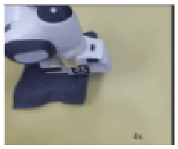
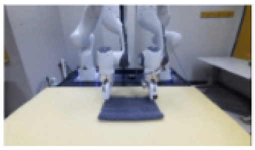
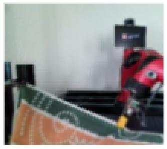
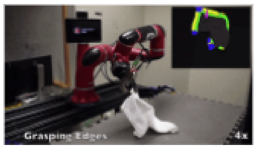
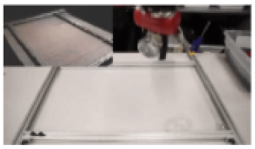

Cloth smoothing


Bimanual cloth folding


Cloth sliding



Cloth edge/corner grasping


Grasping transparent objects

Pouring water

Research Goals

David Held; Personal Website; Lab Website; CV

1 Perceptual Robot Learning

Despite the tremendous recent progress in computer vision and machine learning, robots today are typically confined to interacting with rigid, opaque objects with known object models. However, the objects in our daily lives are often non-rigid, can be transparent or reflective, and are diverse in shape and appearance. I believe that this gap occurs in part because perception and planning are still treated as entirely separate fields, worked on by separate researchers, with little communication between them. To truly advance robotics, I believe that we need to take a more holistic view of perception and planning, designing each module with an awareness of how it will be used within the context of the robotic system.

For example, consider a crumpled cloth. What is the best representation of this cloth for a manipulation task? A standard computer vision output might be a cloth segmentation. However, for manipulation, we need to be able to predict how the cloth will move in response to a robot's actions. A major challenge is in finding a representation for this task that is both sufficient for the manipulation task and also one that can be estimated from visual observations of the crumpled cloth, which may contain severe self-occlusions and ambiguities. There are various representations that can be used for this task, such as a mesh, a latent vector, and many more; we explore novel representations in our work (described below) and show how these new representations lead to greatly improved performance on cloth manipulation tasks.

My research goal is to enable robots to manipulate previously unseen, perceptually challenging, and deformable objects. I am pursuing this goal along three directions: First, I am developing a set of methods for *3D relational decision making*: How can robots reason about the 3D relationships between objects in their environment (or the relationship between parts of a single object) to learn to perform manipulation tasks? I have applied this idea of 3D relational decision making to the challenging tasks of enabling robots to manipulate cloth and articulate unseen objects. Second, I am developing novel approaches for *task-based self-supervised learning*, to enable robots to understand their environment and achieve their objective without requiring manual annotations. Finally, I am developing methods for *active perception*, in which a robot must take actions to better perceive its surroundings. Below, I describe our approaches to each of these challenges.

1.1 3D Relational Reasoning for Robot Manipulation

Recently, learning techniques have achieved impressive results for sequential decision-making tasks for robotics; however, many such results have been shown on locomotion tasks [23, 9, 4, 22, 8, 15, 10, 20]. This is because deep reinforcement approaches usually lack a sophisticated representation of the structures in the environment that are needed for robot manipulation; a common approach for such techniques is to compress the entire environment into a single latent vector and then regress to low-level robot actions. I believe that, in order to apply data-driven decision making techniques to robot manipulation, robots will need to learn to reason about the 3D relationships between objects in their environment and the relationship between parts of a single object. Below, I will describe how I apply this idea of *3D relational decision making* to a variety of robot manipulation tasks.

Deformable object manipulation: Manipulating deformable objects has long been a challenge in robotics due to their high dimensional state representation, complex dynamics, and partial observability due to self-occlusions. However, existing benchmarks used by the reinforcement learning community include only tasks with simple low-dimensional dynamics, such as those with rigid objects. To enable researchers to make more rapid progress on developing methods for deformable object manipulation, we developed SoftGym (CoRL 2020) [11], an open-source simulated benchmark for manipulating deformable objects). Our benchmark enables reproducible research in this important area. SoftGym is quickly growing in popularity and has already been used to enable new and exciting research in deformable object manipulation, both by our group as well as by many others [6, 14, 30].

Building on SoftGym, we explored how to learn a dynamics model for predicting the effect of a robot's actions on a crumpled cloth. Most current deep-learning based methods would compress an image of the cloth into a single latent vector and then train a model to predict the next latent state. However, when trained on a limited amount of data with a network of limited capacity, such latent vector representations tend to either fail to capture important details of the cloth (such as wrinkles or folds) or fail to generalize to novel cloth configurations.

To overcome these challenges, we have developed a novel mesh-based dynamics model that we estimate from a point cloud observation of a crumpled cloth (Figure 1). To reason about the cloth structure from a partial point cloud observation, we infer which *visible* points are connected on the underlying cloth mesh. We then learn a dynamics model over this graph of visible points. Compared to previous learning-based approaches, our mesh-based representation poses a strong inductive bias for learning a realistic dynamics that captures the underlying cloth physics; this inductive bias enables our method to generalize to unseen cloth shapes (e.g. we can train on a square towel and evaluate on a t-shirt), whereas previous learned cloth representations do not achieve such generalization. We show that our method greatly outperforms previous state-of-the-art model-based and model-free reinforcement learning methods for cloth smoothing. Furthermore, we demonstrate zero-shot sim-to-real transfer where we deploy the model trained in simulation on a Franka arm and show that the model can successfully smooth cloths of different materials, geometries, and colors from crumpled configurations (CoRL 2021) [13]. We have followed up on this work with new methods that reason more explicitly about occlusions of a crumpled cloth by combining learning and optimization (RSS 2022) [7].

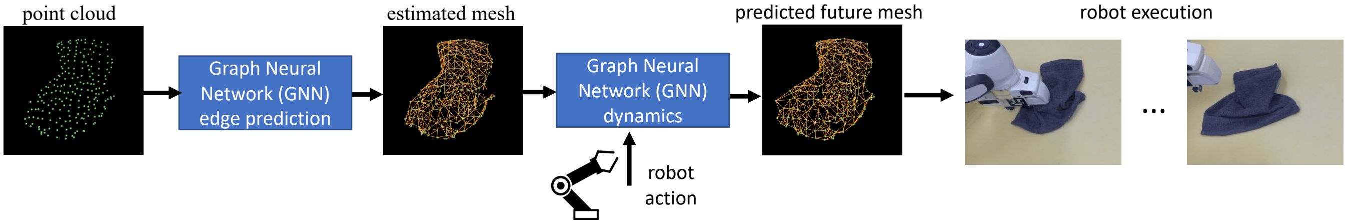

point cloud
Graph Neural Network (GNN) edge prediction
estimated mesh
Graph Neural Network (GNN) dynamics
robot action
predicted future mesh
robot execution
...

Figure 1: Our method achieves robot cloth smoothing by planning using a mesh dynamics model over an estimated visible connectivity graph.

We also tackled the problem of bimanual goal-conditioned cloth folding, a challenging task due to the deformability of cloth. Our insight is that optical flow, a technique normally used for motion estimation in video, can also provide an effective representation for computing the correspondences across a depth image cloth observation and a depth image observation of a goal. Our method, FabricFlowNet (FFN), is a cloth manipulation policy that leverages these estimated correspondences as both an input and as an action representation (Figure 3). FabricFlowNet also elegantly switches between dual-arm and single-arm actions based on the desired goal. We show that FabricFlowNet significantly outperforms state-of-the-art cloth folding policies. Finally, we show that our method generalizes when trained on a single square cloth to be able to fold other cloth shapes, such as T-shirts and rectangular cloths (CoRL 2021) [29]. We are also exploring the use of tactile sensing for cloth manipulation, such as sliding along the edge of a cloth [31] or grasping a specified number of cloth layers from a stack [25].

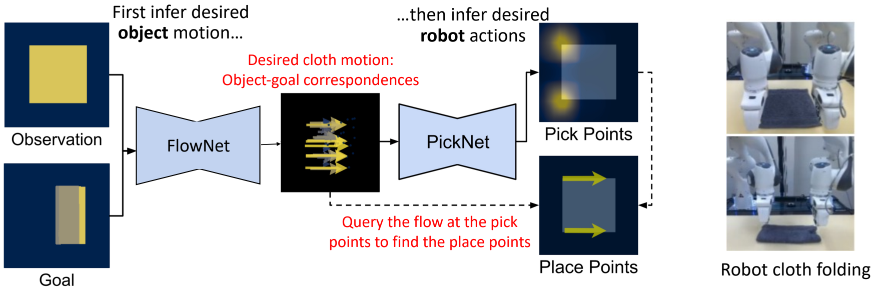

First infer desired **object** motion...
...then infer desired **robot** actions
Desired cloth motion: Object-goal correspondences
Observation
FlowNet
PickNet
Pick Points
Goal
Query the flow at the pick points to find the place points
Place Points
Robot cloth folding

Figure 2: Our method for cloth folding using a bimanual flow-based policy.

In recent work, we have explored how to plan over longer horizons for deformable object manipulation tasks. In this project, our goal is to manipulate dough into a target configuration using cutting, sliding, and rolling actions (with the eventual goal of teaching a robot to making pastries or dumplings). We have developed a novel approach that reasons over both state and temporal abstractions: we first reason about the different dough pieces and their relationship to the goal configuration. We represent each dough piece in latent space and plan over high-level actions to manipulate the dough pieces. Finally, we convert the plan back into a detailed point cloud representation in order to precisely execute the plan. This transformation from point cloud to latent space and back enables us to obtain the benefits of fast planning over spatial and temporal abstractions without losing the ability to precisely manipulate object details [12, 21] (and ongoing work).

Articulated object manipulation: Building on the insights from our work on bimanual cloth folding [29], we next explored how to enable a robot to manipulate unseen articulated objects, such as microwaves, drawers, refrigerators, toilet seats, containers with a lid, and more. The traditional approach for such a task is to segment the object into parts, estimate the connectivity of the parts, estimate the articulation of each joint, and finally to use motion planning to actuate the parts. However, such an approach is brittle due to its many potential points of failure. A naive "deep reinforcement learning" approach might be to learn an end-to-end policy from pixels to robot arm torques. We find that both of these approaches perform poorly on this task.

Instead, our insight is to divide up the problem into two steps: first, to predict object affordances, and then to actuate the object based on its estimated affordances. Based on an analysis of the force needed to most efficiently actuate an object, we predict a 3D flow vector per point which estimates the amount that point would move if the corresponding part were to be opened by a small amount. Our method then selects the point with the largest estimated flow vector which has a flat surface to which a suction gripper can be attached. The proposed affordance estimation system can generalize to both unseen object instances as well as to novel object categories in the real world, significantly outperforming prior work. Results show that our system achieves state-of-the-art performance in both simulated and real-world experiments (RSS 2022; Best Paper Finalist) [3].

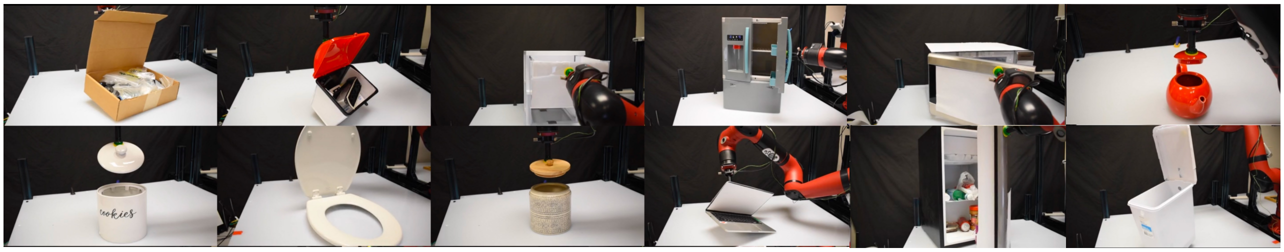


Figure 3: Our method enables robots to manipulate unseen articulated objects using a a single model by estimating 3D articulation flow [3].

1.2 Task-based Self-supervised Learning for Robotics

Robots need to learn task-specific representations to achieve their objectives; however, to enable robots to achieve many diverse tasks, robots need to be able to learn these representations on their own without requiring large-scale human annotations of training datasets. We have developed several approaches for self-supervised learning from real data that enable robots to achieve tasks without requiring human annotations; tasks that we have worked on include grasping transparent and specular objects in clutter, pouring transparent liquid, and self-supervised perception for autonomous driving.

As one example, to enable robots to grasp transparent or reflective objects (which are not perceived well by depth sensors that are used for most grasping methods), our method does not need ground-truth labels and does not require any real grasp attempts for training; instead, we use transfer learning from a pre-existing depth-based grasping model (trained on synthetic data), transferred using paired multi-modal (depth+image) observational data (without grasps) for training. Our experiments demonstrate that our approach is able to reliably grasp transparent and reflective objects in dense clutter (RA-L 2020) [28].

We have also developed a method for segmenting transparent liquid (water) without requiring any manual annotations. Instead, we train a generative model to translate images of colored liquid into synthetically generated images of transparent liquid in the same configuration. Segmentation labels of colored liquids are obtained automatically using background subtraction; these labels are then applied to the synthetically generated images of transparent liquid. Our experiments show that we are able to accurately segment transparent liquids, and we demonstrate our method in a robotic pouring task (ICRA 2022) [18].

For more general manipulation of novel objects, we have developed methods for zero-shot pose estimation that learn to compare a model to an observation of a cluttered scene without requiring any object-specific training; our method can can thus be applied to estimate the pose of previously unseen objects, given an object model, without requiring any retraining (ICRA 2021) [19], (RA-L 2022) [5].

Finally, we have a series of projects on self-supervised perception for autonomous vehicles. It is easy to collect large amounts of unlabeled data in driving contexts by placing sensors on human-driven vehicles. We have developed methods to learn from such large unlabeled datasets, such as self-supervised scene flow estimation (CVPR 2020 Oral; Selection rate 5.7%) [16], point cloud completion (BMVC 2021 Oral, Selection rate 3.3%) [17], 3D data association (IROS 2020) [26], and 3D object detection (3DV 2021) [27] - all of which can learn from unlabeled data.

A somewhat different form of learning from unlabeled data is reinforcement learning, in which robots learn policies from reward signals. We have developed novel methods for offline reinforcement learning in a latent action space (CoRL 2020 Plenary talk; Selection rate 4.1% - mentioned also above) [31] and combining model-based and model-free reinforcement learning with policy-guided planning (CoRL 2021 - Best Paper Finalist) [24].

1.3 Active Perception

To safely navigate unknown environments, robots must accurately perceive the location of obstacles. Instead of directly measuring the scene depth with a LiDAR sensor, we explore the use of a sensor developed at CMU known as "programmable light curtains." Programmable light curtains are inexpensive and high resolution but require a user to select the locations that they want to sense. We have developed a series of methods for active perception using light curtains for object detection (ECCV 2020 Spotlight; Selection rate 5.3%) [1] and for safety envelope estimation (RSS 2021) [2], as shown in Figure 4. More generally, I am developing methods for active perception to ask: how can robots take actions to better perceive the environment? Active perception is especially beneficial for environments in which there is perceptual ambiguity from just a single observation, such as crumpled clothing or mixtures of liquids and solids in a stew; a robot can manipulate objects in the environment to obtain more information about the scene.

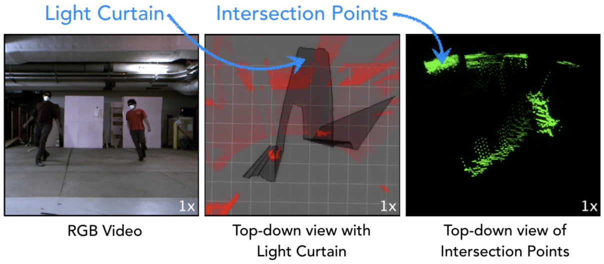

Light Curtain
Intersection Points
1x
1x
1x
RGB Video
Top-down view with Light Curtain
Top-down view of Intersection Points

Figure 4: Our novel active perception algorithm uses programmable light curtains to track moving people and estimates a "safety envelope" that represents a surface that divides free space from occupied space.

1.4 Future Research Goals

My long-term research goal is to enable robots to reliably *manipulate objects with diverse properties*, such as folding cloth (a deformable mesh), manipulating liquids, dough, or granular material (deformable objects without internal structure), peeling fruits and vegetables (reasoning about object layers), scooping stews (mixtures of solids and liquids), and manipulating articulated objects (objects with kinematic constraints). Our past research has suggested that new representations will likely be required to enable robots to reason about its interaction with objects with varying kinematic and dynamic properties. My work and ideas towards this research goal enabled me to recently receive the NSF CAREER Award.

Further, many manipulation tasks require *reasoning about object-object interactions*, such as hanging clothes on a hanger or putting a lid on a jar. Such tasks will require robots to reason about the dynamic properties of individual objects as well as how these objects interact with each other. Similarly, robots that use tools will need to reason about the interaction between the tool and the object that the tool is interacting with. I believe that new methods will be needed to enable robots to reason about such interactions.

Finally, I am excited to develop methods to enable robots to learn from humans. I believe that the object-centric manipulation approaches that I have been developing will enable robots to *learn from visual demonstrations*. I hope to develop methods by which a robot can observe many videos of humans performing a task, after which the robot should be able to quickly perform such a task after limited additional interaction with its environment, generalizing to new positions of the objects as well as generalizing to other objects with similar properties. This ability would enable robots to be able to quickly acquire new skills, and it would enable users without programming experience to be able to easily teach robots how to perform new tasks. I am excited about the future of perceptual robot learning and what it can enable robots to achieve.

References

[1] Siddharth Ancha, Yaadhav Raaj, Peiyun Hu, Srinivasa G. Narasimhan, and David Held. Active perception using light curtains for autonomous driving. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 751–766, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58558-7.

[2] Siddharth Ancha, Gaurav Pathak, Srinivasa G. Narasimhan, and David Held. Active safety envelopes using light curtains with probabilistic guarantees. In *Proceedings of Robotics: Science and Systems*, July 2021.

[3] Ben Eisner*, Harry Zhang*, and David Held. Flowbot3d: Learning 3d articulation flow to manipulate articulated objects. In *Robotics: Science and Systems (RSS)*, 2022.

[4] Alejandro Escontrela, Xue Bin Peng, Wenhao Yu, Tingnan Zhang, Atil Iscen, Ken Goldberg, and Pieter Abbeel. Adversarial motion priors make good substitutes for complex reward functions. *arXiv preprint arXiv:2203.15103*, 2022.

[5] Qiao Gu, Brian Okorn, and David Held. Ossid: Online self-supervised instance detection by (and for) pose estimation. *IEEE Robotics and Automation Letters (In press)*, 2022.

[6] Huy Ha and Shuran Song. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. In *Conference on Robot Learning*, pages 24–33. PMLR, 2022.

[7] Zixuan Huang, Xingyu Lin, and David Held. Mesh-based dynamics model with occlusion reasoning for cloth manipulation. In *Robotics: Science and Systems (RSS)*, 2022.

[8] Gwanghyeon Ji, Juhyeok Mun, Hyeongjun Kim, and Jemin Hwangbo. Concurrent training of a control policy and a state estimator for dynamic and robust legged locomotion. *IEEE Robotics and Automation Letters*, 7(2):4630–4637, 2022.

[9] Ashish Kumar, Zipeng Fu, Deepak Pathak, and Jitendra Malik. Rma: Rapid motor adaptation for legged robots. *RSS*, 2021.

[10] Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning quadrupedal locomotion over challenging terrain. *Science robotics*, 5(47):eabc5986, 2020.

[11] Xingyu Lin, Yufei Wang, Jake Olkin, and David Held. Softgym: Benchmarking deep reinforcement learning for deformable object manipulation. In *Conference on Robot Learning*, 2020.

[12] Xingyu Lin, Zhiao Huang, Yunzhu Li, David Held, Joshua B. Tenenbaum, and Chuang Gan. Diffskill: Skill abstraction from differentiable physics for deformable object manipulations with tools. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=Kef8cKdHWpP`.

[13] Xingyu Lin, Yufei Wang, Zixuan Huang, and David Held. Learning visible connectivity dynamics for cloth smoothing. In *Conference on Robot Learning*, pages 256–266. PMLR, 2022.

[14] Xiao Ma, David Hsu, and Wee Sun Lee. Learning latent graph dynamics for deformable object manipulation. *arXiv preprint arXiv:2104.12149*, 2021.

[15] Gabriel Margolis, Tao Chen, Kartik Paigwar, Xiang Fu, Donghyun Kim, Sangbae Kim, and Pulkit Agrawal. Learning to jump from pixels. *Conference on Robot Learning*, 2021.

[16] Himangi Mittal, Brian Okorn, and David Held. Just go with the flow: Self-supervised scene flow estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[17] Himangi Mittal, Brian Okorn, Arpit Jangid, and David Held. Self-supervised point cloud completion via inpainting. *British Machine Vision Conference (BMVC), 2021*, 2021.

[18] Gautham Narasimhan, Kai Zhang, Ben Eisner, Xingyu Lin, and David Held. Transparent liquid segmentation for robotic pouring. In *2022 International Conference on Robotics and Automation (ICRA) (In press)*, 2022.

[19] Brian Okorn, Qiao Gu, Martial Hebert, and David Held. Zephyr: Zero-shot pose hypothesis rating. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 14141–14148. IEEE, 2021.

[20] Xue Bin Peng, Erwin Coumans, Tingnan Zhang, Tsang-Wei Edward Lee, Jie Tan, and Sergey Levine. Learning agile robotic locomotion skills by imitating animals. In *Robotics: Science and Systems*, 07 2020. doi: 10.15607/RSS.2020.XVI.064.

[21] Carl Qi, Xingyu Lin, and David Held. Learning closed-loop dough manipulation using a differentiable reset module. In *Intelligent Robots and Systems (IROS), IEEE/RSJ International Conference on*. IEEE, 2022.

[22] Nikita Rudin, David Hoeller, Philipp Reist, and Marco Hutter. Learning to walk in minutes using massively parallel deep reinforcement learning. In *Conference on Robot Learning*, pages 91–100. PMLR, 2022.

[23] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[24] Harshit Sikchi, Wenxuan Zhou, and David Held. Learning off-policy for online planning. In *Conference on Robot Learning*, 2021.

[25] Sashank Tirumala, Thomas Weng, Daniel Seita, Oliver Kroemer, Zeynep Temel, and David Held. Learning to singulate layers of cloth based on tactile feedback. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022.

[26] Jianren Wang, Siddharth Ancha, Yi-Ting Chen, and David Held. Uncertainty-aware self-supervised 3d data association. In *IROS*, 2020.

[27] Jianren Wang, Haiming Gang, Siddharth Ancha, Yi-ting Chen, and David Held. Semi-supervised 3d object detection via temporal graph neural networks. *International Conference on 3D Vision*, 2021.

[28] Thomas Weng, Amith Pallankize, Yimin Tang, Oliver Kroemer, and David Held. Multi-modal transfer learning for grasping transparent and specular objects. *IEEE Robotics and Automation Letters*, 5(3):3791–3798, 2020. doi: 10.1109/LRA.2020.2974686.

[29] Thomas Weng, Sujay Man Bajracharya, Yufei Wang, Khush Agrawal, and David Held. Fabricflownet: Bimanual cloth manipulation with a flow-based policy. In *Conference on Robot Learning*, pages 192–202. PMLR, 2022.

[30] Zhenjia Xu, Cheng Chi, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Dextairity: Deformable manipulation can be a breeze. *arXiv preprint arXiv:2203.01197*, 2022.

[31] Wenxuan Zhou, Sujay Bajracharya, and David Held. Plas: Latent action space for offline reinforcement learning. In *Conference on Robot Learning*, 2020.