KITchen: A Real-World Benchmark and Dataset for 6D Object Pose Estimation in Kitchen Environments

Abdelrahman Younes and Tamim Asfour

***Abstract*— Despite the recent progress on 6D object pose estimation methods for robotic grasping, a substantial performance gap persists between the capabilities of these methods on existing datasets and their efficacy in real-world grasping and mobile manipulation tasks, particularly when robots rely solely on their monocular egocentric field of view (FOV). Existing real-world datasets primarily focus on table-top grasping scenarios, where a robot arm is placed in a fixed position and the objects are centralized within the FOV of fixed external camera(s). Assessing performance on such datasets may not accurately reflect the challenges encountered in everyday grasping and mobile manipulation tasks within kitchen environments such as retrieving objects from higher shelves, sinks, dishwashers, ovens, refrigerators, or microwaves. To address this gap, we present KITchen, a novel benchmark designed specifically for estimating the 6D poses of objects located in diverse positions within kitchen settings. For this purpose, we recorded a comprehensive dataset comprising around 205k real-world RGBD images for 111 kitchen objects captured in two distinct kitchens, utilizing a humanoid robot with its egocentric perspectives. Subsequently, we developed a semi-automated annotation pipeline, to streamline the labeling process of such datasets, resulting in the generation of 2D object labels, 2D object segmentation masks, and 6D object poses with minimal human effort. The benchmark, the dataset, and the annotation pipeline are publicly available at https://kitchen-dataset.github.io/KITchen.**

I. Introduction

Recent work in robot navigation in indoor environments shows remarkable advances for mobile robots to navigate towards a goal position following different modalities such as 2D points [1], [2], object's image [3], [4], language instruction [5], [6], and acoustic signals [7], [8]. However, expanding the capabilities of these robots beyond navigation to perform tasks that require physical interaction with the surrounding objects in the environments remains a harder challenge. Therefore, understanding the 3D surroundings and objects' 6D pose estimation are essential pre-tasks for any robotic grasping and manipulation task [9], [10], [11].

Current advances in tackling the 6D pose estimation problem focus on developing new models and approaches [12], [13] to achieve the best results on the BOP challenge[1] datasets [14], [15], [16], [17], [18], [19], [20], [21], [22], [23]. While this paradigm boosted the research on 6D pose estimation, however, the available real-world datasets primarily focus on serving the table-top robotic grasping setup, featuring a robotic arm fixed in a position above objects, close to them, and often the objects are centered within the robot's FOV and in some cases with multiple cameras setup [24].

Fig. 1. Challenging kitchen locations that our dataset covers in contrast with the currently available datasets. The objects are distributed across diverse locations such as fridge, drawer, sink, higher shelves, microwave, dishwasher, oven, etc.

These datasets do not cover the challenging scenarios that mobile manipulators face inside indoor environments, especially in kitchens, where objects are normally placed in different not-centered positions with respect to the robot's field of view (FOV) such as on higher shelves, inside fridges, microwaves, dishwashers or ovens or in sinks. These locations not only impose challenging 6D poses with respect to the robot's camera but also cover more diverse and challenging surroundings such as transparent shelves in the case of refrigerators, see-through shelves in the case of dishwashers, and reflective backgrounds in the case of sinks, these challenges are not covered in the currently available real-world datasets [24]. These gaps and the not-

The research leading to these results has received funding from the Baden-Württemberg Ministry of Science, Research and the Arts (MWK) as part of the state's "digital@bw" digitization strategy in the context of the Real-World Lab "Robotics AI", the Carl Zeiss Foundation through the JuBot project and the German Federal Ministry of Education and Research (BMBF) under the Robotics Institute Germany (RIG). The authors are with the Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany. {younes,asfour}@kit.edu

[1] https://bop.felk.cvut.cz

covered scenarios do not provide a reliable indication of the performance of the developed methods on these real-world datasets in the context of mobile manipulation tasks with monocular egocentric FOV.

In addition to that, the current top 10 models on the BOP leaderboard train a model for each dataset [25], or even for each object [12], [26], which makes it hard to use for robotic applications, where the robots have to deal with a large number of objects under constrained resources. Furthermore, the average inference time of these top 10 approaches is 0.0283 frames per second (fps) with the best being $4.386 fps$. This makes these approaches not reliable for real-time applications, such as mobile manipulation where the 6D pose estimate is only a preliminary step of object grasping which is followed by a set of actions needed to execute the grasp such as grasp selection, motion planning, etc. To

Fig. 2. The humanoid robot ARMAR-6, leveraged for its adjustable torso height and various camera angles provided by its adjustable roll-yaw neck, to enrich our dataset.

overcome the limitations of the current 6D pose estimation methods, we introduce KITchen, the first-of-its-kind large-scale real-world dataset recorded using the humanoid robot ARMAR-6 [27] as shown in Fig. 2, which has adjustable height and roll-yaw neck, in 2 different kitchen environments covering 111 kitchen objects from the robots' egocentric perspective to cover the objects in the challenging kitchens' locations as shown in Fig. 1. KITchen offers 2D bounding boxes, object segmentation, and 6D poses annotated with a semi-automated annotation pipeline to minimize the need for manual labeling.

The main contributions of our work are: (i) we introduce a large real-world annotated RGBD dataset for 111 objects with their 2D bounding boxes, segmentation masks, and 6D poses. (ii) we propose a semi-automated annotation pipeline to annotate the objects in the dataset to facilitate the creation of more real-world datasets and make it publicly available to other researchers to create such large-scale datasets. (iii) we introduce a new benchmark and competition, where the focus is to solve the object 6D pose estimation problem depending solely on the monocular FOV of robots and limiting the submissions to approaches that offer at least $5 fps$ to encourage further work on this problem while taking into consideration real-time applicability.

II. Related Work

A. Objects Datasets

Current research on 6D pose estimation leverages several datasets categorized into two main groups: instance-level object datasets and category-level object datasets. Instance-level datasets offer 6D pose annotations for specific objects, serving as benchmarks for many object pose estimation methods. In contrast, category-level datasets aim to extend object pose estimation approaches to estimate the pose of different instances within the same category. In this work, we focus on instance-level object pose estimation. This subsection provides an overview of currently available real-world datasets for instance-level 6D object pose estimation.

LineMOD (LM) [14] comprises 15 texture-less objects with diverse shapes, colors, and sizes. LM provides approximately 1.2K real-world test images for each object in cluttered scenes, totaling 18241 images. *LineMOD-Occluded (LMO)* [15] offers pose annotations for only eight objects from the LineMOD dataset under severe occluded conditions. *T-LESS* [16] consists of 30 industrial texture-less, symmetric, and similar objects with 1296 real-world images per object, totaling around 39K images. *ITODD* [17] provides 6D pose annotations for 28 industrial objects with less than 1K publicly available Gray-Depth validation images. *Homebrewed-Database (HB)* [18] comprises less than 5K real-world images as validation set for 33 objects, with only 8 of them being household objects. *HOPE* [19] consists of 28 toy grocery objects that could be utilized in kitchen environments, but it provides only 238 real-world images in 50 scenes. *IC-BIN* [21] also offers only 177 real-world test images for only 3 out of its 8 objects in multi-objects cluttered scenes with heavy occlusion to be used for the BOP challenge. *TUD-L* [23] provides around 11K real-world images for 3 objects not placed on tables which differs this dataset from the others. *MP6D* [28] consists of 20.1K real-world frames for 20 symmetrical specular-reflective objects in cluttered multi-object setups with occlusion. *ClearPose* [29] offers about 355K real images for 63 transparent symmetrical objects in 51 cluttered scenes with diverse backgrounds and occlusion. *YCB-video (YCB-V)* [20] provides 134K real-world images for 21 objects from the original YCB dataset [30].

GraspNet-1Billion [31] contains around 97.3K RGBD images for 88 objects recorded with 2 different cameras for table-top grasping scenario with one robot arm. *KIT object models database* [32] was originally introduced in 2012 and offers 3D CAD models for more than 100 diverse objects, the majority of which are kitchen-related groceries. However, it only offers very few images for each object, which makes

Dataset	Objects	Images	Annotated Objects/Image	Multi-object	Multi-instance	Mobile Robot's FOV
LineMOD/LineMOD-Occluded [14], [15]	15	18.2K	≤ 8	✓	✗	✗
T-LESS [16]	30	39K	≤ 10	✓	✓	✗
ITODD [17]	28	1K	≤ 8	✓	✓	✗
Homebrewed-Database [18]	33	5K	≤ 8	✓	✗	✗
HOPE [19]	28	238	5-20	✓	✓	✗
ICBIN [21]	3	177	≤ 3	✓	✓	✗
TUD-L [23]	3	11K	1	✗	✗	✗
MP6D [28]	20	20.1K	≤ 8	✓	✗	✗
ClearPose [29]	63	355K	≤ 10	✓	✗	✗
YCB-video (YCB-V) [20], [30]	21	134K	5	✓	✗	✗
GraspNet-1Billion [31]	88	97.3K	10	✓	✗	✗
KITchen (ours)	111	205K	10-50	✓	✓	✓

TABLE I

OVERVIEW OF AVAILABLE DATASETS FOR INSTANCE-LEVEL 6D POSE ESTIMATION

it hard to use this dataset for 6D pose estimation with the current state-of-the-art (SOTA) data-driven 6D pose estimation approaches. *KIT bimanual manipulation dataset* [33] provides rich data for learning models of bimanual manipulation tasks from human demonstrations. It includes accurate whole-body motion data, hand configurations, and 6D object poses captured using various sensors. The dataset features 12 bimanual actions for 21 kitchen-related objects.

An overview of available datasets for instance-level 6D pose estimation is given in Table I. The overview highlights key metrics including the number of covered objects in the dataset, total image count, number of annotated objects per image, presence of multi-object setups, availability of multiple instances of the same objects, and whether the dataset was captured using a mobile robot's field of view.

In this work, we carefully selected 111 kitchen-related objects from the YCB, KIT object dataset, and the KIT bimanual manipulation dataset to record the first-of-its-kind large-scale real-world RGBD dataset featuring multi-objects in structured cluttered setups with diverse backgrounds and lighting conditions recorded using a humanoid robot.

B. 6D Pose Estimation Methods

The current landscape of 6D pose estimation methods is diverse, ranging from traditional techniques such as template matching [34], [35], [36], [37] and correspondences with locally invariant features [38], [39], [40] to the current advanced deep learning SOTA render & compare approaches [13], [41]. These approaches provide the 6D poses of novel objects by rendering many views of the object during inference using its 3D CAD model and then passing these rendered views with the received cropped image of the object obtained by any 2D object detectors [42], [43], [44], [45], [46] to a coarse model which classifies which rendered image best matches the input image. Finally, they pass the initial pose to a refiner network to estimate an updated 6D pose of the object. In this work, we leverage MegaPose [13], Segment Anything [47], and YOLOv8 [48] to annotate our dataset.

III. THE KITCHEN DATASET

A. Dataset's Objects

We aim to create a large-scale real-world dataset that covers objects that are commonly used in kitchen environments. Although some of the existing object datasets already offer objects that are commonly used in kitchens, they lack enough diverse RGBD annotated images to train on [20] or no annotated RGBD at all [30], [33], [32], [49]. Therefore, we decided to reuse the already available kitchen-related objects from these datasets and provide a large real-world RGBD annotated dataset for them to facilitate research on 6D pose estimation for kitchen objects. These objects vary from toy vegetables and fruits from [30] to kitchen tools such as knives, spoons, cups, mugs, bowls, cutting board, egg whisk, frying pan, plate, etc. from [30], [20], [33], [32], [49] to kitchen groceries objects from [30], [32].

B. Dataset Recording

We recorded the dataset using our humanoid robot ARMAR-6 [50] inside two distinct kitchen environments as seen in Fig. 3 the first kitchen, referred to as the *Main Kitchen*, includes typical kitchen appliances such as a fridge, counter with drawers, table, sink, microwave, dishwasher, and oven. The second kitchen, named *Mobile Kitchen*, features a counter with drawers, sink, dishwasher, fridge, and three tables. To enhance diversity, we utilized four different table-top colors (red, white, gray, and blue) and varied the camera's heights (150cm, 177cm, and 185cm) using ARMAR-6's torso as shown in Fig. 4. Additionally, we recorded data under three different pitch angles (10 degrees, 37 degrees, and 49 degrees down) and six different lighting conditions as shown in Fig. 5. We shuffled the objects with each change of lighting, camera's height, or camera's angle to enrich the diversity of the recorded scenes. To avoid similar and repetitive frames, we limited our recording to 5 fps. To the best of our knowledge, this is the first of its kind dataset that covers this amount of different robots' fields of view.

C. Annotation Pipeline

Annotating objects with their ground truth 6D poses is a labor-intensive and time-consuming task. Although some of

Fig. 3. The two distinguished kitchens where we recorded our dataset. On the left side is the Main Kitchen while on the right side is the Mobile Kitchen.

Fig. 4. Diverse robot and camera heights realized through different torso positions of ARMAR-6. The images display heights of 145cm, 177cm, and 185cm from left to right, illustrating the varied perspectives captured in the datasets and the different placements of objects relative to the robot's field of view.

Fig. 5. Variation in robot neck pitch angle. The images depict angles of 10, 37, and 49 degrees from left to right, showcasing a diverse range of perspectives.

the recently published datasets attempted to semi-automate the annotation process. For instance, GraspNet-1Billion's approach [31] relies on manually annotating the first frame of each scene, then leveraging recorded camera poses to calculate the objects' poses in the following frames. However, this method was not optimal for our dataset, as KITchen has many more diverse scenes per kitchen compared to the simple setup used in GraspNet-1Billion, resulting in significantly more effort required for manual annotation. Another attempt to semi-automate the annotation process was presented by HANDAL [51], but their approach assumes a single object in the scene, making it unsuitable for our dataset, which contains $10 - 50$ objects in each scene. To overcome the limitations of existing approaches and streamline the annotation process, we propose a semi-automated annotation pipeline. This pipeline generates three types of annotations: 2D object bounding boxes, 2D segmentation masks, and 6D poses, see Fig. 6.

1) 2D Objects Bounding Boxes Annotation: The pipeline starts by receiving the collected 3D CAD object models for the dataset, then it generates around 100K annotated photo-realistic synthetic RGBD images with 2D bounding boxes using BlenderProc2 [52]. These synthetic images are used to finetune a pretrained YOLOv8 model [48] for 2D object detection. Subsequently, the trained model is applied to our real-world data, and manually classified images are inspected to distinguish correctly labeled ones. The model is then fine-tuned iteratively until all real-world data is accurately labeled with 2D object labels.

2) 2D Objects Segmentation Masks: For segmenting the objects and producing the 2D segmentation masks, we leverage Segment Anything [47], by passing the images as well as the 2D bounding boxes generated from the previous step.

3) 6D Object Poses: To generate the 6D poses for the objects in the images, we pass the 2D bounding boxes which are generated using the fine-tuned YOLOv8 object detection

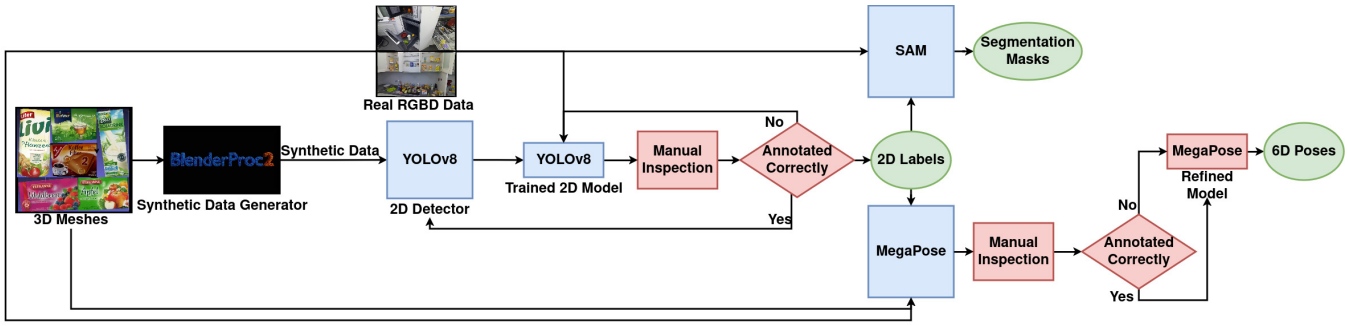

Fig. 6. Our proposed annotation pipeline. The pipeline starts with 3D meshes of dataset objects as input, which are processed by BlenderProc2 to generate synthetic data with 2D bounding boxes. This annotated 2D data is used to train a YOLOv8 2D object detector. Subsequently, real-world recorded data is fed into the trained model, and the output is manually inspected for correct and incorrect labeling. The correctly labeled images are used for model refinement, which is then validated on the incorrectly labeled images. This iterative process continues until all images are correctly labeled. The correctly labeled images are then passed to Segment Anything (SAM) to generate masks. Finally, the images, along with the 2D labels and 3D meshes, are fed into MegaPose to generate 6D poses for detected objects. Manual inspection of poses is performed using contour and mesh overlay images, and corrected annotations are used to iteratively fine-tune MegaPose until the entire dataset is accurately annotated.

model alongside the 3D CAD models of the detected objects with the input image into MegaPose [13]. The output 6D poses are used to overlay contours and meshes on the images for manual inspection. The MegaPose model is fine-tuned with corrected labeled data iteratively until all data are accurately annotated. The entire annotation pipeline is illustrated in Fig. 6 and several illustrative examples of the output of each step are demonstrated in Fig. 7.

D. Comparison to Existing Datasets

When compared to currently available datasets, the KITchen dataset stands out in several key aspects. With a diverse collection of 111 objects, our dataset offers a significantly wider range than the average number of objects found in existing datasets, surpassing the average by a factor of four. This expansive variety is crucial for training robust pose estimation models capable of handling a multitude of real-world scenarios. Moreover, the KITchen dataset offers a total of 205K RGBD images. This surpasses the average number of annotated images in existing datasets covered in Table I by over threefold, providing more data for training and evaluation purposes. Furthermore, our dataset has a remarkably larger number of annotated objects per image compared to the existing datasets with an unprecedented number of objects reaching 50 per image. This exceeds any available dataset by a significant margin, enabling more comprehensive analysis and training of instance-level 6D pose estimation models. Additionally, the KITchen dataset is unique in its capture methodology. It is the only dataset to have been recorded using the field of view of a humanoid robot with adjustable heights, camera angles, and lighting conditions. Unlike existing datasets that predominantly focus on tabletop scenes, our dataset features challenging locations within kitchen environments including refrigerators, ovens, sinks, higher shelves, microwaves, and dishwashers, offering a broader scope of real-world scenarios for pose estimation research. An overview of the dataset comparison is given in Table I.

IV. The KITchen Benchmark

Our proposed KITchen benchmark aims to encourage researchers in both computer vision and robotics to test their developed methods on a diverse and challenging multi-object dataset while considering the resource constraints of robots. To this end, we impose specific guidelines for leaderboard submissions to ensure practical applicability. Specifically, submissions must utilize a single model for all objects and maintain a minimum processing frequency of $5 fps$ during inference. The above conditions enhance the likelihood of the applicability of these methods in robotics. Aligning these criteria with those of the BOP Benchmark [53], we observe remarkable differences. Among the top 10 methods on the leaderboard, only two meet to the requirement of utilizing a single model per dataset rather than per object. Moreover, none of these methods achieves the required performance of 5 fps, with the closest reaching 4.3 fps. This discrepancy underscores a critical gap between current state-of-the-art approaches and the requirements of time-critical robotics applications, as evidenced by the average processing speed of the top 10 approaches on the BOP leaderboard, which is only 0.03 fps.

A. Problem Statement

The benchmark is designed to address the object 6D pose estimation problem, where the model receives an image I from the dataset D, where D is a set of RGBD images. The image I contains a set of objects $\{o\}_{i=0}^{n}$. The model has access to the M, where M is a set of 3D meshes of all objects O in the dataset D. The objective is to estimate the pose P of all objects $\{o\}_{i=0}^{n}$ in each image I, where $P = [R, T; 0, 1]$, where R is a 3×3 rotational matrix that describes the rotation of each of detected objects $\{o\}_{i=0}^{n}$ to the robot camera's frame and T is the translation vector to the origin of robot camera's coordinate system.

B. Datasets

Our benchmark leverages the KITchen dataset introduced in Sec. III. Notably, this dataset stands out as the first

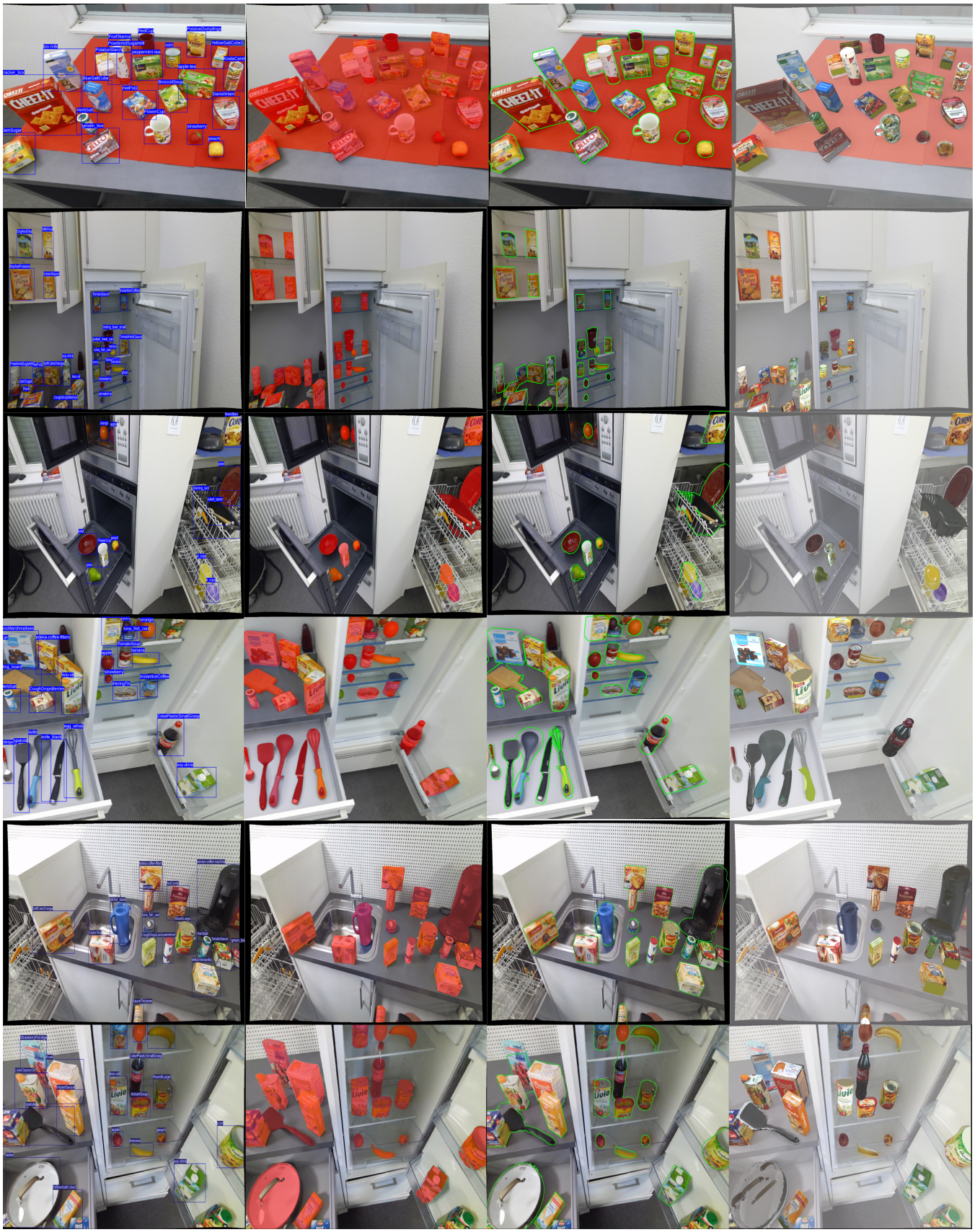

Fig. 7. Examples of the results generated by our proposed annotation pipeline. Sequentially from left to right: output of the 2D detector, segmentation masks, contour overlay, and mesh overlay.

of its kind, captured from the perspective of a humanoid robot, and encompasses varying heights and pitch angles, making it more suited to cover robotic mobile manipulation scenarios in kitchen environments. We split the dataset to training/validation/test sets with a 70/20/10 ratio.
Although our benchmark primarily focuses on the KITchen dataset, we invite other robotics research groups to record datasets in kitchen environments using their own robots and leverage our proposed annotation pipeline in Sec. III-C to annotate their data efficiently. Our vision for this benchmark extends beyond our dataset alone, we see it as a dynamic community platform where diverse research groups can collectively work to advance the field of robotic perception and pose estimation by testing their methods on a variety of datasets and providing their own datasets for other researchers to test on.

C. Pose Error Calculation

We utilize the same pose error function used by the BOP challenge [53]. The estimated pose is considered correct if the pose error function e calculated between the annotated pose P and the estimated pose $\hat{P}$ is lower than a predefined threshold θ_e, where $e \in \{e_{VSD}, e_{MSSD}, e_{MSPD}\}$, where e_{VSD} is the Visible Surface Discrepancy error function which focuses on the visible part of the object and evaluates poses with indistinguishable shapes as equivalent, disregarding the color information, e_{MSSD} is the Maximum Symmetry-Aware Surface Distance that calculates the surface deviation between vertices in the 3D, calculating the maximum distance between model vertices is crucial to know the chance of a successful grasp, while e_{MSPD} is the Maximum Symmetry-Aware Projection Distance that considers the object symmetries and calculate the difference in X, Y axes which makes it suitable for methods that rely on RGB data only. Finally, the Recall is defined as the ratio of correctly estimated poses with a total pose error e lower than the threshold θ_e across all objects. The Average Recall is then computed by averaging these recall values across various threshold settings.

V. Conclusion

We introduce KITchen, a novel object 6D pose estimation benchmark tailored to tackle this task within challenging kitchen environments using only monocular vision from robots' FOV, with a specific emphasis on real-time performance. To serve this benchmark, we recorded a large-scale real-world dataset, captured from different perspectives of a humanoid robot, featuring multi-objects in structured cluttered scenes in two distinct kitchen environments with diverse lighting conditions. Lastly, we proposed a semi-automated annotation pipeline aimed at streamlining the annotation of such datasets while minimizing manual human effort. We envision our benchmark to promote the development of novel approaches to solve the 6D pose problem on resource-constrained platforms, with an emphasis on real-time and real-world applicability.

Acknowledgment

We would like to thank Diana Burkart and Lisa Joosten for their contributions and assistance during the annotation process of the dataset.

References

[1] J. Ye, D. Batra, E. Wijmans, and A. Das, "Auxiliary tasks speed up learning point goal navigation," in *Conference on Robot Learning*. PMLR, 2021, pp. 498–516.

[2] S. Datta, O. Maksymets, J. Hoffman, S. Lee, D. Batra, and D. Parikh, "Integrating egocentric localization for more realistic point-goal navigation agents," in *Conference on Robot Learning*. PMLR, 2021, pp. 313–328.

[3] D. S. Chaplot, D. P. Gandhi, A. Gupta, and R. R. Salakhutdinov, "Object goal navigation using goal-oriented semantic exploration," *Advances in Neural Information Processing Systems*, vol. 33, pp. 4247–4258, 2020.

[4] A. Pal, Y. Qiu, and H. Christensen, "Learning hierarchical relationships for object-goal navigation," in *Conference on Robot Learning*. PMLR, 2021, pp. 517–528.

[5] M. Chang, T. Gervet, M. Khanna, S. Yenamandra, D. Shah, S. Y. Min, K. Shah, C. Paxton, S. Gupta, D. Batra *et al.*, "Goat: Go to any thing," *arXiv preprint arXiv:2311.06430*, 2023.

[6] P. Anderson, Q. Wu, D. Teney, J. Bruce, M. Johnson, N. Sünderhauf, I. Reid, S. Gould, and A. Van Den Hengel, "Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3674–3683.

[7] A. Younes, D. Honerkamp, T. Welschehold, and A. Valada, "Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 928–935, 2023.

[8] C. Chen, C. Schissler, S. Garg, P. Kobernik, A. Clegg, P. Calamia, D. Batra, P. Robinson, and K. Grauman, "Soundspaces 2.0: A simulation platform for visual-acoustic learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 8896–8911, 2022.

[9] C. Pohl, F. Reister, F. Peller-Konrad, and T. Asfour, "MAkEable: Memory-centered and affordance-based task execution framework for transferable mobile manipulation skills," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[10] F. Reister, M. Grotz, and T. Asfour, "Combining navigation and manipulation costs for time-efficient robot placement in mobile manipulation tasks," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9913–9920, 2022.

[11] T. Birr, C. Pohl, A. Younes, and T. Asfour, "AutoGPT+P: Affordance-based task planning with large language models," in *Robotics Science and Systems (RSS)*, 2024.

[12] Y. Su, M. Saleh, T. Fetzer, J. Rambach, N. Navab, B. Busam, D. Stricker, and F. Tombari, "Zebrapose: Coarse to fine surface encoding for 6dof object pose estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 6738–6748.

[13] Y. Labbé, L. Manuelli, A. Mousavian, S. Tyree, S. Birchfield, J. Tremblay, J. Carpentier, M. Aubry, D. Fox, and J. Sivic, "Megapose: 6d pose estimation of novel objects via render & compare," *arXiv preprint arXiv:2212.06870*, 2022.

[14] 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.

[15] 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 13*. Springer, 2014, pp. 536–551.

[16] T. Hodan, 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," in *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2017, pp. 880–888.

[17] B. Drost, M. Ulrich, P. Bergmann, P. Hartinger, and C. Steger, "Introducing mvtec itodd-a dataset for 3d object recognition in industry," in *Proceedings of the IEEE international conference on computer vision workshops*, 2017, pp. 2200–2208.

[18] R. Kaskman, S. Zakharov, I. Shugurov, and S. Ilic, "Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[19] 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," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 13 081–13 088.

[20] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes," *arXiv preprint arXiv:1711.00199*, 2017.

[21] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.-K. Kim, "Recovering 6d object pose and predicting next-best-view in the crowd," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3583–3592.

[22] 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13*. Springer, 2014, pp. 462–477.

[23] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis *et al.*, "Bop: Benchmark for 6d object pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 19–34.

[24] S. Thalhammer, D. Bauer, P. Hönig, J.-B. Weibel, J. García-Rodríguez, and M. Vincze, "Challenges for monocular 6d object pose estimation in robotics," *arXiv preprint arXiv:2307.12172*, 2023.

[25] Y. Hu, P. Fua, and M. Salzmann, "Perspective flow aggregation for data-limited 6d object pose estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 89–106.

[26] 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.

[27] T. Asfour, M. Wächter, L. Kaul, S. Rader, P. Weiner, S. Ottenhaus, R. Grimm, Y. Zhou, M. Grotz, and F. Paus, "Armar-6: A high-performance humanoid for human-robot collaboration in real world scenarios," *IEEE Robotics & Automation Magazine*, vol. 26, no. 4, pp. 108–121, 2019.

[28] L. Chen, H. Yang, C. Wu, and S. Wu, "Mp6d: An rgb-d dataset for metal parts' 6d pose estimation," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 5912–5919, 2022.

[29] X. Chen, H. Zhang, Z. Yu, A. Opipari, and O. Chadwicke Jenkins, "Clearpose: Large-scale transparent object dataset and benchmark," in *European Conference on Computer Vision*. Springer, 2022, pp. 381–396.

[30] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Benchmarking in manipulation research: Using the yale-cmu-berkeley object and model set," *IEEE Robotics & Automation Magazine*, vol. 22, no. 3, pp. 36–52, 2015.

[31] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 444–11 453.

[32] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization and manipulation in service robotics," *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.

[33] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The kit bimanual manipulation dataset," in *IEEE/RAS International Conference on Humanoid Robots (Humanoids)*, 2021, pp. 499–506.

[34] F. Jurie and M. Dhome, "A simple and efficient template matching algorithm," in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 2. IEEE, 2001, pp. 544–549.

[35] S. Hinterstoisser, S. Holzer, C. Cagniart, S. Ilic, K. Konolige, N. Navab, and V. Lepetit, "Multimodal templates for real-time detection of texture-less objects in heavily cluttered scenes," in *2011 international conference on computer vision*. IEEE, 2011, pp. 858–865.

[36] M. Sundermeyer, M. Durner, E. Y. Puang, Z.-C. Marton, N. Vaskevicius, K. O. Arras, and R. Triebel, "Multi-path learning for object pose estimation across domains," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 13 916–13 925.

[37] V. N. Nguyen, Y. Hu, Y. Xiao, M. Salzmann, and V. Lepetit, "Templates for 3d object pose estimation revisited: Generalization to new objects and robustness to occlusions," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 6771–6780.

[38] A. Collet and S. S. Srinivasa, "Efficient multi-view object recognition and full pose estimation," in *2010 IEEE International Conference on Robotics and Automation*. IEEE, 2010, pp. 2050–2055.

[39] A. Collet, M. Martinez, and S. S. Srinivasa, "The moped framework: Object recognition and pose estimation for manipulation," *The international journal of robotics research*, vol. 30, no. 10, pp. 1284–1306, 2011.

[40] K. Pauwels and D. Kragic, "Simtrack: A simulation-based framework for scalable real-time object pose detection and tracking," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2015, pp. 1300–1307.

[41] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," *arXiv preprint arXiv:2312.08344*, 2023.

[42] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 781–10 790.

[43] X. Long, K. Deng, G. Wang, Y. Zhang, Q. Dang, Y. Gao, H. Shen, J. Ren, S. Han, E. Ding *et al.*, "Pp-yolo: An effective and efficient implementation of object detector," *arXiv preprint arXiv:2007.12099*, 2020.

[44] Z. Zhang, X. Lu, G. Cao, Y. Yang, L. Jiao, and F. Liu, "Vit-yolo: Transformer-based yolo for object detection," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2799–2808.

[45] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie *et al.*, "Yolov6: A single-stage object detection framework for industrial applications," *arXiv preprint arXiv:2209.02976*, 2022.

[46] C.-Y. Wang, I.-H. Yeh, and H.-Y. M. Liao, "Yolov9: Learning what you want to learn using programmable gradient information," *arXiv preprint arXiv:2402.13616*, 2024.

[47] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, and R. Girshick, "Segment anything," *arXiv:2304.02643*, 2023.

[48] R. Varghese and M. Sambath, "Yolov8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*. IEEE, 2024, pp. 1–6.

[49] C. Mandery, O. Terlemez, M. Do, N. Vahrenkamp, and T. Asfour, "Unifying representations and large-scale whole-body motion databases for studying human motion," *IEEE Transactions on Robotics*, vol. 32, no. 4, pp. 796–809, 2016.

[50] T. Asfour, M. Waechter, L. Kaul, S. Rader, P. Weiner, S. Ottenhaus, R. Grimm, Y. Zhou, M. Grotz, and F. Paus, "Armar-6: A high-performance humanoid for human-robot collaboration in real-world scenarios," *IEEE Robotics & Automation Magazine*, vol. 26, no. 4, pp. 108–121, 2019.

[51] A. Guo, B. Wen, J. Yuan, J. Tremblay, S. Tyree, J. Smith, and S. Birchfield, "Handal: A dataset of real-world manipulable object categories with pose annotations, affordances, and reconstructions," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 11 428–11 435.

[52] M. Denninger, D. Winkelbauer, M. Sundermeyer, W. Boerdijk, M. Knauer, K. H. Strobl, M. Humt, and R. Triebel, "Blenderproc2: A procedural pipeline for photorealistic rendering," *Journal of Open Source Software*, vol. 8, no. 82, p. 4901, 2023.

[53] 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 *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*. Springer, 2020, pp. 577–594.