large-Scale LiDAR Point Clouds Understanding Network Leveraging Correlated Feature Mining and Geometric-Aware Modelling. 2020.
[40] Semantic Segmentation on S3DIS. Avaliable online: https://paperswithcode.com/sota/semantic-segmentation-on-s3dis