UniGarmentManip: A Unified Framework for Category-Level Garment Manipulation via Dense Visual Correspondence

Ruihai Wu*[1,4] Haoran Lu*[2,1] Yiyan Wang [3,1] Yubo Wang [2,1] Hao Dong[1,4]
[1]CFCS, School of CS, PKU [2]School of EECS, PKU [3]School of CS&T, BIT
[4]National Key Laboratory for Multimedia Information Processing, School of CS, PKU

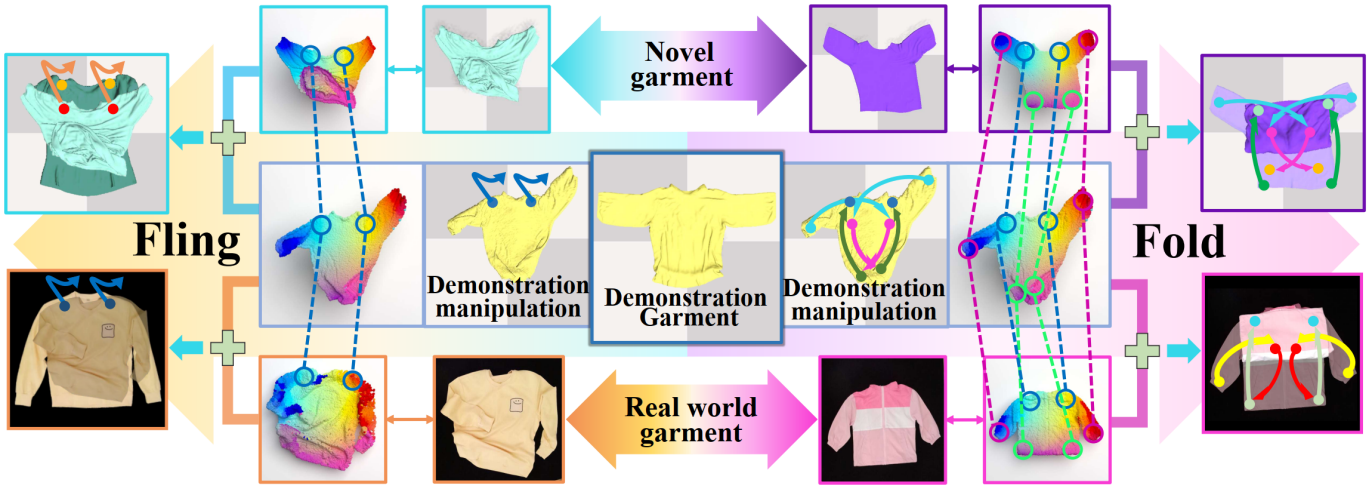

Fig. 1: **Our Proposed Framework** broadcasts the support relations recursively from the target object using local dynamics between adjacent objects, and uses the support relation graph to efficiently guide the step-by-step target object retrieval.

***Abstract*—Garment manipulation (*e.g.*, unfolding, folding and hanging clothes) is essential for future robots to accomplish home-assistant tasks, while highly challenging due to the diversity of garment configurations, geometries and deformations. Although able to manipulate similar shaped garments in a certain task, previous works mostly have to design different policies for different tasks, could not generalize to garments with diverse geometries, and often rely heavily on human-annotated data. So we leverage the property that, garments in a certain category have similar structures, and then learn the topological dense (point-level) visual correspondence among garments in the category level with different deformations in the self-supervised manner. The topological correspondence can be easily adapted to be functional to guide the manipulation policies for various downstream tasks, within only one or few-shot demonstrations.**

***Index Terms*—Garment, One-/Few-shot Manipulation**

I. Introduction

Next-generation robots should have the abilities to manipulate a large variety of objects in our daily life, including rigid objects, articulated objects [6] and deformable objects [19]. Compared with rigid and articulated objects, deformable objects are much more difficult to manipulate, for the highly large and nearly infinite state and action spaces, and complex kinematic and dynamics. Garments, such as shirts and trousers, are essential types of deformable objects, due to the potentially wide-range applications for both industrial and domestic scenarios. Manipulating garments, such as unfolding, folding and dressing up, has garnered significant interest in robotics.

There have been a long range of studies on manipulating relatively simple shaped deformable objects, such as square-shaped cloths [13], [19], [21], ropes and cables [15], [19], [21], and bags [2], [4]. Nevertheless, manipulating garments presents a substantial challenge, as it necessities the comprehensive understanding of more diverse geometries (garments in a certain category have different shapes, let alone in different categories), more complex states (various geometries with diverse self-deformations), and more difficult goals (*e.g.*, garments require multiple fine-grained actions fold step by step). Many existing studies on garment manipulation rely on large-scale annotated data [1], [3], which is labor-intensive and time-consuming, hindering the scalability in the scenarios of real-world applications. Besides, many works design quite different methods to tackle different specific tasks [1], [3], [18], [23], making it difficult to efficiently share and reuse information among different tasks.

*Equal contribution.

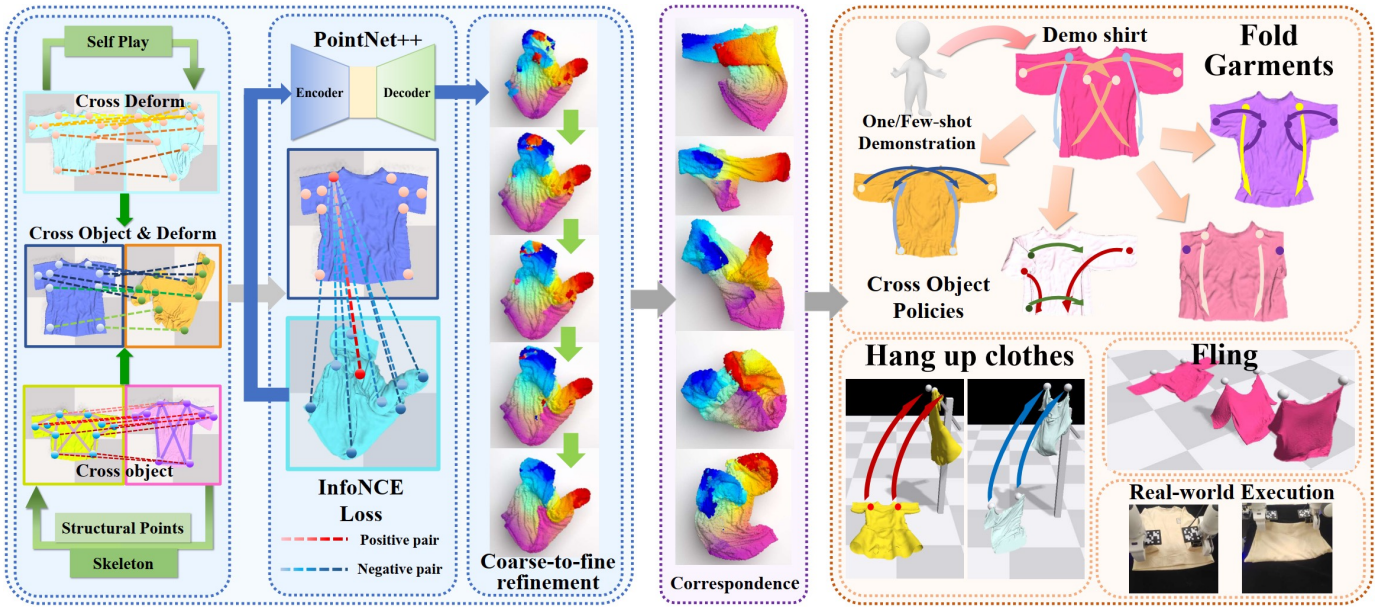

Fig. 2: **Our Proposed Learning Framework for Dense Visual Correspondence. (Left)** We extract the cross-deform correspondence and cross-object correspondence point pairs respectively using self-play and skeletons, and train the per-point correspondence scores in the contrastive manner, with the Coarse-to-fine module refines the quality. **(Middle)** Learned correspondence demonstrates point-level similarity across different garments in different deformations. **(Right)** The learned point-level correspondence can facilitates multiple diverse downstream tasks using one or few-shot demonstrations.

Different from other object types, garments possess a property that, in a certain category, while different garments may have different geometries, they usually share the same structure. For example, tops (such as T-shirts, jackets and jumpers), are composed of certain components (a body with two sleeves and a collar), and the topological structures of the components are usually the same, even though the length, width and geometries of a certain component in different garments may be quite different. Thanks to such similarity in structure shared among garments in the category level, it is easy for humans to fulfill a task on unseen novel garments using the experience of manipulating only one or a few garments in the same category. Therefore, we empower robots with the above one/few-shot generalization ability humans have in diverse tasks, by leveraging such structural similarity among garments.

Among multiple ways to describe and represent garments (*e.g.*, poses [5], [24], lines [8], [26] and keypoints [25]), **skeleton** [16], *i.e.*, a graph of keypoints covering significant points on garment edges and joints to represent the topology of 3D objects, is suitable for describing the above-mentioned structures shared among garments. The skeleton points are sparse, distinct and ordered, and thus (1) exist on each garment and (2) can easily distinguish with other skeleton points, making them easy to learn. Therefore, we use skeleton points to build structural correspondence among garments. Moreover, as different-extent self-deformations make the garments to be quite complex, while previous works only studied skeleton points on rigid [16], articulated [22] or fixed-posed deformable objects in the canonical view [25], we further extend skeleton points to garments at any deformation states, making a step to more realistic scenarios for garment manipulation.

While skeleton points build topological correspondence between different garments in the skeleton keypoint level, the state and action spaces of garments are exceptionally large and each point on the garment could be the manipulation point, making the sparse skeleton points unable to fully represent garments for manipulation. To represent objects with large state and action spaces, dense (*i.e.*, point-level or pixel-level) object representations, including dense object descriptors [7] and dense visual actionable affordance [14], which indicate the actionable information on each point of the object, have demonstrated its superiority on rigid [7], articulated [20], and simple-shaped deformable object manipulation [19]. We further extend dense object representations to garments, with the awareness of garment correspondences, using the proposed skeleton points, and thus achieve fine-grained manipulation for complicated garments.

With dense visual correspondence aware of garment structures, one demonstration can roughly guide manipulating a novel garment by indicating corresponding action points and policies. Furthermore, as manipulation for specific tasks rely on not only garment structures but also task-specific knowledge, we further transform the representation from task-agnostic structural to task-specific functional for more accurate manipulation in various downstream tasks, using few-shot demonstrations to achieve this adaptation.

II. FRAMEWORK

A. Overview

Our framework first learns topological dense visual correspondence aware of different garment deformations and shapes respectively using self-play and skeleton points (Section II-B), with further coarse-to-fine refinement (Section II-C). Therefore, the proposed framework could facilitate manipulating unseen novel garments on various tasks using one or few-shot demonstrations (Section II-D).

B. Self-supervised Topological Dense Visual Correspondence

The diversity of garments in different states mainly comes from two perspectives: self-deformations, and styles of objects in the same category. To empower the Dense Visual Correspondence with the alignment ability for different garments in different states, we decouple the learning process into two parts, respectively learning cross-deformation correspondence and cross-object correspondence.

1) Cross-Deformation Correspondence: Many tasks, such as unfolding and hanging, require manipulating the garment at any random states (*e.g.*, after a random drop). As demonstrated in [9], while garments have complex states and infinite deformations, the manipulation policies (manipulation points) are usually invariant to deformations. To empower the model with the ability to handle garments in different deformations, we introduce learning correspondence across deformations of the same garment.

Given two partial observations O and O' of the same garment in different deformations generated by self-play, and a visible point p on O, we can easily get its corresponding position point p' in O' using point tracing in simulation. If p' is visible, the representations f_p and $f_{p'} \in \mathbb{R}^{512}$ of p and p' extracted by the backbone network $\mathbf{F}$, should be the same, as the representations are agnostic to self-deformations. We normalize point representations to be unit vectors, and thus the similarity between f_p and $f_{p'}$ can be computed by the dot product of f_p and $f_{p'}$, *i.e.*, $f_p \cdot f_{p'}$. For p on O, we use p' on O' as the positive point, and sample m negative points ($m = 150$): $p'_1,\ p'_2,\ ...,\ p'_m$. We pull close f_p and $f_{p'}$, while push away f_p and other point representations. Following InfoNCE [12], a widely-used loss function in one-positive-multi-negative-pair contrastive representation learning, we identify the positive p' amongst m negative samples.

2) Cross-Object Correspondence: In a certain category, while garments highly vary in original shapes, such as sizes, length-width ratios, sleeve lengths and styles, they share the same topological structure. The awareness of such structures will make it easy to manipulate unseen novel garments with demonstrations.

To leverage the shared structural information and generalize to novel shapes, we propose to use skeleton, *i.e.*, a graph of keypoints that represents topology of the 3D object, as the shared bridge for different garments with similar structures. The reasons for using skeleton include:

- Skeleton points are **distinct** and **sparse**, thus easy to learn and generalize, compared to complicated representations;
- Skeleton points are **distinct** and **ordered**, making it easy to build topological correspondence between two objects by aligning each specific skeleton point on them;

As skeleton points are **ordered**, given observation O of a flat garment with one of its skeleton point p, we can get the corresponding skeleton point $\tilde{p}$ on the observation $\tilde{O}$ of another flat one, by applying the skeleton network on $\tilde{O}$ and get the skeleton point in the same order of p in O. Then, the topological correspondence between **flat garments** have been built in the **skeleton-point level**. As the features extracted by neural networks are continuous when point positions continuously change, and skeleton points cover the whole garment, the feature of any point can be reflected by its nearby skeleton points (like interpolation) with topological information. Therefore, the representation of each point on the garment will reflect its topology, and **dense correspondence** between flat garments has been naturally built.

3) Integration of Cross-Deformation and Cross-Object Correspondence: Since we have designed dense correspondence between **the same garment in different deformations**, and dense correspondence between **different flat garments**, the next step is to aggregate them into one dense representation system on diverse garments in any deformation states.

We first project skeletons of garments in their flat states to any deformation states using point tracing in simulation. Thus, given the observation O in random deformation with one of its skeleton point p, we can get the corresponding skeleton point $\tilde{p}$ on the observation $\tilde{O}$ of another garment in random deformation. If $\tilde{p}$ is visible on $\tilde{O}$, f_p and $f_{\tilde{p}}$ should be the same. For p on O, we use $\tilde{p}$ on $\tilde{O}$ as the positive point, and sample m negative points ($m = 150$): $\tilde{p}'_1,\ \tilde{p}'_2,\ ...,\ \tilde{p}'_m$. We use contrastive learning for training.

C. Coarse-to-fine Correspondence Refinement

Although above framework can learn the general distributions of all points' representations using offline randomly collected data, some difficult details (such as the boundaries between the folded sleeve on the garment body) should be paid more attention by the model, and there may exist inaccurate representations on some points or areas. The above phenomenon is also demonstrated in previous dense correspondence learning studies for 3D objects [10], [11], [17].

Therefore, we propose the Coarse-to-fine (C2F) Correspondence Refinement procedure to make the model more focused on difficult points on the garment, and eliminate inaccurate predictions, by refining the offline trained model using its online prediction failures.

D. Manipulation Policy Generation

As shown in Figure 1, for novel garments over different downstream tasks, we can easily generate manipulation policies by selecting the picking and placing points that are most close to the demonstrations in the correspondence space.

REFERENCES

[1] Yahav Avigal, Lars Berscheid, Tamim Asfour, Torsten Kröger, and Ken Goldberg. Speedfolding: Learning efficient bimanual folding of garments. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8. IEEE, 2022.

[2] Arpit Bahety, Shreeya Jain, Huy Ha, Nathalie Hager, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Bag all you need: Learning a generalizable bagging strategy for heterogeneous objects. *IROS*, 2023.

[3] Alper Canberk, Cheng Chi, Huy Ha, Benjamin Burchfiel, Eric Cousineau, Siyuan Feng, and Shuran Song. Cloth funnels: Canonicalized-alignment for multi-purpose garment manipulation. In *International Conference of Robotics and Automation (ICRA)*, 2022.

[4] Lawrence Yunliang Chen, Baiyu Shi, Daniel Seita, Richard Cheng, Thomas Kollar, David Held, and Ken Goldberg. Autobag: Learning to open plastic bags and insert objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3918–3925. IEEE, 2023.

[5] Cheng Chi and Shuran Song. Garmentnets: Category-level pose estimation for garments via canonical space shape completion. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[6] Yushi Du, Ruihai Wu, Yan Shen, and Hao Dong. Learning part motion of articulated objects using spatially continuous neural implicit representations. In *British Machine Vision Conference (BMVC)*, November 2023.

[7] Peter Florence, Lucas Manuelli, and Russ Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. *Conference on Robot Learning*, 2018.

[8] Antonio Gabas and Yasuyo Kita. Physical edge detection in clothing items for robotic manipulation. In *2017 18th International Conference on Advanced Robotics (ICAR)*, pages 524–529. IEEE, 2017.

[9] Aditya Ganapathi, Priya Sundaresan, Brijen Thananjeyan, Ashwin Balakrishna, Daniel Seita, Jennifer Grannen, Minho Hwang, Ryan Hoque, Joseph E Gonzalez, Nawid Jamali, et al. Learning dense visual correspondences in simulation to smooth and fold real fabrics. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 11515–11522. IEEE, 2021.

[10] Oshri Halimi, Or Litany, Emanuele Rodola, Alex M Bronstein, and Ron Kimmel. Unsupervised learning of dense shape correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4370–4379, 2019.

[11] Ling Hu, Qinsong Li, Shengjun Liu, and Xinru Liu. Efficient deformable shape correspondence via multiscale spectral manifold wavelets preservation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14536–14545, 2021.

[12] Cheng-I Lai. Contrastive predictive coding based feature for automatic speaker verification. *arXiv preprint arXiv:1904.01575*, 2019.

[13] Xingyu Lin, Yufei Wang, Zixuan Huang, and David Held. Learning visible connectivity dynamics for cloth smoothing. In *Conference on Robot Learning*, 2021.

[14] Kaichun Mo, Leonidas J Guibas, Mustafa Mukadam, Abhinav Gupta, and Shubham Tulsiani. Where2act: From pixels to actions for articulated 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6813–6823, 2021.

[15] Daniel Seita, Pete Florence, Jonathan Tompson, Erwin Coumans, Vikas Sindhwani, Ken Goldberg, and Andy Zeng. Learning to rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4568–4575. IEEE, 2021.

[16] Ruoxi Shi, Zhengrong Xue, Yang You, and Cewu Lu. Skeleton merger: an unsupervised aligned keypoint detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 43–52, 2021.

[17] Matthias Vestner, Zorah Lähner, Amit Boyarski, Or Litany, Ron Slossberg, Tal Remez, Emanuele Rodola, Alex Bronstein, Michael Bronstein, Ron Kimmel, et al. Efficient deformable shape correspondence via kernel matching. In *2017 international conference on 3D vision (3DV)*, pages 517–526. IEEE, 2017.

[18] Yufei Wang, Zhanyi Sun, Zackory Erickson, and David Held. One policy to dress them all: Learning to dress people with diverse poses and garments. In *Robotics: Science and Systems (RSS)*, 2023.

[19] Ruihai Wu, Chuanruo Ning, and Hao Dong. Learning foresightful dense visual affordance for deformable object manipulation. In *IEEE International Conference on Computer Vision (ICCV)*, 2023.

[20] Ruihai Wu, Yan Zhao, Kaichun Mo, Zizheng Guo, Yian Wang, Tianhao Wu, Qingnan Fan, Xuelin Chen, Leonidas Guibas, and Hao Dong. VAT-mart: Learning visual action trajectory proposals for manipulating 3d ARTiculated objects. In *International Conference on Learning Representations*, 2022.

[21] Yilin Wu, Wilson Yan, Thanard Kurutach, Lerrel Pinto, and Pieter Abbeel. Learning to manipulate deformable objects without demonstrations. In *16th Robotics: Science and Systems, RSS 2020*. MIT Press Journals, 2020.

[22] Chi Xu, Lakshmi Narasimhan Govindarajan, Yu Zhang, and Li Cheng. Lie-x: Depth image based articulated object pose estimation, tracking, and action recognition on lie groups. *International Journal of Computer Vision*, 123:454–478, 2017.

[23] Han Xue, Yutong Li, Wenqiang Xu, Huanyu Li, Dongzhe Zheng, and Cewu Lu. Unifolding: Towards sample-efficient, scalable, and generalizable robotic garment folding. In *7th Annual Conference on Robot Learning*, 2023.

[24] Han Xue, Wenqiang Xu, Jieyi Zhang, Tutian Tang, Yutong Li, Wenxin Du, Ruolin Ye, and Cewu Lu. Garmenttracking: Category-level garment pose tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21233–21242, 2023.

[25] Bingyang Zhou, Haoyu Zhou, Tianhai Liang, Qiaojun Yu, Siheng Zhao, Yuwei Zeng, Jun Lv, Siyuan Luo, Qiancai Wang, Xinyuan Yu, Haonan Chen, Cewu Lu, and Lin Shao. Clothesnet: An information-rich 3d garment model repository with simulated clothes environment. *ICCV*, 2023.

[26] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 512–530. Springer, 2020.