GarmentPile: Point-Level Visual Affordance Guided Retrieval and Adaptation for Cluttered Garments Manipulation

Ruihai Wu[1*] Ziyu Zhu[2,1*] Yuran Wang[1*] Yue Chen[1] Jiarui Wang[1] Hao Dong[1†]
[1]CFCS, School of Computer Science, PKU [2]School of EECS, PKU

***Abstract*— Cluttered garments manipulation poses significant challenges due to the complex, deformable nature of garments and intricate garment relations. Unlike single-garment manipulation, cluttered scenarios require managing complex garment entanglements and interactions, while maintaining garment cleanliness and manipulation stability. To address these demands, we propose to learn point-level affordance, the dense representation modeling the complex space and multi-modal manipulation candidates, while being aware of garment geometry, structure, and inter-object relations. Additionally, as it is difficult to directly retrieve a garment in some extremely entangled clutters, we introduce an adaptation module, guided by learned affordance, to reorganize highly-entangled garments into states plausible for manipulation. Our framework demonstrates effectiveness over environments featuring diverse garment types and pile configurations in both simulation and the real world. Project page: https://garmentpile.github.io/.**

I. Introduction

Garments, such as shirts, dresses, and socks, are essential in daily life and pose significant challenges for human-assistive robots. Most studies focus on single-garment manipulation, such as unfolding [3], folding [11] and hanging [8]. However, many real-life scenarios involve multiple cluttered garments, such as arranging clothes on a bed or retrieving items from a washing machine. In these cases, it is crucial to maintain cleanliness and avoid disturbing adjacent garments (failure cases in Figure 1).

Manipulating cluttered garments presents greater challenges than single-garment due to the complex states in the clutters and the complicated interrelations between garments. Moreover, garment piles often involve multiple plausible retrieval garments (Figure 1, *row 1*), further increasing the demands on the multi-modal representation capability of the learned manipulation policy.

Point-level affordance, derived from 3D point cloud input and representing the **per-point actionability** on the object for downstream tasks, is a suitable representation for cluttered garments manipulation. First, the per-point space supports representing complex states of cluttered scenes. Also, the per-point score can easily represent the multi-modal policy outputs (Figure 1, *row 1*). Most importantly, the feature of each point is extracted from local to global, capable of representing the local geometry information for grasping, the structural information of each garment, and the interrelations between garments (Figure 1, *row 2*). For unseen garment clutters, the above extracted information (garment geometry, structure and relations) is consistent across scenes, making the representation easily generalize to novel scenarios.

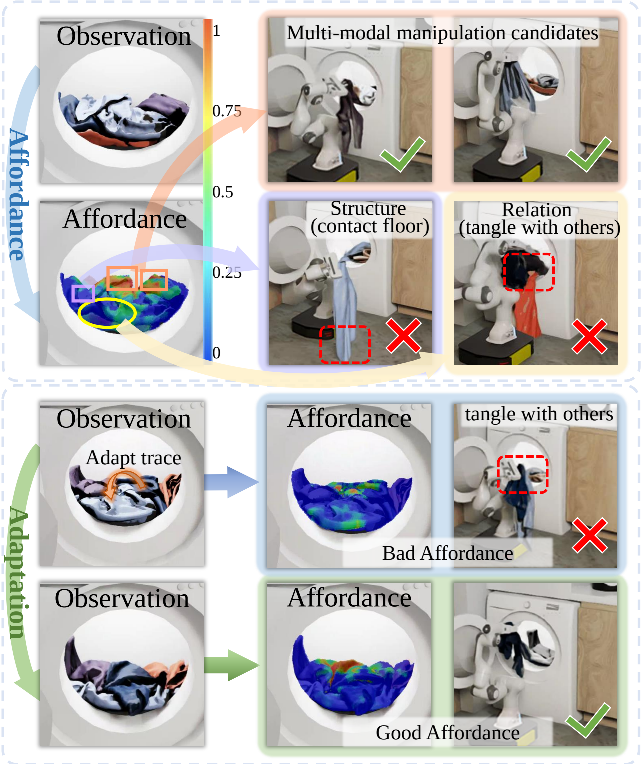

Fig. 1: **Point-Level Affordance for Cluttered Garments**. A higher score denotes the higher actionability for downstream retrieval. ***Row 1***: per-point affordance simultaneously reveals 2 garments suitable for retrieval. ***Row 2***: it is aware of garment structures (grasping edges leads other parts contacting floor) and relations (retrieving one garment while dragging nearby entangled garments out), and thus avoids manipulating on points leading to such failures. ***Row 3* and *4***: highly tangled garments may not have plausible manipulation points, affordance can guide reorganizing the scene, and thus garments plausible for manipulation will exist.

However, affordance alone is not a universal solution. In extreme cases, such as highly tangled garments, there may not exist manipulatable positions (Figure 1, *row 3*), and thus need the robot to first reorganize the garments to a new state plausible for manipulation (Figure 1, from *row 3* to *row 4*). Therefore, we further introduce a novel adaptation module. By iteratively executing the pick-and-place actions, the adaptation module can use learned point-level affordance as the signal to efficiently reorganize cluttered garments.

The absence of suitable simulation environments also partly obstructs the research on cluttered garment manipulation. Previous works have primarily focused on the simulation and manipulation of single garments or simpler deformable objects [1], [6], [7], [9], [10], [12], [13], rather than tackling the challenges posed by cluttered scenarios. To address this, we propose a new evaluation environment based on GarmentLab [4], including 9 garment categories with various deformations and 3 representative scenarios: sofa, washing machine, and basket. Both qualitative and quantitative results from simulations and real-world experiments demonstrate the effectiveness of our framework.

In conclusion, our contributions mainly include:

- We propose to study the novel task of cluttered garments manipulation, and build the pioneering environment with diverse scenarios covering different garment categories.
- We introduce point-level affordance learning for cluttered garments manipulation, with multiple novel designs to efficiently represent highly complex state and action spaces, and multi-modal policy outputs.
- We further develop the adaptation module guided by learned affordance, to efficiently adapt the cluttered garments to states easy to successfully manipulate.

II. Problem Formulation

Given a clutter of k garments and its 3D point cloud observation $O \in \mathbb{R}^{N\times 3}$, we study garment retrieval, aiming to retrieve k garments one-by-one while avoiding 2 common issues that may lead to uncleanliness or unsafety:

- The target garment contacts the floor during the retrieval. (Figure 1, *row 2*, *column 2*).
- When the retrieving one garment, others are dragged out. (Figure 1, *row 2*, *column 3*).

We use pick-and-place as action primitive. We use the grasp point $p_{retrieve} \in \mathbb{R}^3$ with heuristic retrieval orientation as **retrieval action**. In case plausible retrieval garment is not available, we use pick-and-place action ($p_{pick} \in \mathbb{R}^3$ and $p_{place} \in \mathbb{R}^3$) as **adaptation action** to reorganize the scene.

We define point-level retrieval / pick / place affordance maps $A^{retrieve}$, A^{pick}, $A^{place}_{p_{pick}} \in \mathbb{R}^N$, each digit normalized to [0,1], indicating per-point actionability for retrieval / pick / place. The point with highest score will be selected.

III. Method

A. Point-Level Affordance for Retrieval

As described in Introduction and Problem Formulation, the Retrieval Affordance Module (also denoted as Affordance Module for simplicity) $\mathcal{M}_{retrieve}$ predicts the per-point score map $\mathcal{A}^{retrieve}$ for each point. Taking as input the point cloud observation O of the garment clutter, we extract the per-point feature using PointNet++ [5] backbone feature extractor $\mathbf{F}_{retrieve}$. The per-point feature of PointNet++ aggregates the information of local geometry, global structure and garment relations, each of which is essential for predicting whether the manipulation on the target point will succeed. For the point p, we get the feature $f_p^{retrieve} \in \mathbb{R}^{128}$, and parse it into Multi-Layer Perceptrons (MLPs) with sigmoid [2] activation function for normalization, we can get 1-dimension affordance prediction $\hat{g}_p^{retrieve}$ on p. We define the ground truth retrieval affordance score $g_p^{retrieve}$ on p as 1 (success) or 0 (failure), by directly executing the retrieval action on p and acquiring the manipulation result. We use Binary Cross Entropy (BCELoss) $L_{retrieve}$ to calculate the loss.

With trained $\mathcal{M}_{retrieve}$, given the 3D point cloud observation $O \in \mathbb{R}^{N\times 3}$, we can first infer the point-level retrieval affordance map $A^{retrieval} \in \mathbb{R}^N$ and select the point $p_{retrieval}$ with the highest score for the retrieval action.

B. Retrieval Affordance Guided Adaptation

Garments in clutters might be highly entangled, making it difficult to retrieve one garment without disturbing others in some situations, where all points would have low affordance scores, indicating that no point could be manipulated. To deal with this situation, people often reorganize the garments (by picking-placing or stirring), until finding a plausible scene where the subsequent manipulation could be successful. Therefore, we mimic what people often do and propose the adaptation module by iteratively executing the pick-and-place actions to reorganize the scene. The construction of the adaptation module depends on the learned affordance, as it indicates whether the scene is plausible for manipulation.

As pick-and-place composites a large action space, which is difficult for learning, we divide each adaptation action into first predicting the pick point p_{pick} and then the place point p_{place} conditioned on p_{pick}. As the state after placing can be estimated by the learned manipulation (retrieval) affordance, we first learn the actionability for placing on each point given a specific pick point, supervised by the retrieval affordance after the execution of (p_{pick}, p_{place}) action (Section III-B.1). Then, with the learned place affordance for adaptation, we learn the affordance for p_{pick} (Section III-B.2).

1) Place Affordance: The adaptation action is composed of a pick point p_{pick} and a place point p_{place}. With a pick point p_{pick}, the Place Affordance Module M_{place} rates the actionability of each point p on whether placing p_{pick} on p will improve the scene. The pick action p_{pick} is difficult to directly get supervision, due to the diversity of the following place actions. On the contrary, the place action p conditioned on p_{pick} can get the direct feedback from the adapted scene by checking the scene actionability (*i.e.*, retrieval affordance) improvement. Therefore, we first train M_{place}.

For a target place point p, given as input the 3D point cloud O and p_{pick}, two PointNet++, $\mathbf{F}^1_{place}$ and $\mathbf{F}^2_{place}$, respectively extracts the point feature $f^{place_1}_{p_{pick}}$ and $f^{place_2}_p$. Then, their feature concatenation is parsed into MLPs to predict the place affordance $\hat{g}^{place}_{p|p_{pick}}$ normalized to [0, 1]. We execute the pick-and-place action from p_{pick} to p and get the new point cloud O', with the new affordance map. If the new affordance map exceeds the initial one by a margin, the ground truth place affordance $g^{place}_{p|p_{pick}}$ is set as 1 otherwise 0. We use BCELoss L_{place} to calculate the loss between $g^{place}_{p|p_{pick}}$ and $\hat{g}^{place}_{p|p_{pick}}$.

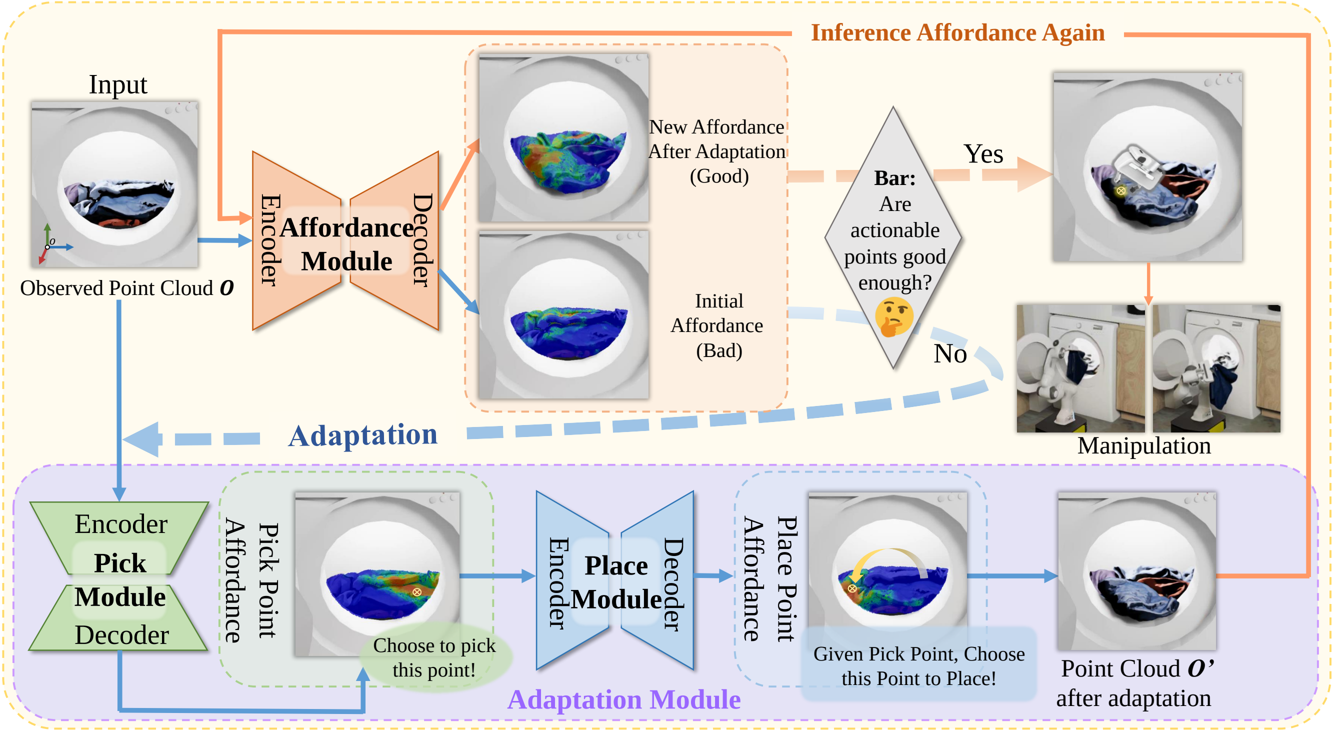

Fig. 2: **Framework Overview.** Given the observed point cloud, the Affordance Module predicts the initial point-level manipulation (retrieval) affordance score. When actionability is not good enough, the framework proposes the adaptation pick-place action. It first predicts per-point pick affordance, and selects the pick point with the highest score, conditioned on which it predicts place affordance and selects the place point. After executing adaptation action, it receives a new point cloud and generates new affordance. When actionability is good enough, the robot retrieves on the point with the highest affordance score. This loop is executed until all garments are retrieved.

With the trained $\mathcal{M}_{place}$, given the 3D point cloud observation $O \in \mathbb{R}^{N\times3}$ and a specific p_{pick}, we can infer the point-level place affordance map $A^{place}_{p_{pick}} \in \mathbb{R}^N$ and select the point p_{place} with the highest score for the place action.

2) Pick Affordance: With a stable Place Module that rates the actionability for each place point conditioned on any pick point, we can further train the Pick Module that rates the actionability g^{pick}_p for each point p, supervised by the best following place action conditioned on p.

Taking as input the point cloud observation O of the garment clutter, we extract the per-point feature using PointNet++ backbone feature extractor $\mathbf{F}_{pick}$. For the point p, we get the feature $f^{pick}_p \in \mathbb{R}^{128}$, and parse it into MLPs to predict the pick affordance $\hat{g}^{pick}_p$ normalized to [0, 1]. To get the groud truth score of p, we use $\mathcal{M}_{place}$ to find the most suitable p_{place} corresponding to p, and then execute the pick-and-place action from p to p_{place} to get the new point cloud O', with the new affordance map. If the new affordance map exceeds the initial one by a margin, the ground truth pick affordance g^{pick}_p is set as 1 otherwise 0. We use BCELoss L_{pick} to calculate the loss between g^{pick}_p and $\hat{g}^{pick}_p$.

With the trained $\mathcal{M}_{pick}$, given the 3D point cloud observation $O \in \mathbb{R}^{N\times3}$, we can infer the point-level pick affordance map $A^{pick} \in \mathbb{R}^N$ and select the point p_{pick} with the highest score for the pick action.

C. Inference and Training Details

Figure 2 (caption) describes **inference** pipeline and details.

For **training data, epoches and computing resources**, we use NVIDIA GeForce 4090 for training. We set batch size to be 128 to train (Retrieval) Affordance and Pick Affordance. While for Place Affordance, we set batch size to be 64 because there are two PointNet++ networks. We collect 20,000 pieces of data and train Retrieval Affordance for 120 epoches, as well as 8,000 pieces of data and train Pick and Place Affordance for 80 epoches. It takes fewer than 24 hours to train each module.

We further use **online data** to boost the robustness during **training**. Since cluttered garments have exceptionally diverse states, the model trained on offline data might not work well in some unseen clutters. Therefore, based on the offline trained models, we first gather a set amount of online data, and then combine it with an equal amount of offline data to train the model. This approach enhances the model's robustness while preserving the knowledge learned from the offline data.

We iterate this process as the sampled mistake distributions might have changes, until the model shows consistent performance with low variance. As we can acquire the manipulation or adaptation execution results for each module, the online adaptation proceeded for all modules.

References

[1] 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, 2023. ii

[2] Stefan Elfwing, Eiji Uchibe, and Kenji Doya. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *CoRR*, 2017. ii

[3] Huy Ha and Shuran Song. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. In *Conference on Robot Learning*. PMLR, 2022. i

[4] Haoran Lu, Ruihai Wu, Yitong Li, Sijie Li, Ziyu Zhu, Chuanruo Ning, Yan Shen, Longzan Luo, Yuanpei Chen, and Hao Dong. Garmentlab: A unified simulation and benchmark for garment manipulation. In *Advances in Neural Information Processing Systems*, 2024. ii

[5] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. ii

[6] Vedant Raval, Enyu Zhao, Hejia Zhang, Stefanos Nikolaidis, and Daniel Seita. Gpt-fabric: Folding and smoothing fabric by leveraging pre-trained foundation models. *arXiv preprint arXiv:2406.09640*, 2024. ii

[7] 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 *ICRA*, 2021. ii

[8] Ruihai Wu, Haoran Lu, Yiyan Wang, Yubo Wang, and Hao Dong. Unigarmentmanip: A unified framework for category-level garment manipulation via dense visual correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2024. i

[9] 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. ii

[10] Yilin Wu, Wilson Yan, Thanard Kurutach, Lerrel Pinto, and Pieter Abbeel. Learning to manipulate deformable objects without demonstrations. *RSS*, 2020. ii

[11] Han Xue, Yutong Li, Wenqiang Xu, Huanyu Li, Dongzhe Zheng, and Cewu Lu. Unifolding: Towards sample-efficient, scalable, and generalizable robotic garment folding. In *CoRL*, 2023. i

[12] Kaifeng Zhang, Baoyu Li, Kris Hauser, and Yunzhu Li. Adaptigraph: Material-adaptive graph-based neural dynamics for robotic manipulation. In *RSS*, 2024. ii

[13] Sun Zhaole, Jihong Zhu, and Robert B Fisher. Dexdlo: Learning goal-conditioned dexterous policy for dynamic manipulation of deformable linear objects. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 16009–16015. IEEE, 2024. ii