Robust Pose Estimation through Failure Explanation and Mitigation
Extended Abstract

***Abstract*— Robust estimation of object poses in robotic manipulation is often addressed using foundational general estimators, that aim to handle diverse error sources implicitly within a single model. Still, they struggle due to environmental uncertainties, while requiring long inference times and heavy computation. In contrast, we propose a modular, uncertainty-aware framework that attributes pose estimation errors to specific error sources and applies targeted mitigation strategies only when necessary. Instantiated with Iterative Closest Point (ICP) as a simple and lightweight pose estimator, we leverage our framework for real-world robotic grasping tasks. By decomposing pose estimation into failure detection, error attribution, and targeted recovery, we significantly improve the robustness of ICP and achieve competitive performance compared to foundation models, while relying on a substantially simpler and faster pose estimator. We illustrate our framework and show examples of real-world manipulation in our video available at `https://drive.google.com/file/d/1iQAVXFynVutyZwhQ5WAnvzMOQ3GHWU0d/view?usp=sharing`.**

I. INTRODUCTION

Object pose estimation is a central problem in robotic perception [1]. It aims to recover the full six degrees of freedom (6-DoF) pose of objects from sensory input and remains challenging due to environmental uncertainty [2]. Occlusion can obscure geometric cues, while depth sensors are particularly prone to structured noise artifacts caused by dark or reflective surfaces [3], [4], [5].

A wide range of pose estimation methods has been proposed [6]. A classical approach that remains widely used is Iterative Closest Point (ICP) [7], which aims to align model and scene point clouds. However, it is sensitive to initialization, noise, and incomplete observations [8]. More recently, deep-learning-based methods, including foundation models such as FoundationPose (FP) [9], achieve strong performance by combining multiple modalities and learned representations. Despite this progress, these approaches remain vulnerable to real-world uncertainty and often require significant computational resources [10].

Beyond pose estimation itself, prior work has investigated identifying and addressing failure sources such as noise, occlusion, and poor initialization [11]. Existing approaches include attribution methods based on explainability techniques or learned classifiers [12], [13], as well as mitigation strategies such as denoising, improved initialization, and viewpoint planning [14], [15], [16]. However, these directions are typically studied in isolation.

Instead of aiming for a single universally robust estimator, we propose a complementary approach: explicitly detecting failures, attributing their causes, and applying targeted mitigation strategies within a unified framework, as illustrated in Fig. 1.

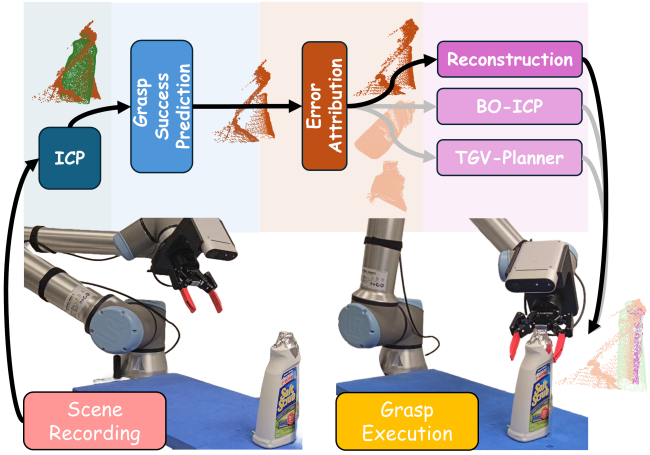

Fig. 1. Overview of our framework. After the scene is recorded, ICP provides an initial pose estimate. A grasp success predictor detects possible grasp failure, which queries an error attribution system. Based on the detected error, one of the mitigation strategies will be selected.

We instantiate this idea using ICP as a lightweight pose estimator augmented with failure detection, error attribution, and mitigation modules. This design enables efficient handling of low-uncertainty cases while improving robustness in challenging scenarios, achieving performance comparable to state-of-the-art methods such as FP [9].

In summary, our contributions are:

- A grasping framework integrating ICP-based pose estimation with failure detection, error attribution, and targeted mitigation.
- A point cloud reconstruction method tailored to address ICP failures caused by real-world noise.
- A transformer-based error attribution model that improves classification accuracy and resolves confusion between occlusion and poor initialization.
- A real-world evaluation demonstrating competitive performance with, and improved robustness over, FoundationPose [9] in noisy settings.

II. METHODOLOGY

We leverage our framework for pose estimation in robotic manipulation tasks. A lightweight point-to-point ICP is used for pose estimation, followed by grasp success prediction. Targeted mitigation strategies are deployed only when specific error sources are detected. Fig. 2 shows the overall framework.

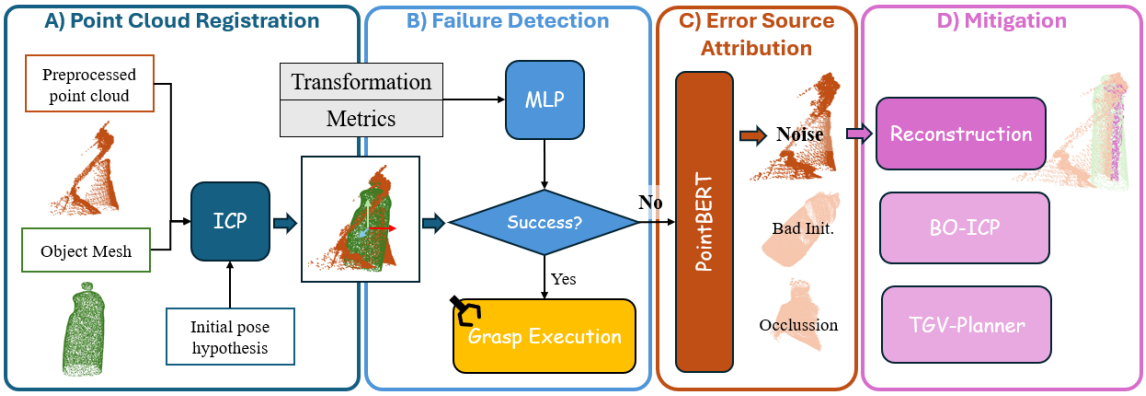

Fig. 2. Overview of the deployed framework. (A) A point cloud sampled from the object mesh is aligned to the recorded scene point cloud via ICP. (B) An MLP predicts grasp success based on the calculated transformation estimate and alignment metrics. (C) In the case of predicted grasp failure, a PointBERT [17] classifier attributes the failure to a specific source, here shown for noise. (D) Based on the classification, targeted mitigation strategies - BO-ICP [18], the TGV-Planner [16], and a custom point cloud reconstruction model - are applied to recover the pose estimate. In the illustrated case, the clean point cloud is reconstructed from the noisy point cloud and used for ICP alignment, which corrects the pose estimate.

A. Scene Recording and Pre-processing

The environment is recorded using a depth sensor to obtain a raw point cloud. The scene is cropped and segmented to remove irrelevant background, resulting in a processed point cloud $P \in \mathbb{R}^{N\times 3}$. A model point cloud P_M is sampled from the object mesh and aligned to P using ICP with an initial pose hypothesis. The resulting alignment is then evaluated using geometric quality metrics.

B. Failure Detection and Error Source Attribution

After ICP registration, we predict whether the result will lead to a successful grasp. We compute alignment quality metrics, including fitness ratios F at different thresholds, an inlier RMSE, and a bidirectional point cloud distance (Chamfer distance).

These metrics, together with the estimated transformation T, are fed into a multi-layer perceptron (MLP) to predict a binary success label:

$$S = \mathrm{MLP}(F, \mathrm{RMSE}_{\mathrm{inlier}}, D_{\mathrm{Chamfer}}, T).$$

If failure is predicted, grasp execution is aborted and error attribution is performed using a PointBERT [17] classifier on the scene point cloud P, producing probabilities over the discrete error classes bad initialization, noise, and occlusion.

C. Mitigation Strategies

Based on the predicted error class, we apply targeted mitigation strategies.

1) Point Cloud Reconstruction: ICP is robust to small surface jitter but fails under structured noise artifacts that may appear, for example, due to reflective distortions. Such structured artifacts can hardly be removed by classical filtering approaches. We address this with a reconstruction model based on PoinTr [19]. The scene point cloud is partitioned into local patches, encoded into tokens via a PointNet-style encoder and passed through a transformer architecture which results in scene tokens $\mathcal{V}$. In parallel, mesh features are extracted using DGCNN [20] and subsampled via farthest-point sampling (FPS), generating mesh proxies $\mathcal{O}$.

We compute pairwise weights between scene tokens $v_i \in \mathcal{V}$ and mesh proxies $o_j \in \mathcal{O}$:

$$w_{i,j} = \mathrm{Linear}(v_i, o_j).$$

Each token is fused with the mesh features using a weighted sum:

$$u_i = \mathrm{Linear}\left(v_i, \sum_j w_{i,j} o_j\right).$$

A transformer decoder predicts displacement vectors for patch centers, which are added to those center points to obtain reconstructed points. The displacement is propagated to neighboring points via distance-weighted averaging. The reconstructed point cloud is then used for ICP alignment.

2) Registration Initialization: ICP is sensitive to initialization and may converge to local minima. We mitigate this using BO-ICP [18], which optimizes the initial transformation.

Each candidate transformation is evaluated by running ICP and computing an objective:

$$\mathcal{L} = \begin{cases} F - \mathrm{RMSE}_{\mathrm{inlier}}, & \text{if overlap exists,} \\ -D_{\mathrm{Chamfer}}, & \text{otherwise.} \end{cases}$$

This enables recovery from large misalignments by searching for transformations that maximize overlap. However, ambiguities due to symmetry or missing geometry cannot be resolved.

3) Occlusion: Occlusions lead to incomplete observations and degraded ICP performance. To address this, we actively change the viewpoint using the Target-Guided View Planner (TGV-Planner) [16], which generates smooth camera trajectories based on a velocity field. This allows continuous view optimization and improved visibility compared to discrete next-best-view methods [21].

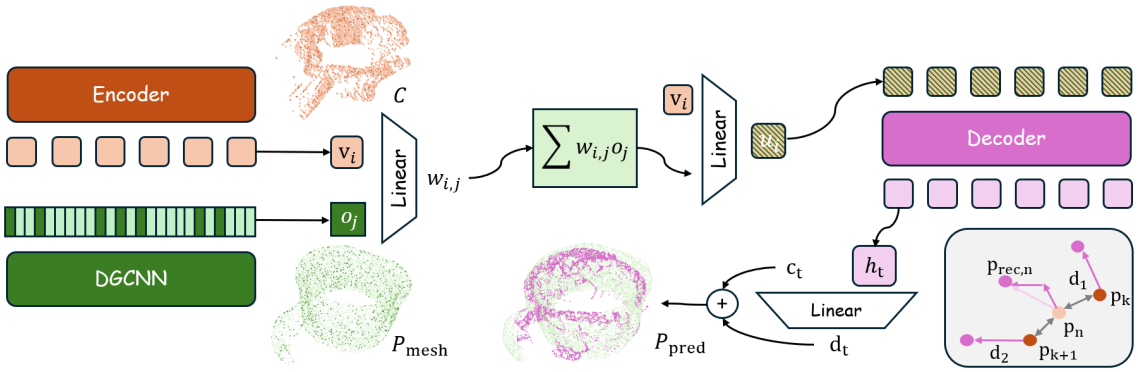

Fig. 3. Point cloud reconstruction module. A transformer-based encoder creates tokens from point patches of the noisy point cloud, while a DGCNN [20] encodes point-wise features of the mesh. From the point-wise mesh features, a subset is selected via FPS. The tokens from the noisy point cloud are merged with the selected point-wise mesh features by calculating a weighted sum with learned weights. An encoder predicts displacement vectors for each patch center point of the noisy point cloud, which is propagated to the surrounding points.

TABLE I

ACCURACY OF ERROR ATTRIBUTION ON REAL-WORLD SCENES.

Model	Mean Acc.	Std. Dev.
PointBERT	83.83%	2.27%
DGCNN	71.49%	2.55%
PointNet	50.64%	3.40%

TABLE II

GRASP SUCCESS COMPARISON OF OUR FRAMEWORK AGAINST FP [9] FOR EACH ERROR CASE.

Error Case	ICP only	ICP&Mitigation	FP
Bad Init.	0%	55%	**85%**
Noise	15%	**80%**	30%
Occlusion	10%	**70%**	**70%**

III. EXPERIMENTAL RESULTS

We evaluate our framework on real-world scenes constructed from nine YCB objects [22], covering bad initialization, noise artifacts, and occlusion (Fig. 4). Initial poses and grasp configurations are fixed per object.

Bad initialization is created by placing objects far from their initial pose hypothesis. Noise artifacts are induced using reflective aluminum foil, creating structured depth distortions. Occlusion is introduced by partially blocking the object while maintaining minimal visibility.

We collect 81 scenes with multiple viewpoints and filter them to obtain a balanced dataset of 200 samples, from which 50 are used for evaluation on unseen objects.

A. Error Attribution

We evaluate PointBERT for error attribution on the dataset of [13]. It achieves a mean accuracy of 99.47%, significantly outperforming DGCNN (85.20%). In addition, it almost completely resolves the confusion between bad initialization and occlusion that was previously reported by [13], reducing misclassification rates from over 20% to around 1%.

On 50 real-world samples, PointBERT achieves a mean accuracy of 83.83%, outperforming DGCNN (71.49%) and PointNet [23] (50.64%) as shown in Tab. I. This supports the use of transformer-based models for robust attribution.

B. Framework Evaluation and Baseline Comparison

Finally, we evaluate our complete framework on grasping tasks. For each error case, we create 20 real-world scenes and execute a grasping task using our framework of ICP pose estimation, success prediction, error attribution and mitigation, and grasp execution. As a baseline, we compare grasp success rates to using FP [9] for pose estimation instead of our framework. While our framework allows targeted mitigation after failure detection, FP performs single-pass pose estimation and leverages multimodal input, including RGB information.

a) Grasp success: The results are displayed in Table II. Generally, the targeted mitigation of error cases led to a significant increase in grasp success, while still only relying on ICP for pose estimation. As FP does not rely on an initial pose hypothesis, it is not affected by bad initialization, achieving 85% success rate in this case compared to 55% success rate using our framework. The success rate dropped for occlusion cases but was still as high as 70%, on par with our framework. In the case of noise, FP was not able to predict a sufficiently accurate pose for grasp success in most cases, with a grasp success rate of only 30% compared to 80% success rate of our framework. Overall, the grasp success rate of FP was 61.7% and as such only slightly higher than the 60% achieved by our framework.

We report the success rate or ICP without mitigation separately in Table II. The sensitivity of ICP to the investigated error cases is evident with 0% grasp success for bad initialization, 10% for occlusion, and 15% for noise.

The results show that our approach of detecting uncertainty, attributing it to specific sources, and performing targeted mitigation can significantly increase the performance of the otherwise limited ICP. Our framework was able to

Fig. 4. Overview of the real-world scenes used for evaluation and training. For each object, we show one example of each error case.

achieve comparable grasp success rates to FP, all while relying only on point clouds as data modality and ICP as a lightweight and simple classical pose estimator.

IV. Conclusion

We presented a structured alternative to general pose estimation methods by decomposing the problem into pose estimation, error attribution, and targeted mitigation. Our framework significantly improves grasp success compared to ICP alone and achieves performance comparable to FoundationPose [9], while remaining computationally efficient.

We further demonstrated improved real-world error attribution using PointBERT [17], alongside a lightweight success predictor and a reconstruction module for handling structured noise.

These results indicate that robust pose estimation does not necessarily require heavy foundation models. Instead, explicitly modeling and mitigating error sources enables simpler methods to achieve competitive performance.

Future work should extend this approach to modern pose estimators and further investigate systematic error modeling and mitigation strategies.

References

[1] Y. Jin, N. Funk, V. Prasad, Z. Li, M. Franzius, J. Peters, and G. Chalvatzaki, "Se(3)-poseflow: Estimating 6d pose distributions for uncertainty-aware robotic manipulation," *arXiv preprint arXiv:2511.01501*, 2025.

[2] S. Thalhammer, D. Bauer, P. Hönig, J.-B. Weibel, J. García-Rodríguez, and M. Vincze, "Challenges for monocular 6-d object pose estimation in robotics," *T-RO*, vol. 40, pp. 4065–4084, 2024.

[3] T.-C. Hsiao, H.-W. Chen, H.-K. Yang, and C.-Y. Lee, "Confronting ambiguity in 6d object pose estimation via score-based diffusion on se(3)," in *CVPR*, 2024.

[4] Z. Huang, K. Yao, S. Z. Zhao, C. Pan, and A. Y. Yang, "Robust 6dof pose estimation against depth noise and a comprehensive evaluation on a mobile dataset," in *CVPRW*, 2025.

[5] S. Sajjan, M. Moore, M. Pan, G. Nagaraja, J. Lee, A. Zeng, and S. Song, "Clear grasp: 3d shape estimation of transparent objects for manipulation," in *ICRA*, 2020.

[6] Y. Wu, H. Zhang, P. Vandewalle, P. Slaets, and E. Demeester, "A comprehensive review on advances in instance-level 6d object pose tracking," *Computer Vision and Image Understanding*, vol. 264, p. 104667, 2026.

[7] P. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[8] Z. Deng, Y. Yao, B. Deng, and J. Zhang, "A robust loss for point cloud registration," in *ICCV*, 2021.

[9] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "FoundationPose: Unified 6d pose estimation and tracking of novel objects," in *CVPR*, 2024.

[10] V. N. Nguyen, S. Tyree, A. Guo, M. Fourmy, A. Gouda, T. Lee, S. Moon, H. Son, L. Ranftl, J. Tremblay, E. Brachmann, B. Drost, V. Lepetit, C. Rother, S. Birchfield, J. Matas, Y. Labbe, M. Sundermeyer, and T. Hodan, "BOP challenge 2024 on model-based and model-free 6D object pose estimation," in *CVPRW, CV4MR Workshop*, 2025.

[11] A. Censi, "An accurate closed-form estimate of icp's covariance," in *ICRA*, 2007.

[12] Z. Qin, J. Lee, and R. Triebel, "Towards explaining uncertainty estimates in point cloud registration," *arXiv preprint arXiv:2412.20612*, 2024.

[13] J. A. Gaus, L. Schneider, Y. Shi, J. Lee, R. Rayyes, and R. Triebel, "Human-interpretable uncertainty explanations for point cloud registration," *arXiv preprint arXiv:2509.18786*, 2025.

[14] X. Yu, Y. Rao, Z. Wang, Z. Liu, J. Lu, and J. Zhou, "Pointr: Diverse point cloud completion with geometry-aware transformers," in *ICCV*, 2021.

[15] J. Yang, H. Li, D. Campbell, and Y. Jia, "Go-icp: A globally optimal solution to 3d icp point-set registration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 11, pp. 2241–2254, 2016.

[16] Y. Shi, D. Wen, G. Chen, E. Welte, S. Liu, K. Peng, R. Stiefelhagen, and R. Rayyes, "Viso-grasp: Vision-language informed spatial object-centric 6-dof active view planning and grasping in clutter and invisibility," in *IROS*, 2025.

[17] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *CVPR*, 2022.

[18] H. Biggie, A. Beathard, and C. Heckman, "Bo-icp: Initialization of iterative closest point based on bayesian optimization," in *ICRA*, 2023.

[19] X. Yu, Y. Rao, Z. Wang, J. Lu, and J. Zhou, "Adapointr: Diverse point cloud completion with adaptive geometry-aware transformers," *TPAMI*, vol. 45, no. 12, pp. 14 114–14 130, 2023.

[20] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Trans. Graph.*, vol. 38, no. 5, 2019.

[21] M. Breyer, L. Ott, R. Siegwart, and J. J. Chung, "Closed-loop next-best-view planning for target-driven grasping," in *IROS*, 2022.

[22] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set," *RAM*, vol. 22, no. 3, pp. 36–52, 2015.

[23] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," *arXiv preprint arXiv:1612.00593*, 2016.