End-to-End Scene Augmentation for Robust Robot Manipulation Learning

Chengbo Yuan[*]
student ID: 2024211275

Shaoting Zhu[*]
Student ID: 2024311565

Suraj Joshi[*]
Student ID: 2024280205

1 Background

We have seen huge progress in developing foundation models in the domain of vision and language. The foundation models are very generalizable and can be used for multiple tasks with zero-shot or few-shot learning. However, we have not seen good foundation models for robotics which can be attributed to the lack of large-scale diverse datasets for robotics. Collecting large-scale datasets with large variations in the scenes requires engineering heavy automation or laborious teleoperations using humans. Some methods aim to solve this issue by generating datasets using simulated, but simulated datasets fail to capture the underlying distribution of the real world and there also exists some gap in sim2real transfer.

In this work, we aim to use diffusion models to augment new samples into an existing dataset by manipulating background scenes so that learned policies using the augmented dataset can be applied to a broad range of robot learning tasks. We finally aim to train a policy network using the augmented dataset, which can be utilized in diverse downstream tasks.

2 Related works

Visual Data Augmentation for robot learning. Some of the early techniques focus on studying how different perturbations on the visual dataset e.g. changing lighting, scaling, and much more can lead to better generalization[1, 2]. However, these techniques can only robustly handle the same scene under varied conditions. Some of the more modern techniques use diffusion models to diversify the scenes by adding new objects and distractors to the scenes in the existing dataset[3, 4]. They are also not very scalable as they are not completely generating new scenes instead manipulating the existing scenes by adding some objects. Also, they require textual input to modify the scene in the given image, making these processes time-consuming. GreenAug[5] tries to completely generate new scenes from the existing scene using chroma key algorithms. Still, it requires data to be collected with green backgrounds and the scenes overlayed are some predefined scenes that might not be semantically compatible with the object to be manipulated. We aim to generate completely new scenes that capture the underlying physics of the world and are also semantically compatible with the object to be manipulated in the scene.

3 Proposed Method

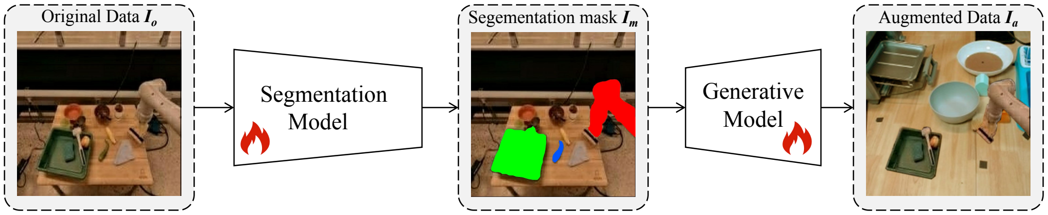

Figure 1: Overview pipeline of our proposed method. We train a segmentation model and a generative model, and augment the robot dataset by changing the image's background.

3.1 Math Definition

Given an unseen robot manipulation dataset $\mathcal{D}$, our goal is augmenting the $\mathcal{D}$ to $\mathcal{D}_a$ using a trained segmentation model and a generative model. We change the background while making the robot arm and the object part the same. For downstream tasks, we may use this augmented dataset to train a generalizable robust robot action policy, $\pi(\mathcal{D}_a)$.

[*] Equal contribution (authors listed in alphabetical order).

Preprint. Course project.

3.2 Basic Pipeline

Specifically, we handle this task in following three steps.

1) We first create a dataset of <RGB image, segmentation mask> pair. We use segmentation models to obtain masks for real-world datasets, while for simulated datasets, masks can be obtained from the simulator.

2) We then fine-tune a segmentation model and a generative model, especially for the robotic arm images. The segmentation model takes in an RGB image and outputs a segmentation mask. The generative model takes in the segmentation mask and outputs an RGB image. In addition, we can use framework in CycleGAN[6] to improve the consistency between original dataset and augmented dataset.

3) Finally, we augment an unseen robot manipulation dataset to expand the background. We first use segmentation model to get the mask of object and robot arm $\mathcal{I}_m(\mathcal{M}_i)$, and invert it to get the mask of the background. Then generative model is used to generate images $\mathcal{I}_o$ containing different types of background. The generated images and the original images form the augmented dataset $\mathcal{D}_a$ together.

4 Resource Survey

In this section, we provide a brief introduction to the datasets, segmentation labeling, and diffusion model resources that will be used in our project.

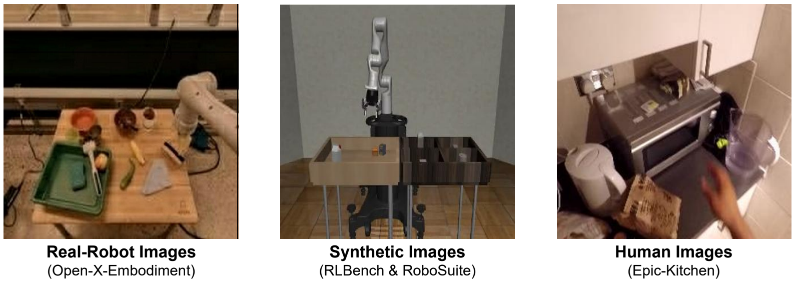
Real-Robot Images
(Open-X-Embodiment)
Synthetic Images
(RLBench & RoboSuite)
Human Images
(Epic-Kitchen)

Figure 2: Three types of data resource we used in our project.

4.1 Dataset Resources

Due to the scarcity of real-world robot datasets, we plan to utilize a combination of real-world robot datasets, synthetic datasets, and egocentric human video datasets to train our model (Figure 2).

Real-world Robot Images (Open-X-Embodiment [7]). For real-world robot images, we will use the Open-X-Embodiment (OXE, [7]) dataset, which is the largest collection of real-world robot data that over 20 robot datasets. For each task in OXE, we randomly select one trajectory and sample 8 images from it, resulting in a total of 4,310 third-person view images.

Sythetic Robot Images (RLBench[8] and RoboSuite[9]). Generating synthetic images from simulation is more convenient, though such images often lack realistic textures and physics. We aim to generate approximately 1,500 simulated images.

Egocentric Human Images (Epic-Kitchens [10]). Egocentric human videos provide valuable insights into hand-object interactions in the physical world, which share similarities with robot manipulation. As a supplementary data source, we will use the Epic-Kitchens dataset [10], extracting around 4,000 images to enhance model training.

4.2 Segmentation Labeling

For real-robot images, we find that the segmentation performance of off-the-shelf models like GroundingDINO [11] and the Segment Anything Model (SAM) [12] is suboptimal. So we decided to train our own segmentation model. We will manually annotate the real-robot images for segmentation model training, leveraging the open-source ISAT tool [13] in combination with SAM [12] to improve labeling efficiency. For synthetic data, we can directly obtain robot and object masks from simulation. For human videos, we will use the state-of-the-art hand-object segmentation model EgoHOS [14] to generate masks for hands and manipulation objects.

4.3 Diffusion Model

Once segmentation for the robot and manipulated objects is complete, the diffusion model will be used to generate entire images, with background inpainting. To achieve this, we plan to train ControlNet [15], one of the most widely used models for conditional image generation. Additionally, more advanced models such as Uni-ControlNet [16] may also be considered.

References

[1] Misha Laskin, Kimin Lee, Adam Stooke, Lerrel Pinto, Pieter Abbeel, and Aravind Srinivas. Reinforcement learning with augmented data. *Advances in neural information processing systems*, 33:19884–19895, 2020.

[2] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3153–3160. IEEE, 2024.

[3] Zhao Mandi, Homanga Bharadhwaj, Vincent Moens, Shuran Song, Aravind Rajeswaran, and Vikash Kumar. Cacti: A framework for scalable multi-task multi-scene visual imitation learning. *arXiv preprint arXiv:2212.05711*, 2022.

[4] Tianhe Yu, Ted Xiao, Austin Stone, Jonathan Tompson, Anthony Brohan, Su Wang, Jaspiar Singh, Clayton Tan, Jodilyn Peralta, Brian Ichter, et al. Scaling robot learning with semantically imagined experience. *arXiv preprint arXiv:2302.11550*, 2023.

[5] Eugene Teoh, Sumit Patidar, Xiao Ma, and Stephen James. Green screen augmentation enables scene generalisation in robotic manipulation. *arXiv preprint arXiv:2407.07868*, 2024.

[6] 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.

[7] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, Albert Tung, Alex Bewley, Alex Herzog, Alex Irpan, Alexander Khazatsky, Anant Rai, Anchit Gupta, Andrew Wang, Andrey Kolobov, Anikait Singh, Animesh Garg, Aniruddha Kembhavi, Annie Xie, Anthony Brohan, Antonin Raffin, Archit Sharma, Arefeh Yavary, Arhan Jain, Ashwin Balakrishna, Ayzaan Wahid, Ben Burgess-Limerick, Beomjoon Kim, Bernhard Schölkopf, Blake Wulfe, Brian Ichter, Cewu Lu, Charles Xu, Charlotte Le, Chelsea Finn, Chen Wang, Chenfeng Xu, Cheng Chi, Chenguang Huang, Christine Chan, Christopher Agia, Chuer Pan, Chuyuan Fu, Coline Devin, Danfei Xu, Daniel Morton, Danny Driess, Daphne Chen, Deepak Pathak, Dhruv Shah, Dieter Büchler, Dinesh Jayaraman, Dmitry Kalashnikov, Dorsa Sadigh, Edward Johns, Ethan Foster, Fangchen Liu, Federico Ceola, Fei Xia, Feiyu Zhao, Felipe Vieira Frujeri, Freek Stulp, Gaoyue Zhou, Gaurav S. Sukhatme, Gautam Salhotra, Ge Yan, Gilbert Feng, Giulio Schiavi, Glen Berseth, Gregory Kahn, Guangwen Yang, Guanzhi Wang, Hao Su, Hao-Shu Fang, Haochen Shi, Henghui Bao, Heni Ben Amor, Henrik I Christensen, Hiroki Furuta, Homanga Bharadhwaj, Homer Walke, Hongjie Fang, Huy Ha, Igor Mordatch, Ilija Radosavovic, Isabel Leal, Jacky Liang, Jad Abou-Chakra, Jaehyung Kim, Jaimyn Drake, Jan Peters, Jan Schneider, Jasmine Hsu, Jay Vakil, Jeannette Bohg, Jeffrey Bingham, Jeffrey Wu, Jensen Gao, Jiaheng Hu, Jiajun Wu, Jialin Wu, Jiankai Sun, Jianlan Luo, Jiayuan Gu, Jie Tan, Jihoon Oh, Jimmy Wu, Jingpei Lu, Jingyun Yang, Jitendra Malik, João Silvério, Joey Hejna, Jonathan Booher, Jonathan Tompson, Jonathan Yang, Jordi Salvador, Joseph J. Lim, Junhyek Han, Kaiyuan Wang, Kanishka Rao, Karl Pertsch, Karol Hausman, Keegan Go, Keerthana Gopalakrishnan, Ken Goldberg, Kendra Byrne, Kenneth Oslund, Kento Kawaharazuka, Kevin Black, Kevin Lin, Kevin Zhang, Kiana Ehsani, Kiran Lekkala, Kirsty Ellis, Krishan Rana, Krishnan Srinivasan, Kuan Fang, Kunal Pratap Singh, Kuo-Hao Zeng, Kyle Hatch, Kyle Hsu, Laurent Itti, Lawrence Yunliang Chen, Lerrel Pinto, Li Fei-Fei, Liam Tan, Linxi "Jim" Fan, Lionel Ott, Lisa Lee, Luca Weihs, Magnum Chen, Marion Lepert, Marius Memmel, Masayoshi Tomizuka, Masha Itkina, Mateo Guaman Castro, Max Spero, Maximilian Du, Michael Ahn, Michael C. Yip, Mingtong Zhang, Mingyu Ding, Minho Heo, Mohan Kumar Srirama, Mohit Sharma, Moo Jin Kim, Naoaki Kanazawa, Nicklas Hansen, Nicolas Heess, Nikhil J Joshi, Niko Suenderhauf, Ning Liu, Norman Di Palo, Nur Muhammad Mahi Shafiullah, Oier Mees, Oliver Kroemer, Osbert Bastani, Pannag R Sanketi, Patrick "Tree" Miller, Patrick Yin, Paul Wohlhart, Peng Xu, Peter David Fagan, Peter Mitrano, Pierre Sermanet, Pieter Abbeel, Priya Sundaresan, Qiuyu Chen, Quan Vuong, Rafael Rafailov, Ran Tian, Ria Doshi, Roberto Mart'in-Mart'in, Rohan Baijal, Rosario Scalise, Rose Hendrix, Roy Lin, Runjia Qian, Ruohan Zhang, Russell Mendonca, Rutav Shah, Ryan Hoque, Ryan Julian, Samuel Bustamante, Sean Kirmani, Sergey Levine, Shan Lin, Sherry Moore, Shikhar

Bahl, Shivin Dass, Shubham Sonawani, Shubham Tulsiani, Shuran Song, Sichun Xu, Siddhant Haldar, Siddharth Karamcheti, Simeon Adebola, Simon Guist, Soroush Nasiriany, Stefan Schaal, Stefan Welker, Stephen Tian, Subramanian Ramamoorthy, Sudeep Dasari, Suneel Belkhale, Sungjae Park, Suraj Nair, Suvir Mirchandani, Takayuki Osa, Tanmay Gupta, Tatsuya Harada, Tatsuya Matsushima, Ted Xiao, Thomas Kollar, Tianhe Yu, Tianli Ding, Todor Davchev, Tony Z. Zhao, Travis Armstrong, Trevor Darrell, Trinity Chung, Vidhi Jain, Vikash Kumar, Vincent Vanhoucke, Wei Zhan, Wenxuan Zhou, Wolfram Burgard, Xi Chen, Xiangyu Chen, Xiaolong Wang, Xinghao Zhu, Xinyang Geng, Xiyuan Liu, Xu Liangwei, Xuanlin Li, Yansong Pang, Yao Lu, Yecheng Jason Ma, Yejin Kim, Yevgen Chebotar, Yifan Zhou, Yifeng Zhu, Yilin Wu, Ying Xu, Yixuan Wang, Yonatan Bisk, Yongqiang Dou, Yoonyoung Cho, Youngwoon Lee, Yuchen Cui, Yue Cao, Yueh-Hua Wu, Yujin Tang, Yuke Zhu, Yunchu Zhang, Yunfan Jiang, Yunshuang Li, Yunzhu Li, Yusuke Iwasawa, Yutaka Matsuo, Zehan Ma, Zhuo Xu, Zichen Jeff Cui, Zichen Zhang, Zipeng Fu, and Zipeng Lin. Open X-Embodiment: Robotic learning datasets and RT-X models. `https://arxiv.org/abs/2310.08864`, 2023.

[8] Stephen James, Zicong Ma, David Rovick Arrojo, and Andrew J. Davison. Rlbench: The robot learning benchmark & learning environment. *IEEE Robotics and Automation Letters*, 2020.

[9] Yuke Zhu, Josiah Wong, Ajay Mandlekar, Roberto Martín-Martín, Abhishek Joshi, Soroush Nasiriany, and Yifeng Zhu. robosuite: A modular simulation framework and benchmark for robot learning. In *arXiv preprint arXiv:2009.12293*, 2020.

[10] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

[11] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[13] Shuwei Ji and Hongyuan Zhang. ISAT with Segment Anything: An Interactive Semi-Automatic Annotation Tool, 2023. Updated on 2023-06-03.

[14] Lingzhi Zhang, Shenghao Zhou, Simon Stent, and Jianbo Shi. Fine-grained egocentric hand-object segmentation: Dataset, model, and applications. In *European Conference on Computer Vision*, pages 127–145. Springer, 2022.

[15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[16] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K. Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 2023.