arXiv:2603.01205v1 [cs.CV] 1 Mar 2026

CoSMo3D：通过大语言模型引导的规范空间建模实现开放世界可提示三维语义部件分割

Li Jin[1]*、Weikai Chen[2]*、Yujie Wang[3]†、Yingda Yin[2],Zeyu HU[2],Runze Zhang[2]Keyang Luo[2], Shengju Qian[2],Xin Wang[2],Xueying Qin[1]†

[1]SDU [2]LIGHTSPEED [3]北卡罗来纳大学教堂山分校

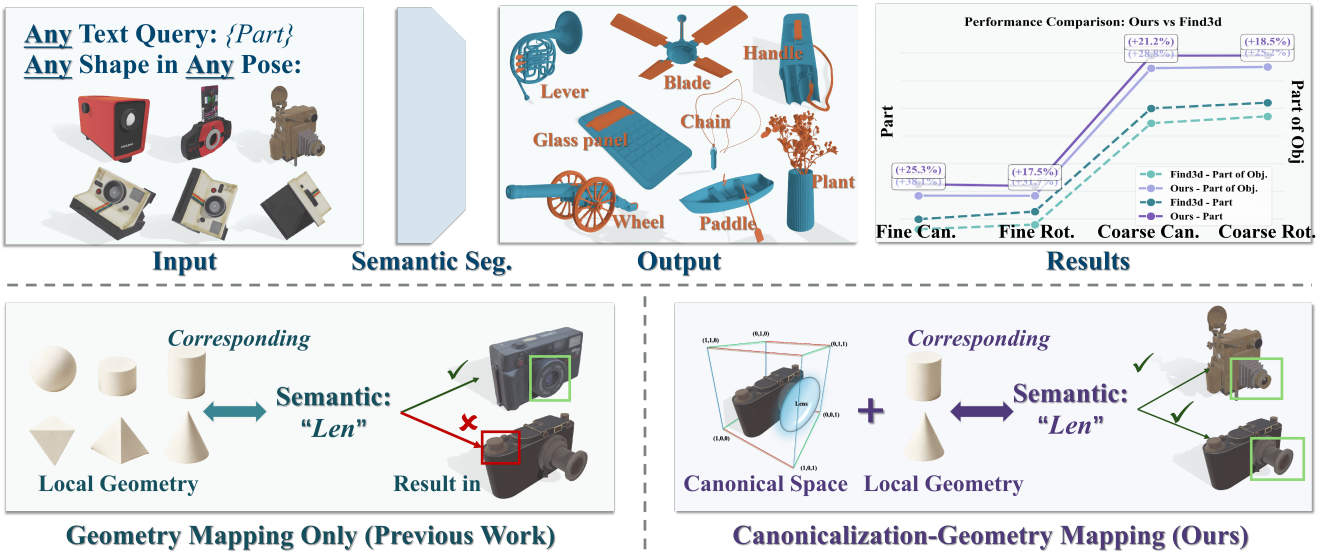


图1. 我们提出了 CoSMo3D，这是一种开放世界可提示三维语义分割方法。该方法引入规范空间感知，以突破任意姿态和形状带来的限制，在多种设置下达到最先进的性能，并显著优于仅几何映射方法。

摘要

开放世界可提示三维语义分割由于在输入传感器坐标系中推断语义，仍然十分脆弱。然而，与之相反，人类会在规范空间中依据功能作用来理解部件——翅膀向两侧伸展，把手向侧面突出，而腿部则从下方提供支撑。心理物理学证据表明，我们会在心理上将物体旋转至规范坐标系，以揭示这些作用。为弥补这一空缺，我们提出了 *CoSMo3D*，通过直接从数据中学习潜在规范参考坐标系，实现规范空间感知。通过构造，我们利用大语言模型引导的类别内及跨类别对齐，创建统一规范数据集，揭示 *200* 个类别中的规范空间规律。通过归纳，我们实现了模型内部的规范性：采用带有规范映射锚定和规范框校准的双分支架构，将姿态变化和对称性压缩为稳定的规范嵌入。从输入姿态空间转向规范嵌入，使部件语义具有更高的稳定性和可迁移性。实验结果表明，*CoSMo3D* 在开放世界可提示三维分割领域达到了新的最先进水平。项目主页：*h t tps://github.com/JinLi998/CoSMo3D/tree/main*

1. 引言

人类通过综合推理几何形状与空间语义来感知三维物体。当我们识别出“椅子的腿”时，依靠的不仅是其细长形状，还包括其规范位置 ——位于座椅下方并支撑着物体。心理物理学研究 [17] 进一步表明，

*共同贡献。
†通讯作者。

人类会进行心理旋转，将观察到的物体与规范姿态对齐，从而能够在不同姿态和类别中一致地识别部件。这种规范空间推理 是人类三维理解的核心组成部分，但在当前的分割模型中仍基本缺失。

近期，开放世界三维理解领域的进展开始弥合这一差距。特别是，Find3D [13] 提出了开放世界可提示三维分割任务，其中模型根据自由形式的文本查询（例如“把手”“翅膀”或“桨”）对任意三维物体进行分割。这是迈向可扩展三维理解的重要一步：模型不再将分割局限于固定类别，而是能够理解开放式提示，并将知识迁移到未见类别。Find3D 通过学习几何特征与语言嵌入之间的直接对齐来实现这一点，并由大规模自动标注数据集提供支持。它朝着通用三维感知迈出了关键一步，展现出令人印象深刻的零样本泛化能力和灵活的可提示性。

尽管取得了成功，Find3D 及类似方法仍然受到其对几何—文本匹配依赖的根本限制。它们假设几何形状相似的形状往往具有相似的语义，但这种相关性在实践中经常失效。例如，椅子的扶手和腿在几何形状上可能相似，但承担着不同的功能；而飞机机翼和鸟类翅膀的形状不同，却具有相同的语义。如果没有基于规范空间规律的显式推理，这些模型就无法意识到某个语义部件相对于整个物体应出现在哪里，从而在姿态变化、对称性或跨类别情况下产生不一致的预测。例如，在我们的预告示例（图1）中，Find3D 将一个圆柱体误分类。尽管数据增强提供了有限的鲁棒性，这些模型仍然缺乏对空间语义的内在理解，而空间语义是人类感知的核心要素。

我们认为，实现类人的三维理解需要具备规范空间感知能力：能够在不同形状和类别之间内化共享的规范参考空间，并相对于这一参考系来解读部件语义，而非依赖原始输入姿态。为此，我们推出*CoSMo3D*，这是一个大语言模型引导的开放世界可提示三维分割框架，通过**从数据中诱导潜在参考坐标系**实现**规范空间感知** 。这一原则在 CoSMo3D 中体现于两个层面。1）在外部，我们通过大语言模型引导的类别内与跨类别规范化流程构建统一规范数据集，使规范性不再仅限于单个类别，而是在不同物体族之间共享。这为我们提供了一个共同的基础 ，使泛化能力超越类别边界。2）在内部，我们通过直接从数据中诱导潜在规范参考坐标系来实现规范性，而非手动指定规范姿态。在此框架下，我们提出一种双分支架构：主分支保留 Find3D 的几何—语言对齐能力，而规范分支则推动点嵌入靠近规范代码，并通过两个规范空间目标（规范映射锚定和规范框校准）来收紧部件级空间范围。由此，规范空间成为一种涌现的吸引子流形：同一功能部件的不同姿态、对称形态和形状变体都会收敛到单一的规范嵌入，使语义具有空间依据且天然不受姿态影响。

大量实验表明，CoSMo3D 的性能显著优于现有方法。通过在统一框架下整合语言驱动的语义、几何理解和规范空间推理，CoSMo3D 弥合了几何与文本对齐和类人空间感知之间的差距。总体而言，我们的贡献可总结如下：

•我们将开放世界三维分割重新定义为基于规范空间规律的推理，而非在输入姿态下将几何形状与文本进行匹配。

• 我们将规范性构建为可学习的潜在结构，并通过与大语言模型对齐的规范数据集和双分支规范正则化予以实现。

• 在准确率、空间稳定性和跨类别泛化方面，CoSMo3D 将开放世界三维分割的性能提升至最先进水平。

2. 相关工作

传统的数据驱动部件分割 [5, 12, 29, 30]以二维图像或三维形状为输入。其核心在于预先定义物体类别和部件标签，然后进行逐像素或逐点分类。然而，由于高度依赖预定义标签，这类方法的泛化能力有限，无法适应开放世界场景。在开放世界场景中，二维分割技术率先取得突破。其中，以 SAM [8, 16]为代表的类无关分割无需预定义类别，兼顾了分割准确率和泛化能力。SAM3 进一步引入提示语义分割，通过建立开放词汇与部件之间的映射，提升了与人类语义认知的对齐程度和可控性，为三维分割技术的发展提供了参考。

类无关三维分割以三维形状为输入，旨在不预先定义类别的情况下，根据几何特征实现物体的多层级分解。它主要包括两种技术路径：一种是基于二维转换的方法 [19, 25], ，通过渲染将三维形状转换为二维图像，使用 SAM 等二维模型进行分割，再将结果反投影至三维空间，最后缓解诸如

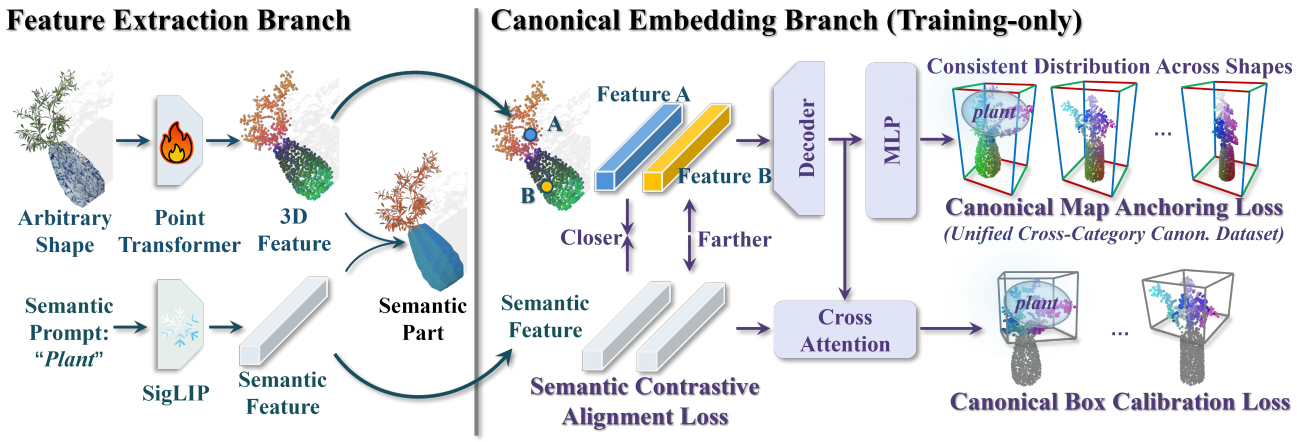


图 2. 我们提出了一种用于开放世界可提示三维语义分割的双分支框架：特征提取分支对三维形状特征（通过 Point Transformer）和文本语义特征（通过 SigLIP）进行编码，以实现跨模态部件分割。随后，仅在训练阶段使用的规范嵌入分支通过语义对比对齐、规范映射锚定和规范框校准损失，确保规范空间感知的一致性，从而能够对任意姿态下的任意形状进行稳健推理。

多视图一致性和自遮挡问题的后处理方法。然而，这一路径依赖转换和后处理，导致速度较慢，且无法避免多视图一致性问题。另一种是纯三维方法 [4, 11, 24, 26, 33]: Sampart3D [26] 通过结合三维特征提取器与测试时自适应（TTA）实现分割，但 TTA 会增加计算成本；Partfield [11] 及其后续改进方法 [4, 24, 33]通过对比学习扩大部件间的特征差异，再通过聚类完成分割。尽管凭借更强大的网络架构和更多数据取得了更好的性能，但这些方法都局限于低层级部件分割，缺乏高层级语义理解，交互性和可控性较差。此外，物体特征在任意姿态下不具备一致性，因此难以建立与语义的映射关系。

可提示三维语义分割以三维形状和语义文本为输入，核心是通过文本条件建立语义与三维部件之间的映射，从而输出目标分割结果。此类方法分为两种：一类是基于二维的方法 [1, 10, 18, 20, 31, 32], ，采用 GLIP [15] 等二维语义分割模型，着重解决将二维结果聚合到三维空间时的一致性问题。然而，由于依赖二维分割，这类方法只能处理处于直立姿态的物体（例如，物体倒置时便无法处理），且转换过程会显著增加耗时。另一类是纯三维方法 [13],，基于三维数据集训练语义提示模型，能够以较强的鲁棒性支持对“任意文本、任意物体和任意姿态”进行分割，无需进行二维转换，因此速度显著提升。然而，在打破规范空间的约束后，网络仅通过“文本-部件相似性”的约束，学习“文本与几何结构”之间的对应关系。尽管语义相似的部件通常具有相似的几何结构（例如，椅子的腿大多较为细长），但相似的几何结构可能对应不同的语义（例如，椅子的扶手和腿都较为细长，但语义不同）。因此，仍需具备额外的感知能力，以解决开放世界分割中的语义歧义。

3. 方法

3.1. 概述

给定一个三维形状和一条文本提示，我们的方法通过编码形状的几何特征和文本的语义特征，计算跨模态相似度以将文本与形状区域关联起来，并将这些关联解码为部件级标签，从而实现可提示式三维语义分割。概述见图2。

为了使框架具备可迁移的跨类别规范空间理解能力，以支持开放世界三维语义部件推理，我们从外部和内部两个方面实现这一目标。**数据与监督（外部）**（第3.2节）。我们通过大语言模型引导的类别内与跨类别规范化流程，构建了一个跨类别规范数据集。该语料库提供规范映射、部件框和语义关联等监督信号，使模型能够从数据中归纳出潜在的规范参考坐标系。**双分支框架与规范空间目标（内部）**（第3.3节）。如图2所示，我们的双分支架构包括：（i）在训练和推理阶段均使用的特征提取分支，以及（ii）仅用于训练、用于预测规范映射和部件级边界框的规范分支

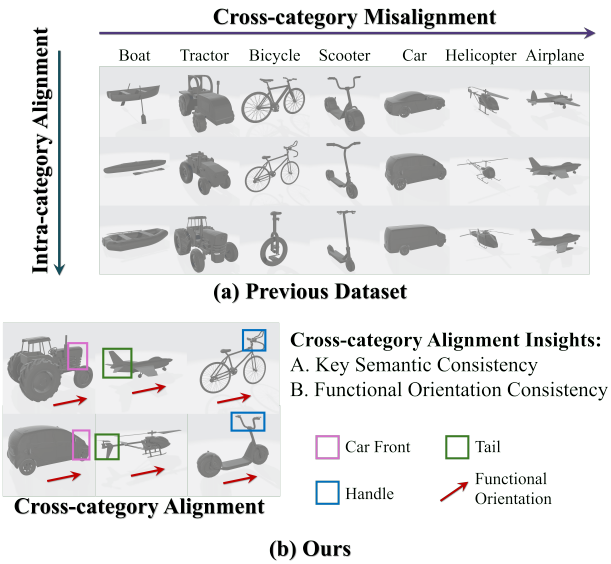


图 3. **(a)** 以往工作进行类别级规范化，对齐类别内形状，却忽略了跨类别一致性。**(b)** 我们通过大语言模型对类别进行聚类，并依据关键语义部件和功能一致性对齐不同类别。

我们的训练损失结合了几何-文本对齐项和两个规范空间正则项。几何-文本对齐项用于稳定跨模态映射；两个规范空间正则项，即规范映射锚定和规范边界框校准，则通过将点嵌入拉近规范代码，并收紧规范空间中语义部件的空间范围，确保与归纳得到的参考坐标系保持一致。

3.2. 统一的跨类别规范数据集

学习可迁移的规范空间先验，对于在开放世界环境中推广至多样且可能未见过的类别至关重要。然而，现有规范数据集 [3, 21] 将各类别单独处理，未进行跨类别对齐（如图3(a) 所示），这限制了其在开放世界场景中的扩展性。为解决这一问题，我们构建了一个 统一规范数据集，涵盖大多数常见物体类别，并进行跨类别对齐（图3 (b)）。该过程包含两个关键步骤：(i) 类别内规范化，将每个类别中的实例对齐到共享规范空间；以及 (ii) 跨类别规范化，对齐不同类别中语义对应的部件或地标。

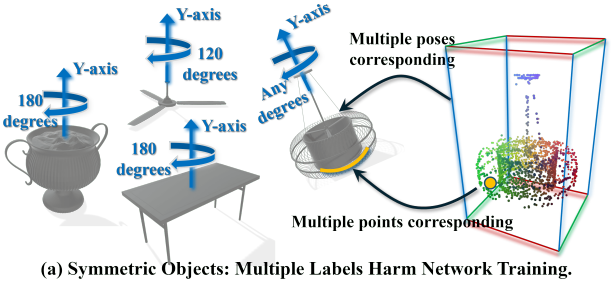
由于类别内规范化已在现有工作中得到广泛研究 [2, 7, 22], ，我们主要关注更具挑战性的跨类别规范化任务。由于各类别在几何形状和功能上存在巨大差异，这一任务十分困难。例如，叉子、自行车和树干等类别在形状和用途上都有显著差异。为解决这一问题，我们设计了一种分层对齐策略：我们首先将物体类别归入语义上连贯的类别簇，依据共享功能或使用情境；然后依次进行簇内对齐与跨簇对齐，逐步在所有类别上构建统一的规范空间。

具体而言，我们构建了一个类别级规范数据集 [2], 包含 200 个常见类别和 1.7 万种形状。首先，我们使用大型语言模型（例如 GPT）将这 200 个类别聚类为 19 个语义连贯的组（例如交通运输、工具）。在每个簇内，根据共同的功能特征进行对齐。例如，交通运输簇中的自行车和飞机的转向相关部件会进行对齐，以确保方向一致。由于该数据集 [2] 已统一了每个类别的核心语义方向，因此对同一簇内的多个类别进行对齐通常只需简单的离散旋转（例如 90°、180° 或 270°），从而提高了计算效率。接下来，通过验证高层语义的一致性来进行跨簇对齐。例如，我们确保交通运输簇与动物簇具有一致的向前移动方向。此外，为进一步丰富形状多样性，我们对物体应用轴对齐形变。最终，我们获得了一个在类别之间及类别内部均保持对齐一致的统一规范数据集。更多数据集详情和可视化示例请参见补充材料。

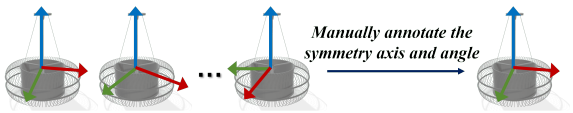
3.3. 规范感知双分支框架

架构。 我们设计了一个用于可提示三维语义理解的双分支框架。如图2所示，我们采用特征提取分支，从输入点云中预测三维特征，仅需进行一次推理，不同于依赖二维渲染的既有三维聚合方法。规范嵌入分支是在训练过程中引入的、我们设计中的独特组成部分，它生成由规范空间相关信号监督的中间表示。这增强了模型对规范空间的感知能力，从而改善跨形状和提示的对齐。特征提取分支基本遵循 Find3D 的设计。它采用 Pt3 [23] 作为点云编码的骨干网络，并采用 SigLIP [28]进行文本特征提取。一个轻量级的三层多层感知机将点特征投影到与文本特征相同的嵌入空间中。

规范嵌入分支包含两个头：一个用于规范映射预测，另一个用于语义边界框预测。对于规范映射预测，我们以三维形状特征作为输入，并受三维生成方法启发 [6, 9], ，回归三个连续标量场（编码为 RGB色彩图），而不是离散的逐点值，以更好地保留空间连续性。对于语义边界框预测，文本特征作为查询，从形状特征中提取相关区域，并输出一个表示边界框的 6 维向量。



(a) Symmetric Objects: Multiple Labels Harm Network Training.



(b) Previous Work: Category-Specific Processing for Unique Labels.



(c) Ours: Learning the Canonical Map via Semantic Parts Enables Open-World Adaptation, No Category-Specific Design Needed.

图 4. 处理对称物体以锚定规范映射。(a) 对于对称形状，多种有效姿态会导致点级规范标签存在歧义，使得直接进行点级监督并不可靠。(b) 先前的工作依赖特定类别的处理方式和人工对称性标注，限制了其在开放世界中的可扩展性。(c) 我们则对 RGB编码的规范坐标应用置换不变集合损失，在匹配规范空间中语义部件整体布局的同时，对称姿态歧义也不会影响其鲁棒性。

3.3.1. 语义对比对齐损失

在开放世界可提示部件分割中，语义并非互斥：同一个部件可以根据材质、功能、同义词选择或标签粒度来指称。瓶子瓶颈上的一个点，既可以有效地标记为“瓶颈”（细粒度），也可以标记为“瓶身”（粗粒度），因此，单点单标签监督十分脆弱。遵循 Find3D [13], ，我们可以采用对比学习来编码软点-文本亲和度：

$$\mathcal{L}_h = \frac{1}{M}\sum_{i=1}^{M} -\log \frac{f(\bar{\mathbf{p}}_i, \mathbf{t}_i)}{\sum_{k\in\mathcal{B}} f(\bar{\mathbf{p}}_i, \mathbf{t}_k)}, \tag{1}$$

其中， $\bar{\mathbf{p}}_i$ 表示 i个语义部件掩码内点样本的平均特征，$\mathbf{t}_i$ 是文本嵌入。 M 是部件标签的数量， $\mathcal{B}$ 表示小批次中的语义标签集合，而 $f(\bar{\mathbf{p}}_i, \mathbf{t}_i) = \exp(\bar{\mathbf{p}}_i^{\top} \cdot \mathbf{t}_i/\tau)$，温度为 τ。我们会对每个部件中的点进行策略性采样，详见下文。

我们观察到，在每个部件内进行均匀点采样会使式（2）偏向区块-文本一致性，却可能忽略单个点的偏差，导致部件边界附近出现噪声掩码，并减缓收敛。为了解决这一问题，我们提出一种困难负样本采样策略，即从部件间边界采样更多具有判别性的负样本，以提高分割准确率。困难双向对比损失定义如下：

$$\mathcal{L}_h = \frac{1}{2M}\sum_{i=1}^{M} -\log \frac{f(\bar{\mathbf{p}}_i, \mathbf{t}_i)}{\sum_{k\in\mathcal{B}_i} f(\bar{\mathbf{p}}_i, \mathbf{t}_k)} - \log \frac{f(\mathbf{t}_i, \bar{\mathbf{p}}_i)}{\sum_{n\in\mathcal{P}_i} f(\mathbf{t}_n, \hat{\mathbf{p}}_n)},$$
$$\hat{\mathbf{p}}_n = \frac{1}{W_n}\left(\sum_{\mathbf{p}_j\in\Omega_n} \mathbf{p}_j + (1+\alpha)\sum_{\mathbf{p}_e\in\mathcal{E}_n} \mathbf{p}_e\right), \tag{2}$$

其中 $\alpha = 0$ 用于 $n = i$， $\alpha > 0$ 用于 $n \neq i$。 $\mathbf{p}j$ 和 $\mathbf{p}e$ 表示各个点特征。Ω_n 和 $\mathcal{E}_n$ 分别表示与语义标签 n 相关联的部件的内部区域和边缘区域。归一化权重定义为 $W_n = |\Omega_n| + (1+\alpha)|\mathcal{E}_n|$。

3.3.2. 规范空间正则化

基于双分支设计和精选的跨类别规范数据集，我们的方法利用规范空间线索对点特征进行正则化。我们引入两项损失（详见下文），以对齐各部件的规范分布，并强制实现稳健的空间范围。

规范映射锚定损失。我们的目标是确保具有相同（或密切相关）语义的语义部件在规范空间中展现出一致的空间分布，无论是在物体类别内部还是跨物体类别。我们借助精选数据集所提供的规范化元数据与语义部件标签，锚定各部件的规范映射，从而实现这一目标。然而，在实践中，我们观察到轴向与平面对称性常常会引入对应关系歧义（图4 （a））：对称物体的多种姿态在规范空间中同样有效，从而使逐点规范监督变得不可靠。现有规范空间感知方法试图通过手动标注对称轴或施加特定类别约束来解决该问题（图4 （b）），但此类设计无法扩展至开放世界分割。

我们的核心洞见是彻底避免对应关系。我们不再逐点监督规范坐标，而是将每个语义部件视为规范空间中的一个分布。随后，我们通过双向倒角距离将预测规范映射与真实规范分布进行匹配，以此对其进行正则化。由于这一目标比较的是分布的形状，而非单个坐标，因此对称构型在规范空间中会变得等价。于是，对称点会自动收敛到同一规范区域，无需使用对称性标签或类别级轴标注。

形式化地，令 $\mathcal{G}_m^p = \{\mathbf{a}_i\}$和 $\mathcal{G}_m^t = \{\mathbf{b}_j\}$分别表示规范空间 $\Omega \subset \mathbb{R}^d$ （通常为 $d \in \{2, 3\}$）中部件 m的预测点集和真实点集。我们将规范映射锚定损失定义为：

$$\mathcal{L}_{ca} = \frac{1}{M}\sum_{m=1}^{M}\Big(\frac{1}{|\mathcal{G}_m^p|}\sum_{\mathbf{a}_i\in\mathcal{G}_m^p} \min_{\mathbf{b}_j\in\mathcal{G}_m^t} \|\mathbf{a}_i - \mathbf{b}_j\|_p + \frac{1}{|\mathcal{G}_m^t|}\sum_{\mathbf{b}_j\in\mathcal{G}_m^t} \min_{\mathbf{a}_i\in\mathcal{G}_m^p} \|\mathbf{b}_j - \mathbf{a}_i\|_p\Big). \tag{3}$$

$\|\cdot\|_p$ 表示 p-范数（默认使用 p=2）。通过对齐分布而非点集，该损失促使模型学习具有旋转不变性且对称性鲁棒的规范布局。

规范框校准损失。在推理时，部件通过将逐点特征与来自用户提示的文本嵌入相匹配来进行分割。尽管该方法较为灵活，但容易受到局部噪声的影响，并可能产生模糊或不规则的边界，部分原因在于前述损失更强调部件级分布对齐，而非逐点准确率。为增强边界一致性，规范分支在规范空间中为每个语义部件预测一个三维边界框。这些框提供了粗略但稳定的空间先验，从而锐化边界并抑制虚假激活（见图5（中））。该损失表述为：

$$\mathcal{L}_{\text{cb}} = \frac{1}{M}\sum_{m=1}^{M}\frac{1}{6}\left\|\mathbf{b}_m^{\text{p}} - \mathbf{b}_m^{\text{t}}\right\|_1, \quad \mathbf{b}_m^{(\cdot)} \in \mathbb{R}^6. \tag{4}$$

此处 M 表示语义部件的数量。对于部件 m，预测框与真实框为$\mathbf{b}_m^{\text{p}}$，$\mathbf{b}_m^{\text{t}} \in \mathbb{R}^6$，并参数化为 $[x_{\min}, y_{\min}, z_{\min}, x_{\max}, y_{\max}, z_{\max}]$。该损失鼓励部件在规范空间中占据连贯的空间范围，并以更紧致的几何正则化补充分布级锚定。

3.3.3. 完整训练目标

完整训练目标的公式如下：

$$\mathcal{L}_{\text{total}} = \lambda_h \cdot \mathcal{L}_h + \lambda_{ca} \cdot \mathcal{L}_{ca} + \lambda_{cb} \cdot \mathcal{L}_{cb}, \tag{5}$$

其中，$\lambda_h = 1$、$\lambda_{ca} = 10$和 $\lambda_{cb} = 3$ 表示平衡权重。为确保稳定收敛，我们采用两阶段训练方案：在第一阶段，我们仅使用对齐损失 $\mathcal{L}_h$ 训练框架，直至收敛。随后，在第二阶段加入 $\mathcal{L}_{ca}$ （规范映射锚定损失）和 $\mathcal{L}_{cb}$ （规范框校准损失），并继续训练直至收敛。

4. 实验

4.1. 实现细节

训练数据集。我们基于3Dcompat200语料库 [2], 构建统一规范数据集，该语料库提供了涵盖200个类别的 ∼17K个形状及其语义部件标注。该数据集覆盖了超过日常物体类别，并支持强泛化。为了训练框架，我们将每个物体归一化至单位边界框，并对每个物体均匀采样5,000个表面点，同时保留RGB颜色和法向量。

框架。 我们采用 PointTransformerV3 [23]作为骨干网络，提取 768 维逐点特征，并使用 SigLIP-Base/16-224 [28] 获得 768 维文本嵌入。更多实现细节请参阅补充材料。

4.2. 比较

为比较可提示语义分割，我们设计了涵盖数据分布、输入状态和查询形式的评估，详情如下：

评估数据集。 我们在多个公开的三维部件分割基准测试上评估了我们的方法和具有代表性的最先进的基线方法。首先，我们报告了3Dcompat200数据集 [2], 中的3Dcompat-Coarse和3Dcompat-Fine测试集上的结果。该数据集包含200个物体类别，以及约2,000个形状，分别以粗粒度和细粒度标注了部件标签。此外，我们还在两个常用基准测试ShapeNet-Part [27]和PartNet-E [14], 上评估性能，它们分别包含16个和45个物体类别。

测试设置。 为了在开放世界条件下全面评估所提出的方法，我们考虑了实际应用中常见的两个关键因素：提示变化和物体姿态不确定性。为评估对提示歧义的鲁棒性，我们为文本查询设计了两种输入模式：“{部件}”表示单词提示（例如腿），而“{部件} 属于 {类别}”表示组合短语提示（例如椅子的腿）。为模拟开放世界场景中的姿态变化，我们在两种设置下评估分割效果：规范设置下，物体处于规范姿态；旋转设置下，物体经过随机旋转，以模拟任意朝向。

在 3Dcompat-Coarse 和 3Dcompat-Fine 上的结果。表 1 报告了 3Dcompat200 数据集粗粒度和细粒度测试集上的定量结果。我们使用平均交并比（mIoU）评估性能，具体做法是先计算每个对象的部件交并比平均值，再对所有实例取平均。我们的方法始终优于所有基线方法，平均比排名第二的方法 Find3D 高出 25.55% 。与基于2D渲染的方法（例如 PartSLIP ++ [31]、PointCLIPV2 [32]）相比，我们的模型不仅依托前馈机制实现了显著更快的推理速度（0.9 秒/个形状，而 PartSLIP 需要 2.5 分钟++），还取得了更高的分割精度。此外，与三维前馈方法 Find3D 相比 [13], ，为确保公平，我们同时报告了其原始作者发布模型和在我们精选数据集上重新训练的变体的结果；相比之下，我们的方法在所有测试设置下都表现出稳定的提升。在粗粒度数据集上，无论是规范姿态还是旋转姿态，我们的模型相较 Find3D 的绝对 mIoU 均提升了 8% − 11%。在细粒度数据集上，相同设置下我们观察到4% − 7% 的提升。这些改进凸显了我们以规范空间为引导的学习策略的有效性，表明该策略能够在姿态变化和不同提示形式下增强语义理解。

ShapeNet-Part 和 PartNet-E 上的结果。 表1 还报告了 ShapeNet-Part 和

表 1. 可提示语义分割的定量评估（以语义标签为参照，并报告平均交并比）。上：粗粒度数据集和细粒度数据集；下：ShapeNet-Part 数据集和 PartNet-E 数据集。Find3D* 表示使用我们构建的 3D 数据重新训练；PartSLIP++* 表示以规范视图为输入，在 PartNet-E 数据集上进行微调。

平均交并比 (%)	3D 前馈	3Dcompat-Coarse				3DCompat-Fine			
		规范 {物体}的 {部件}	{部件}	旋转 {物体} 的 {部件}	{部件}	规范 { {物体的部件}	{部件}	旋转 { {物体的部件}	{部件}
PointCLIPV2	×	14.09	14.16	13.17	13.39	7.25	7.18	7.43	7.20
PartSLIP++	×	6.12	17.62	5.89	15.82	3.79	5.68	3.41	4.76
Find3D	✓	31.72	25.36	32.47	26.23	12.69	12.57	12.73	12.79
Find3D*	✓	36.89	45.00	38.14	46.02	17.72	24.97	18.57	26.35
Ours	✓	**47.51**	**54.52**	**47.74**	**54.55**	**24.48**	**31.29**	**24.46**	**30.97**
平均交并比 (%)	3D 前馈	ShapeNet-Part				PartNet-E			
		规范 {部件}的 {物体}	{部件}	旋转 {部件} 属于 {物体}	{部件}	规范 {部件} 属于 {物体}	{部件}	旋转 {部件} 属于 {物体}	{部件}
OpenMask3D	×	8.94	10.37	6.75	14.56	12.54	11.24	11.93	11.67
PartSLIP++*	×	1.43	6.46	0.94	6.03	5.12	**32.71**	3.87	**23.03**
PointCLIPV2	×	16.91	20.22	16.88	18.19	11.28	9.70	10.32	10.22
Find3D	✓	28.39	24.09	29.64	23.71	16.86	16.38	17.62	17.16
Ours	✓	**36.16**	**33.31**	**34.20**	**32.84**	**17.59**	17.19	**18.48**	18.17

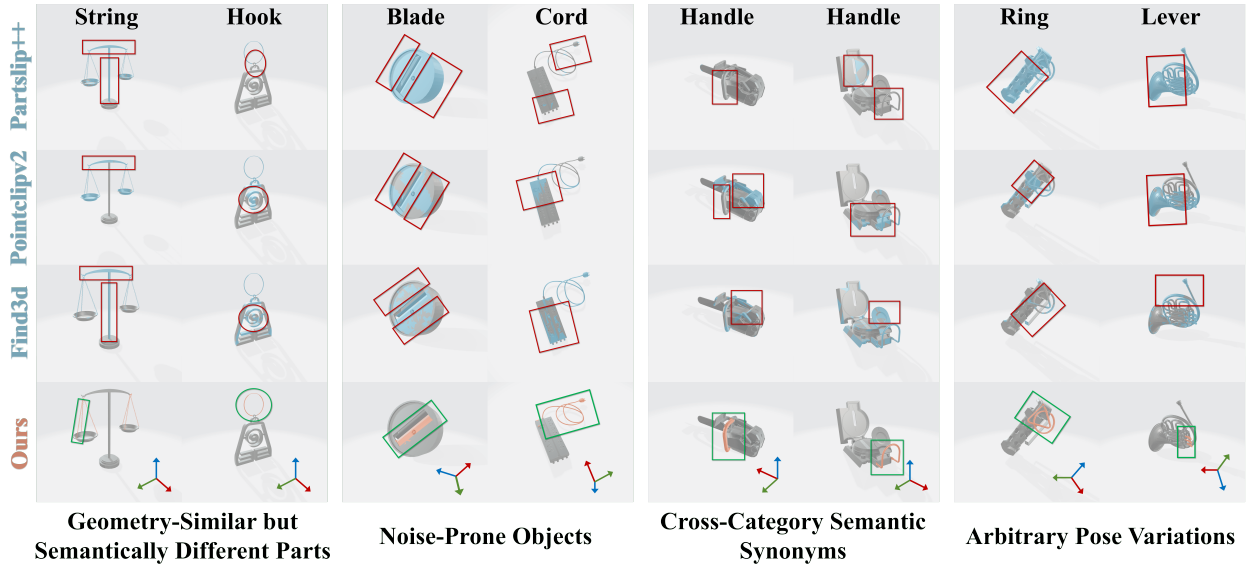


图5. 可提示3D部件分割的定性比较。在各种具有挑战性的情形中（几何形状相似但语义不同、易受噪声干扰的物体、跨类别语义以及任意姿态），与现有基线方法相比，我们的方法能够产生更准确且更一致的部件定位。

PartNet-E 数据集。我们的方法在这两个基准测试上均达到了最先进水平（SOTA），相较于表现最佳的基线方法，在 ShapeNet-Part 上平均提升 29.89%，在 PartNet-E 上平均提升 5.01%。值得注意的是，大多数方法在 PartNet-E 上表现一般，因为其标注侧重于物体材质和细粒度细节，而这些特征通常不会出现在其他三维语义数据集中。有趣的是，作为一种基于二维渲染的方法，PartSLIP++ 在单词提示设置下取得了相对更好的表现。这可能归功于其骨干网络 GLIP：该模型使用约 2700 万个图文对进行预训练，并在 PartNet-E 数据集的规范数据上进一步微调。尽管基于二维渲染的方法受限于速度和多-

视角不一致，这些观察结果凸显了以下需求：更好地从大型 2D数据集中挖掘结构化语义先验，同时考虑其噪声标签和领域差距，并将这些知识与数量有限但标注完善的三维分割数据高效结合。

定性比较。 图5 展示了与Find3D以及基于二维渲染的基线方法在四种场景下的定性比较。对于几何形状相似但语义不同的部件（左），现有方法常因细长形状相近而将绳子和钩子与物体主体混淆，而我们的模型能够准确定位每个目标部件。对于易受噪声干扰的物体，细小或纤薄的部件容易受到周围杂物干扰，基线方法要么漏检该部件，要么预测出过大的区域，而我们的

表2。定量结果：　　消融研究。所有数值均为在所有类别的所有实例中，针对各部件计算的mIoU分数　　s.

变体	困难负样本采样	规范映射锚定	跨类别规范化 规范化（数据）	规范框校准	规范 {物体} 的 {部件}	规范 {部件}	旋转 {物体} 的 {部件}	旋转 {部件}
A					36.89	45.00	38.14	46.02
B	✓				38.30	48.30	39.28	48.76
C	✓	✓			41.71	52.01	42.48	52.57
D	✓	✓	✓		43.34	54.15	42.63	53.70
完整模型	✓	✓	✓	✓	**47.51**	**54.52**	**47.74**	**54.55**

* 所有设置均在类内规范化形状上训练。* “跨类别规范化（数据）”表示通过我们的跨类别规范化流程处理数据集。

预测结果依然紧致且稳定。对于跨类别的语义同义词，例如出现在不同物体类别上的“把手”，先前方法难以在不同形状之间迁移，并经常漂移到邻近区域，而我们的方法能够生成与预期语义提示一致的分割结果。最后，在任意姿态变化下，对比方法对姿态变化较为敏感，而我们的框架借助规范空间引导，能够在不同姿态下保持鲁棒定位。

4.3. 消融研究与分析

消融研究。我们通过评估变体 A–D 和完整模型，考察各个关键组件的作用。结果汇总于表2中，其中变体 A 作为基线。困难负样本采样（变体 B）小幅提升了对比对齐的鲁棒性，而规范地图锚定损失（变体 C）通过强化姿态鲁棒且具备规范空间感知能力的部件特征，带来了更显著的提升。尽管变体 A–C 均在类内规范化实例上训练，但进一步将我们的跨类别规范化流程应用于训练数据（变体 D），能在所有测试设置下持续提升性能，凸显了跨类别规范监督的优势。最后，加入规范边界框校准损失后，模型取得了最佳整体结果；该损失通过约束规范分支预测的各部件规范边界框，对提取的特征进行正则化。不同变体的定性分析见补充材料。

三维形状特征分析。 为了更好地理解学习到的表示，我们在图 6 中可视化了不同方法得到的逐点特征。类别无关的分解方法 PartField [11] 生成了清晰的部件划分，但其特征在不同形状或姿态之间并不一致。Find3D [13]展现出更强的跨形状和跨姿态一致性，但不同形状间的特征仍未完全对齐。此外，Find3D 在单个物体内得到的特征缺乏清晰的部件边界，相邻部件上的点也常常混杂在一起。相比之下，我们的框架生成的特征既具有语义对齐性，结构上也分离良好：不同部件呈现出清晰的边界，而不同形状和姿态中的对应部件则具有高度相似的特征模式。这种表现源于

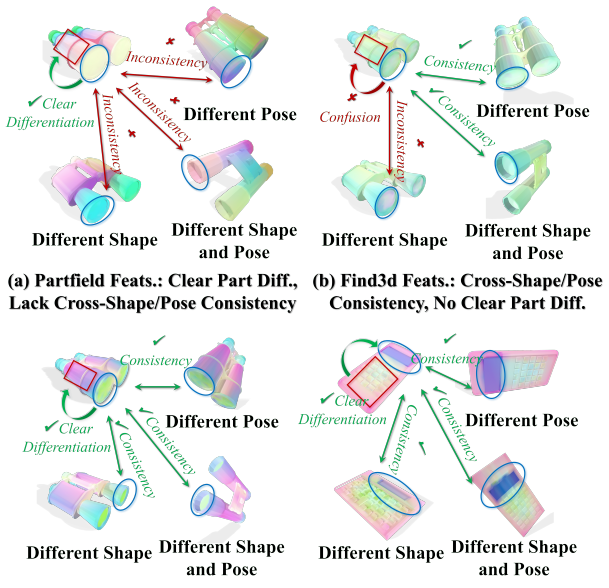


图 6. 形状特征分析。（a）Partfield [11] 能够产生清晰的部件划分，但缺乏跨形状和跨姿态的一致性。（b）Find3D [13] 提高了一致性，但生成的部件边界较为模糊。（c）我们的方法同时实现了清晰的部件分离，以及跨形状和姿态的一致特征。

在语义部件分割训练中结合了我们的规范空间正则化，该正则化明确鼓励网络在共享规范空间中对齐几何形状与语义。因此，学习到的逐点特征既具有跨形状和跨姿态的一致性，也能细致地区分各个部件，这表明我们的架构有望成为强大的三维形状特征学习骨干网络，而不局限于特定的语义分割任务。

5. 结论

我们重新审视了开放世界可提示三维分割，并发现规范空间是缺失的表征变量。虽然我们的方案通过潜在规范参考坐标系将这一理念付诸实践，但我们认为其影响远不止于可提示分割。一旦规范性成为可学习的对象，规范空间感知便可支持更丰富的任务：组合式3D查询

应答、跨CAD/视频领域的跨模态基础对齐，以及能够先在规范空间中运行、再于欧几里得空间中行动的新一代三维智能体。我们认为，CoSMo3D 不仅仅是一个系统，更是迈向更有原则的三维理解栈的一步；在这个理解栈中，规范参考系是首要的表征层。

参考文献

[1] Ahmed Abdelreheem、Ivan Skorokhodov、Maks Ovsjanikov和Peter Wonka。Satr：三维形状的零样本语义分割。见《IEEE/CVF国际计算机视觉大会论文集》，第15166–15179页，2023年。3[2] Mahmoud Ahmed、Xiang Li、Arpit Prajapati和Mohamed Elhoseiny。3Dcompat200：面向组合式识别的语言对齐的大规模三维视觉数据集。《神经信息处理系统进展》，37：135770–135782，2024年。4，6[3] Angel X Chang、Thomas Funkhouser、Leonidas Guibas、Pat Hanrahan、Qixing Huang、Zimo Li、Silvio Savarese、Manolis Savva、Shuran Song、Hao Su等。ShapeNet：信息丰富的三维模型库。arXiv预印本 arXiv:1512.03012，2015。4[4] Ken Deng、Yunhan Yang、Jingxiang Sun、Xihui Liu、Yebin Liu、Ding Liang和Yan-Pei Cao。GeoSAM2：释放SAM2在3D部件分割中的能力。arXiv预印本 arXiv:2508.14036，2025年。3[5] Ross Girshick、Jeff Donahue、Trevor Darrell和Jitendra Malik。用于精确目标检测与分割的基于区域的卷积网络。《IEEE模式分析与机器智能汇刊》，38 158，2015 1 142 2 ( )：– 。[6] 混元3D团队，Shuhui Yang、Mingxin Yang、Yifei Feng、Xin Huang、Sheng Zhang、Zebin He、Di Luo、Haolin Liu、Yunfei Zhao等。混元3D 2.1：从图像到具备可投产基于物理的渲染材质的高保真三维资产。arXiv预印本 arXiv:2506.15442，2025年。4[7] Li Jin、Yujie Wang、Wenzheng Chen、Qiyu Dai、Qingzhe Gao、Xueying Qin和Baoquan Chen。基于几何与语义一致性的单样本三维物体规范化。见《计算机视觉与模式识别会议论文集》，第16850–16859页，2025年。4[8] Alexander Kirillov、Eric Mintun、Nikhila Ravi、Hanzi Mao、Chloe Rolland、Laura Gustafson、Tete Xiao、Spencer Whitehead、Alexander C Berg、Wan-Yen Lo等。分割一切。见《IEEE/CVF国际计算机视觉大会论文集》，第4015–4026页，2023年。2[9] Weiyu Li、Jiarui Liu、Hongyu Yan、Rui Chen、Yixun Liang、Xuelin Chen、Ping Tan和Xiaoxiao Long。Craftsman3D：结合三维原生生成与交互式几何细化器的高保真网格生成。arXiv预印本 arXiv:2405.14979，2024年。4[10] Minghua Liu、Yinhao Zhu、Hong Cai、Shizhong Han、Zhan Ling、Fatih Porikli和Hao Su。PartSLIP：借助预训练图像-语言模型对三维点云进行低样本部件分割。见《IEEE/CVF con-

计算机视觉与模式识别会议，页码 21736–21746，2023年。3[11] Minghua Liu 、 Mikaela Angelina Uy 、 Donglai Xiang、Hao Su、Sanja Fidler、Nicholas Sharp 和 Jun Gao。Partfield：学习用于部件分割及其他任务的三维特征场。载于 IEEE/CVF国际计算机视觉大会论文集，页码 9704–9715，2025年。 3, 8[12] Xu Ma、Can Qin、Haoxuan You、Haoxi Ran 和 Yun 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。arXiv 预印本 arXiv:2202.07123，2022年。 2[13] Ziqi Ma、Yisong Yue 和 Georgia Gkioxari。在三维中寻找任意部件。载于 IEEE/CVF国际计算机视觉大会论文集，页码 7818–7827，2025年。 2, 3, 5, 6, 8[14] Kaichun Mo、Shilin Zhu、Angel X Chang、Li Yi、Subarna Tripathi、Leonidas J Guibas 和 Hao Su。PartNet：用于细粒度、层次化三维物体理解的大规模基准测试。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 909–918，2019年。 6[15]Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。通过自然语言监督学习可迁移视觉模型。载于国际机器学习大会论文集，页码 8748–8763。PMLR，2021年。 3[16] Nikhila Ravi、Valentin Gabeur、Yuan-Ting Hu、Ronghang Hu、Chaitanya Ryali、Tengyu Ma、Haitham Khedr、Roman R¨adle、Chloe Rolland、Laura Gustafson 等。SAM 2：在图像和视频中分割一切。arXiv 预印本 arXiv:2408.00714，2024年。 2[17] Roger N. Shepard 和 Jacqueline Metzler。三维物体的心理旋转。《科学》，171 703，1971 3972 701 1 ( ): – 。[18] Ayc.a Takmaz、Elisabetta Fedele、Robert W Sumner、Marc Pollefeys、Federico Tombari 和 Francis Engelmann。Open-mask3d：开放词汇三维实例分割。arXiv 预印本 arXiv:2306.13631，2023年。 3[19] George Tang、William Zhao、Logan Ford、David Benhaim 和 Paul Zhang。分割任意网格：通过将 Segment Anything 2 提升至三维实现零样本网格部件分割。arXiv 电子印本，页码 arXiv–2408，2024年。 2[20] Ardian Umam、Cheng-Kun Yang、Min-Hung Chen、Jen-Hui Chuang 和 Yen-Yu Lin。PartDistill：通过视觉语言模型蒸馏进行三维形状部件分割。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 3470–3479，2024年。 3[21] He Wang、Srinath Sridhar、Jingwei Huang、Julien Valentin、Shuran Song 和 Leonidas J Guibas。用于类别级六维物体姿态与尺寸估计的归一化物体坐标空间。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 2642 2651，2019 4 – 。[22] Tong Wu、Jiarui Zhang、Xiao Fu、Yuxin Wang、Jiawei Ren、Liang Pan、Wayne Wu、Lei Yang、Jiaqi Wang、Chen Qian 等。OmniObject3D：大词汇量三维物体数据集，适用于

逼真感知、重建与生成。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 803–814，2023年。 4[23] Xiaoyang Wu、Li Jiang、Peng-Shuai Wang、Zhijian Liu、Xi-hui Liu、Yu Qiao、Wanli Ouyang、Tong He 和 Hengshuang Zhao。Point Transformer v3：更简单、更快速、更强大。载于 IEEE/CVF计算机视觉与模式识别会议论文集，页码 4840–4851，2024年。 4, 6[24] Xinhao Yan、Jiachen Xu、Yang Li、Changfeng Ma、Yunhan Yang、Chunshi Wang、Zibo Zhao、Zeqiang Lai、Yunfei Zhao、Zhuo Chen 等。X-part：高保真且结构连贯的形状分解。arXiv 预印本 arXiv:2509 08643，2025 3 . 。[25] Yunhan Yang、Xiaoyang Wu、Tong He、Hengshuang Zhao 和 Xihui Liu。Sam3d：分割三维场景中的一切。arXiv 预印本 arXiv:2306.03908，2023年。 2[26] Yunhan Yang、Yukun Huang、Yuan-Chen Guo、Liangjun Lu、Xiaoyang Wu、Edmund Y Lam、Yan-Pei Cao 和 Xihui Liu。Sampart3D：分割三维物体中的任意部件。arXiv 预印本 arXiv:2411.07184，2024年。 3[27]Li Yi、Vladimir G Kim、Duygu Ceylan、I-Chao Shen、Mengyan Yan、Hao Su、Cewu Lu、Qixing Huang、Alla Sheffer 和 Leonidas Guibas。用于三维形状集合区域标注的可扩展主动框架。ACM图形学汇刊 (ToG)，35(6):1–12，2016年。 6[28] Xiaohua Zhai、Basil Mustafa、Alexander Kolesnikov 和 Lucas Beyer。用于语言图像预训练的 Sigmoid损失。载于 IEEE/CVF国际计算机视觉大会论文集，页码 11975–11986，2023年。 4, 6[29] Qi Zhang、Jibin Peng、Zhao Huang、Wei Feng 和 Di Lin。用于语义点云分割的生成式困难样本增强。载于计算机视觉与模式识别会议论文集，页码 22205 22214，2025 2 – 。[30] Hengshuang Zhao、Li Jiang、Jiaya Jia、Philip HS Torr 和 Vladlen Koltun。Point Transformer。载于 IEEE/CVF国际计算机视觉大会论文集，页码 16259–16268，2021年。 2[31] Yuchen Zhou、Jiayuan Gu、Xuanlin Li、Minghua Liu、Yun-hao Fang 和 Hao Su。PartSLIP++：通过多视图实例分割和最大似然估计增强低样本三维部件分割。arXiv 预印本 arXiv:2312.03015，2023年。 3, 6[32] Xiangyang Zhu、Renrui Zhang、Bowei He、Ziyu Guo、Ziyao Zeng、Zipeng Qin、Shanghang Zhang 和 Peng Gao。Point-clip v2：通过提示 CLIP 和 GPT 实现强大的三维开放世界学习。载于 IEEE/CVF国际计算机视觉大会论文集，页码 2639–2650，2023年。 3, 6[33] Zhe Zhu、Le Wan、Rui Xu、Yiheng Zhang、Honghua Chen、Zhiyang Dou、Cheng Lin、Yuan Liu 和 Mingqiang Wei。Partsam：一种在原生三维数据上训练的可扩展、可提示部件分割模型。arXiv 预印本 arXiv: . . 3 21965，2025 2509