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

Parts-Based Articulated Object Localization in Clutter Using Belief Propagation

Jana Pavlasek Stanley Lewis Karthik Desingh Odest Chadwicke Jenkins

Abstract— **Robots working in human environments must be able to perceive and act on challenging objects with articulations, such as a pile of tools. Articulated objects increase the dimensionality of the pose estimation problem, and partial observations under clutter create additional challenges. To address this problem, we present a generative-discriminative parts-based recognition and localization method for articulated objects in clutter. We formulate the problem of articulated object pose estimation as a Markov Random Field (MRF). Hidden nodes in this MRF express the pose of the object parts, and edges express the articulation constraints between parts. Localization is performed within the MRF using an efficient belief propagation method. The method is informed by both part segmentation heatmaps over the observation, generated by a neural network, and the articulation constraints between object parts. Our generative-discriminative approach allows the proposed method to function in cluttered environments by inferring the pose of occluded parts using hypotheses from the visible parts. We demonstrate the efficacy of our methods in a tabletop environment for recognizing and localizing hand tools in uncluttered and cluttered configurations.**

I. Introduction

Robot assistants operating in real-world environments should be capable of performing maintenance and repair tasks. Going beyond pick-and-place actions, we aim to enable robots to use the diversity of objects it might encounter. The ability to use commercial, off-the-shelf hand tools is critical for robots to perform tasks in unstructured, everyday environments. In order to accomplish this, robots must be able to identify and localize tools in an arbitrary cluttered scene to plan appropriate actions toward performing a task.

Recognizing hand tools and localizing their pose remains challenging in common human environments. These challenges arise from uncertainty caused by physical clutter and the high-dimensionality of the space of poses multiple objects in contact may occupy. Many hand tools are articulated, adding complexity to the localization problem by introducing additional degrees-of-freedom. Figure 1 shows one example of hand tools in a cluttered scene that could be typical in a work area.

State-of-the-art object and pose recognition methods have been proposed that estimate the six degree-of-freedom (6D) pose of objects using convolutional neural networks (CNNs) [1], [2]. Other methods have accomplished pose estimation using probabilistic inference [3], [4]. However, localizing *articulated* objects remains a challenge for these methods due to both the added degrees of freedom that arise from articulations and occlusions due to clutter. Parts-based representations [5] have the potential to achieve higher levels of robustness under these conditions than whole-object based approaches. For such methods to be suitable for robot manipulation tasks, they must be able to localize 6D pose with reasonable computational efficiency. We suggest that generative inference methods, if made more computationally efficient, offer compelling and complementary benefits to modern deep learning. Additionally, a parts-based representation can provide information about the affordances of an object, because robot actions are typically applied to the object parts.

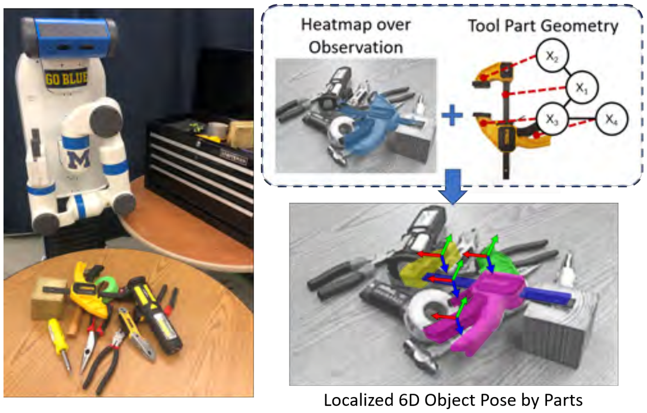


Fig. 1: The ProgressLab Fetch robot perceiving a cluttered scene with tools (articulated objects): The top right image shows the pixelwise heatmap over the RGB observation for the clamp object, which misses the top part of the clamp as it is occluded by another object. Our parts-based localization method is able to use this partially informed heatmap along with the clamp geometry to localize the 6D pose of the entire object (bottom right image).

In this paper, we present a method for recognition and localization of articulated objects in clutter suited to robotic manipulation of object affordances. We formulate the problem of articulated object pose estimation as a Markov Random Field (MRF), representing the 6D poses of each rigid object part and the articulation constraints between them. We propose a method to perform inference over the MRF based on message passing. We are inspired by work by Desingh et al. [6], in which parts-based articulated object localization is facilitated by combining information from both the observation as well as the compatibility with neighbouring parts within the inference process. Our method is informed jointly by a learned likelihood modelled by a CNN, as well as by the known articulation constraints between each component part. We assume known object mesh models and kinematic constraints in the form of a Unified Robot Description Format

J. Pavlasek, S. Lewis, K. Desingh and O. C. Jenkins are with the Department of Electrical Engineering and Computer Science, Robotics Institute, University of Michigan, Ann Arbor, MI, USA 48109
{pavlasek, stanlew, kdesingh, ocj}@umich.edu

This work was supported in part by NSF award IIS-1638060.

978-1-7281-6211-9/20/$31.00 ©2020 IEEE

(URDF) file, a standard geometrical object representation in the field of robotics.

By employing generative inference to integrate both data-driven techniques and domain knowledge about the object models, we leverage the speed and representational abilities of deep CNNs, while retaining the ability to reconcile noisy results and provide structure and context to the estimate. Methods we present emphasize novel synthesis of (1) efficient discriminative-generative inference via nonparametric belief propagation for pose estimation of articulated objects, and (2) a learned part-based likelihood to evaluate hypotheses of articulated object pose against RGB observations. We present results using a custom dataset made up of commercial, off-the-shelf hand tools with robot observations containing varying levels of clutter.

II. Related Work

Pose estimation has received considerable attention in robotics. Here, we discuss related work that focuses on rigid body, parts-based, and articulated object pose estimation.

A. Rigid Body Pose Estimation

Methods that tackle the problem of rigid body pose estimation include geometry-based registration approaches [7], generative approaches [3], [8], approaches combining discriminative and generative methods [4], [9]–[13], and end-to-end learning approaches [1], [2]. Here, we focus on the discriminative-generative methods and end-to-end learning methods that are most relevant to this work.

Combining the discriminative power of feature-based methods with generative inference has been successful under challenging conditions such as background and foreground clutter [4], [9], [13], adversarial environment conditions [11], and uncertainty due to robot actions [10], [12]. We are inspired by the success of the above approaches in taking advantage of the speed of discriminative methods to perform analysis and synthesis based generative inference.

Xiang et al. propose an end-to-end network for estimating 6D pose from RGB images [1]. This work was further extended to use synthetic data generation and augmentation techniques to improve performance [2]. Wang et al. [14] propose an end-to-end network that uses depth information along with RGB information. These methods rely significantly on the textured appearance of objects. More importantly, the state representation used in these methods assume rigidity. Our attempts to adapt these methods for articulated objects required considerably more training data and computation time. In addition, estimates from the end-to-end methods can be noisy, especially in challenging cluttered scenarios. We believe estimates from these methods will be a good prior to help generative methods recover under challenging scenarios. Hence, in this work, we learn a likelihood function over the observation to inform the generative inference.

B. Parts-Based Pose Estimation

Understanding objects in terms of their parts paves the way to meaningful and purposeful action execution, such as tool-use. Parts-based representations have been proposed to aid scene understanding and action execution [5], [15], [16], and have recently garnered attention within the robotics and perception communities [17], [18]. Parts-based localization has led to research in recognizing objects and their articulated parts [19]. Parts-based perception for objects in human environments is often limited to recognition and classification tasks. Parts-based pose estimation is often considered for human body pose [20] and hand pose [21] estimation with fixed graphical models. Here, we propose a general framework for estimating pose of articulated objects, such as hand tools, that includes parts with fixed transforms as constraints.

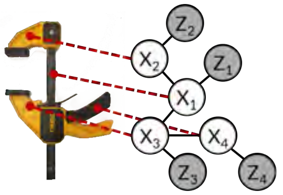


Fig. 2: The Markov Random Field representation of the clamp object. The clamp is broken up into four parts to fully represent both its affordances and articulations. The hidden nodes (white) represent the pose of each part, X_i. The observed nodes (grey) correspond to the sensor observation.

C. Articulated Pose Estimation and Tracking

Probabilistic inference is a popular technique in robot perception for articulated body tracking [22]–[24], where filtering-based approaches alongside novel observation models have been proposed. These tracking frameworks are either initialized to the ground truth poses of objects, or applied to robot manipulators, where the inference is informed by joint encoder readings. In this work, we aim to perform pose estimation of multiple articulated objects using a single RGB-D frame with weak initialization from pixel-wise segmentations.

Interactive perception [25] for articulated object estimation [26] has been a problem of interest in the robotics community. Various works [27]–[29], propose methods for estimating kinematic models from demonstration of manipulation or articulation examples. We instead focus on using known kinematic models to estimate the objects in challenging cluttered environments.

Li et al. [30] explore category-level localization of articulated bodies in a point cloud, however their method does not consider clutter and occlusions from the environment. Michel et al. [31] perform one-shot pose estimation of articulated bodies using 3D correspondences with optimization over hypotheses. Desingh et al. consider pose estimation of articulated objects in cluttered scenarios using efficient belief propagation [6], but do not consider RGB information. All of these approaches consider large, primarily planar objects that cover significant portion of the observation as opposed to the small objects in clutter in this work.

Li et al. [32] developed techniques to handle the challenges of hand tools and small objects with no articulation, however the techniques proposed require multi-viewpoint information, as opposed to the single image approach that we propose.

III. Problem Statement

Given a scene containing objects O, such that $\{O_k\}_{k=1}^{K}$ is the set of K relevant objects, we wish to localize each

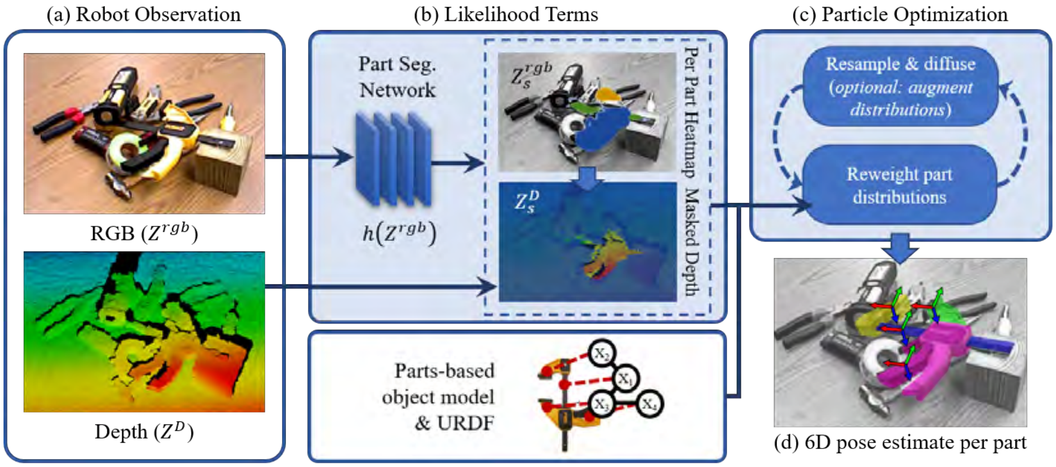

(a) Robot Observation
RGB (Z^{rgb})
Depth (Z^D)
(b) Likelihood Terms
Part Seg. Network
$h(Z^{rgb})$
Z_s^{rgb}
Z_s^D
Per Part Heatmap Masked Depth
Parts-based object model & URDF
X_1
X_2
X_3
X_4
(c) Particle Optimization
Resample & diffuse (optional: augment distributions)
Reweight part distributions
(d) 6D pose estimate per part

Fig. 3: The inference pipeline. (a) The robot observes a scene as an RGB-D image, $\mathcal{Z} = (Z^{rgb}, Z^D)$. (b) The RGB image is passed through a trained part segmentation network, $h(Z^{rgb})$, that generates a pixel-wise heatmap for the P_k parts of an object class of interest, $\{Z_s^{rgb}\}_{s=1}^{P_k}$ (in this example, the clamp, which has one fully occluded part). The heatmaps are used to generate masked depth images, $\{Z_s^D\}_{s=1}^{P_k}$ (c) The inference is initialized with part poses using these heatmaps and the depth image. Hypotheses are iteratively reweighed using Equation 4, and resampled with importance sampling. (d) The inference process generates an estimate of the 6D pose of each part. (Best viewed in color).

object O_k. The state of an object O_k is represented by the set of part poses $\mathcal{X} = \{X_i\}_{i=1}^{P_k}$, where X_s is the 6D pose of an articulating rigid part s of O_k, with P_k parts. Each object O_k in the scene is estimated independently.

This estimation problem is formulated as a Markov Random Field (MRF). Let $G = (V, E)$ denote an undirected graph with nodes V and edges E. An example MRF is illustrated in Figure 2. The joint probability of the graph G is expressed as:

$$p(\mathcal{X}, \mathcal{Z}) \propto \prod_{(s,t)\in E} \psi_{s,t}(X_s, X_t) \prod_{s\in V} \phi_s(X_s, Z_s) \qquad (1)$$

where $\mathcal{X}$ denotes the hidden state variables to be inferred and $\mathcal{Z}$ denotes the observed sensor information in the form of an RGB-D image. The function $\psi_{s,t}$ is the *pairwise potential*, describing the correspondence between part poses based on the articulation constraints, and ϕ_s is the *unary potential*, describing the correspondence of a part pose X_s with its observation Z_s. The problem of pose estimation of an articulated model O_k is interpreted as the problem of estimating the marginal distribution of each part pose, called the belief, $bel(X_s)$.

In addition to the sensor data, the articulation constraints and 3D geometry of the object, in the form of a Unified Robot Description Format (URDF), and the 3D mesh models of the objects are provided as inputs. We assume that the object articulations are produced by either fixed, prismatic or revolute joints. We consider scenes which contain only one instance of an object. In Section IV, our proposed inference mechanism is detailed, along with a description of our modelling of the potentials in Equation 1.

IV. Methodology

Belief propagation via iterative message passing is a common approach to infer hidden variables while maximizing the joint probability of a graphical model [6], [21]. We adopt the sum-product iterative message passing approach to perform inference [33], where messages are passed between hidden variables until their beliefs converge. A message, denoted by $m_{ts}^n(X_s)$, can be considered as the belief of the receiving node s as informed by its neighbor t at iteration n. An approximation of the message, denoted by $\hat{m}_{ts}^n(X_s)$, is computed using the incoming messages to t:

$$\hat{m}_{ts}^n(X_s) = \sum_{X_t\in\mathbb{X}_t} \phi_t(X_t, Z_t)\psi_{s,t}(X_s, X_t) \prod_{u\in\rho(t)\setminus s} \hat{m}_{ut}^{n-1}(X_t) \qquad (2)$$

where $\rho(t)$ denotes neighboring nodes of t, and $\mathbb{X}_t$ denotes the particle set of node t.

The marginal belief of a hidden node is a product of all the incoming messages weighted by the node's unary potential:

$$bel_s^n(X_s) \propto \phi_s(X_s, Z_s) \prod_{t\in\rho(s)} \hat{m}_{ts}^n(X_s) \qquad (3)$$

Our particle optimization algorithm aims to approximate the joint probability of the MRF, as in Equation 1, by maintaining the marginal belief, as in Equation 3 for each object part. The belief of a rigid part pose, $bel_s(X_s)$, is represented nonparametrically as a set of N weighted particles $\mathbb{X}_s = \{X_s^{(i)}, w_s^{(i)}\}_{i=1}^N$.

Section IV-A describes the message passing algorithm. Section IV-B describes how the function $\phi_s(X_s, Z_s)$ is represented, and Section IV-C describes the function $\psi_{s,t}(X_s, X_t)$.

A. Belief Propagation via Message Passing

Our method adopts the traditional reweight and resample paradigm for particle refinement methods. The particles are first *reweighted* using an approximated sum-product message. The particles are then *resampled* using importance sampling based on the calculated weights.

The high-dimensional nature of the estimation problem and the cluttered settings with similar parts and partial observations make the inference prone to convergence to local minima. To mitigate this problem while computing messages, we can optionally add an augmentation step before the reweight step to accommodate different proposals. The augmentation technique is adapted from Pacheco et al. [34] and is discussed in Section IV-D.

The overall system is summarized in Figure 3.

1) Reweighting and Resampling steps: Each particle $X_s^{(i)} \in \mathbb{X}_s$ is reweighted as follows:

$$w_s^{(i)} = \phi_s(X_s^{(i)}, Z_s) \prod_{t \in \rho(s)} \hat{m}_{ts}^n(X_s^{(i)}) \tag{4}$$

where $\hat{m}_{ts}^n(X_s^{(i)})$ is the sum-product message:

$$\hat{m}_{ts}^n(X_s^{(i)}) = \sum_{X_t^{(j)} \in \mathbb{X}_t} \psi_{s,t}(X_s^{(i)}, X_t^{(j)}) \phi_t(X_t^{(j)}, Z_t) \tag{5}$$

which only takes into account the immediate neighbors of the node. Since the number of parts in each object is small, this approximation has negligible effect in practice and saves computation time. For numerical stability, the log-likelihoods are used in practice. The weights are normalized and then the particles are resampled using importance sampling. The object pose estimate is made by selecting the maximum likelihood estimates (MLE) from each of the marginal beliefs.

B. Unary Likelihood

The unary potential represents the compatibility of each pose hypothesis with the RGB-D observation, $\mathcal{Z}$. The RGB and depth portions of the observation, Z^{rgb} and Z^D, are treated as independent such that the unary likelihood is:

$$\phi_s(X_s, Z_s) = \phi_s^{\text{rgb}}(X_s, Z_s^{rgb}) \phi_s^D(X_s, Z_s^D) \tag{6}$$

where ϕ_s^D and ϕ_s^{rgb} are the likelihoods with respect to depth and RGB parts of the observations.

1) RGB Unary Likelihood: The RGB portion of the unary likelihood makes use of the Dilated ResNets architecture [35]. This architecture maintains a high dimensional feature space which is beneficial for semantic segmentation tasks.

The CNN outputs a pixelwise score for each object part class s. We apply a sigmoid function so the final scores lie between zero and one. This constitutes a learned heatmap $Z_s^{rgb} = h_s(Z^{rgb})$ over an RGB observation Z^{rgb}, where h is the Dilated ResNets model trained on parts and $h_s(\cdot)$ is the output indexed at class s. For each particle hypothesis X_s, we generate a mask $\mathcal{M}_s$ over the image for the object part at the hypothesis pose. We transform the mesh model of the part to pose X_s and use the camera parameters to obtain a corresponding binary mask in image space. We represent the likelihood of a particle X_s over the heatmap Z_s^{rgb} using the Jaccard index [36], commonly called the Intersection over Union (IoU), between the heatmap and the rendered mask:

$$\phi_s^{\text{rgb}}(X_s, Z_s^{rgb}) = \frac{|\mathcal{M}_s \cap Z_s^{rgb}|}{|\mathcal{M}_s| + |Z_s^{rgb}| - |\mathcal{M}_s \cap Z_s^{rgb}|} \tag{7}$$

The CNN is trained using the analagous intersection over union (IoU) loss, and as such, $\phi_s^{\text{rgb}}(X_s)$ represents a learned likelihood function over the image.

2) Depth Unary Likelihood: For a given part s, depth observation Z_s^D is generated using a threshold over the heatmap Z_s^{rgb} to mask the depth image Z^D. For a particle X_s, $\phi^D(X_s, Z_s^D)$ is the exponential of the negative average pixelwise error between Z_s^D and the mesh model of part s, rendered at pose X_s. The error is only evaluated over areas in which the two depth images overlap. If there is no overlap between the masked observation and the hypothesis, we assign a maximum error instead, which is a chosen constant.

C. Pairwise Likelihood

The pairwise likelihood between neighbouring particles $\psi_{t,s}(X_t, X_s)$ measures how compatible X_s is with respect to X_t. If X_s falls within the joint limits of s with respect to t at pose X_t, then $\psi_{t,s}(X_t, X_s) = 1$. Otherwise, the likelihood is the exponential of the negative error between X_s and the nearest joint limit. We refer to [6] for further details.

D. Particle Augmentation

At each node s, the particle set $\mathbb{X}_s$ can be augmented by drawing particles from various proposal distributions. Given N particles in $\mathbb{X}_s$, Gaussian noise is first added to the current particles, then the distribution is augmented to $\mathbb{X}_s^{prop} = \mathbb{X}_s \cup \mathbb{X}_s^{aug}$, where $\mathbb{X}_s^{aug}$ represents the particles generated from the augmentation procedure q. The set $\mathbb{X}_s^{prop}$ contains αN particles, where $\alpha > 1$. Various proposals q_s^{pair}, q_s^{unary}, and q_s^{rand}, as described below can be used to augment the particle set. This optional variant is evaluated and discussed in the results section.

Pairwise: The pairwise proposal distribution $q_s^{pair}(X_s) \propto \psi_{s,t}(X_s, \tilde{X}_t)$ is conditioned on a sample $\tilde{X}_t$, drawn from neighboring node t. Using the known geometric relationship between nodes t and s, a compatible proposal for node s, $\tilde{X}_s$, is generated from $\tilde{X}_t$.

Unary: The unary proposal distribution $q_s^{unary}(X_s) \propto \phi_s(X_s, Z_s)$ draws samples based on the unary potential ϕ_s.

Random: The random proposal distribution $q_s^{rand}(X_s) \propto \mathcal{N}(X_s, \Sigma)$ draws additional noisy samples. This can be used to avoid the belief falling into a local minima due to the high dimensionality of the orientation space, and to account for mirror symmetry in some objects.

V. Experiments

We evaluate our methods for articulated object localization in uncluttered and cluttered scenes. We run experiments on each component of our method and provide an analysis of their effects. These results provide quantitative and qualitative evidence of the accuracy and practicality of our methods.

We test on 20 uncluttered and 17 cluttered test scenes, unseen in the training data. We localize 196 total object instances in these scenes. We do not include results on objects which are severely or fully occluded such that there is no clear observation of any part. We remove 19 objects which fall into this category. An example of such a case is shown in the highly cluttered scene in Figure 8a, where the flashlight (behind the hammer) and lineman's pliers (behind the clamp) are almost entirely occluded.

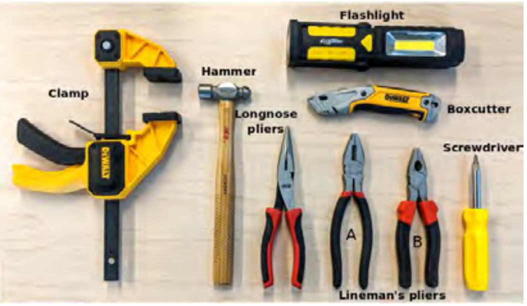


Fig. 4: Objects in our custom dataset. The clamp has one prismatic joint. The three sets of pliers each have one revolute joint. The other four objects are treated as rigid. The objects are separated into parts based on the presence of both affordances and articulations.

A. Dataset & Training

Our custom dataset consists of hand tools with eight distinct tool instances: hammer, clamp, boxcutter, flashlight, screwdriver, longnose pliers, and two instances of lineman's pliers (see Figure 4). We collect videos of both cluttered and uncluttered scenes using the Fetch Mobile Manipulator's onboard Primesense Carmine 1.09 sensor. The articulated hand tools span the full range of possible articulations in the data. Semantic masks and 6D poses for the objects are labelled using Label Fusion [37], which generates annotations for each video once the first scene is manually labelled. Semantic part masks and part poses are calculated using the object URDFs. The pixels in the images which do not correspond to a tool part are given class label "background." After downsampling to remove adjacent frames in the videos, the dataset contains $\sim 6k$ RGB-D images of 640×480 pixel resolution.

We train the Dilated ResNets *DRN-D-22* architecture [35] to perform semantic segmentation on 90% of the dataset, and reserve 10% for validation. We further augment the training images with random crops, flips, and rotations. We increase the training set size by applying two transforms per image. The backbone is pre-trained on ImageNet, and the last layers are finetuned on our dataset. We employ the Intersection over Union (IoU) loss with an Adam optimizer. We train for 10 epochs on a RTX 2080 Max-Q GPU.

B. Implementation Details

Our implementation performs efficient unary potential computation on the GPU, to evaluate the heatmap from DRN and to generate binary masks and depth images for pose hypotheses. The current implementation is vectorized and processes all object parts in $\sim 0.5 - 2s$ for one iteration with 300 particles. The computation time could be further reduced with more efficient implementation.

The x and y locations of particle poses are initialized randomly in areas corresponding to high heat pixels of the heatmap over the RGB observation. The z-axis is initialized to the corresponding depth in the observed depth image. The initial orientations are uniformly distributed. For completely occluded parts which do not appear on the segmentation mask, we generate compatible poses from the neighbour initializations.

C. Evaluation Metric

For evaluation, we use the average point matching error proposed by Hinterstoisser et al. [38], which measures the average point pairwise distance between the rigid object model's point cloud in the ground truth and estimated poses:

$$m(P_{gt}, \hat{P}) = \frac{1}{\mathbb{N}} \sum_{(p_{gt}, \hat{p}) \in (P_{gt}, \hat{P})} ||\hat{p} - p_{gt}|| \tag{8}$$

where $(p_{gt}, \hat{p}) \in (P_{gt}, \hat{P})$ are corresponding points in the ground truth and estimated point clouds respectively, each with $\mathbb{N}$ points in the rigid object model. We also report the symmetric point matching error, which measures the average pairwise distance between points in the estimated point cloud and the *nearest* point in the ground truth point cloud:

$$m_{\text{sym}}(P_{gt}, \hat{P}) = \frac{1}{\mathbb{N}} \sum_{\hat{p} \in \hat{P}} \min_{p_{gt} \in P_{gt}} ||\hat{p} - p_{gt}|| \tag{9}$$

The symmetric matching error represents the error in symmetric objects, such as the screwdriver, better by not penalizing estimates rotated around a degree of symmetry in the object. However, it tends to provide artificially low errors for incorrect estimates.

D. Baselines

We implement two baselines, described below.

Segmentation with ICP (DRN+ICP): We initialize the 3D position of the particle hypotheses using the depth image and segmentation mask generated by Dilated ResNets, with random orientations. We use Iterative Closest Point (ICP) [7] to find the transform from the initialized point cloud to the observed point cloud. ICP works best on local refinements, and is prone to failure when the initial orientation is incorrect. To accommodate for this failure, we generate N proposal poses per part, perform ICP, and select the one with the best final fitness score as the estimate. In our experiments, $N = 20$. A similar method is used by Wong et al. [39].

Part-based Particle Filter (Parts-PF): This baseline consists of independent particle filters at each tool part. We use the unary potential from Equation 6 to calculate the weights for each hypothesis, and use importance sampling to select particles at each iteration. We use 300 particles per part, and run for 85 iterations.

For both *DRN+ICP* and *Parts-PF* baselines, if a part is completely occluded in the image, a pose estimate cannot be generated from the segmentation. In such cases, we randomly select a neighboring part for which an estimate was made and use the object model to generate a corresponding pose for the occluded part. If the edge between the parts is articulated, a joint value is uniformly sampled within the joint limits.

E. Parts-Based Pose Estimation

To fully understand the performance of our method, we perform an ablation study over the components of the proposed method. We focus on three factors: the message passing (*MP*), the use of *RGB* only vs. the inclusion of depth (*RGB-D*) in the unary potential, and the augmentation step (*Aug*). We use 300 particles (before augmentation)

TABLE I: Average matching errors m (cm) and symmetric matching errors m_{sym} (cm)

Method	**Cluttered**		**Uncluttered**		**Overall**	
	m	m_{sym}	m	m_{sym}	m	m_{sym}
DRN+ICP	6.59 ± 3.36	3.31 ± 2.69	6.38 ± 8.00	3.93 ± 10.84	6.47 ± 6.32	3.65 ± 8.22
Parts-PF	6.03 ± 4.64	2.73 ± 4.04	4.55 ± 2.57	1.43 ± 1.33	5.23 ± 3.74	2.03 ± 2.98
MP+RGB	4.57 ± 3.96	2.83 ± 2.95	3.60 ± 2.75	2.04 ± 1.95	4.06 ± 3.40	2.41 ± 2.50
MP+RGB+ICP	4.77 ± 3.89	3.03 ± 2.88	3.85 ± 2.65	2.29 ± 1.90	4.28 ± 3.32	2.64 ± 2.44
MP+RGB-D+Aug	4.80 ± 3.80	2.90 ± 2.60	3.71 ± 2.55	2.18 ± 1.88	4.21 ± 3.23	2.51 ± 2.27
MP+RGB-D	**3.58** ± 2.56	**1.89** ± 1.84	**2.65** ± 2.01	**1.30** ± 1.21	**3.08** ± 2.32	**1.57** ± 1.56

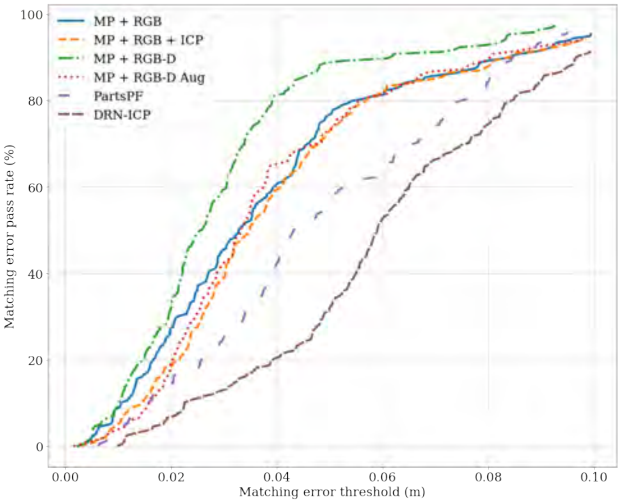
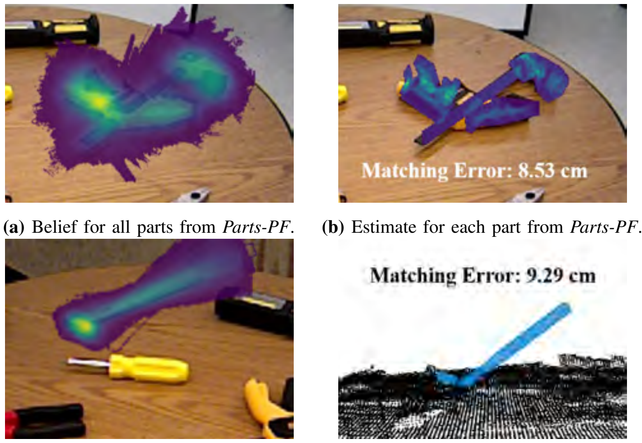


Fig. 5: Average pairwise distance pass rate for each described method. All tools use the matching error described in Equation 8, except the screwdriver, which due to its symmetrical nature, uses the symmetrical form (m_{sym}).



(a) Belief for all parts from *Parts-PF*. **(b)** Estimate for each part from *Parts-PF*.

(c) Belief for all parts from *MP+RGB* **(d)** Estimate for each part from *MP+RGB*

Fig. 6: Common failure cases. The first row is a common failure of the baseline method, *Parts-PF*, which does not use message passing. In (a) the parts are clustered in the correct positions in the image, but (b) shows that the part estimates are oriented incorrectly in an incompatible configuration. The second row is a common failure case in *MP+RGB*. In (c), the particles have converged in the 2D image, but in (d), the hammer 6D pose is rotated about an axis which is not well represented in the 2D image plane.

for all experiments. The choice of particles was observed qualitatively to achieve sufficient results in most cases. While representation of the underlying belief improves with more particles, computation becomes intractable for very large particle sets. We run each method for 100 iterations, after which we observe little change in the estimate.

The results of each method are shown in Table I. Figure 5 shows the results for all scenes. Message passing leads to superior results compared to the baselines, *DRN-ICP* and *Parts-PF*, which do not use message passing. Best performance is achieved by using the full RGB-D observation (*MP-RGB-D*). Further description and analysis of each method is provided below.

Message Passing: RGB Unary (MP+RGB): To test the effect of the depth component of the unary potential, we evaluate using only the RGB component, informed by the heatmap, such that Equation 6 becomes $\phi_s(X_s, Z_s) = \phi_s^{\text{rgb}}(X_s, Z_s^{rgb})$. We use message passing to calculate the final likelihood for each particle using the pairwise potential, as described in Equation 4. Using only the RGB image, we obtain lower accuracy on the pose estimates. The RGB image captures the position and orientation of the objects well *in image space* (see Figure 6(c)), but is prone to falling into local minima in the axes which are not well represented by the image, namely z, pitch, and roll (see Figure 6(d)).

Message Passing: RGB Unary and ICP (MP+RGB+ICP): To attempt to recover from the errors in z, pitch, and roll, we add an ICP step on the final estimate from *MP+RGB* to align it to the masked depth image. We estimate the offset in the z-axis based on the depth image. Since ICP is a local refinement method which relies on an accurate estimation of the initial transform, the ICP step does not always reconcile the orientation error in pitch and roll, for which the initial transform is unknown.

Message Passing: RGB-D Unary (MP+RGB-D): We hypothesize that by including depth information in the unary potential, we can more reliably estimate the full 6D pose of the parts and make up for missing information in the 2D image. The depth term improves the estimation accuracy by discouraging all unoccluded particles from deviating from the depth image at each iteration. This performs better than *MP+RGB+ICP* because the latter only attempts to align to the depth image in the final iteration, where it often has converged to a local minimum. This is the best performing method. Selected qualitative results are shown in Figures 7 and 8.

Message Passing with Augmentation (MP+RGB-D+Aug): Using the unary informed by both RGB and depth, as well as message passing, we augment the particle set at the beginning of each iteration as described in Section IV-A. We use $\alpha = 1.5$, with 5% of the additional particles drawn from the unary distribution. At the first iteration, the

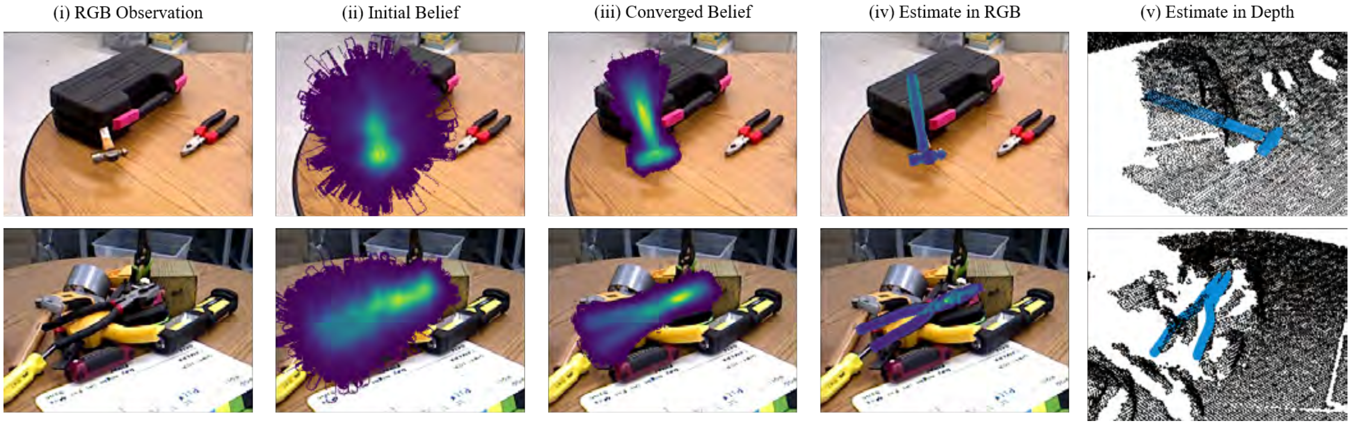
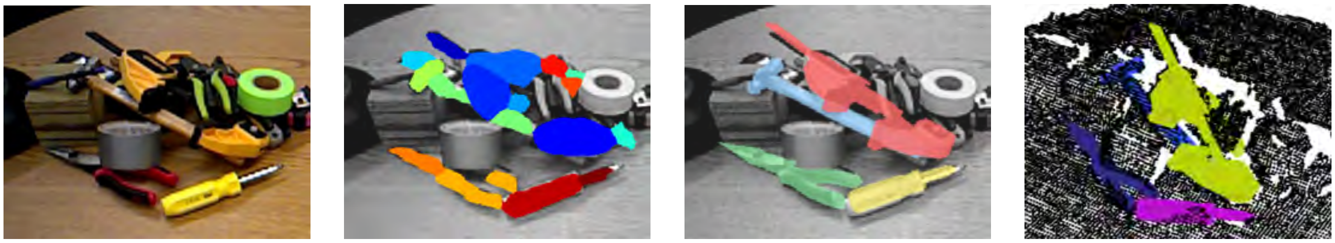


Fig. 7: Qualitative results for each stage of the *MP+RGB-D* method on cluttered scenes. The method results in accurate pose estimates despite significant occlusions (top) and articulations (bottom).



(a) RGB input image

(b) Segmentation mask from CNN

(c) Localized objects in RGB image

(d) Localized objects in depth point cloud

Fig. 8: Qualitative results for localization of each object in a cluttered scene using the *MP + RGB-D* method. Although the segmentation map (b) is missing information for occluded parts, our iterative metho is able to recover the 6D pose of the parts (c), (d).

remaining 95% of the particles are drawn from q^{rand}. The percentage of particles drawn from q^{pair} is increased by 10% every 5 iterations, and the percentage from q^{rand} is decreased, up to a maximum of 90% of particles from q^{pair}.

Qualitatively, we observe that the augmentation step leads to quicker convergence in some highly cluttered scenarios. On average, this method performs worse than *MP+RGB-D* in some cases because it is susceptible to propagating incorrect estimates with artificially inflated pairwise scores, due to the addition of perfectly compatible pose estimates. However, these results depend on careful selection of parameters, and might be improved by further tuning. Further analysis of the effect of the parameters on the final estimate is left to future work.

F. Analysis on Tool Classes

The results for the *MP+RGB-D* and *MP+RGB-D+Aug* methods for each object in the dataset are shown in Table II. We present the percentage of class indices which have error under 4 cm. We observed high error in the flashlight which is likely due to its symmetrical nature. The unary potential does not explicitly encode texture information, so geometrically symmetric parts can tend to flip. The clamp is among the most difficult objects to localize due to its high-dimensionality and significant self-occlusions.

G. Qualitative Analysis

Selected examples of the *MP+RGB-D* method are shown in Figure 7. We show a selected scene which demonstrate the effectiveness of the *MP+RGB-D* method at localizing hand tools, even under partial or full occlusion of some of their parts, in Figure 8. While the heatmap may provide little to no information for some parts, by leveraging geometric information through message passing, we are able to resolve the pose of all the visible tools in the scene. More qualitative results are available in the attached video.

TABLE II: Fraction of tools with matching error less than 4 cm. All objects are evaluated with the average pairwise matching error m, except the screwdriver which uses the symmetric matching error, m_{sym}.

Class	*MP+RGB-D*	*MP+RGB-D+Aug*
clamp	0.677	0.387
hammer	0.920	0.400
longnose pliers	1.000	0.938
lineman's pliers A	0.909	0.818
lineman's pliers B	0.917	0.833
boxcutter	0.833	0.750
flashlight	0.640	0.640
screwdriver (m_{sym})	*0.759*	*0.689*

VI. Conclusion

In this work, we present an inference technique for estimating articulated parts-based object pose in clutter. We model the part poses of each articulated object as a Markov Random Field (MRF) and perform efficient particle-based belief propagation. We use articulation constraints between parts and a novel learned likelihood function to perform message passing in the MRF. We perform a thorough analysis of our method and show that it performs well on both uncluttered and cluttered scenes. We demonstrate that the message passing step is highly beneficial in terms of enforcing geometric consistency to inform pose estimation in the high dimensional space of 6D articulated object pose.

ACKNOWLEDGEMENTS

We would like to thank Zhen Zeng for helpful discussions and feedback, and Adrian Röfer and Zhiming Ruan for creating the 3D tool models. Zhiming Ruan also contributed to initial software development.

REFERENCES

[1] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6D object pose estimation in cluttered scenes," in *Robotics: Science and Systems (RSS)*, 2018.

[2] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Conference on Robot Learning (CoRL)*, 2018.

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

[4] X. Deng, A. Mousavian, Y. Xiang, F. Xia, T. Bretl, and D. Fox, "PoseRBPF: A Rao-Blackwellized particle filter for 6D object pose tracking," *Proceedings of Robotics: Science and Systems*, 2019.

[5] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, pp. 1627–1645, 2009.

[6] K. Desingh, S. Lu, A. Opipari, and O. C. Jenkins, "Efficient nonparametric belief propagation for pose estimation and manipulation of articulated objects," *Science Robotics*, vol. 4, no. 30, 2019.

[7] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. International Society for Optics and Photonics, 1992, pp. 586–606.

[8] K. Desingh, O. C. Jenkins, L. Reveret, and Z. Sui, "Physically plausible scene estimation for manipulation in clutter," in *International Conference on Humanoid Robots (Humanoids)*. IEEE, 2016, pp. 1073–1080.

[9] V. Narayanan and M. Likhachev, "Deliberative object pose estimation in clutter," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 3125–3130.

[10] Z. Sui, Z. Zhou, Z. Zeng, and O. C. Jenkins, "Sum: Sequential scene understanding and manipulation," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3281–3288.

[11] X. Chen, R. Chen, Z. Sui, Z. Ye, Y. Liu, R. I. Bahar, and O. C. Jenkins, "GRIP: Generative robust inference and perception for semantic robot manipulation in adversarial environments," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 3988–3995.

[12] Z. Zeng, Y. Zhou, O. C. Jenkins, and K. Desingh, "Semantic mapping with simultaneous object detection and localization," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 911–918.

[13] C. Mitash, A. Boularias, and K. Bekris, "Robust 6d object pose estimation with stochastic congruent sets," *arXiv preprint arXiv:1805.06324*, 2018.

[14] C. Wang, D. Xu, Y. Zhu, R. Martín-Martín, C. Lu, L. Fei-Fei, and S. Savarese, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019, pp. 3343–3352.

[15] P. F. Felzenszwalb and D. P. Huttenlocher, "Pictorial structures for object recognition," *International journal of computer vision*, vol. 61, no. 1, pp. 55–79, 2005.

[16] Y. Xiang and S. Savarese, "Estimating the aspect layout of object categories," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2012, pp. 3410–3417.

[17] K. Mo, S. Zhu, A. X. Chang, L. Yi, S. Tripathi, L. J. Guibas, and H. Su, "PartNet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019.

[18] C. Lu, H. Su, Y. Li, Y. Lu, L. Yi, C. K. Tang, and L. J. Guibas, "Beyond holistic object recognition: Enriching image understanding with part states," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 6955–6963.

[19] L. Yi, H. Huang, D. Liu, E. Kalogerakis, H. Su, and L. Guibas, "Deep part induction from articulated object pairs," *SIGGRAPH Asia*, 2018.

[20] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard, "Tracking loose-limbed people," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004, pp. 421–428.

[21] E. B. Sudderth, M. I. Mandel, W. T. Freeman, and A. S. Willsky, "Visual hand tracking using nonparametric belief propagation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04)*, 2004, pp. 189–189.

[22] C. G. Cifuentes, J. Issac, M. Wüthrich, S. Schaal, and J. Bohg, "Probabilistic articulated real-time tracking for robot manipulation," *Robotics and Automation Letters*, vol. 2, no. 2, pp. 577–584, 2016.

[23] T. Schmidt, R. A. Newcombe, and D. Fox, "DART: dense articulated real-time tracking," in *Robotics: Science and Systems (RSS)*, 2014.

[24] T. Schmidt, K. Hertkorn, R. Newcombe, Z. Marton, M. Suppa, and D. Fox, "Depth-based tracking with physical constraints for robot manipulation," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 119–126.

[25] J. Bohg, K. Hausman, B. Sankaran, O. Brock, D. Kragic, S. Schaal, and G. S. Sukhatme, "Interactive perception: Leveraging action in perception and perception in action," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1273–1291, Dec 2017.

[26] K. Hausman, S. Niekum, S. Osentoski, and G. S. Sukhatme, "Active articulation model estimation through interactive perception," in *International Conference on Robotics and Automation (ICRA)*. IEEE, May 2015, pp. 3305–3312.

[27] R. M. Martin and O. Brock, "Online interactive perception of articulated objects with multi-level recursive estimation based on task-specific priors," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2014, pp. 2494–2501.

[28] J. Sturm, C. Stachniss, and W. Burgard, "A probabilistic framework for learning kinematic models of articulated objects," *Journal of Artificial Intelligence Research*, vol. 41, pp. 477–526, 2011.

[29] J. Sturm, *Approaches to Probabilistic Model Learning for Mobile Manipulation Robots*, ser. Springer Tracts in Advanced Robotics (STAR). Springer, 2013.

[30] X. Li, H. Wang, L. Yi, L. Guibas, A. L. Abbott, and S. Song, "Category-level articulated object pose estimation," in *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020.

[31] F. Michel, A. Krull, E. Brachmann, M. Y. Yang, S. Gumhold, and C. Rother, "Pose estimation of kinematic chain instances via object coordinate regression." in *BMVC*, 2015, pp. 181–1.

[32] C. Li, J. Bohren, E. Carlson, and G. D. Hager, "Hierarchical semantic parsing for object pose estimation in densely cluttered scenes," in *International conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 5068–5075.

[33] M. J. Wainwright, M. I. Jordan *et al.*, "Graphical models, exponential families, and variational inference," *Foundations and Trends® in Machine Learning*, vol. 1, no. 1–2, pp. 1–305, 2008.

[34] J. Pacheco, S. Zuffi, M. Black, and E. Sudderth, "Preserving modes and messages via diverse particle selection," in *International Conference on Machine Learning (ICML)*, 2014, pp. 1152–1160.

[35] F. Yu, V. Koltun, and T. Funkhouser, "Dilated residual networks," in *Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017.

[36] R. Moulton and Y. Jiang, "Maximally consistent sampling and the jaccard index of probability distributions," in *International Conference on Data Mining (ICDM)*. IEEE, 2018, pp. 347–356.

[37] P. Marion, P. R. Florence, L. Manuelli, and R. Tedrake, "Label fusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes," in *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3325–3242.

[38] 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 *Computer Vision – ACCV 2012*, K. M. Lee, Y. Matsushita, J. M. Rehg, and Z. Hu, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 548–562.

[39] J. M. Wong, V. Kee, T. Le, S. Wagner, G.-L. Mariottini, A. Schneider, L. Hamilton, R. Chipalkatty, M. Hebert, D. M. Johnson *et al.*, "SegICP: Integrated deep semantic segmentation and pose estimation," in *International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5784–5789.