Representing clothing items for robotics tasks

Marco Moletta, Michael C. Welle, Alexander Kravchenko, Anastasia Varava and Danica Kragic

***Abstract*— We study the problem of clothing item representation and build upon our recent work to perform a comparative study of the most commonly used representations of clothing items. We focus on visual and graph representations, both extracted from images. Visual representations follow the current trend of learning general representations rather than tailoring specific features relevant for the task. Our hypothesis is that graph representations may be more suitable for robotics tasks employing task and motion planning, while keeping the general properties and accuracy of visual representations. We rely on a subset of DeepFashion2 dataset and study performance of developed representations in an unsupervised, contrastive learning framework using a downstream classification task. We demonstrate the performance of graph representations in folding and flattening of different clothing items in a real robotic setup with a Baxter robot.**

I. Introduction

We address the problem of representing clothing items in the context of robotic manipulation tasks. Perceiving, representing and manipulating clothing items is studied in computer vision [11], machine learning [48], materials science [8]. However, robotic applications often require representations that are related to downstream tasks commonly not addressed in other scientific areas. For example, classification of clothing items [31] may be relevant for a sorting or recycling robot, but the representation used for classification may not apply when the items are to be picked up [5], folded [13], flattened [43], or used in an assistive dressing task [16].

In this paper, we build upon our recent work on clothing item classification, representation and manipulation [33], [29], [30], [50], and perform a comparative study of the most commonly used representations of clothing items. The goal is to study the representations in the context of robotic manipulation tasks, and identify challenges toward employing same representations for several manipulation tasks. In particular, we study visual and graph based representations. Visual representations are based on RGB images and graph representations are built from segmented and binarized RGB images. Visual representations follow the current trend of learning general representations rather than tailoring specific features relevant for the task. Our hypothesis is that graph representations may be more suitable for robotics tasks employing task and action planning, while still keeping the general properties and thus accuracy of visual representations.

To this end, we compare the performance of visual and graph representations using contrastive learning on a downstream classification task. In addition, we test the applicability of graph representations for flattening and folding and point to our previous work [29], [30] for using visual representations in the context of folding. One important challenge we face in our study is the availability of appropriate datasets relevant for robotics tasks. While computer vision and machine learning communities commonly resort to large databases of RGB images (ImageNet[12], FFHQ[24], IG-1T[47], Deepfashion[32], DeepFashion2 [19]), in robotic applications we commonly work with RGBD data. However, no widely accepted real-world datasets of robots performing tasks exist although various simulation scenarios have been proposed ([2], [17]). In robotics, there has been work on developing benchmarks for deformable object manipulation [18], but the works focus on defining evaluation metrics on aspects of employed perception, prior knowledge and ability to complete the underlying task, rather than on the representation itself. Thus, we resort to DeepFashion2, contributing to bridging the gap between the communities given that we use these robotics tasks of folding and flattening. In [29], [30], we focused on learning visual representations for folding and concluded that other types of representations may be needed for an easier transfer between different robotic tasks. We hope that better understanding of graph representations is a natural step towards achieving that.

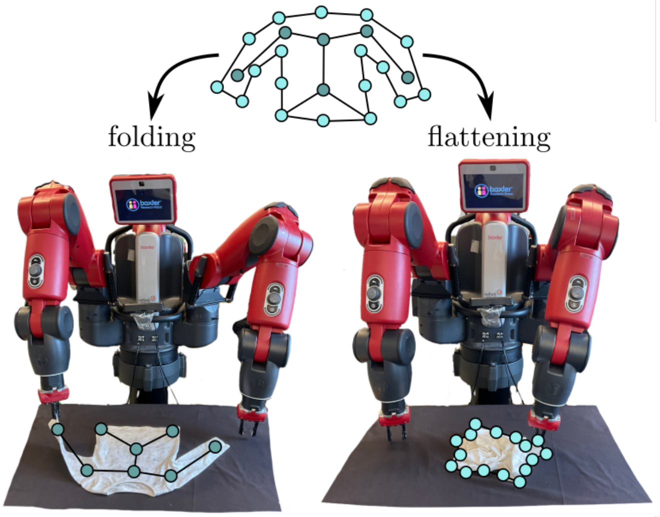

Fig. 1. Although rather similar in nature and relying typically on camera images, robotic tasks such as flattening and folding may resort to rather different state representations to enable effective task execution. For example, in folding we may benefit from a skeleton-like representation and use nodes to generate efficient manipulation plans. For flattening, relying on the outer contour of the clothing item may be sufficient.

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

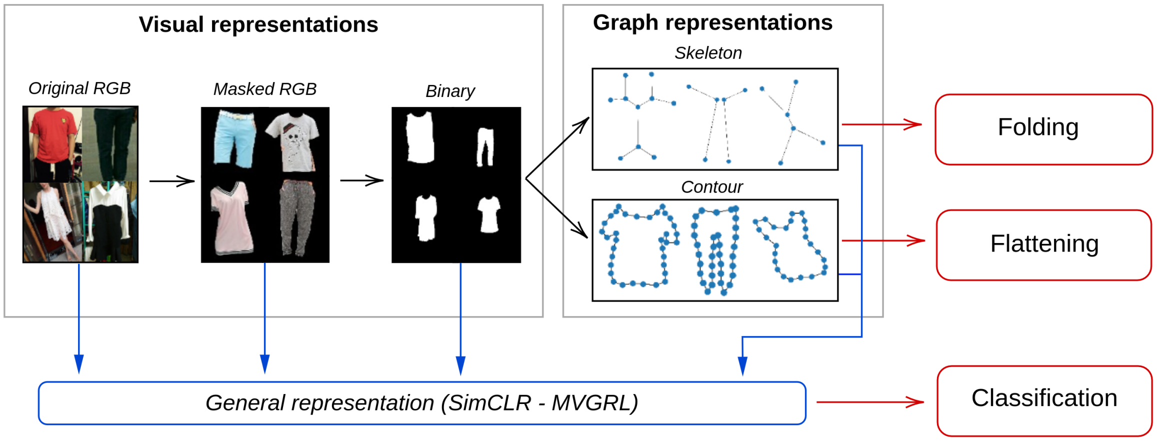

Fig. 2. Overview of the pipeline used to extract the representations, the evaluation frameworks and tasks used in this work. The black arrows identify the processing steps, while the red arrows define which representation is used for which task.

II. Related Work

Work on clothing items representations can be surveyed from the the perspective of perception [1], representation learning [27], task dependency [40], to name some. Our recent survey focuses on modelling and learning of deformable objects, considering also simulation and control aspects of these [49]. In this paper, we study clothing item representations and their use in folding and flattening, and thus survey the most relevant work with respect to these.

Clothing items are often represented using graphs and meshes, extracted from simulation [3], [28] or RGBD data/landmarks [41], [15] that can be learned [50], [34] or extracted [35], [13]. End-to-end approaches resort to learning internal representations from images or videos [14], [22], [45], often as a byproduct of the task to accomplish.

In terms of tasks related to clothing items, most of the work considers folding or flattening, handcrafting representations for the task. There are examples of extracting geometric cues from images or using fiducial markers to identify grasping points [4], [35], [6]. Flattening examples involve lifting the item and exploiting gravity to perform the task [20], [37], [25]. There are also examples of detecting wrinkles to guide the flattening [43], [42]. Recent work on end-to-end learning also addresses folding and flattening [29], [30], [36], [22], [45], [23]. One challenge of these methods is the dependency on retraining when something new occurs and the learned representation is no longer valid. For example, our own work on folding [29], showed that data-driven low-dimensional latent space representations can be employed to address complexity of planning actions on deformable objects. Our work considered learning of structured visual representations that allowed for real-time planning of folding tasks. However, as in most of the above mentioned related work, such representations are not easy to transfer to new tasks, even if these may seem similar. One of the advantages of using graph representations studied in this paper is the fact that they represent the geometry of the clothing item, allowing then to specify manipulation plans using graph nodes. We hypothesise that graph representations may be a better for transferring between tasks and study them in comparison to visual representation and in the context of folding and flattening.

III. Representations and contrastive learning

Fig. 2 illustrates our evaluation pipeline. We evaluate visual and graph representations in two contrastive learning frameworks SimCLR [9] and MVGRL [21]. The three visual representations are: i) *Original RGB*, ii) *Masked RGB* with removed background, and iii) *Binary* images. In terms of graphs, *i)* Skeleton and *ii)* Contour graphs, both extracted from the *Binary* images.

The motivation for choosing the three visual representations is rather straightforward - these are commonly used in computer vision, and most approaches addressing end-to-end learning employ these. We thus assess the benefit of background removal and binarization to contrastive representation learning. Regarding graph representations, skeletons have been a common representation when encoding human bodies [39] and are thus closely related to clothing items. Skeletons are invariant to changes in color or texture, and may be well suited for robotic folding or assistive dressing tasks as these represent the geometric structure on which planning and control may be defined. Here, we first perform *skeletonization* on the binary image using equidistance to the boundaries ([26], [44]) and then transform the result into a graph by adding nodes and edges as in [38].

The second type of graph representations are Contour graphs. We use Open-CV [7] for contour extraction on binary images, use a downsampled contour to generate graph nodes and generate an adjacency matrix by connecting the first-neighbours nodes of the contour.

A. Contrastive learning on visual and graph representations

SimCLR [9] and MVGRL [21] are unsupervised, contrastive representation learning methods, designed to produce general representations, normally assessed with k-nearest neighbour or linear classification. SimCLR [9], for example,

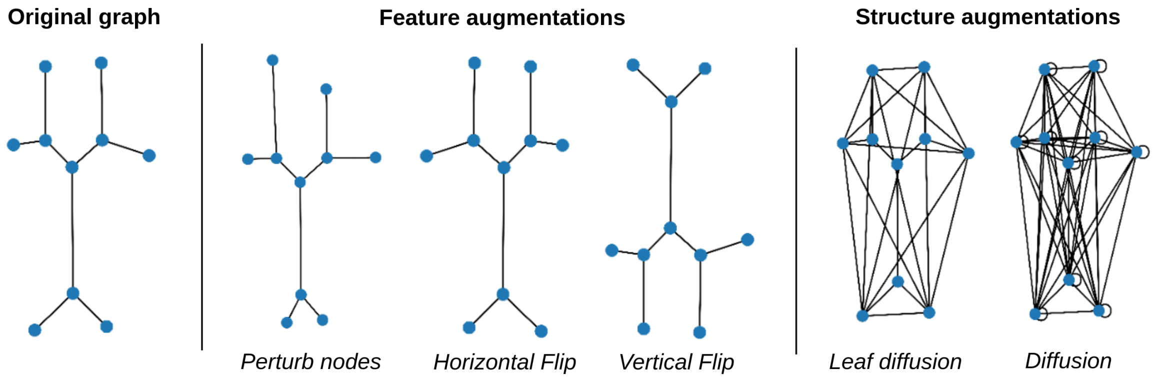

Fig. 3. The different graph augmentations evaluated in the MVGRL framework, created from the original graph. We show the augmentations for a skeleton graph extracted from a *short sleeve top* item.

showed that composition of data augmentations plays a critical role in obtaining general representations and that contrastive learning benefits from larger batch sizes and more training steps in comparison to supervised learning. In the experimental evaluation, we assess how background removal and binarization affect the representation learned using SimCLR and MVGRL on a downstream classification task. The motivation is that if the representation learned from a binarized image retains its general properties, we can use it to further extract the graph representation from it.

Visual Augmentations: *Random-resize-crop* or *color-jitter* augmentation have been proposed [9], [10] as common data augmentation techniques for RGB images in a contrastive learning framework. We have previously shown that *random-crop* is especially important in the case where the resulting image patches intersecting each other [46]. In addition to these two, in our experiments we also apply *Random horizontal flip* and *Random grayscale* augmentations. For binary images, *color-jitter* and *Random grayscale* cannot be applied and are therefore not employed.

Graph Augmentations: Augmentations applied to graphs can be divided into two categories: *i)* Structure-augmentations, and *ii)* Feature-augmentations. Structure-augmentations are performed on the graph structure itself, such as adding, removing, or modifying connectivity. Feature-augmentations, on the other hand, affect the node features directly, in our case this translates to shifting the position of the nodes.

In MVGLR only the Structure-augmentation *diffusion* is applied, which augments the adjacency matrix of the skeleton graph with additional edges. For *skeletons*, we also employ and analyse the effect of an additional structure-augmentation called *Leaf diffusion*, that allows to increase the connectivity of the graph by adding edges between the leaf nodes of the *skeleton*. We also employ a number of feature-augmentations:

- *Perturb nodes:* the nodes features are displaced from their original position by a certain amount. This displacement slightly differs in implementation between skeletons and contours: for skeletons, the magnitude of the perturbation is relative to the euclidean distance (in the pixel-space) between the leaf and the parent node. The displacement of the leaf node in the horizontal and vertical direction is sampled in the interval $[0, d(v_l, v_p)]$, where v_l and v_p are the features respectively of the leaf and the parent node. This is done to ensure that no overlapping of leaf and parent nodes is present. For contours, this interval is instead the distance between two subsequent nodes.
- *Horizontal Flip:* the nodes features are modified such that the graph is mirrored with respect to the vertical axis of the image.
- *Vertical Flip:* the nodes features are modified such that the graph is mirrored with respect to the horizontal axis of the image.

An overview of the employed augmentations can be seen in Fig. 3.

B. Dataset

The dataset used for training and evaluation consists of a subset of the Deepfashion2 dataset [19]. The Deepfashion2 dataset consists of a total of 491k images of 13 categories of clothing items. In addition to having segmentation landmarks that we use for background segmentation, each item presents further attributes such as *scale, occlusion, zoom-in, viewpoint*.

We use a subset of this dataset and name it *Rep-fashion* where we removed four under-represented categories (*short sleeve outwear*, *sling*, *long sleeve dress* and *sling dress*) and focus on images with attributes: *scale* = *moderate, occlusion* = *no/slight, zoom-in* = *no* and *viewpoint* = *frontal*. The images in the dataset are also all downsampled and padded to 160*160 pixels. In total, the Rep-fashion dataset consists of 25102 images where 21766 are used for training and 3336 for testing. The exact composition of the Rep-fashion dataset can be seen in Table I.

IV. Analysis

The first question we want to answer is *"Are graph representations comparable to visual representations in terms of encoding general information?"* To this end, SimCLR

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

TABLE I

CATEGORIES AND COMPOSITION OF TRAINING DATASET (NUMBER OF SAMPLES - PERCENTAGE).

and MVGRL are first applied on both visual and graph representations, and a comparison is performed using a classification task on the respective representations.

The second question is *"Are graph representations useful for robotic flattening and folding?"* The goal of the latter is to support the hypothesis that graph representations may generalize better than visual representations *over* several robotics tasks on different clothing items. All task execution videos can be found in the project website[1].

A. Representation in the context of downstream classification

The performance comparison between representations obtained with SimCLR (*raw RGB*, *RGB masked*, *Binary masked*) and representations obtained with the MVGRL model (*skeleton-* and *contour-graphs*) is shown in Table II. The results, from the best performing augmentations, are reported using the KNN evaluation protocol for SimCLR [9] and the linear evaluation protocol for [21] for MVGRL. In this experiment, the batch size is set to 64 and both models are trained for 1000 epochs. Both models are trained with the loss functions used in the original papers (NT-Xent loss [9] for SimCLR and Jensen-Shannon divergence (JSD) [21].

Model	Top 1 Acc.	Top 5 Acc.
Original-RGB SimCLR	54.1 %	84.8 %
Masked-RGB SimCLR	**71.6 %**	**92.5 %**
Binary SimCLR	65.0 %	85.2 %
Skeleton MVGRL	52.4 %	90.7 %
Contour MVGRL	48.6 %	90.9 %

TABLE II

REPRESENTATION CLASSIFICATION RESULTS USING THE KNN EVALUATION PROTOCOL FOR SIMCLR USING RGB-RAW, RGB-MASKED, AND BINARY-MASKED AS INPUTS, AND THE LINEAR EVALUATION PROTOCOL FOR MVGRL USING THE SKELETON- AND CONTOUR-GRAPHS AS INPUTS USING THE AUGMENTATION *Diffusion - Horizontal Flip - Vertical Flip*.

From the results, we observe that SimCLR achieves the best results on RGB-masked images. To some extent an expected result, given that RGB-masked is the best balance between keeping the relevant information (texture, shape) and removing irrelevant one (background). While classification rates decrease further for graph based representations, the results are rather comparable between skeleton and contour graphs and not much worse than for visual representations. Thus, even if graphs are simpler in terms of the dimensionality, their performance on downstream classification tasks are comparable to high-dimensional visual representations. The results need to be put in the context of tasks such as folding or flattening that require to generate an action plan for a robot to execute. The underlying dimensionality of visual representations is several orders larger than that of graphs, which commonly requires to embed the representation in some latent space on which the planning can be performed.

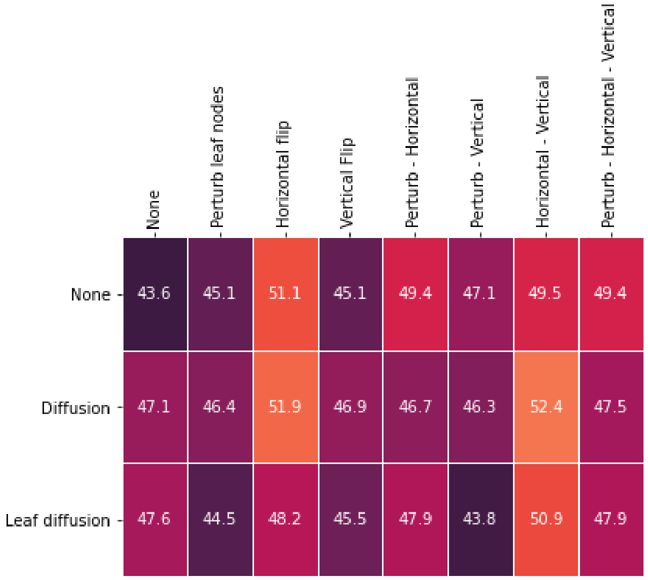

Fig. 4. Ablation study of the combination of different graph augmentation for the skeleton graph using the MVGRL framework.

B. Graph Augmentations

Compared to visual representations, graph representations are lower-dimensional and one way of obtaining general representations with MVGRL is to apply various augmentations to original graphs. We compared graph augmentation techniques by applying a structure-augmentation to one view of the instance and a combination of different feature augmentations to the other view. We do not combine *Diffusion* and *Leaf diffusion* as these are redundant. Fig. 4 and Fig. 6 summarize the results, former for the skeleton and latter for the contour graphs. The X-axis represents the feature space augmentation and Y-axis the structure space augmentation. The values are the top 1 accuracy performance of the linear evaluation protocol with the different augmentation combinations. Overall, the performance increases slightly with the employed augmentations. In particular, both graph representation seem to benefit from *structure-augmentation* suggesting that increasing the connectivity of the graph is important to obtain general representations. For skeleton graphs, the *horizontal flip* results in best performance, while this is not the case for contour graphs where the *vertical flip*

[1] https://cloth-representation.github.io/web/

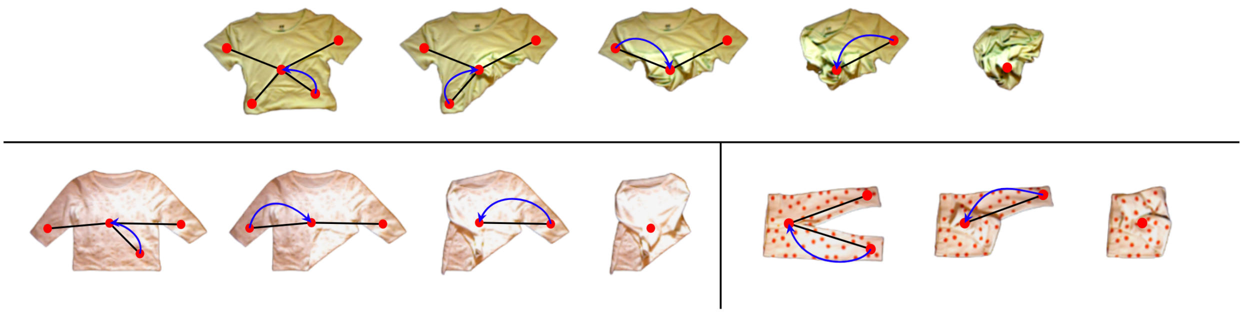

Fig. 5. Demonstration of folding task performed by a Baxter robot using the *skeleton* representation on the *short sleeve top* (top row) *long sleeve top* (bottom row - left) and *trousers* (bottom row - right).

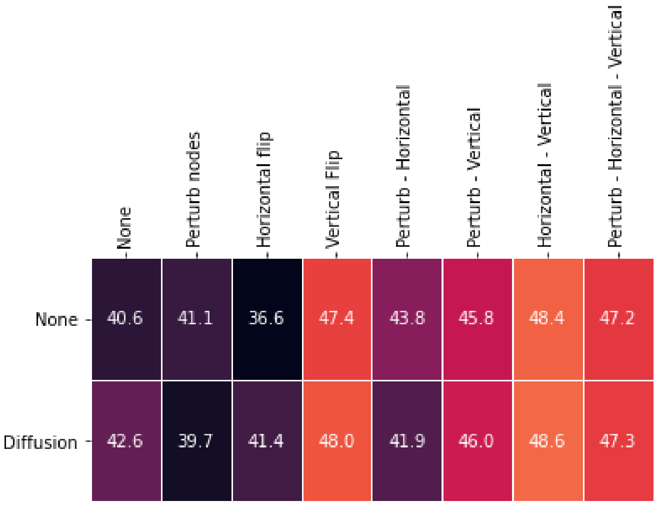

Fig. 6. Ablation study of the combination of different graph augmentation for the contour graph using the MVGRL framework.

helps more. Another interesting finding is that *perturb nodes* augmentation worsens the performance on both representations, which is in line with the findings shown in [21].

C. Graph representations for folding and flattening

We study graph representations in the context of folding and flattening using a Baxter robot, see Fig. 2. For folding, we start by defining a *folding-plan* similar to [13] consisting of node pairs (pick-up node, put-down node), see Fig. 5 for an example. We follow the structure of the skeleton-graph: leaf nodes are folded on top of their parent nodes. The folding plan for the *short sleeve top* consists of four folding steps, for *long sleeve top* three and for *trousers* two. The number of steps in the folding plans are easily changed based on the complexity of the extracted skeletons. In the current evaluation, we assume that the garment is in the similar starting position. Table III shows results for *skeletons* on three clothing items. The reported scores are the percentage of correctly picked and placed nodes, averaged among 5 iterations. The task execution videos can be found here[1].

	Correctly placed nodes
long sleeve top	3 steps - 93.3%
short sleeve top	4 steps - 90.0%
trousers	2 steps - 100.0%

TABLE III
FOLDING RESULTS - SKELETON

Flattening relies on the contour graphs and a flattening plan to reach a target state of a clothing item given an initial state, see Fig. 7. The target state is predefined for each item and maintained fixed for the 5 different trials. We keep the number of nodes same during the whole flattening sequence consisting of several pick-up-node and put-down-node steps. Put-down-nodes position are the ones defined by the target state of the clothing item. A flattening error e is calculated after each step and defined as the distance (in pixels) between the *contour* in the current state and the *contour* in the target state, as $e = \sum_{i \in I} ||v_i^{current}, v_i^{target}||_1/|I|$ where I is the set of indices of the nodes of the *contour* and $v = (x, y)$ the pixel coordinates of the nodes. The *score* is then calculated as $score = (e^{initial} - e^{final})/e^{initial}$, where $score = 1$ is the maximum possible score, which consists of reaching the target state where the item is perfectly flattened.

From the results in Table IV, we can see that the representation allows to consistently obtain a positive score, corresponding to a state closer to the final state. The results are indicative of that flattening simpler, more convex items (short sleeve top) is easier than for more complex ones.

	Avg. Score	**Max. score**
long sleeve top	0.20 ± 0.10	0.40
short sleeve top	0.52 ± 0.15	0.67
trousers	0.39 ± 0.24	0.77

TABLE IV
FLATTENING RESULTS - CONTOUR

While *contours* could be used for *folding*, the opposite is not valid for *skeletons* in *flattening*. This is due to the fact that the *skeleton* of a crumpled item is substantially different from a flattened one, both in number of nodes and in connectivity between them. Hence, we argue that setting up a *flattening* task using *skeleton* representations is not indicated.

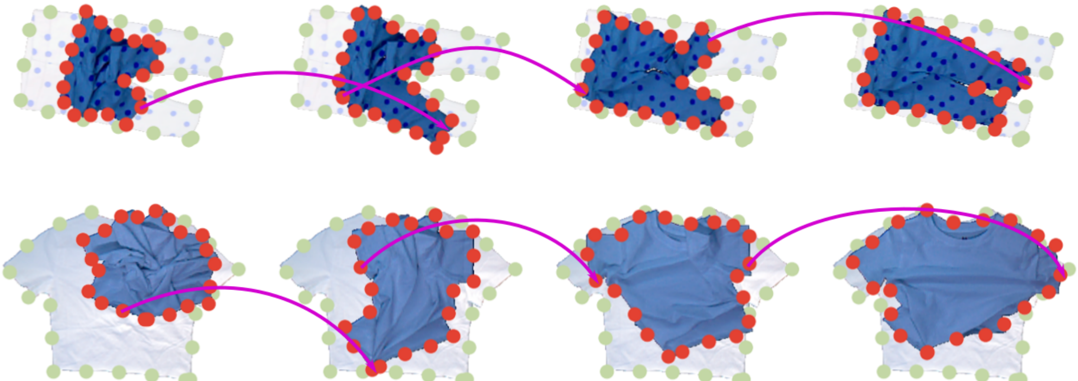

Fig. 7. Demonstration of flattening task performed using the contour representation on *trousers* (top row) and *short sleeve top* (bottom row)

V. Conclusion

We addressed the problem of representing clothing items for robotics tasks, and assessed graphs based representations in relation to commonly used visual representations. Graphs are extracted from binarized RGB images, thus removing texture and color information. An important question was to what extent graphs still retain properties of general representations learned from image data. Given lower dimensionality of graphs and their demonstrated usefulness in robotics tasks such as flattening and folding, we find these as more suitable representations for robotic tasks requiring task and motion planning. For the assessment, we employed a subset of a commonly used dataset DeepFashion2 and learned representation using contrastive learning frameworks SimCLR (for visual representations) and MVGRL (for graph representations). The results showed that while there is a slight drop in performance for the suggested graph representations in comparison to visual representations, there is no such tremendous difference that would motivate the use of raw RGB images and work with representations learned directly from them. We showed how graph representations can be used for flattening and folding and we conclude that these are a promising, low dimensional representation for cloth manipulation that retains a high degree of general information. We plan to further use them for learning representations that transfer over several robotic tasks in the context of clothing item manipulation.

References

[1] K. E. Ak, A. A. Kassim, J. H. Lim, and J. Y. Tham, "Learning attribute representations with localization for flexible fashion search," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7708–7717.

[2] R. Antonova, P. Shi, H. Yin, Z. Weng, and D. K. Jensfelt, "Dynamic environments with deformable objects," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

[3] 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.

[4] C. Bersch, B. Pitzer, and S. Kammel, "Bimanual robotic cloth manipulation for laundry folding," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2011, pp. 1413–1419.

[5] J. Borràs, G. Alenyà, and C. Torras, "A grasping-centered analysis for cloth manipulation," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 924–936, 2020.

[6] J. Borràs, G. Alenya, and C. Torras, "A grasping-centered analysis for cloth manipulation," 2020.

[7] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[8] J. Cao, R. Akkerman, P. Boisse, J. Chen, H. Cheng, E. De Graaf, J. Gorczyca, P. Harrison, G. Hivet, J. Launay, *et al.*, "Characterization of mechanical behavior of woven fabrics: experimental methods and benchmark results," *Composites Part A: Applied Science and Manufacturing*, vol. 39, no. 6, pp. 1037–1053, 2008.

[9] 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.

[10] T. Chen, S. Kornblith, K. Swersky, M. Norouzi, and G. E. Hinton, "Big self-supervised models are strong semi-supervised learners," *Advances in neural information processing systems*, vol. 33, pp. 22 243–22 255, 2020.

[11] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion meets computer vision: A survey," *ACM Computing Surveys (CSUR)*, vol. 54, no. 4, pp. 1–41, 2021.

[12] 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*, 2009, pp. 248–255.

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

[14] 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," 2018.

[15] C. Elbrechter, R. Haschke, and H. Ritter, "Folding paper with anthropomorphic robot hands using real-time physics-based modeling," in *2012 12th IEEE-RAS International Conference on Humanoid Robots (Humanoids 2012)*, 2012, pp. 210–215.

[16] Z. Erickson, H. M. Clever, G. Turk, C. K. Liu, and C. C. Kemp, "Deep haptic model predictive control for robot-assisted dressing," in *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2018, pp. 4437–4444.

[17] Z. Erickson, V. Gangaram, A. Kapusta, C. K. Liu, and C. C. Kemp, "Assistive gym: A physics simulation framework for assistive robotics," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 169–10 176.

[18] 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.

[19] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo, "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5337–5345.

[20] 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.

[21] K. Hassani and A. H. K. Ahmadi, "Contrastive multi-view representation learning on graphs," *CoRR*, vol. abs/2006.05582, 2020. [Online]. Available: https://arxiv.org/abs/2006.05582

[22] R. Hoque, D. Seita, A. Balakrishna, A. Ganapathi, A. K. Tanwani, N. Jamali, K. Yamane, S. Iba, and K. Goldberg, "Visuospatial foresight for physical sequential fabric manipulation," *Autonomous Robots*, pp. 1–25, 2021.

[23] B. Jia, Z. Hu, J. Pan, and D. Manocha, "Manipulating highly deformable materials using a visual feedback dictionary," 2019.

[24] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[25] Y. Kita, T. Ueshiba, E. S. Neo, and N. Kita, "A method for handling a specific part of clothing by dual arms," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 4180–4185.

[26] T.-C. Lee, R. L. Kashyap, and C. N. Chu, "Building skeleton models via 3-d medial surface/axis thinning algorithms," *CVGIP Graph. Model. Image Process.*, vol. 56, pp. 462–478, 1994.

[27] Y. Li, Y. Luo, and Z. Huang, "Fashion recommendation with multi-relational representation learning," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 3–15.

[28] X. Lin, Y. Wang, Z. Huang, and D. Held, "Learning visible connectivity dynamics for cloth smoothing," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 256–266. [Online]. Available: https://proceedings.mlr.press/v164/lin22a.html

[29] 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.

[30] ——, "Enabling visual action planning for object manipulation through latent space roadmap," *arXiv preprint arXiv:2103.02554*, 2021.

[31] J. Liu and H. Lu, "Deep fashion analysis with feature map upsampling and landmark-driven attention," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.

[32] 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.

[33] A. Longhini, M. C. Welle, I. Mitsioni, and D. Kragic, "Textile taxonomy and classification using pulling and twisting," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 7564–7571.

[34] X. Ma, D. Hsu, and W. S. Lee, "Learning latent graph dynamics for deformable object manipulation," 2021.

[35] J. Maitin-Shepard, M. Cusumano-Towner, J. Lei, and P. Abbeel, "Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding," in *2010 IEEE International Conference on Robotics and Automation*, 2010, pp. 2308–2315.

[36] 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.

[37] F. Osawa, H. Seki, and Y. Kamiya, "Unfolding of massive laundry and classification types by dual manipulator," *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol. 11, no. 5, pp. 457–463, 2007.

[38] F. Reinders, M. E. D. Jacobson, and F. H. Post, "Skeleton graph generation for feature shape description," in *Data Visualization 2000*, W. C. de Leeuw and R. van Liere, Eds. Vienna: Springer Vienna, 2000, pp. 73–82.

[39] B. Ren, M. Liu, R. Ding, and H. Liu, "A survey on 3d skeleton-based action recognition using learning method," *arXiv preprint arXiv:2002.05907*, 2020.

[40] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, "Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey," *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, 2018.

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

[42] K. Sun, G. Aragon-Camarasa, P. Cockshott, S. Rogers, and J. Siebert, "A heuristic-based approach for flattening wrinkled clothes," vol. 8069, 08 2013.

[43] L. Sun, G. Aragon-Camarasa, S. Rogers, and J. P. Siebert, "Accurate garment surface analysis using an active stereo robot head with application to dual-arm flattening," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 185–192.

[44] H. Sundar, D. Silver, N. Gagvani, and S. Dickinson, "Skeleton based shape matching and retrieval," in *2003 Shape Modeling International.*, 2003, pp. 130–139.

[45] A. Wang, T. Kurutach, K. Liu, P. Abbeel, and A. Tamar, "Learning robotic manipulation through visual planning and acting," *arXiv preprint arXiv:1905.04411*, 2019.

[46] M. C. Welle, P. Poklukar, and D. Kragic, "Batch curation for unsupervised contrastive representation learning," *arXiv preprint arXiv:2108.08643*, 2021.

[47] I. Z. Yalniz, H. Jégou, K. Chen, M. Paluri, and D. Mahajan, "Billion-scale semi-supervised learning for image classification," 2019.

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

[49] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021.

[50] 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.