FruitBin: a tunable large-scale dataset for advancing 6D pose estimation in fruit bin-picking automation

Guillaume Duret, Mahmoud Ali, Nicolas Cazin, Danylo Mazurak, Anna Samsonenko, Alexandre Chapin, Florence Zara, Emmanuel Dellandréa, Liming Chen, Jan Peters

FruitBin: a tunable large-scale dataset for advancing 6D pose estimation in fruit bin-picking automation

Guillaume Duret[1,3], Mahmoud Ali[1], Nicolas Cazin[1], Danylo Mazurak[1], Anna Samsonenko[1], Alexandre Chapin[1], Florence Zara[2], Emmanuel Dellandrea[1], Liming Chen[1], and Jan Peters[3]

[1] Centrale Lyon, CNRS, LIRIS, UMR5205, F-69130 Ecully, France
[2] UCBL, CNRS, LIRIS, UMR5205, F-69622 Villeurbanne, France
[3] Intelligent Autonomous Systems Lab, Technical University of Darmstadt, 64289 Darmstadt, Germany {guillaume.duret, liming.chen, nicolas.cazin, alexandre.chapin, emmanuel.dellandrea, liming.chen}@ec-lyon.fr {danylo mazurak, anna samsonenko}@etu.ec-lyon.fr florence.zara@liris.cnrs.fr jan.peters@tu-darmstadt.de

Abstract. Bin picking, essential in various industries, depends on accurate object segmentation and 6D pose estimation for successful grasping and manipulation. Existing datasets for deep learning methods often involve simple scenarios with singular objects or minimal clustering, reducing the effectiveness of benchmarking in bin picking scenarios. To address this, we introduce FruitBin, a dataset featuring over 1 million images and 40 million 6D poses in challenging fruit bin scenarios. FruitBin encompasses all main challenges, such as symmetric and asymmetric fruits, textured and non-textured objects, and varied lighting conditions. We demonstrate its versatility by creating customizable benchmarks for new scene and camera viewpoint generalization, each divided into four occlusion levels to study occlusion robustness. Evaluating three 6D pose estimation models—PVNet, DenseFusion, and GDRNPP—highlights the limitations of current state-of-the-art models and quantitatively shows the impact of occlusion. Additionally, FruitBin is integrated within a robotic software, enabling direct testing and benchmarking of vision models for robot learning and grasping. The associated code and dataset can be found on: https://gitlab.liris.cnrs.fr/gduret/fruitbin.

Keywords: 6D pose estimation · occlusion robustness · robotics · dataset and benchmark

1 Introduction

Bin picking, a fundamental process where objects are retrieved from containers or bins, is widely utilized in various industries such as manufacturing, logistics, and warehousing. Common approaches rely on object instance segmentation and

Fig. 1: 15 initial scenes from a single viewpoint illustrating the domain randomization.

6D pose estimation for the objective of grasping [11,18]. A relevant application of this process in agriculture, food industry, and household robot assistance is fruit bin picking. It is an uncovered challenge that involves various types of objects and textures where the availability of data is pivotal for further progress [42]. It illustrates an example case where any imprecision in vision could result in irreversible grasping-based damage, creating a need for precision and robustness in existing 6D pose estimation. This leads to a need for precise benchmarking of 6D pose estimation in this context.

Numerous datasets have been created for the purpose of 6D pose estimation and have gained attention with the increasingly popular BOP challenge [35]. This challenge has supported significant improvements in 6D pose estimation methods, mainly due to the introduction of large-scale synthetic data. However, current benchmarks for 6D pose estimation predominantly portray tabletop scenes, neglecting the specific challenges posed by bin-picking scenarios characterized by multiple object instances, significant occlusions, and clutter. Additionally, benchmarks only offer partial robotic environments and overlook the crucial stage of robot learning for manipulation, which involves mastering the intricate interactions between robots and objects, necessitating a linked training of vision and manipulation models [6].

In this paper, we introduce FruitBin, a large-scale dataset consisting of simulated data tailored to facilitate robot learning, specifically emphasizing the demanding task of fruit bin picking. Illustrative examples of FruitBin can be seen in Figures 1-2. Table 1 offers a comparative overview of FruitBin in relation to state-of-the-art datasets. FruitBin is constructed upon PickSim [8], a recently introduced open-source simulation pipeline for robotics. PickSim is based on Gazebo [20], a widely adopted open-source 3D robotics simulation software used in robotics research and development [5]. The versatility of PickSim allows the vision model designed using this dataset to be directly transferred for dynamic tasks like grasping in the Gazebo simulator. It seamlessly integrates with robotics frameworks such as ROS [32] and Moveit [4] compared to state-of-the-art synthetic data generators such as Blenderproc [7] and Kubric [12].

The proposed dataset comprises over 1 million images, along with 40 million instance-level 6D pose annotations. It encompasses symmetric and asymmetric

fruits, with and without texture, high occlusion, clustering, different viewpoints, and lighting conditions, capturing all major 6D pose estimation challenges [33,34] within a single dataset across more than 70,000 scenes with 15 points of view. FruitBin boasts comprehensive annotations and metadata, covering 6D pose, depth, segmentation masks, point clouds, 2D and 3D bounding boxes, and occlusion rates. The amalgamation of this extensive annotation set, its substantial scale, and its diversity of challenges positions FruitBin as an adaptable dataset for generating benchmarks in challenging bin-picking scenarios.

To demonstrate its potential, we propose two distinct types of benchmarks for evaluating 6D pose estimation models: new scene generalization and new camera viewpoint generalization. Each benchmark encompasses four levels of difficulty, incorporating occlusion scenarios. We evaluated the performance of three foundational 6D pose estimation models: PVNet [30], DenseFusion [39], and GDRNPP [23,40]. To the best of our knowledge, FruitBin stands as a dataset meticulously tailored to address the demanding task of fruit bin picking [42]. It represents the largest-scale dataset available for 6D pose estimation, offering bin-picking challenges that can be finely tuned, using our proposed pipeline, to create custom benchmarks for 6D pose estimation. Finally, FruitBin can also be employed for different computer vision problems such as multi-view 3D reconstruction, new view synthesis [27], or camera pose estimation and 6D pose-based robotic grasping.

In the subsequent sections, Section 2 reviews prior work related to data generators and 6D pose datasets. This work presents a complete pipeline developed to be flexible and can be used either to generate other types of datasets or custom benchmarks for 6D pose estimation over our Fruitbin dataset, hopefully being useful for the community. Particularly, Section 3 outlines the first part of the pipeline: generating the Fruitbin dataset using the PickSim software and its statistics. Secondly, Section 4 describes the tunability of Fruitbin and the benchmark generation with the relative statistics. The outcomes of the baseline 6D pose estimation models across benchmarks are detailed in Section 5. Section 6 discusses certain limitations, while Section 7 concludes the article by providing insights and outlining future directions.

2 Related Work

The first step is data generation. In this work, targeting robotics applications, the choice has been oriented to PickSim [8]. The motivation is that although general vision models are making tremendous progress [24], robotic tasks are still very complex, and fine-tuning on a targeted dataset typically improves the performance of vision models [21] for robotics manipulation. This highlights the importance of a dataset that can be directly used in a robotic environment for grasping benchmarks. To our knowledge, state-of-the-art data generators such as Blenderproc [7] and Kubric [12] do not allow robotics integration without introducing a domain gap. PickSim, a recent pipeline, offers comprehensive annotation generation features, as illustrated in Figure 3, making it well-suited

for applications in robotics learning and 6D pose estimation. Employing robotic software, such as Gazebo, to generate synthetic computer vision data brings forth several advantages. Firstly, it allows for the seamless integration of physical engines, leading to realistic outcomes and surpassing 6D pose datasets that feature simple objects rendered against random backgrounds [16]. Secondly, it simplifies the integration of robots and sensors, equipped with native robot control capabilities. Lastly, it unleashes the potential to craft datasets and benchmarks tailor-made for robotic tasks, leveraging diverse open-source robotic motion planning libraries like MoveIt [4], integrated within Gazebo. PickSim further streamlines this process by providing user-friendly setup files for domain randomization, dataset recording, and generation. In the case of FruitBin, focusing on 6D pose, it enables the possibility to directly test 6D pose-based grasping outcomes [18]. The specific generation of FruitBin is described in Section 3.

Dataset	**type**	**#samples**	**#scenes**	**#6D pos**	**challenges**	**occ**	**C**	**rob-env**
LINEMOD [13]	R	18k	15	15k	TL	No	*	No
O-LINEMOD [2]	R	1214	15	120k	TL	*	*	No
APC [38]	R	10k	12	~240k	L	No	*	No
T-LESS [14]	R	49k	20	47k	TL/MI	*	*	No
YCB-V [43]	R-S	133k	92	613k	L	*	*	No
FAT [36]	S	60k	3	205k	L	*	*	No
BIN-P [19]	R-S	206k	12	20M	MI/BP	***	***	No
ObjectSynth [16]	S	600k	6	21M		*	*	No
HomebrewedDB [17]	S	17.4k	13	56k	L	*	*	No
GraspNet-1B [26]	R	97k	190	970k	-	**	**	No
RobotP [44]	S	4k	-	-	TL	*	*	No
HOPE [37]	R	2k	50	~30k	MI/L/BP	*	*	No
MetaGraspNet [9]	R-S	217k	6.4k	3M	MI/BP	**	**	**Yes**
SynPick [31]	S	503k	300	10M	BP	*	*	**Yes**
StereOBJ-1M [22]	R	396k	183	1.5M	L	*	No	No
DoPose [10]	R	3k	301	11k	BP	*	*	No
FruitBin	S	**1M**	**70k**	**40M**	**MI/BP/TL/L**	***	***	**Yes**

Table 1: Comparison of 6D pose datasets with their diverse challenges (R: Real, S: Synthetic, Occ: Occlusion, C: Clutter, MI: Multiple Instances, BP: Bin Picking, TL: Textureless, L: Light variety). Rob-Env indicates whether the dataset is integrable for application in a robotic environment.

Numerous datasets have been established for 6D pose estimation. Table 1 offers a comprehensive comparison of these datasets, encompassing various characteristics such as data nature (real or synthetic), size (including the number of image samples, scenes, and 6D pose annotations), and the specific 6D pose challenges included. It also indicates whether the datasets are integrable into a robotic environment. 6D pose challenges are quantified for each dataset, including bin-picking scenarios, multiple instances, texture-less objects, and lighting variety. Occlusion and clutter levels are also detailed to differentiate the varying complexities across datasets. The proposed FruitBin dataset distinguishes itself by being the only dataset covering all current 6D pose estimation challenges within a single dataset. It is notable that the majority of existing datasets do not cover the bin-picking scenario, which is one of the most challenging cases, making our data particularly relevant for hard-case scenarios. Significantly, FruitBin expands the sample size, offering 2 to 1,000 times more samples, and scales the

number of scenes from 6.4k to 70k. This considerable enhancement in dataset size holds critical implications, especially in addressing the challenge of generalization to unknown scenes. With over 40 million 6D pose annotations, FruitBin not only outperforms other datasets in terms of scale but also excels in the number of scenes and challenges covered, all while being seamlessly integrable within a robotic environment. Creating expansive, varied, and meticulously annotated benchmarks for 6D pose estimation is a demanding and time-intensive undertaking. The presence of thorough and well-annotated datasets holds immense significance in advancements in 6D pose estimation. However, as illustrated in Table 1, each dataset introduces distinct challenges that warrant tailored datasets to effectively address those specific challenges. These complexities encompass aspects like bin picking, scene diversity, viewpoint variety, diverse lighting conditions, occlusion, and multiple instances, all of which are gathered in FruitBin, making it suitable as a database for benchmark generation.

3 Raw data generation process of FruitBin using PickSim

This section outlines the process of generating FruitBin by harnessing the capabilities of PickSim [8]. In Section 3.1, we present the four key steps involved in the generation of FruitBin using the PickSim pipeline. Additionally, Section 3.2 provides relevant statistics. It is important to note that the data generation for FruitBin can be reproduced for any type of data, and the pipeline is made available for use by the community.

3.1 PickSim generation of FruitBin

Pre-processing. For FruitBin, PickSim employed eight raw meshes representing some of the most common fruits: apple, apricot, banana, kiwi, lemon, orange, peach, and pear. To maintain the distinct characteristics of each fruit and meet the requirements of a 6D pose dataset, no randomization was applied to the meshes or textures. Through this automated process, SDF files were generated, which are essential for Gazebo simulation. These SDF files contain crucial metadata, such as the category ID, necessary for future dataset recording.

Scene randomization. PickSim [8] includes domain randomization techniques [3, 28, 29], utilized to generate diverse scenes for fruit bin picking. By using configuration files, users can easily customize object counts, cameras, and lighting conditions, eliminating the need for additional code and simplifying the creation of randomized Gazebo world files. In the FruitBin dataset, scene randomization targets the bin, lighting, and fruits. The bin undergoes randomization with rotations and color variations, while the lighting setup includes randomized positions, intensities, and colors. The fruits are subjected to position randomization atop the bin. To maintain statistical consistency, the number of instances for each fruit category was randomly set between 0 and 30, ensuring a consistently full bin. This design ensures significant diversity in terms of lighting and overall scene configuration, as illustrated in Figure 1.

Camera randomization. The final step of randomization involves camera settings, utilizing the orbiter sampler within PickSim to introduce variability in the distance (ranging from 0.55m to 1m) between the camera and the orbiter center, as well as varying angles to ensure optimal scene viewpoints. This seamless setup facilitates the generation of fully randomized scenes that are both physically realistic and well-suited for fruit bin-picking scenarios. Figure 2 illustrates the impact of these camera parameters with 15 viewpoints of a scene.

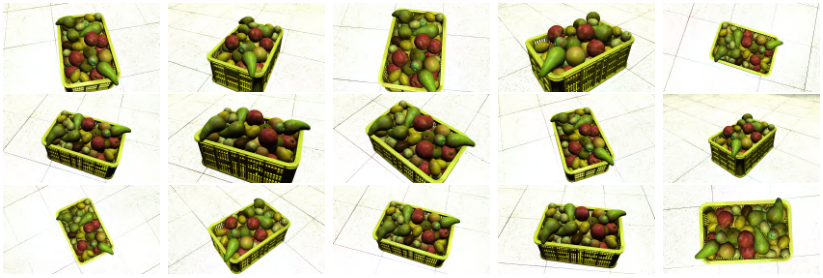
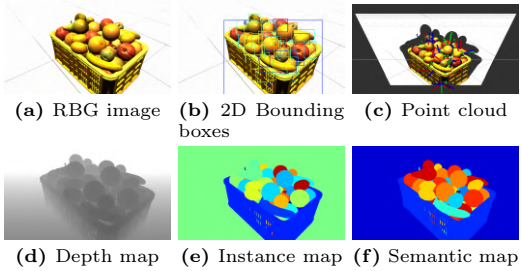


Fig. 2: 15 camera viewpoints of a single scene from the dataset FruitBin.

Data recording. Simulations in Gazebo can be effortlessly launched using the generated world files. These simulations yield datasets with recorded annotations from real simulated camera parameters, such as the RealSense D415 in our case. PickSim adds the generation of 6D pose features such as instance and semantic segmentation, bounding boxes, occlusion rates, 6D pose estimations, depth maps, point clouds, and normals, as illustrated in Figure 3.



(a) RBG image **(b)** 2D Bounding boxes **(c)** Point cloud

(d) Depth map **(e)** Instance map **(f)** Semantic map

Fig. 3: Examples of annotations generated with PickSim.

3.2 Description of the full dataset

For FruitBin, which comprises eight different fruits, the described randomization process was executed 10,000 times with 15 cameras, yielding 150,000 data frames.

This process was repeated seven times. The aggregation of these seven parts forms the entirety of FruitBin, containing over 1 million frames across 70,000 scenes and 105 different camera viewpoints, making it suitable for benchmark creation as described in Section 4. An overview of data statistics and insights into the distribution of 6D poses among various fruit categories is presented in Figure 4. It shows the distribution of fruit categories present in one image over our randomized process described in Section 3.1. This distribution, coming from random fruit pose initialization, logically ends up in a Gaussian distribution of instance numbers in images, ensuring an equitable representation of each fruit category. It can be highlighted that the majority of images have a relatively low number of instances, while few images have a high number of instances, as would be the case in a real-world scenario of random bin picking. Additionally, although the generation is random, all fruit categories are well represented.

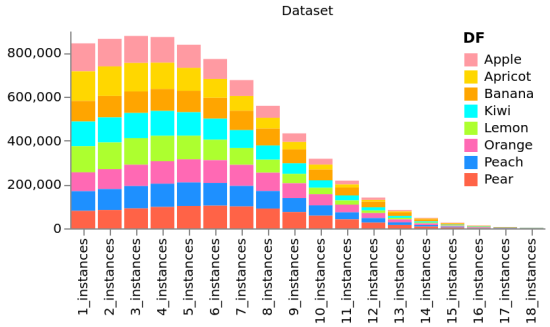


Fig. 4: Statistics of the complete dataset for each category, indicating the instance count for each image. The one million images are categorized for each fruits category by its instance number.

4 FruitBin: a tunable large-scale dataset for Fruit Bin Picking benchmark generation

The FruitBin dataset, enriched with extensive annotations, presents a multitude of challenges and has been created to have exceptional tunability for benchmarking. The dataset's extensive scale permits the creation of sub-datasets customized for specific purposes or for executing ablation studies. As described in Section 3, the sub-datasets are generated with 15 fixed points of view and 10,000 different scenes and lighting conditions. This setup allows for the generation of our targeted benchmarks. In practice, the benchmarks are generated by sampling the data from the global dataset of FruitBin and can be formulated for the targeted scene generalization and camera generalization scenarios, both with four levels of occlusion robustness benchmarks, resulting in eight benchmarks as it will be

specified in Section 4.1. Concretely, the raw data of the FruitBin dataset, as described in Section 3, is sampled to obtain smaller and more precise benchmarks. Users can also take FruitBin and use our pipeline to generate custom benchmarks with others parameters.

4.1 A tunable large-scale dataset

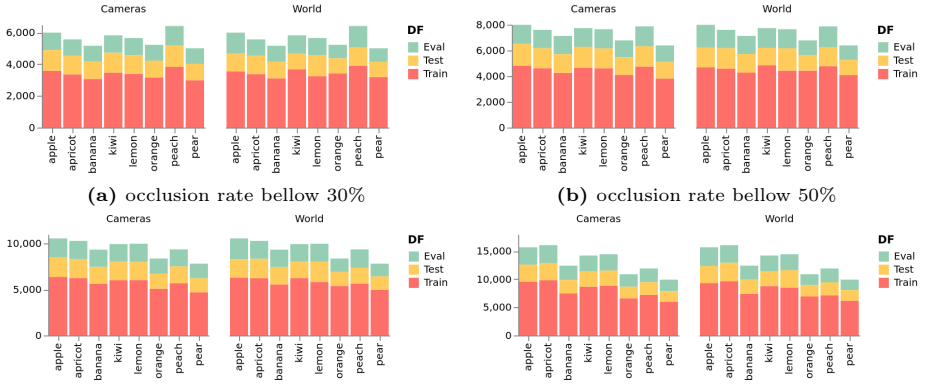
Eight benchmarks have been created based on the dataset. This section presents the benchmark generation from the FruitBin dataset (generated in the previous Section 3). In practice, a provided script is included to generate specific benchmarks for 6D pose estimation, incorporating user-defined parameters such as occlusion range, desired instance count, preferred viewpoints, and scene selections. While we demonstrated this process with the example of eight benchmarks, users can create their own benchmarks tailored to their specific needs. Lastly, the benchmark generation follows a default training, evaluation, and testing data split of 60%, 20%, and 20%, respectively. It is worth noting that the generated benchmarks are BOP format compatible, meeting the usual dataset format requirements for 6D pose estimation training [35]. All post-processing scripts to process the dataset format of our baseline are provided. The tunability feature of FruitBin is exemplified by its utilization in addressing two distinct types of 6D pose estimation benchmarks: scene generalization and camera viewpoint generalization, each encompassing four different levels of occlusion. The following Section 4.2 describes the specifics of the proposed benchmarks.

4.2 FruitBin benchmark generation

Camera and scene generalization scenarios. To investigate scene and camera point-of-view generalization, we established two distinct benchmarks for single-instance 6D pose estimation. The approach involves the careful sampling of the FruitBin dataset to generate scenario-specific benchmarks. The sampling process was performed using the initial portion of the dataset, encompassing 10,000 distinct scenes and 15 fixed camera viewpoints. In the scene-oriented scenario, data was extracted from the extensive dataset to form training, evaluation, and testing subsets. Specifically, 60% of the samples, equivalent to 6,000 scenes with all 15 camera viewpoints, were allocated for training, 20% were designated for evaluation, and the remaining 20% were reserved for testing, with each partition containing distinct scenes. A parallel methodology was applied to address the camera-oriented scenario. Here, 9 initial viewpoints were assigned for training, 3 for evaluation, and the last 3 for testing. Throughout the dataset filtering process, all image samples were categorized based on their respective object categories, priming the data for future 6D pose estimation tasks.

Occlusion robustness scenarios. To conduct a detailed examination of occlusion robustness, we extended this analysis to the two aforementioned types of benchmarks. Specifically, we incorporated four levels of difficulty related to occlusion, leveraging occlusion rate annotations. Instead of utilizing the entire benchmark

dataset, a filtering process was applied based on the occlusion rate associated with each object. The first version of the benchmark concentrates on objects with occlusion rates below 30%, followed by subsequent versions with occlusion rates of 50%, 70%, and up to 90%, respectively, progressively representing more challenging scenarios. To delve even deeper into the study of occlusion impact, a partition within the testing phase could be established to provide an occlusion-aware performance analysis, as discussed in Section 5. Figure 5 visually illustrates the data splitting in terms of image counts and the distribution among training, evaluation, and testing subsets for both types of benchmarks across the four levels of occlusion ranges. It demonstrates that no fruit category is overrepresented in the dataset, ensuring a balanced representation of all categories.



(a) occlusion rate bellow 30% **(b)** occlusion rate bellow 50%

(c) occlusion rate bellow 70% **(d)** occlusion rate bellow 90%

Fig. 5: Statistical figures depict the image counts for each fruit category across the four occlusion ranges for both types of benchmarks (scene and camera generalization), further segmented by the train (Train), evaluation (Eval), and testing partitions (Test).

5 Experiments

Baseline methods. To evaluate current state of the art methods on FruitBin benchmarks, we conducted an in-depth assessment using three distinct state-of-the-art 6D pose estimation models utilizing different data modalities:

a) PVNet. The first method, known as PVNet [30], employs an RGB image and 3D model information of objects as input to predict the 6D pose. This approach consists of two stages: initially, it identifies the 2D keypoint locations of objects through a series of convolution and deconvolution blocks, followed by a RANSAC-based voting mechanism. Subsequently, the 6D pose is derived by solving an uncertainty-driven Perspective-n-Point (PnP) problem, utilizing the 2D keypoints and the 3D model.

b) DenseFusion. The second method, referred to as DenseFusion [39], takes an RGB image, depth information (RGB-D input), and a semantic mask of the scene as inputs. In the case of DenseFusion, it initially generates binary masks for each object, which are then employed to crop the image and point cloud within the region of interest (ROI). Each ROI serves as input to a 2D feature extractor and a point cloud extractor, leading to the acquisition of color and geometry embeddings. These embeddings are concatenated for each point and fused to generate 'local' and 'global' information, ultimately resulting in the dense fused features. The 6D pose is estimated via a pose predictor model that progressively refines the pose through iterative steps.

c) GDRNPP. The third method, called GDRNPP [23], an improved version of GDR-Net [40], achieved the best results in the BOP challenge in 2022 and remains the best open-source model for the latest benchmark of 2023 [15, 35], proving to be one of the best models in terms of 6D pose estimation. It is designed as an end-to-end pipeline from RGB images to 6D pose by predicting three geometry features from a ROI: Dense Correspondences Map (M_{2D-3D}), the Surface Region Attention Map (M_{SRA}), and the Visible Object Mask of the object (M_{vis}). In practice, the correspondence map predicts at a pixel level the normalized 3D geometry of the targeted object. Additionally, the surface region classifies the pixels to create a geometry-aware attention map dealing with uncertainty. Finally, these two feature maps are used by a patch network to regress the final 6D pose of the object.

These 3 models have been trained and evaluated on established 6D pose estimation datasets such as LINEMOD and YCB-Video, as discussed in Section 2. They represent 3 distinct approaches [25,45] to 6D pose estimation: DenseFusion relies on the RGB-D modality, PVNet on keypoint detection, and GDRNPP on geometric-based 6D pose regression. They consequently provide an overview of 6D pose estimation methods performances on our dataset.

The baseline models are evaluated using the ADD metric [13] (*average distance*) for non-symmetrical objects and ADD-S [43] (*average closest point distance*) for symmetrical objects. In the case of FruitBin, Apple, Apricot, Kiwi, Lemon, Orange, and Peach objects are considered as symmetrical while Banana and Pear are non-symmetrical. ADD is defined as the mean distance between the transformed 3D model points using the estimated pose $[\hat{R}|\hat{t}]$ and the ground truth pose $[R|t]$. ADD-S, on the other hand, calculates the mean distance between each transformed model point and the nearest point on the ground truth transformed model, which accounts for symmetrical ambiguities. Based on the computed distance, the estimated pose is considered correct if the distance is less than 10% of the model's diameter, where the diameter represents the longest distance between any two points on the object.

Benchmark results. Using the three baseline models and the metrics presented earlier, we trained each model on the eight benchmarks outlined in Section 4.1. Table 2 presents results for DenseFusion, PVNet, and GDRNPP over the different benchmarks, while Figure 6 presents the evaluated models' performance with respect to occlusion.

Models	apple	apricot	pear*	kiwi	lemon	orange	peach	banana*	avg
Benchmark scene generalisation									
Occlusion from 0% to 30%									
Densefusion	**0.997**	**0.993**	0.674	**0.991**	**0.996**	**1.000**	**1.000**	0.490	0.899
PVNet	0.505	0.422	0.762	0.501	0.486	0.572	0.640	**0.858**	0.593
GDRNPP	0.922	0.862	**0.97**	0.936	0.938	0.965	0.982	0.828	**0.925**
Occlusion from 0% to 50%									
Densefusion	**0.995**	**0.950**	0.636	**0.948**	**0.956**	**1.000**	**1.000**	0.526	**0.882**
PVNet	0.430	0.432	0.793	0.503	0.473	0.572	0.685	**0.880**	0.596
GDRNPP	0.801	0.698	**0.897**	0.827	0.844	0.916	0.934	0.701	0.827
Occlusion from 0% to 70%									
Densefusion	**0.981**	**0.950**	0.570	**0.894**	**0.933**	**0.997**	**0.998**	0.414	**0.849**
PVNet	0.533	0.431	0.763	0.475	0.492	0.581	0.649	**0.879**	0.600
GDRNPP	0.629	0.582	**0.832**	0.673	0.695	0.802	0.857	0.576	0.706
Occlusion from 0% to 90%									
Densefusion	**0.844**	**0.713**	0.306	**0.656**	**0.726**	**0.896**	**0.903**	0.278	**0.676**
PVNet	0.445	0.363	**0.761**	0.487	0.481	0.561	0.621	**0.864**	0.573
GDRNPP	0.443	0.352	0.724	0.495	0.519	0.674	0.707	0.468	0.548
Benchmark camera generalisation									
Occlusion from 0% to 30%									
Densefusion	**0.983**	**0.872**	0.669	**0.968**	**0.957**	**1.000**	**0.999**	0.588	0.888
PVNet	0.590	0.516	0.862	0.631	0.594	0.701	0.784	**0.952**	0.704
GDRNPP	0.943	0.891	**0.973**	0.956	0.959	0.973	0.988	0.863	**0.943**
Occlusion from 0% to 50%									
Densefusion	**0.978**	**0.900**	0.606	**0.974**	**0.980**	**0.999**	**0.999**	0.592	**0.887**
PVNet	0.606	0.524	0.834	0.611	0.597	0.693	0.819	**0.941**	0.703
GDRNPP	0.85	0.766	**0.945**	0.886	0.888	0.932	0.956	0.789	0.876
Occlusion from 0% to 70%									
Densefusion	**0.983**	**0.922**	0.553	**0.887**	**0.864**	**0.995**	**0.997**	0.530	**0.850**
PVNet	0.577	0.475	0.810	0.602	0.588	0.748	0.773	**0.935**	0.688
GDRNPP	0.694	0.634	**0.879**	0.775	0.763	0.833	0.877	0.666	0.765
Occlusion from 0% to 90%									
PVNet	**0.519**	**0.447**	**0.827**	**0.580**	**0.568**	0.673	**0.753**	**0.939**	**0.663**
GDRNPP	0.485	0.442	0.777	0.556	0.529	**0.679**	0.712	0.514	0.587

Table 2: Success rates of DenseFusion, PVNet, and GDRNPP models on scene and camera benchmarks with varying occlusion levels. The upper part shows scene generalization, and the lower part shows camera generalization. Asymmetric objects are marked with an asterisk(*), and bold numbers indicate the best results for each benchmark. The last column shows the average performance across all fruits.

a) Scenarios discussion. The scene generalization and point-of-view generalization achieved acceptable scores, indicating that the models can adapt to new scenes and viewpoints, estimating the object pose in the camera frame. However, we observed that camera generalization results are generally better, especially for PVNet, which improved from an average score of 0.59 to 0.69. This can be attributed to PVNet's sensitivity to new lighting conditions due to its keypoint detection step. In the camera generalization scenario, lighting conditions encountered during testing were also seen during training, unlike the scene generalization scenario, where lighting conditions were novel. In comparison, DenseFusion and GDRNPP demonstrated the ability to generalize to new scenes, viewpoints, and lighting conditions.

b) Symmetry and texture discussion. FruitBin includes two non-symmetrical objects with distinct textures, which significantly impacted the results. DenseFusion's performance decreased by more than 20% for banana and pear compared to other fruits, due to its heavy reliance on depth and geometry, which can lead to local minima. For these fruits, the non-symmetry mainly arises from characteristic textures, while the geometry is nearly symmetrical. Symmetric objects, where the metric accounts for symmetry, do not face this issue. Conversely, PVNet, which relies on characteristic keypoints, significantly performed better with these two objects due to their specific textures compared to texture. GDRNPP, with its learned approach, shows more robustness to object characteristics. For both scenarios, GDRNPP achieved very good results, with scores in the range of [0.8, 1.0], showing no significant changes for symmetry and textures.

c) Occlusion discussion. Occlusion is a major challenge in 6D pose estimation, as confirmed by this study. DenseFusion's effectiveness is significantly influenced by object occlusion. It achieves a success rate of 90% or higher when the object's occlusion is below 30%. However, as occlusion increases to 90%, DenseFusion's performance declines to a success rate of 67%, failing to meet the refinement threshold for high occlusion levels in the camera scenario. Figure 6 illustrates performance across various occlusion ranges. Performance remains satisfactory with slight occlusions, achieving a 90% success rate with occlusion levels below 10%. However, the success rate drops significantly to 30% with occlusion levels between 80% and 90%. PVNet exhibits different characteristics. Its reliance on keypoints reduces its dependency on occlusion but results in less satisfactory overall performance. GDRNPP's performance also depends on occlusion levels, as illustrated in Figure 6. The performance drops linearly with respect to occlusion. Compared to PVNet, which relies on fixed keypoints, GDRNPP's reliance on feature prediction can be more challenging in high occlusion contexts, potentially leading to more misleading predictions.

Overall, DenseFusion, which relies on depth information, generally exhibits superior performance compared to PVNet and GDRNPP, which utilize only RGB images. However, it remains sensitive to both occlusion and object symmetry. PVNet shows the best robustness to occlusion but relies heavily on textures and has the lowest overall results. GDRNPP manages to significantly reduce dependency on object characteristics and can outperform DenseFusion without using depth in low occlusion cases. However, it is also sensitive to occlusion.

These experiments highlight the significant challenge of the FruitBin dataset due to its integration of key challenges in 6D pose estimation into a single dataset, naturally present in bin picking scenarios. The baseline methods do not consistently show satisfactory results across all fruit categories, occlusion levels, or lighting conditions, and no baseline performs satisfactorily in all challenges, demonstrating the dataset's difficulty in bin picking scenarios. It is important to note that these benchmarks encompass only two types of challenges and occlusion levels across all described fruits. FruitBin provides ample opportunity for increased difficulty, such as the addition of multi-instance and multiview 6D pose estimation or the integration of grasping success into the benchmark.

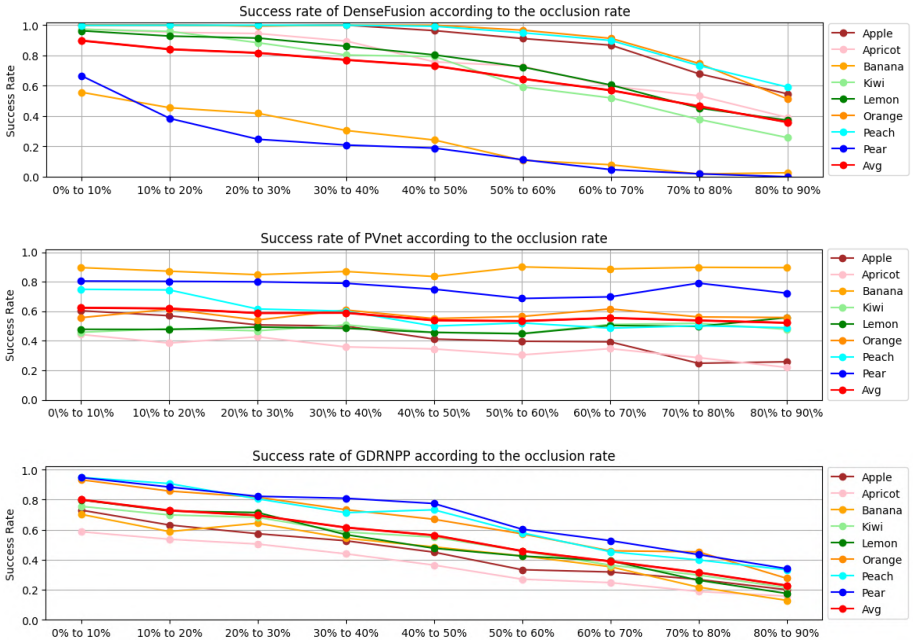


Fig. 6: Precise evaluation of DenseFusion, PVNet and GDRNPP models, trained on the 0-90% occlusion benchmarks, across different occlusion level partitions.

6 Discussion

This work introduces a benchmark for 6D pose estimation by presenting a dataset specifically designed for this purpose. However, there are limitations related to the fruit meshes used in our dataset. Given the inherent uniqueness of each fruit instance, the current dataset may not fully capture the variability within each category. Expanding the dataset to include a broader range of category-level 6D pose estimations would address this limitation. A logical next step would be to incorporate additional vision annotations into the open-source software PickSim such as NOCS [41], which is widely utilized for category-level 6D pose estimation. This enhancement would improve the dataset and enable a more comprehensive evaluation and analysis of category-level 6D pose estimation methods.

Furthermore, while the primary focus of this study is benchmarking, addressing the sim2real gap between our simulator and real-world fruits is crucial. We recognize the need for extensive studies on domain randomization techniques to bridge this gap effectively, especially in the context of robotics applications. By investigating and refining domain randomization methods, we can enhance simulation realism and improve the transferability of models trained in the simulator to real-world scenarios. As an initial step to mitigate the sim2real gap, we propose using diffusion models [1] to replace image backgrounds. This approach

contextualizes the images, significantly improving the sim2real gap and allowing for more diverse image data augmentation. Examples of generated images using theses methods are shown in Figure 7. It is worth noting that this method generates diverse, realistic backgrounds, including shadows, without altering the bin, thereby preserving the validity of the annotations.

Fig. 7: The left image shows the original input. The middle and right images demonstrate examples of background generation using diffusion models.

Finally, this paper highlights the current limitations of 6D pose estimation models in the context of bin picking for robotic applications. Given this goal, integrating grasping success metrics into the benchmark results would provide a quantitative measure of how 6D pose estimation accuracy impacts grasping performance. This addition would be a promising improvement to further validate and enhance the practical utility of the dataset.

7 Conclusion

We introduced FruitBin, the largest dataset for fruit bin picking, featuring over 40 million 6D pose annotations and 1 million images. This dataset gathers major challenges in 6D pose estimation. It addresses complexities such as bin picking, occlusion, symmetry, texture-less objects and lighting conditions, as examined in the dataset comparison in Section 2.

To cover various facets of 6D pose estimation, we devised 2 benchmarks evaluating scene and camera viewpoint generalization across four occlusion levels. Although the current baseline models exhibit individual strengths, none achieve satisfactory performance across all categories and benchmarks, presenting an intricate challenging dataset for the research community.

Carefully curated for 6D pose estimation in challenging bin picking scenario, this dataset is also applicable to other research problems such as 3D reconstruction, NeRF reconstruction, and multi-view 6D pose estimation. Its integration within a robotic simulator facilitates advancements in robotics learning, bridging computer vision and robotics. Researchers can leverage this dataset to assess models in simulations, advancing grasping and reinforcement learning. Our aim is for this dataset to catalyze improvements in 6D pose estimation models for robotics learning.

Acknowledgements

This work was in part supported by the French Research Agency, l'Agence Nationale de Recherche (ANR), through the projects Learn Real (ANR-18-CHR3-0002-01), Chiron (ANR-20-IADJ-0001-01), Aristotle (ANR-21-FAI1-0009-01). It was granted access to the HPC resources of IDRIS under the allocation 2023-[AD011013894], 2024-[AD011015271] and 2024-[AD011015591] made by GENCI.

References

1. Background difusion model. https://www.promeai.com/background-diffusion, accessed: 2023-08-25
2. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: 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. pp. 536–551. Springer (2014)
3. Chen, X., Hu, J., Jin, C., Li, L., Wang, L.: Understanding domain randomization for sim-to-real transfer. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=T8vZHIRTrY
4. Coleman, D., Sucan, I.A., Chitta, S., Correll, N.: Reducing the barrier to entry of complex robotic software: a moveit! case study. ArXiv **abs/1404.3785** (2014)
5. Collins, J., Chand, S., Vanderkop, A., Howard, D.: A review of physics simulators for robotic applications. IEEE Access **9**, 51416–51431 (2021). https://doi.org/10.1109/ACCESS.2021.3068769
6. Dasari, S., Ebert, F., Tian, S., Nair, S., Bucher, B., Schmeckpeper, K., Singh, S., Levine, S., Finn, C.: Robonet: Large-scale multi-robot learning. CoRR **abs/1910.11215** (2019), http://arxiv.org/abs/1910.11215
7. Denninger, M., Winkelbauer, D., Sundermeyer, M., Boerdijk, W., Knauer, M., Strobl, K.H., Humt, M., Triebel, R.: Blenderproc2: A procedural pipeline for photorealistic rendering. Journal of Open Source Software **8**(82), 4901 (2023). https://doi.org/10.21105/joss.04901, https://doi.org/10.21105/joss.04901
8. Duret, G., Cazin, N., Ali, M., Zara, F., Dellandréa, E., Peters, J., Chen, L.: PickSim: A dynamically configurable Gazebo pipeline for robotic manipulation. In: Advancing Robot Manipulation Through Open-Source Ecosystems - 2023 IEEE International Conference on Robotics and Automation (ICRA) Conference Workshop (May 2023), https://hal.science/hal-04074800
9. Gilles, M., Chen, Y., Robin Winter, T., Zhixuan Zeng, E., Wong, A.: MetaGraspNet: A Large-Scale Benchmark Dataset for Scene-Aware Ambidextrous Bin Picking via Physics-based Metaverse Synthesis. In: 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE). vol. 2022-Augus, pp. 220–227. IEEE (aug 2022). https://doi.org/10.1109/CASE49997.2022.9926427, https://ieeexplore.ieee.org/document/9926427/
10. Gouda, A., Ghanem, A., Reining, C.: Dopose-6d dataset for object segmentation and 6d pose estimation. In: 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA). pp. 477–483 (2022). https://doi.org/10.1109/ICMLA55696.2022.00077
11. Grard, M., Dellandréa, E., Chen, L.: Deep multicameral decoding for localizing unoccluded object instances from a single rgb image. International Journal of Computer Vision **128** (05 2020). https://doi.org/10.1007/s11263-020-01323-0

12. Greff, K., Belletti, F., Beyer, L., Doersch, C., Du, Y., Duckworth, D., Fleet, D.J., Gnanapragasam, D., Golemo, F., Herrmann, C., Kipf, T., Kundu, A., Lagun, D., Laradji, I., Liu, H.T., Meyer, H., Miao, Y., Nowrouzezahrai, D., Oztireli, C., Pot, E., Radwan, N., Rebain, D., Sabour, S., Sajjadi, M.S., Sela, M., Sitzmann, V., Stone, A., Sun, D., Vora, S., Wang, Z., Wu, T., Yi, K.M., Zhong, F., Tagliasacchi, A.: Kubric: A scalable dataset generator. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. vol. 2022-June, pp. 3739–3751. IEEE (jun 2022). https://doi.org/10.1109/CVPR52688.2022.00373, https://github.com/https://ieeexplore.ieee.org/document/9880070/
13. Hinterstoisser, S., Lepetit, V., Ilic, S., Holzer, S., Bradski, G., Konolige, K., Navab, N.: Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In: Computer Vision–ACCV 2012: 11th Asian Conference on Computer Vision, Daejeon, Korea, November 5-9, 2012, Revised Selected Papers, Part I 11. pp. 548–562. Springer (2013)
14. Hodan, T., Haluza, P., Obdrzalek, S., Matas, J., Lourakis, M., Zabulis, X.: 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). pp. 880–888. IEEE (mar 2017). https://doi.org/10.1109/WACV.2017.103, http://ieeexplore.ieee.org/document/7926686/
15. Hodan, T., Sundermeyer, M., Labbe, Y., Nguyen, V.N., Wang, G., Brachmann, E., Drost, B., Lepetit, V., Rother, C., Matas, J.: Bop challenge 2023 on detection segmentation and pose estimation of seen and unseen rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5610–5619 (2024)
16. Hodaň, T., Vineet, V., Gal, R., Shalev, E., Hanzelka, J., Connell, T., Urbina, P., Sinha, S., Guenter, B.: Photorealistic image synthesis for object instance detection. IEEE International Conference on Image Processing (ICIP) (2019)
17. Kaskman, R., Zakharov, S., Shugurov, I., Ilic, S.: HomebrewedDB : RGB-D Dataset for 6D Pose Estimation of 3D Objects Technical University of Munich , Germany Siemens Corporate Technology , Germany. ICCV Workshop (2019)
18. Kleeberger, K., Bormann, R., Kraus, W., Huber, M.F.: A Survey on Learning-Based Robotic Grasping. Current Robotics Reports **1**(4), 239–249 (2020). https://doi.org/10.1007/s43154-020-00021-6, https://doi.org/10.1007/s43154-020-00021-6
19. Kleeberger, K., Landgraf, C., Huber, M.F.: Large-scale 6D Object Pose Estimation Dataset for Industrial Bin-Picking. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 2573–2578. IEEE (nov 2019). https://doi.org/10.1109/IROS40897.2019.8967594, https://ieeexplore.ieee.org/document/8967594/
20. Koenig, N., Howard, A.: Design and use paradigms for Gazebo, an open-source multi-robot simulator. 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) **3**, 2149–2154 (2004). https://doi.org/10.1109/iros.2004.1389727
21. Lin, Y.C., Zeng, A., Song, S., Isola, P., Lin, T.Y.: Learning to see before learning to act: Visual pre-training for manipulation. 2020 IEEE International Conference on Robotics and Automation (ICRA) pp. 7286–7293 (2020), https://api.semanticscholar.org/CorpusID:214129334
22. Liu, X., Iwase, S., Kitani, K.M.: StereOBJ-1M: Large-scale Stereo Image Dataset for 6D Object Pose Estimation. In: 2021 IEEE/CVF International Conference on

Computer Vision (ICCV). pp. 10850–10859. IEEE (oct 2021). https://doi.org/10.1109/ICCV48922.2021.01069, https://ieeexplore.ieee.org/document/9711414/
23. Liu, X., Zhang, R., Zhang, C., Fu, B., Tang, J., Liang, X., Tang, J., Cheng, X., Zhang, Y., Wang, G., Ji, X.: Gdrnpp. https://github.com/shanice-l/gdrnpp_bop2022 (2022)
24. Majumdar, A., Yadav, K., Arnaud, S., Ma, Y.J., Chen, C., Silwal, S., Jain, A., Berges, V.P., Abbeel, P., Batra, D., Lin, Y., Maksymets, O., Rajeswaran, A., Meier, F.: Where are we in the search for an artificial visual cortex for embodied intelligence? In: Workshop on Reincarnating Reinforcement Learning at ICLR 2023 (2023), https://openreview.net/forum?id=NJtSbIWmt2T
25. Marullo, G., Tanzi, L., Piazzolla, P., Vezzetti, E.: 6d object position estimation from 2d images: A literature review. Multimedia Tools and Applications **82**(16), 24605–24643 (2023)
26. Maximilian, G., Chen, Y., Winter, T.R., Zeng, E.Z., Wong, A.: MetaGraspNet: A large-scale benchmark dataset for scene-aware ambidextrous bin picking via physics-based metaverse synthesis. In: IEEE International Conference on Automation Science and Engineering (CASE) (2022)
27. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020)
28. Mishra, S., Panda, R., Phoo, C.P., Chen, C.F.R., Karlinsky, L., Saenko, K., Saligrama, V., Feris, R.S.: Task2sim: Towards effective pre-training and transfer from synthetic data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9194–9204 (June 2022)
29. Muratore, F., Ramos, F., Turk, G., Yu, W., Gienger, M., Peters, J.: Robot learning from randomized simulations: A review. Frontiers in Robotics and AI **9** (2021)
30. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4561–4570 (2019)
31. Periyasamy, A.S., Schwarz, M., Behnke, S.: SynPick: A Dataset for Dynamic Bin Picking Scene Understanding. In: 2021 IEEE 17th International Conference on Automation Science and Engineering (CASE). vol. 2021-Augus, pp. 488–493. IEEE (aug 2021). https://doi.org/10.1109/CASE49439.2021.9551599, https://ieeexplore.ieee.org/document/9551599/
32. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y., et al.: Ros: an open-source robot operating system. In: ICRA workshop on open source software. vol. 3, p. 5. Kobe, Japan (2009)
33. Sahin, C., Garcia-Hernando, G., Sock, J., Kim, T.K.: A review on object pose recovery: From 3D bounding box detectors to full 6D pose estimators. Image and Vision Computing **96**, 103898 (2020). https://doi.org/https://doi.org/10.1016/j.imavis.2020.103898, https://www.sciencedirect.com/science/article/pii/S0262885620300305
34. Sahin, C., Kim, T.K.: Recovering 6D Object Pose: A Review and Multi-modal Analysis. In: Leal-Taixé, L., Roth, S. (eds.) Computer Vision – ECCV 2018 Workshops. pp. 15–31. Springer International Publishing, Cham (2019)
35. Sundermeyer, M., Hodaň, T., Labbe, Y., Wang, G., Brachmann, E., Drost, B., Rother, C., Matas, J.: Bop challenge 2022 on detection, segmentation and pose estimation of specific rigid objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2784–2793 (2023)

36. Tremblay, J., To, T., Birchfield, S.: Falling Things: A Synthetic Dataset for 3D Object Detection and Pose Estimation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). vol. 2018-June, pp. 2119–21193. IEEE (jun 2018). https://doi.org/10.1109/CVPRW.2018.00275, https://ieeexplore.ieee.org/document/8575443/
37. Tyree, S., Tremblay, J., To, T., Cheng, J., Mosier, T., Smith, J., Birchfield, S.: 6-DoF Pose Estimation of Household Objects for Robotic Manipulation: An Accessible Dataset and Benchmark. IEEE International Conference on Intelligent Robots and Systems **2022-Octob**, 13081–13088 (2022). https://doi.org/10.1109/IROS47612.2022.9981838
38. University, R.: Rutgers apc rgb-d dataset (2016), https://robotics.cs.rutgers.edu/pracsys/rutgers-apc-rgb-d-dataset/
39. Wang, C., Xu, D., Zhu, Y., Martín-Martín, R., Lu, C., Fei-Fei, L., Savarese, S.: Densefusion: 6d object pose estimation by iterative dense fusion. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3338–3347 (2019). https://doi.org/10.1109/CVPR.2019.00346
40. Wang, G., Manhardt, F., Tombari, F., Ji, X.: GDR-Net: Geometry-guided direct regression network for monocular 6d object pose estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16611–16621 (June 2021)
41. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
42. Wang, Z., Hirai, S., Kawamura, S.: Challenges and opportunities in robotic food handling: A review. Frontiers in Robotics and AI **8**, 789107 (2022)
43. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In: Robotics: Science and Systems XIV. Robotics: Science and Systems Foundation (jun 2018). https://doi.org/10.15607/RSS.2018.XIV.019, http://www.roboticsproceedings.org/rss14/p19.pdf
44. Yuan, H., Hoogenkamp, T., Veltkamp, R.C.: RobotP: A benchmark dataset for 6D object pose estimation. Sensors (Switzerland) **21**(4), 1–26 (2021). https://doi.org/10.3390/s21041299
45. Zhu, Y., Li, M., Yao, W., Chen, C.: A review of 6d object pose estimation. In: 2022 IEEE 10th Joint International Information Technology and Artificial Intelligence Conference (ITAIC). vol. 10, pp. 1647–1655. IEEE (2022)