Semantic State Estimation in Cloth Manipulation Tasks

Georgies Tzelepis\textsuperscript{1}, Eren Erdal Aksoy\textsuperscript{2}, Júlia Borràs\textsuperscript{1}, and Guillem Alenyà\textsuperscript{1}

Abstract—Understanding of deformable object manipulations such as textiles is a challenge due to the complexity and high dimensionality of the problem. Particularly, the lack of a generic representation of semantic states (e.g., crumpled, diagonally folded) during a continuous manipulation process introduces an obstacle to identify the manipulation type. In this paper, we aim to solve the problem of semantic state estimation in cloth manipulation tasks. For this purpose, we introduce a new large-scale fully-annotated RGB image dataset showing various human demonstrations of different complicated cloth manipulations. We provide a set of baseline deep networks and benchmark them on the problem of semantic state estimation using our proposed dataset. Furthermore, we investigate the scalability of our semantic state estimation framework in robot monitoring tasks of long and complex cloth manipulations. We also release the dataset and source code of the baseline models: https://github.com/BOBaraki/state-estimation-for-continuous-cloth-manipulation

I. INTRODUCTION

Textiles are an important part of our daily living objects both in domestic, public health, and industrial scenarios. While rigid object manipulation has achieved maturity, cloth manipulation remains in its infancy due to its high complexity, and only recently it is becoming a very active research topic. Recent results include novel manipulation solutions, mainly focused on cloth state estimation, grasp point selection, and efficient representations [1–9], but the high-level understanding of 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. In robotics, recognizing the semantic state of a textile in a continuous manipulation is essential for the subsequent tasks, such as monitoring, task planning, learning from human demonstrations, and action execution.

In this paper, we classify semantic cloth states of rectangular cloths during the course of a continuous manipulation task (see Fig. 1). Cloth state estimation has been usually focused in the past to estimate the deformation state, and more in particular, the corresponding mesh [4, 10] or interesting grasping points [5, 7]. In this work, we introduce a high-level semantic description of the cloth state that includes information on not only the deformation state but also the grasping state and the contacts with the environment. We use the definition of the grasp type introduced in [11], which describes textile grasps based on the geometry of the prehension contacts, that can be either grippers or environmental objects that interact with the manipulation, such as the table where the manipulation takes place. In addition, our semantic cloth states include not only abstract cloth deformation types (e.g. crumpled, flat, folded), but also tags representing where the cloth is grasped from (e.g. right/left corners, edge, etc.). This allows us to define a sequence of semantic states during a continuous manipulation task. By being able to autonomously identify such states opens the door to learning plans of manipulations from human demonstrations, monitoring tasks, and closing the loop of a high-level planing.

To investigate the scalability of our semantic state concept, we collected a new large scale RGB image dataset captured during human demonstrations of different uni- and bi-manual
cloth manipulation tasks such as folding diagonally and lifting with two grippers. Each captured frame is annotated by one of ten semantic states described in Table I. To solve the state estimation problem, we employed state-of-the-art network models (e.g. EfficientNet [12] and DeiT [13]) pre-trained on Imagenet [14]. Different neural networks were used for the task with the aim of providing an initial baseline on the dataset. To show the generalization of our framework, we further recorded relatively long and complex cloth manipulation tasks performed both by humans and by two Kinova robot arms, where we can monitor the manipulations in totally new scene contexts.

To summarize, our contributions are threefold:

• We introduce a novel fully-annotated, uni- and bi-manual cloth manipulation dataset with 33.6K RGB images involving 10 different semantic states and using 18 different textile objects.
• We propose a cloth state estimation benchmark on this dataset and provide baseline experiments using state-of-the-art neural networks.
• We perform experimental evaluations showing that our state estimation framework can be used for monitoring robot and human demonstrations in new scene contexts.

II. RELATED WORK
A. Cloth Manipulation

Research on the perception of textile objects and their grasping points has been done for 2D and 3D data. For instance, by detecting task-oriented grasping points of the collar of a shirt with the use of a 3D descriptor, simple hanging tasks have been performed [6]. In [7], corners of the cloth are detected in RGB-D images to perform the folding tasks. To avoid multiple grasping strategies, active search with the use of different neural networks has been employed to recognize two grasping points [5]. Finally, semantic area segmentation and domain adaptation were used to identify grasping points from a single shot image with the help of synthetic data [15]. Significant progress has been made on tasks such as folding, unfolding, and spreading by learning policies in a simulated environment and then transferring them into real world manipulators [8, 16], or already performing all these tasks in simulation [17].

In the context of image-based learning approaches, various methods have been introduced, which particularly rely on the Euclidean distance between pixels to identify equal states [2, 3, 7, 18, 19]. More similar to our work, the gripper states [9] or the robotic arm joints [20] were jointly used with the cloth information. These works, however, omit the interaction between the gripper and the cloth.

Persistent homology was used to extract topological features of deformables with nontrivial topology in a simulated environment [21]. Other works rather focused on understanding the deformation by reconstructing the scene from single shot images [4, 22] or from the detected edges [23].

Along the same lines with our here presented work, abstract semantic representations of states in a manipulation task, based on contact interactions between the object, the hands, and the environment have also been introduced in the past in [24] using rigid objects, with the application to the manipulation recognition, segmentation [25] and robot execution tasks [26].

B. Deformable Object datasets

In the context of deformable objects and more particularly textiles, several attempts have been made to create various datasets. Large-scale static cloth image datasets were introduced for the category classification tasks [27–29], which have also been extended by introducing landmarks to enhance the classification performance [30].

Human demonstrations of folding tasks have been recorded with the corresponding skeletal labels of the person performing the manipulation. These works focused on action recognition rather than state estimation [31]. Other datasets for manipulation tasks had a limited spectrum of actions, i.e., they omitted the gripper interactions [32]. Category estimation and cloth part segmentation have also been performed jointly by employing images coupled with grasping point descriptors [6, 33].

There also exist several datasets addressing the state estimation problem. Synthetic data of hanging garments [34] from multi-view points generated at large scale was used for convolutional neural networks. A different approach generated a mesh of various 3D deformable objects by energy minimization from RGBD images [35]. Unlike ours, these datasets are, however, not available for public use and thus limit their applications in different use-cases.

III. METHOD

Following the work in [36], we focus on human demonstrations which are mainly performed with two point grippers, each is handled by a different subject to grasp and manipulate a cloth simultaneously (see Fig. 1). The deformable object manipulated in our experiments is either a towel or a kitchen cloth with various patterns, sizes, and stiffness. The cloth size is the only restriction that is imposed, since we assume that the cloth should fit within the table edges while being completely open and flat as shown in the first row in Table I. The position of the cloth is usually placed to the center of the table, unless the action is performed solely with one gripper. In this uni-manual grasping case, the cloth is placed closer to the human demonstrator holding the gripper to ensure that the cloth is within the gripper’s reach.

A. Cloth Manipulation Tasks and Semantic States

Seven different uni- and bi-manual cloth manipulation tasks such as folding diagonally and lifting with two grippers are performed on eighteen different cloths. Since not all cloths are square shaped, we consider the manipulations that involve diagonally folding only with cloths that are square. Table I shows in each column one of the seven manipulation types. There are also in total ten semantic states, each of which defines a unique deformation type of the cloth, as introduced in the grasping-centered framework in [11, 36].
TABLE I: Manipulation tasks and semantic states.

| State          | DoC | One-hand Manipulations                                                                 | Two-hand Manipulations                                                                 |
|----------------|-----|----------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|
| S1: Flat       | π_ε | Folding: Sideways, Folding: Diagonal, Dropping, Lifting                               | Folding, Lifting, Edge                                                                 |
| S2: Flat semi-lifted 1 gripper | π_ε | PP: Bi-manual pinch, π_ε: Extrinsic contact with the table                               |                                                                                         |
| S3: Crumpled semi-lifted 1 gripper | π_ε | PP: Bi-manual pinch, π_ε: Extrinsic contact with the table                               |                                                                                         |
| S4: Flat semi-lifted 2 grippers | π_ε | DPP: Bi-manual pinch, π_ε: Extrinsic contact with the table                               |                                                                                         |
| S5: Folded sideways | π_ε |                                                                                         |                                                                                         |
| S6: Folded diagonally | π_ε |                                                                                         |                                                                                         |
| S7: Crumpled   | π_ε |                                                                                         |                                                                                         |
| S8: Lifted w. 1 gripper | π_ε |                                                                                         |                                                                                         |
| S9: Lifted w. 2 grippers | π_ε |                                                                                         |                                                                                         |
| S10: Middle edge grasp | π_ε |                                                                                         |                                                                                         |

1: For each state, we define the grasp type, the location of the grasps and the semantic description of deformation. Following [11], PP: pinch grasp, 2PP: bi-manual pinch, and π_ε: 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, that can be reversed to its previous state.
3: This manipulation is repeated at different distances from the corner grasp, always on the same edge.

The definition of the states takes inspiration from works like [25] where each change of contact interaction between hand, object and environment was considered a different scene state, however, in our case changes in deformation category are also considered. Rows in Table I depict these semantic states with the definition of the corresponding grasp type, location of grasp in the cloth, and deformation category.

The definition of the states takes inspiration from works like [25] where each change of contact interaction between hand, object and environment was considered a different scene state, however, in our case changes in deformation category are also considered. Rows in Table I depict these semantic states with the definition of the corresponding grasp type, location of grasp in the cloth, and deformation category.

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 I) 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 I, 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 I is shown with a different textile object from our proposed dataset.

B. Data Collection and Annotation

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

At the start of each demonstration, the cloth is placed in the initial state, i.e., lying flat on the table. 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 scale of the dataset, each of the seven manipulation tasks is performed at least two times by altering the speed, the initial position, the grasping points, and the manipulation trajectory. In order to introduce a higher volume of deformation, eighteen different garments were introduced of different size or shape, while each has a unique color texture and pattern as shown in Fig. 2. Seven of these garments are of squared shape, ten of rectangular, and one is squared but with smoothed corners. The table, the grippers, and the background remain the same for all manipulations. The RGB camera position is most of the time static, with minimal changes across the manipulations.

After having 22 human demonstration recordings using seven different manipulation types and eighteen different garments, we collect in total 33.6K fully annotated RGB images of 10 various semantic states.

The annotation has been done manually through human observation. To ease the heavy workload of data labelling, the images are annotated once it exhibits a state change. Otherwise, the remaining image frames are automatically labelled with the last adjacent state name. Furthermore, some short and flickering states, which either last less than 3 frames or are not in our state list given in Table I, are omitted.

C. Semantic State Estimation

Given the annotated dataset introduced in section III-B, we employ state of the art neural networks to estimate the semantic states of the manipulated deformable objects. For
this purpose, we particularly use networks pre-trained on Imagenet [14] with the help of transfer learning.

The networks of our choice are relying on convolutional operations such as EfficientNet [12], ResNet-50 [37] and ResNeXt-50 [38]. We also use a more lightweight vision transformer-based model DeiT [13]. Note that we append the same classifier layer to these network models and employ the cross-entropy loss function. In addition, we augment the data by randomly translating, flipping around the y-axis, and cropping. We also apply random rotation with a 15 degree restriction. The optimizer of our choice was stochastic gradient with warm restarts [39].

IV. EXPERIMENTS

We evaluate the performance of semantic state estimation networks in two use cases: monitoring of human and robot manipulations. In both cases, we measure the correct classification rate (i.e., the accuracy scores) as the evaluation metric. Furthermore, we generate class activation maps of the networks to ensure that the network attention is on similar features in both human demonstrations and robot executions.

A. Human Demonstrations for Training

We trained the four networks in section III-C by dividing our 33.6K annotated dataset with a 75-25 stratified split. To verify that the trained networks are not biased towards any specific cloth type, we exclude all the manipulations with some specific cloth types from the training data and employ them only for testing purposes. For instance, all demonstrations with the cloth White shown in Fig. 1 are reserved as the new unseen test data, while all the remaining data are used for training. The same leave one out testing protocol is also used for the Orange and Grid garments shown in Fig. 1.

Table II shows accuracy scores in percentage (%) for the validation and individual test cases with these three cloths. As shown in Table II, EfficientNet [12] 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 EfficientNet is not biased with the cloth types in the training data. The sample activation maps provided in Fig. 1 depict regions (such as corners, edges, etc.) where EfficientNet pays more attention when predicting the correct states, in these three test cases. Note that the network tends to focus more on the grasp points, cloth edges and corners (reddish zones) rather than the cloth textures and patterns. This is a strong evidence indicating that the state estimates heavily rely on manipulation-related features such as grasp points and cloth edges instead of irrelevant cues such as texture.

Fig. 3 depicts the confusion matrix for the test case of the White cloth in Table II. This figure clearly shows that there is no major confusion between the predicted semantic states even when the network is exposed to a new test cloth.

B. Long and Complex Human Demonstrations for Testing

To show the generalization of our state estimation network, we further recorded relatively long and complex cloth manipulation scenarios performed by humans. Our ultimate aim here is to diagnose the capacity of our network in totally new scene contexts.

For this purpose, we have recorded four chained manipulation scenarios each is composed of different number of manipulation tasks and semantic states defined in Table I.

TABLE II: Quantitative Evaluation.

| Networks          | Validation Scores | Test Scores       |
|-------------------|-------------------|-------------------|
|                   | White | Orange | Grid | Average |
| ResNet-50 [37]    | 97.77 | 96.27  | 91.79| 95.54   | 94.53   |
| ResNeXt-50 [38]   | 97.50 | 95.67  | 92.30| 95.81   | 94.49   |
| EfficientNet [12] | 97.71 | 95.53  | 93.16| 96.61   | 95.10   |
| DeiT [13]         | 98.20 | 93.85  | 86.18| 91.60   | 90.54   |

Fig. 4: The four long manipulation scenarios presented as a graph of the semantic states. Following the state definitions in Table I, in these four 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. Red arrows correspond to state transitions that are not present in our proposed dataset.
We used the same three cloths (White, Orange, and Grid) in these test scenarios. Fig. 4 illustrates those four scenarios as a graph sequence. Here, graph nodes represent the semantic states, whereas edges describe the manipulation tasks to go from one state to the next. For instance, the first scenario represents folding a cloth sideways by following a path (P1 in Fig. 4) with three states: S9, S4 and S5, whereas the third scenario follows P3 with six states to reach the same goal: S9, S4, S1, S10, S4 and S5. Note that all scenarios start with the same state S9: Lifted with two grippers to reach two different goal states, either S5: Folded sideways or S6: Folded diagonally.

In these four novel scenarios, we obtained 52.16% average accuracy score for all semantic states. The reason of this substantial drop in accuracy is mainly due to having totally new deformation and manipulation types, such as Add flat table contact and Edge tracing (shown in red arrows in Fig. 4), neither of which can be encapsulated by one of our trained semantic states or manipulation tasks. However, after fine-tuning the EfficientNet model with only two additional human demonstrations of these new deformation types using garments from the training set, the overall accuracy increased up to 84.80%. We here note that during these four scenarios, we collected more than 6K image frames, and only 350 RGB images from a garment that is not part of the three testing garments were used for the fine-tuning operation.

This experimental finding clearly shows the complexity of the state estimation problem: having unseen cloth deformations can still lead to errors due to high dimensionality in the deformation of textile objects. Incrementally refining the already trained network can, however, boost the performance to a large extent, as explored in our experiments.

Fig. 5 illustrates the EfficientNet network performance on the manipulation scenario P4 described in Fig. 4. The colored blocks clearly show that the network predictions are very similar to the human defined ground truth. There still exist false positive predictions which particularly emerge around borderline cases, as highlighted by the red frame in Fig. 5. For instance, the network has difficulty to distinguish states S1: Flat and S2: Flat semi-lifted with 1 gripper, in particular, while the subject is about to release the garment. The provided sample activation maps in Fig. 5 also depict that the network pays more attention to the grasp points, cloth edges and corners as expected.

C. Robot Manipulations

We executed the first two scenarios (P1 and P2) of these four long test scenarios described in Fig. 4 using 2 Kinova robot arms. In both scenarios, the goal is to fold the garment sideways by initially holding it in the air with both grippers.

We recorded the robot executions of P1 and P2 using 5 garments, 3 of them being the same test garments: White, Orange, and Grid. We collected in total 24K frames from robot executions. For the test garments, we had about 14.5K frames. For performance measurement of our network, we follow the same leave one out testing protocol employed in section IV-A. In each test case, we excluded data including one of White, Orange, and Grid garments respectively and used all the remaining recordings for fine-tuning our EfficientNet model already trained with the human demonstration data in section IV-B.

Note that differences in scene contexts between the robot and human demonstrations introduce the domain shift problem. Therefore, our EfficientNet trained with the human demonstration data in section IV-B cannot be directly employed here to monitor these robot executions. To easily overcome this problem, we applied this additional fine-tuning operation before testing with three garments. Note also that
Fig. 6: Human and robot demonstrations of the manipulation scenario P2 described in Fig. 4. The colored blocks in both cases represent the network predictions vs human labeled ground truth. Sample images with the corresponding class activation heatmaps are depicted on the top. The images in the red frames represent false positive predictions.

The recorded images were first cropped to have similar views with those recorded in human demonstrations. Finally, we obtained 97.52% average accuracy on the unseen three test garments: White, Orange, and Grid. The reason of obtaining such as a high accuracy is also due to the fact that robot motions are slow and less noisy than that of humans.

Fig. 6 shows the network performance for human and robotic demonstrations of the same manipulation scenario P2 described in Fig. 4. Each image is accompanied by the activation map taken from the network’s final convolution layer, which helps us observe the similarities between both demonstrations. It is evident that the network focuses on the lower boundary of the garment while being held in the air. Once the garment has contact with the table, the attention shifts to the grippers’ position with respect to the corners and the edges of the garment. The red frames depict incorrect predictions, which emerged, for instance, when the garment just switched from states $S9$: lifted with two grippers to $S4$: flat semi-lifted with two grippers. The colored blocks in Fig. 6 clearly show that such false positive predictions are borderline cases where the state is either about to change or just switched to the next.

Fig. 7 displays the network performance together with the human labeled ground truth for all human and robot demonstrated long manipulation scenarios described in Fig. 4. This side-by-side comparison indicates that the network can...
successfully predict semantic states even in the case of having unseen textiles. Note that the lengths of manipulation scenarios are normalized for the sake of clarity in the display. This figure also shows that the state predictions in robot executions is less noisy than that of human demonstrations, since humans follow more natural motion patterns. Note that the false network predictions in other paths (P3-Human and P4-Human in Fig. 7) also emerge mostly around state transitions.

V. DISCUSSION

We first would like to highlight the fact that in this work, we do not propose any novel network model. We rather show that the state-of-the-art models (e.g., EfficientNet [12]) can already handle the challenging state estimation problem with the help of transfer learning. Our reported high accuracy scores in Table II already show that there is no need to focus on designing new deep network architectures. Instead, we should address the domain shift problem. Our results show how the network loses accuracy very fast when presented with new unseen deformations, due to the high dimensionality of the cloth state estimation problem. We, however, show that fine-tuning the network with very few new frames can again boost the performance. Despite the domain shift, the class activation heatmaps in Figs. 5 and 6 also clearly ensure that our network focuses on similar manipulation relevant regions such as the grasp points, cloth edges and corners instead of the cloth textures and patterns.

In our dataset, human demonstrations were performed by two different individuals who had grippers attached to their hands and executed the planned actions through vocal commands. This process introduced additional noisy states to the expected state transitions in Fig. 4. For instance, a noisy state S2 appeared in Figs. 5 since one person released the gripper faster than the other subject. These naturally emerging noisy states in human demonstrations are kept in our dataset to make the dataset more challenging.

It is evident from Figs. 5, 6, and 7 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 (e.g., see P3-Human in Fig. 7), can easily be solved by incorporating temporal state information during the monitoring task. Since our networks perform frame-wise predictions, the temporal cue is omitted in this work.

Furthermore, we pose the following questions to better understand our network performance:

**What if we exclude the fine-tuning process?** In this case, we observe substantial accuracy drops both in human and robot test manipulations. Our ablation study shows that the fine-tuning boosts the performance by 32.64% and 45.06% for human and robot demonstrations, respectively.

**What if we train only with the robot demonstrations?** In case of excluding the human demonstrations and fine-tuning the network only using robot demonstration also leads to an increase in the average accuracy by 1.08%. However, the generated activation class heatmaps become extremely noisy. Therefore, we can conclude that this slight increase in accuracy comes with the cost of over-fitting to the observed scene. Instead of collecting more robot executions, learning the semantic states from more natural human demonstrations is needed to regularize the network and solve the overfitting problem.

**What if there exists no fine-tuning data?** We argue that such cases can be handled by injecting the uncertainty estimation to the prediction. However, such approaches are going beyond the scope of this paper.

VI. CONCLUSIONS

In this paper, we presented and evaluated a novel framework to solve the problem of semantic state estimation in continuous cloth manipulation tasks in an end-to-end manner. Our semantic state definition differs from classic rigid object approaches in that we introduce a high-level semantic description of the cloth state which couples the cloth deformation type, the grasping state and the contacts with the environment. To validate our approach, we benchmarked four different networks on our new dataset of 33.6K annotated RGB images.

As a 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 to autonomously define unseen textile deformations and plan a proper grasping accordingly.

We hope that the here presented dataset and benchmarks will be adopted by the cloth manipulation community and trigger further contributions in robotics.

REFERENCES

[1] A. Doumanoglou, J. Stria, G. Peleka, I. Mariolis, V. Petrık, A. Karagkos, L. Wagner, V. Hlaváč, T.-K. Kim, and S. Malassiotis, “Folding clothes autonomously: A complete pipeline,” *IEEE Transactions on Robotics*, vol. 32, no. 6, pp. 1461–1478, 2016.

[2] R. Hoque, D. Seita, A. Balakrishna, A. Ganapathi, A. K. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg, “Visuospatial foresight for multi-step, multi-task fabric manipulation,” *arXiv preprint arXiv:2003.09044*, 2020.

[3] M. Lippi, P. Poklukar, M. C. Welle, A. Varava, H. Yin, A. Marino, and D. Kragic, “Latent space roadmap for visual action planning of deformable and rigid object manipulation,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 5619–5626.

[4] A. Pumarola, A. Agudo, L. Porzi, A. Sanfeliu, V. Lepetit, and F. Moreno-Noguer, “Geometry-aware network for non-rigid shape prediction from a single view,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4681–4690.

[5] E. Corona, G. Alenyà, A. Gabas, and C. Torras, “Active garment recognition and target grasping point detection using deep learning,” *Pattern Recognition*, vol. 74, pp. 629–641, 2018.

[6] A. Ramisa, G. Alenyà, F. Moreno-Noguer, and C. Torras, “A 3d descriptor to detect task-oriented grasping points in clothing,” *Pattern Recognition*, vol. 60, pp. 936–948, 2016.

[7] 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 *Robotics Research: The 19th International Symposium ISRR*, 2018.
D. Seita, A. Ganapathi, R. Hoque, M. Hwang, E. Cen, A. K. Tanwani, A. Balakrishna, B. Thananjeyan, J. Ichnowski, N. Jamali et al., “Deep imitation learning of sequential fabric smoothing from an algorithmic supervisor,” in 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2020, pp. 9651–9658.

J. Matas, S. James, and A. J. Davison, “Sim-to-real reinforcement learning for deformable object manipulation,” in Conference on Robot Learning. PMLR, 2018, pp. 734–743.

Y. Li, Y. Wang, Y. Yue, D. Xu, M. Case, S.-F. Chang, E. Grinspun, and P. K. Allen, “Model-driven feedforward prediction for manipulation of deformable objects,” IEEE Transactions on Automation Science and Engineering, vol. 15, no. 4, pp. 1621–1638, 2018.

J. Borrós, G. Alenyà, and C. Torras, “A grasping-centered analysis for cloth manipulation,” IEEE Transactions on Robotics, vol. 36, no. 3, pp. 924–936, 2020.

M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in International Conference on Machine Learning. PMLR, 2019, pp. 6105–6114.

H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, “Training data-efficient image transformers & distillation through attention,” in International Conference on Machine Learning. PMLR, 2021, pp. 10 347–10 357.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

R. Ren, M. G. Rajesh, J. Sanchez-Riera, F. Zhang, Y. Tian, A. Agudo, Y. Demiris, K. Mikolajczyk, and F. Moreno-Noguer, “Grasp-oriented fine-grained cloth segmentation without real supervision,” arXiv preprint arXiv:2110.02903, 2021.

K. Puthuveetil, C. Kemp, and Z. Erickson, “Bodies uncovered: Learning to manipulate real blankets around people via physics simulations,” IEEE Robotics and Automation Letters, 2022.

D. Seita, P. Florence, J. Tompson, E. Coumans, V. Sindhwani, K. Goldberg, and A. Zeng, “Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks,” in 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021, pp. 4568–4575.

W. Yan, A. Vangipuram, P. Abbeel, and L. Pinto, “Learning predictive representations for deformable objects using contrastive estimation,” arXiv preprint arXiv:2003.05436, 2020.

R. Jangir, G. Alenyà, 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.

P.-C. Yang, K. Sasaki, K. Suzuki, K. Kase, S. Sugano, and T. Ogata, “Repeatable folding task by humanoid robot worker using deep learning,” IEEE Robotics and Automation Letters, vol. 2, no. 2, pp. 907–913, 2016.

R. Antonova, A. Varava, P. Shi, J. F. Carvalho, and D. Kragic, “Sequential topological representations for predictive models of deformable objects,” in Learning for Dynamics and Control. PMLR, 2021, pp. 348–360.

D. Fuentes-Jimenez, D. Casillas-Perez, D. Pizarro, T. Collins, and A. Bartoli, “Deep shape-from-template: Wide-baseline, dense and fast registration and deformable reconstruction from a single image,” arXiv preprint arXiv:1811.07791, 2018.

A. Gabas and Y. Kita, “Physical edge detection in clothing items for robotic manipulation,” in 2017 18th International Conference on Advanced Robotics (ICAR). IEEE, 2017, pp. 524–529.

F. Wörgötter, E. E. Aksoy, N. Krüger, J. Pfister, A. Ude, and M. Tomasini, “A simple ontology of manipulation actions based on hand-object relations,” IEEE Transactions on Autonomous Mental Development, vol. 5, no. 2, pp. 117–134, 2013.

E. E. Aksoy, A. Abramov, J. Dörk, K. Ning, B. Dellen, and F. Wörgötter, “Learning the semantics of object-action relations by observation,” The International Journal of Robotics Research, vol. 30, no. 10, pp. 1229–1249, 2011.

M. J. Acín, E. E. Aksoy, and F. Wörgötter, “Library of actions: Implementing a generic robot execution framework by using manipulation action semantics,” The International Journal of Robotics Research, vol. 38, no. 8, pp. 910–934, 2019.

Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, “Deepfashion: Powering robust clothes recognition and retrieval with rich annotations,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1096–1104.

L. Sun, S. Rogers, G. Aragon-Camarasa, and J. P. Siebert, “Recognising the clothing categories from free-configuration using gaussian-process-based interactive perception,” in 2016 IEEE international conference on robotics and automation (ICRA). IEEE, 2016, pp. 2464–2470.

L. Sun, G. Aragon-Camarasa, S. Rogers, R. Stolkin, and J. P. Siebert, “Single-shot clothing category recognition in free-configurations with application to autonomous clothes sorting,” in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2017, pp. 6699–6706.

T. Ziegler, J. Butepage, M. C. Welle, A. Varava, T. Novkovic, and D. Kragic, “Fashion landmark detection and category classification for robotics,” in 2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC). IEEE, 2020, pp. 81–88.

A. Verleysen, M. Biondina, and F. Wytels, “Video dataset of human demonstrations of folding clothing for robotic folding,” The International Journal of Robotics Research, vol. 39, no. 9, pp. 1031–1036, 2020.

J. Schulman, A. Lee, J. Ho, and P. Abbeel, “Tracking deformable objects with point clouds,” in 2013 IEEE International Conference on Robotics and Automation. IEEE, 2013, pp. 1130–1137.

A. Ramisa, G. Alenyà, F. Moreno-Noguer, and C. Torras, “Using depth and appearance features for informed robot grasping of highly wrinkled clothes,” in 2012 IEEE International Conference on Robotics and Automation. IEEE, 2012, pp. 1703–1708.

I. Mariolis, G. Peleka, A. Kargakos, and S. Malassiotis, “Pose and category recognition of highly deformable objects using deep learning,” in 2015 International conference on advanced robotics (ICAR). IEEE, 2015, pp. 655–662.

B. Willimon, S. Hickson, I. Walker, and S. Birchfield, “An energy minimization approach to 3d non-rigid deformable surface estimation using rgbd data,” in 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 2012, pp. 2711–2717.

J. Borrás, G. Alenyà, and C. Torras, “Encoding cloth manipulations using a graph of states and transitions,” arXiv preprint arXiv:2009.14681, 2020.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.

S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1492–1500.

I. Loshchilov and F. Hutter, “Sgdr: Stochastic gradient descent with warm restarts,” arXiv preprint arXiv:1608.03983, 2016.