Supplementary material
Pix2Pose: Pixel-Wise Coordinate Regression of Objects for 6D Pose Estimation

1. Detail parameters

1.1. Data augmentation for training

Add(each channel)	Contrast normalization	Multiply	Gaussian Blur
$\mathcal{U}$(-15, 15)	$\mathcal{U}$(0.8, 1.3)	$\mathcal{U}$(0.8, 1.2)(per channel chance=0.3)	$\mathcal{U}$(0.0, 0.5)

Table 1. Color augmentation

Type	Random rotation	Fraction of occluded area	
Dataset	All	LineMOD	LineMOD Occlusion, T-Less
Range	$\mathcal{U}$(-45°, -45°)	$\mathcal{U}$(0, 0.1)	$\mathcal{U}$(0.04, 0.5)

Table 2. Occlusion and rotation augmentation

1.2. The pools of symmetric poses for the transformer loss

I: Identity matrix, R_a^{Θ}: Rotation matrix about the a-axis with an angle Θ.

- LineMOD and LineMOD Occlusion - eggbox and glue: $sym = [I, R_z^{\pi}]$
- T-Less - obj-5,6,7,8,9,10,11,12,25,26,28,29: $sym = [I, R_z^{\pi}]$
- T-Less - obj-19,20: $sym = [I, R_y^{\pi}]$
- T-Less - obj-27: $sym = [I, R_z^{\frac{\pi}{2}}, R_z^{\pi}, R_z^{\frac{3\pi}{2}}]$
- T-Less - obj-1,2,3,4,13,14,15,16,17,18,24,30: $sym = [I]$, the z-component of the rotation matrix is ignored.
- Objects not in the list (non-symmetric): $sym = [I]$

1.3. Pose prediction

- Definition of non-zero pixels: $||I_{3D}||_2 > 0.3$, I_{3D} in normalized coordinates.
- PnP and RANSAC algorithm: the implementation in OpenCV 3.4.0 [1] is used with default parameters except the re-projection threshold θ_{re}= 3.
- List of outlier thresholds

	ape	bvise	cam	can	cat	driller	duck	eggbox	glue	holep	iron	lamp	phone
θ_o	0.1	0.2	0.2	0.2	0.2	0.2	0.1	0.2	0.2	0.2	0.2	0.2	0.2

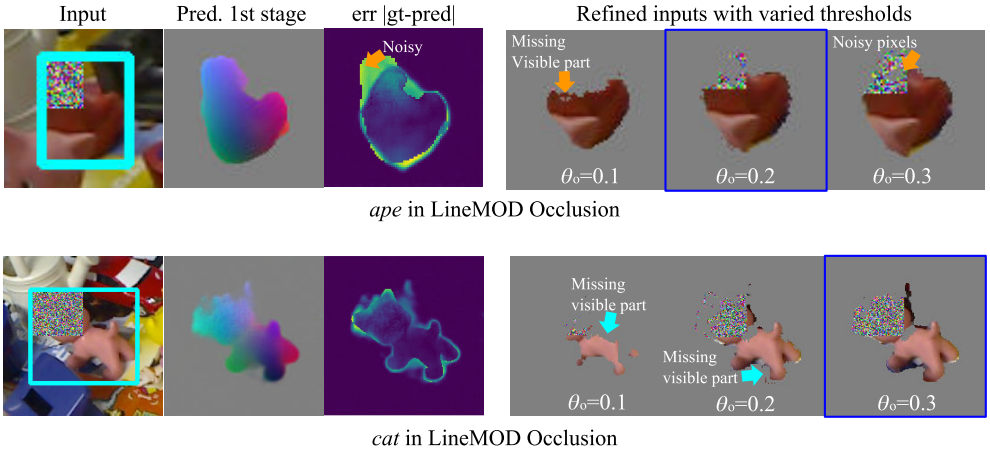
Table 3. Outlier thresholds θ_o for objects in LineMOD

	ape	can	cat	driller	duck	eggbox	glue	holep
θ_o	0.2	0.3	0.3	0.3	0.2	0.2	0.3	0.3

Table 4. Outlier thresholds θ_o for objects in LineMOD Occlusion

	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
θ_o	0.1	0.1	0.1	0.3	0.2	0.3	0.3	0.3	0.3	0.2	0.3	0.3	0.2	0.2	0.2	0.3	0.3	0.2	0.3	0.3	0.2	0.2	0.3	0.1	0.3	0.3	0.3	0.3	0.3	0.3

Table 5. Outlier thresholds θ_o for objects in T-Less



ape in LineMOD Occlusion

cat in LineMOD Occlusion

Figure 1. Examples of refined inputs in the first stage with varied values for the outlier threshold. Values are determined to maximize the number of visible pixels while excluding noisy predictions in refined inputs. Training images are used with artificial occlusions. The brighter pixel in images of the third column represents the larger error.

2. Details of evaluations

2.1. T-Less: Object-wise results

Obj.No	01	02	03	04	05	06	07	08	09	10
VSD Recall	38.4	35.3	40.9	26.3	55.2	31.5	1.1	13.1	33.9	45.8
Obj.No	11	12	13	14	15	16	17	18	19	20
VSD Recall	30.7	30.4	31.0	19.5	56.1	66.5	37.9	45.3	21.7	1.9
Obj.No	21	22	23	24	25	26	27	28	29	30
VSD Recall	19.4	9.5	30.7	18.3	9.5	13.9	24.4	43.0	25.8	28.8

Table 6. Object reall ($e_{\text{vsd}} < 0.3, \tau = 20mm, \delta = 15mm$) on all test scenes of Primesense in T-Less. Objects visible more than 10% are considered. The bounding box of an object with the highest score is used for estimation in order to follow the test protocol of 6D pose benchmark [2].

2.2. Qualitative examples of the transformer loss

Figure 2 and Figure 3 present example outputs of the Pix2Pose network after training of the network with/without using the transformer loss. The *obj-05* in T-less is used.

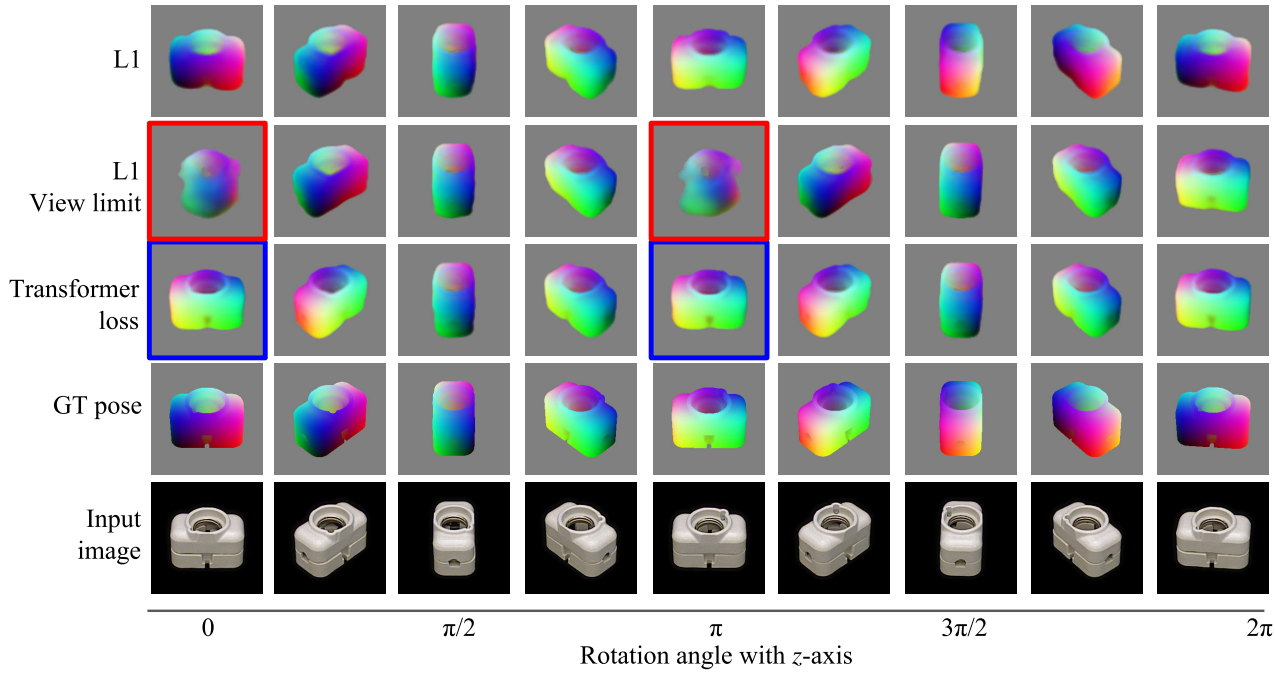


Figure 2. Prediction results of varied rotations with the z-axis. As discussed in the paper, limiting a view range causes noisy predictions at boundaries, 0 and π, as denoted with red boxes. The transformer loss implicitly guides the network to predict a single side consistently. For the network trained by the L1 loss, the prediction is accurate when the object is fully visible. This is because the upper part of the object provides a hint for a pose.

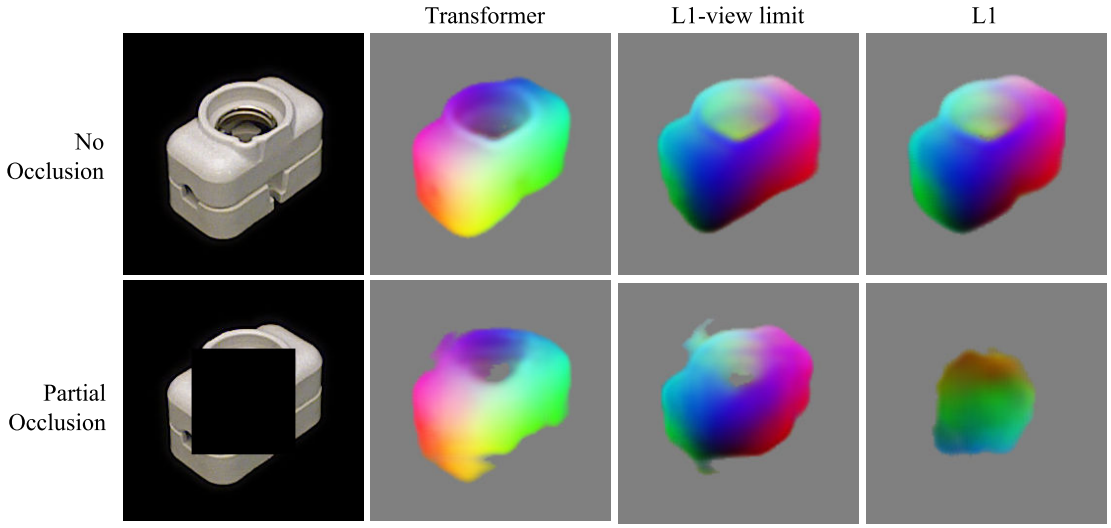


Figure 3. Prediction results with/without occlusion. For the network trained by the L1 loss, it is difficult to predict the exact pose when the upper part, which is a clue to determine the pose, is not visible. The prediction of the network using the transformer loss is robust to this occlusion since the network consistently predicts a single side.

2.3. Example results on LineMOD

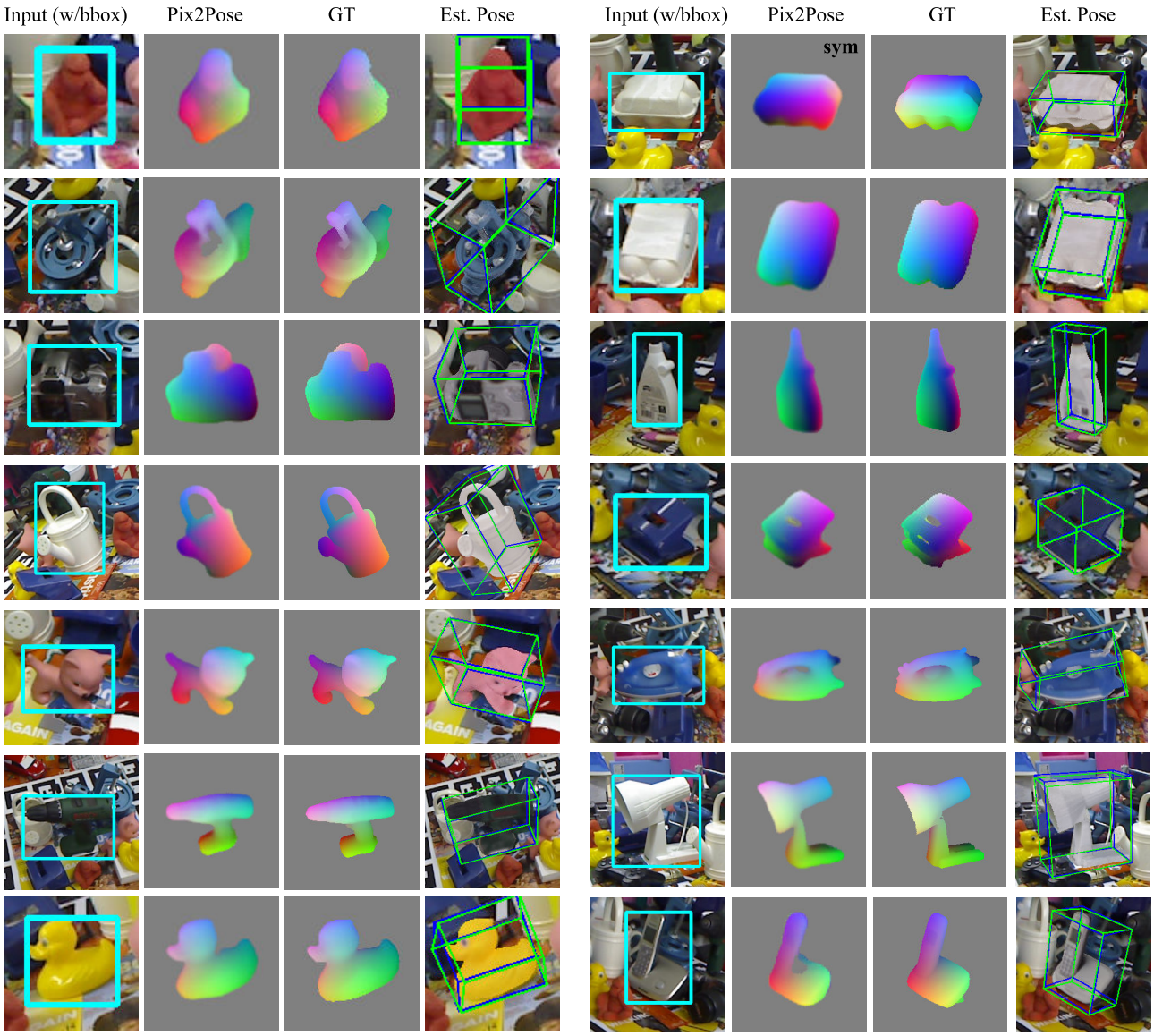


Figure 4. Example results on LineMOD. The result marked with *sym* represents that the prediction is the symmetric pose of the ground truth pose, which shows the effect of the proposed transformer loss. Green: 3D bounding boxes of ground truth poses, blue: 3D bounding boxes of predicted poses.

2.4. Example results on LineMOD Occlusion

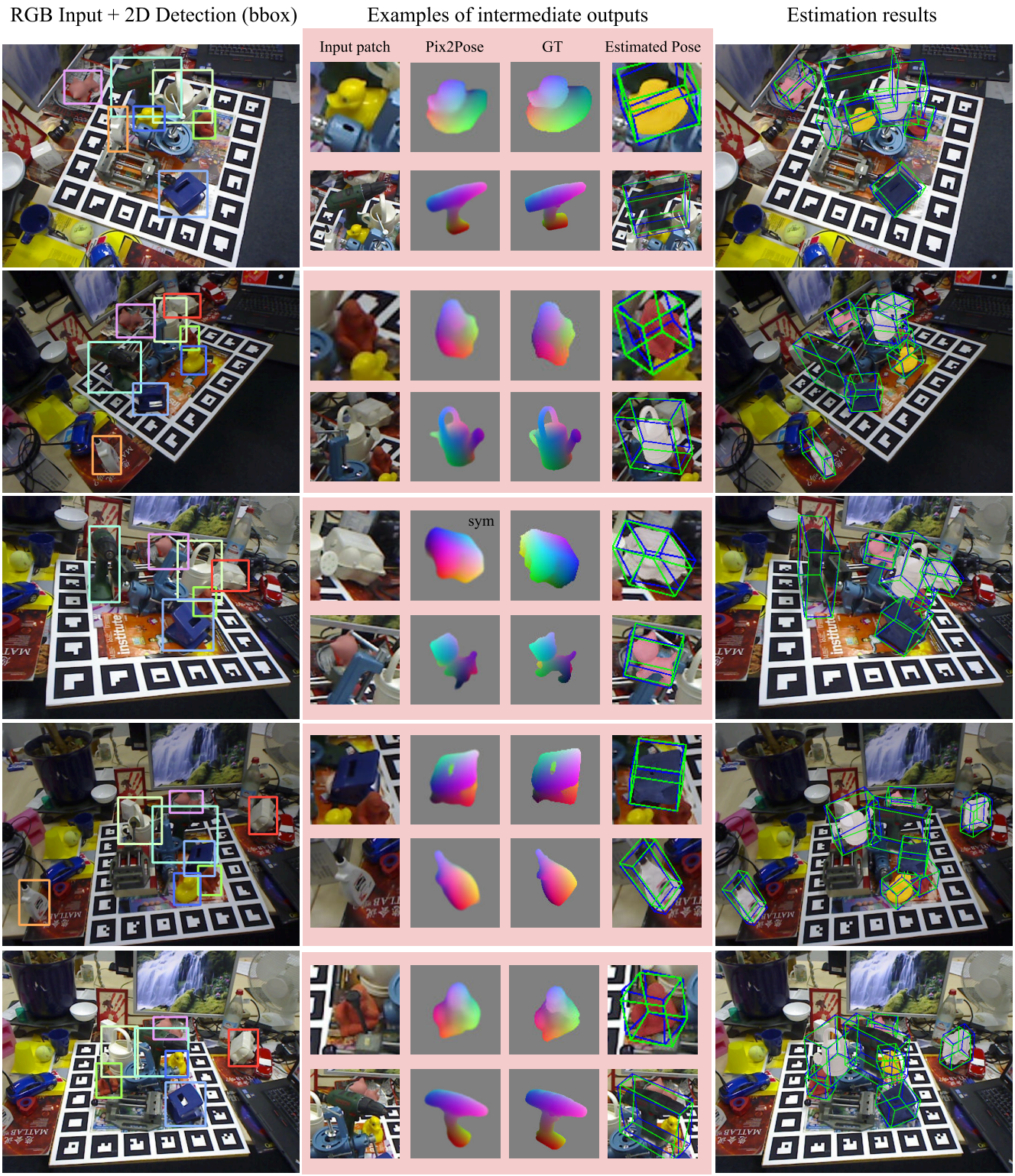


Figure 5. Example results on LineMOD Occlusion. The precise prediction of occluded parts enhances robustness.

2.5. Example results on T-Less

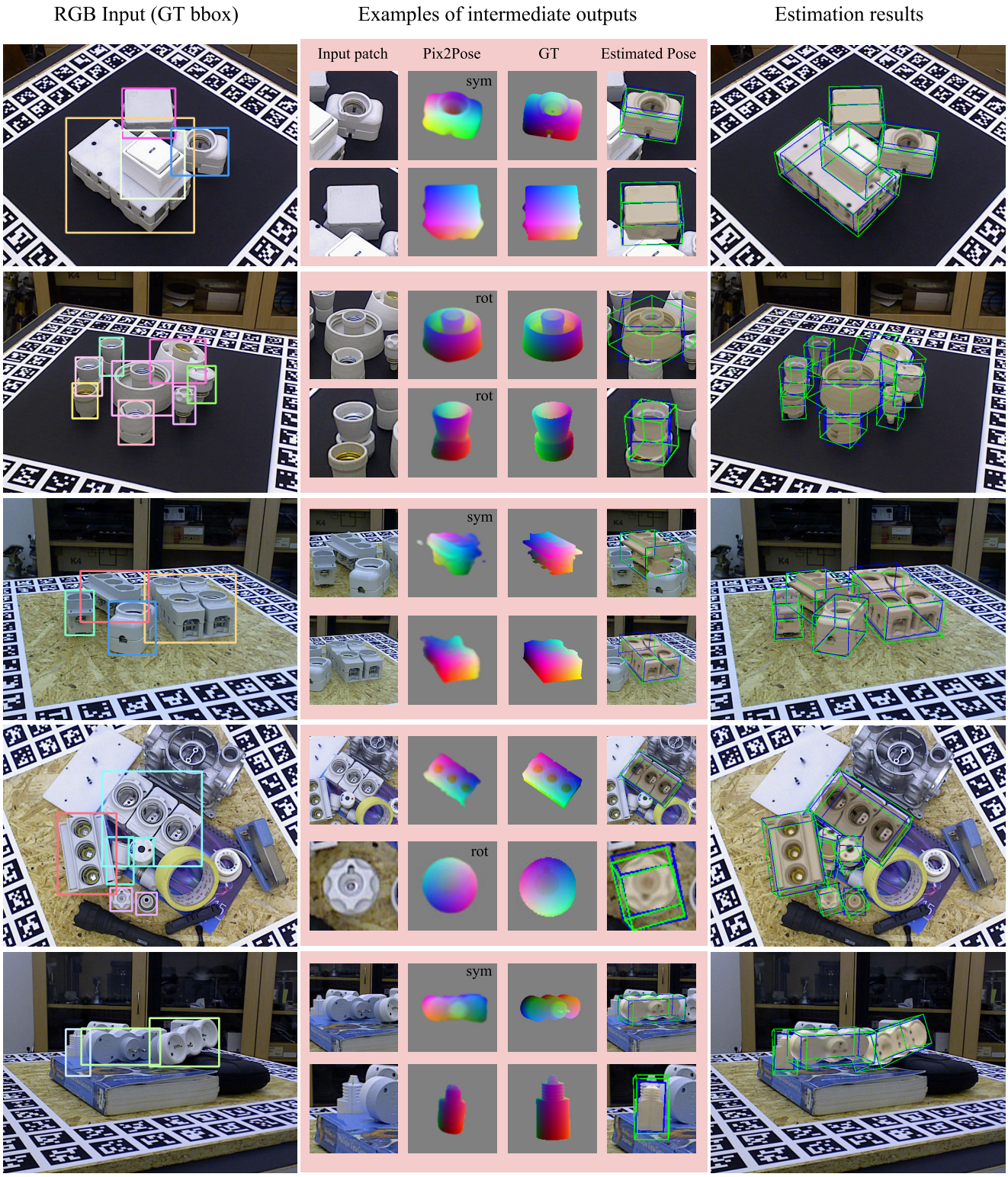


Figure 6. Example results on T-Less. For visualization, ground-truth bounding boxes are used to show pose estimation results regardless of the 2D detection performance. Results with *rot* denote estimations of objects with cylindrical shapes.

3. Failure cases

Primary reasons of failure cases: (1) Poses that are not covered by real training images and the augmentation. (2) Ambiguous poses due to severe occlusion. (3) Not sufficiently overlapped bounding boxes, which cannot be recovered by the bounding box adjustment in the first stage. The second row of Fig. 7 shows that the random augmentation of in-plane rotation during the training is not sufficient to cover various poses. Thus, the uniform augmentation of in-plane rotation has to performed for further improvement.

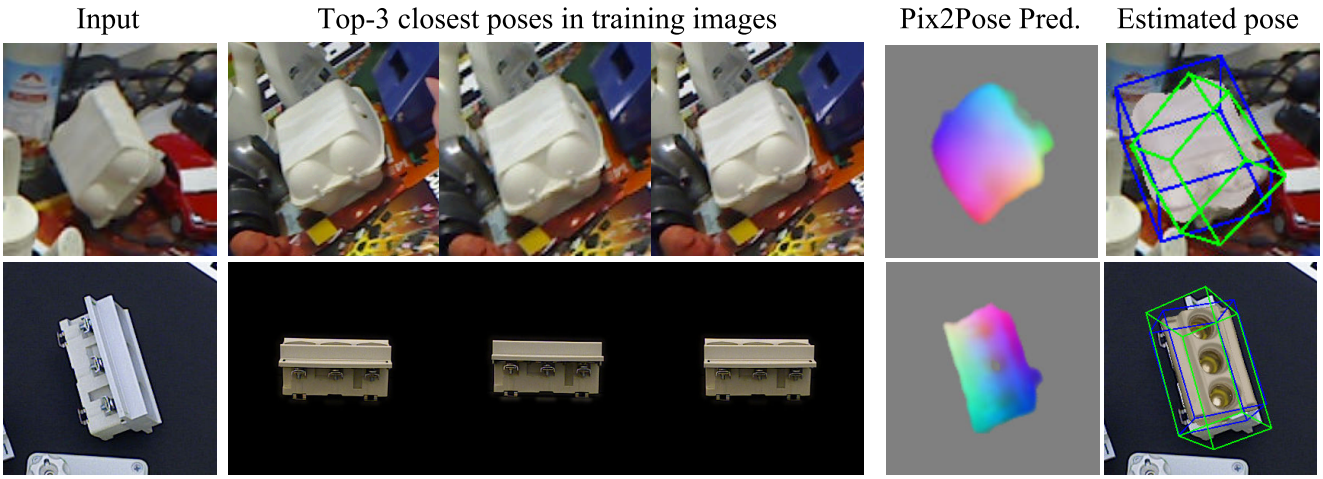


Figure 7. Examples of failure cases due to unseen poses. The closest poses are obtained from training images using geodesic distances between two transformations (rotation only).

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 1

[2] T. Hodan, F. Michel, E. Brachmann, W. Kehl, A. GlentBuch, D. Kraft, B. Drost, J. Vidal, S. Ihrke, X. Zabulis, C. Sahin, F. Manhardt, F. Tombari, T.-K. Kim, J. Matas, and C. Rother. Bop: Benchmark for 6d object pose estimation. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2