

XYZ-IBD: Benchmarking Robust 6D Object Pose Estimation under Real-World Industrial Complexity

Junwen Huang^{1,2}, Jiaqi Hu¹, Peter KT Yu^{3,4}, Slobodan Ilic¹,
Martin Sundermeyer⁵ Benjamin Busam^{1,2}

¹Technical University of Munich ²Munich Center for Machine Learning (MCML)
³XYZ Robotics ⁴ROBOX ⁵Google



Fig. 1: XYZ-IBD is an industrial-grade dataset for object detection and 6D pose estimation, providing **sub-millimeter annotation precision** for reliable evaluation. Unlike household datasets, XYZ-IBD focuses on the perception bottleneck of robotic manipulation: tackling multi-instance ambiguity, randomly stacked scenarios, heavy occlusion, geometric symmetries, and extreme specular reflections. We benchmark the recent SOTA object detection and pose estimation under both object-specific and unseen-object setups, revealing a significant performance gap, challenging existing SOTA methods to bridge the divide between laboratory benchmarks and real-world factory complexity.

Abstract. While current 6D pose estimation benchmarks have reached near-saturation on household objects, they often fail to capture the stochastic and optical complexities of industrial environments. We introduce XYZ-IBD, a high-precision benchmark for object detection and 6D pose estimation specifically designed for industrial bin-picking. XYZ-IBD addresses the domain gap by providing 75 multi-view real-world scenes containing approximately 273k annotated instances of metallic, symmetrical, and specular objects. Unlike existing datasets, our benchmark features

high-density stochastic stacking and multi-instance ambiguity, reflecting authentic robotic manipulation challenges. We employ a rigorous multi-stage and semi-automatic annotation pipeline, ensuring sub-millimeter annotation accuracy. The annotations are validated through our designed error quantification scheme, securing the reliability of the annotation quality. In addition to real-world evaluation data, we provide a large-scale complementary synthetic training set that is rendered under realistic bin-picking simulation. Benchmarking state-of-the-art (SOTA) methods for 2D detection and 6D pose estimation reveals a significant performance degradation compared to standard household benchmarks, highlighting the unsolved challenges of industrial vision. XYZ-IBD establishes a new frontier for robust pose estimation in complex, high-occlusion, and reflective scenarios. The dataset and benchmark are publicly available at <https://xyz-ibd.github.io>.

1 Introduction

Robust 6D object pose estimation is a fundamental prerequisite for autonomous robotic manipulation, particularly in industrial bin-picking [31]. While state-of-the-art (SOTA) frameworks achieve near-saturation on standard household benchmarks, a critical performance gap emerges in real-world industrial deployment [33]. Household datasets [12, 18, 38] typically feature objects with rich textures, semantic cues, and minimal clutter. In contrast, industrial environments present a severe perception bottleneck: parts are often highly reflective, texture-less, and geometrically symmetric [7]. Furthermore, industrial manipulation dictates strict sub-millimeter spatial tolerance; the centimeter-level errors acceptable in household scenarios inevitably lead to catastrophic grasping failures.

Although recent datasets have introduced texture-less or industrial objects [1, 3, 6, 11, 17, 23], they frequently fail to replicate authentic factory complexity, missing high-density stochastic stacking, intense multi-instance ambiguity, and severe specular reflections. Crucially, obtaining accurate ground truth for highly reflective metallic parts is notoriously difficult. Standard depth sensors suffer from severe noise and geometric distortion on specular surfaces, causing existing industrial datasets to inherit measurement biases that compromise their reliability as rigorous evaluation benchmarks.

To bridge this divide, we introduce XYZ-IBD, a novel, high-precision RGB-D benchmark engineered specifically for industrial bin-picking. XYZ-IBD comprises texture-less, metallic industrial parts arrayed across 75 densely cluttered, multi-view real-world scenes, yielding approximately 273k 6D pose annotations. To ensure robust data modalities, we capture scenes using a multi-sensor suite (structured light, laser scanner, and stereoscopic cameras) mounted on an industrial robotic arm. To guarantee the sub-millimeter precision required for industrial-grade evaluation, we employ a rigorous semi-automatic annotation pipeline. We utilize an anti-reflection coating to suppress specularity [14, 15, 39], enabling high-fidelity multi-view depth fusion and precise point cloud reconstruction. Furthermore, we explicitly quantify our cumulative annotation error by

replicating the physical multi-view setup in a controlled simulation environment. This simulation-to-real validation confirms that our ground truth annotations achieve metrology-grade, sub-millimeter accuracy. We supplement this real-world data with a physics-rendered synthetic dataset of 45k views that fully recap the multi-instance randomness the simulation environment.

Extensive benchmarking of representative SOTA 2D detection and 6D pose estimation frameworks [20, 34, 37] on XYZ-IBD reveals a stark performance degradation. Models that excel on existing datasets struggle significantly with the visual ambiguity and optical complexity inherent to our scenes. By exposing this domain gap, XYZ-IBD provides a critical, high-precision testbed to drive the development of robust perception algorithms for industrial automation.

In summary, our key contributions are:

- We introduce a challenging real-world industrial bin-picking dataset that captures the extreme complexities of multi-instance stochastic stacking, specular reflection, and geometric ambiguity.
- We deploy a rigorous semi-automatic annotation pipeline to provide sub-millimeter precision labels suitable for industrial-grade evaluation, which is further verified through a noise-recovered quantification process, securing the reliability of our dataset as a real-world benchmark.
- We comprehensively benchmark SOTA instance-specific and generalizable frameworks for object detection and 6D pose estimation, demonstrating a significant performance gap and establishing a new frontier for robust 6D pose estimation.

2 Related Work

2.1 Household Datasets

A large number of object pose and scene depth datasets have been developed to address everyday scenarios involving household objects. Datasets such as LineMOD [10], LineMOD-Occlusion [1], YCB-V [38], HomebrewedDB [18], and TUD-L [12] are widely used in benchmarks for model-based object pose estimation [12, 33], and have driven progress on key challenges such as handling texture-less objects [10], occluded targets [1], and typical household environments [18, 38]. The HOPE dataset [32] extends this focus to robotic manipulation scenarios with varied lighting and occlusion conditions. IC-BIN [6] introduces an early bin-picking setup with randomly placed objects, but it includes only two textured objects and suffers from low annotation quality. StereoOBJ-1M [23] improves annotation precision through structure-from-motion (SfM) with checkerboards and offers a large number of RGB images, yet it lacks object diversity and does not include depth data. NOCS [35] presents the first category-level 6D pose dataset, covering six household object categories. More recent datasets such as PhoCal [36], HouseCat6D [16], Booster [29], and SCRREAM [14] focus on more complex scenes involving transparent or highly reflective objects and utilize a range of sensor modalities, including RGB, depth, and polarization

Table 1: Comparison of datasets for object pose estimation from different dimensions.

Dataset	Modalities	Number of Objects	Object Diversity	Object Diameter (mm)	Frames	Instances	Instances per Scene	Accurate Depth G.T	Occlusion	Reflection	Labeling Error (mm)
DIMO [3]	RGB-D	6	+	75~302	31.2k	100k	<10	✗	+	✗	2.7
T-LESS [11]	RGB-D	30	+++	63~152	147k	100k	<10	✗	++	✗	11.3
ITODD [7]	RGB-D	28	+++	24~270	800	5k	<10	✗	++	✗	1.8
ROBI [39]	RGB-D	7	+	24~76	8k	600k	>10	✓	+++	✓	1.8
StereOBJ-1M [23]	RGB	18	++	-	396k	1.5M	<10	✗	+	✗	2.3
IPD [17]	RGB-D+Polar	20	+++	80~240	30k	100k	<10	✗	+	✓	N/A
XYZ-IBD (Ours)	RGB-D	15	+++	54~300	22.5k	273k	22	✓	+++	✓	0.99

images. While those datasets provide high-quality depth and pose annotations, they lack typical scene properties found in industrial environments. Therefore, existing datasets featuring household objects do not fully capture the challenges inherent in industrial applications, which involve both object-level and scene-level complexity.

2.2 Industrial Datasets

In industrial applications, the working environment is quite different from the household scenario. Firstly, unlike household objects, industrial parts are usually texture-less and often symmetric and highly reflective [7, 17, 39]. Consequently, networks trained on household objects hardly generalize to industrial datasets. Secondly, the required pose accuracy in industrial robotics is usually higher than in household robotics or AR/VR applications. The robotic arm is expected to not only pick up singulated objects, but typically needs to pick objects from a filled container and place them at a target pose or assemble them. Even though bin-picking is a typical setup for industrial applications, only a few publicly available datasets target this scenario which severely hampers the usability of pose estimation pipelines in industrial practice. T-LESS dataset [11] features texture-less industrial objects with symmetries but does not present challenging lighting conditions, and the annotation quality is not mm-accurate, a requirement in many industrial applications. Only a few scenes present the complexity of real bin-picking configurations where similar objects occlude each other. ITODD [7] collects industrial parts with challenging geometry and lighting conditions but does not feature bins filled with objects. The consistently low pose estimation scores on ITODD [7] in the BOP challenge [33] also demonstrate the need for industrial bin-picking datasets. Other datasets such as DIMO [3] and ROBI [39] focus on metallic objects for bin-picking setups, but they focus on a limited number of objects whose size and shape are not representative of the diversity in real applications. The recent dataset IPD [17] leverages multiple sensors to collect data from industrial objects but presents little clutter, stacking and occlusions which simplifies the setup compared to real industrial scenarios. Table 1 compares the characteristics of current industrial datasets.

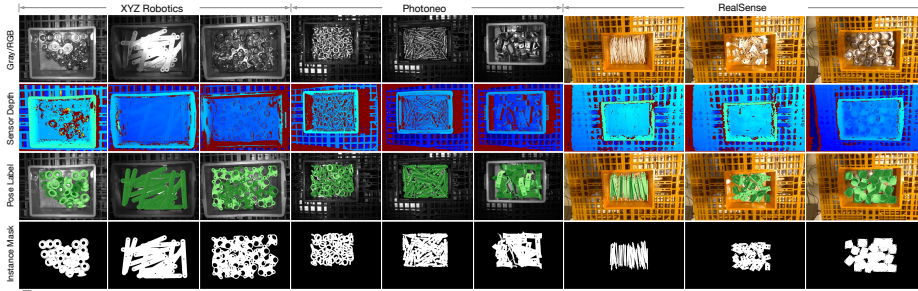


Fig. 2: Multi-sensor Data and High-Precision Annotations in XYZ-IBD. Synchronized RGB, grayscale, and depth streams acquired via a tri-sensor suite capture severe industrial perception bottlenecks. The scenes feature geometrically ambiguous, texture-less instances under complex illumination, stochastically stacked to induce high-density clutter and severe occlusion. Overlaid are our high-fidelity 6D pose and instance mask annotations, demonstrating the sub-millimeter precision required for rigorous industrial benchmarking.

3 The XYZ-IBD Dataset

As shown in Figure 2, XYZ-IBD establishes a benchmark for industrial bin-picking by capturing data under authentic factory conditions. It advances prior work through four perspectives: **(1) Industrial-Grade Setup:** Data is acquired using industry-standard robotic arms (FANUC M10iD/8L) and multi-modal sensors (RGB/depth/grayscale) mounted at industrial working distances, replicating real application conditions. **(2) Challenging Objects:** fifteen reflective, texturless, and mostly symmetric industrial parts that present rich geometrical shapes and sizes (54–300 mm scale), introducing academic challenges for pose estimation. **(3) Multi-instance Dense Clutter:** Objects are randomly and densely arranged in a container with multiple repeat instances, creating more ambiguity for instance detection and alignment. **(4) Precise Annotation:** Our annotation pipeline achieves <1 mm positional and $<1^\circ$ angular annotation accuracy, validated with simulated environment.

This benchmark consists of **75 real-world scenes** (five configurations per object), encompassing approximately **273k annotated instances** with **22 instances per scene on average**, and **up to 60 instances in some scenes**. In addition, it includes **45k synthetic training set** generated with BlenderProc [4] through physics-based object interactions, simulating a realistic bin-picking setup.

3.1 Objects and Hardware

Objects and Instance Distribution. Our dataset comprises fifteen representative industrial parts with diameters ranging from 54 mm to 300 mm, including components like sheet metal parts, bolts, pins, covers, and many other kinds of machined metal objects. As shown in Figure 3, these objects exhibit challenging visual properties such as high reflectivity and symmetry that are common in



Fig. 3: The real industrial parts.

manufacturing environments yet problematic for vision algorithms. The original CAD models provided by industrial partners ensure micron-level geometric accuracy for both real-world captures and synthetic renderings. All real-world data is collected in bins with sensor-to-object distances carefully calibrated between 600 and 1000mm. We randomly stacked multiple instances of each object into the bin, with severe occlusion and clutter. Figure 4 illustrates the distribution of instance counts across different datasets. Compared to existing industrial object datasets, ours exhibits the highest density of visible objects per image. This high density specifically addresses key challenges in object detection and pose estimation, particularly in scenarios characterized by significant visual ambiguity.

Sensor Setup. For precise and repeatable data acquisition, we employ an industrial-grade FANUC M10iD/8L robotic arm with $\pm 0.06\text{mm}$ repeatability to position our multi-sensor array. Three complementary vision systems are rigidly co-mounted on the end-effector (see Figure 5): the *Intel RealSense D415 stereoscopic camera* provides aligned RGB (1920×1080) and depth streams at 30 FPS, offering baseline color-depth registration for general scene understanding; the *XYZ Robotic AL-M DLP structured-light camera* delivers high-precision grayscale (1440×1080) and depth maps (0.08mm resolution) through projected pattern deformation analysis, particularly effective for matte surfaces; the *Photoneo PhoXi M 3D scanner* utilizes laser triangulation to generate high-accuracy depth data (up to 2064×1544 resolution) with 0.1 mm voxel precision, complemented by synchronized grayscale imagery. All sensors are positioned at optimized working distances of 600-1000 mm based on object size and bin geometry, maintaining consistent fields-of-view across the industrial container. The fixed relative positions between cameras enable direct cross-modality calibration, while the robotic arm’s precise positioning ensures reproducible viewpoint acquisition throughout the data collection process.

3.2 Data Acquisition Pipeline.

As shown in Figure 6, our data acquisition pipeline integrates three sequential stages: viewpoint sampling and calibration, multi-pass scene capture for depth

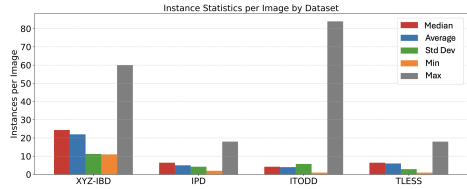


Fig. 4: Comparison of per-image instance distribution among different industrial datasets. Our dataset presents the highest number of instances per image, introducing challenges of visual ambiguity.

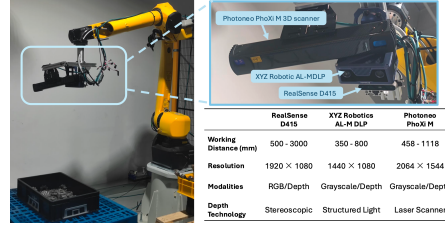


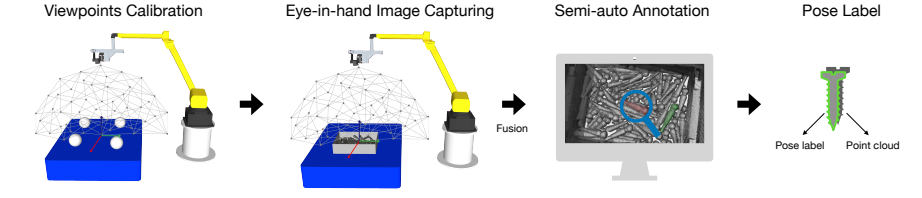
Fig. 5: The robotic setup and sensor parameters for data collection, following the real-world factory-level working layout.

ground truth, and a hybrid annotation protocol combining manual and algorithmic refinement.

Multi-view Sampling and Calibration. Beginning with the bin’s centroid as the origin, we define a spherical sampling surface spanning elevation angles of 45° to 90° to balance perspective diversity and robotic arm operability. Fifty viewpoints are randomly distributed across this surface to ensure comprehensive spatial coverage. Following the calibration framework of [39], we place four precisely machined calibration spheres on the working plane. During an initial calibration pass, the robotic arm captures multi-modal images of these spheres across all viewpoints. The cameras are firstly undistorted and obtain the initial camera poses with hand-eye calibration [2], then pose refinement via iterative closest point (ICP) alignment on the spheres’ point clouds establishes relative transformations between viewpoints with around 0.248 mm average root mean square error (RMSE), resolving the 6 DoF relationships between 49 secondary viewpoints and a primary reference view. These transformations enable subsequent multi-view data fusion and label projection with sub-millimeter consistency. After calibration, the calibration spheres are removed, and the robotic arm systematically revisits each pre-calibrated viewpoint to capture cluttered industrial bin-picking scenes. At each viewpoint, three rigidly mounted cameras (Intel RealSense D415, XYZ Robotic DLP, Photoneo PhoXi) acquire synchronized RGB, grayscale, and depth data. To maximize scene diversity, we perform five complete capture cycles per object, randomly shuffling parts between cycles.

Multi-Pass Scene Capture. To address depth sensing challenges from reflective surfaces, we employ a dual-phase capture strategy. The first phase applies a temporary anti-reflective coating (Acksys SP-102) to suppress specularity, enabling high-fidelity ground truth depth acquisition. After allowing 15 minutes for complete evaporation under controlled ambient conditions ($25^\circ\text{C} \pm 1^\circ\text{C}$), we execute an identical second capture pass to record the scene’s native optical properties. Both phases maintain pixel-wise spatial correspondence through robotic arm pose repetition (± 0.06 mm precision), providing high-quality depth to fuse

Real-world Data Acquisition



Annotation Quantification

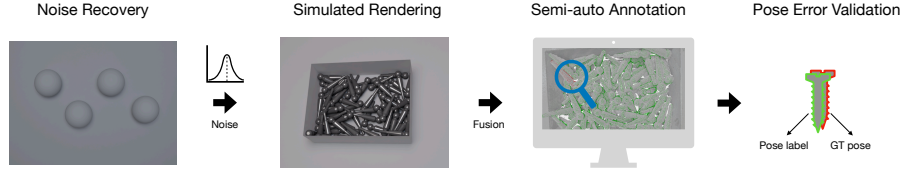


Fig. 6: The real-world industrial data collection pipeline (upper) and the annotation error quantification pipeline in the simulated environment(lower). We deploy a semi-auto pipeline to ensure a high-precision annotation protocol. The annotations are validated through our designed error quantification scheme, securing the reliability of the annotation quality.

the scene point cloud, as shown in Figure 7, thus resulting in more accurate pose annotation and also aligned datasets of enhanced and raw depth for algorithm benchmarking.

6D Pose Annotation. The annotation derives from a hierarchical process beginning with 3D fusion of spray-enhanced depth data into a unified scene point cloud. Annotators coarsely align CAD models to this reconstruction using our developed constrained GUI (± 1 mm translational, $\pm 1^\circ$ rotational increments), followed by multi-scale ICP refinement. The ICP pipeline first aligns downsampled point clouds for global adjustment and then iteratively optimizes with full-resolution data to achieve sub-millimeter accuracy. Finalized poses are propagated to all 50 viewpoints using pre-calibrated transformations that we obtained from the first stage, ensuring label consistency across perspectives without manual per-view annotation. This protocol yields around 273k annotated object instances in total, with a minimum of 11 and a maximum of 60 instances per scene, resulting in an average of 22 instances per image. We choose the XYZ Robotic DLP camera as the primary camera and perform annotations on the data it collected. The annotations of the other two cameras are projected through the calibrated relative transformation between the cameras.

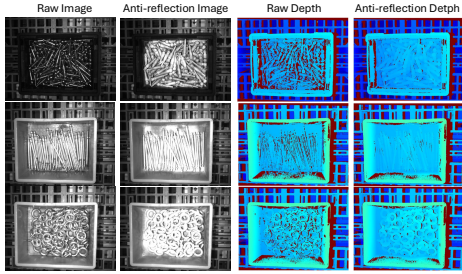


Fig. 7: Comparison between raw depth and anti-reflection depth.

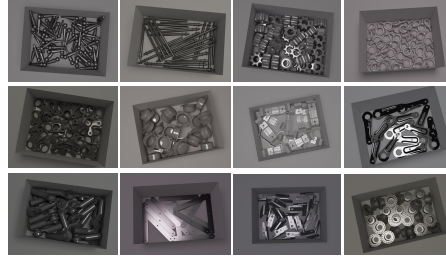


Fig. 8: The complementary synthetic training images.

3.3 Annotation Error Quantification.

As illustrated in Figure 6, to quantify the cumulative annotation error, we replicated the data collection and annotation process in a simulated environment, recovering the sensor error, calibration error and the human annotation error, and compared the resulting annotations against ground truth poses. Our evaluation framework comprises three stages: data noise recovery, simulated data collection, and 6D pose quantification.

Data Noise Recovery. We simulate the multi-view calibration procedure in a synthetic environment using identical calibration spheres and camera parameters as in the real-world setup. 50 calibration views are sampled, and varying levels of Gaussian noise are added to the rendered depth images. The corresponding RMSE of the point cloud is then computed using ICP, revealing the relationship between noise magnitude and calibration error (see Table 3). When Gaussian noise $\sigma = 0.26$ mm is applied, the computed RMSE reached 0.248 mm, matching the error observed during real-world calibration. This provided the chosen noise level to best represent the cumulative error introduced by the sensor, robotic system, and multi-view calibration process.

Simulated Data Collection. To ensure realistic, cluttered arrangements, we generate synthetic counterparts using the same CAD models within a simulated bin-picking environment rendered with physically-based rendering in Blender-Proc [4]. Objects are randomly dropped into the bin via a free-fall simulation, and any that fall outside the bin are removed. Multi-view synthetic images are rendered using a complementary noise model derived in the last step. The same annotation pipeline used for real data, incorporating multi-view fusion, manual adjustments, and multi-scale ICP refinement, is also applied to the synthetic scenes. As ground truth poses are available in the simulation, this setup allows for direct comparison between annotated and true object poses.

To complement the real dataset, we additionally render a large-scale synthetic dataset as the training data. We programmatically vary rendering conditions,

Table 2: Specifications for data collection error sources.

Source	Specify	Error (mm)
Viewpoints Calibration	Sensor Calibration	0.10
	Sensor Temporal noise	0.10
	Sensor Distortion	N/A
	Robot arm Repeatability	0.06
	Viewpoints Calibration RMSE (Total)	0.245
Depth fusion	TSDf	N/A
Manually annotate	Human, ICP	N/A
Overall		0.99

Table 3: Effect of Gaussian noise on average viewpoints calibration RMSE.

Gaussian Noise σ (mm)	RMSE (mm)
0.0	0.199
0.1	0.209
0.2	0.233
0.26	0.248
0.3	0.259

including lighting, material properties, object quantity, and pose configurations, closely replicating real-world setups to ensure cross-domain consistency. For each of the 15 objects, we perform 120 free-fall simulations, with random variations in material and lighting, resulting in a total of approximately 45,000 frames in the synthetic training dataset. As shown in Figure 8, this parallel real-synthetic collection supports robust benchmarking while preserving strong visual alignment between domains.

6D Pose Quantification. To assess annotation quality, we systematically investigated several error sources: inherent sensor inaccuracies, robotic arm repeatability, viewpoint calibration discrepancies, and annotator subjectivity. Specifically, the sensor error encompasses both the camera calibration inaccuracies and measurement noise due to sensor characteristics and environmental conditions. This, together with the robot arm repeatability, is manifested in the overall multi-view pose calibration error. We compute pose errors through nearest-neighbor matching between annotated and GT poses using Hungarian assignment on 3D centroid distances. We analyze up to 60 samples per scene $\times$ 3 scenes per object, revealing a mean positional error of 0.999 mm ($\sigma = 0.12$ mm) and an angular error of 0.432° ($\sigma = 0.08^\circ$). Per-object error averages across 15 industrial parts demonstrate sub-millimeter precision even for challenging geometries. This synthetic validation confirms that our real-world annotations achieve <1 mm positional and $<1^\circ$ angular accuracy relative to physical GT.

4 Benchmarks

4.1 Dataset Split

Synthetic Training set. As showcased in Fig. 8, the synthetic training dataset is generated using high-fidelity CAD models of the industrial components. To simulate realistic bin-picking scenarios, we employ a physics engine to model the free-fall and stochastic settling of multiple object instances range from 10 to 60. For each of the 120 generated scenes per object, we render 25 frames under diverse illumination conditions and varying material properties, utilizing camera intrinsics that match the XYZ structured-light sensor. This process yields approximately 3,000 frames per object, totaling 45k RGB-D samples. The resulting dataset provides comprehensive ground truth, including instance masks,

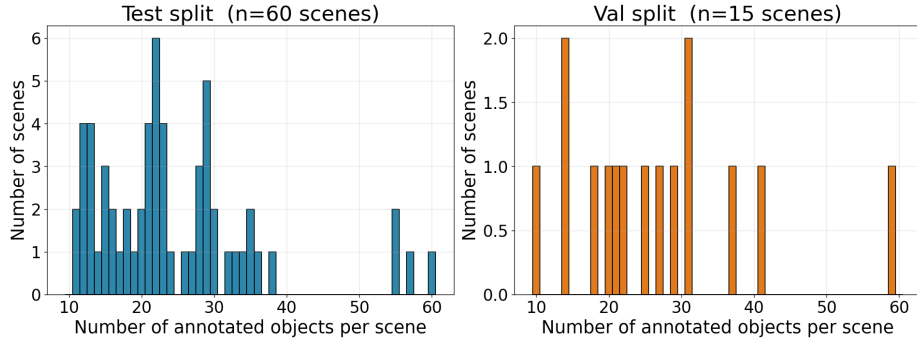


Fig. 9: The distribution of per-scene instance density in the XYZ-IBD dataset’s test and validation splits.

depth maps, and 6D poses, facilitating supervised training for depth estimation, 2D object detection, and 6D pose estimation tasks.

Real-world Validation and Testing and Set. For each industrial part, we captured five distinct scenes by systematically varying instance counts, stacking configurations, and illumination conditions with 3 above mentioned type of cameras. Among the 75 real-world scenes, 15 scenes with 2250 different frames are designated for the validation set, while the remaining 60 scenes with 9000 different frames are reserved for testing. As the per-scene instance histogram shown in Figure 9, most scenes contain **20–30** instances, tailing to **60**. Figure 4 also indicates that we have the highest mean density among existing industrial datasets. The achieved sub-millimeter annotation precision ensures reliable evaluation metrics, even within this highly challenging setup. The synthetic training dataset comprises approximately 80 GB in size. The validation and test sets occupy roughly 8 GB and 3 GB, respectively. We will publicly release the object CAD models along with the fully annotated training and validation sets. While the test set images will be made public, the corresponding ground truth will be hosted on our evaluation server to facilitate an open leaderboard for benchmarking.

4.2 Evaluation Setups

Our dataset provides high-precision 6D object pose and depth annotations, enabling the establishment of a comprehensive benchmark for **object detection and pose estimation**. In order to align the evaluation protocol with the widely used Benchmark for 6D Object Pose Estimation (BOP) and challenges, we adopt their evaluation protocols to assess performance on our dataset. For the 2D detection and 6D pose estimation tasks, we evaluate representative methods under both **seen and unseen object settings**. In the seen object setup, models are trained on our synthetic dataset and evaluated on the real-world test split. In the unseen object setup, we directly use off-the-shelf generalizable methods,

which have been pretrained on large-scale external datasets, to infer our real test scenes without extra finetuning. We benchmark several recent state-of-the-art methods across **five BOP-Core datasets** (LM-O [1], T-LESS [11], YCB-V [6], IC-BIN [6], TUD-L [12]) in compare with XYZ-IBD, under **four tasks** (model-based seen/unseen object 2D detection and model-based seen/unseen object 6D detection). All seen-object baseline methods are trained on our synthetic training dataset and evaluated on the real testing split. For all the unseen baselines, we directly use the pretrained model to infer on the testing split.

4.3 Evaluation Criteria

Object 2D Detection Metics. For the object 2D detection task, we follow the model-based 2D detection task defined in BOP 2024-2025 Challenge [33]. The objective is to generate a set of non-overlapping 2D binary instance masks with associated confidence scores from an RGB-D input image that contains multiple object instances from a given dataset. To evaluate performance, we adopt the Average Precision (AP) metric, following the protocol used in the COCO 2020 challenges [21]. AP is calculated by averaging the precision scores at several Intersection-over-Union (IoU) thresholds, ranging from 0.5 to 0.95 in increments of 0.05. Each object’s AP score reflects its detection quality across these thresholds. To obtain an overall dataset-level performance measure, the mean Average Precision (mAP) is computed by averaging the AP scores across all object categories. This evaluation strategy comprehensively captures both the accuracy of object localization and the effectiveness of category-level recognition, ensuring alignment with established benchmarking standards.

Object 6D Detection Metics. For the 6D pose estimation task, we adopt the model-based 6D object detection metric defined in BOP 2024-2025 Challenge [33], evaluating detection accuracy using symmetry-aware Average Precision (AP) scores. For each predicted pose $\hat{P}$ and its corresponding ground truth pose P_{GT} , we compute two error metrics: *Maximum Symmetry-Aware Surface Distance (MSSD)* and *Maximum Symmetry-Aware Projection Distance (MSPD)*. MSSD measures the maximum 3D surface deviation under object symmetries, defined as $e_{MSSD} = \max_{x \in M} \min_{S \in \mathcal{S}} |\hat{P}x - S(P_{GT}x)|$, where M is the object mesh and $\mathcal{S}$ is the set of predefined symmetry transformations. MSPD evaluates the maximum 2D projection deviation considering object symmetries, computed as $e_{MSPD} = \max_{u \in U} \min_{S \in \mathcal{S}} |\Pi(\hat{P}x_u) - \Pi(S(P_{GT}x_u))|$, where Π denotes the camera projection function and U the set of visible mesh vertices. A pose estimate is deemed correct when the error e falls below a threshold θ_e . For each error type $e \in \text{MSSD, MSPD}$ and object $o \in O$, we compute the object-level AP score as $AP_{e,o} = \frac{1}{|\Theta_e|} \sum_{\theta \in \Theta_e} P_o(\theta)$, where Θ_e is the set of threshold values and $P_o(\theta)$ is the precision at threshold θ . The final AP score aggregates over all objects and both error types as $AP = \frac{1}{2|O|} \sum_{o \in O} \sum_{e \in \text{MSSD, MSPD}} AP_{e,o}$.

Table 4: Performance comparison of 2D and 6D detection SOTA methods on previous core datasets in BOP and XYZ-IBD dataset. We report the AP scores on the main BOP datasets YCB-V, T-LESS, and LM-O, and the average AP on the 5 BOP core datasets (YCB-V, T-LESS, LM-O, IC-BIN and TUD-L) under the four tasks of seen object 2D detection, unseen object 2D detection, seen object 2D detection, and unseen object 2D detection.

Method	Task	Unseen Object	YCB-V	T-LESS	LM-O	BOP-Core 5	XYZ-IBD
YOLOX [8]	2D Detection	$\times$	0.877	0.707	0.894	0.798	0.774
CNOS(SAM) [26]		$\checkmark$	0.490	0.395	0.330	0.361	0.275
SAM-6D(SAM) [20]		$\checkmark$	0.518	0.465	0.437	0.449	0.296
NIDS-Net [25]		$\checkmark$	0.650	0.496	0.439	0.494	0.258
GDRN [34]	6D Detection	$\times$	0.906	0.852	0.775	0.827	0.266
SurfEmb [9]		$\times$	0.799	0.828	0.760	0.758	0.247
MatchU [13]		$\checkmark$	0.756	0.668	0.680	0.705	0.529
FoundationPose [37]		$\checkmark$	0.889	0.646	0.756	0.734	0.564
SAM6D [20]		$\checkmark$	0.845	0.515	0.699	0.704	0.578

4.4 Evaluation of Object 2D Detection

Seen Object 2D Detection. YOLOX [8] is a widely used advanced real-time object detection model that builds upon the YOLO [30] series. We follow the implementation of the SOTA method for object pose estimation GDRNet [34] to train the YOLOX model on our synthetic training data and test on the real test split.

Unseen Object 2D Detection. CNOS [26] is a model-based method that uses vision foundation models SAM [19] and DINOv2 [27] for novel object segmentation and detection without re-training. It renders object templates from a CAD model and ranks SAM-generated segments by comparing their DINOv2 class token features with those of the templates. NIDS-Net [25] uses a similar workflow but takes Grounding DINO [22] as the backbone. SAM-6D [20] detects the objects with a similar strategy as CNOS [26] but computes a weighted score including semantics, appearance, and geometry to match the query object template with the segments extracted from SAM [19].

Results. As shown in Table 4, for the 2D detection task, YOLOX [8] is trained on the training split of each dataset and tests on the test split for the same objects (seen objects), achieving comparable results across different datasets under this overfitting setup. However, for the unseen object detection methods, we directly deploy the pretrained model to inference on our dataset. CNOS and SAM6D show a clear drop compared to other BOP datasets. Specifically, SAM6D degrades by more than 33% of the AP on average. These results highlight the increased difficulty of our dataset for 2D detection, due to heavy occlusion, repeated object instances, and strong surface reflections. We show a qualitative comparison in Figure 10.

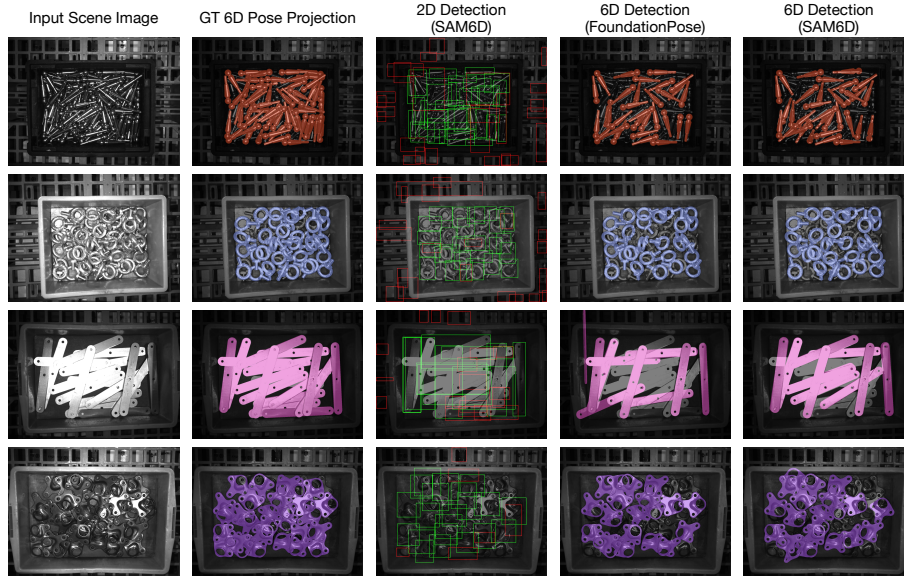


Fig. 10: Example qualitative results of SOTA methods for unseen object 2D and 6D detection tasks on XYZ-IBD dataset. In the detection results, green boxes indicate correct object IDs, while red boxes indicate mismatches with the ground truth ID.

4.5 Evaluation of Object 6D Detection

Seen Object 6D Detection. SurfEmb [9] learns per-object dense 2D–3D correspondence distributions over object surfaces using contrastive learning in an unsupervised fashion. It achieves strong performance on BOP and handles visual ambiguities effectively. GDRNet [34] is a recent state-of-the-art framework that processes zoomed-in RoIs from RGB images to predict intermediate geometric features: dense 2D-3D correspondences, surface region attention maps, and visible object masks. These features guide a Patch-PnP module to directly regress the 6D pose in a differentiable manner.

Unseen Object 6D Detection. FoundationPose [37] supports both model-based and model-free settings using neural implicit representations for view synthesis. Trained on large-scale synthetic data with transformer-based coarse-to-fine design, it generalizes well and outperforms prior methods across benchmarks. MatchU [13] extracts texture and geometric features from the input RGB-D and CAD models to fuse the descriptors for 3D registration. Similarly, SAM-6D [20] uses the Segment Anything Model for segmentation and applies ViT [5] and GeoTransformer [28] to extract features from RGB-D input and CAD models. Trained on a large-scale synthetic dataset, it achieves strong performance in model-based 6D pose estimation.

Results. Unseen object methods assume the availability of object segmentation or detection as a prior for pose estimation. Accordingly, we use segmentation masks produced by SAM-6D for fair comparison among these methods, while seen object methods utilize detection results from YOLOX. As shown in Table 4, all methods struggle on our dataset. Specifically, both GDRNet and SurfEmb, representing seen object methods, fail to predict accurate poses, despite being trained on synthetic data. In contrast, unseen object methods demonstrate relatively better performance, with SAM6D achieving state-of-the-art results. Compared to existing household datasets, [31, 33], our dataset introduces greater challenges for pose estimation due to the complexity of object materials, geometric variations, and severe scene clutter. Figure 10 shows a qualitative comparison of the baseline results on XYZ-IBD benchmark.

5 Conclusion

We introduce XYZ-IBD, a high-precision benchmark for industrial bin-picking designed to capture the complexities of real-world manufacturing, including stochastic scene configurations, multi-instance ambiguity, and specular reflectivity. The benchmark comprises a diverse set of industrial components captured under authentic factory conditions using three distinct sensors, yielding 273k annotated real-world samples. This is augmented by a 45k-frame synthetic dataset that simulates realistic bin-picking environments. Utilizing a multi-stage, semi-automatic annotation protocol, we provide high-fidelity 6D pose labels. We rigorously quantify annotation error through simulations that model sensor and calibration noise, achieving a mean pose error below 1 mm. We believe XYZ-IBD brings real-world industrial vision problems to the academic community and helps bridge the gap between academic research and practical application. Potential future work lies to expand the scale of object categories and working conditions, covering more typical but challenging vision tasks for the evaluation of recent foundational industrial solutions.

6 Acknowledgement

We thank XYZ Robotics for providing the industrial parts and environments used for dataset collection. We are especially grateful to Jizhong Liang for his valuable efforts in annotating the dataset during his internship at XYZ Robotics. We also thank Tomáš Hodaň for his help in integrating the XYZ-IBD dataset into the BOP benchmark and Médéric Fourmy for assisting with the evaluation system integration.

Appendix

A Benchmark Setup

In the test set, we follow the BOP Challenge 2025’s setup and provide both single-view and multi-view evaluation protocols. For multi-view evaluation, 5 viewpoints are selected per scene based on the maximum spatial spread. One view is designated as the target view for evaluation, while the other 4 views serve as reference views. Relative camera poses among the 5 views are provided to enable multi-view methods to exploit spatial context. For single-view methods, only the target view is used.

B Implementation Details for the Baselines

All seen-object baseline methods are trained on our synthetic training dataset and evaluated on the real testing split. For all the unseen baselines, we directly use the pretrained model to infer on the testing split.

B.1 YOLOX

We train a YOLOX [8] model for object detection following the configuration used in GDRNPP [24]. Training is performed on a single NVIDIA RTX 4090 GPU with a batch size of 24 for 30 epochs. Data augmentation is applied during the first 15 epochs, consistent with the GDRNPP setup. The complete training process on the synthetic PBR dataset takes approximately 18 hours.

B.2 GDRNet

GDRNet [34] is trained on all the objects of our synthetic dataset with a batch size of 24 for 10 epochs, totaling roughly 490,000 training steps. The training is conducted on a single NVIDIA RTX 4090 GPU and completes in approximately 24 hours.

B.3 SurfEmb

SurfEmb [9] is trained on each object of our synthetic dataset with a batch size of 24 for 500,000 steps. Training is performed on a single NVIDIA RTX 4090 GPU and takes approximately 20 hours.

C More Visualizations of Data Samples

We visualize more data samples in this section. Figure 11 shows more examples for the scenes that were recorded with the XYZ camera, and Figure 12 shows more examples from the RealSense camera. The dataset follows the BOP dataset format as shown in Figure 13.

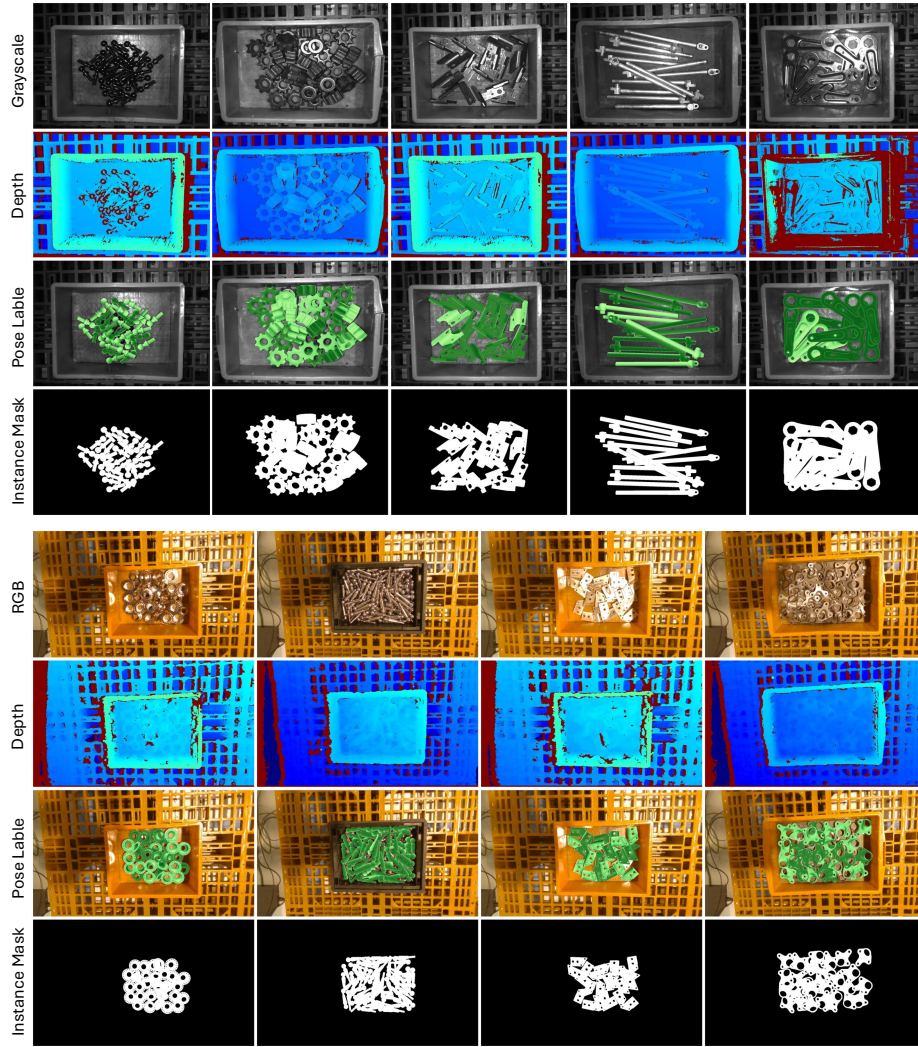


Fig. 12: More data samples from the RealSense camera.

References

1. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6d object pose estimation using 3d object coordinates. In: ECCV (2014)
2. Daniilidis, K.: Hand-eye calibration using dual quaternions. *The International Journal of Robotics Research* **18**(3), 286–298 (1999)
3. De Roovere, P., Moonen, S., Michiels, N., Wyffels, F.: Dataset of industrial metal objects. *arXiv preprint arXiv:2208.04052* (2022)
4. Denninger, M., Sundermeyer, M., Winkelbauer, D., Olefir, D., Hodan, T., Zidan, Y., Elbadrawy, M., Knauer, M., Katam, H., Lodhi, A.: Blenderproc: Reducing the

- reality gap with photorealistic rendering. In: International Conference on Robotics: Science and Systems(RSS) (2020)
5. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
 6. Dumanoglou, A., Kouskouridas, R., Malassiotis, S., Kim, T.K.: Recovering 6d object pose and predicting next-best-view in the crowd. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3583–3592 (2016)
 7. Drost, B., Ulrich, M., Bergmann, P., Hartinger, P., Steger, C.: Introducing mvtec itodd - a dataset for 3d object recognition in industry. In: ICCVW (2017)
 8. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: YOLOX: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
 9. Haugaard, R.L., Buch, A.G.: Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings. In: CVPR (2022)
 10. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Asian conference on computer vision. pp. 548–562. Springer (2012)
 11. Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X.: T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. WACV (2017)
 12. Hodan, T., Michel, F., Brachmann, E., Kehl, W., GlentBuch, A., Kraft, D., Drost, B., Vidal, J., Ihrke, S., Zabulis, X., Sahin, C., Manhardt, F., Tombari, F., Kim, T.K., Matas, J., Rother, C.: Bop: Benchmark for 6d object pose estimation. In: ECCV (2018)
 13. Huang, J., Yu, H., Yu, K.T., Navab, N., Ilic, S., Busam, B.: Matchu: Matching unseen objects for 6d pose estimation from rgb-d images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10095–10105 (2024)
 14. Jung, H., Li, W., Wu, S.C., Bittner, W., Brasch, N., Song, J., Pérez-Pellitero, E., Zhang, Z., Moreau, A., Navab, N., Busam, B.: SCRREAM : SCan, register, REnd and map: A framework for annotating accurate and dense 3d indoor scenes with a benchmark. In: The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2024), <https://openreview.net/forum?id=NHob4eMg7R>
 15. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., et al.: On the importance of accurate geometry data for dense 3d vision tasks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 780–791 (2023)
 16. Jung, H., Wu, S.C., Ruhkamp, P., Schieber, H., Wang, P., Rizzoli, G., Zhao, H., Meier, S.D., Roth, D., Navab, N., et al.: Housecat6d—a large-scale multi-modal category level 6d object pose dataset with household objects in realistic scenarios. arXiv preprint arXiv:2212.10428 (2022)
 17. Kalra, A., Stoppi, G., Marin, D., Taamazyan, V., Shandilya, A., Agarwal, R., Boykov, A., Chong, T.H., Stark, M.: 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 (CVPR). pp. 22691–22701 (June 2024)
 18. Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. ICCVW (2019)

19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
20. Lin, J., Liu, L., Lu, D., Jia, K.: 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. pp. 27906–27916 (2024)
21. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer vision—ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part v 13. pp. 740–755. Springer (2014)
22. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)
23. Liu, X., Iwase, S., Kitani, K.M.: Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10870–10879 (2021)
24. Liu, X., Zhang, R., Zhang, C., Wang, G., Tang, J., Li, Z., Ji, X.: Gdrnpp: A geometry-guided and fully learning-based object pose estimator. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
25. Lu, Y., Guo, Y., Ruozzi, N., Xiang, Y., et al.: Adapting pre-trained vision models for novel instance detection and segmentation. In: 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13341–13348. IEEE (2025)
26. Nguyen, V.N., Groueix, T., Ponimatin, G., Lepetit, V., Hodan, T.: Cnos: A strong baseline for cad-based novel object segmentation. In: ICCV (2023)
27. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
28. Qin, Z., Yu, H., Wang, C., Guo, Y., Peng, Y., Ilic, S., Hu, D., Xu, K.: Geotransformer: Fast and robust point cloud registration with geometric transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(8), 9806–9821 (2023)
29. Ramirez, P.Z., Costanzino, A., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L.: Booster: a benchmark for depth from images of specular and transparent surfaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(1), 85–102 (2023)
30. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 779–788 (2016)
31. Sundermeyer, M., Hodaň, T., Labbe, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J.: Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In: CVPR (2023)
32. Tyree, S., Tremblay, J., To, T., Cheng, J., Mosier, T., Smith, J., Birchfield, S.: 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 13081–13088. IEEE (2022)
33. Van Nguyen, N., Tyree, S., Guo, A., Fourmy, M., Gouda, A., Lee, T., Moon, S., Son, H., Ranftl, L., Tremblay, J., et al.: Bop challenge 2024 on model-based and model-free 6d object pose estimation. arXiv e-prints pp. arXiv-2504 (2025)
34. Wang, G., Manhardt, F., Tombari, F., Ji, X.: Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: CVPR (2021)

35. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2642–2651 (2019)
36. Wang, P., Jung, H., Li, Y., Shen, S., Srikanth, R.P., Garattoni, L., Meier, S., Navab, N., Busam, B.: Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 21222–21231 (2022)
37. Wen, B., Yang, W., Kautz, J., Birchfield, S.: Foundationpose: Unified 6d pose estimation and tracking of novel objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17868–17879 (2024)
38. Xiang, Y., Schmidt, T., Narayanan, V., , D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems* (2018)
39. Yang, J., Gao, Y., Li, D., Waslander, S.L.: Robi: A multi-view dataset for reflective objects in robotic bin-picking. In: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9788–9795. IEEE (2021)



Fig. 13: Directory structure of the dataset.