Supplementary Material for
Learning deep network for detecting 3D object keypoints and 6D poses

1. $5cm5^\circ$ results on LINEMOD dataset

We use the $\mathbf{5cm5^\circ}$ metric to evaluate our method in real world. With this metric, a pose is considered correct if the translation and rotation errors are below 5cm and 5° respectively.

Table 1. The pose detection accuracy ($\mathbf{5cm5^\circ}$) on the LINEMOD dataset for single object category.

Training data	RGB with Relative Transformation		RGB with 3D CAD Models			RGB with 3D Annotation	
	OK-POSE	keypointnet+bbox	keypointnet+mask	SSD6D	AAE	Brachmann	BB8
Ape	38.6	9.2	26.3	-	-	34.4	80.2
Benchvise	29.2	13.3	8.4	-	-	40.6	81.5
Cam	38.6	6.3	14.3	-	-	30.5	60.0
Can	28.4	3.6	12.4	-	-	48.4	76.8
Cat	34.7	6.4	17.8	-	-	34.6	79.9
Driller	31.8	7.6	20.3	-	-	54.5	69.6
Duck	31.4	5.2	21.5	-	-	22.0	53.2
Eggbox	38.1	5.5	17.9	-	-	57.1	81.3
Glue	35.3	4.3	14.5	-	-	23.6	54.0
Holepuncher	16.8	14.8	8.6	-	-	47.3	73.1
Iron	34.6	6.1	17.4	-	-	58.7	61.1
Lamp	38.9	4.4	21.4	-	-	49.3	67.5
Phone	21.6	12.3	15.9	-	-	26.8	58.6
Mean	32.15	7.4	16.7	-	-	40.6	69.0

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, and the FPS will drop obviously. For trade-off between accuracy and FPS, we use 10 keypoints for pose estimation.

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

Keypoint Number	5	10	15	20
Mean (ADD)	23.28	30.16	31.20	31.81
FPS	22	18	15	11

3. Qualitative results on LINEMOD dataset

We show some more clear qualitative results of our method on LINEMOD dataset which demonstrates that our method can get accurate poses in cluttered scenes.

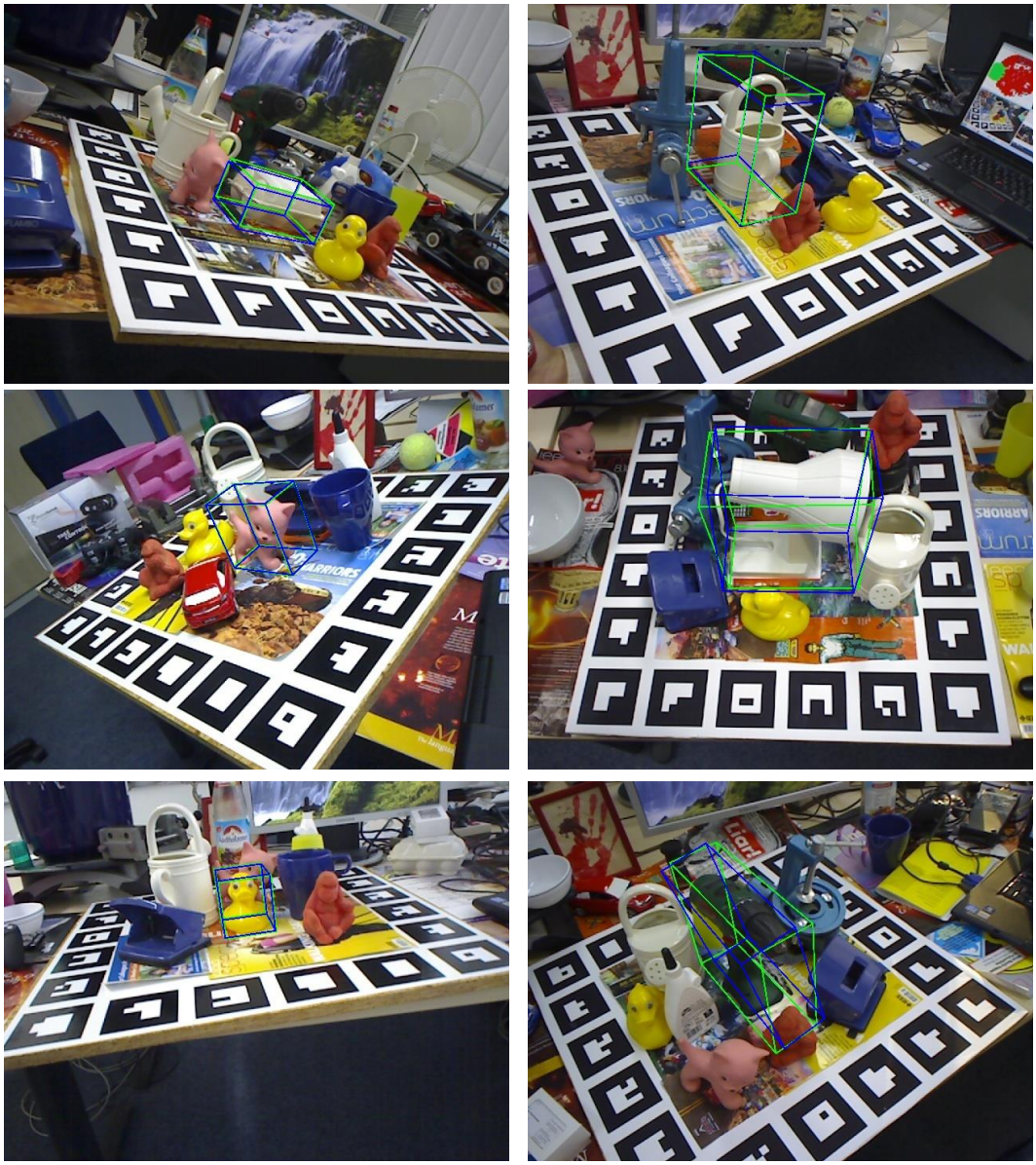

Figure 1. Qualitative results for the single object pose estimation on the LINEMOD dataset. The green bounding boxes correspond to the ground truth poses, and the blue bounding boxes to the poses estimated with our method.

4. Keypoint detection results on occlused objects

Figure 2. Keypoint detection results on occlused objects. Our method learn the geometric relationship between keypoints, so that the occluded keypoints can be robustly recovered by the visible keypoints.

5. Qualitative results on OCCLUSION dataset

We also provide some examples on OCCLUSION dataset by our method.

Figure 3. Qualitative results for the pose estimation on the multiple objects dataset.

6. Detailed 2D metric result on OCCLUSION dataset

We provide the detailed accuracy using 2D pose metric result by our method on each category of OCCLUSION dataset.

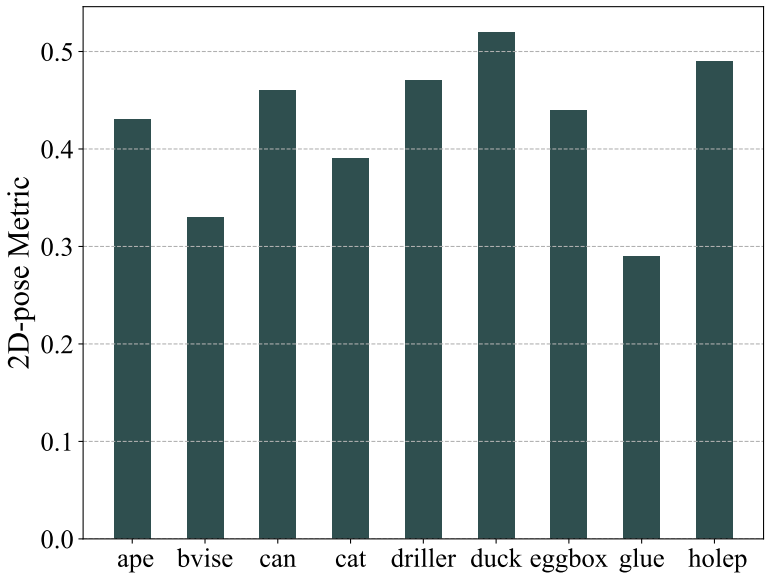

Figure 4. The accuracy (2D-pose metric) of 9 categories on OCCLUSION dataset.

7. Examples of keypoint detection in unseen background.

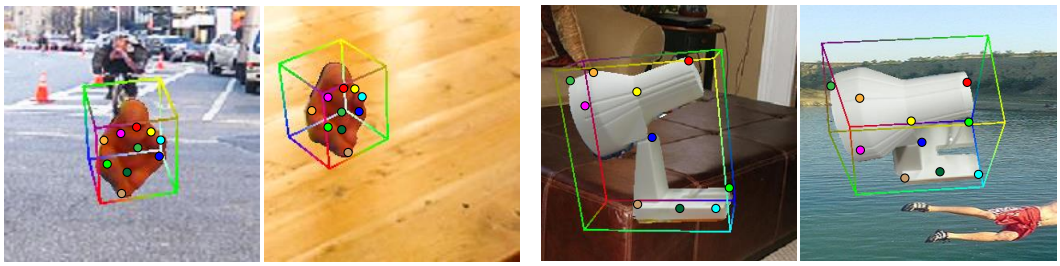

Figure 5. Examples of detection in unseen background.

The input of our keypoint branch is the Region of Interest (ROI) which are trained to focus on objects without the background. This will significantly alleviate the influence of the background. On the other hand, the cross-view consistency loss and distinctiveness loss will also encourage the network to find consistent points in different views and backgrounds. As shown in Fig 5, our method is still able to detect correctly in unseen backgrounds.

8. The reference image IDs in LineMOD

We report the picked 1, 3, 9 image IDs in each sequence as follows:

ape

1: *000200.png*, **3:** *000200.png 000521.png 000393.png*, **9:** *000200.png 000294.png 000521.png 000693.png 000826.png 000939.png 000966.png 001082.png 001155.png*

benchvise

1: *000349.png*, **3:** *000349.png 000644.png 000766.png*, **9:** *000110.png 000247.png 000381.png 000517.png 000644.png 000766.png 000881.png 00963.png 001048.png*

cam

1: *000380.png*, **3:** *000380.png 000600.png 000896.png*, **9:** *000179.png 000380.png 000471.png 000600.png 000708.png 000760.png 000896.png 00988.png 001046.png*

can
1: *000141.png*, **3:** *000141.png 000694.png 001125.png*, **9:** *000141.png 000458.png 000470.png 000694.png 000741.png 0001004.png 0001082.png 001125.png 001145.png*
cat
1: *000039.png*, **3:** *000039.png 000704.png 000859.png*, **9:** *000039.png 000344.png 000363.png 000494.png 000521.png 000704.png 000771.png 000859.png 001057.png*
driller
1: *000087.png*, **3:** *000087.png 000619.png 000962.png*, **9:** *000087.png 000200.png 000305.png 000438.png 000507.png 000619.png 000763.png 000962.png 001070.png*
duck
1: *000190.png*, **3:** *000190.png 000661.png 0001033.png*, **9:** *000190.png 000278.png 000359.png 000486.png 000620.png 000661.png 000687.png 000832.png 001033.png*
eggbox
1: *000411.png*, **3:** *000411.png 000302.png 001209.png*, **9:** *000199.png 000302.png 000411.png 000562.png 000674.png 000778.png 000953.png 001035.png 001209.png*
glue
1: *000298.png*, **3:** *000298.png 000483.png 001069.png*, **9:** *000266.png 000298.png 000351.png 000483.png 000583.png 000672.png 000884.png 000914.png 001069.png*
holepuncher
1: *000224.png*, **3:** *000224.png 000535.png 000926.png*, **9:** *000224.png 000236.png 000327.png 000547.png 000535.png 000668.png 000890.png 000926.png 001058.png*
iron
1: *000746.png*, **3:** *000746.png 000641.png 001142.png*, **9:** *000257.png 000437.png 000641.png 000746.png 000782.png 000836.png 000932.png 001012.png 001142.png*
lamp
1: *000459.png*, **3:** *000459.png 000722.png 000949.png*, **9:** *000109.png 000299.png 000459.png 000643.png 000722.png 000866.png 000949.png 001023.png 001120.png*
phone
1: *000061.png*, **3:***000061.png 000615.png 001002.png*, **9:** *000061.png 000173.png 000386.png 000452.png 000615.png 000721.png 000801.png 001002.png 001089.png*

9. The experimental results on T-LESS dataset

We also report the experimental results on T-LESS dataset. For T-LESS, we use Visible Surface Discrepancy (VSD) as a metric which also is employed by AAE and Pix2Pose.

Table 3. The accuracies 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 and Pix2Pose are cited from their papers.

Training data	RGB with Relative Transformation	RGB with 3D CAD Models	RGB with 3D Annotation
Method	OK-POSE	AAE	Pix2Pose
Mean	27.8	18.35	29.5