2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

IndustryShapes: An RGB-D Benchmark dataset for 6D object pose estimation of industrial assembly components and tools

Panagiotis Sapoutzoglou, Orestis Vaggelis, Athina Zacharia, Evangelos Sartinas, Maria Pateraki

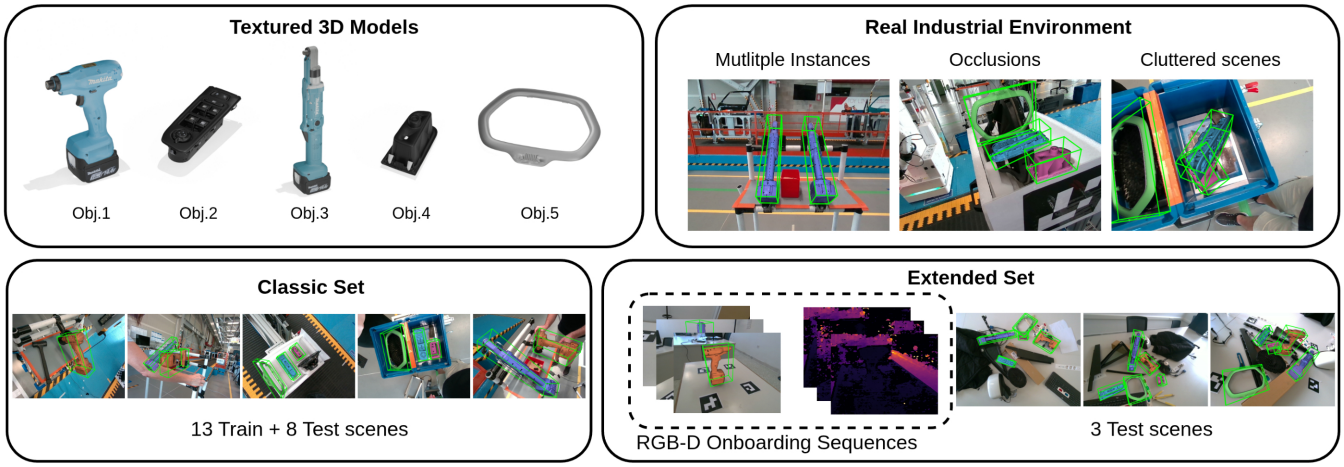

Fig. 1: Overview of the IndustryShapes dataset. Five industrial tools/components with challenging properties (weak/absent texture, symmetries, thin and reflective parts) are captured in realistic industrial environment scenes. Data are organized into two complementary sets: *Classic* set for instance-level evaluation and *Extended* set tailored for novel-object methods, including RGB-D static onboarding sequences.

***Abstract*— We introduce *IndustryShapes*, a new RGB-D benchmark dataset of industrial tools and components, designed for both instance-level and novel object 6D pose estimation approaches. The dataset provides a realistic and application-relevant testbed for benchmarking these methods in the context of industrial robotics bridging the gap between lab-based research and deployment in real-world manufacturing scenarios. Unlike many previous datasets that focus on household or consumer products or use synthetic, clean tabletop datasets, or objects captured solely in controlled lab environments, *IndustryShapes* introduces five new object types with challenging properties, also captured in realistic industrial assembly settings. The dataset has diverse complexity, from simple to more challenging scenes, with single and multiple objects, including scenes with multiple instances of the same object and it is organized in two parts: the *classic set* and the *extended set*. The *classic set* includes a total of 4,6k images and 6k annotated poses. The *extended set* introduces additional data modalities to support the evaluation of model-free and sequence-based approaches. To the best of our knowledge, *IndustryShapes* is the first dataset to offer RGB-D static onboarding sequences. We further evaluate the dataset on a representative set of state-of-the art methods for instance-based and novel object 6D pose estimation, including also object detection, segmentation, showing that there is room for improvement in this domain. The dataset page can be found in https://pose-lab.github.io/IndustryShapes.**

This work was funded by the HEU programme SOPRANO (GA No 101120990) and PANDORA (GA No 101135775). Special thanks to Stellantis—Centro Ricerche FIAT (CRF) for supporting the dataset collection.

All authors are with the National Technical University of Athens (NTUA), Greece (e-mail:{psapoutzoglou, orestisvaggelis, azacharia, vsartinas, mpateraki}@mail.ntua.gr)

I. INTRODUCTION

6D pose estimation is a cornerstone of robotic perception, enabling robots to accurately estimate the position and orientation of objects relative to their camera(s). It is essential for robust manipulation in tasks such as pick and place and precise assembly/ disassembly and constitutes a foundational capability for perception-driven robotic systems in manufacturing industries. As modern factories evolve toward flexible automation, robust 6D pose estimation becomes increasingly important not only for repetitive tasks in robotic workcells but also for dynamic, collaborative environments. Despite its importance, 6D pose estimation in industrial environments remains challenging due to object and scene characteristics [1]. Many tools and components are textureless, metallic, or reflective, while symmetric, thin and geometrically similar objects introduce ambiguities in pose estimation. Moreover, occlusions, clutter, variable and often unconstrained lighting conditions further degrade performance. Deep learning methods have advanced the field. However, the effectiveness of these methods depends on the availability of diverse, and task-relevant data. Without realistic datasets, deep learning models often fail to generalize beyond clean, controlled environments. This limitation is especially pronounced in the industrial domain, where object properties and scene dynamics are particularly demanding. As emphasized in recent reviews [1] if the datasets' complexity does not reflect the relevant real-world complexity of robotic scenarios, this

will lead to a performance discrepancy at deployment. On the other hand an exhaustive annotation of all possible object configurations in complex scenes is impractical. While existing datasets like T-LESS [2], ITODD [3], or YCB-V [4] have contributed significantly to the field, they are often limited to controlled laboratory conditions or constrained domains such as bin-picking and household objects. These datasets are typically optimized for instance-level methods where object CAD models are known during training and broader support for novel object 6D pose estimation methods is underrepresented, including also modalities like onboarding sequences required for some of these methods.

IndustryShapes addresses this critical gap by introducing an RGB-D dataset designed specifically for 6D pose estimation in industrial environments, forming a benchmark for evaluating developed methods under real-world conditions and bridging the gap between lab-based research and deployment in real-world manufacturing scenarios. It provides rich annotations and challenging object characteristics but also data formats that directly support modern pipelines, including model-based, model-free, and sequence-driven approaches. The key contributions of the IndustryShapes dataset include:

- Challenging object set in realistic industrial conditions: The dataset provides five industrial objects characterized by complex geometries, reflective surface, symmetries, and lack of texture, with the majority captured under realistic industrial conditions with varied complexity, from single-object setups to cluttered scenes involving multiple object instances and occlusions.
- Two complementary sets: the *classic set*, designed primarily for instance-level methods, and the *extended set*, tailored specifically for novel object pose estimation. Altogether, the dataset encompasses over 26k annotations. A unique feature of IndustryShapes is the inclusion of RGB-D static onboarding sequences, making it the first dataset explicitly developed to support model-free pose estimation methods.
- Benchmarking Framework: We evaluate IndustryShapes using representative state-of-the-art baselines for instance-level methods (e.g., EPOS [5], DOPE [6], ZebraPose [7]) and latest novel object methods (e.g., FoundPose [8], FoundationPose [9]), as well as state-of-the-art object detection and segmentation methods (e.g., CNOS [10], SAM-6D [11]). Our results demonstrate scope for advancement in current methodologies, reinforcing the dataset's relevance as a robust benchmark for ongoing research.

II. Related work

A. Datasets

Diverse datasets have been developed and exploited in the past years to support research and benchmarking in 6D pose estimation across various application domains. The BOP-Classic-Core corpus, is a well-established collection that has served as the foundation for the BOP (Benchmark for 6D Object Pose Estimation) challenges since 2019 and comprises of seven datasets: LM-O [12], T-LESS [2], ITODD [3], HB [13], YCB-V [4], IC-BIN [14], and TUD-L [15]. Complementing the core datasets, the BOP-Classic-Extra group includes datasets such as LM [16], HOPEv1 [17], RU-APC [18], IC-MI [19], and TYO-L [15]. These datasets extend the benchmark's diversity by covering a broader range of object types and acquisition setups, primarily consumer and household objects and sequences designed to evaluate robustness to varying illumination, collected in both realistic and controlled laboratory environments. The BOP-Industrial group expands the benchmark with datasets specifically tailored for industrial use such as bin picking and automated inspection. These include XYZ-IBD [20], which features cluttered bin-picking scenes with occlusions and industrial parts with reflective and low-textured surfaces; ITODD-MV [3], which extends ITODD scenes with additional images in a multiview setup; and IPD [21], which captures RGB, high-resolution depth and polarization images of industrial parts in realistic setups like trays and conveyors. These datasets capture key challenges encountered in industrial environments, such as weak textures, symmetric geometries, and lighting variability, though the focus is primarily on bin picking scenarios.

In addition to the above instance-specific datasets, recent efforts within the BOP benchmark have begun addressing the more challenging task of 6D pose estimation of unseen or novel objects, where systems must generalize beyond a predefined set of models known during training. Notable datasets supporting this paradigm include T-LESS, originally part of BOP-Classic-Core, which is often used in this context due to its inclusion of multiple similar-looking, texture-less industrial parts that test a model's generalization capabilities. The HOPE dataset [17] by NVIDIA includes 50 scenes of 28 toy grocery objects captured in household/office environments. The HANDAL dataset [22] - as released in BOP- includes 40 objects from 7 categories of hardware tools and kitchen utensils. The HOT3D dataset [23] stands out from others by focusing on egocentric, real-world hand-object interactions, capturing synchronized RGB and grayscale imagery from head-mounted devices. Outside the BOP benchmark, the Objectron dataset [24] provides annotated RGB-only video sequences of everyday objects captured from mobile devices in real-world settings, totaling 4M images, for category-level 3D object detection for AR applications. Moreover, the PACE dataset [25] includes 238 real-world objects, designed to advance the development and evaluation of pose estimation methods in cluttered scenarios.

Overall the majority of datasets target household objects and consumer goods (YCB-V, TUD-L, IC-MI, RU-APC, TYO-L, HOPE, HANDAL) supporting applications in domestic robotics and AR. Datasets like LM-O, IC-BIN, TUD-L, PACE and parts of HB adopt controlled laboratory setups, where lighting, background, and object arrangements are optimized for repeatable benchmarking under simplified or constrained conditions. Datasets that focus on industrial-relevant objects include T-LESS, ITODD, ITODD-MV, XYZ-IBD,

IPD, and to a lesser extent HB, offering representative challenges such as weak texture, symmetries, and clutter. Notably, XYZ-IBD, IPD, and ITODD(-MV) are specifically designed to support bin-picking scenarios, featuring scenes with multiple instances and occlusions. Table I provides an overview of the datasets.

Unlike lab-captured datasets or those focused solely on bin-picking scenarios, IndustryShapes prioritizes industrial realism and challenging properties over object quantity, offering more general industrial assembly scenes featuring objects placed on holders, boxes, trolleys or in mixed-use environments that closely resemble real robotic workstations and human-robot collaboration settings. It introduces new object types with challenging properties, while also supporting research on novel object pose estimation by providing not only still images, but also RGB-D image sequences with close-range depth data and multiple objects per scene. Although IndustryShapes includes a limited number of objects, this design choice is intentional, aiming to concentrate multiple, industrially relevant sources of difficulty within a compact benchmark. As reflected in our benchmark results (Sec. IV-C), even recent instance-level and novel-object approaches experience substantial performance drops, indicating that IndustryShapes exposes fundamental limitations that may remain hidden when evaluating on larger but less demanding datasets.

B. Methods

6D object pose estimation has witnessed rapid progress in the last years, especially with the advent of deep learning. In contrast to the surveys of [26], [27] that primarily focused on instance-level methods, Liu et al. [28] systematically covers instance-level, category-level, and unseen object pose estimation and provides a multidimensional analysis of methods considering input modalities (RGB, RGBD, depth), object properties (textureless, symmetric, transparent, occluded), inference mode stages (e.g. segmentation, correspondence prediction, template matching, pose regression, pose refinement). It also reviews evaluation criteria according to the pose DoF (3DoF, 6DoF, 9DoF).

Instance-level 6D object pose estimation. Here, we provide a brief overview of relevant recent deep learning-based methods, categorizing them into *correspondence-based* and *direct pose regression* methods, highlighting representative examples in each. For a broader and more detailed analysis of the full spectrum of 6D pose estimation research, the reader is referred to recent dedicated surveys.

Correspondence-based methods predict 2D–3D associations between the image pixels and points on the object's 3D model, typically followed by a Perspective-n-Point (PnP) solver to recover the pose. Keypoint-based methods, a specific case of these approaches, focus on predicting the 2D image projections of a small set of predefined 3D keypoints, as done in DOPE [6]. PVNet [29] introduced pixel-wise voting mechanisms for keypoint localization aiming to increase the number and quality of point correspondences under partial occlusions. Dense correspondence-based approaches generalize this idea by predicting either visible 3D object coordinates for each pixel, as in Pix2Pose [30], or multiple correspondence hypotheses per pixel, as in EPOS [5]. ZebraPose [7] further introduces a coarse-to-fine hierarchical surface encoding.

Direct pose regression methods attempt to predict the 6D pose parameters directly from an input image. Early works such as PoseCNN [4] regressed translation vectors and discretized rotation angles into classification bins, while SSD-6D [31] integrated viewpoint classification into a detection framework. Methods like GDR-Net [32] incorporate geometric guidance into the regression process, using dense feature representations to improve robustness and accuracy. Despite their simplicity, direct regression methods historically struggled with rotation representation, instabilities and ambiguities, caused by symmetries and occlusions, often leading to reduced accuracy compared to correspondence-based approaches [28].

Novel object 6D pose estimation has emerged only recently as a distinct and challenging subfield and enables pose prediction for entirely new objects without retraining and can be categorized as either model-based or model-free. *Model-based methods* [8], [11], [9], [33], [34], [35], [36], [37] typically use a 3D CAD model of the object at test time, rendering it from multiple viewpoints to generate templates that represent different poses. These templates are then compared to the observed scene to identify the best alignment with the object's viewpoint, using feature descriptors, image similarity, or 3D point clouds in the case of RGB-D or depth input. In contrast, *model-free methods* [38], [39], [40], [41] do not rely on explicit 3D representations but instead operate on reference images or videos of the object. They often employ feature matching, visual localization techniques to build an internal object representation. Model-free pipelines typically run in two phases: onboarding and inference. During onboarding, the system observes a short sequence of the novel object, the so-called onboarding sequence, and constructs a representation using methods such as SfM, or neural rendering. In the inference phase, this learned representation is used to estimate the object's 6D pose from a single frame.

Object Detection and Segmentation. Recent advancements in foundation models have significantly influenced novel object segmentation. Many pipelines for pose estimation now rely on a preliminary detection or segmentation stage. Methods like CNOS [10] leverage the Segment Anything Model (SAM) [42] to generate numerous segmentation proposals from an RGB image. These proposals are then matched against pre-rendered templates of a given CAD model by comparing feature descriptors extracted by DINOv2 [43]. This approach allows for training-free segmentation of novel objects specified by their CAD models. Similarly, SAM-6D adapts this approach for RGB-D data, where its Instance Segmentation Model (ISM) uses SAM to generate initial object proposals. However, it introduces a more detailed object matching score that evaluates proposals based on a combination of semantics, appearance, and geo-

TABLE I: Overview of datasets for 6D object pose estimation

Dataset	Object category	Input	Occl.	Textureless objects	Mult. objects	Mult. instance	Onboard. videos	Application domain
LM-O [12]	8 household, workshop tools	RGB-D	✓	✓	✓	✗	✗	general rob., AR
T-LESS [2]	30 industry-relevant	RGB-D	✓	✓	✓	✓	✗	industrial rob.
ITODD-MV [3]	28 industry-relevant	Gray-D	✓	✓	✓	✓	✗	industrial rob. (bin picking)
HB [13]	33 objects	RGB-D	✓	✓	✓	✗	✗	AR, general rob.
YCB-V [4]	21 household objects	RGB-D	✓	✗	✓	✓	✗	general rob., AR
ICBIN [14]	2 objects	RGB-D	✓	✓	✓	✓	✗	robotics (bin picking)
TUD-L [15]	3 household objects	RGB-D	✗	✓	✗	✗	✗	general rob. - light
IC-MI [19]	6 household objects	RGB-D	✓	✓	✓	✓	✗	general rob., AR
RU-APC [18]	14 consumer products	RGB-D	✓	✗	✓	✓	✗	robotics in warehouse
TYO-L [15]	21 household objects	RGB-D	✗	✗	✗	✗	✗	general rob. - light
IPD [21]	22 industrial parts	RGB-D & polar.	✓	✓	✓	✓	✗	industrial rob. (bin picking) - light
XYZ-IBD [20]	17 industrial parts	Gray-D	✓	✓	✓	✓	✗	industrial rob. (bin picking) - light
HOPE [17]	28 toy grocery objects	RGB-D	✓	✗	✓	✓	✓	general rob., AR
HANDAL [22]	40 hardware tools, utensils	RGB-D	✓	✓	✓	✗	✓	general rob. (part. industrial)
HOT3D [23]	33 household objects	RGB	✓	✗	✓	✓	✓	AR, egocentric hand-object inter.
Objectron [24]	9 categories (mult)	RGB	✗	✗	✗	✗	✗	AR, 3D object detection
PACE [25]	238 objects	RGB-D	✓	✓	✓	✗	✗	general rob., AR
IndustryShapes	5 industrial objects	RGB-D	✓	✓	✓	✓	✓	industrial rob.

metric properties to identify valid instances of novel objects in cluttered scenes.

III. The IndustryShapes dataset

A. Objects and 3D models

The dataset provides high-quality 3D CAD models (Fig. 1) of electric screwdrivers (Obj. 1 and 3) and car components (Obj. 2, 4, and 5) used in robotic manipulation tasks like pick-and-place operations. Furthermore, these models include photorealistic textures allowing for the generation of synthetic data in complex scenes with distractor objects using rendering tools like BlenderProc [44]. The objects were deliberately selected to exhibit a range of challenging characteristics that are underrepresented in existing benchmarks: lack of texture, complex geometries, reflective surfaces, thin structures, and symmetries. This focus on challenging, domain-relevant properties provides a more demanding test for modern algorithms than many larger-scale datasets featuring simpler household objects, a fact reflected in our benchmark performance analysis in IV-C.

B. Dataset structure and Composition

The dataset comprises of two sets: the *classic set* and the *extended set*. The *classic set* was originally created to support instance-level pose estimation methods and includes a variety of real and lab captured scenes. The real scenes were primarily captured in a realistic industrial assembly setting under realistic lighting conditions and feature varying scene complexities, from single-object setups to cluttered scenes with multiple objects, occurring instances of the same object, and occlusions (Fig. 2, top rows). These were complemented with images of single objects acquired in laboratory conditions using a turn-table setup and systematically sampling views of the object using a fixed camera configuration [45]. Data in both industrial and laboratory settings were captured using the Intel RealSense D455 RGB-D camera at a resolution of 640 x 480, selected to balance spatial detail with real-time processing demands in robotic environments, while satisfying the minimum depth distance of the sensor, i.e. 0.52 m. In addition, synthetic images were generated using an OpenGL-based rendering pipeline [46], exploiting the photorealistic object texture to produce RGB-D data.

Classic set. The *classic set* includes a total of 21 scenes, with 13 scenes used for training and 8 scenes for test. The training data include 1217 images from 4 scenes captured in laboratory conditions and 1122 images from 8 scenes captured from the industrial assembly setting. All scenes with the exception of one from this setting feature a single object per scene. In addition to these, 1361 synthetically rendered images for Object 3, are included in the training data, totaling 3,7k frames. The test set consists of 923 images from challenging scenes in the real industrial environment, featuring occlusions, multiple object instances and diverse spatial configurations. In these scenes, objects may appear in different physical locations or in different configurations with other objects, in boxes, trays or on tool holders. The rationale for using primarily single-object scenes for training is that it enables accurate supervision for learning accurate object-specific 6D pose representations without interference

from background clutter, while testing on challenging world scenes enables robust evaluation of a model's generalization ability under realistic deployment conditions. Furthermore, this design lifts practical constraints as capturing and annotating all possible object configurations and placements in real industrial environments is time-consuming and impractical. In total, the *classic* set contains 4623 images and approximately 6k annotated poses (Table II). The annotations per object and the number of train and test scenes depicting each object are listed in Table III. Scenes feature object-to-camera distances from 200 mm to 1000 mm, with the majority falling within the 400 to 800 mm range (Fig. 4). This range reflects typical distances used in assembly lines to enable robotic manipulation, while adhering to safety regulations that minimize the risk of collisions with the surrounding environment. Furthermore, specific objects had characteristic placements and orientations within the workstation setups, or were observed from specific viewpoints, as the case of objects placed in boxes, which introduced visibility constraints.

TABLE II: Dataset groups

Group	**Annot.**	**Scenes**
Classic Train	3,9k	13
Classic Test	1,9k	8
Onboarding	6,3k	10
Extended Test	10.3k	3
Total	22.4k	34

TABLE III: Annotations per object & number of scenes per object. TR: Train, TE: Test, ON: Onboarding

Object	**Annot.**	**Classic Sc.**		**Extended Sc.**	
		TR	TE	ON	TE
Obj. 1	4,6k	3	2	2	3
Obj. 2	4.3k	5	5	2	3
Obj. 3	6k	2	2	2	3
Obj. 4	3.9k	2	4	2	3
Obj. 5	3.6k	2	5	2	3
Total	22.4k	13	8	10	3

Extended set.The *extended set* was introduced to support the benchmarking for novel object pose estimation methods, both model-based and model-free methods, as they gained momentum. The set includes RGB-D static onboarding sequences, two per object, with different placements of the objects on a table. One where the object is resting naturally upright on the table and another where the object is laid flat or positioned on its opposite side or rotated to expose different parts of the object. The set also includes three test scenes featuring all five objects in an office environment with unconstrained lighting, various distractors, occlusions, acquired from diverse viewpoints (Fig. 2, bottom row). Combined with the objects' challenging characteristics, these factors make the pose estimation task particularly demanding. The sequences of the set were captured, at a resolution of 640×480 pixels, using a hand-held Intel RealSense D405 RGB-D camera, offering close-range depth information, necessary for the onboarding sequences. It is important to note

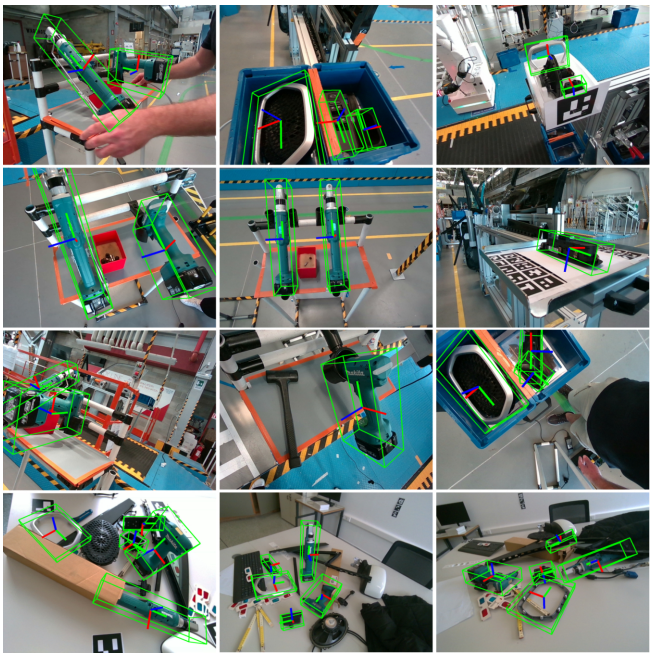

Fig. 2: Sample images with annotated object poses from the *classic set* in the industrial environment (rows 1-3 from top) and the *extended set* in the office environment (bottom row).

that, to the best of our knowledge, IndustryShapes is the first dataset that provides RGB-D static onboarding sequences. The 10 onboarding sequences (2 per object) of the set, provide in average 650 frames each and a total of 6.3k RGB-D frames, while the 3 test scenes include over 2k images and approximately 10.3k annotated object instances. Together with the onboarding sequences, they significantly enrich the dataset, providing dense camera orientations (Fig. 3) and well distributed object-to-camera distances (Fig. 4).

All data include RGB images, depth maps, annotated 6D object poses, segmentation masks, and are formatted according to the BOP specifications. The annotation procedure is detailed in Sec. III-C.

C. 6D pose annotation of real image sequences

Given the diverse types of data in the dataset , different approaches were employed regarding the annotation process. For the lab-captured data, the turn-table setup enabled the use of a marker-based approach, where camera poses were estimated from ArUco markers [45]. Similarly, in the industrial environment, the marker-based method was applied wherever marker installation was feasible. However, in cases where marker placement was not possible, a semi-automatic annotation pipeline was developed using the Structure-from-Motion (SfM) software [47]. In this setup, SfM reconstruction was combined with manually defined anchor points corresponding to known 3D coordinates from the CAD model of the object. These anchor points enabled the establishment of 2D–3D correspondences between the object in the images and its 3D model. The software automatically tracked the 2D marker locations across frames using the computed camera poses. These 2D points, together with the corresponding 3D coordinates, were then used to

solve the PnP problem, yielding the object's pose relative to the camera. While this method requires less manual effort and supervision, its accuracy is comparable to fully manual pose annotation. Considerable effort was devoted in ensuring annotation quality, as inaccurate ground truth poses can result in erroneous evaluations. The accuracy of the annotations on real images was evaluated by comparing captured depth data (d_c) with rendered depth images (d_r) from the ground truth poses. After filtering for valid pixels and excluding outliers, the absolute depth error between the annotated poses and the captured depth was found to be less than 12 mm for the classic set and approximately 5 mm for the extended set. This represents a relative error of less than 5% when compared to the objects mean diameter of 254 mm.

TABLE IV: Depth difference statistics of the $\delta = d_c - d_r$ distribution (in mm)

	μ_δ	σ_δ	$\mu_{\lvert\delta\rvert}$	$med_{\lvert\delta\rvert}$
Classic (D455)	-1.85	11.6	9.3	7.15
Extended (D405)	-1.01	5.2	10.2	8.95

IV. Benchmark

A. Evaluation metrics

We adhere to the BOP challenge protocol [48], evaluating methods using four pose error metrics based on the estimated pose and the ground-truth pose: the Visible Surface Discrepancy (VSD), computing differences from renderings of the estimated and ground-truth poses only over the visible surface areas; the Maximum Symmetry-Aware Surface Distance (MSSD), measuring the maximum surface distance considering object symmetry; the Maximum Symmetry-Aware Projection Distance (MSPD) measuring the maximum 2D projection error in pixels, considering symmetry; and the Average Distance for distinguishable (ADD) objects which quantifies the average misalignment between the model's vertices in the true and estimated pose. Although not symmetry-aware, ADD is widely used in prior literature and provides a valuable reference point for evaluating pose accuracy. In line with the BOP protocol, the Average Recall (AR), used to summarize the overall performance, is computed as the mean recall of VSD, MSSD, and MSPD. In addition to pose estimation metrics, we also report the mean Average Precision (mAP) for both detection and segmentation, following common practice in the BOP challenge.

B. Baselines and evaluation setup

We evaluate object detection, segmentation and pose estimation methods using both instance-level and novel object baselines. For instance-level methods, we deliberately selected EPOS [5], ZebraPose [7], and DOPE [6]. EPOS was chosen for its efficiency and robustness, as it handles object symmetries better than early BOP baselines such as Pix2Pose [30], supports multi-object training within a unified network and remains competitive in the BOP benchmark, particularly under the VSD and MSSD metrics. ZebraPose was included as a state-of-the-art representative from the 2023 BOP challenge and DOPE was selected as a representative keypoint-based approach. In contrast, we excluded detector-dependent methods such as GDRNet [32], whose accuracy is tightly coupled to bounding-box quality and whose per-object training time made it impractical for our benchmarking scope, as reported in [1]. All selected instance-level methods were trained and tested on the Classic IndustryShapes dataset. For novel-object methods, we included FoundPose [8] (a model-based approach) and FoundationPose [9] (supporting both model-based and model-free setups). We chose to exclude other methods such as Gen6D [40] due to the significant manual effort required to adapt it to custom objects, and OnePose++ [39] as it relies on an iOS-only mobile app for data capture that does not scale well to large datasets. Both selected methods were evaluated directly on the Classic and Extended datasets without retraining. Publicly available code was used for all methods.

For object detection, we include CNOS [10], which leverages contrastive learning to generalize to unseen objects without retraining. For segmentation, we adopt SAM-6D [11], a recent approach that combines the Segment Anything Model (SAM) with a 6D pose estimation head, enabling robust segmentation-driven object localization.

C. Performance evaluation

a) 6D Pose Estimation: The results, summarized in Table V, provide a direct comparison of the baseline methods across the diverse object types and varying scene complexities. The challenging nature of the dataset is evident from the generally low performance across most methods. Notably, the strength of recent novel object pose estimation approaches is emphasized, as they achieve comparable performance to instance-level methods. In addition, an analysis of the AR per object reveals significant variability in performance depending on the specific object and scene context (Table VI). This is evident in the novel-object case, where no training is performed on the target objects. Certain objects consistently prove to be more challenging than others. For instance, Obj. 5, which is reflective with a thin structure, shows significantly lower performance compared to Obj. 1. This trend is also observable in the results of instance-level methods, indicating that object-specific factors contribute in pose estimation accuracy. The domain shift between training and test data is a persistent issue with objects appearing in train scenes with more complex background and clutter to exhibit better accuracy at test time The relatively low performance of DOPE can be attributed to its reliance solely on image data and the corresponding projected bounding cuboids, without utilizing CAD models. While CAD models could be used to generate additional synthetic data and increase viewpoint coverage and scene diversity, as proposed by the authors, we chose not to incorporate such data to maintain a consistent evaluation across all compared methods.

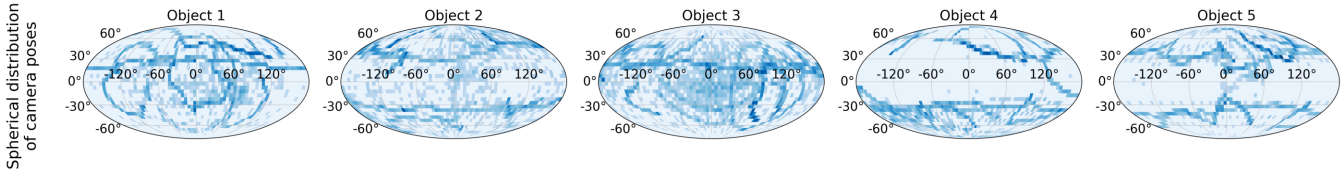

Fig. 3: Pose distribution per Object. Visualization of the overall spherical viewpoint coverage of the complete IndustryShapes dataset in Mollweide projection, indicating the density and pose variation.

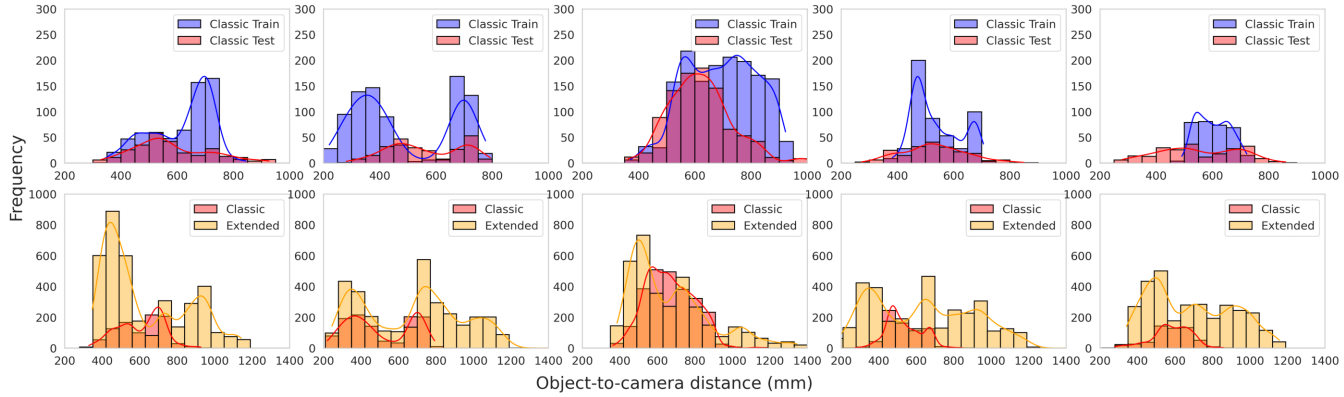

Fig. 4: Distribution of object-to-camera distances for annotated poses, grouped by object (1 to 5 from left to right). Top row: annotated poses in the training (blue) and test (magenta) data of the *classic set*. Bottom row: annotated poses of the *classic set* (orange) and the *extended set* (yellow).

TABLE V: Performance of baseline methods on IndustryShapes dataset, *classic* and *extended* sets. The table shows Average Recall (AR) scores per method based on [48], along with the recall rates for the ADD, MSPD, MSSD and VSD error metrics.

		IndustryShapes Classic					**IndustryShapes Extended**				
		ADD	**MSPD**	**MSSD**	**VSD**	**AR**	**ADD**	**MSPD**	**MSSD**	**VSD**	**AR**
Inst. level	EPOS	0.64	0.54	0.56	0.41	**0.51**	-	-	-	-	-
	DOPE	0.13	0.09	0.09	0.05	0.08	-	-	-	-	-
	Zebrapose	0.61	0.54	0.55	0.40	0.50	-	-	-	-	-
Novel objects	FoundPose	0.55	0.31	0.39	0.22	0.30	0.34	0.41	0.25	0.18	0.28
	FoundationPose (MB)	0.78	0.73	0.74	0.53	**0.67**	0.81	0.83	0.74	0.50	**0.69**
	FoundationPose (MF)	0.44	0.26	0.31	0.18	0.25	0.48	0.40	0.37	0.22	0.33

TABLE VI: Average Recall (AR) per object per baseline method across the IndustryShapes dataset.

		IndustryShapes Classic					**IndustryShapes Extended**				
		Obj 1	**Obj 2**	**Obj 3**	**Obj 4**	**Obj 5**	**Obj 1**	**Obj 2**	**Obj 3**	**Obj 4**	**Obj 5**
Inst. level	EPOS	0.66	**0.26**	0.59	**0.25**	0.00	-	-	-	-	-
	DOPE	0.18	0.05	0.03	0.10	0.00	-	-	-	-	-
	Zebrapose	**0.84**	0.11	0.49	**0.39**	**0.11**	-	-	-	-	-
Novel objects	FoundPose	0.43	0.34	0.25	0.11	0.15	0.48	0.17	0.13	0.23	0.04
	FoundationPose (MB)	**0.78**	**0.49**	**0.64**	**0.47**	**0.31**	**0.79**	**0.54**	**0.42**	**0.54**	**0.56**
	FoundationPose (MF)	0.53	0.32	0.12	0.09	0.23	0.47	0.26	0.07	0.02	0.38

b) Detection and segmentation: The detection and segmentation results are summarized in Table VII. Both CNOS [10] and SAM-6D [11] achieve moderate performance, with consistently higher AP values on the extended dataset compared to the classic set. The main difference between the methods arises in segmentation on the classic set, where SAM-6D shows a clear advantage over CNOS, consistent with its SAM-based design for more precise mask generation. These findings highlight the complementary strengths of the two approaches and underline the continued challenges of accurate object localization in cluttered industrial scenes.

V. Conclusions

The IndustryShapes dataset features five challenging industrial components and tools and supports a wide range of

evaluation scenarios, from single-object scenes to complex, cluttered environments with occlusions and multiple objects. The dataset is divided into two parts: the classic set and the extended set. The extended set is designed to support the evaluation of novel object pose estimation methods, particularly for industrially relevant objects, and includes RGB-D static onboarding sequences. Representative state-of-the-art methods were evaluated and the results demonstrate the strengths and current limitations of these methods in the industrial-related domain.

Limitations and future work. Some test scenes in the classic set comprise of fewer frames than others which may lead to uneven object representation across the dataset. The discrepancy in training scene complexity, with some objects trained on cluttered, realistic setups and others on simpler ones affects the performance of instance-level methods. To address data imbalance and ensure fuller pose coverage, we plan to scale the dataset by generating photorealistic synthetic scenes using BlenderProc [44] and incorporating additional real-world industrial data.

TABLE VII: Mean Average Precision (mAP) of baseline methods on IndustryShapes dataset for detection and segmentation.

	BBox – mAP		Segm – mAP	
Method	**Classic**	**Extended**	**Classic**	**Extended**
CNOS	0.240	**0.574**	0.203	**0.512**
SAM6D	**0.270**	0.527	**0.345**	0.453

REFERENCES

[1] S. Thalhammer, D. Bauer, *et al.*, "Challenges for Monocular 6-D Object Pose Estimation in Robotics," *TOR*, pp. 4065–4084, 2024.

[2] T. Hodan, P. Haluza, *et al.*, "T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-Less Objects," in *Proc. WACV*, 2017.

[3] B. Drost, M. Ulrich, *et al.*, "Introducing MVTec ITODD - A Dataset for 3D Object Recognition in Industry," in *Proc. ICCVW*, 2017.

[4] Y. Xiang, T. Schmidt, *et al.*, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," in *Proc. RSS*, 2018.

[5] T. Hodaň, D. Baráth, and J. Matas, "EPOS: Estimating 6D pose of objects with symmetries," in *Proc. CVPR*, 2020.

[6] J. Tremblay, T. To, *et al.*, "Deep object pose estimation for semantic robotic grasping of household objects," in *Proc. CoRL*, 2018.

[7] Y. Su, M. Saleh, *et al.*, "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation," in *Proc. CVPR*, 2022.

[8] E. P. Örnek, Y. Labbé, *et al.*, "Foundpose: Unseen object pose estimation with foundation features," in *Proc. ECCV*, 2024.

[9] B. Wen, W. Yang, *et al.*, "FoundationPose: Unified 6d pose estimation and tracking of novel objects," in *Proc. CVPR*, 2024.

[10] V. N. Nguyen, T. Groueix, *et al.*, "Cnos: A strong baseline for cad-based novel object segmentation," in *Proc. CVPR*, 2023.

[11] J. Lin, L. Liu, *et al.*, "SAM-6D: Segment Anything Model Meets Zero-Shot 6D Object Pose Estimation," in *Proc. CVPR*, 2024.

[12] E. Brachmann, A. Krull, *et al.*, "Learning 6D Object Pose Estimation Using 3D Object Coordinates," in *Proc. ECCV*, 2014.

[13] R. Kaskman, S. Zakharov, *et al.*, "HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects," in *Proc. ICCVW*, 2019.

[14] A. Doumanoglou, R. Kouskouridas, *et al.*, "Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd," in *Proc. CVPR*, 2016.

[15] T. Hodan, F. Michel, *et al.*, "BOP: Benchmark for 6D Object Pose Estimation," in *Proc. ECCV*, 2018.

[16] S. Hinterstoisser, V. Lepetit, *et al.*, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Proc. ACCV*, 2013.

[17] Y. Lin, J. Tremblay, *et al.*, "Multi-view fusion for multi-level robotic scene understanding," in *Proc. IROS*, 2021.

[18] C. Rennie, R. Shome, *et al.*, "A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place," *RA-L*, pp. 1179–1185, 2016.

[19] A. Tejani, D. Tang, *et al.*, "Latent-class hough forests for 3d object detection and pose estimation," in *Proc. ECCV*, 2014.

[20] "xyzibd dataset," https://huggingface.co/datasets/bop-benchmark/xyzibd, 2025.

[21] A. Kalra, G. Stoppi, *et al.*, "Towards co-evaluation of cameras hdr and algorithms for industrial-grade 6dof pose estimation," in *Proc. CVPR*, 2024.

[22] A. Guo, B. Wen, *et al.*, "HANDAL: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions," in *Proc. IROS*, 2023.

[23] P. Banerjee, S. Shkodrani, *et al.*, "HOT3D: Hand and object tracking in 3D from egocentric multi-view videos," in *Proc. CVPR*, 2025.

[24] A. Ahmadyan, L. Zhang, *et al.*, "Objectron: A large scale dataset of object-centric videos in the wild with pose annotations," in *Proc. CVPR*, 2021.

[25] Y. You, K. Xiong, *et al.*, "Pace: A large-scale dataset with pose annotations in cluttered environments," in *Proc. ECCV*, 2025.

[26] Z. Fan, Y. Zhu, *et al.*, "Deep Learning on Monocular Object Pose Detection and Tracking: A Comprehensive Overview," *ACMCS*, pp. 1–40, 2022.

[27] G. Marullo, L. Tanzi, *et al.*, "6D object position estimation from 2D images: A literature review," *Multimedia Tools and Applications*, pp. 24 605–24 643, 2023.

[28] J. Liu, W. Sun, *et al.*, "Deep Learning-Based Object Pose Estimation: A Comprehensive Survey," 2024, arXiv:2405.07801.

[29] S. Peng, Y. Liu, *et al.*, "Pvnet: Pixel-wise voting network for 6dof pose estimation," in *Proc. CVPR*, 2019.

[30] K. Park, T. Patten, and M. Vincze, "Pix2Pose: Pixel-wise coordinate regression of objects for 6D pose estimation," in *Proc. ICCV*, 2019.

[31] W. Kehl, F. Manhardt, *et al.*, "SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again," in *Proc. ICCV*, 2017.

[32] G. Wang, F. Manhardt, *et al.*, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proc. CVPR)*, 2021.

[33] Y. Labbé, L. Manuelli, *et al.*, "Megapose: 6d pose estimation of novel objects via render & compare," in *Proc. CoRL*, 2022.

[34] V. N. Nguyen, T. Groueix, *et al.*, "Gigapose: Fast and robust novel object pose estimation via one correspondence," in *Proc. CVPR*, 2024.

[35] I. Shugurov, F. Li, *et al.*, " OSOP: A Multi-Stage One Shot Object Pose Estimation Framework ," in *Proc. CVPR*, 2022.

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

[37] P. Ausserlechner, D. Haberger, *et al.*, "ZS6D: Zero-shot 6D Object Pose Estimation using Vision Transformers," in *Proc. ICRA*, 2024.

[38] J. Sun, Z. Wang, *et al.*, "OnePose: One-shot object pose estimation without CAD models," in *Proc. CVPR*, 2022.

[39] X. He, J. Sun, *et al.*, "Onepose++: Keypoint-free one-shot object pose estimation without CAD models," in *Proc. NeurIPS*, 2022.

[40] Y. Liu, Y. Wen, *et al.*, "Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images," in *Proc. ECCV*, 2022.

[41] B. Wen, J. Tremblay, *et al.*, "BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects," in *Proc. CVPR*, 2023.

[42] A. Kirillov, E. Mintun, *et al.*, "Segment anything," *arXiv:2304.02643*, 2023.

[43] M. Oquab, T. Darcet, *et al.*, "Dinov2: Learning robust visual features without supervision," 2023.

[44] M. Denninger, D. Winkelbauer, *et al.*, "Blenderproc2: A procedural pipeline for photorealistic rendering," *Journal of Open Source Software*, 2023.

[45] A. Papadaki and M. Pateraki, "6d object localization in car-assembly industrial environment," *Journal of Imaging*, 2023.

[46] M. Pateraki, P. Sapoutzoglou, and M. Lourakis, "Crane Spreader Pose Estimation from a Single View," in *Proc. VISAPP*, 2023, p. 796–805.

[47] Agisoft LLC, "Agisoft metashape," https://www.agisoft.com/, 2025.

[48] T. Hodaň, M. Sundermeyer, *et al.*, "BOP challenge 2020 on 6D object localization," in *Proc. ECVA*, 2020.