Scene-level Programming by Demonstration

Zhen Zeng Zheming Zhou Zhiqiang Sui Odest Chadwicke Jenkins

***Abstract*— Scene-level Programming by Demonstration (PbD) is faced with an important challenge - perceptual uncertainty. Addressing this problem, we present a scene-level PbD paradigm that programs robots to perform goal-directed manipulation in unstructured environments with grounded perception. Scene estimation is enabled by our discriminatively-informed generative scene estimation method (*DIGEST*). Given scene observations, *DIGEST* utilizes candidates from discriminative object detectors to generate and evaluate hypothesized scenes of object poses. Scene graphs are generated from the estimated object poses, which in turn is used in the PbD system for high-level task planning. We demonstrate that *DIGEST* performs better than existing method and is robust to false positive detections. Building a PbD system on *DIGEST*, we show experiments of programming a Fetch robot to set up a tray for delivery with various objects through demonstration of goal scenes.**

I. Introduction

We aim to provide a system for a user to effectively program a robot to complete manipulation tasks. We focus on understanding the goal of manipulation tasks. The user only needs to demonstrate once to the robot a desired scene state. As the scene changes, the robot should be able to manipulate the objects in the scene to restore the desired scene state. In tasks such as organizing objects in a household environment, users can program the robot to keep the living room at a tidy state.

The most related field to this objective is Programming by Demonstration (PbD), and similar notions of learning by imitation, which provide a natural way for naive users to convey skills and tasks to robots. To effectively program a robot in a general manner, the robot should be able to understand the goal of the task, i.e., *what to imitate*. With an understood goal, a robot can reason and adapt its actions to reach this goal in various scenarios and changing environments.

While there has been considerable advances in robot learning, scene perception remains a challenge in general goal-directed manipulations. Traditional PbD works using kinesthetic teaching [11], [19], [4], [1], [21] focused on robot learning of a configuration-space control policy for a particular task or skill. Scene perception in workspace is important in that a robot should be able to adapt a learned skill to current scene. The perceptual uncertainty of scene perception has limited robot task-level reasoning to limited domains due to issues of computational tractability.

In this paper, we describe our approach to goal-directed PbD over scenes. Our work is motivated by a service robot scenario, as illustrated in Figure 1. We argue that scene perception is a critical missing piece for general goal-directed PbD. With proposed scene estimation method *DIGEST*, we demonstrate how goal-directed PbD can be performed at the level of scenes in different scenarios of setting up a tray. A critical distinction of our work is that we are estimating goals as workspace scene graphs, represented by object relations and poses, instead of the configuration space of the robot. Thus we focus on estimating the goal as the desired scene state, s.t. any capable robot can perform the task.

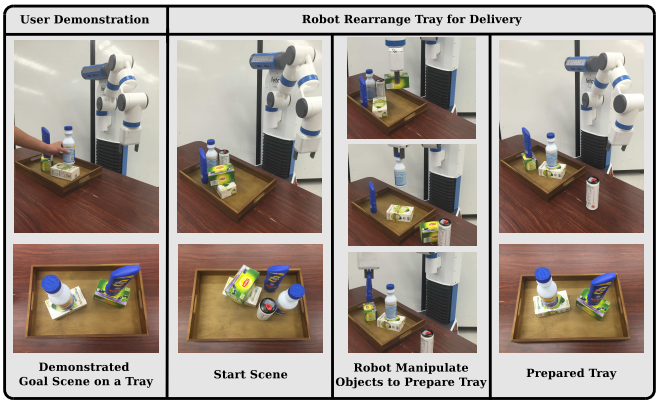

Fig. 1: Our work is motivated by a service robot scenario: the user needs some food items placed with certain configurations on a tray for delivery to his office daily. With our goal-directed PbD paradigm, the user can program a robot to automatically set up the tray for delivery. It is critical for the robot to understand the current and goal scene, so that it can plan actions to set up the tray.

Our contribution is two-fold:

- We propose a discriminatively-informed generative method (*DIGEST*) to estimate object poses in cluttered scenes for goal-directed manipulation
- We present a goal-directed PbD system at the level of scene graphs generated from the estimated object poses, where the robot is an operator in the scene

Our experiments involve 1) evaluating the goal-directed PbD system in tray setup tasks with a Fetch Robotics robot, 2) evaluating *DIGEST* in cluttered scene. And *DIGEST* outperforms the D2P [20] on the household occlusion dataset [2].

II. Related Work

A. Scene Perception for Manipulation

Feature-based methods and analysis-by-synthesis are two traditional approaches for model-based object pose estimation. Feature-based methods try to match hand designed or learned features between models and observations. Spin

Z. Zeng, Z. Zhou, Z. Sui and O.C. Jenkins are with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI, USA, 48109-2121

images [12] and FPFH [22] are two local feature based methods, where the features are extracted and matched between models and observation for pose estimation. OUR-CVFH [3] and VFH [23] are examples of global feature based methods that are more robust to object occlusions. Different features are learned with respect to the object pose. However, the performance of feature-based methods will degrade when the environment becomes more cluttered and key features are occluded.

Analysis-by-synthesis, also referred as rendering and verification method, is a generative approach that renders multiple hypotheses, and finds the hypothesis that best explains the observation. The early works by Stevens et al. [24] proposed an iterative hypothesize, render and verify process to estimate 3D pose of an object. Our previous work APF [26] [27] use MCMC to search for the scene hypothesis that best explains the observation. Similarly, D2P [20] uses A* as the search method to estimate 3 Dof object poses. Both APF and D2P assume object identities are known *a priori*. However, there does not exist an ideal recognition system that can identify all the objects correctly in a cluttered environment yet. *DIGEST* aims to avoid such a strong assumption.

B. Goal-Directed PbD

There have been some works in PbD focus on learning low level control. Grollman and Jenkins [9] simultaneously segment a demonstrated task and learn the policy for each individual subtask. Chernova and Veloso [6] present an interactive algorithm for policy learning from demonstration. Ijspeert et al. [11] and Nakanish et al. [19] teach robot biped locomotion patterns through demonstrations. In an incremental learning system by Calinon and Billard [4], a HOAP-3 humanoid learns different gestures via generalizing over demonstrated joint angle trajectories. Akgun et al. [1] learns generalized trajectory by identifying a set of keyframes. Niekum et al. [21] presents an approach to discover finite-state representations of multi-step tasks. For domains such as locomotion and gesture control, it is sufficient to generalize over the robot configuration space and proprioceptive contacts. While in the domain of robot manipulations, in contrast, it is important to generalize over new scenes in the workspace.

Goal-directed PbD is complementary to aforementioned works, in addition to learning generalized trajectories over control from demonstration, perceiving and inferring goal from demonstrations are also important. In goal-directed PbD, understanding the goal of a demonstrated scene is faced with challenges such as perceptual uncertainty, especially in cluttered scenes involving occlusions and object interactions. Mohan et al. [18] and Kirk et al. [15] present a cognitive system that learns task formulations, and perform goal-directed control with grounded perception. However, they assume that all the objects are solid-colored, so that the visual system can infer the scene state reliably. Similarly, Chao et al. [5] limit the demonstrations to planar objects with distinguished colors. Unlike these works, our perception system does not rely on assumptions of object colors or planar objects.

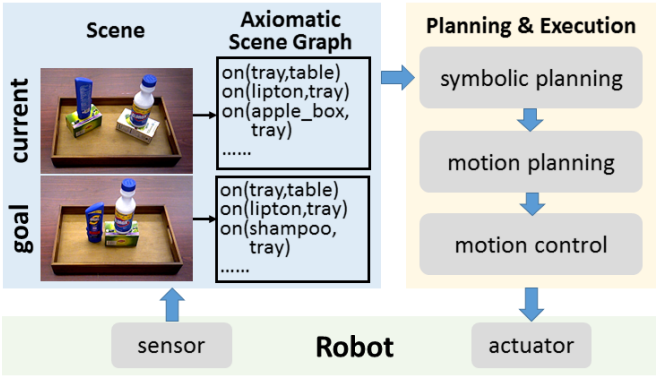

Fig. 2: Our programming by demonstration pipeline. Current and goal scene states are inferred from the sensor observations, and represented by axiomatic scene graph. Given axiomatic scene graphs of current and goal scenes, symbolic planner plans a sequence of pick and place actions to manipulate the objects to achieve the goal scene. Motion planning module is in charge of generating collision-free arm trajectories for each pick and place action.

To deal with scene perception uncertainty, the robot can make a decision and act based on a belief space of the scene, which can be formulated by POMDP [13], but the problem becomes intractable as the state space increases. Kaelbling and Lozano-Perez [14] build their work on blending probabilistic and task inference in a logical representation of belief space. We argue that if viable scene perception can be realized, then decision making and probabilistic inference of the scene can be decoupled as in the domain of autonomous navigation. In our work, we maintain a distribution over the scene state, i.e., the inter-object relations as well as the object poses. We use the maximal likely scene state as the scene estimate, and perform goal-directed reasoning based on the scene estimate of the demonstrated and start scene.

III. Problem Statement

Assume a robot R as a physically capable agent in the scene, a set of geometries $\mathbf{V} = \{v_1, \cdots, v_k\}$ of known objects in the scene, and a set of manipulation actions $\mathbf{A} = \{a_1, \cdots, a_n\}$ with known pre- and post- conditions. Our objective is to infer goal scene state s_G and current scene state s_t at time t from the observations of user's desired scene o_G and current scene o_t, respectively, and plan a sequence of actions $\{a_i, \cdots, a_j\}$ for R to carry out to reorganize the scene such that the scene state transits from s_t to s_G.

We use a list of axiomatic assertions to describe a scene. The scene state at time t is expressed as a scene graph $s_t = \{h_t^i(\mathbf{x})\}_{i=0}^{M}$, where $h_t^i \in \{\text{exist}, \text{clear}, \text{on}, \text{in}\}$ is an axiomatic assertion parametrized by $\mathbf{x_t} = \{w_t^j(q_t^j, v_t^j)\}_{j=0}^{N_c}$, with $w_t^j(q_t^j, v_t^j)$ representing that at time t object w_t^j has pose q_t^j and geometry v^j, N_c being the number of objects, and M is the total number of axiomatic assertions. In our work, the assertions are limited to to spatial relations that can be tested geometrically or physically.

The 6 Dof pose $q_t^j = [x_t^j, y_t^j, z_t^j, \phi_t^j, \psi_t^j, \theta_t^j]$ of each object is estimated, which consists of a 3D position (x_t^j, y_t^j, z_t^j) and

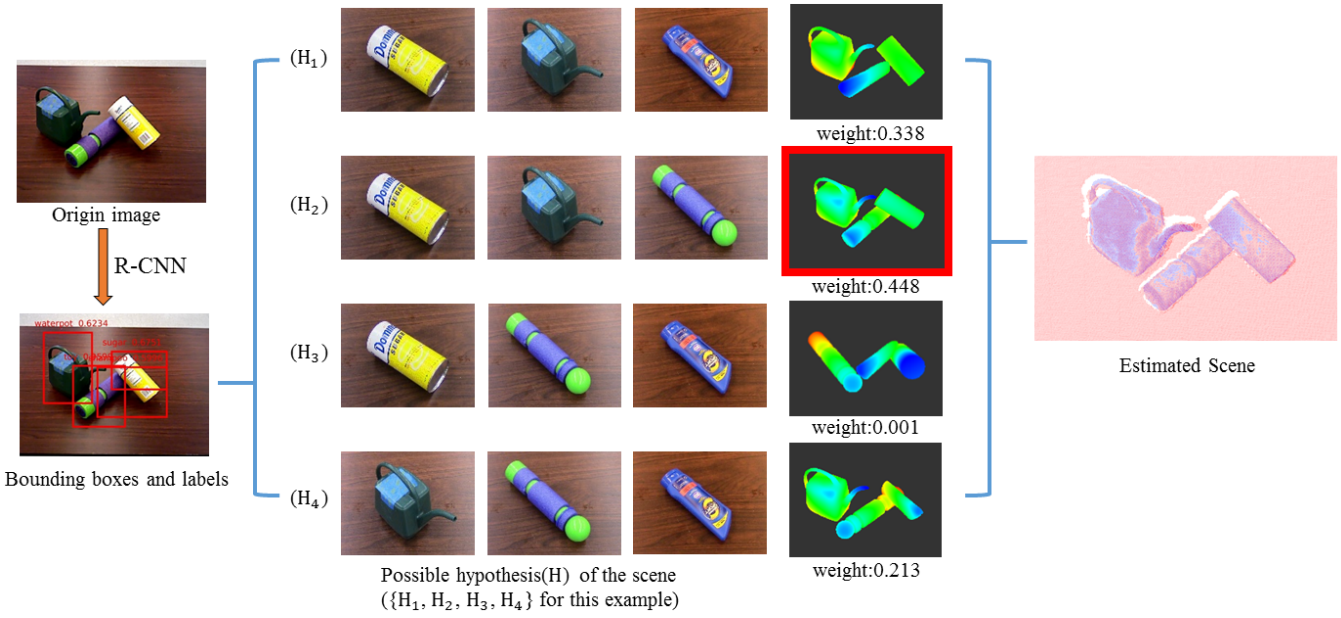

Fig. 3: *DIGEST* pipeline. First, the test RGB image is passed through an R-CNN object detector trained on our dataset and a set of bounding boxes with object labels is generated. Then possible combinations of hypothesized scenes are enumerated and evaluated against the pre-processed observed point cloud. The particle filtering pose estimation method evaluated each scene and produced weight. At last, the result comes from the scene with the max weight.

orientation $(\phi_t^j, \psi_t^j, \theta_t^j)$. The number of objects N_c in the scene is known in advance, but the object categories are not known. Scene graph can be inferred from estimated object poses, as explained later in section IV-B.

IV. Methods

Perception is the core of our approach to PbD. To address the perceptual uncertainty, we aim to decouple probabilistic scene state estimation and action planning s.t. the robot takes actions based on the current estimate of the scene. Our pipeline consists of perception, plan and execution stages, as shown in Figure 2. Given observations of a cluttered scene, the generative process of scene estimation is informed by discriminative object detector. A scene graph that represents the inter-object relations are geometrically inferred from the estimated scene, which is used for high-level goal-directed reasoning in the PbD system.

A. DIGEST: Cluttered Scene Estimation

Given observation o_t as the depth image of a cluttered scene at time t, the objective is to estimate the object poses $q_t^j, j = 1, \cdots, N_c$. We utilize the discriminative power of a pre-trained object detector and recognizer to first obtain a set of bounding boxes and object labels. These bounding boxes are used to guide the generative process of sampling scene hypothesis. The overview of the cluttered scene estimation is as illustrated in Figure 3.

1) Object Detection and Hypothesized Scene Generation: Any object detection method that can produce candidates including bounding boxes and object labels from the observation suits our method. Assume there are m bounding boxes detected from the object detector. Let B_i $(0 \leq i \leq m)$ denote each bounding box, L_j $(1 \leq j \leq k$, k is the number of object class) denote each object label in the bounding box B_i and $v(L_j|B_i)$ is the confidence of object label L_j in the bounding box B_i. For each B_i, we generate a candidate C_i,

$$C_i = \{\arg\max_{L_j} v(L_j|B_i), B_i\} \tag{1}$$

which is a set including the object label with highest confidence score and the current bounding box.

For m generated candidates, the number of hypothesized scene h is defined as:

$$h = \begin{cases} {}^mC_{N_c}, & \text{if } N_c \leq m \\ 1, & \text{otherwise} \end{cases} \tag{2}$$

So if there are more or equal candidates than objects in the scene, each hypothesized scene H_i contains a combination of N_c candidates selected from m candidates. If there are fewer candidates than N_c, just one hypothesized scene with m candidates will be generated.

2) Particle Filtering for Pose Estimation: Each hypothesized scene H_i is modeled as a random state variable x_t, comprised of a set of real-valued object poses. We model the inference of the state x_t from a history of robot observations $z_{1:t}$ as a sequential Bayesian filter,

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \int_{x_t} p(x_t|x_{t-1}, u_{t-1}) p(x_{t-1}|z_{1:t-1}) \mathrm{d}x_{t-1} \tag{3}$$

The posterior belief is formed at time t by updating a prior belief from time $t-1$ with a dynamic resampling and likelihood evaluation step. The sequential Bayesian filter in Eq. 3 is often approximated by a collection of N weighted

particles, $\{x_t^{(j)}, w_t^{(j)}\}_{j=1}^N$, with weight $w_t^{(j)}$ for particle $x_t^{(j)}$, expressed as:

$$p(x_t|z_{1:t}) \propto p(z_t|x_t) \sum_j w_{t-1}^{(j)} p(x_t|x_{t-1}^{(j)}, u_{t-1}) \tag{4}$$

as described by [7]. Inference is then performed by computing the likelihood of each hypothesis, normalizing the weights to one, and drawing N scene hypotheses by importance sampling iteratively.

$$x_t^{(j)} \sim \sum_j w_{t-1}^{(i)} p(x_t|x_{t-1}^{(i)}, u_{t-1}) \tag{5}$$

$$w_t^{*(j)} = p(z_t|x_t^{(j)}) \tag{6}$$

$$w_t^{(j)} = \frac{w_t^{*(j)}}{\sum_k w_t^{*(k)}} \tag{7}$$

Through the z-buffer of a 3D graphics engine, each particle $x_t^{(j)}$ is rendered into a depth image, represented as $\hat{z}_t^{(j)}$. We project back the rendered depth image into a point cloud $\hat{r}_t^{(j)}$ in camera frame given intrinsic parameters. Then, we compute the likelihood for each particle with the observation z_t:

$$p(z_t|x_t^j) = e^{-\lambda_r \cdot \mathrm{d}(z, \hat{r}_t^{(j)})} \tag{8}$$

where λ_r is a constant scaling factor and $\mathrm{d}(R,O)$ is the Euclidean distance between the rendered point cloud and the observation point cloud:

$$\mathrm{d}(R,O) = \sum_{(a,b)\in I} \|(R(a,b) - O(a,b))\| \tag{9}$$

where a and b are indices in rendered point cloud R and observed point cloud O, and $\|\cdot\|$ is the norm of the norm of a 3D vector.

We apply Iterated Likelihood Weighting [17] to bootstrap the scene estimation process, where the state transition in the action model is represented by a zero-mean Gaussian noise. Once the bootstrap filter converges, the scene $\hat{H}_t$ from the most likely particle $\hat{x}_t$ is taken as the scene estimate:

$$\hat{H}_t = \arg\max_{x_t^{(j)}} p(x_t^{(j)}|z_{1:t}) \tag{10}$$

and the unnormalized $w_t^{*(j)}$ weight for the mostly likely scene is taken as the weight for ranking the all the scenes.

3) Final Scene Ranking: After evaluating all hypothesized scenes, we rank them based on the weight computed from our pose estimation method. The scene hypothesis with highest weight is taken as the maximal likely scene estimate.

B. Scene Graph Structure

The objects pose estimation of a cluttered scene can be turned into a scene graph composed by a set of axiomatic assertions that describes the scene. We use $exist(w^j(q^j, v^j))$ to assert that object w^j exists in the scene with pose q^j and geometry v^j. $clear(w^i)$ for assertion that the top of object w^i is clear and no other objects are stacked on it. $on(w^i, w^j)$ for assertion that object w^i is stacked on object w^j. $in(w^i, w^j)$ for assertion that object w^i is in object w^j. An example of a scene graph is given in Figure 4.

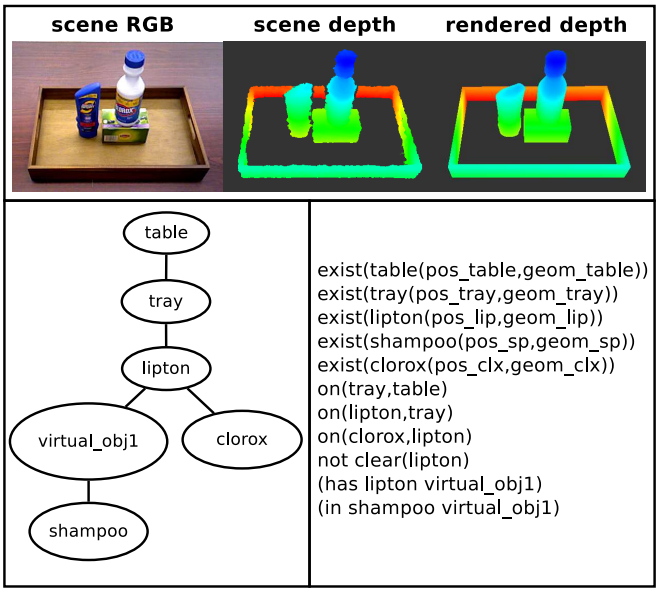

Fig. 4: An example of scene estimation for: *lipton* and *shampoo* are placed next to each other, with *clorox* stacked on *lipton*. **scene RGB**: RGB image; **scene depth**: the depth image of the scene with table cropped out; **rendered depth**: the depth image rendered by the axiomatic scene estimator. ***lower left***: estimated scene graph tree structure with nodes representing objects, and edges representing inter-object relations; ***lower right***: The corresponding scene graph composed by a list of parametrized axiomatic assertions.

We extend the tree-structured scene graph representation used in APF [26] by introducing a scene element (*virtual object*). With a *virtual object* $w^\gamma(q^\gamma, v^\gamma)$, a scene graph can express proximity relations between objects. To assert the proximity relations between two objects w^i, w^j, we add a *virtual object* $w^\gamma(q^\gamma, v^\gamma)$ into the scene graph, with v^γ being an arbitrary shape, and q^γ expressed in the frame of object w^i. Then, the spatial relation between w^i, w^j can be encoded by $\{has(w^i, w^\gamma),\ in(w^\gamma, w^j)\}$, where $has(w^i, w^\gamma)$ asserts that w^i has an *virtual object* w^γ attached to its frame. When the parent object w^i is in a new location, the robot can adapt to the new scenario by placing the child object w^j within the region of w^γ attached to the frame of w^i.

We assume that one object is supported by a single object, versus that one object is supported by multiple objects at the same time. In each 3D object model, the z-axis is the gravitational axis when the object stands upright, the x-axis is the gravitational axis when the object lies on a ground, and y-axis is the cross product of the other two axes. The dimensions $\{h_x, h_y, h_z\}$ of the 3D box that encloses each object model are provided. In order to decide whether object w^i is being supported by another object, two heuristics are tested: 1) if one of the object axes (e.g. x-axis) is aligned with the gravitational axis, then the height h_i of the 3D volume occupied by the object equals to the corresponding dimension (e.g. h_x) of the provided 3D enclosing box. A simple rule $z^i - h^{table} > 0.5h^i$ implies whether object w_i is being supported by another object; 2) if none of the object

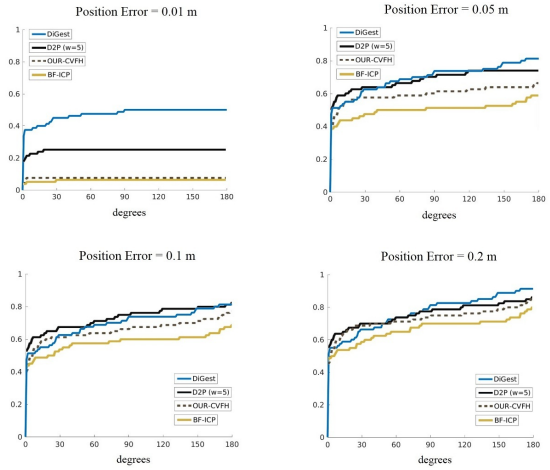

Fig. 5: *DIGEST* evaluation result with baseline method D2P along with OUR-CVFH and Brute force ICP. Note that we did not run the baseline methods and the plots of these three methods are from [20]. The y-axis is the estimation accuracy. The x-axis is the allowed orientation error for an estimation to be considered as accurate.

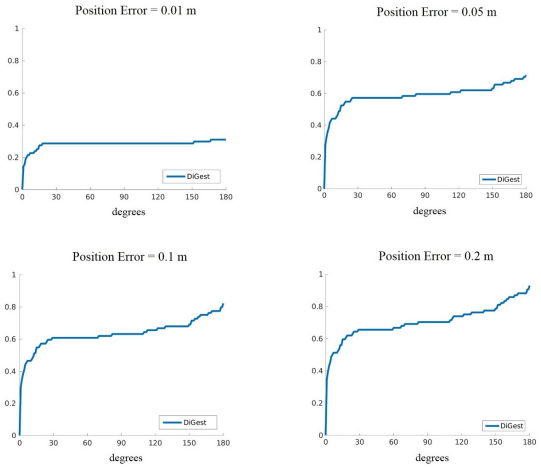

Fig. 6: *DIGEST* Scene estimation for our dataset. The x-axis is the allowed orientation error for an estimation to be considered as accurate.

axes are aligned with the gravitational axis, then w_i is being supported by another object.

The set of objects that are being supported by other objects are denoted as O_s, and the rest of the objects are being denoted as O_r. For object $w^i \in O_s$, a heuristic measure is used to determine which object $w^j \in O_r$ is supporting w^j,

$$\arg\max_j m(r_b(w^i), r_t(w^j))$$

where $m(r_1, r_2)$ measures the overlapping area of two regions r_1, r_2. $r_t(w), r_b(w)$ represent the projected region on the horizontal plane of the top and bottom surface of object w, respectively. With identified supporting relation between a pair of objects w^i, w^j, the corresponding axiomatic assertion is expressed as either $on(w^i, w^j)$ or $in(w^i, w^j)$, depending on the geometry type of the supporting object w^j being convex or concave.

V. Implementation

A. RCNN object detector

We employ RCNN [8] as our discriminative object detector for *DIGEST*. RCNN first generates object proposals given an image and then classifies the proposals using a deep convolutional neural metwork. For the sake of efficiency and performance, we replaced the original selective search [28] with EdgeBox [29] for object proposal generation.

B. Particle Filtering and parallelization

The implementation of particle filtering pose estimation method consists of three modules: *measurement*, *resampling* and *diffusion*. For each object in each particle $x_t^{(j)}$, it is initialized by the candidate C_i in the hypothesized scene, the object label l_i determines which object to sample, and the initial pose is uniformly sampled from the bounding box B_i. Then these particles are given as input to a parallel graphics engine which generates scene estimation depth images rapidly. The measurement module takes a preprocessed point cloud as the observation and compares it with the point cloud back projected from the rendered depth image. Resampling and diffusion will give the new posterior distribution and add noise to the current state. We set the fixed particle filter iteration to 400 and employ 625 particles for all the experiments.

In the particle filtering process, the pose of each object is estimated sequentially. For example, if there are four hypothesized objects and 400 iterations for particle filtering, the pose of the object with the largest recognition confidence is estimated in the first 100 iterations. Then the pose of the object with the second largest recognition confidence is estimated in the next 100 iterations, with the first object fixed at the converged estimated pose. We iterate the same estimation process for the rest objects.

C. Planning

Given a demonstrated goal scene, the robot estimates the goal scene state, i.e., the object poses and desired inter-object relations, and stores the goal scene state by PDDL [16]. Similarly the robot estimates and stores the start scene state by PDDL. Note, that each pose label in PDDL is associated with an object pose in the robot base frame, s.t. the robot knows where to pick or place objects during action execution. For example, as shown in the lower right part of Figure 4, *lipton* is at *pos_lipton*, where *pos_lipton* is a pose label for lipton that is associated with the estimated *lipton* pose in the robot base frame. With sets of PDDL that describe the start and goal scene, the robot uses symbolic planner to plan a series of actions to reorganize the start scene to match the goal scene step-by-step. We use breadth first search for the symbolic planner in our experiments.

D. Manipulation Execution

To execute the planned actions, the robot relies on existing motion planning library to generate collision-free arm trajectories. We use Moveit! [25] in our experiment. In our experiment, the object manipulation actions are essentially a sequence of pick and place actions. To pick up an object, we

pre-defined a set of possible pre-grasp and grasp end-effector poses w.r.t the object, i.e., grasping from the top and from the side. Affordance template [10] can be incorporated to extend our pipeline to deal with more complex manipulation behavior.

VI. Experiments

We evaluated 1) the accuracy of our cluttered scene estimator *DIGEST* in public dataset and our dataset, and 2) the performance the overall goal-directed PbD pipeline in tasks of setting up a tray.

A. DIGEST: Cluttered Scene Estimation

In this section, we examine our *DIGEST* estimator in two datasets: household occlusion dataset by Aldoma et al. [2] and our own scene estimation dataset with more objects and higher occlusion scenario. For the household occlusion dataset, we compare our method with D2P [20] under the same assumption and heuristics. For our own dataset, we trained a RCNN object detector and evaluated our work with the same metrics. All the experiments are ran on a Ubuntu 14.04 system with an Titan X Graphics card and CUDA 7.5.

1) Household Occlusion Dataset: We evaluate generative power of *DIGEST* with D2P, the state-of-art multi-object identification and 3 Dof localization method by Narayanan et, al. [20] on the household occlusion dataset [2]. The dataset contains 36 household objects with 3D geometry models and 22 RGBD test scenes with 80 objects in total. There are at most four objects in the scene. All objects are standing upright in the dataset, thus in this experiment, only 3 Dof object pose (x, y, θ) is estimated instead of 6 Dof, where x, y are the 2D locations of the object and θ is the yaw angle of the object.

Here, we take the same assumption as D2P: known object number and object categories. We also take exactly the same deep learning heuristics generated from D2P and use it in the same way. For each ROI in the depth image, candidates are generated only from the known object in the scene with high confidence score. The threshold is set to 0.2 as in D2P.

Pose estimation accuracy is percentage of correctly localized objects over the total number of objects in the dataset. An object is correctly localized if the pose error falls into a certain error bound. The pose error is computed the same way as in D2P. As shown in Figure 5, *DIGEST* performs better than D2P in small error bounds (position error smaller than 1cm and 10cm) and performs on par with D2P in larger position error bound. For position error bound with 1cm, *DIGEST* performs much better result than D2P since that 1) *DIGEST* explores the state space a lot more than D2P as we do not discretize the state space, and 2) *DIGEST* does not use ICP for local search, which D2P employs for their pose estimation step. *DIGEST* takes around 30 seconds (varying with the number of objects and the size of object geometries), which is faster than 139.74 seconds by D2P.

2) Our Dataset: We collected our own RGBD dataset for 15 household objects and 3D models. It has 20 test scenes with more objects and occlusions in each scene. There are 104 objects in total in all the scenes. We also generated ground truth pose by manually matching geometry models with observed point cloud in Blender. As we only estimate 3 Dof object pose, we take the ground truth value of the other three dimensions $(z, \text{roll}, \text{pitch})$.

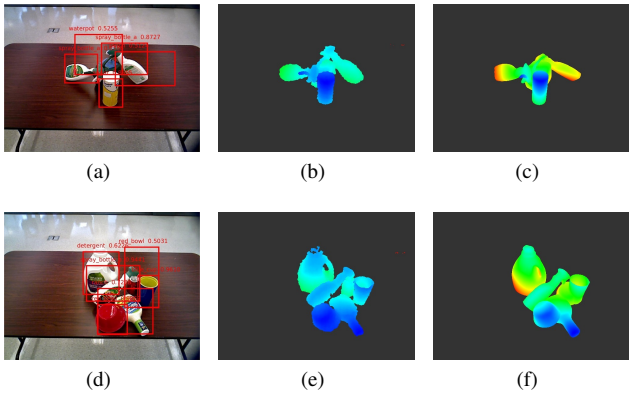

Fig. 7: (a)(d) The input RGB image with bounding boxes and object labels from R-CNN object detectors for generating hypothesized scenes. (b)(e) The observed depth image. (c)(f) The scene estimate result shows that *DIGEST* can identify the right scene from the hypothesized scenes when false detection exists.

We also employ the same criteria from the previous experiment to evaluate the pose errors, as shown in Figure 6. The increasing tail in the plots is because some objects with complex geometry like waterpot or spray bottle sometimes has a flipped orientation in the estimation. The proposed method *DIGEST* has the ability to resolve false positives from the discriminative object detectors by leveraging the robustness of generative estimation, as shown in Figure 7.

B. PbD Pipeline: Set up a Tray

We design our experiments with the service robot scenario in mind, as illustrated in Figure 1. The robot needs to prepare a tray of food such that it can be delivered to an user. We tested our system on scenes of three to five objects plus a tray, with objects stacking, and/or placed next to each other. The goal of the robot is to reproduce the **topological layout** of the goal scene, i.e., to reproduce the same inter-object relations, and the object poses can vary from the goal scene as long as the inter-object relations are satisfied.

1) Proof of Concept: To provide a proof of concept on our overall goal-directed PbD paradigm, we first tested our system with objects that are mostly primitive shaped, and with the assumption that the categories of the objects are known, and each object is standing upright, as in D2P [20]. Given the assumptions above, we used a previous method, i.e., APF [26] for 3 Dof object pose and scene graph estimation. In the later section, we relaxed these assumptions and used *DIGEST* for 6 Dof object pose estimation.

As summarized in Figure 8, we considered object stacking scenario and objects that are placed near each other. Each object has a set of pre-defined grasp pose, i.e., grasping from the top and from the side. During the manipulation experiments, the robot iterates through the provided grasp

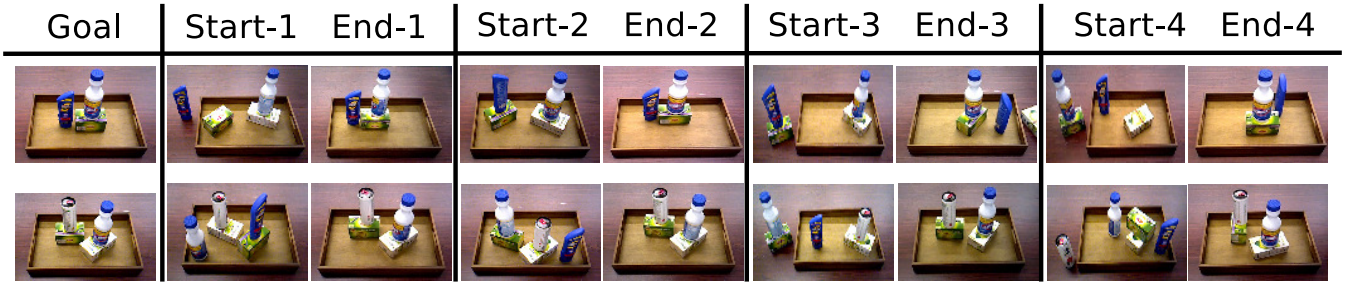

Fig. 8: We tested with primitive shaped objects in for two goal scenes. For each goal scene, four different start scenes are tested. For all 8 experiments, the scene has been successfully programmed through the robot.

poses, and select the one that it can plan a trajectory for.

2) PbD Enabled by DIGEST*:* We relax the assumptions in the previous section, thus the object categories are not known, and the objects can be in arbitrary poses. The test scenes are composed by objects that we have trained the object detector on. We used *DIGEST* to estimate the 6 Dof object poses, and then scene graph is derived from the pose estimation results.

As shown in Figure 10, the goal and start scenes are well estimated, as a collection of 6 Dof poses of objects. Based on the scene graph inferred from the object pose estimates, the robot generated a sequence of symbolic actions to transit the scene from the start state to the goal state. The pose estimation does fall into local minima sometimes, resulting in flipped orientation of objects that is highly symmetric, e.g., spray bottle and clorox bottle.

In these experiments, the grasp pose for each object is defined by the user instead of using existing work for grasp pose localization on objects. The reason for the user-defined grasp pose is that when the robot picks up an object, not all the valid grasp poses will be equivalently good for a later placement action. An example of the robot manipulation action sequence is as shown in Figure 9. A video of the experiments are available at `https://youtu.be/H1OMNR_qlDA`.

VII. Conclusion

We demonstrate a PbD paradigm for users to easily program a robot to complete manipulation tasks, where is goal of the task is represented as workspace scene graphs. We address the scene perception problem using the proposed *DIGEST* method, which is a discriminatively-informed generative method that recognizes objects and estimates their 6 Dof poses for manipulation. We show that *DIGEST*t performs better than D2P in public dataset for 3 Dof pose estimation and also performs well in our own dataset. We also show that users can use our overall PbD pipeline building on *DIGEST* to successfully program a Fetch robot to complete manipulation tasks of setting up a tray.

References

[1] B. Akgun, M. Cakmak, K. Jiang, and A. L. Thomaz. Keyframe-based learning from demonstration. *International Journal of Social Robotics*, 4(4):343–355, 2012.

[2] A. Aldoma, Z.-C. Marton, F. Tombari, W. Wohlkinger, C. Potthast, B. Zeisl, R. B. Rusu, S. Gedikli, and M. Vincze. Point cloud library. *IEEE Robotics & Automation Magazine*, 1070(9932/12), 2012.

[3] A. Aldoma, F. Tombari, R. B. Rusu, and M. Vincze. Our-cvfh–oriented, unique and repeatable clustered viewpoint feature histogram for object recognition and 6dof pose estimation. In *Joint DAGM (German Association for Pattern Recognition) and OAGM Symposium*, pages 113–122. Springer, 2012.

[4] S. Calinon and A. Billard. Incremental learning of gestures by imitation in a humanoid robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction*, pages 255–262. ACM, 2007.

[5] C. Chao, M. Cakmak, and A. L. Thomaz. Towards grounding concepts for transfer in goal learning from demonstration. In *2011 IEEE International Conference on Development and Learning (ICDL)*, volume 2, pages 1–6. IEEE, 2011.

[6] S. Chernova and M. Veloso. Interactive policy learning through confidence-based autonomy. *Journal of Artificial Intelligence Research*, 34(1):1, 2009.

[7] F. Dellaert, D. Fox, W. Burgard, and S. Thrun. Monte carlo localization for mobile robots. In *IEEE International Conference on Robotics and Automation (ICRA 1999)*, May 1999.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition*, 2014.

[9] D. H. Grollman and O. C. Jenkins. Incremental learning of subtasks from unsegmented demonstration. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 261–266. IEEE, 2010.

[10] S. Hart, P. Dinh, and K. Hambuchen. The affordance template ROS package for robot task programming. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6227–6234. IEEE, 2015.

[11] A. J. Ijspeert, J. Nakanishi, and S. Schaal. Movement imitation with nonlinear dynamical systems in humanoid robots. In *Robotics and Automation, 2002. Proceedings. ICRA'02. IEEE International Conference on*, volume 2, pages 1398–1403. IEEE, 2002.

[12] A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on pattern analysis and machine intelligence*, 21(5):433–449, 1999.

[13] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1):99–134, 1998.

[14] L. P. Kaelbling and T. Lozano-Pérez. Integrated task and motion planning in belief space. *The International Journal of Robotics Research*, page 0278364913484072, 2013.

[15] J. R. Kirk and J. E. Laird. Learning task formulations through situated interactive instruction. In *Proceedings of the Second Annual Conference on Advances in Cognitive Systems*, pages 219–236, 2013.

[16] D. McDermott, M. Ghallab, A. Howe, C. Knoblock, A. Ram, M. Veloso, D. Weld, and D. Wilkins. PDDL-the planning domain definition language. 1998.

[17] S. J. McKenna and H. Nait-Charif. Tracking human motion using auxiliary particle filters and iterated likelihood weighting. *Image and Vision Computing*, 25(6):852–862, 2007.

[18] S. Mohan, A. H. Mininger, J. R. Kirk, and J. E. Laird. Acquiring grounded representations of words with situated interactive instruction. In *Advances in Cognitive Systems*, 2012.

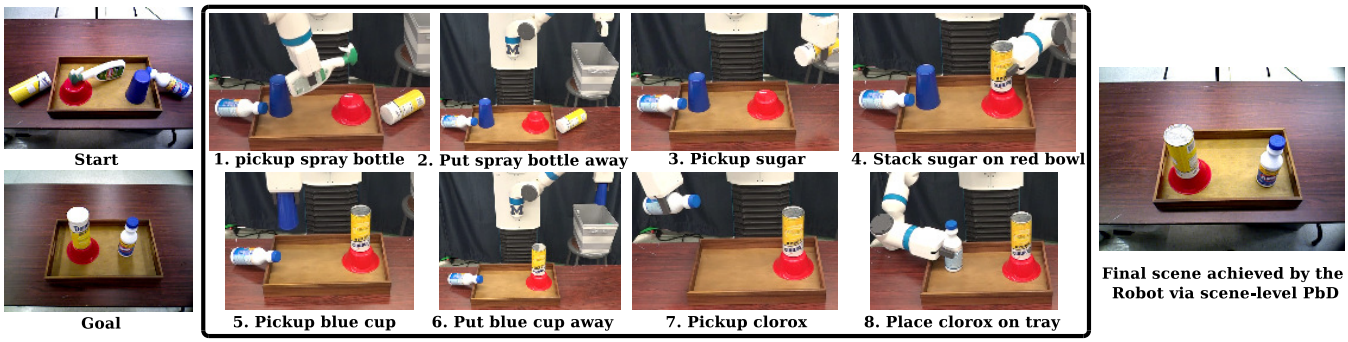

Fig. 9: Example of robot manipulation action sequence that transits the scene from the start state to the goal state via our scene-level PbD.

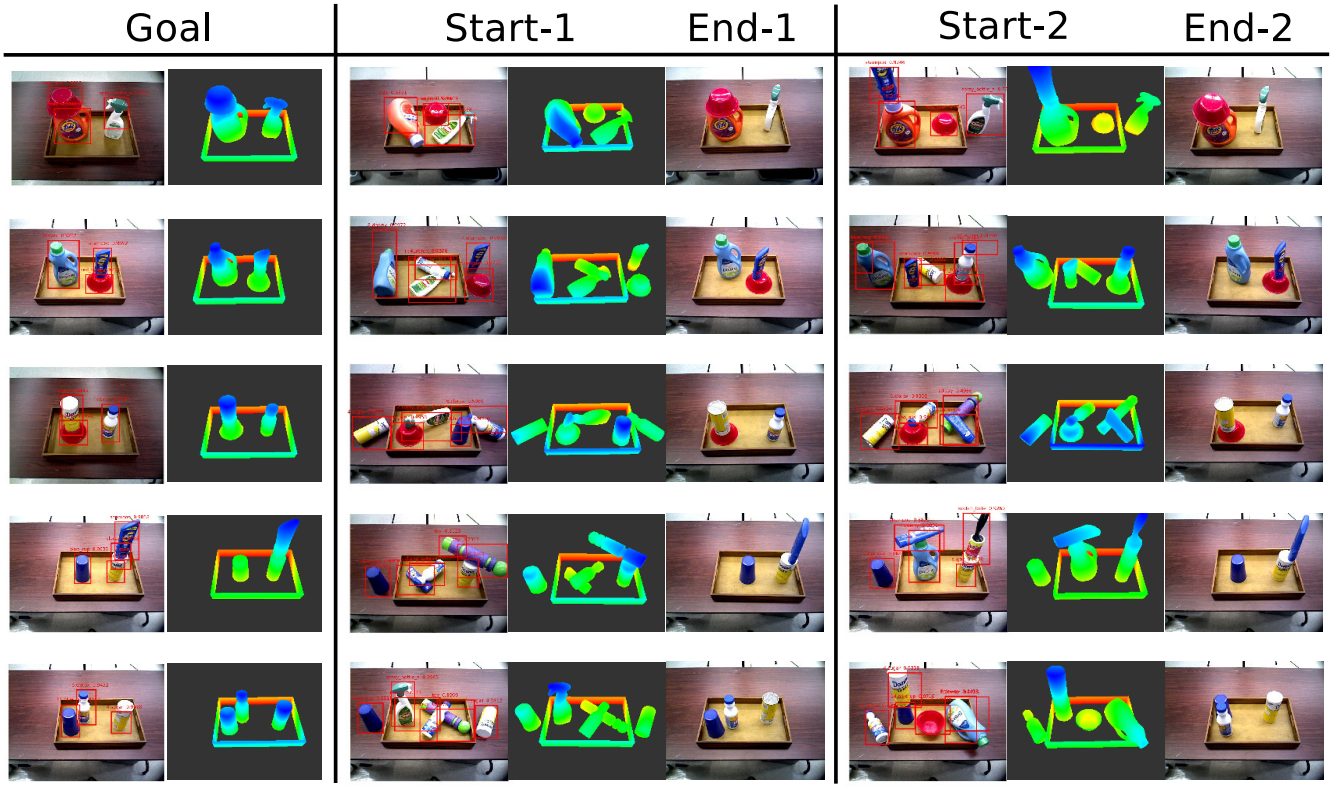

Fig. 10: We tested our pipeline for 5 different goal scenes. For each goal scene, two different starting scenes were tested. The rendered scene based on the object pose estimates are shown as depth images. The first four goal scenes involve object stacking, and the final goal scene involves objects placed near each other. In the experiments, the robot has successfully transited the scene from the start state to the goal state through manipulation actions.

[19] J. Nakanishi, J. Morimoto, G. Endo, G. Cheng, S. Schaal, and M. Kawato. Learning from demonstration and adaptation of biped locomotion. *Robotics and Autonomous Systems*, 47(2):79–91, 2004.

[20] V. Narayanan and M. Likhachev. Discriminatively-guided deliberative perception for pose estimation of multiple 3d object instances. In *Proceedings of Robotics: Science and Systems*, June 2016.

[21] S. Niekum, S. Osentoski, G. Konidaris, S. Chitta, B. Marthi, and A. G. Barto. Learning grounded finite-state representations from unstructured demonstrations. *The International Journal of Robotics Research*, 34(2):131–157, 2015.

[22] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3d registration. In *Robotics and Automation, 2009. ICRA'09. IEEE International Conference on*, pages 3212–3217. IEEE, 2009.

[23] R. B. Rusu, G. Bradski, R. Thibaux, and J. Hsu. Fast 3d recognition and pose using the viewpoint feature histogram. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 2155–2162. IEEE, 2010.

[24] M. R. Stevens and J. R. Beveridge. Localized scene interpretation from 3d models, range, and optical data. *Computer Vision and Image Understanding*, 80(2):111–129, 2000.

[25] I. A. Sucan and S. Chitta. Moveit! *Online Available: http://moveit. ros. org*, 2013.

[26] Z. Sui, O. C. Jenkins, and K. Desingh. Axiomatic particle filtering for goal-directed robotic manipulation. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 4429–4436. IEEE, 2015.

[27] Z. Sui, L. Xiang, O. C. Jenkins, and K. Desingh. Goal-directed robot manipulation through axiomatic scene estimation. *The International Journal of Robotics Research*, 36(1):86–104, 2017.

[28] J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

[29] C. L. Zitnick and P. Dollár. Edge boxes: Locating object proposals from edges. In *European Conference on Computer Vision*, pages 391–405. Springer, 2014.