环顾四周并参考:2D 合成语义 3D 视觉基础的知识提炼

Islam Mohamed Bakr、Yasmeen Alsaedy、Mohamed Elhoseiny阿卜杜拉国王
科技大学 (KAUST) {eslam.abdelrahman, yasmeen.alsaedi,
mohamed.elhoseiny}@kaust.edu.sa

抽象的

已经探索了 3D 视觉基础任务,使用视觉和语言流理解指称语言来识别 3D 场景中的目标对象。然而,大多数现有方法都使用现成的点云编码器将视觉流用于捕捉 3D 视觉线索。我们在本文中解决的主要问题是“我们能否通过从点云合成的 2D 线索来巩固 3D 视觉流并在训练和测试中有效地利用它们?”。主要思想是通过合并丰富的 2D 对象表示来协助 3D 编码器,而无需额外的 2D 输入。为此,我们利用从 3D 点云合成的 2D 线索,并通过经验证明它们有能力提高学习到的视觉表示的质量。我们通过在 Nr3D、Sr3D 和 ScanRefer 数据集上进行全面实验来验证我们的方法,与现有方法相比,我们的性能得到了持续提升。我们提出的模块称为“环视和参考”(LAR),在三个基准测试(即 Nr3D、Sr3D 和ScanRefer)上的表现明显优于最先进的 3D 视觉接地技术。代码可在https://eslambakr.github.io/LAR.github.io/ 上找到。

1 简介

视觉语言建模在过去十年中引起了广泛关注,因为它对各个领域都有影响,包括视觉问答[3, 36]、3D 场景字幕[7]、图像文本匹配[26, 44]、视觉关系检测[25]和场景图生成[42]。因此,理解自然语言话语并将其应用于现实世界场景具有重要意义,因为它们在许多实际应用中都很重要,例如机器人导航和室内环境中的交互[22] [40]。

近年来,人们在 2D 环境下探索了视觉接地 (VG) [16, 30, 46, 55, 57],将物体检测扩展到更复杂的任务。然而,为了让真正的机器人在拥挤的室内或室外场景等具有挑战性的环境中将给定的话语智能地接地到自然场景中,需要 3D 表示来更好地理解周围环境。因此,ReferIt3D [2]和 Scan- Refer [5]提出了3D 视觉接地任务。ReferIt3D [2]引入了两个包含自然和合成语言话语的数据集,

图 1:我们提出的方法概览,我们的目标是根据输入话语和 3D 场景定位所指对象。我们的新方法通过在对象周围放置虚拟摄像头来为场景中的每个对象生成 2D 合成图像。这种新方法使我们能够利用丰富的 2D 语义表示,而无需真实的额外 2D 图像,从而不会限制潜在的应用场景。

第 36 届神经信息处理系统会议(NeurIPS 2022)。

表 1:各种 3D 视觉接地模型的比较。其中 GT 表示使用了地面实况框， Pred. 表示使用预训练检测器来获取物体提案。我们表明 Looking-outside [18]和SAT [52]分别使用了额外的输入,例如整个场景点云Spc和额外的 2D 图像2DImg。Opc指的是物体的点云。

	改装3D [2]	只是。参考 [37]	机构参考 [58]	时间分辨率神经网络[17]	向外看 [18]	D3网络 [6] [13] [59]	三维模型	3DVG	星期六[52]	我们的
目标提案GT语言编码器		燃气调轮	前	前	前	前	燃气调轮	前	燃气调轮	燃气调轮
RNN DistilBert GloVE GloVE RoBERTa GRU DistilBert GloVE BERT BERT视觉输入OP C OP C OP C + SP C OP C OP C + SP C OP C OP C OP C OP C + 2DImg OP C										

即 Nr3D 和 Sr3D,旨在从每个场景的预定义对象集中确定指称对象。相比之下,ScanRefer [5]旨在根据给定的语言描述和 3D 点云预测指称对象的 3D 边界框。ScanRefer [5]和ReferIt3D [2]都是基于 ScanNet 数据集[8] 构建的。关键的区别在于,在每个场景中,几个对象实例与指称对象属于相同的细粒度类别,即干扰项,这使得 ReferIt3D [2]更具挑战性。

由于 3D 点云的稀疏性和混乱性,SAT [52]探索在训练期间利用ScanNet 数据集[8]提供的2D图像来辅助点云语言联合表征学习。SAT [52]提出了三种变体:1)基线,非 SAT,在处理基础任务时仅依赖 3D 点云。2)仅在训练阶段使用 2D 语义并在推理时屏蔽它们,我们将其称为 SAT。3)在两个阶段（即训练和测试)都使用 2D 语义,我们将其称为 SAT-GT。然而,在训练或推理中需要额外的 2D 输入限制了潜在的应用场景。为了解决这个问题,我们假设结合合成的 2D 语义将巩固 3D 学习到的表示,而不需要任何额外的信息。为此,我们提出了一种新颖的 2D 合成图像生成器 (SIG) ,将 3D 点云投影到多视图 2D 图像中,如图 1 所示。据我们所知,LAR 是第一种使用合成 2D 语义增强 3D 任务学习的方法。此外,我们提出了一种基于多模态 Transformer 的新型架构。我们提出的 Nr3D 架构 LAR 优于 SAT [52] ,准确率提高了 1.6% ,同时超过了所有现有的除了 3D 点云之外没有使用任何额外信息的方法,差距更大,即最接近的 5.0%。此外,LAR 架构在 Sr3D 和 ScanRefer 基准测试中分别取得了 59.35% 和 54.6% 的最佳结果。另一个变体是从我们的架构中定制的,以与 SAT-GT 进行比较,其中几何信息和真实类标签被连接到 2D 图像;在我们的案例中是合成图像,在 SAT 案例中是原始图像。我们的LAR 架构比 SAT-GT 高出 11.7% （我们的方法为 62%,而该方法为 50.3%）。我们的贡献总结如下:

·我们针对仅依赖 3D 点云的 3D 视觉接地任务提出了一种新颖的端到端多模态变压器架构 LAR,该架构利用了我们的 SIG 模块生成的合成 2D 图像。 ·我们通过各种基准（即 Nr3D、Sr3D 和 ScanRefer)的大量实验验证了我们的 LAR 架构的有效性,这
些实验表明,利用我们的 2D合成语义可以显着提高结果。 · 通过详细分析,我们检查了我们方法的内部行为和有效性。

2 相关工作

我们的新颖架构在多个领域取得了成功,包括 2D 视觉基础、3D 视觉基础和 3D 任务中的 2D 语义。

–2D 视觉接地。2D视觉接地是根据自然语言描述定位对象。Flickr30k [32]和 ReferItGame [20]为图像中的特定对象生成引用表达式,用于识别所引用的对象。先前的研究[56、27、29]试图通过关注边界框级别,根据给定的表达式定位这些引用的对象。相比之下， [53、10]则专注于像素级预测。另一种 2D 视觉接地分类法是单阶段,也称为单阶段框架,与两阶段框架相对。单阶段方法（例如[38、49、50])将语言查询与图像的每个像素/块融合,并生成一组可能的边界框候选集。相比之下,两阶段框架[45、54、56]使用视觉图像语义来生成对象提议。之后，

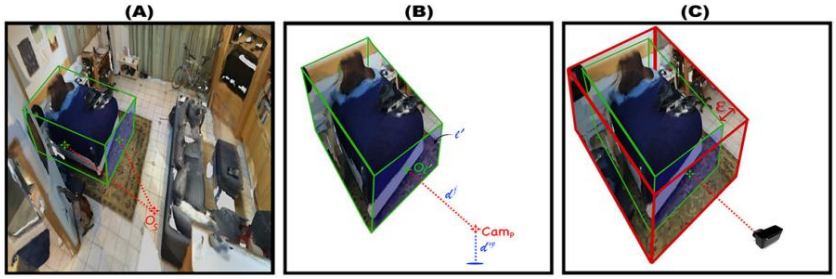

图 2:我们的 2D 合成图像生成器 (SIG) 模块的简化概览。首先,我们确定每个物体相对于场景中心的突出面。然后,将相机放置在距离该面d f 的位置,距离房间地板 d 的位置。最后,我们将感兴趣的区域随机扩展。

通过将每个提议与语言查询进行比较来选择基础预测。因此， 2D 视觉基础方法可以指导学习 3D 关系。在此推动下,我们探索使用2D 合成语义来巩固 3D 视觉流。

–3D 视觉接地。表 1 总结了现有的 3D 视觉接地 (3D VG) 模型,无论是使用地面实况对象提议还是使用预训练检测器来生成提议。此外,它总结了是否与对象的 3D 点云一起使用了额外信息。3D视觉接地领域在过去几年中经历了广泛的研究并取得了长足的进步[2、5、12、17、59、1、37、13、58、19、52、6、19] 。3D 视觉接地任务使用语言指令和 3D 场景[8]作为输入,以估计将最相关对象分配给所提供表达的边界框。然而,文献提供了一些扩展,例如使用辅助输入来改进 3D 对象的定位[52]。先前的研究[2, 12, 17, 59, 1, 13, 58, 19] 侧重于通过采用两阶段框架来改进语言和 3D 场景融合。第一阶段,给定表示场景的 3D 点云,输出 3D 对象提案。其次,融合视觉和语言特征以计算每个 3D 对象提案的置信度得分。可以直接使用真实提案而不是预测它们[52] [13] [2],从而跳过第一阶段。相反， [6, 19, 59]不假设可以访问真实对象,而是使用对象检测器来提取提案。直观地讲,如果检测器失败,则引用性能将显著下降。因此,我们专注于与基础任务相关的主要训练目标,即通过假设可以访问真实框来学习 3D 视觉-语言联合表示。此外,我们还探索了利用检测器生成的提案的其他设置,如附录 5 中所述。

–3D 任务中的 2D 语义。2D语义包含可用于检测 3D 对象的有用信息。视觉场景的表示由点云、网格或体积组成,而 2D 图像由像素网格组成。结合两个领域的优点可以增强 3D 物体定位。研究人员试图通过采用不同的 RGB-D 数据表示方式来解决 3D 检测问题[34, 48, 23, 15, 41, 24]。通过将 2D 图像投影到 3D 场景上[34, 48, 24]或将 2D 特征与 3D 特征融合[23, 15, 43, 41], [33]。SAT [52]利用 2D 语义执行 3D VG 任务。但是,需要额外的 2D 输入限制了潜在的应用场景。相比之下,在本研究中,我们在训练阶段生成合成的二维图像,并将其用于推理阶段。合成图像使我们能够通过二维语义整合三维视觉流,而无需额外的二维输入。

3 环顾四周并参考（LAR）

LAR 旨在增强视觉模块,从而提高最终目标,即参考准确性。通过将合成的 2D 语义与 3D 点云结合起来,LAR 固有地捕获了 2D 和 3D 流之间的相关性。为了设计一个适合任何 3D 应用程序的通用框架,当输入是 3D 点云时,应该绕过对额外信息（例如 2D 语义)的需求。为此,引入了一个 2D 合成图像生成器,即 SIG,这将在第 3.1 节中讨论。然后,我们演示如何通过将我们的 SIG模块合并到我们提出的架构中,通过将生成的 2D 语义与 3D 语义对齐来学习鲁棒的表示,如第 3.2 节所示。

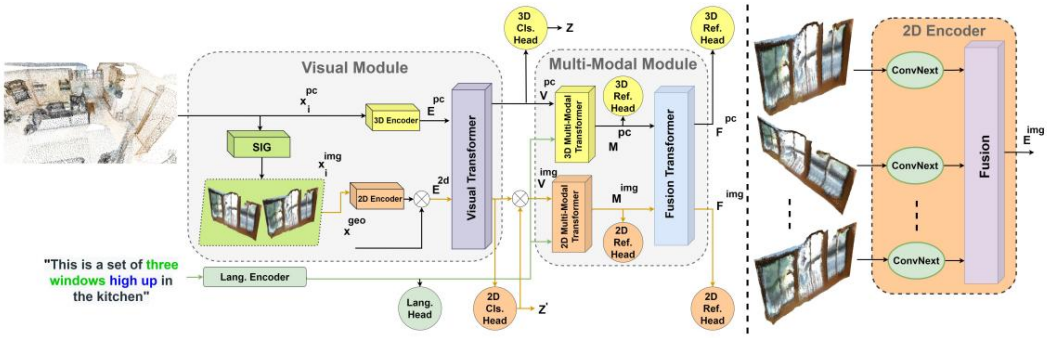

图 3:LAR 的详细概述。可视化模块使用 Visual Transformer 将提取的 3D 对象点的丰富 3D 表示与 2D 合成图像特征结合起来。2D合成图像首先由 SIG 提取,然后由共享的 ConvNext 主干进行处理。

同时,语言描述被转换成标记并嵌入到特征向量中。

然后,多模式模块通过两个 Transformer 分别融合视觉模块的输出。

最后,多模态变换器的输出由融合变换器 (Fusion Transformers) 处理。

3.1 环顾四周:2D 合成图像生成器

LAR 利用合成 2D 语义来强化学习到的 3D 表示。因此,本节将演示我们的 2D 合成图像生成器 (SIG) 模块,该模块将表示对象的任意 3D 点云投影到 2D 图像上。首先,我们介绍模块在单个图像上的基本操作,然后展示如何扩展它以实现多视图设置。

3D 场景,分别为 $S \in R^{N\times 6}$,用具有空间和颜色信息的N个点表示,即XYZ,和 RGB。根据前人的研究[2] [1] [37] [52],每个场景中的对象提议x都是已知的,要么由现成的 3D 对象检测器生成,$^{个人电脑}_{我}$要么手动注释。对象提议x是整个场景的一个子集,其中N < N且N 表示表示对象 i 的点数。

$$^{个人电脑}_{我}\ 6\in R^{N\times}$$

给定一个场景S和一个对象提案x代表对象i。主要思$^{个人电脑}_{我,}$ 我们的 SIG 模块生成 2D 合成图像x $^{img}_{我}$想是在馈送的对象提案前面放置一个虚拟摄像机,

然后将其点x投影到 2D 平面以生成相应的 2D ^{个人电脑}图像x相机的位置应仔细设置,以便为对象i 生成合适的图像表示。如图2 中的部分 (A) 所示,在从馈送的对象提案中^{img}排除上部和下部人脸后,我们得到剩余四个人脸的中心,然后计算每个人脸中心之间的距离Ol

^{我,}以及场景中心Os,其中l 的范围从 0 到 3。距离场景中心Os最近的面l是突出的面,在图 2 中用蓝色标记,这保证我们从房间内部看到物体i 。摄像机距离突出面l的距离为d f ,距离房间地板的上方距离为d ;因此,摄像机位置被解释为Camp = [d f df y , dup],如图 2 中的部分 (B) 所示。摄像机框架的方向定义如下。z 轴与摄像机的主轴对齐,另外两个正交方向 (x, y)与图像平面的相应轴对齐。当摄像机位于房间外时,图 5 显示了失败的情况。

x ,

要将点投影到x 4×4,变换矩阵M ∈ R ^{个人电脑}需要定义表示物体i 的从场景坐标到相机坐标的变换矩阵。给定相机位置Camp和突出脸部的中心Camp Ol ,变换矩阵M可以解释为,其中相机方向由三个向量定义; V f camera

使用统一相机模型 (UCM),它有五个参数I = [γx, γy, cx, cy, ζ]

V r V上升V f 0 0

0 1

, 电压 , 和V向上。我们定义我们的虚拟

按照 UCM,投影定义如下:π(x $^{个人电脑}_{我}$,我)= $\frac{\gamma x\ \delta d+zy\ yy}{\delta d+z} + {}^{障容}_{赛}$,

之后,将投影的 3D 点云分配到 2D 网格地图。为了减轻过度拟合,我们随机扩展了感兴趣的区域,如图 2 中的部分 (C) 所示。

点云不同于 2D 图像,2D 图像中的每个像素都包含空间和纹理细节。此外,与依赖于视图的图像不同,对象以视图不变的方式表示。为了缓解图像表示中的这一缺陷,构建了一个多视图设置,如图1 所示,其中v 个摄像机以圆圈形式安装在突出面l周围。算法 1 附在补充材料中,演示了我们的 SIG 模块的伪代码。

3.2 参考:Visiolinguistic Transformer 架构

我们展示了我们的方法（如图 3 所示）,用于学习与语言教学相关的 3D 场景的多模态表示。首先,所提出的视觉模块利用 2D 合成图像生成器（即第 3.1 节中描述的 SIG 模块)丰富了 3D 表示。得益于此,我们在训练或测试期间不需要访问任何图像传感器。

然后,多模态模块学习针对最终任务的联合表示,即根据语言指令引用场景中的对象。

–视觉模块。给定一个 3D 场景S ∈ R N×6 结合不可知的多个 3D 对象提案

电脑个人 6∈R N× ，我们的目标是通过预测语义标签z 来对每个对象提案进行分类。首先，

对于场景 S 中的每个对象i ,我们从 3D 对象的点云x生成多视图合成 2D 图像x ∈ R ,其中v是每个对象的视图数， H和W是生成图像的空间维度。其次,我们使用 PointNet++ [35]作为 3D ∈ R 1×dimpc编码器,产生 3D 特征E。因此,采用

Tiny-ConvNext [28]作为2D 编码器,产生y 。其中H和W表示空间特征的粗空间网格,每个特征用 dimimg 表示。然后,空间维度被压缩为o 。由于我们的 2D 编码器对同一个对象i 编码了多个视图v ,因此需要一种融合机制来融合视图的特征。我们探索了各种可能性来确定最佳的融合机制。我们可以将其解释为完全连接的层,以完全捕获视图和跨通道交互。相反,为了学习视图的交互并忽略跨通道关系,可以使用 1D-Conv。以避免像完全连接的使用ksize = v的 1D-Conv。作为融合层产生E img ∈ R 1×dimimg 情况那样涉及大量参数。

，其中ksize是核大小。融合的 2D 特征E img 和 3D

特征E i 连接起来并馈送到由多个堆叠的 Transformer 层组成的视觉转换器。视觉转换器用于捕获 2D和 3D 特征之间的互相关性,输出精细特征V pc和V img,由三个加权辅助损失（即 V pc和 V img之间的对齐损失)和两个视觉分类损失引导。

–2D 语义。受 SAT [52] 的启发,我们探索了不同的 2D 语义信息类型组合。我们将 2D 语义表述如下：

而2d = LN(ϕ(E 图片 , LN(W1x (1)其中， z是预测的类标签，W1是学习到的投影矩阵,LN 是层归一化， ϕ表示连接。我们使用不同的语义信息类型对每个对象单独编码,即视觉语义E img ∈ R 1×30。几何语义x i表示对象在场景空间中的边界框坐标,并对虚拟相机参数（即在生成 2D img图像x时使用的内在和外在参数)进行编码,以捕获对象之间的交互。相比之下,SAT [52]利用了真实的 2D 语义,其中z表示从图像传感器 [8] 捕获的真实 2D 图像生成的嵌入i 。

，和几何语义x 。因此,它可以被解释为视觉变换器的位置编码,其中

表示手动注释的类标签， E img

–多模态模块。使用预先训练的 BERT 模型[11] ,然后采用来自 Referit3D [2] 的语言分类头,我们嵌入了单词查询Q。使用上述视觉模块嵌入每个视觉V img特征后,我们使用两个不同的变换器同时融合输入模态Q、 V pc ;3D 多模态变换器和 2D 多模态变如图 3 所示。在 3D 多模态变换器中,我们联合编码两个输入模态V pc和Q ,以更好地理解捕捉场景中不同对象之间共 ， 换器，现关系的指称表达。同时,相同的单词查询Q与V img一起编码以增强上下文感知表示。采用来自 Referit3D 的单独 2D 和 3D 指称头,它们是辅助任务。最后,多模态变换器的输出由融合变换器处理,该变换器从融合特征 F pc和 F img中选择指称对象

。

3.3 损失

我们的架构称为 LAR,除了一个语言流外,还包含两个视觉流,即 3D 和 2D。为了正确地引用输入话语引导的对象,我们必须首先准确地对对象进行分类。因此,采用两个辅助损失来优化视觉模块;3D 3D分类损失L此外,我们采用了 Referit3D [2]辅助分类

和 2D 分类损失L cls 。 2D

语言分类损失L lang配合视觉基础任务,根据描述识别所指对象类别。因此,两个辅助基础损失（L

和L 二维 埃雷夫, 之前利用3D Eref （称为早期引用损失）。由于这些早期引用损失,该模型分别学习语言描述和视觉特征之间的相关性,从而提高了性能。此外,我们采用了 SAT 中的对象对应损失Lcor来鼓励 3D 模型从 2D流中提取知识。最后,在融合转换器之后添加了两个基础损失,即L和L来预测每个提案的分数。我们以端到端的方式优化整个模型,与SAT 不同,后者分别学习 2D 特征 E img ,具有以下损失函数：

在融合转换器

3D

$$L = \lambda cls \quad (L^{3D}_{分类} + L^{二维}_{cls}) + \lambda Eref \quad (L^{3D}_{埃雷夫} + L^{二维}_{Eref}) + \lambda ref \quad (L^{3D}_{考} + L^{二维}_{参ref}) + \lambda cor \quad L\,cor + \lambda lang \quad L\,lang \quad (2)$$

其中， λcls是对象分类损失权重， λEref是早期引用损失权重， λref是最终引用损失权重， λcor是对象对应性损失权重， λlang是语言分类损失权重。我们的实施细节详见附录 5

4 实验

4.1 数据集

我们在三个基准数据集Nr3D [2]、Sr3D [2] 和 ScanRefer [5]上评估了我们提出的 3D 视觉接地方法 LAR。Nr3D ,3D 自然参考数据集[2],由人类使用两个人之间的引用游戏收集的 41.5K 自然、自由形式的话语组成。相比之下,Sr3D [2] 包含 83,5K 合成话语。因此,ScanRefer为 800 个 3D 室内场景提供了 11K 个对象的 51.5K 话语。

4.2 评估指标

我们遵循了与前人研究[2] [52] 相同的指标,主要使用三个评估指标,即引用准确度、视觉分类准确度和语言分类准确度。对于 Nr3d 和 Sr3d 数据集,我们遵循前人研究的惯例,假设提案是由 GT 对象生成的。根据这一假设,3D 基础问题被重新表述为分类问题,其中模型目标是正确地将引用对象与其他对象进行分类,称为引用准确度。引用准确度是根据模型是否从一组 X 个提案中挑选出正确的提案来计算的。除了主要目标（即引用准确度）之外,还定义了两个辅助指标来分别评估每个流的性能,即视觉分类准确度和语言分类准确度。给定不可知的 GT 提案,视觉分类准确度是根据模型是否从预定义的一组可能类别中正确预测对象的类别来衡量的。同样,语言分类准确度是另一个评估语言分支性能的辅助指标。该分支旨在根据输入的话语正确预测所指对象的类别。

4.3 实现细节

– 输入配置。对于 3D 输入,我们从其点云段中随机抽取每个提案 1024 个点。此外,对于合成的 2D 输入,我们遵循算法 1,该算法在第 3.1 节中进行了仔细检查。在测试阶段,相机距离突出面部 2 米,即dfl = 2,距离地板 1 米,即dup = 1 。此外,感兴趣区域扩展比（如图 2 部分 (C) 所示）设置为 2%。生成的图像是一个大小为32 像素的正方形。每个对象生成五个视图,其中每个相机之间的角度θ为30° ,如图 1 所示。因此,相机设置为以下角度（ 0 °、30° ,

,

－30 ◦ 、60 ◦ 、－60 ◦ ）。

表 2:2D 合成图像分辨率的消融研究
以及 Nr3D 上的多视图设置 [2]。

图像尺寸	多-视图	编码器	参考3D [2]		非 SAT [52]	
			参考	账户 类别 账户	参考	账户 类别 账户
-	-	3D	35.6	57.4	37.7	59.1
32	1		28.0	47.34	30.9	49.31
64	1 2D 1 3		32.0	56.93	33.7	56.14
128			34.4	62.15	35.5	63.32
32			29.7	50.11	31.1	51.40
64	3 二维		32.9	63.20	34.2	64.02
128 3 32 5			35.1	64.85	36.4	65.13
64			31.1	55.18	33.7	56.72
	5 二维 128 5		34.1	64.05	35.6	64.52
			36.0	65.19	38.2	65.40

表 3:不同类型 2D
语义信息如Nr3D上的等式1所示[2]。

+x地理+x	cls +xROI	SAT † [52]			我们的		
		参考	账户 类别 账户		参考	账户 类别 账户	
(a) - 37.7 (b) - - ✓ 42.3 (c) ✓ - ✓ 46.2 (d) - ✓ ✓ 43.2 (e) ✓ ✓ ✓ 47.3		59.1	38.1	58.2			
		61.0	42.2	62.2			
		61.0	46.3	63.3			
		61.2	44.5	62.1			
		60.6	48.9	65.42			

– 网络和训练配置。在 ImageNet 数据集[9]上预先训练的模型Tiny- ConvNext [28]用于提取我们合成的 2D 图像的特征。我们的 2D 主干共享
在不同视角下。我们采用与之前研究相同的 3D 编码器[1] [2] [13] [37] [18]
[52]、 PointNet++ [35],对每个提案的采样点进行编码。对于语言编码器,我们
采用三层的 BERT 编码器[11] 。我们采用了与
[52] [51]为我们的视觉、多模态和融合转换器。所有模型都经过 100 个 epoch 的训练
从头开始,使用[14]中描述的权重初始化策略。如 SAT [52] 中所述,
初始学习率设置为10−4 ,每 10 个 epoch 减少 0.65。 Adam 优化器[21]
并且每个 GPU 的小批量大小为 32,用于训练我们的所有模型。我们将损失权重设置为
因此, λcls = 5、 λEref = 0.5、 λref = 5、 λcor = 10 和λlang = 0.5。

–软件和硬件细节。我们使用 Python 增强框架 Albumentations [4],
用于图像增强。为了在训练时进行在线增强,我们实现了
我们在第 3.1 节中描述的 2D 合成图像生成器 (SIG),采用纯 Python 方式
无需使用任何 3D 库,这也加快了训练速度。我们的模块实现于
使用 PyTorch 框架和四个 Nvidia V100 GPU 的 Python。

4.4 消融研究

我们进行了五项消融研究,以评估所提出的架构并验证贡献
每个组件都应具有良好的性能。这些组件是 1) 2D 多视图和 2D
分辨率。2)2D 语义信息的类型。3)相机与图像增强 4)SIG 模块
与 2D 真实线索相比。5) 外部和内部噪声相机。附录 5 中列出了更多消融。

–2D 多视图和 2D 分辨率。为了评估我们的 2D 合成图像生成器 (SIG),我们
故意删除了 3D 流,在图 3 中用黄色标出;3D 编码器、3D 头和依赖
仅在我们生成的 2D 图像上处理引用任务。因此,我们的架构是
简化为仅包括 2D 编码器、2D 多模态转换器和 2D 头。我们训练了
此 2D 架构变体具有不同的图像分辨率和多视图设置组合。
如表 2 所示,随着图像尺寸的增加,引用准确率呈线性增加,或者
增加了投影图像的视图数量。但是,我们的多视图设置具有
占主导地位,而仅使用 64 幅图像分辨率,五个视图超过一张照片
128 分辨率。此外,我们发现使用五种不同视图的组合,每个视图都有 128
分辨率,超越了 Referit3D 和非 SAT 变体,它们都只使用 3D 语义,
通过我们新颖的合成 2D 图像生成器 (SIG) 展示学习到的表示的质量。

–2D语义信息的类型。受SAT [52]的启发,我们分析了不同2D
语义,遵循公式 1。如表 3 所示,第一行 (a) 代表我们
架构中仅包含 3D 编码器、3D 多模态变换器和 3D 头。
因此,该变体模仿了 Non-SAT [52]配置。从 b 行到 e 行,我们的
合成的二维语义与三维语义结合在一起。此外,我们还与 SAT 进行了比较
[52]强调每个语义信息含义上的关键差异。我们的 ROI
架构是使用我们的 SIG 模块从 3D 点云生成的合成 2D 图像。在
相比之下,SAT [52]在训练阶段使用原始 ScanNet 图像。与 SAT 不同,SAT
在训练期间使用真实分类标签,并在测试期间屏蔽它们,我们使用

表 4:相机增强与
Nr3D 上的图像增强[2]。

访问权限	类别访问权限		
		45.9	60.33
		48.2	62.69
	✓	47.8	63.97
✓✓	✓	48.9	65.42

表 5:我们的 SIG 模块与
SAT 在 Nr3D 上使用的真实 2D 线索 [2]。

方法。	训练推理摄像头 8 月图片 8 月参考		参考帐户
SAT 真实 2D 图像 47.3			
SAT 真实 2D 图像 真实 2D 图像 47.9			
LAR	说	说	48.9
LAR 真实 2D 图像 真实 2D 图像 46.8			

对相机的稳健性
增强。

培训测试参考权限	类别权限		
		48.2	62.69
		46.7	61.32
	✓	48.9	65.42
✓✓	✓	48.8	65.22

表 7:基准 - 表 6:显示 LAR
表 8:基准测试结果
做[5]。

方法参考 Acc.
改装3D 46.9%
非 SAT 48.2%
SAT 53.8%
SAT † 52.3%
我们的54.6%

在 Nr3D [2] 上,同时使用
地面真实分类。

方法参考	访问类别	访问
SAT-GT 50.3% 61%		
我们的62% 64%		

来自二维分类头z的预测标签。因此,我们不需要掩盖
2D 语义,因为我们使用的是合成语义。如 (e) 行所示,实现了最佳准确率
当我们将这三种语义类型结合在一起时,我们还注意到
分类准确率,其中我们的合成二维图像在两个阶段都协助三维分类器;训练
和测试。

–相机与图像增强。我们在训练过程中有两种增强:1)图像
增强。2)相机增强。

对于图像增强,我们遵循了主流计算机
视觉应用,如水平和垂直翻转、模糊、颜色抖动、随机缩放、随机
移动和随机裁剪。对于相机增强,我们将d f 的范围设定为 1.5 至 4 米， d向上
从 0.5 米到 2.5 米,从 2% 到 15%。通过改变上述变量,
增强了相机的外部参数。此外,还向相机的本征参数中添加了随机噪声
参数。表 4 展示了每种增强类型单独使用的效果以及融合的效果
两者都有。利用这两种增强类型可以实现最佳准确度。

–SIG 模块与 2D 真实线索。

为了验证我们的 SIG 模块,我们定制了我们的架构变体,其中我们的 SIG 模块
被 Faster R-CNN [39]产生的实际二维线索所取代。如表 5 所示,用我们的
使用 Faster R-CNN [39] 生成的特征的 SIG 模块会降低 2% 的性能。在
此外,即使我们与使用额外二维信息的方法进行了不公平的比较,
SAT 的一个变体,在训练和测试过程中使用真实图像（第 2 行）,我们仍然比它表现更好
1%。

–相机噪声外部和内部。由于利用
相机增强,如表 4 所示,进行了详细的分析,以探索
相机变化的鲁棒性。因此,我们的 2D 合成分支学习如何融合
以数据驱动的方式将信息传输到不同的摄像头。因此,我们可以训练模型
对噪声相机模型具有鲁棒性,如表6所示。

相机死亡和噪声参数实验都可以被认为是无关紧要的,因为 LAR 利用
虚拟相机。因此,不会引入相机丢失或噪音。但是,我们的 SIG 模块
如第 3.1 节所述,当这些稳健性
抗噪声特性对于这些应用至关重要。例如,在自主
驾驶世界,我们的 SIG 模块可用于将密集的激光雷达点云转换为合成
图像,这些图像将用于训练任意神经网络,以处理特定任务（例如,物体
在 Cocoon 设置中分割[31] [47]）。然后,将使用真实相机而不是
从激光雷达生成合成图像输入。

表 9:Nr3D 和 Sr3D 数据集上的基准测试结果[2]。所有报告的技术均使用真实数据
对象建议。附加输入列强调除了对象点云之外使用的额外信息，
其中Spc和2DImg分别表示整个场景点云和额外的 2D 图像。 †符号
表示我们重新训练模型以进行公平的比较。

方法	额外的输入	没有3D				三维Sr			
		总体(σ)简单 困难 与视图相关 与视图无关	总体(σ)简单 困难 与视图相关 与视图无关						
ReferIt3D [2]	-	35.6% 43.6% 27.9% 32.5% 37.3% 44.2%	37.1%	40.8%	44.7% 31.5% 39.2% 48.5%		40.8%		
文本引导的 GNN [17]	-	30.6% 35.8% 38.8% 46.0% 31.8% 34.5%	38.0%	45.0%	36.9% 45.8% 51.1% 40.5%		45.0%		
实例参考[58]	特殊控制	39.0% 46.4% 32.0% 34.7% 40.8% 48.5%	41.9%	48.0%	45.4% 50.7% 38.3% 44.3%		48.1%		
3DRefTransformer [1]	-	34.8% 34.8% 41 .7% 48.2% 35.0% 37.1%	41.2%	47.0%	54.2% 44.9% 44.6%		47.1%		
3DVG-Transformer [59]	-	42.1% 48.5% 36.0% 36.5% 43.9% 51.0%	43.7%	51.4%			51.7%		
FFL-3狗 [12]	-	36.6% 41.7% 37.7% 44.5% 31.2% 34.1%	44.7%	-	-	-	-	-	-
转诊3D [13]	-	49.2% 56.3 42.4 47.3% 55.8% 41.4%	44.9%	57.4%	60.5% 50.2% 49.9% 58.9%		57.7%		
语言参考 [37]	-	46.9% 48.9%±0.2 58.4% 42.3% 47.4%	45.0%	56.0%	49.3% 49.2%		56.3%		
非 SAT [52]	-		39.5%	43.9%	-	-	-	-	-
星期六 [52]	2D毫克	46.9	50.4	57.9%	61.2 50.0	49.2	58.3		
SAT † [52]	2D毫克		50.4%	56.6% 60.6% 49.7% 48.7% 59.35%±0.1			57.4%		
LAR （我们的）	-		52.1%	63.0% 51.2% 50.0%			59.1%		

4.5 与最先进技术的比较

此外,我们在三个著名的 3D 视觉接地基准（即 Nr3D [2]、Sr3D [2] 和 ScanRefer [5] ）上取得了最佳效果。通过有效利用我们的 2D 合成图像，
LAR 的表现优于所有不使用任何额外训练数据集的现有方法，
Nr3D 和 Sr3D的差距较大,分别为 5.0%和 1.9% ,如表 9 所示。其中，
LanguageRefer [37]达到了 43.9%,而 LAR 在 Nr3D 上达到了 48.9%。对于 Sr3D,我们的模型
达到 59.35%,比 TransRefer3D [13]提高了 1.9% 。此外,对于 ScanRefer
[2]基准,我们也取得了最佳成果,准确率比
最接近的对应架构;SAT [52],如表 7 所示。

为了进行公平比较,我们重新训练了 SAT [52] ,因为我们注意到复制和验证 SAT 存在困难
结果。虽然[1]我们在训练过程中不使用额外的输入,但 SAT 使用的是原始的
ScanNet 2D 图像[8],我们在 Nr3D、Sr3D 和 ScanRefer 上的SAT 分别超过了 1.6%、 2.7%和 2.3% ，
分别。

4.6 讨论与分析

首先,我们将展示我们的定性分析。然后,我们进行了两个实验来探究
通过我们提出的方法 LAR 学习到的表示的稳健性。

–定性分析。图 5 描述了我们的投影模块 (SIG) 中的两个极端情况。虽然
将投影的三维点分配到二维网格中,多个点可能共享精确的空间
网格上的位置,这意味着只有一个点将存储在网格中,其余的点将
丢弃。如图 5 中的部分 (A) 所示,在左列中,地板与其他物体重叠
这是由于在分配步骤中点之间的冲突造成的。为了克服这种行为,一个简单的
根据每个点的高度给出优先级。B 部分显示了另一个失败案例,其中
突出的脸部被错误地定位,导致将相机放在房间外面或里面
其他物体。因此,我们将突出的脸部定位在房间中心,以缓解这种
效果。图 6 的 A 部分显示了一个模糊描述的例子。该模型错误地
预测“右边的椅子,而不是左边的椅子”。模型和人工注释
有时会与视图相关描述中的对象产生混淆。
成功案例可以在图 6 B 部分中看到。有关更多定性可视化,请参阅
附录 5.

–使用真实类别标签进行评估。

表 8 显示了另一种变体,它是根据我们的架构量身定制的
与 SAT-GT 进行比较。预测的类别标签为
被真实数据取代。我们的 LAR 架构表现优于
SAT-GT 提高 12%。

–零镜头相机测试失败。我们故意将相机掉落
在测试过程中验证我们所学表征的稳健性

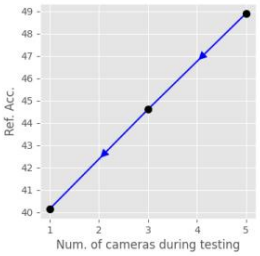

图 4:相机丢失。

1参考官方 SAT 实施中的第 1、3 和2 个问题。

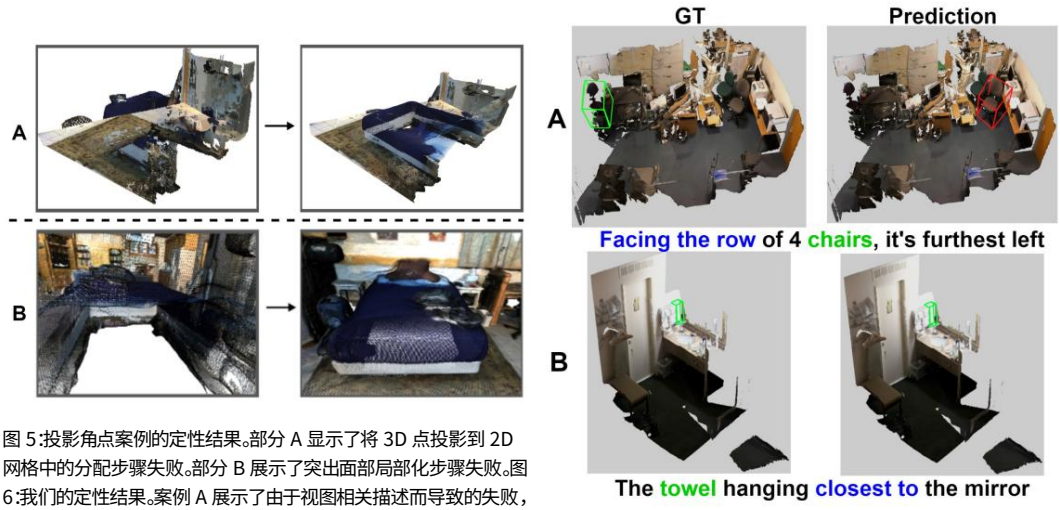

图 5:投影角点案例的定性结果。部分 A 显示了将 3D 点投影到 2D 网格中的分配步骤失败。部分 B 展示了突出面部局部化步骤失败。图 6:我们的定性结果。案例 A 展示了由于视图相关描述而导致的失败，而案例 B 展示了成功案例。

对抗相机变化。如图 4 所示，从使用五台相机训练的最佳模型开始，当我们在测试期间仅使用三台相机和一台相机时，参考准确度分别下降到 44.6% 和 40.15%。受这些结果的驱动，可以调用相机丢弃技术来克服性能下降，方法是在训练期间故意丢弃一个随机相机；用另一个相机的重复视图替换其输入。

– 局限性。按照 Referit3D [2] 的标准设置，我们假设可以访问每个 3D 场景的对象提议，类似于 SAT [52]、 LanguageRefer [37]、TransRefer3D [13]和 3DVG-Trans [59]。因此，我们在实验中使用了地面实况提议。根据这一假设， 3D 地面问题被重新表述为分类问题，其中模型目标是正确地将所指对象与其他干扰项进行分类。但是，LAR与使用基于检测的提议的设置兼容。

5 结论

我们提出了一种高效的 3D 基础架构，利用 2D 知识来强化学习到的 3D 表示。为了避免在训练和测试期间都不需要 2D 语义，我们引入了一个 2D 合成图像生成器，称为 SIG。我们的 SIG 模块可生成高质量的 2D 合成图像。LAR，即 3D 视觉基础的 2D 合成语义知识提炼，由两个阶段组成，即视觉和多模态模块。LAR 保留了将合成的 2D 知识提炼到 3D 流而无需任何额外信息的优势。

我们的实验表明，LAR 的性能显著优越，在三个不同的 3D 接地基准上取得了最佳结果，远远超过不依赖额外训练数据的现有方法，在 Nr3D、Sr3D 和 ScanRefer 上分别高出 5.0%、 1.9%和 2.3% 。因此，尽管与使用额外 2D 图像的 SAT 相比存在不公平的比较，但 LAR在 Nr3D、Sr3D 和 ScanRefer 上分别比 SAT 高出 1.6%、 2.7%和 2.3% 。

此外，我们通过对相机内在和外在参数以及即将死亡的相机的零样本实验验证了 LAR 的鲁棒性及其泛化能力。补充材料中附有更多定性结果。

致谢和资金披露

这项工作由 KAUST BAS/1/1685-01-0 资助，并得到 SDAIA- KAUST 数据科学和人工智能卓越中心 (SDAIA-KAUST AI) 的部分支持。