2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Information-driven Affordance Discovery for Efficient Robotic Manipulation

Pietro Mazzaglia[*]
Qualcomm AI Research[†]
Ghent University

Taco Cohen
Qualcomm AI Research[‡]

Daniel Dijkman
Qualcomm AI Research

***Abstract*— Robotic affordances, providing information about what actions can be taken in a given situation, can aid robotic manipulation. However, learning about affordances requires expensive large annotated datasets of interactions or demonstrations. In this work, we argue that well-directed interactions with the environment can mitigate this problem and propose an information-based measure to augment the agent's objective and accelerate the affordance discovery process. We provide a theoretical justification of our approach and we empirically validate the approach both in simulation and real-world tasks. Our method, which we dub IDA, enables the efficient discovery of visual affordances for several action primitives, such as grasping, stacking objects, or opening drawers, strongly improving data efficiency in simulation, and it allows us to learn grasping affordances in a small number of interactions, on a real-world setup with a UFACTORY xArm 6 robot arm.**

I. Introduction

Given the success of learning-based approaches across various domains [1], [2], [3], [4], applying learning-based approaches to robotics represents a sensible and appealing idea, in order to develop more general and robust approaches. However, despite recent progress [5], [6], learning generalizable robotic systems remains challenging. In particular, robotic manipulation policies can be learned via imitation [7] or reinforcement learning [8], but these struggle to work outside of their training conditions.

In order to produce more generalizable manipulation policies, the robot's perception should provide meaningful representations about how to interact with the environment. *Affordances*, defined as the properties that determine how things in the environment can be used [9], provide strong cues on how to operate things in the environment. More specifically, visual affordance models generate a visual representation of the environment that maps observed points to a set of possible actions that can be executed at that point [10]. However, how to efficiently learn visual affordance models remains a challenge.

Learning visual affordances can be seen as a supervised learning problem, where the agent needs to predict whether a point in the scene is actionable or not, given a certain action primitive. However, if we tackle the problem the classical way, learning an affordance model from a given dataset, training a general and robust model requires large chunks of annotated experiences that are expensive to collect via human interaction or teleoperation [11], [12]. Alternatively,

[*]Qualcomm AI Research is an initiative of Qualcomm Technologies, Inc.
[†]Work done during an internship at Qualcomm AI Research.
[‡]Work completed while employed at Qualcomm AI Research.

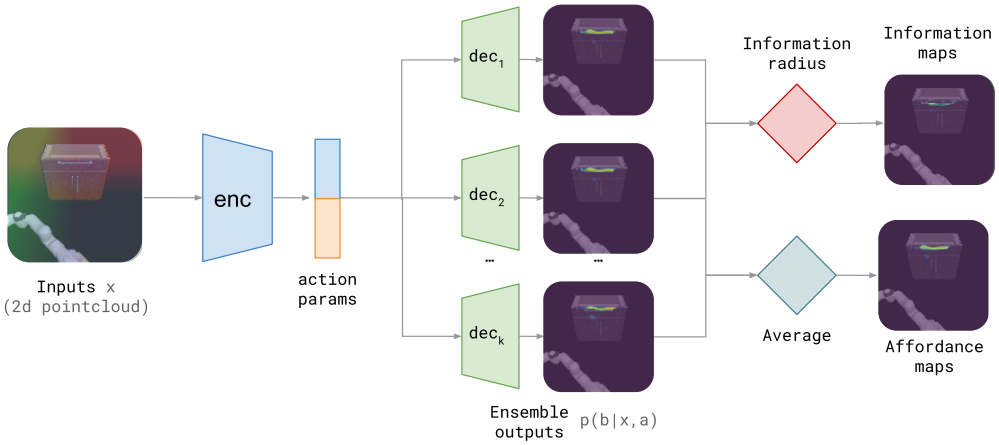


Fig. 1: **Information-driven affordance discovery.** The model processes inputs from the environment (2D point cloud) using a single encoder, concatenates action parameters and decodes visual affordance maps with multiple decoders (ensemble). Averaging the outputs of these networks, we can extract reliable affordance maps, thanks to the ensemble diversity. Computing the information radius, we can obtain information gain maps about affordances in the scene, to drive considerate explorative interactions. Images represent actual model outputs from IDA in the ManiSkill2 Open Drawer environment.

979-8-3503-8457-4/24/$31.00 ©2024 IEEE

synthetic data can be used to rapidly obtain large datasets to train on [10], [13], with the risk that the model may not behave as expected in more realistic or complex settings. In this work, we aim to overcome these limitations, by tackling the affordance discovery problem as an active learning problem. Actively learning what the agent can do in the environment, by continuously interacting with the environment and learning from the new data, should enable faster and more effective learning of affordances, compared to passive data collection from humans or the use of synthetic data [14].

We propose **IDA**, an {I}nformation-{D}riven method for visual {A}ffordance discovery, that is able to learn what's feasible in the environment in a small number of interactions, by casting the problem of affordance learning as a contextual bandit problem and exploring actions that are both rich in information and likely to succeed.

Contributions. Our contributions can be summarized as:

- we propose an information-driven measure to augment the agent's objective for visual affordance discovery in interactive settings, and we provide motivation for our approach based on information theory;
- we validate IDA in simulation, where the agent quickly learns to grasp, stack objects, and open drawers, strongly outperforming previous methods trained on large synthetic datasets [13]. In this setting, we also show the importance of well-grounded exploration to sensibly improve performance as the number of interactions increases over time;
- we demonstrate the applicability of IDA in a real robotic setup, with a UFACTORY xArm 6, where our agent learns to grasp objects in a small number of interactions, without any prior information.

II. Related Work

Visual robotic manipulation. Learning to perform manipulation from high-dimensional visual inputs is challenging. Previous work has successfully applied deep learning vision models for robotics, using supervised learning [15], reinforcement learning [16], [17], or world models [8]. Fully convolutional models, similar to those applied for segmentation [18], have shown good performance in learning visual tasks such as pushing, grasping [19], [20] and rearranging objects [21]. U-Net-like models [22] have been shown to allow efficient reinforcement learning from pixels [23]. The success of fully convolutional models for manipulation could be attributed to the improved capacity of these models to process positional information [24], which is crucial for manipulation tasks.

Affordance learning. Previous work on affordances generally focuses on two aspects of the problem: learning the perception modules, that allow detecting affordances in the scene [10], [25], and learning the action modules [26], [27], which allow translating affordance abstractions into actions. One of the closest works to ours is Where2Act [10], which proposes a perception model and learning-from-interaction framework that combines random and adaptive sampling. The method uses a large number of interactions, simulated through a "flying" gripper, to learn visual pushing and pulling affordances. To bootstrap learning of the affordance module, random sampling is used until obtaining 10k successful interactions, a procedure that requires sampling hundreds of thousands of trajectories. In contrast, IDA achieves greater data efficiency, by actively sampling interactions rich in information.

Exploratory grasping. Grasping remains a challenging problem in robotics [28] due to the high level of complexity required in the interaction between the agent and the objects. Moreover, the large space of potential grasping configurations makes the problem hard to explore. Successful results have been obtained leveraging large synthetic datasets [29], [30], [13], but they still struggle to generalize [31]. In order to develop more data-efficient approaches, the problem of exploratory grasping [32], i.e. active learning of grasping policies, has been recently studied, leveraging pre-trained models as prior [33] or analytical solutions [34] to ease the learning process. Some works also employed a bandit formulation to tackle the problem [35], [32], [36]. In our work, we present the idea of actively learning multiple kinds of affordances, including grasping, and present a reliable solution to efficiently learn such affordances by interacting with the environment, without any prior data or knowledge.

III. Information-driven Affordance Discovery

Setting. In contextual bandits problems, at each iteration, the agent observes a context $x \in X$ and selects an action $a \in A$ to perform, which is rewarded by the environment. In this work, we propose to tackle the affordance discovery problem as a contextual bandit problem. The context is the current state of the environment, which is observed by the agent through a camera. For each kind of affordance, e.g., stacking, grasping, opening, actions are represented by parameterizable primitives [37]. For a certain primitive, actions are represented as $a = [p, q]$, where p is the position where the primitive is applied and q is the orientation to apply the primitive. Rewards represent the affordance success, and so the possible values are $\{0, 1\}$.

Information gain. Let Θ be the space of parameters for an affordance model of the environment, then $p(\theta)$ represents the probability density function of a certain set of parameters to be the best fit for the affordance function. In order to improve our knowledge about what could the best affordance model, we adopt the information gain criterion for exploration, i.e. we want to try actions that maximize the reduction in the entropy with respect to Θ achieved by learning the outcome of the affordance transition (x, a, b), with outcome $b \in \{0 = \text{fail}, 1 = \text{success}\}$. The information gained with one transition can be written as:

$$I(b, x, a) = D_{\text{KL}}[p(\Theta|b, x, a) \| p(\Theta)], \tag{1}$$

that is the Kullback-Leibler divergence between the posterior about the model's parameters given the transition and the prior about the same parameters.

We are interested in measuring the expected amount of information to be gained from performing action a in state x. Observing that $p(b, x, a|x, a) = p(b|x, a)$, we can rewrite the above as:

$$I(x, a) = \mathbb{E}_{p(b|x,a)}[I(b, x, a)]. \tag{2}$$

In Appendix, we show that this equals the Jensen-Shannon Divergence (JSD) between the distributions $p(b|x, a, \theta)$ for all θ:

$$\begin{aligned} I(x, a) &= JSD(p(b|x, a, \theta)|\theta \sim \Theta) = \\ &= H(\mathbb{E}_{\theta\sim\Theta}[p(b|x, a, \theta)]) - \mathbb{E}_{\theta\sim\Theta}[H(p(b|x, a, \theta))], \end{aligned} \tag{3}$$

The JSD, also known as information radius [38], represents the total divergence to the average in a space of distributions. It can also be interpreted as the amount of information that is carried by evidence about the model [39]. In our context, the JSD captures the amount of information that a context-action pair (x, a) carries about which model parameterization θ fits the data best. Thus, we can use it as a measure of the utility of the pair in order to improve the affordance model. Crucially, as we also discuss in Section IV, computing the JSD only requires sampling multiple sets of parameters, and thus requires no additional data from the environment or updates of the model.

Affordance discovery. During the interaction and training stage, we aim to tackle the affordance discovery problem by sampling actions that are informative to learn a better affordance model but also likely to be successful. For the former, we can rely on the JSD presented in Equation 3 to measure the information we expect to gain with a new interaction. For the latter, we should consider the probability $p(b|x, a)$ of an affordance to be successful.

As our setting is analogous to a contextual bandit problem, we can combine our information-driven measure with the expected "reward" from the environment, by adopting an upper confidence bound (UCB) strategy [40], [41]. This implies that the agent samples actions according to an overestimate of the expected reward:

$$\arg\max_a [\hat{r}(x, a) + c_{\text{expl}} \cdot I(x, a)] \tag{4}$$

where the coefficient $c_{\text{expl}} \geq 0$ controls the tradeoff between exploration and exploitation. This means that practically the agent, being in context x, should sample actions that satisfy:

$$\arg\max_a [p(b|x, a) + c_{\text{expl}} \cdot JSD(p(b|x, a, \theta)|\theta \sim \Theta)] \tag{5}$$

that, given that the JSD is non-negative, complies with the upper confidence bound condition of the agent's criterion being greater or equal than $\hat{r}(x, a)$.

IV. Implementation

Model. In order to obtain visual affordance maps of the environment, we adopt an auto-encoder architecture based on a fully convolutional U-Net model [22]. The input x to the model is a 2D point cloud representation of the scene, which is obtained by projecting the depth information from the camera into the world frame, using the camera parameters. The input goes through the encoder obtaining a latent vector, to which we concatenate the action orientation parameters vector q. After going through the decoder, where the skip-connections of the U-Net are applied, we obtain a map of the same size as the input, which we pass through a sigmoid layer to obtain outputs in $[0, 1]$. At each point, the output represents the parameter of a Bernoulli distribution that gives the probability of successfully applying the affordance primitive with the corresponding pixel position p and orientation q. For computing information gain estimates for affordance discovery, we instantiate a lightweight ensemble, where the ensemble shares the encoder but multiple decoders. We found that sharing the encoder overall simplifies training (higher accuracy) while still allowing us to compute the information gain estimates with reduced computation overhead [42].

We refer to θ as the set of parameters of one member of the ensemble, which represents our discrete space of parameters Θ. Each member of the ensemble is trained by minimizing a binary cross-entropy loss, where the targets are the success labels of the actions.

$$\begin{aligned} \mathcal{L}(\theta) = \sum_n & -\log p(b = \text{success}|x_n, a_n, \theta) \cdot b_n \\ & -\log p(b = \text{fail}|x_n, a_n, \theta) \cdot (1 - b_n) \end{aligned} \tag{6}$$

Following previous work [43], [44], we do not employ bootstrapping, as it can degrade the single model's performance.

Information-driven sampling. In our model, a set of model parameters θ can be obtained by uniformly sampling models from the set of ensemble parameters Θ, so that the information radius can be rewritten as:

$$\begin{aligned} I(x, a) &= JSD(p(b|x, a, \theta)|\theta \sim \Theta) = \\ &= H\left(\frac{1}{N}\sum_{i=1}^{N} p(b|x, a, \theta_i)\right) - \frac{1}{N}\sum_{i=1}^{N} H(p(b|x, a, \theta_j)). \end{aligned} \tag{7}$$

Crucially, the entropy computation is done per pixel, with minimal computation burden.

In order to foster exploration further, we opt for sampling the expected reward $\hat{r}(x, a)$ using a randomly sampled model from the ensemble. This technique, which can be interpreted as a form of Thompson sampling [45] with ensembles [46], [47], can particularly impact the initial phase of learning, when the differently initialized models tend to focus on different areas of the scene.

Robust evaluation. After the training stage, for evaluation, we can use the information gain as a measure of the amount of missing information with respect to a certain interaction. This measure can be used to select more robust actions, following a strategy of "pessimism" [48], [49].

The information $I(x, a)$, as described in Equation 3, can be subtracted from the expected rewards, obtaining a more conservative estimate of the action's value:

$$\arg\max_a [\hat{r}(x, a) - c_{\text{eval}} \cdot I(x, a)] \tag{8}$$

where the coefficient $c_{\text{eval}} \geq 0$ controls how conservative the actions should be. Conversely with respect to Equation 4, this equation satisfies a lower confidence bound with respect

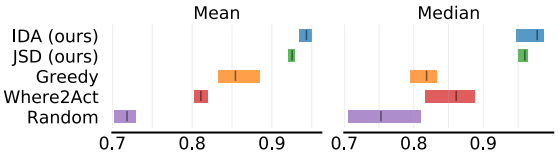


Fig. 2: **ManiSkill2 performance.** Affordance success aggregated across ManiSkill2 tasks and runs.

	IDA (ours)	JSD (ours)	Greedy	Where2Act	Random	C-GraspNet
Grasp Cube	**0.99**±0.01	0.96±0.02	**0.97**±0.04	0.95±0.01	0.81±0.04	0.99
Grasp YCB	**0.91**±0.02	0.85±0.03	0.82±0.03	0.66±0.05	0.50±0.05	0.43
Grasp EGAD	**0.86**±0.02	**0.83**±0.03	0.76±0.11	0.70±0.04	0.57±0.08	0.52
Stack Cube	**0.99**±0.01	**0.99**±0.01	**0.99**±0.01	0.88±0.03	0.96±0.04	-
Open Drawer	**0.98**±0.05	**1.00**±0.00	0.73±0.14	0.86±0.06	0.75±0.09	-

TABLE I: **Performance per task.** Dissecting performance per task on the ManiSkill2 tasks (mean ± standard deviation over 5+ seeds).

to the expected reward, which entails pessimism in the face of lack of information [50]. During the evaluation process, in order to sample our most robust expected reward predictions, we use the mean of the affordance probability maps, obtained by averaging the probability outputs of all models in the ensemble, i.e. $\hat{r}(x,a) = \frac{1}{N}\sum_{i=1}^{N} p(b|x,a,\theta_i)$.

V. Experiments

With our empirical evaluation, we aim to answer the following questions: *i)* what are the effects of adopting IDA for the affordance discovery problem in terms of data efficiency and final performance, *ii)* what is the impact of employing the JSD term for sampling informative actions. We perform experiments both in simulated and real-world settings. We also perform some ablations on the components of IDA.

A. Simulation experiments

To test our approach in an interactive simulation setting, we adopt environments from the ManiSkill2 benchmark [51]. The scene is recorded by an external RGBD camera that points at the scene, showing the robot and workspace in front of it. During training the agent alternates between environment interactions, where it tests the action primitive in an *interaction*, and updates the affordance model after each interaction. An interaction roughly requires 100 simulation steps. Every n interactions, where $n \in \{$ 50, 250, 500, 2500, 10000 $\}$, we evaluate the agent's performance.

We consider the following set of primitives and corresponding affordances:

- *Grasping:* success is achieved if the object can be lifted without falling out of the gripper.
- *Stacking:* success is achieved if the object stably remains on the other object after the gripper releases the held object and withdraws.
- *Opening:* success is achieved if a part of the object is pulled and that classifies as an opening part, e.g., the drawer in a cabinet.

For all primitives, we adopt a discrete set of orientations with 6-DoF for grasping (429 angles) and 3-DoF for stacking and opening (8 angles). Actions are executed by performing (linear) motion planning. Further details are given in Appendix.

Results. We present evaluation results, in terms of affordance success rate, for IDA and the following baselines:

- *JSD (ours)*: this method only samples actions according to information gain during training, and uses the ensemble predictions for evaluation. It represents an ablation of IDA, using no UCB and no robustness during evaluation and showcases the performance of the information gain objective;
- *Where2Act*: this baseline mimics the Where2Act sampling strategy [10], while employing our U-net based architecture. The agent randomly samples actions during the first 5000 training steps, while during the latter 5000 training steps, it samples from the softmax probability distribution over all possible actions, essentially doing a form of Boltzmann exploration [52]. Note that in our experiments we use a mounted robot arm, as opposed to the flying gripper adopted in the original work.
- *Greedy*: this method uses a U-Net and selects the highest probability affordance actions, both for training and evaluation;
- *Random*: the agent randomly samples affordance actions during training and selects the highest probability affordance action for evaluation, like Greedy.

For the grasping tasks, we also compare to *Contact-GraspNet* [13], a grasping approach that has been trained on 17.7 million simulated grasps across 8872 meshes, for which we use the original open-source pre-trained models and code.

In Figure 2, we present bootstrapped statistics and confidence intervals for the mean and the median performance after 10k steps aggregated across all tasks and runs (5+ seeds per method and task), using 50k repetitions (see [53]). We observe that IDA reaches the highest mean, with high confidence, and the highest median, followed by JSD (ours). Both our methods largely outperform all the other baselines.

In Table I, we present the mean and standard deviation statistics per task. We observe that on the most difficult tasks, Grasp EGAD (large number of variations) and Open Drawer (harder exploration), IDA and JSD have the largest advantage, corroborating the idea that information-driven affordance discovery leads to higher final performance. Compared with Contact-GraspNet in the grasping tasks, we also observe that all interactive learning methods, except for Random, eventually outperform pre-training on large synthetic datasets.

Finally, in Figure 3, we compare performance over time aggregated across all tasks and runs, using bootstrapped statistics and confidence intervals with respect to the mean [53]. We observe that while JSD's final performance is close to IDA's one, the Greedy approach performs better in the early stages of learning. This shows that the reward term, which we incorporate in IDA using a UCB, can be useful when sampling actions. We speculate the reason for

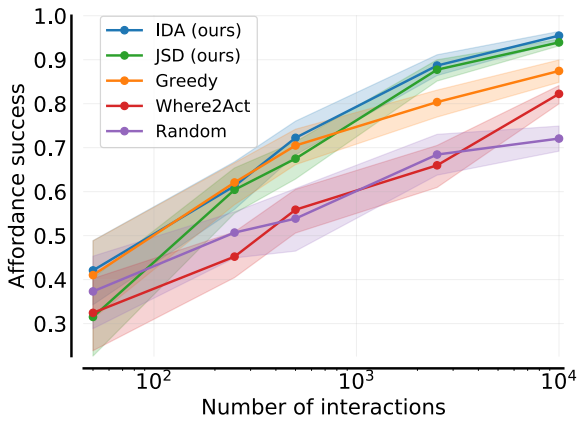


Fig. 3: **Performance over time.** The affordance success rate in the evaluation stage increases over the number of interactions, averaged over all tasks (5+ seeds per task).

this being that during the first stages of learning, it allows the agent to learn and consolidate knowledge about easy affordances, before moving to more complex ones.

Reward-free ablation. The objective of IDA is made of two components: a reward-driven term and an information-driven term. *What happens when we ignore the reward-driven term?* In that case, the other approaches (including Where2Act) rely on random sampling, while IDA relies on the JSD strategy. We more detailedly compare these two strategies over time on the three ManiSkill grasping tasks (Cube, YCB and EGAD), along with a *Random + Ensemble* strategy, which aims to reduce the gap between the two methods, by learning an ensemble model for evaluation. In Figure 4, we observe that although the performance of Random increases, during the later stages, when employing an ensemble for evaluation, the JSD sampling strategy largely remains the most effective one.

Interestingly, JSD's performance is lower at the first evaluation (after only 50 interactions). This is probably due to

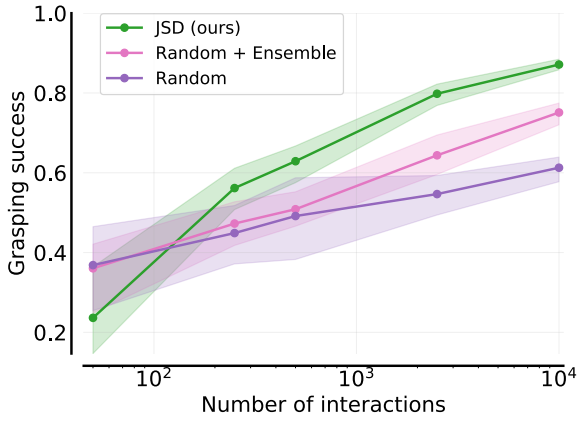
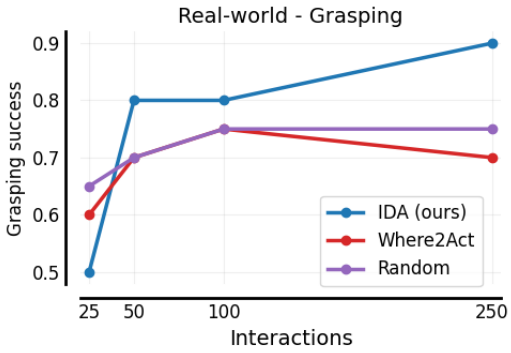


Fig. 4: **Reward-free ablation.** Comparing reward-free affordance discovery methods over time. (5+ seeds).



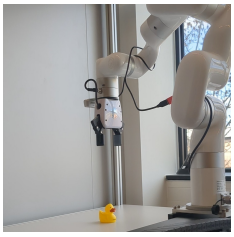


Fig. 5: **Real-world results and setup.** IDA learns to grasp objects faster than other approaches, achieving up to 90% grasping success, on a UFACTORY xArm 6 platform.

the fact that random samples possess greater diversity at the beginning of training when the JSD information estimates, and thus action sampling, are not yet well-directed. As we show in Table II, the use of rewards and of ensemble sampling in IDA both help in mitigating the issue.

IDA	IDA (no ens. sampl.)	JSD
0.45	0.39	0.25

TABLE II: Grasping success (avg) after 50 interactions.

B. Real-world experiments

To confirm that our simulation results hold in the real world, we deployed and tested our method on a grasping task using a UFACTORY xArm 6 with UFACTORY gripper.

We studied the grasping primitive using a set of four toy objects of varying size. The objects were presented repeatedly in order during both training and evaluation and manually placed at random locations and orientations in the workspace of the robot arm before each grasp attempt.

Differently from the simulation setup, the grasping angle was restricted to vertical grasps with eight discrete rotations (3 DoF) and a wrist-mounted RGBD camera is used to obtain the inputs. This also shows our method works indipendently of the viewpoint. The affordance model is always trained from random initialization (no prior knowledge) and evaluated at 25, 50, 100, 250 episodes. Evaluation is done for 20 iterations, i.e., each of the four objects is presented five times per evaluation round. The entire experiment, including training and evaluation, takes around 90 minutes.

Results. In Figure 5 we compare IDA with two baselines: the Where2Act and the Random action sampling strategies. We observe the real-world results for grasping closely match

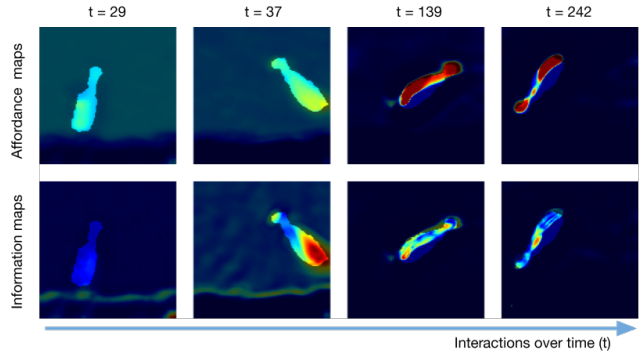


Fig. 6: **Real-world affordance and information maps.** Affordance and information maps, showing for each pixel the highest value across all possible gripper orientations.

the simulation results, as our agent efficiently learns to grasp the objects. The final performance obtained is 90% grasping success, which is considerably higher than the baselines. Nonetheless, we experienced lower performance than Random and Where2Act (also using random sampling during the initial steps) after only 25 interactions. This result reflects what we found in the reward-free ablation experiments, and how to further mitigate this early-learning effect will be the object of future studies.

Visualizations. To provide additional insights into how IDA learns visual affordances and estimates information gain over time, in Figure 6, we show how the affordance and information maps from our method evolve over time while learning to grasp (a bowling pin) on the real robot arm. We observe that the affordance probabilities are uniformly distributed at the beginning of the training ($t = 29$). Later, the information maps recommend exploring grasping point towards the edge of the object ($t = 37$, $t = 139$), leading the agent to eventually learn that areas close to the edges are easier to grasp ($t = 242$) as they have a less slippery surface.

VI. Conclusion

We presented IDA, a method that fosters the discovery of affordances for robotics manipulation. IDA showcases strong performances across several tasks in ManiSkill2 and the ability to quickly learn to grasp objects in the real world. We empirically showed the importance of well-directed action sampling, to achieve higher affordance success and we analyzed several components of our method. One limitation of the approach tested is that it relies on motion planning for precise affordance execution. While this facilitates exploration, especially in the early stages of learning, as the actions executed by the agent are more stable and reliable, the issue should be addressed in future work, aiming to provide more adaptive policies, e.g. using reinforcement learning. We also aim to extend our work towards the development of an end-to-end system able to solve longer horizon tasks, potentially instantiating a hierarchical controller on top of the possible affordance actions [54] or employing large language models to decide which affordances should be executed towards the solution of tasks [5], [55].

Appendix

A. Information radius derivation

As per Equation 2, we are interested in measuring the expected information given by a certain state-action:

$$I(x,a) = \sum_b p(b|x,a) D_{\mathrm{KL}}[p(\Theta|b,x,a) \| p(\Theta)],$$

where we can use the KL definition to obtain:

$$I(x,a) = \sum_b p(b|x,a) \sum_\theta p(\theta|b,x,a) \log \frac{p(\theta|b,x,a)}{p(\theta)}.$$

Using the Bayes rule ($p(\theta|b,x,a) = \frac{p(b|x,a,\theta)p(\theta|x,a)}{p(b|x,a)}$) and the observation $p(\theta|x,a) = p(\theta)$ (as an incomplete transition provides no information about the affordance model) we get:

$$I(x,a) = \sum_b \sum_\theta p(b|x,a,\theta) p(\theta) \log \frac{p(b|x,a,\theta)}{p(b|x,a)}.$$

By splitting the logarithm and applying the law of total probability we derive:

$$I(x,a) = \sum_\theta p(\theta) \sum_b p(b|x,a,\theta) \log p(b|x,a,\theta) - \sum_b \sum_\theta p(\theta) p(b|x,a,\theta) \log \Big(\sum_\theta p(\theta) p(b|x,a,\theta) \Big),$$

that given entropy $H(Y) = -\sum_y p(y) \log p(y)$ becomes:

$$I(x,a) = -\mathbb{E}_{p(\theta)}[H(p(b|x,a,\theta))] + H(\mathbb{E}_{p(\theta)}[p(b|x,a,\theta)])$$

that is the JSD we present in Equation 3.

B. Training details

Model. The ensemble model is a U-Net-like model made of a shared encoder and multiple decoders. The encoder has a convolutional block with 8 filters, kernel size 3 and stride 1, followed by two blocks with 16 filters, kernel size 5 and stride 2. The decoder follows the same structure reversed, but replacing stride with bilinear upsampling layers and applying skip connections. The inputs' resolution for the model is 128x128. The number of networks in the ensemble is 5. The model is updated 5 times after each interaction, using ADAM with a learning rate of $3{\cdot}10^{-4}$. The batch size used is 256. To start filling the replay buffer with experience, 10 random actions are sampled at the beginning of training.

Action sampling. To avoid sampling invalid points we filter out some regions of the environment from the output of the model, including points that are behind and above the robot's base and points that are on the surface of the workspace or below. Before summing the information gain values with the predicted affordance success reward, we normalize the information gain maps to have values in [0,1]. As default values, we use $c_{\text{expl}} = 0.3$ and $c_{\text{eval}} = 0.1$.

C. Evaluation details

- *Grasp Cube*: 100 randomized positions.
- *Grasp EGAD*: 1 interaction for each of the 1600 objects.
- *Grasp YCB*: 5 interactions for each of the 74 objects.
- *Stack Cube*: 100 randomized interactions
- *Open Drawer*: 5 interactions for each cabinet (6 models)

REFERENCES

[1] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, T. Lillicrap, and D. Silver, "Mastering atari, go, chess and shogi by planning with a learned model," *Nature*, vol. 588, no. 7839, pp. 604–609, dec 2020.

[2] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," 2021.

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[4] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[5] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, M. Yan, and A. Zeng, "Do as i can, not as i say: Grounding language in robotic affordances," 2022.

[6] T. Haarnoja, B. Moran, G. Lever, S. H. Huang, D. Tirumala, M. Wulfmeier, J. Humplik, S. Tunyasuvunakool, N. Y. Siegel, R. Hafner, M. Bloesch, K. Hartikainen, A. Byravan, L. Hasenclever, Y. Tassa, F. Sadeghi, N. Batchelor, F. Casarini, S. Saliceti, C. Game, N. Sreendra, K. Patel, M. Gwira, A. Huber, N. Hurley, F. Nori, R. Hadsell, and N. Heess, "Learning agile soccer skills for a bipedal robot with deep reinforcement learning," 2023.

[7] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," 2022.

[8] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "Daydreamer: World models for physical robot learning," 2022.

[9] J. J. Gibson, *The ecological approach to visual perception: classic edition*. Psychology press, 2014.

[10] K. Mo, L. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," 2021.

[11] J. Borja-Diaz, O. Mees, G. Kalweit, L. Hermann, J. Boedecker, and W. Burgard, "Affordance learning from play for sample-efficient policy learning," 2022.

[12] P. Mandikal and K. Grauman, "Learning dexterous grasping with object-centric visual affordances," 2021.

[13] M. Sundermeyer, A. Mousavian, R. Triebel, and D. Fox, "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes," 2021.

[14] R. Held and A. Hein, "Movement-produced stimulation in the development of visually guided behavior." *Journal of comparative and physiological psychology*, vol. 56, no. 5, p. 872, 1963.

[15] S. Levine, C. Finn, T. Darrell, and P. Abbeel, "End-to-end training of deep visuomotor policies," 2016.

[16] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, "Mt-opt: Continuous multi-task robotic reinforcement learning at scale," 2021.

[17] Y. Lu, K. Hausman, Y. Chebotar, M. Yan, E. Jang, A. Herzog, T. Xiao, A. Irpan, M. Khansari, D. Kalashnikov, and S. Levine, "Aw-opt: Learning robotic skills with imitation and reinforcement at scale," 2021.

[18] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2015.

[19] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," 2018.

[20] S. Ferraro, T. Van de Maele, P. Mazzaglia, T. Verbelen, and B. Dhoedt, "Computational optimization of image-based reinforcement learning for robotics," *Sensors*, vol. 22, no. 19, p. 7382, 2022.

[21] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, A. Wahid, V. Sindhwani, and J. Lee, "Transporter networks: Rearranging the visual world for robotic manipulation," 2022.

[22] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015.

[23] S. James and A. J. Davison, "Q-attention: Enabling efficient learning for vision-based robotic manipulation," 2022.

[24] R. Liu, J. Lehman, P. Molino, F. P. Such, E. Frank, A. Sergeev, and J. Yosinski, "An intriguing failing of convolutional neural networks and the coordconv solution," 2018.

[25] K. Mo, Y. Qin, F. Xiang, H. Su, and L. Guibas, "O2o-afford: Annotation-free large-scale object-object affordance learning," 2021.

[26] S. Belkhale and D. Sadigh, "Plato: Predicting latent affordances through object-centric play," 2022.

[27] A. Khazatsky, A. Nair, D. Jing, and S. Levine, "What can i do here? learning new skills by imagining visual affordances," 2021.

[28] R. Platt, "Grasp learning: Models, methods, and performance," 2022.

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

[30] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/scirobotics.aau4984

[31] D. Wang, D. Tseng, P. Li, Y. Jiang, M. Guo, M. Danielczuk, J. Mahler, J. Ichnowski, and K. Goldberg, "Adversarial grasp objects," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, 2019, pp. 241–248.

[32] M. Danielczuk, A. Balakrishna, D. S. Brown, S. Devgon, and K. Goldberg, "Exploratory grasping: Asymptotically optimal algorithms for grasping challenging polyhedral objects," 2020.

[33] H. Y. Li, M. Danielczuk, A. Balakrishna, V. Satish, and K. Goldberg, "Accelerating grasp exploration by leveraging learned priors," 2020.

[34] L. Fu, M. Danielczuk, A. Balakrishna, D. S. Brown, J. Ichnowski, E. Solowjow, and K. Goldberg, "Legs: Learning efficient grasp sets for exploratory grasping," 2022.

[35] M. Laskey, J. Mahler, Z. McCarthy, F. T. Pokorny, S. Patil, J. van den Berg, D. Kragic, P. Abbeel, and K. Goldberg, "Multi-armed bandit models for 2d grasp planning with uncertainty," in *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, 2015, pp. 572–579.

[36] Q. Lu, M. V. der Merwe, and T. Hermans, "Multi-fingered active grasp learning," 2020.

[37] M. Dalal, D. Pathak, and R. Salakhutdinov, "Accelerating robotic reinforcement learning via parameterized action primitives," 2021.

[38] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 14, no. 2, pp. 149–160, Jun 1969. [Online]. Available: https://doi.org/10.1007/BF00537520

[39] N. Jardine and R. Sibson, "Mathematical taxonomy, john wiley and sons: London," 1971.

[40] P. Auer, N. Cesa-Bianchi, and P. Fischer, "Finite-time analysis of the multiarmed bandit problem," *Machine Learning*, vol. 47, no. 2, pp. 235–256, May 2002. [Online]. Available: https://doi.org/10.1023/A:1013689704352

[41] J.-Y. Audibert, R. Munos, and C. Szepesvári, "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits," *Theoretical Computer Science*, vol. 410, no. 19, pp. 1876–1902, 2009.

[42] O. Laurent, A. Lafage, E. Tartaglione, G. Daniel, J.-M. Martinez, A. Bursuc, and G. Franchi, "Packed-ensembles for efficient uncertainty estimation," 2023.

[43] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," 2017.

[44] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. V. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," 2019.

[45] D. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "A tutorial on thompson sampling," 2020.

[46] X. Lu and B. V. Roy, "Ensemble sampling," 2023.

[47] I. Osband, C. Blundell, A. Pritzel, and B. V. Roy, "Deep exploration via bootstrapped dqn," 2016.

[48] A. Kumar, A. Zhou, G. Tucker, and S. Levine, "Conservative q-learning for offline reinforcement learning," 2020.

[49] L. Shi, G. Li, Y. Wei, Y. Chen, and Y. Chi, "Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity," 2022.

[50] P. Rashidinejad, B. Zhu, C. Ma, J. Jiao, and S. Russell, "Bridging offline reinforcement learning and imitation learning: A tale of pessimism," 2023.

[51] J. Gu, F. Xiang, X. Li, Z. Ling, X. Liu, T. Mu, Y. Tang, S. Tao, X. Wei, Y. Yao, X. Yuan, P. Xie, Z. Huang, R. Chen, and H. Su, "Maniskill2: A unified benchmark for generalizable manipulation skills," 2023.

[52] N. Cesa-Bianchi, C. Gentile, G. Lugosi, and G. Neu, "Boltzmann exploration done right," 2017.

[53] R. Agarwal, M. Schwarzer, P. S. Castro, A. Courville, and M. G. Bellemare, "Deep reinforcement learning at the edge of the statistical precipice," *Advances in Neural Information Processing Systems*, 2021.
[54] D. Hafner, K.-H. Lee, I. Fischer, and P. Abbeel, "Deep hierarchical planning from pixels," 2022.
[55] K. Lin, C. Agia, T. Migimatsu, M. Pavone, and J. Bohg, "Text2motion: From natural language instructions to feasible plans," 2023.