Dexterous Robotic Grasping with Object-Centric Visual Affordances

Priyanka Mandikal
UT Austin
mandikal@cs.utexas.edu

Kristen Grauman
UT Austin & Facebook AI Research
grauman@fb.com

Abstract

Dexterous robotic hands are appealing for their agility and human-like morphology, yet their high degree of freedom makes learning to manipulate challenging. We introduce an approach for learning dexterous grasping. Our key idea is to embed an object-centric visual affordance model within a deep reinforcement learning loop to learn grasping policies that favor the same object regions favored by people. Unlike traditional approaches that learn from human demonstration trajectories (e.g., hand joint sequences captured with a glove), the proposed prior is *object-centric* and *image-based*, allowing the agent to anticipate useful affordance regions for objects unseen during policy learning. We demonstrate our idea with a 30-DoF five-fingered robotic hand simulator on 40 objects from two datasets, where it successfully and efficiently learns policies for stable grasps. Our affordance-guided policies are significantly more effective, generalize better to novel objects, and train 3× faster than the baselines. Our work offers a step towards manipulation agents that learn by watching how people use objects, without requiring state and action information about the human body. Project website: http://vision.cs.utexas.edu/projects/graff-dexterous-affordance-grasp.

1 Introduction

Robot grasping is a vital prerequisite for complex manipulation tasks. From wielding tools in a mechanics shop to handling appliances in the kitchen, grasping skills are essential to everyday activity. Meanwhile, common objects are designed to be used by human hands (see Fig. 1). Hence, there is increasing interest in dexterous, anthropomorphic robotic hands with multi-jointed fingers [5, 16, 7, 25, 1, 14, 2]. Unlike simpler end effectors such as a parallel-jaw gripper, a dexterous hand has the potential for fine-grained manipulation. Furthermore, because its morphology agrees with that of the human hand, in principle it is readily compatible with the many real-world objects built for people's use. Of particular interest is *functional grasping*, where the robot should not merely lift an object, but do so in such a way that it is primed to use that object [4, 9]. For instance, picking up a pan by its base for cooking or gripping a hammer by its head for hammering is contrary to functional use.

Learning to perform functional grasping with a dexterous hand is highly challenging. Traditional parallel jaw grippers can be trained to grasp objects by simply predicting a 6-DoF pose for the end effector followed by model-based planning [23, 13, 11, 21]. Typical dexterous hand models, however, have 24 degrees of freedom (DoF) across the articulated joints, presenting high-dimensional state and action spaces to master. As a result, a reinforcement learning approach trained purely on robot experience faces daunting sample complexity. Existing methods attempt to control the complexity by concentrating on a single task and object of interest (e.g., Rubik's cube [1]) or by incorporating explicit human demonstrations [19, 5, 16, 7, 25, 20, 15]. For example, a human "teacher" wearing a glove instrumented with location and touch sensors can supply trajectories for the agent to imitate [16, 7, 15]. While inspiring, this strategy is nonetheless expensive in terms of human

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

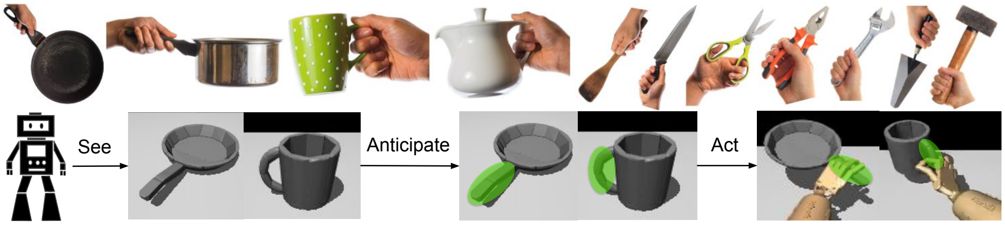

Figure 1: **Main idea.** We aim to learn deep RL grasping policies for a dexterous robotic hand, injecting a visual affordance prior that encourages using parts of the object used by people for functional grasping. Given an object image (left), we predict the affordance regions (center), and use them to influence the learned policy (right). The key upshots of our approach are better grasping, faster learning, and generalization to successfully grasp objects unseen during policy training.

time, the possible need to wear specialized equipment, and the close coupling between the person's hand trajectory and the target object of interest, which limits generalization.

Towards overcoming these limitations, we propose a new approach to learning to grasp with a dexterous robotic hand. Our key insight is to shift from *person-centric* physical demonstrations to *object-centric* visual affordances. Rather than learn to mimic the sequential states/actions of the human arm/hand as it picks up an object, we learn the regions of objects most amenable to a human interaction, in the form of an image-based affordance prediction model. We embed this visual affordance model (a convolutional neural network) within a deep reinforcement learning framework in which the agent is rewarded for touching the afforded regions with its hand. In this way, the agent has a “human prior" for how to approach an object, but is free to discover its exact grasping strategy through closed loop experience. Aside from accelerating learning, a critical advantage of the proposed object-centric design is generalization: the learned policy generalizes to unseen object instances because the image-based module can anticipate their affordance regions (see Figure 1).

In experiments with 40 objects, we show that our approach yields significantly better quality grasps compared to other pure RL models unaware of the human affordance prior. The learned grasping policies are stable under hostile external forces and robust to changes in the objects' physical properties (mass, scale). Furthermore, our approach significantly improves the sample efficiency of learning process, for a 3× speed up in training despite having no state-action demonstrations. Finally, we show our agent generalizes to pick up object instances never encountered in training. Our results offer a promising step towards agents that learn by *watching* how people use real-world objects, without requiring information about the human operator's body.

2 Approach

Our goal is to learn dexterous robotic grasping policies influenced by object-centric grasp affordances from images. Our proposed model, called GRAFF for *Grasp-Affordances*, consists of two stages (Fig. 2). First, we train a network to predict affordance regions from static images (Sec. 2.1). Second, we train a dynamic grasping policy using the learned affordances (Sec. 2.2).

2.1 Affordance Anticipation From Images

We first design a perception model to infer object-centric grasp affordance regions from static images (Fig. 2a), which has the advantage of providing human priors while generalizing to new objects. We train the affordance model with images having ground truth functional grasp regions obtained from ContactDB [3]. ContactDB contains 3D scans of 50 household objects along with real-world human contact maps captured using thermal cameras. We consider contact maps corresponding to the *use* intent and exclude objects having bimanual grasps, which yields 16 total objects. We port the 3D models into the MuJoCo physics simulator [22] and render them on a tabletop to create an image training set of 15k image-affordance pairs $\{X, Y\}$ (Suppl A.1). Our goal now is to train an affordance anticipation network $G : X \to Y$ that will infer the grasp affordance regions from an individual image. We pose the affordance learning problem as a segmentation task to predict binary per-pixel labels, and approximate G with a CNN (results in Suppl Figs. A & D). We now have a simple but effective model to infer object-centric grasp affordances from static images, which we will use below to guide a dexterous grasping policy.

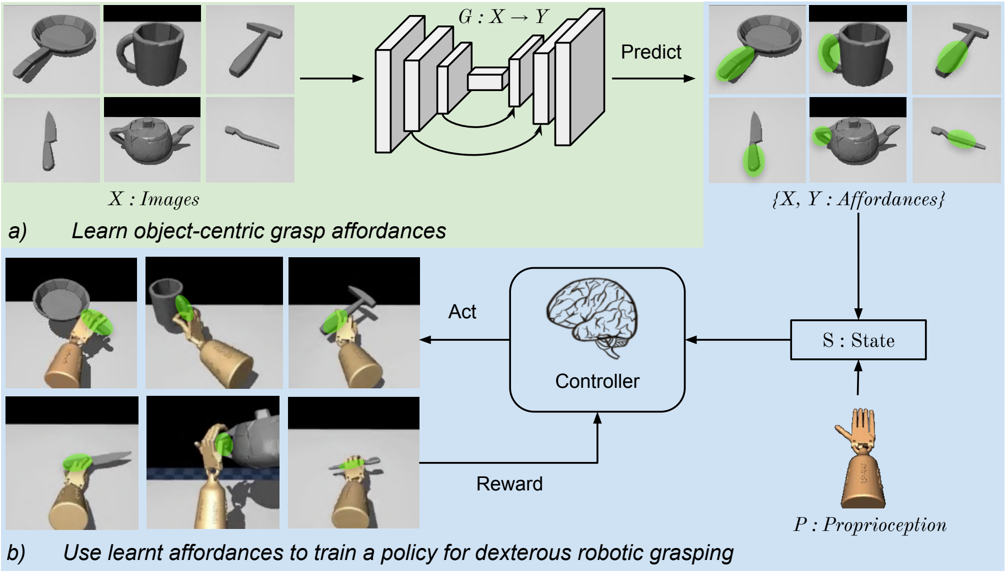
$G : X \to Y$
Predict
$X : Images$
$\{X, Y : Affordances\}$
a) Learn object-centric grasp affordances
Act
Controller
S : State
Reward
$P : Proprioception$
b) Use learnt affordances to train a policy for dexterous robotic grasping

Figure 2: **Overview of our GRAFF model.** a) In Stage I, we train an affordance prediction model that predicts object-centric grasp affordances given an image. b) In Stage II, we train an RL policy that leverages these affordances along with other visuomotor sensory inputs (RGB-D image + hand joint variables) to learn a stable grasping policy.

2.2 Dexterous Grasping using Visual Affordances

We want a controller that can intelligently process sensory inputs and execute successful grasps for a variety of objects with diverse geometries. Towards this end, we develop a deep model-free reinforcement learning model for dexterous grasping. Our robot model assumes access to visual sensing and proprioception, as well as 3D point tracking. However, the agent does not have access to world dynamics, full object state, or the reward function. Given the large action and state spaces, sample efficiency is a significant challenge. We show how the visual affordance model streamlines policy exploration to focus on object regions most amenable to grasping. See Fig. 2b & Suppl A.2.

Problem formulation We pose the problem of grasp acquisition as a finite-horizon discounted Markov decision process (MDP), with state space $\mathcal{S}$, action space $\mathcal{A}$, state transition dynamics $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$, initial state distribution ρ_0, reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and discount factor $\gamma \in (0, 1]$. The objective is to maximize the expected discounted reward to determine the optimal policy. We use an actor-critic model to estimate state values and policy distribution at each time step.

State space The work space of the robot consists of an object on a table at a random orientation. The state space consists of the visuomotor inputs used to train the control policy: $\mathcal{S} = \{X, Y, P, D\}$. The visual input at time t consists of an RGB-D image $x_t \in X$ captured by an egocentric hand-mounted camera. The affordance input $y_t \in Y$ is the binary affordance map inferred from the image, $y_t = G(x_t)$. The proprioception input $p_t \in P$ is the positions and velocities of each DoF in the hand. The distance input $d_t \in D$ is the distance between the agent's hand and the object affordance region.

Action space We use a 30-DoF position-controlled anthropomorphic hand from the Adroit platform [10] as our manipulator. It consists of a 24-DoF five-fingered hand attached to a 6-DoF arm. Hence, our action space consists of 30 continuous position values, which are predicted by sampling from a multivariate Gaussian whose parameters are returned by the policy π.

Reward function The reward function should not only signal a successful grasp, but also guide the exploration process to focus on graspable object regions. To realize this, we combine two rewards: R_{succ} (positive reward when the object is lifted off the table) and R_{aff} (negative reward denoting the hand-affordance contact distance), with an additional entropy term. Our total reward function is: $r = \alpha R_{succ} + \beta R_{aff} + \eta R_{entropy}$. Through R_{aff}, the agent is incentivized to explore areas of the object that lie within the affordance region. The object-centric formulation poses no constraints on the hand pose, and can be seen as softer supervision than that employed in imitation learning for manipulation which requires kinesthetic teaching [18, 19] or tele-operation [6, 16, 12].

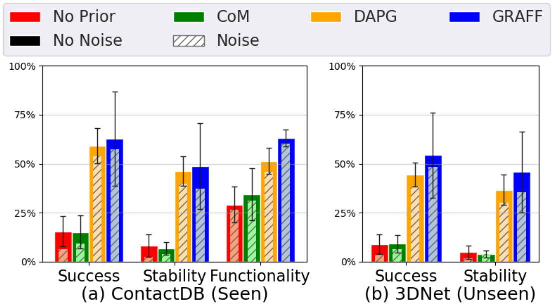

Figure 3: **Performance on ContactDB and 3DNet**.

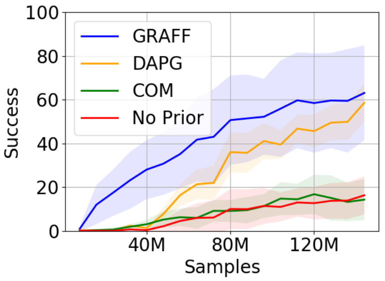

Figure 4: **Training curves.**

3 Experiments

Datasets We validate our approach on 40 objects from two datasets: ContactDB [3] and 3DNet [24]. We use objects from 3DNet that roughly align with the objects in ContactDB. We classify 3DNet objects into *known* (classes present in ContactDB) and *unknown* (classes related to ContactDB but not present). We train a single policy across all ContactDB objects.

Comparisons We devise two pure RL baselines that lack the proposed affordances: (**1**) NO PRIOR: uses the lifting success reward only. (**2**) COM: uses the center of mass as a prior, which may lead to stable grasps [17, 8], by penalizing the hand-CoM distance. (**3**) DAPG [16]: This is a hybrid imitation+RL model that uses motion-glove demonstrations (strong state-action supervision, ref. Supp C). GRAFF uses inferred object-centric affordances to guide the policy.

Metrics We use two metrics: (**1**) Grasp Success: Object has been lifted off the table by the hand for at least the last 50 time steps (**2**) Grasp Stability: After episode completion, we apply perturbing forces of 5 Newtons in six orthogonal directions to the object, and check if the object remains held.

Grasping seen objects from ContactDB GRAFF outperforms both pure RL baselines consistently on both metrics while also beating the more intensely supervised imitation+RL method DAPG [16] (Fig. 3a). From qualitative results (Fig. 5a), GRAFF can successfully grasp objects at the anticipated affordance regions, while the baselines fail to grasp objects with complex geometries. Our method also more effectively executes functional grasps on unseen objects thanks to its image-based model. Furthermore, GRAFF displays better sample efficiency than the RL baselines (Fig. 4).

Grasping unseen objects from 3DNet We evaluate the policy on generalization to unseen objects from 3D-Net (Fig. 3b). We outperform all three baselines by a large margin in both grasp success and stability. The key factor is our visual affordance idea: the anticipation model generalizes sufficiently to new object shapes so as to provide a useful object-centric prior. Fig. 5b shows samples.

4 Conclusion

We present an approach to learn dexterous robotic grasping with object-centric visual affordances. Breaking away from the norm of expert demonstrations, our GRAFF method uses an image-based affordance model to focus the agent's attention on “good places to grasp". To our knowledge, ours is the first work to demonstrate closed-loop RL policies learned with visual affordances. The key advantages of our design are its learning speed and ability to generalize policies to unseen (visually related) objects. We see the results as encouraging evidence for manipulation agents learning faster with more distant human supervision.

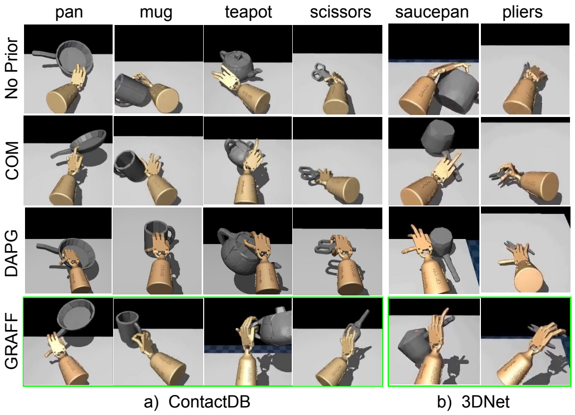

Figure 5: **Grasping performance.** Example frames from a) seen objects in ContactDB and b) novel objects in 3D-Net. Our affordance-based GRAFF is able to grasp both seen and novel objects at their functional grasp locations, while the two pure-RL baselines either fail to learn successful grasps (mug, teapot, cup, saucepan) or grasp at non-functional regions (pan, knife, scissors). Despite GRAFF's weaker supervision, it grasps as well as DAPG on known objects and, thanks to the affordance model, generalizes better to unseen ones. Visit the project page for full episodes.

References

[1] Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[2] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *The International Journal of Robotics Research*, 39(1):3–20, 2020.

[3] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[4] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. ContactGrasp: Functional multi-finger grasp synthesis from contact. *arXiv:1904.03754*, 2019.

[5] Abhishek Gupta, Clemens Eppner, Sergey Levine, and Pieter Abbeel. Learning dexterous manipulation for a soft robotic hand from human demonstrations. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[6] Ankur Handa, Karl Van Wyk, Wei Yang, Jacky Liang, Yu-Wei Chao, Qian Wan, Stan Birchfield, Nathan Ratliff, and Dieter Fox. Dexpilot: Vision based teleoperation of dexterous robotic hand-arm system. *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[7] Divye Jain, Andrew Li, Shivam Singhal, Aravind Rajeswaran, Vikash Kumar, and Emanuel Todorov. Learning deep visuomotor policies for dexterous hand manipulation. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[8] Dimitrios Kanoulas, Jinoh Lee, Darwin G Caldwell, and Nikos G Tsagarakis. Center-of-mass-based grasp pose adaptation using 3d range and force/torque sensing. *International Journal of Humanoid Robotics*, 15(04):1850013, 2018.

[9] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters*, 2020.

[10] Vikash Kumar, Zhe Xu, and Emanuel Todorov. Fast, strong and compliant pneumatic actuation for dexterous tendon-driven hands. In *IEEE international conference on robotics and automation*, 2013.

[11] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International Journal of Robotics Research*, 37(4-5):421–436, 2018.

[12] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. *arXiv preprint arXiv:1811.02790*, 2018.

[13] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *IEEE International Conference on Computer Vision (ICCV)*, 2019.

[14] Anusha Nagabandi, Kurt Konoglie, Sergey Levine, and Vikash Kumar. Deep dynamics models for learning dexterous manipulation. *Conference on Robot Learning (CoRL)*, 2019.

[15] I. Radosavovic, X. Wang, L. Pinto, and J. Malik. State-only imitation learning for dexterous manipulation. *arXiv:2004.04650v1*, 2020.

[16] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *Robotics: Science and Systems (RSS)*, 2018.

[17] Máximo A Roa and Raúl Suárez. Grasp quality measures: review and performance. *Autonomous robots*, 38(1):65–88, 2015.

[18] Pratyusha Sharma, Lekha Mohan, Lerrel Pinto, and Abhinav Gupta. Multiple interactions made easy (mime): Large scale demonstrations data for imitation. In *CoRL*, 2018.

[19] Pratyusha Sharma, Deepak Pathak, and Abhinav Gupta. Third-person visual imitation learning via decoupled hierarchical controller. In *NeurIPS*, 2019.

[20] Jaeyong Sung, Seok Hyun Jin, and Ashutosh Saxena. Robobarista: Object part-based transfer of manipulation trajectories from crowd-sourcing in 3d pointclouds. In *International Symposium on Robotics Research (ISRR)*, 2015.

[21] Andreas ten Pas, Marcus Gualtieri, Kate Saenko, and Robert Platt. Grasp pose detection in point clouds. *The International Journal of Robotics Research*, 36(13-14):1455–1473, 2017.

[22] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.

[23] Jonathan Tremblay, Thang To, Balakumar Sundaralingam, Yu Xiang, Dieter Fox, and Stan Birchfield. Deep object pose estimation for semantic robotic grasping of household objects. *Conference on Robot Learning (CoRL)*, 2018.

[24] Walter Wohlkinger, Aitor Aldoma Buchaca, Radu Rusu, and Markus Vincze. 3DNet: Large-Scale Object Class Recognition from CAD Models. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2012.

[25] Henry Zhu, Abhishek Gupta, Aravind Rajeswaran, Sergey Levine, and Vikash Kumar. Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3651–3657. IEEE, 2019.