2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Introducing CEA-IMSOLD: an Industrial Multi-Scale Object Localization Dataset

Boris Meden[1,2], Pablo Vega[1], Fabrice Mayran de Chamisso[1] and Steve Bourgeois[1]

***Abstract*— We introduce the CEA Industrial Multi-Scale Object Localization Dataset (CEA-IMSOLD), a new BOP format dataset for 6-DoF object localization, crucial for robotics. This dataset aims to evaluate the current localization methods with respect to a new difficulty: large variations in observation distance and, consequently, large variations in image appearance. Compared to the other publicly available datasets, our dataset provides both images with objects small and completely visible in the image, and images where objects are observed close enough so they appear larger than the field of view of the camera. We also propose to consider the observation distance in the evaluation process and introduce new metrics to do so. Finally, our dataset contains a large variety of industrial objects, from small and simple objects such as bolts to sizable and complex ones such as large car parts. We provide baseline results and the dataset is made publicly available to support the community at https://cea-list.github.io/CEA-IMSOLD/.**

I. INTRODUCTION

Object 6D localization is a task required for several applications, such as interactive robotics (*e.g.* object manipulation), quality inspection, augmented reality, *etc*. Depending on the application, observation conditions can vary drastically, from isolated objects observed in a controlled setting by a robot-mounted camera to symmetrical objects subjected to occlusions and observed with a hand-held camera with uncontrolled trajectory. This variety of conditions makes 6D localization algorithms benchmarking complex, simply due to the sheer difficulty of creating a dataset that covers all possible situations/use cases. Consequently, datasets are usually biased due to the setup used for their constitution (sensors quality and type, objects nature, lighting set-up, viewpoint variability, ground truth quality, *etc*.). It is then crucial to evaluate methods on a collection of datasets covering different set-ups in order to get an overall assessment. Indeed, a biased benchmark can lead to biased evaluations, but ultimately it can also bias a research field since new methods are more likely to be published if they perform well on established benchmarks, hence the need for benchmarks with variability.

The BOP Challenge [1] was created to reach this goal, progressively including new and diverse datasets to cover more conditions and to provide better algorithm performance evaluations. Our work follows this approach since we provide a new public dataset that covers aspects that were until now underconsidered:

1) Camera to object distance variation: as illustrated in Fig. 1, our dataset includes large observation distance variations while, as illustrated in Fig. 3, this variation is limited in BOP challenge datasets. This dataset is then more suitable for use cases inluding such varations (inspection, augmented reality, mobile robotics, *etc*.).
2) Depth sensor quality variety: as [2] our dataset includes depth images provided by both a high-end sensor and a low-cost sensor, covering both industrial and consumer applications.
3) Variation of complexity of objects: our dataset includes both simple (*e.g.* bottle cap), and complex objects (*e.g.* car engine), covering use cases such as pick- and-place and quality inspection.
4) High accuracy ground-truth: the annotation process based on robotic arm and high-end depth sensors allows to reach high accuracy ground-truth.
5) A new metric taking into account distance to objects in performance evaluation.

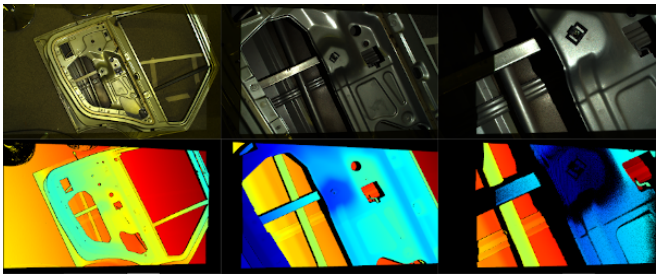


Fig. 1: Varying the distance of observation implies image scale variation, a challenge for object pose estimation.

II. RELATED WORK

A. Datasets and Bias

LINEMOD [3], one the first publicly-available datasets for object localization, is constituted of objects observed from a collection of viewpoints that can be approximated by a sphere. While the dataset includes cluttered objects to distract the evaluated algorithms, said objects are daily-life low complexity ones, mostly textureless, with uniform and specific color, never occluded (or to a very limited extent), observed in constant lighting conditions, with viewpoints mostly located on a hemisphere, and with a single object instance per scene. These specificities representing many biases, additional datasets were developed to target them. Objects occlusions have been added with LINEMOD Occluded [4] and IC-BIN [5]. To evaluate the ability to localize objects based on their shape exclusively, and the robustness to multiple instances of the same object, T-LESS [6] and

[1]All authors are with Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France [2]boris.meden@cea.fr

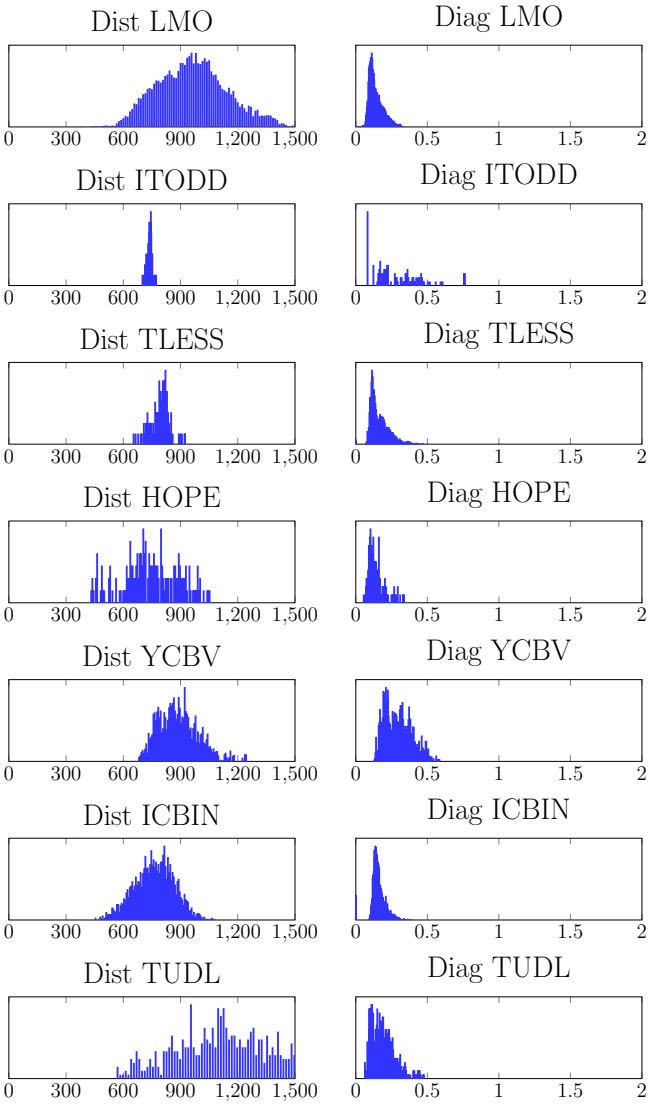


Fig. 2: Distributions of objects distances to the camera (Dist) expressed in mm (from 40cm to 1.5m) and objects scales in the image (Diag), expressed as the diagonal of object bounding box over the diagonal of the image, for BOP datasets. We note that objects scales are mainly below 0.25.

ITODD [2] were introduced, the former considering textureless white plastic objects while the latter considering metallic industrial ones. Finally, lighting conditions changes were introduced in TUD-Light, Toyota Light [1] and HOPE [7].

Among the biases that are not yet tackled by the community, and at the origin of this new dataset, is the very limited object observation distance, and consequently the very limited object size variation in the image. As illustrated in Fig. 2 and Fig. 3, we measured the distance of observation and the ratio between the object 2D bounding box diagonal length and the image diagonal length. These calculations were performed for each object of each scene of each dataset in the BOP Challenge. While the first one provides an overview of the variation of observation distance, the second one provides an overview of the object size variation in the

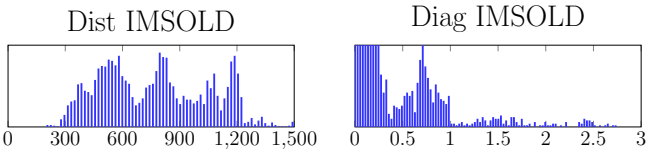


Fig. 3: Distributions of objects distances to the camera (Dist) expressed in mm (from 40cm to 1.5m) and objects scales in the image (Diag), expressed as the diagonal of object bounding box over the diagonal of the image, for IMSOLD. We note that objects scales go up to 2.6, which means that less than half of the object represents the whole image.

image. We notice that observation distance and object size variations are limited in most datasets and that object size in the image is always smaller than the field of view of the sensor. Based on these observations, one can question the correlation between this bias and the fact that most recent RGB-only object 6D pose estimation approaches rely almost always exclusively on deep learning methods with 2D bounding box detection or 2D object segmentation as first stage. While bounding box detection and 2D segmentation are adapted for small objects in the image, their working hypotheses might be challenged by larger-than-field-of-view objects.

B. Datasets and Ground-truth Annotations

Ground-truth image annotations are essential since they directly affect evaluation quality. Usually, annotations rely on three steps:

1) 3D model object reconstruction.
2) Camera pose estimation w.r.t. world coordinate frame, for each dataset image.
3) Object pose estimation w.r.t. world coordinate frame, for each dataset scene.

Regarding 3D model quality, it can vary from one dataset to another. Indeed, many datasets [3], [4] [6] rely on a volumetric 3D reconstruction process [8], [9] with consumer-grade RGB-D sensors, or manual modeling of CAD models [6]. Such processes provide 3D reconstructions with very limited accuracy. Therefore, more recent datasets decided to use high-end 3D scanners [10], [7], [11] to provide high quality 3D models.

Regarding camera pose estimation w.r.t. the world coordinate frame, most datasets rely on visual pose estimation based on markers [3], [4], [6], [5], [12]. While this method is simple, pose accuracy remains limited. More recent datasets suggest to use a robotic arm to provide an accurate trajectory of the camera [11]. Such a method is more expensive and complex since accurate robotic arms are expensive and hand-eye calibration is needed.

With respect to 3D object registration in world coordinates, a classic solution consists in using a consumer-grade RGB-D camera and a volumetric reconstruction algorithm [8], [9] to get a 3D model of the whole scene [3], [4], [6], [10]. Each object is registered to the scene either manually [6] or automatically [10] with the help of an object

localization method followed by ICP refinement. However, due to the limited quality of the sensor used for the 3D scene mapping, the resulting registration quality remains limted. To go a step further in terms of accuracy, recent datasets such as [11] use a robotic arm with a tip effector to measure scene objects specific 3D points in order to recover their 3D pose.

To obtain an accurate ground-truth annotation, as detailed in section III-A, we followed an approach similar to [11], using a high-end scanner or native CAD for object model creation, a robotic arm for camera localization, and a high-end depth sensor for automatic and accurate object registration in the scene.

C. Datasets and Evaluation Metrics

Currently, the most commonly used metrics to qualify pose estimation are:

1) Average Distance metric or ADD, and its variations for symmetrical objects (ADD-S [13], ADD-H [7]). These metrics target to measure the mean euclidean error between the surface points of the object transformed with the estimated pose and the ground truth pose.
2) Maximum Symmetry-Aware Surface Distance or MSSD [14] which is similar to ADD but considers the maximal error instead of the mean error and symmetry management.
3) Visual Surface Discrepancy or VSD [1]. This metric measures the ratio of pixels for which the discrepency between their value in the distance map computed with the ground truth pose and their values in the distance map computed with the pose to evaluate.
4) Mean Projection Distance or MPD [4], which measures the mean reprojection error between the projection of the visible points with the ground-truth pose and their corresponding projection with the pose to evaluate.
5) Maximum Symmetry-Aware Projection Distance or MSPD [14]. This metric similar to MPD but uses maximal distance instead of the mean and manages symmetries.

Since MPD and MSPD are based on reprojection error, they do not take into account the discrepancy along the optical axis. This property makes these metrics adapted to RGB-only setups since such sensors do not observe the position along the optical axis. On the other hand, the ADDs and MSSD metrics consider euclidean distance in 3D space while VSD rely exclusively on the discrepancy along the Z axis. In both cases, these metrics do not take into account observation distance even though depth sensor accuracy is usually affected by this distance. Finally, all these metrics are computationally expensive since they iterate on all 3D points of the model. In section IV), we suggest a new method to efficiently compute an approximation of MPD metric.

III. Dataset Acquisition Pipeline

The following section details the hardware setup and the process to create our dataset.

A. Objects and their 3D Models

The dataset is composed of 25 objects (cf. fig 4) with different characteristics in terms of surface (reflecting *vs*. lambertian), symmetries (with *vs*. without full rotational symmetry), complexity (primitive shapes *vs*. complex objects), flatness (flat *vs*. voluminous), details (with *vs*. without very fine details on their surface), compactness (long *vs*. compact), and size (diameters from 23 mm to 1278 mm).

Reagrding object 3D models, native CAD models were used for the small objects. Said CAD models were provided by manufacturers or distributors on their websites. For big and complex objects, such as the car door or car engine, an industrial- grade scanner (HandyScan 3D from Creaform) was used to reconstruct the 3D meshes with high precision and accuracy.

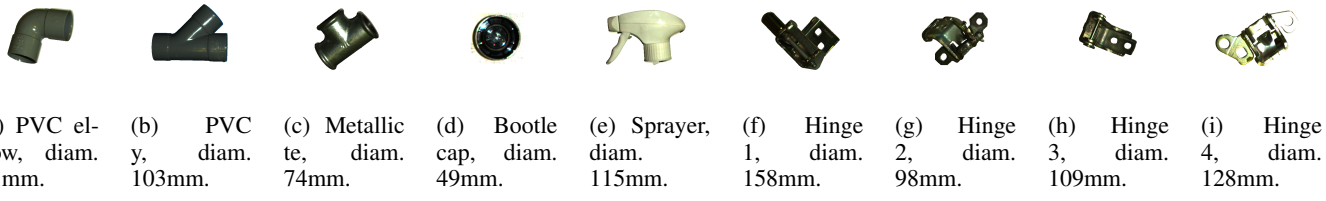
B. Scene acquisitions

Each scene was acquired with a calibrated setup consisting of an industrial grade structured light RGB-D sensor (Zivid 2, with spatial resolution of 0.39mm at 700mm), a consumer- grade active stereo RGB-D sensor (Realsense D415) and an industrial RGB camera (FLIR camera with resolution 2048x1536), all three rigidly mounted on a 6 DOF robotic manipulator (UR10e from Universal Robots, with a position repeatability of 0.05 mm). Each scene was acquired from multiple viewpoints and with large observation distance variations as illustrated in Fig. 5. The robotic arm is first manually operated in gravity compensation mode, and the joints positions are recorded at a given time step. The trajectory is then replayed by the robot, making a pause at each recorded location to capture the sensors data in static position, avoiding any risk of motion blur or sensor synchronization problems. At each location, sensor data such as the location and orientation of the robot effector is recorded. Each sensor was first calibrated intrinsically. The extrinsic calibrations with respect to the robotic manipulator were then achieved with a hand-eye calibration process using MoveIt and ChAruCo markers.

C. Ground truth annotation

Regarding sensor pose, the ground truth was simply computed using the robotic effector pose (recorded during sensor acquisitions) and the corresponding hand-eye calibration. The resulting calculation gives the camera pose estimation in the robotic manipulator coordinate frame. For object ground-truth pose determination, the process relied first on an accurate 3D reconstruction of the whole scene in the robotic manipulator coordinate frame. This was achieved by fusing multiple viewpoint acquisitions made with the Zivid2 sensor. Each object was then automatically detected and localized in the robotic manipulator coordinate frame with [15]. The pose was then refined with an ICP refinement on the scene reconstruction and the registration quality was validated by a human operator.

2D object masks for each view were obtained by configuring a virtual camera with the corresponding intrinsic and extrinsic parameters, and rendering the 3D model of the

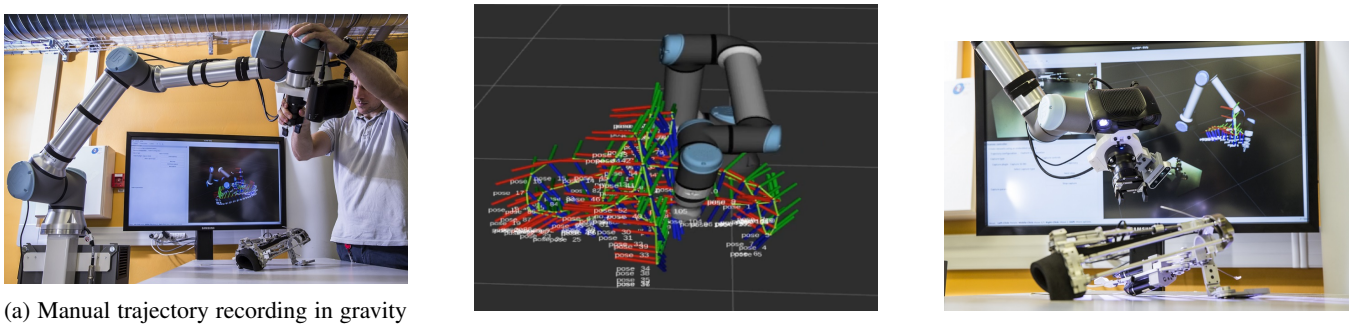


(a) PVC elbow, diam. 91mm. (b) PVC y, diam. 103mm. (c) Metallic te, diam. 74mm. (d) Bootle cap, diam. 49mm. (e) Sprayer, diam. 115mm. (f) Hinge 1, diam. 158mm. (g) Hinge 2, diam. 98mm. (h) Hinge 3, diam. 109mm. (i) Hinge 4, diam. 128mm.

(j) Cylinder 133 63, diam. 146mm. (k) Cylinder 14 78, diam. 79mm. (l) Cylinder 26 106, diam. 108mm. (m) Small spacer, diam. 23mm. (n) Axis, diam. 67mm. (o) Big gear, diam. 86mm. (p) Small gear, diam. 66mm. (q) Nut, diam. 26mm. (r) Spacer, diam. 24mm.

(s) Plaque, diam. 173mm. (t) Cylinder head, diam. 473mm. (u) Car door, diam. 1278mm. (v) Bypass, diam. 693mm. (w) Car seat, diam. 1205mm. (x) Car engine, diam. 904mm. (y) Rail, diam. 302mm.

Fig. 4: Images of the 25 objects used in the dataset.



(a) Manual trajectory recording in gravity compensated mode with timestep sampling.

(b) Trajectory generated in RViz environment.

(c) Replay of the trajectory with data acquisition at each step.

Fig. 5: Pipeline for scene acquisition.

target object at its location. To get occlusion-aware masks, all the other objects of the scene were rendered to the z-buffer the considered object. As [6], [10], Table I evaluates the quality of pose annotations by comparing real and rendered object depth maps.

Dataset	μ_δ	σ_δ	$\mu_{\lvert\delta\rvert}$	$med_{\lvert\delta\rvert}$
T-LESS [6]	0.6300	0.8224	0.87197	0.9191
HomebrewedDB [10]	0.61593	0.8089	0.8564	0.9070
Ours Zivid	0.5678	0.7476	0.8258	0.8762

TABLE I: Statistics of differences between the depth of object models at the ground truth poses and the captured depth (in mm). μ_δ and σ_δ is the mean and the standard deviation of the differences, $\mu_{|\delta|}$ and $med_{|\delta|}$ is the mean and the median of the absolute differences.

IV. Evaluation metrics

A. Metric approximation for fast evaluation

Performance assessment of a 3D pose estimation algorithm is usually obtained with the recall of the localization, meaning the ratio of localizations with an error below a specific threshold. Therefore, if the evaluation is performed with a small threshold on the metric error, which is usually the case considering the current performances of the algorithm, the metric used only needs to be accurate for a pose located in the neighborhood of the ground truth to provide a fair evaluation. However, current metrics are currently computed accurately even for poses far from the ground truth, implying long and useless computation time. To reduce this computation time while keeping a fair evaluation, we suggest to approximate the metric around the ground truth pose with its second order approximation.

We want to back-propagate the uncertainty of the observa-

tion (the measures) to the pose. We follow the formulation of [16]: let $\mathbf{p}_i = g(\mathbf{q}_i, \mathbf{x})$ be the function that transforms 3D CAD points $\mathbf{q}_i$ into measurement points $\mathbf{p}_i$ (either image pixels or 3D points) at pose $\mathbf{x}$, assuming independence between measure points and a noise distribution given by covariance matrix $\mathbf{C_P}$.

$$\mathbf{p}_i + \Delta\mathbf{p}_i = g(\mathbf{q}_i, \mathbf{x}_{est} + \Delta\mathbf{x}) \approx g(\mathbf{q}_i, \mathbf{x}_{est}) + \left[\frac{\partial g}{\partial \mathbf{x}}\right]^{\top}_{\mathbf{q}_i, \mathbf{x}_{est}} \cdot \Delta\mathbf{x} \quad (1)$$

This simplifies to:

$$\mathbf{p}_i \approx g(\mathbf{q}_i, \mathbf{x}_{est}) \implies \Delta\mathbf{p}_i = \left[\frac{\partial g}{\partial \mathbf{x}}\right]^{\top}_{\mathbf{q}_i, \mathbf{x}_{est}} \cdot \Delta\mathbf{x} = \mathbf{J}_i \Delta\mathbf{x} \quad (2)$$

where $\mathbf{J}_i$ is the Jacobian of g, evaluated at $(\mathbf{q}_i, \mathbf{x}_{est})$. Combining all measurement points, we obtain:

$$\begin{pmatrix} \Delta\mathbf{P}_1 \\ \vdots \\ \Delta\mathbf{P}_n \end{pmatrix} = \begin{pmatrix} \mathbf{J}_1 \\ \vdots \\ \mathbf{J}_n \end{pmatrix} \Delta\mathbf{x} \implies \Delta\mathbf{P} = \mathbf{J}\Delta\mathbf{x} \quad (3)$$

$$\Delta\mathbf{x} = (\mathbf{J}^{\top}\mathbf{J})^{-1}\mathbf{J}^{\top}\Delta\mathbf{P} \quad (4)$$

The covariance matrix of $\mathbf{x}$ is given by the outer product :

$$\mathbf{C_x} = (\mathbf{J}^{\top}\mathbf{J})^{-1}\mathbf{J}^{\top} E(\Delta\mathbf{P}\Delta\mathbf{P}^{\top})((\mathbf{J}^{\top}\mathbf{J})^{-1}\mathbf{J}^{\top})^{\top} \quad (5)$$

$$\mathbf{C_x} = \mathbf{J}^{-1} \begin{pmatrix} \mathbf{C_P} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \mathbf{C_P} \end{pmatrix} (\mathbf{J}^{-1})^{\top} \quad (6)$$

where $E(\Delta\mathbf{P}\Delta\mathbf{P}^{\top})$ is the covariance matrix on the measure points. We consider the g function that maps the CAD model and the image.

$$g(\mathbf{q}_i, \mathbf{x}) = \Pi(\mathrm{K} P_{world2cam} Q) \quad (7)$$

where $\Pi : (x, y, z) \to (\frac{x}{z}, \frac{y}{z})$ is the perspective projection, K is the intrinsic camera matrix, $P_{world2cam}$ transforms points from the world to points in the camera coordinate frame and Q is a 3D point in the world coordinate frame. The $\mathbf{C_P}$, expressing noise level on data points becomes a 2×2 matrix, expressed in pixels.

The inverse of the Covariance Matrix (also named Information Matrix) corresponds to the second order approximation of the mean squared of the reprojection error. We can then approximate the Mean SPD from the pose deviation $delta_{pose}$ from the ground truth with our new pose metric e_{cov}, defined as the Mahalanobis distance between the pose estimate and the ground truth pose, ponderated by the inverse of the Covariance Matrix of the ground truth pose.

$$e_{\mathrm{cov}}(\mathbf{X}_{est}, \mathbf{X}_{gt}) = \sqrt{(\mathbf{X}_{est} - \mathbf{X}_{gt})^{\top} {\mathbf{C_x}}^{-1} (\mathbf{X}_{est} - \mathbf{X}_{gt})} \quad (8)$$

where $\mathbf{X}_i$ are the SE(3) poses, and $\mathbf{C_x}$ is the covariance matrix computed with equation 6 around the ground truth pose $\mathbf{X}_{gt}$.

Assuming the Covariance Matrices computed, based on the ground truth pose, this new metric is far more efficient than the different BOP metrics, which need either to iterate over all object 3D points or to render it to qualify the quality of $\mathbf{X}_{est}$.

B. Comparison to BOP Metrics

To assess the quality of the e_{cov} metric, we plot its evolution for a constant 3D error on the pose, with the object being at different distances. We compare the evolution of the error with the other projection errors MSPD (MaxSPD) and MPD (MeanSPD) in Fig. 6.

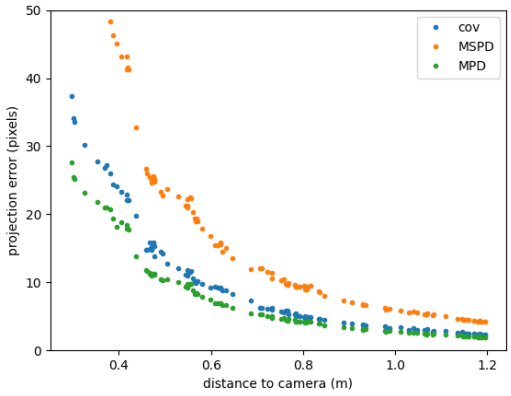


Fig. 6: Projection errors for a constant 3D error on the pose, as a function of different observation distances from 20cm to 1.2m. We note that e_{cov} (cov) is a good approximation of MPD.

V. Benchmark

A. Evaluation of the Metric on LMO

First, we chose the LMO dataset to test the new e_{cov} metric. Table II presents the processing of top BOP contenders results, evaluated with e_{cov}.

Methods	$\mathbf{AR_{VSD}}$	$\mathbf{AR_{MSSD}}$	$\mathbf{AR_{MSPD}}$	$\mathbf{AR_{Cov}}$
GDRNPP [17], [18]	0.6300	0.8224	0.87197	0.9191
Surfemb [19]	0.61593	0.8089	0.8564	0.9070
Cosypose [20]	0.5678	0.7476	0.8258	0.8762

TABLE II: Some of the top BOP contenders (2022 edition) evaluated through the BOP metrics and the new covariance metric on the LMO dataset.

As discussed in previous section, we note that $\mathbf{AR_{MSPD}}$ and $\mathbf{AR_{Cov}}$ are close, yet e_{MSPD} complexity is $O(n)$ and e_{cov} is $O(1)$ as it can be precomputed, with n the number of points of the CAD model. Thus, $\mathbf{AR_{Cov}}$ is much quicker to compute.

B. Baseline Evaluations of IMSOLD

To provide first baseline results on our new dataset, we chose the simplicity with Depth-based registration methods, less affected by scale changes, evaluated on the Zivid2 data. Average Recalls of Table III leave room for new registration methods to address the challenge of scale variation for 6D object pose estimation.

Finally, Fig. 7 provides some examples of the challenges proposed by IMSOLD.

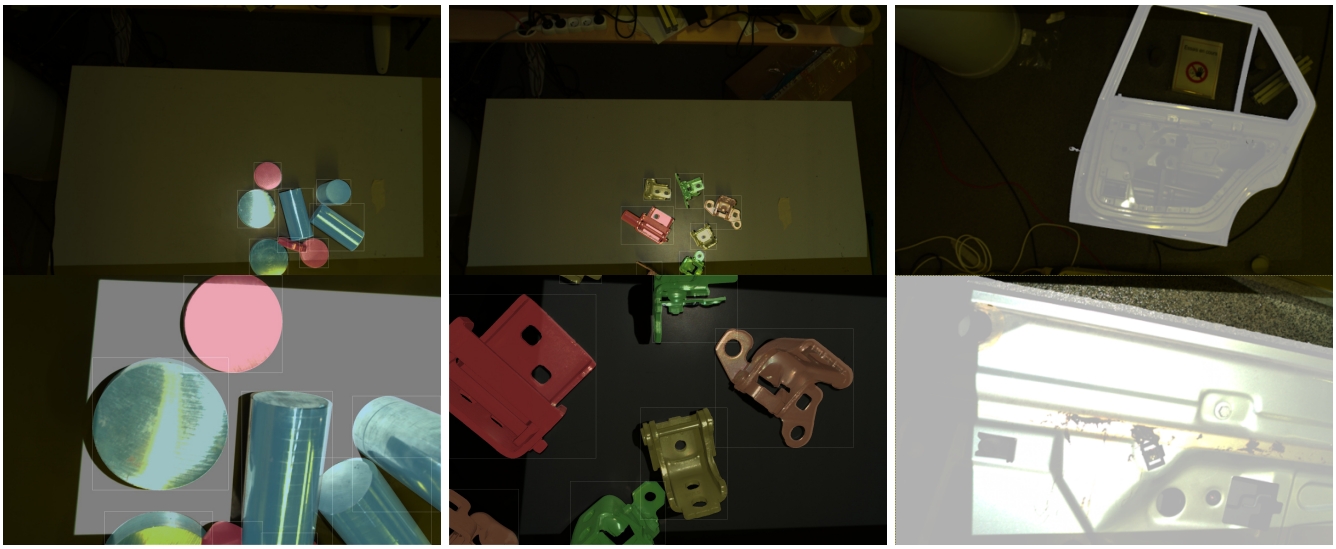


Fig. 7: Examples of ground truth annotations for a variety of objects constituting IMSOLD. Challenges lie in objects geometry, material and appearance, and mainly scale variation in the image.

Methods	$\mathbf{AR_{VSD}}$	$\mathbf{AR_{MSSD}}$	$\mathbf{AR_{MSPD}}$	$\mathbf{AR_{Cov}}$
HSPA[15]	0.4111	0.3039	0.3145	0.4283
Halcon23[21]	0.3494	0.2852	0.3052	0.4036

TABLE III: Average recall of the BOP metrics and $\mathbf{AR_{Cov}}$ for the IMSOLD test set of the Zivid2 data.

VI. CONCLUSION

We have introduced the CEA Industrial Multi-Scale Object Localization Dataset (CEA-IMSOLD) for 3D object detection and localization. We focused on industrial challenging objects, with a specific attention to variation of objects scales. The dataset provides both industrial grade 3D sensing and kinect-like quality. We hope that this dataset will successfully complement existing ones, and will encourage the community to consider industrial challenges during the design and development of new methods.

ACKNOWLEDGMENT

The authors would like to thank Bertram Drost for providing Halcon scripts for baseline evaluation, as well as valuable discussions.

REFERENCES

[1] T. Hodaň, F. Michel, E. Brachmann, W. Kehl, A. G. Buch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T.-K. Kim, J. Matas, and C. Rother, "BOP: Benchmark for 6D Object Pose Estimation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, Sept. 2018, pp. 19–35.

[2] B. Drost, M. Ulrich, P. Bergmann, P. Härtinger, and C. Steger, "Introducing MVTec ITODD - A Dataset for 3D Object Recognition in Industry," in *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, Oct. 2017, pp. 2200–2208.

[3] S. Hinterstoisser, V. Lepetit, S. Ilic, S. Holzer, G. Bradski, K. Konolige, and N. Navab, "Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes," in *Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11*. Springer, 2013, pp. 548–562.

[4] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother, "Learning 6D Object Pose Estimation Using 3D Object Coordinates," in *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, Sept. 2014, pp. 536–551.

[5] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, "Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA: IEEE, June 2016, pp. 3583–3592.

[6] T. Hodaň, P. Haluza, Š. Obdržálek, J. Matas, M. Lourakis, and X. 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.

[7] S. Tyree, J. Tremblay, T. To, J. Cheng, T. Mosier, J. Smith, and S. Birchfield, "6-DoF Pose Estimation of Household Objects for Robotic Manipulation: An Accessible Dataset and Benchmark," Dec. 2022.

[8] S. Izadi, R. A. Newcombe, D. Kim, O. Hilliges, D. Molyneaux, S. Hodges, P. Kohli, J. Shotton, A. J. Davison, and A. Fitzgibbon, "KinectFusion: Real-time dynamic 3D surface reconstruction and interaction," in *ACM SIGGRAPH 2011 Talks*. Vancouver British Columbia Canada: ACM, Aug. 2011, pp. 1–1.

[9] F. Steinbrucker, J. Sturm, and D. Cremers, "Volumetric 3D mapping in real-time on a CPU," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*. Hong Kong, China: IEEE, May 2014, pp. 2021–2028.

[10] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic, "HomebrewedDB: RGB-D Dataset for 6D Pose Estimation of 3D Objects," in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 2767–2776.

[11] P. Wang, H. Jung, Y. Li, S. Shen, R. P. Srikanth, L. Garattoni, S. Meier, N. Navab, and B. Busam, "PhoCaL: A Multi-Modal Dataset for Category-Level Object Pose Estimation with Photometrically Challenging Objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 222–21 231.

[12] A. Tejani, D. Tang, R. Kouskouridas, and T.-K. Kim, "Latent-Class Hough Forests for 3D Object Detection and Pose Estimation," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and

T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, vol. 8694, pp. 462–477.
[13] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes," in *Robotics: Science and Systems (RSS)*, May 2018.
[14] T. Hodaň, M. Sundermeyer, B. Drost, Y. Labbé, E. Brachmann, F. Michel, C. Rother, and J. Matas, "BOP challenge 2020 on 6D object localization," in *European Conference on Computer Vision*. Springer, 2020, pp. 577–594.
[15] F. Mayran de Chamisso, B. Meden, and M. Tamaazousti, "HSPA: Hough Space Pattern Analysis as an Answer to Local Description Ambiguities for 3D Pose Estimation," in *2022 British Machine Vision Conference (BMVC)*, Nov. 2022.
[16] W. Hoff and T. Vincent, "Analysis of head pose accuracy in augmented reality," *IEEE Transactions on Visualization and Computer Graphics*, vol. 6, no. 4, pp. 319–334, 2000.
[17] X. Liu, R. Zhang, C. Zhang, B. Fu, J. Tang, X. Liang, J. Tang, X. Cheng, Y. Zhang, G. Wang, and X. Ji, "GDRNPP," 2022.
[18] G. Wang, F. Manhardt, F. Tombari, and X. Ji, "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 16 611–16 621.
[19] R. L. Haugaard and A. G. Buch, "Surfemb: Dense and continuous correspondence distributions for object pose estimation with learnt surface embeddings," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6749–6758.
[20] Y. Labbe, J. Carpentier, M. Aubry, and J. Sivic, "CosyPose: Consistent multi-view multi-object 6D pose estimation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, Aug. 2020.
[21] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model Globally, Match Locally: Efficient and Robust 3D Object Recognition," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010.