Downstream Applications of Open-Vocabulary Visual Perception: Scene Understanding

Xinyu Chen
chenxy835@mail2.sysu.edu.cn

Abstract

With the rapid development of deep learning and computer vision technologies, visual perception systems have made significant progress in tasks such as image recognition, object detection, and scene understanding. However, traditional visual perception models often rely on predefined closed vocabulary sets, limiting their ability to handle unseen objects and scenes. To overcome this limitation, open-vocabulary visual perception has emerged, enabling the recognition and understanding of a large, potentially unlimited, set of new vocabularies. This significantly enhances the generality and flexibility of visual perception systems. Among the various applications of open-vocabulary visual perception, scene understanding is a key research area. Scene understanding requires not only recognizing individual objects in an image but also understanding the relationships between objects, the semantic structure of the scene, and the context of the overall environment, which is crucial for applications such as autonomous driving, intelligent surveillance, and augmented reality. This paper reviews the downstream applications of open-vocabulary visual perception in scene understanding, focusing on the research presented in three papers: From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models, OpenMask3D: Open-Vocabulary 3D Instance Segmentation, and OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation. These papers introduce technologies related to open-vocabulary visual perception in scene understanding, covering various techniques and providing insights into the field.

1. Introduction

1.1. OVVP

Open-Vocabulary Visual Perception (OVVP) refers to the ability of computer vision systems to recognize and process new objects and concepts beyond a predefined set of vocabulary. Traditional visual perception systems typically rely on fixed vocabularies for object detection and recognition, exhibiting significant limitations when faced with unseen objects. The goal of OVVP is to enhance the system's adaptability and generality in real-world applications by integrating visual and language models, enabling the system to handle an infinitely expanding set of vocabulary. Key techniques for achieving OVVP include:

Vision-Language Models (VLMs) VLMs establish semantic connections between visual and textual data through joint training, allowing for the understanding and recognition of new vocabulary. For example, the CLIP (Contrastive Language-Image Pretraining) model maps images and text to a common vector space and uses the similarity between image and text embeddings for object recognition and classification. By leveraging large-scale image-text contrastive learning, VLMs have demonstrated excellent performance in open-vocabulary object recognition tasks.

Zero-Shot Learning Zero-shot learning is a crucial technique in OVVP, aiming to recognize categories that have not been seen during training. Zero-shot learning utilizes the semantic descriptions of categories (e.g., textual descriptions or attribute information) to map the semantic information of new categories to the visual feature space, enabling the recognition of unseen objects.

1.2. Scene Understanding

Scene understanding is a key task in computer vision, aiming to extract meaningful semantic information from images or videos, including object recognition, inference of object relationships, and the semantic structure of the overall scene. Scene understanding involves detecting and classifying individual objects, understanding complex relationships and contexts between objects, and typically requires the completion of the following tasks:

Object Detection and Recognition: Identifying individual objects in images or videos and determining their categories and locations.

Scene Classification: Classifying entire images to identify the scene category they belong to, such as streets, kitchens, forests, etc.

Scene Graph Generation (SGG): Generating graph structures that describe the objects and their relationships within a scene, where nodes represent objects, and edges represent relationships between objects.

Semantic Segmentation: Classifying each pixel in an image to identify different regions and objects within the image.

Instance Segmentation: Similar to semantic segmentation but further distinguishes between different instances of the same object category within the image.

1.3. Challenges and Technical Difficulties

The application of OVVP technology in scene understanding holds great potential but also faces numerous challenges and technical difficulties:

Data Annotation and Acquisition: Training OVVP models requires large-scale, high-quality image-text data. The data annotation process is time-consuming and labor-intensive, often plagued by poor annotation quality and noise. Efficiently acquiring and annotating large-scale multimodal data is a critical challenge.

Model Generalization Ability: Despite the excellent performance of existing VLMs in open-vocabulary object recognition, their generalization ability remains limited when faced with completely unknown or complex scenes and objects.

Computational Resources and Efficiency: Training large-scale VLMs demands significant computational resources and memory, and the efficiency of the inference process also needs optimization. Balancing model performance and computational efficiency is another challenge for the practical application of OVVP technology.

1.4. Overview of the Articles

This paper discusses the specific application techniques of OVVP in scene understanding based on three significant articles. The overviews of the three articles are as follows:

From Pixels to Graphs:Open-Vocabulary Scene Graph Generation with Vision-Language Models [1]:

This paper proposes a method for generating open-vocabulary scene graphs from pixels using VLMs, transforming images into semantically rich scene graphs. This approach enhances scene understanding by enabling the system to generalize over unseen objects and relationships, producing more accurate and comprehensive scene graphs.

OpenMask3D: Open-Vocabulary 3D Instance Segmentation [2]:

This paper presents a method for open-vocabulary 3D instance segmentation, capable of detecting and segmenting unseen object instances in 3D scenes, significantly improving 3D scene understanding. By combining multimodal data and advanced segmentation algorithms, the system achieves high-precision object recognition and segmentation in complex 3D environments.

OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation [3]:

This paper proposes an iterative vision-and-language navigation method combining open-vocabulary detection and structured representation, enhancing the navigation performance of robots in complex environments. Through open-vocabulary detection technology, robots can recognize and understand unseen environmental features and use structured representation for path planning and decision-making.

2. Techniques for Scene Understanding

2.1. Open-Vocabulary Scene Graph Generation

The paper ”From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models” introduces a novel framework for open-vocabulary scene graph generation (SGG). The primary goal of SGG is to parse images into graph representations that describe object entities and their relationships within visual scenes. These generated scene graphs serve as structural and interpretable representations for various vision-language tasks, enabling a deeper understanding of the visual content.

Most existing methods have focused on closed-world SGG problems, which cover only a limited subset of the diverse visual relationships found in the real world. This limitation leads to incomplete scene representations and creates domain gaps in downstream vision-language tasks. Traditional SGG methods often suffer from data biases and insufficient labels, resulting in suboptimal performance. While some approaches attempt to address open-vocabulary SGG using vision-language pretraining models, they usually focus on simplified open-vocabulary settings, such as predicate classification for new entities or given entity pairs.

To address these issues, the authors propose an open-vocabulary SGG framework based on image-to-sequence generation, termed PGSG (Pixels to Scene Graph Generation). This framework formulates the SGG task as an image-to-sequence generation problem, leveraging the powerful capabilities of generative vision-language models (VLMs) to generate scene graph sequences. It introduces a fine-tuning strategy based on generative VLMs, offering a more efficient way to utilize the rich visual-language knowledge of pretrained VLMs for relationship-aware representations without additional VLM pretraining.

The PGSG framework comprises three main components:

Scene graph prompting: This component generates sequence representations with relationship-aware tokens, effectively guiding the VLM in capturing relevant visual rela-

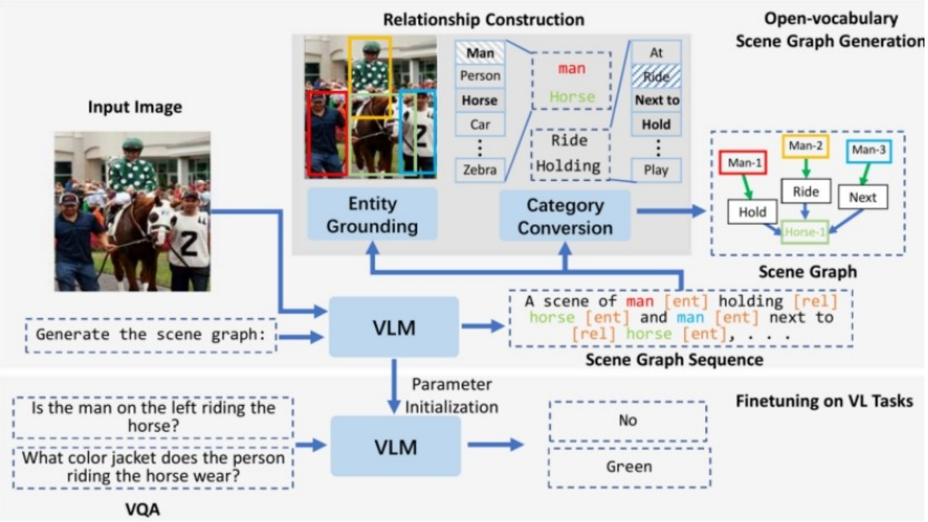


Figure 1. overall pipeline of our PGSG

tionships.

Pretrained VLM generating scene graph sequences: The pretrained VLM generates corresponding scene graph sequences for each input image, translating visual information into structured graph formats.

Relationship construction module: This plug-and-play module uses relationship triplets for entity localization and category conversion. Entity localization predicts entity bounding boxes using an encoder-decoder architecture, while category conversion predicts the transformation from vocabulary space to category space, ultimately generating the output scene graph.

D	S	M	Novel+base		Novel
			mR50/100	R50/100	mR50/100
VG	PCls	CaCao	10.3 / 12.6	-	-
		SVRP	8.3 / 10.8	**33.5 / 35.9**	-
		PGSG*	**10.8 / 13.9**	26.9 / 33.9	5.2 / 7.7
	SGCls	SVRP	3.2 / 4.5	19.1 / 21.5	-
		PGSG*	**8.4 / 11.0**	**22.6 / 27.2**	4.8 / 6.0
	SGDet	VS3	5.1 / 5.7	11.0 / 12.8	0.0 / 0.0
		SGTR†	3.5 / 5.4	12.6 / 18.2	0.0 / 0.0
		PGSG	**6.2 / 8.3**	**15.8 / 19.1**	**3.7 / 5.2**
PSG		SGTR†	6.4 / 8.4	14.2 / 18.2	0.0 / 0.0
		PGSG	**13.5 / 16.4**	**18.0 / 20.2**	**7.4 / 11.3**
OIv6		SVRP	-	-	-
		SGTR†	11.0 / 16.7	36.1 / 38.4	0.0 / 0.0
		PGSG	**20.8 / 23.0**	**41.3 / 43.3**	**3.8 / 8.9**

Figure 2. The open-vocabulary scene graph generation on VG, PSG, and OIV6 datasets

The authors validate their framework on three SGG benchmark datasets 2 (Panoptic Scene Graph, OpenImages-V6, and Visual Genome), achieving state-of-the-art performance. Additionally, they apply the SGG-based VLM to various vision-language tasks such as visual question answering, image captioning, and visual grounding, demonstrating consistent performance improvements. This highlights the effectiveness of their relationship knowledge transfer paradigm in enhancing the interpretability and functionality of vision-language systems.

In summary, this paper proposes a new framework based on generative VLMs to address the general open-vocabulary SGG problem. It introduces scene graph prompting and a plug-and-play relationship-aware transformation module, enabling more efficient model learning and application. The framework demonstrates significant performance improvements in various downstream vision-language tasks, providing new directions and ideas for future research in open-vocabulary visual perception.

2.2. Open-Vocabulary 3D Instance Segmentation

The paper ”OpenMask3D: Open-Vocabulary 3D Instance Segmentation” introduces an innovative technique for open-vocabulary 3D instance segmentation grounded in open-vocabulary visual perception. Traditional 3D instance segmentation methods are constrained to recognizing only predefined object categories that are annotated in the training datasets. This limitation hinders their effectiveness in real-world applications where new and unseen objects frequently appear.

OpenMask3D addresses this limitation through a zero-shot approach, enabling the segmentation of 3D instances based on open-vocabulary descriptions. This method can handle queries that describe previously unseen or novel ob-

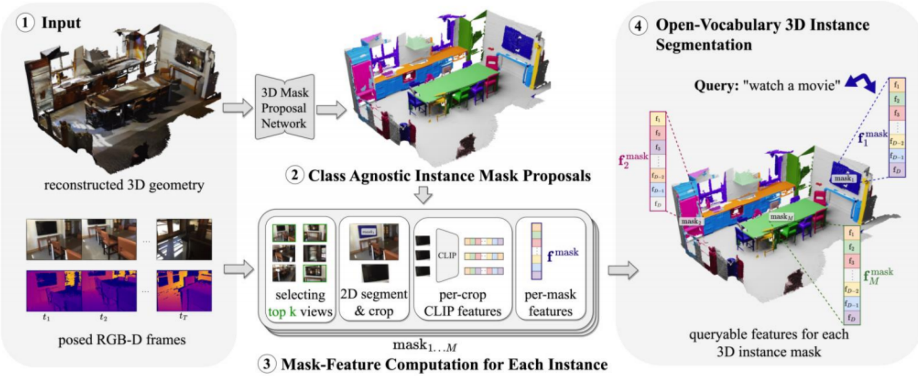

1 Input
reconstructed 3D geometry
posed RGB-D frames
t_1
t_2
t_T
3D Mask Proposal Network
2 Class Agnostic Instance Mask Proposals
selecting top k views
2D segment & crop
CLIP
per-crop CLIP features
$\mathbf{f}^{\text{mask}}$
per-mask features
$\text{mask}_{1\dots M}$
3 Mask-Feature Computation for Each Instance
4 Open-Vocabulary 3D Instance Segmentation
Query: "watch a movie"
$\mathbf{f}_1^{\text{mask}}$
$\mathbf{f}_2^{\text{mask}}$
$\mathbf{f}_M^{\text{mask}}$
mask_1
mask_2
mask_M
f_1
f_2
f_3
f_{D-2}
f_{D-1}
f_D
queryable features for each 3D instance mask

Figure 3. overview of OpenMask3D approach

ject attributes, such as geometric shapes, uses, and materials. By doing so, OpenMask3D extends the capabilities of instance segmentation beyond the predefined concepts typically encountered in training data.

OpenMask3D consists of three computational stages:

Instance Mask Proposal: Utilizing a pretrained 3D instance segmentation model's mask module, OpenMask3D computes category-agnostic instance mask proposals. This step generates initial masks that are not tied to specific object categories, providing a flexible foundation for further processing.

Mask Feature Computation: For each predicted instance mask, OpenMask3D calculates a task-agnostic feature representation using the CLIP model. This process involves selecting the top-k views of the object, obtaining multi-level cropped images, and extracting CLIP features to create a comprehensive feature representation for each mask. This approach ensures that the feature representation captures various aspects of the object from different perspectives.

Concept Query Computation: The final stage involves computing the concept query to obtain the 3D instance segmentation from the previously computed features. This stage leverages the rich, task-agnostic features to match and retrieve object instances based on the similarity to the provided query.

Distinct from existing point-based methods, OpenMask3D emphasizes instance-based feature computation. This focus enhances the system's ability to retrieve object instance masks effectively based on the similarity of features to the query, allowing for more accurate and flexible instance segmentation.

The authors validate the effectiveness of OpenMask3D on the Replica 4 and ScanNet200 5 datasets. The results demonstrate that OpenMask3D outperforms existing methods, particularly in handling long-tail distributions of object categories. Qualitative experiments further showcase its capability to segment objects based on diverse attributes and free-form queries, underscoring its versatility and robustness in real-world scenarios.

Model	AP	AP_{50}	AP_{25}
OpenScene [52] (2D Fusion) + masks	10.9	15.6	17.3
OpenScene [52] (3D Distill) + masks	8.2	10.5	12.6
OpenScene [52] (2D/3D Ens.) + masks	8.2	10.4	13.3
OpenMask3D (Ours)	**13.1**	**18.4**	**24.2**

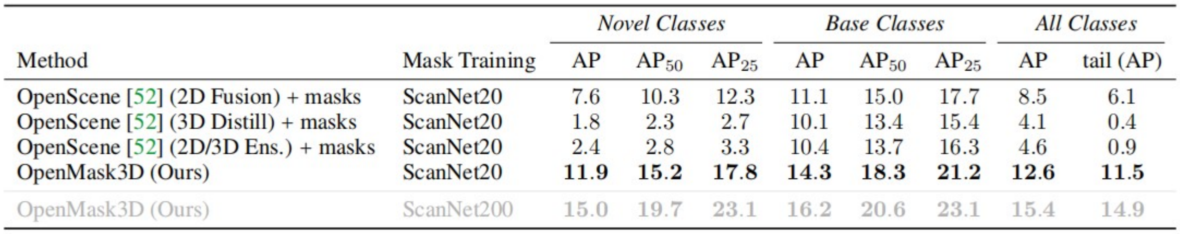
Figure 4. 3D instance segmentation results on the Replica dataset

In summary, similar to the PGSG approach, OpenMask3D leverages vision-language models to transcend the limitations of closed-vocabulary systems. This technique exemplifies the synergistic role of vision-language model methods in advancing open-vocabulary visual perception. Notably, OpenMask3D is the first zero-shot 3D instance segmentation method, significantly broadening the spectrum of recognizable object categories. This advancement enhances the ability of robots to interact with and navigate through unknown environments. It aligns with the objectives of "OVER-NAV," which also improves vision-language navigation through open-vocabulary detection. Together, these methods demonstrate substantial applicability in dynamic and unstructured environments, paving the way for more intelligent and adaptable robotic systems.

2.3. Vision-and-Language Navigation

The paper "OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation" introduces a groundbreaking framework for vision-and-language navigation (VLN), termed OVER-NAV. VLN focuses on developing intelligent agents capable of navigating unfamiliar environments by following natural language instructions.

Traditional VLN benchmarks often overlook the agent's memory capabilities, thereby failing to fully leverage cumulative navigation information. While iterative vision-



Method	Mask Training	*Novel Classes* AP	*Novel Classes* AP_{50}	*Novel Classes* AP_{25}	*Base Classes* AP	*Base Classes* AP_{50}	*Base Classes* AP_{25}	*All Classes* AP	*All Classes* tail (AP)
OpenScene [52] (2D Fusion) + masks	ScanNet20	7.6	10.3	12.3	11.1	15.0	17.7	8.5	6.1
OpenScene [52] (3D Distill) + masks	ScanNet20	1.8	2.3	2.7	10.1	13.4	15.4	4.1	0.4
OpenScene [52] (2D/3D Ens.) + masks	ScanNet20	2.4	2.8	3.3	10.4	13.7	16.3	4.6	0.9
OpenMask3D (Ours)	ScanNet20	**11.9**	**15.2**	**17.8**	**14.3**	**18.3**	**21.2**	**12.6**	**11.5**
OpenMask3D (Ours)	ScanNet200	15.0	19.7	23.1	16.2	20.6	23.1	15.4	14.9

Figure 5. 3D instance segmentation results on the ScanNet200 dataset

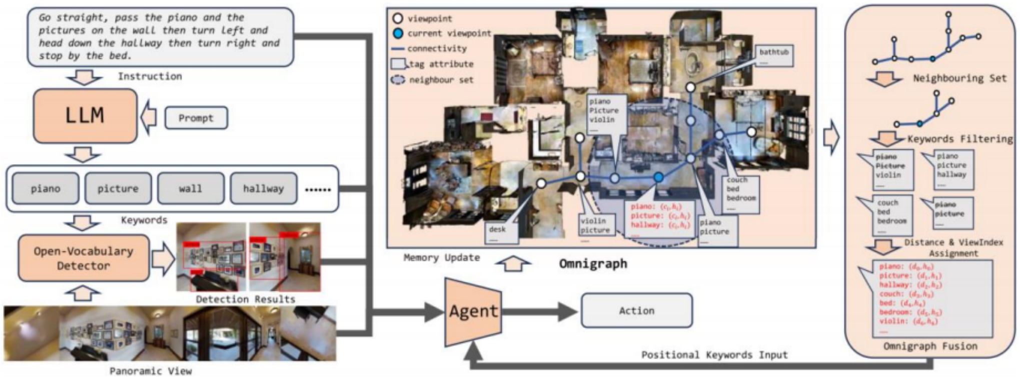


Figure 6. overview of OVER-NAV method

and-language navigation (IVLN) introduces the concept of long-term memory, it still grapples with the challenges of utilizing highly unstructured navigation memory and sparse supervision signals. The OVER-NAV framework effectively addresses these challenges by integrating large language models (LLMs) and open-vocabulary detection (OVD). This integration allows for the extraction and fusion of multimodal information, leading to the construction of a structured memory known as Omnigraph. This significantly enhances the navigation agent's capability to operate in unseen environments.

The core of the OVER-NAV framework is constructing Omnigraph through OVD. The process includes:

Keyword extraction: LLMs are employed to extract keywords from each navigation instruction. This step ensures that the essential components of the instructions are identified and isolated for further processing.

Keyword detection: As the agent traverses the environment, it continually sends the extracted keywords along with observations made along its path to the OVD. The OVD then detects relevant elements within the environment corresponding to the keywords.

Omnigraph construction: The results from the keyword detection process are stored within the Omnigraph. This structured memory aids the agent in recalling and utilizing parts of the environment's observations, thus facilitating more informed decision-making.

Navigation decision-making: The real-time computation of scene memory graph information involves feeding the Omnigraph-generated keywords as a text modality into the VLN agent program. This allows the agent to predict navigation actions with greater accuracy in both continuous and discrete environments.

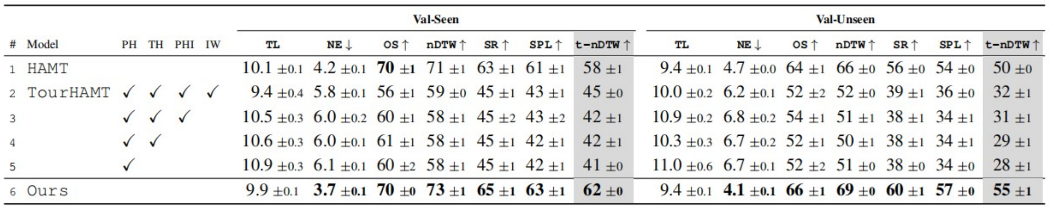
Experimental results demonstrate the superior performance of OVER-NAV on challenging benchmarks such as IR2R 7 and IR2R-CE 8. The framework also proves effective in discrete environments, as evidenced by its performance on the REVERIE benchmark. Moreover, the method's effectiveness is validated when applied to HAMT and MAP-CMA base models, showcasing its versatility and robustness.

In summary, OVER-NAV introduces an innovative framework that incorporates LLMs and OVD into the IVLN paradigm. By introducing a structured representation encoded within Omnigraph, OVER-NAV effectively integrates multimodal information. This leads to significant performance improvements in both discrete and continuous environments, underscoring its potential for advancing the field of vision-and-language navigation.

3. Applications of Scene Understanding

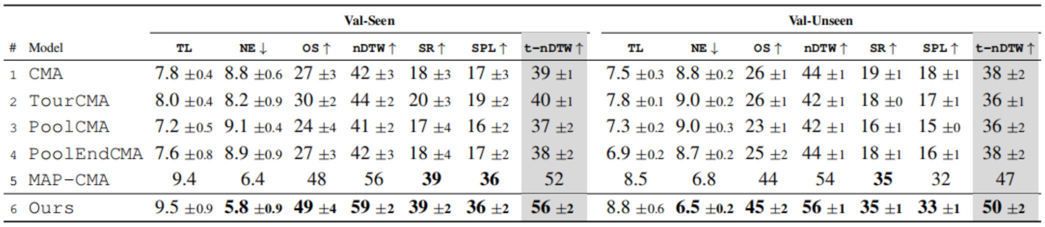
Scene understanding is important in several practical applications, and the following are a few key application areas:

Autonomous Driving: In the field of autonomous driving, vehicles require real-time and accurate perception and



#	Model	PH	TH	PHI	IW	Val-Seen TL	NE ↓	OS ↑	nDTW ↑	SR ↑	SPL ↑	t-nDTW ↑	Val-Unseen TL	NE ↓	OS ↑	nDTW ↑	SR ↑	SPL ↑	t-nDTW ↑
1	HAMT					10.1 ±0.1	4.2 ±0.1	**70** ±1	71 ±1	63 ±1	61 ±1	58 ±1	9.4 ±0.1	4.7 ±0.0	64 ±1	66 ±0	56 ±0	54 ±0	50 ±0
2	TourHAMT	✓	✓	✓	✓	9.4 ±0.4	5.8 ±0.1	56 ±1	59 ±0	45 ±1	43 ±1	45 ±0	10.0 ±0.2	6.2 ±0.1	52 ±2	52 ±0	39 ±1	36 ±0	32 ±1
3		✓	✓	✓		10.5 ±0.3	6.0 ±0.2	60 ±1	58 ±1	45 ±2	43 ±2	42 ±1	10.9 ±0.2	6.8 ±0.2	54 ±1	51 ±1	38 ±1	34 ±1	31 ±1
4		✓	✓			10.6 ±0.3	6.0 ±0.1	61 ±1	58 ±1	45 ±1	42 ±1	42 ±1	10.3 ±0.3	6.7 ±0.2	52 ±1	50 ±1	38 ±1	34 ±1	29 ±1
5		✓				10.9 ±0.3	6.1 ±0.1	60 ±2	58 ±1	45 ±1	42 ±1	41 ±0	11.0 ±0.6	6.7 ±0.1	52 ±2	51 ±0	38 ±0	34 ±0	28 ±1
6	Ours					9.9 ±0.1	**3.7** ±0.1	**70** ±0	**73** ±1	**65** ±1	**63** ±1	**62** ±0	9.4 ±0.1	**4.1** ±0.1	**66** ±1	**69** ±0	**60** ±1	**57** ±0	**55** ±1

Figure 7. comparison between OVER-NAV, HAMT and TourHAMT on IR2R



#	Model	Val-Seen TL	NE ↓	OS ↑	nDTW ↑	SR ↑	SPL ↑	t-nDTW ↑	Val-Unseen TL	NE ↓	OS ↑	nDTW ↑	SR ↑	SPL ↑	t-nDTW ↑
1	CMA	7.8 ±0.4	8.8 ±0.6	27 ±3	42 ±3	18 ±3	17 ±3	39 ±1	7.5 ±0.3	8.8 ±0.2	26 ±1	44 ±1	19 ±1	18 ±1	38 ±2
2	TourCMA	8.0 ±0.4	8.2 ±0.9	30 ±2	44 ±2	20 ±3	19 ±2	40 ±1	7.8 ±0.1	9.0 ±0.2	26 ±1	42 ±1	18 ±0	17 ±1	36 ±1
3	PoolCMA	7.2 ±0.5	9.1 ±0.4	24 ±4	41 ±2	17 ±4	16 ±2	37 ±2	7.3 ±0.2	9.0 ±0.3	23 ±1	42 ±1	16 ±1	15 ±0	36 ±2
4	PoolEndCMA	7.6 ±0.8	8.9 ±0.9	27 ±3	42 ±3	18 ±4	17 ±2	38 ±2	6.9 ±0.2	8.7 ±0.2	25 ±2	44 ±1	18 ±1	16 ±1	38 ±2
5	MAP-CMA	9.4	6.4	48	56	**39**	**36**	52	8.5	6.8	44	54	**35**	32	47
6	Ours	9.5 ±0.9	**5.8** ±0.9	**49** ±4	**59** ±2	**39** ±2	**36** ±2	**56** ±2	8.8 ±0.6	**6.5** ±0.2	**45** ±2	**56** ±1	**35** ±1	**33** ±1	**50** ±2

Figure 8. The performance of OVER-NAV on IR2R-CE

understanding of the surrounding environment to make safe driving decisions. Scene understanding technologies assist autonomous driving systems in identifying objects such as vehicles, pedestrians, and traffic signs on the road, and understanding their relationships, thereby enhancing the safety and reliability of autonomous driving.

Intelligent Surveillance: Intelligent surveillance systems monitor specific areas using cameras, requiring real-time detection and recognition of anomalies such as intrusion and fighting behaviors. Scene understanding technologies improve the automation of surveillance systems, reduce manual intervention, and enhance monitoring efficiency and accuracy.

Augmented Reality (AR): In augmented reality (AR) applications, scene understanding technologies are used to identify objects and environments in the real world and overlay virtual information on them. For example, in educational and entertainment applications, AR systems can recognize images in textbooks and display related 3D models or animations, enhancing user experience.

Robot Navigation: Robots performing tasks in complex environments rely on scene understanding technologies to perceive and comprehend the surrounding environment for path planning and obstacle avoidance. For instance, service robots in home environments need to recognize and locate objects such as furniture and appliances to complete tasks like delivery and cleaning.

4. Conclusion

This review explores the downstream applications of open-vocabulary visual perception technologies in scene understanding, analyzing three pivotal papers: ”OpenMask3D: Open-Vocabulary 3D Instance Segmentation,” ”From Pixels to Graphs: Open-Vocabulary Scene Graph Generation with Vision-Language Models,” and ”OVER-NAV: Elevating Iterative Vision-and-Language Navigation with Open-Vocabulary Detection and Structured Representation.” These papers demonstrate the latest advancements in the field and their wide-ranging applications.

Open-Vocabulary Scene Graph Generation

The introduction of a scene graph generation framework from image-to-text in open-vocabulary settings has significantly enhanced both the accuracy of scene graph generation and the performance in downstream vision-language tasks, showcasing the robust capabilities of open-vocabulary systems in complex scene understanding.

Open-Vocabulary 3D Instance Segmentation

OpenMask3D proposes a two-stage method based on the CLIP model for open-vocabulary 3D instance segmentation. It achieves this through category-agnostic mask proposals and multi-view feature aggregation, overcoming the limitations of traditional closed-vocabulary methods and significantly improving system performance in handling novel and diverse objects.

Open-Vocabulary Navigation

OVER-NAV integrates large language models and open-vocabulary detection techniques to propose a new iterative vision-and-language navigation framework. By constructing the structured memory Omnigraph, this approach markedly enhances the navigation capabilities of agents in unknown environments. This method underscores the importance of multimodal information fusion and demonstrates its potential applications in navigation tasks.

In summary, open-vocabulary visual perception technologies exhibit tremendous potential and extensive applications in scene understanding. By leveraging advanced techniques such as vision-language models, open-vocabulary detection, and structured representations, these methods surpass the limitations of traditional closed-vocabulary systems, providing more flexible and intelligent solutions. Future research should continue exploring these technologies across various domains to further advance the development of visual perception and artificial intelligence.

References

[1] 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. 2

[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. 2

[3] 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