IFOR: Iterative Flow Minimization for Robotic Object Rearrangement

Ankit Goyal[1,2]*, Arsalan Mousavian[1], Chris Paxton[1], Yu-Wei Chao[1], Brian Okorn[1,3]*
Jia Deng[2], Dieter Fox[1]
[1]NVIDIA, [2]Princeton University, [3]Carnegie Mellon University

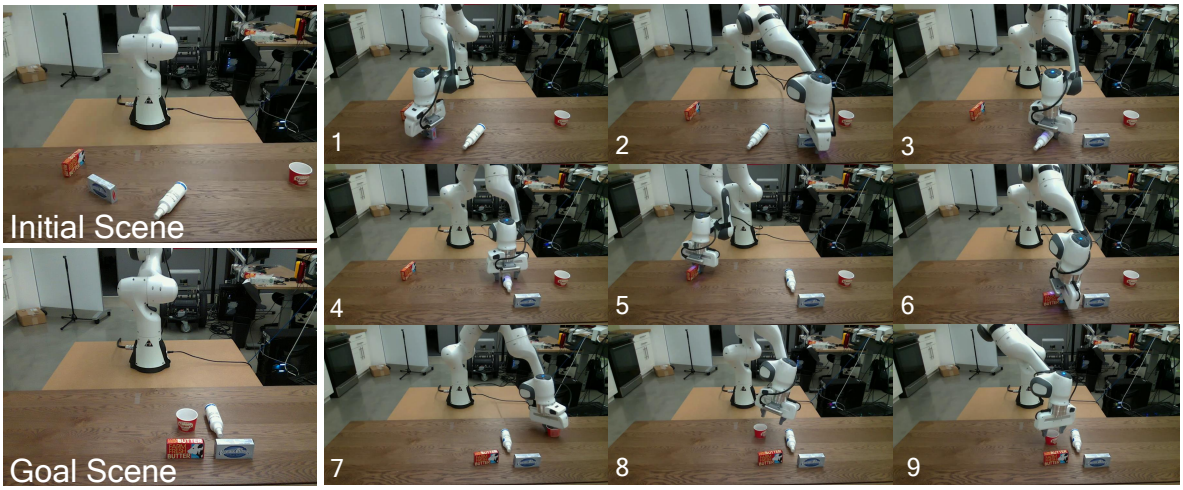

Figure 1. An example of *IFOR* being applied to real data. The initial and goal scenes are shown on the left. Our approach allows the robot to repeatedly identify transformations that will minimize the flow for various objects between the current and goal scenes. It can then repeatedly grasp, move, and place objects, rotating as necessary, in order to achieve the configuration in the goal scene.

Abstract

Accurate object rearrangement from vision is a crucial problem for a wide variety of real-world robotics applications in unstructured environments. We propose IFOR*, Iterative Flow Minimization for Robotic Object Rearrangement, an end-to-end method for the challenging problem of object rearrangement for unknown objects given an RGBD image of the original and final scenes. First, we learn an optical flow model based on RAFT to estimate the relative transformation of the objects purely from synthetic data. This flow is then used in an iterative minimization algorithm to achieve accurate positioning of previously unseen objects. Crucially, we show that our method applies to cluttered scenes, and in the real world, while training only on synthetic data. Videos are available at* https://imankgoyal.github.io/ifor.html.

1. Introduction

Object rearrangement is the capability of an embodied agent to physically re-configure the objects in a scene into a desired goal configuration [1]. It is an essential skill in day-to-day activities like setting a dining table, putting away groceries, and organizing a desk.

With varying task setups, the desired goal state can be provided in different forms, for instance, a compact state representation [7, 14] or natural language descriptions [9, 11]. In this work, we address the rearrangement task where the goal state is specified by an RGB-D image [8,10], as shown in Fig. 1. This setup lends itself well to many scenarios where the goal state can be snapped once, either in the first place or from a one-time demonstration.

Traditionally, object rearrangement problems have been studied in the robotics community, often in the context of Task and Motion Planning (TAMP) [4]. Despite much recent progress [3,5,6], most TAMP approaches still rely on a strict set of assumptions on the perception front. Recent efforts in robotics have attempted to relax these constraints by leveraging the power of deep learning. A recent approach called NeRP, proposed by Qureshi et al. [10], has allowed for rearranging objects unseen at the training time, by representing the observed objects with learned embeddings. It also removes the need of explicit object pose estimation for planning by leveraging recent progress on learning-based grasp planners [12] and collision detectors [2]. However, NeRP only allows moving objects with 2D in-plane transla-

*Work done while authors were interns at NVIDIA

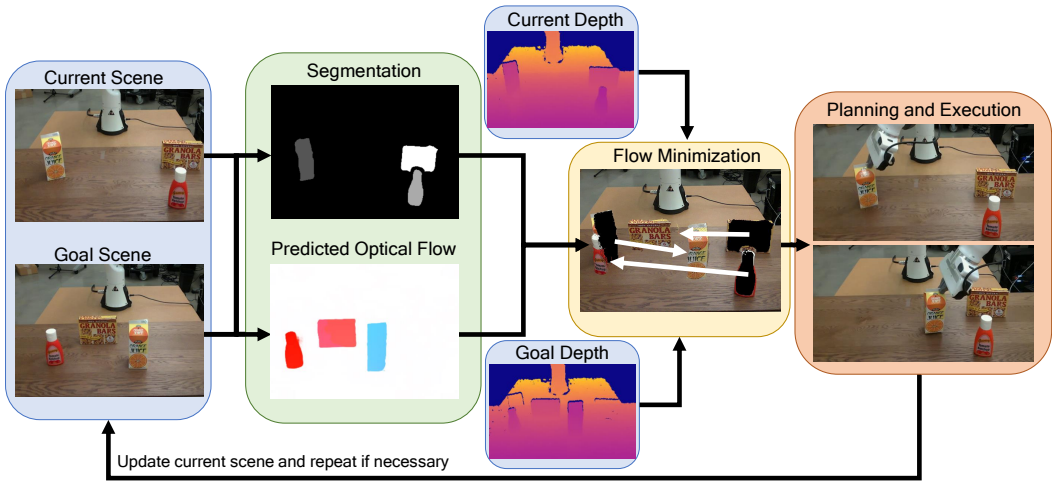

Figure 2. Overview of the *IFOR* algorithm. *IFOR* takes as input RGB+D images of a current and a goal scene, and uses these to make predictions as to which objects should move and by which transformations, using RAFT to estimate optical flow. This is then sent to a robot planning and execution pipeline which is capable of grasping of unknown objects and motion planning in scenes with unknown geometry.

tions on the table surface and allows no change in their orientation. This prevents its applications in realistic scenarios that require moving objects with more complex transformations such as those shown in Fig. 1.

We propose a new approach to image-guided robotic object rearrangement with RGB-D input. It achieves, for the first time to the best our of knowledge, the ability to handle unknown objects with translation as well as planar rotations. The key to our method is re-formulating object rearrangement as an iterative minimization of optical flow between the current observed image and the goal image. By using optical flow as an intermediate representation, we can capitalize on the cutting edge development in flow estimation models [13]. Using this estimated flow, together with the depth input and generic object segmentation models [15], we can obtain dense 3D correspondences for each object. This provides a general representation that allows us to solve for the desired transformation of objects with simple optimization. Furthermore, with such a general representation, our method trained entirely on synthetic data transfers well to the real world in a zero shot manner.

2. Method

IFOR takes as input the RGB-D images of the current and the goal scene, and iteratively generates a pick-and-place action for one object at a time. At each iteration, the RGB-D image of current and goal scene is passed through two components: (1) perception and (2) planning (Fig. 2). The perception component is responsible for estimating the relative transformation of all objects between the current and goal scene. Given the estimated transforms, the planning component selects an object to be moved along with the required transformation, by taking into account collision and kinematic feasibility. Finally, after executing the planned pick-and-place action, the system will take a new observation of the scene and repeat the process.

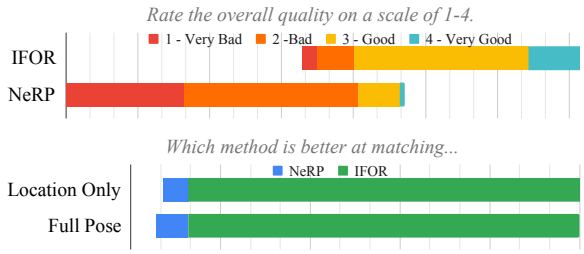

Figure 3. User scores for *IFOR* vs. NeRP [10]. When asked to rate performance of the two methods on a scale of 1-4, users preferred *IFOR* by a wide margin. Users chose *IFOR* over NeRP in almost all situations, when looking at either position only (94%) or full pose (position and orientation, 92%).

3. Experiments

We evaluated on 6 scenes, where each scene has between 2 to 5 objects in the initial configuration and a distinct goal configuration. In order to quantitatively evaluate the performance of the methods in the real world, we conducted a user study with 10 users, where we asked users to select the method that performed better: *IFOR* or NeRP [10]. We also asked users to rate *IFOR* and NeRP on a scale of 1-4, where 1 is "very bad" and 4 is "very good".

All the components of the pick-and-place system were the same for both *IFOR* and NeRP, except the estimation of the objects' final pose. Since NeRP does not handle changing the orientation, we asked users to rank the two methods only based on translation as well as considering both rotation and translation. Fig. 3 shows that users significantly prefer *IFOR* over NeRP on both the settings.

References

[1] Dhruv Batra, Angel X. Chang, Sonia Chernova, Andrew J. Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, Manolis Savva, and Hao Su. Rearrangement: A challenge for embodied AI. *arXiv preprint arXiv:2011.01975*, 2020. 1

[2] Michael Danielczuk, Arsalan Mousavian, Clemens Eppner, and Dieter Fox. Object rearrangement using learned implicit collision functions. In *ICRA*, 2020. 1

[3] Danny Driess, Jung-Su Ha, and Marc Toussaint. Deep visual reasoning: Learning to predict action sequences for task and motion planning from an initial scene image. In *RSS*, 2020. 1

[4] Caelan Reed Garrett, Rohan Chitnis, Rachel Holladay, Beomjoon Kim, Tom Silver, Leslie Pack Kaelbling, and Tomás Lozano-Pérez. Integrated task and motion planning. *Annual Review of Control, Robotics, and Autonomous Systems*, 4(1):265–293, 2021. 1

[5] Caelan Reed Garrett, Chris Paxton, Tomás Lozano-Pérez, Leslie Pack Kaelbling, and Dieter Fox. Online replanning in belief space for partially observable task and motion problems. In *ICRA*, 2020. 1

[6] Brian Ichter, Pierre Sermanet, and Corey Lynch. Broadly-exploring, local-policy trees for long-horizon task planning. *arXiv preprint arXiv:2010.06491*, 2020. 1

[7] Jennifer E. King, Marco Cognetti, and Siddhartha S. Srinivasa. Rearrangement planning using object-centric and robot-centric action spaces. In *ICRA*, 2016. 1

[8] Yann Labbé, Sergey Zagoruyko, Igor Kalevatykh, Ivan Laptev, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Monte-Carlo Tree Search for efficient visually guided rearrangement planning. *RA-L*, 5(2):3715–3722, 2020. 1

[9] Weiyu Liu, Chris Paxton, Tucker Hermans, and Dieter Fox. StructFormer: Learning spatial structure for language-guided semantic rearrangement of novel objects. *arXiv preprint arXiv:2110.10189*, 2021. 1

[10] Ahmed H. Qureshi, Arsalan Mousavian, Chris Paxton, Michael C. Yip, and Dieter Fox. NeRP: Neural rearrangement planning for unknown objects. In *RSS*, 2021. 1, 2

[11] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. CLIPort: What and where pathways for robotic manipulation. In *CoRL*, 2021. 1

[12] Martin Sundermeyer, Arsalan Mousavian, Rudolph Triebel, and Dieter Fox. Contact-GraspNet: Efficient 6-DoF grasp generation in cluttered scenes. In *ICRA*, 2021. 1

[13] Zachary Teed and Jia Deng. RAFT: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 2

[14] Luca Weihs, Matt Deitke, Aniruddha Kembhavi, and Roozbeh Mottaghi. Visual room rearrangement. In *CVPR*, 2021. 1

[15] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A SimulAted Part-based Interactive ENvironment. In *CVPR*, 2020. 2