Sparse to Dense: Robotic Perception of Deformable Objects via Foundation Models

Alessio Caporali, Kevin Galassi, Matteo Pantano, Gianluca Palli

***Abstract*—Despite their prevalence, interacting with Deformable Objects (DOs) poses challenges for robotic systems due to their complex perception. This paper introduces a method for pixel-level labeling of DOs, starting from sparse key point annotations, enabling the creation of real-world datasets with minimal human effort. The approach involves three steps: 1) Collecting images using a camera-equipped robotic arm, 2) Sparse annotation of key points by the user on one image, and 3) Converting sparse annotations into dense labels using a foundation model in zero-shot settings for segmentation tasks. Validation on cloth and rope-like objects demonstrates practicality and efficiency, laying the groundwork for seamless integration of deep learning perception into robotic agents.**

***Index Terms*—Deformable Objects, Semantic Segmentation, Dataset Generation, Deep Learning, Garment Perception, Cloth**

I. Introduction

Deformable Objects (DOs) refer to objects with the ability to change their shape when subjected to external forces. They are commonly encountered in everyday life, such as clothes and garments, which are commonly referred to as Deformable Planar Objects (DPOs), or cables, wires, and ropes, known as Deformable Linear Objects (DLOs). These objects are also prevalent in various fields, including the medical [1], agricultural [2], and industrial domains [3], [4].

The perception of DOs poses challenges due to their inherent deformability, which makes their shape unpredictable, as well as the limited (or possibly lack of) relevant features to be used in common computer vision approaches [4]–[6]. To address these challenges, new perception methodologies based on deep learning are crucial, particularly concerning the segmentation of DOs. However, these approaches require training data [7]. The size and quality of datasets greatly affect the performance of existing data-driven approaches, particularly in the DOs domain [8], making the development of efficient data collection and labeling procedures desirable.

In previous work, we introduced *DLO-WSL*, a method for labeling DLOs such as cables and wires with minimal effort using a spatial sensor and an eye-in-hand robot camera [4]. However, the approach relied on knowledge of the target DLO (e.g., diameter) and a specifically designed learned label-tuning algorithm for error correction. Therefore, *DLO-WSL* is not readily applicable to other DOs and is subject to domain shift problems when dealing with DLOs with quite different textures, such as ropes. In this paper, we focus on eliminating these constraints to convert sparse annotations into dense masks in the most general and versatile manner possible. Fig. 1 provides an overview of the approach. By harnessing the capabilities of a pre-trained foundational model [9], specifically the Segment Anything Model (SAM) [10], we can convert any sparse annotation into a dense one without requiring fine-tuning steps or domain-specific knowledge. Additionally, the utilization of an eye-in-hand camera robot allows us to expand the set of samples collected with just one annotation. This enhances the method's portability and efficiency by simplifying the approach and making it more cost-effective, thus increasing the likelihood of adoption by robotics practitioners.

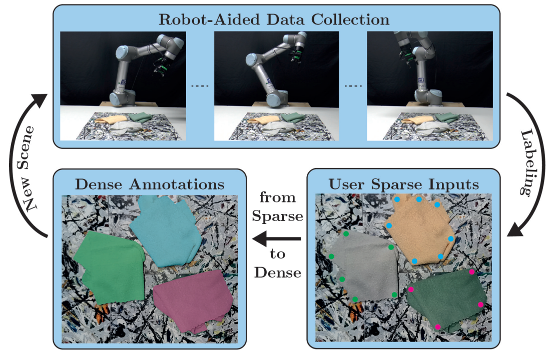


Fig. 1: Robot equipped with eye-in-hand camera collects multiple camera samples. User sparsely annotates each object in one image, and dense annotations are extracted for all images using a pre-trained foundation model.

II. Method

A. Dataset Collection and Sparse Key-points Input

1) Data Collection: To collect the set of images, knowledge of the images and the camera's position in the world coordinate system is essential [4], [11]. This is accomplished by using a calibrated 2D RGB camera mounted on the flange of a robotic arm in an eye-in-hand setup. With this information, an ellipsoidal robot trajectory is executed to collect visual samples of the DOs, ensuring that the object remains at the trajectory center while the camera is inward-facing, see [4].

An illustration of the ellipsoidal trajectory with several reference frames (e.g., $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$) is shown in Fig. 2. For clarity, only the z-axis is shown in several frames. The camera frame is denoted as $\mathcal{F}_c$.

Alessio Caporali, Kevin Galassi, and Gianluca Palli are with the Department of Electrical, Electronic and Information Engineering, University of Bologna, IT-40136 Bologna, Italy.

Matteo Pantano is with the Department of Electrical and Computer Engineering, Technical University of Munich, D-80333 Munich, Germany.

This work was partially supported by the Horizon Europe project *IntelliMan - AI-Powered Manipulation System for Advanced Robotic Service, Manufacturing and Prosthetics* [grant number 101070136].

Corresponding author: *alessio.caporali2@unibo.it*

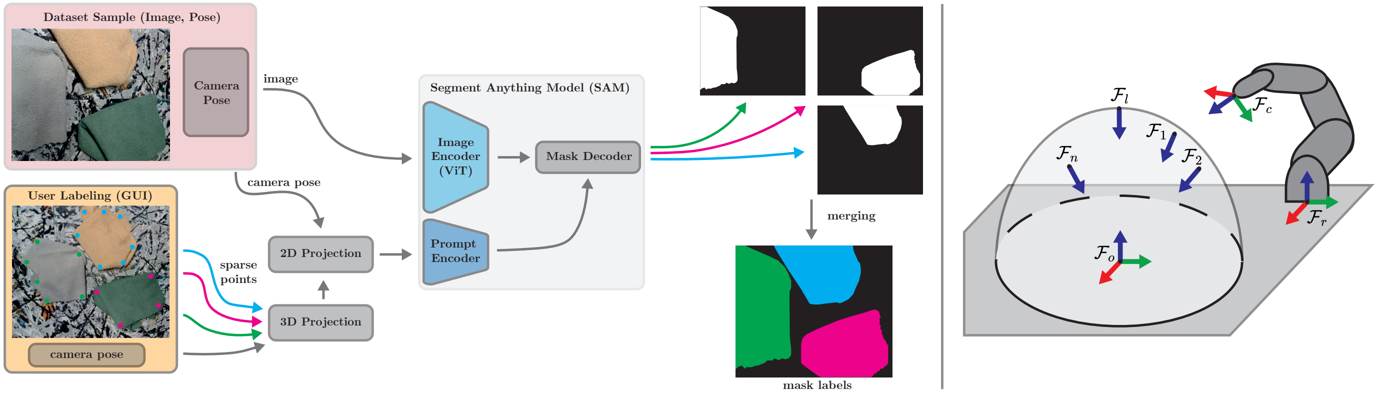


Fig. 2: On the left side, dataset labeling with SAM involves projecting sparse user labels into world coordinates, prompting SAM with these points for each dataset sample, and then merging the results to derive final labels. On the right side, a schematic view of the robotic setup, ellipsoidal trajectory, and main reference frames utilized.

2) Key Points-based User Sparse Annotation: Next, a top-view perspective image from the collected dataset is used to generate sparse labels for the entire dataset, denoted as $\mathcal{F}_l$ in Fig. 2.

Users are then instructed to trace a sequence of key points along the object's shape, aided by an intuitive visualization to facilitate the labeling process. The labeling methodology differs slightly between DLOs and DPOs. For DLOs, key points are roughly traced along the centerline, while for DPOs, they are marked along the interior border, following the object's perimeter. This approach ensures an intuitive and efficient labeling process for both DLOs and DPOs.

Subsequently, with knowledge of the camera pose and the specific camera perspective parallel to the working plane, each input key point is projected into Cartesian space as outlined by Eq. 1

$$\begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} = {}^{c}T_r \begin{bmatrix} \frac{u_i - c_x}{f_x} \\ \frac{v_i - c_y}{f_y} \\ 1 \\ 1 \end{bmatrix}, \forall i \qquad (1)$$

Where u_i and v_i are the labeled pixels, c_x, c_y, f_x, and f_y are the camera parameters obtained from the camera intrinsics matrix, ${}^{c}T_r$ is the extrinsic matrix of the camera obtained by knowing the camera position in the world coordinate frame, and x_i, y_i, and z_i are the world coordinates of the labeled point.

B. Dataset Labeling via Foundation Models

1) Sparse Inputs Projection: First, the 3D points computed in Sec. II-A2 are projected onto the specific image plane. Indeed, with the world coordinates for the labeled points established, each captured sample from Sec. II-A1 can be labeled without further user input. Specifically, by employing the inverse relation to that utilized in Eq. 1, the key points provided by the user are projected back onto the required image. In other words, the manual generation of key points for each image in the dataset is replaced by the camera-equipped robot and its associated camera-robot transformation, enabling seamless conversion between image coordinates and world coordinates.

2) Transforming Sparse Inputs to Dense Labels: Given the 2D key points, dense masks are directly generated by leveraging SAM. The SAM network consists of three primary components: 1) an RGB image encoder, which utilizes a ViT transformer; 2) a prompt encoder capable of accepting bounding boxes or key points as input; and 3) a mask decoder responsible for computing the output mask based on the embedded image and prompt.

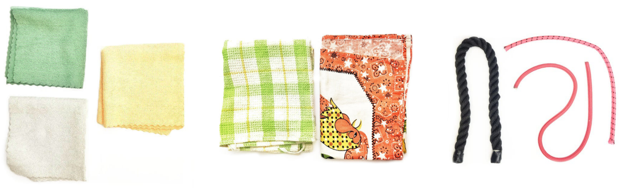
In this paper, we apply the readily available pre-trained weights of SAM, thereby avoiding costly and unnecessary fine-tuning procedures. Currently, SAM does not support multi-object prompting. Therefore, when labeling multiple objects is necessary, SAM must be prompted separately for each. However, the most computationally intensive task, i.e. the image embedding via the vision transformer, needs to be performed only once and can be saved for subsequent use with different prompts. In contrast, the prompt encoder and mask decoder are relatively small and efficient models.

In practice, the raw unthresholded prediction is obtained by SAM by setting an explicit flag. When labeling D objects, such as $D = 3$ as illustrated in Fig. 2, each object undergoes specific prompting, and the resulting masks are concatenated. Consequently, an overall mask of dimensions $H \times W \times D$ is produced, where H and W represent the height and width of the image, respectively. An additional empty mask, filled with zeros, is appended to accommodate the background "class". Subsequently, the softmax activation function is applied to the concatenated masks to derive probability values across the dimension D. Finally, the merged mask is obtained by executing the *argmax* function along the last dimension, as depicted in Fig. 2.

III. Experiments

A. Data Collection

The proposed approach is validated by exploiting various types of DOs (see Fig. 3), including cloth-like materials and rope-like ones: *Group A* consists of three soft cloths with



(a) Group A (b) Group B (c) Group C

Fig. 3: Sets of test deformable objects.

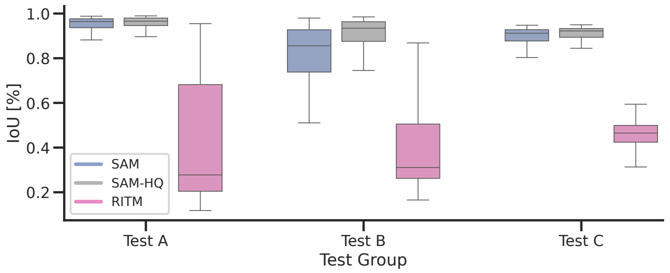


Fig. 4: Comparison between dense label generation employing the sparse or p̂oints and SAM, SAM-HQ or RITM.

uniform colors; *Group B* comprises two soft cloths with complex colors and textures; *Group C* is composed of three different ropes of varying colors and diameters.

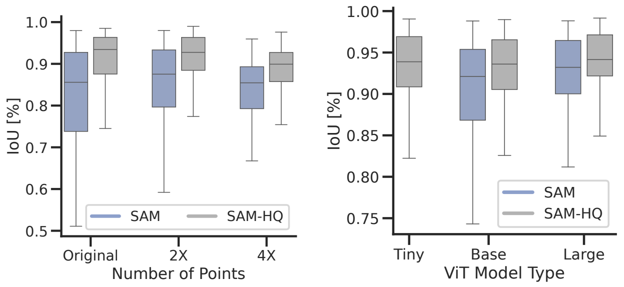
The data samples are acquired using the robotic configuration depicted in Fig. 1, comprising a UR5 robot manufactured by Universal Robots equipped with an eye-in-hand OAK-1 camera provided by Luxonis. The camera resolution is set at 1080×1920 pixels, and it has been both intrinsically and extrinsically calibrated relative to the robot flange.

The ellipsoidal trajectory detailed in Sec. II-A is executed with the following parameters selected taking into consideration the robot workspace and the camera field of view, specifically: $a = 0.35$, $b = 0.35$, $c = 0.35$, elevation angle steps 5, maximum elevation angle 50, heading angle steps 5.

To validate the quality of the labels, four backgrounds (two uniform colors, two complex shapes and clutter) are utilized to gather test data samples. Specifically, 5 samples per background are selected resulting in 20 test samples for each group, denoted as *Test A*, *Test B*, and *Test C*. Each test image is precisely annotated by a human expert to obtain accurate pixel-level ground truth data.

B. Dense Labels Quality

Starting from the key points annotated by a user (sparse points), the input is propagated to the other five images of the scenario following the proposed pipeline (Sec. II). The dense annotation starting from the sparse input points is performed using SAM [10]. As alternative approaches, the SAM-HQ model [12] and RITM [13] are also investigated. Specifically, all three methods require only the set of points as input to produce the segmentation mask. Therefore, the same set of user input points is provided to all methods, allowing a comparison of their performances.



(a) Effects number of points (b) Effects model size

Fig. 5: (a) Effects on the number of prompt points on Test B. (b) Comparison of different model sizes across the test set.

The intersection over union (IoU) score is employed as a metric for comparing the annotated masks to the ground truth data [4]. The results comparing the accuracy of SAM, SAM-HQ and RITM are shown in Fig. 4. Both SAM and SAM-HQ models consistently provide accurate results, with SAM-HQ demonstrating improved accuracy across the different test objects. RITM often fails to accurately interpret an object's entire shape and may even merge different objects together. Additionally, small errors in the projected points, for instance, due to camera calibration inaccuracies, can lead to significant deviations with RITM. In contrast, both SAM and SAM-HQ effectively address these problems.

The effect of the number of input points on accuracy is tested in Fig.5a. In the plot, *Original* refers to the set of input points provided by the users, *2X* and *4X* denote the conditions of sampling additional points in the middle of existing ones. Notably, the *2X* setting appears to improve accuracy, while no real benefits are observed with *4X*.

Ultimately, the impact of ViT model complexity on annotation accuracy is assessed in Fig. 5b. This involves evaluating both the *base* and *large* ViT models for both SAM and SAM-HQ. Additionally, SAM-HQ introduces a *tiny* variant. The figure highlights SAM-HQ's enhanced performance, even with its smallest and most resource-efficient variant.

IV. Limitations and Conclusions

In conclusion, this paper addresses the challenge of labeling Deformable Objects (DOs) to generate a real-world, task-specific dataset for use in data-driven methods for robotic perception. The proposed method offers an effective pixel-level labeling approach for DOs in images, utilizing sparse annotations of key points as a starting point. The utilization of the Segment Anything Model (SAM) enables us to obtain accurate dense masks without the need for specific fine-tuning objectives or domain-specific knowledge. Moreover, the proposed approach drastically improves the usability while reducing the labeling effort.

In future work, we will investigate implementing autonomous regeneration of scene configurations using a robotic arm to increase data variance during sample collection.

References

[1] J. Pile, G. B. Wanna, and N. Simaan, “Force-based flexible path plans for robotic electrode insertion,” in *Proc. of the ICRA*, 2014, pp. 297–303.

[2] R. J. M. Masey, J. O. Gray, T. J. Dodd, and D. G. Caldwell, “Guidelines for the design of low-cost robots for the food industry,” *Industrial Robot: An International Journal*, 2010.

[3] J. Trommnau, J. Kühnle, J. Siegert, R. Inderka, and T. Bauernhansl, “Overview of the state of the art in the production process of automotive wire harnesses, current research and future trends,” *Procedia CIRP*, 2019.

[4] A. Caporali, M. Pantano, L. Janisch, D. Regulin, G. Palli, and D. Lee, “A weakly supervised semi-automatic image labeling approach for deformable linear objects,” *IEEE Robotics and Automation Letters*, 2023.

[5] A. Caporali, K. Galassi, B. L. Žagar, R. Zanella, G. Palli, and A. C. Knoll, “RT-DLO: Real-time deformable linear objects instance segmentation,” *IEEE Transactions on Industrial Informatics*, 2023.

[6] A. Caporali, K. Galassi, and G. Palli, “Deformable linear objects 3D shape estimation and tracking from multiple 2D views,” *IEEE Robotics and Automation Letters*, 2023.

[7] R. Zanella, A. Caporali, K. Tadaka, D. De Gregorio, and G. Palli, “Auto-generated wires dataset for semantic segmentation with domain-independence,” in *2021 International Conference on Computer, Control and Robotics (ICCCR)*. IEEE, 2021, pp. 292–298.

[8] H. G. Nguyen, R. Habiboglu, and J. Franke, “Enabling deep learning using synthetic data: A case study for the automotive wiring harness manufacturing,” *Procedia CIRP*, 2022.

[9] R. Firoozi, J. Tucker, S. Tian, Majumdar *et al.*, “Foundation models in robotics: Applications, challenges, and the future,” *arXiv preprint arXiv:2312.07843*, 2023.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland *et al.*, “Segment anything,” *arXiv preprint arXiv:2304.02643*, 2023.

[11] M. Pantano, V. Klass, Q. Yang, A. Sathuluri *et al.*, “Simplifying robot grasping in manufacturing with a teaching approach based on a novel user grasp metric,” in *5th Int. Conf. on Industry 4.0 and Smart Manufacturing*, 2023.

[12] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu *et al.*, “Segment anything in high quality,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[13] K. Sofiiuk, I. A. Petrov, and A. Konushin, “Reviving iterative training with mask guidance for interactive segmentation,” in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3141–3145.