Visual and Physical Plausibility of Object Poses for Robotic Scene Understanding

Dominik Bauer, Timothy Patten, Markus Vincze

Abstract

Humans use the relations between objects in a scene to determine how they may interact with, grasp and manipulate them. For robots, such an object-based scene understanding not only allows interaction with objects but also allows humans to interpret the robot's perception and actions. To gain a higher-level understanding of an observed scene, knowledge of the objects' poses is crucial. The poses, when combined with 3D models of the objects, allow for easy derivation of the interactions between objects, enabling reasoning about occlusion, collisions, support and, finally, manipulation by the robot. However, most related work does not consider scene-level object interactions but rather focuses on finding the pose of a single object in a given frame. Object interactions are considered only to augment training data or in post hoc verification steps. In contrast, we show that such scene-level information should be exploited during the estimation of the object poses themselves. Our main assumption is that all object hypotheses need to be plausible in terms of their visual observation and the physical scene in which they exist. In this chapter, we present our work on investigating the exploitation of this visual and physical plausibility for robust, accurate estimation and understandable explanation of object poses.

Keywords

robot vision, object pose estimation, object pose refinement, hypothesis verification, explainability

1 Introduction

The ability of a robot to explain its actions – or reasons why it might have failed – is an important building block for establishing and maintaining human trust [Lomas et al. 2012; de Graaf and Malle 2017; de Graaf et al. 2018]. For example, interactive explanations are an effective way to gain a deeper understanding of the reasoning provided [Dunne et al. 2005; Walton 2007; Arioua et al. 2017; Madumal et al. 2019]. But to provide such interactive explanations, the robot must attain a thorough understanding of the scene it inhabits. This may include the scene's objects, their location and their relationship to one another, for example expressed as their class, pose and spatial relations, respectively [Naseer et al. 2018]. Moreover, such an understanding enables the robot to perform tasks, such as grasping and manipulating objects, in the first place [Srinivasa et al. 2010; Chitta et al. 2012; Tremblay et al. 2018].

We hypothesize that, for the robot to provide an effective explanation of its understanding of a scene and its interactions with it, it must resolve to human-understandable reasoning approaches, such as how well the robot's understanding visually aligns with its camera images or how physically plausible an object's pose would be in a simulation of its estimated scene. We conjecture that both the visual and physical plausibility of the robot's scene understanding must be jointly considered and we examine their application to the object pose estimation task.

https://doi.org/10.34727/2022/isbn.978-3-85448-052-5_4

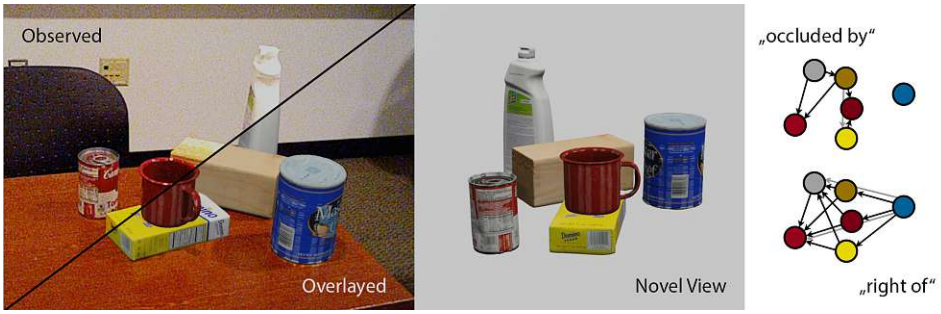


Figure 1 Applications of an object-based scene understanding. Left: Rendering the objects' models under their estimated poses allows to overlay and compare the robot's perception to the observed scene. Mid: Similarly, a novel view of the observed scene may be rendered. Right: Using the estimated poses, also the relations between the objects in the scene may be derived.

The poses, when combined with 3D models, allow the robot to manipulate the scene and explain it in terms of objects and their relations as illustrated in Figure 1.

This chapter provides an overview of our work exploring these hypotheses. In Section 2, we define visual plausibility through rendering and physical plausibility through simulation or evaluation of the static equilibrium. We present two different approaches for exploiting plausibility in object pose estimation. The methods we propose in Section 3 only require the 3D models of the objects and augment existing pose refiners. In Section 4, we propose novel object pose refinement methods based on reinforcement learning. These methods may jointly consider both aspects of plausibility that are discussed in this chapter. In Section 5, we present reasoning strategies that exploit this information for explanations in human-robot interaction. Finally, in Section 6, we discuss our findings and draw conclusions for future work.

2 Defining Visual and Physical Plausibility of Object Poses

A scene understanding represented by (semantically annotated) 3D models and their object poses allows to derive information about the scene that can be used for explanation and improvement of the poses themselves. For example, spatial relations between objects may be derived or a rendering of the estimated scene may be compared to the robot's camera image, as shown in Figure 1. Furthermore, the latter allows a robot to determine the plausibility of its scene understanding and subsequently explain why its actions might have succeeded or failed.

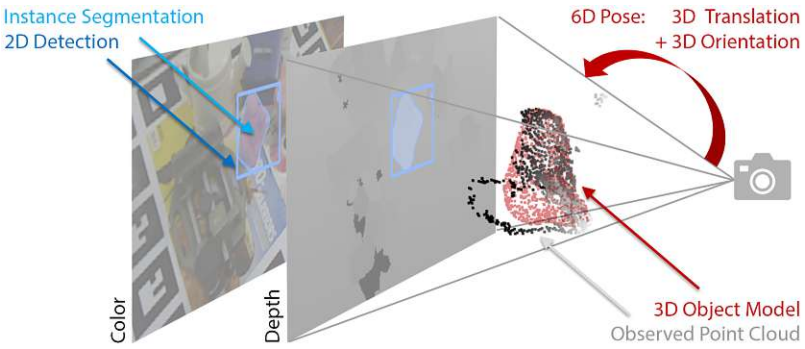


Figure 2 An object pose estimation pipeline. Left: A known object of interest is detected in the observed image. Mid: Using the instance segmentation mask, a cloud of all points predicted to belong to the object is generated from the corresponding depth image. Right: The task is to determine the 6D pose of the 3D model of the object such that it aligns to the observed image or point cloud.

The task of object pose estimation is to find the transformation T that aligns a 3D model of the object with its observation, as illustrated in Figure 2. We need to estimate this transformation by $\hat{T} = [\hat{R} \in SO(3), \hat{t} \in \mathbb{R}^3]$, i.e., a rotation $\hat{R}$ and a translation $\hat{t}$.

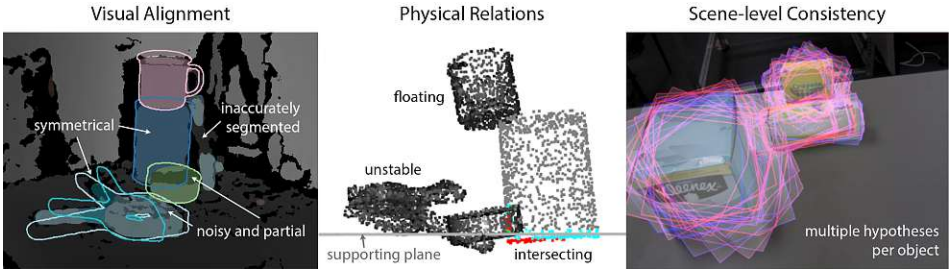


Figure 3 Challenges in object pose estimation. Left: Limited visibility, noise and inaccurate segmentation result in inaccurate pose estimates. Mid: The physical object relations in the estimated scenes violate the assumptions of plausible, static scenes. Right: Considering all scene-level interactions of multiple object under multiple (inaccurate) pose hypotheses quickly grows intractable.

The observation may be in the form of RGB or depth images. It is therefore only a partial and noise afflicted view of the object due to limited visibility from a single view and sensor limitations, as shown in Figure 3 (left). This problem is exacerbated in cluttered scenes and affects all parts of the perception pipeline – from detection, to segmentation and pose estimation. As a result, we might end up with multiple inaccurate pose hypotheses, as illustrated in Figure 3 (right). On the

one hand, to prevent failure, we want to verify and select the best available object pose before executing any robotic actions. On the other hand, we want to be able to explain why the robot selects a certain pose or why it decides that the pose is sufficiently accurate to base its interactions on it. In this section, we propose two approaches to this, based on visual alignment and physical plausibility.

2.1 Rendering-based Visual Plausibility

Object pose estimation and evaluation thereof are commonly based on the alignment of a 3D object model [Hodaň et al. 2020]. The Average Distance of Model Points (ADD) [Hinterstoisser et al. 2012] is the most used metric in related work. It measures the mean distance between corresponding model points $x \in X$ under estimated pose $\hat{T}$ and ground-truth pose T, or formally

$$ADD = \text{avg}_{x \in X} ||\hat{T}x - Tx||_2. \tag{1}$$

In contrast, the Visual Surface Discrepancy (VSD) [Hodaň et al. 2016, 2018], considers the discrepancy between the rendered depth images of the object under estimated pose $\hat{I}_d(\hat{T})$ and ground-truth pose $\hat{I}_d(T)$ by

$$VSD = \text{avg}_{p \in V(\hat{T}) \cup V(T)} \begin{cases} 0, & \text{if } p \in V(\hat{T}) \cap V(T) \text{ and } \Delta(p) < \tau, \\ 1, & \text{otherwise.} \end{cases} \tag{2}$$

The visibility under a given pose V is computed with respect to the observed depth image I_d and $\Delta(p)$ is the absolute difference between the rendered images at a pixel p.

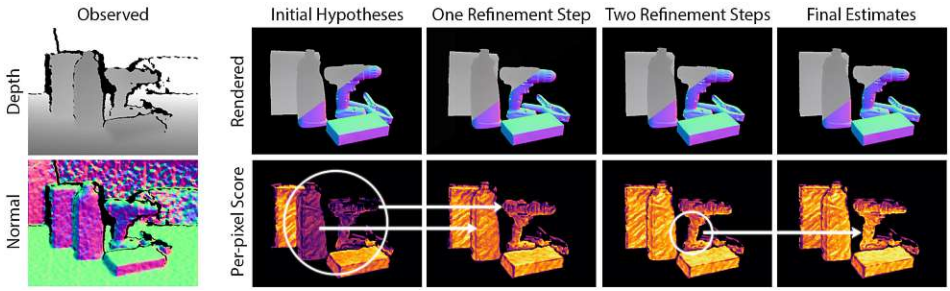


Figure 4 Example of the visual-alignment score. The observed depth and surface normals (left) are compared to the rendered objects under estimated pose (right). The resulting score for different sets of pose hypotheses (columns) is visualized below, where a more yellow color indicates better alignment with the observation. Adapted from [Bauer et al. 2022].

When estimating the pose of an object, the ground-truth pose is unknown and thus these metrics cannot be used to measure the quality of the pose estimate. Building on the idea of VSD, however, we suggest that the rendered view of a scene should be compared to the *observation* (i.e., the robot's camera view), as it can be considered a noisy version of the rendered object under the ground-truth pose T. If both align, we consider the estimate to be visually plausible. We define the visual-alignment score $\bar{a}$ in [Bauer et al. 2020c] that quantifies the average alignment between the object in the observed and rendered depth and normal images under the estimated pose $\hat{T}$. As illustrated in Figure 4, $\bar{a}$ is computed over all pixels with valid depth values, defined as $V = I_d > 0 \cup \hat{I}_d(\hat{T}) > 0$, by

$$\bar{a} = \frac{1}{2}\left(\text{avg}_{p \in V} a_d(p) + \text{avg}_{p \in V} a_n(p)\right), \tag{3}$$

with depth-based alignment a_d and normal-based alignment a_n per pixel p defined as

$$a_d(p) = \begin{cases} 1 - \frac{|d-\hat{d}|}{\tau}, & \text{if } |d-\hat{d}| < \tau \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

$$a_n(p) = \begin{cases} 1 - \frac{1-n\cdot\hat{n}}{\alpha}, & \text{if } 1 - n\cdot\hat{n} < \alpha \\ 0, & \text{otherwise,} \end{cases} \tag{5}$$

where $d \in I_d$ is the depth value and $n \in I_n$ is the corresponding normal at pixel p in the observation. The corresponding values in the rendered image are denoted by $\hat{d} \in \hat{I}_d(\hat{T})$ and $\hat{n} \in \hat{I}_n(\hat{T})$. The parameters τ and α limit the maximal admissible discrepancy.

2.2 Contact- and Simulation-based Physical Plausibility

Visual alignment alone may result in ambiguity under partial observability. We suggest that physical plausibility is able to resolve visually ambiguous cases. We define the physical plausibility of a scene as the combination of feasibility (non-intersecting, non-floating) and static stability of the objects therein, as illustrated in Figure 5.

Contact-based Formulation: We define these conditions based on two sets of critical points in [Bauer et al. 2020a], the intersecting points $\mathcal{I}$ and the contact points $\mathcal{C}$. These point sets depend on the signed distance δ between the object of interest and the scene. δ is computed for uniformly random sampled points $\hat{X}$ on the surface of the model X under an estimated pose $\hat{T}$. We compute these point sets with respect to a slack variable ε, accounting for inaccuracy due to the mesh representation and random sampling.

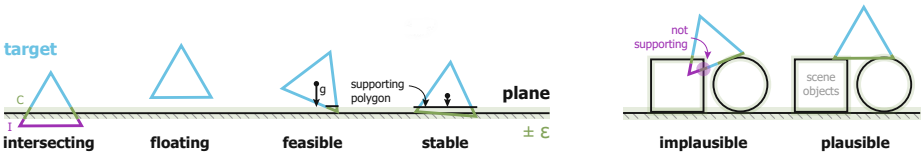


Figure 5 Definition of physical plausibility based on critical points for a single object (left) and a scene (right). If feasible, the center of mass projected in gravity direction must intersect the support polygon (convex hull of supported points) to be considered stable. Reprinted from [Bauer et al. 2022].

Intersecting points lie inside the scene objects' surface and contact points are within a small distance from them. Formally, we define

$$\mathcal{I} = \{\hat{x} \in \hat{X} : \delta(\hat{x}) < -\varepsilon\}, \tag{6}$$

$$\mathcal{C} = \{\hat{x} \in \hat{X} : |\delta(\hat{x})| < \varepsilon\}. \tag{7}$$

Based on these point sets we define an object to be

$$\text{not floating, if} \quad |\mathcal{C}| > 0, \tag{8}$$

$$\text{not intersecting, if} \quad |\mathcal{I}| = 0 \tag{9}$$

and feasible, if both conditions are satisfied.

Additionally, we consider the stability of the object, i.e., we determine whether it would be in static equilibrium (SE) under the estimated pose $\hat{T}$. To be in SE [Del Prete et al. 2016; Hauser et al. 2018], the object must satisfy the conditions of

$$\text{force balance} \quad \sum_i f_i + f_{ext} = \sum_i f_i + mg = 0, \tag{10}$$

$$\text{torque balance} \quad \sum_i (c_m - \hat{x}_i) \times f_i = 0 \quad \text{and} \tag{11}$$

$$\text{admissible contact force} \quad f_i \in \mathcal{K}, \tag{12}$$

where m is the mass of the object, c_m its center of mass, f_i is the contact force at contact point $\hat{x}_i \in \mathcal{C}$ and $\mathcal{K}$ is a friction cone.

The stability constraints may be approximated using the "support polygon principle" [Or and Rimon 2010]. The *support polygon* is defined as the convex hull of the projection of the contact points $\mathcal{C}$ onto the supporting plane. If the projection of the center of mass falls within the support polygon, the object is considered to be in SE [Or and Rimon 2010; McGhee and Frank 1968].

In static cluttered scenes (where gravity is the only external force acting upon objects), certain contact points of an object may not provide support in the gravity direction. Thus, they may result in an overestimation of its static stability, as the support polygon is enlarged by those contacts. Hence, as a compromise between the simplicity of the support polygon principle and the accuracy of solving for conditions (10)–(12), we consider the support polygon with respect to the *supported* points defined [Bauer et al. 2022] as

$$\mathcal{S} = \{\hat{x} \in C : \frac{n_{y(\hat{x})} \cdot g}{||n_{y(\hat{x})}||\,||g||} < 0\}, \tag{13}$$

where $y(\hat{x})$ is the closest point to $\hat{x}$ in the scene and $n_{y(\hat{x})}$ is its surface normal. Therefore, only the subset of contacts is considered onto which a force may be exerted in gravity direction g. See Section 4.2 for an application of this contact-based definition.

Simulation-based Formulation: Instead of evaluating physical plausibility based on contact points, we may also initialize the estimated scene in a physics simulation and evaluate its dynamic progression over time. Intuitively, a plausible configuration of a static scene should not be subject to any change due to gravity in the simulation. Since the 3D models used in the simulation and their physical parameters are inherently approximates of the real objects, we will observe at least slight displacement. Hence, rather than determining *whether* an object moved within the simulation, we want to determine *by how much* it moved over a (varying) period of time. To determine a stable pose, for example, we may want to simulate until the object no longer moves. In the simulation, resolving intersections typically generates an impulse that displaces the involved objects, causing the scene to “explode” in the worst case. To deal with estimated poses that result in intersecting objects, we might only simulate for a few steps at a time before setting the objects’ velocities back to 0 again. See Section 3 for an application of this simulation-based definition.

3 Enforcing Plausibility through Rendering and Simulation

To consider new objects, the methods presented in this section only require 3D models through using rendering and physics simulation . The proposed approaches enforce plausibility, exploit it to limit the search space given multiple pose hypotheses and improve initial poses. In Section 3.1, we present a simple approach for exploiting simulation for pose estimation. In Section 3.2, we present an integrated approach for improving refinement and augmenting it by verification.

3.1 Stable Object Pose Estimation

A simple proof-of-concept pose estimator [Bauer et al. 2020b] demonstrates the predictive power of considering plausibility for this task. It assumes only approximate object meshes and segmentation masks to be given; no additional training is required for pose estimation. This allows us to consider novel instances more easily than with end-to-end trained estimators. We derive a small set of physically plausible poses per object through physics simulation and clustering. Using the visual-alignment score, we are able to determine the visually most plausible candidate.

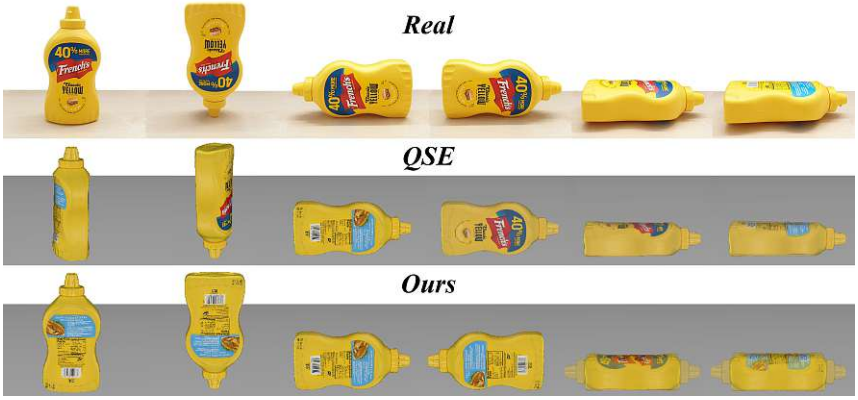


Figure 6 Stable object poses. Top to bottom: The real object, QSE [Goldberg et al. 1999] and our approach for isolated objects (ours). Multiple representatives of the same stable pose are transparently overlayed for QSE and ours. Reprinted from [Bauer et al. 2020a].

To determine the stable poses of an object, it is initialized under a uniformly random rotation in a physics simulator and dropped onto a plane. This assumption is motivated by the observation that objects in static scenes typically rest on horizontal planes, such as tables or shelves. Alternatively, more complex simulation scenes may be used for this purpose. Once the simulated object no longer moves, it has reached a stable pose. This process is repeated multiple times to sample a large number of potential stable poses. However, the resulting poses are highly redundant. First, multiple poses represent the same stable pose, albeit under in-plane rotation. Second, the object resting on different neighboring faces of the locally planar 3D model introduces a slight pose variance. To prune these superfluous poses, we discard in-plane rotation and cluster potential stable poses based on their angular distance. Each resulting stable pose represents the mean rotation and z-translation per cluster, with the plane normal defining the z-axis. Figure 6 shows a a comparison with the related probabilistic *quasi-stable estima-*

tion (QSE) approach [Goldberg et al. 1999] and real-world observations. While both our approach and QSE are able to reliably find all stable poses of an object resting on a horizontal plane, ours leverages a more general simulation-based approach. This would allow us to consider geometrically more complex simulation scenes or further physical properties of the object, beyond its shape and center of mass as in QSE.

To determine the pose of this object in an observation, we generate a pool of stable pose hypotheses by uniformly sampling in-plane rotations for each stable pose. Note that these hypotheses are inherently physically plausible for planar support. Given a segmented depth observation of the object, we may moreover estimate its in-plane translation as an offset from the rendered hypothesis. Among this pool of physically-plausible pose hypotheses, we need to find the visually most plausible pose. This is achieved by computing the visual-alignment score (3) for each hypothesis.

		simulation			
		$\tilde{C}_1$	$\tilde{C}_2$	$\tilde{C}_3$	$\tilde{C}_4$
visual	$\tilde{O}_1$	**51.5**	50.8	48.3	49.1
	$\tilde{O}_2$	51.6	**50.7**	48.4	49.0
	$\tilde{O}_3$	51.4	50.4	**47.8**	48.4
	$\tilde{O}_4$	48.9	48.6	45.2	**44.5**

Table 1 Influence of approximate object meshes on the *visual*-alignment score and *simulation*-based hypotheses generation. Results indicate the AR metric on Occluded LINEMOD.

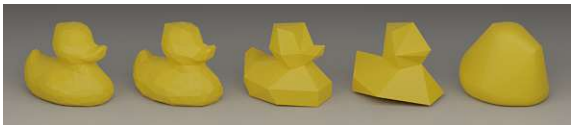


Figure 7 Approximate *duck* models $\tilde{O}_i$ with 704, 352, 70 and 34 faces and the convex hull $\tilde{C}$ of the full-resolution mesh $\tilde{O}$. Reprinted from [Bauer et al. 2020b].

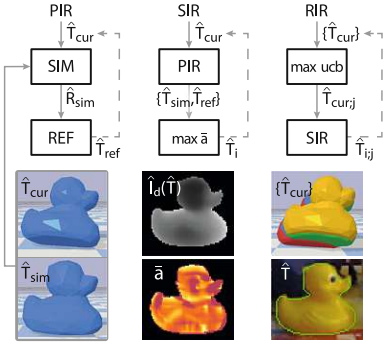
This simple approach achieves competitive pose accuracy on LINEMOD [Hinterstoisser et al. 2012] and Occluded LINEMOD [Brachmann et al. 2016], while also offering a general method to consider novel objects for pose estimation as it only depends on non-textured object meshes. To highlight the robustness of this approach, Table 1 shows our results on the Occluded LINEMOD dataset [Brachmann et al. 2016] using approximations of the object meshes as shown in Figure 7. We evaluate the impact of using the decimated meshes $\tilde{O}_i$ on the visual-alignment score (3) and the influence of using their convex hulls $\tilde{C}_i$ for the simulation-based stable pose generation. The decimated meshes are generated in Blender using the *decimate-collapse* operation. The reported Average Recall (AR) [Hodaň et al.

2020] is computed using the full-resolution object mesh and thus is solely dependent on the accuracy of the estimated pose. As shown per column in Table 1, our hypothesis scoring approach is highly robust to the decimated meshes, producing similarly accurate poses using the first three approximations. Shown per row, the stable pose hypotheses generated using our approach become increasingly inaccurate when the approximated resting shapes deviate farther from the original shape, i.e., with approximations $\tilde{C}_3$ and $\tilde{C}_4$.

3.2 Integrated Object Pose Refinement and Verification

An important step in object pose estimation pipelines is pose refinement. In pipelines yielding multiple pose hypotheses, the best hypothesis must be selected through pose scoring. Moreover, we want to verify the plausibility of the estimated object pose when using it for robotic manipulation, leveraging the pose scoring. With VeREFINE [Bauer et al. 2020c], we integrate iterative refinement, physics simulation and visual-alignment scoring in a joint optimization. We evaluate this approach on pose estimation datasets and in real-world grasping experiments.

(a) Initial pose estimates in the simulation environment (top) are improved using VeREFINE (mid, bottom), enabling successful robotic grasping.



(b) PIR: Integration of physics simulation (SIM) and iterative refinement (REF). SIR: Supervision using verification score $\bar{a}$. RIR: Regret minimization.

Figure 8 Grasping YCB objects with a Toyota HSR (a) and the iterative approaches proposed in VeREFINE (b), given an initial object pose estimate ($\hat{T}_{cur}$). Adapted from [Bauer et al. 2020c].

During refinement, we would like both discussed aspects of plausibility to inform one another. We achieve this by interleaving physics simulation steps with iterative refinement steps, as illustrated in Figure 8b (Physics-guided Iterative Refinement, *PIR*). Thereby, simulation guides refinement towards physically more

plausible poses, while alignment-based refinement improves visual plausibility. Both steps work complementary, improving each other's initialization.

However, either step might diverge, for example, due to bad initialization. The simulated object might topple over and move away from its true pose. Local refinement may determine incorrect correspondences and move toward a false pose. To contain these issues, we embed the visual-alignment score (3) in the refinement loop, as shown in Figure 8b (Supervised Iterative Refinement, *SIR*). Note that this also facilitates pose verification.

Generally, we might have to refine more than one object pose hypothesis. For example, with the pose estimator proposed in Section 3.1, multiple in-plane hypotheses need to be considered per stable pose hypothesis. With a growing number of hypotheses, simply refining and scoring all of them becomes computationally expensive. Rather, we want to spend a fixed budget of refinement iterations. We propose to consider the efficient allocation of the refinement budget as a multi-armed bandit problem. To minimize the regret of choosing to refine a sub-optimal hypothesis with respect to its visual-alignment score, we employ the Upper Confidence Bound policy (UCB) [Auer et al. 2002], as shown in Figure 8b (Regret-minimizing Iterative Refinement, *RIR*). The policy balances exploitation of high-scoring hypotheses with exploration of alternative, potentially better hypotheses.

We extend our approach to multiple objects per scene, considering the scene-level interactions of objects. We cluster scene objects based on their support relationships, with each cluster starting from a base object in contact with the supporting plane. The clusters are then ordered from front to back, i.e., starting from the least occluded base object. To yield physically plausible configurations, we iteratively add objects from the ordered clusters to the simulated scene during refinement. Each object's pose hypotheses are refined as before, albeit considering the visual plausibility of the whole scene. The highest scoring hypothesis per object is added to the simulation scene used for the subsequent objects, allowing the consideration of occlusions and support relationships between them.

Table 2 shows the results of the different single- and multi-hypotheses approaches we propose in VeREFINE [Bauer et al. 2020c] on the YCB-Video dataset [Xiang et al. 2018]. The dataset contains scenes of 3-6 YCB objects [Calli et al. 2015], that are occluded and stacked upon each other in clutter. Initial pose hypotheses are generated using DenseFusion (DF) [Wang et al. 2019] and its associated refinement network (DF-R) is used as implementation of *REF* (see Figure 8b). Figure 9 depicts an ablation study to show the influence of the initial rotation and translation error as wel as the impact of partial depth data. For these

	AR	#ref/obj
DF-R	73.9	2
PIR	74.7	2
SIR	76.5	2
VF$_b$	77.6	10
VF$_d$	**77.8**	10

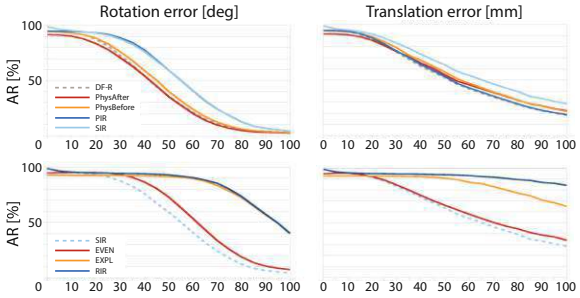
(a) Comparison on YCB-VIDEO.

	mustard	spam	foam	jello	banana	success	found	#ref/obj
DF-R	10	3	1	7	0	42%	46%	2
SIR	9	7	2	7	0	50%	70%	2
DF-R	10	6	5	9	1	62%	70%	10
MCTS	9	10	2	6	0	54%	78%	10
RIR	**10**	**10**	**9**	**10**	**4**	**86%**	**90%**	10

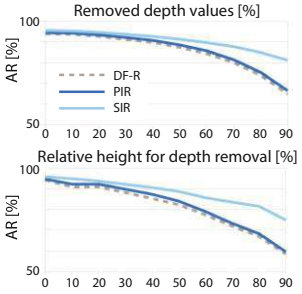
(b) Results of grasping experiments in percentage of *found* collision-free grasp poses and *success*ful grasp attempts.

Table 2 Evaluation of the methods in VeREFINE [Bauer et al. 2020c] (bold). Initial poses from DenseFusion [Wang et al. 2019], sampled to 1/5 hypotheses per object and refined with a budget of two refinement iterations per hypothesis and object for a total of 2/10 iterations.

experiments, the initial poses are generated by adding a uniformly random error of varying magnitude on top of the ground-truth poses.



(a) Using single hypotheses (top) and five hypotheses (bottom). EVEN and EXPL use our verification score to determine the best estimate and PIR for refinement. PhysBefore and PhysAfter apply simulation before and after refinement.



(b) Robustness to missing depth values using a single hypothesis with a fixed error magnitude of 5mm and 5deg.

Figure 9 Ablations on LM. Average Recall (AR) [Hodaň et al. 2020] values are reported at 5mm/deg steps (a) and every 10% (b), respectively, and are linearly interpolated in between. Adapted from [Bauer et al. 2020c].

The integration of physics simulation in the iterative refinement loop (PIR) improves the achieved accuracy by providing better initialization in each step. In Figure 9a (top left) we see how alternative ways of combining simulation with refinement may even reduce the performance. The use of the visual-alignment score (SIR) significantly improves accuracy, as indicated in Table 2a. It also improves

the robustness to partial depth data, as shown in Figure 9b. Our motivation for using a multi-armed bandit formulation for considering multiple hypotheses (RIR) is to balance exploration of the different hypotheses with exploitation of known high-scoring hypotheses. In the extreme case, the former would spend the budget of refinement iterations evenly among hypotheses (EVEN), while the latter would use it to refine a single hypothesis (EXPL). Figure 9a (bottom row) shows the benefit of using multiple hypotheses and our regret-minimizing approach. These findings also transfer to real-world grasping experiments with a Toyota HSR and using the GRASPA layouts [Bottarel et al. 2020] for reproducibility, illustrated in Figure 8a. As indicated by the results in Table 2b, both our single hypothesis (SIR) and multi-hypothesis approaches (RIR) significantly improve grasp success compared to the baseline refiner (DF-R) and a competing approach that uses a combination of physics simulation and refinement in a Monte Carlo tree search (MCTS) scheme [Mitash et al. 2018].

4 Enforcing Plausibility in Learning-based Approaches

The methods presented in Section 3 consider the visual and physical aspects of plausibility separately. For example, in Section 3.2, enforcing physical plausibility through simulation competes with enforcing visual plausibility through iterative refinement, illustrated by the experiments in Figure 9a (top). Instead, the influence of both plausibility aspects should be dynamically adapted depending on the scene configuration and refinement state. We want to leverage the contact-based constraints (8)–(12) directly for refinement. This motivates the design of a learning-based, plausible pose refinement approach.

4.1 Reinforced Point Cloud Registration

As the first step in this direction, we propose a novel approach to the related task of point cloud registration [Bauer et al. 2021]. We pose the iterative registration task as determining a policy that selects basic registration actions in each step, as illustrated in Figure 10. Inspired by [Shao et al. 2020], we use discrete steps per axis, separately for rotation and translation. These actions, for example, translate the source by a small offset in x direction. Our registration agent (ReAgent) is trained using imitation and reinforcement learning. Its formulation allows the incorporation of additional constraints – such as physical plausibility – by including them in the agent's reward.

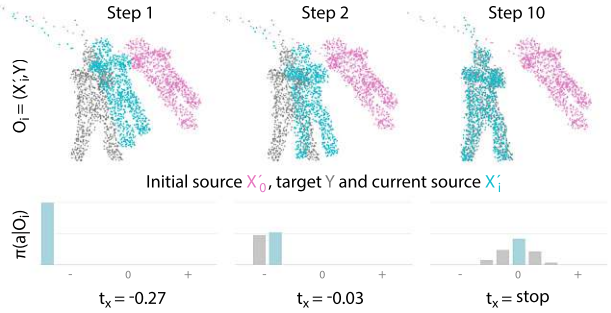


Figure 10 Iterative registration using ReAgent. The source point cloud (cyan) is aligned to the target point cloud (gray), starting from an initial source (magenta). ReAgent follows policy π by taking action $a_i = \arg\max_a \pi(a|O_i)$ given the current observation O_i, improving registration step-by-step. Reprinted from [Bauer et al. 2021].

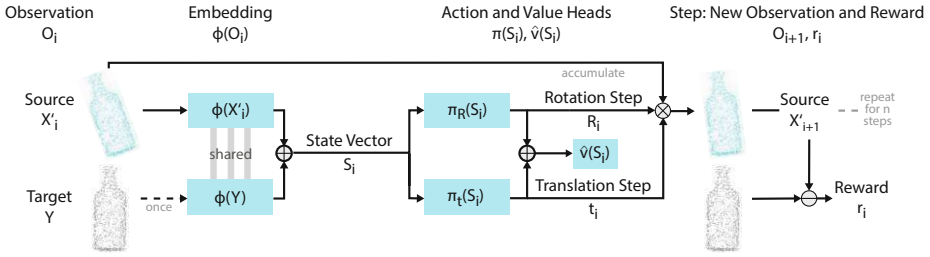


Figure 11 Architecture overview for one iteration of ReAgent. Reprinted from [Bauer et al. 2021].

The agent is implemented as a neural network, illustrated in Figure 11. The observed point clouds are embedded into a state space to reduce their dimensionality. The embedding uses a siamese PointNet-like architecture [Qi et al. 2017], generating a global feature that represents each point cloud. Two policy heads then predict the discrete distribution representing the policies for the rotation and translation action to be selected next. This process is also visualized in Figure 10 (bottom).

Since jointly learning the embedding and the registration policies from scratch using reinforcement learning (RL) might not converge (quickly), we opt for a hybrid training approach that also includes imitation learning (IL). Through IL, the agent should learn to replicate the behavior of an expert. We define an expert registration policy with perfect information (ground-truth transformation T) and, in each iteration, selects the actions that take the largest step toward alignment. Additionally, the agent is reinforced by a symmetry-aware point-cloud alignment

reward. The resulting loss is a combination of a cross-entropy loss for IL and the Proximal Policy Optimization (PPO) loss [Schulman et al. 2017] for RL.

	PoseCNN	DeepIM	Multi-ICP	ReAgent (IL)	ReAgent (IL+RL)
AD $< 0.10d$ (↑)	62.8	88.6	92.1	**98.7**	**98.7**
AD $< 0.05d$ (↑)	26.9	69.2	68.6	90.6	**91.1**
AD $< 0.02d$ (↑)	3.3	30.9	19.0	38.8	**39.6**

Table 3 Comparison of object pose refinement methods on LINEMOD (mean over per-class results) using PoseCNN [Xiang et al. 2018] for initial object pose and segmentation.

Figure 12 Qualitative examples on LINEMOD using ReAgent (IL+RL). In the top row, 1024 points are sampled within the estimated segmentation mask. The black box indicates the zoomed-in view. Outlines are shown for target (gray), initial (magenta) and current source pose (cyan). The last column shows a failure case. Reprinted from [Bauer et al. 2021].

In [Bauer et al. 2021], we show that our lightweight approach achieves faster inference as well as improved accuracy and robustness to noise and initialization as compared to related learning-based approaches on ModelNet40 [Wu et al. 2015] and ScanObjectNN [Uy et al. 2019]. Experiments on LINEMOD [Hinterstoisser et al. 2012], moreover, show high accuracy when applying ReAgent to the pose refinement task. Table 3 shows the comparison of our method to DeepIM [Li et al. 2018] and a rendering-based multi-hypothesis approach (Multi-ICP) [Xiang et al. 2018], employing initial poses and segmentation mask estimated using PoseCNN [Xiang et al. 2018]. When applied to the pose refinement task, our point cloud registration method achieves state-of-the-art performance on the LINEMOD dataset. The results obtained with tighter AD thresholds indicate the benefit of the combined IL and RL approach. Furthermore, Figure 12 illustrates the sampling of the source point cloud and qualitative examples of the accuracy of our ReAgent approach.

4.2 Reinforced Object Pose Refinement and Verification

When we apply the method from Section 4.1 (ReAgent) to cluttered scenes such as the ones observed in the YCB-Video dataset, we must cope with partial point clouds that may contain outliers from neighboring objects due to occlusion and inaccurate segmentation. Additionally, the initial pose estimates are affected by these challenges and are, in general, less accurate than in the single object case previously evaluated.

As we suggested in Section 2.2, additional consideration of physical plausibility allows us to resolve the resulting visual ambiguities. To this end, for *SporeAgent* [Bauer et al. 2022], we integrate our contact-based formulation from [Bauer et al. 2020a] with ReAgent. We modify it further to consider object symmetries, outlier points and visual-alignment scores. As a result, a learning-based approach similar to VeREFINE [Bauer et al. 2020c] (Section 3.2) is achieved that jointly considers both aspects of plausibility.

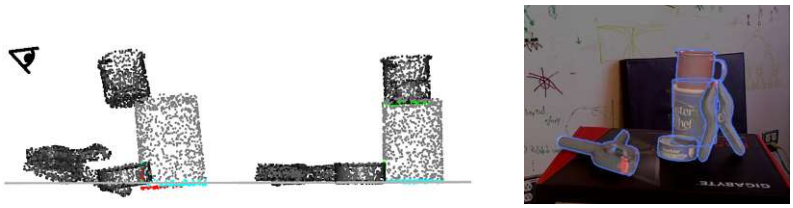


Figure 13 Initial scene representation (left) and refined poses using SporeAgent (mid and right). The critical points for one target object (gray) are shown – intersecting (red), contact (green) and supported (cyan). Adapted from [Bauer et al. 2022].

Physical plausibility is considered at two points in the refinement pipeline. First, we define an additional reward term that reinforces the agent to reach SE, approximated using the support polygon principle for the supported points $\mathcal{S}$ (as defined in Section 2.2). Second, we discover that the surface distance $\delta(\hat{x})$ is a useful input signal for the agent. It provides the underlying information required to determine the SE and, in addition, orients the object within the scene by including the distance to the supporting plane. As illustrated in Figure 13, these extensions allow the agent to resolve implausible configurations.

Visual plausibility with respect to the point clouds is already considered by the refinement itself. Additionally, to evaluate the iterative results, we leverage the visual-alignment score (3). Thereby, we are able to determine the overall most plausible (and accurate) object poses. This reduces the effect of the agent oscillating between two similarly fitting poses for fine alignment, as we observed in our

experiments, and allows to resort to the initial pose should the refinement diverge. Figure 4 shows a qualitative example for scoring.

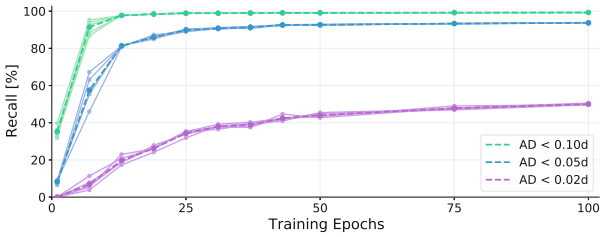
To further adapt the method to the task of object pose refinement in clutter, we introduce an outlier-removal subnetwork. Based on a concatenation of local and global features, this subnetwork is tasked with labeling geometrical outliers and is trained under an artificial segmentation error. The latter is an input augmentation that samples a coherent patch from the ground-truth segmentation mask, simulating occlusion and potentially including background pixels. The outlier predictions prune these geometrical outliers before the computation of the global feature used in the state vector (see Figure 11). Moreover, we adapt the expert policy to consider symmetrical objects by following the shortest trajectory to any symmetrical pose. To this end, we propose a canonical object frame in which the symmetry axes coincide with the origin, allowing symmetrical poses to be reduced to rotations. As a result, the symmetry-aware expert policy tends toward the symmetrical pose with minimal rotation from the current pose estimate.

	PoseCNN	ICC-ICP	P2Pl-ICP	w/ VeREFINE	Multi-ICP	SporeAgent
ADD AUC (↑)	51.5	67.5	68.2	70.1	77.4	**79.0**
AD AUC (↑)	61.3	77.0	79.2	81.0	86.6	**88.8**
ADI AUC (↑)	75.2	85.6	87.8	88.8	92.6	**93.6**

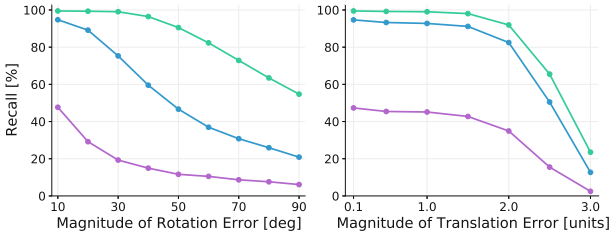
Table 4 Comparison of depth-based refinement methods on YCB-VIDEO (mean over per-class results) using PoseCNN [Xiang et al. 2018] for initial object pose and segmentation.

Table 4 shows the improved accuracy of SporeAgent compared to related depth-based refinement methods on YCB-Video [Xiang et al. 2018]. All compared methods use initial poses and segmentation masks estimated using PoseCNN [Xiang et al. 2018]. We compare our method to Iterative Collision Check with ICP (ICC-ICP) [Wada et al. 2020], vanilla Point-to-Plane ICP (P2Pl-ICP) [Chen and Medioni 1992; Zhou et al. 2018], P2Pl-ICP augmented by single-hypothesis VeREFINE [Bauer et al. 2020c] and a rendering-based multi-hypothesis approach (Multi-ICP) [Xiang et al. 2018]. While VeREFINE is able to significantly improve the results of the simple ICP approach by combining physics simulation with visual-alignment scoring, it is still inherently limited by the performance of the underlying refinement approach. In contrast, SporeAgent is able to exploit both sources of information to achieve state-of-the-art accuracy.

Figure 14a shows the training convergence of SporeAgent for five different random seeds on LINEMOD. For all evaluated thresholds, there is minimal variation in the recall beyond 50 epochs. Figure 14b shows an ablation study that highlights the robustness of SporeAgent to the quality of the initialization. For example, in



(a) Convergence of the mean recall per epoch (dashed) on LINEMOD for five random seeds (solid).



(b) Varying initialization error in rotation (left) and translation (right).

Figure 14 Ablations on LM. AD recalls with thresholds as fraction of the object diameter d [Hinterstoisser et al. 2012]. Reprinted from [Bauer et al. 2022].

the case of a translation error, the accuracy starts to decline only at a magnitude of around 2.0 units, which is limited by the number of iterations and the largest translation-step size.

5 Explaining Plausibility Violations

The consideration of plausibility offers not only a technical advantage but also supports users' understanding of the robot's perception and actions, thereby fostering trust. In [Papagni et al. 2021], we investigate how human interaction partners perceive plausibility-based explanations of robotic failure. Our proposed online study evaluates the impact of different explanation strategies on users' understanding of the robot and their trust in it after the interaction.

Participants in the study are instructed to assist a robot in locating and removing objects from a table, as shown in Figure 15 (top left). They are informed that their human-robot team may earn up to eight points in this task, one per object. This is to give the participants "something at stake" in the interaction. They are given a description of the next object to be be removed and are requested to provide the

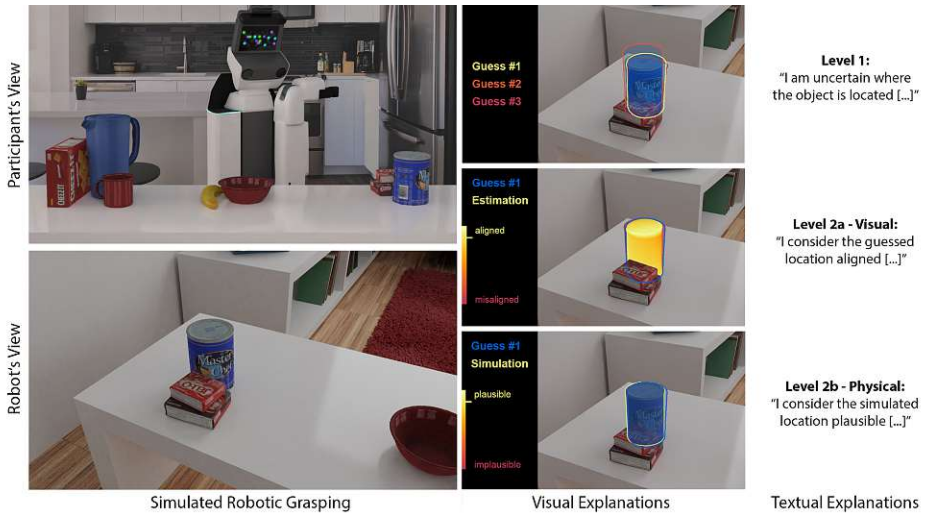


Figure 15 Rendered interaction from the view of the participants (top left) and the robot (bottom left). Example textual explanations are shown together with visualizations of uncertainty (top right), visual plausibility (mid right) and physical plausibility (bottom right). Adapted from [Papagni et al. 2021].

robot with an initial location. While hovering the cursor over the correct object, a circle indicates the corresponding location area. As a result, we aim to increase the perceived involvement of the participants in this human-robot interaction. After providing a location, they are shown rendered videos of the robot performing its task. Initially, robot (and hence *the team*) succeeds twice.

The third grasp attempt of the robot fails and the participants are shown different types of explanations, depending on the experimental condition to which they are assigned, as shown in Figure 15 (right). In a 2-by-2 study design, we modify the interactivity (*single-shot* or *multiple levels*) and the reasoning strategy (*visual alignment of the rendering* or *displacement in the physics simulation*) of the provided explanation. Participants then report their understanding of the explanation and answer short questionnaires regarding trust.

A technical pilot study has already highlighted the importance of the design of the visual explanations. Based on the findings of a currently ongoing user study, we will be able to further improve the visualizations and explanations provided by our object pose estimation approaches for deployment in human-robot interaction scenarios.

6 Conclusion and Future Work

This chapter discussed our definition of visual and physical plausibility, its technical benefit in object pose estimation and robotic grasping as well as its application in generating understandable explanations for human-robot interaction.

We showed that, by jointly considering these two aspects of plausibility, we are able to achieve increased pose accuracy in situations when each aspect alone would be ambiguous. We propose a set of object pose estimation and refinement approaches that are solely based on the 3D model of the objects and may be directly used to augment existing pipelines. Further exploiting the combined visual and physical plausibility information, we present a learning-based pose refinement method that considers the intersecting and supported points between interacting objects. Finally, we give an outlook on ongoing work investigating the exploitation of the plausibility information computed by our approaches to generate human-understandable explanations of robotic failure.

Nevertheless, many of the objects that robots have to deal with are not yet covered by the rigidity and static-scene assumptions of the proposed methods. Dealing with articulated (or even deformable) objects, potentially being manipulated by a human hand or robotic gripper and exposing high intra-class variance in texture and shape, is beyond the scope of this work. To this end, the visual plausibility considerations could be extended to include color information to deal with texture, thereby increasing the robustness of the methods to partial depth data. Considering, for example, hand-object contacts would allow the physical plausibility definition to be extended to these dynamic cases. Moreover, a robotic prototype that employs the presented methods to generate a scene explanation and the corresponding explanations of its actions (and failures) would allow for further evaluation of our approach in the ever-changing environments that the robots' human interaction partners inhabit.

Bibliography

Abdallah Arioua, Patrice Buche, and Madalina Croitoru. 2017. Explanatory dialogues with argumentative faculties over inconsistent knowledge bases. *Expert Systems with Applications* 80 (2017), 244–262.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. 2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning* 47, 2-3 (2002), 235–256.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2020a. Physical Plausibility of 6D Pose Estimates in Scenes of Static Rigid Objects. *European Conference on Computer Vision Workshops*, 648–662.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2020b. Scene Explanation through Verification of Stable Object Poses. *ICRA 2020 Workshop on Perception, Action, Learning.*

Dominik Bauer, Timothy Patten, and Markus Vincze. 2020c. VeREFINE: Integrating object pose verifi cation with physicsguided iterative refinement. *IEEE Robotics and Automation Letters* 5, 3, 4289–4296.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2021. ReAgent: Point Cloud Registration using Imitation and Reinforcement Learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14586–14594.

Dominik Bauer, Timothy Patten, and Markus Vincze. 2022. SporeAgent: Reinforced Scenelevel Plausibility for Object Pose Refinement. *IEEE Winter Conference on Applications of Computer Vision*, 654–662.

Fabrizio Bottarel, Giulia Vezzani, Ugo Pattacini, and Lorenzo Natale. 2020. GRASPA 1.0: GRASPA is a robot arm grasping performance benchmark. *IEEE Robotics and Automation Letters* 5, 2 (2020), 836–843.

Eric Brachmann, Frank Michel, Alexander Krull, Michael Ying Yang, Stefan Gumhold, et al. 2016. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3364–3372.

Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. 2015. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robotics and Automation Magazine* 22, 3 (2015), 36–52.

Yang Chen and Gérard Medioni. 1992. Object modelling by registration of multiple range images. *Image and Visual Computing* 10, 3 (1992), 145–155.

Sachin Chitta, E Gil Jones, Matei Ciocarlie, and Kaijen Hsiao. 2012. Mobile manipulation in unstructured environments: Perception, planning, and execution. *IEEE Robotics and Automation Magazine* 19, 2 (2012), 58–71.

Maartje MA de Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). *AAAI Fall Symposium Series.*

Maartje MA de Graaf, Bertram F Malle, Anca Dragan, and Tom Ziemke. 2018. Explainable robotic systems. *Companion of the ACM/IEEE International Conference on Human-Robot Interaction*, 387–388.

Andrea Del Prete, Steve Tonneau, and Nicolas Mansard. 2016. Fast algorithms to test robust static equilibrium for legged robots. *International Conference on Robotics and Automation,* 1601–1607.

Paul E Dunne, Sylvie Doutre, and Trevor Bench-Capon. 2005. Discovering inconsistency through examination dialogues. *International Joint Conference on Artificial Intelligence,* 1680–1681.

Ken Goldberg, Brian V Mirtich, Yan Zhuang, John Craig, Brian R Carlisle, and John Canny. 1999. Part pose statistics: Estimators and experiments. *IEEE Transactions on Robotics and Automation* 15, 5 (1999), 849–857.

Kris Hauser, Shiquan Wang, and Mark R Cutkosky. 2018. Efficient equilibrium testing under adhesion and anisotropy using empirical contact force models. *IEEE Transactions on Robotics* 34, 5 (2018), 1157–1169.

Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. 2012. Model based training, detection and pose estima-

tion of textureless 3d objects in heavily cluttered scenes. Asian Conference on Computer Vision, 548–562.

Tomáš Hodaň, Jiří Matas, and Štěpán Obdržálek. 2016. On evaluation of 6D object pose estimation. *European Conference on Computer Vision*, 606–619.

Tomáš Hodaň, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. 2018. BOP: Benchmark for 6D object pose estimation. *European Conference on Computer Vision,* 19–34.

Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. 2020. BOP Challenge 2020 on 6D Object Localization. *European Conference on Computer Vision Workshops* (2020).

Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. 2018. Deepim: Deep iterative matching for 6d pose estimation. *European Conference on Computer Vision,* 683–698.

Meghann Lomas, Robert Chevalier, Ernest Vincent Cross, Robert Christopher Garrett, John Hoare, and Michael Kopack. 2012. Explaining robot actions. *ACM/IEEE International Conference on Human-Robot Interaction,* 187–188.

Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A Grounded Interaction Protocol for Explainable Artificial Intelligence. *International Conference on Autonomous Agents and Multiagent Systems,* 1033–1041.

Robert B McGhee and Andrew A Frank. 1968. On the stability properties of quadruped creeping gaits. *Mathematical Biosciences* 3 (1968), 331–351.

Chaitanya Mitash, Abdeslam Boularias, and Kostas E Bekris. 2018. Improving 6D pose estimation of objects in clutter via physics-aware Monte Carlo tree search. *International Conference on Robotics and Automation,* 3331–3338.

Muzammal Naseer, Salman Khan, and Fatih Porikli. 2018. Indoor scene understanding in 2.5/3d for autonomous agents: A survey. *IEEE access* 7 (2018), 1859–1887.

Yizhar Or and Elon Rimon. 2010. Analytic characterization of a class of three-contact frictional equilibrium postures in three-dimensional gravitational environments. International Journal on Robotics Research 29, 1 (2010), 3–22.

Guglielmo Papagni, Dominik Bauer, Sabine Köszegi, and Markus Vincze. 2021. A Study Design for Evaluation of Trust and Understandability through Interactive Multi-Modal Explanations of Robotic Failure. *HRI 2021 WYSD Workshop.*

Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. 2017. Pointnet: Deep learning on point sets for 3d classification and segmentation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 652–660.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

Jianzhun Shao, Yuhang Jiang, Gu Wang, Zhigang Li, and Xiangyang Ji. 2020. PFRL: Pose-Free Reinforcement Learning for 6D Pose Estimation. *IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 11454–11463.

Siddhartha S Srinivasa, Dave Ferguson, Casey J Helfrich, Dmitry Berenson, Alvaro Collet, Rosen Diankov, Garratt Gallagher, Geoffrey Hollinger, James Kuffner, and Michael Vande Weghe. 2010. HERB: A home exploring robotic butler. *Autonomous Robots* 28, 1 (2010), 5.

Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stanley T Birchfield. 2018. Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects. *Conference on Robotic Learning,* 306–316.

Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. 2019. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. *International Conference on Computer Vision,* 1588–1597.

Kentaro Wada, Edgar Sucar, Stephen James, Daniel Lenton, and Andrew J Davison. 2020. Morefusion: multi-object reasoning for 6d pose estimation from volumetric fusion. IEEE/CVF *Conference on Computer Vision and Pattern Recognition,* 14540–14549.

Douglas Walton. 2007. Dialogical Models of Explanation. *Explanation-aware computing: Papers from the 2007 AAAI workshop.* Technical Report WS-07-06 (pp. 1–9). Menlo Park, CA: AAAI Press.

Chen Wang, Danfei Xu, Yuke Zhu, Roberto Martín-Martín, Cewu Lu, Li Fei-Fei, and Silvio Savarese. 2019. DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 338–3347.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3d shapenets: A deep representation for volumetric shapes. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1912–1920.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. 2018. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems.*

Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. 2018. Open3D: A modern library for 3D data processing. *arXiv preprint arXiv:1801.09847* (2018).