Robotics: Science and Systems 2021
Held Virtually, July 12–16, 2021

INVIGORATE: Interactive Visual Grounding and Grasping in Clutter

Hanbo Zhang[1,2*], Yunfan Lu[2*], Cunjun Yu[2], David Hsu[2], Xuguang Lan[1†] and Nanning Zheng[1]
[1]Xi'an Jiaotong University, [2]National University of Singapore

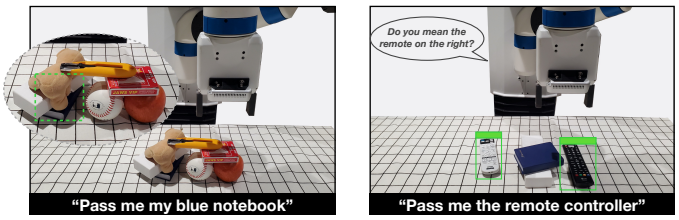
***Abstract*—This paper presents INVIGORATE, a robot system that interacts with human through natural language and grasps a specified object in clutter. The objects may occlude, obstruct, or even stack on top of one another. INVIGORATE embodies several challenges: (i) infer the target object among other occluding objects, from input language expressions and RGB images, (ii) infer object blocking relationships (OBRs) from the images, and (iii) synthesize a multi-step plan to ask questions that disambiguate the target object and to grasp it successfully. We train separate neural networks for object detection, for visual grounding, for question generation, and for OBR detection and grasping. They allow for unrestricted object categories and language expressions, subject to the training datasets. However, errors in visual perception and ambiguity in human languages are inevitable and negatively impact the robot's performance. To overcome these *uncertainties*, we build a partially observable Markov decision process (POMDP) that integrates the learned neural network modules. Through approximate POMDP planning, the robot tracks the history of observations and asks disambiguation questions in order to achieve a near-optimal sequence of actions that identify and grasp the target object. INVIGORATE combines the benefits of model-based POMDP planning and data-driven deep learning. Preliminary experiments with INVIGORATE on a Fetch robot show significant benefits of this integrated approach to object grasping in clutter with natural language interactions. A demonstration video is available online[1].**

I. Introduction

Robots are gradually, but surely entering into our daily life. To become effective human helpers, robots must understand our physical world through visual perception and interact with humans through natural languages. Consider the robot task of following a human instruction to retrieve an object from a cluttered kitchen table (Fig. 1). This seemingly simple task presents multiple challenges:

- Infer the target object among other occluding objects from input language expressions and images;
- Infer object blocking relationships from images;
- Synthesize a multi-step plan to ask questions, if necessary, to disambiguate the target object, and retrieve it successfully despite other obstructing objects.

Advances in deep learning provide powerful neural network (NN) models to process complex visual and language inputs and thus address the first two challenges. However, they alone are not sufficient for two main reasons. First, visual inputs are complex and noisy, which invariably cause errors in perceptual processing. Cluttered scenes are inherently *partially observable* and exacerbate the difficulty of perceptual processing. For example, the target object may not be detected at all because of visual occlusions (Fig. 1*a*). Second, despite their richness, human languages are sometimes ambiguous. Two distinct objects may perfectly match the language specification (Fig. 1*b*). A natural question then arises: *How can we harness the power of these learned NN models for perceptual and language processing and achieve robust robot performance?*



(*a*) (*b*)

Fig. 1: Interactive visual grounding and grasping in clutter. The robot receives from the human a natural language instruction to retrieve an object. It tries to identify the target object visually, asks questions to disambiguate the target object, if necessary, and eventually grasps the object. (*a*) Perceptual uncertainties. The object detection module fails to detect the target object because of visual occlusion. (*b*) Language ambiguity. The instruction is ambiguous, as there are two objects both satisfying the instruction. The robot then asks questions to disambiguate.

To this end, we have developed and experimented with a robot system, *INteractive VIsual GrOunding and gRAsp in clutTEr* (INVIGORATE). See Fig. 1 for examples. INVIGORATE integrates data-driven learning and model-based planning. To handle complex visual inputs and language interactions, we train separate NN models for object detection, for visual grounding, for question generation, and for object relationship detection and grasping. We then build a partially observable Markov decision process (POMDP) that integrates the learned NN modules. In the INVIGORATE POMDP, we model the NN outputs—the detected target object, other objects, and object blocking relationships—as noisy observations and learn a probabilistic observation model of detection failures. Through POMDP planning, INVIGORATE tracks the history of observations over time and obtains a robust probabilistic estimate of the true underlying state, despite uncertainties in perceptual and language processing. Further, we introduce

† Correspondence to Xuguang Lan: xglan@mail.xjtu.edu.cn.
* Equal Contributions.
[1]https://youtu.be/zYakh80SGcU

an explicit action for asking disambiguation questions in the POMDP model. If the language instruction for the target object is ambiguous and there are multiple candidates, INVIGORATE may ask questions and gather information actively for disambiguation. It reasons systematically about the uncertainty of the estimated target object and trades off the benefit of additional information for disambiguation against the cost of asking questions.

We deployed INVIGORATE on a Fetch robot. Experimental results show that INVIGORATE achieves an overall success rate of 83% on our test dataset and consistently outperforms a baseline without POMDP integration. Ablation studies further confirm the importance of reasoning about uncertainties in dealing with noisy visual perception and language ambiguity.

One main contribution of this work is to demonstrate a principled approach that integrates data-driven deep learning and model-based planning for a complex robot task. We build a POMDP model connecting three key elements: robot perception, action, and human interaction. The learned NN models enable INVIGORATE to handle complex visual inputs and language interactions. Model-based POMDP planning enables INVIGORATE to achieve robust performance overall, despite uncertainties in perceptual and language processing.

II. Related Work

1) Visual grounding: Visual grounding has been studied extensively in the computer vision community [28]. Early work usually restricts the language expressions or visual concepts [17, 26, 27]. With advances in deep learning, significant progress has been achieved in visual grounding of objects in the open world [4, 20, 25, 31, 35, 40, 41], but usually objects are clearly visible with few occlusions. In contrast, INVIGORATE focuses on visual grounding in clutter. It achieves robust performance by integrating historical observations and actively interacting with the world.

2) Human-robot language interaction: Visual grounding allows the human to specify an object for grasping through the natural language. However, the robot may fail to identify the intended object because of uncertainties in visual perception or language ambiguities. To disambiguate, the robot needs additional information, which can be acquired by asking questions. Some earlier HRI systems can interact with humans verbally [12, 30, 37]. However, they typically use predefined visual or linguistic concepts for interaction. In contrast, INVIGORATE takes advantage of recent advances in deep learning for visual grounding and generates object-specific questions to facilitate goal-directed grasping.

3) Goal-directed object grasping in clutter: Object grasping is a long-standing challenge in robotics [1]. In recent years, both data-driven deep learning and model-based planning helped bring about significant progress in object grasping in general [9, 21, 22, 29] and goal-directed object grasping in particular [8, 13, 24, 42, 44]. Several recent methods aim at goal-directed object grasping in clutter scenes [6, 18, 44]. Specifically, Zhang et al. [44] propose to learn object blocking relationships for grasping. However, they do not address the issue of language interactions between the robot and the human. To this end, Chen et al. [5] and Hatori et al. [11] propose to directly fuse visual and linguistic features in neural networks. Shridhar et al. [31] formulate a POMDP to ask disambiguating questions in interactive grasping tasks. Mees and Burgard [23] propose a robot system capable of grounding language instructions for both object picking and placement. They, however, do not consider visual occlusion and physical obstruction in dense object clutter. INVIGORATE tackles both challenges together: object grasping in clutter and natural language interaction with the human.

4) Integrating learning and planning: INVIGORATE benefits significantly from the integration of learning and planning, an active research direction that has attracted much attention recently. Learning and planning interact in many interesting ways. One very common idea is to learn models for planning. To scale up complex decision making, both Silver et al. [32, 33, 34] and Cai et al. [2] use planning in the short term and use learning for long-term prediction. They aim to avoid the prohibitive cost of long-horizon search and improve computational efficiency. Another idea is to embed a planning algorithm in the neural network and train it end-to-end [15, 36]. These differentiable algorithm networks [16] are structured, interpretable, task-driven, and robust, thus combing the benefits of model-based planning and data-driven learning. INVIGORATE uses a model-based approach to integrate multiple learned NN modules and reasons about their uncertainties systematically in order to achieve robust robot performance.

III. Overview

INVIGORATE takes from the human a natural language instruction that refers to an object of interest. It uses both the referring expression and the image from the visual sensor to identify the target object in a clutter. If the referring expression is ambiguous, INVIGORATE asks the human simple questions for disambiguation and eventually grasps the target object without unnecessarily perturbing other objects. See Fig. 2 for an overview of the system.

INVIGORATE integrates data-driven deep learning with model-based POMDP planning. We train four NN modules, O-Net, R-Net, G-Net, and Q-Net, from data. At each time step, O-Net extracts from the input image I a set of object proposals. Based on these object proposals, R-Net further processes I and extracts pairwise blocking relationships among the objects, as well as candidate object grasps. G-Net uses the input referring expression E and the object proposals for visual grounding and outputs a set of candidates for the target object. If E is ambiguous, there may be multiple candidates. INVIGORATE uses Q-Net to generate a referring expression to a candidate. It fits the generated expression into a question template and asks the human a disambiguation question, e.g., "Do you mean the cup on top?".

The trained NN models are powerful and allow for unrestricted object categories and language expressions, subject to the training datasets. However, the outputs of O-Net, R-Net, and G-Net are all noisy, because of sensor noise, visual

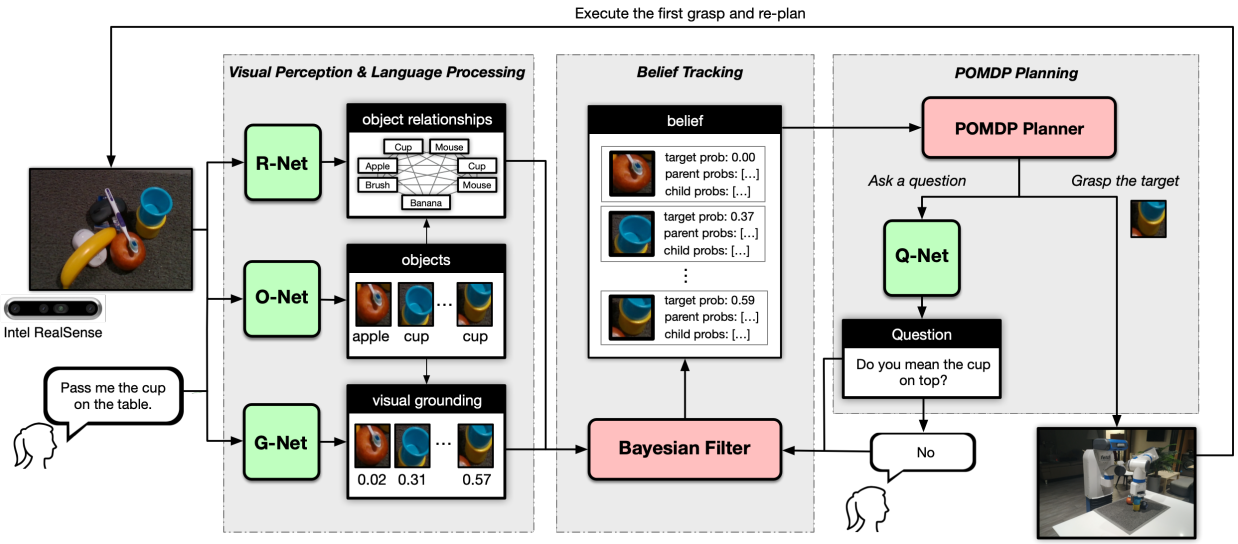


Fig. 2: An overview of INVIGORATE.

occlusions, and ambiguity in human languages. To achieve robust robot performance despite these uncertainties, we build a POMDP model that integrates the learned NN modules. INVIGORATE maintains a *belief*, i.e., a distribution over the underlying state, which consists of the target object, the other objects, and their blocking relationships. It treats the NN outputs as noisy *observations* on the state. At each step, INVIGORATE updates the belief with new observations and actions, through Bayesian filtering. The belief summarizes the history of observations and actions; it quantifies the uncertainties probabilistically and provides the basis for principled decision-making. Given a belief, INVIGORATE performs POMDP planning through look-ahead search to choose the best action. INVIGORATE models two types of actions: ask a disambiguation question or grasp an object. If the uncertainty on the target object is high according to the belief, then INVIGORATE invokes Q-Net to generate a question to gather additional information. If the uncertainty is low, INVIGORATE grasps either the target object directly or an obstructing object according to the estimated object blocking relationships. By reasoning about the belief, POMDP planning enables INVIGORATE to choose a near-optimal sequence of actions.

IV. Neural Networks for Visual Perception and Language Interaction

INVIGORATE takes advantage of deep learning to build back-end perceptual and interaction modules. We describe each of these modules here.

A. O-Net for Object Detection and Tracking

Object detection provides a set of object proposals $\mathcal{B}$ for INVIGORATE. Noticeably, our states and beliefs are all object-based. Thus, it is necessary to not only detect objects but also track objects across different steps so as to update the belief according to the observations.

In INVIGORATE, we apply the well-known Cascade R-CNN [3] as the base detector O-Net to provide object proposals for every single step. The detector is trained on the union of COCO [19] and VMRD [43] to support a wide variety of objects while keeping good generality. To track objects across multiple steps, we maintain an object pool $\mathcal{B}$. In each step, we first feed the raw image to the base detector to obtain a set of proposals $\mathcal{B}^D$. Then, we feed all historical proposals in $\mathcal{B}$ to the object detector to re-classify them and get a historical set $\mathcal{B}^H$. Subsequently, we merge $\mathcal{B}^H$ into $\mathcal{B}^D$ using Hungarian algorithm with the cost function defined as:

$$H(B_i, B_j) = \alpha_1 IoU(B_i, B_j) + \alpha_2 \left\| Score_i - Score_j \right\| \quad (1)$$

where $IoU(B_i, B_j)$ is the *intersection of union* (IoU) between B_i and B_j, and $Score$ means the normalized class scores given by the object detector. Intuitively, bounding boxes with large IoU and the same category will be merged into one. Finally, the object pool $\mathcal{B}$ will be updated by $\mathcal{B}^D \cup \mathcal{B}^H$ using Hungarian algorithm again with the same cost function defined in Eq. 1. Such a detection procedure is more robust against false positives and false negatives. The merging process based on the Hungarian algorithm also enables object tracking across different steps, which is a prerequisite for belief update.

B. G-Net for Visual Grounding

G-Net takes an image I, a referring expression E, and detected object proposals in $\mathcal{B}$ to estimate the matching scores between each detected object i and the referring expression E:

$$f_i^g = f_\theta^g(i, E|I) \quad (2)$$

where f_θ^g denotes G-Net and f_i^g denotes the output matching score. In INVIGORATE, we train G-Net following Yu et al. [41] on RefCOCO dataset. G-Net splits the user expression into three parts, subject description, locational description, and relational description. It extracts the visual feature for each proposal and performs separate visual linguistic matching. To illustrate, given the expression "the blue cup to the right of the book", such a sentence would be decomposed into a subject description, "the blue cup", a locational description "to the right of" and a relational description "the book".

An embedding is obtained for each description through a language attention network. The image is fed into a CNN-based neural network to generate a visual feature map, which is then pooled into region features using proposals in $\mathcal{B}$. After that, we estimate a similarity score between phrases and pooled visual features together with their location information with a trainable layer. Finally, an attention-based summation of similarity scores for three phrases is used to determine the referred object. All layers are trained in an end-to-end manner. As a result, G-Net is scalable and can handle almost unrestricted referring expressions. e.g., "the red apple", "the apple on top", "the blue cup to the right of the book", etc.

C. Q-Net for Question Generation

Q-Net generates object-specific questions when needed. In INVIGORATE, we apply relational captioning to generate object descriptions since self-referential captioning [31], which describes an object in terms of its attributes, e.g. name, color or shape, suffers severely from occlusions in clutters.

Our Q-Net takes as input the visual and spatial features of the object of interest i and a context object i_c, which may be the whole image. It then generates a relational caption in an auto-regressive manner, i.e., words are generated one by one based on the input features and previous words. To ensure that the description is informative, we feed all possible pairs consisting of the object of interest and a context object to Q-Net and selects the caption that has the largest probability.

$$S^* = \arg\max_{S_c \in \mathcal{S}} P(S_c | i, i_c, I) \tag{3}$$

where $\mathcal{S}$ includes all pair-wise captions w.r.t. object i. We follow Shridhar et al. [31] and Nagaraja et al. [25] to train our LSTM-based Q-Net on RefCOCO dataset using Multi-Instance Learning.

Subject to the dataset, it typically generates descriptions such as "the apple on the right of the cup" and "the apple in the back of the image".

D. R-Net for OBR and Grasp Detection

In INVIGORATE, a single network R-Net outputs both grasps and OBRs of detected objects in $\mathcal{B}$.

For OBR detection, we formulate it as a classification problem, which takes object pairs as inputs and classifies pair-wise OBRs. Following [43], there are three kinds of OBRs: "parent", "child", and "none". "Parent" relation between A and B means A should be grasped after B, and vice versa for the "child" relation. To classify OBR, we first represent each object by a pooled feature with a fixed size (7×7). Then we form all possible pair-wise permutations of object features. The feature of an object pair (i, j) includes the features of i, j, and the union bounding box. Finally, the pair-wise OBR for object (i, j) is directly classified based on the corresponding pair-wise feature and results in an OBR score f^r_{ij}. For grasp detection, since our task is goal-directed, the grasp should be object-specific. To do so, we detect grasps on each object instead of the input scene. Concretely, the grasp detector regresses grasps using the 7×7 pooled feature of each object with a few convolutional layers.

We follow Zhang et al. [44] to train our R-Net on VMRD [43], which contains around 4300 images and 100k grasps. In practice, we found that the grasp detector sometimes returns unstable grasps. Therefore, based on the detection result, we finetune the grasp pose through local search. In detail, we do a grid search by discretizing the area along five dimensions (x, y, z, w, θ) near the detected grasp, where (x, y, z) is the center of the grasp, w is the width of the gripper, and θ represents the rotation angle w.r.t. the approaching vector. We traverse all possible grasp poses to find the best one, whose closing area contains more points of the object.

V. INVIGORATE POMDP

A. State Space

To grasp the specified target in clutter, the state of INVIGORATE can be decomposed into two parts, the visual grounding state s^g and OBR state s^r, i.e., $s = s^g \cup s^r$. $s^g = \cup_{i=1}^{N_{obj}} s^g_i$ is an object-oriented state [7, 39], with each s^g_i indicating whether object i is a target. $s^r = \cup_{ij=1}^{N_{obj}} s^r_{ij}$ is a graph of all pair-wise OBRs, i.e., the correct grasping order of detected objects, with each s^r_{ij} denoting the true OBR between object i and j. Since the underlying true state is not available, we maintain a belief b_t at time step t over the state s_t, which represents a distribution over the state space. Similarly, $b_t = b^g_t \cup b^r_t$.

B. Action Space and Transition Model

To handle possible ambiguity, INVIGORATE allows active interaction with human to gather more information. Therefore, INVIGORATE has two types of actions: 1) asking a question; 2) grasping.

For each object i, action a^q_i means asking the human whether i is the target. The question follows a template "Do you mean S_i?", where S_i is a relational caption from Q-Net.

Grasp actions are defined by grasp macros. Each grasp macro is a sequence of grasps resulting in a terminal state. Assuming N_{obj} objects is detected, there will be $N_{obj} + 1$ grasp macros, including N_{obj} goal-directed choices and 1 clearing choice. Each goal-directed grasp macro a^g_i targets at object i. According to b^r, it sequentially removes exposed objects that most likely blocks object i, until it retrieves i. The clearing grasp macro a^g_{-1} is used to remove all detected objects when none of them is the target. According to b^r, it sequentially removes the most exposed objects. It is non-trivial to analytically find the most exposed object blocking one specified target since all relations are probabilistic. Therefore, we apply Monte Carlo method to estimate the probability of each object to be exposed and block the target, and then select the most probable one. As a result, the size of action space $|A| = 2N_{obj} + 1$. Note that in practice, for each step, we only execute the first grasp and then do re-planning, which helps to improve robustness.

TABLE I: Linguistic Observation Model $Z^l(ans_r|s_i^g, a^q)$

	$P(ans_r \in Res_p)$	$P(ans_r \in Res_n)$
$s_i^g = 1, a^q = a_i^q$	1	0
$s_i^g = 0, a^q = a_i^q$	0	1
$s_i^g = 1, a^q \neq a_i^q$	0	1
$s_i^g = 0, a^q \neq a_i^q$	ϵ	$1-\epsilon$

Since we assume that human does not change their mind about the target object, for any a_i^q, the transition model is:

$$T(s'|s, a_i^q) = \begin{cases} 1, s' = s \\ 0, s' \neq s \end{cases} \tag{4}$$

On the other hand, any grasp macro results in a terminal state. Therefore, we simply ignore the associated transition model.

C. Visual Observations

INVIGORATE takes the output of the G-Net and R-Net as the visual observations after each grasping action. At time step t, we denote the visual grounding observation from G-Net as o_t^g and OBR observation from R-Net as o_t^r, which are also object-oriented and accord with the state, i.e., $o_t^g = \cup_{i=1}^{N_{obj}} f_{i,t}^g$ and $o_t^r = \cup_{ij=1}^{N_{obj}} f_{ij,t}^r$. Accordingly, our visual observation model captures the distribution over o_t^g and o_t^r using visual grounding observation model Z^g and OBR observation model Z^r in a factorized way. Formally:

$$Z^g = P(f_i^g|s_i^g) \qquad Z^r = P(f_{ij}^r|s_{ij}^r) \tag{5}$$

where f_i^g is the output of G-Net and f_{ij}^r is the output of R-Net.

Unfortunately, Z^g and Z^r cannot be specified directly. Instead, we resort to data-driven methods. Specifically, we collect a dataset in clutter using G-Net, in which each data is represented by $\{f_i^g, \hat{g}_i\}$ where $\hat{g}_i$ is a binary label that indicates whether the object i is the referred target. Similarly, we collect a dataset in clutter using R-Net containing tuples $\{f_{ij}^r, \hat{r}_{ij}\}$, where $\hat{r}_{ij}$ is the ground truth OBR between object i and j. We then apply Gaussian kernel density estimation to learn an approximate model for Z^g and Z^r.

D. Linguistic Observations

After the robot asks a question, it receives an answer ans from the human, which is the linguistic observation in INVIGORATE. Each linguistic observation is an unrestricted natural language expression including a response phrase ans_r (e.g. "Yes" or "No") that may be followed by an additional description ans_d (e.g. "No, the left one"). To reduce the computation cost of POMDP planning, during the forward search, we use a simplified observation model that effectively ignores the additional description.

$$Z^l(o|s, a^q) = Z^l(ans|s, a^q) \approx Z^l(ans_r|s, a^q) \tag{6}$$

where ans_r belongs to either positive phrases Res_p ={"Yes", "Yeah", "Yep", "Sure"} or negative phrases Res_n ={"No", "Nope"}. During the belief update, we handle ans_r according to Eqn. 6, but merge ans_d into E which will be used for visual grounding in subsequent steps.

We assume that the human is trustful, and once the human confirms a target, the robot needs not consider other objects anymore. Under this assumption, the factorized observation model for asking questions is shown in Table I. ϵ is a small positive constant, and in practice, it is set to 0.01. Intuitively, a positive answer for object i makes object i the only target while a negative answer eliminates object i as a target but does not affect the belief of other objects.

E. Reward

We want the robot to grasp the correct target while asking a minimal number of questions. Thus, we impose a small penalty, i.e. a reward of -2, when it asks a question, and a large penalty when it fails the task (e.g. grasping the wrong object). When multiple objects seemingly satisfy the user expression, the robot cannot accurately differentiate between ambiguity and multi-target. Thus, to encourage disambiguation and avoid grasping wrong targets in such cases, we empirically engineer the reward for goal-directed grasp macros $R(s, a_i^g)$:

$$R(s, a_i^g) = \begin{cases} -10 + \dfrac{10}{\sum s^g}, & s_i^g = 1 \\ -10, & s_i^g = 0 \end{cases} \tag{7}$$

If there is only one object satisfying the human's instruction, grasping it will result in no penalty. Otherwise, to encourage disambiguation, the reward of grasping decreases as the number of targets increases. The robot receives a reward of -10 if it fails to grasp the target.

For the clearing grasp macro a_{-1}^g, the reward is:

$$R(s, a_{-1}^g) = \begin{cases} 0, & \forall s_i^g = 0 \\ -10, & otherwise \end{cases} \tag{8}$$

That is, the robot will not be penalized only if all detected objects are not the target. Otherwise, it receives a reward of -10 since it removes the target without passing it to the human.

F. Belief Tracking

Based on the imperfect observation o_t in each step, we update the belief b_t to obtain a more accurate estimate of the underlying true state. Since our state is object-oriented, it can naturally be factorized. We factor target belief b_t^g into belief over each object $b_t^g(s_i^g)$, and relationship belief b_t^r into belief over each pair of object $b_t^r(s_{ij}^r)$. This factorization allows us to perform belief tracking on each object and object pair separately.

The robot receives visual observations o_t^g and o_t^r after it performs a grasping and a linguistic observation ans after it asks a question. As mentioned, we factor o_t^g into target observation over each objects, $o_t^g = \cup_{i=1}^{N_{obj}} f_{i,t}^g$ and o_t^r into relationship observation over each pair of objects, $o_t^r = \cup_{ij=1}^{N_{obj}} f_{ij,t}^r$. We then track each factorized belief using Bayesian filter:

$$\begin{aligned} b_{t+1}^g(s_i^g) &\propto Z^g(f_i^g|s_i^g, a) b_t^g(s_i^g) \\ b_{t+1}^r(s_{ij}^r) &\propto Z^r(f_{ij}^r|s_{ij}^r, a) b_t^r(s_{ij}^r) \end{aligned} \tag{9}$$

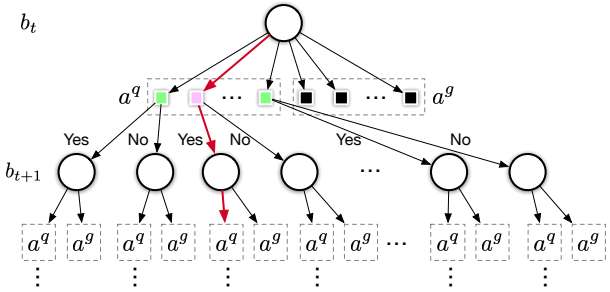


Fig. 3: An overview of policy tree search. Circles denote beliefs and squares denote possible actions. It searches all possible trajectories to find the optimal one (noted as the red path). Then the robot will execute the first action (noted as the pink square) with the highest expected cumulative reward.

where Z^g and Z^r are learned observation model for target and relation respectively. Since the human's answer does not affect OBR, ans is only used to update b^g

$$b_{t+1}(s_i^g) \propto Z^l(o_t|s_i^g, a)b_t(s_i^g) \tag{10}$$

G. POMDP Planning

Intuitively, our POMDP planner evaluates the trade-off between gathering more information and directly retrieving the target. This setting is similar to the Tiger problem[14]. Therefore, we utilize the policy tree search introduced by [14]) as the POMDP planner.

As shown in Fig. 3, our POMDP planner takes the current belief b_t as the input, and performs look-ahead search for the optimal action that maximizes the cumulative reward:

$$a^* = \arg\max_a E\left[\sum_t^\infty R(s_t, a_t)\right] \tag{11}$$

In our policy tree, each node b_t represents a belief. The parent node b_t and child node b_{t+1} are connected with an observation-action pair. Since a^g results in a terminal state, observation-action pairs are all based on a^q during planning. The maximum search depth is set to 3 to limit the number of questions the robot can ask. By traversing all possible trajectories, the planner returns an optimal trajectory that maximizes the expected cumulative reward (denoted as the red path in Fig. 3). The robot then executes the first action in the optimal trajectory (denoted as the pink square). If the action is a grasp macro, the robot grasps the first object in the grasp sequence. If the action is to ask a question, the robot simply says the caption generated by Q-Net. After the action is performed, the robot transits into the next step where it receives a new observation, updates its belief, and performs the search again.

VI. Experiments

Our experiments aim to investigate three questions:

Q1. Does INVIGORATE perform well overall in interactive visual grounding and grasping tasks?

Q2. What are the main contributors to INVIGORATE's performance?

Fig. 4: Test dataset, consisting of 10 scenes in toal. The test dataset will be available online.

Q3. Does INVIGORATE perform well in visual grounding in clutter, a key component of the system?

For Q1, we compare the performance of INVIGORATE with a pure deep-learning method without POMDP planning. Results show that INVIGORATE outperforms the baseline substantially and achieves an overall 83% success rate. For Q2, we conduct several ablation studies to evaluate various aspects of INVIGORATE. Results show that language interaction and observation histories boost overall success. For Q3, we compare INVIGORATE with ViLBERT [20], the current state-of-the-art visual grounding algorithm and show INVIGORATE consistently outperforms.

A. Implementation Details

We deploy INVIGORATE on a Fetch robot under the framework of Robot Operating System (ROS). All deep neural networks run on a single external NVIDIA Titan X GPU. We use Intel Realsense D435 camera to capture RGB images for visual inputs and point cloud for grasping and Google Cloud APIs to translate human verbal instructions into texts as well as synthesize speech for generated questions.

B. Experimental Settings

1) Dataset: To ensure fair comparisons between different variants of the system, we run all experiments on a test dataset consisting of 10 cluttered scenes shown in Fig. 4. We generate 100 test cases by recruiting 10 participants and asking them to select a target object and give a corresponding description for each scenario. For a comprehensive evaluation, we split test cases into two parts:

1) **Test 1:** Targets are selected before the participants see the clutter but are described after the clutter is shown. Participants therefore do not know where the target object will be located at the time of target selection.
2) **Test 2:** Targets are selected by participants after they see the clutter. Participants exactly know which object is challenging for the robot to grasp.

As we encourage participants to choose challenging targets, *Test 2* is generally harder than *Test 1*.

2) Baseline: INVIGORATE combines data-driven learning and model-based planning. Given the success of deep learning, the natural tendency is to use it directly. We build up the baseline method based purely on learned NN modules. The

TABLE II: Comparison of overall performance.

	Success Rate		
	Test 1	Test 2	Overall
MAttNet+VMRN	0.76	0.60	0.68
INVIGORATE	**0.86**	**0.80**	**0.83**

baseline, called *MAttNet+VMRN*, utilizes MAttNet [41] for visual grounding and VMRN [44] for OBR and grasp detection. It greedily follows the output of MAttNet to locate the most probable target while following the most likely OBRs to plan the grasp sequence.

3) Ablations: INVIGORATE POMDP consists mainly of three components: interaction, belief tracking, and policy tree search. In ablation studies, we aim to determine their respective contribution. Our ablation studies include:

- **w/o interaction:** the robot never asks questions but maintains the visual history to track the belief.
- **w/o history:** the robot remembers neither visual nor QA history. The POMDP planner works on the belief estimated only by the current observation.
- **w/o visual history:** the robot remembers QA history but not historical visual observations. The POMDP planner works on the belief estimated by the current visual observation and QA history.
- **w/o tree search:** it utilizes a heuristic method instead of tree search. Concretely, we apply two-class K-Means on the expected rewards of all grasp macros to check if multiple grasp macros have similar expected rewards. If that is the case, the robot will ask a question. Otherwise, it will execute the grasp macro with the max reward.

4) Procedures: We conduct two experiments on the real robot using the collected dataset.

The first experiment aims to compare the overall performance of INVIGORATE against the baseline and conduct ablation studies. Each variant of the system receives the same initial image and expression from the dataset as input. It then computes an action for the robot to execute. In each experimental scene, the experimenter is only required to describe one of the objects freely using its name, without any further detailed instructions to avoid possible interaction biases. During the process, if a question is asked, the experimenter will provide an answer (e.g., "yes/no") according to whether the object being asked is the true target. Though the experimenter is allowed to give additional descriptions when being asked, we found that in our experiments they did not tend to do so. Therefore, if there exist multiple ambiguous objects, the robot might ask several rounds of questions to disambiguate. Since grasp failures are not handled by any variant of the system and do not offer a meaningful comparison, if the robot fails to grasp an object, the experimenter would manually remove it. We record the success rate of each variant. A test case is regarded as a success only if the robot retrieves the true target. For ablation studies, we in addition record the *Normalized Cumulative Reward* and *Number of Questions* to give a comprehensive comparison.

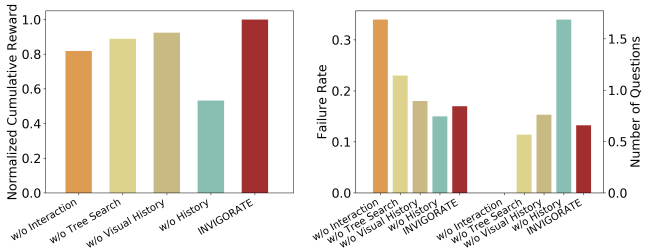


Fig. 5: Ablation studies. Total reward, failure rate, and the number of questions asked are averaged over 100 test cases from our dataset.

The second experiment aims to compare INVIGORATE's visual grounding performance against the SOTA method. We run INVIGORATE and ViLBERT side-by-side. No action is planned or executed by INVIGORATE. The experimenter instead manually removes blocking objects sequentially to retrieve the final target. In each step, we record the target probabilities estimated by both systems. Since ViLBERT is trained with cross-entropy loss, we directly apply exponential on its output to get the target probabilities.

C. Results

a) Does INVIGORATE perform well overall for interactive visual grounding and grasping?: Table II shows that INVIGORATE outperforms the baseline with an overall success rate of 83% ($p < 0.01$ in t-test). And in both Test 1 and Test 2, INVIGORATE achieves higher success rates. On average, INVIGORATE asks 0.65 questions and spends 0.5 additional grasp steps per scenario. This shows INVIGORATE achieves a higher success rate without a large number of redundant actions.

Furthermore, INVIGORATE's performance is more stable than the baseline. While the baseline *MAttNet+VMRN* achieves an average success rate of 76% on Test 1, its performance drops severely to 60% when applied on the harder Test 2. In contrast, the performance of INVIGORATE only drops by 6%. A closer look at the experiment result shows that the baseline's performance drop in Test 2 is mainly due to the increase in target detection failures. In fact, the baseline nearly fails in all cases where the target object is not visible or not detected at the beginning. In such cases, without a probabilistic estimate of the true underlying state, the baseline simply chooses the most likely target among visible objects and retrieves it for the user. On the other hand, INVIGORATE is able to reason that the target is not directly visible and would choose the clearing action to look for the target at the bottom.

b) What are the main contributors to INVIGORATE 's performance?: Fig. 5 shows the results of ablation studies. We found that interaction significantly improves the overall success rate ($p < 0.01$ in t-test). *w/o Interaction* suffers about 17% success rate loss, mainly from grasping the wrong target. The information gathered from interaction greatly helps to obtain an accurate belief and prevents the robot from target failures.

TABLE III: Comparison of visual grounding performance.

	Mean Accuracy			Mean L1 Loss		
	Test 1	Test 2	Overall	Test 1	Test 2	Overall
ViLBERT	0.855	0.817	0.831	0.084	0.108	0.099
INVIGORATE	**0.879**	**0.873**	**0.875**	**0.049**	**0.050**	**0.050**

In addition, we conclude that history reduces the number of questions. Compared to INVIGORATE, *w/o History* and *w/o Visual History* ask more questions (both with $p < 0.01$ in t-test). In *w/o Visual History*, the robot uses only the current observation to estimate the belief over the state which is less accurate. Therefore, it has to ask more questions to refine its belief. Besides, *w/o History* asks the most number of questions as the robot does not remember previous answers from the human. Though it achieves a high success rate, the system's behavior is annoying, resulting in a low cumulative reward.

For comparison between INVIGORATE and *w/o tree search*, we found that only 100 experiments do not show statistically significant difference due to high variance. Therefore, we conducted 100 more experiments with the same procedure. Compared to *w/o tree search*, INVIGORATE asks slightly more questions (with $p < 0.01$) but leads to fewer failures (with $p < 0.05$). Noteworthily, INVIGORATE shows a higher cumulative reward than *w/o tree search*. In our experiments, we also observed that the behavior of *w/o tree search* is more aggressive, which means that it tends to be confident about itself judgment without asking questions. The intrinsic reason should lie in the two-class K-Means policy, which is quite close to the one-step planning and might be myopic. Unfortunately, 200 experiments still fail to show some significant difference. We will conduct more experiments in the future to explore the effects of the planner.

c) Does INVIGORATE perform well in visual grounding in clutter, which is a key component of the system?: We compare target probability L1 loss and mean average accuracy between INVIGORATE and ViLBERT. Results are shown in Table III. In order to calculate the accuracy of both systems, we treat visual grounding as a binary classification problem. The object is regarded as the target once its target probability is higher than a certain threshold. The mean accuracy reported is computed by averaging accuracies computed on 9 different thresholds (0.1 to 0.9 with interval 0.1).

Our results show that the visual grounding performance of INVIGORATE consistently outperforms ViLBERT in clutter. Despite its SOTA performance on visual grounding in uncluttered scenes, ViLBERT suffers from visual occlusions and language ambiguities in our test dataset and becomes inaccurate and unstable. On the other hand, INVIGORATE treats neural network's outputs as noisy observations. It learns an observation model and uses the Bayesian filter to constantly update its belief of the state across multiple steps. Our results confirm that such a principled approach for visual grounding exhibits more robust performance in clutter.

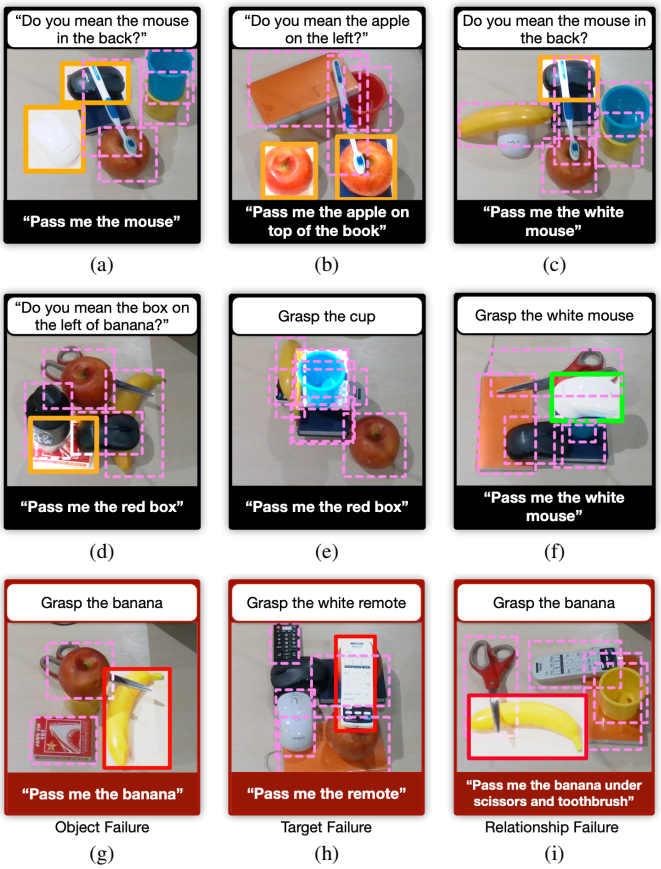


Fig. 6: Qualitative results. (*a-f*): some selected successful cases of INVIGORATE with different kinds of uncertainty and ambiguity. (*g-i*): some selected failures. The sentence on top is the action of the robot and the sentence at the bottom is the human's instruction. Best viewed in color.

D. Examples

Fig. 6 shows examples of INVIGORATE. In Fig. 6(a), the user gives an ambiguous expression. As there are two mouses in the scene, INVIGORATE asks a question to disambiguate. Fig. 6(b-d) show some complex scenarios where the target cannot be easily identified. In Fig. 6(b), the relational clue object "book", which is the blue book lying under the right apple, is not detected. In Fig. 6(c), the black mouse is covered by a white toothbrush while the target white mouse is not detected. In Fig. 6(d), the red box is detected, but its visual features are not strong since it is occluded by the bottle and mouse on top. To tackle these difficulties, INVIGORATE asks questions to query for more information. Fig. 6(e) shows a case where the target is not detected, INVIGORATE therefore removes the cup on top to look for the target at the bottom. Fig. 6(f) shows a simple case where the target is not occluded or obstructed, INVIGORATE therefore directly grasps the target without asking any question.

Fig. 6(g) shows that the object detector fails to detect the scissors that are blocking the true target banana. INVIGORATE directly grasps the banana and thus violates the true OBR. In

Fig. 6(h), the user gives an ambiguous expression "remote" while the true target is the black remote in the back. Due to occlusion, the visual grounding module places a very low score on the true target, INVIGORATE therefore grasps the white remote controller directly, believing that it is the only target in the scene. Fig. 6(i) shows a case of relationship detection failure. INVIGORATE directly grasps the banana although it is blocked by the scissor, violating the true OBR.

VII. Conclusion

INVIGORATE enables the robot to interact with human through the natural language and perform goal-directed object grasping in clutter. It takes advantage of a POMDP model that integrates the learned neural network models for visual perception and language interaction. By integrating data-driven deep learning and model-based POMDP planning, INVIGORATE successfully tackles complex visual inputs and language interactions and achieves strong overall performance, despite the inevitable errors of the learned neural network models in perceptual and language processing.

Many exciting challenges lie ahead. First, the neural network models for visual perception and language interaction are trained independently. It would be interesting to embed the INVIGORATE POMDP into the network and apply end-to-end training. Previous work demonstrates that such an end-to-end architecture may bring considerable performance improvement [15, 36]. Second, INVIGORATE cannot fully address the systematic errors from the deep neural network models for visual perception. Calibration of the learned models [38] can potentially improve the uncertainty estimate for INVIGORATE in the future. Finally, INVIGORATE assumes that the human would get annoyed if the robot asks more than 3 questions. It is reasonable simplification, but neglects the nuance of human-robot interaction. For humans, seamless interaction depends on conventions shaped by shared experiences and culture. For robots to achieve the same, it may be necessary to model the human cognitive state and adapt information exchange accordingly [10], an interesting yet challenging direction for future work.

Acknowledgments

This work has benefitted greatly from discussions with Panpan Cai. It is supported in part by NSFC under grant No.91748208, No.62088102, No.61973246, Shaanxi Project under grant No.2018ZDCXLGY0607, the program of the Ministry of Education of China, Singapore A*STAR under the National Robotics Program (Grant No. 192 25 00054), and Singapore National Research Foundation under its AI Singapore Program (AISG Award No. AISG2-RP-2020-016).

References

[1] Jeannette Bohg, Antonio Morales, Tamim Asfour, and Danica Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on Robotics*, 30(2):289–309, 2013.

[2] Panpan Cai, Yuanfu Luo, Aseem Saxena, David Hsu, and Wee Sun Lee. Lets-drive: Driving in a crowd by learning from tree search. *arXiv preprint arXiv:1905.12197*, 2019.

[3] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018.

[4] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer, 2020.

[5] Yiye Chen, Ruinian Xu, Yunzhi Lin, and Patricio A Vela. A joint network for grasp detection conditioned on natural language commands. *arXiv preprint arXiv:2104.00492*, 2021.

[6] Michael Danielczuk, Andrey Kurenkov, Ashwin Balakrishna, Matthew Matl, David Wang, Roberto Martín-Martín, Animesh Garg, Silvio Savarese, and Ken Goldberg. Mechanical search: Multi-step retrieval of a target object occluded by clutter. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 1614–1621. IEEE, 2019.

[7] Carlos Diuk, Andre Cohen, and Michael L Littman. An object-oriented representation for efficient reinforcement learning. In *Proceedings of the 25th international conference on Machine learning*, pages 240–247, 2008.

[8] Kuan Fang, Yunfei Bai, Stefan Hinterstoisser, Silvio Savarese, and Mrinal Kalakrishnan. Multi-task domain adaptation for deep learning of instance grasping from simulation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3516–3523. IEEE, 2018.

[9] Neha P Garg, David Hsu, and Wee Sun Lee. Learning to grasp under uncertainty using pomdps. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 2751–2757. IEEE, 2019.

[10] Michael A Goodrich, Alan C Schultz, et al. Human–robot interaction: A survey. *Foundations and Trends® in Human–Computer Interaction*, 1(3):203–275, 2008.

[11] 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 *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3774–3781. IEEE, 2018.

[12] Sachithra Hemachandra and Matthew R Walter. Information-theoretic dialog to improve spatial-semantic representations. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5115–5121. IEEE, 2015.

[13] Eric Jang, Sudheendra Vijayanarasimhan, Peter Pastor, Julian Ibarz, and Sergey Levine. End-to-end learning of semantic grasping. In *Conference on Robot Learning*, pages 119–132. PMLR, 2017.

[14] Leslie Pack Kaelbling, Michael L Littman, and Anthony R Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

[15] Peter Karkus, David Hsu, and Wee Sun Lee. Qmdp-net: Deep learning for planning under partial observability. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[16] Peter Karkus, Xiao Ma, David Hsu, Leslie Pack Kaelbling, Wee Sun Lee, and Tomás Lozano-Pérez. Differentiable algorithm networks for composable robot learning. In *Robotics: Science and Systems*, 2019.

[17] Jayant Krishnamurthy and Thomas Kollar. Jointly learning to parse and perceive: Connecting natural language to the physical world. *Transactions of the Association for Computational Linguistics*, 1:193–206, 2013.

[18] Andrey Kurenkov, Joseph Taglic, Rohun Kulkarni, Marcus Dominguez-Kuhne, Animesh Garg, Roberto Martín-Martín, and Silvio Savarese. Visuomotor mechanical search: Learning to retrieve target objects in clutter. *arXiv preprint arXiv:2008.06073*, 2020.

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[20] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[21] Jeffrey Mahler and Ken Goldberg. Learning deep policies for robot bin picking by simulating robust grasping sequences. In *Conference on robot learning*, pages 515–524. PMLR, 2017.

[22] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017.

[23] Oier Mees and Wolfram Burgard. Composing pick-and-place tasks by grounding language. In *International Symposium on Experimental Robotics (ISER)*, 2021.

[24] Adithyavairavan Murali, Arsalan Mousavian, Clemens Eppner, Chris Paxton, and Dieter Fox. 6-dof grasping for target-driven object manipulation in clutter. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6232–6238. IEEE, 2020.

[25] Varun K Nagaraja, Vlad I Morariu, and Larry S Davis. Modeling context between objects for referring expression understanding. In *European Conference on Computer Vision*, pages 792–807. Springer, 2016.

[26] Dejan Pangercic, Benjamin Pitzer, Moritz Tenorth, and Michael Beetz. Semantic object maps for robotic housework-representation, acquisition and use. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4644–4651. IEEE, 2012.

[27] Claudia Pateras, Gregory Dudek, and Renato De Mori. Understanding referring expressions in a person-machine spoken dialogue. In *1995 International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 197–200. IEEE, 1995.

[28] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 2020.

[29] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1316–1322. IEEE, 2015.

[30] Stephanie Rosenthal, Joydeep Biswas, and Manuela M Veloso. An effective personal mobile robot agent through symbiotic human-robot interaction. In *AAMAS*, volume 10, pages 915–922, 2010.

[31] Mohit Shridhar, Dixant Mittal, and David Hsu. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2-3):217–232, 2020.

[32] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529 (7587):484–489, 2016.

[33] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[34] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

[35] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[36] Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel. Value iteration networks. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2154–2162, 2016.

[37] Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. Asking for help using inverse semantics. 2014.

[38] Juozas Vaicenavicius, David Widmann, Carl Andersson, Fredrik Lindsten, Jacob Roll, and Thomas Schön. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3459–3467. PMLR, 2019.

[39] Arthur Wandzel, Yoonseon Oh, Michael Fishman, Nishanth Kumar, Lawson LS Wong, and Stefanie Tellex. Multi-object search using object-oriented pomdps. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7194–7200. IEEE, 2019.

[40] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring

expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016.

[41] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1307–1315, 2018.

[42] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. *The International Journal of Robotics Research*, page 0278364919868017, 2019.

[43] Hanbo Zhang, Xuguang Lan, Xinwen Zhou, Zhiqiang Tian, Yang Zhang, and Nanning Zheng. Visual manipulation relationship network for autonomous robotics. In *2018 IEEE-RAS 18th International Conference on Humanoid Robots (Humanoids)*, pages 118–125. IEEE, 2018.

[44] Hanbo Zhang, Xuguang Lan, Site Bai, Lipeng Wan, Chenjie Yang, and Nanning Zheng. A multi-task convolutional neural network for autonomous robotic grasping in object stacking scenes. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6435–6442. IEEE, 2019.