Generalizable Category-Level Topological Structure Learning for Clothing Recognition in Robotic Grasping

Xingyu Zhu[1,2], Yan Wu[3], Zhiwen Tu[1,2], Haifeng Zhong[1,2], Yixing Gao[1,2,*]

***Abstract*— Recognizing various types of clothing is crucial for robotic clothing manipulation tasks. Existing classification models primarily focus on clothing color and texture while overlooking structural features, limiting their ability to distinguish between deformable clothing categories with similar color and texture. Moreover, these models heavily rely on manually annotated labels, making it difficult to accurately recognize unseen clothing items with new colors or textures. To address these challenges, we propose a novel topological structure representation and optimization strategy for category-level clothing structural feature learning. Then, we introduce a fabric-specific grasping position estimation method and develop a corresponding robotic grasping system capable of selecting and grasping specified clothing items based on user instructions. Extensive real-world robotic experiments demonstrate the effectiveness of our system.**

I. Introduction

Accurate recognition of various deformable clothing is crucial for embodied intelligent application scenarios such as clothing arrangement and robot-assisted dressing [1]–[4]. Some recent research has attempted to enhance recognition performance by leveraging point cloud feature or temporal information from video frames [5]–[7]. However, in real-world robotic clothing manipulation scenarios, garments are often in deformed states, which significantly limits the effectiveness of conventional classification models [5], [7].

Specifically, existing classification models often overlook structural features, making it difficult for the model to accurately distinguish between different types of deformed clothing that share similar color and texture characteristics. As a result, these models are also constrained to the specific datasets they are trained on and struggle to generalize to unseen clothing items with new colors or textures. As illustrated in Figure 1, the baseline model fails to differentiate between ***top*** and ***shorts*** with the same color and texture. Additionally, it is unable to correctly identify the category of unseen clothing items with novel colors or textures.

In this paper, we address the aforementioned challenges by compelling the model to learn category-level topological structural features of deformable clothing rather than relying on color or texture. To achieve this, we first introduce a novel clothing topological structural feature representation and optimization strategy. Based on this, we train a clothing classification model that operates independently of color and texture features. Furthermore, we design a multi-clothing classification framework that leverages the Segment Anything Model (SAM) [8]. Our framework can effectively differentiate between clothing items with identical colors and textures while also generalizing to unseen garments with novel colors or textures.

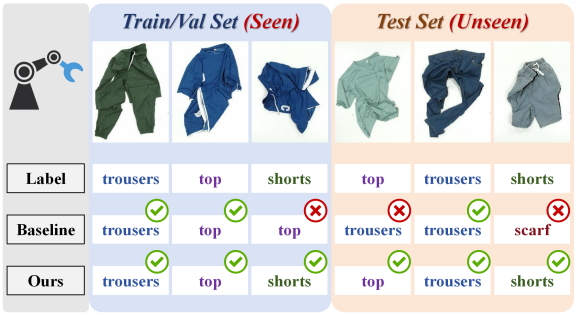

Fig. 1: Our classification method not only distinguishes different categories of clothing with the same color and texture, but also accurately generalizes to unseen targets.

Building on the multi-clothing classification framework, we developed a robotic system for verification and testing. This system enables the robot to grasp any clothing item on the table based on user instructions while demonstrating strong generalization to unseen targets. To evaluate the performance of our approach, we conducted extensive user-specified real-world clothing grasping experiments. The comprehensive experimental results demonstrate the effectiveness and superiority of our proposed approach.

II. Methods

A. Topological Structure Representation and Optimization

Figure 2 shows the overview of our strategy. The model training process of the method is divided into two stages: *Topological Feature Representation and Aggregation*; *Unified Topological Feature Learning*.

1) Topological Feature Representation and Aggregation.

Topological Feature Representation. At this stage, we first defined the topological structures of different categories of clothing according to their morphological characteristics. These topological structures include two attributes: points and edges. For example, for **Short-sleeved Top (S-top)**, we define the point set and edge set as $\{P_0^K, P_1^K, P_2^K, P_3^K, P_4^K\}$ and $\{(P_0^K, P_1^K), (P_1^K, P_2^K), (P_1^K, P_3^K), (P_3^K, P_4^K)\}$ respectively, where K represents the category label. In addition, we set constraints to avoid generating unusable topological structure images, that is, the length of the edge is a fixed range $W = [w_s, w_e]$, where w_s and w_e are pixel values, and there cannot be intersections between edges. For an image

[1] School of Artificial Intelligence, Jilin University, China.
[2] Engineering Research Center of Knowledge-Driven Human-Machine Intelligence, Ministry of Education, China.
[3] Robotics and Autonomous Systems Division, Institute for Infocomm Research (I^{2}R), A*STAR, Singapore.
* Corresponding author. Email: gaoyixing@jlu.edu.cn

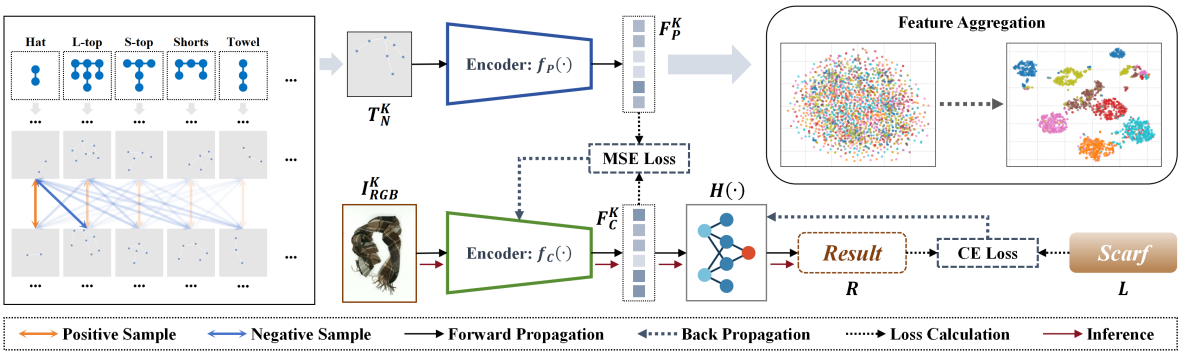

Fig. 2: Topological Structure Representation and Optimization.

of 128*128 pixels, we set w_s and w_e to 30 and 50 respectively. In our work, for each clothing category, we randomly generated 1200 images to enable the model to fully learn the representation and aggregation of topological structural features. We define the generated topological structure image set as T_N^K, where K represents the category identifier and $N \in [1, 1200]$, represents the image sequence number. We define the encoder as $f_P(\cdot)$ and the extracted features as F_P^K, then the forward inference process is expressed as follows:

$$f_P(T_N^K) \Longrightarrow F_P^K \tag{1}$$

Topological Feature Aggregation. We constructed an optimization strategy based on unsupervised contrastive learning. The input data comes from random sampling of image pairs from the image set T_N^K. We set the image pair from the same category as a positive example, defined as (T_n^K, T_m^K), and the image pair from different categories as a negative example, defined as $(T_i^K, T_j^{\overline{K}})$, where n, m, i and j represent the randomly sampled image indexes respectively. Our optimization goal is to minimize the feature distance between positive pairs and maximize the feature distance between negative pairs. We first define the feature relevance of positive and negative pairs based on cosine similarity calculation:

$$S_{positive} = \frac{F_n^K \cdot F_m^K}{\|F_n^K\|\|F_m^K\|}, S_{negative} = \frac{F_i^K \cdot F_j^{\overline{K}}}{\|F_i^K\|\|F_j^{\overline{K}}\|} \tag{2}$$

F_n^K, F_m^K, F_i^K, and $F_j^{\overline{K}}$ represent the features of topological structure images T_n^K, T_m^K, T_i^K, and $T_j^{\overline{K}}$ decoded by model $f_P(\cdot)$, respectively. The loss function and optimization process we constructed are expressed as follows:

$$\mathcal{L} = -\log(S_{positive}) + \log(1 + \exp(S_{negative})), \tag{3}$$

$$\min_\theta \mathcal{L} = \min_\theta \Big[-\frac{1}{B} \sum_{n,m} \log \left(\frac{F_n^K \cdot F_m^K}{\|F_n^K\|\|F_m^K\|} \right) + \frac{1}{B} \sum_{i,j} \log \left(1 + \exp \left(\frac{F_i^K \cdot F_j^{\overline{K}}}{\|F_i^K\|\|F_j^{\overline{K}}\|} \right) \right) \Big] \tag{4}$$

where B represents the batch size of training and θ represents the parameters of encoder $f_P(\cdot)$. Finally, the encoder we built learned the representation of the topological structural features of different categories. The visualization aggregation of the feature space is shown in Figure 2.

2) Unified Topological Feature Learning.

In the Topological Feature Representation and Aggregation stage, we finally obtain a trained encoder $f_P(\cdot)$. The encoder can already extract different topological features for different categories of clothing. At this stage, our goal is to obtain a classification model that can extract color-texture-independent topological structural features from the input clothing images and map them to corresponding categories. To achieve this goal, we first construct an encoder with the same structure as $f_P(\cdot)$, defined as $f_C(\cdot)$, and additionally build a classification head based on a fully connected network, defined as $H(\cdot)$. Then we design the following training strategy: for the same category, the topological structure image features obtained by the encoder $f_P(\cdot)$ and the clothing image features obtained by the encoder $f_C(\cdot)$ should be brought closer; on the other hand, the classification head $H(\cdot)$ should learn how to map the features to categories. Based on the above training strategy, encoder $f_C(\cdot)$ can learn the unified topological structural features of different categories of clothing. The features decoded by encoders $f_P(\cdot)$ and $f_C(\cdot)$ are defined as F_P^K and F_C^K respectively. The classification result and category label are defined as R and L respectively. For the optimization of encoder $f_C(\cdot)$, we calculate the distance between two output features based on the MSE (Mean Square Error) loss function. The loss function optimization process is expressed as follows:

$$\min_{\theta_C} \mathcal{L}_{\text{MSE}} = \sum_{K=1}^{K_{\max}} \|F_P^K - F_C^K\|_2^2 \tag{5}$$

For the optimization of the classification head $H(\cdot)$, we calculate the distance between the category output and the label based on the CE (Cross Entropy) loss function. The loss function optimization process is expressed as follows:

$$\min_{\theta_H} \mathcal{L}_{\text{CE}} = -\frac{1}{B} \sum_{i=1}^{B} \sum_{j=1}^{K_{\max}} L_{i,j} \log R_{i,j} \tag{6}$$

θ_C and θ_H are defined as the parameters of encoders $f_C(\cdot)$ and $H(\cdot)$, respectively. B is defined as the sample batch size, i and j are defined as the sample index and class index, respectively. $L_{i,j}$ is defined as the ground-truth of the i-th sample, and $R_{i,j}$ is defined as the probability that the classification head predicts the i-th sample as the j-th class. We eventually implemented the training of deformation

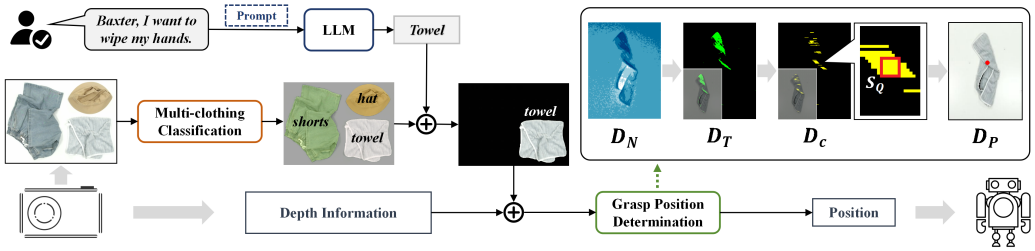

Fig. 3: Robotic Clothing Grasping System.

clothing classification model based on topological structure representation and optimization.

In the model inference and testing phase, we only use the encoder $f_C(\cdot)$ and the classification head $H(\cdot)$. The input of the model is an image I^K_{RGB} containing only a single target clothing item, and the output is the classification result R. The process is expressed as follows:

$$f_C(I^K_{RGB}) \Longrightarrow F^K_C, H(F^K_C) \Longrightarrow R \tag{7}$$

B. Robotic Clothing Grasping System

The overview of our robotic clothing grasping system is shown in Figure 3. Considering the situation where the robot grasps clothing according to user instructions, the input of our entire system consists of the following parts: text data indicated by the user; RGB images for multi-clothing classification; depth information provided by the depth camera for grasping position estimation. The multi-clothing classification framework perceives the RGB image and outputs regional category information. According to the category hints output by the LLM, we filter out the specified category region, and then further obtain the depth information within the category region to achieve a preliminary graspable range. Based on our proposed method for estimating the grasp position on the fabric surface, the grasp position information is finally sent to the real-world robot to achieve user-specified clothing grasping.

Grasping Position Estimation. The core of our method is to search for suitable wrinkles on the clothing surface. As shown in Figure 3, we first normalize the depth image for more convenient calculation. We define the normalized depth information image as D_N. For D_N, we only consider the areas in the top 10% of the pixel value distribution of the entire image to filter out relatively protruding areas. We define the protruding area information as D_T. For the filtered protrusion areas, we further search for parts with wrinkle characteristics that are easy to grasp. We first consider the mapping of gripper width to depth image area in the real world to search for candidate grasping positions in D_T that match the gripper width, and the search result is defined as D_C. For these candidate grasping positions D_C, our goal is to search for parts that are continuous and conform to the characteristics of vertical wrinkles relative to the grippers. To achieve the above effect, we adopt a maximum square search strategy. Specifically, we search for the largest square in the area D_C composed of candidate grasping positions. Such a square shape ensures the continuity and wrinkle characteristics of the area near the finally selected grasping position. We define the largest square to be searched as S_Q, and select the geometric center of S_Q as the final grasping position defined as D_P.

TABLE I: Experimental results against MobileNetV2.

Methods	Classification		Grasping	
	Seen↑	Unseen↑	Seen↑	Unseen↑
MobileNetV2 [9]	97.11%	53.48%	90.72%	53.57%
Ours	**94.84%**	**84.01%**	**89.29%**	**79.29%**

TABLE II: Experimental results against other methods.

Methods	Seen↑	Unseen↑
GarFusion [7]	76.67%	65.00%
Shehawy et al. [10]	76.67%	74.52%
Chen et al. [2]	53.81%	52.86%
Ours	**89.29%**	**79.29%**

III. Experiments

We set 7 different clothing categories: **Hat**, **Scarf**, **Top-short (Short-sleeved tops)**, **Top-long (Long-sleeved tops)**, **Shorts**, **Trousers**, **Towel**. The total size of our dataset is 1140, which means there are 38 different clothing in total, and 30 samples are collected for each piece of clothing. We divide the dataset into training and test sets, and we stipulate that clothing that appear in the training set will not appear in the test set to verify the generalization of our method. We also construct a baseline classification method that only retains the MobileNetV2 [9] and the classification head. As shown in Table I, our method not only maintained high performance on seen clothing, but also achieved 84.01% classification and 79.29% grasping performance on unseen clothing, respectively, confirming the high accuracy and generalization of our proposed method.

We performed comparative evaluations as shown in Table II. GarFusion [7] is an advanced method for classifying clothing using continuous video frames. It needs to perform grasping and then classification in the scene. Although GarFusion has achieved improved performance for unseen clothing, the overall performance is still low and requires additional long-term video inference and grasping attempts one by one. [10] was impossible to generate the optimal position when selecting the grasping position based on the specific wrinkle quantization threshold. Therefore, [10] achieved lower performance, and our method based on candidate position screening can find a relatively optimal result. For [2], it is designed for collar grasping and is suitable for tops with visible collars or similar structures such as hat edges, so its performance is limited in our scenario.

REFERENCES

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

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

[3] S. Kotsovolis and Y. Demiris, "Model predictive control with graph dynamics for garment opening insertion during robot-assisted dressing," in *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 883–890, 2024.

[4] F. Zhang and Y. Demiris, "Learning garment manipulation policies toward robot-assisted dressing," *Science robotics*, p. eabm6010, 2022.

[5] J. Stria and V. Hlavác, "Classification of hanging garments using learned features extracted from 3d point clouds," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5307–5312, 2018.

[6] L. Duan and G. Aragon-Camarasa, "A continuous robot vision approach for predicting shapes and visually perceived weights of garments," *IEEE Robotics and Automation Letters*, pp. 7950–7957, 2022.

[7] L. Huang, T. Yang, R. Jiang, X. Tian, F. Zhou, and Y. Chen, "Deforming garment classification with shallow temporal extraction and tree-based fusion," *IEEE Robotics and Automation Letters*, pp. 1114–1121, 2023.

[8] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, *et al.*, "Segment anything," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520, 2018.

[10] H. Shehawy, P. Rocco, and A. M. Zanchettin, "Estimating a garment grasping point for robot," in *IEEE International Conference on Advanced Robotics (ICAR)*, pp. 707–714, 2021.