Supplementary Material for Keypoint-graph-driven learning framework for object pose estimation

1 Detailed Analysis on the Quality of Object Detection

Since our method needs to crop the objects in images based on bounding boxes of objects predicted by object detection network, we present the detailed quantitative evaluation in this section. In our work, we use Faster-RCNN [10] with Resnet-101 [3] backbone to crop the objects in images for 6D pose estimation. To improve the performance of Faster-RCNN on real images, we pre-train Resnet-101 on a classification task on ImageNet [2] and freeze the Resnet-101 after block3 for follow-up training like in Hinterstoisser [4]. Besides, we follow the structure of domain adaptive Faster-RCNN [1] that uses two domain adaption components to reduce the domain discrepancy on image and instance level. The training images are same as the training images for DAKDN, where only the synthetic images have labels. To evaluate the object detection, we calculated IoU between the predicted bounding boxes and the ground-truth bounding boxes on LineMOD test set. From Table 1, we can see that the mean average percentage of correctly predicted 2D bounding boxes (IoU>0.5) is 84.3%. It shows that the object detection network can provide precise cropped objects for pose estimation

Table 1: Detailed evaluation of predicted bounding boxes. We report the mean average percentages of correct 2D bounding boxes (IoU>0.5) on LINEMOD test set

	Ape	Benchvise	Cam	Can	Cat	Driller	Duck	Eggbox	Glue	Holepuncher	Iron	Lamp	Phone	Mean
Accuracy	86.2	90.4	80.5	83.3	88.5	79.3	82.1	85.7	81.9	82.6	91.3	80.7	83.1	84.3

2 Sensitiveness analysis of the keypoint number

When analysing the sensitiveness of the keypoint number on pose estimation, we train our network to detect 5, 10, 15 and 20 keypoints, respectively. From table 2, we can find that the accuracy of pose estimation increases with the keypoint number. But the gap between ”10”, ”15” and ”20” is negligible. For reducing the computational cost and computing time, we use 10 keypoints for pose estimation.

Table 2: The mean ADD using different number of keypoints on the LINEMOD dataset.

Keypoint Number	5	10	15	20
Mean (ADD)	60.4	68.2	68.4	68.5

3 Evaluation results on LINEMOD dataset using different metrics

We also evaluate our method on LINEMOD using 2D Projection metric and $5cm5^\circ$ metric in Table 3. With the $5cm5^\circ$ metric, a pose is considered correct if the translation and rotation errors are below 5cm and 5° respectively. For the 2D Projection metric which measures pose error in 2D, we computes the mean distance between the projections of 3D model points given the estimated and the ground truth pose. The estimated pose is accepted if the distance is less than 5 pixels. The result shows that our method outperforms YOLO6D [13] by 1.63%, BB8 [9] by 8.5% and PoseCNN [15] by 22.2% in 2D projection metric. In $5cm5^\circ$ metric, our method outperforms PoseCNN [15] by 39.63%.

Table 3: The accuracies of our method and the baseline methods on the LINEMOD dataset using different metrics.

labels	w/o manual pose labels				w/ manual pose labels			
Training data	Syn		Syn+Real		Real			
Method	SSD-6D [6]	AAE [12]	Self6D [14]	Ours	YOLO6D [13]	BB8[9]	PoseCNN [15]	CDPN [7]
2D Projection	-	-	-	92.4	90.37	83.9	70.2	98.10
$5cm5^\circ$	-	-	-	56.33	-	-	19.4	94.31

4 Qualitative results

We show some more clear qualitative results of our method on LINEMOD dataset, OCCLUSION dataset and Homebrewed dataset in Figure 1, Figure 2 and Figure 3.

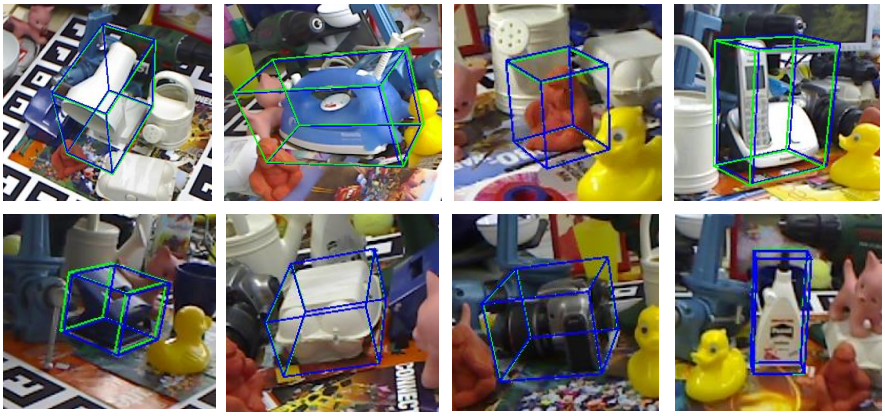

Figure 1: Qualitative results on LINEMOD dataset. The green bounding boxes correspond to the ground truth poses, and the blue bounding boxes to the poses estimated with our method.

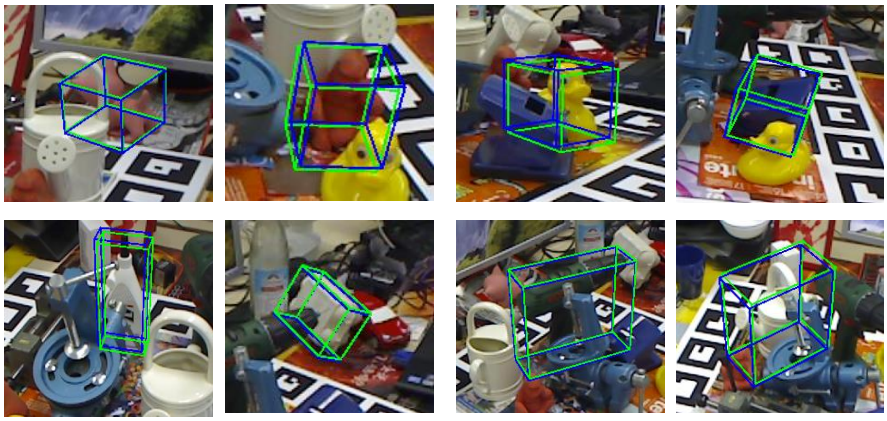

Figure 2: Qualitative results on OCCLUSION dataset. The green bounding boxes correspond to the ground truth poses, and the blue bounding boxes to the poses estimated with our method.

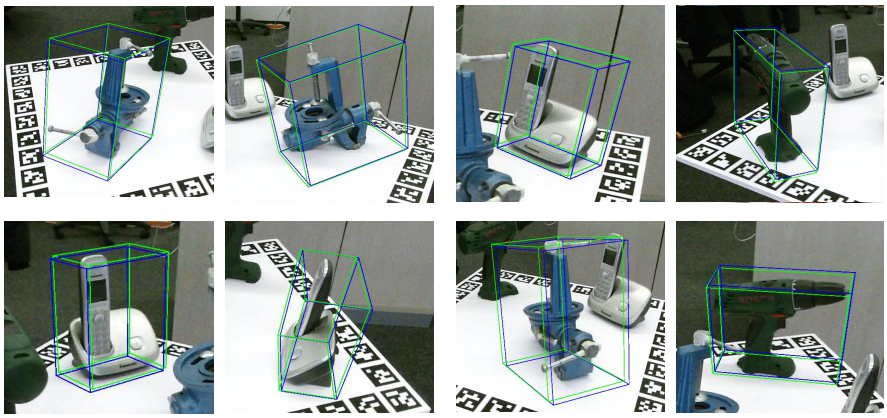

Figure 3: Qualitative results on Homebrewed dataset. The green bounding boxes correspond to the ground truth poses, and the blue bounding boxes to the poses estimated with our method.

5 The experimental results on T-LESS dataset

We also report the experimental results on the T-LESS dataset which is a particularly challenging 6D object detection benchmark containing texture-less, symmetric objects as well as clutter and serve occlusions using Visible Surface Discrepancy (VSD) as in BOP benchmark [5]. Table 4 shows the full T-LESS results on each object tested on all scene views of the Primesense test set. The result shows that our method outperforms the state-of-the-art methods that do not require manual pose labels and is comparable with pix2pose in some objects. It is because our method aligns the domain shift between synthetic and real images, and optimize the network for 6D pose estimation task by learning the domain-invariant structure as a constraint.

Table 4: The accuracy of our method and the baseline methods on the T-LESS dataset in terms of the ($e_{VSD} < 0.3, \tau = 20mm$) on all test scenes using PrimeSense. Results of AAE, MP-AAE and Pix2Pose are cited from their papers.

labels	w/o manual pose labels			w/ manual pose labels
Training data	Syn		Syn+Real	Real
Method	AAE [12]	MP-AAE [11]	Ours	pix2pose [8]
1	9.48	5.56	6.47	38.4
2	13.24	10.22	9.89	35.3
3	12.78	14.74	15.22	40.9
4	6.66	6.23	5.61	26.3
5	36.19	37.53	39.50	55.2
6	20.64	30.36	28.35	31.5
7	17.41	14.62	26.28	1.1
8	21.72	10.73	14.94	13.1
9	39.98	19.43	29.37	33.9
10	13.37	32.75	30.12	45.8
11	7.78	20.34	12.33	30.7
12	9.54	29.53	36.64	30.4
13	4.56	12.41	18.75	31.0
14	5.36	21.30	22.54	19.5
15	27.11	20.82	14.62	56.1
16	22.04	33.20	24.72	66.5
17	66.33	39.88	35.71	37.9
18	14.91	14.16	15.60	45.3
19	23.03	9.24	14.27	21.7
20	5.35	1.72	3.18	1.9
21	19.82	11.48	15.42	19.4
22	20.25	8.30	7.83	9.5
23	19.15	2.39	11.26	30.7
24	4.54	8.66	6.78	18.3
25	19.07	22.52	24.21	9.5
26	12.92	30.12	28.81	13.9
27	22.37	23.61	24.28	24.4
28	24.00	27.42	26.97	43.0
29	27.66	40.68	45.40	25.8
30	30.53	56.08	47.34	28.8
Mean	19.26	20.53	21.41	29.5

References

[1] Y. Chen, W. Li, C. Sakaridis, et al. Domain adaptive faster r-cnn for object detection in the wild. In *CVPR*, 2018. 1

[2] J. Deng, W. Dong, R. Socher, et al. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 1

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] S. Hinterstoisser, V. Lepetit, P. Wohlhart, and K. Konolige. On pre-trained image features and synthetic images for deep learning. In *ECCV*, 2018. 1

[5] T. Hodan, F. Michel, E. Brachmann, et al. Bop: Benchmark for 6d object pose estimation. In *ECCV*, 2018. 3

[6] W. Kehl, F. Manhardt, F. Tombari, S. Ilic, and N. Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, 2017. 2

[7] Z. Li, G. Wang, and X. Ji. Cdpn: Coordinates-based disentangled pose network for real-time rgb-based 6-dof object pose estimation. In *ICCV*, 2019. 2

[8] K. Park, T. Patten, and M. Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *ICCV*, 2019. 4

[9] M. Rad and V. Lepetit. Bb8: A scalable, accurate, robust to partial occlusion method for predicting the 3d poses of challenging objects without using depth. In *ICCV*, 2017. 2

[10] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: towards real-time object detection with region proposal networks. In *NIPS*, 2015. 1

[11] M. Sundermeyer, M. Durner, E. Puang, et al. Multi-path learning for object pose estimation across domains. In *CVPR*, 2020. 4

[12] M. Sundermeyer, Z. Marton, M. Durner, M. Brucker, and R. Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *ECCV*, 2018. 2, 4

[13] B. Tekin, S. Sinha, and P. Fua. Real-time seamless single shot 6d object pose prediction. In *CVPR*, 2018. 2

[14] G. Wang, F. Manhardt, J. Shao, X. Ji, et al. Self6d: Self-supervised monocular 6d object pose estimation. In *ECCV*, 2020. 2

[15] Y. Xiang, T. Schmidt, et al. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2