DLODepth: Real-time Depth Recovery for 3D Reflective Deformable Linear Object

Anonymous Authors

***Abstract*—An end-to-end monocular 3D recovery framework for Deformable Linear Object (DLO) is proposed in this paper. The fragmented and unreliable 3D point clouds caused by the thin profile and reflective surfaces of DLOs have been a critical challenge in 3D DLO perception. Conventional algorithms circumvent these issues by relying on simplified background environments or expensive multi-sensor systems, yet such constraints severely limit their practical downstream applications. This paper proposes a mixed body of multiple branches, including RGB segmentation, relative depth estimation, relative-to-metric scaling transformation, and a recovery fusion module. Our framework achieves state-of-the-art performance in recovering DLO from highly noisy inputs, recovering 93.8% (median) of the target point cloud within a 5cm error band, with a mean distance error of 4.3cm. An open-sourcing implementation of the proposed algorithm, a GUI-based data collection tool, and a ready-to-use dataset have also been provided for the benefit of the community.**

***Index Terms*—DLO Perception, Depth Recovery, Sensor Fusion**

I. Introduction

Deformable Linear Objects (DLOs), including power cables, network cables, and surgical sutures, play crucial roles in diverse applications such as industrial assembly [1] [2] and surgical operations [3]. Robotic manipulations of DLOs are typically grasping [4], pulling [5], knotting [6] and unknotting [7]. As a vital prerequisite for these tasks, reconstructing the cable's configuration has become an increasingly important research topic in recent years. In this task, the depth data is severely affected by geometrical and optical errors [8] [9] due to its thin diameter and reflective surface. Approaches, such as deploying higher-resolution sensors [10] or multi-view stereo systems [11] can sometimes reduce perception noise to a level negligible for subsequent manipulation tasks. However, they are often impractical for many DLO manipulation scenarios due to the significant resource requirements. Recently, learning models have paved a potential way to assist the DLO recovery by "predicting" missing DLO geometry. Inspired by recent success in depth completion and relative depth estimation, a new algorithm is proposed, named DLODepth, which learns from complete point cloud references and inferring reliable DLO point cloud even from highly noisy RGB-D inputs, as pictorially demonstrated in Fig. 1.

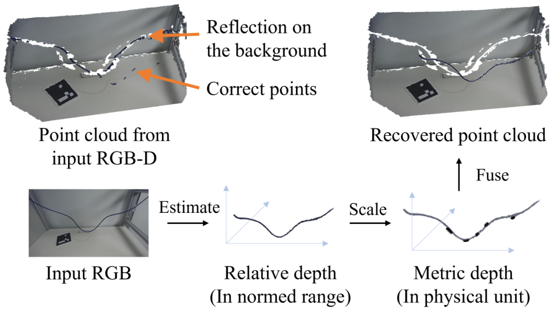
Reflection on the background
Correct points
Point cloud from input RGB-D
Recovered point cloud
Fuse
Input RGB
Estimate
Relative depth (In normed range)
Scale
Metric depth (In physical unit)

Fig. 1. Illustration of the proposed algorithm through analysis on a challenging DLO perception case. Due to the cable's highly reflective nature, the majority of DLO points in the raw point cloud data are erroneously projected onto the background, with only a sparse, noisy set of points remaining in the foreground. The proposed algorithm estimates the relative depth of the DLO from the RGB channels, learns a relative-to-metric scaling factor from the available (albeit noisy) foreground points, and fuses the recovered depth values with the raw data, replacing erroneous measurements. As such, A refined point cloud is generated that accurately reconstructs the true spatial configuration of the DLO.

The contributions of this paper can be listed as follows:

1) A novel formulation for DLO depth recovery task is introduced, designed for correcting sensing-erroneous depth data, as opposed to those with missing data.
2) The first end-to-end framework is proposed that recovers the accurate 3D shape of a DLO from monocular data. The algorithm achieves state-of-the-art results on challenging reflective DLOs, with a mean distance error of 4.3cm (1.4cm median) and a mean recovery rate of 69.4% (93.8% median).
3) A projection loss is introduced to compensate for the discrepancy between the point-to-camera distance and the depth value reported.
4) The implementation of the proposed algorithm, a data collection tool with a GUI front-end, and the ready-to-use dataset used in this paper are all open-sourced, at: https://anonymous.4open.science/r/701A.

II. Proposed Algorithm

A. Network Structure

The input depth image is denoted as $D \in \mathbb{R}^{H\times W}$, its corresponding RGB image denoted as $I \in \mathbb{R}^{H\times W\times 3}$, and the intrinsic matrix of the sensor is known. The proposed network is four-folded: RGB Segmentation, Relative Depth Estimation, Relative-to-Metric Scaling Transformation, and Recovery Fusion. The systematic flowchart is provided in Fig. 2.

RGB Segmentation. The main challenge of identifying foreground DLO points from RGB image is the extreme foreground-background imbalance and finding a solution that prioritises a low false-positive rate. PP-LiteSeg [12] is the implementation. The segmentation confidence $M_{\text{conf}} = \text{RGBSeg}(I)$ and predictive binary mask $\hat{M} = \text{argmax}(\text{softmax}(M_{\text{conf}}))$.

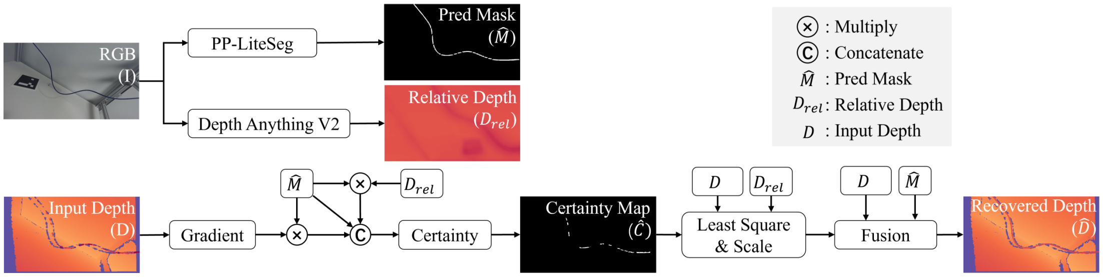

Fig. 2. Flowchart of the proposed algorithm. All visual elements used in this figure are results from an actual experiment, not conceptual illustrations.

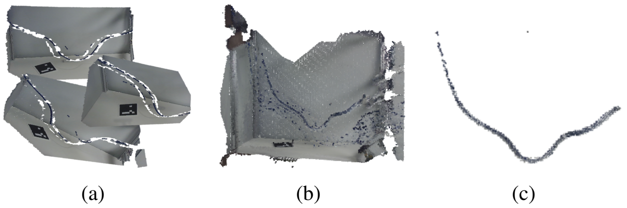

Fig. 3. Demonstration of the DLO dataset annotation. (a) Point clouds from single frames. (b) Merged point cloud. (c) The annotated DLO point cloud.

Relative Depth Estimation. Relative depth estimation is to assign a depth value to each pixel in an RGB image ordinally correct (i.e., closer points have lower values) in a normalised to a fixed range ($[0,1]$). In this paper, this module is implemented by DepthAnythingV2 [13], i.e., $D_{\text{rel}} = \text{RelativeDepth}(I)$.

Relative-to-Metric Scaling Transformation. A learning model with a ResNet-18 [14] backbone and a DPT [15] decoder is constructed to predict a *certainty map* $\hat{C} \in \mathbb{R}^{H\times W}$,

$$\hat{C} = \text{Certainty}\left(\left[(D_{\text{norm}} \odot \hat{M}, G \odot \hat{M}, \hat{M}]\right)\right) \tag{1}$$

where $D_{\text{norm}} = D/\max(D)$, G is the gradient of D_{norm} calculated using a Sobel operator $\text{Gradient}(D_{\text{norm}})$, $\odot$ denotes element-wise multiplication, and $[\cdot]$ denotes concatenation.

A set of *anchor points* $\mathscr{A}$ is then collected, containing all the points that have high certainty to be the foreground points:

$$\mathscr{A} = \{(u,v) \mid (\hat{C}(u,v) > 0.5) \wedge (\hat{M}(u,v) = 1)\} \tag{2}$$

A least-square optimisation problem is formulated to estimate the scale factor s that best aligns the relative depth with the corresponding raw depth at the anchor points,

$$s = \operatorname*{argmin}_{s} \sum_{(u,v)\in\mathscr{A}} \|s \cdot D_{\text{rel}}(u,v) - D(u,v)\|_2^2 \tag{3}$$

Recovery Fusion. The recovered metric depth image $\hat{D}$ is obtained by fusing the raw depth D and the predicted data $s \cdot D_{\text{rel}}$ based on the segmentation mask $\hat{M}$,

$$\hat{D} = D \odot (1 - \hat{M}) \oplus (s \cdot D_{\text{rel}}) \odot \hat{M} \tag{4}$$

where $\oplus$ denotes element-wise addition.

B. Loss Functions

Loss terms $\mathscr{L}_{\text{rel}}$, $\mathscr{L}_{\text{proj}}$, $\mathscr{L}_{\text{cont}}$, $\mathscr{L}_{\text{cert}}$, and $\mathscr{L}_{\text{seg}}$ are adopted in this framework, where $\mathscr{L}_{\text{rel}}$ is the scale-invariant logarithmic loss [16] to supervise D_{rel} with normalized ground truth depth, Focal loss for certainty prediction $\mathscr{L}_{\text{cert}} = \text{FocalLoss}(\hat{C}, C^*)$, and weighted CE loss to supervise segmentation branch $\mathscr{L}_{\text{seg}} = \text{CE}(M_{\text{conf}}, M^*)$.

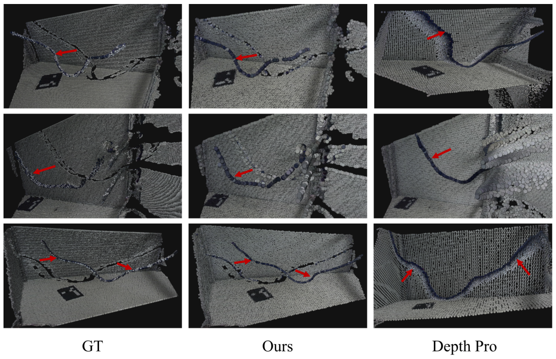

Fig. 4. Illustration of test results of the compared algorithms on the blue cable dataset. Each row shows the results from the same DLO configuration, where red arrows highlight the predicted DLO.

Projection Loss. A novel projection loss is introduced to eliminate the bias between the perpendicular distance from the point to the image plane and the geometric distance from the point to the camera centre,

$$\mathscr{L}_{\text{proj}} = \frac{1}{|\hat{\Omega} \cup \Omega^*|} \sum_{(u,v)\in\hat{\Omega}\cup\Omega^*} \left|\hat{D}(u,v) - D^*(u,v)\right| \cdot \|\mathbf{n}(u,v)\| \tag{5}$$

where $\mathbf{n}(u,v) = \left[\frac{u-c_x}{f_x},\ \frac{v-c_y}{f_y},\ 1\right]^T$, (c_x, c_y) is the coordinate of the principle point, and f_x and f_y are the focal lengths of the camera, and $\hat{\Omega}$ and Ω^* are the set of foreground pixel coordinates in the predicted mask $\hat{M}$ and GT mask M^* [1].

Continuity Loss. Huber [17] penalty is applied to encourage local smoothness of DLO depth, $\mathscr{L}_{\text{cont}} = \frac{1}{|\hat{\Omega}|}\sum_{(u,v)\in\hat{\Omega}} H_\delta\left(\sigma(u,v)\right)$, where $\sigma(u,v)$ is the local depth variance within a $k \times k$ window, and H_δ is the Huber function.

III. Reflective DLO Dataset

This dataset is collected by a handheld Orbbec Gemini 336L camera. An ArUco tag is placed at a random but fixed pose in the scene. During collection, the operator moves the camera around the cable whilst keeping the ArUco tag clearly visible. This dataset includes 20 distinct configurations for both blue and yellow Ethernet cables (diameter: 0.6cm), containing 6,840 and 4,636 RGB-D pairs, respectively.

[1] The construction of M^* will also be described in Section III.

TABLE I
COMPARISON RESULTS ON TEST DATASET

Methods	Train	Depth	Fusion	Success Num	MAD(m) ↓		CR@5cm(%) ↑		AbsRel ↓	RMSE ↓	δ_1 ↑	Time (s) ↓
					mean	median	mean	median				
Raw depth	-	Met.[1]	-	0(0.0%)	0.310	0.305	48.5	47.6	0.009	0.045	0.988	-
AdaBins [18]	Train	Met.	-	230 (17.57%)	0.081	0.075	66.7	71.9	0.107	0.104	0.904	0.027
AdaBins [18]	Train	Met.	SAM 2 [19]	230 (17.57%)	0.081	0.075	66.7	71.9	0.003	0.016	0.995	0.027+0.143[2]
Metric3Dv2 [20]	Zero[3]	Met.	-	17 (1.30%)	0.351	0.341	2.9	0.0	0.241	0.178	0.644	0.050
Depth Pro [21]	Zero	Met.	-	155 (11.84%)	0.140	0.123	25.1	18.2	0.140	0.126	0.805	0.404
UniDepthV2 [22]	Zero	Met.	-	0(0.00%)	0.542	0.497	0.3	0.0	0.498	0.364	0.270	0.042
Depth Anything V2 [13]	Zero	Rel(lstsq)	-	214 (16.35%)	0.130	0.109	26.6	17.2	0.200	0.181	0.574	0.015
[23](with depthFM [24])	Zero	Met.	completion	1 (0.08%)	0.223	0.220	30.0	23.8	0.044	0.064	0.959	14.64
Ours	Train	Met.	PP-LiteSeg [12]	**957 (73.11%)**	**0.043**	**0.014**	**69.4**	**93.8**	0.005	0.034	0.994	0.031

[1] ”Met”: metric depth. ”Rel”: relative depth. ”lstsq”: least squares fitting.
[2] The 0.143s is the average single frame inference time of Grounded SAM 2 on NVIDIA GeForce RTX 4090 GPU.
[3] ”Zero”: zero-shot. ”Train”: fine-tuned on the proposed DLO dataset.

The ground-truth of the RGB mask is semi-automatically annotated using SAM 2 [19] with manual refinement, denoted as M^*. For pseudo-GT depth annotation, the frames of each DLO configuration are first aligned with the ArUco tag pose to construct a complete point cloud (Fig. 3 (a) and (b)), and then manually annotated using MATLAB Lidar Labeller tool. The pseudo-GT depth image is generated by simulating the raytracing from the camera's optical centre. The pseudo-GT depth image D^* is

$$D^*(u,v) = \begin{cases} D_{\text{proj}}(u,v), & (u,v) \in \Omega^* \\ D(u,v), & \text{otherwise} \end{cases} \tag{6}$$

The pseudo-GT certainty map C^* is hereby defined as

$$C^*(u,v) = \begin{cases} 1, & (u,v) \in \Omega^* \text{ and } |D(u,v) - D^*(u,v)| \leq 0.05m \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

The dataset statistics show that the RGB mask of the DLO constitutes only 1.5% of the whole image, and fewer than 20% of these points can be recognised as anchor points. These underscore severe class imbalance problems in the DLO recovery task.

IV. EXPERIMENTS

Metrics. The accuracy and completeness metrics are used to evaluate the DLO recovery. The *Accuracy Distance* (AD) is defined as the distance from the point to its closest ground truth point [25], and the *Mean Accuracy Distance* (MAD) metric is then computed as the average AD value across all predicted DLO points. An RGB-D frame is recognised as "success" if the MAD is within 5cm. The *CR@5cm* metric [26] is defined as the proportion of ground truth DLO points for which the distance to the nearest predicted DLO point is less than 5cm.

Comparative Experiments. The proposed algorithm is evaluated against several categories of algorithms that researchers might consider for the DLO recovery task, including pure RGB-based DLO segmentation [27], monocular depth estimation [18], [20], [21], [22], [13], and depth completion [23] algorithms. See demonstrations in Fig. 4 and results in Table I. The proposed algorithm achieves a mean MAD of 0.043m and a median MAD of 0.014m, successfully recovering 957 out of 1309 DLO instances in the dataset. In contrast, all baseline algorithms perform significantly worse, with the best competitor reaching a mean MAD no lower than 8cm and a median MAD no lower than 7cm. Regarding the recovery completeness, the proposed algorithm achieves a mean *CR@5cm* of 69.4% and a median of 93.8%. This performance surpasses all compared algorithms, among which the best AdaBins [18] attains a mean *CR@5cm* of 66.7% and a median of 71.9%.

TABLE II
GENERALIZATION RESULTS OF TRAINING DATA NUMBER

DLO	Train/Test	Success Num	MAD (m) ↓		CR@5cm (%) ↑	
			mean	median	mean	median
Blue Cable	16 / 4	957 (73.11%)	**0.043**	0.014	69.4	**93.8**
	10 / 10	**2639 (78.36%)**	0.055	**0.012**	**72.1**	92.8
	5 / 15	3819 (72.87%)	0.065	0.016	65.1	85.9
Yellow Cable	16 / 4	**823 (86.91%)**	**0.041**	**0.015**	**76.3**	**87.9**
	10 / 10	1937 (85.52%)	0.049	0.016	73.3	84.9
	5 / 15	2498 (71.41%)	0.077	0.025	60.1	71.2

Evaluation on more objects. The proposed algorithm is trained and tested on a different yellow rope. Results reveal that the proposed algorithm maintains a mean MAD of 4.1cm and a median MAD of 1.5cm, which remains sufficient for DLO manipulation. These results are collected in Table II.

Evaluation on fewer-shot cases. Furthermore, the algorithm's performance proves robust to the size of the training set. Reducing the train-test ratio from 16:4 to 10:10, the proposed algorithm still maintains a mean MAD below 5cm and a mean *CR@5cm* greater than 69%. These verify the robustness and effectiveness of our approach across different training data volumes and DLOs.

V. CONCLUSION

The proposed algorithm in this paper integrates relative depth estimation through a delicate fusion strategy, addressing the problem of reconstructing thin and reflective DLOs under severe optical artifacts. Experimental results have validated that the proposed algorithm has been ready for plug-and-play deployment into existing DLO manipulation pipelines. A Python implementation has been provided for the benefit of the research community.

References

[1] S. Bartelt and B. Kuhlenkötter, “Evaluation of the design of a tool for the automated assembly of preconfigured wires,” in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2258–2263, IEEE, 2024.

[2] G. Laudante, M. Mirto, O. Pennacchio, K. Galassi, A. Govoni, A. Pasquali, S. Pirozzi, Ž. Gosar, and G. Palli, “Mechatronic integration of a dual-arm robotic system for wiring harness manufacturing,” *IEEE/ASME Transactions on Mechatronics*, 2025.

[3] A. Attanasio, B. Scaglioni, E. De Momi, P. Fiorini, and P. Valdastri, “Autonomy in surgical robotics,” *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 4, no. 1, pp. 651–679, 2021.

[4] J. Sanchez, J.-A. Corrales, B.-C. Bouzgarrou, and Y. Mezouar, “Robotic manipulation and sensing of deformable objects in domestic and industrial applications: a survey,” *The International Journal of Robotics Research*, vol. 37, no. 7, pp. 688–716, 2018.

[5] Z. Wang and A. H. Qureshi, “Deri-igp: Learning to manipulate rigid objects using deformable linear objects via iterative grasp-pull,” *IEEE Robotics and Automation Letters*, 2025.

[6] Y. Yamakawa, A. Namiki, and M. Ishikawa, “Dynamic high-speed knotting of a rope by a manipulator,” *International Journal of Advanced Robotic Systems*, vol. 10, no. 10, p. 361, 2013.

[7] X. Huang, D. Chen, Y. Guo, X. Jiang, and Y. Liu, “Untangling multiple deformable linear objects in unknown quantities with complex backgrounds,” *IEEE Transactions on Automation Science and Engineering*, vol. 21, no. 1, pp. 671–683, 2023.

[8] K. P. Cop, A. Peters, B. L. Žagar, D. Hettegger, and A. C. Knoll, “New metrics for industrial depth sensors evaluation for precise robotic applications,” in *IROS*, pp. 5350–5356, IEEE, 2021.

[9] A. Caporali and G. Palli, “Robotic manipulation of deformable linear objects via multiview model-based visual tracking,” *IEEE/ASME Transactions on Mechatronics*, 2025.

[10] M. Yu, K. Lv, H. Zhong, S. Song, and X. Li, “Global model learning for large deformation control of elastic deformable linear objects: An efficient and adaptive approach,” *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 417–436, 2022.

[11] X. Li, Y. Guo, Y. Tu, Y. Ji, Y. Liu, J. Ye, and C. Zheng, “Textureless deformable object tracking with invisible markers,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[12] J. Peng, Y. Liu, S. Tang, Y. Hao, L. Chu, G. Chen, Z. Wu, Z. Chen, Z. Yu, Y. Du, *et al.*, “Pp-liteseg: A superior real-time semantic segmentation model,” *arXiv preprint arXiv:2204.02681*, 2022.

[13] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, “Depth anything v2,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 21875–21911, 2024.

[14] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[15] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in *CVPR*, pp. 12179–12188, 2021.

[16] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” *NeurIPS 2014*, vol. 27, 2014.

[17] P. J. Huber, “Robust estimation of a location parameter,” in *Breakthroughs in statistics: Methodology and distribution*, pp. 492–518, Springer, 1992.

[18] S. F. Bhat, I. Alhashim, and P. Wonka, “Adabins: Depth estimation using adaptive bins,” in *CVPR*, pp. 4009–4018, 2021.

[19] N. Ravi, V. Gabeur, Y.-T. Hu, R. Hu, C. Ryali, T. Ma, H. Khedr, R. Rädle, C. Rolland, L. Gustafson, *et al.*, “Sam 2: Segment anything in images and videos,” in *The Thirteenth International Conference on Learning Representations*, 2025.

[20] M. Hu, W. Yin, C. Zhang, Z. Cai, X. Long, H. Chen, K. Wang, G. Yu, C. Shen, and S. Shen, “Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[21] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun, “Depth pro: Sharp monocular metric depth in less than a second,” in *International Conference on Learning Representations*, 2025.

[22] L. Piccinelli, C. Sakaridis, Y.-H. Yang, M. Segu, S. Li, W. Abbeloos, and L. V. Gool, “UniDepthV2: Universal monocular metric depth estimation made simpler,” *arXiv preprint arXiv:2502.20110*, 2025.

[23] L. Hyoseok, K. S. Kim, K. Byung-Ki, and T.-H. Oh, “Zero-shot depth completion via test-time alignment with affine-invariant depth prior,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 3877–3885, 2025.

[24] M. Gui, J. Schusterbauer, U. Prestel, P. Ma, D. Kotovenko, O. Grebenkova, S. A. Baumann, V. T. Hu, and B. Ommer, “Depthfm: Fast generative monocular depth estimation with flow matching,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, pp. 3203–3211, 2025.

[25] G. Yang, X. Zhou, C. Gao, X. Chen, and B. M. Chen, “Learnable cost metric-based multi-view stereo for point cloud reconstruction,” *IEEE Transactions on Industrial Electronics*, vol. 71, no. 9, pp. 11519–11528, 2023.

[26] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, “Tanks and temples: Benchmarking large-scale scene reconstruction,” *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, pp. 1–13, 2017.

[27] S. Zhaole, H. Zhou, L. Nanbo, L. Chen, J. Zhu, and R. B. Fisher, “A robust deformable linear object perception pipeline in 3d: From segmentation to reconstruction,” *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 843–850, 2023.