Towards Co-Evaluation of Cameras, HDR, and Algorithms for Industrial-Grade 6DoF Pose Estimation

Agastya Kalra Guy Stoppi Dmitrii Marin Vage Taamazyan Aarrushi Shandilya
Rishav Agarwal Anton Boykov Tze Hao Chong Michael Stark
Intrinsic Innovation LLC

Abstract

6DoF Pose estimation has been gaining increased importance in vision for over a decade, however it does not yet meet the reliability and accuracy standards for mass deployment in industrial robotics. To this effect, we present the Industrial Plenoptic Dataset (IPD): the first dataset for the co-evaluation of cameras, HDR, and algorithms targeted at reliable, high-accuracy industrial automation. Specifically, we capture 2,300 physical scenes of 20 industrial parts covering a $1m \times 1m \times 0.5m$ working volume, resulting in over 100,000 distinct object views. Each scene is captured with 13 well-calibrated multi-modal cameras including polarization and high-resolution structured light. In terms of lighting, we capture each scene at 4 exposures and in 3 challenging lighting conditions ranging from 100 lux to 100,000 lux. We also present, validate, and analyze robot consistency, an evaluation method targeted at scalable, high accuracy evaluation. We hope that vision systems that succeed on this dataset will have direct industry impact. The dataset and evaluation code are available at https://github.com/intrinsic-ai/ipd.

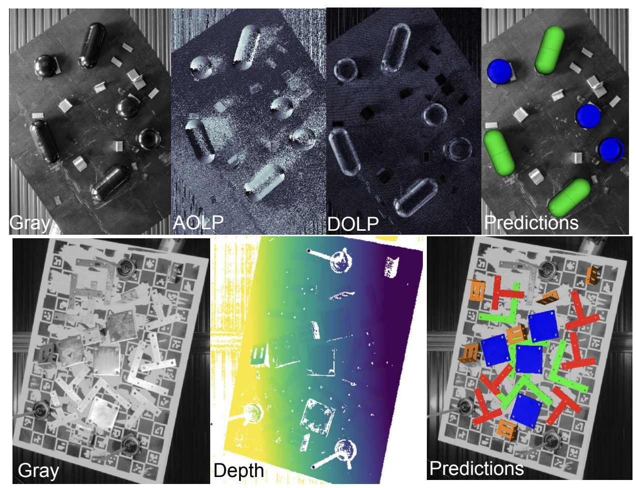

Figure 1. **The Industrial Plentopic Dataset (IPD) affords co-evaluation of cameras, HDR, and pose estimation algorithms.** Our dataset contains polarization (top) which provides high contrast for dark objects, and high-res structured-light (bottom).

1. Introduction

Only a small fraction of the thousands of robot arms currently deployed in factories use vision, with even fewer using 6DoF object pose estimation [7, 43]. Instead, the majority of factories still rely on mechanical fixtures and pre-planned robot motions that need to be re-fabricated and reprogrammed whenever the work product changes. This is due to the fact that computer vision in general, and 6DoF pose estimation in particular, is not seen as reliable enough for industrial applications [43], where requirements include sub-millimeter accuracy in pose estimation, robustness to challenging surface properties and drastic changes in factory lighting conditions, and guaranteed levels of recall. This may seem surprising, considering that 6DoF pose estimation has been an active area of research in terms of both algorithms [9, 11–13, 15, 16, 31, 38, 42, 44, 47, 48, 52, 54, 65, 70, 72] and benchmarks [21, 22, 26, 32–34, 37, 44, 45, 55, 61, 63, 67, 71] for over a decade. We argue that research has so far been lacking in one key ingredient to enable the step to real-world factory floors: *the co-design of camera systems, high dynamic range imaging (HDR), and corresponding pose estimation algorithms*.

In particular, we find that existing public 6DoF pose estimation benchmarks rarely provide enough variety in terms of cameras, viewpoints, exposure settings, and modalities (e.g., in the form of RGB, depth, and polarization) to allow the joint optimization of cameras, HDR, and algorithms that enables industrial applications with high precision requirements. Some datasets collect all data with only a single camera [31, 45, 63, 70]. Others use multiple, but derive ground truth poses from just a single camera, which is prone to generating biased labels [22, 32, 37]. Only two datasets consider polarization [34, 67], however only from a single-camera, without complex lighting or HDR, and in the context of household objects.

At the same time, existing benchmarks do not measure industrial-grade 6DoF accuracy. Most benchmarks for 6DoF pose estimation [22, 33, 37, 71] define an acceptable pose estimate typically as 10% of the object diameter (5-45mm) in ADD [31], at distances of 50-100cm. This does not meet industrial requirements. Our work focuses on industrial-grade performance, meaning an acceptable estimate to be $<$ 3mm MVD (defined in Sec. 3.1) at distances of 150-200cm at 99% recall. Note, MVD is the upper bound of ADD. Measuring 3mm MVD at distances of 150-200cm is difficult using the existing methods [33] of measuring accuracy relative to (inaccurate) annotated ground truth.

This paper aims to address both limitations, as follows.

First, we introduce a novel dataset of 20 industrial objects and 2,300 scenes, tailored towards real-world tasks with high precision requirements. We capture our dataset with a total of 13 calibrated cameras at large baselines, including RGB, structured light, and polarization cameras (see Tab. 1(b)), and include challenging lighting conditions in multiple exposures, allowing researchers to co-design HDR and 6DoF algorithms to be invariant to lighting. Polarization as a modality is critical for mitigating glare [58, 73], enhancing HDR [69, 74], identifying edges of highly reflective/translucent surfaces [36, 68], and improving 6DoF pose [27]. Our 4 polarization cameras capture polarized HDR images and are placed at multiple locations to allow for experimentation with multiple baselines.

Second, we introduce *Robot Consistency*, inspired by [35], to the scientific community as a means of evaluating and comparing the performance of highly precise 6DoF pose estimation algorithms at scale. In contrast to methods based on (inaccurate) ground truth, it leverages consistency between estimates of the pose estimation algorithm under test to assess performance. We theoretically and empirically demonstrate the validity of this approach, and highlight its favorable properties for the high-precision setting.

The rest of this paper is organized as follows. Sec. 2 reviews related work. Sec. 3 introduces Robot Consistency and establishes its validity. Sec. 4 introduces our novel dataset of industrial parts. Sec. 5 highlights directions of co-design afforded by our dataset and Sec. 6 concludes the paper. Additional results are included in the supplement.

2. Related Work

We begin with an overview of related work, focusing on industrial datasets (Sec. 2.1) and evaluation methods (Sec. 2.2) for 6DoF pose estimation.

2.1. Industrial Datasets

While a large number of datasets for 6DoF pose estimation exist, a majority of them is targeted primarily at household objects. These datasets include the widely used LineMOD [31] and YCB-Video [70], the recently introduced PhoCaL [67] and HouseCat6D [34] among others such as LabelFusion [48], Occluded LineMOD [9], HOPE [63], TOD [44], Wild6D [26], NOCS [65], and [21, 55, 61]. While household objects often pose challenges in terms of geometry and appearance variations, they tend to be less demanding in terms of lighting or required accuracy.

In contrast, only a handful of existing datasets is concerned with the industrial high precision settings that provide the basis for our exploration (Tab. 1 compares key characteristics of those existing datasets to ours). T-LESS [32] features texture-less industrial objects with more than $100,000$ images for all sensors combined, but provides limited ground truth pose accuracy at around 5mm. HomebrewedDB [37] includes 8 industrial parts out of 33 total objects, with a moderate ground truth pose accuracy of around 2mm. MVTec ITODD [22] covers a wide range of industrial parts but it has only around 100 to 200 instances per part and does not provide publicly available ground truth poses. ROBI [71] has 7 parts in heavy clutter. DIMO [57] provides two types of lighting conditions and different types of backgrounds for 6 parts. StereOBJ-1M [45] covers 11 different scene types (including 3 outdoor scenes), and more than $396,000$ captured images. However, it does not provide any depth maps. Finally, Fraunhofer IPA Bin-Picking [39] is mostly focused on simulated scenes.

In summary, none of these datasets comes close to ours (last line, Tab. 1) in terms of provided cameras (13), modalities (RGB, depth, polarization images, HDR) and 3 lighting conditions with 4 exposures, making ours the first to enable co-evaluation of cameras, HDR, and pose estimation algorithms.

2.2. Evaluation methods

Existing methods for evaluating 6DoF pose estimation algorithms typically rely on some notion of ground truth data.

Synthetic Data Generation. Synthetically generated datasets [20, 62] simplify the process of accurate ground truth generation, but at a cost: First, they often sacrifice the fidelity of the rendered data, notably in scenarios involving polarized light. Second, content creation becomes a bottleneck [48] and the lack in variability and unpredictability might lead to overfitting in trained models [20, 33, 45].

3D Point Cloud Registration. This method typically relies on a 3D reconstruction of a physical scene, achieved through Time-of-Flight [25, 32, 33, 55] or Structured Light sensors [32]. It involves extracting object poses by aligning the object's CAD model with the scene's 3D reconstruction, either manually [25, 26, 32, 33, 55], semi- [22, 34, 63, 67, 70], or fully-automatically [37]. However, this method has fundamental limitations. First, its accuracy depends on the

Industial Dataset	**Parts**	**Cams**	**Modalities**	**Frames**	**Object Instances**	**Lighting Conditions**	**HDR**	**Working Distance, cm**	**Claimed Annotation Accuracy, mm**	**Scaled Annotation Accuracy, mm**
DIMO [57]	6	4	RGB-D	31,200	100k	2	No	<50	0.3	2.7
ROBI [71]	7	2	RGB-D	8,000	600k	1	No	<50	0.2	1.8
HomebrewedDB [37]	8	2	RGB-D	20,000	100k	1	No	<140	2.0	2.3
ITODD [22]	28	5	RGB-D	800	5k	1	No	<50	0.2	1.8
T-LESS [32]	**30**	3	RGB-D	147k	100k	1	No	<100	5.0	11.3
StereOBJ-1M [45]	18	2	RGB	**396k**	**1.5M**	2	No	<150	2.3	2.3
Ours	20	**13**	**RGB-D+Polar**	30,000	100k	**3**	**Yes**	**150-200**	N/A	N/A

(a)

Camera	**#**	**Resolution**	**Modalities**
Basler-LR [2]	3	1920 x 1200	RGB
Basler-HR [1]	5	2592 x 1944	RGB
FLIR-MonoP [3]	4	2448 x 2048	Gray, Polar
Photoneo [5]	1	2064 x 1544	RGB, Depth

(b)

Table 1. **Industrial datasets overview.** (a) Comparing existing industrial datasets with ours. Scaled annotation accuracy is an estimated annotation accuracy at 150 cm with a quadratic z-error decay on distance [4]. (b) Cameras used in the proposed dataset.

3D reconstruction's precision, often requiring sensors more accurate than the one being evaluated. Depth sensors have difficulty with transparent or reflective surfaces [44, 45] and are less effective in densely populated scenes. To mitigate these issues, some approaches resort to scanning spray [71] or the use of a robot/handheld wand [34, 67], but these increase the data collection effort. Even (semi-)automatic methods often fail to achieve the desired accuracy, necessitating manual verification [61, 63, 71]. Moreover, these methods are sensitive to environmental factors like lighting [33] and sensor-object distance [71]. Second, this approach presupposes the availability of accurate CAD models or meshes of the objects (e.g. [32–34, 37]).

2D Keypoint-based Pose Estimation. This method capitalizes on the relative ease of annotating two-dimensional keypoints on images. It involves capturing a scene from various angles, followed by manual annotation of correspondences between the CAD model and selected images. Poses are then either extracted by using triangulation [44, 45] or the Perspective-n-Point (PnP) algorithm [63]. Notable challenges include human labeling accuracy, the necessity of a multi-view camera setup, adding the complexity of inter-camera calibration [44, 45], and the re-projection of manually annotated labels from one view to others, particularly when considering occlusions. This tends to be biased towards the camera for which the labeling is done, making comparing cameras difficult.

Fixture Based Detection. This approach attaches a highly detectable marker to the object, precisely calibrating its relative position. Variants include using an active target with a laser tracking system for extreme precision [49] or a mechanical fixture that measures its position and orientation [49]. Despite their precision, these methods have drawbacks, such as high cost, limited scene complexity [49], and not being reusable for different objects [49].

3. Scalable, Accurate Evaluation

In this section, we present the Robot Consistency evaluation pipeline, inspired by [35], that is both scalable and allows the accurate measurement of small errors. To that end, we first describe the methodology (Sec. 3.1) and provide theoretical justification for its validity (Sec. 3.2). We then highlight its favorable properties for high-precision applications in extensive experiments on synthetic data (Sec. 3.3).

3.1. Robot Consistency

Robot Consistency relies on a well calibrated robot to generate a sequence of *visual scenes* (physical scenes of objects imaged from different viewing directions) and exploits the known relative rigid transforms between them to estimate standard pose metrics. Prior work focuses on evaluating against annotated ground truth, which is likely to both be biased and inaccurate. Robot Consistency, however, relies on accurate robot transforms, making it valuable for high-precision requirements. Algorithm 1 lists all relevant steps.

Assume there is a fixed camera C observing a working volume. There is a robot arm R with a gripper G. We rigidly mount object O on the gripper (Sec. 3.1.2 extends this to multiple objects). Let T_{CR} be the transform from the robot base frame to the camera frame given from hand-eye calibration, and T_{GO} be the unknown transform from the object frame to the gripper (end effector) coordinates. We capture images of the object in N different robot configurations. For the i-th capture, we record the predicted 6DoF pose ${}^{pred}T^i_{CO}$, which is the transform from the object frame to the camera coordinates, and the transform from the gripper frame into robot coordinates T^i_{RG}. Our goal is to evaluate the accuracy of predicted object poses ${}^{pred}T^i_{CO}$. Since the ground truth ${}^{gt}T^i_{CO}$ is not available, most of prior literature uses annotated ground truth ${}^{ann}T^i_{CO}$:

$$E_{ann}(\mathcal{T}_{\text{pred}}) = \frac{1}{N}\sum_i d({}^{pred}T^i_{CO}, {}^{ann}T^i_{CO}) \tag{1}$$

where d is a pose metric and $\mathcal{T}_{pred} = \{{}^{pred}T^i_{CO}\}_i$ is the set of all pose predictions.

We propose measuring the pose error against the robot instead. Intuitively, this means that if the robot arm moves in a particular way, then the pose estimates should move the same. This is enabled by the kinematic calibration of the robot, which is up to 0.1 mm accurate.

Note, that the unknown object pose can be expressed as

$$T_{GO} = (T^i_{RG})^{-1} T^{-1}_{CR} \{{}^{pred}T^i_{CO}\} = T^i_{GR} T_{RC} \{{}^{pred}T^i_{CO}\} \tag{2}$$

where we used $T_{AB} \equiv T_{BA}^{-1}$ for any reference frames A, B. Eq. (2) provides N distinct measurements of unknown T_{GO}. Inspired by the well-known statistical law of large numbers [46] one can hypothesise that a mean[1] of these measurements will be a good estimate of T_{GO}:

$$T_{GO}^{*} = \arg\min_{\mu \in SE(3)} \sum_{i=1}^{N} \rho^2(\mu, T_{GR}^{i} T_{RC} \{{}^{pred}T_{CO}^{i}\}) \tag{3}$$

where ρ is a distance between transforms. It turns out that any choice of ρ (subject to minor constraints) converges to T_{GO} at the same rate. This is true for unbiased pose estimates ${}^{pred}T_{CO}^{i}$ but also for some forms of bias when T_{RG}^{i} is chosen in a particular way, see Sec. 3.2. We also validate the method on real pose estimators in Sec. 3.3. In practice, we use the sample average for the translation and quaternion averaging [28] for the rotation in Eq. (3). Then, we can approximate any pose metric d:

$$E_{rc}(\mathcal{T}_{\text{pred}}) = \frac{1}{N} \sum_{i} d({}^{pred}T_{CO}^{i}, T_{CR} T_{RG}^{i} T_{GO}^{*}) \tag{4}$$

Discussion Measuring Eq. (4) introduces a bias as we reduce the number of degrees of freedom by estimating T_{GO} through minimization. However given large N and diverse robot poses, we show theoretically and empirically that Robot Consistency converges to the true value of the metric and is more accurate than annotations.

3.1.1 Pose Accuracy Metrics

The standard metric d is ADD [31]:

$$ADD({}^{pred}T_{CO}, {}^{gt}T_{CO}) = \frac{1}{N} \sum_{k} ||{}^{pred}T_{CO} v_k - {}^{gt}T_{CO} v_k|| \tag{5}$$

where $V = \{v_k\}_k$ is a set of vertices on object mesh. However, we prefer using another metric, namely Maximum Vertex Distance (MVD):

$$MVD({}^{pred}T_{CO}, {}^{gt}T_{CO}) = \max_{i} ||{}^{pred}T_{CO} v_k - {}^{gt}T_{CO} v_k||. \tag{6}$$

This is motivated by industrial peg-in-hole insertion tasks that fail if any portion of the object exceeds the error bound. MVD is also independent of vertex sampling, whereas ADD is heavily dependent on the distribution of vertices.

3.1.2 Scenes with Multiple Objects

When there are multiple objects mounted to the robot, there is a correspondence problem, that is, we do not know which prediction in scene i corresponds which object j on the robot. To resolve this, we provide annotations ${}^{ann}T_{GO}^{j}$ which are used only for determining the correspondence. For each object j, the closest prediction to ${}^{ann}T_{GO}^{j}$ in each scene is grouped together before calculating MVD using Eq. (4) and Eq. (3).

[1] Due to the structure of the rigid transforms $SE(3)$ we cannot use standard arithmetic mean but have to resort to a more general Fréchet mean.

Algorithm 1 Robot Consistency.

0. **Hand-Eye Calibration.** Obtain hand-eye calibration T_{CR} (transform between camera and robot base) using checkerboard pose estimation.
1. **Setup.** Mount an object O rigidly to the robot arm.
2. **Data Capture.** Move the robot arm to $1..i..N$ different gripper poses, recording the transform T_{RG}^{i} from gripper G to robot base R.
3. **Prediction.** Run the pose estimation algorithm under test on the desired camera setup to yield pose predictions ${}^{pred}T_{CO}^{i}$.
4. **Conversion.** Convert all pose predictions to robot base coordinates using T_{CR} and then to gripper coordinates using the recorded T_{RG}^{i}.
5. **Evaluation.** Calculate the pose error E_{rc} using Eq. (3) and Eq. (4).

Semi-Automated Annotations: We compute ${}^{ann}T_{GO}^{j}$ using a semi-automated pipeline. Specifically, for a sequence of N scenes, we run the best performing pose estimator on all scenes, yielding K_i pose predictions ${}^{pred}T_{CO}^{i,k_i}$ for scene $i \in \{1, \dots, N\}$. which we transform to the gripper coordinate system using Eq. (2):

$$\mathcal{H} = \{T_{GR}^{i} T_{RC} ({}^{pred}T_{CO}^{i,k_i}) \mid 1 \le i \le N, 1 \le k_i \le K_i\} \tag{7}$$

We then cluster $\mathcal{H}$ using DBSCAN with pose distance threshold ϵ, such that each cluster corresponds to a set of spatially consistent object hypotheses. Any cluster larger than N' is considered valid[2], while others are rejected as outliers. All predictions in each cluster are averaged to create an estimate of ${}^{ann}T_{GO}^{j}$. A manual inspection and filtering step concludes the computation of reference pose annotations. This must be done once per dataset.

3.2. Theoretical Validation

We discuss basic statistical properties of Robot Consistency, assuming for simplicity the single object case. First, we consider the case where the pose estimates $\Phi(t_i) = {}^{pred}T_{CO}^{i}$ are unbiased and independent. Here Φ is the pose estimator under evaluation and t_i is the image capture for scene $i \in \{1, \dots, N\}$. Under mild conditions the poses estimated by Robot Consistency Eq. (3) converge in probability to the actual ground truth as N increases. We can obtain a good proxy to the ground truth via Eq. (3) by increasing the number of captures N. Likewise, approximations of common pose metrics Eq. (4) also converge to their true value. Second, we discuss an extension to a biased case.

Robot Consistency can be seen as a pose estimation method, which is built on top of Φ. In general, an estimator is called *consistent* if it converges in probability to

[2] In practice, we set N' as 25% of N.

the estimated value. An estimator is called *unbiased* if its mathematical expectation equals the estimated value. Assume that the original pose estimates $\Phi(t_i)$ are statistically independent from each other. Then, assume that pose estimator Φ is unbiased. We consider Eq. (3) separately for the rotation and translation components. In case of translation, we assume the Euclidean distance. Hence, the translation component of Eq. (3) is the sample mean.

Proposition 1. *The Robot Consistency pose estimator is consistent and unbiased.*[3]

Proof. The translation component is unbiased and consistent due to standard properties of the sample mean estimate. The rotation component is unbiased and consistent due to [8, 24]. Moreover, an extension of the Central Limit Theorem for manifolds [8, 24] applies giving the rate of convergence $N^{-\frac{1}{2}}$ in both components. □

Biased Posed Estimators Let $P \in S$ denote camera pose, where S defines the domain of the camera pose distribution. Let camera pose P be defined as a rigid transform from the world to the camera frame of reference. Let further $\Phi_T(t_P)$ be the translation component of pose estimator result $\Phi(t_P)$ in the world reference frame for image t_P corresponding to the camera at pose P. Let P^* be the true pose of the target object in the world reference frame. We assume the world reference frame is chosen such that $P^* = I$ is identity. $\mathbb{E}$ denotes mathematical expectation. Here randomness is due to the unknown camera pose and inaccuracy of the pose estimator Φ.

Prop. 2 below shows that under these conditions the expected Robot Consistency pose estimate Eq. (3) is unbiased.

Proposition 2. *If (1) the camera pose distribution is spherical w.r.t. to the target object, and (2) the pose estimator's* translation *error has the mean value of $\mu \neq 0$ in the reference frame of the camera for any camera pose (bias is only w.r.t. camera),* i.e.*, the expectation of predicted pose translation given camera pose is $\mathbb{E}(P\Phi_T(t_P)|P) = \mu$. Then, the expected pose estimate translation $\hat{P} = \mathbb{E}\Phi_T(t_P)$ is zero.*

Proof.

$$\begin{aligned}\hat{P} = \mathbb{E}\Phi(t_P) = \mathbb{E}\{\mathbb{E}(\Phi(t_P)|P)\} = \mathbb{E}\{P^{-1}\mathbb{E}(P\Phi(t_P)|P)\} \\ = \mathbb{E}\{P^{-1}\mu\} = \mathbb{E}\{P^{-1}\}\mu = 0 \cdot \mu = 0 \qquad (8)\end{aligned}$$

□

Discussion While this theory does not account for robot errors, calibration errors, many types of biases, etc., it shows that Robot Consistency does not introduce new biases, in some cases improves evaluated pose estimator biases, and converges to the ground truth-based evaluation.

[3] See additional mild conditions in [8, 24].

3.3. Synthetic Validation

Having established the validity of Robot Consistency in theory (Sec. 3.2), we now proceed to highlighting its favorable properties in comparison to approaches based on annotated ground truth, in particular for high-precision settings. To that end, we conduct an extensive experimental study in a controlled setting, using synthetic data.

Synthetic data and pose estimation algorithms We render 20 physical scenes with 10 object instances each from 30 different viewpoints, using 4 different cameras of 5 MP resolution. Camera setup and pose distribution closely match our dataset (Sec. 4). We train and evaluate a population of 24 different key point-based models differing in NN backbone and degree of convergence, each with and without edge-alignment [17] (see Sec. 5.1), which we deem representative of algorithms of different performance levels.

Evaluation methods under test. Since we have access to ${}^{gt}T_{CO}$ of each object in our synthetic scenes, we compare the accuracy of robot consistency E_{rc} to evaluation methods based on different levels of human annotation quality. Specifically, we generate 3 degradations of the true poses by adding random jitters in pose until an MVD of 1, 2, or 3 mm is reached. From Tab. 1, existing datasets show 2-3mm annotation accuracy when scaled to our dataset's working distance (150-200cm) using n^2 z-decay [4].

Testing methodology. We quantify the degree to which different evaluation methods under test capture the true performance of a pose estimation algorithm in three ways: (1) as the correlation between the performance estimate provided by the evaluation method under test and the true performance, as measured w.r.t. true synthetic pose ${}^{gt}T_{CO}$, on the population of representative pose estimation algorithms (Fig. 2 (a)), (2) as the mean absolute difference between the performance estimate provided by the evaluation method under test and the true performance. (Fig. 2 (b)), and (3) as the empirical probability with which an evaluation method under test provides the correct ordering of the mean error of two pose estimation algorithms (Fig. 2 (c)).

Results. From Fig. 2 (a), we see that Robot Consistency is better correlated with true model performance than methods relying on inaccurate ground truth annotations. Interestingly, those methods tend to systematically overestimate the model error (points lie under the diagonal), and this effect tends to worsen as true performance improves, i.e., for the high precision settings that are relevant for industrial applications. Our method scales as model performance improves, allowing us to correctly estimate the error of highly

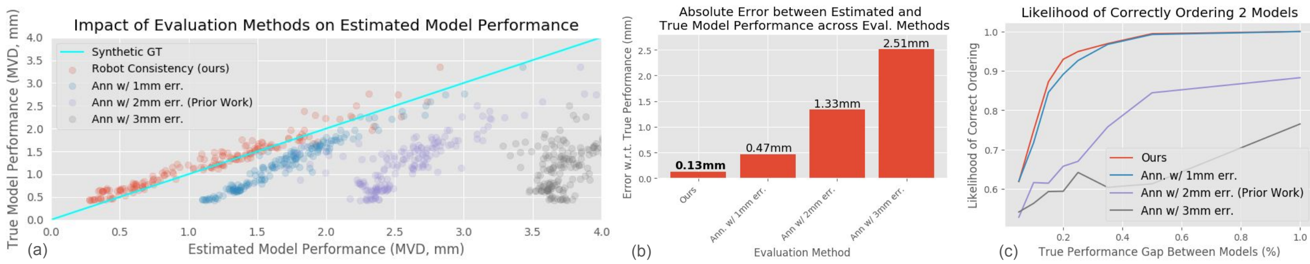

Figure 2. **Our method of evaluating 6DoF poses is much closer to (synthetic) true ground truth than using mm-level human-labeled annotations.** (a) Shows model performance estimates using different evaluation methods against the true model performance from synthetic GT. (b) Shows the absolute difference between model performance estimates from different evaluation methods against the true model performance. (c) Shows the likelihood of ordering two models correctly using different evaluation methods.

accurate algorithms. Fig. 2 (b) shows this favorable behavior also reflected in terms of the mean absolute difference from the true error (0.13 for our method vs. 0.47 mm for 1 mm jitter). Fig. 2 (c) shows that Robot Consistency also has a slight edge in terms of ordering the performance of different pose estimation methods correctly, which is critical for benchmarking. It maintains the highest level of probability among all compared evaluation methods, in particular for high precision settings.

4. Industrial Plenoptic Dataset

In this section we describe our novel Industrial Plenoptic Dataset aimed at providing the basis for the co-design of camera systems, HDR, and 6DoF pose estimation algorithms. It is targeted at high accuracy pose estimation with moderate clutter, as it is typical in industrial applications such as machine tending.

4.1. Setup

Physical scenes. We select 20 industrial objects from the inventory of a mechanical parts vendor that we deem representative of industrial applications (Fig. 3 (d)). They range from metal gears to metal brackets and baskets to highly reflective, black surfaces. The difficulty of estimating accurate 6DoF poses for these objects varies according to which camera, HDR approach, and pose estimation algorithm is being used. Each physical scene is rigidly mounted to a UR5e robot arm that has been kinematically calibrated, resulting in less than 100μm relative pose accuracy.

Capture. For each physical scene we capture 30 robot configurations (corresponding to distinct visual scenes, Fig. 3 (c)). They differ in up to 30 degrees in pitch and roll and 360 degrees in yaw. We also allow the robot to move up to 50cm in Z, and 1m in X and Y, respectively, to represent a large working volume typical of industrial applications.

Our setup includes a total of 13 cameras in order to afford the comparison of pose estimation approaches based on structured light, multi-view key points, and polarization (Fig. 3 (c)): 4 Mono-Polar FLIR Cameras [3] at 5MP resolution with a baseline of 50cm to 1m, 8 Basler RGB cameras [1, 2], 5 at 8MP and 3 at 2MP with baselines varying from 10cm to 1m, and a Photoneo XL [5], which gives $\approx 500\mu$m accurate depth maps at a distance of 2m. Furthermore, each visual scene is captured by FLIR and Basler cameras using four exposures (1ms, 30ms, 80ms, and 200ms) to enable HDR experimentation. For Photoneo, a 12-bit HDR image is captured, allowing for tone mapping exploration.

Lighting. We capture 3 different lighting conditions of varying difficulty for single exposures, some of them posing challenges even to HDR-enabled approaches (Fig. 3 (b)). *Roomlight:* lux level of $1,000$ to $2,000$, offering the friendliest lighting conditions. *Spotlight:* simulates sunlight with stark shadows, creating scenes with extremely bright ($100,000$ lux) and dark (100 lux) regions, which are challenging for HDR. *Daylight:* our robot arm is positioned close to a large window that exposes the scene to directional sunlight with large variability due to changes in weather.

5. Towards co-evaluation for co-design

In this section, we demonstrate the unique property of our novel dataset to enable co-evaluation of cameras, HDR, and pose estimation algorithms to support the co-design that is required for highly precise industrial pose estimation.

To that end, we conduct four different ablation experiments. First (Sec. 5.2, Tab. 2 (a)), we evaluate average pose estimation performance of different camera setups across all 20 parts in our dataset for different lighting conditions. Second (Sec. 5.2, Tab. 2 (b)), we analyze the performance of different cameras on specific parts, some that are challenging in terms of material and geometry. Third (Sec. 5.3, Tab. 2 (c)), we quantify the impact of different HDR variants on performance. Fourth (Sec. 5.4, Tab. 2 (d)), we demonstrate significant performance gains from polarization when applying specialized pose refinement. Lastly

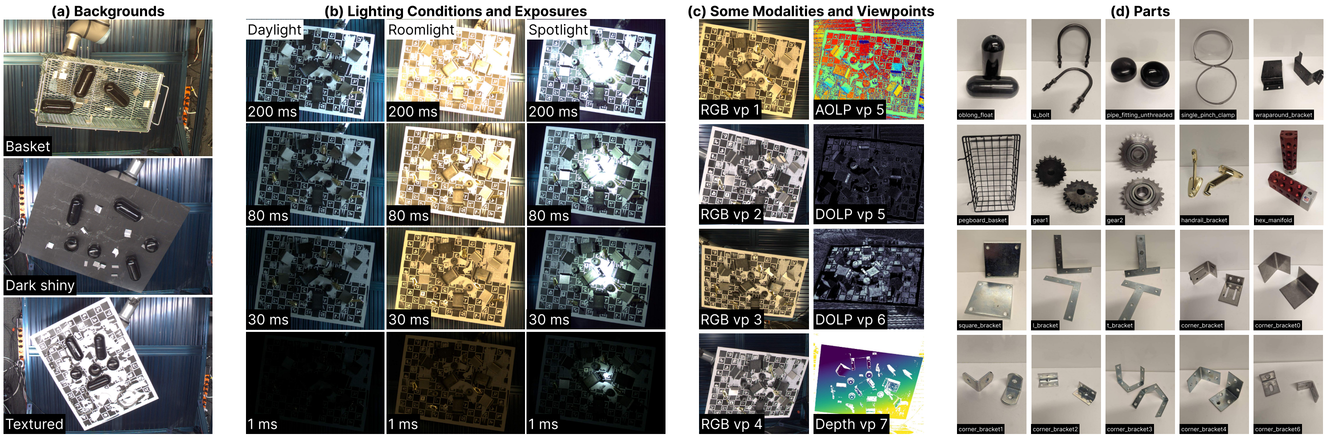

Figure 3. **Industrial Plenoptic Dataset**. Every physical scene is captured (a) against 3 different backgrounds, (b) under 3 different lighting conditions, and with 4 exposures. (c) A subset of the 13 cameras fixed at different viewpoints, representing various modalities (RGB, depth, and polarization). (d) The 20 parts that make up physical scenes in the dataset (row 1: challenging parts).

Camera	Basler LR		Basler HR		FLIR-monoP		Photoneo	
Lighting	MVD	Recall	MVD	Recall	MVD	Recall	MVD	Recall
daylight	5.361	0.66	3.078	**0.83**	**4.808**	**0.70**	6.348	0.60
roomlight	**4.580**	**0.69**	**2.640**	**0.83**	4.875	0.69	**5.676**	**0.63**
spotlight	4.898	0.60	3.061	0.77	5.743	0.62	6.646	0.61

(a) Performance of cameras across lighting conditions.

Part (see Fig. 3d)	Hex Manifold		Pegboard basket		Oblong float		Gear 2	
Camera	MVD	Recall	MVD	Recall	MVD	Recall	MVD	Recall
Basler LR	2.359	0.84	6.280	0.27	4.574	**0.97**	2.580	0.99
Basler HR	1.883	**0.87**	**2.435**	**0.42**	**3.141**	0.95	**1.652**	**1.00**
FLIR-monoP	2.713	0.81	3.681	0.21	4.952	0.85	2.521	0.98
Photoneo	**1.633**	0.57	10.201	0.14	8.128	0.06	6.521	0.93

(b) Performance of cameras on representative parts.

Lighting	HDR Method	Median MVD	Recall	Precision
Spotlight	No HDR	3.271	0.72	0.960
	Debevec [19]	3.300	0.74	0.951
	Robertson [56]	3.351	0.74	0.952
	Mertens [50]	**3.061**	**0.77**	**0.963**

(c) Performance of HDR in challenging light.

Camera	Part (see Fig. 3d)	Modality for Refinement [17]	Median MVD	Recall	Precision
FLIR-MonoP	Pipe Fitting	None	9.92	0.98	1.00
		Gray	9.66	0.98	1.00
		AOLP / DOLP	**8.43**	0.98	1.00
	Oblong Float	None	6.17	0.94	1.00
		Gray	5.91	0.94	1.00
		AOLP / DOLP	**5.41**	0.94	1.00

(d) Performance of Polarization on dark, highly reflective parts.

Table 2. **Our dataset allows for the evaluation of cameras, modalities, HDR, and algorithms for 6DoF pose estimation against a variety of geometries, materials, and lighting.** We show ablations that are only possible using our dataset.

(Sec. 5.5), we highlight particularly challenging parts.

5.1. Experimental Setup

Pose estimation algorithms. The following experiments share a set of pose estimation algorithms as the basis for evaluation. Common across cameras is an object detection step (standard Mask-RCNN [29] pipeline), followed by 2D key point estimation and some mechanism for lifting the 2D key points to 6DoF object poses. This mechanism depends on the camera setup and consists either of some form of PnP (for multi-view images, candidates include [18, 23, 40, 41, 59]) or ICP (for structured light sensors, candidates include [23, 30, 53, 60, 72]). For the former, we found key points from [14] (heatmap regression with a High Resolution Network [66]) to work best. We apply the same key point network to multiple views and combine it with multi-view re-projection error minimization [18], implemented in Open-CV [10] and Ceres [6]. We select candidate key points using the standard farthest point method [53]. For the latter, we use the same key points, but in combination with 3D-3D least-squares optimization [30]. We generate training data with a rendering pipeline using NVISII [51] and generate data similar to BlenderProc [20].

We experiment with both traditional [17] (applied to gray-scale images of any camera) and polarization-aware edge refinement (applied to edges from the angle and degree of linear polarization (AOLP / DOLP), Tab. 2 (d)).

Evaluation. We report median MVD (mm, Sec. 3.1), precision, and recall. True positives are determined as objects within 10mm and 10 degrees of ${}^{ann}T_{GO}$.

5.2. Camera Ablation

In Tab. 2 (a, b) we compare the performance of different cameras for various lighting conditions, object geometries, and materials. We find that high-resolution multi-view (Basler-HR) is the most robust, however in some cases high-resolution structured light (Photoneo) is more accurate.

Lighting conditions. In Tab. 2(a), high-resolution multi-view cameras (Basler-HR) consistently outperform other cameras on average across all lighting conditions and metrics. However, all cameras struggle in spotlight conditions when compared to performance in roomlight. Multi-view RGB cameras face challenges due to dark and bright regions being close together, while Photoneo's use of a red/NIR laser leads to ambient light interference [64] with the high-intensity spotlight. This underscores the need for future research on integrating HDR algorithms and pose estimators to improve performance in challenging lighting conditions.

Object geometry. Tab. 2 (b) gives pose estimation results for different object geometries and materials. As before, Basler-HR has the edge in terms of performance for most parts. *Hex Manifold* is one instance where high-resolution structured light is more accurate as it is a large part with many holes as geometric features. On the contrary, Photoneo has the worst performance on a large part like *Pegboard basket*, because it is made up of thin structures that are detected better with the additional RGB resolution of Basler-HR (Recall 0.42 vs. 0.14 of Photoneo and 0.27 of Basler-LR). Finally, *Gear 2* is a simple part where all cameras achieve high recall. Results for all other parts are available in the supplement.

Object material. Tab. 2 (b) highlights the performance of *Oblong float*, a particularly challenging object with a dark, highly reflective surface, obtaining poor performance among all objects in the table. Structured light (Photoneo) is particularly impacted by the photometric properties (recall of 0.06), as the laser of the structured light scanner will not reflect back into the camera. Multi-view camera setups (Basler-LR, Basler-HR, FLIR) are less impacted by this effect, and polarization can further improve (Tab. 2 (d)).

5.3. HDR Ablation

Tab. 2 (c) compares the performance of three standard HDR algorithms [19, 50, 56] available in OpenCV [10] on the best-performing Basler-HR cameras, in the challenging spotlight lighting condition. We find that Mertens has the best performance. Visually, we also find that it retains most of the contrast in bright lighting (Fig. 5).

5.4. Polarization Ablation

Tab. 2 (d) shows that polarization can significantly improve the quality of 2D edge refinement [17] in dark-on-dark scenes leading to improved MVD. We run this experiment on two difficult objects with dark, highly reflective surfaces in our dataset, *Pipe fitting* and *Oblong float*, and the dark and shiny background category (Fig. 1 (top)). For both objects, edge refinement [17] on the angle and degree of linear polarization edges outperforms the grayscale edges.

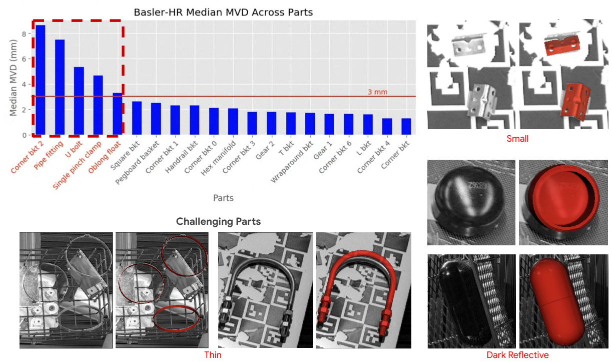

Figure 4. **Our dataset contains industry relevant challenging parts.** We compare median MVD across different parts and show the challenging ones leading to poor pose accuracy.

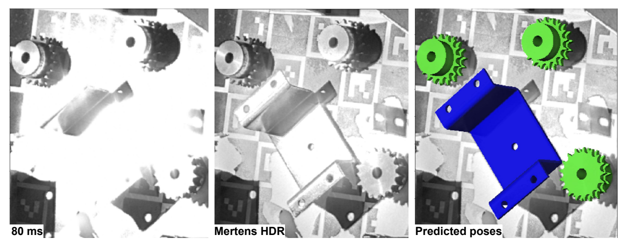

Figure 5. **Mertens HDR (middle) improves lighting robustness compared to a single exposure (left).** Our dataset offers multiple exposures to allow co-design of pose estimation and HDR.

5.5. Challenging Parts

In Fig. 4, we order all 20 parts of our dataset according to the performance of our best-performing camera on average, Basler-HR. We highlight the 5 parts with more than 3mm MVD, and show they all fall into three categories: *thin, small, and dark reflective*. We feel that these three challenges remain in industrial pose estimation, and we believe the co-design of camera-configurations, HDR, and pose estimation algorithms will be important to address them.

6. Conclusions

In this paper, we have introduced a novel Industrial Plenoptic Dataset of 20 parts, paired with a scalable and accurate Robot Consistency evaluation methodology, to enable co-evaluation of camera systems, HDR, and 6DoF pose estimation algorithms. To that end, we have made two important contributions. First, we have validated the evaluation method both in theory and in synthetic experiments. And second, we have highlighted initial directions of co-evaluation based on our novel dataset that we hope will inspire future research and facilitate the step to real-world, high-precision industrial applications.

References

[1] Basler ace 2. Model: a2a2590-22gcbas. https://www.baslerweb.com/en/shop/a2A2590-22gcBAS/, . 3, 6

[2] Basler ace. Model: aca1920-40gc. https://www.baslerweb.com/en/shop/aca1920-40gc/, . 3, 6

[3] Flir Blackfly S USB3. Model: BFS-U3-51S5P-C. https://www.flir.com/products/blackfly-s-usb3/?model=BFS-U3-51S5P-C. 3, 6

[4] Nerian depth error calculator. https://nerian.com/support/calculator/. 3, 5

[5] Photoneo PhoXi 3D Scanner XL. https://www.photoneo.com/products/phoxi-scan-xl/. 3, 6

[6] Sameer Agarwal, Keir Mierle, et al. Ceres solver. 2012. 7

[7] Janis Arents and Modris Greitans. Smart industrial robot control trends, challenges and opportunities within manufacturing. *Applied Sciences*, 12(2):937, 2022. 1

[8] Rabi Bhattacharya and Vic Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds—II. *The Annals of Statistics*, 33(3):1225–1259, 2005. 5

[9] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten Rother. Learning 6d object pose estimation using 3d object coordinates. In *ECCV*, 2014. 1, 2

[10] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 7, 8

[11] Yannick Bukschat and Marcus Vetter. Efficientpose: An efficient, accurate and scalable end-to-end 6d multi object pose estimation approach, 2020. 1

[12] Benjamin Busam, Hyun Jung, and Nassir Navab. I like to move it: 6d pose estimation as an action decision process. *arXiv:2009.12678*, 2020.

[13] Peter Carr, Yaser Sheikh, and Iain Matthews. Monocular object detection using 3d geometric primitives. In *ECCV*, 212. 1

[14] Bo Chen, Alvaro Parra, Jiewei Cao, Nan Li, and Tat-Jun Chin. End-to-end learnable geometric vision by backpropagating pnp optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8109, 2020. 7

[15] Dengsheng Chen, Jun Li1, Zheng Wang, and Kai Xu. Learning canonical shape space for category-level 6d object pose and size estimation. In *CVPR*, 2020. 1

[16] Kai Chen and Qi Dou. Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In *ICCV*, 2021. 1

[17] Changhyun Choi and Henrik I Christensen. 3d textureless object detection and tracking: An edge-based approach. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3877–3884. IEEE, 2012. 5, 7, 8

[18] Alvaro Collet and Siddhartha S Srinivasa. Efficient multiview object recognition and full pose estimation. In *2010 IEEE International Conference on Robotics and Automation*, pages 2050–2055. IEEE, 2010. 7

[19] Paul Debevec, Erik Reinhard, Greg Ward, and Sumanta Pattanaik. High dynamic range imaging. In *ACM SIGGRAPH 2004 Course Notes*, pages 14–es. 2004. 7, 8

[20] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Dmitry Olefir, Tomas Hodan, Youssef Zidan, Mohamad Elbadrawy, Markus Knauer, Harinandan Katam, and Ahsan Lodhi. Blenderproc: Reducing the reality gap with photorealistic rendering. In *International Conference on Robotics: Sciene and Systems, RSS 2020*, 2020. 2, 7

[21] Andreas Doumanoglou, Rigas Kouskouridas, Sotiris Malassiotis, and Tae-Kyun Kim. Recovering 6d object pose and predicting next-best-view in the crowd. In *CVPR*, 2016. 1, 2

[22] Bertram Drost, Markus Ulrich, Paul Bergmann, and Philipp Hartinger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. In *ICCV Workshop*, 2017. 1, 2, 3

[23] Fabian Duffhauss, Tobias Demmler, and Gerhard Neumann. Mv6d: Multi-view 6d pose estimation on rgb-d frames using a deep point-wise voting network. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3568–3575. IEEE, 2022. 7

[24] Benjamin Eltzner and Stephan F Huckemann. A smeary central limit theorem for manifolds with application to high-dimensional spheres. *The Annals of Statistics*, 47(6):3360–3381, 2019. 5

[25] Hao-Shu Fang, Chenxi Wang, Minghao Gou, and Cewu Lu. Graspnet-1billion: A large-scale benchmark for general object grasping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11444–11453, 2020. 2

[26] Yang Fu and Xiaolong Wang. Category-level 6d object pose estimation in the wild: A semi-supervised learning approach and a new dataset. *arXiv:2206.15436*, 2022. 1, 2

[27] Daoyi Gao, Yitong Li, Patrick Ruhkamp, Iuliia Skobleva, Magdalena Wysocki, HyunJun Jung, Pengyuan Wang, Arturo Guridi, and Benjamin Busam. Polarimetric pose prediction. In *European Conference on Computer Vision*, pages 735–752. Springer, 2022. 2

[28] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation averaging. *IJCV*, 103:267–305, 2013. 4

[29] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 7

[30] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020. 7

[31] Stefan Hinterstoisser, Stefan Holzer, Cedric Cagniart, Slobodan Ilic, Kurt Konolige, Nassir Navab, and Vincent Lepetit. Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes. In *2011 International Conference on Computer Vision*, pages 858–865, 2011. 1, 2, 4

[32] Tomáš Hodaň, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-LESS: An RGB-D dataset for 6D pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017. 1, 2, 3

[33] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders Glent Buch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, Caner Sahin, Fabian Manhardt, Federico Tombari, Tae-Kyun Kim, Jiri Matas, and Carsten Rother. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018. 2, 3

[34] HyunJun Jung, Shun-Cheng Wu, Patrick Ruhkamp, Hannah Schieber, Pengyuan Wang, Giulia Rizzoli, Hongcheng Zhao, Sven Meier, Daniel Roth, Nassir Navab, and Benjamin Busam. Housecat6d – a large-scale multi-modal category level 6d object pose dataset with household objects in realistic scenarios. *arXiv:2212.10428v4*, 2023. 1, 2, 3

[35] Agastya Kalra, Achuta Kadambi, and Kartik Venkataraman. Systems and methods for characterizing object pose detection and measurement systems. In *US patent no. 11580667*, 2020. 2, 3

[36] Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8602–8611, 2020. 2

[37] Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *ICCV Workshop*, 2019. 1, 2, 3

[38] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017. 1

[39] Kilian Kleeberger, Christian Landgraf, and Marco F Huber. Large-scale 6d object pose estimation dataset for industrial bin-picking. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2573–2578. IEEE, 2019. 2

[40] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16*, pages 574–591. Springer, 2020. 7

[41] Alan Li and Angela P Schoellig. Multi-view keypoints for reliable 6d object pose estimation. *arXiv preprint arXiv:2303.16833*, 2023. 7

[42] Jiehong Lin, Zewei Wei, Zhihao Li, Songcen Xu, Kui Jia, and Yuanqing Li. Dualposenet: Category-level 6d object pose and size estimation using dual pose network with refined learning of pose consistency. *CoRR*, abs/2103.06526, 2021. 1

[43] Sanneman Lindsay and Julie Fourie. The state of industrial robotics: Emerging technologies, challenges, and key research directions. 2020. 1

[44] Xingyu Liu, Rico Jonschkowski, Anelia Angelova, and Kurt Konolige. Keypose: Multi-view 3d labeling and keypoint estimation for transparent objects. In *CVPR*, 2020. 1, 2, 3

[45] Xingyu Liu, Shun Iwase, and Kris M. Kitani. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In *ICCV*, 2021. 1, 2, 3

[46] Michel Loeve. *Probability theory 1 (4th ed.)*. Springer, 1977. 4

[47] Fabian Manhardt, Manuel Nickel, Sven Meier, Luca Minciullo, and Nassir Navab. CPS: class-level 6d pose and shape estimation from monocular images. *CoRR*, abs/2003.05848, 2020. 1

[48] Pat Marion, Peter R. Florence, Lucas Manuelli, and Russ Tedrake. Labelfusion: A pipeline for generating ground truth labels for real rgbd data of cluttered scenes. In *ICRA*, 2018. 1, 2

[49] Jeremy A. Marvel, Joe Falco, and Tsai Hong. Ground truth for evaluating 6 degrees of freedom pose estimation systems. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems*, page 69–74, New York, NY, USA, 2012. Association for Computing Machinery. 3

[50] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum*, pages 161–171. Wiley Online Library, 2009. 7, 8

[51] Nathan Morrical, Jonathan Tremblay, Yunzhi Lin, Stephen Tyree, Stan Birchfield, Valerio Pascucci, and Ingo Wald. Nvisii: A scriptable tool for photorealistic image generation. *arXiv preprint arXiv:2105.13962*, 2021. 7

[52] Keunhong Park, Arsalan Mousavian, Yu Xiang, and Dieter Fox. Latentfusion: End-to-end differentiable reconstruction and rendering for unseen object pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[53] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4561–4570, 2019. 7

[54] Cody J. Phillips, Matthieu Lecce, and Kostas Daniilidis. Seeing glassware: from edge detection to pose estimation and shape recovery. In *RSS*, 2016. 1

[55] Colin Rennie, Rahul Shome, Kostas E Bekris, and Alberto F De Souza. A dataset for improved rgbd-based object detection and pose estimation for warehouse pick-and-place. *IEEE Robotics and Automation Letters*, 1(2):1179–1185, 2016. 1, 2

[56] Mark A Robertson, Sean Borman, and Robert L Stevenson. Dynamic range improvement through multiple exposures. In *Proceedings 1999 international conference on image processing (Cat. 99CH36348)*, pages 159–163. IEEE, 1999. 7, 8

[57] Peter De Roovere, Steven Moonen, Nick Michiels, and Francis Wyffels. Dataset of industrial metal objects. *arXiv:2208.04052v1*, 2022. 2, 3

[58] Tushar Sandhan and Jin Young Choi. Anti-glare: Tightly constrained optimization for eyeglass reflection removal. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1241–1250, 2017. 2

[59] Ivan Shugurov, Ivan Pavlov, Sergey Zakharov, and Slobodan Ilic. Multi-view object pose refinement with differentiable renderer. *IEEE Robotics and Automation Letters*, 6(2):2579–2586, 2021. 7

[60] Chen Song, Jiaru Song, and Qixing Huang. Hybridpose: 6d object pose estimation under hybrid representations. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 431–440, 2020. 7

[61] Alykhan Tejani, Danhang Tang, Rigas Kouskouridas, and Tae-Kyun Kim. Latent-class hough forests for 3d object detection and pose estimation. In *ECCV*, 2014. 1, 2, 3

[62] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018. 2

[63] Stephen Tyree, Jonathan Tremblay, Thang To, Jia Cheng, Terry Mosier, Jeffrey Smith, and Stan Birchfield. 6-dof pose estimation of household objects for robotic manipulation: An accessible dataset and benchmark. In *International Conference on Intelligent Robots and Systems (IROS)*, 2022. 1, 2, 3

[64] Sophie Voisin, Sebti Foufou, Frédéric Truchetet, David Page, and Mongi Abidi. Study of ambient light influence for three-dimensional scanners based on structured light. *Optical Engineering*, 46(3):030502–030502, 2007. 8

[65] He Wang, Srinath Sridhar, Jingwei Huang, Julien Valentin, Shuran Song, and Leonidas J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *CVPR*, 2019. 1, 2

[66] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *TPAMI*, 2019. 7

[67] Pengyuan Wang, HyunJun Jung, Yitong Li, Siyuan Shen, Rahul Parthasarathy Srikanth, Lorenzo Garattoni, Sven Meier, Nassir Navab, and Benjamin Busam. Phocal: A multi-modal dataset for category-level object pose estimation with photometrically challenging objects. In *CVPR*, 2022. 1, 2, 3

[68] Lawrence B Wolff. Polarization vision: a new sensory approach to image understanding. *Image and Vision computing*, 15(2):81–93, 1997. 2

[69] Xuesong Wu, Hong Zhang, Xiaoping Hu, Moein Shakeri, Chen Fan, and Juiwen Ting. Hdr reconstruction based on the polarization camera. *IEEE Robotics and Automation Letters*, 5(4):5113–5119, 2020. 2

[70] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *Robotics: Science and Systems (RSS)*, 2018. 1, 2

[71] Jun Yang, Yizhou Gao, Dong Li, and Steven L. Waslander. Robi: A multi-view dataset for reflective objects in robotic bin-picking. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9788–9795, 2021. 1, 2, 3

[72] Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Dpod: 6d pose object detector and refiner. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1941–1950, 2019. 1, 7

[73] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4786–4794, 2018. 2

[74] Chu Zhou, Yufei Han, Minggui Teng, Jin Han, Si Li, Chao Xu, and Boxin Shi. Polarization guided hdr reconstruction via pixel-wise depolarization. *IEEE Transactions on Image Processing*, 32:1774–1787, 2023. 2