Effective Robotic Cloth Grasping Through Suppressing False Discoveries

Xingyu Zhu[1,2], Zhiwen Tu[1,2], Yan Wu[3], Shan Luo[4], Hechang Chen[1,2], Yixing Gao[1,2*]

[1]School of Artificial Intelligence, Jilin University, China
[2]Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Jilin University, China
[3]Institute for Infocomm Research, A*STAR, Singapore
[4]Department of Engineering, King's College London, United Kingdom
{xingyu24, tuzw23}@mails.jlu.edu.cn, wuy@a-star.edu.sg, shan.luo@kcl.ac.uk, {chenhc, gaoyixing}@jlu.edu.cn

Abstract

Enabling robots to grasp disorganized cloth for efficient storage is valuable in robot-assisted room organization. Diverse deformations of cloth and the stacking of multiple items limit grasping-pose estimation that relies on annotations. This necessitates segmenting each cloth item in an unsupervised manner before estimating the grasping position. However, existing segmentation methods primarily focus on improving metrics such as Intersection-over-Union and Pixel Accuracy, which cannot effectively measure the segmentation errors of the cloth area and thus lead to failure grasping position estimation. To address this challenge, we use False Discovery Rate (FDR) as a novel measure of segmentation errors and analyze its impact on grasping success. Our preliminary study reveals a negative correlation between segmentation FDR and grasping success rate, highlighting the need for more reliable segmentation in cluttered cloth scenarios. Therefore, we propose an unsupervised cloth segmentation network based on feature distance-weighted constraints, designed to reduce the false discovery rate in cloth area perception without requiring expensive pixel-level manual annotations. Additionally, to estimate the grasping position on the perceived cloth area, we introduce a strategy based on cloth surface wrinkle analysis, which operates without the need for annotations or training. By integrating the proposed segmentation network and grasping strategy, we develop a robotic system capable of sequentially grasping cluttered cloth from a table. Extensive real-world robotic experiments demonstrate the effectiveness of our approach, outperforming multiple baseline methods in segmentation FDR and grasping success rate.

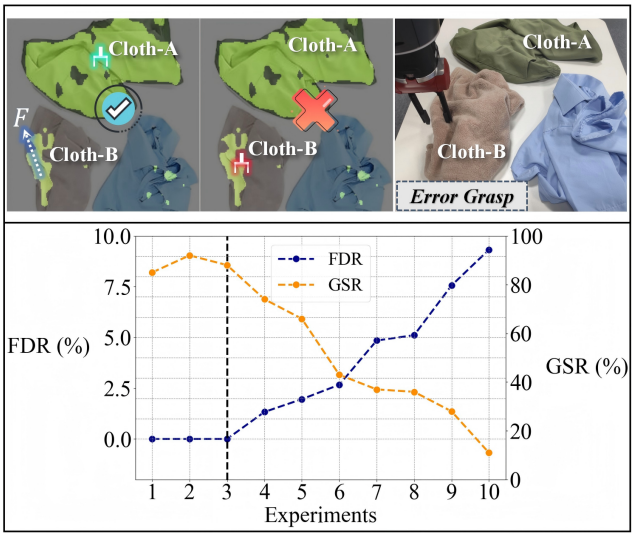

Figure 1: **Motivation for Our Approach.** Misperception of cluttered cloth by the current method increases the risk of robot grasp failures. To quantify this issue, we employ False Discovery Rate (FDR) as a measure of misperception. Our preliminary study reveals that a higher FDR correlates with a decline in Grasping Success Rate (GSR). To address this challenge, we propose a cloth area segmentation method that achieves an FDR of 0%, ensuring effective robotic cloth grasping in cluttered scenes.

1 Introduction

Automated grasping and storage of cluttered cloth is a key step to achieving robot-assisted room tidying (Shehawy, Rocco, and Zanchettin 2021; Wu et al. 2023, 2025). Unlike rigid objects with stable shapes (Rahmatizadeh et al. 2018; Li et al. 2025), the deformation characteristics of clothing lead to its morphological diversity and a wider surface (Shehawy, Rocco, and Zanchettin 2021; Zhu et al. 2022; Blanco-Mulero et al. 2024; Wu et al. 2025). This limits grasp pose estimation methods that rely on annotation and training. In addition, in real-world robotic manipulation, the quantity and category of cloth are unknown, requiring low-data-cost and unsupervised methods. Consequently, the cluttered cloth storage requires the robot to first perceive the area of all cloth items in the scene based on the unsupervised segmentation, then estimate the grasping position on the perceived cloth surface in an unsupervised manner, and finally grasp each piece of cloth one by one and place them into the basket.

In recent years, image segmentation has been widely studied in many tasks such as medical image analysis and autonomous driving, and has made significant progress (Wu et al. 2021; Ceola et al. 2022; Cao et al. 2022; Cai et al. 2024). However, current image segmentation methods still have limited performance in robotic cloth perception and grasping tasks. In our preliminary study, we analyzed the

*Corresponding author.

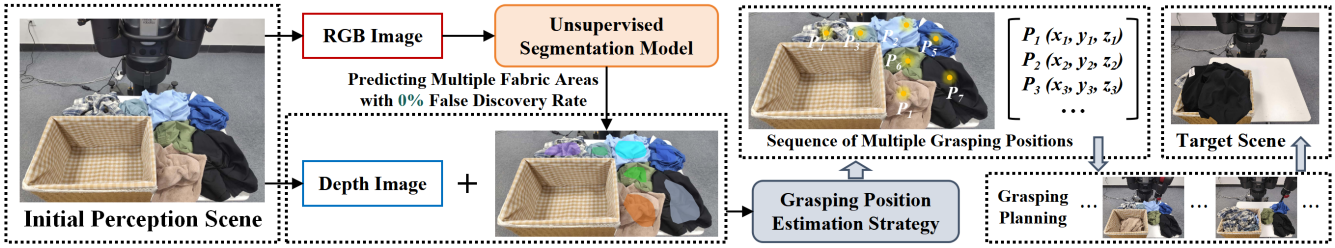

Figure 2: **Overview of the Proposed Robotic Cloth Grasping System.** Our system simultaneously acquires the RGB image and the depth image of the scene. The proposed unsupervised segmentation model predicts multiple cloth areas with a false discovery rate of 0%. Then, combined with the depth information corresponding to the perceived cloth areas, the grasping position estimation strategy generates a grasping position sequence. The robot performs action planning based on the obtained position sequence and completes the task of cluttered cloth storage.

grasping failure cases, as shown in Figure 1. Although existing methods can achieve significant segmentation accuracy, there are still areas with incorrect segmentation. For example, area F in Figure 1 should be perceived as **Cloth-B** but is perceived as **Cloth-A**. Suppose we attempt to grasp **Cloth-A**, but the grasping position estimation strategy outputs a pose in area F, which leads to grasping failure. This is because the current methods tend to improve metrics such as Intersection-over-Union (IoU) or Pixel Accuracy (PA), and cannot effectively guarantee that there are no segmentation errors, which brings the risk of grasping failure. To quantify this issue, we use False Discovery Rate (FDR) to measure the area of incorrect segmentation. The FDR is an important indicator in classification, which is used to measure the proportion of negative samples among all samples predicted to be positive samples (Krylov et al. 2016). In image segmentation, FDR measures the proportion of pixels predicted to be target classes that are actually non-target classes. The calculation of FDR is as follows:

$$FDR = \frac{FP}{FP + TP} \tag{1}$$

where FP represents the number of samples that are actually negative but predicted as positive samples, and TP represents the number of samples that are actually negative but correctly predicted as negative samples. We further conducted robotic grasping experiments on a small pile of cluttered cloth to investigate this issue. As shown in Figure 1, even a slight increase in the FDR leads to a decline in the grasping success rate, while a further rise in FDR can cause the grasping system to fail. In this paper, we focus on reducing the FDR by proposing an unsupervised cloth segmentation network based on feature distance-weighted constraints. Our method ultimately achieves cloth area segmentation with an FDR of 0%, providing a more robust foundation for robotic cloth grasping.

To estimate the grasping position on the perceived cloth area in an unsupervised manner, we introduce a strategy based on cloth surface wrinkle analysis. Unlike the grasping pose estimation of rigid objects (Rahmatizadeh et al. 2018; Li et al. 2025), grasping cloth presents additional challenges due to its larger surface and morphological inconsistencies caused by deformation (Shehawy, Rocco, and Zanchettin 2021; Zhu et al. 2022; Blanco-Mulero et al. 2024; Wu et al. 2025). We leverage the spatial information provided by depth images to search for graspable wrinkles with morphological consistency on the cloth surface. Our proposed grasping position estimation strategy consists of three steps. First, we construct a unified representation of wrinkle morphology based on a mathematical model. Second, we map the wrinkle morphology information into a feature sequence determined by hyperparameters. Finally, we develop an algorithm that traverses the depth image of the target cloth and matches the easy-to-grasp wrinkles based on feature similarity, thereby realizing the estimation of the grasping position. Our grasp position estimation method is a general strategy that does not require training and annotation, and can be more efficiently applied to real-world robotic systems.

Based on the proposed unsupervised cloth segmentation network and grasping position estimation strategy, we developed a robotic cloth grasping system. The overall framework of the system we built is shown in Figure 2. In our real-world experimental setup, a Baxter robot equipped with this system successfully grasped pieces of cloth from a table one by one and placed them into a storage basket, completing the cluttered cloth storage task without any data annotation cost. Extensive real-world experiments demonstrate the outstanding performance of our system. We built multiple baselines based on current advanced image segmentation and cloth grasping methods for comparative experiments, which further confirmed the superiority of our proposed method.

Our contributions can be summarized as follows:

- We propose an unsupervised cloth segmentation network based on feature distance-weighted constraints. This network can perceive cloth areas in the scene with a 0% false discovery rate, thus providing a robust basis for effective grasping.
- We leverage the spatial information provided by depth images and introduce a grasping position estimation strategy based on cloth surface wrinkle analysis, which requires no annotations or training.
- We developed a robotic cloth grasping system for cluttered cloth storage. Extensive real-world robot tests and comparative experiments with baseline methods demonstrate the superiority of our approach.

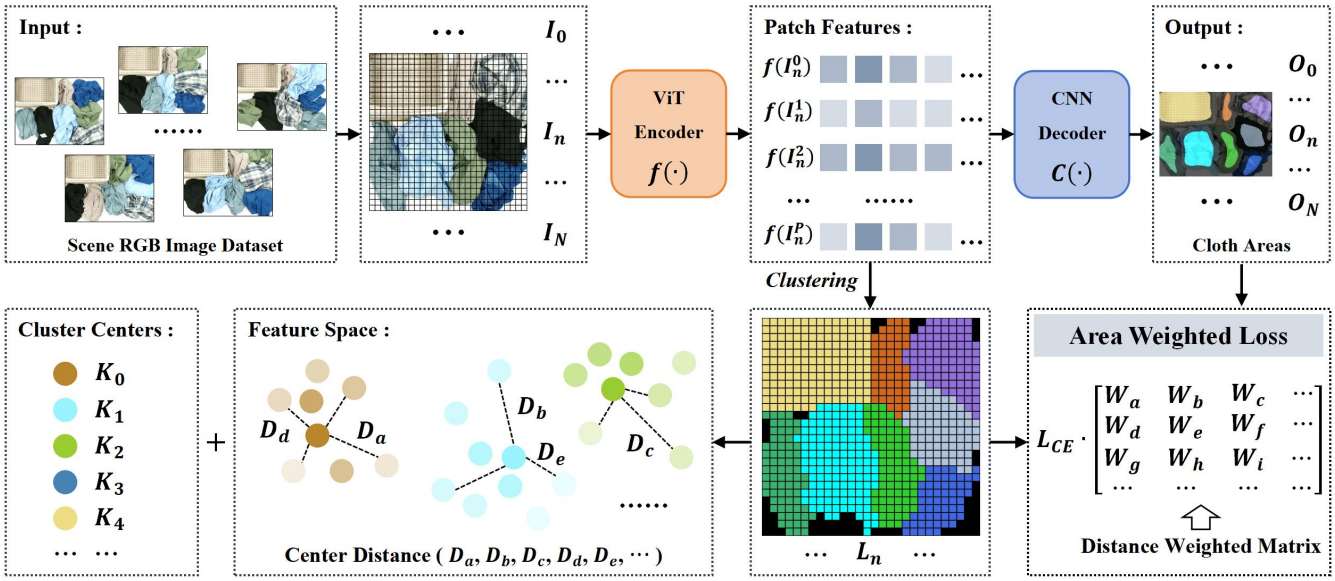

Figure 3: **Unsupervised Cloth Segmentation Network.** We adopt an encoder-decoder architecture, where the encoder is a Vision Transformer and the decoder is a Convolutional Neural Network combined with upsampling. We then perform patch-level feature extraction on the input image and obtain the corresponding initial pseudo-labels based on feature clustering. Considering the distance between the patch feature and its cluster center in the feature space, we construct a weighted constraint for different segmentation area confidences. Based on the constructed new optimization strategy, we obtain a robust segmentation model for cloth area perception with a false discovery rate of 0%.

2 Related Work

2.1 Image Segmentation

As a computer vision technology that can realize area perception, image segmentation has been widely studied and applied in multiple scenarios, such as healthcare (Cao et al. 2022), autonomous driving (Cai et al. 2024), and robot interaction (Wu et al. 2021; Ceola et al. 2022; Zhu et al. 2023). Recently, some excellent image segmentation research has achieved remarkable results on multiple benchmarks (Kirillov et al. 2023; Ni et al. 2024; Zhang et al. 2023; Kim et al. 2024). On the other hand, image segmentation that relies on costly pixel-level annotations is detrimental to real-world robotics applications. Therefore, unsupervised image segmentation has also been widely studied (Ji, Henriques, and Vedaldi 2019; Ouali, Hudelot, and Tami 2020; Cho et al. 2021; Bielski and Favaro 2022). However, the above segmentation methods focus on the segmentation indicators Intersection-over-Union (IoU) and Pixel Accuracy (PA) rather than the risk of segmentation errors. In the robotic cloth grasping scenario, the False Discovery Rate (FDR) of cloth area segmentation can lead to grasping failure cases. Therefore, in this paper, we propose an unsupervised cloth segmentation method without manual annotations that specifically suppresses the false discovery rate in cloth areas to ensure reliable grasping.

2.2 Robot Cloth Grasping

Robotic cloth grasping has always been a significant research topic because it is an essential part of robotic cloth manipulation scenarios such as cloth arrangement and assisted dressing (Wang et al. 2023; Zhang and Demiris 2022, 2020). Unlike grasping rigid targets based on pose estimation, the large surface area and deformable characteristics of cloth pose challenges to grasping (Shehawy, Rocco, and Zanchettin 2021; Zhu et al. 2022; Blanco-Mulero et al. 2024; Wu et al. 2025). Some studies relied on model-based labeled data-driven or human-involved learning of specific grasping positions to achieve automatic robotic grasping of cloth (Zhang and Demiris 2020; Fu et al. 2023; Lee et al. 2024; Tabernik et al. 2024). Other methods target specific cloth and require that the grasping positions are predefined and visible, such as the edge of a towel (Qian et al. 2020; Galassi et al. 2024; Wang et al. 2024; Longhini et al. 2024), the edge of unfolded cloth (Clark et al. 2023), and the collar of a shirt (Ramisa et al. 2016; Chen et al. 2023). Some studies also designed targeted grasping methods based on specific cloth folding or unfolding tasks (Ha and Song 2022; Wang et al. 2022; Mo et al. 2022; Canberk et al. 2023; Garcia-Camacho et al. 2024; Islam et al. 2024). (Caporali and Palli 2020) determined the grasping position of wrinkled cloth surfaces by setting hyperparameters, which reflects the possibility of designing an accurate grasping strategy based on depth information. In this paper, we propose a position estimation strategy based on cloth surface wrinkle analysis for accuracy grasping, which does not require annotation and training.

3 Unsupervised Cloth Segmentation Network

In this section, we introduce our proposed unsupervised segmentation network for cloth area perception. Our goal is to

train a vision model which can separate the scene image into different categories of areas without any manual pixel-level annotation. Then we obtain a result containing rich area information with a false discovery rate of 0% for effective robotic cloth grasping and cluttered cloth storage.

Network Structure. The overview of the network is shown in Figure 3. Following previous studies (Hamilton et al. 2022; Seong et al. 2023), our unsupervised segmentation network adopts an encoder-decoder structure. For the encoder, we use the Small-Version of the Vision Transformer (ViT) (Dosovitskiy et al. 2021). For the decoder, we design a segmenter based on Convolutional Neural Networks (CNN) and upsampling to restore the features to the same resolution as the input RGB image. We define the encoder and decoder as $f(\cdot)$ and $C(\cdot)$ respectively. At the same time, we define the input scene RGB image as $I_n, n \in [0, N]$, the corresponding features extracted by the encoder as $f(I_n^p), p \in [0, P]$, and the clothing area segmentation result output by the decoder as O_n, where N represents the scale of the dataset and P represents the number of patches in each image.

Pseudo Label Generation. As shown in Figure 3, we first consider generating a batch of initialized pseudo labels for the entire dataset, so the entire scene RGB images are used as input for training. The vision transformer inputs are in the form of patches. We take the RGB image I_n as an example and decompose it into a batch of patches.

$$I_n \Longrightarrow \{I_n^0, I_n^1, I_n^2, ..., I_n^{P-1}, I_n^P\} \tag{2}$$

We define I_n^p as the pth patch in the RGB image I_n. Then we get a set of patch features F_n, which is defined as follows.

$$F_n = \{f(I_n^0), f(I_n^1), f(I_n^2), ..., f(I_n^{P-1}), f(I_n^P)\} \tag{3}$$

Since the encoder pre-training parameters come from the result of self-supervised training on large-scale datasets, they are valid priors for extracting image features, so the extracted patch features can be used for pseudo-label initialization (Hamilton et al. 2022; Seong et al. 2023). We process the patch feature set based on the clustering strategy to obtain the initialized pseudo-label. For the clustering strategy, we adopt the K-means clustering algorithm and use cosine similarity to calculate the distance in the feature space. Specifically, we assign the clustering result of each patch feature to the patch in the form of a pixel value category label, thereby generating a pixel label with the same resolution as the patch, and then concatenate the pixel labels corresponding to all patches according to their original positions to obtain a pseudo label. We define the generated pseudo-label as L_n. The pseudo-label generation process is defined as follows.

$$L_n = Kmeans(F_n), n \in [0, N] \tag{4}$$

Feature Distance Weighting. To reduce the false discovery rate of the segmented cloth areas to 0%, we construct a new constraint that enables the model to learn to discard regions that are prone to segmentation errors and retain regions with high confidence. We represent different regions of the scene RGB image based on the granularity of the current patch division, and then measure the confidence of each patch for segmentation and downstream tasks. Since the extracted patch feature set already contains rich and valid priors, we can obtain a set of cluster centers based on clustering the feature set. As shown in Figure 3, we define the cluster center set as K_{id}, where id represents the category label. In this case, the distance from the feature corresponding to each patch to the cluster center in the feature space can be used to represent its correlation with the category represented by the cluster center. In other words, this distance reflects the confidence with which the current patch is assigned the current category label based on clustering. The smaller the feature distance of the patch, the stronger the correlation of the patch area, and the more attention should be paid to it during the model training process. Patches with large feature distance and weak correlation should be ignored and represented as background, so as to achieve the false discovery rate of 0%. Taking image I_n as an example, the calculation process of the cluster center distance can be expressed as follows:

$$D = \sqrt{(f(I_n^p) - K_{id})^2}, p \in [0, P], id = 0, 1, 2, ... \tag{5}$$

where D is defined as the cluster center distance, $f(I_n^p)$ represents the patch feature in image I_n, and K_{id} represents the cluster center corresponding to the patch feature. For a complete input scene RGB image, we first record the distance value of each patch to obtain a center distance matrix, which we define as M_D. We then take the exponential operation of the matrix M_D and normalize it to scale the values to the range of $(0, 1)$ to achieve strong correlation with short distance values. We obtain the new distance weighted matrix M_W through the following process.

$$M_W = \frac{e^{-M_D} - \left(e^{-M_D}\right)_{min}}{\left(e^{-M_D}\right)_{max}} = \begin{bmatrix} W_a & W_b & W_c & ... \\ W_d & W_e & W_f & ... \\ W_g & W_h & W_i & ... \\ ... & ... & ... & ... \end{bmatrix} \tag{6}$$

Loss Function Construction. We define the cross-entropy loss function as $L_{CE}(O, L)$, where O represents the segmentation result and L represents the pixel-level label. Therefore, the loss function we initially constructed is defined as follows:

$$L_{CE}(O_n, L_n) = -\frac{1}{V} \sum_{v=0}^{V-1} \sum_{r=0}^{R-1} y_{v,r}^L \cdot \log\left(y_{v,r}^O\right) \tag{7}$$

where V is the total number of pixels in the image, $y_{v,r}^L$ is the vth pixel in L_n, and $y_{v,r}^O$ is the probability that the vth pixel in O_n belongs to class r. The final area weighted loss function is defined as follows.

$$L_{area-weighted} = M_W \cdot L_{CE}(O_n, L_n) \tag{8}$$

Based on the above process, we finally achieve cloth area perception results with a false discovery rate of 0%, providing effective and robust results for subsequent grasping position estimation.

4 Grasping Position Estimation Strategy

After completing the cloth areas perception based on the cloth segmentation model, we need to estimate the appropriate grasping position for the area corresponding to each

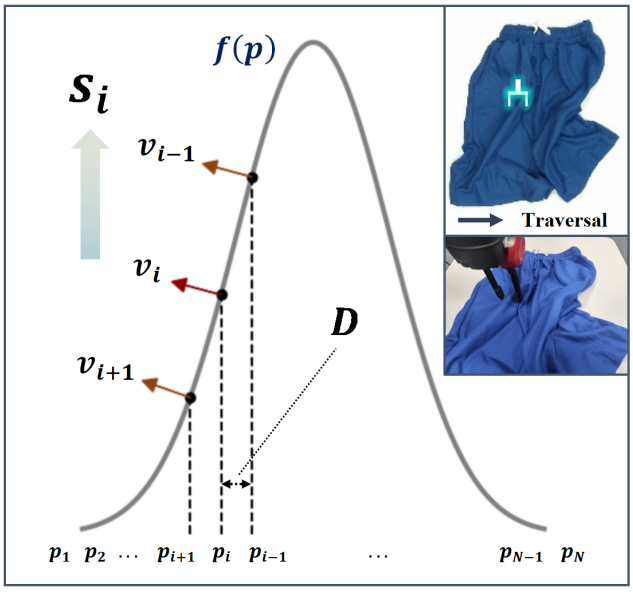

Figure 4: **Grasping Position Estimation Strategy.** We first model the cross-sectional morphology of surface wrinkles and then construct the sequence features corresponding to the morphology based on the normal vector similarity. Subsequently, we build a search algorithm to traverse the depth image to determine the optimal grasping position.

cloth item. Different from the pose estimation of rigid object grasping, cloth has a large surface and its shape is inconsistent due to deformation. In this work, we consider wrinkles with appropriate characteristics on the cloth surface as the graspable positions for cloth of different forms, and search for wrinkle positions based on the spatial information provided by depth images. The proposed position estimation strategy consists of three steps. First, we quantify wrinkles of different shapes based on mathematical models. Second, we convert the wrinkle shape information into feature vectors. Third, we construct a search algorithm to traverse the cloth surface to match appropriate position based on feature similarity.

Wrinkle Modeling. Given that the wrinkles suitable for vertical downward grasping have different protrusions, we use the normal distribution probability density curve to simulate the cross-section of the wrinkles as shown in Figure 4. The function curve is represented as follows:

$$f(p) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(p-\mu)^2}{2\sigma^2}} \tag{9}$$

where p denotes the relative pixel position of the cross-section of the wrinkle corresponding to the image. The value of μ is independent of the wrinkle morphology, and we set it to 0 for easy calculation. The value of σ can be used to approximate the degree of protrusion of the wrinkle.

Feature Mapping. We convert the current curve expression into a p-value-independent vector to represent the position-independent wrinkle morphology. As shown in Figure 4, we first define the set of positions $P = \{p_1, p_2, p_3, ..., p_{N-1}, p_N\}$, where N is the number of pixel positions. Then, we calculate the normal vectors of the points on the curve corresponding to position p to obtain the set of normal vectors $V = \{v_1, v_2, v_3, ..., v_{N-1}, v_N\}$. We calculate the sum of the cosine similarity of each normal vector to its left and right adjacent normal vectors, and the process is expressed as follows:

$$s_i = \frac{v_i \cdot v_{i-1}}{\|v_i\|\|v_{i-1}\|} + \frac{v_i \cdot v_{i+1}}{\|v_i\|\|v_{i+1}\|} \tag{10}$$

We define the result of this calculation as the neighborhood similarity of the normal vector p_i. We finally constructed the corresponding neighborhood similarity set $S = \{s_1, s_2, s_3, ..., s_{N-1}, s_N\}$. We set the vertical sampling step size of the corresponding pixel position on the curve to D. In this way, our neighborhood similarity set can be expressed as a feature sequence determined by parameters σ and D:

$$\mathbf{F}(\sigma, D) = [s_1, s_2, s_3, ..., s_{N-1}, s_N] \tag{11}$$

We use this sequence to represent the morphology of the wrinkles.

Optimal Position Searching. We search for the optimal wrinkles by traversing a continuous sequence of pixel values in the depth image. Considering the correspondence between the actual robot gripper opening width and the image size, we set the sequence length to 17. We set the traversing sequence to be horizontal and calculate the cosine similarity between each sequence searched and the modeled feature sequence $\mathbf{F}(\sigma, D)$. We traversed the entire cloth dataset and adjusted the parameter values of σ and D to determine the case with the highest average similarity. We found that when σ=0.25 and D=0.1, the average similarity of the wrinkles in the dataset was the highest. Therefore, we chose to match the optimal grasping position under this parameters setting.

Figure 5: **Overview of Our Cloth Dataset.** We collected a batch of cloth-image data and performed pixel-level area annotation only for performance evaluation and comparative experiments.

5 Experiments

5.1 Dataset, Model Training and Evaluation

We first collected a batch of data for the robotic cloth storage scene, totaling 1080 RGB images. All images were ac-

Figure 6: **Real-World Robotic Cluttered Cloth Storage.** For the cluttered cloth piled on the table, our framework can accurately perceive the area of each piece of cloth and estimate the corresponding wrinkle position that is easy to grasp. The robot places each piece of cloth on the table into the storage basket sequentially according to the obtained grasping position sequence, confirming the real-world effectiveness of our method.

Methods	Annotation-Free	mFDR ↓	Success Rate ↑
CGRSeg	×	4.24%	74%
STEGO	✓	3.23%	80%
HP	✓	5.13%	66%
Ours	✓	**0%**	**94%**

Table 1: Experimental results compared to current advanced image segmentation methods. Our proposed annotation-free method achieves a significant success rate of 94% with an mFDR of 0%.

Methods	Training-Free	Success Rate ↑
Shehawy et al. 2021	×	76%
Chen et al. 2023	×	44%
Ours	✓	**94%**

Table 2: Experimental results compared to current grasping position estimation methods for robotic cloth manipulation.

quired with a Kinect v2 camera. Throughout the experiment, we kept the camera directly above the table to capture the entire desktop scene. Then, we randomly divided the training set and validation set of the dataset into 640 and 440 respectively to further evaluate the model. Although our cloth segmentation method is annotation-free, we still annotated the RGB images at the pixel level to obtain labels for performance evaluation and comparative experiments. An overview of our scene dataset is shown in Figure 5, which includes seven items of cloth (a towel, three tops, and three shorts), a storage basket, and a table as the background.

We built our model based on the Pytorch framework, using an RTX 3090 GPU for model training and evaluation. We use bilinear interpolation to scale the three-channel RGB image and the nearest neighbor algorithm to scale the single-channel labeled image. For the training of the decoder, the initial learning rate is set to 0.01, and for the training of the vision transformer encoder, the initial learning rate is set to 0.0001. We set the batch size to 16 and use the Adam (Kingma and Ba 2015) optimizer. We use mFDR (mean False Discovery Rate) to measure the mis-segmentation in image segmentation, and mFDR is defined as follows:

$$mFDR = \frac{1}{C}\sum_{c=1}^{C} FDR \tag{12}$$

where C represents the number of segmented targets in the scene and FDR represents the false discovery rate introduced in section **Introduction**. We performed a total of 50 epochs to complete the training. We calculated the evaluation results on the collected scene dataset according to the evaluation metric. Our model has an mFDR of **0%**, thereby avoiding mis-segmentation and providing a robust basis for effective grasping. Experimental results demonstrate our robust segmentation for effective grasping and cluttered cloth storage.

5.2 Real-world Robot Test

We deployed the cloth segmentation model and the grasping position estimation strategy to a Baxter robot to construct a cloth grasping system for testing. We kept the overhead configuration of the Kinect v2 camera to ensure consistency in the test. We stipulate that the robot would be considered to have successfully stored cloth when it grasps all items of cloth on the table in sequence and puts them completely into the storage basket. The grasping order is the numerical category index order generated by clustering. In each test, we kept all the cloth in the dataset were piled on the table to ensure the integrity of the experiment. We randomly adjusted the position and shape of each cloth item and conducted 50 tests to calculate the storage success rate. Part of the real-world experimental process is shown in Figure 6. The overall storage success rate was as high as **94%**, demonstrating the effectiveness of our robotic cloth grasping system.

5.3 Comparison with Multiple Baselines

To demonstrate the superiority of both our proposed cloth segmentation method and our grasping position estimation strategy, we built multiple baselines for comparative experiments. The experimental settings are based on our collected cloth dataset and the evaluation strategy described in section **Real-world Robot Test** to ensure fair comparison. The quantitative comparison results are shown in Table 1 and Table 2, and the segmentation visualization comparison results are shown in Figure 7.

We first compared with the state-of-the-art image segmentation methods. The CGRSeg (Ni et al. 2024) is supervised, and the STEGO (Hamilton et al. 2022) and HP (Seong et al. 2023) are unsupervised. Although these methods have significant results on mIoU and PA, they ignore the risk of grasping failure caused by FDR, which limits the performance of cluttered cloth storage. In contrast, our method has an mFDR of **0%**, providing a robust ba-

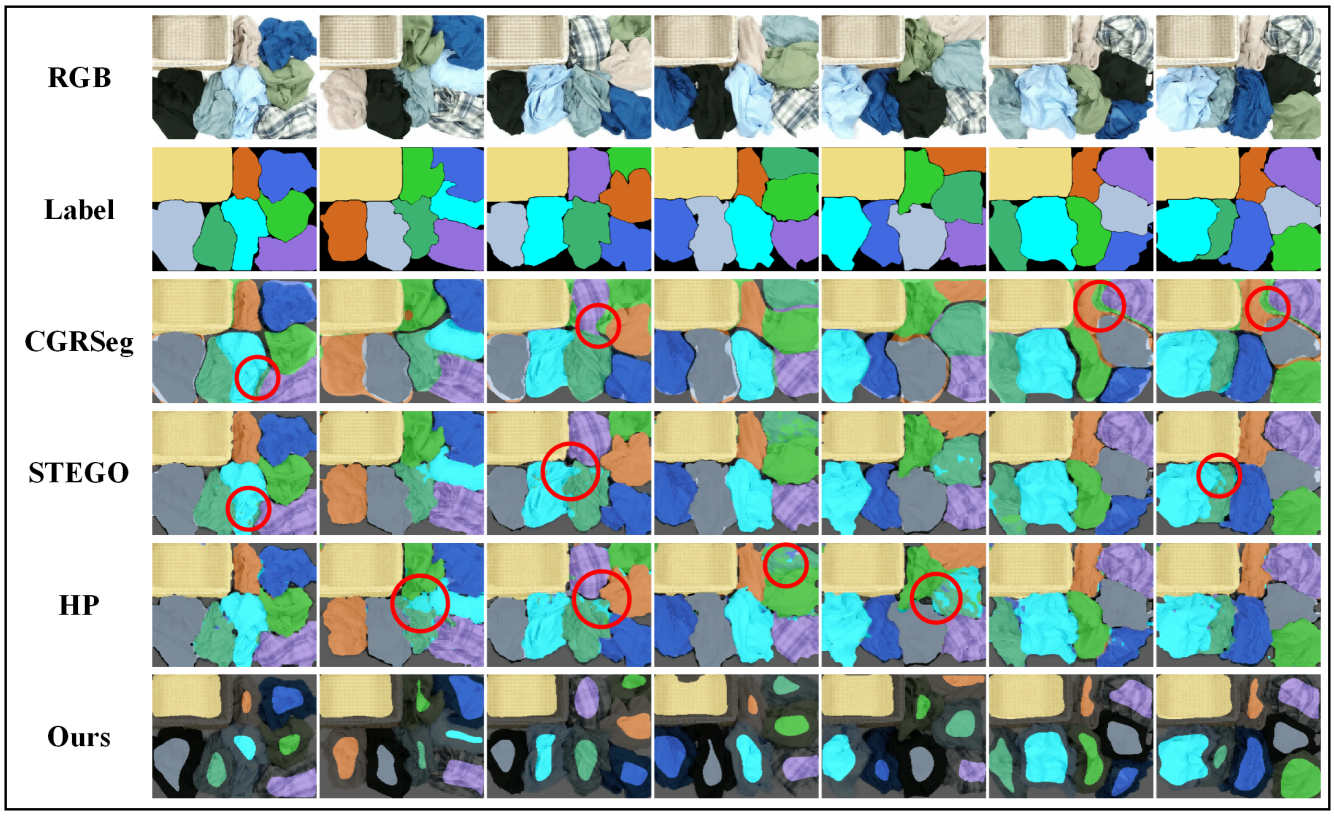

Figure 7: **Visualization Results of Cluttered Cloth Perception.** The unsupervised cloth segmentation method we proposed ensures that the perceived area is correct as long as it is perceived, which provides a robust foundation for effective grasping. In contrast, the comparative image segmentation methods exhibit segmentation errors, leading to an increased risk of grasping failure and ultimately limiting the performance of the cluttered cloth storage task.

sis for grasping, and ultimately improves the success rate of cluttered cloth storage by **+20%**, **+14%**, and **+28%** compared with CGRSeg, STEGO, and HP, respectively. Moreover, our method is annotation-free, which is conducive to further research on continuous learning of robots. Based on the cloth segmentation method we proposed, we also compared two cloth grasping position estimation methods. (Shehawy, Rocco, and Zanchettin 2021) is based on setting hyperparameters to filter out wrinkles that meet the standards to estimate the grasping position. Based on real-world experimental observations, we found that this fixed parameter setting method may lead to the inability to search for the optimal grasping position, resulting in grasping failure. (Chen et al. 2023) is based on the recognition of collars for grasping, which is only applicable to situations where the collar or collar-like structures are visible, thus limiting performance in the cluttered cloth storage task. Experimental results show that compared with (Shehawy, Rocco, and Zanchettin 2021) and (Chen et al. 2023), our grasping position estimation strategy achieves a higher success rate on the cluttered cloth storage task, demonstrating the superiority of our proposed method. In particular, our grasping position estimation method is training-free, which is more efficient for real-world robotics applications.

6 Conclusion

In this paper, we propose an unsupervised cloth segmentation network that enables the robot to perceive and grasp highly cluttered cloth for storage tasks. Considering the false discoveries of current methods, which can lead to grasping failure, we utilize the distances from scene image patch features to cluster centers to measure area criticality. Then we construct an optimization strategy based on feature-distance weighted constraint to train the model to obtain segmentation results with a false discovery rate of 0%. Additionally, we introduce a label-free and learning-free grasp position estimation strategy that quantitatively analyzes surface wrinkles to select the optimal position for precise grasping. We deployed the proposed unsupervised cloth segmentation model with the grasping position estimation strategy on a real-world Baxter robot and develop a robotic grasping system for cluttered cloth storage. Extensive real-world grasping experiments demonstrate the high-performance of our robotic cloth grasping system. Comparative experiments with both segmentation and grasping baselines demonstrate the superiority of our proposed method. In the future, we plan to explore more diverse cloth categories and scenarios, including robotic automatic clothing folding and human-robot interactive clothing manipulation.

Acknowledgments

This research was supported by the National Natural Science Foundation of China under Grant Nos. 62203184 and W2421093, and the International Cooperation Project of Jilin Province under Grant No. 20250205079GH. This research was also supported by the National Natural Science Foundation of China under Grant Nos. 62476110 and U2341229.

References

Bielski, A.; and Favaro, P. 2022. Move: Unsupervised movable object segmentation and detection. *Advances in Neural Information Processing Systems*, 35.

Blanco-Mulero, D.; Barbany, O.; Alcan, G.; Colomé, A.; Torras, C.; and Kyrki, V. 2024. Benchmarking the sim-to-real gap in cloth manipulation. *IEEE Robotics and Automation Letters*, 2981–2988.

Cai, J.; Li, Q.; Shen, Y.; Pan, J.; and Liu, W. 2024. Efficient semantic segmentation for compressed video. In *IEEE International Conference on Robotics and Automation (ICRA)*, 4266–4272.

Canberk, A.; Chi, C.; Ha, H.; Burchfiel, B.; Cousineau, E.; Feng, S.; and Song, S. 2023. Cloth funnels: Canonicalized-alignment for multi-purpose garment manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 5872–5879.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision (ECCV)*, 205–218.

Caporali, A.; and Palli, G. 2020. Pointcloud-based identification of optimal grasping poses for cloth-like deformable objects. In *IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 581–586.

Ceola, F.; Maiettini, E.; Pasquale, G.; Meanti, G.; Rosasco, L.; and Natale, L. 2022. Learn fast, segment well: fast object segmentation learning on the icub robot. *IEEE Transactions on robotics*, 38(5): 3154–3172.

Chen, W.; Lee, D.; Chappell, D.; and Rojas, N. 2023. Learning to grasp clothing structural regions for garment manipulation tasks. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 4889–4895.

Cho, J. H.; Mall, U.; Bala, K.; and Hariharan, B. 2021. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16794–16804.

Clark, A. B.; Cramphorn-Neal, L.; Rachowiecki, M.; and Gregg-Smith, A. 2023. Household clothing set and benchmarks for characterising end-effector cloth manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 9211–9217.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: transformers for image recognition at scale. *ICLR*.

Fu, T.; Bai, Y.; Li, C.; Li, F.; Wang, C.; and Song, R. 2023. Human-robot deformation manipulation skill transfer: sequential fabric unfolding method for robots. *IEEE Robotics and Automation Letters*, 8454–8461.

Galassi, K.; Wu, B.; Perez, J.; Palli, G.; and Renders, J.-M. 2024. Attention-based cloth manipulation from model-free topological representation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 18207–18213.

Garcia-Camacho, I.; Longhini, A.; Welle, M.; Alenyà, G.; Kragic, D.; and Borràs, J. 2024. Standardization of cloth objects and its relevance in robotic manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 8298–8304.

Ha, H.; and Song, S. 2022. Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding. In *Conference on Robot Learning (CoRL)*, 24–33.

Hamilton, M.; Zhang, Z.; Hariharan, B.; Snavely, N.; and Freeman, W. T. 2022. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations (ICLR)*.

Islam, S.; Owen, C.; Mukherjee, R.; and Woodring, I. 2024. Wrinkle detection and cloth flattening through deep learning and image analysis as assistive technologies for sewing. In *International Conference on PErvasive Technologies Related to Assistive Environments (PETRA)*, 233–242.

Ji, X.; Henriques, J. F.; and Vedaldi, A. 2019. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 9865–9874.

Kim, J.; Shim, K.; Lee, I.; and Shim, B. 2024. Expand-and-quantize: unsupervised semantic segmentation using high-dimensional space and product quantization. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2768–2776.

Kingma, D.; and Ba, J. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 4015–4026.

Krylov, V. A.; Moser, G.; Serpico, S. B.; and Zerubia, J. 2016. False discovery rate approach to unsupervised image change detection. *IEEE Transactions on Image Processing*, 4704–4718.

Lee, R.; Abou-Chakra, J.; Zhang, F.; and Corke, P. 2024. Learning fabric manipulation in the real world with human videos. In *IEEE International Conference on Robotics and Automation (ICRA)*, 3124–3130.

Li, S.; Liu, F.; Cui, L.; Lu, J.; Xiao, Q.; Yang, X.; Liu, P.; Sun, K.; Ma, Z.; and Wang, X. 2025. Safe planner: Empowering safety awareness in large pre-trained models for robot task planning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 14619–14627.

Longhini, A.; Welle, M. C.; Erickson, Z.; and Kragic, D. 2024. AdaFold: Adapting folding trajectories of cloths via feedback-loop manipulation. *IEEE Robotics and Automation Letters*, 9183–9190.

Mo, K.; Xia, C.; Wang, X.; Deng, Y.; Gao, X.; and Liang, B. 2022. Foldsformer: Learning sequential multi-step cloth manipulation with space-time attention. *IEEE Robotics and Automation Letters*, 760–767.

Ni, Z.; Chen, X.; Zhai, Y.; Tang, Y.; and Wang, Y. 2024. Context-guided spatial feature reconstruction for efficient semantic segmentation. In *European Conference on Computer Vision (ECCV)*, 239–255.

Ouali, Y.; Hudelot, C.; and Tami, M. 2020. Autoregressive unsupervised image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 142–158.

Qian, J.; Weng, T.; Zhang, L.; Okorn, B.; and Held, D. 2020. Cloth region segmentation for robust grasp selection. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9553–9560.

Rahmatizadeh, R.; Abolghasemi, P.; Behal, A.; and Bölöni, L. 2018. From virtual demonstration to real-world manipulation using LSTM and MDN. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Ramisa, A.; Alenya, G.; Moreno-Noguer, F.; and Torras, C. 2016. A 3D descriptor to detect task-oriented grasping points in clothing. *Pattern Recognition*, 936–948.

Seong, H. S.; Moon, W.; Lee, S.; and Heo, J.-P. 2023. Leveraging hidden positives for unsupervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19540–19549.

Shehawy, H.; Rocco, P.; and Zanchettin, A. M. 2021. Estimating a garment grasping point for robot. In *IEEE International Conference on Advanced Robotics (ICAR)*, 707–714.

Tabernik, D.; Muhovič, J.; Urbas, M.; and Skočaj, D. 2024. Center direction network for grasping point localization on cloths. *IEEE Robotics and Automation Letters*, 8913–8920.

Wang, W.; Li, G.; Zamora, M.; and Coros, S. 2024. Trtm: Template-based reconstruction and target-oriented manipulation of crumpled cloths. In *IEEE International Conference on Robotics and Automation (ICRA)*, 12522–12528.

Wang, X.; Zhao, J.; Jiang, X.; and Liu, Y.-H. 2022. Learning-based fabric folding and box wrapping. *IEEE Robotics and Automation Letters*, 5703–5710.

Wang, Y.; Sun, Z.; Erickson, Z.; and Held, D. 2023. One Policy to Dress Them All: Learning to dress people with diverse poses and garments. In *Robotics: Science and Systems (RSS)*.

Wu, J.; Antonova, R.; Kan, A.; Lepert, M.; Zeng, A.; Song, S.; Bohg, J.; Rusinkiewicz, S.; and Funkhouser, T. 2023. Tidybot: Personalized robot assistance with large language models. *Autonomous Robots*, 1087–1102.

Wu, R.; Zhu, Z.; Wang, Y.; Chen, Y.; Wang, J.; and Dong, H. 2025. GarmentPile: Point-Level Visual Affordance Guided Retrieval and Adaptation for Cluttered Garments Manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 6950–6959.

Wu, Y.; Jones, O. P.; Engelcke, M.; and Posner, I. 2021. APEX: Unsupervised, object-centric scene segmentation and tracking for robot manipulation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 3375–3382.

Zhang, D.; Li, C.; Li, H.; Huang, W.; Huang, L.; and Zhang, J. 2023. Rethinking alignment and uniformity in unsupervised image semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, 11192–11200.

Zhang, F.; and Demiris, Y. 2020. Learning grasping points for garment manipulation in robot-assisted dressing. In *IEEE International Conference on Robotics and Automation (ICRA)*, 9114–9120.

Zhang, F.; and Demiris, Y. 2022. Learning garment manipulation policies toward robot-assisted dressing. *Science robotics*, eabm6010.

Zhu, J.; Cherubini, A.; Dune, C.; Navarro-Alarcon, D.; Alambeigi, F.; Berenson, D.; Ficuciello, F.; Harada, K.; Kober, J.; Li, X.; et al. 2022. Challenges and outlook in robotic manipulation of deformable objects. *IEEE Robotics & Automation Magazine*, 67–77.

Zhu, X.; Wang, X.; Freer, J.; Chang, H. J.; and Gao, Y. 2023. Clothes grasping and unfolding based on RGB-D semantic segmentation. In *IEEE International Conference on Robotics and Automation (ICRA)*, 9471–9477.