BoxDreamer: Dreaming Box Corners for Generalizable Object Pose Estimation

Supplementary Material

1. YCB-Video Reference Database Construction

Table 1 shows the selected reference video sequences for the YCB-Video dataset [8]. The reference video sequences are chosen based on the following criteria: (1) Most-overlapping: the video contains the most overlapping viewpoints based on the object coordinates; (2) Occlusion-minimizing: the video sequences with a high overlapping ratio and without obvious occlusions. More specifically, the most overlapping video sequences are automatically selected based on the average angular difference between consecutive frames, with higher average angular differences indicating more overlapping views. The Occlusion-minimizing video sequences are manually selected from the video sequences sorted in descending order by overlapping ratio.

	Reference Video Sequence	
Objects	Most-overlapping	Occlusion-minimizing
002_master_chef_can	0091	0014
003_cracker_box	0007	0007
004_sugar_box	0089	0074
005_tomato_soup_can	0008	0003
006_mustard_bottle	0008	0008
007_tuna_fish_can	0008	0039
008_pudding_box	0070	0076
009_gelatin_box	0003	0000
010_potted_meat_can	0008	0014
011_banana	0010	0010
019_pitcher_base	0009	0041
021_bleach_cleanser	0008	0006
024_bowl	0007	0007
025_mug	0007	0070
035_power_drill	0010	0010
036_wood_block	0090	0081
037_scissors	0010	0016
040_large_marker	0010	0089
051_large_clamp	0010	0010
052_extra_large_clamp	0003	0003
061_foam_brick	0081	0081

Table 1. **Reference video sequences selection on the YCB-Video dataset.**

2. Additional Quantitative Results

2.1. OnePose and OnePose-LowTexture Dataset

We evaluated the performance of our method on the OnePose [7] and OnePose-LowTexture [2] datasets by computing the pose error success rate for different thresholds. Table 2 compares our approach with OnePose++ [2] and Gen6D [4] under both five-view and ten-view settings—the latter being the minimum number of views required for OnePose++ to reconstruct all objects successfully. Although OnePose++ achieves higher accuracy thanks to its dense matching strategy, it failed to recover most query views, yielding less than 50% accuracy for the 30cm-30deg threshold. For low-texture objects, the 30cm-30deg accuracy dropped to only 10.1%. In contrast, our method demonstrates robustness on both datasets, achieving a higher low-threshold success rate. As for Gen6D, its detection module failed to identify object regions in both datasets (see Fig. 1), highlighting its lack of robustness even when query and reference views exhibit minimal scale variations. Even when provided with ground-truth detection results, Gen6D could not deliver satisfactory performance.

Figure 1. **Detection results of Gen6D on the OnePose dataset.** The red bounding boxes indicate the Gen6D detection results.

2.2. YCB-Video Dataset

In this section, we present detailed quantitative results on the YCB-Video dataset. Table 3 and Table 4 display the performance on the Most-overlapping and Occlusion-minimizing reference databases against Gen6D and OnePose++, respectively. With only five reference views, our method outperforms both baselines in most cases, especially on the Occlusion-minimizing reference database with fewer occluded views, achieving higher accuracy. Although OnePose++ performs well with dense reference views (200 views), its performance still lags behind our method due to incomplete point reconstruction. As for Gen6D, even with dense reference views, its performance

Ref. images	Method	OnePose						OnePose-LowTexture					
		1cm-1deg	3cm-3deg	5cm-5deg	10cm-10deg	20cm-20deg	30cm-30deg	1cm-1deg	3cm-3deg	5cm-5deg	10cm-10deg	20cm-20deg	30cm-30deg
5	OnePose++‡	**14.8**	**31.6**	36.4	42.8	44.9	46.3	**0.8**	2.0	2.6	4.2	7.6	10.1
	Gen6D†	0.1	1.8	4.6	-	-	-	0.1	2.2	5.5	-	-	-
	Ours	0.5	14.1	**39.3**	**78.5**	**91.7**	**92.9**	0.3	**7.9**	**25.2**	**60.5**	**83.3**	**88.8**
10	OnePose++	**40.6**	**67.7**	**73.4**	78.8	80.7	81.4	**8.3**	**25.5**	**34.4**	44.6	51.9	56.0
	Gen6D†	0.1	1.9	4.9	-	-	-	0.2	2.5	5.7	-	-	-
	Ours	0.7	19.0	48.4	**81.8**	**90.6**	**92.0**	0.3	9.3	28.3	**67.5**	**88.0**	**91.5**

Table 2. **Performance comparison on OnePose and OnePose-LowTexture datasets.** ‡ indicates onepose++ have several objects failed to reconstruct and † indicates provide ground-truth detection results for Gen6D. The best results are highlighted in bold.

remains similar to that under the sparse view setting, highlighting an intrinsic limitation in handling occlusion scenarios.

	Gen6D†	OnePose++		Gen6D‡		Gen6D†		OnePose++		Ours	
Ref. images	25		25		200		200		200		5	
Metrics	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD
002_master_chef_can	78.8	31.8	64.8	29.2	58.7	27.9	77.1	30.5	70.4	32.4	66.7	23.2
003_cracker_box	43.6	16.0	24.3	8.6	8.3	1.0	39.7	13.1	57.0	51.8	72.7	50.3
004_sugar_box	43.5	7.3	16.0	6.9	19.9	2.5	45.9	8.1	31.7	13.5	67.1	7.9
005_tomato_soup_can	49.4	17.0	8.1	2.3	27.7	8.9	76.5	49.4	40.0	13.3	56.3	41.1
006_mustard_bottle	76.1	47.6	51.6	37.1	68.7	46.0	86.6	50.4	72.3	47.3	67.8	44.3
007_tuna_fish_can	87.2	52.3	0.4	0.1	69.0	37.8	42.4	7.3	49.0	30.7	82.2	50.1
008_pudding_box	42.8	4.2	4.1	0.5	4.8	0.6	64.3	36.1	22.4	5.5	61.2	13.3
009_gelatin_box	66.0	37.6	0.0	0.0	34.3	13.4	61.5	39.8	66.7	48.1	28.1	14.8
010_potted_meat_can	61.4	36.6	41.6	30.8	51.3	31.2	37.4	25.1	64.7	49.8	80.2	61.6
011_banana	37.5	25.0	5.5	0.8	19.2	12.0	76.8	19.8	22.2	7.5	67.0	48.1
019_pitcher_base	76.5	18.5	27.1	20.2	25.7	7.7	35.8	17.4	56.9	18.9	86.9	74.1
021_bleach_cleanser	36.2	18.3	59.7	41.8	11.5	4.0	45.7	8.7	53.4	35.0	59.5	41.1
024_bowl	46.8	7.0	12.3	1.2	12.6	2.9	84.9	59.7	-	-	44.2	3.2
025_mug	83.5	63.4	6.1	1.9	79.6	54.7	33.7	7.7	61.0	37.9	90.2	79.3
035_power_drill	34.2	6.4	24.3	14.8	1.4	0.0	41.7	2.2	31.9	20.4	70.6	52.5
036_wood_block	41.5	2.5	0.6	0.0	5.8	0.0	21.6	8.1	4.8	0.0	0.0	0.0
037_scissors	21.5	7.4	0.4	0.0	4.2	2.0	64.2	55.1	14.7	5.1	24.1	10.8
040_large_marker	65.0	55.1	6.1	4.3	48.9	40.9	33.3	7.5	46.4	38.4	67.5	55.7
051_large_clamp	36.0	9.1	7.7	1.1	17.4	1.8	35.7	11.1	13.9	2.4	56.4	17.1
052_extra_large_clamp	34.3	8.7	31.2	8.3	13.2	0.3	35.7	11.1	70.1	27.7	78.2	36.2
061_foam_brick	19.3	6.5	0.1	0.0	28.3	10.4	22.1	7.3	16.3	5.2	56.1	20.1
MEAN	51.5	22.8	18.7	10.0	29.1	14.6	51.1	22.9	43.3	24.5	65.5	35.4

Table 3. **Performance of the YCB-Video dataset (Most-overlapping).** † indicates providing ground-truth detection results for Gen6D, ‡ indicates the background has been masked based on the ground-truth mask to help Gen6D to achieve better performance.

	Gen6D†	OnePose++		Gen6D‡		Gen6D†		OnePose++		Ours	
Ref. images	25		25		200		200		200		5	
Metrics	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD	ADD-S	ADD
002_master_chef_can	73.1	43.1	22.0	7.7	69.6	40.5	72.4	43.7	81.2	43.3	67.8	33.2
003_cracker_box	40.9	13.9	22.0	6.3	8.4	1.0	39.7	13.1	56.8	50.7	75.5	47.8
004_sugar_box	47.0	9.0	14.6	7.7	12.8	4.7	47.9	8.3	29.9	14.8	66.2	8.6
005_tomato_soup_can	57.6	34.8	43.6	20.7	57.0	30.7	56.3	32.9	68.4	40.6	80.4	49.5
006_mustard_bottle	76.0	47.6	52.4	37.8	68.8	46.1	76.5	49.4	73.1	48.1	57.6	38.8
007_tuna_fish_can	75.4	37.9	3.4	1.2	53.9	25.3	76.4	38.5	54.7	23.5	67.3	31.7
008_pudding_box	52.8	4.7	4.4	1.9	3.0	0.0	52.4	5.0	36.8	33.4	70.4	47.7
009_gelatin_box	77.0	28.1	0.0	0.0	11.3	1.2	75.9	28.1	43.0	15.8	75.6	56.1
010_potted_meat_can	59.2	35.8	49.0	37.4	44.5	29.9	58.8	35.2	64.7	49.2	69.0	51.8
011_banana	37.6	25.2	6.6	1.0	19.1	11.9	37.4	25.1	20.4	7.6	67.5	45.9
019_pitcher_base	83.8	17.2	30.2	19.9	29.3	10.7	83.7	16.4	54.0	17.6	85.8	71.9
021_bleach_cleanser	62.1	34.9	45.8	35.5	19.1	13.2	62.4	35.0	50.2	35.1	64.2	43.5
024_bowl	45.3	5.7	12.2	1.4	12.6	2.9	45.7	8.7	-	-	29.8	1.4
025_mug	42.4	12.0	10.1	1.8	62.9	19.3	47.7	14.6	28.8	5.7	80.5	62.9
035_power_drill	34.5	6.7	23.0	13.0	1.4	0.0	33.7	7.7	32.6	20.5	74.6	57.4
036_wood_block	24.2	1.2	14.0	0.0	3.6	0.0	20.6	0.6	20.4	0.0	57.3	0.5
037_scissors	37.2	18.2	0.3	0.0	1.9	0.8	34.0	15.9	5.4	0.9	46.0	24.3
040_large_marker	35.0	19.7	0.4	0.1	22.0	6.1	42.1	20.9	16.1	10.3	48.2	39.0
051_large_clamp	37.1	9.7	8.4	1.1	17.4	1.8	33.3	7.5	12.9	2.3	58.7	18.8
052_extra_large_clamp	35.4	9.4	39.4	12.4	13.2	0.2	35.7	11.1	69.1	27.3	72.2	33.2
061_foam_brick	20.0	6.4	0.1	0.0	28.3	10.5	22.1	7.3	16.1	5.1	14.6	2.3
MEAN	50.2	20.1	19.1	11.3	26.7	12.2	50.2	20.2	41.7	22.6	66.9	37.8

Table 4. **Performance of the YCB-Video dataset (Occlusion-minimizing).** † indicates providing ground-truth detection results for Gen6D, ‡ indicates the background has been masked based on the ground-truth mask to help Gen6D to achieve better performance.

2.3. LINEMOD and Occluded LINEMOD Dataset

This section provides a detailed comparison of results on the LINEMOD [3] dataset. As shown in Table 6, our method outperforms baselines when using only five reference views. Under the 25-reference setting, our method surpasses OnePose++ in most cases and achieves comparable performance with Gen6D†, which benefits from ground-truth detection results and was trained on a different subset of LINEMOD objects. Additionally, our method performs on par with BB8 [6], an instance-level method that must be trained for each specific object.

Furthermore, Table 5 presents additional results on the Occluded LINEMOD [1] dataset, including performance with dense reference views for Gen6D and OnePose++. Despite using dense reference views, both Gen6D and OnePose++ perform poorly in occluded scenarios. In particular, with only 25 reference views, our method achieves a 32.4% improvement in ADD (s)-0.1d and a 39.3% improvement in Proj-2d@5px compared to OnePose++.

Ref. images	Method	Objects								Avg.
		ape	can	*cat*	*driller*	*duck*	eggbox*	glue*	holepuncher	
		ADD(s)-0.1d								
25	**Ours**	21.7	61.6	54.7	53.1	30.4	27.6	57.5	41.9	43.6
Full	OnePose++	0.0	0.0	2.7	0.0	4.2	43.8	20.0	19.1	11.2
	Gen6D	14.9	29.6	9.6	4.2	20.2	23.9	16.2	36.4	19.4
	Gen6D†	17.4	36.5	12.7	25.6	21.5	40.4	34.7	44.9	29.2
		Proj-2d@5px								
25	**Ours**	59.7	58.9	54.7	27.8	56.3	1.9	53.6	71.4	47.9
Full	OnePose++	0.0	0.0	9.3	0.0	9.2	22.2	0.0	28.3	8.6
	Gen6D	43.4	38.9	29.7	4.3	46.3	4.0	12.5	57.0	29.5
	Gen6D†	57.9	51.0	41.9	21.2	52.4	5.0	34.6	73.7	42.2

Table 5. **Additional results on Occluded LINEMOD.** Metrics ADD(s)-0.1d and Proj-2d@5px are reported. † indicates that Gen6D was provided with ground-truth detection results, and objects in *italic* are included in the Gen6D training set.

3. More Analysis

3.1. Advantages of Bounding Box Corner Representation

In this section, we present an additional comparison of different object pose representations. Specifically, we compare the box corner heatmap representation with the ray representation under Plücker coordinate [5] and the vector pose representation. For the vector pose, we adopt the

Ref. images	Method	Objects													Avg.
		ape	benchwise	cam	can	cat	driller	duck	eggbox*	glue*	holepuncher	iron	lamp	phone	
ADD(s)-0.1d															
5	OnePose++	-	0.0	0.0	2.0	0.0	3.3	-	-	-	0.0	2.7	1.9	1.0	-
	Gen6D	-	24.4	21.6	-	17.8	14.8	11.6	51.8	32.5	-	-	41.4	-	-
	Gen6D†	-	39.0	26.1	-	22.8	32.6	15.2	71.4	40.2	-	-	51.6	-	-
	Ours	33.1	81.0	44.0	68.6	41.9	69.8	21.9	89.0	60.3	15.4	45.3	60.1	37.0	51.3
25	OnePose++	6.3	74.0	57.0	44.3	25.2	70.1	16.5	96.8	21.5	25.4	70.4	76.0	39.5	47.9
	Gen6D	-	75.1	60.2	-	59.1	61.4	37.2	66.6	47.0	-	-	86.6	-	-
	Gen6D†	-	83.9	65.8	-	59.8	83.5	46.8	96.9	82.9	-	-	92.9	-	-
	Ours	31.6	86.6	66.1	81.0	49.8	82.9	43.9	83.8	90.0	50.0	67.2	85.1	58.1	67.4
	BB8	40.4	91.8	55.7	64.1	62.6	74.4	44.3	57.8	41.2	67.2	84.7	76.5	54.0	62.7
	PVNet	43.6	99.9	86.9	95.5	79.3	96.4	52.6	99.2	95.7	81.9	98.9	99.3	92.4	86.3
Proj-2d@5px															
5	OnePose++	-	0.0	0.0	1.7	0.0	2.6	-	-	-	0.0	1.5	1.4	2.4	-
	Gen6D	-	23.1	29.5	-	33.4	15.7	29.5	33.0	38.2	-	-	36.6	-	-
	Gen6D†	-	36.1	36.3	-	42.0	34.5	37.8	39.6	54.1	-	-	46.2	-	-
	Ours	66.6	45.8	35.7	47.3	57.1	25.6	61.9	72.3	54.5	39.3	21.9	19.8	19.9	43.8
25	OnePose++	35.2	86.4	90.4	75.1	58.2	75.5	58.2	92.3	26.8	55.4	79.8	80.1	65.5	67.6
	Gen6D	-	78.4	84.8	-	65.2	77.3	92.3	92.0	98.0	-	-	94.5	-	-
	Gen6D†	-	90.1	92.9	-	96.5	91.7	94.4	90.9	95.6	-	-	93.7	-	-
	Ours	96.0	71.1	81.6	90.0	96.0	66.3	97.1	87.3	98.6	95.2	60.0	67.6	69.1	82.9
	BB8	96.6	90.1	86.0	91.2	98.8	80.9	92.2	91.0	92.3	95.3	84.8	75.8	85.3	89.3
	PVNet	99.2	99.8	99.2	99.9	99.3	96.9	98.0	99.3	98.6	100.0	99.2	98.3	99.4	99.0

Table 6. **Additional comparison on the LINEMOD dataset** † indicates provide ground-truth detection results for Gen6D.

6D representation [10] for the rotation and a 3D translation and additionally include the principal point (2D) and focal length (1D) as inputs to the network—resulting in an 11-dimensional vector. For the Plücker ray, our implementation is based on the source code from Camera as Rays [9]. Regardless of the representation, the same decoder architecture is used; the primary differences lie in the input and output projection layers.

All experiments were conducted on the OnePose dataset, and we report the pose error success rate using a threshold of 5 cm and 5°. The networks for each representation were trained exclusively on the OnePose training set for 100 epochs. Ten reference views were used in the inference phase.

	Vector Pose	Plücker Ray	Box Corner Heatmap
5cm-5deg	28.9	33.3	**50.4**

Table 7. **Performance of different regression target representations.**

As shown in Table 7, the box corner heatmap representation outperforms the other two methods. This result highlights its key advantage: it is independent of camera intrinsic parameters and is well-suited for learning by vision transformers.

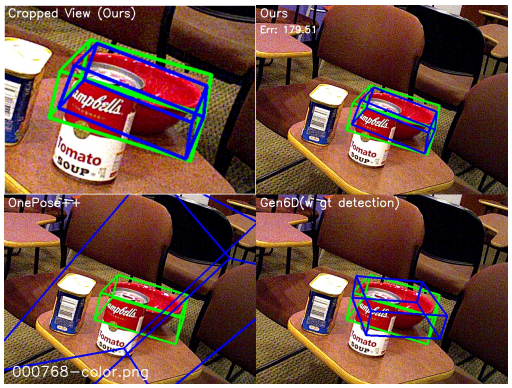

Figure 2. **Failure with symmetric objects.** BoxDreamer, along with OnePose++ and Gen6D, fails to predict the correct rotation for a fully symmetric object.

3.2. Failure Cases

Due to intrinsic ambiguities in the rotation of fully symmetric objects, our method may sometimes fail to predict the correct orientation. As shown in Fig. 2, our method incorrectly estimates the rotation for a symmetric object. Notably, both OnePose++ and Gen6D exhibit similar failures in such cases.

In another scenario, our method struggles with extreme lighting changes that significantly alter the object's surface color. As illustrated in Fig. 3, the disparity between the reference and target object colors leads to inaccurate pose

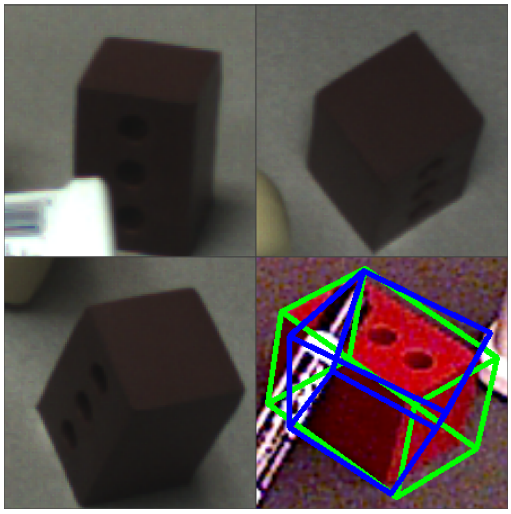

Figure 3. **Failure under extreme lighting changes.** The bottom right shows the predicted object pose, while the other panels display the reference views. The significant color differences caused by lighting variations lead to incorrect pose estimation.

predictions.

References

[1] Eric Brachmann, Alexander Krull, Frank Michel, Stefan Gumhold, Jamie Shotton, and Carsten 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*, pages 536–551. Springer, 2014. 2

[2] Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without CAD models. In *Advances in Neural Information Processing Systems*, 2022. 1

[3] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *Asian conference on computer vision*, pages 548–562. Springer, 2012. 2

[4] Yuan Liu, Yilin Wen, Sida Peng, Cheng Lin, Xiaoxiao Long, Taku Komura, and Wenping Wang. Gen6d: Generalizable model-free 6-dof object pose estimation from rgb images. In *ECCV*, 2022. 1

[5] Bruce D Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision. In *IJCAI'81: 7th international joint conference on Artificial intelligence*, pages 674–679, 1981. 2

[6] Mahdi Rad and Vincent Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *Proceedings of the IEEE international conference on computer vision*, pages 3828–3836, 2017. 2

[7] Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. OnePose: One-shot object pose estimation without CAD models. *CVPR*, 2022. 1

[8] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018. 1

[9] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras as rays: Pose estimation via ray diffusion. In *International Conference on Learning Representations (ICLR)*, 2024. 3

[10] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5738–5746, 2018. 3