

Learning Fabric Manipulation in the Real World with Human Videos

Robert Lee¹, Jad Abou-Chakra¹, Fangyi Zhang¹, Peter Corke¹

Abstract—Fabric manipulation is a challenging area in robotics due to the vast state space and complex dynamics. Learning methods show promise as they enable learning behaviors directly from data. While most previous methods rely on simulation or large datasets, an alternative is learning fabric manipulation from human demonstrations. In this work, we collect demonstrations directly from humans performing the task for a fast data collection pipeline. Using a small number of demonstrations, we learn a pick-and-place policy deployable on a real robot without additional robot data collection. We demonstrate our approach on a fabric manipulation task involving smoothing and folding, successfully reaching folded states from crumpled initial configurations. Videos available at: <https://sites.google.com/view/foldingbyhand>

Index Terms—Deep Learning in Grasping and Manipulation, Visual Learning, Perception for Grasping and Manipulation

I. INTRODUCTION

Fabric manipulation is a task with complex dynamics and high-dimensional state, which makes it difficult for traditional robot control methods, but a good candidate for learning based methods. However, existing approaches often rely on large datasets and simulations, which can be time-consuming and suffer from the sim-to-real gap. We propose a novel method for learning multi-step fabric manipulation tasks directly from a small number of demonstrations of humans performing the task, avoiding both the sim-to-real gap and extensive data collection.

Our approach leverages an off-the-shelf hand-tracking model to recover human pick-and-place actions from videos, which are then used to train a robot policy. We introduce a sample-efficient architecture for learning pick-and-place policies from human data by predicting place heatmaps conditioned on pick location, extending the idea of pick-conditioned placing [1] to imitation learning with spatial action spaces.

We demonstrate our method on a challenging task of folding a cloth from a crumpled configuration. Unlike prior works that separate smoothing and folding tasks [2], [3], our approach learns an effective manipulation policy for the whole task using only 15 demonstrations. We show that our spatial action space, pick-conditioned policy approach, effectively learns fabric manipulation with limited demonstration data.

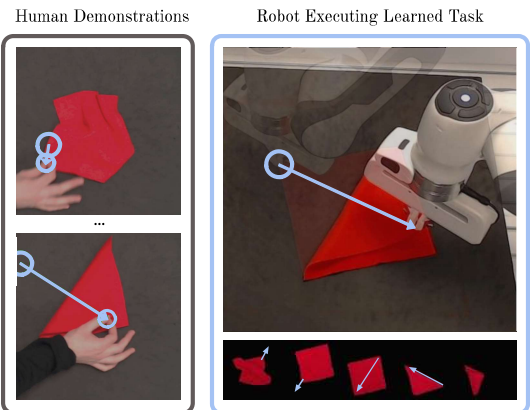


Fig. 1. Our method learns a policy from 15 demonstrations of a human performing the task. The resulting policy can then reach the folded state from new crumpled configurations at test time.

II. RELATED WORK

A. Deformable Object Manipulation

Manipulation of cloth has been a long-standing challenge for robotics [4]–[8]. Early approaches made use of geometric cues [9] and engineered features [10] for grasp detection for smoothing and folding. However, recent progress in cloth manipulation has been impressive due to advancements in learning-based approaches [11]–[14]. In particular, spatial pick-and-place action spaces allow for the use of Fully Convolutional Networks for learning action affordance heatmaps [15] as well as dynamics models [16] to achieve sample efficiency. Such architectures have also been used to learn fabric smoothing via flinging [17], bimanual stretching [18] and pick-and-place [3], as well as one-step fabric folding policies [19]. [1] learns a pick-conditioned placing value function for fabric smoothing, improving sample efficiency. Our method extends the idea to a fully-convolutional network for imitation learning with few demonstrations.

[2] and [3] learn T-shirt folding using self-supervised and human-annotated data or imitation learning combined with analytic methods. Both methods divide the task into separate smoothing and folding stages. Our approach learns a single policy from human hands for manipulating a cloth from a crumpled to a folded state.

B. Learning from Human Videos

Robotics research has advanced in learning from human-performed tasks. [20] uses pre-trained human pose estimation models for real-time robot arm control, while other work learns from human videos for tasks like cooking [21]. Human video demonstrations combined with limited robot learning have

¹Authors are with the Queensland University of Technology (QUT) Centre for Robotics, 2 George Street, Brisbane City, 4000, Queensland, Australia. (email: r21.lee@hdr.qut.edu.au; peter.corke@qut.edu.au)

been explored for mobile manipulation [22]. Our work focuses on learning fabric manipulation pick-and-place policies directly from human videos without additional robot training or data collection.

III. APPROACH

A. Problem Definition

Our study focuses on enabling a single robot arm to transform a crumpled fabric into a folded state. The robot utilizes an overhead camera to observe the fabric on a flat surface and execute pick-and-place actions based on image locations. This spatial action space [23], is common in deformable object manipulation [3], [15], [16], [19]. We aim to identify the best pick-and-place action from the image observation. Unlike previous methods that only smooth [1], [24], [25]; assume a pre-smoothed state for folding [15], [19]; or use separate smoothing-folding policies with switching criteria [2], [3]; we seek a unified policy for the entire smoothing-folding process.

B. Human Demonstrations

Fabric manipulation is a complex, long-horizon task, and learning behaviors from self-supervision without human guidance is difficult. Although engineered biases like corner grasping [15], [19] help, learning long-horizon tasks remains challenging. Human demonstrations are a natural alternative. Instead of using robot-controlling user interfaces [3], [26], we aim to learn directly from videos of humans using their hands for fabric manipulation.

We use Mediapipe Hands [27], a real-time hand tracking system, to estimate digit positions. It is trained on 30,000 real-world hand images and localizes 21 hand coordinates, including fingertip locations. We track grasps by monitoring the distance between the thumb and index fingertips. When below a threshold, we record the average fingertip location as the pick location. The place location is the average of the thumb and index fingertip locations when the grasp is released, and the distance surpasses the grasp threshold.

We record observations before and after actions are performed, by identifying periods of minimal pixel change. The human removes their hand, an image is captured, and then they perform an action. This allows for natural data collection for fabric manipulation tasks.

C. Learning to Imitate Pick-and-Place Actions

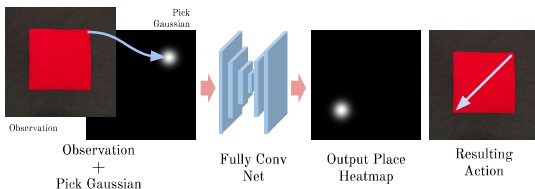


Fig. 2. A visualization of our pick-conditioned place prediction approach. We concatenate the observation image with a candidate pick, taken from the cloth mask, represented as a 2D Gaussian image. The network is trained against the place heatmap label.

To efficiently learn pick-and-place behaviors from human demonstration data, we need a model that can be trained with minimal data, reducing data collection costs. Prior work [1]

has shown that learning only placing actions, conditioned on pick locations, improves sample efficiency for fabric smoothing policies. We extend this approach for supervised imitation learning by predicting place heatmaps conditioned on pick locations. Our fully-convolutional network is conditioned on pick locations, using an image channel with a 2D Gaussian centered on the pick location, similar to [19]. We train the model to predict place locations via heatmap image labels created with 2D Gaussians centered on the place location. We can then estimate the place location by taking the argmax pixel location over the place heatmap, conditioned on a candidate pick location. Since this method would only work for known pick locations, we create artificial negative samples by randomly selecting pick locations on the cloth and training against a heatmap of zeros. This allows the model to output higher placing probabilities for correct pick locations and lower probabilities for incorrect ones. We find the pick action maximizing place probability by taking the argmax over all possible pick locations in the cloth mask. During training, we evaluate our model by sub-sampling the cloth mask by a factor of 2 to reduce training time and memory. At test time, we evaluate every possible pick location in the mask for full-resolution pick selection. Our approach adapts the idea from [1] for supervised, imitation learning settings with spatial action spaces.

IV. EXPERIMENTS

We demonstrate our approach on a real-world fabric manipulation task, folding a cloth in half twice into a small triangle, beginning from a crumpled initial configuration. Folding a square fabric into this shape, assuming a smooth initial state, has been shown in prior work [15], [19], [28]. To the best of our knowledge, we are the first to train a single policy that can reach a folded state from a crumpled state.




A. Experimental Setup

Our experimental setup consists of a Franka Emika Panda robot workstation, with a RealSense L515 camera. We make use of depth for observation, and infrared for cloth masking, while RGB is used for visualization only. The workspace is 40×40 cm square of PVA foam. The fabric is a 22×22 cm square of red polar fleece.

Using our hand-tracking approach, we collect a dataset of 15 demonstration episodes, with 10 for training and 5 for validation. Each episode consists of a human demonstrating single-handed manipulation of the full crumpled-to-folded task. We use depth images for generalization across various cloth appearances and mask the images to isolate the cloth [2], [15], [17], [19], [28]. Infrared images help isolate cloths from the background easily. To augment the dataset by a factor of 20, we apply rotation, flips, scaling, and Gaussian noise to depth images, along with depth scaling for cloth thickness robustness.

We execute the trained policy on the real robot, transforming pixel coordinates to the robot’s workspace using a linear transform. The robot adjusts its wrist angle for consistent actions, moves down until it senses contact, grasps, and executes the action. A 3D-printed sloped gripper enables effective cloth pinching. We use a grasping width heuristic to differentiate

TABLE I
EVALUATION RESULTS FROM CRUMPLED INITIAL CONFIGURATION.
SUCCESS IS OUT OF 5 RUNS.

| | Method | Success | IoU | ISC |
|---|---------------------------|---------|--------------|--------------|
|  | Human | 5 | 0.883 | 0.98 |
| | <i>PickToPlace (Ours)</i> | 5 | 0.843 | 0.965 |
| | Pick+Place | 2 | 0.752 | 0.846 |
|  | Human | 5 | 0.876 | 0.98 |
| | <i>PickToPlace (Ours)</i> | 4 | 0.801 | 0.958 |
| | Pick+Place | 0 | 0.776 | 0.851 |
|  | Human | 5 | 0.86 | 0.982 |
| | <i>PickToPlace (Ours)</i> | 4 | 0.811 | 0.978 |
| | Pick+Place | 0 | 0.684 | 0.909 |
| Mean | Human | 5 | 0.873 | 0.981 |
| | <i>PickToPlace (Ours)</i> | 4.334 | 0.818 | 0.967 |
| | Pick+Place | 0.667 | 0.737 | 0.869 |

between top-layer grasping for smoothing and grasping all layers for folding, adjusting the finger width to be wider at the edge of the mask and narrower otherwise.

B. Task

We evaluate the policy on a set of initial crumpled configurations, following [24], for repeatability. A human resets the cloth, and we test three configurations with five trials each. The policy attempts to smooth and fold the cloth autonomously within 15 timesteps.

Deformable object manipulation metrics are challenging to define due to difficulties in perceiving the object state. As such, we report several metrics:

Success: We qualitatively assess folding success, considering a successful fold when the triangle is reached by folding the smoothed square twice, allowing one minor defect.

Intersection Over Union (IOU): We report the IOU of the fabric mask, aligning it with a final folded state mask. We take the highest IOU across all timesteps for each episode.

Intermediate Smoothing Coverage (ISC): We report the fabric coverage ratio compared to a smoothed cloth mask, taking the highest coverage score across all timesteps, indicating how well the fabric was smoothed during the episode.

C. Baselines

To evaluate our method’s performance and validate the effectiveness of our pick-conditioned place model, we compare it against human performance and a baseline architecture.

Human: We show the performance of a human using a single hand to solve the task. This represents an approximate upper bound for our method, which is trained to mimic human demonstrations.

Pick+Place: We demonstrate a simple FCN network that outputs two heatmaps directly for pick-and-place, based on the policy from [3]. The network architecture is otherwise the same as ours, but the place is not conditioned on the pick. If the pick location is predicted off the fabric, we find the closest point on the fabric mask to the predicted pick location.

D. Results

In the robot evaluation experiments, we assess our policy’s ability to consistently reach a folded configuration from initially crumpled states using a small number of human

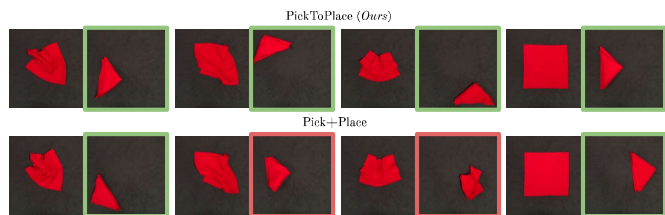


Fig. 3. Comparison of runs by our approach (top) and Pick+Place baseline (bottom). Initial configurations and best states based on IoU are shown. Successes in green, failures in red.

demonstrations and compare our pick-conditioned place model to a baseline model. The results for the full task are shown in Table I.

Our proposed method was successful in 13 of the 15 runs, achieving an average score of 4.334/5 or 86.7%. The policy consistently performs well at smoothing, with high intermediate smoothing coverage (ISC) and IOU scores. However, in one failure case, the policy chose to fold before the fabric was adequately smoothed, resulting in a crumpled final folded state.

The baseline network architecture succeeded in smoothing and folding the cloth twice out of 15 runs, with an average score of 0.667/5 or 13.34% across tasks. We hypothesize its low success rate is due to the difficulty of learning both pick-and-place outputs with limited data and the lack of conditioning between the pick-and-place locations, leading to poor action choices.

E. Generalization



Fig. 4. Our method can generalize to fabrics of a variety of appearances and textures. We show the initial, smooth state before folding, and final configurations from successful folding episodes.

Our method, trained on depth images, demonstrates generalization capabilities by handling a variety of cloths with different visual properties, material properties, and shapes, as shown in Figure 4. The model is robust to differences in thickness, texture, and cloth shape, with only the grasping width of the robot fingers needing adjustment for various materials.

V. CONCLUSION

We present a method for learning fabric manipulation from a small number of human demonstrations collected directly from human hands, achieving over 85% success in folding tasks. Our approach outperforms the baseline pick-and-place architecture and generalizes to unseen fabrics. This work is a step towards leveraging large, freely available online video data for robotic manipulation. Future work includes expanding to more tasks, incorporating bimanual manipulation, and reducing reliance on unobstructed cloth visibility.

REFERENCES

- [1] Y. Wu, W. Yan, T. Kurutach, L. Pinto, and P. Abbeel, "Learning to manipulate deformable objects without demonstrations," in *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020.
- [2] Y. Avigal, L. Berscheid, T. Asfour, T. Kröger, and K. Goldberg, "Speedfolding: Learning efficient bimanual folding of garments," *arXiv preprint arXiv:2208.10552*, 2022.
- [3] R. Hoque, K. Shivakumar, S. Aeron, G. Deza, A. Ganapathi, A. Wong, J. Lee, A. Zeng, V. Vanhoucke, and K. Goldberg, "Learning to fold real garments with one arm: A case study in cloud-based robotics research," *arXiv preprint arXiv:2204.10297*, 2022.
- [4] Y. Li, X. Hu, D. Xu, Y. Yue, E. Grinspun, and P. K. Allen, "Multi-sensor surface analysis for robotic ironing," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 5670–5676.
- [5] —, "Multi-sensor surface analysis for robotic ironing," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 5670–5676.
- [6] F. Osawa, H. Seki, and Y. Kamiya, "Clothes folding task by tool-using robot," *Journal of Robotics and Mechatronics*, vol. 18, no. 5, pp. 618–625, 2006.
- [7] M. Cusumano-Towner, A. Singh, S. Miller, J. F. O'Brien, and P. Abbeel, "Bringing clothing into desired configurations with limited perception," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 3893–3900.
- [8] 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.
- [9] 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 *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 2308–2315.
- [10] B. Willimon, S. Birchfield, and I. Walker, "Model for unfolding laundry using interactive perception," in *IEEE/RSJ Int'l Conf. on Intelligent Robots and Systems*, 2011, pp. 4871–4876.
- [11] R. Jangir, G. Alenya, and C. Torras, "Dynamic cloth manipulation with deep reinforcement learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 4630–4636.
- [12] J. Matas, S. James, and A. J. Davison, "Sim-to-real reinforcement learning for deformable object manipulation," in *Conf. on Robot Learning*, 2018, pp. 734–743.
- [13] Y. Tsurumine, Y. Cui, E. Uchibe, and T. Matsubara, "Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation," *Robotics and Autonomous Systems*, vol. 112, pp. 72–83, 2019.
- [14] J. Hietala, D. Blanco-Mulero, G. Alcan, and V. Kyrki, "Learning visual feedback control for dynamic cloth folding," 2022.
- [15] R. Lee, D. Ward, A. Cosgun, V. Dasagi, P. Corke, and J. Leitner, "Learning arbitrary-goal fabric folding with one hour of real robot experience," *arXiv preprint arXiv:2010.03209*, 2020.
- [16] R. Lee, M. Hamaya, T. Murooka, Y. Ijiri, and P. Corke, "Sample-efficient learning of deformable linear object manipulation in the real world through self-supervision," *IEEE Robotics and Automation Letters*, vol. 7, no. 1, pp. 573–580, 2022.
- [17] H. Ha and S. Song, "Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 24–33. [Online]. Available: <https://proceedings.mlr.press/v164/ha22a.html>
- [18] Z. Xu, C. Chi, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Dextairity: Deformable manipulation can be a breeze," in *Proceedings of Robotics: Science and Systems (RSS)*, 2022.
- [19] T. Weng, S. Bajracharya, Y. Wang, K. Agrawal, and D. Held, "Fabricflownet: Bimanual cloth manipulation with a flow-based policy," in *Conference on Robot Learning*, 2021.
- [20] A. Sivakumar, K. Shaw, and D. Pathak, "Robotic telekinesis: learning a robotic hand imitator by watching humans on youtube," *arXiv preprint arXiv:2202.10448*, 2022.
- [21] Y. Yang, Y. Li, C. Fermuller, and Y. Aloimonos, "Robot learning manipulation action plans by watching unconstrained videos from the world wide web," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 29, no. 1, 2015.
- [22] S. Bahl, A. Gupta, and D. Pathak, "Human-to-robot imitation in the wild," 2022.
- [23] D. Wang, R. Walters, X. Zhu, and R. Platt, "Equivariant q-learning in spatial action spaces," in *Conference on Robot Learning*. PMLR, 2022, pp. 1713–1723.
- [24] S. Sharma, E. Novoseller, V. Viswanath, Z. Javed, R. Parikh, R. Hoque, A. Balakrishna, D. S. Brown, and K. Goldberg, "Learning switching criteria for sim2real transfer of robotic fabric manipulation policies," *arXiv preprint arXiv:2207.00911*, 2022.
- [25] 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.
- [26] B. Waymouth, A. Cosgun, R. Newbury, T. Tran, W. P. Chan, T. Drummond, and E. Croft, "Demonstrating cloth folding to robots: Design and evaluation of a 2d and a 3d user interface," in *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. IEEE, 2021, pp. 155–160.
- [27] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.-L. Chang, and M. Grundmann, "Mediapipe hands: On-device real-time hand tracking," *arXiv preprint arXiv:2006.10214*, 2020.
- [28] R. Hoque, D. Seita, A. Balakrishna, A. Ganapathi, A. K. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg, "Visuospatial foresight for multi-step, multi-task fabric manipulation," *arXiv preprint arXiv:2003.09044*, 2020.