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$\times$ 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 [12, 37, 15, 53, 1, 31, 3]. 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 [7, 19]. 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. Typical hand models 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 [44, 12, 37, 15, 53, 45, 36]. For example, a human "teacher" wearing a glove instrumented with location and touch sensors can supply trajectories for the agent to imitate [37, 15, 36]. While inspiring, this strategy is nonetheless expensive in terms of human time, the possible need to wear specialized equipment, and the close coupling between the person's arm/hand trajectory and the target object of interest, which limits generalization.

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 it 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.

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 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).

Our main contribution is to learn closed loop dexterous grasping policies with object-centric visual affordances. We demonstrate our idea with the 30 DoF AdroitHand model [20] in the MuJoCo physics simulator [48]. We train the visual affordance model from images annotated for human grasp regions [6]. Importantly, image annotations are a much lighter form of supervision than state-action trajectories from expert demonstrations.

In experiments with 40 objects, we show 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$\times$ 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. For example, though trained to pick up a hammer, the model leverages partial visual regularities to pick up an axe. 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 Related Work

Grasping with planning Traditional analytical approaches use knowledge of the 3D object pose, shape, gripper configuration, friction coefficients, etc. to determine an optimal grasp [5, 8]. With the advent of deep neural nets, learning-based approaches to grasping have gained traction. A common protocol estimates the 6-DoF object pose, followed by model-based grasp planning [49, 28, 22, 46]. Image modules trained to detect successful grasps by parallel jaw grippers can accelerate the robot's learning [21, 38, 34, 22, 24, 29]. The above strategies are typically employed for simple pick-up actions (not functional grasps) with simple end-effectors like parallel jaw grippers or suction cups, for which a control policy is easier to codify. Some recent work explores related open-loop strategies with complex controllers, but, unlike our method, assume access to the full 3D model of the objects [10, 2, 4, 7, 42].

Reinforcement learning for closed-loop grasping Reinforcement learning (RL) models offer a counterpoint to the planning paradigm. Rather than break the task into two steps—static grasp synthesis followed by motion planning—the idea is to use closed-loop feedback control based on visual and/or contact sensing so the agent can dynamically update its strategy while accumulating new observations [16, 35, 26]. The proposed model is also closed-loop RL. However, unlike prior

work, we inject an object-centric affordance prior learned from human grasps. It boosts sample efficiency, particularly important for the complex action space of dexterous robotic hands.

Some impressive RL-based systems for dexterous manipulation tackle a specific task with a specific object, like solving Rubik's cube [1], shuffling Baoding balls [31], or reorienting a cube [3]. In contrast, our focus is on grasping and lifting objects, including novel instances, and again our injection of object-centric human affordances is distinct.

Learning manipulation with imitation To improve sample complexity, imitation learning from expert demonstrations is frequently used, whether for non-dexterous [44, 41, 43, 45] or dexterous [12, 37, 15, 53, 36] end effectors. Though advancing the state of the art in dexterous manipulation, the latter approaches rely on "person-centric" human demonstrations with motion capture gloves. Aside from gloves, demonstrations may be captured via teleoperation and video [13] or paired video and kinesthetic demos [43, 44]. In any case, expert demonstrations can be expensive, are specific to the end effector of the demonstration, and their trajectories need not generalize to novel objects.

In contrast, the proposed object-centric affordances sidestep these issues, at the cost of instead supervising the predictive image model. We use supervision from thermal image "hotspots" where people hold objects to use them [6], though other annotation modes are possible. ContactGrasp [7] leverages thermal image data to rank GraspIt [27] hand poses for a model-based optimization approach. In contrast, our approach 1) learns a closed-loop RL policy for grasping, and 2) incorporates a *predictive* image-based affordance model that allows generalization to *unseen* objects. Furthermore, once trained, our policy runs in real-time on new objects, whereas ContactGrasp takes about 4 hours to sample GraspIt poses for each unseen object.

Visual affordances A few methods infer visual affordances for grasping with simple grippers [38, 21, 22, 19] and explore non-robotics affordances [30, 9, 32, 11]. Whereas traditionally supervision comes from labeled image examples [30, 9] or a robot's grasp success/failure [38, 21, 22], recent work explores weaker modes of supervision from video [32, 11]. Visual models can help focus attention for a pick and place robot [51, 52]. All of the prior methods make use of simple grippers in an open-loop control setting [38, 21, 22, 51, 52]. To our knowledge, ours is the first work to demonstrate closed-loop RL policies learned with visual affordances.

3 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. 3). First, we train a network to predict affordance regions from static images (Sec. 3.1). Second, we train a dynamic grasping policy using the learned affordances (Sec. 3.2). All of our experiments are conducted on a simulated tabletop environment using a 30 DoF dexterous hand as the robotic manipulator (detailed below). We next detail each of these stages.

3.1 Affordance Anticipation From Images

We first design a perception model to infer object-centric grasp affordance regions from static images. As discussed above, an object-centric approach has the key advantage of providing human intelligence about how to grasp while forgoing demonstration trajectories. Furthermore, by predicting affordances from images, we open the door to generalizing to new objects the robot has not seen before.

Thermal image contact training data We train the affordance model with images with ground truth functional grasp regions obtained from ContactDB [6]. ContactDB contains 3D scans of 50 household objects along with real-world human contact maps captured using thermal cameras. Participants grasped each object using two different post-grasp functional intents—*use* and *hand-off*—and a thermal camera on a turntable recorded the multi-point "hotspots" where the object was touched. Whereas past vision models often take supervision from traditional image annotations drawn with a mouse by people imagining how they would touch an object [9, 11, 30, 33], ContactDB derives its "annotations" directly from real human interactions, an advantage for realism and ease. However, our model could be similarly trained with manual image annotations.

We consider contact maps corresponding to the *use* intent and exclude objects having bimanual grasps, which yields 16 total objects. Since each object has thermal maps captured from 50 different

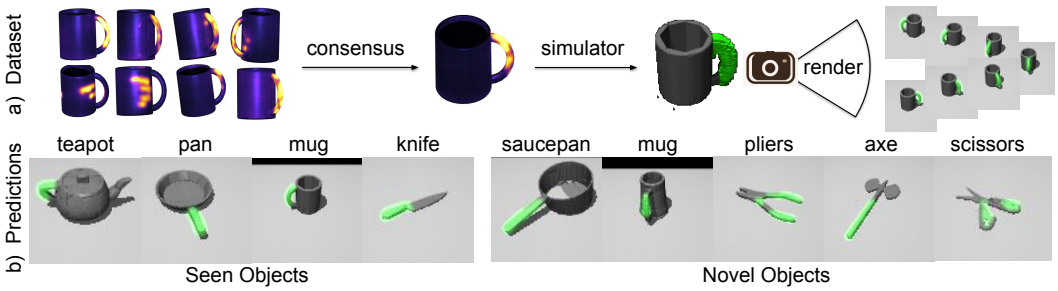

Figure 2: **Affordance anticipation.** a) Training images generated from 3D thermal maps from ContactDB. Green denotes label masks overlaid on images. b) Sample predictions for seen and novel objects from ContactDB and 3DNet, respectively. Our anticipation model predicts meaningful functional affordances for novel objects and viewpoints (e.g., graspable handles and rings).

participants, we use k-medoids clustering to obtain a representative thermal map for each object. Specifically, for a given object, we cluster the XYZ values of mesh points with a contact strength value above 0.5 (following [7]), then take the medoid of the largest cluster as our representative contact map for that object. We port the 3D models into the MuJoCo physics simulator [48] and render them on a tabletop to create an image training set.

For each object, we obtain a set of image-affordance pairs (x_i, y_i) by rendering the 3D object and the 3D contact map, respectively. See Fig. 2a. We rotate each object randomly within a 0-180°range of the camera viewing angle and augment the dataset with varying camera positions. Finally, we obtain a dataset of $\sim$15k training pairs, which we divide into an 80:10:10 train/val/test split.

Image affordance prediction model Let X represent the domain of object images, and let Y be the object-centric grasp affordances. Our goal is to learn a mapping $G : X \rightarrow Y$ that will infer the grasp affordance regions from an individual image. During training, we have labelled (image, mask) pairs $\{x_i, y_i\}_{i=1}^{N}$. We pose the affordance learning problem as a segmentation task to predict binary per-pixel labels, and approximate G with a convolutional neural network. We adapt the Feature Pyramid Network (FPN) [23] to perform semantic segmentation and use an ImageNet pretrained ResNet-50 [14] as the backbone. See Fig 3a, and Supp for details.

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. On the ContactDB test split, the segmentation accuracy averages 80.4% in IoU. Fig. 2b shows sample predictions for both ContactDB and 3DNet [50] (for which we do not have ground truth; see Sec. 4 for dataset info). Our affordance anticipation model is able to predict meaningful functional affordances for novel objects and viewpoints. For example, it faithfully infers graspable handles of saucepan, axe, and pliers despite not having encountered these categories in the training set.

3.2 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. 3b.

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 $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$, initial state distribution ρ_0, reward function $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$, horizon T, and discount factor $\gamma \in (0, 1]$. Hence, we are interested in maximizing the expected discounted reward $J(\pi) = \mathbb{E}_\pi[\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)]$ to determine the optimal stochastic policy $\pi : \mathcal{S} \rightarrow \mathbb{P}(\mathcal{A})$. We use an actor-critic model to estimate state values $V_\theta(s_t)$ and policy distribution $\pi_\theta(a_t|s_t)$ at each time step.

State space The work space of the robot consists of an object positioned on a table at a random orientation. The state space consists of the visuomotor inputs used to train the control policy:

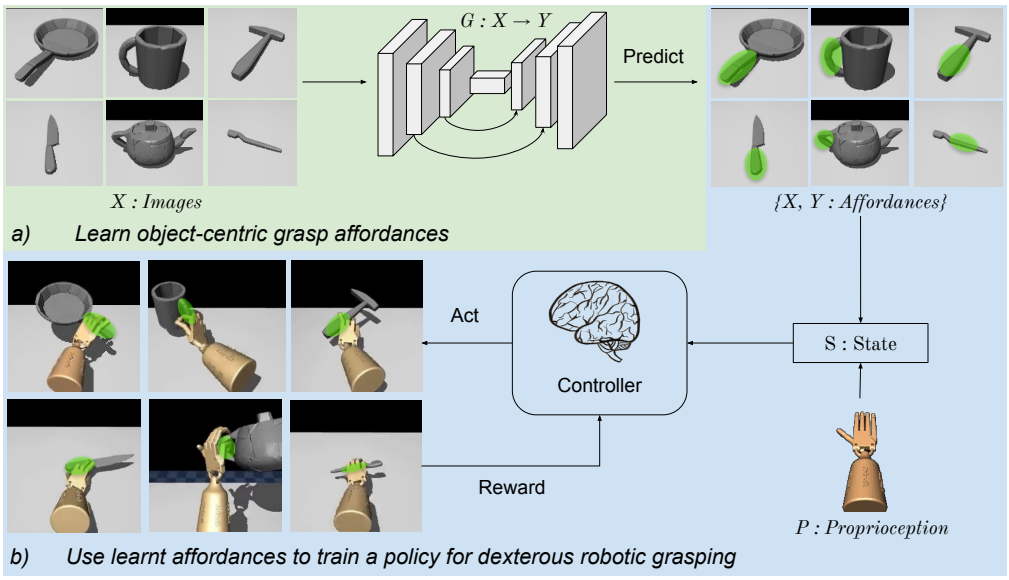

Figure 3: **Overview of our GRAFF model.** a) In Stage I, we train an affordance prediction model that predicts object-centric grasp affordances given an image. b) In Stage II, we train an RL policy that leverages these affordances along with other visuomotor sensory inputs (RGB-D image + hand joint variables) to learn a stable grasping policy.

$\mathcal{S} = \{X, Y, P, D\}$ (see Fig. 4). The visual input at time t consists of an RGB-D image $x_t \in X$ captured by an egocentric hand-mounted camera that translates with the hand but does not rotate. The affordance input $y_t \in Y$ is the binary affordance map inferred from the image, $y_t = G(x_t)$. The proprioception input $p_t \in P$ consists of the positions and velocities of each DoF in the hand actuator.

The distance input $d_t \in D$ is the distance between the agent's hand and the object affordance region. We compute it as the pairwise distance between M fixed points on the hand and N points sampled from the backprojected affordance map. We obtain the latter by backprojecting y_0 to 3D points in the camera coordinate system using the depth map at $t = 0$, then tracking those points throughout the rest of the episode. Hence we do not assume access to the full object state (we do not know the object mesh or mass), but we do assume perfect tracking of the affordance region that was automatically detected in the agent's first video frame. We leave it as future work to relax the tracking assumption, e.g., by strengthening the segmentation model in the presence of occlusions.

Action space We use a 30-DoF position-controlled anthropomorphic hand from the Adroit platform [20] 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. We combine two rewards to realize this. The agent gets a positive reward R_{succ} of +1 for each time step that the object is lifted off the table, and a negative reward R_{aff} on the hand-affordance contact distance to incentivize the agent to explore areas of the object that lie within the affordance region. R_{aff} is computed as the Chamfer distance between the M and N points described earlier. We also include an entropy maximization term, $R_{entropy}$, to encourage exploration of the action space [40]. Our total reward function is:

$$r = \alpha R_{succ} + \beta R_{aff} + \eta R_{entropy}. \tag{1}$$

The hand-affordance contact reward says the agent should try to position itself close to the affordance region. Since it is expressed in an object-centric manner, there are no constraints on how the hand is posed when it reaches those regions nor which finger lines up where. Conceptually, this can be seen as softer supervision than that employed in imitation learning for manipulation which requires kinesthetic teaching [43, 44] or tele-operation [13, 37, 25] to obtain expert trajectories.

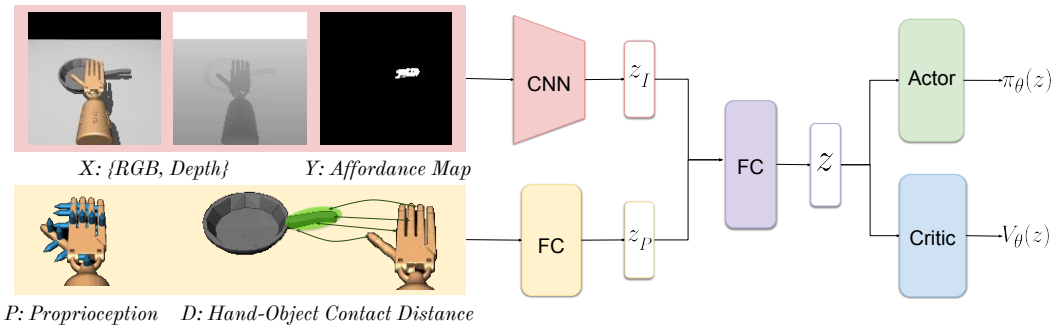

Figure 4: **Grasp policy learning architecture.** See text for details.

Implementation details We implement our approach with the architecture shown in Fig. 4. The affordance network is optimized using Dice loss for 20 epochs with a learning rate of $1e-4$ and minibatch size of 8. We preprocess the affordance map by computing its distance transform, which helps densify the affordance input. The CNN encoder consists of three 2D convolutional layers with filters of size [8,4,3], and a bottleneck layer of dimension 512, with ReLU activations between each layer. The proprioception and hand-object distance inputs are processed using a 2-layer fully-connected encoder of dimension [512,512]. For the hand-object contacts, we use $M = 10$ and $N = 20$ uniformly sampled points. The CNN and FC embeddings are concatenated and further processed (FCs) before predicting the action values. We optimize the network using the Adam optimizer [18] with a learning rate of $5e-5$. The full network is trained using PPO [40]. We train a single policy for all ContactDB objects for 60M agent steps with an episode length of 200 time steps. With each step being 2 ms long, this amounts to 30 hours of learning experience. The coefficients in the reward function (Eq. 1) are set as: $\alpha = 1, \beta = 1, \eta = 0.001$. We train for four random seed initializations. Complete network architecture details are in Supp.

4 Experiments

Datasets We validate our approach with two datasets: ContactDB [6] and 3DNet [50]. We train a single policy across all 16 objects from ContactDB with one-hand grasps. First we evaluate grasping on these 16 *seen* objects. Then, we test on 24 *novel* object meshes from 3DNet, a CAD model database with multiple meshes per category. We use the meshes from 9 categories that roughly align with the objects in ContactDB. We classify the unseen objects into *known* (for classes that are present in ContactDB, e.g., pan, hammer, mug) and *unknown* (for classes that are related to those in ContactDB but not present, e.g., axe, saucepan, wrench, pliers). Full details are in Supp.

Comparisons We first devise two pure RL baselines that lack the proposed affordances: (**1**) NO PRIOR: uses the lifting success and entropy rewards only. (**2**) COM: uses the center of mass as a prior, which may lead to stable grasps [39, 17], by penalizing the hand-CoM distance for R_{aff}. Both pure RL methods use our same architecture (Fig. 4), allowing apples-to-apples comparisons. (**3**) DAPG: We also compare to DAPG [37], a hybrid imitation+RL model that uses motion-glove demonstrations. We train DAPG on the 16 ContactDB objects using the authors' provided demonstrations for object relocation and adapting their code (details in Supp).[1] We stress that DAPG is a strongly-supervised approach with access to full motion trajectories of expert actions, whereas our approach uses inferred object-centric affordances to guide the policy. A practical advantage of our method is to replace heavy demonstrations (state-action pairs) with image-based affordances.

Metrics We use two metrics: (**1**) Grasp Success: For a given episode, a successful grasp has been executed if the object has been lifted off the table by the hand for at least the last 50 time steps (a quarter of the episode length) to allow time to reach the object and pick it up. (**2**) Grasp Stability: After an episode completes, we apply perturbing forces of 5 Newtons in six orthogonal directions to the object. If the object remains held, the grasp is deemed stable. We execute 100 episodes per object using the policies that attained the highest training reward, with the objects placed at different orientations ranging from [0,180°], and report success and stability rates over these 100 test episodes.

[1] We are not aware of any existing demonstration data for ContactDB objects. While DAPG's demonstrations use a ball, we found that with adequate fine-tuning to our data, it can perform well on additional objects.

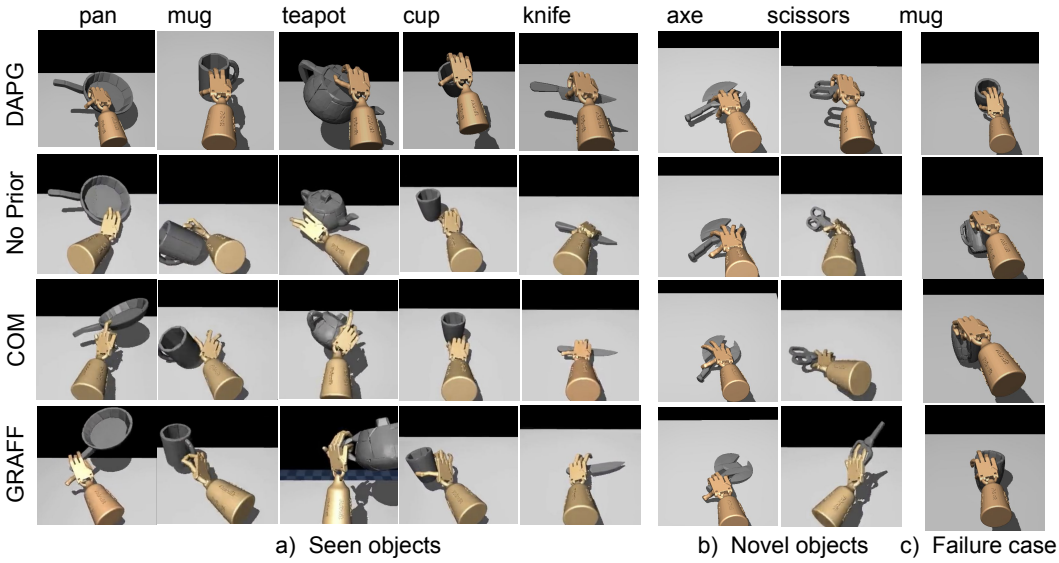

Figure 5: **Grasping performance.** Example frames from a) seen objects in ContactDB and b) novel objects in 3D-Net. Our affordance-based GRAFF is able to successfully grasp both seen and novel objects at their functional grasp locations, while the two pure-RL baselines either fail to learn successful grasps (mug, teapot, cup, axe) 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 image-based model, generalizes better to unseen ones. c) Failure cases can arise from difficult initial object orientations where posing the hand around handles is a challenge. See Supp video.

Grasping seen objects from ContactDB Table 1 shows the results on ContactDB. GRAFF (our model) outperforms both pure RL baselines consistently on both metrics. Fig. 5a shows qualitative examples; please see the Supp video for full episodes. GRAFF can successfully grasp the objects at the anticipated affordance regions (handle of pan, mug, teapot, knife, scissors), while the baselines fail to grasp objects with complex geometries (pan, mug, teapot). This shows the effectiveness of the affordance-guided policy in learning stable functional grasps. Fig. 5c depicts failed grasps.

Our method also fares well compared to the more intensely supervised imitation+RL method DAPG [37], outperforming it on average for the success rate. This is a very encouraging result: not only does our method outperform its RL counterparts, it is also competitive with a method that leverages expert trajectories. DAPG has more difficulty when the objects deviate from that of the original demonstrations (ball), reinforcing our advantage of an image-based prior that can adapt its guidance to new object instances. Compared to DAPG, our method more effectively executes functional grasps on unseen objects thanks to its image-based model (axe, scissors).

Robustness to physical properties of the objects To evaluate robustness to changes in object properties, we apply the our policy to a range of object masses and scales not encountered during training. Fig. 6 shows 3D plots. Here, $m_0 = 1kg$ and $s_0 = 1$ are the mass and scale values used during training. GRAFF remains fairly robust across large variations, which we attribute to GRAFF's preference for stable human-preferred regions.

Grasping unseen objects from 3DNet Next we push the robustness challenge further by requiring the agent to generalize its grasp behavior to object instances it has not encountered before (the 24 3DNet [50] objects). We first render the objects and predict grasp affordances (cf. Fig. 2b). We then apply the policy trained on ContactDB to execute grasps. Table 1 (bottom) shows the results. 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 sample grasps. GRAFF successfully executes grasps at the anticipated affordance regions (e.g., handle of axe and finger rings on scissors), whereas the baselines may grasp the scissors at its blades or fail to lift the axe.

Training time Fig. 7 shows the grasp success rate versus training time. Not only does our model learn more successful policies, it also has a sharper learning curve. While the baselines reach a maximum success rate of 30% in 1800 training iterations (30 hours of robot experience), our method

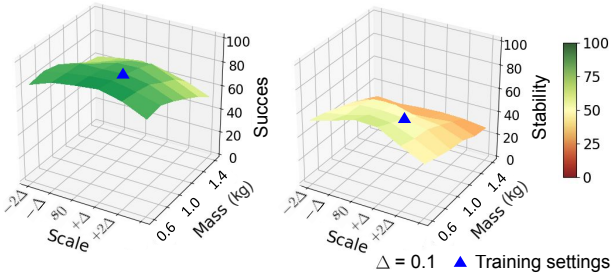

Figure 6: **Robustness to changes in physical properties**.

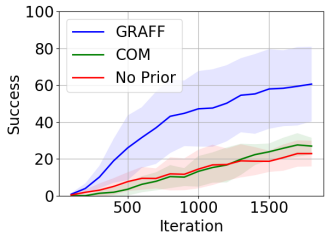

Figure 7: **Training curves.**

Table 1: **Grasp success and stability (%) on both datasets**. Our GRAFF model leverages visual object-centric affordances to outperform the other methods and generalize best to the unseen 3DNet objects (which have no thermal training images). Only DAPG uses human demonstration trajectories. * Note: These high success rates are caused by the hand piercing the object to lift it up (due to loopholes in the simulator physics), and are not actually successful grasps.

Dataset	Category	Grasp Success				Grasp Stability			
		Pure RL			Imit+RL	Pure RL			Imit+RL
		No Prior	CoM	GRAFF	DAPG [37]	No Prior	CoM	GRAFF	DAPG [37]
ContactDB	mug	21	10	**56**	31	15	1	**38**	22
	scissors	38	65	**97**	94	4	58	**94**	89
	teapot	1	**85***	57	67	0	74	31	**79**
	flashlight	12	25	**98**	57	17	12	**89**	57
	cell phone	9	0	54	**68**	3	0	38	**74**
	hammer	1	14	**88**	58	2	3	**44**	41
	knife	11	8	**66**	49	7	3	42	**48**
	apple	0	0	**100**	**100**	0	0	89	**100**
	light bulb	23	46	74	**94**	0	29	45	**77**
	mouse	44	72	74	**77**	8	49	**60**	55
	pan	16	29	75	**95***	3	27	**50**	89
	cup	0	0	**100**	21	0	0	**61**	16
	stapler	51	**71**	18	59	3	10	**42**	33
	door knob	1	0	79	**100**	0	0	49	**100**
	toothbrush	7	30	**84**	51	1	11	**71**	39
	toothpaste	52	45	**84**	72	2	19	61	**71**
	mean	17	27	**78**	69	4	18	56	**62**
3DNet	known	14	25	**72**	57	6	15	**53**	48
	unknown	10	18	**63**	49	4	12	**49**	42
	mean	12	22	**68**	53	5	14	**51**	45

reaches the same success rate in only 600 iterations (10 hours) — a 3× speedup. Recall that 30 hours is a one-time cost: we train a single policy for all ContactDB objects, and simply execute that trained policy when encountering an unseen object. Thus our affordance prior meaningfully improves sample efficiency for dexterous grasping, while convincingly outperforming the other pure RL methods.

Conclusion Our approach learns dexterous grasping with object-centric visual affordances. Breaking away from the norm of expert demonstrations, our GRAFF approach 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. While there is much more work to do in this direction, we see the results as encouraging evidence for manipulation agents learning faster with more distant human supervision.

In future work, we are interested in expanding the visual affordance model and investigating manipulations beyond grasping (e.g., open, sweep). Translating the policies to real-world robots is also important future work, and we are encouraged by recent successes with dexterous robots [3, 47].

Broader Impact

Our work addresses dexterous robot grasping. Robots capable of fine-grained manipulation have valuable future applications in service robotics (e.g., a home health or eldercare robot assistant) and manufacturing. Furthermore, learning methods that require minimal human intervention for training such manipulation robots are appealing for real-world applications where the robot must generalize to objects and tasks it has not seen before, without expensive re-training. As robots become increasingly autonomous, there could be potential negative consequences in terms of replacing human workers at certain tasks. On the other hand, the advances would also open up other jobs for technological development. We discuss the broader scientific impact in the introduction of the paper.

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] Heni Ben Amor, Oliver Kroemer, Ulrich Hillenbrand, Gerhard Neumann, and Jan Peters. Generalization of human grasping for multi-fingered robot hands. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012.

[3] 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.

[4] Yunfei Bai and C Karen Liu. Dexterous manipulation using both palm and fingers. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2014.

[5] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *IEEE International Conference on Robotics and Automation*, 2000.

[6] 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.

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

[8] M. Ciocarlie, C. Goldfeder, and P. Allen. Dimensionality reduction for hand-independent dexterous robotic grasping. In *IROS*, 2007.

[9] Thanh-Toan Do, Anh Nguyen, Darwin G Caldwell Ian Reid, and Nikos G Tsagarakis. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2017.

[10] Mehmet Dogar and Siddhartha Srinivasa. Push-grasping with dexterous hands: Mechanics and a method. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2010.

[11] Kuan Fang, Te-Lin Wu, Daniel Yang, Silvio Savarese, and Joseph J Lim. Demo2vec: Reasoning object affordances from online videos. In *CVPR*, 2018.

[12] 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.

[13] 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.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[15] 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.

[16] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *Conference on Robot Learning (CoRL)*, 2018.

[17] 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.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[20] 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.

[21] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[22] 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.

[23] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[24] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *Robotics: Science and Systems (RSS)*, 2017.

[25] 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.

[26] Hamza Merzić, Miroslav Bogdanović, Daniel Kappler, Ludovic Righetti, and Jeannette Bohg. Leveraging contact forces for learning to grasp. In *International Conference on Robotics and Automation (ICRA)*, 2019.

[27] A. T. Miller and P. K. Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 2004.

[28] 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.

[29] Adithyavairavan Murali, Lerrel Pinto, Dhiraj Gandhi, and Abhinav Gupta. Cassl: Curriculum accelerated self-supervised learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6453–6460. IEEE, 2018.

[30] Austin Myers, Ching L Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381. IEEE, 2015.

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

[32] Tushar Nagarajan, Christoph Feichtenhofer, and Kristen Grauman. Grounded human-object interaction hotspots from video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8688–8697, 2019.

[33] Anh Nguyen, Dimitrios Kanoulas, Darwin G Caldwell, and Nikos G Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. In *IROS*, 2017.

[34] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *2016 IEEE international conference on robotics and automation (ICRA)*, pages 3406–3413. IEEE, 2016.

[35] Deirdre Quillen, Eric Jang, Ofir Nachum, Chelsea Finn, Julian Ibarz, and Sergey Levine. Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6284–6291. IEEE, 2018.

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

[37] 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.

[38] Joseph Redmon and Anelia Angelova. Real-time grasp detection using convolutional neural networks. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1316–1322. IEEE, 2015.

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

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

[41] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine. Time-contrastive networks: self-supervised learning from video. In *ICRA*, 2018.

[42] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020.

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

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

[45] 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.

[46] 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.

[47] Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*, 2017.

[48] 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.

[49] 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.

[50] 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.

[51] Lin Yen-Chen, Andy Zeng, Shuran Song, Phillip Isola, and Tsung-Yi Lin. Learning to see before learning to act: Visual pre-training for manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[52] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois R Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, et al. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1–8. IEEE, 2018.

[53] 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.