EmbodiedSAM:在线分割任何 3D 物体即时的

Xiuwei Xu1 , Huangxing Chen1 , Linqing Zhao1 ，Ziwei Wang2， Jie Zhou1，Jiwen Lu1
1清华大学， 2南洋理工大学

抽象的

具身任务要求代理在探索三维场景的同时充分理解三维场景,因此迫切需要一个在线、实时、细粒度和高度泛化的三维感知模型。由于高质量的三维数据有限,直接在三维环境中训练这样的模型几乎是不可行的。同时,视觉基础模型 (VFM) 以卓越的性能彻底改变了二维计算机视觉领域,这使得利用 VFM 辅助具身三维感知成为一个有希望的方向。然而,大多数现有的 VFM 辅助三维感知方法要么是离线的,要么太慢,无法应用于实际的具身任务。在本文中,我们旨在利用任意分割模型 (SAM)在线设置中进行实时三维实例分割。这是一个具有挑战性的问题,因为在输入的流 RGB-D 视频中没有未来的帧,并且一个实例可能会在几帧中观察到,因此需要在帧之间进行对象匹配。为了应对这些挑战,我们首先提出了一个几何感知查询提升模块,用 3D 感知查询表示 SAM 生成的 2D 掩码,然后通过双层查询解码器对其进行迭代细化。这样， 2D 掩码就被转移到 3D 点云上的细粒度形状。利用3D 掩码的查询表示,我们可以通过高效的矩阵运算计算不同视角的 3D 掩码之间的相似度矩阵,从而实现实时推理。在 ScanNet、ScanNet200、SceneNN 和3RScan 上的实验表明,我们的方法即使与离线方法相比也实现了领先的性能。我们的方法还在几个零样本数据集传输实验中表现出了很强的泛化能力,在开放词汇和数据高效设置中展现出巨大潜力。代码和演示可在此处获取,训练和评估只需要一个RTX 3090 GPU。

1 简介

具身任务,例如机器人操作和导航[20; 3; 36; 35],需要代理理解 3D 场景、推理人类指令并通过自我行动做出决策。

在流程中,体现视觉感知是各种下游任务的基础。
在具体场景中,我们希望三维感知模型具备以下特点：(1)在线性,输入数据是流式的RGB-D视频,而非预先采集的视频,视觉感知与数据采集同步进行；(2)实时性,需要较高的推理速度；(3)细粒度。

它应该能够识别场景中出现的几乎所有物体；(4)高度通用。一个模型可以应用于不同类型的场景,并与不同的传感器参数（如相机固有参数)兼容。由于高质量 3D 数据有限,在纯 3D 中训练这样的模型几乎是不可行的。

受大型语言模型 (LLM) [38; 4; 1] 巨大成就的启发,一系列视觉基础模型 (VFM) 如 SAM [12]和 SEEM [40]应运而生。VFM凭借其细粒度、精确和可泛化的分割能力,正在彻底改变二维计算机视觉领域

通讯作者。

预印本。正在进行中。

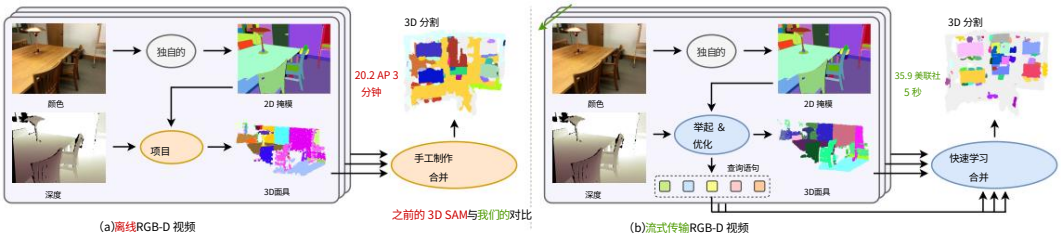

图 1:与以前的 3D SAM 方法[33； 31； 34]将 2D 掩码投影到 3D 并与手工制作的策略合并不同,ESAM 将 2D 掩码提升到 3D 查询并迭代细化它们以预测准确的 3D 掩码。借助 3D 查询,ESAM 还能够通过简单的矩阵运算快速合并不同帧中的 3D 掩码。以 SAM3D [33]为例,我们的 ESAM性能比其高出 15.7% AP,速度提高了 20 倍以上。

基于图像像素的 VFM 研究尚不成熟,但针对 3D 领域的 VFM 开发研究较少。
由于与二维数据相比,高质量带注释的三维数据少得多,因此探索现有二维 VFM 的适应或扩展以实现体现三维感知具有很大的前景。

最近,有一些研究[33; 34; 17]采用 SAM 在3D 场景的多视角图像上自动生成蒙版,并通过投影和迭代合并在 3D 中合并蒙版。
虽然这些方法实现了具有高泛化能力的细粒度 3D 实例分割,但它们仍然存在一些阻碍其应用的严重问题：(1)它们将 SAM 应用于单个图像,并直接将 2D 掩模投影到具有相机参数的 3D 点云。因此,预测不具有几何感知能力,这可能会在不同视图中产生不一致的结果；(2)它们使用手工制作的策略在 3D 中合并每帧掩模预测。例如,计算所有掩模对之间的几何相似性并根据阈值合并它们,这种方法不准确且非常慢；(3)它们中的大多数是基于预先收集的 RGB-D 帧并进行 3D 重建的离线方法。

在本文中,我们提出了一个 VFM 辅助的 3D 实例分割框架,即 Embodied- SAM (ESAM),它利用 SAM 的强大功能对 3D 场景中的任何内容进行在线分割,具有高精度、快速性和强泛化能力。如图 1 所示,与以前的3D SAM 方法[33; 31; 34]将 2D 掩码投影到 3D 并与手工制作的策略合并不同,ESAM 将 2D 掩码提升到 3D 查询,并通过迭代查询细化预测时间和几何一致的 3D掩码。受益于 3D 查询表示,ESAM 还能够通过简单的矩阵运算快速合并不同帧中的 3D 掩码。具体而言,我们从深度图像投影的点云中提取点状特征。然后,我们将SAM 生成的 2D 掩码视为超点,用于指导我们提出的几何感知池化模块进行掩码聚合,从而生成与 SAM 掩码一一对应的 3D 查询。我们进一步提出了一个双层查询解码器来迭代细化 3D 查询,这使得查询能够有效地利用超点特征并生成细粒度的点状掩码。由于每个 3D 实例掩码都与一个查询相关联,我们可以通过并行高效的矩阵乘法计算新预测的 3D 掩码与之前的掩码之间的相似性并准确地合并它们。为了增强查询特征的判别能力,我们设计了三个代表性的辅助任务来估计几何、对比和语义相似性。我们在 ScanNet、ScanNet200、SceneNN 和 3RScan 数据集上进行了广泛的实验。与之前的 VFM 辅助 3D 实例分割方法相比,我们在保持强大泛化能力的同时,大幅提高了准确率和速度。此外,ESAM 可以轻松扩展到开放词汇分割。当使用有限的数据进行训练时,它在数据高效的设置中也显示出巨大的潜力。

2 相关工作

VFM 辅助 3D 场景分割:在 2D 领域,视觉基础模型 (VFM) [23; 12; 15]呈爆炸式增长。得益于大量带注释的视觉数据,2D VFM 表现出极高的准确性和极强的泛化能力,这使得它们在零样本场景中表现良好。由于 3D 视觉领域的高质量带注释数据比 2D 视觉领域少得多,因此使用 2D VFM 辅助 3D 场景感知成为一个有前途的方向[25; 33;

31; 34]。UnScene3D [25]考虑来自 DINO [23]的 2D 自监督特征来生成初始伪掩码,然后通过自训练迭代细化。SAM3D [33]采用 SAM [12]生成2D 实例掩码,然后通过深度和相机参数将其投影到 3D 空间,并根据几何形状进行合并。SAMPro3D [31]将场景中的 3D 点映射到多视角 2D图像作为 3D 提示,用于对齐 SAM 生成的 2D 掩码并将 3D点聚类为实例掩码。SAI3D [34]在重建的 3D 网格上生成 3D 图元。然后采用语义 SAM 获取具有语义分数的 2D 掩码,将其与 3D图元连接并通过基于图的区域增长策略合并。我们的方法还利用 SAM来辅助 3D 实例分割。不同的是,我们使 2D 到 3D 投影和 3D 掩模合并的过程可学习且可在线进行。这样,我们的 ESAM 能够预测更准确的 3D掩模并应用于实际的实时在线任务。

在线 3D 场景感知:为了实现具身 AI,机器人导航[3; 36]和操控[20]等现实世界应用受到越来越多的关注。在线 3D 场景感知从流式 RGB-D 视频中精确理解周围的 3D 场景,成为这些机器人任务的视觉基础。早期的在线 3D 感知方法分别处理 2D 图像并将预测投影到 3D 点云,然后进行融合步骤以合并来自不同帧的预测[18; 21]。然而,2D 图像上的预测不具有几何和时间感知能力,这使得融合步骤困难且不准确。融合感知3D-Conv [37]和 SVCNN [11]构建数据结构以维护先前帧的信息并进行基于点的 3D 聚合以融合 3D 特征以进行语义分割。

INS-CONV [16]将稀疏卷积[9; 5] 扩展到增量 CNN,以有效提取全局3D 特征,用于语义和实例分割。为了简化在线 3D 感知模型的设计并利用先进的离线 3D 架构的强大功能,MemAda [32]提出了一种新的在线 3D 场景感知范例,通过基于内存的多模态适配器为离线模型提供在线感知能力。与以前的研究不同,我们的ESAM 将 SAM 生成的 2D 掩码提升为精确的 3D 掩码和相应的查询,这使我们能够高效地合并高精度的每帧预测。

3 方法

给定一个具有已知姿势的 RGB-D 图像序列Xt = {x1, x2, ..., xt} ,我们的目标是分割相应 3D 场景中的任何实例。正式地， xt = (It, Pt) ,其中It是彩色图像， Pt是通过将深度图像投影到具有姿势参数的 3D 空间获得的点云。我们的方法是预测观察到的 3D 场景St = Pi的实例掩码。此外,我们希望在线解决这个问题;也就是说,在任何时刻,未来帧xi i > t都是未知的,并且应该在每个时刻预测St的时间一致的 3D 实例掩码。

的=1

,

概述。我们的方法概述如图 2 所示。我们以增量方式解决在线 3D 实例分割问题,以实现实时处理。在时刻,我们仅预测当前帧Pt的实例掩码Mcur 。然后我们将Mcur合并到St-1的前一个实例掩码Mpre ,并得到St的更新实例掩码Mpre 。

吨 吨

t−1 吨

3.1 查询提升与细化

假设模型正在接收第 t 个 RGB-D 帧xt = (It, Pt),我们首先采用 SAM 自动掩码生成从It中获取 2D 实例掩码M2d 。在本节中,我们忽略下标 t 以便更清楚地表述。

吨

几何感知查询提升。由于 SAM 不利用先前帧的信息,也不利用深度图像中的 3D 信息,因此直接将M2d投影到P会导致不准确且时间不一致的 3D 掩码。相反,我们的目标是将每个 2D 掩码提升为 3D 查询特征,这使我们能够进一步细化查询以生成 3D 实例掩码。由于 2D二元掩码信息量较少,我们建议从场景中提取点云特征,然后将 2D 掩码视为索引,将点云聚类为超点,这样就可以简单地从超点特征中选择查询。假设点云P ∈ R， M2d中有M 个掩码,我们首先根据颜色深度对应关系将M2d映射到P ,以获得超点,其中S中的每个元素都落在[0, M) 中。然后我们将P馈送到 3D 稀疏 U-Net [5]

×3

指数S ∈ Z N ,

,

使用基于内存的适配器[32]提取时间感知的3D特征FP∈R 密度×碳 . 有了FP和

S,我们可以将逐点特征池化为超点特征FS∈RM × C 。

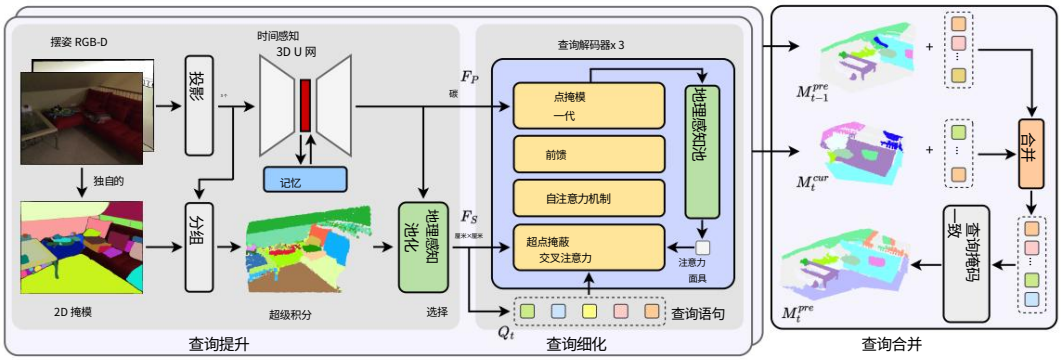

图 2:ESAM 概览。在新的时刻,我们首先采用 SAM 生成 2D 实例掩码M2d t 。我们提出了一个几何感知查询提升模块,将M2d提升为 3D 查询Qt ,同时保留细粒度形状信息。Qt由双层解码器细化,从而实现高效的交叉注意并从Qt 生成细粒度的逐点掩码Mcur 。然后通过快速查询合并策略将Mcur合并到先前的掩码Mpre中。

然而,诸如最大或平均池化之类的简单操作可能会降低FS的表示能力。为了更好地保留每个超点内的点特征,我们取⊆ P 的几何形状,i ∈ [0, M),我们考虑每个超点的归一化计算。对于超点P ,所有点pj ∈ P相对于超点中心ci的相对位置p 。

这样， j | pj ∈ P i}表示该

定Pi = {p rj max(pj) − min(pj)直径为 1 且中心为原点。然后我们

计算每个点的局部和全局特征：

$$z_{\text{全局}} = \text{Agg}(z_{\text{局部}}) \in R^{C}, \quad z_{\text{局部}} = \text{MLP}(Pi) \in R^{|Pi| \times C} \text{其中MLP} \tag{1}$$

在每个单独的点上执行， Agg是用通道最大池化实现的聚合函数。局部和全局特征表示点和形状之间的相关性,因此我们将两个特征连接起来并将它们提供给另一个 MLP 来预测逐点权重:z global]

$$wj = \text{Sigmoid}(\text{MLP}(zj)) \in R^{(0,1)}, \; zj = [z_{\text{当地的}}, \tag{2}$$

最后,我们聚合点特征 F 使用加权平均池化进入第 i 个超点：

$$F_{\text{等代}} = G(F \hat{P}) + z_{\text{全球的}}, \; G(F \hat{P}) = \text{平均值}(F_{\text{磷}} [w1, ..., w|Pi|]) \text{全} \tag{3}$$

特征增强了池化超点特征。每个超点的计算可以与逐点 MLP 和散射 局充分结合形状级几何注意我们用z特征和场景级 3D U-Net 函数 [8] 并行化,因此这种几何感知池化实际上是有效的。

双层查询解码器。池化后， M 个2D 实例掩码M2d被提升为 3D 超点特征FS。然后,我们从FS初始化一系列 3D 实例查询Q0 ,这些查询由多个基于转换器的查询解码器层迭代细化,并用于预测 3D 掩码。在训练期间,我们随机抽取FS的 0.5 到 1 之间的比例来构建Q0以进行数据增强。而在推理时,我们只需设置Q0 = FS。

每个查询解码器在查询和场景表示之间采用掩蔽交叉注意力来聚合每个查询的实例信息：

$$Q^{l} = \text{Softmax}(Q \cdot K^{T} / \sqrt{\text{中}} + \text{铝}) \cdot V, \quad Al(i,j) = \begin{cases} 0 & \text{如果Mcur l,则 (i, j) = True} \\ \text{否则} -\infty & \end{cases}, \; l=0,1,2 \tag{4}$$

其中·表示矩阵乘法， Q是Ql的线性投影， K和V是场景表示F的线性投影。F可以是点特征FP或超点特征FS。Al是从第 l 个解码器层中预测的 3D 实例掩码Mcur得出的注意掩码。 (i, j)表示第 i 个查询关注第 j 个点或超点。然后我们将Q l馈送到自注意层和前馈网络以得到Ql+1,接着由掩码生成模块来预测每个查询的实例掩码：

$$\text{麦克尔} = \text{Sigmoid}(\text{(Ql)} \cdot F^{l}) > \phi, \; l = 0, 1, 2, 3 \tag{5}$$

其中 ϕ 是线性层。如果 F = FP ,则Mcur是点掩码,否则是超点掩码。

查询解码器的一种常见做法[27; 28; 14]是采用相同级别的场景表示进行交叉注意和掩码生成。然而,由于 SAM 已经输出了高级语义感知掩码,我们观察到M ≪ N。如果我们对查询解码器采用点式场景表示FP ,则由于点的数量太多,交叉注意操作将非常耗费内存。而如果我们使用超点特征FS,预测的 3D 实例掩码将只是超点的组合,因此无法细化到更细的粒度。为了兼顾两全其美,我们的查询解码器被设计为双层的。对于等式 (4) 中的交叉注意,我们设置F = FS以实现高效的交互。而对于等式 (5) 中的掩码预测,我们设置F = FP以实现细粒度的掩码生成。为了支持掩蔽注意力,我们在等式 (4) 之前将点掩码池化为超点掩码： Mcur ← G(Mcur) > ϕ (6)其中G是等式 (3) 中的几何感知池化。我们可以重用等式 (2) 中预先计算的权重来减少计算量。这样,经过3×查询解码器后,我们获得了准确的点掩码Mcur以及相应的查询Q3,在以下小节中将其表示为Mcur和Qt 。

作为

我们对Mcur执行 mask-NMS以过滤掉冗余的 mask 以及对应的查询。

3.2 高效的在线查询合并

将二维掩码M2d提升为精确的三维掩码Mcur后,我们将Mcur合并到前一个实例掩码Mpre中以获得Mpre 。注意当t = 1时,我们有Mpre = Mcur作为初始化。然而,以前的工作中合并实例掩码的主流解决方案[33; 34; 16; 21; 17]是遍历Mcur中的所有掩码,并将Mcur中的每个掩码与Mpre中所有先前的掩码进行比较。这个过程非常慢,因为为了准确地决定是否应该将新的掩码合并到前一个掩码中,需要在两个掩码的点云上计算几何相似性,例如掩码IoU或CD距离。相似度的计算涉及每个掩码中的所有点,计算复杂度很高。更糟糕的是,上述操作很难并行计算,因为每个掩码中的点数不同,我们需要根据掩码逐个挑选出每个实例的点云。为此,我们建议用固定大小的向量表示每个掩码,并使用有效的矩阵运算来计算相似度。

于我们的架构,对于Mcur中的每个掩码,我们都有相应的查询特征。查询特征本身是一个固定大小的向量表示,但简单地计算它们之间的相似性信息量较少,性能很低。因此,我们根据查询特征设置了几个有代表性的辅助任务,以学习不同度量下的向量表示,用于计算几何、对比和语义相似性。

首先,对于几何相似性,我们观察到模型能够仅通过部分观察来学习整个几何形状。然而,由于分割的限制,预测只能在现有点上进行,模型无法表达其对整个几何形状的知识。因此,我们通过引入边界框预测的辅助任务使模型能够表达其全部知识。我们采用 MLP根据每个查询的中心（即中心ci)预测边界框回归

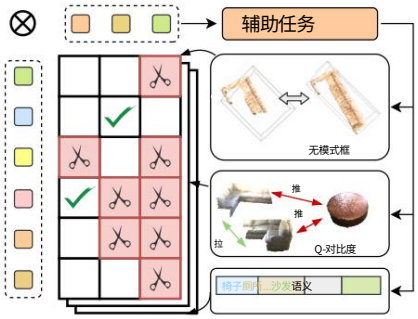

图 3:我们高效查询合并策略的细节。我们提出了三种代表性辅助任务,它们以向量的形式生成几何、对比和语义表示。然后可以通过矩阵乘法高效地计算相似度矩阵。我们进一步修剪相似度矩阵并采用二分匹配来合并实例。

得到框B∈R,然后可以通过两个框之间的IoU来计算两个掩码之间的几何相似度。

我们忽略了框的方向,因为两组轴对齐的边界框之间的IoU 矩阵可以通过简单的矩阵运算来计算。

其次,对于对比相似性,我们的目标是学习一个特定于实例的表示,其中来自同一实例的特征应该被拉到一起,否则就被推开。这种表示可以通过在两个相邻帧之间进行对比训练来学习:我们使用 MLP 将查询特征Qt映射到对比特征ft。然后对于出现在第 t 和第(t + 1) 帧中的实例 ,我们选择

t , fit+1)作为位置对,从其他实例中采样该实例的两个特征(f(f t ,f t+1)作为负对。详细的损失函数将在下一小节中展示。

最后,对于语义相似度,我们只需采用 MLP 来预测每个类别的概率分布S ∈ R K,其中K是预定义类别的数量。此任务还有其他选择。
例如,如果我们采用语义 SAM [15]而不是 SAM,我们可以直接利用2D 掩码的语义预测作为相应查询的 S。

通过这种方式,可以利用Mpre和Mcur对应的几何、对比和语义表示高效地计算出它们之间的相似度矩阵 C：

$$C = \mathrm{IoU}(B \quad + \quad \cdot \quad () \quad + \quad \cdot \quad () \tag{7}$$

其中IoU(·, ·)表示两组轴对齐边界框之间的 IoU 矩阵。我们通过将小于阈值 的元素设置为−∞来修剪C。然后对Mpre和Mcur执行成本为−C 的二分匹配。如果新掩码无法与任何先前的掩码匹配,我们将为该掩码注册一个新实例。否则,我们将合并这两个掩码以及它们的B、 f和S。掩码合并可以通过取并集简单地实现。 ，个掩码 分配给Mpre中的一个掩码,该掩码将分配给Mcur中的每 。

对于其他表示,我们通过以下方式进行加权平均： B [一] = [] [] 等等。我们假设第j个新掩码与第i个前一个掩码合并。n是合并的次数,表示已经合并到Mpre的掩码数量 [我]。

3.3 损失函数

我们在每个 RGB-D 帧上都有语义和实例标签。在每个 RGB-D 视频中,不同帧的实例标签是一致的。给定注释,我们根据每个查询的预测计算每帧损失。由于查询Qt是以一对一的方式从 2D SAM 掩码中提取的,我们忽略了复杂的标签分配步骤,直接利用 2D 掩码上的注释来监督相应查询的预测。我们假设 2D SAM 掩码只能属于一个实例,因此我们可以为每个查询获取真实语义标签和 2D 实例掩码。我们利用与深度图像的像素对应关系将 2D 实例掩码映射到 3D 点云,并根据 3D 实例掩码计算真实轴对齐边界框。利用上述注释,我们计算二分类损失L和交叉熵来区分前景和背景实例。预测的 3D 掩码由二元交叉熵L监督边界框和语义预测的损失

loss Liou 以及 Dice 损失L 和二 般子。 元交叉熵 L 分别。

除了上述每帧损失之外,我们还制定了相邻帧之间的对比损失：

$$\text{连续} \quad = - \quad \frac{1}{\text{当}} \quad \text{日志} \quad \frac{\quad (\quad /\tau)}{\quad (\quad /\tau) + e} \tag{8}$$

其中 ·, · 是余弦相似度。因此最终总损失公式为：

$$L = \quad \frac{1}{} \quad (\alpha L \quad + \quad + \quad + \beta \quad + \quad + \quad + \quad) \tag{9}$$

其中Lcont 和Lcont 设置为0。

4 实验

在本节中,我们首先描述我们的数据集和实现细节。然后,我们将我们的方法与最先进的 VFM 辅助 3D 实例分割方法和在线 3D 分割方法进行比较,以验证其有效性。我们还在开放词汇和数据高效的环境中应用 ESAM 来展示其应用潜力。最后,我们进行消融研究,对我们的设计进行全面分析。