OPFormer: Object Pose Estimation leveraging foundation model with geometric encoding

Supplementary Material

A. Ablation studies

A.1. Number of reference templates

Model number	1	7	9	10	11	12
Num. templates	42	42	12	12	162	162
Neighb. templates	✓	✗	✗	✓	✗	✓
LM-O	57.2	56.9	55.7	53.9	57.3	**57.8**
YCB-V	65.1	65.5	65.0	60.4	65.8	**66.5**
T-LESS	47.1	45.1	42.8	42.1	45.4	**47.4**
Avg	56.5	55.8	54.5	52.1	56.2	**57.2**
Time [s]	0.59	0.52	0.47	0.57	0.76	0.81

Table 4. Impact of the inference setting with varied number of templates and neighborhood templates. Models 1 and 7 are taken over from Tab. 3.

In Tab. 4 we explore impact of inference setting while considering the number of templates based on the subdivision of icosahedron and inclusion of the prediction's six neighboring templates. The models 1 and 7 are transferred from Tab. 3. We observe that our model benefits from an increase number of templates at cost of additional computational resources. Furthermore, adding the 6 neighborhood templates increases performance if the template sampling is sufficiently dense.

A.2. Image resolution

Model number	1	13	14	15	16	17	18	19	20
Template image size	420	420	420	280	280	280	140	140	140
Test image size	420	280	140	420	280	140	420	280	140
LM-O	57.2	**57.5**	54.4	56.6	56.6	53.4	51.6	51.5	48.7
YCB-V	65.1	**66.2**	64.4	63.4	65.1	64.2	60.9	60.8	57.3
T-LESS	**47.1**	45.6	37.1	46.6	45.1	37.7	38.3	37.9	32.4
Avg	**56.5**	56.4	52.0	55.5	55.6	51.8	50.3	50.1	46.1
Time [s]	0.59	0.47	0.43	0.51	0.46	0.42	0.49	0.45	0.42

Table 5. Impact of the template image resolution and test image resolution on the coarse 6D pose estimation. Models 1 is taken over from Tab. 3

We study the influence of the template and test image resolutions. The results of this experiment are shown in Tab. 5. For this experiment, resolutions were selected as multiples of the DinoV2 patch size (14 pixels). The resolution is defined as $r = 14s, s \in \{10, 20, 30\}$. For textureless objects present in the T-LESS dataset, higher input resolution is crucial because performance suffers greatly with low-resolution images. Even with the lowest resolution setting for both test image and reference templates, our method achieves higher average recall when compared to [34, 50, 55, 57] with coarse-only strategy. Further quantitative results, which we do not include into the table, show that performance plateaus at a resolution of $(420, 420)$ and slightly degrades at higher resolutions (e.g., $(700, 700)$). We attribute this to two factors. First, at very high resolutions, a single 14×14 patch captures a smaller and less descriptive region of an image than it would at lower resolutions. Second, due to limited computational resources, our model was trained on resolutions only up to $(420, 420)$ and therefore does not generalize to the longer token sequences generated by these out-of-distribution inputs.

A.3. 2D detection

2D detection source	LM-O	YCB-V	T-LESS	Avg	Time
Ground truth	**68.3**	**69.7**	**70.6**	**69.5**	-
CNOS FastSAM [54]	57.2	65.1	47.1	56.6	0.59
NIDS [48]	58.3	66.3	52.3	58.9	0.76
SAM6D FastSAM [42]	57.6	65.3	48.2	57.0	**0.57**

Table 6. Impact of the 2D detection's quality on the coarse 6D pose estimation of the OPFormer. The CNOS FastSAM 2D detector was used for evaluation in the rest of the experiments.

Tab. 6 presents the impact of different 2D detection sources on OPFormer's performance. We compare results from several proposed methods [42, 48, 54] against the performance achieved using ground-truth bounding boxes. Erroneous 2D detections are a primary cause of pose estimation failures, leading to a significant drop in Average Recall (AR) scores. This issue is particularly pronounced on the T-LESS dataset, where some objects have a similar appearance, causing the detector to frequently mismatch object categories. When comparing 2D detection methodologies, the NIDS detector [48] yields the highest detection accuracy, which in turn improves the final 6D pose estimation. However, this accuracy comes with a trade-off: an approximately 30% reduction in inference speed.

B. Limitations

Our analysis identifies two primary failure modes. First, the most significant performance degradation occurs when the 2D detector fails to produce a bounding box or misclassifies the object category. We provide a discussion and a quantitative analysis of this effect in Appendix A.3 (Tab. 6). Second, the model faces challenges with symmetric or textureless objects. For such objects, the descriptor vectors for

symmetric parts are highly similar, which leads to ambiguous or incorrect feature matching. Additionally, as the other RGB-based methods the prediction precisions degrades depending on the input image size resolution as the input size changes from down-sampling to up-sampling, which inherently introduces noise and artifacts.

The quality of the NeRF-based 3D reconstruction is highly sensitive to the input of multi-view images with known poses. Deficiencies in this input data, such as an insufficient number of views, poor image quality, inaccurate camera poses, or faulty background segmentation, prevent the model from learning a high-fidelity representation. This directly results in the generation of flawed RGB, NOCS, and depth templates. Since our method's accuracy relies on establishing robust 2D-3D correspondences between a test image and these templates, relying on inaccurate templates inevitably leads to poor feature matching and a significantly degraded final 6D pose estimation.

C. BOP results in detail

Dataset	Refiner	AR_{VSD}	AR_{MSPD}	AR_{MSSD}	AR
LM-O	✗	43.3	73.1	55.3	57.2
	1 hyps.	47.4	72.9	60.9	60.4
	5 hyps.	47.4	73.1	60.8	60.4
T-LESS	✗	41.9	55.5	44.1	47.1
	1 hyps.	48.8	56.8	50.0	51.9
	5 hyps.	49.6	58.0	50.9	52.8
TUD-L	✗	54.9	85.3	66.5	68.9
	1 hyps.	55.9	85.6	67.0	69.5
	5 hyps.	56.8	86.3	68.0	70.3
IC-BIN	✗	40.7	52.7	43.3	45.6
	1 hyps.	44.9	55.5	48.8	49.7
	5 hyps.	44.5	55.7	48.8	49.7
ITODD	✗	30.1	52.7	33.3	38.7
	1 hyps.	33.6	52.3	37.6	41.2
	5 hyps.	35.1	54.6	38.9	42.8
HB	✗	70.7	80.3	74.1	75.0
	1 hyps.	72.9	80.9	76.1	76.6
	5 hyps.	72.9	81.0	76.2	76.7
YCB-V	✗	54.9	78.4	62.1	65.1
	1 hyps.	59.1	80.9	66.2	68.7
	5 hyps.	59.0	81.0	66.2	68.7

Table 7. Detailed scores of OPFormer from Tab. 1 for 7 BOP-Core-Classic datasets.

Tab. 7 presents scores that complement those in Tab. 1, detailing result for the initial coarse estimation and subsequent pose refinements. As refinement we deploy MegaPose render-and-compare refiner on both single best hypothesis and the top five hypotheses. Furthermore, Tab. 8 provides detailed results to Tab. 2, showing the coarse estimation and refinement of single best hypothesis. Both tables provide the results based on the error metrics described in Sec. 4.1 and Appendix D.

Dataset	Task	Refiner	AP_{MSPD}	AP_{MSSD}	AP
HOPEv2	Model-based	✗	42.5	31.4	37.0
		✓	42.6	33.3	37.9
	Model-free	✗	47.1	26.0	36.6
		✓	46.5	24.8	35.6
HANDAL	Model-based	✗	41.1	32.4	36.7
		✓	41.9	35.7	38.8
	Model-free	✗	37.2	24.5	30.9
		✓	37.8	27.0	32.4
HOT3D	Model-based	✗	35.2	28.0	31.6
		✓	34.2	26.4	30.3

Table 8. Detailed scores of OPFormer from Tab. 2 on the BOP-H3 datasets in model-based and model-free 6D detection tasks.

D. Metric description

The objective of the evaluation metrics is to calculate the precision of the prediction, given the estimated pose $\hat{\mathbf{P}}$ and the ground truth $\bar{\mathbf{P}}$ for a model $\mathcal{M}$ with set of vertices $\mathcal{V}_{\mathcal{M}}$ and set of symmetries $S_{\mathcal{M}}$ in image I. For our evaluation we follow the BOP metrics as described by [29]. The calculation of both their Average Recall and Average Precision is performed as follows

$$\mathrm{AR} = (\mathrm{AR}_{\mathrm{VSD}} + \mathrm{AR}_{\mathrm{MSPD}} + \mathrm{AR}_{\mathrm{MSSD}})/3, \quad (7)$$

$$\mathrm{AP} = (\mathrm{AP}_{\mathrm{MSPD}} + \mathrm{AP}_{\mathrm{MSSD}})/2. \quad (8)$$

The errors for each metric are calculated as

$$e_{\mathrm{VSD}}(\hat{D}, \bar{D}, \hat{V}, \bar{V}, \tau) = \mathrm{avg}_{p \in \hat{V} \cup \bar{V}} \begin{cases} 0 & \text{if } p \in \hat{V} \cap \bar{V} \\ & \wedge \mid \hat{D}(p) - \bar{D}(p) \mid < \tau, \\ 1 & \text{otherwise,} \end{cases} \quad (9)$$

$$e_{\mathrm{MSSD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}, S_{\mathcal{M}}, \mathcal{V}_{\mathcal{M}}) = \min_{\mathbf{S} \in S_{\mathcal{M}}} \max_{\mathbf{x} \in \mathcal{V}_{\mathcal{M}}} \|\hat{\mathbf{P}}\mathbf{x} - \bar{\mathbf{P}}\mathbf{S}\mathbf{x}\|_2, \quad (10)$$

$$e_{\mathrm{MSPD}}(\hat{\mathbf{P}}, \bar{\mathbf{P}}, S_{\mathcal{M}}, \mathcal{V}_{\mathcal{M}}) = \min_{\mathbf{S} \in S_{\mathcal{M}}} \max_{\mathbf{x} \in \mathcal{V}_{\mathcal{M}}} \|\mathrm{proj}(\hat{\mathbf{P}}\mathbf{x}) - \mathrm{proj}(\bar{\mathbf{P}}\mathbf{S}\mathbf{x})\|_2. \quad (11)$$

The **VSD** (Visible Surface Discrepancy) metric is focused on aliment of the visible part of the model, using the the visibility masks $\hat{V}$ and $\bar{V}$ and the distance maps $\hat{D}$ and $\bar{D}$, the image distance map D_I, and the misalignment tolerance τ. The **MSSD** (Maximum Symmetry-Aware Surface Distance) is a strict metric that calculates maximum distance between corresponding points of the estimated and ground truth surfaces, while considering possible object symmetries. The **MSPD** (Maximum Symmetry-Aware Projection Distance) measures the pixel distance between the estimated and ground-truth image projections, while also considering the symmetries.

In contrast, other metrics, like rotation and translation error, do not consider the model's geometric properties, as they only account for the 6D pose $\bar{\mathbf{P}}$ and $\hat{\mathbf{P}}$. 3DIoU only considers the geometry of the model's bounding box which can lead to misclassifications of continuously symmetric objects or, in cases of point or plane symmetries, where 3D bounding boxes can be identical but the model orientations differ.

ADD and ADI consider an object's geometric properties, but due to their calculation of average vertex distance, they are highly dependent on the mesh sampling. ADI accounts for symmetries by measuring the distance to the nearest neighbor; however, this approach becomes problematic with objects exhibiting discrete symmetries.

E. Additional visualization

We provide additional visualization results for all BOP-Classic-Core datasets and the HANDAL and HOPEv2 datasets from BOP-H3 set, for which we report the results in Sec. 4.3. Since the HANDAL dataset lacks both depth and the ground-truth poses, providing 3D visualization would be uninformative, as only the predictions would be presented. Therefore, we present only the 2D projection visualization, as shown in Fig. 5. In case of the HOPEv2 (Fig. 6), HB (Fig. 7), and ITODD (Fig. 8) datasets, which do not provide ground-truth 6D pose but do contain depth, we visualize the estimated poses with the colored point cloud. The remaining BOP-classic core datasets, which provide the ground-truth poses, are visualized in figures 9 - 13.

In these figures, all ground-truth poses present in each scene are presented, and non-visible or highly occluded models in the inference image are not omitted. This explains, in part, why, especially in the bin scenarios, there are more ground-truth models present than in the estimated meshes. Another reason is erroneous 2D detection, as mentioned in Appendix A.3, which also leads to estimations where the detected mesh is positioned far from others (e.g., cases where the mesh is predicted to be under the floor). In a few images, we observe the mesh being directly in front of camera; this is caused by the final stage of our 6D pose estimation pipeline, where RANSAC together with PnP algorithm fails and returns default prediction.

Notably, the visualization of both the 2D projections and the 3D views demonstrates the complexity of the 6D pose estimation from a single RGB image (mentioned in Appendix B), where minor pixel distance variations can result in more substantial Euclidean distance errors. This observation is supported by findings from Tab. 7 and Tab. 8, as MSPD and MSSD scores, although not directly comparable, suggest a preference for the projection error metric.

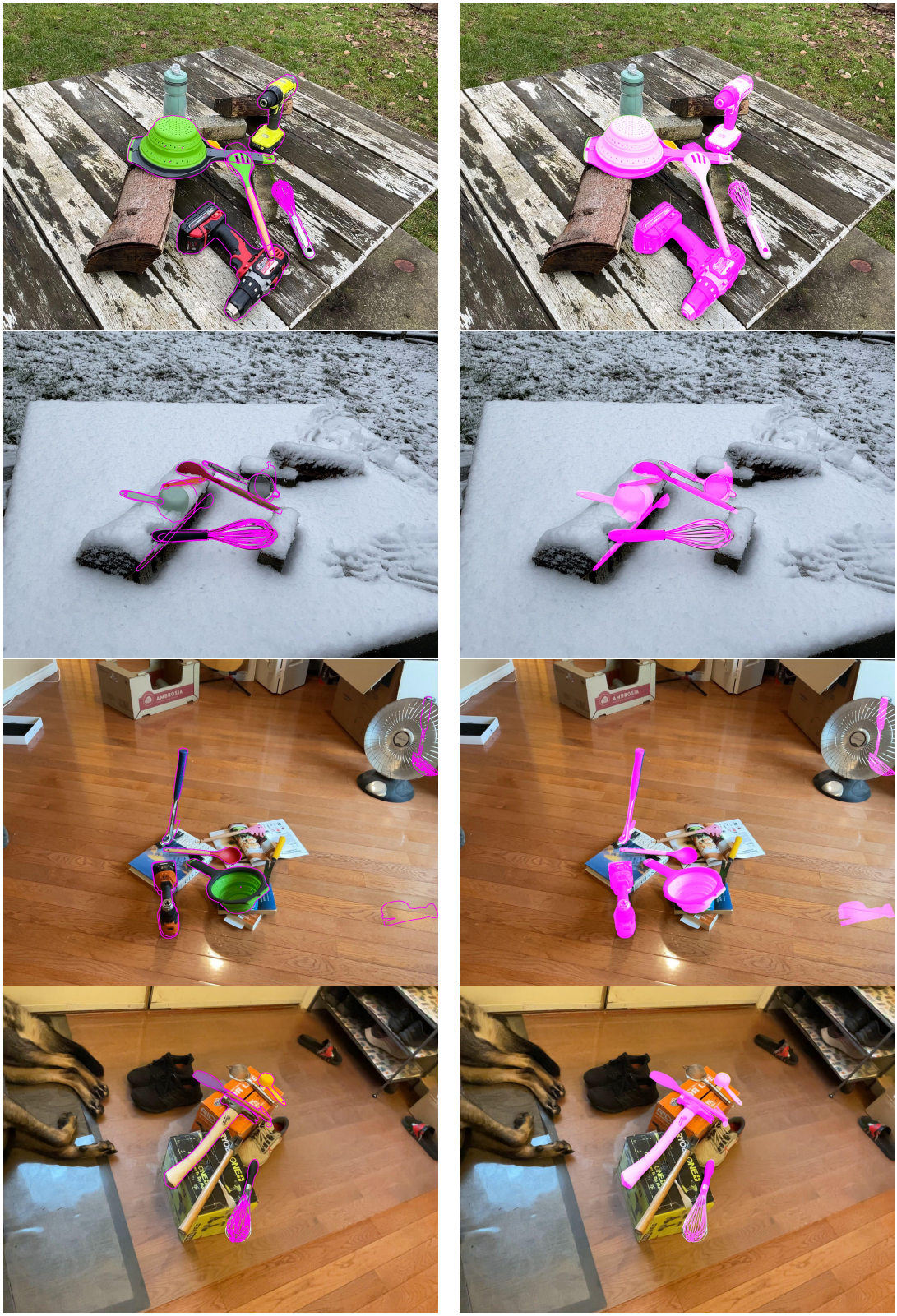

Figure 5. HANDAL dataset visualization of 2D projection made with the predicted 6D poses. In the left column is visualized contour highlight and in the right the projected masks of the prediction

Figure 6. HOPEv2 dataset visualization with estimated (magenta) 6D pose of the meshes and point cloud. The first column shows the test image with a contour of the projection made by the predicted pose. The other two columns show the corresponding 3D view from different viewing angles. The first is captured from approximately the same viewing angle as the image was taken.

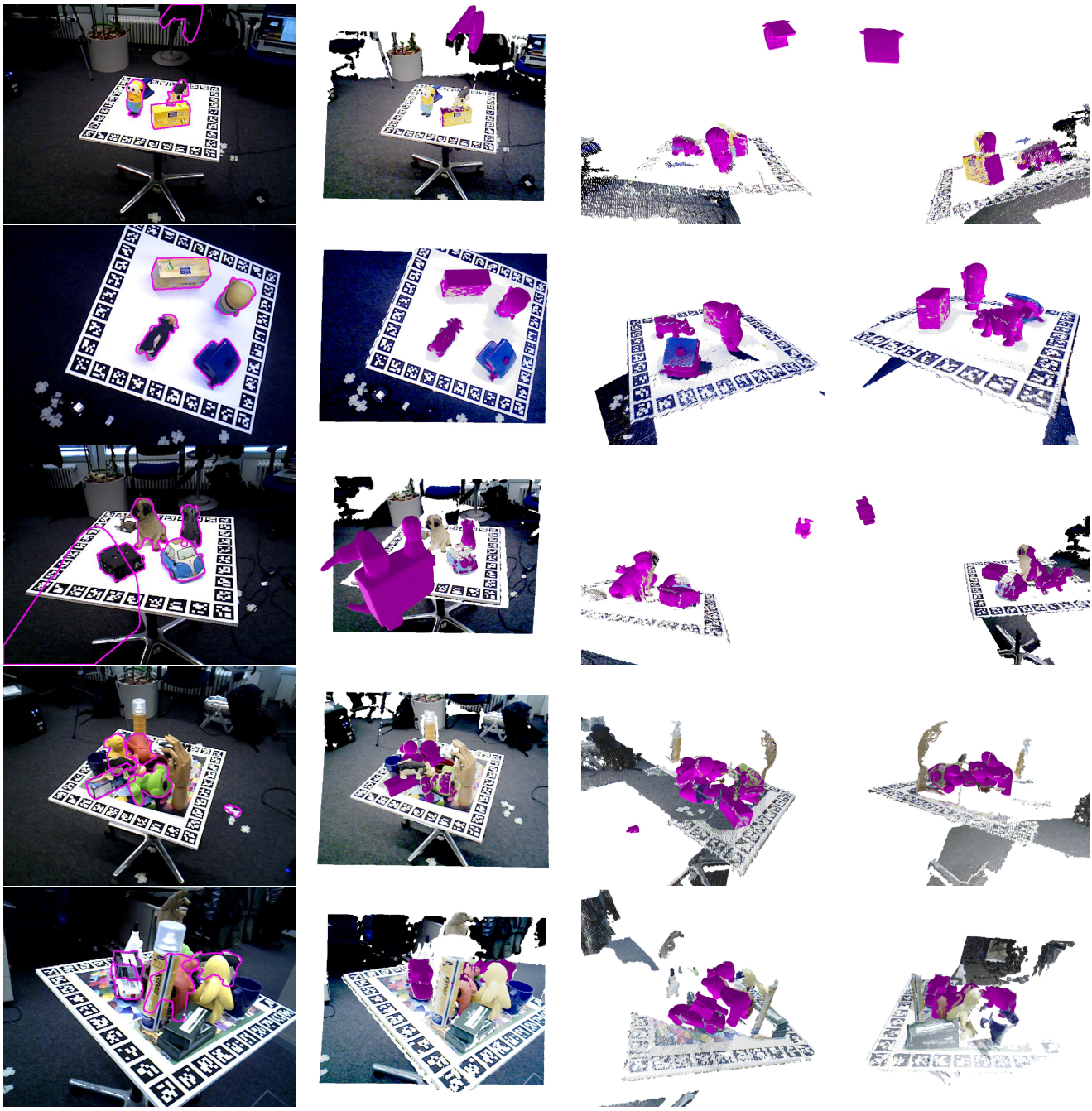

Figure 7. HomebrewedDB (HB) dataset visualization with estimated (magenta) 6D pose of the meshes and point cloud. The first column shows the test image with a contour of the projection made by the predicted pose. The other three columns show the corresponding 3D view from different viewing angles. The first is taken from approximately the same viewing angle as the image was taken.

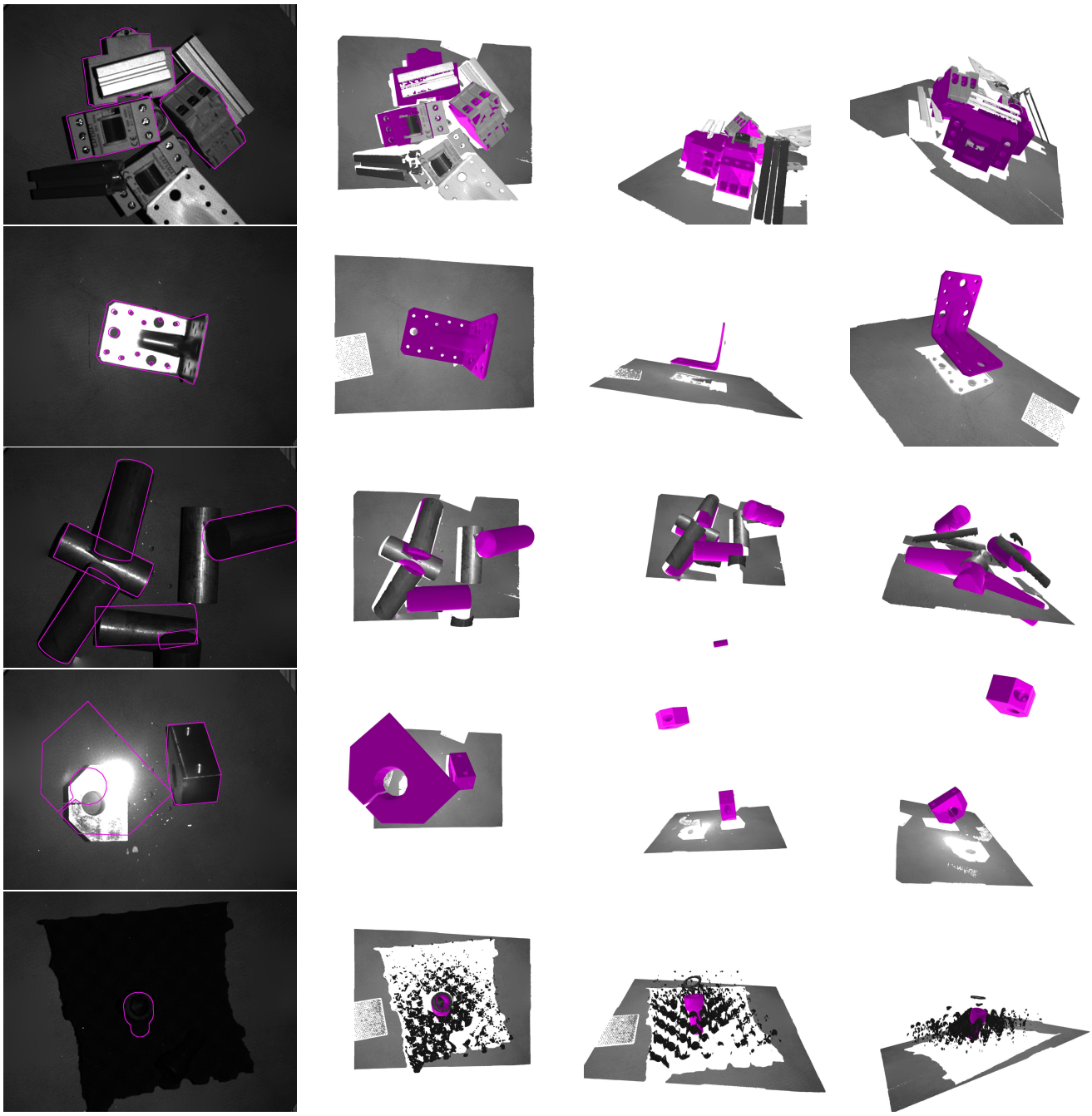

Figure 8. ITODD dataset visualization with estimated (magenta) 6D pose of the meshes and point cloud. The first column shows the test image with a contour of the projection made by the predicted pose. The other three columns show the corresponding 3D view from different viewing angles. The first is taken from approximately the same viewing angle as the image was taken.

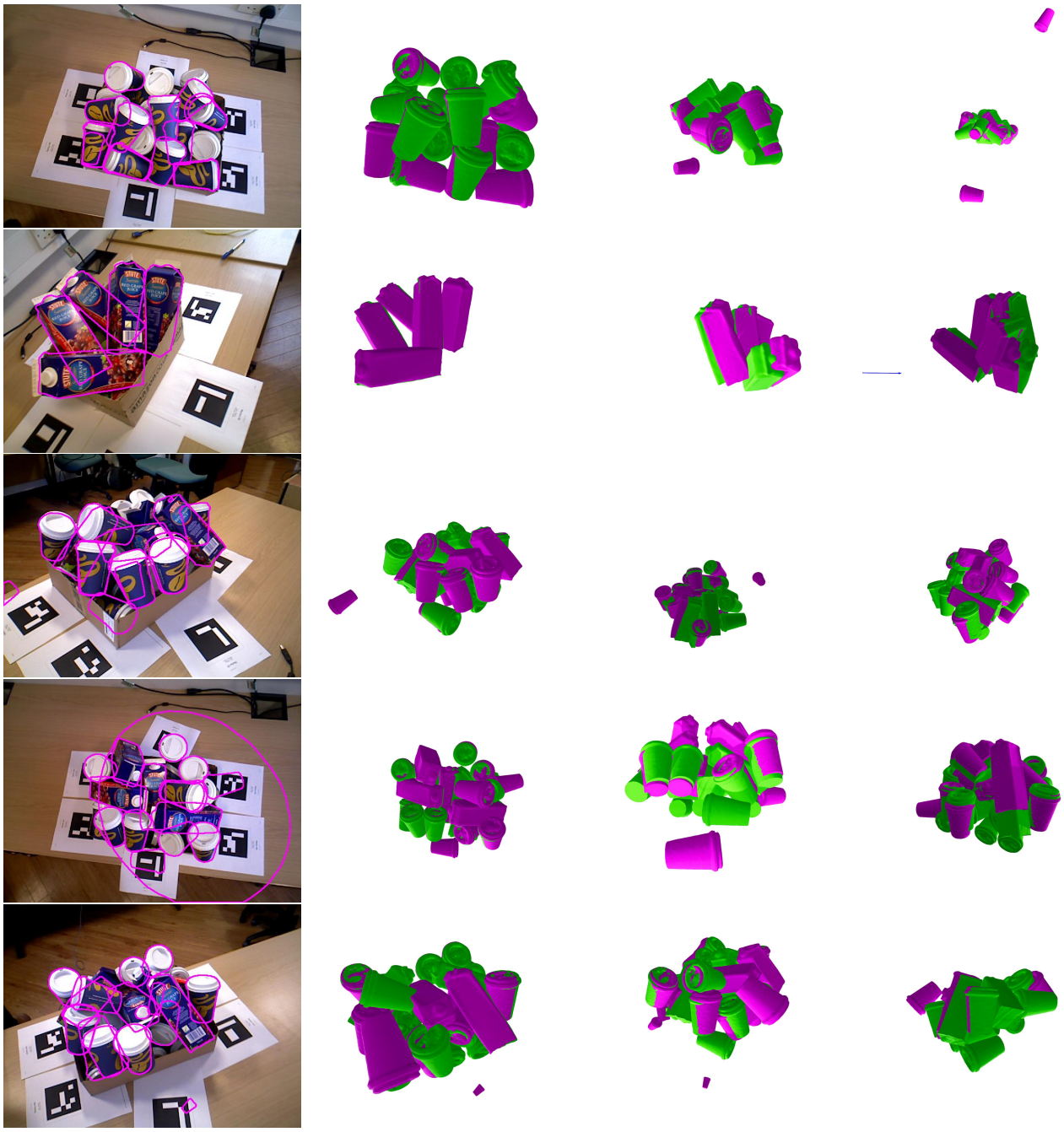

Figure 9. IC-BIN dataset visualization with estimated (magenta) 6D pose of the meshes and ground-truth (green) poses. The first column shows the test image with a contour of the projection made by the predicted pose. The other three columns show the corresponding 3D view from different viewing angles. The first is taken from approximately the same viewing angle as the image was taken.

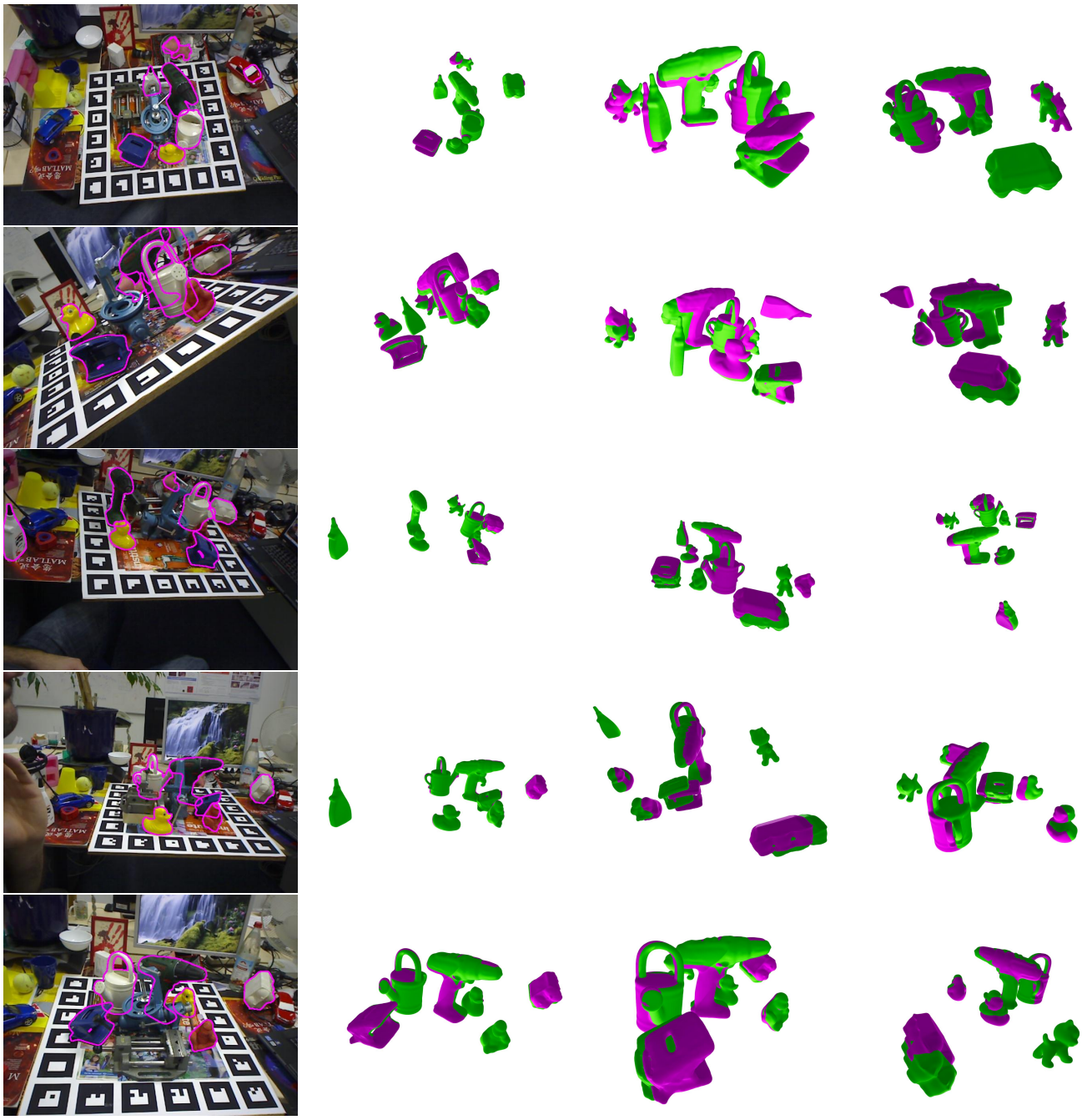

Figure 10. LM-O dataset visualization with estimated (magenta) 6D pose of the meshes and ground-truth (green) poses.. The first column shows the test image with a contour of the projection made by the predicted pose. The other three columns show the corresponding 3D view from different viewing angles. The first is taken from approximately the same viewing angle as the image was taken.

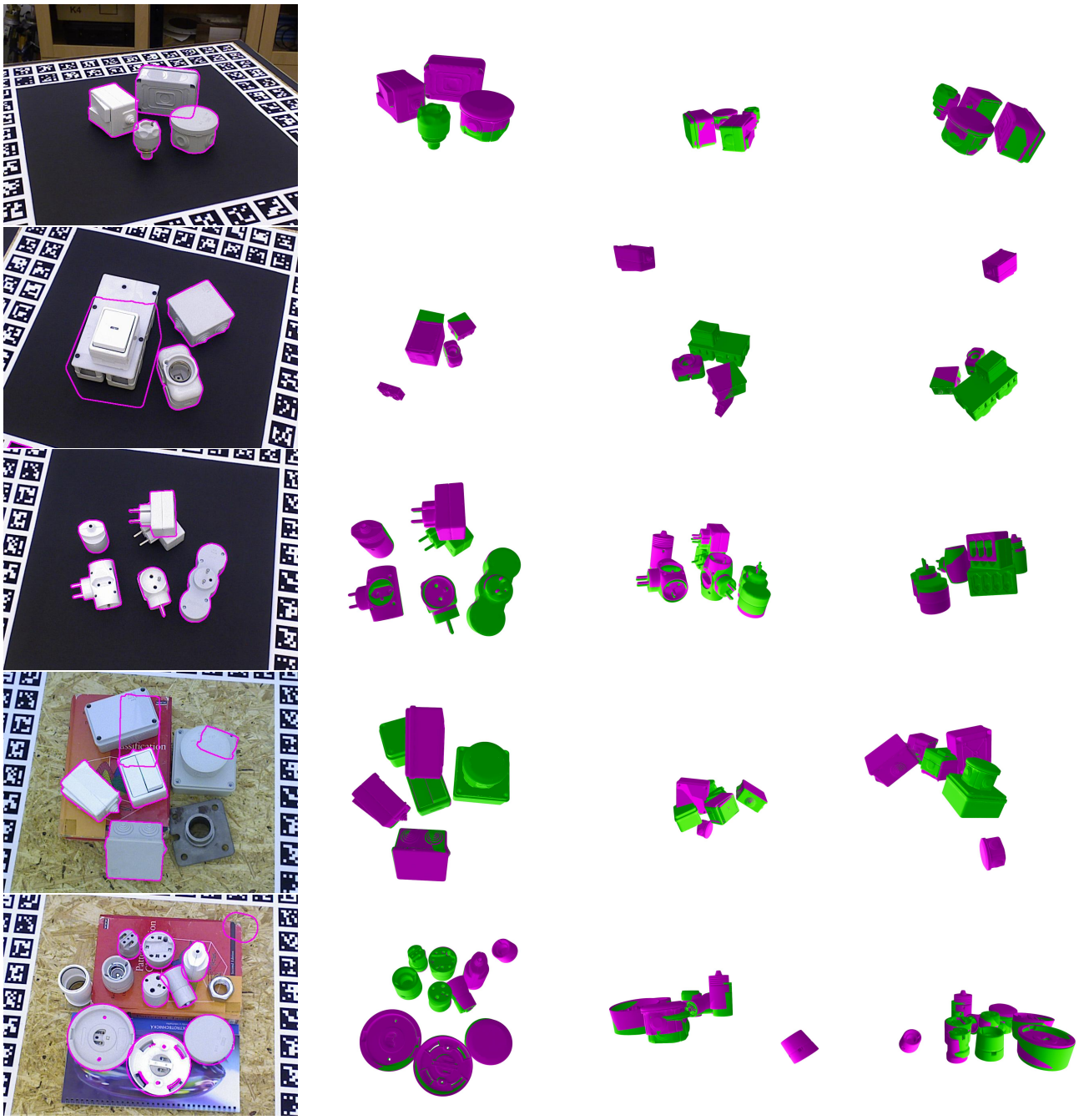

Figure 11. T-LESS dataset visualization with estimated (magenta) 6D pose of the meshes and ground-truth (green) poses. The first column shows the test image with a contour of the projection made by the predicted pose. The other three columns show the corresponding 3D view from different viewing angles. The first is taken from approximately the same viewing angle as the image was taken.

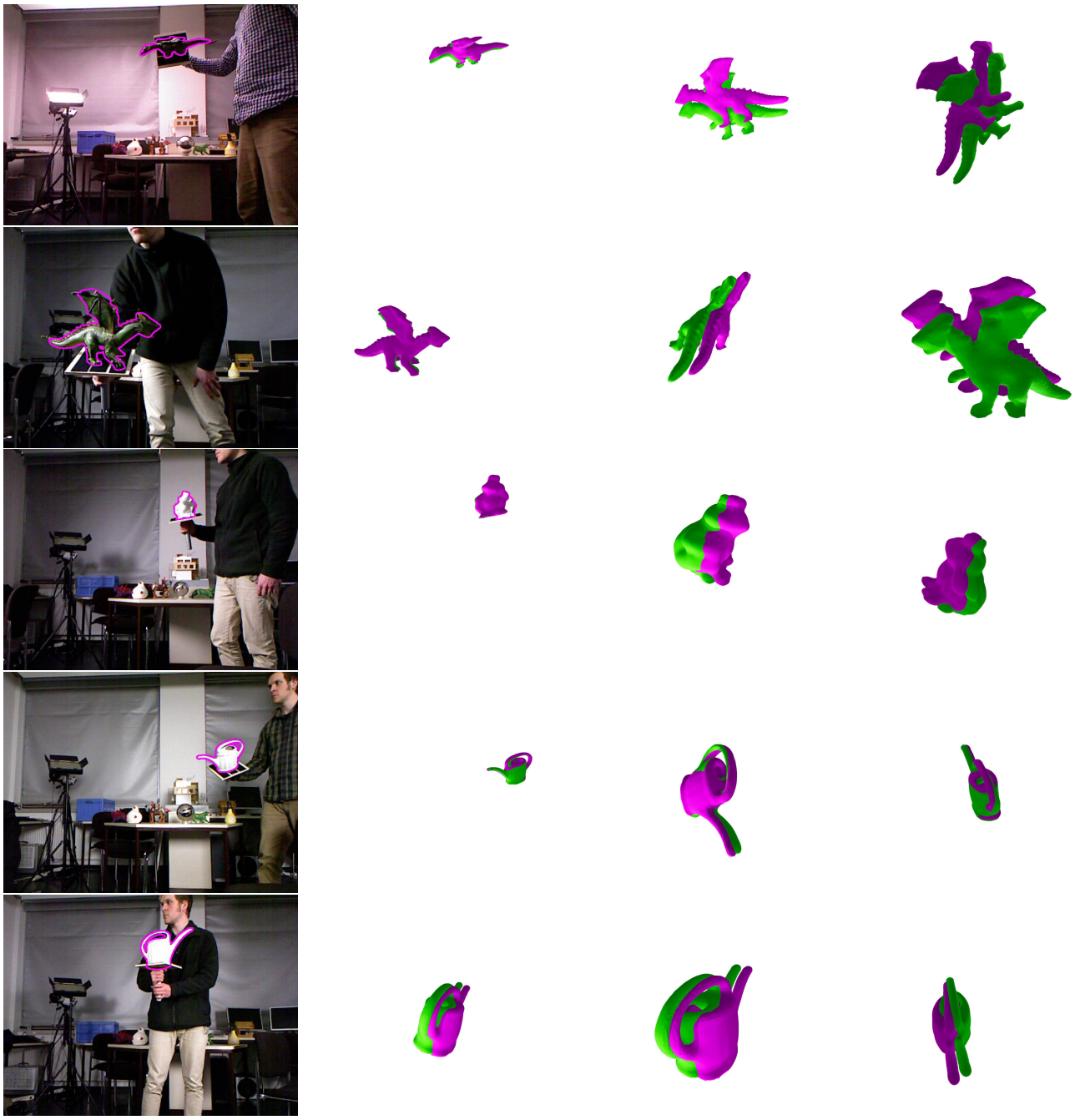

Figure 12. TUD-L dataset visualization with estimated (magenta) 6D pose of the meshes and ground-truth (green) poses. The first column shows the test image with a contour of the projection made by the predicted pose. The other three columns show the corresponding 3D view from different viewing angles. The first is taken from approximately the same viewing angle as the image was taken.

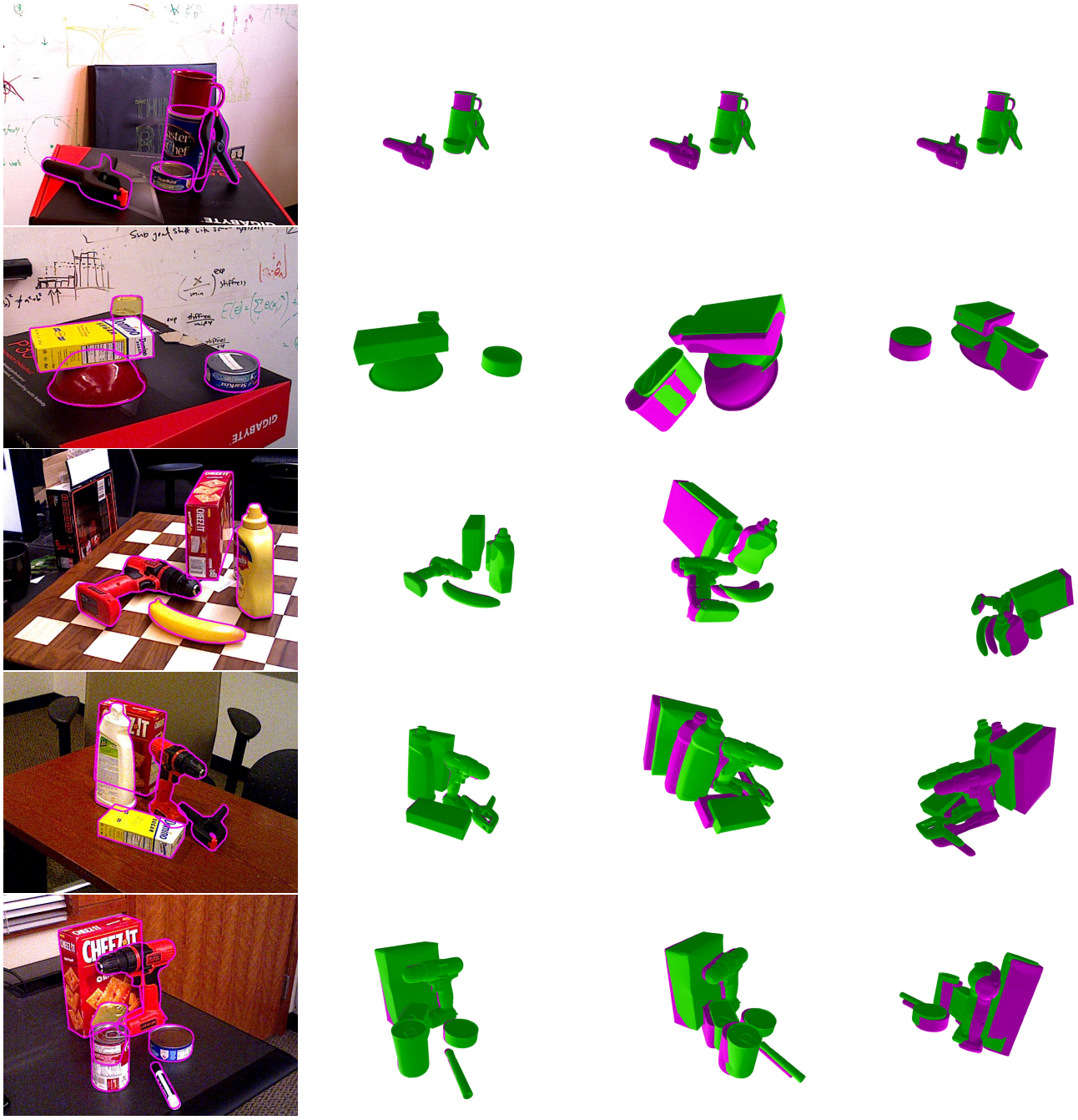

Figure 13. YCB-V dataset visualization with estimated (magenta) 6D pose of the meshes and ground-truth (green) poses. The first column shows the test image with a contour of the projection made by the predicted pose. The other three columns show the corresponding 3D view from different viewing angles. The first is taken from approximately the same viewing angle as the image was taken.