Where2Act: From Pixels to Actions for Articulated 3D Objects

Kaichun Mo[1*], Leonidas Guibas[1], Mustafa Mukadam[2], Abhinav Gupta[2], Shubham Tulsiani[2]

[1] Stanford University [2] Facebook AI Research (*work mostly done while interning at FAIR)

2021 ICCV OCTOBER 11-17 VIRTUAL

Abstract

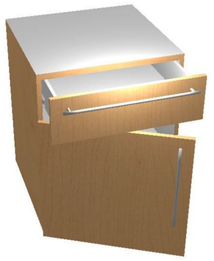
One of the fundamental goals of visual perception is to allow agents to meaningfully interact with their environment. In this paper, we take a step towards that long-term goal -- we extract highly localized actionable information related to elementary actions such as pushing or pulling for articulated objects with movable parts. For example, given a drawer, our network predicts that applying a pulling force on the handle opens the drawer. We propose, discuss, and evaluate novel network architectures that given image and depth data, predict the set of actions possible at each pixel, and the regions over articulated parts that are likely to move under the force. We propose a learning-from-interaction framework with an online data sampling strategy that allows us to train the network in simulation (SAPIEN) and generalizes across categories.

Problem Formulation

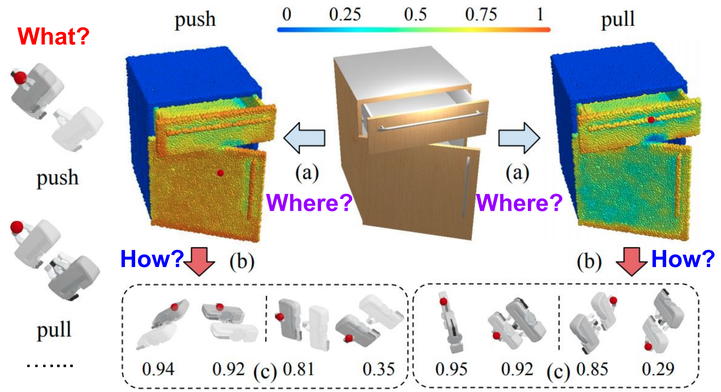
- What actions one can take given this cabinet with articulated parts?

Pull the drawer handle outward

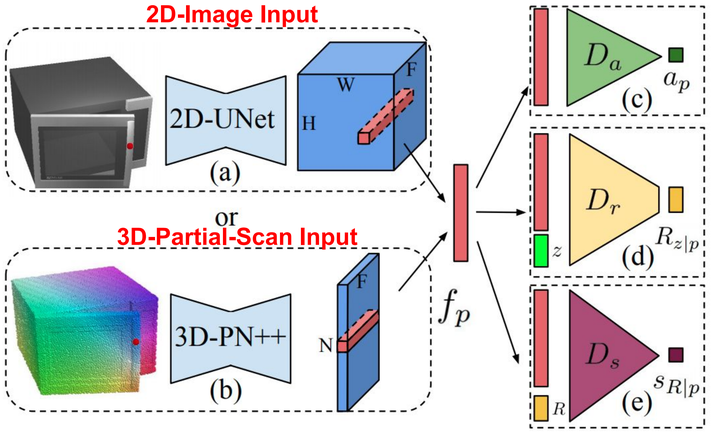
What? Where? How?



- We propose a novel framework to learn Where2Act and How2Act.



Network Architecture



Actionability score: Is this action-type good for this point?

Action Proposals: What actions can be taken?

Action Score: Is this specific action likely to succeed?

Training and Losses

- Train *action scoring module* with binary cross-entropy loss until convergence first;

$$\mathscr{L}_s = -\frac{1}{B}\sum_i r_i \log\left(D_s(f_{p_i|S_i}, R_i)\right) + (1-r_i)\log\left(1-D_s(f_{p_i|S_i}, R_i)\right). \quad (4)$$

Directly supervised by Interaction Data

- Train *action proposal module* with min-of-N loss then;

$$\mathscr{L}_r = \frac{1}{B}\sum_i \min_{j=1,\cdots,100} dist\left(\left(D_r\left(f_{p_i|S_i}; z_j\right)\right), R_i\right), \quad (5)$$

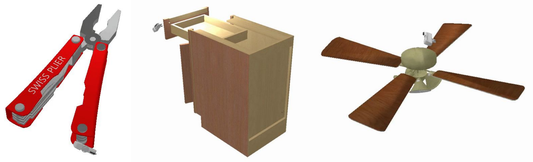
Predictions should cover sampled GTs

- Train *actionability scoring module* with *L2*-regression loss finally.

$$\hat{a}_{p_i|S_i} = \frac{1}{100}\sum_{j=1,\cdots,100} D_s\left(f_{p_i|S_i}, D_r\left(f_{p_i|S_i}, z_j\right)\right); \quad \mathscr{L}_a = \frac{1}{B}\sum_i \left(D_a(f_{p_i|S_i}) - \hat{a}_{p_i|S_i}\right)^2. \quad (6)$$

Inefficient to compute the exact score *(poor-man solution: expected success rate executing a random proposal by Dr)*

- Offline Randomly Simulated Interactions



Low succuss-rate for pulling (1%) → Overfitting to the Few Positive Data

- Online Positive Data Sampling

Pushing	Pulling
- Offline: 11%	- Offline: 1%
- Online: 29%	- Online: 8%

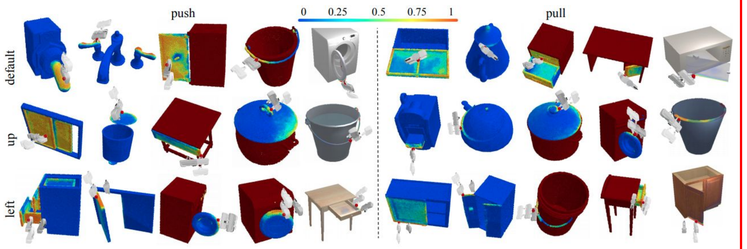
Sample over the Predicted Successful Regions During the Training Process

Results and Visualization

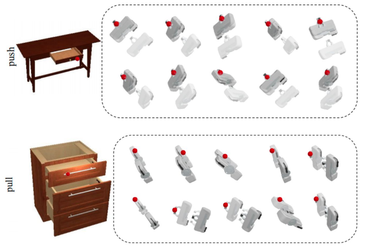
- Quantitative Evaluations and Comparisons to Baselines

		F-score (%)	Sample-Succ (%)			F-score (%)	Sample-Succ (%)
pushing	B-Random	12.02 / 7.40	6.80 / 3.79	pulling	B-Random	2.26 / 3.19	1.07 / 1.55
	B-Normal	31.94 / 17.39	21.72 / 11.57		B-Normal	6.20 / 8.02	3.79 / 4.18
	B-PCPNet	32.01 / 18.21	18.04 / 9.15		B-PCPNet	7.19 / 8.57	4.15 / 3.71
	2D-ours	34.21 / 22.68	21.36 / 10.58		2D-ours	7.04 / 8.98	4.07 / 4.70
	3D-ours	**43.76 / 26.61**	**28.54 / 14.74**		3D-ours	**10.95 / 12.88**	**7.51 / 7.85**
pushing-up	B-Random	4.92 / 3.31	2.70 / 1.62	pulling-up	B-Random	5.01 / 4.13	2.22 / 2.41
	B-Normal	13.37 / 7.56	8.93 / 4.81		B-Normal	13.64 / 9.40	8.67 / 6.08
	B-PCPNet	15.08 / 7.50	8.09 / 4.86		B-PCPNet	14.73 / 10.98	8.37 / 6.19
	2D-ours	15.35 / 8.69	8.70 / 5.76		2D-ours	15.74 / 12.88	9.71 / 7.10
	3D-ours	**21.64 / 11.20**	**12.06 / 6.56**		3D-ours	**22.24 / 16.28**	**13.53 / 9.28**
pushing-left	B-Random	6.18 / 4.05	3.08 / 2.26	pulling-left	B-Random	5.83 / 4.16	3.06 / 2.31
	B-Normal	18.52 / 10.72	11.59 / 5.72		B-Normal	17.52 / 10.51	11.14 / 5.82
	B-PCPNet	18.66 / 10.81	9.69 / 4.43		B-PCPNet	18.89 / 11.00	9.12 / 5.19
	2D-ours	18.93 / 12.04	11.68 / 7.22		2D-ours	16.20 / 10.16	10.15 / 6.05
	3D-ours	**26.04 / 16.06**	**15.95 / 9.31**		3D-ours	**25.22 / 14.49**	**14.25 / 7.10**

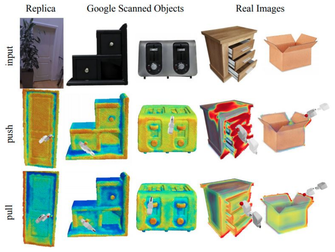
- Qualitative Results for *action scoring module*



- Qualitative Results for *action proposal module*



- Generalization to Real-world Data



Acknowledgements

- This work was supported primarily by Facebook during Kaichun's internship, while also by NSF grant IIS-1763268, a Vannevar Bush faculty fellowship, and an Amazon AWS ML award.
- We thank Yuzhe Qin and Fanbo Xiang for providing helps on setting up the SAPIEN environment.