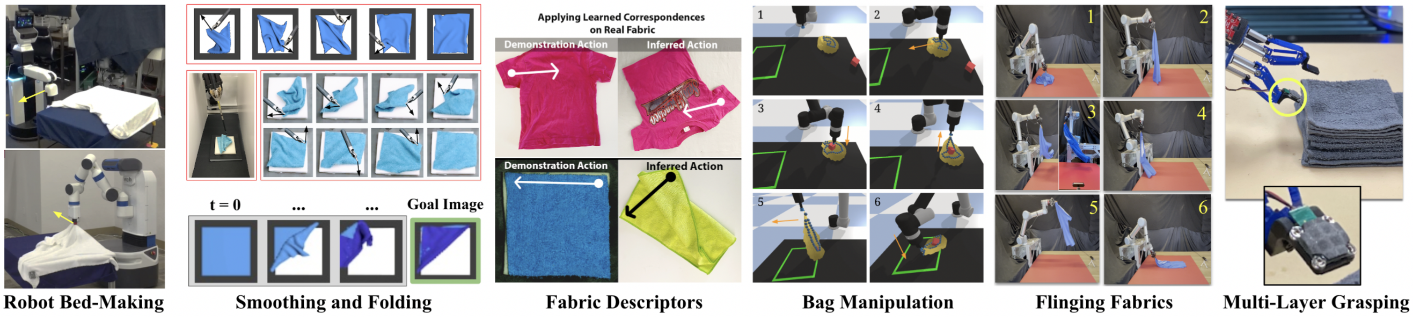
Representations for Robot Manipulation

Daniel Seita
April 13, 2023
http://www.cs.cmu.edu/~dseita/ dseita@andrew.cmu.edu



Robot Bed-Making Smoothing and Folding Fabric Descriptors Bag Manipulation Flinging Fabrics Multi-Layer Grasping

Figure 1: Overview of my prior work on deformable object manipulation. From left to right: robot bed-making [10], fabric smoothing and folding [11, 4], using descriptors as a fabric representation [3], opening a simulated bag, inserting an item, and moving it to a target [12], flinging fabrics [1], and multi-layer fabric grasping [15].

In robot manipulation, it is often assumed implicitly or explicitly that the items being manipulated are *rigid* or approximated as rigid. However, many items that humans interact with on a daily basis are deformable, such as clothing, cables, strings, napkins, food, bags, and fluids. An autonomous robot that can reliably manipulate deformables in unstructured settings could thus address applications as diverse as cleaning an environmental disaster site, tending agricultural crops, or performing assistive dressing. Despite decades of research attention, deformable object manipulation remains challenging. A primary difficulty is estimating the configuration of a deformable object, or a specification of all its points. For example, with a rigid cube, knowing the location of a fixed point relative to the cube's center is sufficient to describe the full cube arrangement. With a fabric, however, parts can remain fixed while other components shift. This makes it difficult to describe a sufficiently expressive representation of the fabric, especially under self-occlusions.

To achieve greater progress in robot manipulation of diverse deformable objects, I advocate for an increased focus on learning and developing appropriate representations. To clarify the terminology, I view *representation learning* as deciding on how observations and actions should be provided or utilized in a machine learning system. A straightforward observational representation is to pass raw images of the object to a deep neural network policy, which then outputs a low-level robot action; however, there may be other representations that lead to more sample-efficient imitation or reinforcement learning, or improved final performance. In Section 1, I describe multiple innovations I have developed in representations for fabric manipulation, dynamic deformable manipulation, object rearrangement, and learning from point clouds. I believe that further innovations will be needed for improving robotic manipulation, and I propose future directions in Section 2.

1 Prior Research

1.1 Manipulation of Deformable Objects using Machine Learning

Classical approaches for fabric smoothing and folding often employed traditional computer vision algorithms such as corner detection for perception, and then leveraged geometric controllers. For example, one tactic was to use a bimanual robot and have one arm lift the fabric in midair while the second arm grasps a hanging corner [8]. While effective, these methods can have limited generalization across fabric configurations, which may motivate using machine learning. In [10], we developed one of the first applications of end-to-end deep learning for fabric manipulation on a quarter-scale robot bed-making task. We used supervised learning on depth images to train a policy to grasp at a blanket corner. Depth images are a useful observational representation because they are invariant to color and can convey cues about fabric edges and wrinkles.

In subsequent work, we addressed two limitations of the bed-making setup: (1) the need for physical data collection, and (2) only learning the pick point, since we predefined the placing point. In [11], we developed a fabric simulator to generate data in simulation, trained a model-free corner-pulling policy for smoothing fabrics, and transferred the policy to a physical robot via domain randomization. While this approach showed promising results, the policy was limited to smoothing and requires an algorithmic supervisor for other tasks, which might be challenging to implement. We subsequently proposed a model-based method, **VisuoSpatial Foresight (VSF)** [4], which uses a large dataset of environment interactions to learn a dynamics model over (simulated) RGB-D images for planning. We demonstrated high-precision, multi-step fabric smoothing and folding under a single policy, and deployed it on a physical robot [5]. In these prior works, we used RGB-D images as the fabric representation, showcasing its utility for model-free and model-based methods.

1.1.1 Visual and Tactile Observational Representations for Fabric Manipulation

Our prior fabric manipulation work showed promising results, but were limited to training a task-specific policy or required training a heavy-duty video prediction model. Consequently, we developed a visual fabric representation by using dense object descriptors [3] which can facilitate multi-task generalization in a model-free manner. This representation led to *pixel-wise correspondence* among pairs of images, which associates the same regions of fabric in the two images. This facilitated learning from demonstrations by generalizing the same (pick-and-place) action across different fabrics.

We have also used *tactile representations*, which offer local information that might not be present in image data due to occlusions. Precise grasping of multiple fabric layers is a task that may benefit from tactile sensing. It is a prerequisite for many downstream manipulation tasks (such as folding fabric in half twice), yet grasping an incorrect number of layers is a common failure case in prior work [3, 11]. We used data from the ReSkin tactile sensor to train a classifier to determine the number of fabric layers grasped [15]. We used this classifier at test time to deliberately grasp a fixed number of fabric layers.

1.1.2 Action Representations for Dynamic Manipulation

Most prior learning-based approaches for deformable manipulation rely on quasistatic pick-and-place actions. This action space is simple to learn and to implement, but is limited in that a robot cannot easily manipulate items outside its reachable workspace. An appealing alternative is to use *dynamic* manipulation, such as tossing. We explored dynamic manipulation of fixed-endpoint cables [17], free-endpoint cables [7], and single-arm fabric flinging [1]. A unifying theme in these works is the choice of the robot action representation: we defined *parabolic trajectory motions* using a small set of parameters, and used self-supervised procedures to learn these parameters. These tasks have a natural reset mechanism in that a robot grasping the cable or fabric can perform a dynamic motion, and then pull it either taut or in midair to reset its state. After each dynamic motion, we obtained labels automatically by detecting the position of the cables or measuring fabric coverage. The takeaway from these works is that we can perform expressive dynamic manipulation while simplifying the learning problem to predicting a small set of parameters.

1.2 Goal-Conditioned Transporter Networks for Object Rearrangement

From manipulating a dining room table to cleaning a room, rearrangement is a canonical task that humans perform, and a challenge is for robots to perform reliable rearrangement. To get robots to perform 2D goal-driven rearrangement with pick-and-place actions, we proposed a novel neural network, **Goal-Conditioned Transporter Networks (GCTNs)** [12]. This architecture extends the Transporter Network [16] from prior work by allowing it to process a separate goal image to specify a desired target rearrangement. This is useful for deformable object manipulation given the difficulty in defining a goal with a set of finite poses.

We built upon the Transporter Network due to its representation advantages. First, it uses Fully Convolutional Networks, enabling it to make use of translation equivariance and to specify multimodal and non-Gaussian distributions over picking and placing locations. Second, it includes a "transporting" operation which performs an efficient deep feature matching search over all pick-and-place actions. These enable Transporters to be extremely sample-efficient at Behavioral Cloning of 2D tabletop manipulation tasks. Critically, our GCTNs *retain the representation advantages* of the vanilla Transporter Network, and we showed that it outperforms alternative methods that use ground-truth pose information instead of images to embed the goal configuration. To evaluate GCTNs and to benchmark progress in deformable manipulation more generally, we also developed an open-source suite of 12 simulated PyBullet tasks called *DeformableRavens*.

1.3 ToolFlowNet: 3D Manipulation by Predicting Tool Flow from Point Clouds

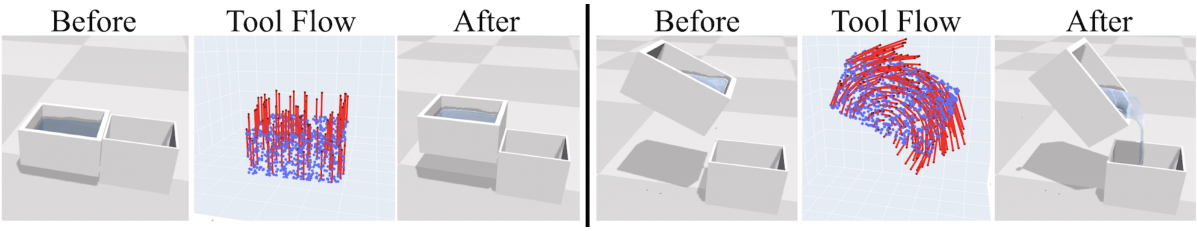


Figure 2: ToolFlowNet applied on a pouring task in simulation, where the tool is the box which contains water. Given a point cloud (colored blue), ToolFlowNet learns dense per-point flow vectors (colored red), which describe the intended 3D motion of each tool point. These are then converted to translation and rotation actions. **Left**: the tool moves upwards. **Right**: the tool rotates to pour water into the target container. In both cases, the flow is subsampled for visual clarity.

Point clouds are a widely available and canonical data modality which preserve the 3D geometry of a scene and are robust to colors, textures, and shadows. Despite significant progress in classification and segmentation from point clouds, policy learning from such a modality remains challenging; most prior works in imitation learning focus on learning policies from images or state information. We proposed a framework for learning policies from segmented point clouds for robotic manipulation with tools [13]. We designed a neural network, **ToolFlowNet**, which predicts *dense per-point flow* on the tool the robot controls, and then uses the flow to derive the transformation that the robot should execute. We call this action representation "tool flow" where each flow vector conveys the 3D movement of each tool point in a point cloud from one time step to the next.

We applied this framework to imitation learning of deformable object manipulation tasks with continuous movement of tools, including scooping and pouring, and demonstrated improved performance over baselines which do not use flow. See Figure 2 for a visualization on the pouring task using FleX simulation. We also used ToolFlowNet to imitate physical scooping demonstrations. From this work, we believe that ToolFlowNet is especially advantageous when learning rotations, as it may be easier to imitate a set of dense flow vectors than representations such as axis-angles and quaternions, even though both represent the same information.

2 Future Research

My prior work has shown that novel observational representations for fabrics (such as with visual descriptors [3] and tactile information [15]) and novel action representations (such as with GCTNs [12], and tool flow [13]) can accelerate and improve policy learning, particularly when robots manipulate diverse and deformable objects. I believe there remains significant room for future work in learning representations and improving robot manipulation. Below, I discuss some major research directions I hope to pursue **in the next 1-5 years**:

Learning from 3D Point Clouds ToolFlowNet [13] is one method to learn from 3D point clouds, and I believe that further progress in this area is critical to advance the state of 3D robot manipulation. In future work, I plan to generalize ToolFlowNet to cases when we do not have ground-truth segmented point clouds and must reason about arbitrary point cloud inputs. Furthermore, I will apply tool flow in scenarios where a robot controls articulated or multiple tools, which could facilitate learning tasks such as cooking or stuffing items in bags [2], where a robot uses items as tools to push at areas of bags to create openings. I will also investigate if per-point flow or alternative action representations can improve sample efficiency of reinforcement learning with point clouds as input.

Reasoning about High-Contact Manipulation I plan to make fundamental progress on tasks that involve frequent contacts between a robot's tool (or end-effector) and objects, or with object-object contact. I believe this is a great fit for robot learning due to the complexities of modeling such contacts, which may motivate a data-driven approach. For example, I plan to explore this in the context of robot food peeling with reasoning about layers, for scooping applications that require reasoning about items to *avoid* when scooping, and for agriculture-related manipulation which could involve retrieving fruits and vegetables entangled in branches. I am also interested in extending applications of my prior work in surgical robotics [6] so that autonomous surgical robots can consider contacts with human tissue in surgical procedures. All these tasks might require novel representations and methods to achieve greater performance capabilities.

Learning Closed-Loop, Visuo-Tactile Policies I will train policies that process *both* visual and tactile representations to obtain the best of both sensor data. I believe this will also accelerate the development of closed-loop policies. For example, I envision a hierarchical framework which uses vision input for a high-level policy and tactile input for a low-level reactive policy. While we have used visual input alone for hierarchical learning [9], using tactile sensing enables a low-level policy to explicitly reason about occlusions which might occur due to the robot gripper or objects in the scene. A task that could make use of such a paradigm might be extending our multi-layer fabric grasping work [15] to allow for closed-loop counting of arbitrary layers, where we use a robot to sequentially grasp and cycle through fabric layers in a pile of fabrics. This approach will also require innovations in action-centric representations to facilitate dexterous maneuvers.

2.1 Long-Term Vision: Multimodal Representations for Deformable Object Manipulation

My **5-10 year vision** is to enable robots to perform fine-grained manipulation tasks of complex deformable objects in unstructured settings. I believe this vision requires fundamental advances in *multimodal representations*. Consequently, I will develop robots that can simultaneously process *multiple sensing modalities as input*, including vision, touch, audio, thermal, language, and variations within each of these categories (such as using different types of tactile sensors). I will propose training procedures that create synergy between the different modalities, so that robots to learn to take actions that preserve the most important sensory modes at a given time. I will also develop robots that can perform *multimodal actions*. For example, I desire robots that can support a variety of actions with appropriate tool representations, so that conditioned on a task, a robot can *decide* on which tool among a set of tools to use, and *how* to use the tool. I believe these multimodal representations will broaden the distribution of tasks which a single robot policy can perform.

From a task perspective, I envision my long-term research facilitating two domains with significant contact-rich, deformable manipulation challenges: *healthcare robots* and *agricultural robots*. For healthcare robots, I hope to develop manipulation procedures for use in robots working with patients, such as in assistive dressing or feeding, or in surgical settings [6]. For agricultural robots, I plan to incorporate my future research into current robot systems [14] to enable them to autonomously harvest fruit from plants in crops and fields.

We are living in an exciting time for robotics, and I look forward to seeing how multimodal representations can improve the manipulation capabilities of future robots.

References

[1] L. Y. Chen*, H. Huang*, E. Novoseller, **D. Seita**, J. Ichnowski, M. Laskey, R. Cheng, T. Kollar, and K. Goldberg. Efficiently Learning Single-Arm Fling Motions to Smooth Garments. In *International Symposium on Robotics Research (ISRR)*, 2022.

[2] L. Y. Chen, B. Shi, **D. Seita**, R. Cheng, T. Kollar, D. Held, and K. Goldberg. AutoBag: Learning to Open Plastic Bags and Insert Objects. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[3] A. Ganapathi, P. Sundaresan, B. Thananjeyan, A. Balakrishna, **D. Seita**, J. Grannen, M. Hwang, R. Hoque, J. Gonzalez, N. Jamali, K. Yamane, S. Iba, and K. Goldberg. Learning Dense Visual Correspondences in Simulation to Smooth and Fold Real Fabrics. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[4] R. Hoque*, **D. Seita***, A. Balakrishna, A. Ganapathi, A. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg. VisuoSpatial Foresight for Multi-Step, Multi-Task Fabric Manipulation. In *Robotics: Science and Systems (RSS)*, 2020.

[5] R. Hoque*, **D. Seita***, A. Balakrishna, A. Ganapathi, A. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg. VisuoSpatial Foresight for Physical Sequential Fabric Manipulation. In *Autonomous Robots*, 2021.

[6] M. Hwang, J. Ichnowski, B. Thananjeyan, **D. Seita**, S. Paradis, D. Fer, T. Low, and K. Goldberg. Automating Surgical Peg Transfer: Calibration with Deep Learning Can Exceed Speed, Accuracy, and Consistency of Humans. In *IEEE Transactions on Automation Science and Engineering (TASE)*, 2022.

[7] V. Lim*, H. Huang*, L. Y. Chen, J. Wang, J. Ichnowski, **D. Seita**, M. Laskey, and K. Goldberg. Planar Robot Casting with Real2Sim2Real Self-Supervised Learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[8] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel. Cloth Grasp Point Detection Based on Multiple-View Geometric Cues with Application to Robotic Towel Folding. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

[9] S. Paradis, M. Hwang, B. Thananjeyan, J. Ichnowski, **D. Seita**, D. Fer, T. Low, J. E. Gonzalez, and K. Goldberg. Intermittent Visual Servoing: Efficiently Learning Policies Robust to Instrument Changes for High-precision Surgical Manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[10] **D. Seita***, N. Jamali*, M. Laskey*, R. Berenstein, A. K. Tanwani, P. Baskaran, S. Iba, J. Canny, and K. Goldberg. Deep Transfer Learning of Pick Points on Fabric for Robot Bed-Making. In *International Symposium on Robotics Research (ISRR)*, 2019.

[11] **D. Seita**, A. Ganapathi, R. Hoque, M. Hwang, E. Cen, A. K. Tanwani, A. Balakrishna, B. Thananjeyan, J. Ichnowski, N. Jamali, K. Yamane, S. Iba, J. Canny, and K. Goldberg. Deep Imitation Learning of Sequential Fabric Smoothing From an Algorithmic Supervisor. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[12] **D. Seita**, P. Florence, J. Tompson, E. Coumans, V. Sindhwani, K. Goldberg, and A. Zeng. Learning to Rearrange Deformable Cables, Fabrics, and Bags with Goal-Conditioned Transporter Networks. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.

[13] **D. Seita**, Y. Wang, S. Shetty, E. Li, Z. Erickson, and D. Held. ToolFlowNet: Robotic Manipulation with Tools via Predicting Tool Flow from Point Clouds. In *Conference on Robot Learning (CoRL)*, 2022.

[14] A. Silwal, F. Yandun, A. Nellithimaru, T. Bates, and G. Kantor. Bumblebee: A Path Towards Fully Autonomous Robotic Vine Pruning. *arXiv preprint arXiv:2112.00291*, 2021.

[15] S. Tirumala*, T. Weng*, **D. Seita***, O. Kroemer, Z. Temel, and D. Held. Learning to Singulate Layers of Cloth Using Tactile Feedback. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[16] A. Zeng, P. Florence, J. Tompson, S. Welker, J. Chien, M. Attarian, T. Armstrong, I. Krasin, D. Duong, V. Sindhwani, and J. Lee. Transporter Networks: Rearranging the Visual World for Robotic Manipulation. In *Conference on Robot Learning (CoRL)*, 2020.

[17] H. Zhang, J. Ichnowski, **D. Seita**, J. Wang, H. Huang, and K. Goldberg. Robots of the Lost Arc: Self-Supervised Learning to Dynamically Manipulate Fixed-Endpoint Cables. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2021.