Learning from Play for Deformable Object Manipulation

Arpit Bahety
Department of Computer Science
Columbia University
ab5232@columbia.edu

Abstract

Deformable object manipulation is a challenging task in robotics due to the complex dynamics of deformable objects as compared to rigid objects and the unlimited degrees of freedom that deformable objects have. In this work, we aim to investigate if learning good representations through playful interactions help in downstream deformable object manipulation tasks such as folding cloth or placing cloth inside a deformable bag (manipulating two deformable objects simultaneously). We use playful interactions with deformable objects to learn visual representations. Then we learn task-specific linear heads on top of the representation. Initial experiments show that learning from play does help downstream deformable object manipulation tasks. To the best of our knowledge, this is the first work on assessing and trying to improve deformable object manipulating using playful interactions. [1]

1 Introduction

Robotic manipulation of rigid objects has received significant interest over the last few decades. However, the objects we interact within our daily lives are not always rigid. From folding clothes to packing a shopping bag, we constantly need to manipulate objects that deform. Deformable object manipulation presents two key challenges for robots. First, unlike rigid objects, there is no direct representation of the state. Second, the dynamics of deformable objects are complex and nonlinear.

One class of techniques that circumvents the challenges in state estimation and dynamics modeling is image-based model-free learning (10; 24). For instance, Matas et al. (17), Seita et al. (26), Wu et al. (32) use model-free methods in simulation for several difficult cloth manipulation tasks. However, model-free learning is notoriously inefficient (6), and often needs millions of samples to learn from. To this end, we aim to investigate if using meaningful representations along with model-free learning helps improve the sample efficiency. If the hypothesis is true, such representations could also help in multi-task learning for deformable objects (a potential future research direction). The model-free learning approach that we use is spatial-action maps (31).

The method that we use for learning visual representations are playful interactions. Research in human development show how children often play with objects to learn various things about out world (21; 4; 28; 30). The advantage of using play for learning visual representations is that play is task-agnostic and it is cheap to obtain. So, how does one collect and learn from playful interactions for deformable object manipulation?

In this work, we present a framework to collect playful interaction with cloths in simulation which uses simple pick-and-place interactions with cloths. Equipped with this data, we then use a self-supervised learning approach, specifically a time-contrastive network (TCN) (27), to learn a visual

[1] This project is related to my ongoing research work at the Columbia Artificial Intelligence and Robotics lab (CAIR Lab)

encoder that can extract visual representations. To demonstrate the usefulness of representations learned through play, we use the representations along with spatial-action maps on downstream tasks.

In summary, the main contributions of this work are as follows:

- We perform playful interactions with deformable objects in order to learn visual representations and hypothesize that learning from play will help deformable object manipulation.
- We investigate the usefulness of the visual representations on two downstream tasks - 1. cloth folding (Figure 1 b.) 2. bagging cloths (Figure 1 c.).

To the best of our knowledge, this is the first work on assessing and trying to improve deformable object manipulating using playful interactions

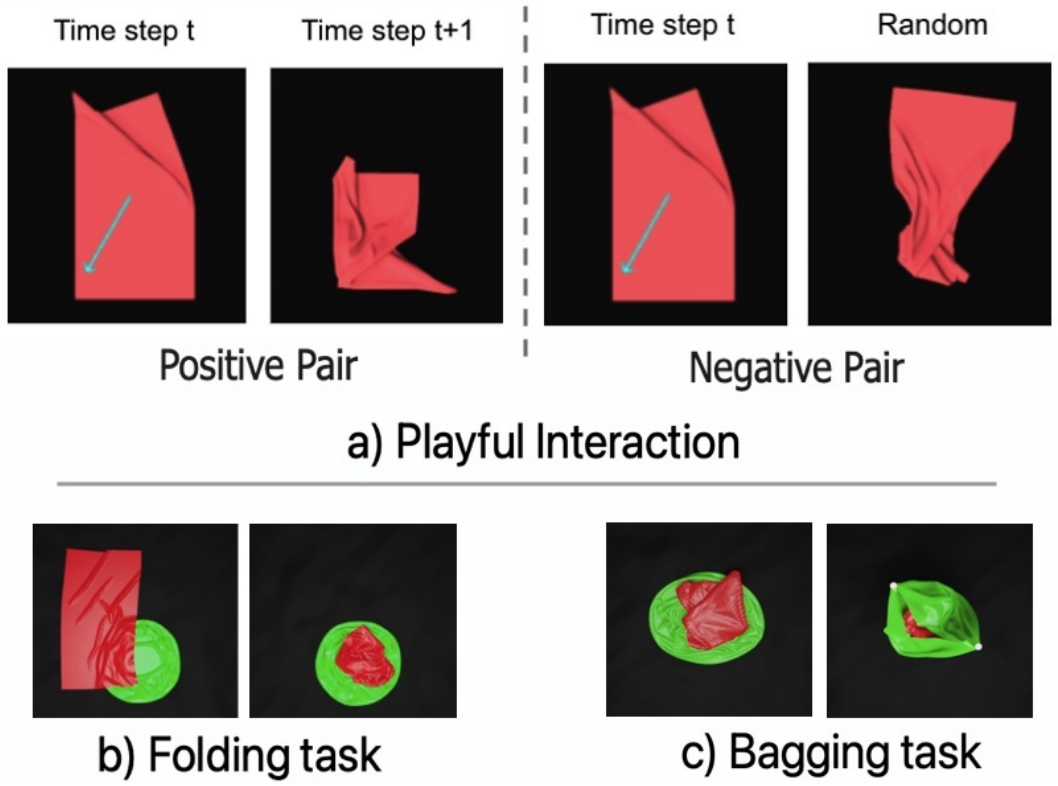

Figure 1: In this work, I aim to learn useful representations from playful interactions with deformable objects and study the usefulness of those representations in downstream tasks. **Part a.** shows how the playful interaction looks like and what's a positive and negative pair in the contrastive learning approach used. (The blue arrow is only for display purpose showing what the action was). **Part b. and c.** show the two downstream tasks that we evaluate the representations on. The bag is in green, cloth in red, robot grippers are the two white circles on the rightmost image in part c.

2 Related Work

2.1 Deformable Object Manipulation

Motion planning has been a popular approach to tackle deformable object manipulation tasks (19; 23). One of the challenges of planning with deformable objects is the large degrees of freedom and hence large configuration space involved when planning. To alleviate this, recent works now use model-free visual learning to perform various deformable object manipulation tasks (25; 32; 9).

2.2 Learning from Play

Learning from play in robotics has been relatively understudied. Here I describe two works that explore play for robotic manipulation. Play-LMP (13) has shown that supervision from teleoperated play data can effectively scale up multi-task learning. In their work, a single goal-conditioned policy is able to perform a variety of user-specified tasks. This demonstrates that playful interactions can learn latent plans capable of task discovery, composition, as well as emergent retrying. Another work (34) focuses on learning visual representations from imitation. They aim to decrease the amount of task-specific, labeled data needed to learn generalizable policies for manipulation tasks. However, none of these works tackle deformable object manipulation which brings its own set of challenges as described in Section 1.

2.3 Self-supervised Representation Learning

Representation learning has long been used in Computer Vision, but interest in this learning technique has recently grown within robotics due to the availability of unlabeled data and its effectiveness in learning tasks. The goal of representation learning is to extract features to improve performance in downstream tasks. The key idea is to exploit information from data without explicit labeling. Unlabeled data is generally first trained on one or more pretext tasks to learn a representation. These tasks can include predicting image rotations and distortions, patches, frame sequence prediction, or instance invariances (7; 5; 18). The idea behind pretraining on pretext tasks is that the learned representations have useful structural meanings and are relevant to downstream tasks. A number of works (2; 8; 1; 35) have demonstrated state-of-the-art performance with unsupervised representation learning. We use Time Contrastive Networks (27) style contrastive network to take the benefit of temporal information in robotic manipulation tasks.

2.4 Representation Learning in Robotics

Recently, interest in self-supervised or semi-supervised representation learning technique has grown within robotics (16) due to the availability of unlabeled data and its effectiveness in visual imitation tasks (15; 3; 20; 34; 36). Self-supervised representation learning has shown impressive results in computer vision. Most of these works cater to rigid object manipulation tasks. Very little work is done in studying the usefulness of visual representations for downstream deformable object manipulation tasks. Yan et al. learn representations for deformable objects but for model-based methods (33). Whereas, in this work, I aim to develop and study visual representations for a model-free method (spatial action maps (31)) for deformable object manipulation tasks. Furthermore, to the best of our knowledge, this is the first work to explore learning visual representations through playful interactions for deformable object manipulation.

3 Approach

3.1 Playful Interaction

We define playful interaction as random pick-and-place of different rectangular cloths in the PyFlex simulation (12; 14)). The cloths differ mainly differ in size and the stiffness. The action of pick-and-place consists of the following prrimitive actions -

1. move the gripper to pre-pickpoint (0.2 meters above the pickpoint)
2. move to pickpoint and grasp the cloth.
3. move back to pre-pickpoint
4. move to pre-placepoint
5. move to placepoint and release the grasp

This type of data is extremely fast and cheap to obtain. Furthermore, it is also task-agnostic meaning that such representations can be used for cloth unfolding (9; 32), goal-conditioned cloth folding (11) or even bagging cloths. Figure 1 a. shows an example of the playful interaction data. The play data would be similar to the downstream tasks in a way, but would involve many more scenarios and

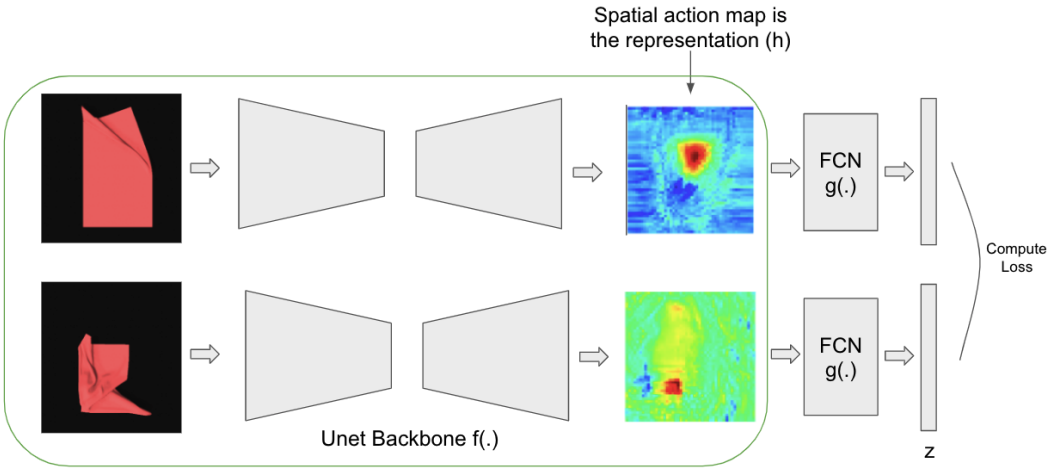

Figure 2: Visual-representation learning architecture

configurations. The data consists of tuples of RGB images at time step t and $t+1$. We collect 30k such tuples.

3.2 Visual-representation Learning from Play

Several prior works (29; 2; 8) have demonstrated success in pretraining models for downstream visual classification tasks. In this work, we aim to show that pretraining models with playful interaction data is effective for downstream robotics tasks. We choose to use time contrastive networks (29) to leverage the temporal association available in videos. Instead of augmenting a copy of the same frame, we augment a frame one timestep away in the same trajectory. Unlike (29), however, we do not require paired viewpoints of the same observation. We learn a representation purely from comparing observations from a single viewpoint at different timesteps.

We train the backbone f in a contrastive learning fashion as shown in Figure 2. We use a UNet backbone (22) as the output representation (spatial-action map) needs to be of the same dimension as the input image. f(.) takes in a single image $I \in R^{3x64x64}$ and output a spatial action map h. The $f(.)$ in the contrastive setting in our case takes in either a positive sample (I_t, I_{t+1}) or a negative sample (I_t, I_{random}). I_t is an augmented version of the frame at timestep t, I_{t+1} is an augmented version of the frame at timestep t + 1 and I_{random} is an augmented version of the frame from a random episode at a random timestep. We then feed h into a MLP projection head $g()$ and return the latent representation $z \in R^{128}$ for each image. Then, we compute a simple L2 loss between these latent representations. The projection head $g()$ is discarded after the self-supervised pretraining phase.

3.3 Downstream Tasks Learning

After training on playful interaction data to learn a meaningful representation, we use this representation for downstream manipulation tasks. The input image, $I \in R^{3x64x64}$ is passed through the backbone to obtain the representation h. h is then passed through a Fully convolutional layer to obtain a final value map as the output. this process is shown in Figure 4. The pixel with the highest value is selected which determines the pickpoint and the placepoint in case of folding task and the lift point in case of the bagging task (further explained in Section 4.1. Note that the lifting policy is a different network than the folding policy.

Our objective for the two downstream tasks are:

1. Cloth folding: minimize the surface area of the cloth
2. Bagging: minimize the area of cloth outside the boundary of the bag opening and then find two liftpoints such that lifting from those points results in the cloth going inside the bag.

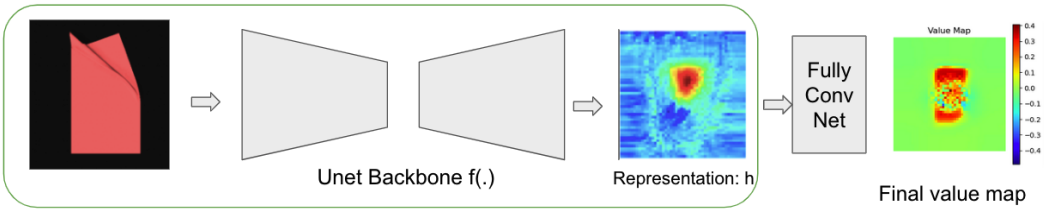

Figure 3: Downstream task learning architecture. The action is chosen from the final value map as explained in Section 4.1

4 Experiments

4.1 Downstream Task Setup

1. Cloth folding: The scene consists of a piece of rectangular cloth (of different sizes and stiffness) on a floor. To calculate the pick and place points, we employ the strategy used in (31) where they exploit rotation invariance properties of this task. We take 12 rotations (0, 30, 60 .. 360) and 8 scales (1, 1.25, 1.5 ...). We create 96 images (transformed images) of the original RGB image with each image being rotated or scaled according to the previous values. These 96 images are stacked and input to the network to output 96 value maps, one for each (rotation, scale). This process is shown in Figure . The pickpoint is the pixel with the highest value out of all 96x64x64 pixels. The place point is fixed to be 10 pixels vertically downwards for each of the transformed image. However, the final and the actual placepoint is obtained after undoing the transformation for the chosen image (thus the placepoint would not always be 10 pixels vertically downwards anymore). An episode ends if 10 steps (pick-and-place) are reached or if the policy chooses a pickpoint on the background (which means that the policy things that it cannot do better and wants to stop). The supervision is computed in a self-supervised manner: $postaction\_coverage - preaction\_coverage$ where post and preaction coverages are the area of the cloth pre and post action. The metric used to evaluate the folding is $mean\_final\_coverage$ which is the area of the cloth at the end of the episode.
2. Bagging: The scene consists of a piece of rectangular cloth (of different sizes and stiffness) on a floor and a bag with its bag opening pulled apart. There are two policies (networks) in play here. First the folding policy fold the cloth with the aim of minimizing the area of the cloth outside the opening of the bag. Next the lifting network decides when the folding network should stop and where should it lift from. The folding network decides the pick and place locations as previously described. The way the lifting network decides the two liftpoints is slightly different. The input and output of the lifting network is the same as the folgin network. However, instead of having the pickpoint as the chosen pixel and the placepoint as the the fixed 10 pixels vertically downwards, the first liftpoints is 5 pixels vertically above the chosen pixel and 2nd lift point is 5 pixels below the chosen pixel. Undoing the transformation as described previously, gives us different liftpoints.
The supervision is binary here and is computed in a self-supervised manner: 1 if the cloth is inside the 3-D convex hull of the bag and 0 otherwise. The metric used to evaluate the lifting is: number of times the cloth goes inside the bag ($\#cloth\_inside\_bag$).

4.2 Results

We evaluate our results on the following models and report the results in Table 4.2 and 4.2:

1. Folding:
   - training from scratch ($fold_{scratch}$)
   - Pretraining using playful interaction data + Finetuning on folding task ($fold_{play}$)
2. Lifting:
   - training both networks from scratch ($fold_{scratch} + lift_{scratch}$)

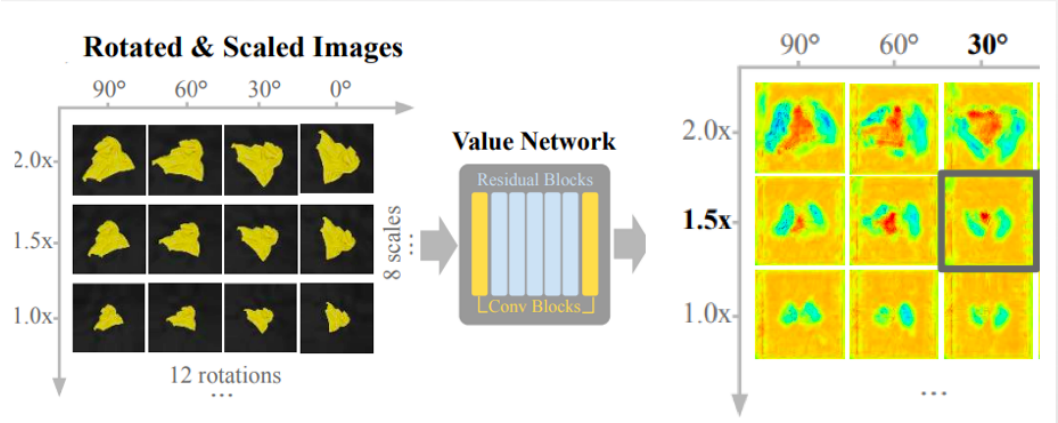

Figure 4: Showing the process of transforming input RGB images using 96 different tuples of (rotations, scales) to obtain the 96 value maps. This scaling and rotating the input image coupled with the explanation in Section 4.1 provides us with the desired action space for the pick-and-place and lifting task

- Pretraining using playful interaction data + Finetuning on folding task and lfting network from scratch($fold_{play} + lift_{scratch}$)
- Pretraining using playful interaction data + Finetuning on folding task ($fold_{scratch} + lift_{play}$)
- Pretraining using playful interaction data + Finetuning on folding task ($fold_{play} + lift_{play}$)

Table 1: Results for downstream task of folding. A $mean_final_coverage$ of 10% means that 10 percent of the falttened cloth area is visible. Lower the better.

Model	$mean\_final\_coverage$
$fold_{scratch}$	55%
$fold_{play}$	**44%**

Table 2: Results for downstream task of Lifting. The metric is out of 100 test tasks how many times does the cloth go inside the bag after folding + lifting. Higher the value, the better.

Model	$\#cloth\_inside\_bag$
$fold_{scratch} + lift_{scratch}$	67
$fold_{play} + lift_{scratch}$	80
$fold_{scratch} + lift_{play}$	69
$fold_{play} + lift_{play}$	**83**

We also report some qualitative results (videos) for $foldscratch$ and $fold_{play}$ for the folding task and $fold_{scratch} + lift_{scratch}$ + $fold_{play} + lift_{play}$ for the lifting task in the supplementary materials.

4.3 Does Training on Playful Interactions Lead to Good Representations?

To test whether self-supervised pretraining with playful interactions can learn a meaningful representation, we first train a model using our collected playful interaction data via TCN. Then, we load the learned backbone and add a fully convolutional layer and fine-tune on the downstream task. If our playful interactions can learn effective visual representations, we expect that this policy will outperform one where the downstream task is directly trained from scratch. As shown in Table 4.2, the performance of $fold_{play}$ is much better than $fold_{scratch}$. Furthermore, Table 4.2 shows that the performance of $fold_{play} + lift_{play}$ is superior to that of $fold_{scratch} + lift_{scratch}$. This shows that playful interactions indeed lead to learning good visual representations. Note however that $lift_{play}$ only slightly helps. Our hypothesis is that the playful interaction is only performed with cloth and not

bag. Thus, it would be interesting to see if adding playful interactions with bag helps the bagging task further.

4.4 Limitations

Overall, learning from play help learn useful representations for deformable object manipulation however, there are failure cases as well. Here we briefly address a few of those.

- The folding network sometimes does bad action after reaching a low cloth coverage. In other words, learning to stop folding is a non-trivial task.
- The lifting network performs relatively poorly on different bags than that it was trained on. One hypothesis is that since there was no playful interaction with bags, it generalizes poorly to bags. it would be interesting to see if adding playful interactions with bag helps.

5 Conclusion

In this work, we have presented an approach for learning downstream deformable object manipulation tasks via self-supervised pretraining on easy-to-obtain playful interaction data. We show that polices that use such a pre-trained backbone and fine-tunes on task-specific data (cloth folding and bagging) performs better than training for the task from scratch. However, it was noticed that playful interaction with cloths did not help the lifting network (which involves bags) a lot. Thus, this leads to potential future work to explore playful interactions with a diverse set of deformable objects and explore its generalization capabilities.

References

[1] CARON, M., MISRA, I., MAIRAL, J., GOYAL, P., BOJANOWSKI, P., AND JOULIN, A. Unsupervised learning of visual features by contrasting cluster assignments. *CoRR abs/2006.09882* (2020).

[2] CHEN, T., KORNBLITH, S., NOROUZI, M., AND HINTON, G. E. A simple framework for contrastive learning of visual representations. *CoRR abs/2002.05709* (2020).

[3] CHEN, X., TOYER, S., WILD, C., EMMONS, S., FISCHER, I., LEE, K.-H., ALEX, N., WANG, S. H., LUO, P., RUSSELL, S., ABBEEL, P., AND SHAH, R. An empirical investigation of representation learning for imitation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)* (2021).

[4] COOK C, GOODMAN ND, S. L. Where science starts: spontaneous experiments in preschoolers' exploratory play. In *Cognition. 2011* (2011).

[5] DOERSCH, C., GUPTA, A., AND EFROS, A. A. Unsupervised visual representation learning by context prediction. *CoRR abs/1505.05192* (2015).

[6] DUAN, Y., CHEN, X., HOUTHOOFT, R., SCHULMAN, J., AND ABBEEL, P. Benchmarking deep reinforcement learning for continuous control. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (2016), ICML'16, JMLR.org, p. 1329–1338.

[7] GIDARIS, S., SINGH, P., AND KOMODAKIS, N. Unsupervised representation learning by predicting image rotations. *CoRR abs/1803.07728* (2018).

[8] GRILL, J., STRUB, F., ALTCHÉ, F., TALLEC, C., RICHEMOND, P. H., BUCHATSKAYA, E., DOERSCH, C., PIRES, B. Á., GUO, Z. D., AZAR, M. G., PIOT, B., KAVUKCUOGLU, K., MUNOS, R., AND VALKO, M. Bootstrap your own latent: A new approach to self-supervised learning. *CoRR abs/2006.07733* (2020).

[9] HA, H., AND SONG, S. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. *CoRR abs/2105.03655* (2021).

[10] HAARNOJA, T., ZHOU, A., HARTIKAINEN, K., TUCKER, G., HA, S., TAN, J., KUMAR, V., ZHU, H., GUPTA, A., ABBEEL, P., AND LEVINE, S. Soft actor-critic algorithms and applications. *ArXiv abs/1812.05905* (2018).

[11] LEE, R., WARD, D., COSGUN, A., DASAGI, V., CORKE, P., AND LEITNER, J. Learning arbitrary-goal fabric folding with one hour of real robot experience. *CoRR abs/2010.03209* (2020).

[12] LI, Y., WU, J., TEDRAKE, R., TENENBAUM, J. B., AND TORRALBA, A. Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids. *arXiv preprint arXiv:1810.01566* (2018).

[13] LYNCH, C., KHANSARI, M., XIAO, T., KUMAR, V., TOMPSON, J., LEVINE, S., AND SERMANET, P. Learning latent plans from play. *CoRR abs/1903.01973* (2019).

[14] MACKLIN, M., MÜLLER, M., CHENTANEZ, N., AND KIM, T.-Y. Unified particle physics for real-time applications. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 1–12.

[15] MANDI, Z., LIU, F., LEE, K., AND ABBEEL, P. Towards more generalizable one-shot visual imitation learning, 10 2021.

[16] MANUELLI, L., LI, Y., FLORENCE, P. R., AND TEDRAKE, R. Keypoints into the future: Self-supervised correspondence in model-based reinforcement learning. *CoRR abs/2009.05085* (2020).

[17] MATAS, J., JAMES, S., AND DAVISON, A. J. Sim-to-real reinforcement learning for deformable object manipulation. *CoRR abs/1806.07851* (2018).

[18] MISRA, I., ZITNICK, C. L., AND HEBERT, M. Unsupervised learning using sequential verification for action recognition. *CoRR abs/1603.08561* (2016).

[19] MOLL, M., AND KAVRAKI, L. Path planning for deformable linear objects. *Robotics, IEEE Transactions on 22* (09 2006), 625 – 636.

[20] PARI, J., SHAFIULLAH, N. M. M., ARUNACHALAM, S. P., AND PINTO, L. The surprising effectiveness of representation learning for visual imitation. *ArXiv abs/2112.01511* (2021).

[21] PIAGET, J. Play dreams and imitation in childhood.

[22] RONNEBERGER, O., FISCHER, P., AND BROX, T. U-net: Convolutional networks for biomedical image segmentation. *CoRR abs/1505.04597* (2015).

[23] SAHA, M., AND ISTO, P. Manipulation planning for deformable linear objects. *Robotics, IEEE Transactions on 23* (01 2008), 1141 – 1150.

[24] SCHULMAN, J., LEVINE, S., MORITZ, P., JORDAN, M., AND ABBEEL, P. Trust region policy optimization. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (2015), ICML'15, JMLR.org, p. 1889–1897.

[25] SEITA, D., FLORENCE, P., TOMPSON, J., COUMANS, E., SINDHWANI, V., GOLDBERG, K., AND ZENG, A. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. *CoRR abs/2012.03385* (2020).

[26] SEITA, D., GANAPATHI, A., HOQUE, R., HWANG, M., CEN, E., TANWANI, A. K., BALAKRISHNA, A., THANANJEYAN, B., ICHNOWSKI, J., JAMALI, N., YAMANE, K., IBA, S., CANNY, J. F., AND GOLDBERG, K. Deep imitation learning of sequential fabric smoothing policies. *CoRR abs/1910.04854* (2019).

[27] SERMANET, P., LYNCH, C., CHEBOTAR, Y., HSU, J., JANG, E., SCHAAL, S., AND LEVINE, S. Time-contrastive networks: Self-supervised learning from video. *Proceedings of International Conference in Robotics and Automation (ICRA)*.

[28] SIM ZL, X. F. Learning higher-order generalizations through free play: Evidence from 2- and 3-year-old children. In *Dev Psychol. 2017* (2017).

[29] VAN DEN OORD, A., LI, Y., AND VINYALS, O. Representation learning with contrastive predictive coding. *CoRR abs/1807.03748* (2018).

[30] WHITEBREAD, D., NEALE, D., JENSEN, H., LIU, C., SOLIS, L., HOPKINS, E., HIRSH-PASEK, K., AND ZOSH, J. The role of play in children's development: a review of the evidence, 11 2017.

[31] WU, J., SUN, X., ZENG, A., SONG, S., LEE, J., RUSINKIEWICZ, S. M., AND FUNKHOUSER, T. A. Spatial action maps for mobile manipulation. *ArXiv abs/2004.09141* (2020).

[32] WU, Y., YAN, W., KURUTACH, T., PINTO, L., AND ABBEEL, P. Learning to manipulate deformable objects without demonstrations. *CoRR abs/1910.13439* (2019).

[33] YAN, W., VANGIPURAM, A., ABBEEL, P., AND PINTO, L. Learning predictive representations for deformable objects using contrastive estimation. *CoRR abs/2003.05436* (2020).

[34] YOUNG, S., PARI, J., ABBEEL, P., AND PINTO, L. Playful interactions for representation learning. *CoRR abs/2107.09046* (2021).

[35] ZBONTAR, J., JING, L., MISRA, I., LECUN, Y., AND DENY, S. Barlow twins: Self-supervised learning via redundancy reduction. *CoRR abs/2103.03230* (2021).

[36] ZHAN, A., ZHAO, R., PINTO, L., ABBEEL, P., AND LASKIN, M. A framework for efficient robotic manipulation. In *Deep RL Workshop NeurIPS 2021* (2021).