Lessons and Winning Solutions in Industrial Object Detection and Pose Estimation from the 2025 Bin-picking Perception Challenge

Ziqin Huang[1,*] Chengxi Li[1,*] Yingyue Li[1,*] Xingyu Liu[1,*] Chenyangguang Zhang[1,†]
Ruida Zhang[1] Bowen Fu[1] Xinggang Hu[2] Yun Qu[1] Mengge Liu[1]
Yixiu Mao[1] Wendong Huang[1] Gu Wang[1,†] Xiangyang Ji[1,†]
[1]Tsinghua University [2]Dalian University of Technology
{huang-zq24@mails.,wanggu1@,xyji@}tsinghua.edu.cn

Abstract

This paper analyzes the challenges encountered in object pose estimation tasks within industrial environments, based on our winning solutions in the 2025 Perception Challenge for Bin-Picking[1]*. We discuss several strategies employed during the competition and highlight two unexpected observations: (1) methods trained on object-specific datasets performed worse than those trained on unseen data; and (2) evaluation results varied significantly depending on the chosen metrics. Through a detailed analysis of these findings, we aim to provide researchers with a deeper understanding of the complexities involved in industrial object pose estimation and offer insights to improve the practical deployment of such systems. Our code is available at* https://github.com/ziqin-h/GBP2025.

1. Introduction

Object pose estimation in industrial environments presents a highly challenging yet critical task in computer vision. With the rapid advancement of deep learning in recent years, many datasets [2, 4, 6, 7] and methodologies [3, 13, 16] tailored for this task have emerged, yielding promising results. However, existing approaches often lack a comprehensive evaluation pipeline that spans from dataset construction to real-world scenario validation. This gap creates a significant divide between academic advancements and their practical applications in industry.

The Perception Challenge for Bin Picking (BPC), organized by OpenCV, addresses this gap by providing an evaluation framework that spans data-driven method development to physical system validation. This paper presents two of our prize-winning approaches from the competition, which secured first and third places.

The paper is structured as follows: In Sec. 2, we describe the detection and segmentation strategies employed. Sec. 3 presents the various pose estimation methods and optimization strategies we explored. Fig. 1 provides an overview of our approaches discussed in Sec. 2 and Sec. 3. In Sec. 4, we provide a overview of the datasets and evaluation metrics used in the competition, present the performance of different methods, and offer a detailed analysis. Finally, in Sec. 5, we summarize the conclusions and discuss potential directions for future research.

2. Object Detection and Segmentation

In the object detection and segmentation stage, we employ a YOLOv12-based object detection pipeline, enhanced with several strategies, including size filtering, edge filtering, and crop & fuse. Following detection, segmentation masks for identified instances are obtained using SAM 2 [11].

YOLOv12 [14] is a SOTA attention-centric YOLO [12] framework, which incorporates three key optimizations, including area attention module (A2), residual efficient layer aggregation networks (R-ELAN), and more architecture-level improvements beyond the standard attention mechanisms to better fit the YOLO system. These combined improvements enable YOLOv12 to deliver both fast inference speed and superior detection accuracy compared to previous versions of the YOLO series. Based on these strengths, we adopt YOLOv12 in our pipeline for object detection. More specifically, we employ the YOLOv12-X variant for higher accuracy.

The inherent characteristics of industrial environments, including the low texture of industrial parts and challenging lighting conditions, present substantial difficulties. Consequently, direct application of YOLOv12-X resulted in degraded performance, manifesting as misclassifications, missed detections of small instances, and increased false positives. To address these issues, we employ several im-

[1]https://bpc.opencv.org/
[*] Equal contribution.
[†] Corresponding authors.

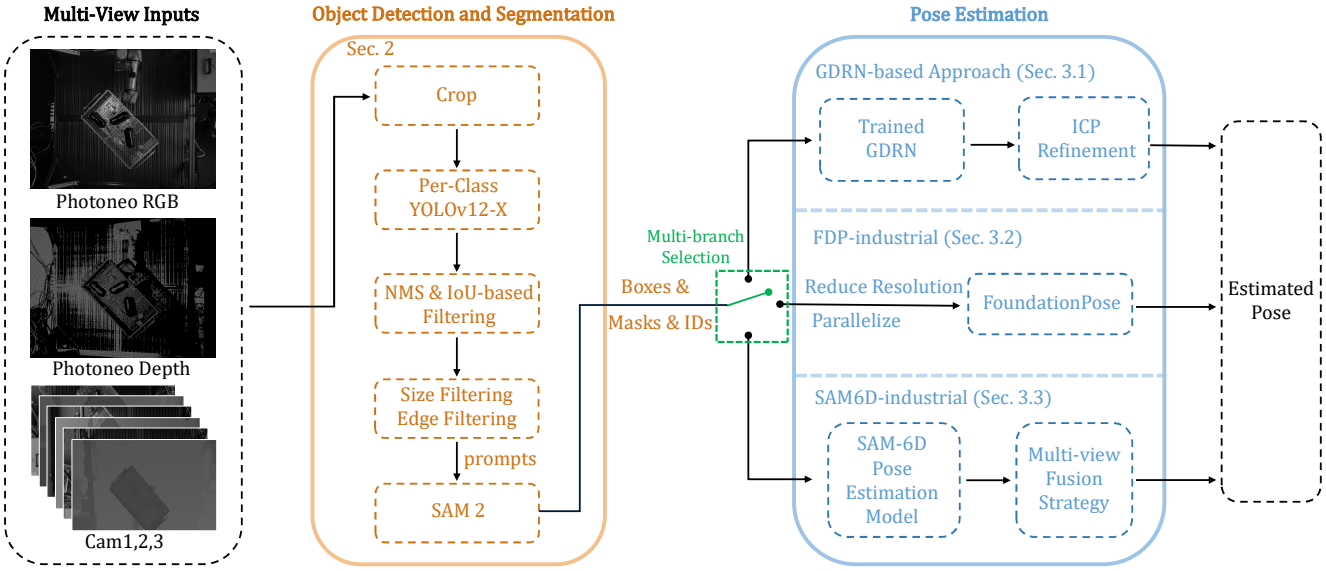


Figure 1. Overall pipeline of our approaches. Each strategy in pose estimation corresponds to a distinct method.

provement strategies, as detailed below. First, to tackle classification errors, we train an individual YOLOv12-X model for each object category, enabling highly class-specific detection. During inference, the outputs of all models were aggregated using per-scene category priors, which effectively allows our approach to distinguish between visually similar categories that were previously prone to confusion.

Second, to address missed detections, we adopt a Crop & Fuse strategy. Specifically, each input image is cropped into four pieces with overlap: top left, top right, bottom left, and bottom right. During inference, detection is performed on both the original image and all crops independently. In the fusion stage, we utilize two filtering techniques: Non-Maximum Suppression (NMS) and our IoU-based filtering strategy. NMS is applied to eliminate redundant bounding boxes, while the IoU-based strategy is designed to address incomplete detections caused by cropping. To this end, we compare overlapping bounding box pairs by computing the ratio of their intersection area to the respective area of each box. If a smaller box exhibits substantial overlap with a larger one and had a lower confidence score, it should be considered a partial detection and subsequently removed. This Crop & Fuse strategy significantly improves the ability of our approach to detect small objects which were otherwise missed in the original resolution.

Finally, to reduce false positives, we employ filtering operations on the detection results. We define a size-based threshold as the ratio of the bounding box area to the image area, and discard all detections whose area ratio exceed the threshold. In addition, considering that the workspace in the photoneo view consistently lies near the image center, we apply edge filtering to remove detections located around the image boundaries.

Furthermore, following GDRNPP [9], we add lighting augmentation to the original YOLOv12 [14] data augmentation, which made our method more robust to varying light conditions. By utilizing aforementioned improvements, our detection pipeline demonstrates strong capability in robust and accurate object detection.

After obtaining detected bounding boxes from our detection pipeline, we use these boxes as prompts for SAM 2 [11] to generate fine-grained, pixel-level instance segmentation results.

3. Pose Estimation

3.1. GDRN-based Approach

For the pose estimation module, our initial attempt was a solution based on an instance-level RGB-based object pose estimation method GDRN [9, 15]. Specifically, we train the GDRN model on synthetic datasets and directly apply it to the testing. However, the method initially underperformed because of the absence of depth information. To fully exploit depth cues, we replace the translation component of the predicted pose of the network with the median depth value within the masked region.

Furthermore, similarly to the fast refinement of GDRNPP [9], we further refine the initial estimated pose using the ICP algorithm (Iterative Closest Point) [1], which matches the back-projected depth points of the masked region with the sampled points from the object model. Considering that back-projected points from the masked depth only represent partial object surfaces, we replace the object model's sampled points with visible points under the

initial pose projection, a strategy validated to enhance performance. Despite these optimizations, our method still exhibits a significant gap from practical industrial deployment.

3.2. FDP-industrial

We hypothesize that the limited performance of GDRN-based methods, which require training from scratch, is primarily due to the low quality of synthetic training data in industrial scenarios. Therefore, we explore FDP-industrial, *i.e*. an approach based on FoundationPose [17] designed for unseen object pose estimation. The primary challenge lies in the implementation of this method under constrained hardware and time budgets of the competition. We optimize its implementation by reducing the input image resolution and parallelizing the original codebase.

We directly use the FoundationPose model that was pre-trained on a large-scale dataset without any fine-tuning. Experimental results demonstrate that this approach outperforms the trained methods, which is consistent with our previous analysis. The large domain gap between synthetic training data and real-world industrial environments undermines the effectiveness of supervised training paradigms of seen object pose estimation.

3.3. SAM6D-industrial

Due to the complex lighting conditions and unique material properties of objects in industrial environments, we observe significant detection failures for certain objects from specific viewpoints. As shown in Fig. 2, the instances within the red and yellow box in viewpoint (a) is hard to identify, whereas the same instances is easily detected from other viewpoints.

Thus we also explore the SAM-6D-based approach, *i.e*. SAM6D-industrial, another alternative unseen-object methodology. Compared to FoundationPose [17], SAM-6D [8] exhibits slightly inferior accuracy but demonstrates significantly lower computational and temporal demands. To mitigate the risk of object miss-detection, we implement a **multi-view fusion strategy** based on the SAM-6D pose estimation model, formalized as follows:

Given N calibrated input views, instance masks, object IDs and corresponding confidence scores ${S_d}_v^i$ are extracted via our detection module. These are then fed into the SAM-6D [8] pose estimator to generate candidate poses P_v^i with pose confidence scores ${S_p}_v^i$, where i represents the index of masks and v represents the index of views. The objective is to obtain a non-redundant, complete, and accurate set of object poses. Our fusion algorithm proceeds in three steps:

1) Score Integration: Compute unified scores from detection confidence and pose confidence $S_v^i = {S_p}_v^i * {S_d}_v^i$.
2) Spatial Alignment: Project all poses to a unified world coordinate system using camera extrinsic.
3) Sort the candidate poses in descending order based on S_v^i and iterate through them. Denote the translation vector of the k-th pose as t_k, the j-th pose is removed if: $\exists i < j, ||t_i - t_j|| < \tau_{dist}$.
Here, τ_{dist} is the threshold which determines whether two poses correspond to the same object. It can be set as a fixed constant or determined by the object's size.

4. Experiments

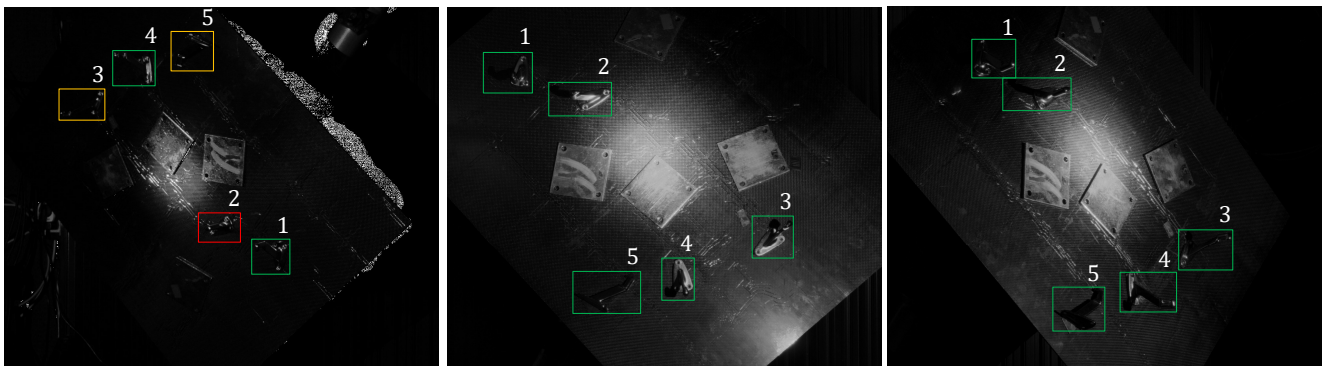
4.1. Datasets and Evaluation Metrics

Datasets The Perception Challenge for Bin Picking is based on the Industrial Plenoptic Dataset (IPD) [7], a multi-view and multimodal dataset specifically designed for high-precision industrial applications. The dataset comprises 2300 physical scenes involving 20 distinct industrial parts, within a working volume of 1m × 1m × 0.5m. Each scene is captured using 13 precisely calibrated cameras placed at different viewpoints, including RGB, depth, and polarization cameras. To simulate diverse lighting in real-world industrial scenarios, each scene is collected under four exposure settings, each with three different lighting conditions. For the single-view approach, we use only the RGB and depth images from the Photoneo camera perspective. In contrast, the multi-view approach utilizes RGB and depth images from all four different perspectives.

To ensure the generalization capabilities of the method, the competition is structured into three phases. In Phase 1, a subset of the IPD dataset is used, focusing on 10 object categories. Phase 2 involves evaluation on 10 entirely new industrial scene objects, with the test dataset not being publicly available. In Phase 3, the same objects from Phase 2 are used, but testing is conducted within a real robotic system in a true bin-picking environment.

Evaluation Metrics Submissions to the Bin-Picking Perception Challenge are evaluated using mean Average Precision (mAP) computed over a range of Maximum Symmetry-aware Surface Distance (MSSD) [5] thresholds. MSSD measures the maximum distance between corresponding surface points of an object under the predicted and ground truth poses, taking into account object symmetries. A pose prediction is considered correct if its MSSD falls below a given threshold. The thresholds range from 2mm to 20mm with a 2mm step, and the final mAP score is obtained by averaging AP over all thresholds and categories.

To avoid bias introduced by inaccurate annotated ground truth poses, BPC evaluates MSSD using Robot Consistency [7], which measures the consistency of pose predictions across multiple robot configurations. Specifically, for a sequence containing N distinct scenes, the transformation T_{CR} between camera and robot base is first obtained through hand-eye calibration. For each scene i, the



(a) (b) (c)

Figure 2. Illustration of multi-view detection results. Viewpoints (a), (b), and (c) correspond to three different views of the same scene. Each object instance is marked with a colored bounding box, and its ID is indicated at the top-right corner. Red boxes indicate instances that are not detected in the current view; yellow boxes indicate instances that are detected but without a valid pose estimation; green boxes indicate instances that are successfully detected and pose-estimated.

Method	Bounding Box	AP
w All strategies	19191	0.809
w/o Crop&Fuse	16550	0.670
w/o Size&Edge Filter	27602	0.807
w/o Individual model	22067	0.614
w/o Lighting Aug.	16044	0.768

Table 1. 2D object detection results on IPD [7].

transform T^i_{RG} from the gripper G to the robot base R is recorded as the robot arm moves to different poses. The object is rigidly mounted on the gripper, ensuring a constant transform T_{GO} from the object frame to the gripper coordinates. Given the predicted object pose ${}^{pred}T^i_{CO}$ in the camera frame for each scene, the corresponding estimated T_{GO} can be recovered as: $T_{GO} = (T^i_{RG})^{-1} T^{-1}_{CR} \{{}^{pred}T^i_{CO}\}$. According to the law of large numbers [10], the average of these N estimates yields a reliable approximation of the true transform, denoted as T^*_{GO}, which serves as a pseudo ground-truth. The final MSSD is measured between each predicted pose ${}^{pred}T^i_{CO}$ and the reprojected pseudo ground-truth pose $T_{CR} T^i_{RG} T^*_{GO}$, averaged over N scenes. In the case of multiple objects mounted on the robot, all estimated T^*_{GO} across all scenes are first clustered, with each cluster corresponding to a distinct object instance.

Since the challenge only provides a single final metric (mAP), we additionally report results obtained using the official BOP evaluation server for reference. Details about the BOP metrics can be found at https://bop.felk.cvut.cz/home.

4.2. Results

Below, we present and analyze the performance of our approach in the competition.

Tab. 1 demonstrates the effectiveness of distinct optimization strategies in our detection pipeline. The results indicate that the Crop & Fuse strategy yields a +13.9% improvement in 2D detection Average Precision (AP). Although the Size & Edge Filter strategy provides only marginal AP gains (+0.8%), it substantially reduces the number of bounding boxes by 8411, thereby accelerating downstream pose estimation. Meanwhile, the per-category training paradigm achieves a significant AP boost of +19.5%, highlighting its efficacy for fine-grained detection. Critically, the removal of lighting augmentation during training resulted in a 4.1% AP degradation, underscoring the necessity of our comprehensive data augmentation strategy.

Tab. 2 benchmarks the performance of diverse pose estimation methods on IPD [7] test dataset (Phase 1). Key observations include: a) Both unseen methods (FDP-industrial, SAM6D-industrial) significantly outperform the trained GDRNPP+ICP baseline in the most stringent metric, AP_{MSSD_mm} ($\delta > 20\%$). b) Further fine-tuning(SAM6D-industrial-S^f) the unseen method on synthetic training data not only failed to yield improvements, but actually led to a degradation in performance metrics. This clearly reflects the significant distributional discrepancy between synthetic and real-world data in industrial settings. c) While our SAM6D-based method in the single-view setting exhibits lower accuracy than FDP-industrial (BPC metric: 0.411 vs. 0.568), SAM6D-industrial leverages multi-view fusion strategy and achieves comparable precision (BPC metric: 0.568 vs.0.581) with superior computational efficiency (average time per image: 11.5 s vs. 29.8 s). Fig. 3 further provides qualitative comparison on phase 1. When objects are easily detectable in single views, both methods achieve comparable performance. However, for targets appearing

Method	AP_{MSPD}	AP_{MSSD}	AP	AP_{MSSD_mm}	BPC metric	$Latency$ (s)
GDRN+ICP	0.811	0.627	0.719	0.367	0.309	6.2
FDP-industrial	0.886	0.825	0.855	0.669	0.581	29.8
SAM6D-industrial-S	0.815	0.720	0.768	0.614	0.411	4.9
SAM6D-industrial-S^f	0.797	0.693	0.745	0.587	-	-
SAM6D-industrial	0.924	0.848	0.886	0.632	0.568	11.5

Table 2. Quantitative comparison on the IPD [7] dataset using BOP and BPC evaluation metrics. SAM6D-industrial-S denotes a single-view variant of SAM6D-industrial that excludes the multi-view fusion strategy, while SAM6D-industrial-S^f represents its further fine-tuned version on IPD training data.

Method	Phase 2 metric	Phase 3
Ours (SAM6D-industrial)	0.31	1_{st}
SEU_WYL_trained	0.32	2_{nd}
Ours (FDP-industrial)	0.54	3_{rd}

Table 3. Results for Phase 2 and Phase 3 of BPC.

less distinct in individual perspectives (column 2 and column 4), SAM6D-industrial demonstrates significant advantages in pose estimation accuracy and robustness thanks to the multi-view fusion strategy.

Due to the competition's online testing format, precise per-image pose estimation times are unavailable. Consequently, the reported timings encompass both detection and image I/O operations. Local benchmarking indicates these overheads are negligible (<0.2s) relative to pose estimation latency (>2s), thus not affecting comparative analysis.

The results of FDP-industrial, SAM6D-industrial and the 2_{nd} place team for Phase 2 and Phase 3[2], are presented in Tab. 3. Although our FDP-industrial approach demonstrates significantly superior performance in Phase 2, it ranks lower in Phase 3's physical grasping evaluation. This discrepancy arises from the fundamental differences in the evaluation criteria: Phase 2 employs stringent millimeter-level accuracy metrics, highlighting FDP-industrial's strong performance under high-precision conditions. In contrast, Phase 3 focuses on real-world performance, evaluating the successful detection, grasping, and placement of physical parts, where the pose estimation accuracy requirements were less stringent. SAM6D-industrial prioritizes higher detection recall, which results in better overall performance in the final evaluation.

5. Conclusion

Based on the analysis of our methodology and experimental results regarding object pose estimation in industrial settings, we summarize the following conclusions: (1) Domain Gap Challenge: Training-based pose estimation methods that use synthetic data tend to underperform compared to template-based unseen pose estimation approaches. This discrepancy is largely due to significant domain gaps between synthetic and real-world data in industrial scenarios, which affect the generalizability of the models. (2) Evaluation Methodology: Conventional dataset accuracy metrics fail to fully capture the performance of pose estimation systems in real-world deployment. Therefore, there is a need for more comprehensive and application-oriented evaluation frameworks that better reflect practical challenges. (3) Multi-View Advantage: The multi-view fusion strategy proves effective in compensating for missed detections from specific viewpoints, thereby improving recall rates. This approach is particularly beneficial in complex industrial environments where objects may be partially occluded or viewed from limited angles.

Limitations and Future Work Although our approaches demonstrate competitive performance in this competition, they still fall short of meeting the accuracy requirements for practical deployment in real-world industrial robotic applications. Additionally, their computational efficiency does not satisfy the low-latency demands critical for industrial environments. Several promising directions remain open for future exploration: (1) evaluating methods in real-world robotic tasks; (2) developing methods using diverse industrial datasets, such as BOP-Industrial[3]; and (3) optimizing processing speed and reducing computational overhead, potentially through algorithmic improvements or hardware acceleration, to meet the stringent demands of industrial settings. We believe these findings will contribute to the successful integration of pose estimation systems in industrial applications.

Acknowledgements

We express our gratitude to the organizers of the Perception Challenge for Bin-picking 2025 for their dedication and hard work in hosting the event. This work is jointly supported by the National Natural Science Foundation of China under Grant No. 62406169, and the China Postdoctoral Science Foundation under Grant No. 2024M761673.

[2] In the final third phase, the organizers did not provide specific success rate results, only the team rankings.

[3] https://bop.felk.cvut.cz/datasets/#BOP-Robotics

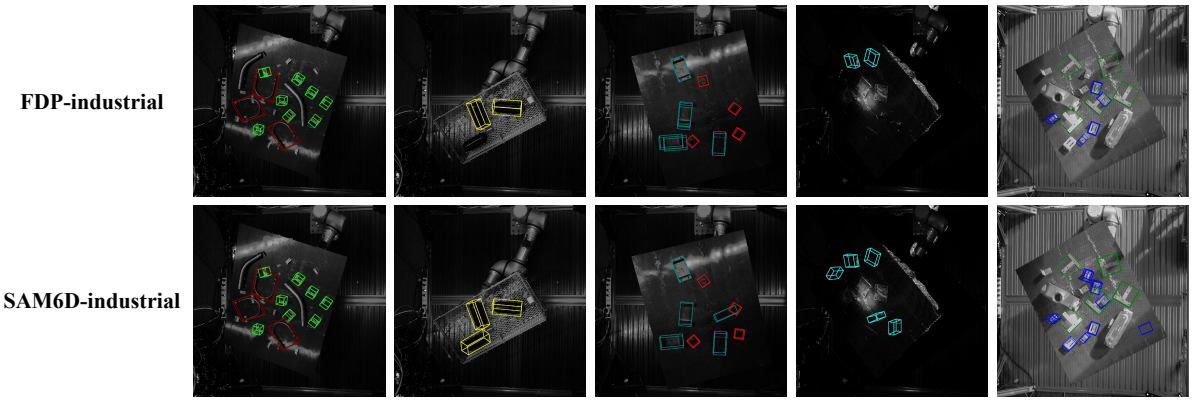


Figure 3. Qualitative comparison of pose estimation results between FDP-industrial and SAM6D-industrial on IPD [7]. 3D bounding boxes of different category are assigned with different colors.

References

[1] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: control paradigms and data structures*, pages 586–606. Spie, 1992. 2

[2] Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd-a dataset for 3d object recognition in industry. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 2200–2208, 2017. 1

[3] Zaixing He, Quanzhi Li, Xinyue Zhao, Jin Wang, Huarong Shen, Shuyou Zhang, and Jianrong Tan. Contourpose: Monocular 6-d pose estimation method for reflective textureless metal parts. *IEEE Transactions on Robotics*, 39(5): 4037–4050, 2023. 1

[4] Tomáš Hodan, Pavel Haluza, Štepán Obdržálek, Jiri Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. In *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 880–888. IEEE, 2017. 1

[5] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 577–594. Springer, 2020. 3

[6] Junwen Huang, Jizhong Liang, Jiaqi Hu, Martin Sundermeyer, Peter KT Yu, Nassir Navab, and Benjamin Busam. Xyz-ibd: High-precision bin-picking dataset for object 6d pose estimation capturing real-world industrial complexity. *arXiv preprint arXiv:2506.00599*, 2025. 1

[7] Agastya Kalra, Guy Stoppi, Dmitrii Marin, Vage Taamazyan, Aarrushi Shandilya, Rishav Agarwal, Anton Boykov, Tze Hao Chong, and Michael Stark. Towards co-evaluation of cameras hdr and algorithms for industrial-grade 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22691–22701, 2024. 1, 3, 4, 5, 6

[8] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024. 3

[9] Xingyu Liu, Ruida Zhang, Chenyangguang Zhang, Gu Wang, Jiwen Tang, Zhigang Li, and Xiangyang Ji. Gdrnpp: A geometry-guided and fully learning-based object pose estimator. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2

[10] Michel Loève and M Loeve. *Elementary probability theory*. Springer, 1977. 4

[11] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1, 2

[12] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 1

[13] Han Sun, Yizhao Wang, Zhenning Zhou, Mingyang Li, Nailong Liu, Randolph Osivue Odekhe, and Qixin Cao. Metal parts' zero-shot 6d pose estimation via foundation model and template update for industrial scenario. *IEEE Transactions on Instrumentation and Measurement*, 2025. 1

[14] Yunjie Tian, Qixiang Ye, and David Doermann. Yolov12: Attention-centric real-time object detectors. *arXiv preprint arXiv:2502.12524*, 2025. 1, 2

[15] Gu Wang, Fabian Manhardt, Federico Tombari, and Xiangyang Ji. Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16611–16621, 2021. 2

[16] Qide Wang, Daxin Liu, Zhenyu Liu, Jiatong Xu, Hui Liu, and Jianrong Tan. A geometry-enhanced 6d pose estimation network with incomplete shape recovery for industrial parts. *IEEE Transactions on Instrumentation and Measurement*, 72:1–11, 2023. 1

[17] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of

novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024. 3