Supplemental Material: A Pose Proposal and Refinement Network for Better 6D Object Pose Estimation

Ameni Trabelsi, Mohamed Chaabane, Nathaniel Blanchard, and Ross Beveridge

Department of Computer Science, Colorado State University

1 Additional Quantitative Results

In this section, we report the detailed evaluation of the different objects of both the YCB-Video and the LINEMOD datasets. We also show the results of the ablation study on the LINEMOD dataset.

1.1 Detailed Results on the YCB-Video Dataset

In Table 1, we show detailed pose estimation results on the YCB-Video dataset[9] in terms of ADD AUC. As mentioned in the main paper, our approach achieves the best results in 12 object classes out of 21 compared to other methods. DeepIM, surpasses other methods on 6 object classes out of 21, and HMap outperforms other methods on 4 object classes.

1.2 Detailed Results on the LINIEMOD Dataset

In Table 2, we compare our approach with existing state-of-the-art methods: Tekin[8], PVNet[6], BB8[7], SS6D[3] and DeepIM[4] on LINEMOD dataset[2]. Compared with other methods, our approach had the highest performance on 9 of the 13 object classes, PVNet had the best performance on 2 object classes, and SSD6D had the best performance on 2 object classes.

1.3 Results of the Ablation Study on the LINEMOD Dataset

In Table 3, we report the results of ablation study on LINEMOD dataset. The ablation study is similar to the one conducted in the main paper on YCB-Video dataset. The results in Table 3 suggest that each component iteratively improves the refinement results, highlighting their effectiveness, but the full importance of each method may be somewhat muted, compared to the results on the YCB-Video dataset, since the experiment took place on the LINEMOD dataset, where accuracy is near the dataset ceiling.

Table 1. Detailed results of our approach and other existing RGB-based methods on the different objects of the YCB-Video dataset in terms of ADD AUC

Methods	HMap[5]	PVNet[6]	DeepIM†[4]	OURS†
002-master-chef-can	**81.6**	-	71.2	72.1
003-cracker-box	**83.6**	-	**83.6**	81.7
004-sugar-box	82.0	-	94.1	**95.7**
005-tomato-soup-can	79.7	-	86.1	**88.2**
006-mustard-bottle	91.4	-	91.5	**94.8**
007-tuna-fish-can	49.2	-	87.7	**88.2**
008-pudding-box	**90.1**	-	82.7	80.2
009-gelatin-box	93.6	-	91.9	**94.5**
010-potted-meat-can	79.0	-	76.2	**82.6**
011-banana	51.9	-	**81.2**	78.7
019-pitcher-base	69.4	-	**90.1**	87.7
021-bleach-cleanser	76.1	-	**81.2**	78.1
024-bowl*	76.9	-	81.4	**83.4**
025-mug	53.7	-	81.4	**81.7**
035-power-drill	82.7	-	85.5	**87.8**
036-wood-block*	55.0	-	81.9	**83.7**
037-scissors	65.9	-	60.9	**67.4**
040-large-marker	56.4	-	**75.6**	71.1
051-large-clamp*	67.5	-	74.3	**75.2**
052-extra-large-clamp*	53.9	-	**73.3**	71.3
061-foam-brick*	**89.0**	-	81.9	82.2
MEAN	72.8	73.4	81.9	**83.1**

† denotes methods that deploy refinement steps.
* denotes symmetric objects.

Table 2. Detailed Results of our approach and other existing RGB-based methods on the different objects of the LINEMOD dataset in terms of ADD metric

Method	Tekin[8]	PVNet[6]	BB8†[7]	SSD6D†[3]	DeepIM†[4]	OURS†
ape	21.62	43.62	40.4	65	77	**84.47**
benchvise	81.80	**99.90**	91.8	80	97.5	98.71
cam	36.57	86.86	55.7	78	93.5	**93.73**
can	68.80	95.47	64.1	86	96.5	**97.84**
cat	41.82	79.34	62.6	70	82.1	**87.33**
driller	63.51	96.43	74.4	73	95	**96.91**
duck	27.23	52.58	44.30	66	77.7	**88.45**
eggbox*	69.58	99.15	57.8	**100**	97.1	98.49
glue*	80.02	95.66	41.2	**100**	99.4	99.5
holepuncher	42.63	81.92	67.20	49	52.8	**84.53**
iron	74.97	98.88	84.7	78	98.3	**99.10**
lamp	71.11	**99.33**	76.5	73	97.5	98.74
phone	47.74	92.41	54.0	79	87.7	**92.53**
MEAN	55.95	86.27	62.7	79	88.6	**93.87**

† denotes methods that deploy refinement steps.
* denotes symmetric objects.

Table 3. Results of the ablation study on different components of our refinement network MARN on LINEMOD dataset

Experiments	flow features	CNN features	Attention maps	ADD	2D-Reproj
Variant 1	None	✓	None	87.32	96.59
Variant 2	✓	✓	None	89.17	97.99
Variant 3	✓	✓	single	91.28	98.56
Variant 4	✓	✓	multiple	**93.87**	**99.19**

1.4 PPN Only: Evaluation on Three Benchmarks

We evaluate the performance of PPN, our pose estima-tion network without refinement, and compare it with state-of-the-art methods that do not use refinement. Results in Table 4 on three benchmarks suggest that PPN alone performs better than HMap and PoseCNN on all three datasets, and performs comparably to PVNet.

Unlike these approaches, PPN has the highest speed (50 fps), is completely end-to-end, and does not require any additional steps such as the PnP algorithm. Thus, we suggest that PPN alone is fast and robust enough to be deployed in real-world applications.

Table 4. Evaluation Results of our PPN compared to other state-of-the-art RGB-based methods that do not use refinement on three datasets: YCB-Video, LINEMOD and Occlusion using the 2D-Proj metric

Methods	PoseCNN[9]	HMap[5]	PVNet[6]	PPN(ours)
YCB-Video	3.72	39.4	47.4	**49.3**
LINEMOD	62.7	-	**99.0**	96.12
Occlusion	17.2	60.9	61.06	**61.10**

2 Additional Qualitative Results

In Fig. 1 to 3, we show qualitative results on the three datasets: YCB-Video[9], LINEMOD[2] and Occlusion[1] datasets. These examples show that our proposed method is robust to severe occlusions, scene clutter, different illumination and reflection.

Fig. 1. Examples of 6D object pose estimation results on the YCB-Video dataset. Each row corresponds to images from one testing video. Red bounding boxes correspond to ground truth poses, cyan bounding boxes correspond to predicted poses using our approach.

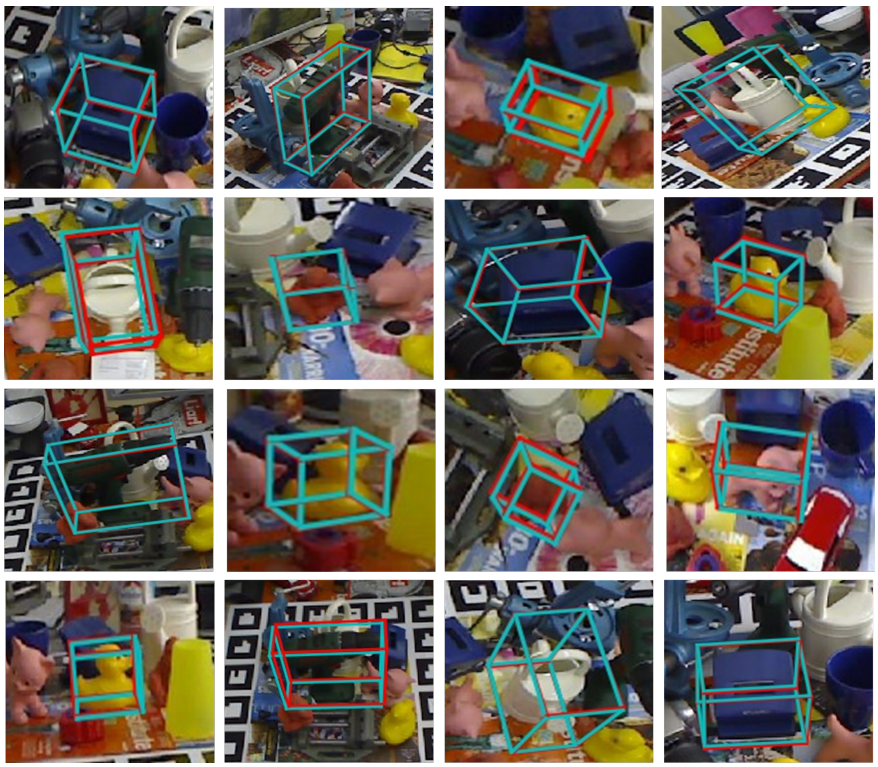

Fig. 2. Examples of 6D object pose estimation results on different objects from the LINEMOD dataset. Objects are: Holepuncher, driller, duck, can, ape, cat. Red bounding boxes correspond to ground truth poses, cyan bounding boxes correspond to predicted poses using our approach.

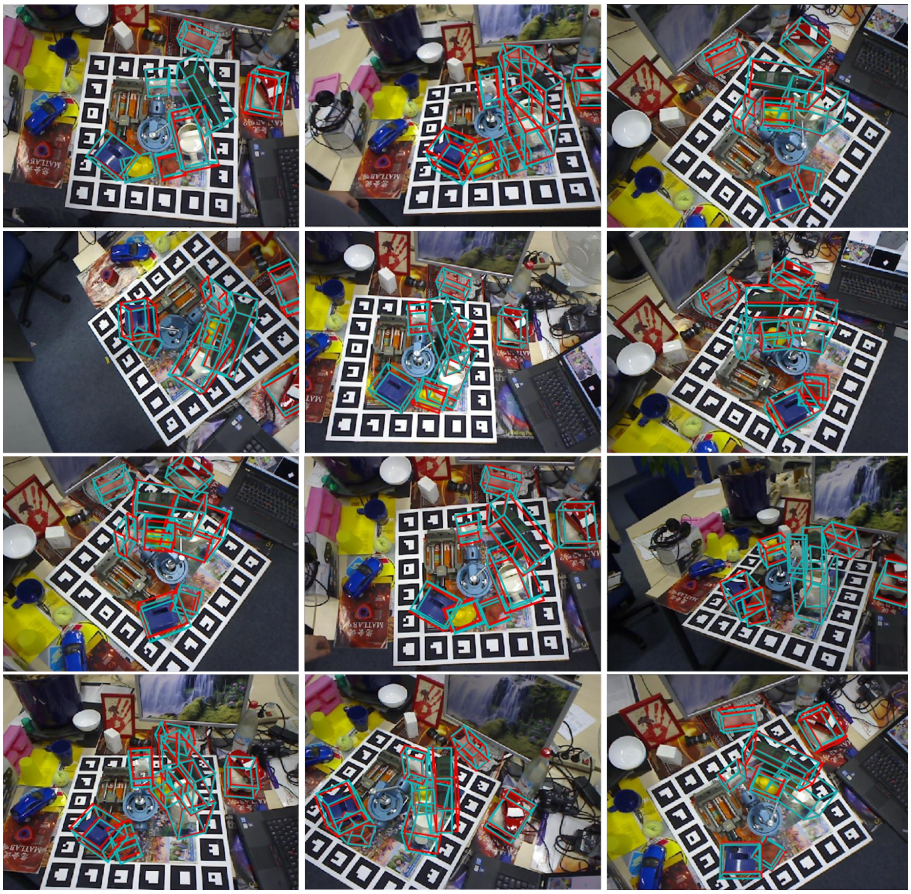

Fig. 3. Examples of 6D object pose estimation results on different objects from the Occlusion dataset. Red bounding boxes correspond to ground truth poses, cyan bounding boxes correspond to predicted poses using our approach.

References

1. Brachmann, E., Krull, A., Michel, F., Gumhold, S., Shotton, J., Rother, C.: Learning 6D object pose estimation using 3D object coordinates. In: European conference on computer vision. pp. 536–551. Springer (2014)
2. 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: Asian conference on computer vision. pp. 548–562. Springer (2012)
3. Kehl, W., Manhardt, F., Tombari, F., Ilic, S., Navab, N.: SSD-6D: Making RGB-based 3D detection and 6D pose estimation great again. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1521–1529 (2017)
4. Li, Y., Wang, G., Ji, X., Xiang, Y., Fox, D.: Deepim: Deep iterative matching for 6D pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 683–698 (2018)
5. Oberweger, M., Rad, M., Lepetit, V.: Making deep heatmaps robust to partial occlusions for 3D object pose estimation. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 119–134 (2018)
6. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6DoF pose estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4561–4570 (2019)
7. Rad, M., Lepetit, V.: 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. pp. 3828–3836 (2017)
8. Tekin, B., Sinha, S.N., Fua, P.: Real-time seamless single shot 6D object pose prediction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 292–301 (2018)
9. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. Robotics: Science and Systems (RSS) (2018)