Learning to Hang Crumpled Garments with Confidence-Guided Grasping and Active Perception

Shengzeng Huo, He Zhang, Hoi-Yin Lee, Peng Zhou, and David Navarro-Alarcon

***Abstract*— In this study, we concentrate on the task of hanging crumpled garments on a rack. This context presents two primary challenges: (1) perceiving and grasping the structural regions of garments that exhibit severe deformations and self-occlusions; (2) adjusting the configuration of garments to fit the supporting components of the rack. We propose a confidence-guided grasping strategy that actively seeks garment collars through handovers between dual robotic arms. The exact grasping pose is determined through depth-aware contour extraction, and its success is evaluated based on a specially designed metric. Furthermore, we formulate the hanging task as one-shot imitation learning with an egocentric view. We propose a two-step hanging strategy that involves coarse approaching followed by fine transformation. We perform comprehensive experiments and show that our framework notably enhances the success rate compared to existing methods.**

I. Introduction

Manipulating garments is challenging due to their infinite state spaces and complex dynamics [1]. Most studies [2]–[4] make strong assumptions about task specifications. such as ideal pre-grasping [4]–[6] and nearly flattened configuration [7], [8]. However, garments often undergo severe self-occlusions due to their deformable nature, complicating accurate state estimation. In addition, hanging a garment on a rack is a common household scenario. Current methods [7], [9] assume prior knowledge of the rack's positions, which limits their practical applications. Thus, we seek to address the task of autonomously hanging a crumpled, collared garment on a rack while imposing only minimal assumptions regarding either the garment's or the rack's configuration.

In this work, we introduce a novel system for robust garment manipulation, utilizing wrist cameras for active perception, as shown in Fig. 1. First, handovers between the dual arms are employed to actively locate the collar of garments with a learned detection model, thereby facilitating a confidence-guided grasping strategy. Second, a two-step active sensing strategy is implemented to adjust the configuration of the grasped garment for alignment with the rack, allowing for the reproduction of the demonstrated interaction trajectory afterward. Extensive real-world experiments validate the robustness and superiority of our methods.

Shengzeng Huo, Ho-Yin Lee, David Navarro-Alarcon are with the Faculty of Engineering, The Hong Kong Polytechnic University, Hong Kong.

He Zhang is with the Tencent Robotics X, China.

Peng Zhou is with the Department of Computer Science, The University of Hong Kong, Hong Kong.

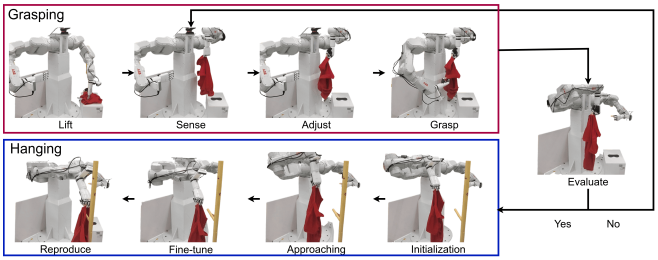

Fig. 1. The complete pipeline for the garment hanging task.

II. Methods

A. Confidence-guided Grasping

Dual arms handover with random grasping position to capture diverse garment configurations. Fig. 2(a) illustrates the autonomous labeling process. The masked depth image D and the collar mask $\mathcal{M}_C$ are fed into Yolo-v8 [10] model for detection. During deployment, the model f_D takes images of garments as input and outputs the bounding box B_i and the confidence score S_i, as illustrated in Fig. 2(b). The angle with the highest confidence is selected for the following grasping pose determination.

We develop a depth-aware strategy for computing the grasping pose, as shown in Fig. 2(c). The principle is to identify the collar's contour and designate the center of this contour as the grasping position. First, we establish a planar polar coordinate frame in the cropped depth image. Then, the pixels with minimum depth are extracted in each partitioned patches. Identifying the largest interval, we find out the center of the contour as the grasp position.

The grasping success is evaluated through comparing the depth value between each element $q_j \in Q$ and the center C of bounding box D'. The geometry interpretation of this evaluation is to determine whether the hole structure has been grasped appropriately.

B. Two-step Hanging

The hanging task is to align the current pose to the recorded situation in the demonstration. However, directly learning the 6-DoF transformation is challenging due to the real-world measurement noise and the coupling effect between translation and rotation. Thus, we propose a two-step reaching strategy that consists of a coarse approaching f_C and a fine alignment f_A, as shown in Fig. 3. Firstly, we adjust the camera's viewpoint to effectively capture the key structure of the supporting item. Secondly, the rack's

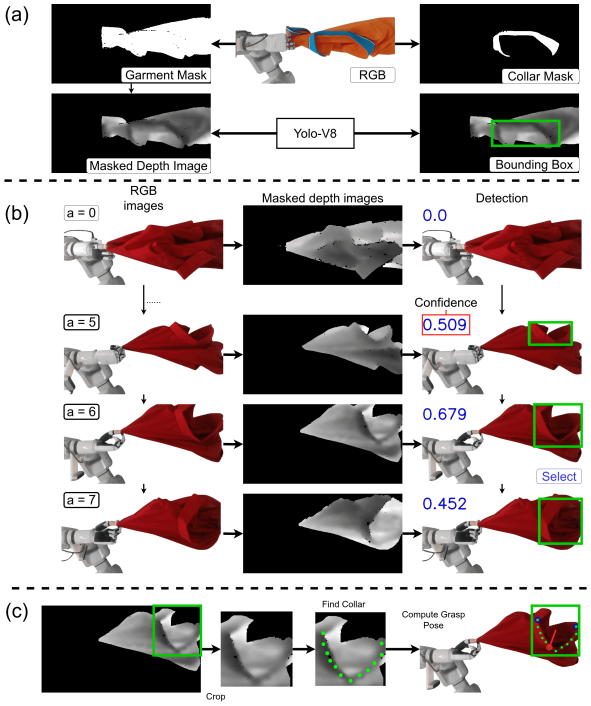

Fig. 2. The details of the confidence-guided grasping strategy. (a) The data processing pipeline. (b) The wrist camera captures several images from different angles. The angle with highest confidence is selected for grasping. (c) The collar region is cropped to determine the grasping pose.

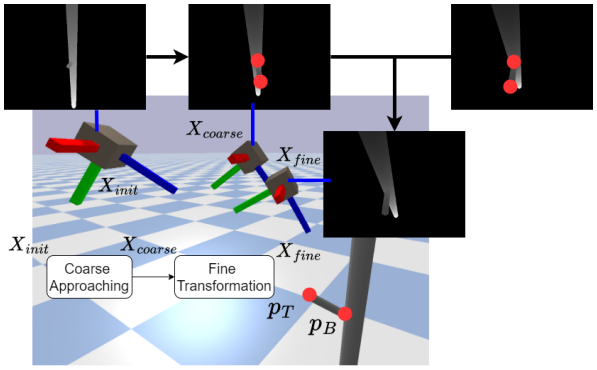

Fig. 3. The procedures of the two-step hanging alignment.

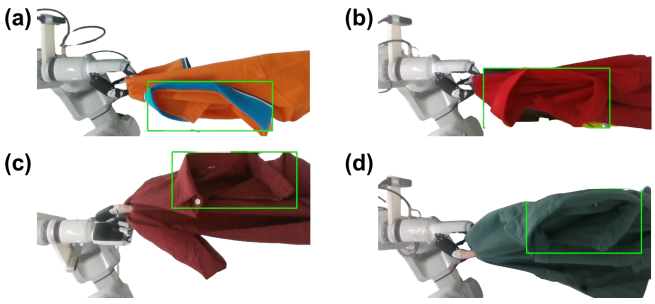

Fig. 4. The typical examples of collar detection across various kinds of garments. (a) TPL. (b) NPL. (c) SS. (d) LS.

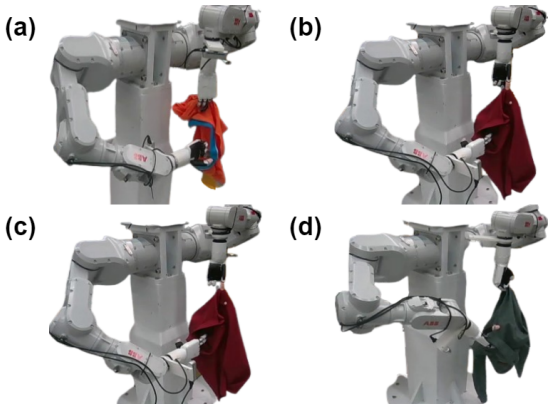

Fig. 5. The typical examples of grasping across various kinds of garments. (a) TPL. (b) NPL. (c) SS. (d) LS.

TABLE I

PERCEPTION EVALUATION RESULTS OF SEEN AND UNSEEN GARMENTS

	TPL	NPL	SS	LS
Precision	0.957	0.762	0.750	0.667
Recall	0.815	0.800	0.750	1.000
Fitness	0.880	0.781	0.750	0.800
μ_{iou}	0.832	0.694	0.751	0.715
σ_{iou}	0.130	0.130	0.074	0.151

μ_{iou} and σ_{iou} are the mean and variance of IoU.

keypoints are detected to minimize the position error with the demonstration.

We generate a synthetic dataset in simulation [11], covering various observations of the rack, as shown in Fig. 3. For the coarse model, the label corresponds to the relative displacement to percept the rack. For the fine model, the label corresponds to the keypoints of the rack. Since the supporting point and direction are critical in the hanging task [12], the endpoint of the supporting part and the connection point to the standing part are chosen. In deployment, the camera first moves to a new position guided by the coarse model. Then, the keypoints are detected based on the fine model and used for pose alignment. The cost function C in the minimization process consists of two components: (1) the distance error associated with the keypoint of the base p_B; (2) the directional error of the vector from the base keypoint p_B to the tip keypoint p_T, computed by f_V. In addition to the alignment error, we also regularize the rotation.

Following the coarse-to-fine alignment strategy, the end-effector replicates the interaction trajectory demonstrated.

III. RESULTS

A. Recognition

We include three kinds of shirts to assess the robustness of the recognition model, as shown in Fig. 4. Table I presents the optimal performance regarding fitness in relation to confidence. These findings indicate that our algorithm is resilient across various garments. Typically, there are two typical recognition failures: 1) the model occasionally misidentifies the sleeve as the collar; and 2) it sometimes fails to detect collars with a small area.

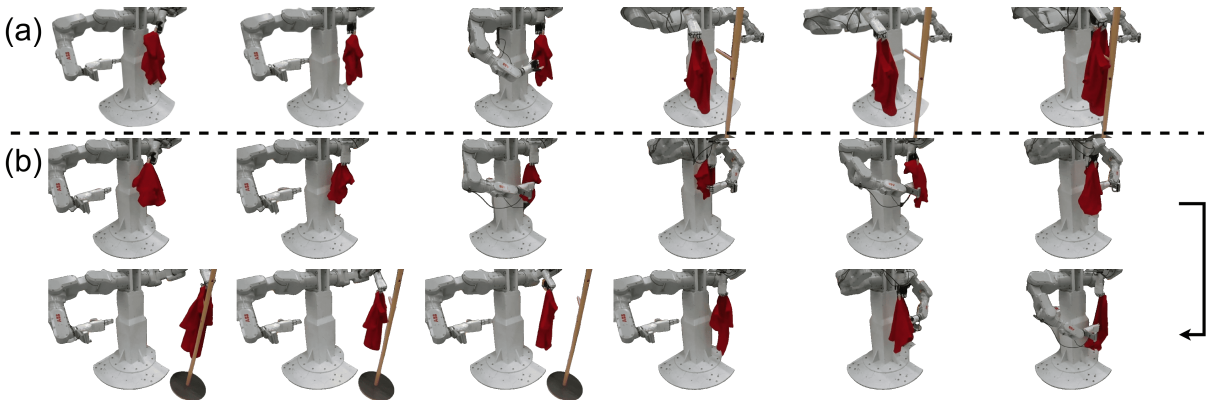

Fig. 6. Two typical examples of the complete pipeline.

TABLE II
SUCCESS RATE FOR GRASPING PHASE

	TPL	NPL	SS	LS
GCSR [9]	53.3	46.7	33.3	36.7
GPGM [14]	46.7	46.7	26.7	20.0
Ours	**93.3**	**93.3**	**90.0**	**86.7**

TABLE III
SUCCESS RATE FOR HANGING PHASE

	Left (%)	Right (%)	All (%)
Dinobot [15]	36.7	40.0	38.3
KOVIS [16]	33.3	30.0	31.7
Ours	**90.0**	**93.3**	**91.7**

TABLE IV
COMPLETE PIPELINE EVALUATION RESULTS

Close-loop grasping	Two-step hanging	Success Rate (%)
✓	×	70.0
×	✓	36.7
✓	✓	**83.3**

B. Grasping

The grasping process starts with crumpled configurations [13] and proceeds until either the evaluation criteria are satisfied or the maximum number of handovers is reached. Two typical baselines (**GCSR** [9] and **GPGM** [14]) are involved in the comparisons. Success is defined as instances where the collar regions are grasped and lifted stably in space. As shown in Table II, our confidence-guided grasping strategy outperform the other baseline methods. **GCSR** operates at the pixel level, requiring a significant amount of training data. **GPGM** identifies the grasping position through supervised learning, ignoring the structural regions of the garment.

Our enhanced performance can be attributed to the active search for the structural region and compensating for false detection through closed-loop evaluation. Fig. 5 illustrates several successful examples across various kinds of garments. The grasping pose is designed to insert into the hole of the collar to achieve a stable grasp. There are two kinds of common failures, including incorrect detection and wrong feedback from the evaluation.

C. Hanging

To assess the effectiveness of our proposed method, we implement two standard baseline approaches (**DINO** [15]) and **KOVIS** [16]). Success is only counted when the garment remains stably positioned on the rack after the hands are released. As shown in Table III, our two-step hanging strategy outperforms the other baselines. Both **DINO** and **KOVIS** perform the task using a single-step alignment, which restricts their effectiveness to situations where the initial configuration is close to the desired pose. There are generally two common failure modes in the hanging process: 1) the coarse model may output a displacement vector that exceeds the arm's reachability; and 2) the keypoint detection provided from the fine model may be inaccurate due to measurement noise from the depth camera.

D. Ablation Study

We examine the contributions of the confidence-guided grasping and the two-step hanging strategy. We eliminate the close-loop evaluation during the grasping phase and the coarse approaching stage in the hanging phase respectively. As shown in Table IV, our algorithm shows a lower success rate when either of the key modules is removed. On one hand, without the closed-loop evaluation, the arm occasionally grasps the incorrect region of the garment. On the other hand, when the coarse approaching stage is omitted, the arm struggles to achieve the desired predefined pose in certain challenging scenarios.

Two typical complete episodes are shown in Fig. 6. Handovers are terminated when the grasp is successful. After grasping the collar, the hanging algorithm guides the end-effector to approach and interact with the rack. One limitation of our algorithm is that a random element of the garment is selected for grasping during each handover. Taking into account the complete structure of the garment to determine the optimal handover point could potentially expedite the search for the collar in challenging scenarios [17].

REFERENCES

[1] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, 2021.

[2] H. Ha and S. Song, "Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding," in *Proc. Conf. Robot Learn.*, 2021.

[3] T. Weng, S. Bajracharya, Y. Wang, K. Agrawal, and D. Held, "Fabricflownet: Bimanual cloth manipulation with a flow-based policy," *ArXiv*, vol. abs/2111.05623, 2021.

[4] Y. Wang, Z. Sun, Z. Erickson, and D. Held, "One policy to dress them all: Learning to dress people with diverse poses and garments," in *Robotics: Science and Systems (RSS)*, 2023.

[5] F. Zhang, A. Cully, and Y. Demiris, "Probabilistic real-time user posture tracking for personalized robot-assisted dressing," *IEEE Trans. on Robotics*, vol. 35, no. 4, pp. 873–888, 2019.

[6] J. Zhu, M. Gienger, G. Franzese, and J. Kober, "Do you need a hand? – a bimanual robotic dressing assistance scheme," *IEEE Trans. on Robotics*, vol. 40, pp. 1906–1919, 2024.

[7] R. Wu, H. Lu, Y. Wang, Y. Wang, and H. Dong, "UniGarmentManip: A unified framework for category-level garment manipulation via dense visual correspondence," *ArXiv*, vol. abs/2405.06903, 2024.

[8] T. Lips, V.-L. De Gusseme *et al.*, "Learning keypoints for robotic cloth manipulation using synthetic data," *IEEE Robot. Autom. Lett.*, 2024.

[9] W. Chen, D. Lee, D. Chappell, and N. Rojas, "Learning to grasp clothing structural regions for garment manipulation tasks," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 4889–4895, 2023.

[10] J. Redmon, S. K. Divvala, R. B. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016.

[11] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," http://pybullet.org, 2016–2021.

[12] C.-L. Kuo, Y.-W. Chao, and Y.-T. Chen, "Skt-hang: Hanging everyday objects via object-agnostic semantic keypoint trajectory generation," *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 15 433–15 439, 2024.

[13] I. Garcia-Camacho, M. Lippi, M. C. Welle, H. Yin, R. Antonova, A. Varava, J. Borras, C. Torras, A. Marino, G. Alenyà, and D. Kragic, "Benchmarking bimanual cloth manipulation," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 1111–1118, 2020.

[14] F. Zhang and Y. Demiris, "Learning grasping points for garment manipulation in robot-assisted dressing," *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 9114–9120, 2020.

[15] N. Di Palo and E. Johns, "Dinobot: Robot manipulation via retrieval and alignment with vision foundation models," *arXiv preprint arXiv:2402.13181*, 2024.

[16] E. Y. Puang, K. P. Tee, and W. Jing, "Kovis: Keypoint-based visual servoing with zero-shot sim-to-real transfer for robotics manipulation," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 7527–7533, 2020.

[17] S. Huo, F. Hu, F. Wang, L. Hu, P. Zhou, J. Zhu, H. Wang, and D. Navarro-Alarcon, "Rearranging deformable linear objects for implicit goals with self-supervised planning and control," *Advanced Intelligent Systems*, vol. 7, no. 2, p. 2400330, 2025.