Spatial Reasoning with Open Set Vocabulary Object Detectors for Robot Perception

Negar Nejatishahidin[1] and Jana Kosecka[1]

***Abstract*— We present a probabilistic approach for Spatial Relation Detection for 3D robotic perception. We exploit the state-of-the-art open set vocabulary object detectors [28] and rich 3D geometric features to localize objects and their spatial relations. We carry out experiments and ablation studies on both real Spatial Sense [25] and synthetic Semantic abstraction dataset [8] and demonstrate challenges on the open-set vocabulary setting and effectiveness of our approach on both synthetic and real data.**

I. INTRODUCTION

The task of spatial relationship detection refers to the ability to localize objects of interest and determine the spatial relationships between them. Spatial reasoning is a foundational cognitive ability essential for human perception and interaction with the environment. Understanding relations play a crucial role in organizing spatial layouts and interpreting the physical world. Consequently, for machines to effectively interact with their surroundings, they must be equipped with the ability to reason about the spatial relationships within a scene. This capability has a wide range of applications across diverse domains, including robotics [3, 17], scene understanding [10, 19] as well as in human-robot interaction and task planning using natural language [17].

With the rapid advancements in Computer Vision [6, 28, 11, 1], Natural Language Processing [5, 20], and Vision and Language Models (VLMs) [16, 13, 24, 12], a large body of previously challenging vision and language tasks marked notable improvements. For these new approaches, reasoning about spatial relationships in a zero-shot manner or by fine-tuning the large vision and language models continues to be a challenging task [22, 15, 7]. We consider a setting where given a triplet consisting of an expression defining target object e_t, a spatial relation e_{rel}, and a reference object e_r in the corresponding image, we aim to accurately locate and identify the target object specified by the expression $< e_t, e_{rel}, e_r >$. This task is closely related to recent formulations demonstrated in previous works [8, 22, 25, 15, 7].

We propose a probabilistic modular approach utilizing the state-of-the-art open-set vocabulary object detector [28, 13] and relying on 3D geometric cues to reason about spatial relations. We can summarize our contributions as follows:

- We study the effectiveness of an open set object detector [28] and propose different variants for grounding the target and referred objects in spatial relation expression.
- We propose a novel geometric Spatial Relation Module, that takes as input the pose and dimensions of the object computed from the 3D point cloud of the objects masked by Detic [28]. We also conduct a thorough ablation study on the effectiveness of different 2D, and 3D geometric features along with language features.
- We introduce a probabilistic ranking module that combines the evidence from object grounding and spatial relationship classification to get the best triplet.

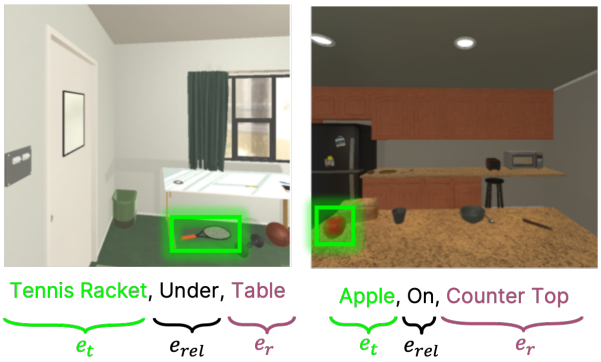

Fig. 1: Examples showing inputs and desired outputs. The input consists of a triplet consisting of a target object e_t, spatial relation e_{rel}, and reference object e_r and corresponding image. The output is the target object bounding box (boxes colored in green).

We demonstrate the performance on both real Spatial Sense [25] dataset and synthetic Semantic Abstraction dataset [11]. Later in Section III, we will describe our approach, and in Section IV, we will go through our results.

II. RELATED WORK

With the advent of large vision-language models, spatial reasoning has attracted additional interest in recent years motivated by the poor zero-shot performance of spatial relationship recognition. Representative approaches typically adopt modular approache [4, 27], or proceed with fine-tuning of large pre-trained Vision-Language Models [15, 9] of adopt some combinations of the two [22, 17, 23]. Although many works have been proposed in this area, several major challenges arise when attempting to

[1] George Mason University, 4400 University Dr, Fairfax, VA nnejatis, kosecka @gmu.edu

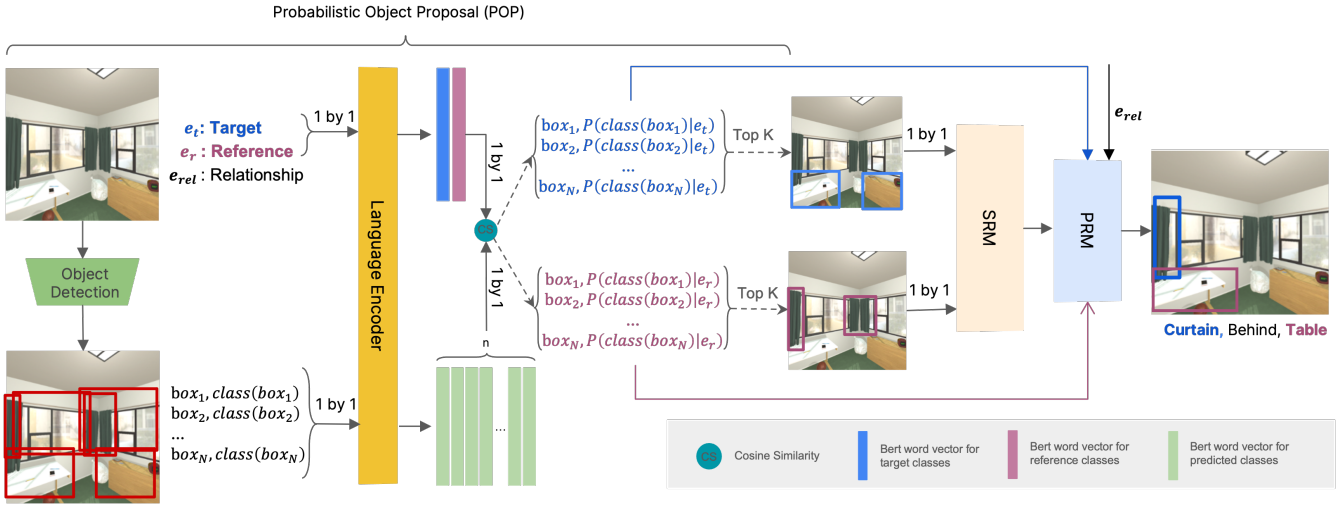

Fig. 2: This is an overview of our pipeline. It consists of three main modules: first, the Probabilistic Object Proposal (POP), which provides a set of boxes as candidates for the target and reference objects; second, the Spatial Relation Module (SRM), which gives a distribution over possible relationships for each pair; and third, the Probabilistic Ranking Module (PRM), which provides the best triplet.

adapt these models to real-world robotic tasks and 3D scenes.

Earlier modular approaches use training splits of the specially curated datasets [4, 27] and often exhibit limitations when encountering changes in the distribution of visual or linguistic data. In [14] authors introduce the Super CLEVR dataset to study various domain shifts between training and test distribution for VQA task and demonstrate improved robustness of the modular approach for several distribution shifts.

Another category tackles the problem of spatial relationship recognition either in zero-shot setting or by fine-tuning large vision and language models [12, 16, 24] that have been pre-trained in a self-supervised manner. The root of the problems in these cases can be identified to some extent using explainability methods [22] that often indicates a lack of grounding of nouns or noun phrases. Furthermore, fine-tuning these models for downstream tasks often results in using overly large and complex models for specific tasks, which may not align well with robotics purposes. As emphasized in [18], robotics research is motivated to employ modules applicable to multiple tasks concurrently and beneficial for various applications. Large pre-trained models [21, 13, 12] have been also exploited in [8, 17, 23] in a zero-shot setting. In [8] authors propose an obscured object localization module using CLIP [21] for computing initial relevancy maps in RGB-D data and introducing Semantic Abstraction dataset. The performance of this approach on spatial relationship recognition was quite low (< 30 %).

Motivated by the robotic application of table top manipulation authors in [17] propose a probabilistic modular technique for visual question answering (VQA) using CLIP. The visual setting of this approach is quite simple, involving tabletop objects, considering only 2D information. The effectiveness of 3D features was demonstrated in [7] presented a 3D geometric-based spatial reasoning approach considering different 3D features of pairs of objects on uniform backgrounds.

We propose a fully modular and explainable probabilistic spatial reasoning pipeline, which leverages state-of-the-art object detectors and language modules [5, 28] and introduce a novel spatial relation prediction module using 3D features. The approach is evaluated on synthetic Semantic Abstraction dataset [8] improving the performance of the CLIP based baseline. We further extended our experiments to real-world data and reconfigured the Spatial Sense dataset [25] to evaluate our approach, thereby demonstrating its effectiveness in tackling real-world challenges.

III. APPROACH

The proposed approach comprises three core components: the Probabilistic Object Proposal (POP), discussed in Section III-B, which utilizes an open-vocabulary object detector to for target and reference objects; the Spacial Relationship Module (SRM) (Section III-C) that predicts spatial relationships between pairs of objects in the scene; and the Probabilistic Ranking Module (PRM) (Section III-D) for final target object ranking. In the following sections, we will elaborate on the problem definition. The approach is visualized in Figure 2.

A. Problem Definition.

An example in our training set consists of an expression e and its corresponding image I. The expression e in natural language consists of three components: a reference object e_r and target object e_t and their spatial relation e_{rel}. The model output is the precise location of

the referred object described in the spatial expression in the image. An example from Semantic Abstraction dataset along with its inputs and desired outputs, is shown in Figure 1.

B. Probabilistic Object Proposals

To ground a referred and target objects we use open set vocabulary object detector f_d, in our case DETIC [28]. Applying f_d on the image I results in N boxes $\{box\}_1^N$, each associated with class names $\{class(box_i)\}_1^N$ and confidence scores $\{P(box_i)\}_1^N$. We calculate the probability of each box being the target object, denoted as $P(box_i|e_t)$. To compute these probabilities, we explore three strategies described below.

First, we use the LVIS vocabulary along with DETIC. Using the target object e_t, we verify whether it exists among the detected box classes $\{class(box_i)\}_1^N$. In this baseline, all boxes with the same name as the target object are considered as candidates for the target object. The same process is applied to the reference object. We refer to this baseline as DETIC. Results are shown in Table III. This limitations of this approach include failures when the exact name of target object class is absent from the LVIS vocabulary or when the object detector fails to detect the object.

To address the issue of missing classes, we select the most similar bounding box based on their predicted classes with the target object e_t. To accomplish this, semantic similarity between detected classes names and text embeddings can be computed. For each box box_i, the predicted class name $class(box_i)$ we compute its and the target object e_t representations using text encoders f_L [5, 2], followed by cosine similarity computation between the two:

$$P(class(box_i)|e_t) = \frac{f_L(class(box_i)) \cdot f_L(e_t)}{\left|f_L(class(box_i))\right| \cdot \left|f_L(e_t)\right|}. \quad (1)$$

A third approach uses only the target and reference object classes names as the object detector vocabulary and is termed as DETIC with the target object. For all the approaches, we sort the predictions based on the computed $P(class(box_i)|e_t)$. The top K boxes with the highest probabilities will be considered as the target object candidates for the subsequent experiments, as shown in Figure 2. This baseline is referred to as DETIC with a language model and the results are reported in Section IV-E, Table V.

C. Spacial Relation Module

For spatial relationship classification, we train a multi-layer perceptron MLP to estimate probabilities of different spatial relationships. In the section III-D, we will use this model to rank the pairs and choose the best matching pair with the expression e. For the pair of boxes one as target and the other as reference object, we first compute 3D geometric features from each box, we refer to this as $\phi(box_i)$. Then, we train our Multilayer Perceptron (MLP), using these 3D geometric features. To compute these 3D geometric representations, we utilize each box's segmentation mask, depth map, and camera intrinsic parameters to reconstruct the 3D point cloud of the object. By employing PCA, a 3D box is fitted into the point cloud. The estimated rotation, translation, box axis sizes, and box center are then concatenated to form the 3D input feature.

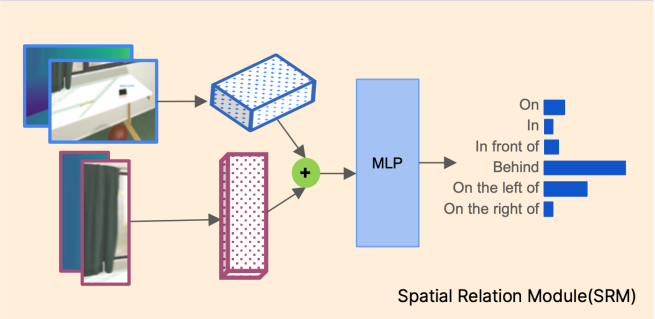

Fig. 3: The Spatial Relation Module. In this part, the cropped target and reference objects are used from both the image and depth data to compute the 3D point cloud. Using PCA, a box is fitted to the point cloud. The center of the box, its dimensions, and the translation and rotation are concatenated together for both objects as inputs to the MLP. The model outputs a distribution over the possible spatial relation classes.

We have also conducted additional experiments to incorporate language priors. The fast-text encoding of candidate box classes was concatenated as a separate input into the MLP. These pieces of information were subsequently be fused together in the architecture to enhance the model's performance.

D. Probabilistic Ranking Module

So far, we have outlined a method to obtain the top K boxes, each with defined $p(class(box_i)|e_t)$. Additionally, we can compute the relation distribution between each pair of boxes using the MLP and obtain the probability $P(\phi(box_i),\phi(box_j)|e_{rel})$. To rank the best pairs of boxes as target and reference objects, we compute the following probability:

$$\begin{aligned} P(box_i, box_j|e) \quad \propto\ & P(box_i)P(class(box_i)|e_t) \\ & P(box_j)P(class(box_j)|e_r) \\ & P(\phi(box_i),\phi(box_j)|e_{rel}) \end{aligned} \quad (2)$$

Where box_i is a candidate for the target object and obx_j is a candidate for the reference object. The probabilities $P(box_i)$ and $P(box_j)$ are derived from the Detic confidence scores.

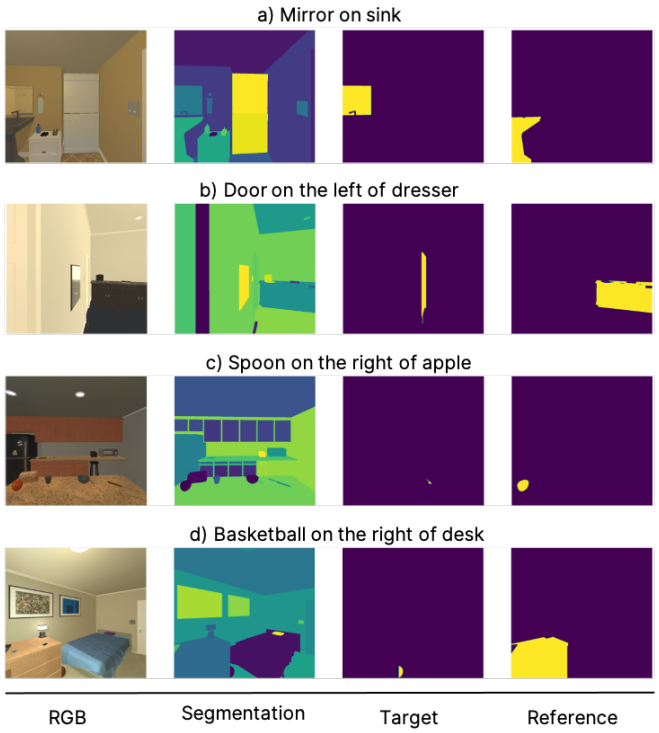

Fig. 4: Sample data from Semantic Abstraction. The dataset showcases challenges arising from small object sizes, occlusions, and clutter. Each data sample from Semantic Abstraction includes semantic segmentation, depth images, and ground truth labels for the target and reference objects, along with expressions provided in triplets.

IV. Experiments

A. Datasets

To investigate the problem of spatial reasoning, we proposed using a synthetic dataset provided by Semantic Abstraction [8]. This dataset comprises a total of 6085 views spread across 100 scenes. The evaluation was conducted on three main subsets: novel visuals, novel synonyms, and novel classes, consisting of 1244, 940, and 597 views, respectively, distributed across 20 test scenes. This dataset is particularly valuable due to the available supervision for both target and reference boxes, in addition to depth information and camera parameters. Furthermore, using the mentioned splits, we can demonstrate the effectiveness of the proposed approach when exposed to domain gaps, both visual and textual. The dataset features six common spatial prepositions: behind, left of, right of, in front, on top of, and inside. Samples of this dataset have been provided in Figure 4.

Moreover, to assess the effectiveness of our pipeline in a real-world setting, we used the SpatialSense [25] dataset. Originally designed for spatial relation recognition, we restructured this dataset to fit for the task of spacial relation detection. The demonstration of real-world data provides insight into the challenges in real-world. The spatial relations available in the dataset are: above, behind, in, in front of, next to, on, to the left of, to the right of, and under. Like Semantic Abstraction, this dataset provides target and reference boxes as supervision. However, depth images and ground truth segmentation masks are not available for the entire dataset. To address this issue, we employed state-of-the-art depth estimation models [26] and Detic for object masks. This dataset provides expressions in the form of triplets similar to Semantic Abstraction. In total, the dataset contains 11569 images, half of which are negative samples that are not useful for our purposes.

B. Evaluation Metric

For the entire pipeline, similar to grounding papers, we propose evaluating our approach based on the accuracy of targeted box predictions with an Intersection over Union (IOU) of more than 0.5 with the ground-truth bounding box. We have also reported the average IOU. Regarding the Spatial Relation Module described in Section III-C, we reported performance based on the accuracy of spatial relation predictions.

C. Training

We utilized Detic with the LVIS vocabulary in a zero-shot setting. In our study, we set the threshold of proposal scores to 0.2 for optimal performance in Probabilistic Object Proposal experiments. For the Spatial Relation Module (SRM), we trained a three-layer MLP with cross-entropy loss and a batch size of 10. The learning rate was set to 0.001 with the ADAM optimizer, and a learning decay of 0.5 after every 3 epochs was applied for a total of 10 epochs on both datasets.

In the last experiment, given that Detic operates in a zero-shot manner, we set the Detic threshold to 0.02. We chose the top 3 boxes based on the Probabilistic Object Proposal. We have reported our results using the 3D geometric Spatial Relation Module (SRM).

D. Spatial Relation Module

In Section III-C, we introduced a classification model designed to categorize relations between every pair of objects. This module was trained using all samples in the training dataset, incorporating language features, 2D geometric features, and 3D geometric features. Specifically, the language-only approach considers FastText embedding of the class of the boxes. The 2D geometric [22] approach only considers the center and box dimensions. For Spatial Sense data, the 2.5D approach incorporates the depth value of the center of the box along with the 2D features. Lastly, the 3D approach considers the 3D box rotation, and translation, along with the center and dimensions of the box, inspiring from [7, 3]. The results of our experiments are presented in Table I.

Our findings indicate a clear advantage of using 3D geometric features over 2D geometric features, especially

Supervision	Approach	Top1 % ↑			Top2 % ↑		
		Visual	Synonyms	Class	Visual	Synonyms	Class
–	Language	35.41	35.42	40.70	60.27	59.99	63.13
2D	Geometric	28.44	28.71	26.12	50.08	45.93	50.09
	Geometric + Language	36.54	35.59	41.26	60.14	60.14	63.47
3D	Geometric	70.33	69.62	71.40	84.31	83.88	86.39
	Geometric + Language	**76.16**	**75.83**	**78.01**	**89.72**	**89.76**	**91.93**

TABLE I: The results for ISRM on Semantic Abstraction Data. The results demonstrate how effectively we can classify the relation between pairs of objects.

Supervision	Approach	Top1 % ↑	Top2 % ↑	Top3 % ↑
–	Language	30.62	40.03	50.41
2D	Geometric	40.42	55.11	66.04
	Geometric + Language	42.35	56.27	66.81
2.5D	Geometric	43.24	57.15	**68.19**
	Geometric + Language	**45.00**	57.75	67.15
3D	Geometric	38.19	52.94	63.03
	Geometric + Language	44.62	**57.76**	68.07

TABLE II: In this table, we have presented the results of the ISRM on real-world data. The results demonstrate how effectively we can classify the relation between pairs of objects in the SpatialSense dataset.

when ground truth depth and camera intrinsics are available. Additionally, we explored the impact of including language features in our classification module, which revealed their potential to enhance performance.

Moreover, our results highlight the robustness of the designed MLP model, even in the presence of visual or textual gaps in the data. This robustness can be attributed to the way the input features are designed, which enables the model to be robust to domain gaps.

Additionally, we conducted the same experiment on the real-world dataset, SpatialSense, table II. Our results reveal that incorporating 3D geometric cues leads to better performance compared to using only 2D cues. However, due to inaccuracies in depth estimation when ground truth camera intrinsics are unavailable (using depth estimation from DepthAnything), the 2.5D approach outperforms the 3D approach. In the 2.5D approach, we calculate the average depth value of points inside the object box and consider it as the depth of the center of the object.

Furthermore, our results indicate that, due to the inherent complexities of real-world data, the classification model generally performs lower accuracy.

E. Probabilistic Object Proposal

In this section, we evaluate Detic as a grounding module, as described in Section III-B. We present three main baselines: Detic with no language model, Detic with a language model, and Detic with the target object class as the vocabulary set. The results are presented in Table III.

An ablation study was conducted on possible language encoders. As shown in the table, we experimented with FastText, BERT using the [CLS] token (BERT+CLS), and BERT with the average of output tokens (BERT+AVG) as the word encoder. We selected the top-ranked box as the targeted box and report the results on the Semantic Abstraction dataset in Table III. The results indicate that the grounding module performs best when the target object is used as the Detic class.

We also applied the two best-performing approaches on SpatialSense and report the results in table V. Based on our experiments, the gap in real data between the two approaches is higher. This is due to the fact that there are more unique classes in the SpatialSense dataset.

Additionally, these results indicate that almost 40% of the grounding process does not require any information about the reference object or relation.

F. Probabilistic Ranking Module

We have demonstrated the performance of the entire pipeline on the Semantic Abstraction dataset in Table IV. Here, we present the top two performing approaches: one utilizing the best-matched box using only Detic (Detic with Top1 and POP), and the other employing the entire pipeline. The results show the robustness and effectiveness of the proposed ranking modules.

Our results are most comparable with the Semantic Abstraction results, presented in Table IV. Although the semantic abstraction results are mean IOU in 3D voxel representations, replicating the results for 2D IOU is non-trivial. However, given the considerable gap, our model performs better.

V. CONCLUSIONS

In conclusion, our paper introduces a probabilistic approach to improve object grounding in Spatial Reasoning (SR). Through extensive experiments on real and synthetic datasets, we demonstrate the effectiveness of integrating 3D geometric cues and novel ranking methods for accurate object localization. Our modular framework addresses challenges like domain shifts and diverse environments, offering a robust solution applicable to robotics and vision-language models.

Approach	Language Model	ACC % ↑			Mean IOU ↑		
		Visual	Synonyms	Class	Visual	Synonyms	Class
Detic	–	24.97	25.31	22.26	0.23	0.24	0.20
	Fast_Text	35.83	35.11	30.43	0.33	0.33	0.28
	Bert+cls	37.92	38.44	30.95	0.35	0.36	0.29
	Bert+avg	38.43	38.51	32.83	0.36	0.36	0.30
	Target Object	**45.76**	**45.46**	**36.27**	**0.42**	**0.41**	**0.33**

TABLE III: This table demonstrates the results of the POP module on Semantic Abstraction data, showcasing the effectiveness of our baseline. The results are shown for three different testing subsets proposed by the dataset. We have reported both the accuracy and the mean IOU to understand the effectiveness of the approach.

Top K	Approach	Language Model	ACC % ↑			Mean IOU ↑		
			Visual	Synonyms	Class	Visual	Synonyms	Class
Target Object	POP	Top 1	53.67	53.86	46.49	0.49	0.49	42.54
	POP+SRM+PRM	Top 3	**55.14**	**55.46**	**47.53**	**0.51**	**0.51**	**0.43**
Semantic Abstraction	–	–	–	–	–	0.19	0.23	0.20

TABLE IV: We have demonstrated the performance of the entire pipeline here. We reported the two best-performing models.

Approach	Classes	Language Model	ACC % ↑	Mean IOU ↑
Detic	Target Object	–	**45.50**	**0.42**
	Lvis	Bert+avg	35.06	0.33

TABLE V: This table demonstrates the results of the POP module on real-world data, Which demonstrates the effectiveness of our baseline.

REFERENCES

[1] Kourosh T Baghaei et al. "Deep Representation Learning: Fundamentals, Technologies, Applications, and Open Challenges". In: *IEEE Access* (2023).

[2] Piotr Bojanowski et al. "Enriching Word Vectors with Subword Information". In: *Transactions of the Association for Computational Linguistics* 5 (2016), pp. 135–146. URL: https://api.semanticscholar.org/CorpusID:207556454.

[3] Boyuan Chen et al. "SpatialVLM: Endowing Vision-Language Models with Spatial Reasoning Capabilities". In: *ArXiv* abs/2401.12168 (2024). URL: https://api.semanticscholar.org/CorpusID:267069344.

[4] Bo Dai, Yuqi Zhang, and Dahua Lin. "Detecting Visual Relationships with Deep Relational Networks". In: *CoRR* abs/1704.03114 (2017). arXiv: 1704.03114. URL: http://arxiv.org/abs/1704.03114.

[5] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *North American Chapter of the Association for Computational Linguistics*. 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[6] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *ArXiv* abs/2010.11929 (2020). URL: https://api.semanticscholar.org/CorpusID:225039882.

[7] Ankit Goyal et al. "Rel3D: A Minimally Contrastive Benchmark for Grounding Spatial Relations in 3D". In: *ArXiv* abs/2012.01634 (2020). URL: https://api.semanticscholar.org/CorpusID:227254225.

[8] Huy Ha and Shuran Song. *Semantic Abstraction: Open-World 3D Scene Understanding from 2D Vision-Language Models*. 2022. arXiv: 2207.11514 [cs.CV].

[9] Ruozhen He et al. "Improved Visual Grounding through Self-Consistent Explanations". In: *ArXiv* abs/2312.04554 (2023). URL: https://api.semanticscholar.org/CorpusID:266055943.

[10] Bowen Jiang and Camillo Jose Taylor. "Hierarchical Relationships: A New Perspective to Enhance Scene Graph Generation". In: 2023. URL: https://api.semanticscholar.org/CorpusID:257496631.

[11] Alexander Kirillov et al. "Segment Anything". In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)* (2023), pp. 3992–4003. URL: https://api.semanticscholar.org/CorpusID:257952310.

[12] Junnan Li et al. "Align before Fuse: Vision and Language Representation Learning with Momentum Distillation". In: *Neural Information Processing Systems*. 2021. URL: https://api.semanticscholar.org/CorpusID:236034189.

[13] Liunian Harold Li et al. “Grounded Language-Image Pre-training”. In: *CoRR* abs/2112.03857 (2021). arXiv: 2112.03857. URL: https://arxiv.org/abs/2112.03857.

[14] Zhuowan Li et al. “Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2022), pp. 14963–14973. URL: https://api.semanticscholar.org/CorpusID:254125164.

[15] Fangyu Liu, Guy Edward Toh Emerson, and Nigel Collier. “Visual Spatial Reasoning”. In: *Transactions of the Association for Computational Linguistics* 11 (2022), pp. 635–651. URL: https://api.semanticscholar.org/CorpusID:248496506.

[16] Jiasen Lu et al. “ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks”. In: *Neural Information Processing Systems*. 2019. URL: https://api.semanticscholar.org/CorpusID:199453025.

[17] Yuchen Mo, Hanbo Zhang, and Tao Kong. “Towards Open-World Interactive Disambiguation for Robotic Grasping”. In: *CoRL 2022 Workshop on Learning, Perception, and Abstraction for Long-Horizon Planning*. 2022. URL: https://openreview.net/forum?id=yrj58pS05_.

[18] Negar Nejatishahidin, Pooya Fayyazsanavi, and Jana Košecka. “Object Pose Estimation using Mid-level Visual Representations”. In: *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2022, pp. 13105–13111. DOI: 10.1109/IROS47612.2022.9981452.

[19] Negar Nejatishahidin et al. “Graph-covis: Gnn-based multi-view panorama global pose estimation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 6458–6467.

[20] Alec Radford et al. “Language Models are Unsupervised Multitask Learners”. In: 2019. URL: https://api.semanticscholar.org/CorpusID:160025533.

[21] Alec Radford et al. “Learning transferable visual models from natural language supervision”. In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[22] Navid Rajabi and Jana Kosecka. “Towards Grounded Visual Spatial Reasoning in Multi-Modal Vision Language Models”. In: *ArXiv* abs/2308.09778 (2023). URL: https://api.semanticscholar.org/CorpusID:261048680.

[23] Sanjay Subramanian et al. “ReCLIP: A Strong Zero-Shot Baseline for Referring Expression Comprehension”. In: *Annual Meeting of the Association for Computational Linguistics*. 2022. URL: https://api.semanticscholar.org/CorpusID:248118561.

[24] Hao Hao Tan and Mohit Bansal. “LXMERT: Learning Cross-Modality Encoder Representations from Transformers”. In: *Conference on Empirical Methods in Natural Language Processing*. 2019. URL: https://api.semanticscholar.org/CorpusID:201103729.

[25] Kaiyu Yang, Olga Russakovsky, and Jia Deng. “SpatialSense: An Adversarially Crowdsourced Benchmark for Spatial Relation Recognition”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), pp. 2051–2060. URL: https://api.semanticscholar.org/CorpusID:160028501.

[26] Lihe Yang et al. “Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data”. In: *CVPR*. 2024.

[27] Hanwang Zhang et al. “Visual Translation Embedding Network for Visual Relation Detection”. In: *CoRR* abs/1702.08319 (2017). arXiv: 1702.08319. URL: http://arxiv.org/abs/1702.08319.

[28] Xingyi Zhou et al. “Detecting Twenty-thousand Classes using Image-level Supervision”. In: *ECCV*. 2022.