On the miscalibration of object 6D pose estimation methods

Matteo Bortolon and Fabio Poiesi

***Abstract*— Uncertainty estimation in 6D object pose estimation is crucial for industrial robotic applications, where short cycle times and near-optimal availability are essential. The primary concern is not just the accuracy of the pose, but the reliability of the confidence scores to prevent costly robotic errors, such as crashes. Despite the great strides in accuracy demonstrated by leading methods in benchmarks like the BOP Challenge, we have observed that their confidence scores are often poorly calibrated. This poor calibration results in a significant gap between a method's reported performance and its true reliability in industrial settings. In this paper, we conduct a experimental study on this issue across multiple BOP datasets. We demonstrate that even after applying state-of-the-art calibration techniques, this miscalibration persists. Our findings highlight the limitations of current post-hoc calibration methods, which we found do not work "out of the box" and have notable drawbacks. This indicates a deeper, systemic issue that requires fundamental changes at the model or foundation model training level. Our work underscores the critical need for new calibration approaches tailored to the unique challenges of 6D pose estimation for robust industrial deployment.**

I. INTRODUCTION

Uncertainty quantification and calibration are fundamental for deploying 6D pose estimation systems in industrial settings [1]. In applications such as robotic manipulation on an assembly line, systems must not only predict poses accurately but also provide reliable confidence measures [2]. Poorly calibrated systems, overly confident in wrong predictions or insufficiently confident in correct ones, can cause collisions, part damage, and production delays. For industrial deployment requiring short cycle times and high availability, systems must flag unreliable predictions for intervention. Lack of reliable uncertainty estimation, rather than pose accuracy alone, hinders adoption of top-performing academic methods in real-world industrial applications.

6D pose estimation has evolved from handcrafted approaches to sophisticated deep learning systems [3]–[7], culminating in foundation model-based methods that estimate poses without task-specific training [8]. Early methods relied on handcrafted features and geometric constraints [9], [10] but suffered from limited robustness to lighting, occlusion, and appearance variations [6]. Deep learning brought improvements through object-specific networks directly regressing pose parameters [11], [12], though with limitations in generalization and precision for symmetric or textureless objects [13], [14]. Hybrid approaches combined neural networks with geometric techniques [13]–[15], while recent methods like Co-op [3] and FoundationPose [4] achieved high accuracy in zero-shot scenarios. Foundation model-based approaches like FreeZe [8] now eliminate training requirements by leveraging pre-trained geometric and vision models, achieving state-of-the-art performance.

Despite substantial advances in pose accuracy, confidence scores predicted by these systems remain poorly calibrated and fail to reflect prediction quality accurately. Neural network miscalibration is a well-documented phenomenon in machine learning [16], where predicted confidence scores poorly reflect actual prediction accuracy. While calibration techniques like Platt scaling [17] and isotonic regression [18] address miscalibration in classification [16], their effectiveness in 6D pose estimation remains unexplored. Our analysis reveals severe miscalibration: across foundation model-based methods, Pearson correlation coefficients between confidence scores and actual accuracy range from 0.165 to 0.792, indicating strong dependence on scoring methodology and object characteristics. This issue is particularly concerning for robotics applications requiring reliable uncertainty estimation for safe operation.

II. OUR ANALYSIS

We analyze confidence score calibration in 6D pose estimation through three components: BOP metrics evaluation for pose quality assessment, Hungarian algorithm-based prediction-ground truth matching, and isotonic regression for calibration analysis. We evaluate state-of-the-art model-based methods across multiple BOP datasets to assess confidence score reliability comprehensively.

A. Pose quality evaluation

We examine the relationship between predicted confidence scores and pose estimation accuracy across state-of-the-art methods that excel in recent BOP challenges. We employ the popular BOP evaluation protocol [5] using three complementary pose-error functions: *Visible Surface Discrepancy (VSD)* quantifies misalignment between visible object surfaces using rendered distance maps, treating perceptually equivalent poses as identical. *Maximum Symmetry-Aware Surface Distance (MSSD)* computes maximum vertex distances after accounting for symmetries, critical for robotic manipulation where surface deviations affect grasp success. *Maximum Symmetry-Aware Projection Distance (MSPD)* measures maximum projection distances in the image plane while handling symmetries, essential for RGB-only evaluation. The final pose quality score AR_e aggregates these metrics $e \in \{\mathrm{VSD}, \mathrm{MSSD}, \mathrm{MSPD}\}$ across multiple thresholds, representing average recall.

Technologies of Vision, Fondazione Bruno Kessler, Trento, Italy {mbortolon, poiesi}@fbk.eu. This work was supported by PNRR FAIR - Future AI Research (PE00000013).

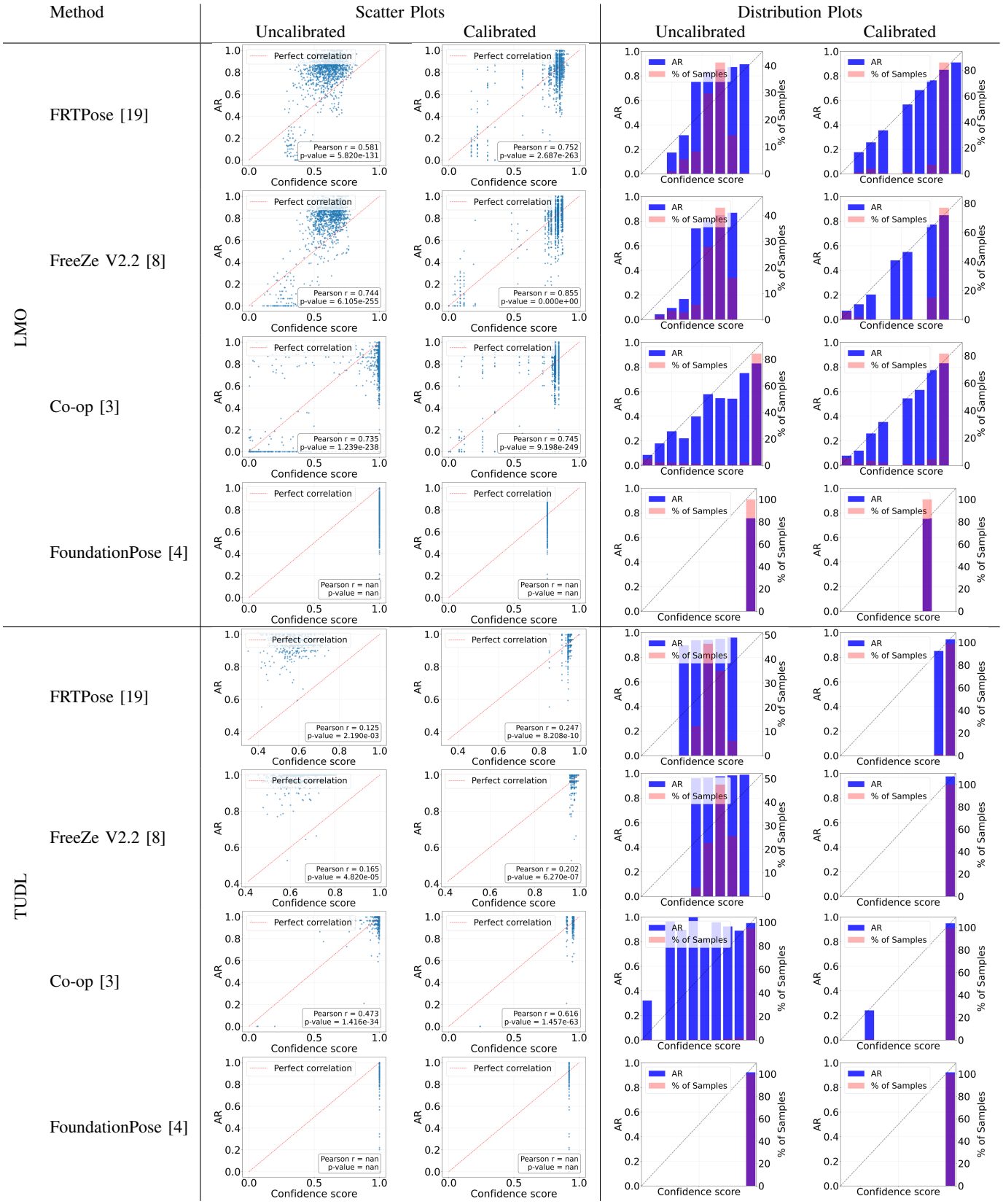

Fig. 1. Analysis of confidence score calibration for LMO and TUDL datasets. Left columns show scatter plots of confidence vs accuracy correlation, right columns show score distributions for correct vs incorrect predictions. Each dataset section shows four methods (FRTPose, FreeZe V2.2, Co-op, FoundationPose) with uncalibrated vs calibrated results.

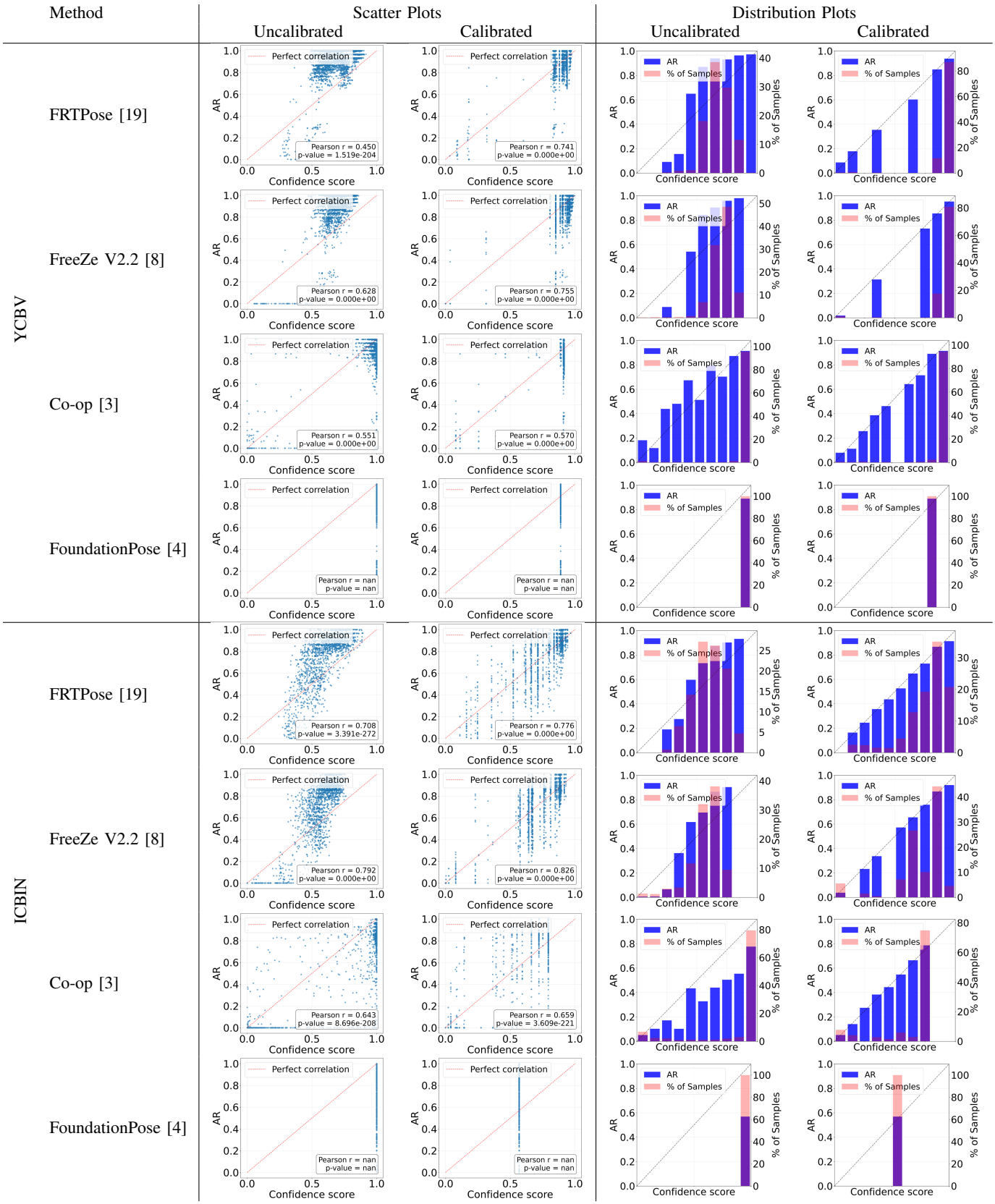

Fig. 2. Analysis of confidence score calibration for YCBV and ICBIN datasets. Left columns show scatter plots of confidence vs accuracy correlation, right columns show score distributions for correct vs incorrect predictions. Each dataset section shows four methods (FRTPose, FreeZe V2.2, Co-op, FoundationPose) with uncalibrated vs calibrated results.

B. Prediction-ground truth assignment

We establish optimal correspondences between predicted detections and ground truth annotations using the Hungarian algorithm [20]. This bipartite matching ensures one-to-one assignments that maximize matching quality while handling multiple object instances per scene.

Each matched prediction is classified as correct or incorrect based on whether its pose error falls below the respective BOP metric threshold. This binary classification, paired with the original confidence score, enables direct calibration analysis by comparing predicted confidence against actual prediction accuracy.

C. Calibration analysis

We apply isotonic regression [21] to study score miscalibration patterns. Following Kuzucu et al. [2] object detection calibration approach, this non-parametric approach preserves confidence score ranking while learning optimal mappings to well-calibrated probabilities. The method fits a monotonic function that transforms original scores based on the empirical confidence-accuracy relationship. This enables quantification of miscalibration severity and assessment of whether standard post-processing techniques can address the identified calibration issues.

III. Experiments

A. Experimental setup

We evaluate four state-of-the-art 6D pose estimation methods: FRTPose [19], FreeZe [8], Co-op [3], and FoundationPose [4], representing both geometric and learning-based approaches. Our analysis spans four BOP Challenge datasets: ICBIN [5] (industrial objects with multiple instances), LMO [5] (diverse objects under occlusion), TUDL [5] (single object with challenging lighting), and YCBV [22] (21 household objects with background diversity).

For each method-dataset combination, we extract confidence scores and pose predictions, applying our Hungarian algorithm-based matching to establish correspondences. We quantify confidence-accuracy relationships using Pearson correlation coefficients, measuring linear association between predicted confidence and actual pose quality from -1 to 1.

B. Calibration analysis

Our scatter plot analysis in Fig. 1 and Fig. 2 reveals systematic calibration failures with distinct patterns across methods and datasets. Correlation coefficients span from strong (r=0.792 for FreeZe V2.2 on ICBIN) to poor (r=0.165 for FreeZe V2.2 on TUDL), demonstrating significant dependence on both method and dataset characteristics.

Geometric score-based methods (FreeZe V2.2, FRTPose) using inlier ratios exhibit high-variance confidence distributions with broader spectral coverage but inconsistent reliability. Learning-based methods (FoundationPose, Co-op) suffer from severe confidence saturation, with FoundationPose clustering at maximum confidence regardless of accuracy. Critical is the systematic overconfidence where zero-precision predictions ($AR = 0.0$) receive high confidence scores across all methods, particularly in YCBV and ICBIN.

Isotonic regression calibration shows limited effectiveness, marginally improving correlations but primarily redistributing rather than meaningfully aligning confidence with accuracy, failing to address confidence saturation issues.

C. Cross-dataset comparative analysis

Cross-dataset comparison reveals confidence patterns strongly correlated with object characteristics, geometric complexity, and imaging conditions.

TUDL exhibits the most severe miscalibration across all methods (r=0.165 for FreeZe V2.2), with predictions clustering in highest confidence bins (0.9-1.0) regardless of accuracy. This creates horizontal banding in calibrated scatter plots, particularly affecting geometric methods where inlier ratios become unreliable under lighting variations.

YCBV demonstrates method-dependent behavior: geometric approaches achieve moderate correlations (r=0.628 for FreeZe V2.2, r=0.450 for FRTPose) with bimodal distributions, while learning-based methods show "L-shaped" patterns where both high-accuracy and zero-accuracy predictions receive high confidence scores.

ICBIN and LMO yield optimal performance, with ICBIN achieving peak correlation (r=0.792 for FreeZe V2.2) and LMO maintaining consistency (r=0.744 for FreeZe V2.2, r=0.581 for FRTPose). Geometric methods demonstrate most reliable behavior with linear confidence-accuracy relationships, yet learning-based methods still exhibit confidence saturation even under favorable conditions. This highlighting that these issues are not merely dataset-dependent but reflect fundamental architectural limitations in the score generation.

Our analysis confirms that geometric scoring provides more interpretable uncertainty estimates despite higher variance, while learning-based approaches suffer from systematic confidence saturation independent of dataset complexity.

IV. Conclusions

Our study reveals systematic confidence score miscalibration in state-of-the-art 6D pose estimation methods across multiple BOP datasets. We identify fundamental differences between geometric methods (FreeZe V2.2, FRTPose), which provide meaningful but high-variance uncertainty estimates, and learning-based approaches (FoundationPose, Co-op), which suffer from severe confidence saturation. We found that recent post-hoc calibration techniques were largely ineffective, suggesting that a more fundamental solution is required. The systematic overconfidence observed poses significant safety concerns for industrial robotics deployment. Our findings also highlight the critical need for calibration-aware architectures and evaluation protocols that assess both accuracy and calibration quality in 6D pose estimation.

References

[1] P. Quentin, D. Knoll, and D. Goehring, "Industrial application of 6d pose estimation for robotic manipulation in automotive internal logistics," in *CASE*, 2023.

[2] S. Kuzucu, K. Oksuz, J. Sadeghi, and P. K. Dokania, "On calibration of object detectors: Pitfalls, evaluation and baselines," in *ECCV*, 2024.

[3] J. Lee *et al.*, "Co-op: Correspondence-based novel object pose estimation," in *CVPR*, 2025.

[4] B. Wen *et al.*, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *CVPR*, 2024.

[5] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. G. Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T.-K. Kim, J. Matas, and C. Rother, "Bop: Benchmark for 6d object pose estimation," 2018.

[6] Y. Zhu, M. Li, W. Yao, and C. Chen, "A review of 6d object pose estimation," in *ITAIC*, 2022.

[7] G. Marullo, L. Tanzi, P. Piazzolla, and E. Vezzetti, "6d object position estimation from 2d images: a literature review," *Multimedia Tools and Applications*, vol. 82, no. 16, 2023.

[8] A. Caraffa *et al.*, "Freeze: Training-free zero-shot 6d pose estimation with geometric and vision foundation models," in *ECCV*, 2024.

[9] D. G. Lowe, "Object recognition from local scale-invariant features," in *CVPR*, 1999.

[10] F. Tombari, S. Salti, and L. di Stefano, "Unique signatures of histograms for local surface description," in *ECCV*, 2010.

[11] A. Amini, A. S. Periyasamy, and S. Behnke, "T6d-direct: Transformers for multi-object 6d pose direct regression," in *DAGM GCPR*, 2021.

[12] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun, "Ffb6d: A full flow bidirectional fusion network for 6d pose estimation," in *CVPR*, 2021.

[13] Y. Di, F. Manhardt, G. Wang, X. Ji, N. Navab, and F. Tombari, "So-pose: Exploiting self-occlusion for direct 6d pose estimation," in *ICCV*, 2021.

[14] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *CVPR*, 2021.

[15] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6d object pose estimation using 3d object coordinates," in *ECCV*, 2014.

[16] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," *ICML*, 2017.

[17] J. Platt *et al.*, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," in *Advances in large margin classifiers*, 1999.

[18] B. Zadrozny and C. Elkan, "Transforming classifier scores into accurate multiclass probability estimates," in *KDD*, 2002.

[19] A. Authors, "Frtpose: 6d object pose estimation method," 2025, submitted to BOP Challenge 2025.

[20] H. W. Kuhn, "The hungarian method for the assignment problem," *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, A. Müller, J. Nothman, G. Louppe, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," 2018.

[22] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," in *CoRR*, 2017.