Learning to predict grasping interaction with geometry-aware 3D representations

Xinchen Yan*, Mohi Khansari†, Yunfei Bai†, Jasmine Hsu‡,
Arkanath Pathak‡, Abhinav Gupta‡, James Davidson‡ and Honglak Lee‡
*University of Michigan, Ann Arbor, MI, xcyan@umich.edu
†X Inc., Mountain View, CA
‡Google Brain, Mountain View, CA

I. Abstract

Learning object interaction is an essential problem in artificial intelligence that involves perception, motion planning and control. In this paper we present our results on the problem of grasp prediction from a single-view RGBD as well as the camera view matrix. We show that learning geometry is at the heart of this type of interaction and propose a geometry-aware grasping procedure with which first we predict a 3D volumetric representation of an object from an image, and then use this together with the image and a grasp pose proposal to predict the grasp success/failure (see Figure 1). We compare our results with a vanilla approach where outcome is a high-order mapping from image and action [3, 4, 2, 1] (see Figure 2a and b).

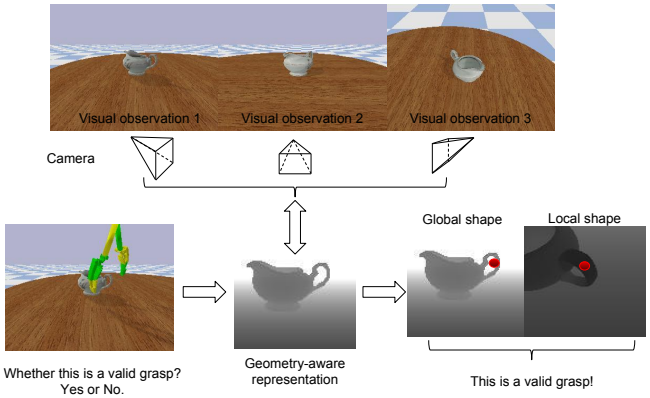

Fig. 1. Learning grasping interactions from demonstrations.

II. Approach

A. Proposed Architecture

Compared to existing deep learning frameworks for grasping [3, 2], we propose a two-stage procedure for learning grasping interaction from demonstrations: (1) the agent learns to understand object geometry from 2D visual input, and (2) the agent learns to predict grasping interaction from demonstrations (see Figure 2c). More specifically, we design an encoder-decoder deep neural network for learning such procedure. Our geometry-aware encoder-decoder network has two components: shape prediction network and grasping outcome prediction network. The shape prediction network has an image encoder, a 3D shape decoder, and a learning-free visual projection layer. The image encoder transforms the 2D visual data into the high-level geometry representation. The shape decoder network takes in the geometry representation and outputs the 3D volume of object. To enable supervision with 2D visual data only, we propose a learning-free visual projection layer similar to [5]. The grasping outcome prediction network has a state encoder and an outcome predictor. The state encoder network transforms the current visual state (e.g., object and gripper) to high-level state representation. The outcome predictor network takes in action, state, and geometry representation and outputs the outcome (e.g., success or failure). The shape and outcome prediction networks are bridged by the global and local shape representation.

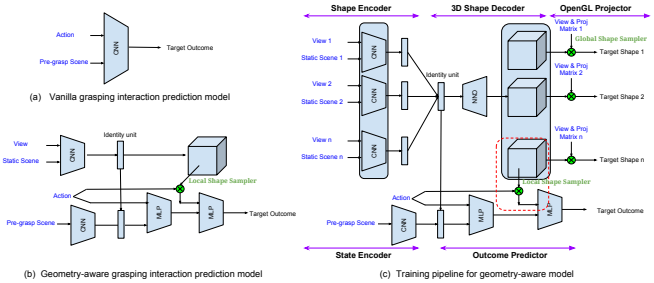

Fig. 2. Illustration of deep geometry-aware grasping network.

B. Dataset

We collected a database consists of 101 everyday objects on seven categories of objects (see Figure 3). In total more than 150K grasps were collected in Virtual Reality (VR) from both human and augmented perturbed demonstrations. For each object, we collect 10-20 successful grasping attempts with a 1-DoF virtual gripper from five right-handed users. For each attempt, we log the pre-grasp status which includes the location, orientation of object and gripper as well as the grasping outcome (e.g., success or failure). Additionally, we augment the data by perturbing the gripper location and orientation around every each human demonstrated grasps. We use Bullet (https://github.com//bulletphysics) to evaluate these augmented grasps in simulator and label them as success/failure based on their outcome.

* Xinchen Yan was an intern at Google Brain during this work.

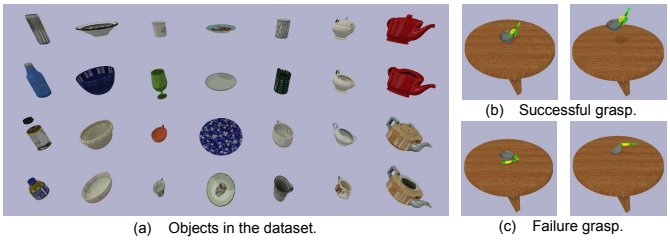

Fig. 3. VR-Grasping-101 dataset.

Method / Category	bottle	bowl	cup	plate	mug	sugarbowl	teapot	all
baseline (15)	72.81	73.36	73.26	66.92	72.23	70.45	66.13	71.42
geo-aware (15)	78.83	79.32	77.60	68.88	78.25	76.09	73.69	76.55
baseline (45)	71.02	74.16	73.50	63.31	74.23	72.70	64.19	71.32
geo-aware (45)	78.77	80.63	78.06	70.13	79.29	77.52	72.88	77.25

TABLE I
OUTCOME PREDICTION ACCURACY FROM SEEN ELEVATION ANGLES.

Method / Category	bottle	bowl	cup	plate	mug	sugarbowl	teapot	all
baseline (30)	71.15	72.98	71.65	61.90	71.01	70.06	61.88	69.50
geo-aware (30)	79.17	77.71	77.23	67.00	75.95	75.06	70.66	75.27
baseline (60)	68.45	73.05	72.50	61.27	74.40	71.30	63.25	70.18
geo-aware (60)	77.40	78.52	76.24	68.13	79.39	76.15	70.34	75.76

TABLE II
OUTCOME PREDICTION ACCURACY FROM NOVEL ELEVATION ANGLES.

C. Training

Learning robust 3D shape representation from single-view 2D sensory input is essentially a challenging task in computer vision due to the shape ambiguity. To reduce sub ambiguity for shape prediction, we assume multiple observations of the scene are available during model training. From the interaction perspective, multi-view observations provide additional input to the system which can be useful as well. Given a series of n observations $\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_n$ of the scene, the 3D reconstruction can be formulated as $f^V : \{\mathcal{I}_i\}_{i=1}^n \rightarrow \mathbf{V}$, where $\mathbf{V}$ corresponds to the volumetric representation of the object. Similarly, the projective operation from i-th viewpoint is $f^D : \mathbf{V} \times \mathbf{P}_i \rightarrow \mathcal{D}_i$, where $\mathcal{D}_i$ and $\mathbf{P}_i$ are the depth and camera transformation matrix from corresponding viewpoint, respectively. The overall loss function is given in Eq. 1, where $\lambda_{\mathcal{D}}$ and $\lambda_{\mathcal{M}}$ are the coefficients of the depth and mask prediction, respectively.

$$\mathcal{L}_\theta^{joint} = \lambda_{\mathcal{D}} \sum_{i=1}^{n} \mathcal{L}_\theta^{depth}(\hat{\mathcal{D}}_i, \mathcal{D}_i) + \lambda_{\mathcal{M}} \sum_{i=1}^{n} \mathcal{L}_\theta^{mask}(\hat{\mathcal{M}}_i, \mathcal{M}_i) \tag{1}$$

Inspired by previous work [4, 2, 1], where outcome is a high-order mapping from observation and action, a straightforward approach is to fit a functional mapping $f^l_{vanilla} : \mathcal{I} \times \mathbf{a} \rightarrow l$ (see Figure 2a). Building upon the vanilla prediction model, we propose a novel geometry-aware prediction model where the agent learns to predict the grasping interaction by taking the geometry-aware repsentation as additional input. Finally, given current obseravtion $\mathcal{I}$, proposed action $\mathbf{a}$, and learned 3D shape representation $\mathbf{V}$, we fit a functional mapping $f^l_{geometry-aware} : \mathcal{I} \times \mathbf{a} \times \mathbf{V} \rightarrow l$, where l is the binary label of whether it is a valid grasp.

We pre-trained the shape prediction model (shape encoder and shape decoder) using ADAM optimizer with learning rate $1e^{-5}$ for 400K iterations with a mini-batch of size 4. In each batch, we sample 4 random viewpoints as our multi-view training. We observed this setting led to a more stable shape prediction performance compared to single-view training. In addition, we used $\mathcal{L}_1$ loss for foreground depth prediction and $\mathcal{L}_2$ loss for silhouette prediction with coefficients $\lambda_{\mathcal{D}} = 0.5$ and $\lambda_{\mathcal{M}} = 10.0$. In the next step, we fine-tuned the state encoder and outcome predictor using ADAM optimizer with learning rate $3e^{-6}$ for 200K iterations with a mini-batch of size 4. We used cross-entropy as our objective function since the grasping prediction is formulated as a binary classification task.

III. RESULTS

A. Visualization: 3D shape prediction

We evaluate the performance of our network by running inference using the shape encoder and decoder network. In our evalutions, we used single-view RGBD as well as camera view matrix as input. As shown in Figure 4, our model demonstrates reasonable generalization ability when applied the same model to novel objects that do not exist in the training set.

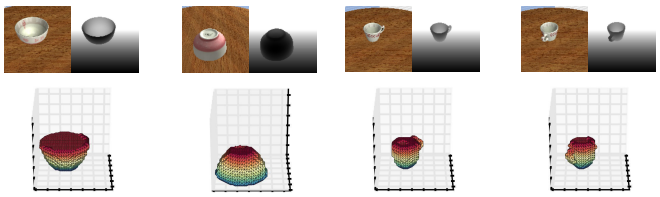

Fig. 4. Visualization: 3D shape prediction from single-view RGBD on testing (unseen) objects.

We adopt the classification accuracy as the evaluation metric and run dense evaluations for both models on the novel objects (from testing set). For each demonstration, we simulate 100 grasps (50% of them are success grasps) and run inference using both our geometry-aware model and baseline model. To investigate the model robustness to viewpoint change, we repeat the evaluations using 4 elevation angles (e.g, 15, 30, 45, and 60 degrees). We summarize the results in Table I and Table II. Overall, the geometry-aware model performs consistently better than vanilla model in outcome classification. Interestingly, we found that "teapot" and "plate" categories are comparatively more challenging than the other categories. This can be understood since "teapot" has irregular components (e.g., tip and handle) while "plate" has a fairly flat shape.

REFERENCES

[1] Alexey Dosovitskiy and Vladlen Koltun. Learning to act by predicting the future. *arXiv preprint arXiv:1611.01779*, 2016.

[2] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Advances in Neural Information Processing Systems*, pages 64–72, 2016.

[3] 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. *arXiv preprint arXiv:1703.09312*, 2017.

[4] Junhyuk Oh, Xiaoxiao Guo, Honglak Lee, Richard L Lewis, and Satinder Singh. Action-conditional video prediction using deep networks in atari games. In *Advances in Neural Information Processing Systems*, pages 2863–2871, 2015.

[5] Xinchen Yan, Jimei Yang, Ersin Yumer, Yijie Guo, and Honglak Lee. Perspective transformer nets: Learning single-view 3d object reconstruction without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 1696–1704, 2016.