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 Samsonenk, Alexandre Chapin, Florence Zara, Emmanuel Dellandrea, Liming Chen and Jan Peters

Laboratoire d'InfoRmatique en Image et Systèmes d'information
LIRIS UMR 5205 CNRS / INSA de Lyon / Université Claude Bernard Lyon 1 / Université Lumière Lyon 2 / Ecole Centrale de Lyon

Introduction

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.

Related works

FruitBin offers multiple advantages:

- The largest 6D-oriented dataset
- A dedicated fruit bin picking dataset
- All major challenges combined in one dataset
- Integration into robotic software

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.

Dataset generation with PickSim

The FruitBin dataset was generated using PickSim with:

- lighting randomization
- random number of objects
- 15 camera positions
- full 6D pose annotations

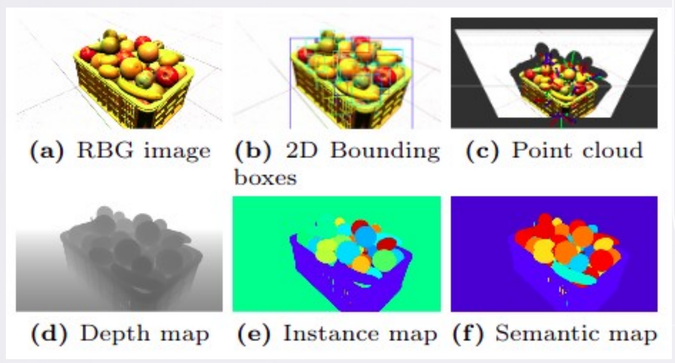

Fig. 1: Examples of annotations generated with PickSim.

FruitBin dataset and benchmarks

The large scale of the dataset allow for sampling specific scenario benchmark generation :

- Scene generalization
- Camera view point generalization
- Occlusion robustness

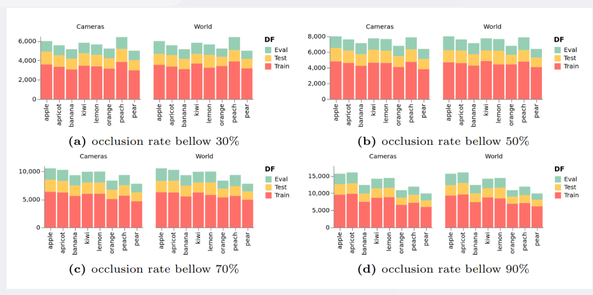

Figure 2: Statistical figures depict the image counts for each fruit category across the four occlusion ranges for both types of benchmarks (scene and camera generalization), further segmented by the train (Train), evaluation (Eval), and testing partitions (Test).

Figure 3: examples of images from Fruitbin dataset, top row are showing diversities of lighting and bottom rows are showing variation of camera views.

Experimentation

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**	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.

Benchmarks scenarios has been evaluated with three 6D pose estimation methods :

- PVnet [1]
- Densefusion[2]
- GDRNPP [3]

FruitBin demonstrates its challenges over texture or texture-less objects, occlusion rates and scene variation.

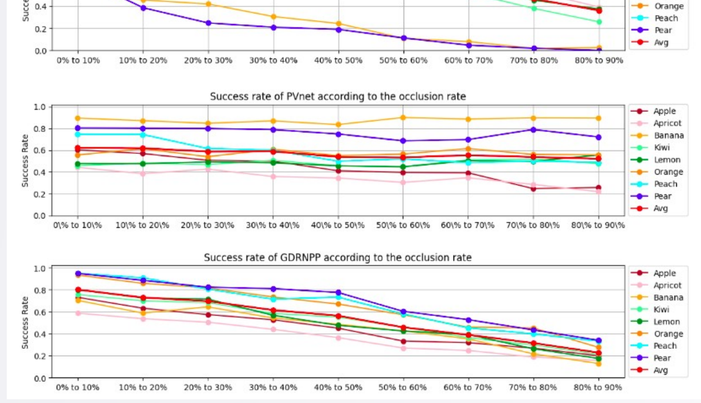

Figure 4: Precise evaluation of DenseFusion, PVNet and GDRNPP models, trained on the 0-90% occlusion benchmarks, across different occlusion level partitions.

Future works

A current limitation is the lack of variation in instances and real-world applications. Future work may consider:

- Category-based 6D pose estimation
- Enhancing rendering techniques
- 6D pose-based robotic grasping

Figure 5 : The left image shows the original input. The middle and right images demonstrate examples of background generation using diffusion models.

Acknowledgment

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] Peng, Sida, et al. "Pvnet: Pixel-wise voting network for 6dof pose estimation." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[2] Wang, Chen, et al. "Densefusion: 6d object pose estimation by iterative dense fusion." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019.
[3] Wang, Gu, et al. "Gdr-net: Geometry-guided direct regression network for monocular 6d object pose estimation." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021.