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

Shengzeng Huo, He Zhang, Hoi-Yin Lee, Peng Zhou, *Member, IEEE,* Hesheng Wang, *Senior Member, IEEE,* and David Navarro-Alarcon, *Senior Member, IEEE*

***Abstract*—Accurately recognizing the structural regions of targeted objects is crucial for successful manipulation. In this study, we concentrate on the task of hanging crumpled garments on a rack, a common scenario in household environments. 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. To address these challenges, we propose a confidence-guided grasping strategy that actively seeks garment collars through handovers between dual robotic arms. In particular, we develop an autonomous data collection procedure in real-world settings to train the collar detection network. 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. To precisely align the collar with the supporting item, we propose a two-layered 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.**

***Index Terms*—Deformable object manipulation, Hanging garments, Active perception, One-shot imitation learning**

I. Introduction

Garments are ubiquitous in our daily lives and have a wide range of applications [1], including folding [2], flattening [3] and assistive dressing [4]. In comparison to rigid objects, garments present significantly greater challenges for manipulation due to their extensive infinite state and action spaces, as well as their complex kinematics and dynamics [5].

Although there have been substantial body of research on garment manipulation [2]–[4], most studies make strong assumptions about task specifications. For instance, [4], [6], [7] assume an ideal pre-grasping configuration at the outset, which is maintained throughout the entire process. Similarly, [8], [9] focus on nearly flattened garments where keypoints are consistently visible. However, garments often undergo severe self-occlusions due to their deformable nature, complicating accurate state estimation. After grasping a garment, hanging it on a rack is a common scenario in household settings. Current methods [8], [10] assume prior knowledge of the rack's positions, which limits their practical applications. To enhance practicality, we aims to address the challenge of hanging a crumpled garment on a rack without relying on strong assumptions about either the garment or the rack. This task is complex, as the robot must reason about the structural regions of the environment (garments and racks) and determine the appropriate actions (grasping pose and hanging trajectory) while remaining robust to variations in their configurations.

In this work, we introduce a novel system for robust garment manipulation, utilizing eye-on-hand cameras integrated with dual-arm end-effectors for active perception. First, handovers between 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-layered 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.

Perceptual feedback is essential for manipulating garments due to their deformable characteristics [11]. Previous works detect specific garment patterns such as wrinkles [12] and corners [13]), for various applications. However, they are task-specific and lack generality. Additionally, [14] explicitly reason about occlusions to reconstruct crumpled garments' meshes. Meanwhile, [15], [16] employ a strategy of lifting the garment prior to recognizing, utilizing gravitational force to aid in untangling. However, challenges arise concerning the acquisition of high-quality data in training and the computational complexity in test-time optimization. We employ a similar strategy to alleviate the burden in structural recognition, but we focus on detecting the collar to facilitate effective grasping and hanging.

The manipulation of garments is heavily influenced by their initial configurations and the contextual factors of the scenario [10]. Studies [17], [18] identify predefined grasping points of garments hung on a rack with supervised learning, whose deformations are relatively straightforward. Additionally, [19] extracts edges and corners from crumpled fabrics to develop a grasping policy. Nevertheless, the structure of fabrics is considerably simpler than that of garments. While [10] at-

The work of Shengzeng Huo was supported by the Tencent Robotics X. This work is supported in part by the Research Grants Council (RGC) of Hong Kong under C4042-23GF, in part by the National Natural Science Foundation (NSFC) of China under 62403211, and in part by the Research Student Attachment Programme 2021/22 of The Hong Kong Polytechnic University (PolyU). *Corresponding author: David Navarro-Alarcon.*

S. Huo, is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong, and also with the Tencent Robotics X, China.

H. Zhang is with the Tencent Robotics X, China.

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

H. Wang is with the Department of Automation, Shanghai Jiaotong University, China.

H.-Y. Lee and D. Navarro-Alarcon is with the Department of Mechanical Engineering, The Hong Kong Polytechnic University, Hong Kong.

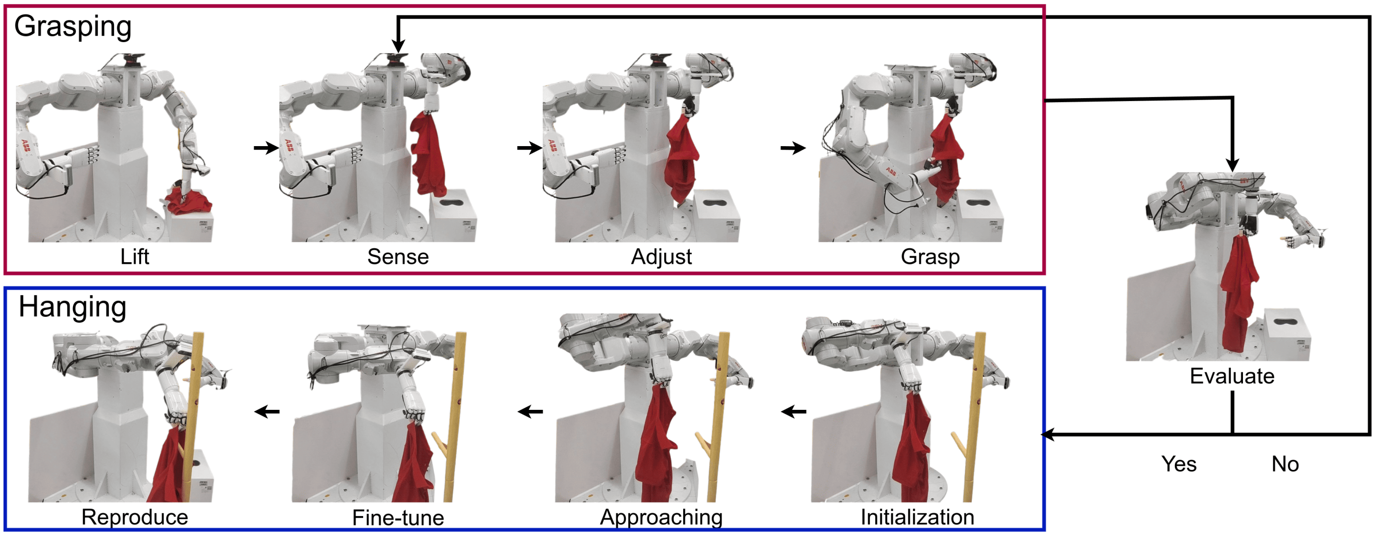

Fig. 1. The complete pipeline for the garment hanging task. **Grasping**: The dual-arm robot lifts the garment and locates its collar for grasping. **Hanging**: The robotic arm that has grasped the garment adjusts its pose to enable successful hanging.

tempts to grasp a garment's collar and hang it on a rack, they assume the collar is initially visible, which is not always true. In contrast, our work does not make any assumptions about garment configurations, thus providing a more comprehensive framework for garment manipulation.

Equipping robots with the ability to hang objects on supporting structures has applications across various domains [20], [21]. While the studies [20], [21] focus on hanging various grasped objects on diverse supports, they depend on a third-person perspective for perception in constrained scenarios. Previous work [10] and [8] address the garment hanging task using a single pick-and-place primitive, assuming the rack's position is known, which greatly restricts the applicability. In contrast, we flexibly adjust the garment's configuration based on an egocentric view to fit the rack, requiring only a rough estimation of its position.

Recently, mounting cameras on robots' wrists has gain popularity, as it allows to actively obtain multiple perspectives of the target object [22]. [23] presents a keypoint-based visual servoing method for fine robotic manipulation. Meanwhile, [24] utilizes vision foundation models to extract corresponding keypoints and apply registration to adjust end-effector's pose. However, both approaches are constrained to planar scenarios with limited degrees of freedom. To facilitate successful hanging in three-dimensional environments, it is crucial to detect contact points between objects and supports to determine their appropriate relative poses [20].

The work in [10] closely resembles our work; however it has two significant limitations: (1) it assumes that the collar of garment is initially visible; (2) its hanging trajectory is predefined. To enhance adaptability, our comprehensive algorithm exclusively utilizes egocentric views to identify the structural regions of both the garments and the supporting rack. The original contributions of this work are as follows:

- We propose a novel pipeline for garment manipulation using dual arms equipped with eye-in-hand cameras.
- We introduce a confidence-guided grasping strategy that actively searches for garments' collars.
- We propose a two-layered hanging strategy for precise alignment between the garment's collar and the rack.
- We conduct an experimental study to validate our solution for hanging garments in crumpled configurations.

The remainder of this paper is organized as follows. Sec. II presents the methods, which include the confidence-guided grasping strategy and the two-layered hanging approach. Sec. III reports the results, while Sec. IV gives the conclusions.

II. Methods

Algorithm 1: Garment Hanging

```
/* Grasping */
Lift the garment
while True do
    Capture the image set $\{I_D\}_{i=1}^{N}$
    Detect the collar $f_D(\{D_i\}_{i=1}^{N}) \rightarrow \{B_i, S_i\}_{i=1}^{N}$
    if the collar is detected then
        Rotate to the $I-th$ angle $\leftarrow$ Eq. 2
        Compute the grasping pose $\leftarrow$ Eq. 3-8
        Execute the grasping
        if the collar is grasped $\leftarrow$ Eq. 9 then
            break
    Handover
/* Hanging */
Initialize the hanging phase $X_{init} \leftarrow$ Eq. 10
Obtain the coarse displacement $\Delta P \leftarrow$ Eq. 11
Detect the keypoints $\{p_k\}_{k=1}^{2} \leftarrow$ Eq. 11
Tune the pose of the camera $R, t \leftarrow$ Eq. 13
Reproduce the interaction trajectory
```

In this section, we outline the details of our methodology. Fig. 1 depicts the complete pipeline of our system, designed to hang a crumpled garment on a rack. The pipeline, detailed in Alg. 1, is divided into two phases, grasping and hanging.

Grasping: Initially, one arm lifts the crumpled garment with a random pick point. We designate the arm grasping the garment as the master arm A_M and the other as the slave arm A_S. The wrist camera on the slave arm A_S then captures multiple images while the master arm A_M rotates, allowing for the collection of information about the garment from various angles. Next, the collar detection network scans these images to locate the collar. The master arm A_M rotates to the angle with the highest confidence, and the slave arm A_S re-senses to determine the optimal grasping pose. The phase concludes if our close-loop evaluation confirms a successful grasp; otherwise, the roles of the dual arms are switched to initiate another search for the collar.

Hanging. In accordance with the principle outlined in [25], the hanging task is formulated as achieving a user-defined pose relative to the object of interest, followed by an open-loop replay of the demonstrated end-effector trajectory. To adapt the collar with the rack in three-dimensional space, we propose a two-layered strategy to achieve the desired pose. First, we estimate a displacement to adjust the camera's viewpoint, improving the clarity of crucial information regarding the rack. Second, we identify the keypoints of the supporting items and calculate a fine-grained transformation to attain the desired relative pose. Finally, the end-effector reproduce the demonstrated interaction trajectory.

In the following, we first outline the details of the confidence-guided grasping strategy, which encompasses the autonomous data collection procedure in real-world settings, the selection of optimal grasping angles, and the determination and evaluation of depth-aware grasping pose. Next, we describe how we leverage simulation data to learn the coarse approaching and fine-grained transformation required to achieve the desired pose.

A. Confidence-guided Grasping

A large dataset is crucial for training an effective and robust deep neural network. One challenge in learning-based manipulation is data acquisition. Traditional methods for data collection demand significant human effort to alter environmental configurations and carry out manual annotations [10]. While there is an growing trend toward generating synthetic images to enhance hybrid datasets [17], inaccuracies in physical engines can create a simulation-to-reality gap [26], especially for crumpled garments.

We propose a data collection paradigm that employs handovers between dual robotic arms to capture a variety of garment configurations. Specifically, the master arm A_M adjusts to a random angle while the slave arm A_S selects a random point on the garment to grasp. This handover process continues until the data collection procedure is complete. Following the methodology described in [19], our data acquisition utilizes a template polo shirt, with the collar distinctly separated from the main body. Fig. 2(a) illustrates the process, where the garment mask $\mathcal{M}_G$ is integrated with the depth image D to eliminate irrelevant background information. The collar mask $\mathcal{M}_C$ is autonomously generated using a color extraction method. The resulting masked depth image D and the collar mask $\mathcal{M}_C$ are then fed into the Yolo-v8 [27] model for object detection.

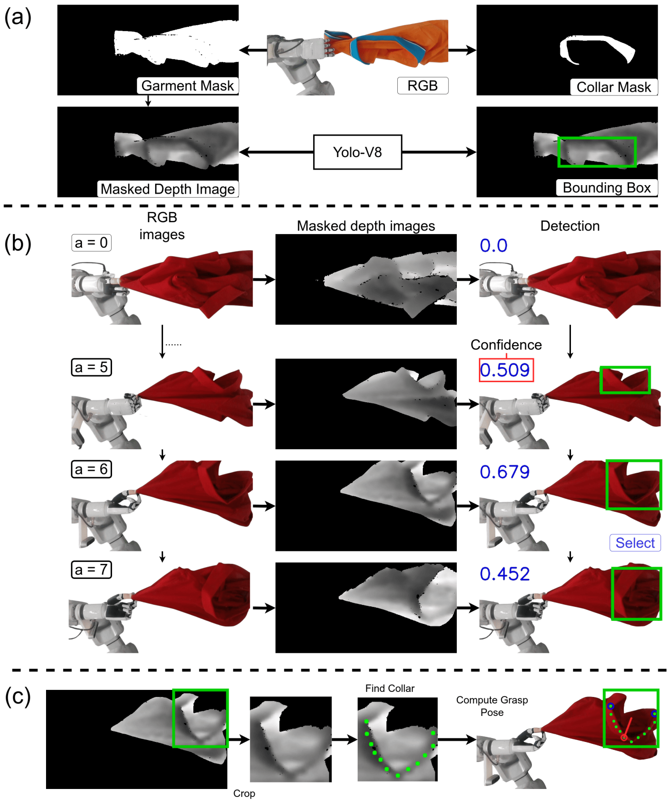

Fig. 2. The details of the confidence-guided grasping strategy. (a) The mask of the garment and its corresponding collar are extracted based on color information. The masked depth image of the garment serves as the input, while the bounding box of the collar is the output of the detection network. (b) As the master arm rotates, the camera capture several images of the garment from different angles. Among the predictions, the angle with highest confidence is selected for grasping. (c) The masked depth image is cropped according to the detection, allowing for the depth-aware collar contour extraction. Finally, the position and the orientation of the grasp are determined.

During deployment, the trained Yolo model f_D receives a masked depth image D as input and produces the corresponding bounding box B_i, along with its confidence score S_i.

$$f_D(D_i) \rightarrow (B_i, S_i) \tag{1}$$

At each step, the camera mounted on the slave arm captures N images of the garment, forming a set represented as $\{D_i\}_{i=1}^{N}$, as illustrated in Fig. 2(b). The detection network then processes these images as input and outputs the detected bounding boxes and their associated scores, denoted as $f_D(\{D_i\}_{i=1}^{N}) \rightarrow \{B_i, S_i\}_{i=1}^{N}$. From the results, we choose the angle with the highest confidence as the optimal angle for the slave arm A_S to determine the grasping pose.

$$I = \arg\max\{S_i\}_{i=1}^{N} \tag{2}$$

The master arm then move to the $I - th$ angle, and the camera mounted on the slave arm senses the environment again for grasping pose determination. A crucial insight for collar detection is to rearrange the garment's configuration to provide easier access to the collar for grasping. Consequently, we have developed a depth-aware strategy for computing the grasping pose, as shown in Fig. 2(c). The objective is to identify the collar's contour and designate the center of this contour as the grasping position.

Based on the output bounding box $B_I = f_D(D_I)$, we crop the depth image D_I to create a smaller image D'_I. Next, we set up a planar polar coordinate frame with the origin at the center C_I of D'_I. By applying an angle threshold φ, we partition the positive pixels in D'_I into several patches $\{R_j\}_{j=1}^{M}$:

$$\mathcal{R} = \{R_j | (r * \cos\varphi_j,\ r * \sin\varphi_j) \in D'_I,\ r \in [1, \infty]\}_{j=1}^{M} \quad (3)$$

where $\varphi_j = j * \varphi, M = 360/\varphi$.

For each sub-region R_j, we extract the pixels that correspond to the minimum depth:

$$Q = \{q_j = \arg\min_i D_I(u, v), (u, v) \in R_j\}_{j=1}^{M} \quad (4)$$

From the set of pixels $Q = \{q_j\}_{j=1}^{M}$, pixels with greater depth are excluded:

$$Q' = \{q_j \in Q | \frac{D_I(q_j) - \min D_I(Q)}{\max D_I(Q) - \min D_I(Q)} < \tau_Q\} \quad (5)$$

Following the processing described above, Q' includes the elements of the collar surrounding D'_I. To find the center θ_{mean}, we begin by identify the largest interval within the sorted Q' based on the corresponding φ_j values. Next, we enhance Q' by adding an additional endpoint:

$$\begin{aligned} Q'_{aug} &= \{Q'[-1] - 360, Q'[0], \cdots, Q'[-1]\} \\ \overline{Q'} &= Q'_{aug}[1:] - Q'_{aug}[0:-1] \end{aligned} \quad (6)$$

We then determine the indices for the start position J_{start} and the end position J_{end} respectively. Consequently, the center position q_M is:

$$\begin{aligned} J_{start} &= \arg\max(\overline{Q'}) \\ q_M &= Q'[J_{start}] + (360 - \overline{Q'}[J_{start}])/2 \end{aligned} \quad (7)$$

In our dexterous grasping system, the Z-axis $\vec{v}_z$ denotes the wrist's pointing direction, while the Y-axis $\vec{v}_y$ indicates the palm direction. The $\vec{v}_z$ axis is defined as pointing forward from the egocentric view, whereas the $\vec{v}_y$ extends from the center, influenced by the collar's hole structure. In summary, the grasping pose η is:

$$\begin{aligned} \eta &= \begin{bmatrix} \vec{v}_y \times \vec{v}_z & \vec{v}_y & \vec{v}_z & q_M \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ \vec{v}_y &= q_M - C_I,\ \vec{v}_z = [0\ 0\ 1] \end{aligned} \quad (8)$$

After each grasping attempt, we evaluate its success. Specifically, we crop the depth image D to a fixed size D', and extract the set Q with Eq. 3 and 4. We compare the depth value between each element $q_j \in Q$ and the center C of bounding box D':

$$f_E(D) = \sum_{j=1}^{M} f_J(D[q_j], D[C]) - \tau \cdot M \quad (9)$$

where $f_J(D[q_j], D[C])$ returns the comparison results between $D[q_j]$ and $D[C]$, τ represents the predefined threshold. The geometry interpretation of this evaluation is to determine whether the hole structure has been grasped appropriately.

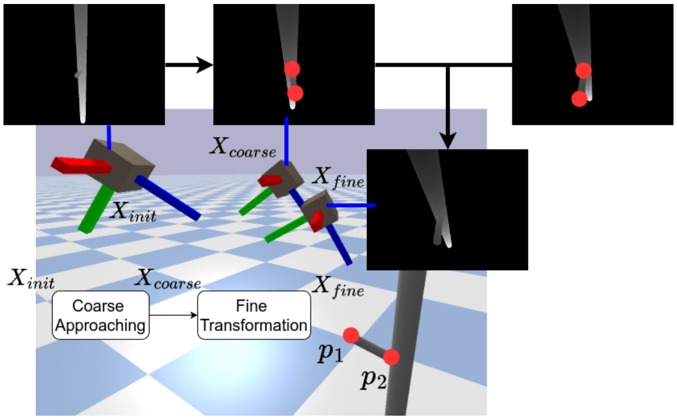

Fig. 3. The procedures of the two-layered hanging alignment. **Coarse Approaching**: Adjust the camera pose to facilitate the detection of keypoints on the rack. **Fine Transformation**: Detect the keypoints of the rack and refine the camera pose to reach the predefined pose in the demonstration.

B. Two-layered Hanging

The target supporting item is partly visible within the initial camera view $\hat{X}_{init}$ w.r.t. the rough estimated position $\hat{X}_{rack}$, which is provided to guide the arm's movement. The objective here is to determine the transformation between the current pose and the pose recorded during the demonstration, through learning from synthetic data. However, directly learning this transformation present two major challenges. Firstly, regressing the 6 degrees-of-freedom relative pose, which includes both translation and rotation, is hard to converge due to the varying modalities involved. Secondly, the depth measurement noise in real-world scenario can be significant, complicating the situation further.

To address these challenges, we propose a two-layered reaching strategy that consists of a coarse approaching f_C and a fine alignment f_A, as shown in Fig. 3. During the coarse phase, we aim to adjust the camera's viewpoint to effectively capture the key structure of the supporting item. Once the position is updated, the camera's orientation is defined as:

$$\vec{v}_z = \hat{X}_{rack} - X, \vec{v}_y = \vec{v}_z - \vec{v}_z \cdot \vec{v}_{down} \quad (10)$$

where $\vec{v}_{down}$ pointing downwards and X is the current position of the camera. Conversely, the fine phase focuses on detecting the rack's keypoints and minimize the error in relation to the observation made during the demonstration. To summarize, the coarse model f_C and the fine model f_F produce the following prediction:

$$f_C(D) \to= (\Delta x, \Delta y, \Delta z), f_F(D) \to \{p_k\}_{k=1}^{2} \quad (11)$$

To learn the models (f_C, f_A), we generate a synthetic dataset of depth images using the physics simulator Pybullet [28]. A 3D model of a rack is constructed, as shown in Fig. 3. A desired camera position X^*_{cam} relative to the rack is predefined. To gather diverse data, the camera is randomly positioned around $\hat{X}_{init}$. For the coarse model, the label corresponds to the relative displacement $X^*_{cam} - X_{init}$. For the fine model, the camera is randomly placed around X^*_{cam}. The supporting point and direction are critical in the hanging task [21]. Consequently, we designate the first keypoint as the

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.

endpoint of the supporting part and the second keypoint as the connection region between the supporting part and the standing part. The ground-truth positions of the keypoints in each observation are automatically obtained through simulation.

After training the models (f_C, f_A) with synthetic data, we implement them in real-world experiments. The coarse model f_C first take the initial observation as input and output the relative displacement $\Delta X = (\Delta x, \Delta y, \Delta z)$. The camera then moves to the new position $X_{init} + \Delta X$. Subsequently, the fine model f_A detects keypoints from the updated observation $P = \{p_k\}_{k=1}^{2}$. With the detected keypoints P, we aim to transform the camera's pose to achieve the predefined one demonstrated earlier. The pose alignment cost function is:

$$C(P, P^*) = |p_1 - p_1^*| + f_V(\overrightarrow{p_1 p_2}, \overrightarrow{p_1^* p_2^*}) \tag{12}$$

The cost function $C(P, P^*)$ in the minimization process consists of two components: (1) the distance error associated with the $1-th$ keypoint; (2) the directional error of the vector from $1-th$ keypoint to $2-th$ keypoint, computed by f_V. In addition to the alignment error, we also regularize the rotation, thus the whole optimization procedure is:

$$\min_{R,t} \; C(R * \hat{P} + t, P^*) + ||R - R_C|| \tag{13}$$

where R, t is the desired relative rotation and the translation, R_C is the orientation of the pose in the coarse phase.

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

III. RESULTS

In this section, we present a series of experiments designed to evaluate the performance of our proposed algorithm across several aspects:

- The recognition performance of our structural region detection model for various types of garments;
- the accuracy and robustness of our grasping and hanging strategies;
- the necessity of the key modules within our complete pipeline.

As illustrated in Fig. 1, the real-world experiments are conducted on two ABB robotic arms with Inspire dexterous hands. Each arm is equipped with a Realsense D435 camera. Inference is performed on a machine running Ubuntu 20.04 powered by an NVIDIA RTX A6000 GPU. The mask segmentation is performed using the method outlined in [29].

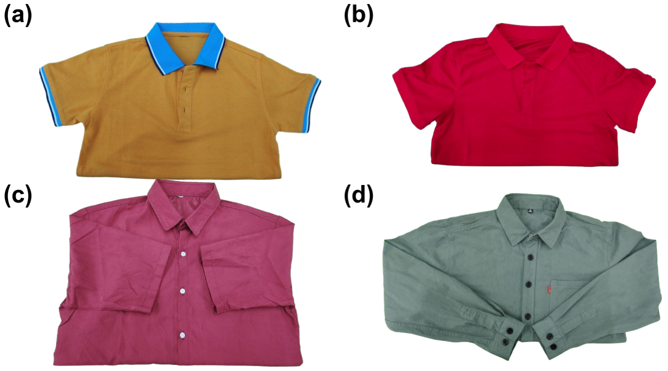

Fig. 4. Four kinds of garments are involved in the experimental study. (a) a template polo shirt (TPL). (b) a new polo shirt (NPL). (c) a short-sleeved shirt (SS). (d) a long-sleeved shirt (LS).

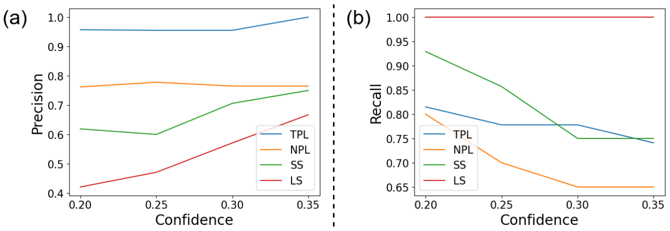

Fig. 5. The detection performance across different confidence thresholds. (a) Precision. (b) Recall.

A. Recognition

As described in Sec. II-A, we utilize a handover method between dual arms to automatically gather the dataset. The training set consists of 130 handover instances featuring a template polo shirt, as shown in Fig. 4(a). The dataset collection takes approximately 1.5 hours. For each handover step, we record 8 images from various angles at $45-$degree interval, yielding a total of 1040 images.

To assess the generalizability and the robustness of the recognition model, we include a new polo shirt (Fig. 4(b)), a short-sleeved shirt (Fig. 4(c)) and a long-sleeved shirt (in Fig. 4(d)) in the test set. Each garment type consists of a test set with 30×8 images. We evaluate detection accuracy using precision, recall and F1 score. Additionally, we report the mean and variance of the Intersection of Union (IoU) to quantify the overlap of positive detections.

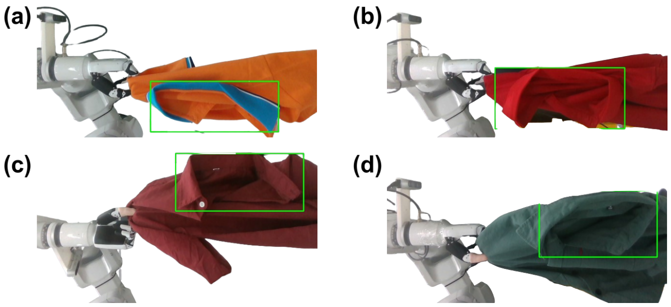

Fig. 6. The typical examples of collar detection across various kinds of garments. (a) TPL. (b) NPL. (c) SS. (d) LS.

The confidence threshold is critical for detection performance. To thoroughly assess the performance, we analyze the detection metrics across thresholds ranging from 0.2 to 0.35. The resulting precision and recall metrics are presented in Fig. 5. It is essential to highlight that performance is assessed based on each episode rather than individual step. In this context, we prioritize recall over precision, as our primary objective is to minimize the risk of missing positive grasping examples, which could lead to additional handovers. Furthermore, the grasping success evaluation (Eq. 9) can help correct the algorithm when negative examples are mistakenly grasped.

The results show that precision is directly related to confidence, whereas recall is inversely related to it. 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.

B. Grasping

All our experiments start with crumpled initial configurations [30]. In each episode, the grasping process proceeds until either the evaluation criteria are satisfied or the maximum number of handovers is reached. Two typical baselines are implemented to evaluate the effectiveness of our proposed method.

- **GCSR** [10] employs semantic segmentation for recognizing structural regions and determines the grasping pose by analyzing skeleton and surface variation.
- **GPGM** [17] identifies the grasping position using a supervised deep neural network and estimates the orientation through normal analysis.

Given that these baselines concentrate on individual images, we have also developed a score network to determine the most suitable grasping angle. In particular, we utilize images with labeled collars as positive samples to train a binary classification network. During the deployment phase, the angle with highest score is selected for grasping.

For each method, we perform 30 grasping trials for each garment shown in Fig. 4 to evaluate the grasping success rate. 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** locates the collar's contour through semantic segmentation. However, this approach operates at the pixel level, requiring a significant amount of data to effectively train a robust network. In the original implementation described in [10], it is assumed that the collar is visible and that a fixed camera is used for sensing. The performance of their algorithm is heavily dependent on this specific simplified scenario. **GPGM** identifies the grasping position through supervised learning, without considering the structural regions of the garment. The original implementation in [10] assumes that the garment is already hung on a rack, which reduces the occurrence of self-occlusion and limits the variations in state.

TABLE II
GRASPING EVALUATION RESULTS

	TPL	NPL	SS	LS
GCSR [10]	53.3	46.7	33.3	36.7
GPGN [17]	46.7	46.7	26.7	20.0
Ours	93.3	93.3	90.0	86.7

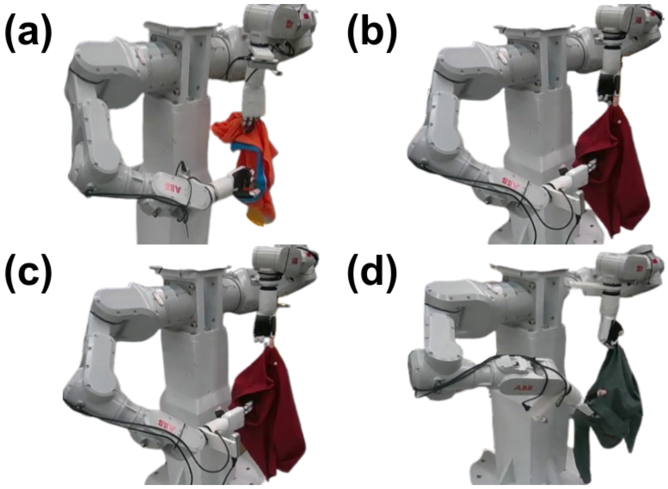

Fig. 7. The typical examples of grasping across various kinds of garments. (a) TPL. (b) NPL. (c) SS. (d) LS.

The enhance performance of our method can be primarily attributed to the active search for the collar by adjusting the configuration of crumpled garments. Additionally, the close-loop success evaluation significantly minimizes the chances of false prediction from the models. Fig. 7 illustrates several successful examples across various kinds of garments. It is crucial to emphasize that the hand pose is specifically designed to insert into the "hole" of the collar to achieve a stable grasp. There are generally two common types of grasp failures: 1) the hand may struggle to grasp the garment stably due to incorrect detection; and 2) the evaluation algorithm may occasionally yield inaccurate feedback regarding the success of the grasp.

C. Hanging

In accordance with the formulation outlined in [25], we begin by gathering individual demonstrations for each of the dual arms respectively. Specifically, we first position the arm that is pre-grasping the collar into a pose that enables clear sensing of the rack's structure, which we refer to as the "bottleneck pose". Following this, we record the subsequent trajectory needed to complete the hanging task.

All experiments in this section begin with a pre-grasping of the collar. To assess the effectiveness of our proposed method, we implement two standard baseline approaches:

- **DINO** [24] employs the large vision model to identify the corresponding keypoints with the demonstration.
- **KOVIS** [23] learns keypoint representations in a self-supervised manner for visual servoing.

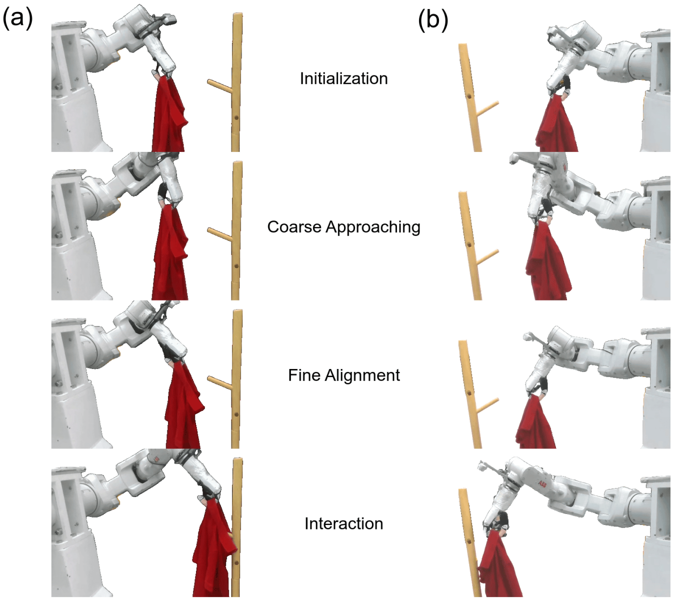

Fig. 8. The typical examples of garment hanging across dual arms. (a) Left arm. (b) Right arm.

TABLE III
HANGING EVALUATION RESULTS

	Left (%)	Right (%)	All (%)
Dinobot [24]	36.7	40.0	38.3
KOVIS [23]	33.3	30.0	31.7
Ours	90.0	93.3	91.7

Using the detected keypoints, we perform registration to align the current observation with the demonstration. For each method, we conduct 30 trials for individual arms to evaluate the hanging success rate. 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-layered 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. In their original implementations, both methods focus primarily on fine manipulation in tabletop settings. However, hanging garments in spatial contexts presents a significantly greater challenge. The position adjustments provided by the coarse approaching enable the camera to achieve a pose that is more conductive to sense the key structures of the supporting item, thus facilitating a more accurate alignment with the demonstrated pose. Fig. 8 illustrates the entire process of our two-layered hanging strategy using dual arms respectively. 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

In this experiment, we examine the contributions of the confidence-guided grasping and the two-layer hanging strategy. Specifically, we eliminate the close-loop evaluation during the grasping phase and the coarse approaching stage in the hanging phase respectively. For each method, we conduct 30 complete trials that encompass both grasping and hanging for three types of garments (NPL, SS, LS). Success rates are recorded only when the collar is successfully grasped and hung on the rack in a stable manner. 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.

TABLE IV
COMPLETE PIPELINE EVALUATION RESULTS

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

Two typical complete episodes are shown in Fig. 9. In Fig. 9(a), only one search attempt is needed to locate and successfully grasp the collar. The success is attributed to the collar being visible to the camera mounted on the slave arm, which facilitates the detection and arrangement of an appropriate grasping pose. After grasping the collar, the hanging algorithm guides the end-effector to approach and interact with the rack. Conversely, as shown in Fig. 9(b), multiple handovers are required in certain cases. This requirement arises when the collar is obscured within the garment and is not visible to the camera. As a result, the dual arms perform handovers until the collar is detected, at which point the grasp pose is established. In our experiments, we limit the maximum number of handovers per episode to 10. This example also highlights the robustness of our algorithm with respect to the crumpled configuration. 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 particularly challenging scenarios.

IV. CONCLUSION

This study presents a novel algorithm for hanging garments from crumpled configurations. By employing handovers between dual robotic arms, we are able to automatically collect a collar detection dataset without the need for human intervention. Utilizing the trained detection model, we develop a confidence-guided grasping algorithm and implement a close-loop judgment algorithm to evaluate the success of the grasping action. Furthermore, we create a two-layered hanging algorithm that aligns the garment with the rack in a coarse-to-fine manner. Through comparative real-world experiments, we demonstrate the effectiveness and superiority of our proposed method.

In future work, we aim to deploy the whole algorithm on a mobile manipulator within a real household setting.

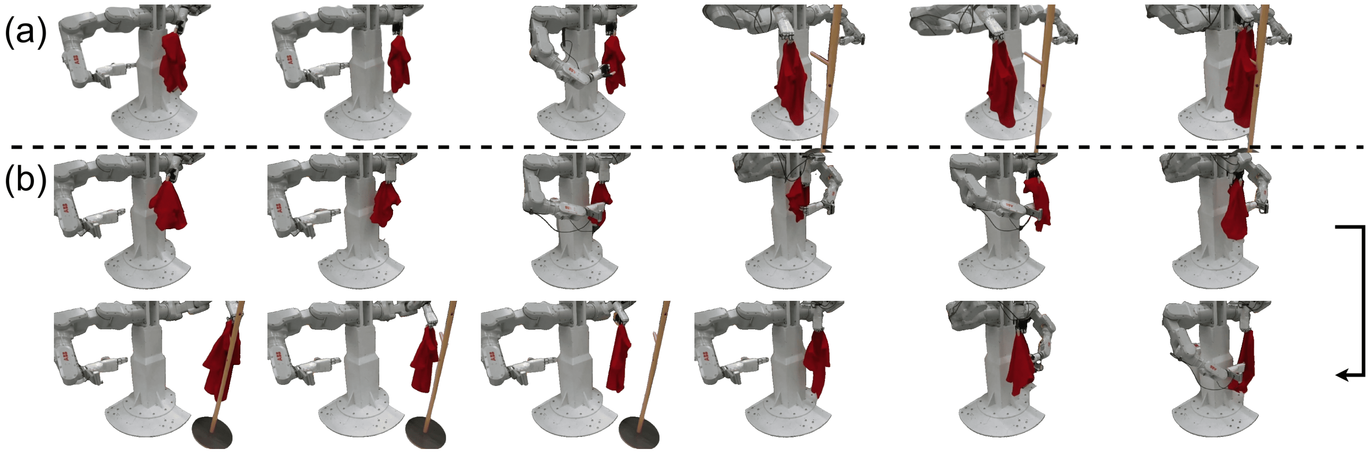

Fig. 9. Two typical examples of the complete pipeline. (a) A common scenario with only a single step of handover is required. (b) A challenging scenario that several handovers are required to adjust garment's configuration to locate the collar.

Additionally, we seek to generalize our algorithm to additional kinds of garments.

References

[1] A. Longhini, Y. Wang, I. Garcia-Camacho, D. Blanco-Mulero, M. Moletta, M. Welle, G. Alenyà, H. Yin, Z. Erickson, D. Held *et al.*, "Unfolding the literature: A review of robotic cloth manipulation," *arXiv preprint arXiv:2407.01361*, 2024.

[2] 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.

[3] H. Ha and S. Song, "Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding," in *Proc. Conf. Robot Learn.*, 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] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, 2021.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] P. Jiménez and C. Torras, "Perception of cloth in assistive robotic manipulation tasks," *Natural Computing*, vol. 19, pp. 409–431, 2020.

[12] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 1–8, 2018.

[13] D. Seita, N. Jamali, M. Laskey, A. K. Tanwani, R. Berenstein, P. Baskaran, S. Iba, J. F. Canny, and K. Goldberg, "Deep transfer learning of pick points on fabric for robot bed-making," in *International Symposium of Robotics Research*, 2018.

[14] Z. Huang, X. Lin, and D. Held, "Mesh-based dynamics with occlusion reasoning for cloth manipulation," *ArXiv*, vol. abs/2206.02881, 2022.

[15] C. Chi and S. Song, "Garmentnets: Category-level pose estimation for garments via canonical space shape completion," in *The IEEE International Conference on Computer Vision (ICCV)*, 2021.

[16] 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 Trans. Autom. Sci. Eng.*, vol. 15, pp. 1621–1638, 2016.

[17] 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.

[18] X.-H. Zhu, X. Wang, J. Freer, H. J. Chang, and Y. Gao, "Clothes grasping and unfolding based on rgb-d semantic segmentation," *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 9471–9477, 2023.

[19] J. Qian, T. Weng, B. Okorn, and L. Zhang, "Cloth region segmentation for robust grasp selection," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 9553–9560, 2020.

[20] Y. You, L. Shao, T. Migimatsu, and J. Bohg, "Omnihang: Learning to hang arbitrary objects using contact point correspondences and neural collision estimation," *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 5921–5927, 2021.

[21] 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.

[22] S. Jauhri, S. Lueth, and G. Chalvatzaki, "Active-perceptive motion generation for mobile manipulation," *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 1413–1419, 2024.

[23] 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.

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

[25] E. Valassakis, G. Papagiannis, N. D. Palo, and E. Johns, "Demonstrate once, imitate immediately (dome): Learning visual servoing for one-shot imitation learning," *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, pp. 8614–8621, 2022.

[26] 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 *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2021.

[27] 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.

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

[29] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. B. Girshick, "Segment anything," *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3992–4003, 2023.

[30] 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.