Preprint Version; Final version avaliable at: https://ieeexplore.ieee.org/document/10185075
DOI: 10.1109/LRA.2023.3296371

Visual-Tactile Learning of Garment Unfolding for Robot-Assisted Dressing

Fan Zhang and Yiannis Demiris

***Abstract*— Assistive robots have the potential to support disabled and elderly people in daily dressing activities. An intermediate stage of dressing is to manipulate the garment from a crumpled initial state to an unfolded configuration that facilitates robust dressing. Applying quasi-static grasping actions with vision feedback on garment unfolding usually suffers from occluded grasping points. In this work, we propose a dynamic manipulation strategy: tracing the garment edge until the hidden corner is revealed. We introduce a model-based approach, where a deep visual-tactile predictive model iteratively learns to perform servoing from raw sensor data. The predictive model is formalized as Conditional Variational Autoencoder with contrastive optimization, which jointly learns underlying visual-tactile latent representations, a latent garment dynamics model, and future predictions of garment states. Two cost functions are explored: the visual cost, defined by garment corner positions, guarantees the gripper to move towards the corner, while the tactile cost, defined by garment edge poses, prevents the garment from falling from the gripper. The experimental results demonstrate the improvement of our contrastive visual-tactile model predictive control over single sensing modality and baseline model learning techniques. The proposed method enables a robot to unfold back-opening hospital gowns and perform upper-body dressing.**

***Index Terms*— Tactile manipulation, model-based learning, robot-assisted dressing.**

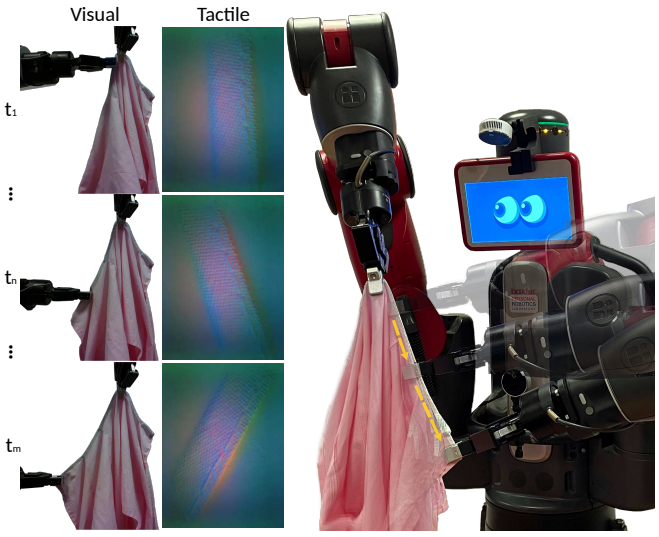

Fig. 1: The robot unfolds the garment by sliding the garment edge inside the gripper without loosing it until the hidden corner is revealed. This is achieved by a visual-tactile-based model predictive controller using contrastive optimization. A RGB-D camera is mounted on the robot for capturing visual data, while an optical tactile sensor is attached on the gripper for tactile sensing.

I. Introduction

Dressing is a challenging basic aspect of the daily life of elderly people, and people who suffer from impairments. Studies have reported that of all the activities of daily living, dressing has shown the highest burden on caregiving staff, but the lowest use of assistive technologies [1]. Recent works have made great progress toward using robots to perform dressing [2-9]. In [10], a full pipeline of dressing bedridden users has been proposed, which includes an intermediate stage of unfolding garment to bring it from an uncertain state into a configuration that facilitates robust dressing.

Most recent research has successfully tackled garment unfolding manipulation by adopting quasi-static prehensile interactions (i.e., grasping) [11]. These approaches usually rely on strong assumptions about the initial stage of the cloth: grasping points are visible or not severely occluded. To expose the key visual features of the cloth for downstream applications, dynamic non-prehensile garment manipulation (i. e., fling motion [12], air-based blowing policy [13]) has been explored to maximize the coverage of the cloth, which has attained great progress. In this work, we adopt another alternative strategy commonly used by humans for garment unfolding, which is to slide the garment edge inside the gripper without loosing it until the hidden corner is revealed, as shown in Fig. 1.

Prior studies on garment edge tracing [14-16] usually rely on only visual feedback, which might suffer from occlusions created by the garment. Inspired in part by the notion that a human can use what he/she feels while performing deformable manipulation by grasping the object between fingers, recent studies has incorporated tactile sensing in various applications including cable following [17], clothing properties recognizing [18] and cloth layer singulation [19]. However, the single tactile sensor modality might not capture the global information that is useful for the task [20]. In this paper, we fuse vision and tactile inputs to complete the task of garment edge tracing for unfolding.

We propose a deep model predictive control (MPC) method, a general framework for model-based deep reinforcement learning, to iteratively perform servoing from raw camera and tactile sensor inputs. The predictive model is formalized as Conditional Variational Autoencoder (CVAE) with contrastive optimization, which jointly learns the underlying visual-tactile latent representations, latent dynamics

The authors are with Personal Robotics Lab, Imperial College London. (e-mail:f.zhang16@imperial.ac.uk; y.demiris@imperial.ac.uk). This research is supported in part by UKRI grant EP/V026682/1, a RAEng Chair in Emerging Technologies to YD, and the Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a Schmidt Futures program.

model for deformable garments, and future predictions of garment states. We first bootstrap the training of the predictive model from demonstration data to guide and accelerate the agent toward good behaviors. Then, MPC with one-step predictions is used to trace the garment edge by minimizing the costs. The predictive model is iteratively updated given the corrective action and the observed current state. Instead of operating on raw sensing data, we calculate the cost using distilled information, including the garment target corner position from the visual perspective and the edge pose from the tactile perspective. The distilled information is learned across offline demonstration data in a supervised network.

We adopt contrastive loss in the predictive model as it has proven to be capable of efficiently learning latent space representations in various applications due to its inherent information maximization objective [21-24]. The contrastive loss could efficiently map pairs of similar observations to nearby points, whereas dissimilar pairs are pushed apart in the embedding space. In our case, the ground truth future state of the garment and its corresponding predicted sample are contrastively encoded to match each other for a better latent dynamics model. Experimental evaluation empirically demonstrates that the proposed predictive model with contrastive optimization outperforms standard model-based learning baselines across garment edge tracing, unfolding, and the consequent dressing tasks. Compared to single visual-sensory input, we also show that the fusion of visual-tactile sensing boosts the performance of the controller over a range of challenging conditions including visual occlusions.

The main contributions can be summarized as follows:

1) A pipeline comprised of garment grasping, unfolding and dressing is proposed, using tactile sensing for cloth unfolding with a dynamic sliding-manipulation strategy.

2) An iterative model predictive controller (MPC) with visual-tactile cost learning is introduced to automate the pipeline. The experimental results not only prove the importance of multimodality, but also explain the contributions of each modality (i. e., tactile preventing cloth dropping and vision data for a mature motion stop).

3) A deep predictive model using contrastive loss in MPC is proposed to jointly learn underlying latent dynamics model for deformable garment and predictions of garment states. The contrastive loss contributes in predicting a representation that matches the encoding of the true future observation.

II. Related Work

A. Robot-Assisted Dressing

Recent works have made great progress toward addressing the challenges of using a robot to perform dressing from multiple perspectives, including dressing different cloth (e. g., scarf [2], T-shirt [3]), dressing assistance for various target users (e. g., healthy users [4], impaired users [6], paralyzed patients [10]), personalized assistive dressing (e. g., taking user capabilities [25] and preferences [26] into account), and learning dressing with a variety of sensing modalities (e. g., vision [3], force [6], capacity [5]). While these different studies have led to promising results of dressing, they usually simplify the setup of their experiments by manually attaching the garment to the robot end-effector. [10] has successfully deployed a complete dressing pipeline, including grasping a hospital gown hung on a rail and fully unfold the gown.

B. Garment Unfolding

Achieving a fully unfolded garment configuration from a severely crumpled initial configuration remains a challenging problem. Most prior works on garment unfolding have formulated this problem as computing suitable grasping points on garments, either through extracting handcrafted features (e. g., wrinkles, edges and corners [27]), or using deep learning methods (e. g., supervised learning [11, 28], reinforcement learning [29]). The garments are usually manually strewn across a flat surface, or hold by a robot gripper in midair using gravity to help expose borders for grasping points detection. Thus these works might fail to generalize to severely self-occluded cloth configurations, when the required key points are not visible.

To reveal the hidden key visual features of the cloth for downstream applications, dynamic manipulation that combines both prehensile and non-prehensile has been explored to spread the cloth. Solutions include flinging motions [12], air-based blowing [13] and swinging [29] to maximize the garment coverage on the table. Some other research has contributed to garment unfolding by sliding the garment edge inside the gripper without loosing it until the hidden corner is revealed. Early works on this topic either rely on custom hardware or only use visual feedback [14-16]. A natural extension is to incorporate tactile modality to imitate human who can manipulate by feel.

C. Tactile Manipulation

Compared to traditional tactile sensing, optical-based tactile sensors, such as GelSight and DIGIT, have proven to be an extremely versatile, high-bandwidth, high-spatial resolution alternative. Some research has exploited these sensors for system property identification through machine learning approaches, including cloth texture identification [18] and liquid classification [30]. Manipulation is another area where optical-based tactile sensors have received attention. Successful execution of ball-rolling in-hand manipulation task has been demonstrated with GelSight and DIGIT sensors using deep tactile model predictive control in [20, 31]. Grasping stability [32], slip detection [33] and object surface following [34] have also been improved using tactile sensing. A few follow ups have extended tactile sensing to deformable object manipulation such as cloth layer singulation [19].

In this work, we focus on deformable object contour following using tactile sensing. [17] has implemented real-time cable following using GelSight sensor. [35] further extends such research by training a reinforcement learning agent with visual-tactile fusion in simulation for cable following. For garment edge tracing, tactile sensing has been successfully deployed by [36] with Linear–Quadratic Regulator controller, and [37] with offline reinforcement learning. However, only

small square cloths, partially strewn across a flat surface, have been investigated in these works.

D. Contrastive Learning in Robot Manipulation

Many prior works have demonstrated that contrastive loss and its variants can efficiently learn latent representations and achieve generalizability in various domains, including vision [22], language [23] and audio representations [24]. Recent works have promisingly utilized contrastive optimization to boost robot manipulation performance. [38] have focused on using Contrastive Predictive Coding [24] to learn an embedding which maps temporally neighboring states close together. Based on the learned latent representations, robot forward dynamic models or behaviour policies are further trained with supervised learning or reinforcement learning. Some other works [21] jointly encode the representation and forward dynamic model in the latent space, so that the predicted and ground truth next state could be mapped close to each other. In our case, we use contrastive learning to get better underlying visual-tactile latent representations and the latent dynamics model, and thus boost the performance of future predictions of garment and sensor states.

III. Visual-Tactile Learning Garment Unfolding

We formulate our robot control task from tactile-visual observations as a partially observable Markov decision process (POMDP). At each time step t, the robot receives the observation $\boldsymbol{o}_t$ that is represented in the latent space as $\boldsymbol{z}_t$, takes an action $\boldsymbol{a}_t$, and moves to the state $\boldsymbol{z}_{t+1}$. Compared to standard model predictive control methods. which use regression over observation-action-observation tuples with L2 reconstruction error, we add a contrastive auxiliary loss to encourage better latent representations and dynamics model.

A. Contrastive Deep Predictive Model

The predictive model consists of the following modules:

$$\begin{aligned} \text{Observation Encoder,} \quad & \boldsymbol{z}_t = f_e(\boldsymbol{o}_t) \\ \text{Latent Forward Dynamics Model,} \quad & \hat{\boldsymbol{z}}_{t+1} = f_g(\boldsymbol{z}_t, \boldsymbol{a}_t) \\ \text{Future Predictor,} \quad & \hat{\boldsymbol{o}}_{t+1} = f_d(\hat{\boldsymbol{z}}_{t+1}) \\ \text{Robot Behaviour Model,} \quad & \boldsymbol{a}_t = \underset{\boldsymbol{a}_s}{\operatorname{argmin}}\, c(\hat{\boldsymbol{o}}_{t+1}) \end{aligned}$$

Observation Encoder: The predictive model is set to predict future visual-tactile sensor observations $\hat{\boldsymbol{o}}_{t+1} = \{\hat{\boldsymbol{I}}^v_{t+1}, \hat{\boldsymbol{I}}^q_{t+1}\}$, a concatenation of the predicted visual and tactile images, conditioned on the current observations of visual image $\boldsymbol{I}^v_t$ and tactile image $\boldsymbol{I}^q_t$, as well as the action $\boldsymbol{a}_t$. As presented in Fig. 2a, we first train an encoder $\boldsymbol{z}_t = f_e(\boldsymbol{o}_t)$ to embed current observations to the latent space.

Latent Forward Dynamics Model: A garment forward dynamics model in the latent space is represented as $\hat{\boldsymbol{z}}_{t+1} = f_g(\boldsymbol{z}_t, \boldsymbol{a}_t)$ to predict next sensor latent state $\hat{\boldsymbol{z}}_{t+1}$ based on the current latent representation and action. We use a contrastive auxiliary loss for training here. The predicted $\hat{\boldsymbol{z}}_{t+1}$ and the ground truth obtained by $\boldsymbol{z}_{t+1} = f_e(\boldsymbol{o}_{t+1})$ close are considered as positive pairs. The contrastive loss results in the positive sample pairs being aligned together but the negative samples pushed further apart, and thus learns stronger and more planable latent representations. Thus our learning objective lies with maximizing mutual information between the predicted encodings and their respective positive samples. We use the InfoNCE contrastive loss [24]:

$$\mathcal{L}_c = -\mathbb{E}\left[\log \frac{s(\hat{\boldsymbol{z}}_{t+1}, \boldsymbol{z}_{t+1})}{\sum_{k=1}^{N} s(\hat{\boldsymbol{z}}_{t+1}, \boldsymbol{z}_k)}\right]$$

where s is a similarity function $s(\hat{\boldsymbol{z}}_{t+1}, \boldsymbol{z}_{t+1}) = \exp(-\|\hat{\boldsymbol{z}}_{t+1} - \boldsymbol{z}_{t+1}\|)^2$, $(\hat{\boldsymbol{z}}_{t+1}, \boldsymbol{z}_k)$ represent negative pairs, $\boldsymbol{z}_k$ is incorrect latent representation of the next state. The final contrastive loss is computed across the positive pair $(\hat{\boldsymbol{z}}_{t+1}, \boldsymbol{z}_{t+1})$ and N samples of negative pairs $(\hat{\boldsymbol{z}}_{t+1}, \boldsymbol{z}_k)$ within minibatch.

Future Predictor: A decoder is then learned to reconstruct visual-tactile predictions form the latent space $\hat{\boldsymbol{o}}_{t+1} = f_d(\hat{\boldsymbol{z}}_{t+1})$ with L2 image reconstruction error $\mathcal{L}_s(\hat{\boldsymbol{o}}_{t+1}, \boldsymbol{o}_{t+1})$.

Overall, we jointly learn the encoder $\boldsymbol{z}_{t+1} = f_e(\boldsymbol{o}_{t+1})$, the latent garment dynamics model $\hat{\boldsymbol{z}}_{t+1} = f_g(\boldsymbol{z}_t, \boldsymbol{a}_t)$, and the decoder $\hat{\boldsymbol{o}}_{t+1} = f_d(\hat{\boldsymbol{z}}_{t+1})$ with a hybrid loss of contrastive loss and L2 Mean Squared Errors with a weighting parameter λ:

$$\mathcal{L} = \mathcal{L}_c + \lambda \mathcal{L}_s$$

Robot Behaviour Model: Having the predictive model, we use the Model Predictive Control (MPC) approach with one-step prediction, where at each time step we randomly sample actions $\boldsymbol{a}_s$ and feed them through the forward model from the current $\boldsymbol{o}_t$, and choose the best action $\boldsymbol{a}_t$ that optimizes the objectives (i. e., the cost functions), as shown in Fig. 2b.

B. Planning Cost Functions

In this section, we discuss how to design cost functions for model predictive control. One naive approach is to use pixel-wise error between a goal image and the predicted image. However, there is an issue with such approaches: large objects in the image (i. e., the robot gripper, garment and shadows) dominate the cost. Common failures occur when the planner matches the robot and the garment positions with their positions in the goal images, while ignoring smaller details of garment edge and its contact with the robot gripper. This failure motivates us to use more sophisticated mechanisms with distilled information to specify cost functions.

Garment edge pose objective: Recent works have used Principal Component Analysis (PCA) to estimate the object pose given the tactile images [17, 31]. Such methods would be challenging when applied to garment edge as the garment is thin and covers a large area, unlike the cables [17] or marbles [31] which create a distinct imprint. Therefore, we propose to train a CNN network to estimate the pose of garment edge (i. e., the starting and ending points of the edge) in a supervised manner. We then use the estimated pixel positions of the two points to calculate the position and the orientation of the garment edge, which are parameterized with respect to the X axis of the tactile sensor with pixel distance y and angle θ, as shown in Fig. 2c. We define the

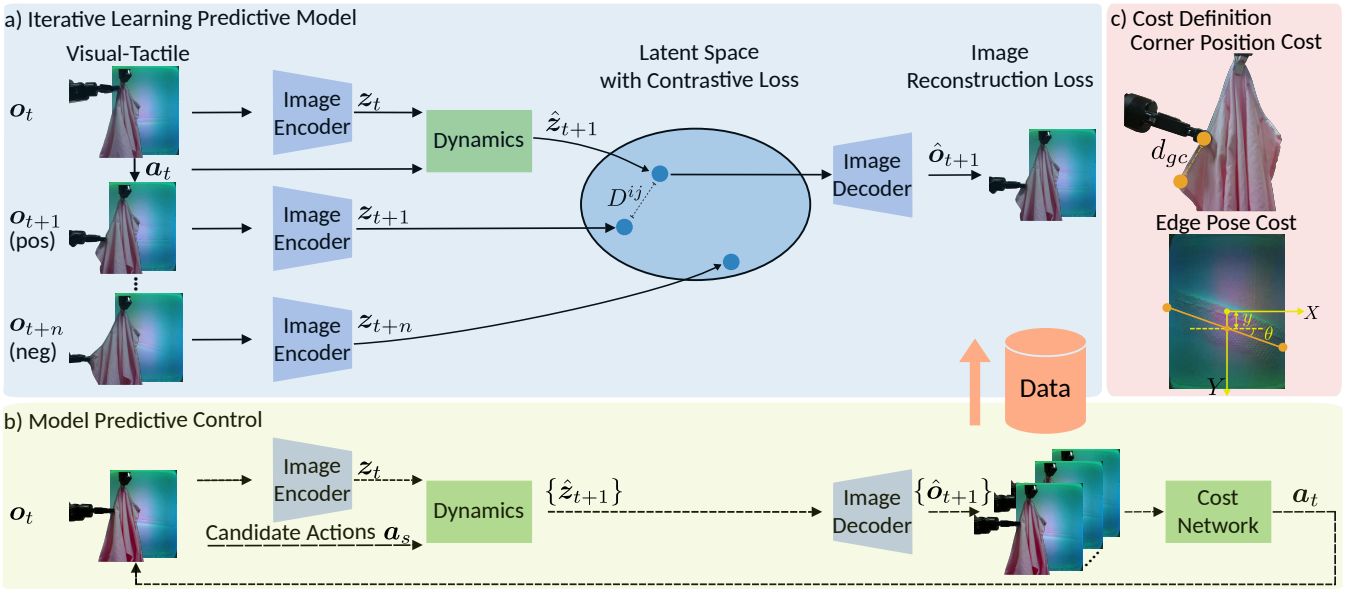

Fig. 2: Framework of iterative deep visual-tactile model predictive control. (a) A predict model is trained using contrastive loss to jointly learn the visual-tactile latent representations, garment dynamic model and future sensor data prediction. (b) Given the current visual-tactile observation, the learned deep predictive model and sampled multiple potential actions, model predictive control with one-step predictions is used to iteratively update the predictive model. The action that attains lowest cost is then applied to the robot. (c) Costs are computed based on the estimated distilled information (i. e., corner target position from visual data and edge pose from tactile sensory).

cost $c_q(\hat{\boldsymbol{I}}^q_{t+1})$ as a weighted sum of y and θ. The objective is to minimize the cost, which maintains the edge position in the center of the tactile sensor, and the orientation of the edge to be parallel to the X axis. This inclines that the gripper could slide the garment along the edge without dropping it.

Garment corner position objective: The objective of garment unfolding task is to slide the edge until the cloth corner. We use CNN networks to respectively locate the garment corner pixel position and the current gripper pixel position on the image at each frame obtained from the vision sensor. The network generates two likelihood heat maps, and the corner and gripper are respectively localized as the pixel with the maximum likelihood in each map. Even in the scenarios when the corner is occluded by the other parts of the cloth which mostly happen in the early stage of edge tracing, we still use the trained network to guess the approximate corner position. Then the cost $c_v(\hat{\boldsymbol{I}}^c_{t+1})$ here is defined by calculating the distance d_{gc} between the gripper pixel position and corner pixel position on the predicted image at each time step, as shown in Fig. 2c. Minimizing this cost guarantees that the gripper would move towards the end of garment edge and stop at the appropriate position.

The overall cost is defined by the sum of edge pose cost and corner position cost $c = c_q + \beta c_v$ with weighting parameter β. The best action $\boldsymbol{a}_t$ from sampled candidates $\boldsymbol{a}_s$ that optimizes this cost is selected at each time step:

$$\boldsymbol{a}_t = \underset{\boldsymbol{a}_s}{\operatorname{argmin}}\, c(\hat{\boldsymbol{o}}_{t+1})$$

C. Implementation Details

We first present the structure details of the predictive model. The custom encoder architecture is a sequence of 2D convolutions kernel sizes [5,5,3,3,3], and filter sizes [32,64,128,128,256] respectively, with ReLU activation, max pooling and batch normalization. The output is flattened and fed into a fully connected layer, followed by a forward model using a multi-layer perceptron (MLP) with two hidden layers of size 64, which is also the dimensionality of the latent representation of visual-tactile sensing. The garment dynamics module is formulated as two dense layers of size 128. The decoder is a dense layer followed by 6 transposed convolutions (reversed structure of the encoder) to upscale back to 256x256 image size. For the InfoNCE contrastive loss, in each batch we include 31 negative samples with one positive pair. The images are scaled to the range of $[0, 1]$. Other components of the neural network include: an Adam optimizer for 100 epochs, a batch size of 32 and a learning rate of 0.0001. We use a Densenet-121 network to estimate the garment corner positions and edge poses.

As discussed in the Section I-Introduction, we first bootstrap the network training from demonstration data, followed by MPC with one-step predictions to iteratively update the predictive model. Demonstration data has been collected in a kinesthetic teaching manner by human users holding the robot arm to perform garment edge sliding. A total of 100 trials of sliding (including 7,862 pairs of visual-tactile images) have been collected. For the iterative MPC, we collect another 100 trials of sliding (6,936 pairs of visual-tactile images). The actions are centered and scaled to the range of [-1, 1], and rescaled back to [-3cm, 3cm] for both x, y and z coordinates. Both the visual and tactile cost networks are trained in a supervised manner using the demonstration data. Five people (ages 25-31, mean: 27, std: 2.21, female: 2), who are familiar with robots, have been involved in sliding

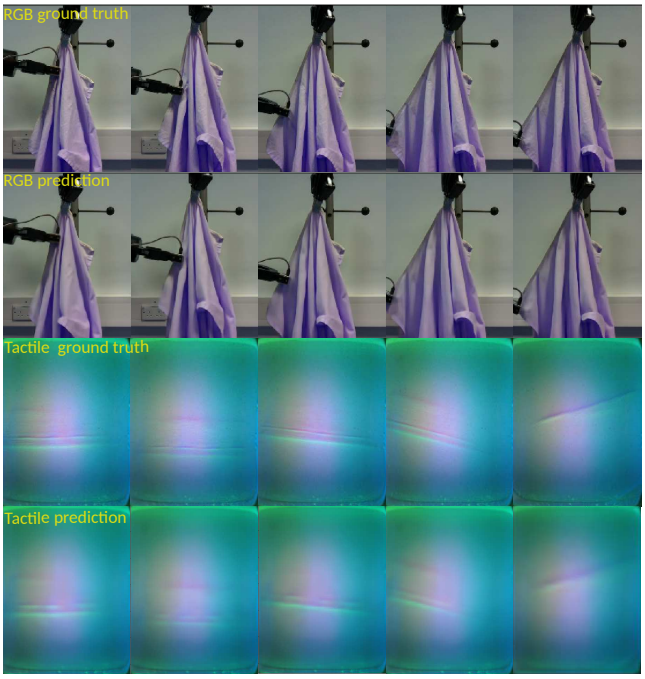

Fig. 3: Examples of ground truth and predicted visual-tactile trajectories of garment edge tracing.

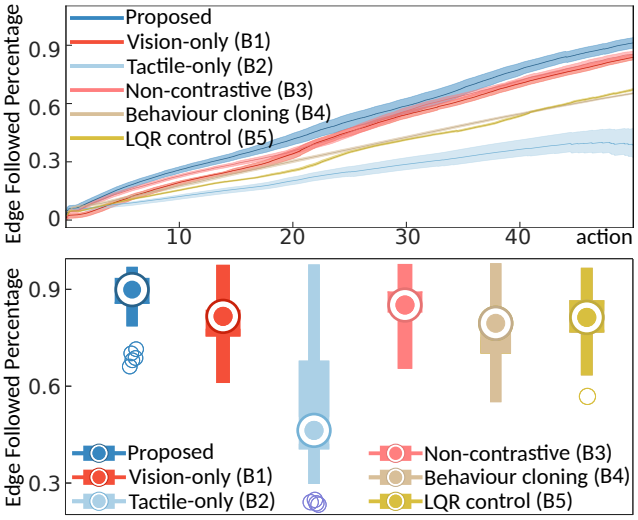

Fig. 4: Results of the garment edge followed ratio using the proposed method and baselines. (top) The ratio of the followed garment edge over actions with 95% confidence interval shaded. (bottom) The boxplot pictures the final ratio of garment edge followed. The central dot corresponds to the median value of the errors, while the sides of the box refer to the first and third quartiles of the data. The outliers are displayed as individual circles.

data collection and cost function data annotation.

We implement our method on the Baxter robot with a RealSense L515 camera mounted for capturing visual data, and a GelSight Mini tactile sensor attached on the Robotiq 2F-85 gripper for tactile sensing. The coordinate transformation between the sensors and the robot has been determined prior to the experiments. Both the camera and tactile sensors are operated at 30Hz. We use Robot Operating System (ROS) to integrate all devices for synchronous data recording. The network takes an average of 0.11 seconds for prediction given the tactile-visual information. We use the latest fusion of observations at each time step for prediction.

IV. Experiments and Results

We conduct three experiments to investigate: 1) how the visual-tactile fusion outperforms the single sensing modality (Visual-Tactile Fusion Evaluation); 2) the effects of contrastive learning on model predictive control (Model Predictive Control Evaluation); and 3) the overall performance of the proposed method on garment unfolding and dressing compared against multiple experimental baselines (Garment Unfolding and Dressing Evaluation).

A. Experimental Setup

We have tested the proposed approach on two back-opening hospital gowns. We compare our work against nine baselines in total. For baseline 1 to baseline 5, we focus on comparing various methods for garment edge tracing. For baseline 6 and 7, we investigate different garment grasping/manipulation strategies and their influence on the dressing performance. For baseline 8 and 9, we explore multiple visual-tactile fusion network structures.

- Baseline 1: MPC but only using vision feedback, as customary in the literature [39].
- Baseline 2: MPC but only using tactile feedback. This baseline corresponds to the method used in literature [31].
- Baseline 3: Visual-tactile MPC without contrastive learning. This is equal to jointly learning a classical autoencoder together with a garment forward dynamics model in the latent space. The autoencoder is trained to minimize the L2-loss between reconstructed and actual image. This baseline and its variants have been adopted in multiple robot manipulation research, for example conditional autoencoder in [20] and sequential VAE in [40].
- Baseline 4: Behaviour cloning. The manipulation policy is directly learned from visual-tactile pixel space, as in [10].
- Baseline 5: Tactile-based LQR control. [17] has proposed to learn a linear dynamics model that finds the relationship between a cable state and pulling angle.
- Baseline 6: Using grasping-only actions for garment unfolding and dressing as in [11, 28].
- Baseline 7: Using a pre-grasp manipulation strategy for garment unfolding and dressing as in [10].
- Baseline 8: A 3D convolution-based visual-tactile fusion deep neural network as in [41].
- Baseline 9: A Transformers-based visual-tactile fusion deep neural network as in [42].

We use three metrics to evaluate performance: 1) the ratio of garment edge followed with respect to the total garment edge length; 2) garment unfolding success rate; and 3) dressing success rate. For the methods using MPC, in the testing phase, the sliding motion stops when the selected optimized cost is smaller than a threshold. Then the robot fully closes the gripper to grasp the garment. While for the

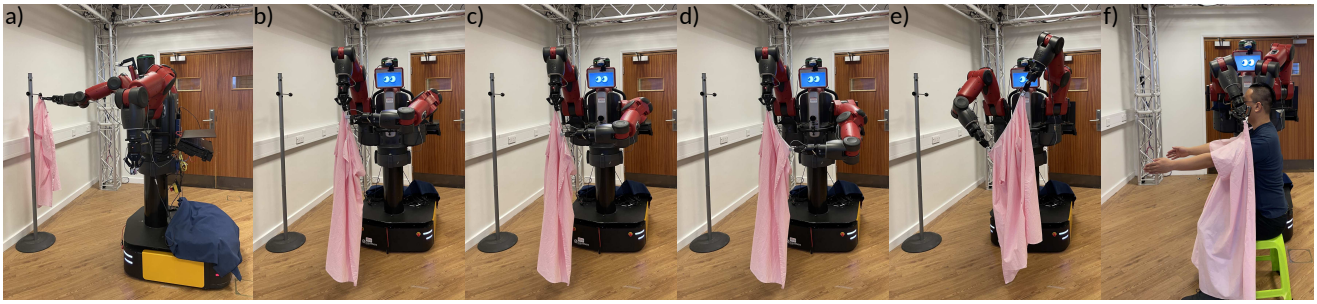

Fig. 5: Pipeline snapshots. (a) The robot first grasps a hospital gown that is naturally hung on a rail. (b) Then the robot grasps a point on the segmented edge next to the first hand. (c-d) The robot traces the edge until the corner. (e) The robot would then slide the edge to the other garment corner in the same manner. (f) The robot would pull the hospital gown through the user's arm to dress the user.

other baselines, the stopping condition is that the predicted action is smaller than a threshold. To be considered as a success, the sliding motion has to stop at the appropriate position near the garment corner with the gripper fully closed. Other circumstances, including the garment dropping from the gripper, or a premature stop (too early or too late) of edge sliding when the corner is not reached, are all regarded as failures. In the case that the robot reaches the corner but continues sliding instead of stopping, we reckon that the edge is fully followed (ratio = 1), but consider it as a failure. In each trial, we record the gripper positions in the whole process until the robot stops or the garment drops to calculate the percentage of edge followed.

B. Visual-Tactile Fusion Evaluation

We first present the ablation studies of using single sensory input. The robot has grasped the garment around the middle of the collar prior to the experiment. Then the robot slides the garment edge to reveal the corner. Fig. 3 visually presents examples of ground truth and predicted sequences of visual-tactile data in one trial of garment sliding experiment.

We compare our proposed method against baseline 1 and 2. For each baseline and our proposed method, 50 trials of garment tracing have been carried out. Fig. 4-top shows the results of garment edge followed ratio over the first 50 actions during the sliding process with 95% confidence interval shaded. Note that the sliding process might stop before or after 50 actions. If it is fewer than 50 actions, only the data before the sliding stops or the garment drops is included. Fig. 4-bottom presents the results of the garment edge followed ratio using the proposed method and baselines.

Three conclusions can be made from Fig. 4:

1) We can see that the robot behaviour changes significantly depending on the input signal. When both sensing modalities are used, the robot achieves best performance (largest median edge followed ratio: 0.89).

2) Baseline 2 (tactile only) achieves a median garment edge followed ratio of 0.44, which is around 49.4% of the results from the proposed method. Tactile data might not capture the global information that is useful for the task. We observe during the experiments that without vision, the robot tends to finish the sliding movement prematurely (too early or too late). This is expected as tactile signals might appear similar at some certain time steps and do not hold much information that allows to identify the end of the garment edge. The tactile cost value could be small during the sliding process and cause an inappropriate stop. Only the vision input allows the model to stop and hold the edge at the appropriate moment.

3) Baseline 1 (vision only) achieves a median garment edge followed ratio of 0.83. The vision-only baseline performs better than the tactile-only baseline, as the vision input allows the model to stop and grasp the edge at the appropriate moment. However, our multi-modal sensory method still outperforms baseline 1. An explanation is that the garment edge pose information, which is more certain in the case of tactile input, is still useful in learning edge tracing to interpret and take advantage of the local information.

To further evaluate the necessity of tactile sensing, we investigate the garment edge pose estimation based on visual inputs and tactile inputs respectively. Similarly to our proposed tactile-based approach, for the vision-only input, we manually label the starting and ending points of the edge on RGB images, and calculate its relevant pose with respect to the gripper through the gripper poses information obtained from ROS topics and URDF. Then a CNN network is trained to estimate the pose of garment edge. We use the labeled data on the tactile images as ground truth as the tactile sensors capture clearer local information. The proposed garment edge pose cost is used as a metric for evaluation. The results show that pose cost using the vision sensor is 1.32 times larger than when using the tactile sensor. The results are expected as the visual feedback might suffer from severe occlusions and low resolutions for the local pose information.

We can make a clear conclusion from these results that the vision plays a crucial role in finishing the task at the end of the garment edge. Tactile sensory is important for the agent to go further along the garment edge without dropping it, especially when the garment edge and corner are occluded.

C. Model Predictive Control Evaluation

In this experiment, we investigate how the contrastive loss improves the performance of model predictive control with respect to the garment edge tracing task. We benchmark the proposed method with classical conditional autoencoder without contrastive loss (baseline 3), behaviour

cloning (baseline 4) and LQR controller (baseline 5). Two conclusions can be made from this Fig. 4:

1) With our proposed method, the robot achieves best performance (largest median edge followed ratio).

2) Baseline 3 achieves a median edge followed ratio of 0.85. A classical autoencoder without contrastive learning is optimized to have pixel-level perfect reconstructions. Thus features, such as lighting and color, must be encoded in the latent space even when they are not needed for garment dynamics model learning. When used along contrastive loss, the system adds auxiliary constraints and induces the latent space to capture information that is maximally useful to predict future garment states. Thus contrastive learning contributes to creating a more robust forward dynamics model. This could also explain the results of baseline 4, 5.

D. Garment Dressing Evaluation

Lastly, we deploy all the above learned control policies (nine baselines and our proposed method) on the physical robot to perform dressing. Two human participants (i. e., the authors) are involved in the dressing task. Note that no expertise in robotics is required for this experiment, and the authors are exempt from the ethics approval. Fig. 5 describes the framework of our garment unfolding task. The robot first grasps a hospital gown that is naturally hung on a rail (Fig. 5a). The grasping point is randomly selected on the segmented garment collar near the hanging point. The segmentation is implemented using Mask R-CNN with R-101-FPN backbone pretrained weights on the COCO dataset, and finetuned on our custom data. The ground truth of garment edge is obtained through color. Then the robot grasps an point on the segmented edge next to the first hand (Fig. 5b) and trace the edge until the corner is reached (Fig. 5c-d). The robot would then slide the edge to the other garment corner (Fig. 5e) and finish the dressing task (Fig. 5f).

The dressing motion is an open-loop process which is updated based on the real-time tracked user posture through the RealSense L515 camera mounted on the top of the robot, using HRNet library [43]. The real-time detected user's hand, elbow and shoulder positions are defined as the waypoints of the dressing trajectories, which are then executed using a standard proportional-derivative (PD) controller. All intermediate stages have to be performed correctly to be considered as a success. For instance, the dressing will only be executed on the condition that the unfolding is successful. Note that most robots lack dexterous grippers to tie the gown on the back, which would be another complicated robot task.

For each learned policy, 50 trials of garment unfolding and dressing have been carried out. Note that we only run experiments of different policies when the robot slides to the first corner (Fig. 5b-d). For the sliding motion to the second corner (Fig. 5e), all trials are executed using visual-MPC only since no tactile sensor is mounted on the second gripper and the sliding distance is relatively short. Fig. 6 shows our method outperforms all other baselines. In baseline 6 and 7, both the grasping and pre-grasp manipulation strategies rely on the assumption that the grasping points are fully or

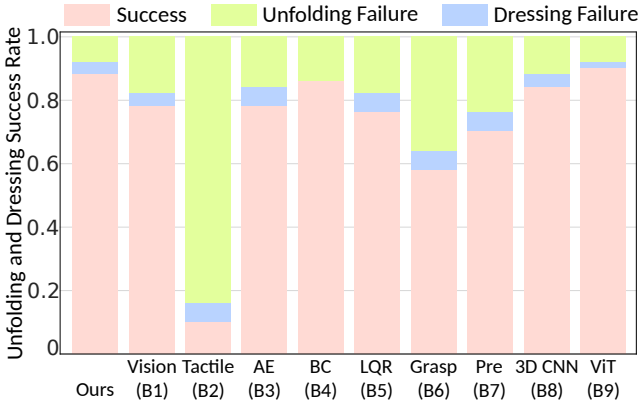

Fig. 6: Results of garment unfolding and dressing success rates.

partially visible. In our experiments, we have included some challenging setups when the grasping point is fully occluded, which are not considered in the previous work [10, 11, 28]. We also observe some dressing failures. In some cases, after the successful unfolding, the garment can tangle itself in the process of the robot approaching the user, which also makes the sleeve opening not easily accessible for the user.

We observe that similar results have been obtained between our method and baseline 8 and 9 when using different neural network structures for visual-tactile fusion. We speculate the reason is that our task may not require much spatial reasoning, which is one advantage of 3D convolution (baseline 8 [41]) and Transformers in (baseline 9 [42]).

V. Conclusions and Future Work

In this paper, we have investigated the use of visual-tactile inputs to complete the garment edge tracing task for unfolding. The tactile sensing provides rich but easy-to-interpret imprints for tracking the garment edge pose. Such local garment pose information are otherwise difficult to be captured from vision system during continuous manipulation, as they are usually occluded and expensive to interpret. The results reveal the contributions of each modality: 1) the tactile sensing contributes to preventing the garment edge from dropping out of the gripper; 2) the visual data helps the robot trace towards the end and stop accurately at the garment corner. Although adding the tactile sensor increases the complexity of the system (e. g., extra cost and time for the software and hardware), no significant difference on the time cost of data collection and network training has been observed when compared to the vision-only method. In the meanwhile, the robot behaviour is indeed boosted around 15% due to tactile sensing. Future extensions include controlling gripping force using tactile feedback to maintain it within a reasonable value for cable sliding.

A deep model predictive controller has been introduced to learn better garment latent dynamics using contrastive optimization. It is envisaged that the method could be readily generalized to held-out garment, given more collected data containing various RGB images of cloth and tactile textures, and other knowledge transfer methods.

REFERENCES

[1] T. L. Mitzner, T. L. Chen, C. C. Kemp, and W. A. Rogers, "Identifying the potential for robotics to assist older adults in different living environments," *International journal of social robotics*, vol. 6, no. 2, pp. 213–227, 2014.

[2] A. Colomé, A. Planells, and C. Torras, "A friction-model-based framework for reinforcement learning of robotic tasks in non-rigid environments," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 5649–5654.

[3] T. Matsubara, D. Shinohara, and M. Kidode, "Reinforcement learning of a motor skill for wearing a t-shirt using topology coordinates," *Advanced Robotics*, vol. 27, no. 7, pp. 513–524, 2013.

[4] E. Pignat and S. Calinon, "Learning adaptive dressing assistance from human demonstration," *Robotics and Autonomous Systems*, vol. 93, pp. 61–75, 2017.

[5] Z. Erickson, H. M. Clever, V. Gangaram, G. Turk, C. K. Liu, and C. C. Kemp, "Multidimensional capacitive sensing for robot-assisted dressing and bathing," in *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*. IEEE, 2019, pp. 224–231.

[6] F. Zhang, A. Cully, and Y. Demiris, "Personalized robot-assisted dressing using user modeling in latent spaces," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 3603–3610.

[7] ——, "Probabilistic real-time user posture tracking for personalized robot-assisted dressing," *IEEE Transactions on Robotics*, vol. 35, no. 4, pp. 873–888, 2019.

[8] F. Zhang and Y. Demiris, "Learning grasping points for garment manipulation in robot-assisted dressing," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9114–9120.

[9] N. Koganti, T. Tamei, K. Ikeda, and T. Shibata, "Bayesian nonparametric learning of cloth models for real-time state estimation," *IEEE Transactions on Robotics*, vol. 33, no. 4, pp. 916–931, 2017.

[10] F. Zhang and Y. Demiris, "Learning garment manipulation policies toward robot-assisted dressing," *Science robotics*, vol. 7, no. 65, p. eabm6010, 2022.

[11] K. Saxena and T. Shibata, "Garment recognition and grasping point detection for clothing assistance task using deep learning," in *2019 IEEE/SICE International Symposium on System Integration (SII)*. IEEE, 2019, pp. 632–637.

[12] 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.

[13] Z. Xu, C. Chi, B. Burchfiel, E. Cousineau, S. Feng, and S. Song, "Dextairity: Deformable manipulation can be a breeze," *arXiv preprint arXiv:2203.01197*, 2022.

[14] H. Yuba, S. Arnold, and K. Yamazaki, "Unfolding of a rectangular cloth from unarranged starting shapes by a dual-armed robot with a mechanism for managing recognition error and uncertainty," *Advanced Robotics*, vol. 31, no. 10, pp. 544–556, 2017.

[15] A. Gabas, Y. Kita, and E. Yoshida, "Dual edge classifier for robust cloth unfolding," *ROBOMECH Journal*, vol. 8, no. 1, pp. 1–12, 2021.

[16] I. Garcia-Camacho, M. Lippi, M. C. Welle, H. Yin, R. Antonova, A. Varava, J. Borras, C. Torras, A. Marino, G. Alenya *et al.*, "Benchmarking bimanual cloth manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1111–1118, 2020.

[17] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[18] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 4842–4849.

[19] S. Tirumala, T. Weng, D. Seita, O. Kroemer, Z. Temel, and D. Held, "Learning to singulate layers of cloth using tactile feedback," *arXiv preprint arXiv:2207.11196*, 2022.

[20] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 818–824.

[21] W. Yan, A. Vangipuram, P. Abbeel, and L. Pinto, "Learning predictive representations for deformable objects using contrastive estimation," *arXiv preprint arXiv:2003.05436*, 2020.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[23] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[24] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[25] Y. Gao, H. J. Chang, and Y. Demiris, "Iterative path optimisation for personalised dressing assistance using vision and force information," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 4398–4403.

[26] G. Canal, G. Alenyà, and C. Torras, "Adapting robot task planning to user preferences: an assistive shoe dressing example," *Autonomous Robots*, vol. 43, no. 6, pp. 1343–1356, 2019.

[27] L. M. Martínez and J. Ruiz-del Solar, "Recognition of grasp points for clothes manipulation under unconstrained conditions," in *Robot World Cup*. Springer, 2017, pp. 350–362.

[28] D. Seita, N. Jamali, M. Laskey, A. K. Tanwani, R. Berenstein, P. Baskaran, S. Iba, J. Canny, and K. Goldberg, "Deep transfer learning of pick points on fabric for robot bed-making," in *The International Symposium of Robotics Research*. Springer, 2019, pp. 275–290.

[29] 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.

[30] H.-J. Huang, X. Guo, and W. Yuan, "Understanding dynamic tactile sensing for liquid property estimation," *arXiv preprint arXiv:2205.08771*, 2022.

[31] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[32] R. Kolamuri, Z. Si, Y. Zhang, A. Agarwal, and W. Yuan, "Improving grasp stability with rotation measurement from tactile sensing," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 6809–6816.

[33] I. H. Taylor, S. Dong, and A. Rodriguez, "Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 10 781–10 787.

[34] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 754–10 761, 2022.

[35] L. Pecyna, S. Dong, and S. Luo, "Visual-tactile multimodality for following deformable linear objects using reinforcement learning," *arXiv preprint arXiv:2204.00117*, 2022.

[36] N. Sunil, S. Wang, Y. She, E. Adelson, and A. R. Garcia, "Visuotactile affordances for cloth manipulation with local control," in *6th Annual Conference on Robot Learning*.

[37] W. Zhou, S. Bajracharya, and D. Held, "Plas: Latent action space for offline reinforcement learning," *arXiv preprint arXiv:2011.07213*, 2020.

[38] R. Gieselmann and F. T. Pokorny, "Expansive latent space trees for planning from visual inputs," 2021.

[39] F. Ebert, C. Finn, S. Dasari, A. Xie, A. Lee, and S. Levine, "Visual foresight: Model-based deep reinforcement learning for vision-based robotic control," *arXiv preprint arXiv:1812.00568*, 2018.

[40] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *International conference on machine learning*. PMLR, 2019, pp. 2555–2565.

[41] S. Cui, R. Wang, J. Wei, F. Li, and S. Wang, "Grasp state assessment of deformable objects using visual-tactile fusion perception," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 538–544.

[42] Y. Chen, A. Sipos, M. Van der Merwe, and N. Fazeli, "Visuo-tactile transformers for manipulation," *arXiv preprint arXiv:2210.00121*, 2022.

[43] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 5693–5703.