Composing Pick-and-Place Tasks By Grounding Language

Oier Mees and Wolfram Burgard*

Abstract Controlling robots to perform tasks via natural language is one of the most challenging topics in human-robot interaction. In this work, we present a robot system that follows unconstrained language instructions to pick and place arbitrary objects and effectively resolves ambiguities through dialogues. Our approach infers objects and their relationships from input images and language expressions and can place objects in accordance with the spatial relations expressed by the user. Unlike previous approaches, we consider grounding not only for the picking but also for the placement of everyday objects from language. Specifically, by grounding objects and their spatial relations, we allow specification of complex placement instructions, e.g. "place it behind the middle red bowl". Our results obtained using a real-world PR2 robot demonstrate the effectiveness of our method in understanding pick-and-place language instructions and sequentially composing them to solve tabletop manipulation tasks. Videos are available at `http://speechrobot.cs.uni-freiburg.de`

1 Introduction

As robots become ubiquitous across human-centered environments the need for natural and effective human-robot communication grows. Natural language provides a rich and intuitive way for humans and robots to interact due to the possibility of referring to abstract concepts. Moreover, many real-world tasks can be effectively described by a series of language instructions. In this work, we aim to develop an approach that enables a robot to solve complex manipulation tasks by understanding a series of unconstrained language expressions characterizing pick-and-place commands. To do so, the robot has to locate unconstrained object categories based on arbitrary natural language expressions, known as referring expression comprehen-

* All authors are with the University of Freiburg, Germany. Wolfram Burgard is also with the Toyota Research Institute, USA. Corresponding author's email meeso@informatik.uni-freiburg.de

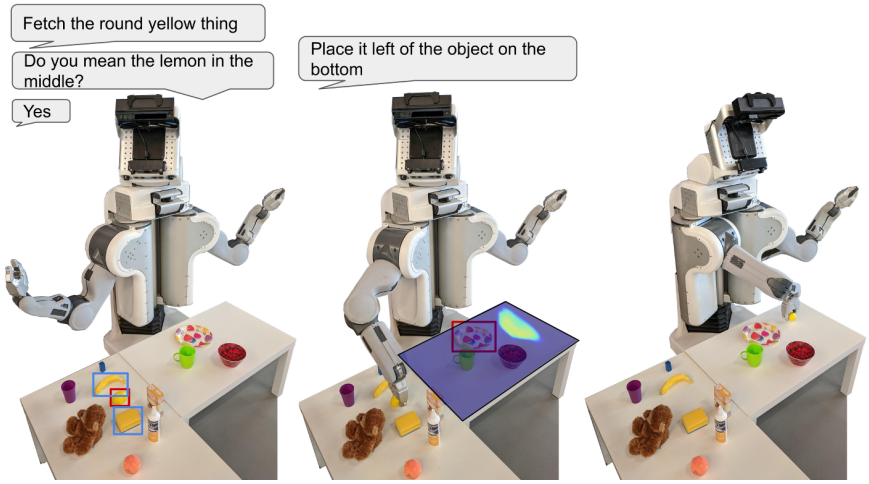

Fig. 1 The goal of our work is to control a robot to perform tabletop manipulation tasks via natural language instructions. Our approach is able to segment objects in the scene, locate the objects referred to in language expressions, solve ambiguities through dialog and place objects in accordance with the spatial relations expressed by the user.

sion, and understand spatial relations to generate object placing locations. In other words, the robot needs to "ground" the referred objects and their spatial relations from language in its world model.

However, understanding unconstrained language instructions is challenging due to the complexity and wide variety of abstract concepts expressed via human language, e.g. "fetch the yellow thing" and "place it left of the bottom object". Moreover, the expression might contain ambiguities because there are several "yellow things" in which case the robot should be able to resolve the ambiguity through dialogue, as shown in Figure 1. Finally, the robot needs to reason about where to place the "yellow thing" relative to the "leftmost container" in order to reproduce the spatial relation "right", which is inherently ambiguous as natural language placement instructions do not uniquely identify a location in a scene.

In this paper, we propose the first comprehensive system for controlling robots to perform tabletop manipulation tasks by sequentially composing unconstrained pick-and-place language instructions. Our approach consists of two neural networks. The first network learns to segment objects in a scene and to comprehend and generate referring expressions. The second network estimates pixelwise object placement probabilities for a set of spatial relations given an input image and a reference object. The interplay between both networks allows for an effective grounding of object semantics and their spatial relationships, without assuming a predefined set of object categories. We demonstrate the effectiveness of our approach by enabling non-expert users to instruct tabletop manipulation tasks to a robot, based on sequences of pick-and-place speech commands.

2 Related Work

Our work is primarily concerned with the task of grounding natural language instructions and spatial relations in the context of the robot's world model [1]. Locating entities in images based on language is closely related to object recognition. Previous works in robotics [2, 3] have addressed semantic object retrieval by training classifiers to recognize predefined object categories. These approaches are limited in real-world scenarios as they are not capable of handling variation in the users natural language descriptions and are restricted to a small number of objects.

Spatial relations also play a crucial role in understanding natural language instructions [4, 5], as objects are often described in relation to others in tasks such as object placing [6, 7, 8] or human robot interaction [2, 9, 10]. Concretely, spatial relations help the robot disambiguate multiple instances of the same object and to define target areas for placing the picked objects. In our previous work, we introduced a novel method to predict pixelwise object placement probability distributions for a set of commonly used prepositions in natural language [7]. In contrast, we relax the assumption of having a single reference object on the tabletop and add a grounding model to effectively place arbitrary objects in a scene that contains multiple objects.

Recently, there has been significant progress made towards systems that can demonstrate their visual understanding by generating or responding to natural language in the context of images [11, 12, 13, 14, 15]. To learn joint visual-linguistic representations, state-of-the-art approaches use convolutional neural networks to encode visual features and recurrent neural networks to process language, replacing traditional handcrafted visual features and language parsers. We leverage advances in modular networks [16, 17, 18] for referential expression comprehension. This allows decomposing language into modular components related to subject appearance, location, and relationship to other objects, flexibly adapting to expressions containing different types of information in an end-to-end fashion.

Most related to our approach are the works by Shridhar *et al.* [9] and Hatori *et al.* [4], as both use an interactive fetching system to localize objects mentioned in referring expressions with bounding boxes. We tackle temporally more extended tasks, using our model which enables complex object placement commands such as "place the cup on top of the leftmost box". Notably, sequentially composing pick-and-place language instructions can lead to desirable high-level behaviours, such as tidying up a tabletop or table setting for example. Finally, in contrast to the template-based picking approaches of prior interactive fetching systems [9, 4] we leverage state-of-the-art methods for grasping novel objects with 6-DOF grasps [19].

3 Method Description

In this section we describe the technical details of our method to control a robot to perform tabletop manipulation tasks via natural language instructions. Our approach relies on two models: a grounding model that identifies the most likely object referred

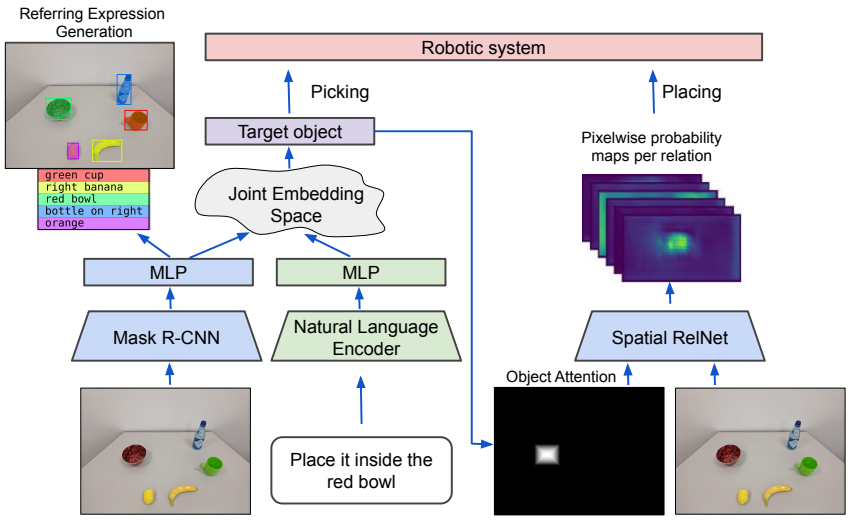

Fig. 2 Overview of the system architecture. Our grounding network processes the input sentence and visual object candidates detected with Mask-RCNN [20] and performs referential expression comprehension. Additionally, it generates referential expressions for each object candidate to disambiguate unclear instructions. Once the reference object of a relative placement instruction has been identified, a second network predicts object placing locations for a set of spatial relations.

by a language instruction and a neural network that predicts object placing locations conditioned on a set spatial relation. An overview of the system is given in Figure 2.

3.1 Target Object Selection

We start off by detecting and segmenting all objects in the scene. We train a semantic segmentation network based on Mask-RCNN [20] with a Resnet-101 backbone, which extracts a set of region proposals or object candidates o_i from an image. After all objects on the scene are recognized, we need to identify which object the user is referring to in its language instruction. Given an input image I and expression r, the target object selection is formulated as a task to find the best bounding box from the set of predicted candidate boxes $O = \{o_i\}_{i=1}^{N}$. Our grounding model is based on MAttNet [16], a modular referring expression comprehension network. To enable human-robot communication in cases of ambiguous instructions, we have extended it to support the generation of self-referential expressions, described in Section 3.2.

The candidate regions are encoded by a neural network consisting of three modular grounding components related to subject appearance, location and relationship to other objects. These modules combine image features encoded by a Resnet-101 network with relational and geometric features pertaining to the neighborhood or context of each candidate region. The language expression r is encoded in a word

embedding layer, which encodes each word in the input sentence to a vector representation, followed by a bi-directional Long Short-Term Memory (LSTM) and a fully-connected (FC) layer. Additionally, the language network learns two types of attention: attention weights that are computed on each word for each module and are summarized as phrase embedding $q^m \mid m \in \{\text{subj}, \text{loc}, \text{rel}\}$, and module weights $[w_{subj}, w_{loc}, w_{rel}]$ that estimate how much each module contributes to the overall expression score. Each visual module computes scores for each object candidate by calculating the cosine similarity between the vector representation of the instruction, and that of the candidate image region. Finally, the output module takes a weighted average of these scores to get an overall matching score $S(o_i \mid r) = w_{subj} S(o_i \mid q^{subj}) + w_{loc} S(o_i \mid q^{loc}) + w_{rel} S(o_i \mid q^{rel})$. During training, we sample triplets consisting of a positive match (o_i, r_i) and two random negative samples (o_i, r_j) and (o_k, r_i), where o_k is some other object and r_j is an expression describing some other object in the same image to apply a hinge loss:

$$\begin{aligned} \mathscr{L}_1 = \sum_i [&\lambda_1 \max(0, m_1 + S(o_i \mid r_j) - S(o_i \mid r_i)) \\ &+ \lambda_2 \max(0, m_1 + S(o_k \mid r_i) - S(o_i \mid r_i))]. \end{aligned} \tag{1}$$

3.2 Resolving Ambiguities

If the referred object cannot be uniquely identified by the grounding model, the system needs to ask for clarification from the human operator. Inspired by recent advances in image caption generation and understanding [13, 21, 22], we incorporate a LSTM based captioning module to our grounding network that allows the robot to describe each detected object with a natural language description. Our referring expression generation module is jointly trained with our grounding network and shares the features used in the three modules related to subject appearance, location and relationship to other objects. Specifically, the visual target object representation v_i^{vis} is modeled by a concatenation of the ResNet-101 C3 and C4 features, followed by one FC layer which is shared with the comprehension network and one exclusive FC layer. To facilitate the generation of referential expressions that contain location information, such as "the cup in the middle", we leverage the representation learned by the location module v_i^{loc}. This module combines a 5-d vector representing the top-left position, bottom-right position and relative area to the image for the candidate object, together with a relative location encoding of up to five surrounding objects of the same category. Finally, we integrate the output of the relationship module v_i^{rel}, which encodes the appearance and localization offsets of up to five category-agnostic objects in the targets surroundings to enable modeling sentences such as "the teddy bear on top of the box". The final visual representation for the target object is then a concatenation of the above features $v_i = [v_i^{vis}, v_i^{loc}, v_i^{rel}]$. The model is trained to generate sentences r_i by minimizing the negative log-likelihood:

$$\mathscr{L}_2 = -\sum_i \log P(r_i \mid v_i). \tag{2}$$

To generate discriminative sentences, we use a Maximal Mutual Information constraint proposed by Mao *et al.* [21] that encourages the generated expression to describe the target object better than the other objects within the image. Concretely, given a positive match (o_i, r_i) we sample a negative (o_k, r_i), where o_k is some other object, and optimize the following max-margin loss:

$$\mathscr{L}_3 = \sum_i [\lambda_3 \max(0, m_2 + \log P(r_i \mid v_k) - \log P(r_i \mid v_i)). \tag{3}$$

In order to detect if an instruction is ambiguous, we leverage the max-margin loss the comprehension model is trained with. Concretely, during training the max-margin loss aims to guarantee that every correct pair of a sentence and an object has scores by a margin m_1 than any other pair with a wrong object or sentence. Therefore, if at test time there are more than one objects within that threshold, we consider them potential targets. For each candidate we generate multiple self-referential expressions via beam search and use the comprehension module to rerank these expressions and select the least ambiguous expression, similar to Yu *et al.* [22]. We then let the system ask the human "Do you mean ...?". After asking the question, the user can respond "yes" to choose the referred object or "no" to continue iterating through other possible objects. Alternatively, the user can provide a specific correcting response to the question, e.g., "no, the banana on the right", in which case we re-run our grounding module.

3.3 Relational Object Placement

Once an object has been picked, our system needs to be able to place it in accordance with the instructions from the human operator. We combine referring expression comprehension with the grounding of spatial relations to enable complex object placement commands such as "place the ball inside the left box". Given an input image I of the scene and the location of the reference item, identified with our aforementioned grounding module, we generate pixelwise object placement probabilities for a set of spatial relations by leveraging the Spatial-RelNet architecture we introduced in our previous work [7]. We consider pairwise relations and express the subject item as being *in relation to* the reference item. We model relations for a set of commonly used natural language spatial prepositions $C = \{\texttt{inside}, \texttt{left}, \texttt{right}, \texttt{in front}, \texttt{behind}, \texttt{on top}\}$. As natural language placement instructions do not uniquely identify a location in a scene, Spatial-RelNet predicts non-parametric distributions to capture the inherent ambiguity. A key challenge to learning such pixelwise spatial distributions is the lack of ground-truth data. Spatial-RelNet overcomes this problem by leveraging a novel auxiliary learning formulation, as shown in Figure 3. During training, pixel locations (u, v) are sampled

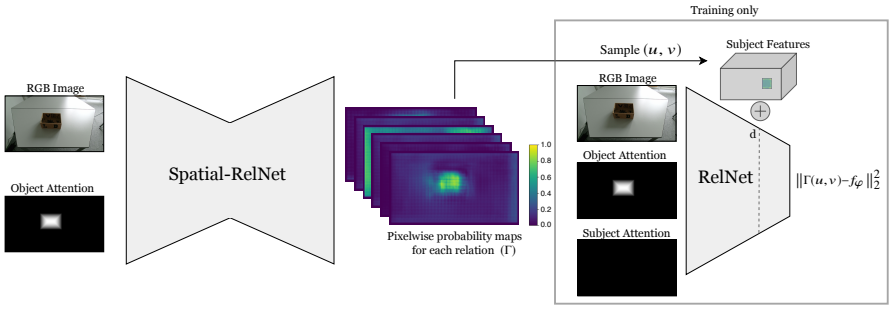

Fig. 3 Our Spatial-RelNet [7] network processes the input RGB image and an object attention mask to produce pixelwise probability maps Γ over a set of spatial relations. During training, we sample locations (u, v) according to Γ, implant inside an auxiliary classifier network at the sampled locations high level features of objects and classify the hallucinated scene representation to get a learning signal for Spatial-RelNet. At test time the auxiliary network is not used.

according to the probability maps Γ produced by Spatial-RelNet. To get the learning signal, high level features of objects are implanted into a pretrained auxiliary classifier f_φ to compute a posterior class probability over relations. This way, we can reason over what relation would most likely be formed if we placed an object at the given location.

4 System Implementation

4.1 Machine Learning Setup

During training, we sample the same triplets for both the object comprehension module and the expression generation module. We set the margin $m_1 = 0.1$ for the comprehension ranking and $m_2 = 1.0$ for the generation loss. We additionally use MAttNet's auxiliary visual attribute classification loss. We use the Adam optimizer to train the joint model with an initial learning rate of 0.0004. For the contrastive pairs, we set $\lambda_1 = 1$, $\lambda_2 = 1$ and$\lambda_3 = 0.1$. We make the word embedding of the comprehension and generation modules shared to reduce the number of parameters. For implementation details of Spatial-RelNet, we refer to the original paper [7].

4.2 Robot Setup

To pick an object from language, we first identify the object with our grounding model and extract the corresponding segmentation mask of the selected object. We use an Amazon Echo Dot device to synthesize the voice instructions. We localize the

object in 3D space and generate grasp poses with Grasp Pose Detection (GPD) [19], which predicts a series of 6-DOF candidate grasp poses given a 3D point cloud for a 2-finger grasp. The reachability of the proposed candidate grasps are checked using MoveIt!, and the highest quality reachable grasp is executed with the PR2 robot. For placing the object, we first sample a location from the spatial distribution predicted by our Spatial-RelNet model. We rely on keyword spotting to select the corresponding predicted distribution. Next, we localize the pixel coordinate in 3D space and plan a top-down grasp pose to the calculated 3D point. Finally, the end-effector is moved above the desired location and then the gripper is opened to complete the placement.

5 Experiments

We evaluate our approach under two settings. First, we evaluate the capability of our approach to comprehend and generate referring expressions for a wide variety of objects on the RefCOCO dataset [11]. Next, we evaluate the ability of our robotics system to follow pick-and-place language instructions in human-robot experiments.

5.1 RefCOCO Benchmark

The RefCOCO dataset contains images and corresponding referring expressions that uniquely identify a wide variety of objects in the images. We compare our grounding networks ability to comprehend and generate referring expressions against several strong baselines on Table 1. For evaluating the comprehension, we compute the intersection-over-union (IoU) of the selected region with the ground-truth bounding box, considering IoU $>$ 0.5 a correct comprehension. To evaluate the generation module, we leverage standard machine translation metrics commonly used in image captioning, such as METEOR and CIDEr. We observe that by jointly training the comprehension and language generation modules, they regularize each other and improve their respective performances, demonstrating the effectiveness of multitask learning [23, 22, 24].

	RefCOCO comprehension			RefCOCO generation			
	val	TestA	TestB	TestA		TestB	
				Meteor	CIDEr	Meteor	CIDEr
Mao [21]	-	63.15	64.21	-	-	-	-
INGRESS [9]	77	76.7	77.7	-	-	-	-
SLR [22, 4]	79.56	78.95	80.22	0.268	0.697	0.329	1.323
MAttNet [16]	85.65	85.26	84.57	-	-	-	-
Ours	**86.15**	**87.18**	**85.36**	**0.29**	**0.753**	**0.33**	**1.33**

Table 1 Referring expression comprehension and generation on the RefCOCO dataset, with human-annotated ground-truth object regions.

5.2 Robot Experiments

We evaluate our approach on two real-world scenarios: picking and placing objects according to user defined object arrangements and a tidy-up task. We will first describe the setup of the object arrangement experiment. Our study involved 4 participants recruited from a university community[2]. The robots workspace contained two tables, as shown in Figure 1. One table contained previously unseen objects in clutter. The second table contained a single reference object. The average number of objects on the cluttered table was 5.6. The participants were asked to instruct a PR2 robot to arrange a desired target scene by picking objects from the cluttered table and using relational expressions to place them on the second table. In addition to the robots RGB-D camera we placed a second camera in front of the cluttered table and performed online registration to compute a global point cloud. The tidy-up task consisted of iteratively picking 4 colored objects from the cluttered table and placing the same colored objects on the left container and the remaining objects on the right container. In this experiment we were interested in evaluating the number of actions the robot has to take to complete the task, given unambiguous instructions.

Table 2 shows the performance of our approach on a PR2 robot for the first experiment. Our approach achieves a 78.3% target object selection accuracy and a 85.7% accuracy on selecting the reference object the placing will be relative to. The

	Target Object Selection	Target Object Grasping	Placing Base Grounding	Placing Success	Avg. Number of Feedback	Pick and Place
Ours	78.3% (47/60)	74.4% (35/47)	85.7% (30/35)	83.3% (25/30)	0.63 (60/95)	63% (60/95)

Table 2 Performance of our approach on a real robot platform following natural language instructions to pick and place objects in a tabletop scenario.

higher accuracy of the latter is due to fewer candidate objects being on the placing table and the participants preferring to use ambiguous expressions for the picking instructions. The robot took $\sim$ 20 seconds to complete an action from the moment the human started to speak. We report a grasping performance of 74.4% with GPD. We find that some objects such as mugs are particularly difficult for GPD as it often fails to find feasible grasps due to either occluded object parts or noisy measurements on thin structures such as rims. Our object placement approach achieves a success rate of 85.7%. We observe some failure cases for large object placements, because of missing 3D priors of the objects to be placed. Thus, when placing a big box left of a small box, it is possible that the chosen placement results in the big box partially ending up on top of the small box. For the tidy up task, we report a mean task length of 14.4 actions, due to several re-grasp attempts. Overall, our results demonstrate the ability of our approach to allow non-expert users to instruct tabletop manipulation tasks based on sequences of pick-and-place speech commands.

[2] Further quantitative experiments were infeasible at time of submission due to COVID-19.

6 Conclusions and Discussion

In this paper, we proposed the first robotic system that allows non-expert users to instruct tabletop manipulation tasks by sequentially composing unconstrained pick-and-place language instructions and can clarify a human operator's intention through dialogue. We demonstrate the effectiveness of our approach to encode high-level behaviours in a highly challenging, realistic environment. Even though we are far from achieving robots that can learn to relate human language to their world model, we hope our work is a step in this direction.

While the experimental results are promising, our approach has several limitations. First, relative object placement instructions do not allow for fine-grained target specification due to its inherent ambiguity. Addressing this issue would require learning user preferences from feedback [25]. Second, we observe some failure cases for large object placements, because of missing 3D priors of the objects to be placed. Integrating 3D priors is a natural extension to enable optimizing placement poses [26] and to reason over the effects of actions on the scene [27]. Third, we find that GPD often fails to find feasible grasps due to either occluded object parts or noisy measurements. Integrating methods that can complete occluded scene regions [28, 29] or generate more diverse grasps [30] might help alleviating these problems. Finally, our approach is limited to tabletop tasks that can be characterized by pick-and-place actions. An exciting area for future work may be one that not only grounds object semantics and spatial relations, but also grounds actions in order to learn complex behaviours with language conditioned continuous control policies [31, 32].

Acknowledgments

This work has been supported partly by the Freiburg Graduate School of Robotics and the German Federal Ministry of Education and Research under contract number 01IS18040B-OML. We thank Henrich Kolkhorst for his contributions to the speech-to-text pipeline and to Andreas Eitel for valuable discussions.

References

1. Herbert H Clark and Susan E Brennan. Grounding in communication. *Prespectives on Socially Shared Cognition*, 1991.
2. Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, and Trevor Darrell. Grounding spatial relations for human-robot interaction. In *IROS*, 2013.
3. Dejan Pangercic, Benjamin Pitzer, Moritz Tenorth, and Michael Beetz. Semantic object maps for robotic housework-representation, acquisition and use. In *IROS*, 2012.
4. Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *ICRA*, 2018.

5. Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. In *RSS*, 2016.
6. Yun Jiang, Marcus Lim, Changxi Zheng, and Ashutosh Saxena. Learning to place new objects in a scene. *IJRR*, 2012.
7. Oier Mees, Alp Emek, Johan Vertens, and Wolfram Burgard. Learning object placements for relational instructions by hallucinating scene representations. In *ICRA*, 2020.
8. Oier Mees, Nichola Abdo, Mladen Mazuran, and Wolfram Burgard. Metric learning for generalizing spatial relations to new objects. In *IROS*, 2017.
9. Mohit Shridhar and David Hsu. Interactive visual grounding of referring expressions for human-robot interaction. In *RSS*, 2018.
10. Dipendra K Misra, Jaeyong Sung, Kevin Lee, and Ashutosh Saxena. Tell me dave: Context-sensitive grounding of natural language to manipulation instructions. *IJRR*, 2016.
11. Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014.
12. Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *ICCV*, 2015.
13. Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *CVPR*, 2016.
14. Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.
15. Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.
16. Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L Berg. Mattnet: Modular attention network for referring expression comprehension. In *CVPR*, 2018.
17. Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. Modeling relationships in referential expressions with compositional modular networks. In *CVPR*, 2017.
18. Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *CVPR*, 2016.
19. Marcus Gualtieri, Andreas Ten Pas, Kate Saenko, and Robert Platt. High precision grasp pose detection in dense clutter. In *IROS*, 2016.
20. Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *CVPR*, 2017.
21. Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, 2016.
22. Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L Berg. A joint speaker-listener-reinforcer model for referring expressions. In *CVPR*, 2017.
23. Rich Caruana. Multitask learning. *Machine learning*, 1997.
24. Oier Mees, Markus Merklinger, Gabriel Kalweit, and Wolfram Burgard. Adversarial skill networks: Unsupervised robot skill learning from videos. In *ICRA*, 2020.
25. Nichola Abdo, Cyrill Stachniss, Luciano Spinello, and Wolfram Burgard. Organizing objects by predicting user preferences through collaborative filtering. *IJRR*, 2016.
26. Joshua Alexander Haustein, Kaiyu Hang, Johannes A Stork, and Danica Kragic. Object placement planning and optimization for robot manipulators. In *IROS*, 2019.
27. Iman Nematollahi, Oier Mees, Lukas Hermann, and Wolfram Burgard. Hindsight for foresight: Unsupervised structured dynamics models from physical interaction. In *IROS*, 2020.
28. Oier Mees, Maxim Tatarchenko, Thomas Brox, and Wolfram Burgard. Self-supervised 3d shape and viewpoint estimation from single images for robotics. In *IROS*, 2019.
29. Jacob Varley, Chad DeChant, Adam Richardson, Joaquín Ruales, and Peter Allen. Shape completion enabled robotic grasping. In *IROS*, 2017.
30. Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *ICCV*, 2019.
31. Corey Lynch and Pierre Sermanet. Grounding language in play. *arXiv preprint arXiv:2005.07648*, 2020.
32. Lin Shao, Toki Migimatsu, Qiang Zhang, Karen Yang, and Jeannette Bohg. Concept2robot: Learning manipulation concepts from instructions and human demonstrations. In *RSS*, 2020.