Semantic State Estimation in Robot Cloth Manipulations Using Domain Adaptation from Human Demonstrations*

Georgies Tzelepis[1], Eren Erdal Aksoy[2], Júlia Borràs[1] and Guillem Alenyà[1]
[1]*Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens i Artigas 4-6, 08028 Barcelona, Spain*
[2]*Halmstad University, Center for Applied Intelligent Systems Research, Halmstad, Sweden*

Keywords: Robotic Perception, Garment Manipulation, Semantics, Cloth, Transfer Learning, Domain Adaptation.

Abstract: Deformable object manipulations, such as those involving textiles, present a significant challenge due to their high dimensionality and complexity. In this paper, we propose a solution for estimating semantic states in cloth manipulation tasks. To this end, we introduce a new, large-scale, fully-annotated RGB image dataset of semantic states featuring a diverse range of human demonstrations of various complex cloth manipulations. This effectively transforms the problem of action recognition into a classification task. We then evaluate the generalizability of our approach by employing domain adaptation techniques to transfer knowledge from human demonstrations to two distinct robotic platforms: Kinova and UR robots. Additionally, we further improve performance by utilizing a semantic state graph learned from human manipulation data.

1 INTRODUCTION

While rigid object manipulation has achieved maturity, cloth manipulation remains in its infancy due to its high complexity. Recent results mainly focused on cloth state estimation, grasp point selection, and efficient representations (Hoque et al., 2020; Lippi et al., 2020; Pumarola et al., 2018; Corona et al., 2018; Qian et al., 2020), but the high-level understanding of the cloth deformation state is still an uncharted challenge. Unlike their rigid and articulated counterparts, where the number of possible states for an object is manageable and can be semantically defined and linked to actions, identifying semantic deformation states of a textile object is a high dimensional problem that has so far been unexplored, to the best of our knowledge.

One of the core challenges in cloth manipulation is the difficulty of obtaining reliable labeled data, hindering the training of AI systems. This becomes even more challenging due to substantial performance drops of such AI systems when switching from one domain (e.g., from humans) to another one involving, for instance, robots with different embodiments running in new scenes. We tackle this challenge by extracting semantic cloth states and employing them to reduce the shift between different domains. Our method for semantic state estimation also requires labeled data, but we show how we can apply our network trained with labeled human demonstrations to different robotic domains using domain shift adaptation techniques, without requiring additional labeled data (See Fig. 1). This opens the door to learning from human demonstrations and automatic data segmentation, allowing the application of tailored techniques for each semantic state.

In the past, cloth state estimation has usually been focused on estimating the corresponding mesh (Li et al., 2018; Pumarola et al., 2018) or interesting grasping points (Seita et al., 2018; Corona et al., 2018). Instead, in this work, we utilize a high-level semantic description of the cloth state, following a framework introduced in (Garcia-Camacho et al., 2022), that includes not only semantic information on the cloth deformation types (e.g. *crumpled, flat, folded*) but also the tags representing where the cloth is grasped from (e.g. right/left corners, edge, etc.) and what contacts the cloth has with the environment.

Unlike techniques such as edge tracking (Schulman et al., 2013; Kumar et al., 2018), trajectory tracing (Lee et al., 2022) or reconstruction (Chi and Song,

*This work receives funding from the Spanish State Research Agency through the project CHLOE-GRAPH (PID2020-118649RB-l00); by MCIN/ AEI /10.13039/501100011033 and by the "European Union (EU) NextGenerationEU/PRTR under the project CO-HERENT (PCI2020-120718-2); and the EU H2020 Programme under grant agreement ERC–2016–ADG–741930 (CLOTHILDE).

Tzelepis, G., Aksoy, E., Borràs, J. and Alenyà, G.
Semantic State Estimation in Robot Cloth Manipulations Using Domain Adaptation from Human Demonstrations.
DOI: 10.5220/0012368200003660

In *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2024) - Volume 4: VISAPP*, pages 172-182
ISBN: 978-989-758-679-8; ISSN: 2184-4321

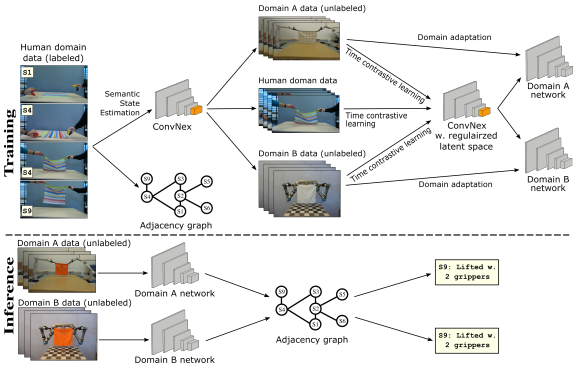


Figure 1: First, we train on fully labeled data from human demonstration, from which we also learn an adjacency graph. In the second step, we regularize the latent space with time contrastive learning in all domains and then perform domain adaptation separately on each robot domain. When doing inference, accuracy is improved by performing a post-process on the predictions using the learned graph.

2021; Bednarik et al., 2018; Bednarik et al., 2020) which incorporate low-level features and time coherence that are difficult to extract to estimate the garment deformation from some target poses in a single frame, the high-level information provided by the semantic states make them ideal for monitoring, data segmentation, and decision-making. Our framework learns to recognize both the garment deformation and its grasping state. Thus, it acts as a link between the low and high-level features.

Our framework's central purpose is knowledge transfer from human demonstrations. Deforming objects through human demonstration facilitates quicker actions with more dynamic garment interactions where robotic data collection poses challenges due to control, grasping feasibility, and manipulation speed complexities. Additionally, static environments arise from limited robot mobility. Demonstration-to-sample ratios underscore these claims: human data (33.6K RGB images, 264 demos), UR (15.5K RGB images, 20 demos), Kinova (32.7K RGB images, 20 demos).

Our contributions (summarized in Fig. 1) include:

- A solution for the semantic state estimation problem in cloth manipulation tasks, evaluating different state-of-the-art network models. Application of time contrastive learning (Sermanet et al., 2018) and domain adaptation (Ganin and Lempitsky, 2015) to generalize from human demonstration to two different robotic domains, without requiring additional labeled data.
- Learning a graph of semantic state adjacencies from our labeled data, to be further used to improve the accuracy of the domain adaptation inferences.
- A novel, fully-annotated dataset for uni- and bimanual cloth manipulation including human and robotic manipulations.

Our experimental findings show our semantic state estimation model trained with human demonstrations can recognize states in completely different robotic domains, using Kinova robotic arms and Universal Robot arms, without requiring additional labeled data. We will release our full dataset and code to encourage further research on the subject.

2 RELATED WORK

Our work is inspired by the abstract semantic representations of manipulation states done for rigid objects, based on contact interactions between the object, the hands, and the environment (Wörgötter et al., 2013), which was applied to manipulation recognition, segmentation (Aksoy et al., 2011) and robot execution tasks (Aein et al., 2019).

Recognition of semantic tags in cloth manipulation mainly consisted of identifying garment or fabric types. In (Mariolis et al., 2015), CNNs were used to first recognize the garment type and its pose. In (Kampouris et al., 2016), attributes like garment or fabric type were identified using multisensorial data. In (Ramisa et al., 2013), the cloth category could be identified using a descriptor from RGB-D data. However, these semantic tags differ from our work because they are not related to the manipulation state. The work in (Verleysen et al., 2022) could monitor the task progression in human folding tasks using time contrastive learning, where task progress steps are similar to semantic states. However, this approach has limitations when applied to tasks where the same semantic state repeats in different time frames. To the best of our knowledge, no other semantic state identification has been applied to cloth manipulation.

Research on the perception of textiles has also focused on other aspects than semantic state estimation. For instance, to detect task-oriented grasping points (Ramisa et al., 2016), to identify corners in RGB-D images (Seita et al., 2018), or to do semantic area segmentation to identify corners and edges from a cloth (Qian et al., 2020) or to estimate the mesh (Pumarola et al., 2018). Others use directly the RGB-D image as the state definition for image-based learning approaches (Hoque et al., 2020; Seita et al., 2018; Jangir et al., 2020; Lippi et al., 2020), sometimes by adding state parameters on gripper states (Matas et al., 2018) or on robotic arm joints (Yang et al., 2016). While all these methods extract visual information through a neural network,

they either focus on static garments or learn control policies for reinforcement learning.

In the context of the robotic manipulation of textiles, several datasets have been released. Human demonstrations of folding tasks have been recorded with the corresponding skeletal labels of the person performing the manipulations (Verleysen et al., 2020). Other datasets include one for tracking clothes (Schulman et al., 2013), for reconstruction purposes (Bednarik et al., 2018), or to identify relevant parts of deformables, with labels obtained from UltraViolet light (Thananjeyan et al., 2022). Except (Verleysen et al., 2020), these datasets have a limited spectrum of actions because they omitted the gripper interactions. They either focus only on one single domain or are very problem-specific, such as the detection of grasping points. Our approach, however, derives the high-level semantic scene representation that can be used to link multiple domains, e.g., human and robotic.

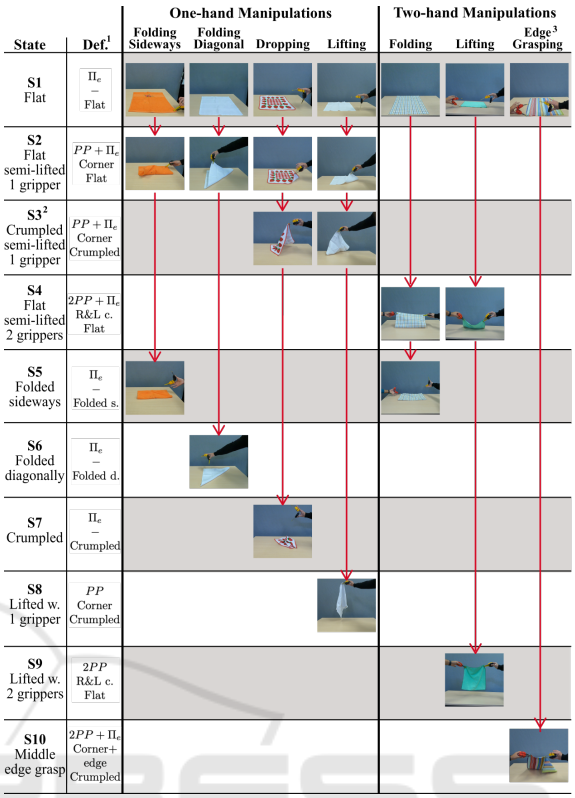
Domain adaptation has also been an active area of research since the emerged popularity and success of deep learning. One of the most popular approaches to unsupervised domain adaptation is Domain Adversarial Neural Networks (DANN) (Ganin and Lempitsky, 2015), which uses a classifier and a discriminator to align the feature representation between the source and target domains. Domain Separation Networks (DSN) (Bousmalis et al., 2016) employed domain confusion loss to encourage the model to learn domain-invariant features. Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017) further extended DANN by incorporating adversarial learning to learn domain-invariant features. Conditional Domain Adaptation Network (CDAN) (Long et al., 2018) employed a conditioned multilinear map to fully capture cross-variance between the feature representation and classifier prediction, resulting in a better alignment of the joint distributions. Finally, Batch Spectral Penalization (BSP) (Chen et al., 2019) penalizes the largest singular values to enhance the feature discriminability of the lower-rank representations.

Time contrastive learning (Sermanet et al., 2018) focuses on capturing temporal relationships and dependencies within sequential data and thus regularizing the latent space.

3 METHOD

Inspired by the work in (Garcia-Camacho et al., 2022) in which complex garment manipulations were broken down into semantic states by observing humans, our research focuses on cloth manipulations, primarily through human demonstrations that utilize a pair of two-point grippers, and later through robotic manipulations executed in new scene contexts using different platforms. Towels and kitchen towels are included in both human and robotic experiments since they share the same topology. To facilitate the evaluation we introduce two restrictions in our experiments: one refers to the cloth size which should fit the edges of the table, while the other is the initial position, which is either flat on the table or is lifted in the air by two grippers (see Fig. 1 and Table 1).

Table 1: Cloth manipulation tasks and semantic states.



1: For each state, we define the grasp type, the location of the grasps and the semantic description of deformation. Following (Borràs et al., 2020), *PP*: pinch grasp, 2*PP*: bi-manual pinch, and Π_e: the extrinsic contact with the table.
2: By crumpled we mean the cloth is deformed enough so that it cannot go back to a flat configuration without additional manipulation, as opposed to flat in the S2 state, which can be reversed to its previous state.
3: This manipulation is repeated at different distances from the corner grasp, always on the same edge.

3.1 Cloth Manipulation Tasks and Semantic States

In our framework, seven different *uni- and bi-manual cloth manipulation* tasks such as *folding diagonally* and *lifting with two grippers* are performed on eighteen different clothes. Since not all clothes are square-shaped, we consider the manipulations that involve diagonally folding only with clothes that are square.

Table 1 shows the seven manipulations considered (one in each column). Ten semantic states are required in total (rows), each of which defines a unique deformation type of the cloth. They are defined by the corresponding grasp type, location of grasp in the cloth, and deformation category (Borràs et al., 2020; Garcia-Camacho et al., 2022). The definition of the states takes inspiration from works like (Aksoy et al., 2011) where each change of contact interaction between hand, object, and environment was treated as a different scene state, however, in our case changes in the deformation category are also considered. The states are enough to represent these manipulations, but we note that other manipulations may require new states. We also follow the same principles in robotic manipulations, excluding only the semantic state *S10: middle edge grasp* in Table 1 due to grasping-related challenges.

As can be observed, each manipulation is composed of a sequence of semantic states. For instance, the uni-manual cloth manipulation *folding sideways* (shown in the first column in Table 1) involves three states: *S1: flat*, *S2: flat semi-lifted with one gripper*, and *S5: folded sideways*. On the other hand, as shown in the last column in Table 1, the bi-manual manipulation *edge grasping* has only two states: *S1: flat* and *S10: middle edge grasp*. Note that for the sake of clarity, each manipulation task in Table 1 is shown with a different textile object from our proposed dataset.

By labeling each frame under a semantic state, we essentially project the relatively complex action recognition problem into the simpler classification task. However, it is worth noting that two different time frames that fall under the same state can look very different as shown in Table 1 and Fig. 4 which makes the problem more ambiguous.

3.2 Data Collection and Annotation

We create a large-scale dataset showing various human manipulation and robotic demonstrations on different cloth types. During each demonstration, an RGB video is recorded. Each extracted frame in human demonstrations is then manually annotated with one of ten semantic states described in Table 1.

At the start of each demonstration, the cloth is placed in the initial state, i.e., lying flat on the table or bi-manually held in the air by two grippers that grasp the opposite corners. For the sake of having more natural scenes, the initial flat position of the cloth also contains slight deformations such as wrinkles.

To increase the amount of deformation, we included a total of eighteen garments with various sizes and shapes in our human demonstrations, while we used four of these garments in our robotic manipulations. Out of these garments, seven are square-shaped, ten are rectangular, and one has smoothed corners but is still square-shaped. During all these manipulation demonstrations, we kept the table, grippers, and background constant, and the RGB camera was mostly stationary with only minor adjustments introduced between manipulations.

In total, we collected 33.6K RGB images featuring 10 different semantic states (Table 1) using 18 distinct textile objects, which were demonstrated by humans. For each garment, a minimum of five or more demonstrations were performed, and the total number of demonstrations was 264.

The annotation has been done manually through human observation. To facilitate the heavy workload of data labeling, the images are annotated once they exhibit a state change. Otherwise, the remaining image frames are automatically labeled with the last adjacent state name. Some short and flickering states, which either last less than 3 frames or are not in our state list in Table 1, are omitted.

3.3 Semantic State Estimation Framework

We treat the semantic state estimation as a classification problem. For that purpose, we employ five dif-

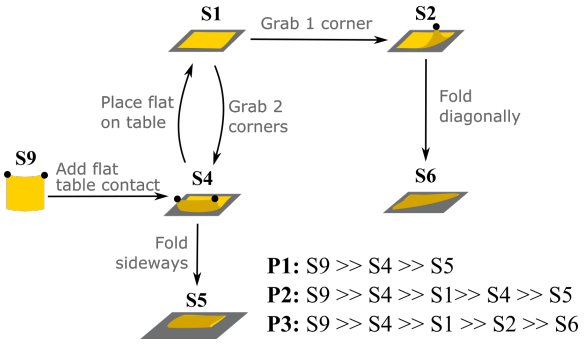


Figure 2: The three sample long manipulation scenarios are presented as a graph of the semantic states. Following the state definitions in Table 1, in these three test demonstrations, the goal states are either *S5: Folded sideways* or *S6: Folded diagonally*. All possible paths to reach the goals are shown in the bottom right corner of the figure.

ferent advanced neural network classifiers (ResNet-50 (He et al., 2016), ResNeXt-50 (Xie et al., 2017), EfficientNet (Tan and Le, 2019), ConvNext (Tan and Le, 2019), DeiT (Touvron et al., 2021)) and train them using our labeled demonstrations in the source domain.

In order to validate the generalization and the scalability of our method, we perform domain adaptation, where we transfer the semantic states learned from human demonstrations into two different robotic platforms. To do so, we record five different robotic manipulations with kitchen clothes and towels. Two of those manipulations are P1 and P2 depicted in Fig. 2, while the other three are the single gripper manipulations *Folding Diagonal*, *Dropping*, and *Lifting*. The domain adaptation is performed individually from the human domain to each of these robotic domains. Therefore, in our method, the human demonstration acts as the source domain, while the robotic manipulations are the target domain.

Given the nature of video data, where time coherence can be learned, prior to commencing the domain adaptation training process, a preliminary step involving time contrastive learning is undertaken to impart regularization to the latent space. This strategic approach is adopted to fine-tune the latent space across both domains while preserving temporal awareness and thus reducing the instability of domain adaptation. We perform time contrastive learning on each manipulation individually by minimizing the time contrastive loss as follows:

$$L_{tc} = max(\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha, 0) \ ,$$

where x_i^a, x_i^p and x_i^n are the anchor, positive and negative samples. The hyperparameter α denotes the *margin* between positive and negative pairs. Notably, its value is domain-dependent, as it is subject to variation across different domains arising from dissimilar manipulation speed characteristics in each domain.

The selection of hyperparameters in the context of both robotic domains depends on the choice of the human domain. For the manipulation task denoted as P1 in Fig. 2, the hyperparameter α is assigned to be two times the positive range, which is 0.2 seconds. Subsequently, the corresponding value of α is adjusted within each respective robotic domain. This adjustment process involves estimating the equivalent value based on the duration of the human demonstration. In practice, the duration of robotic demonstrations plays a crucial role. For instance, if the duration of a robotic demonstration is longer in comparison to a human demonstration, the associated value of α is increased according to that difference in durations. This approach seeks to maintain a consistent relation between the manipulation speeds across domains.

Once the training is completed with known labels in the source domain, we obtain the adjacency matrix $\mathscr{A}$ for the neighboring state graph. For example, the state $S1$ is a neighboring state with $S2$, but not with $S3$ since such a state transition to $S3$ is not possible without visiting $S2$ (see Table 1). We then use the trained network predictions and adjacency matrix $\mathscr{A}$ to determine whether a prediction at time $t+1$ is feasible from time t. Starting at time t, we designate the current frame as an anchor. If the prediction for the frame at time $t+1$ is feasible, we designate it as the new anchor and continue. However, in case the prediction is not feasible according to $\mathscr{A}$, we provide a distance threshold T and examine whether other predictions within that certain distance from the initial prediction are feasible. If there are feasible predictions within the predefined threshold, we choose the prediction that is the closest to the current anchor. Otherwise, we keep the prediction and the anchor remains unchanged. This process allows us to autonomously identify and eliminate wrong predictions in an unsupervised manner. It is worth noting that the graph is generated automatically from the training data.

4 EXPERIMENTS

We evaluate the effectiveness of semantic state estimation networks in two different scenarios which involve monitoring of human and robot manipulations. In both cases, we use accuracy scores as the primary evaluation metric to measure the correct classification rate. For the human demonstrations, we employ supervised training, whereas, for the robotic manipulation, the training is completely unsupervised, as the aim is for the manipulations to be learned from annotated human demonstrations.

4.1 Human Demonstrations for Training

In section 3.3, we trained five networks by partitioning our annotated dataset of 33.6K samples using a 75-25 stratified split. The optimization method of our choice is stochastic gradient descent with warm

Table 2: Quantitative evaluation.

Networks	Validation Scores	Test Scores White	Orange	Grid	Average
ResNet-50 (He et al., 2016)	97.77	96.27	91.79	95.54	94.53
ResNeXt-50 (Xie et al., 2017)	97.50	95.67	92.30	95.51	94.49
EfficientNet (Tan and Le, 2019)	97.71	95.53	93.16	96.61	95.10
DeiT (Touvron et al., 2021)	98.20	93.85	86.18	91.60	90.54
ConvNext (Tan and Le, 2019)	**98.45**	**96.92**	**94.16**	**97.02**	**96.03**

restarts (Loshchilov and Hutter, 2016). We finetune the network for 30 epochs with batch size 32 and apply random cropping, flipping, and 15-degree rotation during data augmentation. We wanted to ensure that the trained networks were not biased towards any particular type of cloth, therefore, we excluded all manipulations involving certain cloth types from the training data and reserved them only for testing purposes. For instance, all demonstrations featuring the cloth *Orange* are reserved as unseen test data, while the remaining data is utilized for training. We apply the same leave-one-out testing practice for the garments *White* and *Grid*. Table 2 shows accuracy scores in percentage (%) for the validation and individual test cases with these three clothes. As shown in Table 2, ConvNext (Liu et al., 2022) performs the best on average in contrast to the other three networks. Having a minor difference between the validation and average test scores confirms that ConvNext is not biased with the cloth types in the training data.

4.2 Domain Adaptation for Robotic Manipulations

We generated data in two different domains by performing *Folding Diagonal*, *Dropping*, and *Lifting* manipulations, as outlined in Table 1 columns, using both two Kinova and two Universal robot arms. These robots act as the target domains for our domain adaptation network. Additionally, we executed two long test scenarios (P1 and P2) described in Fig. 2 using either of the same two robot arms. For instance, the first scenario represents *folding a cloth sideways* by following a path (P1 in Fig. 2) with three states: *S9, S4* and *S5*, whereas the second scenario follows P2 with five states to reach the same goal: *S9, S4, S1, S4* and *S5*. The Universal Robot manipulation involved four different garments (*White*, *Orange*, *Green*, and *Pink*), while for the Kinova manipulator, we added the *Grid* garment as well. A total of $48.2K$ frames were collected from the robot executions, with about $12K$ frames used for testing the garments. Out of the $48,2K$ frames, $15.5K$ were obtained from the UR manipulations and $32.7K$ from the Kinova robot.

Before initiating the train of domain adaptation, we perform time contrastive learning to regularize the latent space. We train the time contrastive loss with Adam (Kingma and Ba, 2014) with initial learning rate at $5e^{-4}$ decreased to $5e^{-5}$. Then we choose for the positive range to be 2 in the human domain, 4 in the UR, and 6 for the Kinova, while the negative samples are chosen outside that range. We perform time contrastive learning for 40 epochs.

As a source domain, we used the human demonstration data that were described in section 4.1, excluding the *Edge Grasping* manipulation. Additionally, we included three long human demonstrations (P1, P2, and P3) described in Fig. 2. For the target domain, like in section 4.1 we applied the same policy by keeping one garment out as the unseen test case and employing the rest to train the discriminator. For our testing, we selected squared-shaped garments, specifically the *Green*, *Orange*, and *Pink* ones, as they allow for diagonal folding. Our models were trained using a DANN architecture (Ganin and Lempitsky, 2015) with a BSP loss (Chen et al., 2019) and combined with a pre-trained ConvNext feature extractor from Imagenet. Since the same number of classes are shared amongst all domains, we opted to use the combination DANN-BSP because we obtained the best results amongst other domain adaptation methods such as CDAN (Chen et al., 2019) and MCD (Saito et al., 2018). We also used Adam with pre-warmed-up initial rate for training the classification branch of DANN for 3 epochs which is decreased to $5e^{-5}$. Due to the instability associated with domain adaptation methods, we trained our network 5 times for 50 epochs and reported the best score obtained for each test scenario in Table 3. This table shows that there is no significant bias among the tested garments, as they all show similar performance. Table 3 also indicates that our proposed post-process (PP) using the adjacency graph improves the network predictions. Note also that the mean accuracy and standard deviation scores in Table 3 increase over 5 training sessions. Table 4 shows that using time contrastive learning on the latent space before performing domain adaptation contributes to stability. Note that Table 4 reports the average scores over 5 training sessions, unlike Table 3 which reveals the best scores obtained.

Figure 3 shows the confusion matrices reporting the network predictions for Kinova and UR robot manipulations. The diagonal of each matrix represents the per-class prediction. We obtain high classification accuracy for each state except one for each robot domain. For instance, *S*6 is mainly mixed with *S*2 in UR robot manipulations, although the average accuracy

Table 3: Quantitative evaluation of domain adaptation. The term *w. PP* stands for "with post-process" using the adjacency graph. *PP. M. STD* stands for "post-process mean and standard deviation" and denotes the mean increase and its standard deviation over 5 training sessions.

	Garments			
	Green	Orange	Pink	Average
Kinova	65.25	72.75	70.05	69.40
w. PP	**66.60**	**73.35**	**71.95**	**70.65**
PP. M. STD	1.28 ± 0.29	0.98 ± 0.14	1.32 ± 0.34	1.19 ± 0.26
UR	80.15	82.25	81.50	81.30
w. PP	**80.45**	**83.65**	**82.45**	**82.20**
PP. M. STD	0.62 ± 0.27	1.91 ± 0.21	1.05 ± 0.17	1.19 ± 0.21

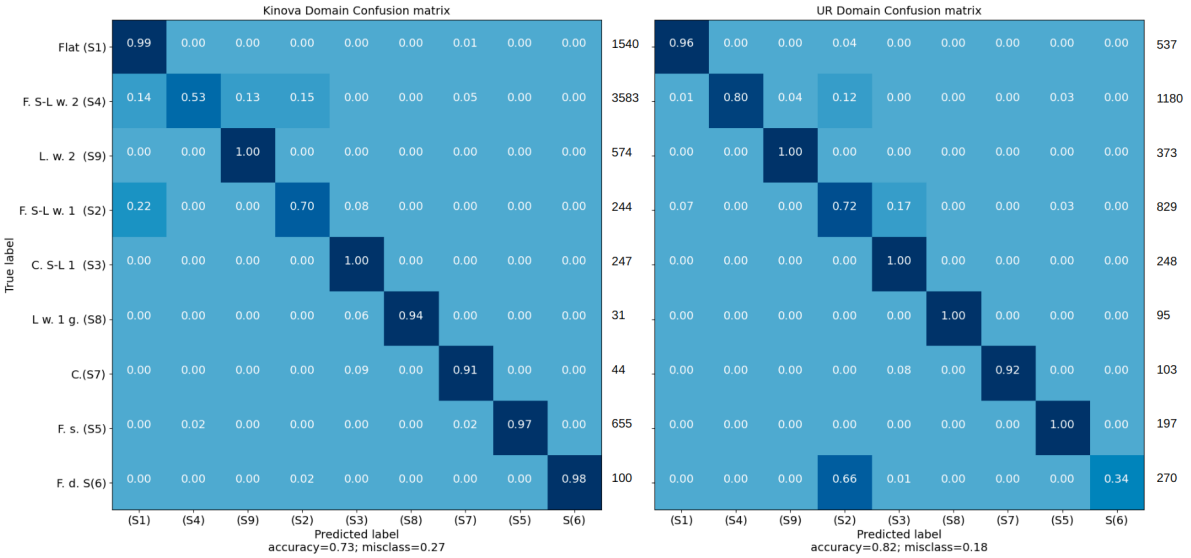


Figure 3: The confusion matrices for the network predictions in both the Kinova (on the left) and UR (on the right) domains. The diagonal of each matrix represents the per-class prediction, while the number of samples is indicated on the right of each matrix. The states are listed in the y-axis from top to bottom as follows: Flat (S1), Flat semi-lifted 2 grippers (S4), Lifted w. 2 grippers (S9), Flat semi-lifted 1 gripper (S2), Crumpled semi-lifted 1 gripper (S3), Lifted w. 1 gripper (S8), Crumpled (S7), Folded sideways (S5), and Folded diagonally (S6).

still reaches up to 82%.

Figure 4 shows the network performance for human and robotic demonstrations of the same manipulation scenario P1 described in Fig. 2. The red frames depict incorrect predictions, which emerged, for instance, before and after applying our knowledge from the graph adjacency matrix, when the garment just switched from states *S9:lifted with two grippers* to *S4:flat semi-lifted with two grippers*. The blue frames depict incorrect predictions only before the use of post-processing. The colored blocks in Fig. 4 clearly show that such false positive predictions are borderline cases where the state is either about to change or just switched to the next. It is worth mentioning that even before applying the post-processing, predictions which are incorrect are very similar to other states (e.g. the state *S4:Semi-lifted w. 2 grippers* shares similarities with the borderline case transition from *S1:Flat* to *S2:Semi-lifted w. 1 gripper* when the grippers are very close to the table).

We make use of the adjacency matrix information in the graph by applying an empirical threshold of 0.05. However, since each prediction at time $t+1$ is dependent only on the anchor or the prediction at time t, we reset the anchor every 10 frames as the most probable prediction. This is done to avoid carrying the information of incorrect predictions and positions on the graph nodes for an extended period.

Table 4: Ablation study of time contrastive learning. Each number represents the average five training sessions. The term TCL stands for "time contrastive learning".

	Garments			
	Green	Orange	Pink	Average
Kinova	53.10	57.25	54.95	55.10
w. TCL	**58.25**	**65.75**	**63.30**	**62.45**
UR	72.60	74.85	73.30	73.60
w. TCL	**74.35**	**79.70**	**76.10**	**76.70**

5 DISCUSSION

We first would like to highlight the fact that in this work, we do not propose any novel network model but rather a novel framework transferring semantic state estimations learned from human demonstrations to robot manipulation tasks without requiring any additional data annotation effort in the target robot domain. We show that the state-of-the-art models (e.g., ConvNext (Liu et al., 2022)) can already handle the challenging state estimation problem. The reported high accuracy scores in Table 2, Fig. 3, and Fig.4 already show that there is no need to focus on designing new deep network architectures. Instead, we address the domain shift problem where we use the rich information of the human-labeled dataset to make inferences in two distinct robotic domains.

During domain adaptation, low per-class accuracies are observed in the states where the garment is *Diagonally Folded*. For instance, in Fig. 3 *S*6 is confused with *S*2 at the end of the manipulation with the

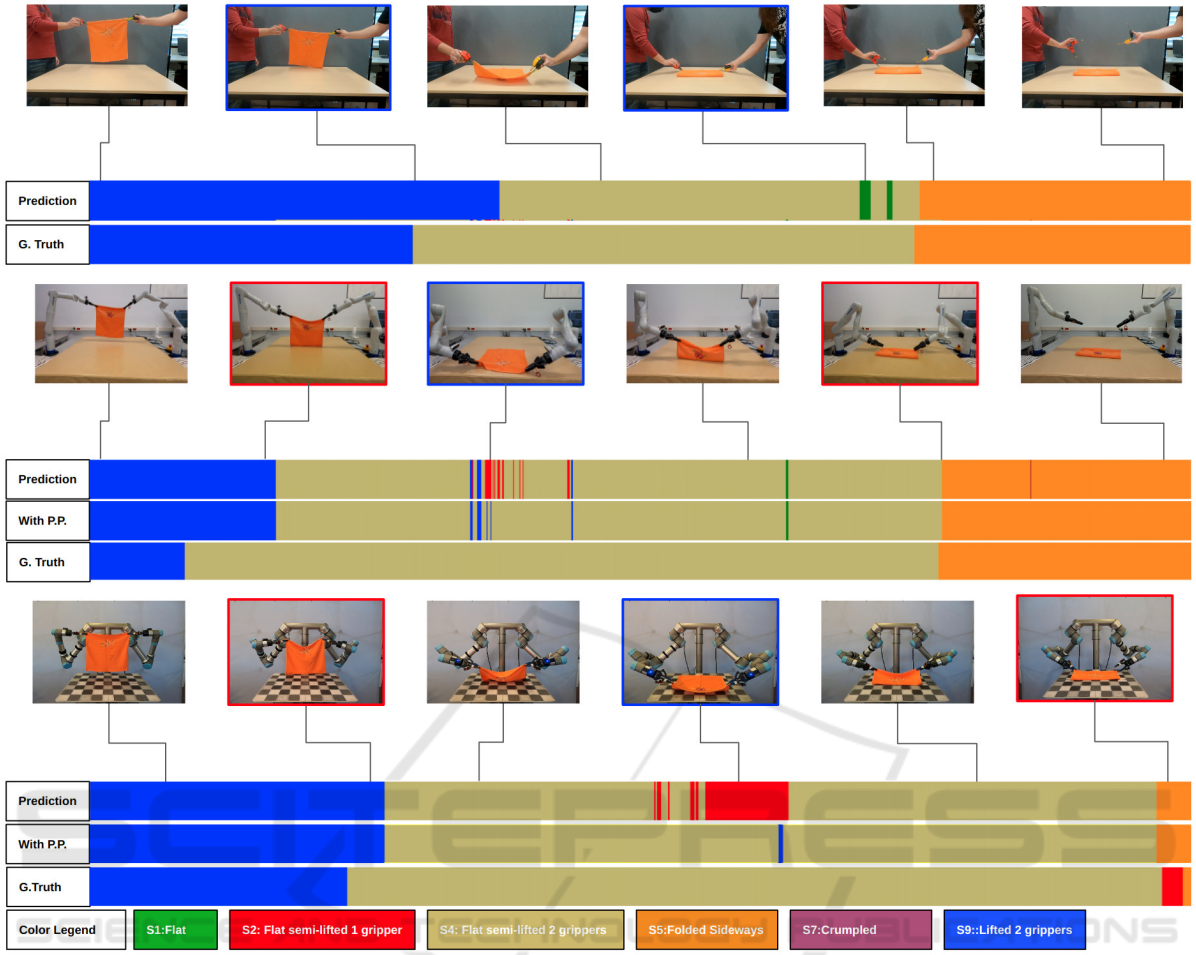


Figure 4: The event plots display the manipulation scenario P1, as described in Fig. 2. On the top, we show the network performance for the human demonstration. Next, prediction results are illustrated for both the Kinova and UR robot manipulators. The colored blocks in both cases represent the predictions made by the network, the predictions after post-processing (PP), and the human-labeled ground truth. The red frames around images indicate false positive predictions before and after post-processing, while the blue image frames represent predictions that are false positives only before post-processing. Note that the post-processing is applied to the target robot domains only.

UR manipulator. *S*2 appears because one gripper releases a corner before the other by just some frames, but it may not be observable if the gripper is opened or not. This leads to false detection of *S*2, probably due to the fact that the network is placing too much emphasis on the position of the gripper, rather than the shape of the garment, particularly when it is not visible whether the gripper is open or closed.

It is evident from Fig. 4 that the network's wrong predictions are mostly due to borderline cases where two consecutive states are very similar to each other. We here note that false state estimates, which are irrelevant to borderline cases, can easily be solved by incorporating temporal state information during the monitoring task. As our network is focused on single-frame predictions, we do not utilize information on optical flow, which limits our ability to capture temporal coherence outside the post-processing stage. This leads to difficulties during domain adaptation, for example, in manipulation scenario P2 the network is unable to differentiate between the state of *S4: Semi-lifted w. 2 grippers* when it occurs due to lowering the garment versus lifting it.

From our experiments, we observe that domain adaptation in dynamic cloth manipulation is highly unstable. Thus, we train the network on average 5 times to achieve optimal performance. The predictions are, however, improved when more target domain data are introduced.

Furthermore, we pose the following questions to better understand the performance of our domain adaptation framework:

Does the Post-Prediction Processing Improve the Performance? In Table 3 and in Fig. 4 it is shown

that introducing prior knowledge from the graph adjacency matrix significantly improves predictions in every test case. Additional information (such as integrating the knowledge of manipulation success or failure, or incorporating information from a predefined plan) can further be utilized to improve the performance.

What Is the Cause of the High Instability? The instability associated with domain adaptation in our framework poses a challenge for investigating improvements in the loss function. The results show that the stability is increased by regularizing the latent space with time contrastive learning. However, it is difficult to ascertain whether any improvement results from a better loss function or a randomly selected initial weight before training. This instability may limit our ability to make meaningful advancements in the domain adaptation process.

How Can We Incorporate Temporal Coherence During Training? Our approach to the semantic state estimation problem is based on treating it as a classification task without incorporating temporal coherence during training. However, the non-linear transition sequence between states in the P2 scenario depicted in Fig. 2, where we return to the same state again, highlights the limitations of techniques such as time contrastive learning that rely solely on single frames. We plan to investigate recurrent networks to capture temporal information.

How Do We Deal with Cases Where the Cloth Is in an Undefined State? Out-of-distribution classes are difficult to handle in classification problems. In our framework, when a new class is introduced during learning, a human demonstration of that specific class is required before the adaptation step is triggered in the robot domain. However, it is not necessary to demonstrate an entire manipulation sequence, only the newly introduced class. Another method of dealing with such cases is stochastic neural networks, where the uncertainty can be quantified and acted upon.

6 CONCLUSION

In this paper, we presented and evaluated a novel framework to solve the problem of semantic state estimation in continuous cloth manipulation tasks. We make use of a convenient high-level semantic description of the cloth state which couples the cloth deformation type, the grasping state, and the contact with the environment. To validate our approach, we benchmarked five different networks on our new dataset coming with 33.6K annotated RGB images.

Furthermore, we show that our approach can be used to learn a representation using labeled human demonstrations, which can be further exploited to predict the semantic states in robotic manipulation tasks in an unsupervised manner. This domain adaptation is evaluated in two distinct robotic domains (Kinova and UR5) and evaluated on unseen garments. To further boost the prediction accuracy, we re-evaluated predictions using the graph adjacency matrix learned from the training data.

In future work, we plan to enlarge our dataset by introducing a higher variety of deformable shapes (semantic states) and more complex manipulation tasks. Furthermore, we would like to incorporate the depth cue which can capture geometrical information in the scene, and thus, play a crucial role in autonomously defining unseen textile deformations and planning a proper grasping accordingly. We hope that the here presented approach will be adopted by the cloth manipulation community and trigger further contributions to robotics.

REFERENCES

Aein, M. J., Aksoy, E. E., and Wörgötter, F. (2019). Library of actions: Implementing a generic robot execution framework by using manipulation action semantics. *The Int. Journal of Robotics Research*, 38(8):910–934.

Aksoy, E. E., Abramov, A., Dörr, J., Ning, K., Dellen, B., and Wörgötter, F. (2011). Learning the semantics of object–action relations by observation. *The Int. Journal of Robotics Research*, 30(10):1229–1249.

Bednarik, J., Fua, P., and Salzmann, M. (2018). Learning to reconstruct texture-less deformable surfaces from a single view. In *Int. Conf. on 3d vision (3DV)*, pages 606–615.

Bednarik, J., Parashar, S., Gundogdu, E., Salzmann, M., and Fua, P. (2020). Shape reconstruction by learning differentiable surface representations. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 4716–4725.

Borràs, J., Alenyà, G., and Torras, C. (2020). A grasping-centered analysis for cloth manipulation. *IEEE Transactions on Robotics*, 36(3):924–936.

Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., and Erhan, D. (2016). Domain separation networks. *Advances in neural information processing systems*, 29.

Chen, X., Wang, S., Long, M., and Wang, J. (2019). Transferability vs. discriminability: Batch spectral penalization for adversarial domain adaptation. In *Int. Conf. on Machine Learning*, pages 1081–1090.

Chi, C. and Song, S. (2021). Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In *IEEE/CVF Int. Conf. on Computer Vision*, pages 3324–3333.

Corona, E., Alenya, G., Gabas, A., and Torras, C. (2018). Active garment recognition and target grasping point detection using deep learning. *Pattern Recognition*, 74:629–641.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In *Int. Conf. on Machine Learning*, pages 1180–1189.

Garcia-Camacho, I., Borràs, J., and Alenyà, G. (2022). Knowledge representation to enable high-level planning in cloth manipulation tasks. In *ICAPS Workshop on Knowledge Engineering for Planning and Scheduling*.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 770–778.

Hoque, R., Seita, D., Balakrishna, A., Ganapathi, A., Tanwani, A. K., Jamali, N., Yamane, K., Iba, S., and Goldberg, K. (2020). Visuospatial foresight for multi-step, multi-task fabric manipulation. In *Robotics: Science and Systems*.

Jangir, R., Alenyà, G., and Torras, C. (2020). Dynamic cloth manipulation with deep reinforcement learning. In *IEEE Int. Conf. on Robotics and Automation*, pages 4630–4636.

Kampouris, C., Mariolis, I., Peleka, G., Skartados, E., Kargakos, A., Triantafyllou, D., and Malassiotis, S. (2016). Multi-sensorial and explorative recognition of garments and their material properties in unconstrained environment. In *IEEE Int. Conf. on Robotics and Automation*, pages 1656–1663.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kumar, S., Cherian, A., Dai, Y., and Li, H. (2018). Scalable dense non-rigid structure-from-motion: A grassmannian perspective. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 254–263.

Lee, R., Abou-Chakra, J., Zhang, F., and Corke, P. (2022). Learning fabric manipulation in the real world with human videos. *arXiv preprint arXiv:2211.02832*.

Li, Y., Wang, Y., Yue, Y., Xu, D., Case, M., Chang, S.-F., Grinspun, E., and Allen, P. K. (2018). Model-driven feedforward prediction for manipulation of deformable objects. *IEEE Transactions on Automation Science and Engineering*, 15(4):1621–1638.

Lippi, M., Poklukar, P., Welle, M. C., Varava, A., Yin, H., Marino, A., and Kragic, D. (2020). Latent space roadmap for visual action planning of deformable and rigid object manipulation. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 5619–5626.

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 11976–11986.

Long, M., Cao, Z., Wang, J., and Jordan, M. I. (2018). Conditional adversarial domain adaptation. *Advances in neural information processing systems*, 31.

Loshchilov, I. and Hutter, F. (2016). Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Mariolis, I., Peleka, G., Kargakos, A., and Malassiotis, S. (2015). Pose and category recognition of highly deformable objects using deep learning. In *Int. Conf. on advanced robotics*, pages 655–662.

Matas, J., James, S., and Davison, A. J. (2018). Sim-to-real reinforcement learning for deformable object manipulation. In *Conf. on Robot Learning*, pages 734–743. PMLR.

Pumarola, A., Agudo, A., Porzi, L., Sanfeliu, A., Lepetit, V., and Moreno-Noguer, F. (2018). Geometry-aware network for non-rigid shape prediction from a single view. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4681–4690.

Qian, J., Weng, T., Zhang, L., Okorn, B., and Held, D. (2020). Cloth region segmentation for robust grasp selection. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 9553–9560.

Ramisa, A., Alenya, G., Moreno-Noguer, F., and Torras, C. (2013). Finddd: A fast 3d descriptor to characterize textiles for robot manipulation. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 824–830.

Ramisa, A., Alenya, G., Moreno-Noguer, F., and Torras, C. (2016). A 3d descriptor to detect task-oriented grasping points in clothing. *Pattern Recognition*, 60:936–948.

Saito, K., Watanabe, K., Ushiku, Y., and Harada, T. (2018). Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732.

Schulman, J., Lee, A., Ho, J., and Abbeel, P. (2013). Tracking deformable objects with point clouds. In *IEEE Int. Conf. on Robotics and Automation*, pages 1130–1137.

Seita, D., Jamali, N., Laskey, M., Tanwani, A. K., Berenstein, R., Baskaran, P., Iba, S., Canny, J., and Goldberg, K. (2018). Deep transfer learning of pick points on fabric for robot bed-making. In *Robotics Research: The 19th Int. Symposium ISRR*.

Sermanet, P., Lynch, C., Chebotar, Y., Hsu, J., Jang, E., Schaal, S., Levine, S., and Brain, G. (2018). Time-contrastive networks: Self-supervised learning from video. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 1134–1141. IEEE.

Tan, M. and Le, Q. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. on Machine Learning*, pages 6105–6114.

Thananjeyan, B., Kerr, J., Huang, H., Gonzalez, J. E., and Goldberg, K. (2022). All you need is LUV: Unsupervised collection of labeled images using uv-fluorescent markings. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 3241–3248.

Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. In *Int. Conf. on Machine Learning*, pages 10347–10357.

Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In

IEEE Conf. on Computer Vision and Pattern Recognition, pages 7167–7176.

Verleysen, A., Biondina, M., and Wyffels, F. (2020). Video dataset of human demonstrations of folding clothing for robotic folding. *The Int. Journal of Robotics Research*, 39(9):1031–1036.

Verleysen, A., Biondina, M., and Wyffels, F. (2022). Learning self-supervised task progression metrics: a case of cloth folding. *Applied Intelligence*, pages 1–19.

Wörgötter, F., Aksoy, E. E., Krüger, N., Piater, J., Ude, A., and Tamosiunaite, M. (2013). A simple ontology of manipulation actions based on hand-object relations. *IEEE Transactions on Autonomous Mental Development*, 5(2):117–134.

Xie, S., Girshick, R., Dollár, P., Tu, Z., and He, K. (2017). Aggregated residual transformations for deep neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1492–1500.

Yang, P.-C., Sasaki, K., Suzuki, K., Kase, K., Sugano, S., and Ogata, T. (2016). Repeatable folding task by humanoid robot worker using deep learning. *IEEE Robotics and Automation Letters*, 2(2):397–403.