SA6D: Self-Adaptive Few-Shot 6D Pose Estimator for Novel and Occluded Objects Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

In appendix, we show additional qualitative results on the predicted target object segmentation including severe occlusion and truncation, qualitative results on the final 6D pose estimation against Gen6D, comparison between ICP and learning-based point cloud registration method, additional ablation studies, and further explanation and analysis on existing methods compared to our method, e.g., the selection of reference images, the effort of annotation, and the practical use case. We also submit a video introducing SA6D in the supplementary material.

A Additional Results

A.1 Gen6D without ground-truth object diameter.

In Tab. 4, we demonstrate that using the object diameter as input is a strong prior knowledge which limits the generalization of Gen6D, by fixing the diameter over all objects with two different values, namely $10\ cm$ and $50\ cm$. Without ground-truth diameter, Gen6D cannot generalize well on any of the datasets.

Diam. (m)	Dataset			
	LM	LMO	FewSOL	HB
0.1	0.06	0.06	0.04	0.10
0.5	0.16	0.05	0.00	0.19
GT	**0.35**	**0.08**	**0.36**	**0.30**

Table 4: Evaluation on Gen6D with different object diameters as prior knowledge. Results are averaged over objects for each dataset.

A.2 Compare ICP with learning-based point cloud registration algorithm

We show a few predicted examples of a state-of-the-art learning-based point cloud registration model, namely RPM-Net, on the LineMOD-OCC/driller in Fig. 6. RPM-Net is prone to the local optimal position for 6D pose estimation, especially for rotation. In our experiments, ICP is more robust to unseen objects.

A.3 SA6D is robust to false positive samples in reference

Using reprojected object center to select positive segments sometimes leads to a false positive sample given the target object center is occluded. An example is shown in Fig. 7a, in which a wrong segment (*yellow rabbit*) is selected as a positive sample for the target object (*milk cow*). However, we find that our online self-adaptation module is robust against false positive samples and is able to learn a correct target-oriented representation. Moreover, SA6D provides explainable confidence

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

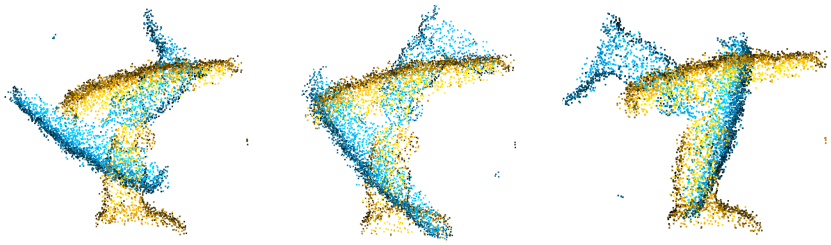
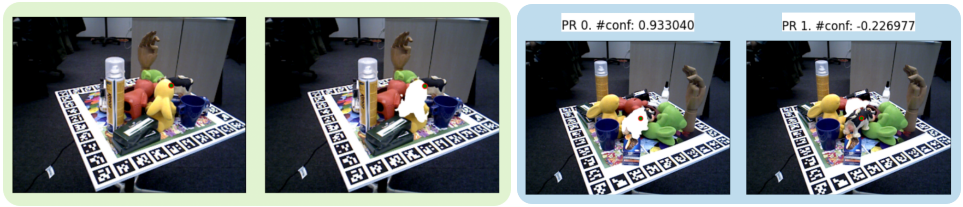


Figure 6: Examples of using RPM-Net for point cloud registration instead of ICP. The yellow point cloud denotes the reconstructed object point cloud model and the blue one denotes the prediction after transformation using the predicted pose from RPM-Net. Better overlapping between two point clouds indicates better performance. RPM-Net cannot generalize on unseen objects and is prone to get stuck in local optima.

scores by computing the cosine similarity between each segment representation and the target object representation. Fig. 7b shows an example of the predicted target (*milk cow*) segments with reasonable induced confidence score though wrong positive samples are given in the reference set.



(a) False positive sample.

(b) Robust prediction.

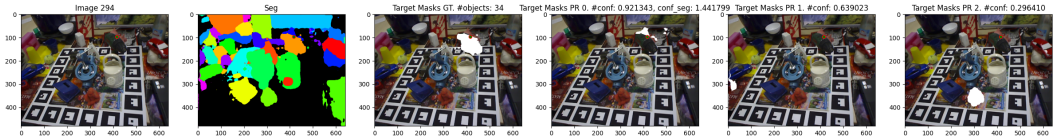
Figure 7: Discussion. (a) A false positive sample is selected given the reprojected center of the target object (*milk cow*) is occluded by another object (*yellow rabbit*). Even though, (b) SA6D provides robust prediction with explainable confidence scores.

A.4 SA6D demonstrates remarkable performance against severe occlusion and truncation

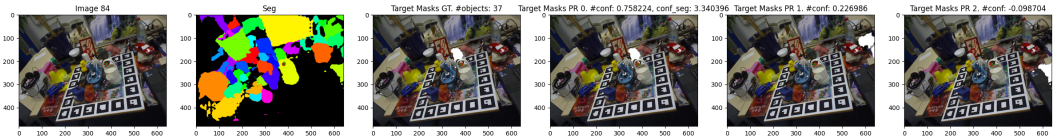
We show superior performance of SA6D on challenging scenes with severe occlusion and truncation in Fig. 8, where the input images, predicted segmentations from the base segmentor φ, ground-truth segmentation of target object based on the reprojected object center, and three predicted candidates with the highest predicted confidence scores are given on each column from left to right. The selected segments are marked in white color. The confidence score *conf* denotes the cosine similarity between the candidate segment representation and the target object representation r^*. The *conf_seg* is computed by dividing the confidence scores between the first and second most similar segment candidates w.r.t. the target object representation. Thus, it can be used in crucial scenarios if the prediction is uncertain among different segments. Note that in Fig. 8a, our method is able to differentiate the target object segment while the provided ground-truth segmentation points to a wrong segment due to the center of target object is occluded.

A.5 Robust and explainable confidence score of the online self-adaptation module

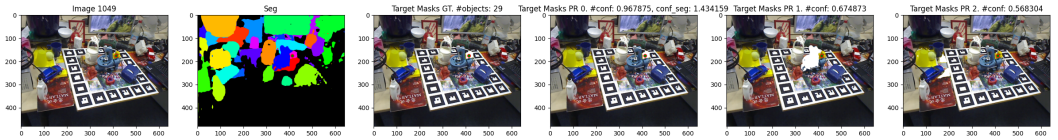
We show more results on the predicted segmentation of the online self-adaptation module in Fig. 9 on LineMOD dataset, Fig. 10 on LineMOD-OCC, and Fig. 11 on HomebrewedDB. Some candidates in Fig. 11 with white background indicate the background segments are selected.



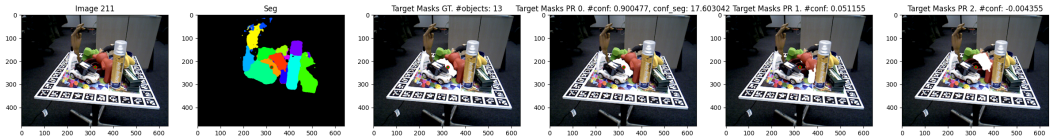
(a) cat



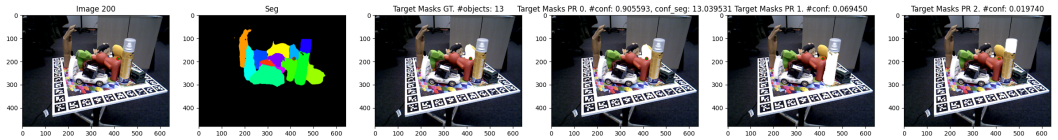
(b) driller



(c) duck



(d) flange



(e) yellow rabbit

Figure 8: Online-Adaptation results on challenging scenes against severe occlusion and truncation. Three candidates with the highest confidence scores are visualized in order.

A.6 More Qualitative Results

We show more qualitative results of the 6D pose prediction and compare our method with Gen6D on LineMOD dataset in Fig. 13, LineMOD-OCC in Fig. 14, HomebrewedDB in Fig. 15 and FewSOL in Fig. 16. The comparison on Wild6D dataset between SA6D and category-level SOTA method RePoNet is shown in Fig. 17.

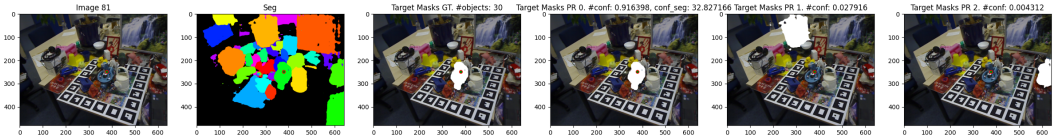
A.7 Failure Cases

We show the examples in Fig. 12 where using ICP leads to a worse prediction than without using ICP in the refinement module. Results are evaluated on the FewSOL dataset, indicating future work on generalizable and learnable point cloud registration is essential to further improve the performance.

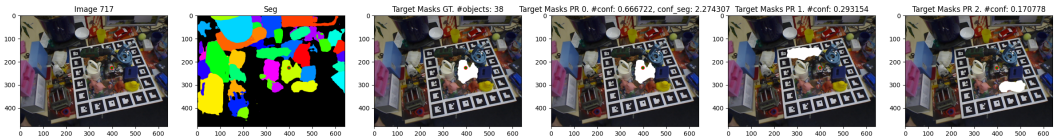
B Additional Explanation

B.1 Selection of Reference Images

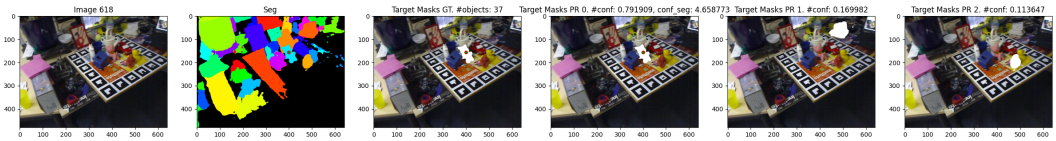
Regarding the selection of the reference images on the LM, LM-O and HB datasets, the original Gen6D selects 64 reference images from a predefined set of images with farthest point sampling (FPS) to make sure that the view distributes evenly among the reference images. We follow the



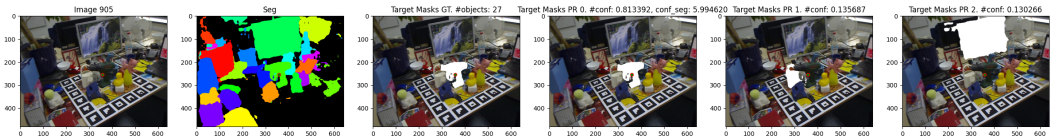
(a) benchvise



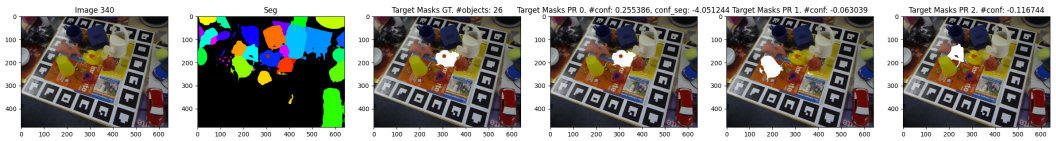
(b) cam



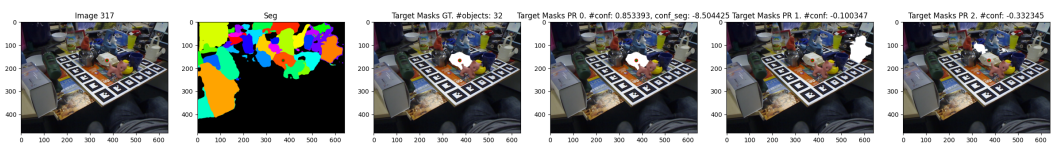
(c) cat



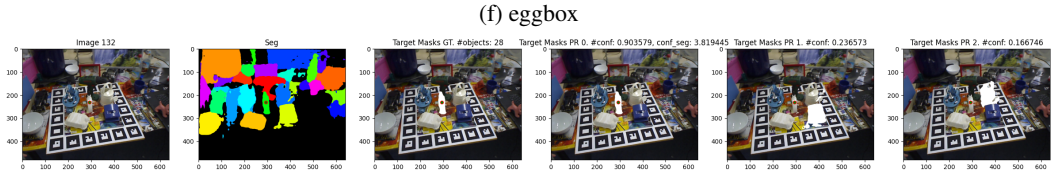
(d) driller



(e) duck

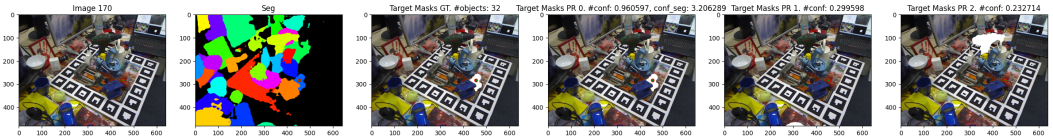


(f) eggbox

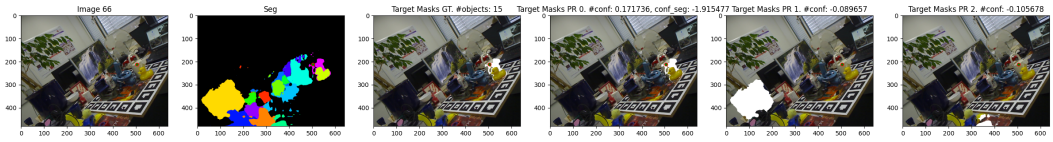


(g) glue

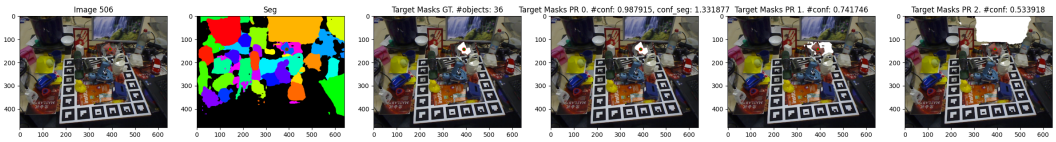
Figure 9: Robust prediction of target segmentation on LineMOD. Three candidates with the highest scores are visualized in order.



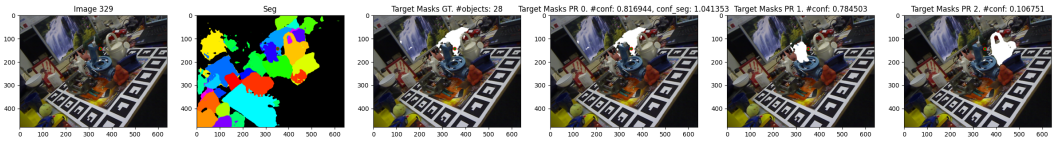
(a) ape



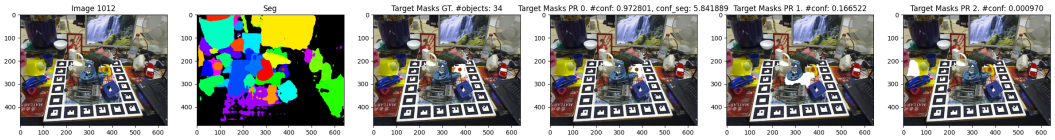
(b) can



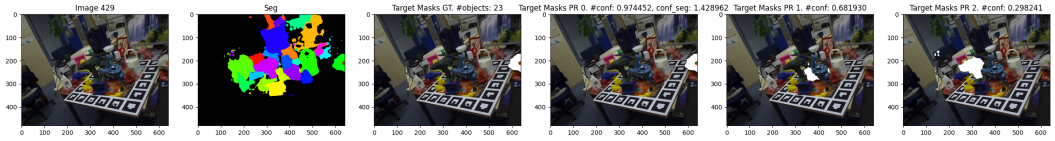
(c) cat



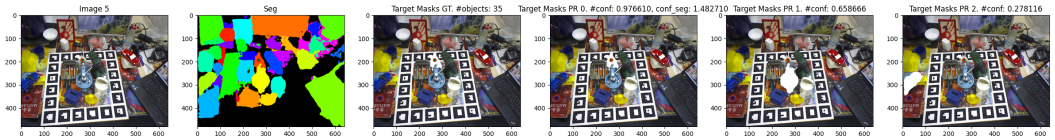
(d) driller



(e) duck

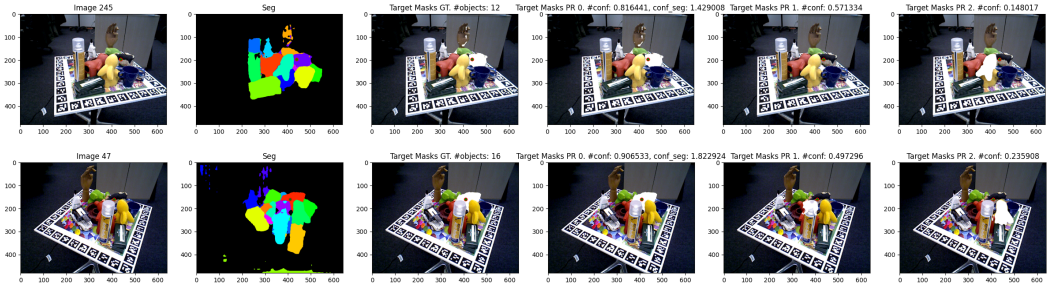


(f) eggbox

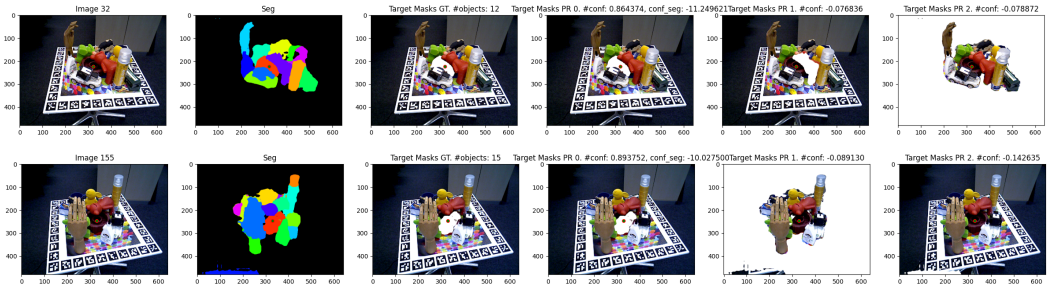


(g) glue

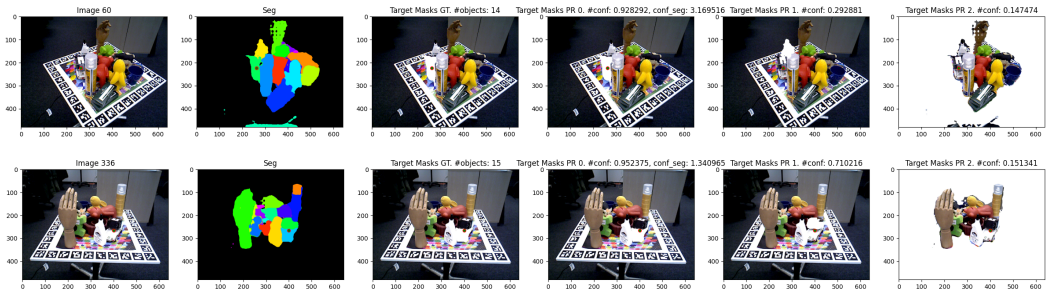
Figure 10: Robust prediction of target segmentation on LineMOD-OCC. Three candidates with the highest scores are visualized in order.



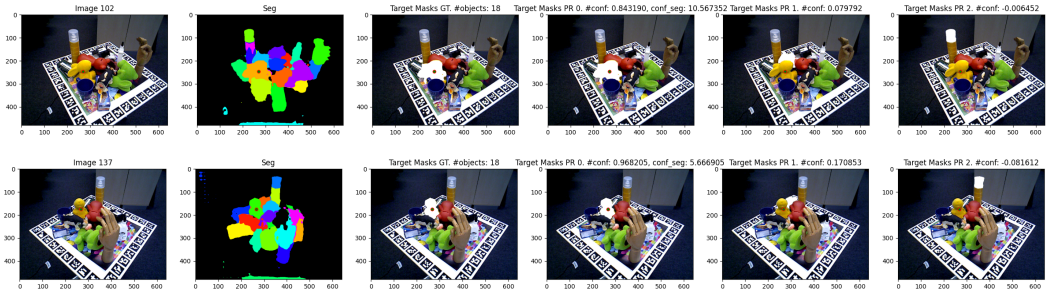
(a) cow



(b) flange



(c) car



(d) yellow rabbit

Figure 11: Robust prediction of target segmentation on HomebrewedDB. Three candidates with the highest scores are visualized in order.

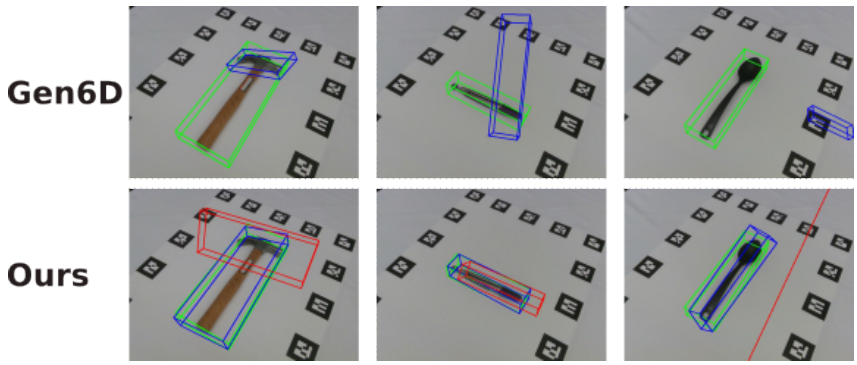


Figure 12: Failure cases. Using ICP in the refinement module leads to a worse prediction than the initial prediction. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

same setup when all models are evaluated with 64 reference images. However, it is not efficient to sample 64 images and it is often that the reference images are not distributed evenly in the real-world. Therefore, we also evaluate all methods by randomly selecting 20 reference images from the dataset, which significantly increases the task difficulty but is more realistic and plausible because it is not always obtainable to collect reference images that could cover all viewpoints.

B.2 Comparison with FS6D and Model-Based Models

Similar to LatentFusion, FS6D [1] also requires object-centric reference images with ground-truth segmentations for cluttered scenes. Considering that its code is not published and we could not reproduce its results, we hence exclude FS6D in our comparisons. Meanwhile, We cannot add the model-based methods [2, 3, 4, 5] into comparison due to their limitation, i.e., the model-based methods can only be applied on the specifically trained object and cannot work in our setup where the results are evaluated on new objects. Also, it is unfair to compare them with our work if we train the model-based methods on the new objects. Moreover, the FewSOL dataset contains only 9 images for each object, which is insufficient to train the model-based methods. Considering all these limitations of the model-based methods, it is also one of our motivations to work on this paper.

B.3 Effort of Annotation Compared with Prior Work

The annotation of a limited number of reference images requires human effort. However, the effort of annotation is also essential in prior work [6, 2, 3, 4, 7, 5, 8] where thousands of annotated images are required for every single object or category. Category-agnostic methods such as our method tremendously reduce human effort by requiring only a small number of annotations. Still, similar to Gen6D and LatentFusion, it is necessary to have a small number of posed reference images for an unseen object to set the canonical object coordinates to further determine the object rotation w.r.t. the camera. Importantly, our method does not require any additional effort compared to existing methods.

B.4 Practical Use Case

Our method can be used in the lifelong robot item picking/sorting in industry. Each time when a new product comes in, the robot only needs to sample a small number of images with ground-truth 6D pose between the new product and the camera by moving the robot arm around the new product where the camera is mounted on the robot arm and the other objects together with the new product are placed on a calibrated picking plate. The pose between the camera and the new product is easily obtainable since the pose of the camera and new product w.r.t. the robot base coordinates

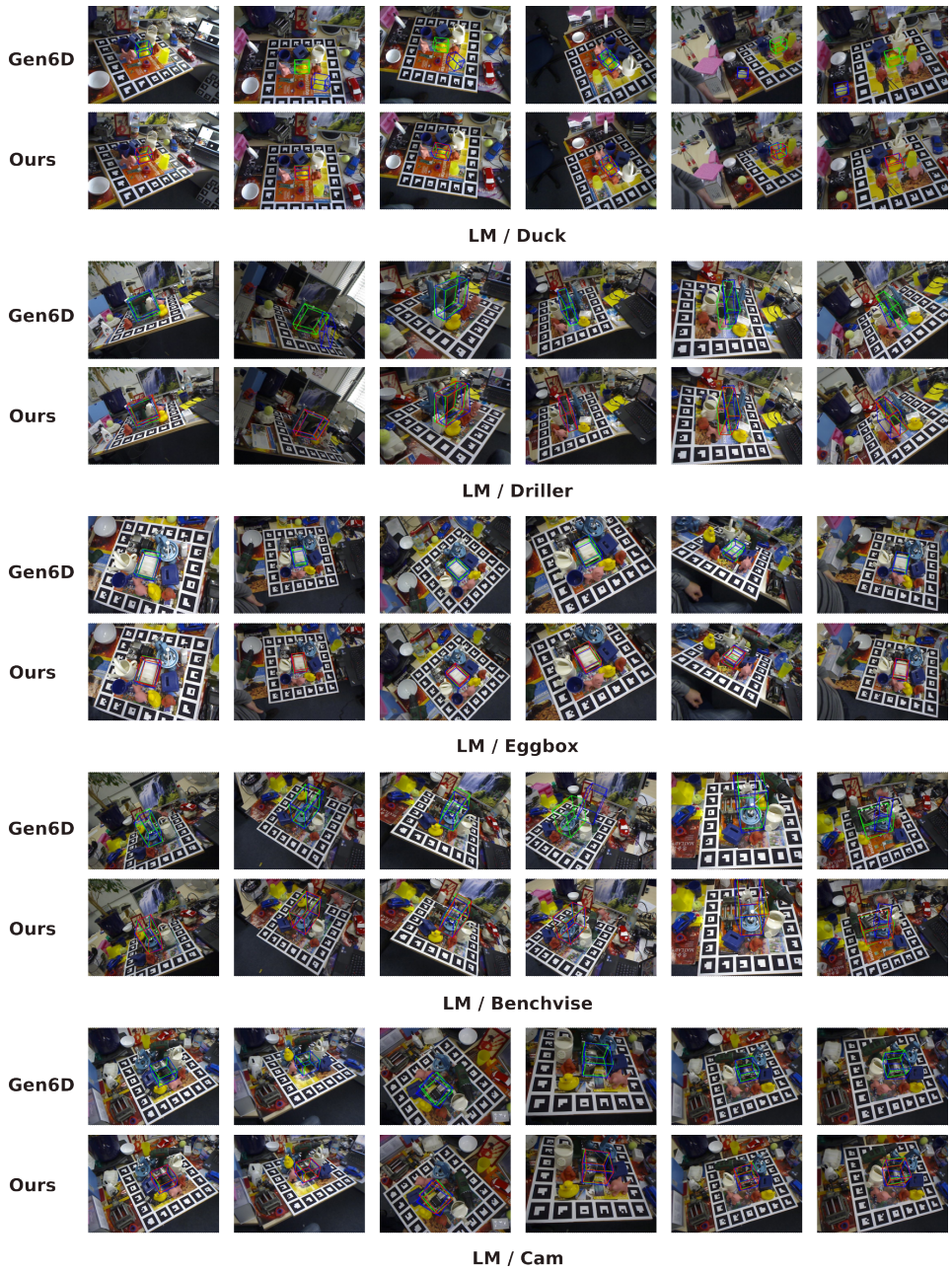


Figure 13: Prediction on LineMOD dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

are known. Thus, the whole system can be fully automatic and does not require further training for new products.

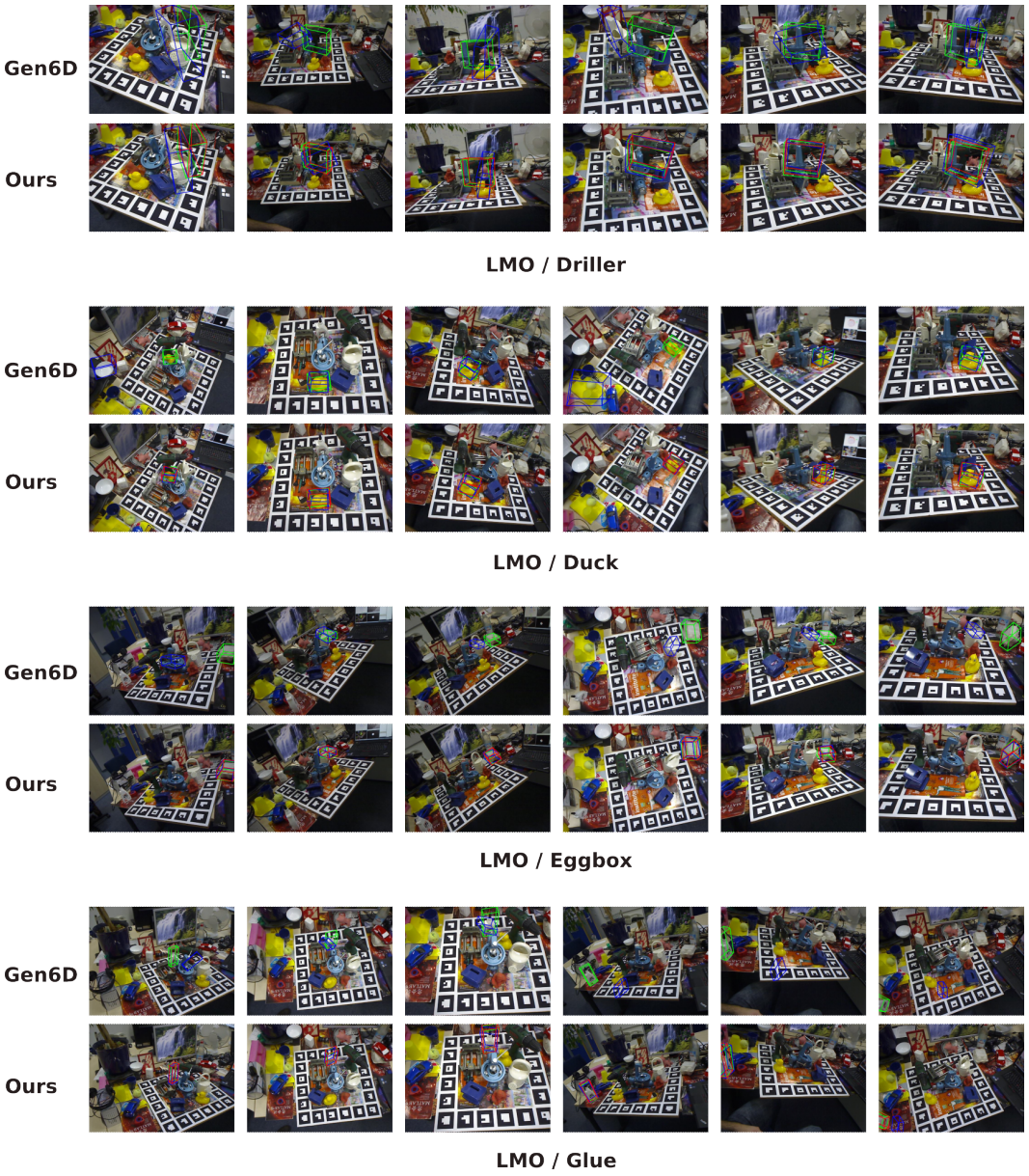


Figure 14: Prediction on LineMOD-OCC dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

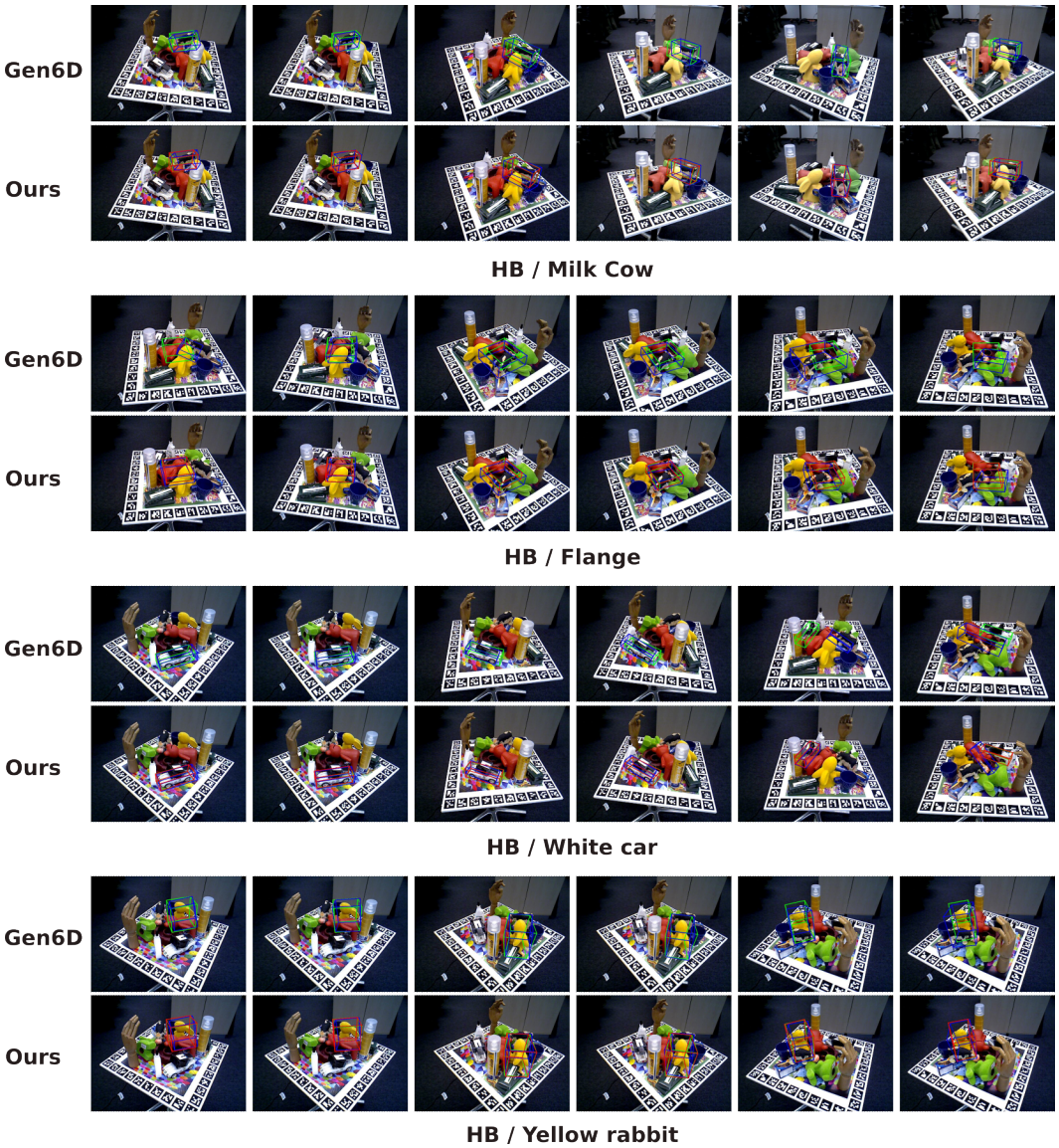


Figure 15: Prediction on HomebrewedDB dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

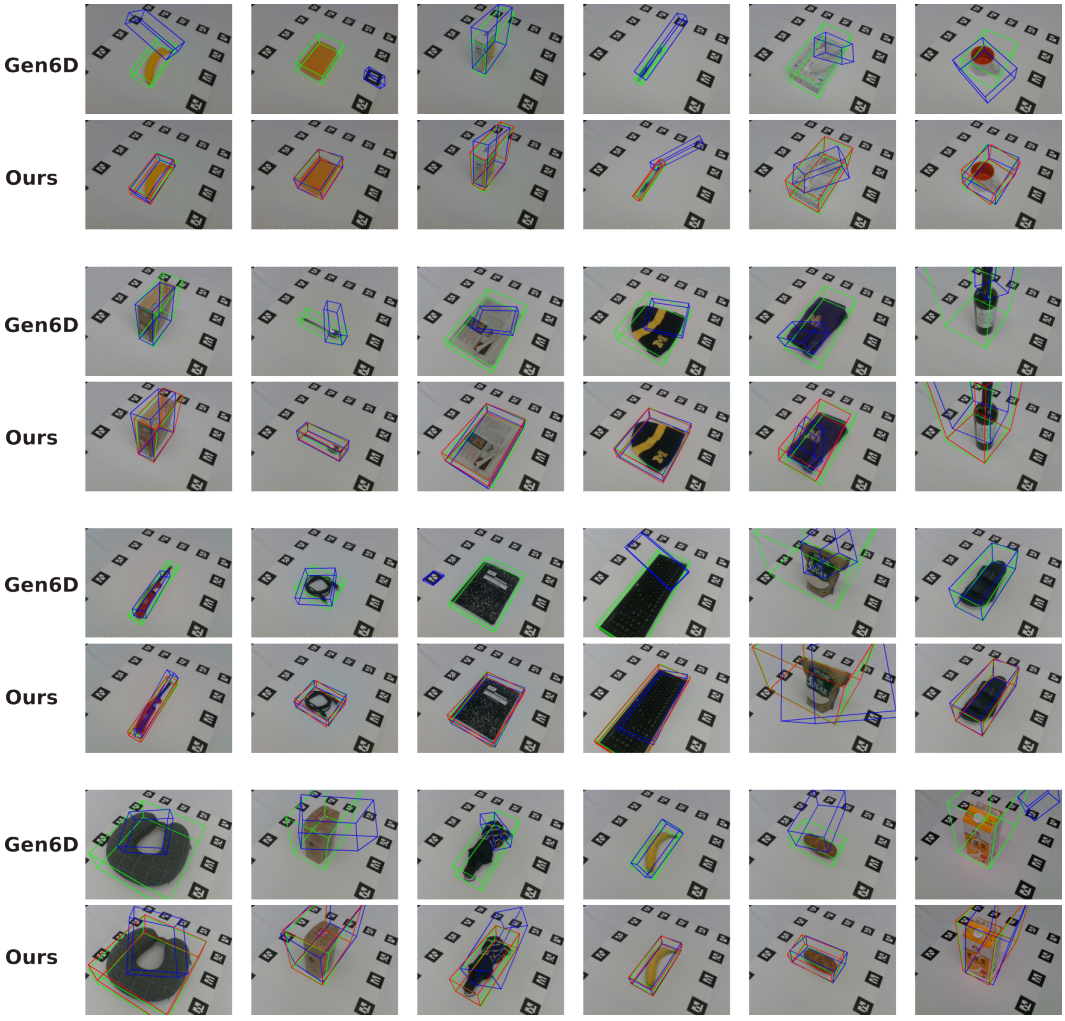


Figure 16: Prediction on FewSOL dataset with 20 reference images. The green bounding box is the ground-truth pose. The blue bounding box denotes the prediction in Gen6D and the prediction before using ICP in the refinement module in our method while the red one denotes the prediction after ICP.

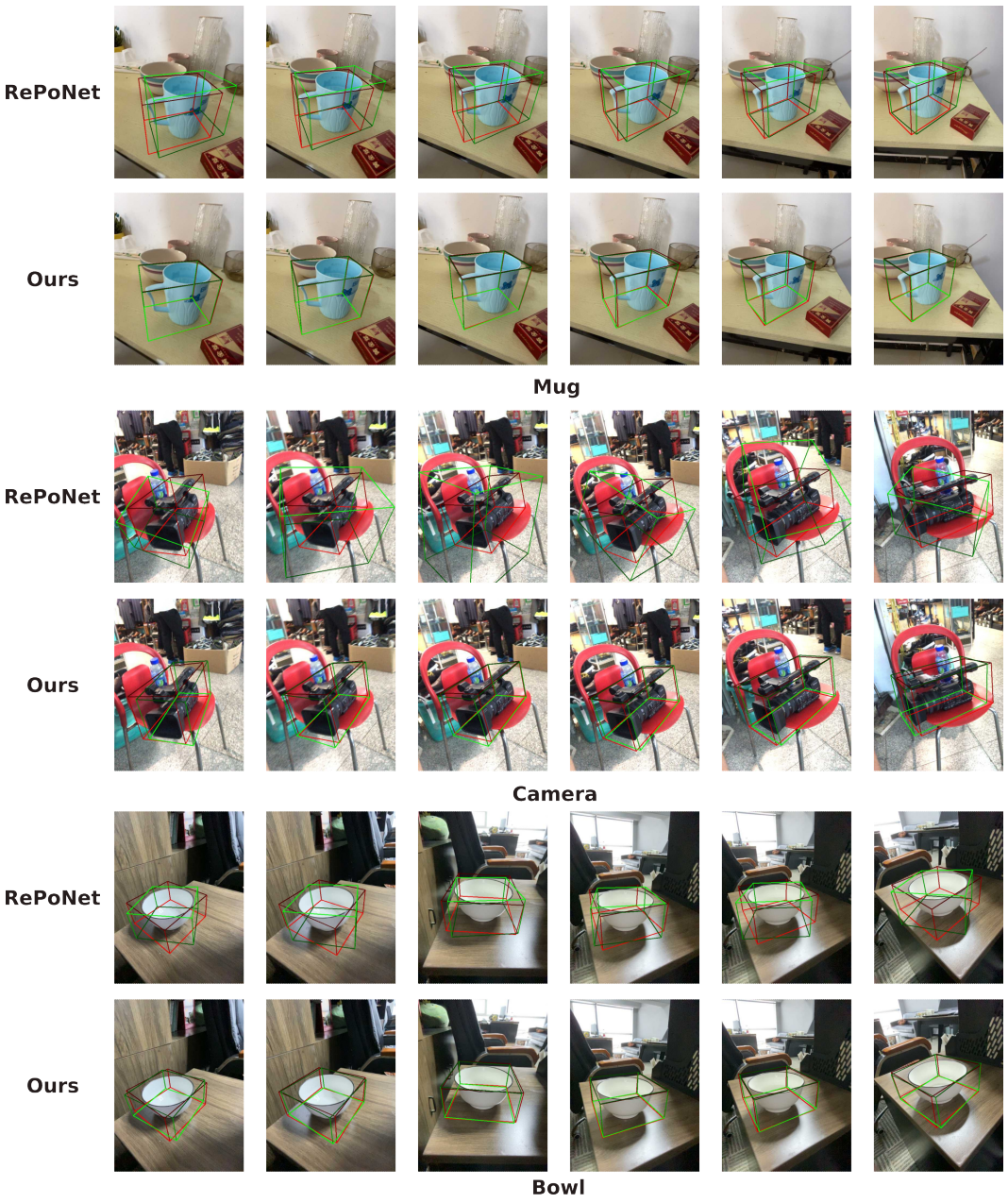


Figure 17: Prediction on Wild6D dataset with 20 reference images. The red bounding box is the ground-truth pose and the green bounding box denotes the prediction.

References

[1] Y. He, Y. Wang, H. Fan, J. Sun, and Q. Chen. Fs6d: Few-shot 6d pose estimation of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6814–6824, June 2022.

[2] Y. He, H. Huang, H. Fan, Q. Chen, and J. Sun. Ffb6d: A full flow bidirectional fusion network for 6d pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3003–3013, June 2021.

[3] Y. He, W. Sun, H. Huang, J. Liu, H. Fan, and J. Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[4] S. Peng, Y. Liu, Q. Huang, X. Zhou, and H. Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[5] C. Wang, D. Xu, Y. Zhu, R. Martin-Martin, C. Lu, L. Fei-Fei, and S. Savarese. Densefusion: 6d object pose estimation by iterative dense fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[6] W. Chen, X. Jia, H. J. Chang, J. Duan, L. Shen, and A. Leonardis. Fs-net: Fast shape-based network for category-level 6d object pose estimation with decoupled rotation mechanism. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1581–1590, June 2021.

[7] M. Tian, M. H. Ang, and G. H. Lee. Shape prior deformation for categorical 6d object pose and size estimation. In A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, editors, *The European Conference on Computer Vision (ECCV)*, pages 530–546, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58589-1.

[8] H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.