Uncertainty-Aware Stereo Grasp Point Selection for Deformable Linear Objects

Cristina Saccani, Alessio Caporali and Gianluca Palli

***Abstract*— Reliable grasp point selection on deformable linear objects, such as cables, requires not only accurate depth estimation but also awareness of prediction reliability. We present a five-stage stereo network for joint *disparity*, *semantic*, and *uncertainty* estimation, and use the predicted uncertainty to filter grasp candidates before geometric ranking. Disparity uncertainty is modeled via a Laplace negative log-likelihood, semantic uncertainty via the entropy of semantic predictions, with an alignment term enforcing consistency between them. Experiments on a synthetic stereo dataset show that uncertainty-aware selection reduces the mean grasp-point depth error from 4.19 mm to 1.55 mm, increases the success rate within a 3 mm tolerance from 74.2% to 88.6%, and lowers the 90th percentile of the failure exceedance above 3 mm from 29.47 mm to 6.77 mm. These results show that uncertainty is an effective cue for safer grasp selection on deformable linear objects.**

I. INTRODUCTION

Deformable Linear Objects (DLOs), such as cables and ropes, are difficult manipulation targets because their thin geometry makes grasp selection highly sensitive to perception errors [1], [2]. In particular, selecting a good grasp point requires not only detecting the cable [3], but also identifying a location supported by reliable depth [4], [5].

Stereo vision provides dense geometric information, but disparity estimation on DLOs is often unstable near thin structures, occlusions, depth discontinuities, and homogeneous background regions that offer little support for stereo matching [6]. Consequently, a grasp point chosen only from geometry and semantics may still be unreliable.

We address this problem by using uncertainty as an explicit cue for grasp selection [7]. We propose an uncertainty-aware stereo framework for DLO grasping and evaluate it on a synthetic dataset with rectified stereo pairs, dense disparity ground truth, semantic labels, and multiple camera heights.

The contributions of this work are: 1) a real-time five-stage stereo network for joint disparity, semantics, and uncertainty estimation; 2) a training objective combining Laplace supervision and uncertainty–entropy alignment; 3) an uncertainty-aware grasp-point selection policy that improves grasp reliability.

II. METHOD

From rectified stereo pairs, we estimate disparity, semantic labels, and per-pixel uncertainties, convert disparity to depth using known stereo geometry, and leverage these uncertainties to improve grasp-point selection on DLOs. A qualitative example is shown in Fig. 1.

The authors are with DEI - Department of Electrical, Electronic and Information Engineering, University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy. Corresponding author: cristina.saccani@unibo.it

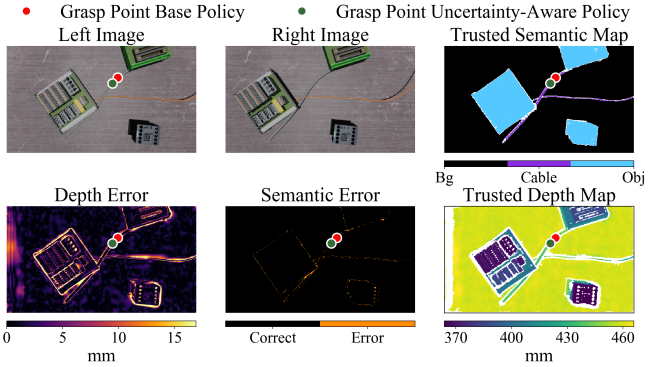

Fig. 1. Qualitative example from a scene with metal background, acquired at 460 mm. Base-policy grasp point (red): depth error 5.98 mm and predicted disparity uncertainty 0.50. Uncertainty-aware policy grasp point (green): depth error 0.48 mm and predicted disparity uncertainty 0.25.

A. Multi-Stage Stereo-Semantic Network

Our model follows a five-stage coarse-to-fine architecture, with stages operating at $1/16$, $1/8$, $1/4$, $1/2$, and full resolution.

At each stage s, disparity-oriented stereo features and semantic features are fused into a hybrid representation $\mathbf{h}_s$ from which the network predicts disparity $\hat{d}_s$, semantic logits $\hat{\mathbf{z}}_s$, and a raw disparity-uncertainty map u_s. At the coarsest stage, the network predicts an absolute disparity map. At finer stages, the previous estimate is upsampled, used to warp the right-view features, and refined through residual predictions. The semantic branch follows the same progressive refinement strategy. Crucially, the disparity uncertainty u_s is predicted from the same hybrid representation $\mathbf{h}_s$, capturing both geometric and semantic ambiguity.

B. Uncertainty Modeling and Training

The raw disparity-uncertainty output $u_s(\mathbf{x})$ is converted into a positive scale. Semantic uncertainty is instead represented by the entropy of the semantic posterior: low values correspond to confident semantic predictions, while high values indicate class ambiguity.

The base supervision exploits Smooth-L_1 for disparity and cross-entropy for semantics, i.e.:

$$\mathcal{L}_{\text{base}} = \sum_{s=1}^{5} \left(\mathcal{L}_{\text{disp}}^{(s)} + \mathcal{L}_{\text{sem}}^{(s)} \right).$$

Disparity uncertainty is learned via a Laplace negative log-likelihood and regularized by an alignment term that encourages consistency between disparity uncertainty and semantic entropy. Since the downstream task is cable grasping, disparity-related losses are treated in a class-aware

manner, assigning greater importance to cable pixels than to background or object regions. Let $\mathcal{S}_{\text{act}} \subseteq \{1, \ldots, 5\}$ denote the set of active stages at the current epoch. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\text{base}}^{\text{act}} + \lambda_{\text{unc}} \sum_{s \in \mathcal{S}_{\text{act}}} \mathcal{L}_{\text{unc}}^{(s)} + \lambda_{\text{align}} \sum_{s \in \mathcal{S}_{\text{act}}} \mathcal{L}_{\text{align}}^{(s)}.$$

During training, the active stage set is progressively expanded, enabling finer refinements only after coarser stages have stabilized.

C. Uncertainty-Aware Grasp Point Selection

Candidate grasp points are extracted from predicted cable pixels and filtered by requiring valid disparity, sufficient distance from image borders, and sufficient distance from object boundaries. For each remaining candidate $\mathbf{x} \in \mathcal{C}$, we compute

$$S_{\text{geo}}(\mathbf{x}) = w_{\text{dt}} \widetilde{\text{DT}}(\mathbf{x}) - w_{\text{grad}} \widetilde{\|\nabla \hat{d}\|}(\mathbf{x}) - w_{\text{var}} \widetilde{\text{Var}_{\mathcal{N}}(\hat{d})}(\mathbf{x}),$$

where $\text{DT}(\mathbf{x})$ is the distance from the nearest cable boundary, $\|\nabla \hat{d}\|(\mathbf{x})$ is the local disparity-gradient magnitude, and $\text{Var}_{\mathcal{N}}(\hat{d})(\mathbf{x})$ is the local disparity variance. Higher scores are assigned to points near the cable centerline and away from depth discontinuities or locally unstable regions.

The baseline policy selects

$$\mathbf{x}^{\text{base}} = \arg\max_{\mathbf{x} \in \mathcal{C}} S_{\text{geo}}(\mathbf{x}).$$

The uncertainty-aware policy first filters $\mathcal{C}$ with class-aware semantic-entropy and disparity-uncertainty trust masks, then selects

$$\mathbf{x}^{\text{unc}} = \arg\max_{\mathbf{x} \in \mathcal{C}_{\text{unc}}} S_{\text{geo}}(\mathbf{x}),$$

where $\mathcal{C}_{\text{unc}} \subseteq \mathcal{C}$ denotes the trusted candidate subset.

Thus, the selected grasp point is required to be not only geometrically plausible, but also supported by reliable depth and semantic predictions. If no trusted candidate is available, the policy abstains.

III. EXPERIMENTS

Experiments are conducted on a synthetic stereo dataset, and results are reported on its test split, which comprises 172 scenes with four background categories, one or two cables, several scene components, and 25 camera heights per scene. Both grasping policies operate on the same network predictions; thus, the observed differences arise solely from uncertainty-aware candidate filtering.

A. Grasp-Point Selection With and Without Uncertainty

Table I shows that uncertainty-aware filtering substantially improves grasp-point quality on the test split. The mean depth error decreases from 4.19 mm to 1.55 mm, while the fraction of selected points lying on the ground-truth cable increases from 96.98% to 99.30%. At the same time, abstention remains low (0.14%, i.e., 6 samples out of 4300).

The improvement is consistent across operating depths, as also reported in Table I: mean depth error decreases from 2.96 mm to 0.51 mm below 0.35 m, from 3.41 mm to 0.98 mm in the 0.35–0.50 m range, and from 5.65 mm to 2.73 mm beyond 0.50 m.

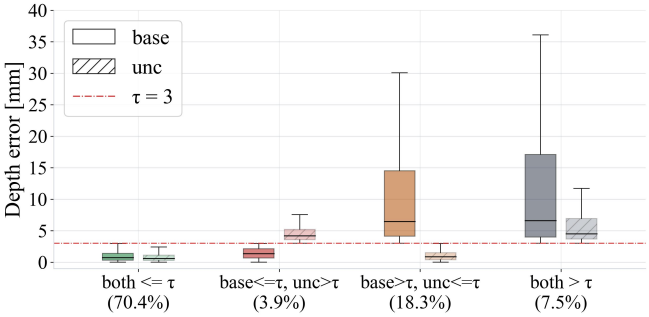

Fig. 2. Paired per-sample boxplots of grasp-point depth error for the baseline and uncertainty-aware policies, partitioned by success/failure outcome under a tolerance of $\tau = 3$ mm.

TABLE I

GRASP-POINT SELECTION RESULTS ON THE TEST SPLIT.

Category	Metric	Interval	Baseline	Uncertainty-aware
Overall	EPE [mm]	–	4.19	**1.55**
	Cable rate [%]	–	96.98	**99.30**
	No-grasp [%]	–	**0.00**	0.14
Depth range	EPE [mm]	0.00–0.35 m	2.96	**0.51**
		0.35–0.50 m	3.41	**0.98**
		>0.50 m	5.65	**2.73**

Using a task-relevant success threshold of $\tau = 3$ mm, uncertainty-aware selection increases the success rate from 74.2% to 88.6%. Among the remaining failures, it also reduces their severity, lowering the 90th percentile of the exceedance above 3 mm from 29.47 mm to 6.77 mm. Overall, the main contribution of uncertainty is not a marginal refinement of easy cases, but the prevention of a subset of high-error grasp selections.

B. Per-Sample Grasp Reliability

Fig. 2 reports boxplots of grasp-point depth error, grouping samples by success or failure of the baseline and uncertainty-aware policies under a tolerance of $\tau = 3$ mm. We adopt $\tau = 3$ mm as a practically reasonable threshold for successful grasp execution. The dominant asymmetric region is the one where the baseline fails but the uncertainty-aware policy succeeds (18.3%), whereas the opposite case is much rarer (3.9%). Moreover, even in joint-failure cases, the uncertainty-aware policy tends to produce smaller errors.

C. Runtime

On an NVIDIA RTX A6000 GPU and at 360×640 resolution, the full inference pipeline runs at 32.4 FPS, including disparity prediction, semantic prediction, and uncertainty-aware grasp-point selection. This supports real-time deployment.

IV. CONCLUSIONS

We presented an uncertainty-aware stereo framework for real-time grasp point selection on DLOs. Uncertainty-aware filtering reduces the mean grasp-point depth error from 4.19 mm to 1.55 mm, increases the success rate at 3 mm tolerance from 74.2% to 88.6%, and reduces failure severity, while the full pipeline runs at 32.4 FPS. These results show that uncertainty improves grasp reliability without compromising real-time performance.

REFERENCES

[1] 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 *IEEE Int. Conf. IROS*, 2021.

[2] A. Caporali, K. Galassi, and G. Palli, "3d dlo shape detection and grasp planning from multiple 2d views," in *2021 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM)*. IEEE, 2021, pp. 424–429.

[3] A. Choi, D. Tong, B. Park, D. Terzopoulos, J. Joo, and M. K. Jawed, "mbest: Realtime deformable linear object detection through minimal bending energy skeleton pixel traversals." *IEEE Robotics Autom. Lett.*, vol. 8, no. 8, pp. 4863–4870, 2023.

[4] L. Huang, T. Yang, X. Tian, R. Jiang, and Y. Chen, "Dlodepth: Real-time depth recovery for 3d reflective deformable linear object," *IEEE Robotics and Automation Letters*, 2026.

[5] 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.

[6] F. Tosi, L. Bartolomei, and M. Poggi, "A survey on deep stereo matching in the twenties," *International Journal of Computer Vision*, vol. 133, no. 7, pp. 4245–4276, 2025.

[7] W. Su, Q. Xu, and W. Tao, "Uncertainty guided multi-view stereo network for depth estimation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7796–7808, 2022.