## 开放词汇视觉感知的下游应用——场景理解

#### Abstract

随着深度学习和计算机视觉技术的飞速发展，视觉感知系统在图像识别、物体检测和场景理解等任务中取得了显著进展。然而，传统的视觉感知模型通常依赖于预定义的闭合词汇集，这限制了它们在处理未见过的物体和场景时的能力。为了克服这一局限，开放词汇视觉感知（Open-Vocabulary Visual Perception）技术应运而生，它能够识别和理解大量甚至无限的新词汇，从而显著提升了视觉感知系统的通用性和灵活性。在开放词汇视觉感知的众多应用中，场景理解（Scene Understanding）是一个关键的研究领域。场景理解不仅要求系统能够识别图像中的个体物体，还需要理解物体之间的关系、场景的语义结构以及整体环境的语境，这对于自动驾驶、智能监控、增强现实等应用具有重要意义。本文将围绕开放词汇视觉感知在场景理解中的下游应用进行综述，主要探讨以下三篇文献中的研究工作，分别是《From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models》、《OpenMask3D: Open-Vocabulary 3D Instance Segmentation》以及《OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation》，这三篇文献介绍的技术都涉及开放词汇视觉感知在场景理解中的应用，覆盖到开放词汇视觉感知的多项技术，通过联系这三篇文献的研究成果，能够使我们对开放词汇视觉感知的场景理解领域更为熟悉。

#### Introduction

1 **开放词汇视觉感知技术**

开放词汇视觉感知（Open-Vocabulary Visual Perception）是指计算机视觉系统能够识别和处理超出预定义词汇范围的新物体和概念的能力。传统的视觉感知系统通常依赖于固定的词汇集来进行物体检测和识别，这在面对未见过的物体时表现出明显的局限性。开放词汇视觉感知的目标是通过融合视觉和语言模型，使系统具备处理无限扩展词汇的能力，从而提高其在实际应用中的适应性和通用性。存在多种技术实现开放词汇视觉感知，常见的关键技术如下：

* 视觉-语言模型（Vision-Language Models, VLMs）

视觉-语言模型通过联合训练图像和文本数据，能够在视觉和语言之间建立起语义联系，从而实现对新词汇的理解和识别。例如CLIP（Contrastive Language-Image Pretraining）模型，其基本思想是将图像和文本映射到同一向量空间，通过计算图像和文本嵌入之间的相似度来进行物体识别和分类，通过大规模的图像-文本对比学习，在开放词汇物体识别任务中表现出色。

* 零样本学习（Zero-Shot Learning）

零样本学习是开放词汇视觉感知的重要技术之一，旨在识别训练集中未见过的类别。零样本学习通过利用类别的语义描述（如文本描述或属性信息），在没有训练样本的情况下，通过将新类别的语义信息映射到视觉特征空间，实现对未见过物体的识别。

**2 场景理解**

场景理解（Scene Understanding）是计算机视觉中的一项关键任务，旨在从图像或视频中提取有意义的语义信息，包括物体识别、物体间关系推断以及整体场景的语义结构。场景理解不仅涉及个体物体的检测和分类，还包括对物体间复杂关系和上下文的理解，通常需要完成以下任务：

* 物体检测和识别（Object Detection and Recognition）：识别图像或视频中的个体物体，并确定其类别和位置。
* 场景分类（Scene Classification）：对整个图像进行分类，识别出图像所属的场景类别，如街道、厨房、森林等。
* 场景图生成（Scene Graph Generation, SGG）：生成描述场景中物体及其关系的图结构，其中节点代表物体，边表示物体间的关系。
* 语义分割（Semantic Segmentation）：将图像中的每个像素进行分类，以识别图像中的不同区域和物体。
* 实例分割（Instance Segmentation）：类似语义分割，但进一步区分图像中相同类别的不同实例。

**3 挑战和技术难点**

开放词汇视觉感知技术在场景理解中的应用具有巨大潜力，但也面临诸多挑战和技术难点：

* 数据标注和获取

开放词汇视觉感知模型的训练需要大规模、高质量的图像-文本数据，而数据标注过程耗时费力，且存在标注质量不佳和噪声等问题。如何高效地获取和标注大规模多模态数据，是一个亟待解决的难题。

* 模型的泛化能力

尽管现有的视觉-语言模型在开放词汇物体识别中表现很好，但面对完全未知或是复杂的场景和物体时，模型的泛化能力仍然有限。

* 计算资源和效率

训练大规模的视觉-语言模型的计算资源和空间消耗很大，且推理过程中的计算效率也有待优化。如何在保证模型性能的前提下，优化计算效率，是开放词汇视觉感知技术在实际应用中需要克服的另一个挑战。

**4 文章概况**

本文对三篇文献中有关开放词汇视觉感知在场景理解中的具体应用技术展开论述，三篇文章的概况如下：

* From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models

提出了一种从像素到图的开放词汇场景图生成方法，利用视觉-语言模型（VLMs）将图像转化为语义丰富的场景图，从而提高了场景理解的效果。通过这种方法，系统能够在未见过的物体和关系上进行泛化，生成更为准确和全面的场景图。

* OpenMask3D: Open-Vocabulary 3D Instance Segmentation

介绍了一种开放词汇3D实例分割方法，能够在3D场景中检测和分割未见过的物体实例，显著提升了3D场景理解的能力。通过结合多模态数据和先进的分割算法，系统能够在复杂的3D环境中实现高精度的物体识别和分割。

* OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation

提出了一种结合开放词汇检测和结构化表示的迭代视觉-语言导航方法，提升了机器人在复杂环境中的导航性能。通过开放词汇检测技术，机器人能够识别和理解未见过的环境特征，并利用结构化表示进行路径规划和决策。

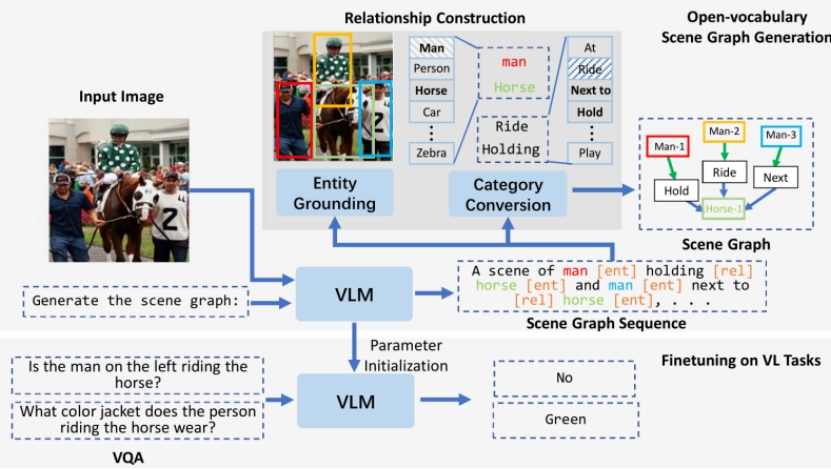
#### Techniques for Scene Understanding

**1 Open-vocabulary Scene Graph Generation**

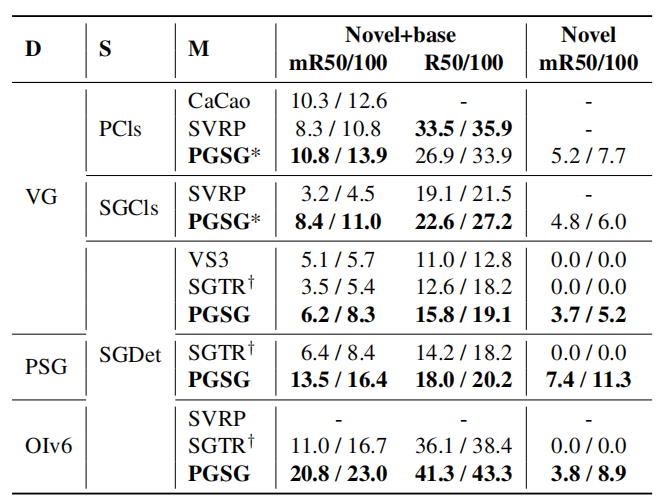
《From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models》介绍了一种新的开放词汇场景图生成（SGG）框架。场景图生成的主要目标是将图像解析为图形表示，以描述视觉场景中的对象实体及其关系。生成的场景图可以作为视觉感知和推理之间的结构性和可解释性表示，应用于各种视觉语言任务中。

大多数现有方法在处理封闭世界场景图生成问题上做出了很多努力，但通常只涵盖现实世界中多样化视觉关系的一个有限子集，导致场景表示的不完整以及在下游视觉语言任务中的域间隙问题。传统的SGG方法大多存在数据偏斜和标签不足等问题，也有一些方法尝试通过视觉-语言预训练模型处理开放词汇SGG问题，但这些方法通常关注简化的开放词汇设置，如仅允许新实体或给定实体对的谓词分类。

为了解决上述问题，作者提出了一种基于图像到序列生成的开放词汇SGG框架，称为PGSG（Pixels to Scene Graph Generation）。该框架通过将SGG任务表述为图像到序列生成问题，利用生成式VLM的强大能力生成场景图序列，并引入了基于生成VLM的微调策略。与依赖图像到文本匹配VLM的方法相比，这种生成式框架提供了一种更有效的利用预训练VLM的丰富视觉语言知识进行关系感知表示的方法，而无需额外的VLM预训练。



PGSG框架包括三个主要组件：场景图提示、预训练VLM生成场景图序列以及关系构建模块。首先，场景图提示通过场景图生成包含关系感知标记的序列表示。接下来，预训练的VLM为每个输入图像生成对应的场景图序列。最后，设计了一个即插即用的关系构建模块，使用关系三元组进行实体定位和类别转换，其中实体定位使用编码器-解码器预测实体的边界框，类别转换则是预测从词汇空间转换为类别空间，最后生成输出场景图。



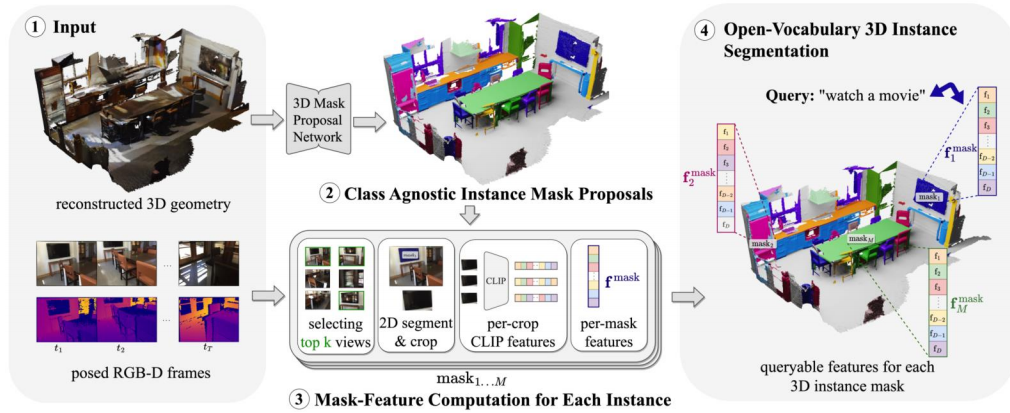
**The open-vocabulary scene graph generation on VG, PSG, and OIV6 datasets** *†* denotes the method is reproduced with BLIP visual encoder ViT as the backbone; D denotes the dataset name; S is the SGG setting; M represents the model; \* denotes that we use the grounding truth assignment on predictions.

作者在三个SGG基准数据集（Panoptic Scene Graph、OpenImages-V6和Visual Genome）上验证了其框架的有效性，取得了最先进的性能。此外，作者还将SGG基础的VLM应用于多个视觉语言任务比如视觉问答、图像标题生成、视觉定位等任务中，获得了一致的性能改进，突出显示了其有效的关系知识转移范式。

总的来看，本篇文献提出了基于生成VLM的新框架来解决一般开放词汇SGG问题，并且引入了场景图提示和即插即用的关系感知转换模块，允许更高效的模型学习，在多种下游视觉语言任务中展示了显著的性能提升，为未来开放词汇视觉感知研究提供了新的方向和思路。

**2 Open-Vocabulary 3D Instance Segmentation**

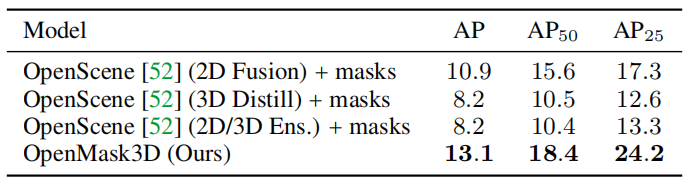
《OpenMask3D: Open-Vocabulary 3D Instance Segmentation》中提出一种基于开放词汇视觉感知的3D实例分割技术，3D实例分割旨在预测3D对象实例掩码及其对象类别，传统3D实例分割通常只能识别预定义的、在训练数据集中进行了注释的对象类别，存在局限性。而本篇文献提出了OpenMask3D思想，是一种零样本方法，能够进行开放词汇3D实例分割，从而处理描述未见过物体或新奇物体属性（如几何形状、用途和材料）的查询。具有超出预定义概念的实例分割能力。



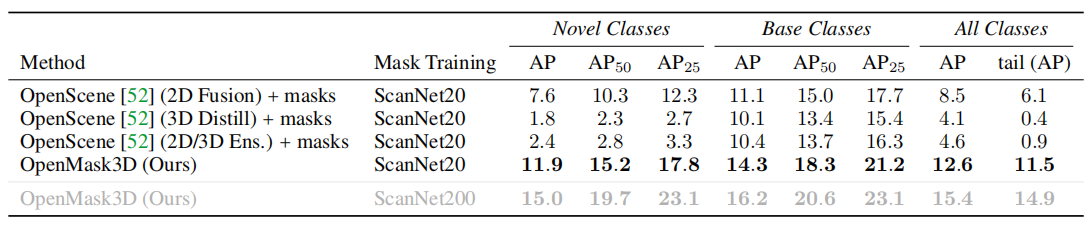
OpenMask3D包括三个计算阶段：

* 利用了一个预训练的3D实例分割模型的掩码模块，计算类别无关的实例掩码proposal
* 掩码特征计算：为每个预测实例掩码，通过CLIP计算一个与任务无关的特征表示，该过程包括选择前k个视角，获取多级裁剪图像，并提取CLIP特征以形成每个掩码的特征表示。
* 计算得到的特征用于概念查询，得到3D实例分割

这种方法不同于现有的基于点的方法，专注于基于实例的特征计算，增强了根据查询相似性检索物体实例掩码的能力。



**Table 2: 3D instance segmentation results on the Replica [61] dataset**. To assess how well our model generalizes to other datasets, we use instance masks from the mask proposal module trained on ScanNet200, and test it on Replica scenes. OpenMask3D outperforms other open-vocabulary counterparts on the Replica dataset.



**Table 3: 3D instance segmentation results using masks from mask module trained on ScanNet20 annotations, evaluated on the ScanNet200 dataset [57].** We identify 53 classes (such as chair, folded chair, table, dining table ...) that are semantically close to the original ScanNet20 classes [10], and group them as “Base”. Remaining 147 classes are grouped as “Novel”. We also report results on the full set of labels, titled “All”.

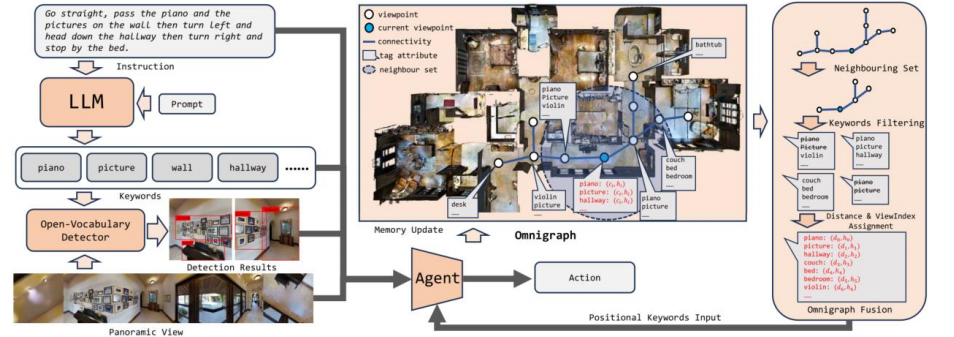
作者在ScanNet200和Replica数据集上的实验，验证了OpenMask3D的有效性。结果显示，OpenMask3D在处理长尾分布的物体类别方面优于现有方法。定性实验展示了其基于各种属性和自由形式查询进行物体分割的能力，突显了其在实际应用中的多功能性和稳健性。

总的来看，OpenMask3D与前面PGSG思想类似，都利用视觉-语言模型突破了封闭词汇系统的限制，说明了视觉-语言模型方法在推动开放词汇视觉感知方面的协同作用，并且OpenMask3D是首个能够进行零样本学习的3D实例分割方法，拓宽了可识别的物体类别范围。这种技术能够有效提升机器人在未知环境中与未见过的物体交互和导航的能力，与下面即将展开的《OVER-NAV》的目标一致，后者也同样通过开放词汇检测，提升视觉-语言导航，展示了它们在动态和非结构化环境中的适用性。

**3 Iterative Vision-and-Language Navigation**

《OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation》介绍了一种用于视觉-语言导航（VLN）的新型框架——OVER-NAV。视觉语言导航旨在构建智能代理程序，能够按照自然语言指令在未知环境中导航。

传统的VLN基准测试忽略了智能体的记忆能力，无法充分利用长时间累积的导航信息。迭代视觉-语言导航（IVLN）虽然引入了长时间记忆，但仍面临如何利用高度非结构化的导航记忆和极其稀疏的监督信号的问题。OVER-NAV框架的提出旨在克服这些挑战，通过集成大型语言模型（LLMs）和开放词汇检测（OVD）来提取和融合多模态信息，构建结构化记忆Omnigraph，从而提升导航智能体在未见过的环境中的导航能力。



OVER-NAV框架的核心在于通过OVD构建Omnigraph。具体过程如下：

1. 关键词提取：LLMs从每个导航指令中提取关键词。

2. 关键词检测：智能体在环境中移动时，将关键词和路径上的观测结果发送给开放词汇检测器OVD进行检测。

3. Omnigraph构建：检测结果存储在Omnigraph中，该结构化记忆帮助智能体记住并利用当前环境的观测部分。

4. 导航决策：实时计算场景记忆图信息，将Omnigraph生成的关键词以文本模态送入VLN代理程序，智能体导航动作预测，在连续和离散环境中更准确地行动。

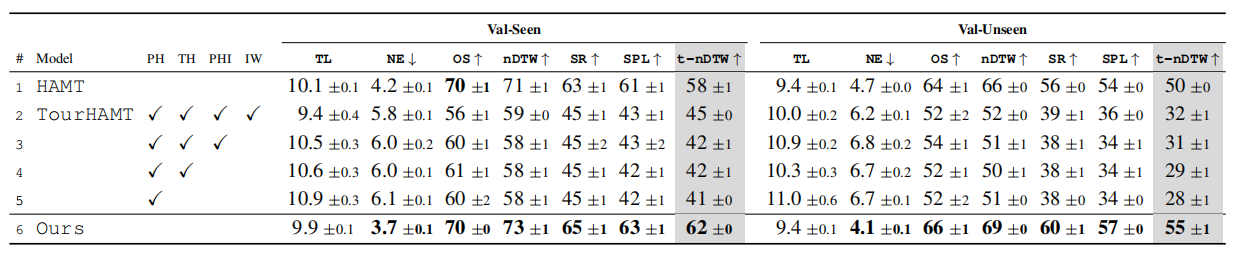


Table 1. The comparison between our method, HAMT and TourHAMT on IR2R.

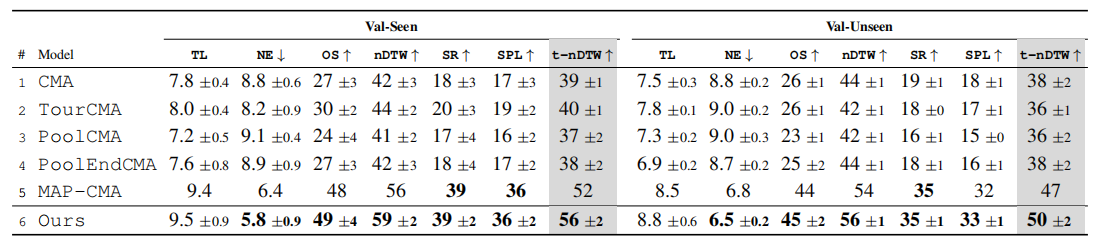


Table 2. The performance of our method on IR2R-CE.

实验结果表明，OVER-NAV在IR2R和IR2R-CE基准测试中表现出色，并且支持在离散环境下的测试性能，在REVERIE基准测试中验证了其有效性，在与HAMT和MAP-CMA两种基础模型比较中验证方法的有效性。

总的来看，OVER-NAV提出了一个新颖的框架，将LLMs和OVD引入IVLN范式；引入了编码Omnigraph的结构化表示，有效整合了多模态信息；在离散和连续环境中均实现了卓越的性能。

#### Conclusion

本文综述了开放词汇视觉感知技术在场景理解中的下游应用，通过分析三篇重要文献——《OpenMask3D: Open-Vocabulary 3D Instance Segmentation》、《From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models》和《OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation》，展示了这一领域的最新进展及其广泛应用。

开放词汇场景图生成中提出了一种从图像到文本生成的场景图生成框架。不仅提高了场景图生成的准确性，还增强了下游视觉语言任务的性能，展示了开放词汇系统在复杂场景理解中的强大能力。

开放词汇3D实例分割中OpenMask3D提出了一种基于CLIP模型的两阶段方法，通过类别无关的掩码提案和多视角特征聚合，实现了3D实例的开放词汇分割。这一方法克服了传统封闭词汇方法的局限性，显著提升了系统在处理新奇和多样化物体时的表现。

开放词汇导航中OVER-NAV结合大型语言模型和开放词汇检测技术，提出了一种新的迭代视觉-语言导航框架，通过构建结构化记忆Omnigraph，显著提升了智能体在未知环境中的导航能力。此方法强调了跨模态信息融合的重要性，并展示了在导航任务中的潜在应用。

综上所述，开放词汇视觉感知技术在场景理解中展现了巨大的潜力和广泛的应用前景。通过结合视觉-语言模型、开放词汇检测和结构化表示等先进技术，这些方法突破了传统封闭词汇系统的限制，提供了更灵活、更智能的解决方案。未来的研究应继续探索这些技术在不同领域中的应用，以进一步推动视觉感知与人工智能的发展。

#### Reference

[1] Ganlong Zhao, Guanbin Li, Weikai Chen, and Yizhou Yu. OVER-NAV: Elevating iterative vision-and-language navigation with open-vocabulary detection and structured representation. arXiv preprint arXiv:2403.17334, 2024.

[2] Aya Takmaz, Jonas Schult, Elisabetta Fedele, Robert W. Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. OpenMask3D: Open-vocabulary 3D instance segmentation. arXiv preprint arXiv:2306.13631, 2023.

[3] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. arXiv preprint arXiv:2404.00906, 2024.