Dexterous Robotic Grasping with Object-Centric Visual Affordances: Supplementary Material

Priyanka Mandikal
UT Austin
mandikal@cs.utexas.edu

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

Please see http://vision.cs.utexas.edu/projects/graff-dexterous-affordance-grasp/ for videos including example episodes.

Contents

A Network Architecture and Training

In this section, we provide additional implementation details for training the affordance anticipation model and dexterous grasping policy.

A.1 Affordance Anticipation

For the affordance anticipation model (Sec. 3.1, main paper), we train a convolutional neural network to perform binary per-pixel segmentation of grasp affordances. We adapt the Feature Pyramid Network (FPN) [8] to perform semantic segmentation and use an ImageNet pretrained ResNet-50 [4] as the backbone. We use sigmoid as our non-linearity function. The network is trained using Dice loss, and optimized using the Adam optimizer [6] with a learning rate of $1e^-4$. We train the entire network for 20 epochs with a batch size of 8. It takes around an hour for the training to complete on a Tesla K40m GPU. The data generation process along with sample predictions are shown in Fig. A.

A.2 Grasp Policy Learning

In addition to the details provided in the main paper (Sec. 3.2: Implementation Details), here we elaborate on the network architecture and training procedure. Fig. B illustrates the policy learning pipeline. The CNN branch that processes the visual inputs (RGB, Depth, Affordance Map) consists of

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

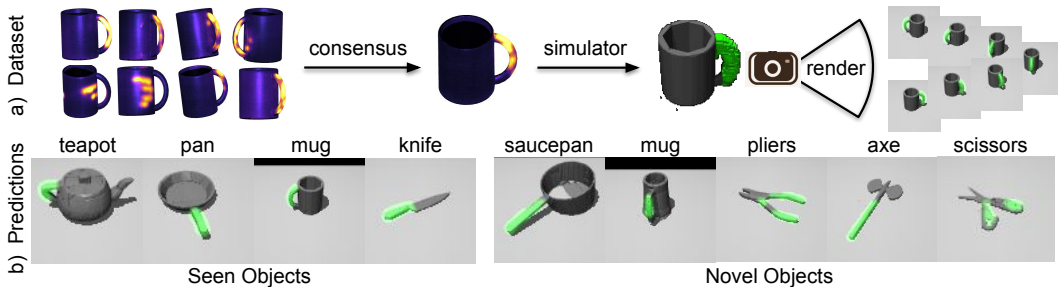

Figure A: **Affordance anticipation.** a) Training images generated from 3D thermal maps from ContactDB. Green denotes label masks overlaid on images. b) Sample predictions for seen and novel objects from ContactDB and 3DNet, respectively. Our anticipation model predicts meaningful functional affordances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

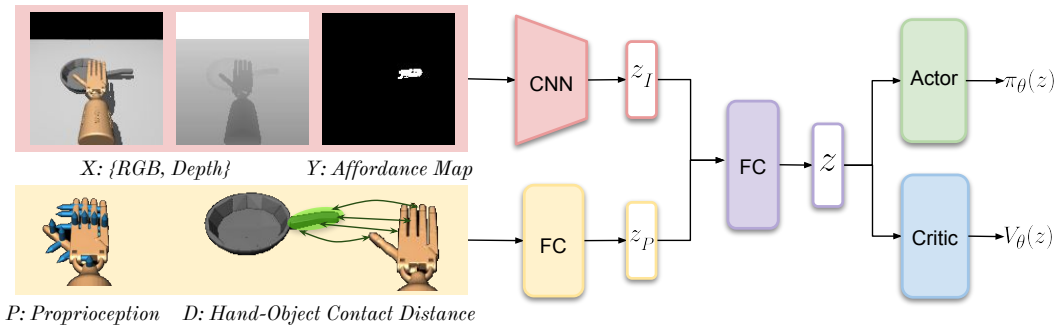

Figure B: **Grasp policy learning architecture.** See text for details.

three 2D convolutional layers with [32, 32, 64] filters of size [8, 4, 3] respectively. It culminates in a fully-connected layer to produce the image latent embedding z_I of 512 dimension. The proprioception and hand-object contact distances are processed by two fully-connected layers of size [512, 512] to produce the proprioception and distance embedding z_P. The embeddings z_I and z_P are concatenated and further processed by two fully-connected layers of size [1024, 512] to produce the final embedding z. Embedding z is then processed by the actor and critic networks (which consist of two FC layers of dimension [512, 512]) to predict action probabilities and state values, respectively. The convolutional layers use the ReLU non-linearity, and the fully-connected layers use the tanh non-linearity. It takes 30 hours to train the policy on a Tesla V100 GPU.

B Additional Results

B.1 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. C 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.

B.2 Examples for Affordance Anticipation

Fig. D shows affordance predictions for the novel objects from 3DNet [12]. We observe that our anticipation model predicts meaningful functional affordances (e.g., graspable handles and rings) for novel objects with variations in viewpoints.

B.3 Noisy sensing and actuation

When deploying trained policies on real robots, we might encounter a number of non-ideal circumstances owing to faulty sensor readings or imperfect robot executions. To better model such realistic settings, we incorporate noise into our training and testing regimes in simulation [1, 14].

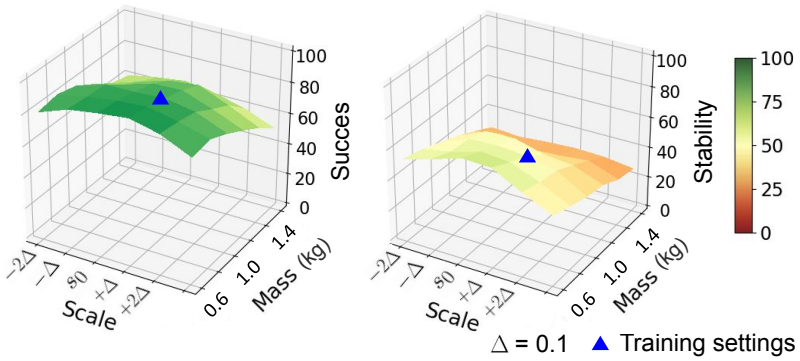

Figure C: Robustness to changes in physical properties. GRAFF shows good generalization across a range of mass and size variations of the objects.

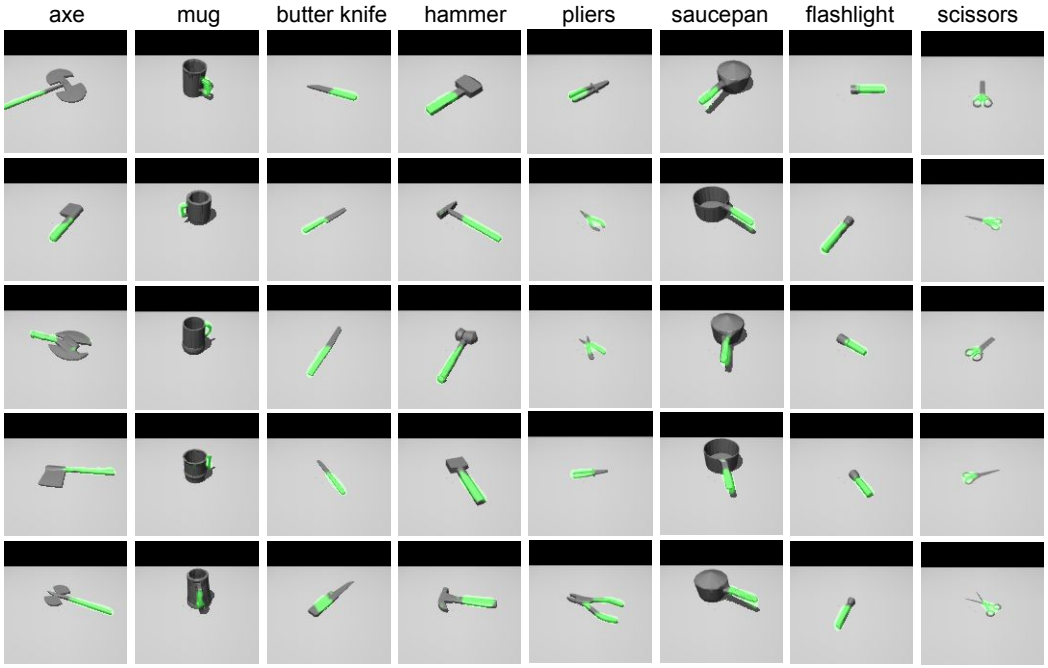

Figure D: **Affordance anticipation results.** Additional prediction results for novel objects from 3DNet. Our anticipation model predicts meaningful functional affordances (e.g., graspable handles and rings) for novel objects with variations in viewpoint and scale.

Following [14], we apply additive Gaussian noise on the proprioceptive sensor readings (robot joint angles and angular velocities), object tracking points, and robot actuation. We also apply pixel perturbations in the range [-5,5] and clip all pixel values between [0,255] (see Fig.3, main paper). We train all versions of all methods under these noisy conditions. We additionally test our method with a tracking failure model that freezes the track for 20 frames at random intervals and find that mean success rate is still at a reasonably high rate of 54%. These empirical results indicate that GRAFF can remain fairly robust to more challenging tracking, sensing, and actuation failures that real robotic systems may encounter.

While our lab does not have access to a real dexterous robotic hand, we believe that (like in [5, 7, 9, 10, 13]) the high quality physics-based simulator together with these noise models offers a meaningful study. We are also encouraged by recent successes translating policies trained only in simulation to real-world dexterous robots [2, 11].

C Modified DAPG Environment

Below, we discuss our setup for adapting DAPG [10] (described in Sec. 4: Comparisons, main paper) to train a grasping policy for ContactDB [3] objects. Recall that DAPG [10] is a hybrid imitation+RL model that uses motion-glove demonstrations for *object relocation*. In order to closely replicate our grasping setup, we modify the DAPG environment accordingly and train a grasping policy on the 16 ContactDB objects. The original DAPG environment consists of lifting up a ball placed on a table to relocate it to a specified target location. The location of the ball and target are varied every episode.

Our modified environment uses ContactDB objects in place of the ball. The settings are the same as the ones used to train the other methods i.e. we keep the object position fixed every episode (no object translation) and rotate the object in [0,180°]. We also keep the target fixed right above the object, i.e. it remains unchanged every episode. These modifications turn the task into a pure grasping task without having to move the object around to relocate it. We use the author's provided code for training the DAPG algorithm. We collect object-specific mocap demonstrations in VR and train separate policies for grasping each ContactDB object (25 demos per object). The policy is first initialized by running behavioral cloning for 5 epochs, followed by an RL loop that runs for 500 iterations, with 200 trajectories drawn per iteration. We train four models initialized with different seeds and report mean and std dev of the metrics across all four seeds. For 3DNet evaluation, we use the trained policy of the ContactDB object that is closest in shape to the one in 3DNet.

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] 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.

[5] 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.

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

[7] T. Li, K. Srinivasan, M. Q. H. Meng, W. Yuan, and J. Bohg. Learning hierarchical control for robust in-hand manipulation. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] Yuechuan Xue and Yan-Bin Jia. Gripping a kitchen knife from the cutting board. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS*, 2020.

[14] Yuke Zhu, Ziyu Wang, Josh Merel, Andrei Rusu, Tom Erez, Serkan Cabi, Saran Tunyasuvunakool, János Kramár, Raia Hadsell, Nando de Freitas, et al. Reinforcement and imitation learning for diverse visuomotor skills. *arXiv preprint arXiv:1802.09564*, 2018.