3D Shape Tracking of Deformable Linear Objects via Tactile-Visual Fusion for Collaborative Assembly

Kejia Chen[1], Celina Dettmering[2], Florian Pachler[2], Zhuo Liu[1], Yue Zhang[1], Tailai Cheng[1], Jonas Dirr[2], Zhenshan Bing[3,1], Alois Knoll[1], Rüdiger Daub[2,4]

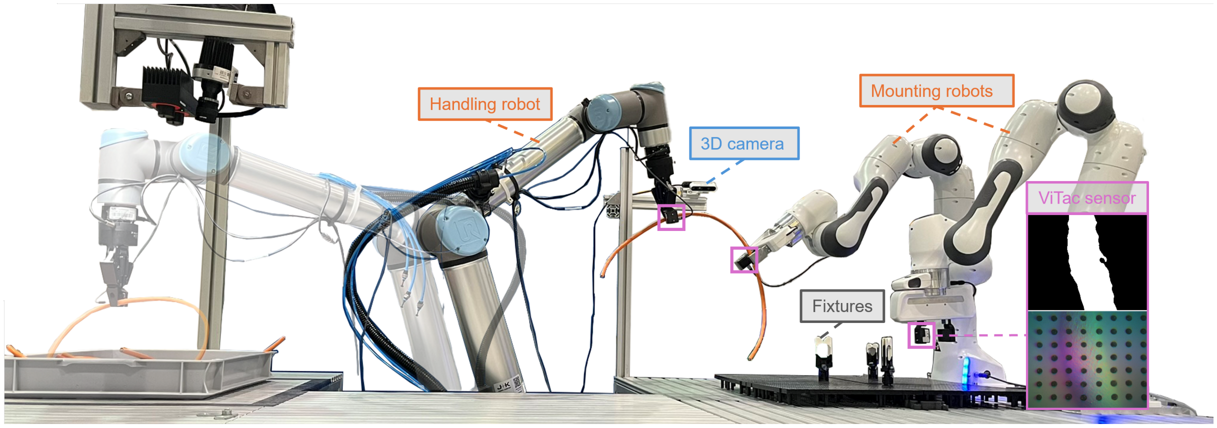

Fig. 1: Overview of the assembly process. The handling robot on the left picks one DLO from a bin full of DLO instances and hands it to one of the mounting robots on the right. The two mounting robots then collaboratively mount the DLO onto designated fixtures.

***Abstract*— Assembly of deformable linear objects (DLOs) such as cables with robots requires accurate tracking of its 3D shape. In this work, we propose a tactile-visual fusion framework for continuous 3D shape tracking during robotic manipulation. Our approach integrates global 3D visual estimation with local tactile sensing from ViTac-equipped grippers to recover occluded shape information and refine global tracking. The fused estimation supports accurate manipulation for collaborative multi-robot tasks. We validate the method in an inter-robot handover scenario, where one robot transfers a grasped DLO and hands it precisely to another. Results demonstrate improved shape tracking accuracy and robust performance under occlusions.**

I. Introduction

The assembly of deformable linear objects (DLOs), such as cables, wires, and hoses, remains a prevalent yet insufficiently automated process in industrial settings. Due to their high number of degrees of freedom and complex physical properties, DLOs exhibit time-varying configurations, making their behavior in dynamic scenarios difficult to model accurately. Consequently, tracking the 3D shape of a DLO during manipulation continues to pose significant challenges. Previous studies have proposed various shape-tracking solutions based on visual perception [1]–[3]. However, in cluttered assembly environments, DLOs often interact extensively with surrounding objects—such as fixtures or other DLO instances [4]—which greatly increases the complexity of visual tracking.

In contrast, once a DLO is grasped by a robot, proprioceptive sensing provides accurate measurements of the grasping position. Furthermore, ViTac sensors mounted on gripper fingertips can capture detailed information about the local contact area when the robot interacts with the DLO. This tactile information enables the reconstruction of the shape of the grasped portion of the DLO, which is typically occluded from visual sensors. Such information has been leveraged in previous work for reconstructing DLO shapes to support contour following [5], [6] and subsequent routing tasks [7].

In this work, we aim to integrate local tactile shape information with global visual estimation to continuously track the DLO's 3D shape during manipulation, even under occlusions. The most comparable work to ours is by Caporali et al. [8], which uses a 2D camera image for the initial grasp and employs local tactile feedback to evaluate and adjust the grasping pose. Our approach differs in that we use tactile feedback not only for local correction but also to refine global shape tracking derived from 3D visual data. This fusion enhances the accuracy of shape estimation and supports motion planning for collaborative manipulation tasks involving multiple robots [9]. We validate our approach through a DLO handling task, in which one robot transfers a grasped DLO across a large distance and accurately hands it over to another robot. This scenario demonstrates the potential of our integrated tracking method to enable robust, collaborative DLO manipulation in complex environments.

[1] Chair of Robotics, Artificial Intelligence and Real-time Systems, School of Computation, Information and Technology, Technical University of Munich, Germany.

[2] Institute for Machine Tools and Industrial Management, School of Engineering and Design, Technical University of Munich, Germany

[3] State Key Laboratory for Novel Software Technology and the School of Science and Technology, Nanjing University (Suzhou Campus), China.

[4] Fraunhofer Institute for Casting, Composite and Processing Technology IGCV, Germany.

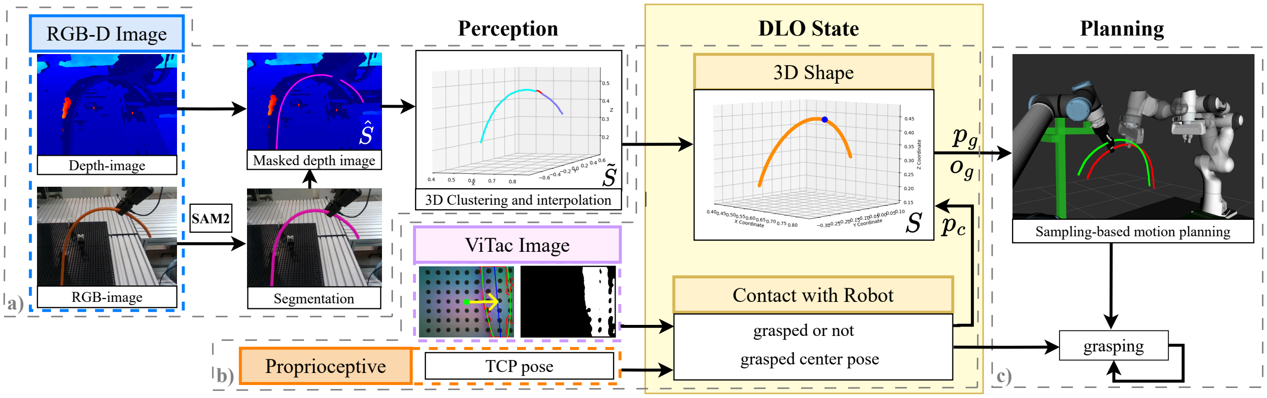

Fig. 2: Handover pipeline with 3D shape Ttacking. a) After extraction from segmentation, the DLO's raw 3D points $\hat{S}$ are clustered by DBSCAN (blue and purple), and the missing part is interpolated (red) to form an initial 3D model $\tilde{S}$. b) The grasping status as well as the grasping center are derived from the in-hand ViTac images and the robot's TCP pose. The yellow arrow on the ViTac image points from TCP to the grasping center. The final 3D shape S (orange) is corrected with grasp center point (dark blue ball). c) Based on the corrected shape (red), the grasping for the second robot is planned.

II. Methodology

We represent the DLO's shape as a sequence of 3D points $S = (\mathbf{p_1}, \mathbf{p_2}, ...\mathbf{p_N})$, where $\mathbf{p_i} = (x_i, y_i, z_i)$ stands for a 3D position. The unit tangent pointing from $\mathbf{p_i}$ to $\mathbf{p_{i+1}}$ is defined as $\mathbf{t_i} = \frac{\mathbf{p_{i+1}} - \mathbf{p_{i-1}}}{|\mathbf{p_{i+1}} - \mathbf{p_{i-1}}|}$. In this section, we align the global visual estimation to the local ground truth from proprioceptive and ViTac observations toward accurate online shape tracking

A. Global Visual Estimation

Primarily, an RGB-D camera tracks the DLO's 3D shape. As the first step, we leverage Segment-Anything Model 2 (SAM2) [10] to obtain the target DLO's mask on RGB image. Applying this mask also on the depth data yields a preliminary set of 3D points $\hat{S}$ (see Fig. 2 (a)). To address the occlusions from the environment and also the noisy depth data, we apply DBSCAN [11] to identify clusters and remove outliers in the depth data.

For each point $\hat{\mathbf{p}}_i$, DBSCAN collects its ε-neighborhood $N_\varepsilon(\hat{\mathbf{p}}_i)$. If the neighborhood contains a higher number of points than a threshold T_n, $\hat{\mathbf{p}}_i$ is classified as a cluster; otherwise, it is labeled as noise. Through this process, DBSCAN yields clusters $\{\hat{c}_i\}$, each representing a denoised, unoccluded DLO segment. The clusters are then ordered based on their projection onto the DLO's principle axis that is computed through principle component analysis (PCA). Subsequently, we apply regression methods to interpolate the missing portions and reconstruct the entire 3D shape $\tilde{S}$.

B. Local Correction

While visual perception provides an overview of the DLO's shape, the extracted world coordinates are often noisy and prone to drift. By contrast, when a robot grasps the DLO, the ground-truth pose of the grasped segment can be acquired directly through proprioceptive and tactile sensing. To enhance overall accuracy, we fuse this precise local measurement with the global visual estimation.

Once the DLO is grasped by a robot equipped with a ViTac sensor, the contact image reveals its in-hand position ${}^{tcp}\mathbf{p}_g$ in the TCP frame (see Fig. 2(b)). Combined with the robot's proprioceptive TCP pose, the DLO's center position in the base frame $\mathbf{p}_c$ is obtained. As this point is occluded in the initial observation $\hat{S}$, we associate it with the nearest cluster center in the DBSCAN-filtered set of missing segments, denoted $\tilde{\mathbf{p}}_c \in \tilde{S}$. To refine the estimated shape, $\tilde{S}$ is aligned to match the local measurement by translating it such that $\tilde{\mathbf{p}}_c$ coincides with $\mathbf{p}_c$:

Algorithm 1 Visual-tactile tracking

```
Require: Raw 3D shape $\hat{S}$ segmented by SAM2, in-hand grasping
    center ${}^{tcp}\mathbf{p}_g$, TCP pose $(\mathbf{p}_{tcp}, \mathbf{o}_{tcp})$
Ensure: Corrected 3D shape $S$
Unoccluded clusters $\{c_i\} = \text{DBSCAN}(\varepsilon, D_n, \hat{S})$
Principle axis $\mathbf{v}_{pca} = \text{PCA}(\{c_i\})$
Ordered clusters $(c_1, c_2, ..., c_N) = \text{Sort}(\{c_i\}, \mathbf{v}_{pca})$
Initialize $\tilde{S} \leftarrow \emptyset$
for $i \leftarrow 1$ to $N-1$ do
    $\tilde{S} \leftarrow (\tilde{S}, c_i)$
    $\tilde{s_i} = \text{PolynomialFitting}(c_i, c_{i+1})$
    $\tilde{S} \leftarrow (\tilde{S}, \tilde{s_i})$
Visual estimated shape $\tilde{S} \leftarrow (\tilde{S}, \tilde{s_N})$
Grasping center $\mathbf{p}_c = \mathbf{o}_{tcp} \cdot {}^{tcp}\mathbf{p}_g + \mathbf{p}_{tcp}$
Corrected shape $S \leftarrow$ Equ. 1
```

$$S = \mathbf{T}_{\text{correct}} \odot \tilde{S}, \;\; \mathbf{T}_{\text{correct}} = \begin{bmatrix} \mathbf{I}_{3\times3} & \mathbf{p}_c - \tilde{\mathbf{p}}_c \\ \mathbf{0}^T & 1 \end{bmatrix}. \tag{1}$$

The corrected shape S is then used as the final tracking result for motion planning, as summarized in Algorithm 1.

C. Handover

Based on the final tracking result (red curve in Fig. 2 (c)) corrected from original visual estimation (green curve), we plan the motion of the second robot to co-grasp the DLO. The grasping position $\mathbf{p}_g$ is set at a certain offset L_g from the first robot's grasping point. Grasping orientation $\mathbf{o}_g$ is selected to align with the DLO's local shape at the grasping point:

$$\mathbf{o}_{g_x} = \mathbf{t_g}, \;\; \mathbf{o}_{g_y} = \mathbf{n_g}, \;\; \mathbf{o}_{g_z} = \mathbf{o}_{g_x} \times \mathbf{o}_{g_y}, \tag{2}$$

where $\mathbf{t_g}$ and $\mathbf{n_g}$ are tangent and normal vectors of the DLO at $\mathbf{p}_g$. For each grasping pose, we solve inverse kinematics to obtain robot joint configurations and employ a sampling-based planner to find collision-free trajectories. The optimal path is selected among the feasible ones, by minimizing the joint path cost.

After executing the planned trajectory, the second robot's TCP may deviate slightly from $\mathbf{p}_g$ due to errors accumulated from e.g., simulation modeling or robot motion control. To compensate, a local grasping correction is applied: as shown by the yellow arrow in Fig. 2(b), the robot moves a small displacement from the current grasp point toward the TCP and retries grasping. This process repeats until sufficient alignment is achieved for the subsequent mounting motion.

III. Experiments

The experimental setup, illustrated in Figure 1, consists of a UR10e robotic manipulator to transport the DLO and hand it over to a Franka Panda robot. Each robotic system is equipped with a GelSight Mini ViTac tactile sensor, embedded in one of the gripper jaws. To oversee the handover process, an Intel RealSense D435 depth camera is strategically mounted at an inclined angle above the workspace. We conduct experiments to evaluate the corrected 3D shape tracking and its impact on the DLO handover process.

A. Tracking Correction

Due to challenges in obtaining ground-truth 3D shapes of the DLO in real-world scenarios, we firstly evaluate the tracking performance, particularly the effectiveness of local correction, by comparing the reconstructed 3D model with and without correction. Once the DLO is grasped, the robot performs random movements to induce deformation. For more complex deformation, we use a more flexible DLO with lower stiffness in this experiment.

Across 10 trials, each consisting of 40 frames, the average adjustment before and after local correction is 2.34 cm. Cases of large or small offsets are depicted in Figure 3 (a) and (b). These results indicate that the local information does refine the reconstructed 3D shape, although the degree of correction may differ from position to position. The average processing time for each frame is around 1.9 seconds, allowing the tracking pipeline to operate online during manipulation.

B. Handover results

We further evaluate the accuracy of DLO handover between robots. After grasping, the UR10e will move the DLO to a certain position, waiting for one Panda robot to take it over. The 3D shape is reconstructed using RGB-D images from the calibrated RealSense camera, the UR10e's proprioceptive information and the ViTac images. Upon initial grasping, the Panda has three chances to adjust its grasping position to align with its TCP with the assistance of ViTac sensors.

We ran handover experiments across four configurations, each involving variations in the UR10e's TCP position and the grasping point on the DLO. In each configuration, we evaluate the grasping offset L_g ranging from 8 cm to 15 cm. The resulting gaps between the grasping point and the Franka's TCP from 32 trials are recorded in Figure 3 (c). An handover with a gap below 0.93 cm is considered successful, which corresponds to the field of view of the GelSight sensor. The results show that the local correction reduces the average gap by 1.54 cm, and remarkably increase the handover success rate from 6.5% to 81.25%. Nevertheless, as the grasping offset grows, the handover accuracy may decrease, even when the correction is enabled. This degradation is likely due to the grasping point shifting closer to the edge of camera view, where depth data experience greater distortion, and the local grasp information becomes less helpful.

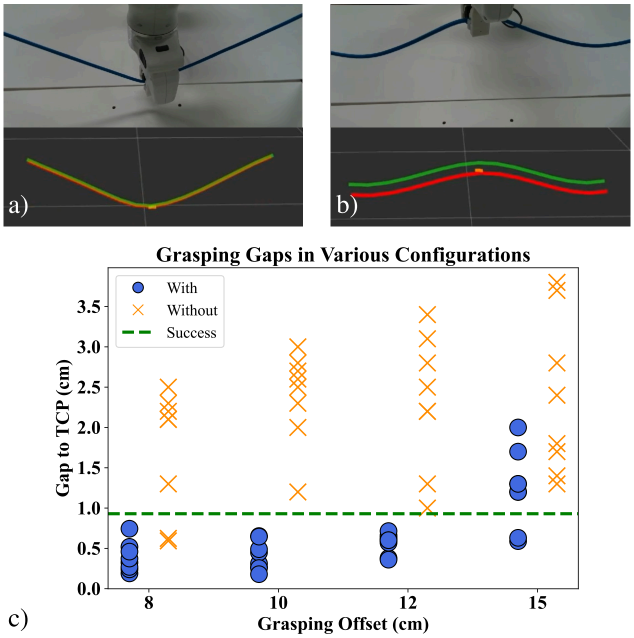

Fig. 3: Handover experiments results. a) and b) Examples of small and large corrections. The raw visual estimation is marked in green, while the corrected model is marked in red. c) Gaps between the grasping point and the robot's TCP with (circle) and without (cross) local correction. Points below the green line are success.

C. Collaborative Assembly

Finally, we validate our framework within the continuous assembly process. The framework is validated in a continuous assembly setup using power cables of length 60 cm, diameter 9.5 mm, and weight 0.1 kg. L_g is set to 1.0 to ensure robust handover while avoiding collisions. Initially, the DLO is retrieved from a bin containing 18 DLOs following the bin picking method. Once extracted successfully, the UR10e transports the DLO to the workspace of Pandas, when its 3D shape is estimated using our proposed approach. One Panda robot moves toward the DLO for grasping, prompting the UR10e to release it after the Panda co-grasped it. Following this, the Panda transports the DLO to the designated fixtures and secures it in cooperation with the second panda within the clip. For more information about the assembly process, please refer to the accompanying video and our project website: https://kejiachen.github.io/DLOAssembly/.

ACKNOWLEDGMENT

The authors acknowledge the financial support from the TUM Integrative Research Fund, provided by the seed funding initiative of the Munich Institute of Robotics and Machine Intelligence (MIRMI), and from the Bavarian State Ministry for Economic Affairs, Regional Development and Energy (StMWi) for the Lighthouse Initiative KI.FABRIK (Phase 1, grant no. DIK0249) as well as the project 'RoLei' (DIE0504/01).

REFERENCES

[1] A. Caporali, K. Galassi, R. Zanella, and G. Palli, "Fastdlo: Fast deformable linear objects instance segmentation," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9075–9082, 2022.

[2] 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*, vol. 19, no. 11, pp. 11 333–11 342, 2023.

[3] J. Xiang, H. Dinkel, H. Zhao, N. Gao, B. Coltin, T. Smith, and T. Bretl, "Trackdlo: Tracking deformable linear objects under occlusion with motion coherence," *IEEE Robotics and Automation Letters*, 2023.

[4] K. Chen, Z. Bing, Y. Wu, F. Wu, L. Zhang, S. Haddadin, and A. Knoll, "Real-time contact state estimation in shape control of deformable linear objects under small environmental constraints," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 2024, pp. 13 833–13 839.

[5] A. Monguzzi, M. Pelosi, A. M. Zanchettin, and P. Rocco, "Tactile based robotic skills for cable routing operations," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3793–3799.

[6] M. Yu, B. Liang, X. Zhang, X. Zhu, L. Sun, C. Wang, S. Song, X. Li, and M. Tomizuka, "In-hand following of deformable linear objects using dexterous fingers with tactile sensing," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 13 518–13 524.

[7] A. Monguzzi, N. Mantegna, A. M. Zanchettin, and P. Rocco, "Potential field-based online path planning for robust cable routing," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 7558–7564.

[8] A. Caporali, K. Galassi, G. Laudante, G. Palli, and S. Pirozzi, "Combining vision and tactile data for cable grasping," in *2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2021, pp. 436–441.

[9] L. Zhang, K. Cai, Z. Sun, Z. Bing, C. Wang, L. Figueredo, S. Haddadin, and A. Knoll, "Motion planning for robotics: A review for sampling-based planners," *Biomimetic Intelligence and Robotics*, vol. 5, no. 1, p. 100207, 2025.

[10] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, E. Mintun, J. Pan, K. V. Alwala, N. Carion, C.-Y. Wu, R. Girshick, P. Dollár, and C. Feichtenhofer, "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Knowledge Discovery and Data Mining*, 1996.