

A Virtual Reality Framework for Human-Robot Collaboration in Cloth Folding

Marco Moletta, Maciej K. Wozniak, Michael C. Welle and Danica Kragic

Abstract—We present a virtual reality (VR) framework to automate the data collection process in cloth folding tasks. The framework uses skeleton representations to help the user define the folding plans for different classes of garments, allowing for replicating the folding on unseen items of the same class. We evaluate the framework in the context of automating garment folding tasks. A quantitative analysis is performed on three classes of garments, demonstrating that the framework reduces the need for intervention by the user. We also compare skeleton representations with RGB images in a classification task on a large dataset of clothing items, motivating the use of the proposed framework for other classes of garments.

I. INTRODUCTION

There is a growing demand for the automation of garment production and recycling processes. Automating cloth manipulation tasks, such as folding or assistive dressing, could provide considerable benefits in terms of decreasing labor expenses and reducing physical effort for workers [1]. Yet, robotic manipulation of deformable objects remains a significant scientific and industrial challenge [2]. Robots have to rely on methods that can cope with significant self-occlusions, complex interaction dynamics, motion and task planning using multimodal data [3]. Recent approaches resort to deep learning models to alleviate these challenges [4], [5]. Unfortunately, such models require a lot of training data, which is expensive to collect for deformable object manipulation. Therefore, there is a need for approaches that enable efficient data collection for garment manipulation tasks and ease collaboration between robots and humans to reduce the effort and intervention by the latter.

In this context, virtual-, augmented-, and mixed-reality (VAM) frameworks can be useful to improve human-robot collaboration, providing a common interface for the human and robot to interact [6]. Such interfaces allow for more intuitive and natural communication, simplifying for the human to give instructions to the robot and for the robot to communicate its intentions [7]. Different frameworks have been used in robotics research for data collection and for automating and facilitating industrial processes [8], [9]. These frameworks have shown to be preferable over traditional 2D screen interfaces, as the latter are often less intuitive and more workload-intensive for human operators [10]. However, few of the current VR interfaces are tailored to clothing manipulation, which currently still requires a high level of human assistance.

The authors are with the Robotics, Perception and Learning Lab, EECS, at KTH Royal Institute of Technology, Stockholm, Sweden moletta, maciejw, mwelle, dani@kth.se

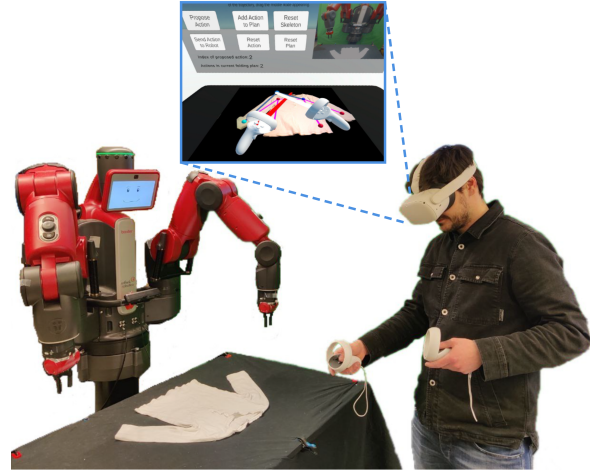


Fig. 1: The proposed VR framework enables human-robot collaboration for data collection in cloth folding tasks. The user defines the folding plan by recording the sequence of pick-place actions to be executed by the robot. The plan is then replicated on unseen items of the same class.

In this paper, we present a framework for facilitating human-robot collaboration for data collection in cloth folding tasks. The concept is visualized in Fig. 1: the interface leverages the immersive experience provided by VR to intuitively define folding plans. The proposed framework uses skeleton representations of garments to allow for automatically replicating folding plans demonstrated by the user on unseen garments of the same class, resulting in increased automation and reduced human intervention. We evaluate the framework by defining folding plans on 3 different classes of garments, requesting the robot to replicate the plans on unseen items of the same class. We assess the level of automation and efficiency of the system by measuring the amount of intervention needed by the user. To further validate the use of skeleton representations for automating folding tasks on novel clothing items, we also employ them in an unsupervised classification task, comparing the results with RGB images. In summary, the contributions of our work are:

- A virtual reality framework based on skeleton representations that automates cloth folding for different classes of garments.
- A quantitative evaluation of our framework for its efficacy in automating cloth folding.
- A comparison between skeleton representations and RGB images in a garments classification task.

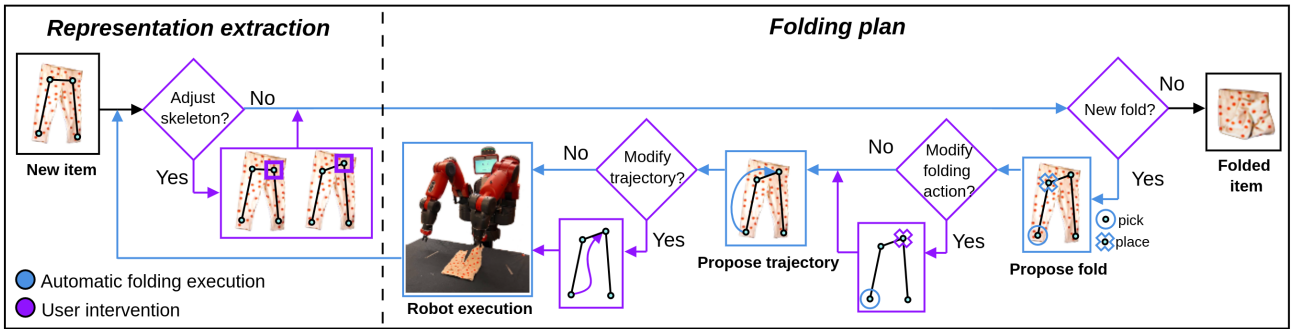


Fig. 2: The framework enables the automation of a folding task (*Automatic Folding Execution*) while allowing the user to intervene (*User Intervention*) if the robot proposes wrong or different folding actions than the ones desired by the user.

II. RELATED WORK

In this section, we provide an overview of the current VAM frameworks used in robotics manipulation as well as of the current approaches to cloth manipulation, highlighting the challenges that hinder the complete automation of cloth folding tasks.

A. VAM frameworks in robotics

VAM frameworks have been applied in a variety of contexts in robotics, including motion planning, control, and human-robot interaction [11], [12]. Studies have also been conducted on the impact of factors such as gender on interactive robotics and teleoperation [13], and virtual gaming platforms have been developed to investigate interpretability and trust in human-robot collaboration [14].

A significant body of work has focused on the use of VR interfaces for control and teleoperation of robotic systems. In these studies, the VR interface is used to manipulate the robotic arm usually via position or velocity control [15], [16] for example to facilitate data collection for training imitation learning policies [17]. In [18], a VR framework is proposed for controlling a multi-robot system composed of various manipulator robots. Additionally, VAM frameworks have been proposed to promote the collaboration between humans and robots in tasks such as handover [19] and to visualize the states and intentions of the robots in delivery tasks [7]. These types of frameworks have also been applied in industrial settings [20], [9] and for interactive programming of robots [21]. Some of these frameworks typically also enable visualization of planning in virtual reality prior to task execution [22], [8].

In the specific context of cloth manipulation, VAM frameworks are not yet well-established. Waymouth et al. [8] did propose a 2D interface and a 3D AR framework for collecting human demonstrations for the task of folding garments. Differently from our framework, that interface neither proposed a folding plan nor allowed for replication, increasing the workload on the user. A similar framework presented in [23] allowed for the automation of data collection to train a deep learning architecture to perform robotics tasks, and was evaluated on folding a small piece of cloth. However, differently from our framework, that interface does

not enable real-time interaction with the robot, which is necessary to enable human-robot collaboration. Borràs et al. [24] proposed a VR framework capable of simulating realistic garments in real-time, allowing the user to collect interactions through handheld controllers. However, this framework does not integrate a physical robot and solely allows the accumulation of synthetic data. Furthermore, it lacks features to automate these interactions.

B. Cloth manipulation approaches and challenges

Manipulating clothing items poses a significant challenge due to the large number of degrees of freedom they exhibit, which causes their configuration space to be infinite-dimensional [2]. To overcome this challenge, current approaches to cloth manipulation resort to learning or extracting representations. For instance, some methods use particle-based representations such as graphs or meshes to discretize the cloth and approximate their dynamics [25], [4], while others focus on learning latent representations to reduce the dimensionality of the configuration space and enable planning [26], [27].

In this context of garment manipulation, the actions are usually defined as pick-place locations on the garment, either by extracting landmarks from images [28], [29], [30] or by selecting nodes of the graph representations. However, since the representations are extracted from data on a particular class of garments, the learned policies and plans typically fail to generalize to unseen shapes, sizes, colors, or properties [31]. Furthermore, most folding methods assume lifting the cloth to a predefined height to complete the folding action [32], [33] as most of the interfaces to collect folding actions do not include the possibility to vary the height or the trajectory. However, the final configuration of the garment varies considerably in relation to different heights of the pick-place actions. It is then important to include this information in the data collection process.

To enable generalization it is necessary to collect a large number of expert demonstrations for each class of garments, as well as to supervise and possibly modify the actions of the robot when deploying these models in unseen conditions. This motivates the framework proposed in this paper, which is explained in detail in the following section.

III. PROPOSED FRAMEWORK

The goal of our VR framework, shown in Fig. 2, is to enhance the autonomy of a robot in the data collection of cloth folding task. To this end, the framework enables a human user to supervise and assist the robot during the most challenging steps currently not extensively addressed in clothing manipulation, which are the definition of a) the sequence of folding actions (folding plan), b) the pick-place locations and c) the trajectory of each folding action. The structure of the framework is composed of two main stages: a representation extraction stage and a folding plan stage, where the latter also enables the automatic replication of the folding task. These two stages are designed with the goal of reducing the workload of the user to define the folding plan for new garments of the same class. The user is still asked to approve all the steps in the pipeline to avoid errors by the robot during the folding execution.

A. Representation extraction

The representation extraction stage consists of creating a skeleton representation of the garment, which is a graph where nodes are used as the pick-place locations that define each folding action. The user can modify the position of the nodes in the VR environment to be in any location on the surface of the garment, as in Fig. 3. The skeleton representation is intended as an aid to increase automation by reducing the need for the user to define all the pick-place locations of the folding actions.

The skeleton representations are extracted from images obtained from an RGB camera. The RGB images are initially background masked and binarized. Next, we perform *skeletonization* on the binary image using the thinning algorithm in [34], [35] and then transform the result into a graph by creating nodes and edges as in [36]. The main advantage of using a skeleton representation is that the proposed action can be defined on the nodes' indices rather than pixel coordinates, making it easier to replicate it on new items given that skeletons extracted from the same class of garments most likely hold the same adjacency matrix.

B. Folding plan

A folding plan is a sequence of consecutive pick-place folding actions defined by the human and executed by the robot. A folding plan is defined for a specific class of garments. To create a new folding action, the user selects and activates a pick- and a place-node of the skeleton (Fig. 4a). This triggers the spawning of a middle-node in between the two, which allows the user to increase or decrease the height of the trajectory of the end effector of the robot (Fig. 4b). Once the action is defined, it can be saved in the current folding plan by using the *Add Action to Plan* button of the interface, visible in Fig. 5. The folding plan can then be replicated on new garments of the same class, which is called *Automatic folding execution*. In the replication process, the framework suggests a series of pick-place actions and relative trajectories based on the new representation extracted from the unseen item to be folded (Fig. 6). The user can ask the

robot to propose a folding action using the *Propose Action* button. The action can be approved by pressing *Send Action to Robot* or modified by the user before execution. The modifications consist of both resetting and changing the pick-place nodes of the action (*Reset Action*) as well as adjusting the intermediate position of the end-effector throughout the movement (the "height" of the trajectory).

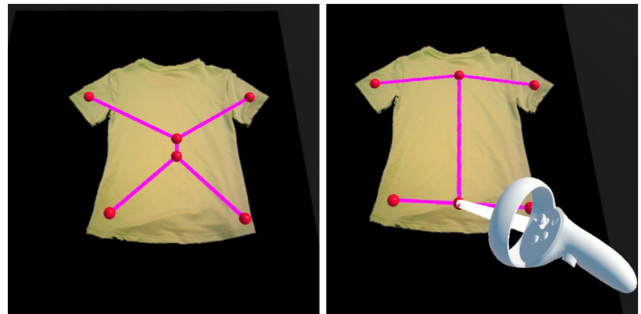


Fig. 3: Modification of nodes locations through joystick interaction.

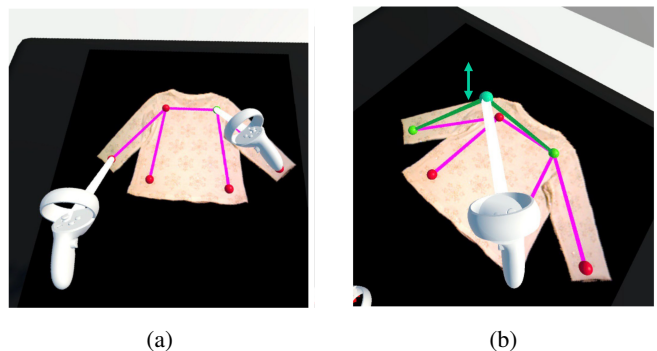


Fig. 4: (a) The folding action is defined by selecting the pick-place nodes. (b) The trajectory can be modified by increasing or decreasing the height of the middle-node.

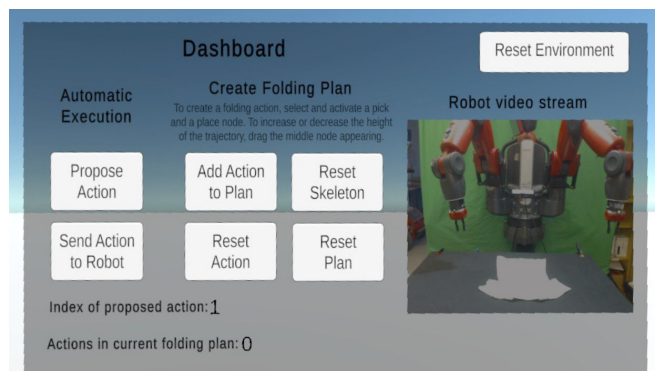


Fig. 5: The dashboard the user uses to create, reset, and save folding actions and plans. The interface allows to visualize the number of actions in the saved folding plan, as well as the real-time video streaming of the robot.

C. Hardware setup

We implemented the user interface in Unity 2020.3 and deploy it on an Oculus Quest 2 headset. We use a Baxter robot, which was already previously used in cloth folding manipulation tasks [26]. A ROS (Robot Operating System) node is used to exchange information between the robot and the user through TCP (Transmission Control Protocol).

IV. EXPERIMENTS

In this section, we evaluate how the proposed framework can facilitate the automation of folding different classes of garments. In particular:

- 1) We provide quantitative results on the success of the robot and the amount of intervention needed by the user when folding plans defined for a cloth of a specific class are automatically replicated on other clothes of the same class.
- 2) To support the hypothesis that skeleton representations are different enough across different classes to allow folding plans to be replicated, we provide the results of an unsupervised classification task on a dataset of clothing items and compare them to RGB images.

A. Automatic folding execution and User intervention

The objective of the experiment is to assess to which extent interventions by the user are needed when using the framework in *automatic folding execution*. We evaluate the automation capability by defining and replicating different folding plans on 3 different classes of clothing items, namely *short sleeve top*, *long sleeve top* and *trousers*, following the taxonomy of Deepfashion2 [37]. The object set used in the experiments is visible in Fig. 7.

The evaluation procedure is visible in Algorithm 1. The procedure is as follows: for each new class of garments, an exemplary folding plan is created and saved by the user on one of the items, which is then replicated on all the items in the class. When a new item is placed in front of the robot to be folded, the skeleton representation is visualized in the VR interface and the user asks the robot to propose the actions present in the folding plan. For each of the actions in the folding plan, if the action displayed is the same that the user defined in the ideal folding plan, the user will accept it and robot executes it, otherwise the user will intervene and modify it.

The accuracy results are reported in Table 1. For each item in each class of garments, the folding plan is replicated 3 times, resulting in $3 \cdot |\text{folding plan}|$ proposals of folding actions by the robot. The errors can be due to both failures in the representations extraction and in the proposal of the folding actions to be executed, where an error means a user intervention. When a failure happens in the representation extraction stage, the user has to intervene and redefine the whole folding plan, as the folding action proposals will not match the ones defined in the exemplary folding plan. Failures due to the robot execution (dropping or not releasing the garment, inverse kinematics errors, etc.) are not reported in the results.



Fig. 6: Different folding actions proposed by the robot during the *automatic folding execution*, pick-nodes displayed in light blue, place-nodes in blue.

Algorithm 1 Evaluation procedure of automatic execution.

```

1:  $R$ : Robot,  $U$ : User.
2: for every Class of Garments do
3:    $U$ : defines exemplary Folding Plan
4:   for every Garment in Class do
5:     Garment is placed in front of Robot.
6:      $R$ : displays Skeleton Representation.
7:     if  $U$ : accepts Skeleton Representation then
8:       for every Action in Folding Plan do
9:          $U$ : Adjusts nodes position (if needed).
10:         $R$ : proposes Folding Action.
11:        if  $U$ : accepts Folding Action then
12:           $R$ : performs Folding Action.
13:        else
14:          Proposal error.
15:           $U$ : defines Folding Action.
16:           $R$ : performs Folding Action.
17:        end if
18:      end for
19:    else
20:      Representation error.
21:       $U$ : adjusts Representation.
22:       $U$ : defines new Folding Plan.
23:       $R$ : executes Folding Plan.
24:    end if
25:  end for
26: end for

```

TABLE I: Results of automatic folding execution.

Class	Folding Plan	Item	Representation Accuracy	Proposal Accuracy
short sleeve top	2	purple	3/3	6/6
		green	3/3	6/6
		white	2/3	4/6
short sleeve top	4	purple	3/3	12/12
		green	3/3	11/12
		white	2/3	8/12
long sleeve top	3	large	2/3	6/9
		small	3/3	8/9
trousers	2	pois	1/3	2/6
		white	3/3	6/6



Fig. 7: The objects set used in the automation experiment: 2 *long sleeve top* (small, large), 3 *short sleeve top* (green, white, purple) and 2 *trousers* (white, pois).

From the results we can conclude that the user intervention is limited, meaning that the framework enables a considerable increase in the level of automation in cloth folding tasks. A video showing different automatic folding executions is present in the supplementary material.

B. Garments unsupervised classification task

Due to the vast diversity and complexity of garment classification, it would be unrealistic to test all the possible classes on our framework, especially considering the lack of uniform labeling standards for clothing items. Our aim is to determine the potential replicability of a defined folding plan for unseen clothing items within the same class. This involves measuring differences in skeleton representations across diverse garment classes not depicted in Fig. 7. High accuracy in distinguishing item classes solely based on skeleton representations increases the likelihood of replicating the plan on novel class items. To make our analysis not biased on specific labels, we examine skeleton representations in an unsupervised classification task, comparing results with original RGB images. We resort to two contrastive learning

TABLE II: classes and composition of the training and test sets (number of samples - percentage).

Id	Name	Training set	Test set
0	short sleeve top	3999 - 18.4%	661 - 19.8%
1	long sleeve top	3999 - 18.4%	661 - 19.8%
2	long sleeve outerwear	1877 - 8.6%	257 - 7.7%
3	vest	1645 - 7.5%	226 - 6.7%
4	shorts	1527 - 7.0%	127 - 3.8%
5	trousers	3059 - 14.0%	176 - 5.2%
6	skirt	983 - 4.5%	282 - 8.4%
7	short sleeve dress	2618 - 12.0%	567 - 16.9%
8	vest dress	2070 - 9.5%	380 - 11.4%

TABLE III: classification results using the KNN evaluation protocol for SimCLR, and the linear evaluation protocol for MVGRL.

Model	Top 1 Acc.	Top 5 Acc.
Original-RGB (SimCLR)	54.1 %	84.8 %
Skeleton (MVGRL)	52.4 %	90.7 %
Random	10.9 %	60.2 %

frameworks: SimCLR [38], and MVGRL [39], respectively, for images and skeleton representations.

The dataset used for the evaluation is a subset of the Deepfashion2 dataset [37]. The Deepfashion2 dataset contains 491k images of 13 categories of clothing items, which include attributes such as *scale*, *occlusion*, *zoom-in*, *view-point*. For our evaluation, we removed four under-represented categories (*short sleeve outerwear*, *slings*, *long sleeve dress* and *slings dress*) and used images with attributes: *scale* = *moderate*, *occlusion* = *no/slight*, *zoom-in* = *no* and *viewpoint* = *frontal*. The images in the dataset are downsampled and padded to 160*160 pixels. In total, the final dataset consists of 25102 images divided into 9 classes, where 85% are used for training and 15% for testing. The composition of the dataset can be seen in Table II.

Since in contrastive learning the composition of data augmentations plays a critical role in increasing the performance of the extracted representation, we provide details on the types of augmentations used both for images and skeleton representations. From [38], for original-RGB we apply *random-resize-crop*, *color-jitter*, *Random horizontal flip* and *random grayscale* augmentations. From [39], we apply *diffusion*, which augments the adjacency matrix of the skeleton graph with additional edges. We also employ *Horizontal Flip* and *Vertical Flip*, which mirror the graph respectively on the horizontal and vertical axis. We experimented also with adding noise but found that it decreases performance, which is also pointed out in [39].

To make the comparison as fair as possible, we present and evaluate the results from the best-performing augmentations and the best-performing evaluation protocols from both types of representations, which in our case is the KNN evaluation

protocol for SimCLR [38] and the linear evaluation protocol for MVGRL [39]. In this experiment, the batch size is set to 64, and both models are trained for 1000 epochs. Both models are trained with the loss functions used in the original papers (NT-Xent loss [38] for SimCLR and Jensen-Shannon divergence (JSD) for MVGRL [39]). For comparison, a random classifier is also implemented as a baseline model, which randomly assigns class labels to instances in the test set without any training. We report the Top 1 and Top 5 accuracies of all the models in Table III.

The classification results for skeleton representations are comparable to original RGB images and considerably better than random. This suggests that skeleton representations of different garments are as different as original RGB images, suggesting that folding plans are likely to be reproducible also on other garments not evaluated in the automatic folding experiment. Moreover, these insights open the possibility of integrating a classification framework based on images or skeletons in future work, for example, to propose some default folding plan based on the class of garment.

V. LIMITATIONS AND FUTURE WORK

One current limitation of the framework is that if the extracted skeleton differs for items of the same class, the folding plan is not automatically replicable and require some user intervention. One way to address this may be to allow rather simple skeletons for each class by limiting the number of skeleton nodes. We also plan to run a user study to collect feedback about the interface and test the automation capabilities of the framework more thoroughly with novice users. We also plan to look into defining bimanual folding actions by pick-placing 2 nodes simultaneously, to further improve the data generation aspect by allowing more sophisticated folding trajectories.

VI. CONCLUSION

We presented a VR framework to automate the data collection process of cloth folding tasks. The framework relies on skeleton representations to help the user to define the folding plan for different classes of garments, allowing the replication of the folding plan on unseen items of the same class. We quantitatively evaluate our framework for its efficacy in automating cloth folding and we compare skeleton representations with RGB images in a garments unsupervised classification tasks. We conclude that the framework enhances the automation of cloth folding tasks, and the classification results suggest that this could be valid also for classes of garments not evaluated in the automatic folding experiments. We also plan to further expand the framework for the automation of flattening tasks, with the aim of creating a complete interface for the collection of cloth manipulation data in a more automated manner through human-robot collaboration.

REFERENCES

[1] 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.

[2] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li, *et al.*, "Challenges and outlook in robotic manipulation of deformable objects," *IEEE Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, 2022.

[3] S. Tirumala, T. Weng, D. Seitla, O. Kroemer, Z. Temel, and D. Held, "Learning to singulate layers of cloth using tactile feedback," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7773–7780, 2022.

[4] A. Longhini, M. Moletta, A. Reichlin, M. C. Welle, D. Held, Z. Erickson, and D. Kragic, "Edo-net: Learning elastic properties of deformable objects from graph dynamics," *arXiv preprint arXiv:2209.08996*, 2022.

[5] Z. Huang, X. Lin, and D. Held, "Mesh-based dynamics with occlusion reasoning for cloth manipulation," *arXiv preprint arXiv:2206.02881*, 2022.

[6] M. Wozniak, C. T. Chang, M. B. Luebbers, B. Ikeda, M. Walker, E. Rosen, and T. R. Groechel, "Virtual, augmented, and mixed reality for human-robot interaction (vam-hri)," in *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction, HRI '23*, (New York, NY, USA), p. 938–940, Association for Computing Machinery, 2023.

[7] K. Chandan, V. Kudalkar, X. Li, and S. Zhang, "Arroch: Augmented reality for robots collaborating with a human," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3787–3793, IEEE, 2021.

[8] 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)*, pp. 155–160, IEEE, 2021.

[9] J. Zubizarreta, I. Aguinaga, and A. Amundarain, "A framework for augmented reality guidance in industry," *The International Journal of Advanced Manufacturing Technology*, vol. 102, no. 9, pp. 4095–4108, 2019.

[10] R. Hetrick, N. Amerson, B. Kim, E. Rosen, E. de Visser, and E. Phillips, "Comparing virtual reality interfaces for the teleoperation of robots," pp. 1–7, 04 2020.

[11] Z. Makhataeva and H. A. Varol, "Augmented reality for robotics: A review," *Robotics*, vol. 9, no. 2, p. 21, 2020.

[12] M. Wonsick and T. Padir, "A systematic review of virtual reality interfaces for controlling and interacting with robots," *Applied Sciences*, vol. 10, no. 24, p. 9051, 2020.

[13] F. Nenna and L. Gamberini, "The influence of gaming experience, gender and other individual factors on robot teleoperations in vr," in *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 945–949, 2022.

[14] M. Mara, K. Meyer, M. Heimpl, H. Pichler, R. Haring, B. Krenn, S. Gross, B. Reiterer, and T. Layer-Wagner, "Cobot studio vr: A virtual reality game environment for transdisciplinary research on interpretability and trust in human-robot collaboration," 2021.

[15] C. Barentine, A. McNay, R. Pfaffenbichler, A. Smith, E. Rosen, and E. Phillips, "A vr teleoperation suite with manipulation assist," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 442–446, 2021.

[16] S. Xu, S. Moore, and A. Cosgun, "Shared-control robotic manipulation in virtual reality," *arXiv preprint arXiv:2205.10564*, 2022.

[17] A. Reichlin, G. L. Marchetti, H. Yin, A. Ghadirzadeh, and D. Kragic, "Back to the manifold: Recovering from out-of-distribution states," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8660–8666, 2022.

[18] F. Kennel-Maushart, R. Poranne, and S. Coros, "Multi-arm payload manipulation via mixed reality," in *2022 International Conference on Robotics and Automation (ICRA)*, pp. 11251–11257, IEEE, 2022.

[19] V. Ortenzi, M. Filipovica, D. Abdikarim, T. Pardi, C. Takahashi, A. M. Wing, M. Di Luca, and K. J. Kuchenbecker, "Robot, pass me the tool: Handle visibility facilitates task-oriented handovers," in *HRI*, pp. 256–264, 2022.

[20] Q. Wang, Y. Cheng, W. Jiao, M. T. Johnson, and Y. Zhang, "Virtual reality human-robot collaborative welding: a case study of weaving gas tungsten arc welding," *Journal of Manufacturing Processes*, vol. 48, pp. 210–217, 2019.

[21] M. Ostanin and A. Klimchik, "Interactive robot programming using mixed reality," *IFAC-PapersOnLine*, vol. 51, no. 22, pp. 50–55, 2018.

- [22] M. K. Wozniak, R. Stower, P. Jensfelt, and A. Pereira, "What you see is (not) what you get: A vr framework for correcting robot errors," *arXiv preprint arXiv:2301.04919*, 2023.
- [23] P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, "Repeatable folding task by humanoid robot worker using deep learning," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 397–403, 2016.
- [24] J. Borràs, A. Boix-Granell, S. Foix, and C. Torras, "A virtual reality framework for fast dataset creation applied to cloth manipulation with automatic semantic labelling," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2023.
- [25] X. Lin, Y. Wang, Z. Huang, and D. Held, "Learning visible connectivity dynamics for cloth smoothing," in *Conference on Robot Learning*, pp. 256–266, PMLR, 2022.
- [26] M. Lippi, P. Poklukar, M. C. Welle, A. Varava, H. Yin, A. Marino, and D. Kragic, "Enabling visual action planning for object manipulation through latent space roadmap," *IEEE Transactions on Robotics*, 2022.
- [27] R. Hoque, D. Seita, A. Balakrishna, A. Ganapathi, A. K. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg, "Visuospatial foresight for physical sequential fabric manipulation," *Autonomous Robots*, pp. 1–25, 2021.
- [28] O. Gustavsson, T. Ziegler, M. C. Welle, J. Bütepage, A. Varava, and D. Kragic, "Cloth manipulation based on category classification and landmark detection," *International Journal of Advanced Robotic Systems*, vol. 19, no. 4, p. 17298806221110445, 2022.
- [29] J. Schulman, A. Lee, J. Ho, and P. Abbeel, "Tracking deformable objects with point clouds," in *2013 IEEE International Conference on Robotics and Automation*, pp. 1130–1137, 2013.
- [30] C. Elbrechter, R. Haschke, and H. Ritter, "Folding paper with anthropomorphic robot hands using real-time physics-based modeling," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, pp. 210–215, 2012.
- [31] V. Petrik and V. Kyrki, "Feedback-based fabric strip folding," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 773–778, IEEE, 2019.
- [32] M. Lippi, M. C. Welle, P. Poklukar, A. Marino, and D. Kragic, "Augment-connect-explore: a paradigm for visual action planning with data scarcity," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 754–761, IEEE, 2022.
- [33] Y. Avigal, L. Berscheid, T. Asfour, T. Kröger, and K. Goldberg, "Speedfolding: Learning efficient bimanual folding of garments," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1–8, IEEE, 2022.
- [34] T.-C. Lee, R. L. Kashyap, and C. N. Chu, "Building skeleton models via 3-d medial surface/axis thinning algorithms," *CVGIP Graph. Model. Image Process.*, vol. 56, pp. 462–478, 1994.
- [35] H. Sundar, D. Silver, N. Gagvani, and S. Dickinson, "Skeleton based shape matching and retrieval," in *2003 Shape Modeling International.*, pp. 130–139, 2003.
- [36] F. Reinders, M. E. D. Jacobson, and F. H. Post, "Skeleton graph generation for feature shape description," in *Data Visualization 2000* (W. C. de Leeuw and R. van Liere, eds.), (Vienna), pp. 73–82, Springer Vienna, 2000.
- [37] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5337–5345, 2019.
- [38] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.
- [39] K. Hassani and A. H. K. Ahmadi, "Contrastive multi-view representation learning on graphs," *CoRR*, vol. abs/2006.05582, 2020.