面向城市场景建筑的 3D 视觉定位数据集

摘要

3D 视觉引导定位 (3D Visual Grounding) 旨在将自然语言描述与三维场景的对象进行匹配，从而实现对特定对象的精确定位。目前用于三维视觉接地的数据集和方法主要侧重于室内场景中识别和定位对象。随着大规模三维城市场景数据集的创建，为语言表达与城市组成部分相结合提供巨大的潜力，尤其是在文本引导的自动驾驶和无人机导航等领域应用。然而，城市级的 3D 视觉定位作为人类交互式理解城市的基础任务仍处于早期发展阶段。为实现上述的目标，我们引入一个新的数据集 UrbanRefer，用于城市实例级的视觉定位。与室内场景大量的文本注释不同，城市规模注释的稀缺性和复杂性都为实现这些应用带来了巨大的挑战。

关键词： 3D 视觉定位；城市点云；多模态融合

1 引言

在当今数字化时代，3D 视觉引导定位领域已成为众多前沿科技应用的核心支撑技术之一。随着人工智能、自动驾驶、无人机技术以及增强现实等领域的迅猛发展，3D 视觉定位的重要性愈发凸显。它不仅是实现人机自然交互的关键环节，更是智能系统在复杂三维环境中精准感知、决策和行动的基础。例如，在自动驾驶场景中，车辆需要精确理解周围的 3D 城市场景，通过视觉定位技术将自然语言指令与实际的建筑物等对象进行匹配，从而实现安全、高效的导航；在无人机应用中，无人机需要根据操作人员的自然语言指令准确识别并定位目标建筑物，完成诸如监测、救援等任务。

当前，现有的 3D 视觉定位方法主要集中在室内场景的研究上，其中 ScanRefer [2]、Sr3D 和 Nr3D [1] 等数据集为室内场景的相关探索提供了重要基础。研究人员在这些数据集上进行了多方面的深入研究，如引导分割、多物体定位以及顺序引导等，取得了一定的成果。然而，随着现实应用场景的不断拓展，城市场景下的 3D 视觉定位需求日益迫切，而现有的研究在这方面存在明显不足。因此，我们引入一个新的数据集 UrbanRefer，用于城市实例级的视觉定位，如图 1所示。

UrbanRefer 数据集的自动生成过程涵盖视觉属性标注与空间关系描述。在视觉属性标注时，利用 Blender 软件对 UrbanBIS [13] 建筑网格数据渲染多角度图像，输入 ChatGPT 并以特定问题引导生成建筑颜色、纹理和形状等描述，实现半自动标注以提升效率、降低成本。对于空间关系描述，划分八个方位并结合模糊距离词，构建建筑物关系图，依预设尺度筛选参考建筑，按单参考模板生成空间表达，保留多参考建筑丰富细节，最终为每栋建筑生成精准全面的空间关系描述，为 3D 城市场景引导定位提供数据支撑。我们的贡献如下：

- 我们引入了一个新的面向大规模城市场景的建筑物视觉定位数据 UrbanRefer。
- 我们将基于室内的 ScanRefer 方法迁移到 UrbanRefer 数据上，进行了实验性的探索，为下一步解决 3D 城市场景引导定位任务打下坚实的基础。

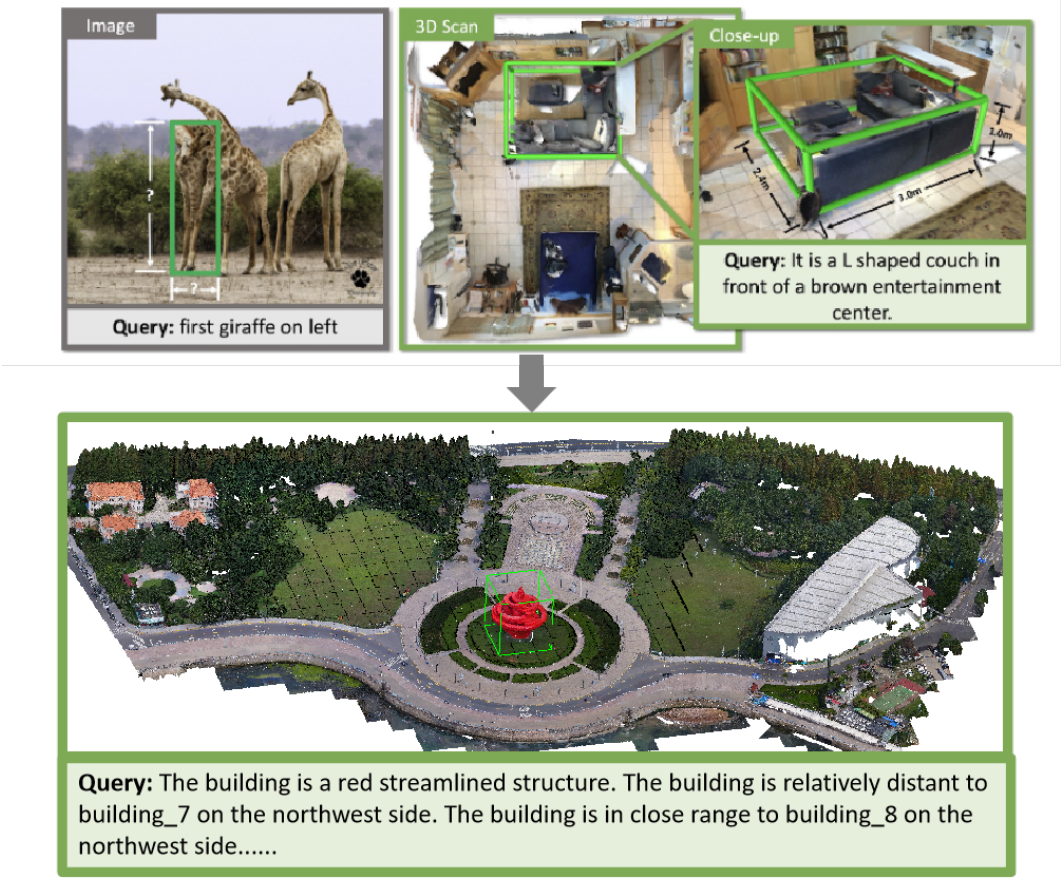

图 1. 现存的 3D 视觉定位数据集通常只包含小尺度的室内场景。为了将室内 3D 场景引入到真实的城市场景，我们将引入了一个新的面向大规模城市场景的建筑物视觉定位数据 UrbanRefer，专注于大规模城市场景中的建筑物视觉定位。给定一个 3D 城市场景和一个自然语言表达式作为输入，预测对应目标建筑物的边界框。

2 相关工作

2.1 视觉定位数据集

2D 视觉定位数据集。我们简要描述了现有的 2D 引导定位的经典数据集。例如，RefCOCO [6]、RefCOCO+ [6]、RefCOCOg [8] 采集于 MSCOCO 图像之上，这些图像全部由小尺度的 2D 自然图像组成。RefCOCO 和 RefCOCO+ 的表达式和包围盒是通过两人博弈生成的。RefCOCOg 是由 MTurk 工人在非交互式环境下收集的。这三个数据集使用单一的表达式来查找对象。RSVG [11] 是一个基于遥感图像的可视化接地数据集。自然图像通常是不具备空间信息的，无法显示广泛的地面信息。而遥感图像可以在非常广泛的范围内捕获地物目标。

3D 视觉定位数据集。 ScanRefer [2] 和 ReferIt3D [1] 从真实世界数据集 ScanNet 中引入了两个由 3D 对象的语言描述组成的数据集。具体来说，ReferenceIt3D 既包含基于对象间空间关系生成的基于模板的描述 (Sr3D)，也包含人工标注的细粒度描述 (Nr3D)。CityRefer [9]

介绍了具有城市尺度 LiDAR 点云及其基线的 3DVG 数据集。然而，它们有一个局限性，即必须将点云分割成更小的子集，以便在单一场景内进行处理。MSSG [3] 是一个开创性的基于 LiDAR 的大规模室外场景三维 VG 模型。然而，由于其粗粒度的跨模态融合和缺乏对上下文信息理解的考虑，它遇到了局限性。

2.2 3D 视觉定位

首先，利用 3D 物体检测器生成候选 3D 物体候选框，而文本特征由语言模型编码。其次，这些先进的方法在对齐目标物体的视觉和文本特征方面进行了关键的努力。其中，TGNN 利用图神经网络推断 3D 物体建议之间的空间关系，并通过文本特征进一步丰富。FFL-3DOG 通过场景图交互捕获文字描述和三维点云的模态内和跨模态关系。最近的努力包含了 Transformer 来建模提案与文本特征的关系。3DVG-Transformer [14] 和 TransRefer3D 通过建立上下文感知的自注意力和交叉注意力协议来实现出色的性能 LanguageReference [10] 无缝地将跨模态任务转换为统一的语言建模挑战，并以预测的对象标签为支撑。

然而，两阶段方法遇到了一个值得注意的障碍：检测阶段忽略了利用语言上下文来优先考虑对指称任务至关重要的对象。解决该问题的另一种方法是单阶段方法，该方法将提取的视觉特征与语言特征直接融合以实现物体的接地。例如，3D-SPS [7] 利用文本特征来指导视觉关键点的选择，从而实现渐进的物体接地。但是 BUTD-DETR [5] 提出了一种类似 DETR 的模型，该模型通过联合注意力机制融合视觉-文本特征，然后从上下文特征中解码出话语中的对象。在此基础上，EDA [12] 提出了一种文本解耦策略，对 3D 物体和相关文本进行密集对齐。

尽管这些方法呈现出了卓越的性能表现，然而其适用范围却被局限于 3D 室内场景之中。当尝试将这些方法迁移到城市场景时，面临着诸多挑战。

3 UrbanRefer 数据集

现存的 3D 视觉定位数据集通常只包含小尺度的室内场景。本文创建了一个新的面向大规模城市场景的建筑物视觉定位数据集 UrbanRefer，专注于大规模城市场景中的建筑物视觉定位。该数据集基于 UrbanBIS 数据集构建，UrbanBIS [13] 是一个大规模三维城市数据集，提供了丰富的城市对象标注，如建筑物、车辆、道路、植被和桥梁等。特别值得一提的是，UrbanBIS 对每栋建筑提供了实例级标注，极大地丰富了不同建筑物类型的形状多样性。建筑作为城市场景中的核心组成部分，具有高度的识别性和多样性。因此我们选择将建筑作为城市场景中的定位对象，为城市场景的理解和分析提供坚实的基础。我们对 UrbanBIS 中的 3370 个建筑进行文本描述，对每栋建筑创建一个包含建筑纹理、形状和空间关系的综合描述。接下来，我们将描述我们的数据收集流程，并对数据集进行统计分析。

在这一部分中，我们将介绍 UrbanRefer 数据集的收集流程。该流程主要包括两个部分：视觉属性和空间关系，分别如图2和图3所示。

3.1 视觉属性收集

建筑的颜色、纹理和形状等视觉属性是识别和区分不同建筑的关键。受 ChatGPT 在语言标注方面取得突破性进展启发，我们设计了一种半自动的方法对建筑进行预标注。为了使

模型聚焦于建筑的颜色和形状，我们利用 Blender 软件对 UrbanBIS 提供的每栋建筑的网格数据进行渲染，生成四个固定视角图像和一个俯瞰图。然后将这些图像作为视觉语言模型的输入，以生成详细的描述。我们通过提出针对性问题（“What is the color of this building?”和“What is the shape of this building?”）来引导模型，分别针对固定视角图像和俯瞰图，从而自动生成关于建筑颜色和形状的具体描述。通过这种方式，我们能够自动生成对建筑的描述和标注，与传统的人工标注相比，可以大幅度降低成本并提高效率。

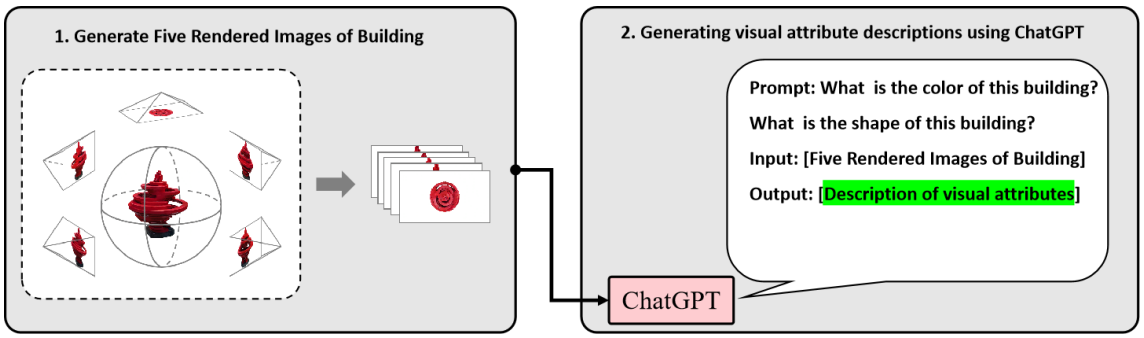

图 2. 视觉属性收集流程

3.2 空间属性收集

空间关系描述了城市场景中建筑物之间的相互联系和排列模式，这种关系对于在场景中寻找目标建筑非常关键。众所周知，我们通常通过描述目标建筑及其周围的参考建筑，以及它们之间的空间关系唯一地定位目标建筑。然而，城市场景中使用的空间关系与室内场景中的空间关系有很大的不同。例如，在室内场景中，我们习惯于使用左、右描述物体的关系。然而，在城市场景中，我们更倾向于使用东、西等真实的方位，并联合不同的距离尺度来描述建筑物间的空间关系。为了更好的细化方位信息，我们将空间方位分为 8 个固定方向：东、西、南、北、东南、东北、西南和西北。同时，根据不同的距离尺度，将建筑物之间的距离描述为“非常接近”、“接近”、“相对较远”。这种模糊的相对推理关系，更贴近人们在实际环境中的感知体验。

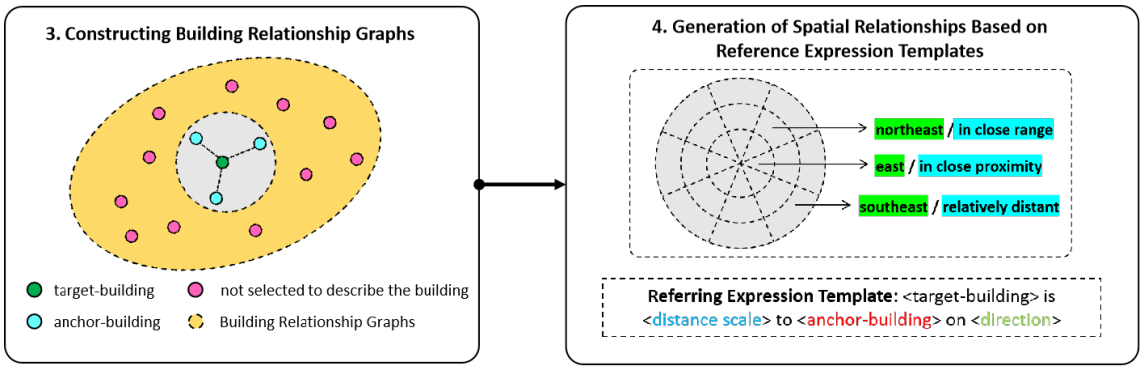

图 3. 空间关系收集流程

空间关系的参考模板是构建我们数据集的关键，负责为两两建筑物之间生成对应准确的空间关系表达。为了获取场景中准确的空间关系，我们设计了一个自动化的程序并依据参考模板来生成每对建筑物之间的空间关系表达。首先，我们会构建一个建筑物关系图，用于表述建筑物之间在场景中的相对位置关系。在这个图中，每个建筑物都被赋予一个独一无二的

ID，以确保标识的准确性。接着，我们根据预设的距离尺度（例如 50 米、150 米、300 米）筛选出每栋建筑周围符合条件的全部参考建筑。确保我们考虑到了不同距离尺度下建筑物之间的空间关系。最后，根据以下的单参考模板逐一生成空间表达。

单参考模板是指依靠一个参考对象来定位目标对象。这种表达方式是一种简单而有效的表达方式，它仅涉及一个参考建筑。例如，一个空间表达为：该建筑与西侧的 3 号建筑距离很近 <This building is in close range to building_3 on the west side.>。building_3 是一个参考建筑。in close range to 表示距离尺度。west 表示空间关系。

为了应对复杂的城市场景，我们保留了所有符合条件的参考建筑，这允许我们在描述中包含更多的空间细节和复杂性。例如，我们可能会描述一个建筑不仅与一个特定的参考建筑接近，还可能与多个建筑在不同方向和距离尺度上有所关联。这样的描述能够提供更全面的视角，帮助我们更好地理解和导航城市空间。通过这种方式，我们能够为城市场景中的每栋建筑生成一个详细的空间关系描述，这些描述不仅包括它们与周围建筑的相对位置，还包括它们之间的距离和方向。

3.3 数据分析

为深入理解 UrbanRefer 数据集，我们进行了全面的数据分析。在 UrbanRefer 数据集中，我们对 3370 个建筑实例进行了全面的数据分析。每栋建筑对应一条唯一的描述。为了分析我们数据集的场景和收集方式等信息，我们对近几年经典数据集的对比。同时，我们在这里提供了两个直方图，如图 4所示，分别对视觉属性和空间关系的文本描述长度进行分析。

Dataset	Year	3D Scene	Task	Text Source	Avg.Text Len
ScanRefer	2020	Indoor Scene	VG	Human	20.27
Nr3D	2020	Indoor Scene	VG	Human	11.5
Sr3D	2020	Indoor Scene	VG	Template	9.7
Multi3DRefer	2023	Indoor Scene	multi-VG	Template + GPT-4	15.1
SG3D	2024	Indoor Scene	SG	GPT-4	70.5
UrbanRefer	2025	Outdoor Scene	VG	GPT-4 + Template	299.41

表 1. UrbanRefer 数据集的对比分析

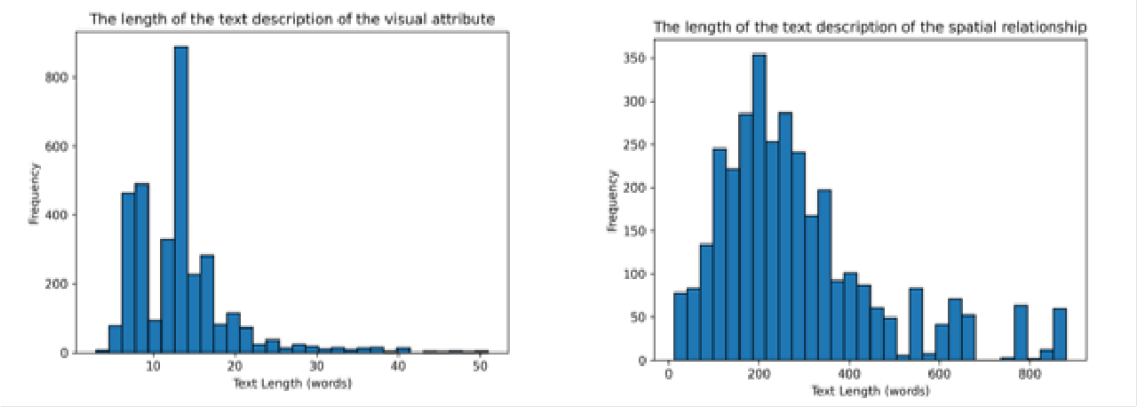

图 4. 文本的平均长度

4 方法

ScanRefer [2] 是一个 two-stage 的模型，第一阶段进行 3D 物体检测和文本的编码，第二阶段进行模态间的融合定位。原始点云在输入时会有两种不同的操作，一种是纯 3D，把点云 XYZ+RGB 作为输入，另外一种则是 2D+3D 形式，采用预训练的 ENet 提取图像特征，并将其投影到点云中，使视觉外观融入到点云中。ScanRefer 的 Pipeline 如图 5所示：

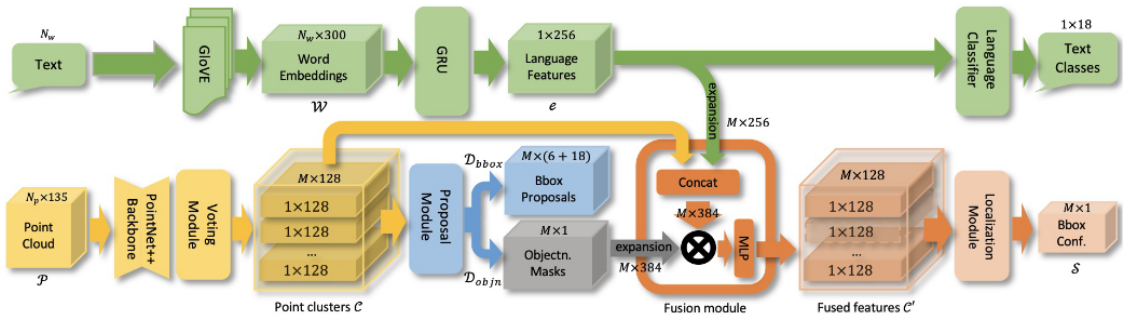

图 5. 方法示意图

检测模块：网络首先采用 PointNet++ 架构作为主干网络从输入的 3D 点云中生成目标物体的候选区域，然后，通过 VoteNet 对点云特征进行聚合，并输出候选物体的边界框及其类别预测。

编码模块：输入的自然语言描述经过 GloVE 嵌入和 GRU 网络编码为语言特征向量。

特征融合模块：将几何特征（3D 点云）和语义特征（语言描述）相结合（实际上是对两种模态的特征进行拼接），通过多层 MLP 对候选区域进行特征增强，同时使用检测模块的目标性预测掩码以滤除无效候选框。

定位模块：旨在预测所提出的边界框中的哪一个与描述相对应。通过对融合后的特征进行评分，网络选出与语言描述最匹配的目标物体的 3D 边界框。

损失函数如下：

$$\boldsymbol{L_{det} = L_{vote-reg} + 0.5L_{objn-cls} + L_{box} + 0.1L_{sem-cls}}$$

其中 $L_{vote-reg}$、$L_{objn-cls}$、L_{box} 和 $L_{sem-cls}$ 分别表示投票回归损失、目标性二分类损失、边界框回归损失以及针对 18 个 ScanNet 基准类别的语义分类损失。

5 复现细节

ScanRefer 已开源，但并不能直接应用于 UrbanRefer 上。例如，在对文本编码和分类时，采用了 18 个 ScanNet [4] 基准类别（室内常见物体类别）的语义分类损失，这并不能直接迁移到城市场景中。因此，我们按照 UrbanBIS 的语义类型依据建筑物的细粒度划分成商业、住宅、办公、文化、交通、市政、临时、未分类等。同时，我们针对 UrbanBIS 进行了一系列的改进和权重生成，如点云 RGB 的均值、建筑物类别的均值大小等诸多权重。

6 实验结果分析

从这些数据可以看出，ScanRefer 在 UrbanRefer 数据集上的性能还有较大的提升空间。无论是验证集还是测试集，数值都相对较低，反映出模型在准确性方面存在一定的局限性，可能存在过拟合、特征提取不充分或模型架构不够优化等问题。

Method	Val_acc@0.25	Val_acc@0.5	Test_acc@0.25	Test_acc@0.5
ScanRefer	21.72	6.72	25.24	8.61

表 2. ScanRefer 在 UrbanRefer 上的实验结果

7 总结与展望

本研究围绕 3D 视觉定位领域展开，随着相关技术发展，3D 视觉定位愈发重要，但现有研究多侧重室内场景，城市场景研究不足。为此引入 UrbanRefer 数据集，基于 UrbanBIS 构建，涵盖丰富城市对象标注，以建筑为定位对象，为城市场景理解提供支撑。同时，将室内的 ScanRefer 方法迁移至 UrbanRefer 数据进行实验，虽取得一定成果，但从实验结果来看，模型在准确性方面存在局限性，在未来我们将继续探索针对城市场景的引导定位的方法以解决可能存在的问题。

参考文献

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020.

[2] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[3] Wenhao Cheng, Junbo Yin, Wei Li, Ruigang Yang, and Jianbing Shen. Language-guided 3d object detection in point cloud for autonomous driving. *arXiv preprint arXiv:2305.15765*, 2023.

[4] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5828–5839, 2017.

[5] Ayush Jain, Nikolaos Gkanatsios, Ishita Mediratta, and Katerina Fragkiadaki. Bottom up top down detection transformers for language grounding in images and point clouds, 2021.

[6] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014.

[7] Junyu Luo, Jiahui Fu, Xianghao Kong, Chen Gao, Haibing Ren, Hao Shen, Huaxia Xia, and Si Liu. 3d-sps: Single-stage 3d visual grounding via referred point progressive selection. *arXiv preprint arXiv:2204.06272*, 2022.

[8] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016.

[9] Taiki Miyanishi, Fumiya Kitamori, Shuhei Kurita, Jungdae Lee, Motoaki Kawanabe, and Nakamasa Inoue. Cityrefer: geography-aware 3d visual grounding dataset on city-scale point cloud data. *arXiv preprint arXiv:2310.18773*, 2023.

[10] Junha Roh, Karthik Desingh, Ali Farhadi, and Dieter Fox. LanguageRefer: Spatial-language model for 3d visual grounding. In *Proceedings of the Conference on Robot Learning*, 2021.

[11] Yuxi Sun, Shanshan Feng, Xutao Li, Yunming Ye, Jian Kang, and Xu Huang. Visual grounding in remote sensing images. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 404–412, 2022.

[12] Yanmin Wu, Xinhua Cheng, Renrui Zhang, Zesen Cheng, and Jian Zhang. Eda: Explicit text-decoupling and dense alignment for 3d visual grounding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[13] Guoqing Yang, Fuyou Xue, Qi Zhang, Ke Xie, Chi-Wing Fu, and Hui Huang. Urbanbis: a large-scale benchmark for fine-grained urban building instance segmentation. In *ACM SIGGRAPH*, pages 16:1–16:11, 2023.

[14] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3DVG-Transformer: Relation modeling for visual grounding on point clouds. In *ICCV*, pages 2928–2937, 2021.