2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

X-Ray: Mechanical Search for an Occluded Object by Minimizing Support of Learned Occupancy Distributions

Michael Danielczuk[1], Anelia Angelova[2], Vincent Vanhoucke[2], Ken Goldberg[1]

***Abstract*— For applications in e-commerce, warehouses, healthcare, and home service, robots are often required to search through heaps of objects to grasp a specific target object. For mechanical search, we introduce X-Ray, an algorithm based on learned occupancy distributions. We train a neural network using a synthetic dataset of RGBD heap images labeled for a set of standard bounding box targets with varying aspect ratios. X-Ray minimizes support of the learned distribution as part of a mechanical search policy in both simulated and real environments. We benchmark these policies against two baseline policies on 1,000 heaps of 15 objects in simulation where the target object is partially or fully occluded. Results suggest that X-Ray is significantly more efficient, as it succeeds in extracting the target object 82% of the time, 15% more often than the best-performing baseline. Experiments on an ABB YuMi robot with 20 heaps of 25 household objects suggest that the learned policy transfers easily to a physical system, where it outperforms baseline policies by 15% in success rate with 17% fewer actions. Datasets, videos, and experiments are available at https://sites.google.com/berkeley.edu/x-ray.**

I. Introduction

Mechanical search – extracting a desired object from a heap of objects – is a fundamental task for robots in unstructured e-commerce warehouse environments or for robots in home settings. It remains challenging due to uncertainty in perception and actuation as well as lack of models for occluded objects in the heap.

Data-driven methods are promising for grasping unknown objects in clutter and bin picking [7, 10, 23, 26, 27], and can reliably plan grasps on the most accessible object without semantic knowledge of the target object. Some reinforcement learning [9, 39] or hierachical [5] mechanical search policies use semantics, but have so far been limited to specific objects or heuristic policies.

In this paper, we draw on recent work on shape completion to reason about occluded objects [29, 35] and work on predicting multiple pose hypotheses [24, 32]. X-Ray combines occlusion inference and hypothesis predictions to estimate an occupancy distribution for the bounding box most similar to the target object to estimate likely poses – translations and rotations in the image plane. X-Ray can efficiently extract the target object from a heap where it is fully occluded or partially occluded (Figure 1).

This paper provides four contributions:

1) X-Ray (maXimize Reduction in support Area of occupancY distribution): a mechanical search policy that minimizes support of learned occupancy distributions.

[1]The Autolab at University of California, Berkeley. [2]Robotics at Google. mdanielczuk@berkeley.edu, anelia@google.com, vanhoucke@google.com, goldberg@berkeley.edu

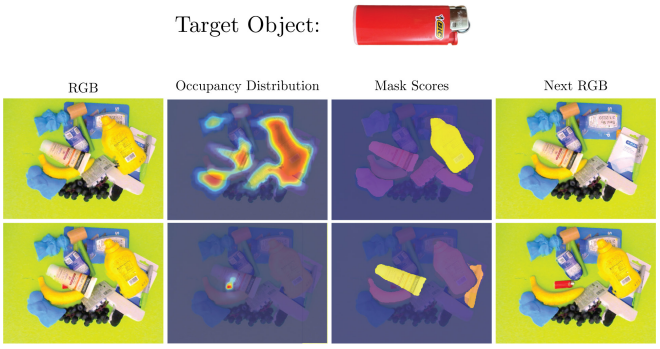

Fig. 1: Mechanical search with a fully occluded target object (top row) and a partially occluded target object (bottom row). We predict the target object occupancy distribution, which depends on the target object's visibility and the heap (second column). Each pixel value in the distribution image corresponds to the likelihood of that pixel containing part of the target object. X-Ray plans a grasp on the object that minimizes the estimated support of the resulting occupancy distribution to minimize the number of actions to extract the target object. We show two nearly-identical heaps; in the fully occluded case, X-Ray grasps the mustard bottle whereas in the partially occluded case, the policy grasps the face lotion (third column), resulting in the respective next states (fourth column).

2) An algorithm for estimating target object occupancy distributions using a set of neural networks trained on a dataset of synthetic images that transfers seamlessly to real images.
3) A synthetic dataset generation method and 100,000 RGBD images of heaps labeled with occupancy distributions for a single partially or fully occluded target object, constructed for transfer to real images.
4) Experiments comparing the mechanical search policy against two baselines in 1,000 simulated and 20 physical heaps that suggest the policy can reduce the median number of actions needed to extract the target object by 20% with a simulated success rate of 87% and physical success rate of 100%.

II. Related Work

A. Pose Hypothesis Prediction

There is a substantial amount of related work in computer vision on 3D and 6D pose prediction of both known and unknown objects in RGB, depth, and RGBD images [8, 13, 18, 36]. Many of these papers assume that the target

objects are either fully visible or have minor occlusions. In addition, many assume that there is no ambiguity in object pose due to self-occlusion or rotational symmetry of the object, as these factors can significantly decrease performance for neural network-based approaches [3]. Recent work has attempted to address the pose ambiguity that results from object geometry or occlusions by restricting the range of rotations [31] predicting multiple hypotheses for each detected object [24, 32]. Rupprecht *et al.* [32] find that refining multiple pose hypotheses to a 6D prediction outperforms single hypothesis predictions on a variety of vision tasks, such as human pose estimation, object classification, and frame prediction. Manhardt *et al.* [24] note that directly regressing to a rotation for objects with rotational symmetries can result in an averaging effect where the predicted pose does not match any of the possible poses; thus, they predict multiple pose hypotheses for objects with pose ambiguities to better predict the underlying pose and show Bingham distributions of the predicted hypotheses. However, only minor occlusions are considered and since ground truth pose distributions are not available for these images and objects, comparisons for continuous distributions can only be made qualitatively. Predicting multiple hypotheses or a distribution to model ambiguity has also been applied to gaze prediction from facial images [30], segmentation [14], and monocular depth prediction [38]. In contrast to these works, we learn occupancy distributions in a supervised manner.

B. Object Search

There has been a diverse set of approaches to grasping in cluttered environments, including methods that use geometric knowledge of the objects in the environment to perform wrench-based grasp metric calculations, nearest-neighbor lookup in a precomputed database, or template matching [1, 20, 25], as well as methods using only raw sensor data [12, 34], commonly leveraging convolutional neural networks [9, 10, 17]. While multi-step bin-picking techniques have been studied, they do not take a specific target object into account [22].

Kostrikov *et al.* [15] learn a critic-only reinforcement learning policy to push blocks in a simulated environment to uncover an occluded MNIST block. Zeng *et al.* [40] train joint deep fully-convolutional neural networks to predict both pushing and grasping affordances from heightmaps of a scene containing multicolored blocks, then show that the resulting policy (VPG) can separate and grasp novel objects in cluttered heaps. The policy can be efficiently trained on both simulated and physical systems, and can quickly learn elegant pushes to expand the set of available grasps in the scene. Yang *et al.* [39] train similar grasping and pushing networks as well as separate explorer and coordinator networks to address the exploration/exploitation tradeoff for uncovering a target object. Their policy learns to push through heaps of objects to find the target and then coordinate grasping and pushing actions to extract it, outperforming a target-centered VPG baseline in success rate and number of actions. Both approaches can generalize to objects outside the training distribution, although they are evaluated on a limited set of novel objects, and Yang *et al.* separate the cases where the target object is partially occluded and fully occluded. Additionally, we focus only on grasping actions, as some mechanical search environments may be constrained or objects may be fragile.

Recently, several approaches to the mechanical search problem have been proposed, both in tabletop and bin picking environments. Price *et al.* [29] propose a shape completion approach that predicts occlusion regions for objects to guide exploration in a tabletop scene, while Xiao *et al.* [37] implement a particle filter approach and POMDP solver to attempt to track all visible and occluded objects in the scene. However, 75% of the objects in Price *et al.*'s evaluation scenes are seen in training and Xiao *et al.*'s method requires models of each of the objects in the scene. We benchmark our policy on a variety of non-rigid, non-convex household objects not seen in training and require no object models. In previous work, Danielczuk *et al.* [5] proposed a general mechanical search problem formulation and introduced a two-stage perception and search policy pipeline. In contrast, we introduce a novel perception network and policy based on minimizing support of occupancy distributions that outperforms the methods introduced in [5].

III. Problem Statement

We consider an instance of the mechanical search problem where a robot must extract a known target object from a heap of unknown objects by iteratively grasping to remove non-target objects. The objective is to extract the target object using the fewest number of grasps.

A. Assumptions

- One known target object, fully or partially occluded by unknown objects in a heap on a planar workspace.
- A robot with a gripper, an overhead RGBD sensor with known camera intrinsics and pose relative to the robot.
- A maximum of one object is grasped per timestep.
- A target object detector that can return a binary mask of visible target object pixels when queried.

B. Definitions

We define the problem as a partially-observable Markov decision process (POMDP) with the 7-tuple $(S, A, T, R, \Omega, O, \gamma)$ and a maximum horizon H:

- **States** (S): A state $\mathbf{s}_k$ at timestep k consists of the robot, a static overhead RGBD camera, and a static bin containing $N+1$ objects, target object $\mathcal{O}_t$ and distractor objects $\{\mathcal{O}_{1,k}, \mathcal{O}_{2,k}, \ldots, \mathcal{O}_{N,k}\}$. No prior information is known about the N distractor objects.
- **Actions** (A): A grasp action $\mathbf{a}_k$ at timestep k executed by the robot's gripper.
- **Transitions** (T): In simulation, the transition model $T(\mathbf{s}_{k+1} \mid \mathbf{a}_k, \mathbf{s}_k)$ is equivalent to that used by Mahler et al. [22] and uses pybullet [4] for dynamics. On the physical system, next states are determined by executing

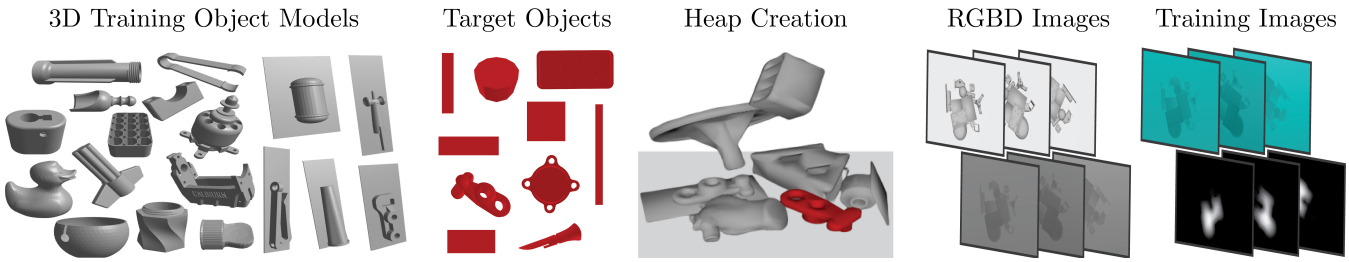

Fig. 2: Training dataset generation for learning the occupancy distribution function. Each dataset image is generated by sampling $N = 14$ object models from a dataset of 1296 CAD models. The target object (colored red) is dropped, followed by the N other objects (colored gray), into a planar workspace using dynamic simulation. Camera intrinsics and pose are sampled from uniform distributions centered around their nominal values and an RGBD image is rendered of the scene. The augmented depth image (top right), consisting of a binary target object modal mask and a two-channel depth image, is the only input used for training for seamless transfer from simulation to real images. The ground truth target object distribution is generated by summing all shifted amodal target object masks whose modal masks correspond with the target object modal mask.

the action on a physical robot and waiting until objects come to rest.

- **Rewards** (R): The reward $r_k = R(\mathbf{s}_k, \mathbf{a}_k, \mathbf{s}_{k+1}) \in \{0, 1\}$ is 1 if the target object is successfully grasped and lifted from the bin, otherwise the reward is 0.
- **Observations** (Ω): An observation $\mathbf{y}_k \in \mathbb{R}_+^{h \times w \times 4}$ at timestep k consists of an RGBD image with width w and height h taken by the overhead camera.
- **Observation Model** (O): A deterministic observation model $O(\mathbf{y}_k \mid \mathbf{s}_k)$ is defined by known camera intrinsics and extrinsics.
- **Discount Factor** (γ): To encourage efficient extraction of the target object, $0 < \gamma < 1$.

We also define the following terms:

- Modal Segmentation Mask $(\mathcal{M}_{m,i})$: the region(s) of pixels in an image corresponding to object $\mathcal{O}_i$ which are visible [11].
- Amodal Segmentation Mask $(\mathcal{M}_{a,i})$: the region(s) of pixels in an image corresponding to object $\mathcal{O}_i$ which are visible or invisible (occluded by other objects in the image) [11].
- The oriented minimum bounding box is the 3D box with the minimum volume that encloses the object, subject to no orientation constraints. We use this box to determine scale and aspect ratio for a target object.
- The *occupancy distribution* $\rho \in \mathcal{P}$ is the unnormalized distribution describing the likelihood that a given pixel in the observation image contains some part of the target object's amodal segmentation mask.

C. Objective

Given this problem definition and assumptions, the objective is to find a policy π_θ^* with parameters θ that maximizes the expected discounted sum of rewards:

$$\theta^* = \arg\max_\theta \ \mathbb{E}_{p(\tau|\theta)} \left[\sum_{k=0}^{H-1} \gamma^k R(\mathbf{s}_k, \pi_\theta(\mathbf{y}_k), \mathbf{s}_{k+1}) \right]$$

where $p(\tau \mid \theta) = \mathbb{P}(s_0) \Pi_{k=0}^{H-1} T(\mathbf{s}_{k+1} \mid \pi_\theta(\mathbf{y}_k), \mathbf{s}_k) O(\mathbf{y}_k \mid \mathbf{s}_k)$ is the distribution of state trajectories τ induced by a policy π_θ [22]. Maximizing this objective corresponds to removing the target object in the fewest number of actions.

D. Surrogate Reward

Because the reward defined in Section III-B is sparse and the transition function relies on complex inter-object and grasp contact dynamics, it is difficult to directly optimize for π_θ. Thus, we instead introduce a dense surrogate reward $\tilde{R}$ describing the reduction of the support of the target object's occupancy distribution:

$$\tilde{R}(\mathbf{y}_k, \mathbf{y}_{k+1}) = |\text{supp}(f_\rho(\mathbf{y}_k))| - |\text{supp}(f_\rho(\mathbf{y}_{k+1}))|,$$

where $f_\rho : \Omega \longrightarrow \mathcal{P}$ is a function that takes an observation $\mathbf{y}_k$ and produces the corresponding occupancy distribution ρ_k for a given bounding box and $\text{supp}(\rho) = \{(i, j) \in \{0, \ldots, h-1\} \times \{0, \ldots, w-1\} \mid \rho(i, j) \neq 0$ is the *support* of the occupancy distribution. Then, $|\text{supp}(\rho)|$ is the number of nonzero pixels in ρ. Section IV discusses a data-driven approximation for the function f_ρ while Section V discusses a greedy policy using the learned f_ρ and $\tilde{R}$.

IV. Learning Occupancy Distributions

We describe a method for estimating the function f_ρ via a deep neural network. Each pixel in the occupancy distribution $\rho \in [0, 1]^{h \times w}$ has a value representing the likelihood of it containing part of the target object's amodal segmentation mask, or the likelihood that some part of the object, in some planar translation or rotation, would occupy that pixel without any occlusions from other objects. We train this pixelwise distribution network on a dataset of augmented depth images and ground-truth occupancy distributions.

A. Dataset Generation

We generate a dataset of 10,000 synthetic augmented depth images labeled with target object occupancy distributions for a rectangular box target object. We choose 10 box targets of various dimensions ranging from $3cm \times 3cm \times 5mm$ to $9.5cm \times 0.95cm \times 5mm$ (aspect ratios varying from 1:1 to 10:1) with equal volume and generate a dataset for each,

Aspect Ratio	Test		Lid		Domino		Flute	
	Bal. Acc.	IoU	Bal. Acc.	IoU	Bal. Acc.	IoU	Bal. Acc.	IoU
1:1	98%	0.91	**93%**	**0.70**	92%	0.74	71%	0.30
2:1	97%	0.90	79%	0.44	**96%**	0.81	84%	0.44
5:1	97%	0.90	66%	0.23	96%	**0.83**	**86%**	**0.49**
10:1	97%	0.87	84%	0.49	82%	0.58	82%	0.41

TABLE I: Balanced accuracy (Bal. Acc.) and Intersection over Union (IoU) metrics for networks trained on various aspect ratio target boxes. The first column is the respective set of 2,000 test images for the network's training dataset. The other columns show how the networks can generalize to unseen objects outside the training distribution. Each dataset contains 1,000 test images for the lid, domino, and flute objects, respectively. These objects are shown in Figure 4 and have approximate aspect ratios of 1:1, 2:1, and 5:1, respectively. Each network performs very well when estimating distributions for its training target object and makes reasonable predictions for target objects with similar bounding box aspect ratios, even for novel target objects at different scales and in the presence of new occluding objects. However, a network trained on a small aspect ratio does not generalize well to higher aspect ratio objects, as it tends to overestimate the occupancy distribution.

resulting in a total of 100,000 dataset images. We choose a relatively small thickness for the target so that it is more likely to be occluded in heaps of objects, as it tends to lie flat on the workspace. We sample a state $\mathbf{s}_0$ by uniformly sampling a set of N 3D CAD models as well as a heap center and 2D offsets for each object from a 2D truncated gaussian. First, $\mathcal{O}_t$ is dropped from a fixed height above the workspace, then the other N objects are dropped one by one from a fixed height and dynamic simulation is run until all objects come to rest (all velocities are zero). Any objects that fall outside of the workspace are removed. N is drawn from a Poisson distribution ($\lambda = 12$) truncated such that $N \in [10, 15]$. The 3D CAD models are drawn from a dataset of 1296 models available on Thingiverse, including "packaged" models, where the original model has been augmented with a rectangular backing, as in [23]. The camera position is drawn from a uniform distribution over a viewsphere and camera intrinsics are sampled uniformly from a range around their nominal values. We use the Photoneo Phoxi S datasheet intrinsics and a camera pose where the camera points straight down at the heap at a height of $0.8m$ for the nominal values. An RGBD image is rendered and augmented depth images are created by concatenating a binary modal mask of the target object with the depth image. Note that if the target object is not visible, the image is equivalent to a two-channel depth image, as the first channel is all zeros. We find that training on these images, as opposed to training on RGBD images directly, allows for seamless transfer between simulated and real images.

To generate the ground-truth occupancy distribution, we find the set of translations and rotations in the image plane for the target object such that an image rendered from the same camera pose with all other objects in the scene in the same respective poses will yield the same target object modal segmentation mask. Thus, when the object is fully visible, the distribution's support collapses to the pixels of the target object modal segmentation mask. However, when the object is partially or fully occluded, then multiple target object translations or rotations may result in the same image and the distribution will spread to reflect where the target could hypothetically be hiding. In practice, we generate this distribution by discretizing the set of possible translations into a 64×48 grid (every 8 pixels in the image) and rotations into 16 bins, then shifting and rotating a target-only depth image to each point on the grid, offsetting by the depth of the bottom of the workspace at that point. By comparing the depths for the set of these shifted and rotated depth images to original depth image, we can determine the modal segmentation mask for the target object as if it were at each location. Any location for which there is intersection-over-union (IoU) greater than 0.9 (or, in cases where the target object has a blank modal mask due to full occlusion, any location for which the modal mask is also blank) is considered to result in the same image. Then, the amodal target object masks from all locations resulting in the same image are summed and the resulting normalized single-channel image is the ground truth occupancy distribution. A visualization of this process is shown in Figure 2. Dataset generation for 10,000 images took about 5 hours on an Ubuntu 16.04 machine with a 12-core 3.7 GHz i7-8700k processor.

B. Occupancy Distribution Model

We split each dataset of 10,000 images image-wise and object-wise into training and test sets (8,000 training images and 2,000 test images, where objects are also split such that training objects only appear in training images and test objects only appear in test images). We train a fully-convolutional network with a ResNet-50 backbone [19] using a pixelwise mean-squared-error loss for 40 epochs with a learning rate of 10^{-5}, momentum of 0.99, and weight decay of 0.0005. The input images were preprocessed by subtracting the mean pixel values calculated over the dataset and transposing to BGR. Training took approximately 2.5 hours on an NVIDIA V100 GPU and a single forward pass took 6 ms on average as compared to 1.5 s for generating the ground-truth distribution.

C. Simulation Experiments for Occupancy Distributions

We benchmark the trained model on the full set of 2,000 test images as well as on 1,000 images with three other simulated target objects shown in Figure 4 - a lid, a domino, and a flute - to test generalization to object shapes, aspect

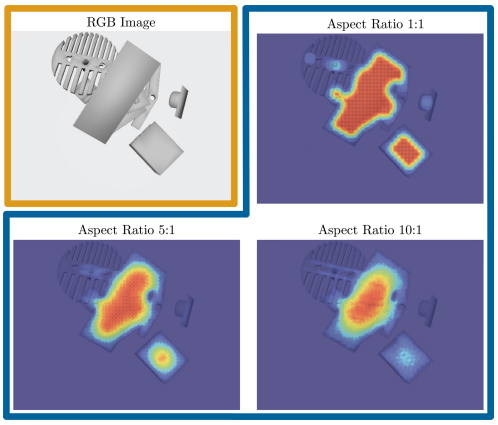

Fig. 3: The ground truth occupancy distributions for a target object of various aspect ratios for the same heap image.

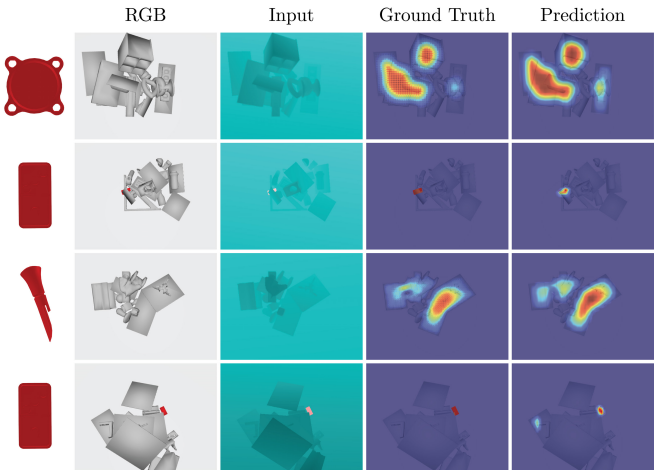

Fig. 4: Example predicted target object occupancy distributions for three target objects, a lid, domino, and flute, unseen during training (far left). Warmer colors indicate a higher likelihood of that pixel containing part of the target object's amodal mask. The network is able to accurately predict a distribution across many objects, a collapsed distribution when the object is partially visible, and multimodal distributions when there are gaps between objects (top three rows). The final row shows a failure mode where the network spuriously predicts an extra mode for the distribution when the target object is partially occluded.

ratios and scales not seen during training. We chose these target objects due to their diversity in scale and object aspect ratio (e.g., the flute is longer, thinner, and deeper, while the lid is nearly square and flat). We report two metrics: balanced accuracy, the mean of pixelwise accuracies on positive and negative pixel labels, and intersection-over-union, the sum of positive pixels in both the ground truth and predicted distribution divided by the sum of total positive pixels in either distribution. We consider true positives as the ground truth pixel having normalized value greater than 0.1 and the predicted value being within 0.2 of the ground truth value. Similarly, we consider true negatives as the ground truth pixel having normalized value less than 0.1 and the predicted value being within 0.2 of the ground truth value. Results are shown in Table I.

Target Object Scale. For objects of different scale than the training target object, we scale the input image by a factor equal to the difference in scale between the box target object and the other target object, feed it through the network, and then rescale the output distribution. We find that this scaling dramatically improves performance with minimal preprocessing of the input image; for example, when testing on the lid object, which is about twice as large as the training box object, we increase balanced accuracy and IoU from 63.0% and 0.186 to 93.1% and 0.697, respectively.

Target Aspect Ratios. We found that, while our network performed well on objects with similar aspect ratios, longer and thinner objects with higher aspect ratios resulted in the model overestimating the support of the distribution. This effect can be seen in Figure 3, which shows ground truth occupancy distributions for target objects of different aspect ratios in the same heap image. Table I suggests that the trained networks can accurately predict occupancy distributions for target objects that have similar aspect ratios to the training boxes, but do not perform as well when tasked with predicting a distribution for objects with dramatically different aspect ratios. In particular, the network trained with a 1:1 box target object tends to overestimate the support for target objects with high aspect ratios, leading to a drop in metrics. This effect is especially visible along corners of occluding objects, where more rotations of a low aspect ratio object are possible, while only one or two rotations of a high aspect ratio object are possible.

Figure 4 shows occupancy distribution predictions with ground truth distributions for the three unseen objects using the network trained on the closest aspect ratio target object and scaled appropriately. Results suggest that the network is able to accurately predict diverse distributions when occluding objects not seen in training are present. Figure 4 suggests not only that the network can predict the correct distribution spanning multiple occluding objects in unimodal and multimodal cases when the target object is fully occluded, but also that it can correctly collapse the distribution to a small area around the visible part of the target object when it is only partially occluded.

V. X-Ray: Mechanical Search Policy

Using the learned occupancy distribution function f_ρ, we propose X-Ray, a mechanical search policy that optimizes for the objective and surrogate reward $\tilde{R}$ defined in Section III. We create both simulated and physical object heaps and generate overhead camera images using an observation model based on the Photoneo PhoXi S depth camera. The heap RGBD image and target object are inputs to the perception system, which uses the network trained on the most similar bounding box to the target object to predict an occupancy distribution for the target. The policy takes the predicted distribution and a set of modal segmentation masks for the scene and computes a grasping action that would

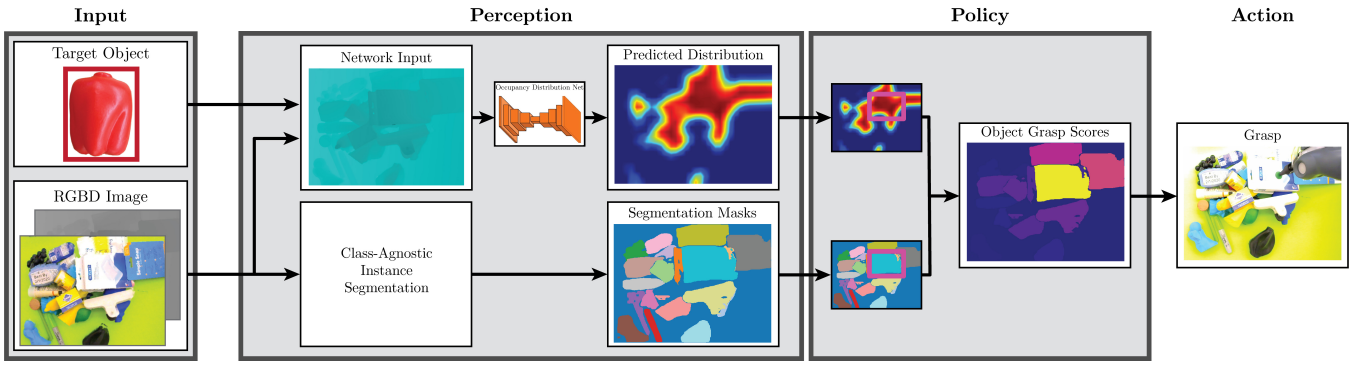

Fig. 5: The perception stage takes as input an RGBD image of the scene and outputs an occupancy distribution prediction using a network based on the target object bounding box dimensions and the created augmented depth image. The perception stage also produces a set of segmentation masks. The X-Ray mechanical search policy then finds the mask that has the most overlap with the occupancy distribution (colored yellow in the grasp scores image) and plans a grasp on that mask.

maximally reduce the support of the subsequent distribution. Specifically, the policy takes an element-wise product of each segmentation mask with the predicted occupancy distribution and sums over all entries in the resulting image, leading to a score for each of the segmentation masks. The policy then plans a grasp on the object mask with the highest score and executes it, as shown in Figure 5.

A. Simulation Experiments with X-Ray

We first evaluate the mechanical search policy with simulated heaps of novel objects. To further test the ability of the learned network to generalize to unseen occluding objects, we use a set of objects unseen in training and validation: 46 YCB objects [2] and 13 "packaged" YCB objects (augmented in the same way as described in Section IV). Initial states were generated as explained in Section IV, first dropping the target object, followed by the other N objects. We use $N = 14$ so each heap initially contained 15 total objects, colorreda similar or larger size to previous bin-picking work [22, 26]. As the focus of this work was not instance segmentation or target detection, we use ground truth segmentation masks and target binary masks in simulation, although we note that any class-agnostic instance segmentation network [6, 16] or object detection network [41] can be substituted. For each grasp, either a parallel jaw or suction cup grasp, we use wrench space analysis to determine whether it would result in the object being lifted from the workspace under quasi-static conditions [20, 21, 28]. If the grasp is collision-free and the object can be lifted, the object is lifted until the remaining objects come to rest using dynamic simulation implemented in pybullet, resulting in the next state. Otherwise the state remains unchanged.

In addition to the policy proposed here, we evaluate two previously proposed baseline policies, **Random** and **Largest** [5]. The **Random** policy that first attempts to grasp the target object, and, if no grasps are available on the target object, grasps an object chosen uniformly at random from the bin. The **Largest** policy that first attempts to grasp the target object, and, if no grasps are available on the target object, iteratively attempts to grasp the objects in the bin according to the size of their modal segmentation mask.

Policy	Success Rate	Number of Actions Quartiles		
Random	42%	4	7	9
Largest	67%	4	**5**	7
X-Ray	**82%**	**3**	**5**	**6**

TABLE II: Evaluation metrics for each policy over 1,000 simulated rollouts. The lower quartiles, medians, and upper quartiles for number of actions are reported for successful rollouts. X-Ray extracts the target at a higher success rate with significantly fewer actions.

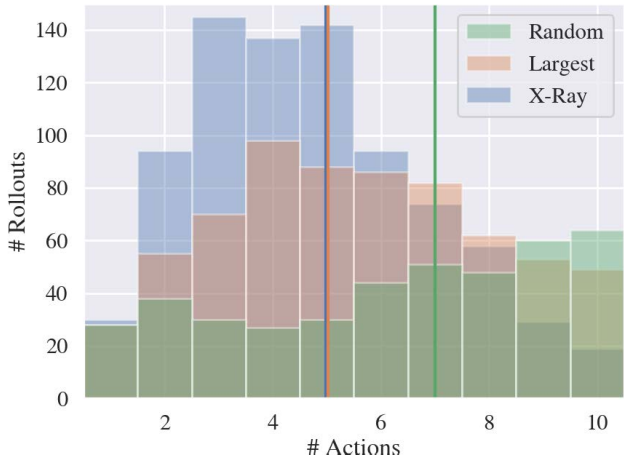

Fig. 6: Histogram of the number of actions taken to extract the target object over the 1,000 simulated rollouts for the three policies tested. The median number of actions for each policy is shown by the corresponding vertical line.

Each policy was rolled out on 1,000 total heaps until either the target object was grasped (successful rollout) or the horizon $H = 10$ was reached (failed rollout). We benchmark each policy using two metrics: success rate of the policy and mean number of actions taken to extract the target object in successful rollouts. Table II and Figure 6 show these metrics and the distribution of successful rollouts over the number of actions taken to extract the target object, respectively.

While the Random and Largest policies occasionally are able to quickly extract the target object, X-Ray consistently extracts the target in fewer actions and succeeds in 15% more heaps than the best-performing baseline. Largest is a reasonable heuristic for these heaps, as shown in [5], as large objects typically have a greater chance of occluding the target, but X-Ray combines this intuition with superior performance when the object is partially occluded. X-Ray outperforms the Largest policy on heaps where the target object is partially occluded by a thin or small object (such as a fork or dice) at some point during the rollout. In these scenarios, a robust grasp is often not available on the target object, and while X-Ray can correctly identify that the occluding object should be removed, the Largest policy will often grasp a larger object further from the target object. In scenarios where there are many large objects, but some are lying to the side, X-Ray will typically grasp objects that are in the more cluttered area of the bin, since they are more likely to reveal the target object. This behavior is a function of weighting the object area by the predicted distribution, which encourages the policy to ignore solitary objects.

B. Physical Experiments with X-Ray

We also evaluate X-Ray with heaps of novel household objects on a physical ABB YuMi robot with a suction cup and parallel jaw gripper, using two target objects. Some examples of the objects used can be seen in Figures 1 and 5. Initial states were generated by placing the target object on the workspace, filling a bin with the N other objects, and then dumping the bin on top of the target object. In these heaps, $N = 24$ was used so that each heap initially contained 25 total objects. We chose 25 total objects because it has been commonly used in cluttered bin-picking environments [23] and objects tend to disperse further on the physical setup. For segmentation masks, we used the class-agnostic instance segmentation network from [6], and for grasp quality analysis, we used FC-GQCNN [33]. To generate binary target masks, we use HSV color segmentation from OpenCV and use red target objects. While we make this assumption for simplicity, we note that we could substitute this process with a target object segmentation method that uses visual features, semantics and shape, such as the one described in [6].

We perform 20 rollouts for each of the three policies. Each policy was rolled out until either the target object was grasped (successful rollout) or the horizon $H = 10$ was reached (failed rollout). We report the same metrics as in the simulated experiments in Table III.

We find that X-Ray outperforms both baselines, extracting the target object in a median 5 actions over the 20 rollouts as compared to 6 actions for the Largest and Random policies while succeeding in extracting the target object within 10 actions in each case. These results suggest that X-Ray not only can extract the target more efficiently than the baseline policies, but also has lower variance. The Largest policy performed comparatively worse with more objects in the heap than in simulation, as it relies heavily on accurate segmentation masks. However, when objects are densely clustered together, segmentation masks are often merged, leading to grasps on smaller objects that do not uncover the target. In this case or in the case of spurious segmentation masks that do not cover objects, X-Ray reduces this reliance on accurate segmentation masks, as the occupancy distribution and segmentation are combined to create a score for the mask. This property of X-Ray causes it to compare favorably to a policy that directly scores segmentation masks based on their relationship to the target object geometry. X-Ray also reduces reliance on the target object binary mask being accurate; if the detector cannot see enough of the target object to generate a detection even when it is partially visible, X-Ray will continue to try and uncover it according to the fully occluded occupancy distribution until more of the target is revealed.

Policy	Success Rate	Number of Actions Quartiles		
Random	85%	**4**	6	7
Largest	85%	4	6	7
X-Ray	**100%**	**4**	**5**	**5.25**

TABLE III: Evaluation metrics for each policy over 20 physical rollouts. The lower quartiles, medians, and upper quartiles for the number of actions are reported across successful rollouts. X-Ray extracts the target with significantly fewer actions, always extracting it within 10 actions.

VI. Discussion and Future Work

We present X-Ray, a mechanical search algorithm that minimizes support of a learned occupancy distribution. We showed that a model trained only on a synthetic dataset of augmented depth images labeled with ground truth distributions learns to accurately predict occupancy distributions for target objects unseen in training. We benchmark X-Ray in both simulated and physical experiments, showing that it can efficiently extract the target object from challenging heaps containing 15-25 objects that fully occlude the target object in 82% - 100% of heaps using a median of just 5 actions.

In future work, we will address some of the failure modes of the system, especially for objects that are significantly non-planar. Currently, the assumption that the object is flat can result in incorrect occupancy distributions for taller objects. Additionally, we will look to add memory to the policy so that if objects shift into previously free space, the distribution will not cover that area, and explore reinforcement learning policies based on a reward of target object visibility.

Acknowledgments

This research was performed at the AUTOLAB at UC Berkeley in affiliation with the Berkeley AI Research (BAIR) Lab. The authors were supported in part by donations from Google. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1752814. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Sponsors. We thank our colleagues and collaborators who provided helpful feedback, code, and suggestions, especially Julian Ibarz, Brijen Thananjeyan, Andrew Li, Andrew Lee, Andrey Kurenkov, Roberto Martín Martín, Animesh Garg, Matt Matl, and Ashwin Balakrishna.

REFERENCES

[1] D. Berenson and S. S. Srinivasa, "Grasp synthesis in cluttered environments for dexterous hands," in *Proc. IEEE-RAS Int. Conf. on Humanoid Robots*, IEEE, 2008, pp. 189–196.

[2] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols," *arXiv preprint arXiv:1502.03143*, 2015.

[3] E. Corona, K. Kundu, and S. Fidler, "Pose estimation for objects with rotational symmetry," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 7215–7222.

[4] E. Coumans and Y. Bai, *Pybullet, a python module for physics simulation, games, robotics and machine learning*, http://pybullet.org/, 2017.

[5] M. Danielczuk, A. Kurenkov, A. Balakrishna, M. Matl, D. Wang, R. Martín-Martín, A. Garg, S. Savarese, and K. Goldberg, "Mechanical search: Multi-step retrieval of a target object occluded by clutter," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2019.

[6] M. Danielczuk, M. Matl, S. Gupta, A. Li, A. Lee, J. Mahler, and K. Goldberg, "Segmenting unknown 3d objects from real depth images using mask r-cnn trained on synthetic data," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, 2019, pp. 7283–7290.

[7] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, IEEE, 2016, pp. 598–605.

[8] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Proc. Asian Conf. on Computer Vision (ACCV)*, Springer, 2012, pp. 548–562.

[9] E. Jang, S. Vijayanarasimhan, P. Pastor, J. Ibarz, and S. Levine, "End-to-end learning of semantic grasping," *arXiv preprint arXiv:1707.01932*, 2017.

[10] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," *arXiv preprint arXiv:1806.10293*, 2018.

[11] G. Kanizsa, *Organization in vision: Essays on Gestalt perception*. Praeger Publishers, 1979.

[12] D. Katz, A. Venkatraman, M. Kazemi, J. A. Bagnell, and A. Stentz, "Perceiving, learning, and exploiting object affordances for autonomous pile manipulation," *Autonomous Robots*, vol. 37, no. 4, pp. 369–382, 2014.

[13] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab, "Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 1521–1529.

[14] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. A. Eslami, D. J. Rezende, and O. Ronneberger, "A probabilistic u-net for segmentation of ambiguous images," in *Proc. Advances in Neural Information Processing Systems*, 2018, pp. 6965–6975.

[15] I. Kostrikov, D. Erhan, and S. Levine, "End to end active perception," 2016.

[16] W. Kuo, A. Angelova, J. Malik, and T.-Y. Lin, "Shapemask: Learning to segment novel objects by refining shape priors," *arXiv preprint arXiv:1904.03239*, 2019.

[17] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. Journal of Robotics Research (IJRR)*, vol. 34, no. 4-5, pp. 705–724, 2015.

[18] Y. Li, G. Wang, X. Ji, Y. Xiang, and D. Fox, "Deepim: Deep iterative matching for 6d pose estimation," in *Proc. European Conf. on Computer Vision (ECCV)*, 2018, pp. 683–698.

[19] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[20] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, 2016, pp. 1957–1964.

[21] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," *Proc. Robotics: Science and Systems (RSS)*, 2017.

[22] J. Mahler and K. Goldberg, "Learning deep policies for robot bin picking by simulating robust grasping sequences," in *Conf. on Robot Learning (CoRL)*, 2017, pp. 515–524.

[23] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, eaau4984, 2019.

[24] F. Manhardt, D. M. Arroyo, C. Rupprecht, B. Busam, N. Navab, and F. Tombari, "Explaining the ambiguity of object detection and 6d pose from visual data," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2019.

[25] M. Moll, L. Kavraki, J. Rosell, *et al.*, "Randomized physics-based motion planning for grasping in cluttered and uncertain environments," *IEEE Robotics & Automation Letters*, vol. 3, no. 2, pp. 712–719, 2017.

[26] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.

[27] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, 2016, pp. 3406–3413.

[28] D. Prattichizzo and J. C. Trinkle, "Grasping," in *Springer handbook of robotics*, Springer, 2008, pp. 671–700.

[29] A. Price, L. Jin, and D. Berenson, "Inferring occluded geometry improves performance when retrieving an object from dense clutter," in *Int. S. Robotics Research (ISRR)*, 2019.

[30] S. Prokudin, P. Gehler, and S. Nowozin, "Deep directional statistics: Pose estimation with uncertainty quantification," in *Proc. European Conf. on Computer Vision (ECCV)*, 2018, pp. 534–551.

[31] M. Rad and V. Lepetit, "Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 3828–3836.

[32] C. Rupprecht, I. Laina, R. DiPietro, M. Baust, F. Tombari, N. Navab, and G. D. Hager, "Learning in an uncertain world: Representing ambiguity through multiple hypotheses," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*, 2017, pp. 3591–3600.

[33] V. Satish, J. Mahler, and K. Goldberg, "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks," *IEEE Robotics & Automation Letters*, vol. 4, no. 2, pp. 1357–1364, 2019.

[34] A. Saxena, L. L. Wong, and A. Y. Ng, "Learning grasp strategies with partial shape information.," in *AAAI*, vol. 3, 2008, pp. 1491–1494.

[35] J. Varley, C. DeChant, A. Richardson, J. Ruales, and P. Allen, "Shape completion enabled robotic grasping," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 2442–2447.

[36] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[37] Y. Xiao, S. Katt, A. ten Pas, S. Chen, and C. Amato, "Online planning for target object search in clutter under partial observability," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, 2019, pp. 8241–8247.

[38] G. Yang, P. Hu, and D. Ramanan, "Inferring distributions over depth from a single image," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2019.

[39] Y. Yang, H. Liang, and C. Choi, "A deep learning approach to grasping the invisible," *arXiv preprint arXiv:1909.04840*, 2019.

[40] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 4238–4245.

[41] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Networks and Learning Systems*, vol. 30, no. 11, pp. 3212–3232, 2019.