Sim2real Image Translation Enables Viewpoint-Robust Policies from Fixed-Camera Datasets

Jeremiah Coholich [1] Justin Wit[1] Zsolt Kira[1] *

Abstract

Vision-based policies for robot manipulation have achieved significant recent success, but are still brittle to distribution shifts such as camera viewpoint variations. One reason is that robot demonstration data used to train such policies often lacks appropriate variation in camera viewpoints. Simulation offers a way to collect robot demonstrations at scale with comprehensive coverage of different viewpoints, but presents a visual sim2real challenge. To bridge this gap, we propose an unpaired image translation method with a novel segmentation-conditioned InfoNCE loss, a highly-regularized discriminator design, and a modified PatchNCE loss. We find that these elements are crucial for maintaining viewpoint consistency during translation. For image translator training, we use only real-world robot play data from a single fixed camera but show that our method can generate diverse unseen viewpoints. We observe up to a 46% absolute improvement in manipulation success rates under viewpoint shift when we augment real data with our sim2real translated data.

1. Introduction

In order for robot imitation learning policies to achieve generalization, they must be trained on a comprehensive robot demonstration dataset covering all variations the robot may encounter. Practically however, it is onerous to collect tabletop manipulation datasets that have a range of camera viewpoints since they often employ fixed, third-person cameras [4, 5, 10]. Indeed, we have observed empirically that when robot polices are trained on fixed-camera datasets, changes in the camera viewpoint during deployment causes a drastic decrease in performance (Table 3). Adapting or generalizing to unseen viewpoints is difficult because the viewpoint affects all objects in the scene, and the model must implicitly estimate the robot's position relative to the new camera position. Therefore, demonstration data with diverse viewpoints must be incorporated into the training dataset.

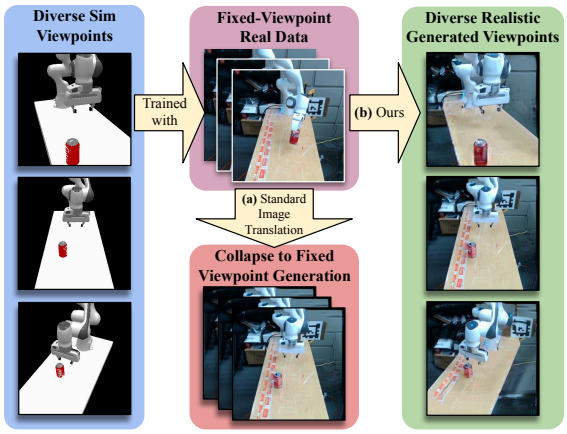

Figure 1. (a) Standard image translation methods fail to generalize to new viewpoints when trained on fixed-viewpoint target domain dataset. (b) Our method enables realistic generation of unseen viewpoints

To this end, we propose to instead *generate* simulated demonstrations covering diverse viewpoints and propose a method for bridging the visual sim2real gap. One option is to add visual domain randomization in simulation for lighting, textures, and colors [1, 2, 9]. Other researchers opt instead to improve simulator realism [6] – a labor-intensive engineering effort which must be done for every scene. In leiu of simulation engineering, we propose learning a generative, segmentation-conditioned unpaired image-to-image translation model mapping from sim to real. Our model is trained on a small, real-world play dataset collected from a single fixed camera and is able to generate a range of diverse viewpoints (Figure 1). The output of our entire pipeline is a dataset of synthetic demonstrations which can be combined with a small amount of real demonstrations and fed into any downstream learning algorithm.

Code and videos can be found at www.sites.google.com/view/sim2real-viewpoints.

*[1]Institute of Robotics and Intelligent Machine, Georgia Institute of Technology, Atlanta, GA, USA. Emails: {jcoholich, jwit3, zkira}@gatech.edu.

Table 1. This table shows the success rates of ACT policies [11] trained on sim image observations translated by a multi-task translation model. Policies labeled "ours" are trained on 100 fixed-camera human demonstrations and 1500 synthetic demonstrations.

	Pick coke		Stack blocks		Stack cup	
Camera Type	**Real Only**	**Ours**	**Real Only**	**Ours**	**Real Only**	**Ours**
Fixed Cam	15/15	14/15	14/15	14/15	8/15	15/15
ID Cam	1/15	2/15	0/15	4/15	0/15	8/15
OOD Cam	0/15	3/15	0/15	1/15	0/15	2/15

2. Viewpoint Sim2Real Image Translation

We propose a novel unpaired sim2real image-to-image translation model which can be trained on only fixed-viewpoint real data and generate unseen real views from simulated image observations. We train our model with the GAN loss [3], a modified InfoNCE [7] loss like in CUT [8], and a highly regularized discriminator that takes randomly sampled and rotated patches. Additionally, we propose a novel segmentation-based InfoNCE loss on generator features.

We leverage the simulator to generate ground-truth segmentation maps for each sim image. We propose an InfoNCE loss which clusters generator decoder features based on segmentation category in order to ensure that semantic segmentation boundaries are preserved during translation. We note that this is a crucial aspect to preserve when training object-centric manipulation policies.

The Segmentation NCE (SegNCE) loss is defined in Equation 1, where ℓ_{NCE} is the NCE contrastive loss. Each sim image contains C segmentation classes and each randomly-sampled feature $\mathbf{z}_i \in \mathcal{Z}$ from encoder layer l generated from datapoint $d \sim D_A$ has an associated class label $y_i \in \mathcal{Y}$. All features that are members of the same segmentation class as the query feature $\mathbf{z}_i$ are positive samples assigned the index j.

$$\ell_{\text{SegNCE}}(l, \mathcal{Z}, i, \mathcal{Y}) = \frac{1}{\left\{ j \middle| \substack{j \in 1..|\mathcal{Z}| \\ y_j = y_i \\ i \neq j} \right\}} \sum_{\left\{ j \middle| \substack{j \in 1..|\mathcal{Z}| \\ y_j = y_i \\ i \neq j} \right\}} \ell_{\text{NCE}}(l, \mathbf{z}_i, \mathcal{Z}, j) \quad (1)$$

Table 2 gives the scores of our proposed method against baselines and ablations. Our proposed method obtains the lowest FID score by a wide margin on ID Camera. The novel patch discriminator D has the largest impact on all metrics and is essential in preventing output collapse to a fixed viewpoint. "No Translation" represents the upper bound for the mIOU and FID Score. An mIOU values less than 1.0 here reflects errors in our test set labeling pipeline.

3. Training Robot Policies

Results for a `pick coke` task are shown in Table 3. We observe that our method is necessary for viewpoint robustness, as compared to ACT models trained on fixed-cam human demonstrations only, which obtain a success rate of 0 under any camera variation. We also observe that our method is necessary to bridge the sim2real gap, since the policies trained on sim demos without translation obtain a similarly negative result on ID and OOD cameras. Finally, we can see that performance scales positively with the number of translated demonstrations produced by our method.

We present preliminary results for different tasks in Table 1, where sim observations for each task are translated with the same model.

Table 2. Unpaired Image Translation FID(↓) and mIOU(↑) scores. Metrics are averaged across two runs.

Method	**ID Camera**		**OOD Camera**	
	FID(↓)	mIOU(↑)	FID(↓)	mIOU(↑)
No Translation	411.2	0.88	379.7	0.85
CUT [8]	312.3	0.54	344.3	0.52
CycleGAN [12]	323.7	0.53	360.0	0.51
Basic D	331.5	0.54	362.0	0.43
Without SegNCE loss	177.6	0.79	226.5	0.70
Without $\tilde{\rho}_l(\cdot)$	171.6	0.82	**212.6**	0.70
Ours	**167.9**	0.82	219.7	0.70
*Ours 360x360	194.4	**0.93**	238.1	**0.76**
*Ours All Tasks 360x360	193.6	0.6	222.7	0.75

Table 3. Success rates of ACT [11] policies on the `pick coke` task. All models were cotrained with 100 fixed-camera human teleoperated demos. Policies are trained with a single image observation and no proprioception. ID and OOD refer to the distribution of camera angles used to train the image translation model. † indicates no translation method was used.

Policy Training Demonstrations		**Evaluation Viewpoints**		
Simulated demos	Sim Camera Randomization	Fixed	ID	OOD
0	-	5/15	0/15	0/15
500†	ID	9/15	0/15	1/15
500†	ID + OOD	2/15	0/15	0/15
500	ID	12/15	4/15	1/15
500	ID + OOD	9/15	2/15	2/15
1500	ID	**14/15**	**6/15**	**4/15**
1500	ID + OOD	14/15	5/15	2/15

References

[1] 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. 1

[2] Ricardo Garcia, Robin Strudel, Shizhe Chen, Etienne Arlaud, Ivan Laptev, and Cordelia Schmid. Robust visual sim-to-real transfer for robotic manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 992–999. ieee, 2023. 1

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[4] Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and match: Supercharging imitation with regularized optimal transport. In *Conference on Robot Learning*, pages 32–43. PMLR, 2023. 1

[5] Michelle A Lee, Yuke Zhu, Krishnan Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. In *2019 International conference on robotics and automation (ICRA)*, pages 8943–8950. IEEE, 2019. 1

[6] Xuanlin Li, Kyle Hsu, Jiayuan Gu, Karl Pertsch, Oier Mees, Homer Rich Walke, Chuyuan Fu, Ishikaa Lunawat, Isabel Sieh, Sean Kirmani, et al. Evaluating real-world robot manipulation policies in simulation. *arXiv preprint arXiv:2405.05941*, 2024. 1

[7] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[8] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 319–345. Springer, 2020. 2

[9] Lerrel Pinto, Marcin Andrychowicz, Peter Welinder, Wojciech Zaremba, and Pieter Abbeel. Asymmetric actor critic for image-based robot learning. *arXiv preprint arXiv:1710.06542*, 2017. 1

[10] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, 2023. 1

[11] Tony Z Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. *arXiv preprint arXiv:2304.13705*, 2023. 2

[12] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2