Supplementary

1. Direct 9D Pose Estimation and Segmentation

In our experiments, we show that our model can work with both segmentation and bounding box masks. Can we locate objects with our voting scheme directly (i.e., without any preprocessing instance detection pipeline)? For some categories like laptop, this is hard since the laptop base can be mixed with the floor in view of point clouds. However, for bowls, this is possible, where the pose estimation is indeed zero-shot without seeing any real-world data during the whole pipeline. We modify Algorithm 1 by sampling point pairs from the whole scene, while keeping the coarse-to-fine procedure. We also use a threshold to generate candidate object locations instead of a simple *argmax* (line 9 in Algorithm 1).

Zero-shot instance segmentation Surprisingly, as a by-product, our method is able to infer the instance-level mask without even seeing any segmentation labels. This is done by counting the contribution of each point, where any point that votes more than v times within a small radius ϵ of the true object center (line 15 in Algorithm 1), is considered lying on the target object. Qualitative 9D pose and segmentation results are given in Figure 1. Quantitative results are listed in Table 1.

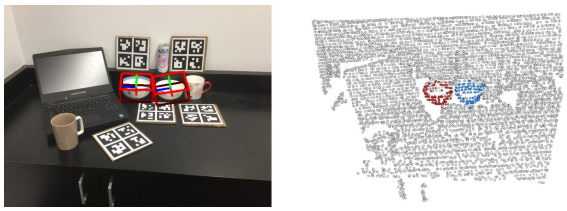

Figure 1. **Our 9D pose prediction and instance segmentation results on NOCS REAL275 test dataset.** Notice that we do not leverage existing instance segmentation models, and the segmentation is generated by our voting model, which is trained on synthetic objects only.

2. More Quantitative Results on SUN RGB-D Dataset

We conduct more zero-shot pose estimation experiments on SUN RGB-D, with the instance masks annotated by SUN RGB-D. Notice that our model is trained solely on ShapeNet synthetic models, and then directly tested on SUN RGB-D frames. Results are listed in Table 2. Qualitative results are given in Figure 2.

	mAP (%)				
	$3D_{25}$	$3D_{50}$	5° 5 cm	10° 5 cm	15° 5 cm
Bowl	43.0	6.9	0.8	10.1	22.6

Table 1. **Zero-shot 9D pose estimation without detection priors.** We report the mAP results for bowls in real-world scenarios with only synthetic training data.

	mAP (%)		
	20° 10 cm	40° 20 cm	60° 30 cm
Bathtub	10.9	38.8	49.8
Bookshelf	0.0	1.0	6.4
Bed	0.0	0.8	3.4
Sofa	0.0	1.7	10.5
Table	0.5	6.9	17.5

Table 2. **Zero-shot pose estimation results using instance masks provided by SUN RGB-D.** Rotation error is evaluated along the gravity axis.

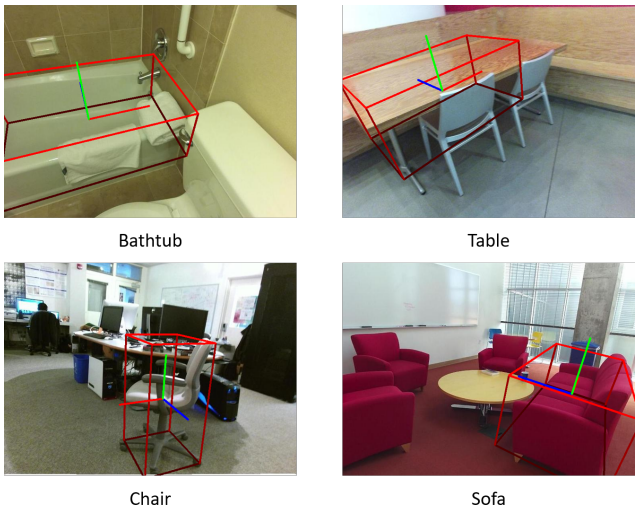

Figure 2. **Some successful pose estimations on SUN RGB-D.** Notice that our model is trained on synthetic ShapeNet objects only.

3. More Ablation Studies

In our sim-to-real pipeline, point clouds are first voxelized and jittered in order to generate the same distribution for both synthetic and real objects. Table 3 gives the ablation studies on these two techniques, where we see that both voxelization and random jittering helps improve the final detection result.

	mAP (%)				
	$3D_{25}$	$3D_{50}$	5° 5 cm	10° 5 cm	15° 5 cm
No Voxelization	76.7	21.6	11.9	37.2	45.6
No Jittering	77.1	24.6	16.2	43.1	49.7
Ours (full)	**78.2**	**26.4**	**16.9**	**44.9**	**50.8**

Table 3. **Ablation results on NOCS REAL275 test set.**

4. Detailed results on each category

We plot detailed comparisons of each category on NOCS REAL275 test set. Rotation AP is given in Figure 3 and Translation AP is shown in Figure 4. Our model achieves the best result on most categories, and outperforms previous self-supervised methods by a large margin.

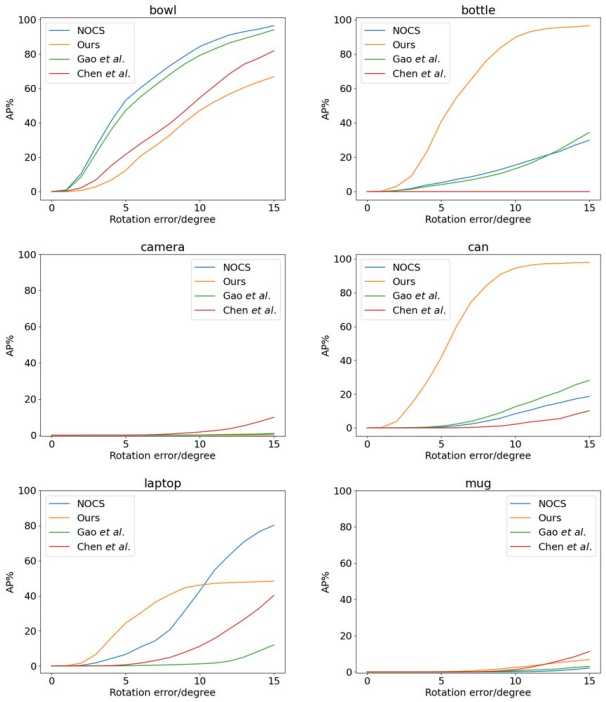

Figure 3. **Rotation AP for each category on NOCS REAL275 test dataset.**

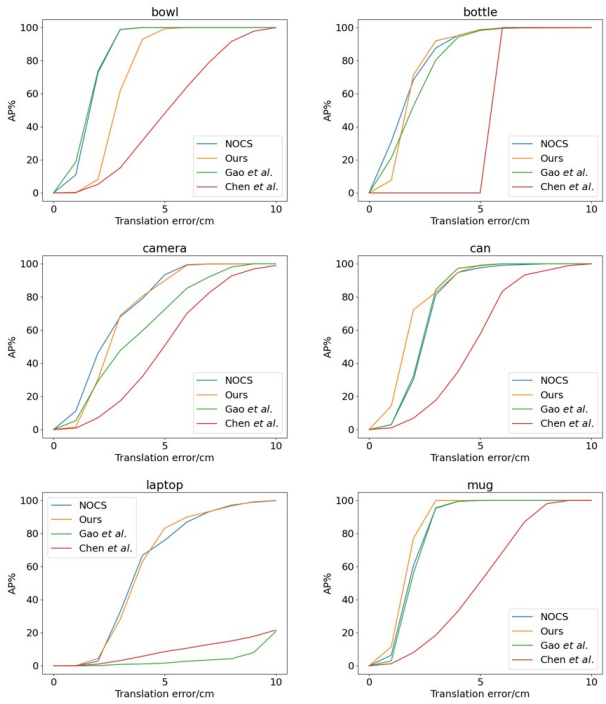

Figure 4. **Translation AP for each category on NOCS REAL275 test dataset.**