武士：形状感知的多模态检索用于 3D 物体识别

Dinh-Khoi Vo *[1,2], Van-Loc Nguyen *[1,2], Minh-Triet Tran [1,2], and Trung-Nghia Le **[1,2]

[1]University of Science, VNU-HCM, Ho Chi Minh City, Vietnam
[2]Vietnam National University, Ho Chi Minh City, Vietnam
{vdkhoi, nvloc}@selab.hcmus.edu.vn, {tmtriet, ltnghia}@fit.hcmus.edu.vn

arxiv:2506.21056v1 中译本

摘要—在复杂室内环境中仅使用遮罩的 2D 图像和自然语言描述检索 3D 物体面临着重大挑战。ROOMELSA 挑战限制了对完整 3D 场景背景的访问，使得关于物体外观、几何形状和语义的推理变得更加复杂。这些挑战因扭曲的视角、无纹理的遮罩区域、模糊的语言提示和噪声分割掩码而加剧。为了解决这个问题，我们提出了 SAMURAI：一种基于形状感知的多模态检索方法用于 3D 物体识别。SAMURAI 结合了基于 CLIP 的语义匹配与从遮罩区域二值轮廓导出的形状引导重新排名，并采用了一种稳健的多数投票策略。专门的预处理管道通过提取最大的连通组件并去除背景噪声来提高掩码质量。我们的混合检索框架利用语言和形状线索，在 ROOMELSA 私有测试集上实现了具有竞争力的表现。这些结果突显了将形状先验与语言理解相结合对于稳健的开放世界 3D 物体检索的重要性。

***Index Terms*—多模态 3D 物体检索，形状先验，语言接地，场景理解。**

I. 介绍

使用自然语言和视觉上下文检索 3D 对象 [1], [2] 已成为计算机视觉、自然语言处理和 3D 场景理解交叉领域中的核心挑战，其在机器人学 [3]、增强现实 [4] 和室内导航 [5] 等方面有广泛的应用。多模态学习的最新突破，如 ULIP[6] 和 OpenScene[7]，展示了对齐 3D、图像和文本模式以实现开放世界的 3D 识别和检索的潜力。

传统的 3D 检索系统 [8], [9] 主要依赖于几何相似性或预定义的对象类别，这限制了它们在开放词汇设置中的灵活性。基于语言的 3D 理解方法 [10], [11] 通过自然语言描述来实现细粒度的对象识别，扩展了这些能力。然而，由于类内变异性、遮挡和空间模糊性，在复杂 3D 环境中将语言进行定位仍然具有挑战性。

ROOMELSA（多模态增强语言和空间辅助最优对象检索）大挑战 [1] 于 2025 年在 SHREC 举行，解决了三维场景理解中的一个关键问题：仅使用自然语言描述和空间环境来检索被遮挡的对象。参与者必须在没有直接视觉访问的情况下推断并定位目标对象，完全依赖于语言线索和空间关系。例如，给定提示“一张现代床，带有简单的床头板、两个枕头以及一条整齐覆盖的白色毯子”再加上一个被遮挡的图像，系统必须在一个三维环境中识别出相应的对象。图 1 显示了一个示例。

最近的研究强调了形状先验在解决杂乱场景中模糊对象引用问题中的重要性。诸如嵌入语言的三维高斯 [12] 和基于变压器的定位方法 [13], [14] 等技术结合了几何和关系线索以提高检索效果。在仅提供遮罩二维图像和文本提示的 ROOMELSA 设置中，形状感知检索发挥着关键作用。这一挑战促进了跨模态检索、空间语言定位以及全面三维场景理解方面的进展。

为了解决 ROOMELSA 挑战中的问题，我们引入了新颖的 SAMURAI 框架（感知形状的多模态检索用于三维物体识别）。我们的系统利用 CLIP 模型 [15] 将自然语言查询与从遮罩 2D 图像和候选对象图像中提取的视觉特征对齐。为减轻遮罩区域中的噪声，我们实现了一个预处理管道，该管道提取最大的连通组件并应用填充以确保完整捕获物体。通过结合形状先验，即通过背景移除技术生成的二值轮廓嵌入，来提高检索准确性。框架采用多种检索策略，包括仅文本、仅形状、先文本

*Both authors contributed equally to this research.
**Corresponding author.

[1]https://aichallenge.hcmus.edu.vn/shrec-2025/smart3droom

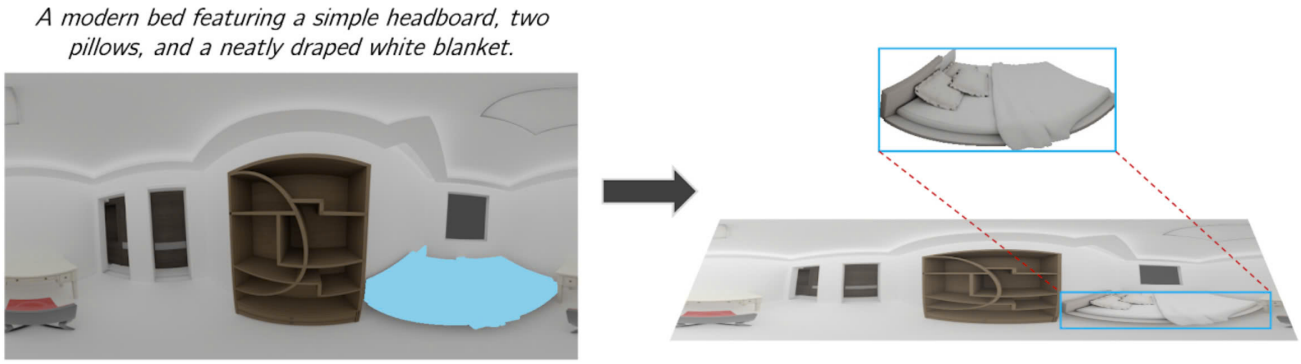

图 1: ROOMELSA 挑战的示例输入和输出。

后形状和多数投票，有效处理多样的查询类型和场景复杂性。通过将语义匹配与形状引导的重排序相结合，我们的系统在 ROOMELSA 私有测试集上实现了具有竞争力的表现，展示了多模态集成对于开放世界三维物体检索的有效性。

我们的 SAMURAI 在 ROOMELSA 挑战中排名前五，实现了 0.93 的 MRR，并且 Recall@5 和 Recall@10 均为完美成绩，与排名第一团队的 0.97 MRR 相当。其无需训练的设计确保了跨数据集和领域的鲁棒性，能够通过语言、形状和视觉线索进行有效的检索。这些结果突显了 SAMURAI 的竞争优势，提供了灵活性并简化了部署过程，无需重新训练。

我们的贡献有三点：

- 我们提出了一种新颖的形状感知检索框架，该框架集成了基于 CLIP 的自然语言和视觉特征嵌入，并通过从二值轮廓中提取的形状先验得到增强。
- 我们开发了一个强大的预处理管道，旨在处理嘈杂或碎片化的掩码，确保高质量的裁剪对象区域以进行特征提取。
- 我们提出了一种多策略检索方法，结合仅文本、仅形状、混合文本后形状以及多数投票方法来应对 ROOMELSA 数据集的多样化挑战。

II. 相关工作

A. 3D 场景理解

3D 场景理解对于机器人和增强现实应用至关重要。王等人。[16] 提出了用于 3D 全视角分割的统一占用方法，而德利察斯等人。[17] 引入了 SceneFun3D，带有 14.8k 交互注释以理解可操作性。王等人。[18] 提出了 EmbodiedScan，一个多模态套件，包含 RGB-D 和语言提示，支持复杂场景解释，如在 ROOMELSA 中所见。周等人。[19] 引入了 HUGS，利用 3D 高斯散斑技术实现整体城市场景理解，能够实时渲染新视角并重建动态场景。

同时，像 ScanRefer [10] 和 ReferIt3D [11] 这样的研究专注于语言引导的 3D 定位，最近的改进如 Chen 等人 [13] 通过空间推理增强了定位。Shi 等人 [12] 提出了嵌入语言的 3D 高斯分布用于开放词汇查询，推进了在杂乱环境中识别物体的技术。

相比之下，我们的方法引入了一个形状感知检索框架，该框架集成了语言和视觉特征，并辅以形状先验。这种方法在复杂环境如 ROOMELSA 中提高了鲁棒性，在这些环境中，遮挡和歧义经常构成重大挑战。

B. 多模态三维检索

近期在三维场景理解方面取得的进展显著推动了多模态检索系统的发展。ULIP [6] 和 MixCon3D [20] 集中于对齐三维对象、图像和文本以实现开放世界的检索。杨等人。[21] 引入了 RegionPLC 用于三维分割，而 OpenScene[7] 将 CLIP 特征投影到点云中，有助于跨模态的对齐，这是 ROOMELSA 挑战的核心。王等人。[22] 探索了多模态关系蒸馏（MRD），这是一种三模态预训练框架，旨在将大型视觉语言模型的知识转移到三维表示中。德尔马斯等人。[23] 通过他们的 PoseEmbroider 方法结合了三维姿态、图像和文本描述以创建更丰富的人体姿态表示，并且有效处理部分信息的能力超越了标准的多模态对齐模型。

我们提出的方法建立在这些进步的基础上，引入了一个形状感知检索框架，该框架集成了语言和视觉特征，并通过形状先验进行了增强。现有的一些工作如 ULIP 和 MixCon3D 专注于对齐三维、文本和图像模态以进行通用对象检索，而我们特别解决了复杂环境中遮挡物体检索的挑战。我们利用二值轮廓来获取形状先验，这对于解决诸如“桌子附近的椅子”这样的查询中的歧义至关重要，正如之前关于细粒度检索的研究 [12], [13] 所强调的那样。通过结合多种检索策略，我们提供了一个更健壮的解决方案，以满足这一挑战的独特需求。

III. 提议的方法

A. 概述

给定一个被遮罩的 3D 场景、一组候选对象和一个自然语言查询，我们的 SAMURAI 框架（图 2）通过联合推理语言描述和来自遮罩区域的部分视觉线索来识别目标对象。该系统由三个关键组件组成：一个查询预处理模块，用于精炼并正确识别与查询对象对应的遮罩；一个特征提取阶段，用于编码语言、形状信息和对象特征；一个多模态检索机制，用于整合文本和视觉特征并对候选对象进行排序。

B. 查询预处理

对于每个场景的掩码 RGB 图像，我们的预处理模块执行四个连续步骤。首先，它使用固定 RGB 颜色（135, 206, 235）识别出掩码对象区域并将其转换为二值掩码。其次，通过连通域标记提取最大的连通组件以消除来自多个掩码区域的噪声。第三，对边界框应用 10 像素填充，并裁剪该区域，确保即使对于离散掩码（如分段的部分床）也能捕获整个对象。第四，使用相同的 RGB 颜色在裁剪后的图像中重新识别出掩码区域，以细化二值掩码，解决离散化问题。此过程生成的二值掩码随后用于提取嵌入，并通过 CLIP 模型处理查询文本提示。

C. 目标特征提取

我们使用 CLIP 模型为每个对象生成两种互补类型的嵌入 [15] 。首先，我们产生一个 RGB 嵌入，编码整个对象图像，捕捉外观、纹理和颜色细节以与查询文本提示进行比较。其次，我们通过使用移除背景库移除背景、隔离前景并将其转换为二值图像来创建一个二值轮廓嵌入，专注于对象的形状。这种背景移除消除了无关的上下文元素，而二值化则突出了几何结构，有助于进行基于形状的稳健比较，这对于在复杂的 3D 场景中检索被遮挡的对象至关重要。

D. 检索策略

我们提出了五种互补的检索策略来处理 ROOMELSA 挑战所提出的多样化的查询类型和场景复杂性：

- 仅文本检索：通过计算对象的 RGB 嵌入与查询文本嵌入之间的余弦相似度对对象进行排名。该策略强调语义和外观线索，使其对于突出颜色、纹理或材料（例如，“一把红色木椅”）的提示有效。
- 仅形状检索：该策略将场景裁剪掩码的二进制嵌入与候选对象的二进制轮廓嵌入进行匹配。通过专注于几何结构，它在视觉线索有限的情况下（例如，“一个圆形桌子”）针对形状驱动查询表现良好。
- 文本然后形状（按形状重新排序）：初始阶段根据文本相似度选择前 15 个候选对象，然后使用形状相似性进一步精炼至前 10，并按形状相关性进行排序。此策略平衡了语言和几何信息，非常适合结合外观和形态的查询（例如，“一个高大的矩形柜子”）。
- 文本然后形状（按文本重新排序）：还是通过文本相似性筛选候选对象并通过形状进行细化，但最终列表会根据文本相关性重新排序。此变体更注重语义保真度，适用于语言精确度比几何契合更重要的查询。
- 多数投票：通过加权投票（例如，Text-then-Shape 变体为+2，其他为+1）整合来自所有四种策略的预测。这种集成方法通过跨策略整合互补优势来增强检索鲁棒性，确保在各种查询场景中保持稳定性能。

IV. 实验与结果

A. 数据集

ROOMELSA 挑战数据集包含两个主要目录：场景和对象。每个场景文件夹包含一个带有蒙版的 RGB 图像以及一个描述该蒙版对象的相关自然语言查询。总的来说，该数据集包括 50 个不同的场景。

对象文件夹包含候选的 3D 物体，每个物体由一张 RGB 图像、一个纹理文件以及一个格式为 .对象的 3D

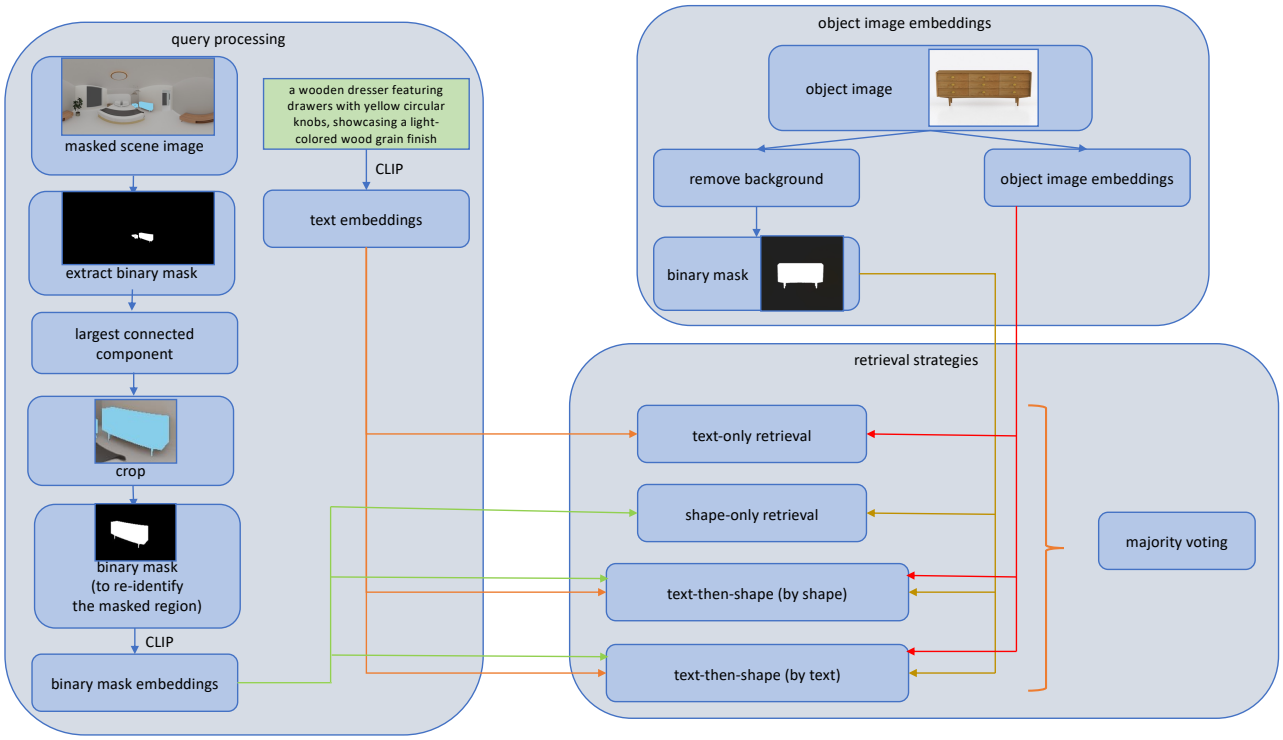

图 2: 提出的 SAMURAI 框架概述。

图 3: 我们五种检索策略的定性结果。

模型表示。对于每一个场景，参与者需要将最有可能的对象匹配进行排名，从最相关到最不相关的前 10 名。

B. 度量标准

检索性能评估在 ROOMELSA 挑战中基于以下标准指标：

表 I: 前五名团队在 ROOMELSA 挑战中使用私有测试集的性能比较。结果使用召回率@1（R@1）、召回率@5（R@5）、召回率@10（R@10）和平均倒数排名（MRR）进行评估。较高的值表示更好的性能。**粗体**表示最佳得分。

团队/方法	R@1 ↑	R@5 ↑	R@10 ↑	平均倒数排名 ↑
Stubborn_Strawberries	**0.94**	**1.00**	**1.00**	**0.97**
Ai-Yahh	0.92	**1.00**	**1.00**	0.96
MealsRetrieval	0.92	**1.00**	**1.00**	0.96
BUCCI_GANG	0.90	**1.00**	**1.00**	0.95
武士（我们提出的）	0.88	**1.00**	**1.00**	0.93

- 召回率@k (R@k)：衡量正确对象出现在前 k 个预测中的查询比例。在这个挑战中，报告了 $k = 1$、5 和 10。
- 平均倒数排名（MRR）：捕捉所有查询中正确对象的平均逆序位。定义为：

$$\mathrm{MRR} = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{\mathrm{rank}_i},$$

其中，Q 是场景（查询）总数，rank_i 表示第 i 个查询中的正确对象在排序列表中的位置。如果正确的对象不在前 10 个预测中，则互为秩设置为 0。

C. 定量结果

在参加 ROOMELSA 挑战的 18 支队伍中，我们的 SAMURAI 框架排名前五。如表 I 所示，我们的方法实现了高达 0.93 的平均倒数排名（MRR），并且在 Recall@5 和 Recall@10 方面都获得了满分。值得注意的是，尽管在方法复杂性上存在差异，这一表现与排名第一队伍的 0.97 MRR 高度可比。

SAMURAI 的一个关键优势在于其无需训练的设计，这消除了对特定任务微调的需求。这使得我们的方法在不同的数据集上具有鲁棒性，并且可以在不重新训练的情况下转移到新领域。它从语言、形状和视觉线索中良好泛化的能力使其即使在具有挑战性的未知场景中也能有效地检索信息。这些结果强调了一个轻量级、模块化管道可以与最先进的系统保持竞争力，同时提供更大的灵活性和更容易部署。

D. 定性结果

图 3 展示了我们的一些定性结果，证明了我们的检索策略的有效性。这些示例使用的查询如下：“一个引人注目的黑色讲台，顶部装有光滑的玻璃，并配以优雅的金色和白色装饰”以及“一张类似于熊猫的卡通风格床的插图，以其独特的黑白配色为特征。”结果显示，大多数我们的检索策略，包括仅文本、仅形状、先文本后形状（按文本和形状排序）和多数投票，在复杂的 3D 环境中表现出色且具有竞争力，成功地在各种条件下识别目标对象。

然而，在某些情况下，仅关注形状是不够的，文本描述变得至关重要。在复杂的背景下，仅凭形状可能无法提供足够的信息来准确检索物体。例如，图 4 描述了一个场景：查询“一款现代简约冰箱，配备两扇从顶部向下开启的门，并采用高端设计，在金属表面带有微妙的纹理图案”，在这种情况下需要依赖文本描述，因为仅凭形状可能无法完全捕捉到物体的独特特征。

V. 限制

尽管我们的方法在无需重新训练的情况下提供了具有竞争力的结果，但仍存在一些限制。在被遮罩的输入图像包含多个空间接近或形状和纹理相似的对象时，模型可能难以准确识别和检索目标对象。此外，高度碎片化或不完整的蒙版可能导致预处理过程中丢失部分对象。在某些极端情况下，物体的真实范围可能会超出边界框 10 像素以上，从而影响检索性能。这些挑战凸显了改进对象分离技术和更强大的蒙版细化技术以提高检索准确性的需求。

VI. 结论与未来工作

在本文中，我们提出了一种形状感知的多模态检索系统，该系统使用 CLIP 集成了语义（文本）和视觉（形状）特征。这种设计促进了灵活的混合策略和高效的实验。虽然我们的方法无需重新训练即可实现具有竞争力的表现，但多对象场景和碎片化的掩码仍然会影响检索准确性。

在未来，我们计划通过解决关键挑战来提高检索性能。例如，可以使用先进的对象补全或图像修复技术来减轻掩码碎片化的问题。此外，整合更深层次的空间上下文可以增强三维空间推理能力，从而更准确地定位目标物体，特别是在复杂场景中。

致谢

本研究得到越南国家大学胡志明市分校信息技术学院科研基金的支持。

图 4: 文本与形状结合进行多模态检索的必要性示例。

References

[1] T.-N. Le et al., "Sketchanimar: Sketch-based 3d animal fine-grained retrieval," *Computers & Graphics*, vol. 116, pp. 150–161, Nov. 2023, ISSN: 0097-8493. DOI: 10.1016/j.cag.2023.07.035.

[2] T.-N. Le et al., "Textanimar: Text-based 3d animal fine-grained retrieval," *Computers & Graphics*, vol. 116, pp. 162–172, Nov. 2023, ISSN: 0097-8493. DOI: 10.1016/j.cag.2023.07.026.

[3] T. Sakaguchi, A. Taniguchi, Y. Hagiwara, L. El Hafi, S. Hasegawa, and T. Taniguchi, "Object instance retrieval in assistive robotics: Leveraging fine-tuned simsiam with multi-view images based on 3d semantic map," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024, in press.

[4] M. Ji and J. Chun, "A sketch-based 3d object retrieval approach for augmented reality models using deep learning," *Journal of Internet Computing and Services*, vol. 21, no. 1, pp. 33–43, Feb. 2020. DOI: 10.7472/JKSII.2020.21.1.33.

[5] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2432–2443. DOI: 10.1109/CVPR.2017.261.

[6] L. Xue et al., "ULIP: Learning a Unified Representation of Language, Images, and Point Clouds for 3D Understanding," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 1179–1189. DOI: 10.1109/CVPR52729.2023.00120.

[7] S. Peng, K. Genova, C. Jiang, A. Tagliasacchi, M. Pollefeys, and T. Funkhouser, "OpenScene: 3D Scene Understanding with Open Vocabularies," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2023, pp. 815–824. DOI: 10.1109/CVPR52729.2023.00085.

[8] J. Hou, A. Dai, and M. Nießner, "Revealnet: Seeing behind objects in rgb-d scans," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2095–2104. DOI: 10.1109/CVPR42600.2020.00217.

[9] J. Chu et al., "Improved semantic representation learning by multiple clustering for image-based 3d model retrieval," *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 18, no. 1, pp. 1–20, 2022. DOI: 10.4018/IJSWIS.297033.

[10] D. Z. Chen, A. X. Chang, and M. Nießner, "Scanrefer: 3d object localization in rgb-d scans using natural language," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 202–221, ISBN: 978-3-030-58565-5.

[11] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. Guibas, "Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds., Cham: Springer International Publishing, 2020, pp. 422–440.

[12] J.-C. Shi, M. Wang, H.-B. Duan, and S.-H. Guan, "Language embedded 3d gaussians for open-vocabulary scene understanding," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 5333–5343.

[13] S. Chen, P.-L. Guhur, M. Tapaswi, C. Schmid, and I. Laptev, "Language conditioned spatial relation reasoning for 3d object grounding," in *Advances in Neural Information Processing Systems*, A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, Eds., 2022.

[14] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, I. Guyon et al., Eds., vol. 30, Curran Associates, Inc., 2017.

[15] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M. Meila and T. Zhang, Eds., ser. Proceedings of Machine Learning Research, vol. 139, PMLR, 18–24 Jul 2021, pp. 8748–8763.

[16] Y. Wang, Y. Chen, X. Liao, L. Fan, and Z. Zhang, "PanoOcc: Unified Occupancy Representation for Camera-based 3D Panoptic Segmentation," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA: IEEE Computer Society, Jun. 2024, pp. 17 158–17 168. DOI: 10.1109/CVPR52733.2024.01624.

[17] A. Delitzas, A. Takmaz, F. Tombari, R. Sumner, M. Pollefeys, and F. Engelmann, "Scenefun3d: Fine-grained functionality and affordance understanding in 3d scenes," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 14 531–14 542.

[18] T. Wang et al., "Embodiedscan: A holistic multi-modal 3d perception suite towards embodied ai," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 19 757–19 767.

[19] H. Zhou et al., "Hugs: Holistic urban 3d scene understanding via gaussian splatting," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 21 336–21 345.

[20] Y. Gao, Z. Wang, W.-S. Zheng, C. Xie, and Y. Zhou, "Sculpting holistic 3d representation in contrastive language-image-3d pre-training," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 22 998–23 008.

[21] J. Yang, R. Ding, W. Deng, Z. Wang, and X. Qi, "Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2024, pp. 19 823–19 832.

[22] H. Wang et al., "Multi-modal relation distillation for unified 3d representation learning," in *Computer Vision – ECCV 2024*, Cham, 2025, pp. 364–381, ISBN: 978-3-031-73414-4.

[23] G. Delmas, P. Weinzaepfel, F. Moreno-Noguer, and G. Rogez, "Poseembroider: Towards a 3d, visual, semantic-aware human pose representation," in *Computer Vision – ECCV 2024*, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., Cham: Springer Nature Switzerland, 2025, pp. 55–73, ISBN: 978-3-031-73209-6.