

Learning Keypoints from Synthetic Data for Robotic Cloth Folding

Thomas Lips, Victor-Louis De Gussemme and Francis wyffels
AI and Robotics Lab (AIRO), IDLab, Ghent University - imec
Thomas.Lips@UGent.be



Fig. 2: Left - Examples of the synthetic data used to train the keypoint detector. Center/Right - Successful folds on towels from the evaluation set. Each image pair shows the initial state with the detected keypoints (yellow dots) and the final state after attempting to grasp and fold.

Abstract—Robotic cloth manipulation is challenging due to its deformability, which makes determining its full state infeasible. However, for cloth folding, it suffices to know the position of a few semantic keypoints. Convolutional neural networks (CNN) can be used to detect these keypoints, but require large amounts of annotated data, which is expensive to collect. To overcome this, we propose to learn these keypoint detectors purely from synthetic data, enabling low-cost data collection. In this paper, we procedurally generate images of towels and use them to train a CNN. We evaluate the performance of this detector for folding towels on a unimanual robot setup and find that the grasp and fold success rates are 77% and 53%, respectively. We conclude that learning keypoint detectors from synthetic data for cloth folding and related tasks is a promising research direction, discuss some failures and relate them to future work. A video of the system, as well as the codebase, more details on the CNN architecture and the training setup can be found at <https://github.com/tlpss/workshop-icra-2022-cloth-keypoints.git>.

Index Terms—Synthetic Data, Procedural Data Generation, Keypoint Detection, Robotic Cloth Folding

I. INTRODUCTION

Robotic manipulation of deformable objects is challenging, both in terms of perception, control and modelling [1]. This is a.o. caused by their high-dimensional state [2]. Nonetheless, any general-purpose robotic manipulation system would encounter such objects. Cloth, for example, is omnipresent in household settings. Although significant progress has been made, cloth manipulation remains a challenging task in terms of perception and control [1]–[4].

Supervised deep learning (DL) has the potential to tackle these perception tasks and to generalise to a wide range of settings. However, DL is far from data-efficient and requires large numbers of labelled data to learn and generalise. Procedural data generation is an appealing alternative to manually collecting these datasets: Synthetic data is cheap to generate, has perfect annotations and can be used to generate all

desired variations. The drawback is that the resulting network often has a reduced performance when transferred to the real world. This is due to the remaining differences between the simulation and the real world, referred to as the reality gap.

In this paper, we tackle towel folding, a standard task for cloth manipulation [5], [6]. We approach this task by using a convolutional neural network (CNN) as keypoint detector to estimate the 2D positions of the towel corners from a single RGB image and executing a scripted, open-loop grasp and quasi-static fold trajectory based on these semantic keypoints. We train the keypoint detector entirely on synthetic data and evaluate the zero-shot sim-to-real performance on a unimanual robotic setup under various conditions. Using the terminology of [3], [7], we assume the towels are unfolded but not necessarily perfectly flattened. Recent progress in cloth unfolding has made this a realistic assumption [8].

Our contributions are as follows:

- We show that synthetic data can be used to train convolutional neural networks to detect keypoints for cloth folding.
- We extensively evaluate the performance of our keypoint detector and unimanual folding system.
- We provide qualitative insights into the failure cases.

II. RELATED WORK

A. Robotic Cloth Folding

Cloth manipulation has been extensively studied [2], [3]. For cloth folding in particular, [4], [6], [9] have devised complete pipelines. These works rely on segmentation and a combination of polygonal approximations and template matching or corner detectors to localise keypoints (landmarks) of unfolded cloth items and subsequently fold them. Although these systems perform very well in their test settings, the

segmentation algorithms assume a known background color distribution, which strongly limits the potential to generalise to diverse environments. Furthermore the perception pipeline can take up to 5 seconds [4] to detect the keypoints.

B. Learning Keypoint Representations for Robotic Control

Recently, CNNs were trained to detect [10]–[13] and even discover [14], [15] semantic keypoints end-to-end for various robotic tasks. As the number of detected keypoints can often vary, keypoint detectors usually output spatial heatmaps instead of cartesian coordinates [16], [17].

C. Synthetic Data for Computer Vision

Procedural data generation for training perception models has been used to learn state representations for various robotic tasks [11], [12], [18]. The main limitation is the induced reality gap, which limits the performance on the real-world target domain. Tobin et al. [19] introduced domain randomization as a way to overcome the reality gap by enlarging the distribution from which the data is generated to ensure it entails the real-world data. However, more recent work has stressed the importance of more realistic image synthesis [18], [20] and showed that over-randomizing can result in loss of context and performance [21]. Most authors find that even with domain randomization, the mismatch between both distributions is still too large and resort to finetuning with real-world data to increase performance [22]. The factors that cause this reality gap are not well-known and one usually attempts to close it with task-specific tuning [18], [22].

III. METHOD

A. Synthetic Dataset Generation

We use Blender [23] and BlenderProc [24] to generate data samples. For each image, we build a new 3D scene and randomize object geometries, materials, lighting, and camera pose. Examples can be found in Fig. 2. The towel geometry is modelled as a rectangular mesh. We created a procedural material that combines a random HSV colour and a Perlin noise texture to omit the need for realistic fabric textures, which are difficult to obtain. Additionally, we sample up to 5 distractor objects from a subset of the Thingi10k dataset [25] and add them randomly to the scene, as this was observed to reduce false positives on the manipulator and other objects present in the scene. The ground plane material combines a random HSV base colour with a texture from PolyHaven¹ to introduce spatial patterns. To generate complex and realistic lighting of the scene, we use 360-degree images as environment textures, which are also obtained from PolyHaven. The position of the camera is sampled inside a spherical cap and its orientation is set to point towards the centre of the scene. Finally, the scene is rendered into a 256x256 image with Cycles, Blender’s physically-based path tracer.

B. Keypoint Detector

To detect the desired keypoints on an RGB image, we use a fully convolutional neural network to predict a single spatial heatmap that contains all corners of the cloth. Inspired

by Vecerik et Al. [10], a U-net architecture [26] is used to combine spatial resolution conserving paths with downsampling paths to obtain a large receptive field. In the spatial bottleneck, ResNet-inspired skip connections [27] are used. Bilinear upsampling is used for the upsampling layers. All hidden layers use the ReLU activation function. The final layer uses a sigmoid, of which the outputs are interpreted as the probability function of having a keypoint centred on that location [16]. Heatmaps for the synthetic data samples are generated using a pixel-wise maximum over 2D Gaussian blobs that are centred on each keypoint [17]. Pixel-wise Binary Cross Entropy is used as a training objective and keypoints are extracted using a max-filter with a configurable receptive field, as implemented in Scikit-Image [28]. We use Pytorch Lightning [29] to train the CNN and Weights and Biases [30] for experiment tracking.

C. Robotic Folding

To initiate a fold, a top-down RGB image is passed through the keypoint detector network. If less than four keypoints are detected, the folding is aborted; otherwise, the four keypoints with the highest probability are extracted and reprojected onto the table plane. Based on these keypoints, a local frame is defined and the scripted pregrasp pose, grasp pose and fold trajectory in this local frame are transformed to the robot frame. The sequence is then executed by the robot. The robot grasps the towel in the middle of the side that needs to be folded. The folding trajectory is an arc from that grasp point to the corresponding point on the other side of the towel.

IV. EXPERIMENTS

A. Evaluation

To evaluate the keypoint detector, two sets of towels are used. All towels were collected by asking a number of lab members to bring random towels to reduce bias. The first set contains 11 towels with various colours and material properties but with uniform textures, as modelled in the synthetic data. These are referred to as the in-distribution towels. The second set consists of 9 towels that are not in the distribution of the synthetic data as they have highly non-uniform textures or very different material properties. We refer to this additional set as the out-of-distribution towels and include them to evaluate to what degree the neural network is able to generalise from the synthetic data.

We use three different environment settings, in which we vary lighting and the presence of 6 random chosen distractor objects. For each setting, the robot attempts to fold each towel twice. The pose of the towel is manually randomised within the workspace of the robot before each attempt. For some larger towels, we already partially fold the towel to reduce their dimensions. The partially-folded towels range from 0.2 m to 0.5 m in size.

As the most informative metric of a robotic system is the task performance, we evaluate the keypoint detector by measuring the grasp success rate. Grasp success is defined as the ability of the system to enclose the cloth at the grasp pose. Note that this is different from [5], where grasps are only considered successful if they are held during the entire

¹<https://polyhaven.com/>



Fig. 3: a) Examples of trials for which the keypoint detection failed, resulting in a failed grasp and fold. b) Examples of trials for which the fold failed after a successful grasp. Each image pair shows the initial state with the detected keypoints (yellow dots) and the final state after attempting to grasp and fold.

manipulation. Fold success is defined, as suggested in [5], as the approximate coincidence of the opposing corners after manipulation.

B. Keypoint Detector

Using the procedure described in Section III-A, a dataset of 30,000 images is generated, of which samples can be found in Fig. 2. The dataset is generated on a Dell XPS 9570 laptop with a low-range Nvidia GTX 1050Ti mobile GPU. Generating a single sample takes 1.8 s on average, of which 0.6 s is spent on building the scene and 1.2 s on rendering.

We then train our keypoint detector to predict all visible corner keypoints using sensible hyperparameters based on previous experience. We train for 15 epochs, which takes 55 minutes on an Nvidia V100 GPU and results in an average precision (with 2 pixel threshold) of 80% on a synthetic validation set. We refer to the accompanying Github page for details about the network architecture and other training parameters.

C. Robotic Setup

We use a UR3e robot and Robotiq 2F-85 gripper. We 3D-printed fingertips with a width of 0.08 m using flexfill TPU 98A filament to ensure compliance of the fingers while sliding underneath the cloth for grasping. The camera is a ZED2i that is mounted 1 m above the table and of which the extrinsics have been determined upfront. The images are cropped to the relevant area and then rescaled to the CNN’s input size of 256x256 pixels.

D. Results

The performance of the system is reported in Table I. The total grasp success rate for the towels that are similar to those modelled in the synthetic data is about 77%. From the success rates, it is also clear that the presence of distractor items or changes in lighting conditions, both of which are important real-world domain shifts, do not influence the performance of the keypoint detector. This shows that the randomizations in the synthetic data generation were effective in this regard.

The fold success rate is about 53%. Examples of successful folds (and hence grasps) for various settings can be seen in Fig. 2. The fold success rate is lower than the grasp success rate due to the limitations of scripted trajectories and single-arm manipulation: About 80% of all fold failures are

Setting	In-distribution Towels		Out-of-distribution Towels	
	Grasp	Fold	Grasp	Fold
Natural Light	17/22	14/22	9/18	8/18
LED Light	16/22	10/22	8/18	8/18
Distractors	18/22	11/22	9/18	4/18
Total	51/66 (77%)	35/66 (53%)	26/54 (48%)	20/54 (37%)

TABLE I: Performance of the system for real-world robotic folding of towels using zero-shot transfer for the keypoint detector.

caused by corner misalignment as the corners bend during the execution of the fold motion (see bottom row of Fig. 3b). This could be resolved with bimanual manipulation or by enlarging the fingertips to better perform a line fold [3]. For towels that are already partially folded, the thickness and material properties can change, which makes the open-loop trajectory no longer suited to fold them (see top row example in Fig. 3b), explaining the remaining fold failures.

Most grasp failures are due to inaccurate or incomplete keypoint detection by the transferred neural network, examples of which can be seen in Fig. 3a. These failures indicate there is still a reality gap that was not covered by the variations introduced in our synthetic dataset, requiring additional tuning to better match the target distribution. For the towels that contained non-uniform textures and were hence not explicitly modelled, the grasp success rate deteriorates to 48% (see Table I). The remaining grasp failures are due to some towels being very light, causing the fingers to push the cloth away instead of reaching underneath and grasping it due to a lack of friction.

Finally, as recommended in [5], we report the execution time of our robot system: A forward pass of the neural network is, even on CPU, negligibly fast ($\ll 1$ s) compared to the time it takes to execute the grasp and fold, which is about 20 s.

V. CONCLUSION AND FUTURE WORK

In this paper, we used synthetic data to train a neural network to detect keypoints as a low-dimensional state representation for cloth folding. We transferred the detector without any finetuning (zero-shot) and measured its performance by using the keypoints for folding towels on a robot setup. The results indicate that using procedural data generation is a viable approach to training keypoint detectors for cloth manipulation. However, more extensive tuning is required to completely overcome the reality gap. In future work, we plan to further explore what factors matter to close this gap. This will enable increased performance and provide more principled guidelines for procedural data generation in general.

Additionally, the gap between the grasp and fold success rates indicates that not only the perception but also the manipulation of cloth remains a challenging task. In future work, we therefore aim to close the control loop to take the cloth properties into account and make the control more robust.

ACKNOWLEDGMENT

The authors wish to thank the members of the *Keypoints Gang*, in particular Rembert Daems and Peter De Roovere, for sharing many insights into Computer Vision and academic research in general. This research is supported by the Research Foundation Flanders (FWO) under Grant numbers 1S56022N (TL) and 1SD4421N (VDG).

REFERENCES

- [1] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021.
- [2] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, 2018.
- [3] J. Borràs, G. Alenyà, and C. Torras, "A grasping-centered analysis for cloth manipulation," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 924–936, 2020.
- [4] A. Doumanoglou, J. Stria, G. Peleka, I. Mariolis, V. Petrík, A. Kargakos, L. Wagner, V. Hlaváč, T.-K. Kim, and S. Malassiotis, "Folding clothes autonomously: A complete pipeline," *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1461–1478, 2016.
- [5] I. Garcia-Camacho, M. Lippi, M. C. Welle, H. Yin, R. Antonova, A. Varava, J. Borràs, C. Torras, A. Marino, G. Alenyà, and D. Kragic, "Benchmarking bimanual cloth manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1111–1118, 2020.
- [6] 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*, 2010, pp. 2308–2315.
- [7] A. Verleysen, M. Biondina, and F. Wyffels, "Video dataset of human demonstrations of folding clothing for robotic folding," *The International Journal of Robotics Research*, vol. 39, no. 9, pp. 1031–1036, 2020.
- [8] H. Ha and S. Song, "Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding," in *Conference on Robot Learning*. PMLR, 2022, pp. 24–33.
- [9] S. Miller, J. van den Berg, M. Fritz, T. Darrell, K. Goldberg, and P. Abbeel, "A geometric approach to robotic laundry folding," *The International Journal of Robotics Research*, vol. 31, no. 2, pp. 249–267, 2012.
- [10] M. Vecerik, J.-B. Regli, O. Sushkov, D. Barker, R. Pevceviciute, T. Rothörl, R. Hadsell, L. Agapito, and J. Scholz, "S3k: Self-supervised semantic keypoints for robotic manipulation via multi-view consistency," in *Conference on Robot Learning*. PMLR, 2021, pp. 449–460.
- [11] Z. Qin, K. Fang, Y. Zhu, L. Fei-Fei, and S. Savarese, "Keto: Learning keypoint representations for tool manipulation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 7278–7285.
- [12] J. Wang, S. Lin, C. Hu, Y. Zhu, and L. Zhu, "Learning semantic keypoint representations for door opening manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6980–6987, 2020.
- [13] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2938–2946.
- [14] T. D. Kulkarni, A. Gupta, C. Ionescu, S. Borgeaud, M. Reynolds, A. Zisserman, and V. Mnih, "Unsupervised learning of object keypoints for perception and control," *Advances in neural information processing systems*, vol. 32, 2019.
- [15] B. Chen, P. Abbeel, and D. Pathak, "Unsupervised learning of visual 3d keypoints for control," in *International Conference on Machine Learning*. PMLR, 2021, pp. 1539–1549.
- [16] T. Jakab, A. Gupta, H. Bilen, and A. Vedaldi, "Unsupervised learning of object landmarks through conditional image generation," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [17] X. Zhou, D. Wang, and P. Krähenbühl, "Objects as points," *arXiv preprint arXiv:1904.07850*, 2019.
- [18] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Conference on Robot Learning*. PMLR, 2018, pp. 306–316.
- [19] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2017, pp. 23–30.
- [20] A. Tsirikoglou, J. Kronander, M. Wrenninge, and J. Unger, "Procedural modeling and physically based rendering for synthetic data generation in automotive applications," *arXiv preprint arXiv:1710.06270*, 2017.
- [21] A. Prakash, S. Boochoon, M. Brophy, D. Acuna, E. Cameracci, G. State, O. Shapira, and S. Birchfield, "Structured domain randomization: Bridging the reality gap by context-aware synthetic data," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7249–7255.
- [22] J. Tremblay, A. Prakash, D. Acuna, M. Brophy, V. Jampani, C. Anil, T. To, E. Cameracci, S. Boochoon, and S. Birchfield, "Training deep networks with synthetic data: Bridging the reality gap by domain randomization," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 969–977.
- [23] B. O. Community, *Blender - a 3D modelling and rendering package*, Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. [Online]. Available: <http://www.blender.org>
- [24] M. Denninger, M. Sundermeyer, D. Winkelbauer, Y. Zidan, D. Olefir, M. Elbadrawy, A. Lodhi, and H. Katam, "Blenderproc," *arXiv preprint arXiv:1911.01911*, 2019.
- [25] Q. Zhou and A. Jacobson, "Thing10k: A dataset of 10,000 3d-printing models," *arXiv preprint arXiv:1605.04797*, 2016.
- [26] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.
- [27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [28] S. Van der Walt, J. L. Schönberger, J. Nunez-Iglesias, F. Boulogne, J. D. Warner, N. Yager, E. Goullart, and T. Yu, "scikit-image: image processing in python," *PeerJ*, vol. 2, p. e453, 2014.
- [29] W. Falcon and The PyTorch Lightning team, "PyTorch Lightning," 3 2019. [Online]. Available: <https://github.com/PyTorchLightning/pytorch-lightning>
- [30] L. Biewald, "Experiment tracking with weights and biases," 2020, software available from wandb.com. [Online]. Available: <https://www.wandb.com/>