Supplementary Material: ZeroGrasp: Zero-Shot Shape Reconstruction Enabled Robotic Grasping

Shun Iwase[1,2] Muhammad Zubair Irshad[2] Katherine Liu[2] Vitor Guizilini[2]
Robert Lee[3] Takuya Ikeda[3] Ayako Amma[3] Koichi Nishiwaki[3] Kris Kitani[1]
Rareș Ambruș[2] Sergey Zakharov[2]
[1]Carnegie Mellon University [2]Toyota Research Institute [3]Woven by Toyota

A. Dataset Details

Table 1. The number of images and 3D models across the different splits of the ReOcS dataset.

Split	# Images	# 3D Models
Easy	479	22
Normal	329	22
Hard	317	20
Total	1125	22

Table 1 shows the number of images and 3D models of the three splits — easy, normal and hard — of the ReOcS dataset. We show examples from the ReOcS and ZeroGrasp-11B datasets in Figures 1 to 3 and Figures 4 to 6, respectively. For physics-based grasp pose validation on the ZeroGrasp-11B dataset, we use the Franka Emika Panda hand. We first move the gripper given a synthesized grasp pose and apply a closing force of 80N. Next, we swing the gripper orthogonally to the direction of the parallel fingers by 60 degrees, repeating this motion four times. A grasp is considered successful only if both fingers maintain contact after the motion. For instance masks, we fine-tune SAM-2 on the ZeroGrasp-11B dataset and the training split of the GraspNet-1B dataset.

B. Experiments

B.1. Implementation Details

We compute the final grasp scores by multiplying the predicted graspness $\mathbf{g}$ and quality $\mathbf{q}$ metrics. The predicted width and depth are clipped to lie within the ranges $[0.0, 0.10]$ and $[0.0, 0.04]$, respectively. The learning rate is decayed by a factor of 0.5 every 5000 iterations. The weight parameters of the loss function are set as follows; $\omega_{\text{occ}} = 1.0$, $\omega_{\text{nrm}} = 1.0$, $\omega_{\text{SDF}} = 1.0$, $\omega_{\text{g}} = 5.0$, $\omega_{\text{q}} = 5.0$, $\omega_{\text{a}} = 1.0$, $\omega_{\text{t}} = 1.0$, $\omega_{\text{w}} = 2.5$, $\omega_{\text{d}} = 2.5$, and $\omega_{\text{KL}} = 0.5$.

We use three layers of 3D CNNs that downsample the resolution by a factor of two at each layer to extract 1-D features from 3D occlusion fields per voxel in the latent space.

B.2. Metrics

Similar to OctMAE [1], we evaluate the reconstruction metrics such as Chamfer distance (CD), F-1 score and normal consistency (NC) with the following equations. Note that the predicted surface points $\mathcal{P}_{\text{pd}}$ are derived from the predicted occupied points, normal vectors, and SDF. The ground-truth surface points is denoted as $\mathcal{P}_{\text{gt}}$.

Chamfer distance (CD). We use the bidirectional Chamfer distance as a more balanced metric to measure the similarity between two point cloud sets.

$$\begin{aligned}\text{CD}(\mathcal{P}_{\text{pd}}, \mathcal{P}_{\text{gt}}) = &\frac{1}{2|\mathcal{P}_{\text{pd}}|} \sum_{\mathbf{x}_{\text{pd}} \in \mathcal{P}_{\text{pd}}} \min_{\mathbf{x}_{\text{gt}} \in \mathcal{P}_{\text{gt}}} ||\mathbf{x}_{\text{pd}} - \mathbf{x}_{\text{gt}}|| \\ &+ \frac{1}{2|\mathcal{P}_{\text{gt}}|} \sum_{\mathbf{x}_{\text{gt}} \in \mathcal{P}_{\text{gt}}} \min_{\mathbf{x}_{\text{pd}} \in \mathcal{P}_{\text{pd}}} ||\mathbf{x}_{\text{gt}} - \mathbf{x}_{\text{pd}}||.\end{aligned} \tag{1}$$

F-1 score. We use the threshold η of 1 cm for all evaluations, in regardless of the object size.

$$\begin{aligned}P &= \frac{\left|\left\{\mathbf{x}_{\text{pd}} \in \mathcal{P}_{\text{pd}} \mid \min_{\mathbf{x}_{\text{gt}} \in \mathcal{P}_{\text{gt}}} \|\mathbf{x}_{\text{gt}} - \mathbf{x}_{\text{pd}}\| < \eta\right\}\right|}{|\mathcal{P}_{\text{pd}}|}, \\ R &= \frac{\left|\left\{\mathbf{x}_{\text{gt}} \in \mathcal{P}_{\text{gt}} \mid \min_{\mathbf{x}_{\text{pd}} \in \mathcal{P}_{\text{pd}}} \|\mathbf{x}_{\text{pd}} - \mathbf{x}_{\text{gt}}\| < \eta\right\}\right|}{|\mathcal{P}_{\text{gt}}|},\end{aligned} \tag{2}$$

$$\text{F-1 score} = \frac{2PR}{P + R}. \tag{3}$$

Normal consistency (NC). We use the symmetrized normal consistency metric, borrowed from ConvONet [2].

$$\begin{aligned} \mathrm{NC}(\mathbf{N}_{\mathrm{pd}}, \mathbf{N}_{\mathrm{gt}}) = & \frac{1}{2|\mathbf{N}_{\mathrm{pd}}|} \sum_{\mathbf{n}_{\mathrm{pd}} \in \mathbf{N}_{\mathrm{pd}}} \left(\mathbf{n}_{\mathrm{pd}} \cdot \mathbf{n}^*_{\mathrm{gt}}\right) \\ & + \frac{1}{2|\mathbf{N}_{\mathrm{gt}}|} \sum_{\mathbf{n}_{\mathrm{gt}} \in \mathbf{N}_{\mathrm{gt}}} \left(\mathbf{n}_{\mathrm{gt}} \cdot \mathbf{n}^*_{\mathrm{pd}}\right), \end{aligned} \tag{4}$$

where $\mathbf{n}^*_{\mathrm{gt}}$ and $\mathbf{n}^*_{\mathrm{pd}}$ represent the closest normal vectors, respectively.

B.3. Qualitative Results

We showcase qualitative results for 3D reconstruction (Figures 7 to 9) and grasp pose prediction (Figures 10 to 12). These examples are randomly sampled from each split or scene, highlighting ZeroGrasp's ability to handle a wide range of target objects. For turntable visualizations and real-robot evaluations, please refer to our website.

B.4. Runtime Analysis

The inference speed on NVIDIA A100 is 212 ms with GPU memory usage below 8GB. Average runtimes for GraspNet, GSNet, and Ma *et al*. are 121 ms, 98 ms, and 238 ms, respectively. Despite additionally reconstructing 3D objects, our runtime remains near real-time and our method still achieves the best AP metric.

B.5. Analysis on Segmentation Masks

We tested ZeroGrasp with ground truth masks but observed only a marginal improvement of +0.45 of AP on average in the GraspNet-1B benchmark. We perform an additional experiment where ZeroGrasp uses a foreground mask instead of an instance mask sand performs reconstruction and grasp pose prediction at the scene level. In practice, the F-1 score for reconstruction on the hard split of the ReOcS dataset drops to 80.23, which is 0.63 lower, primarily due to artifacts merging multiple objects. AP scores on GraspNet-1B show minimal change on average, with 69.91 (-0.62), 62.37 (-0.14), and 27.21 (+0.75) for the seen, similar, and novel splits, respectively. This suggests that foreground masks can replace instance masks, but at the cost of reduced reconstruction quality and 3D instance segmentation.

References

[1] L. K. Iwase, Shun and, V. Guizilini, A. Gaidon, K. Kitani, R. Ambruș, and S. Zakharov, "Zero-shot multi-object scene completion," in *ECCV*, 2024. 1

[2] S. Peng, M. Niemeyer, L. Mescheder, M. Pollefeys, and A. Geiger, "Convolutional Occupancy Networks," in *ECCV*, 2020. 2

[3] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-1billion: A large-scale benchmark for general object grasping," in *CVPR*, 2020. 7, 8

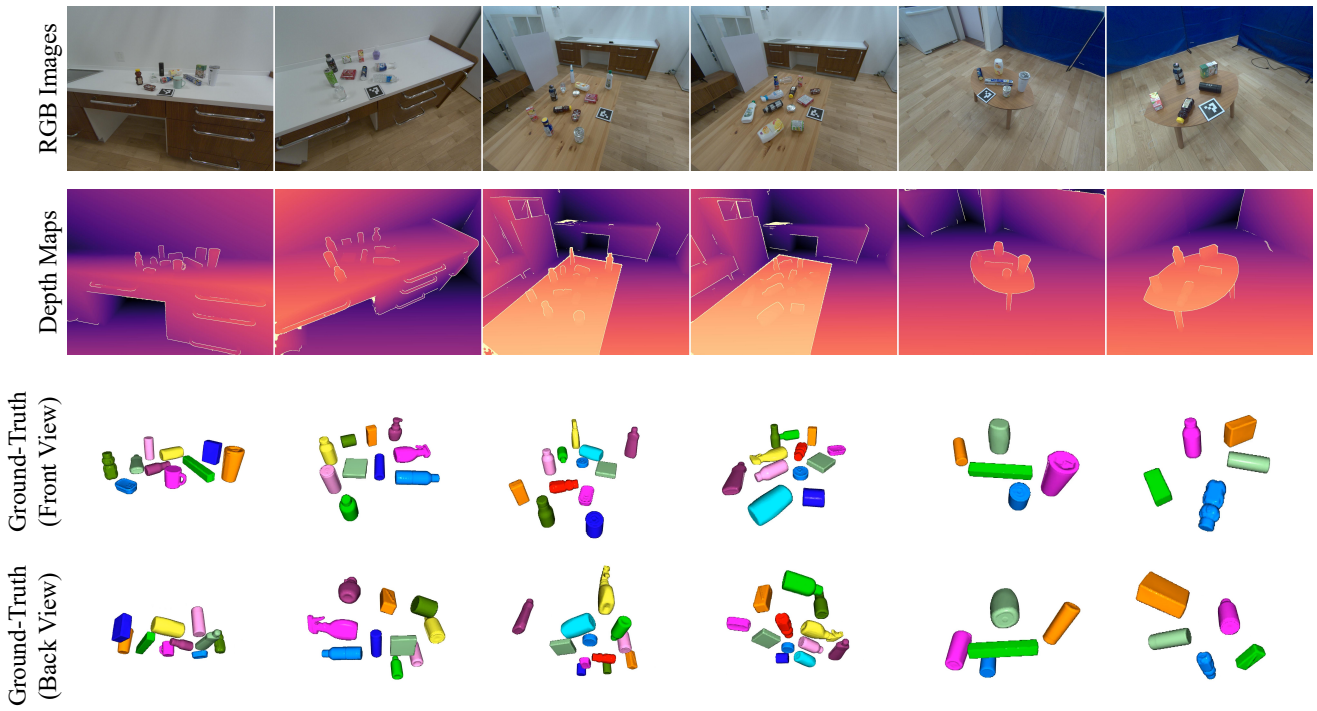

Figure 1. Examples from the easy split of the ReOcS dataset, where ground-truth 3D models are represented as octrees derived from their corresponding meshes.

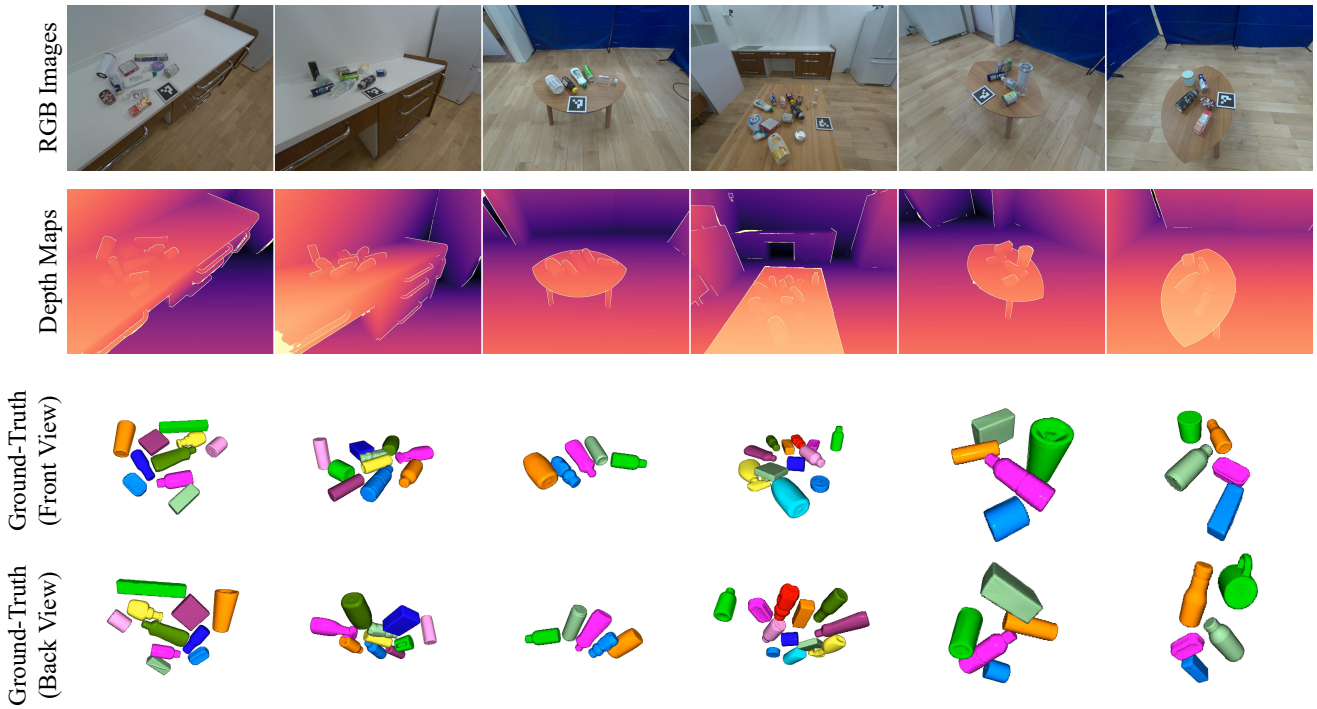

Figure 2. Examples from the normal split of the ReOcS dataset, where ground-truth 3D models are represented as octrees derived from their corresponding meshes.

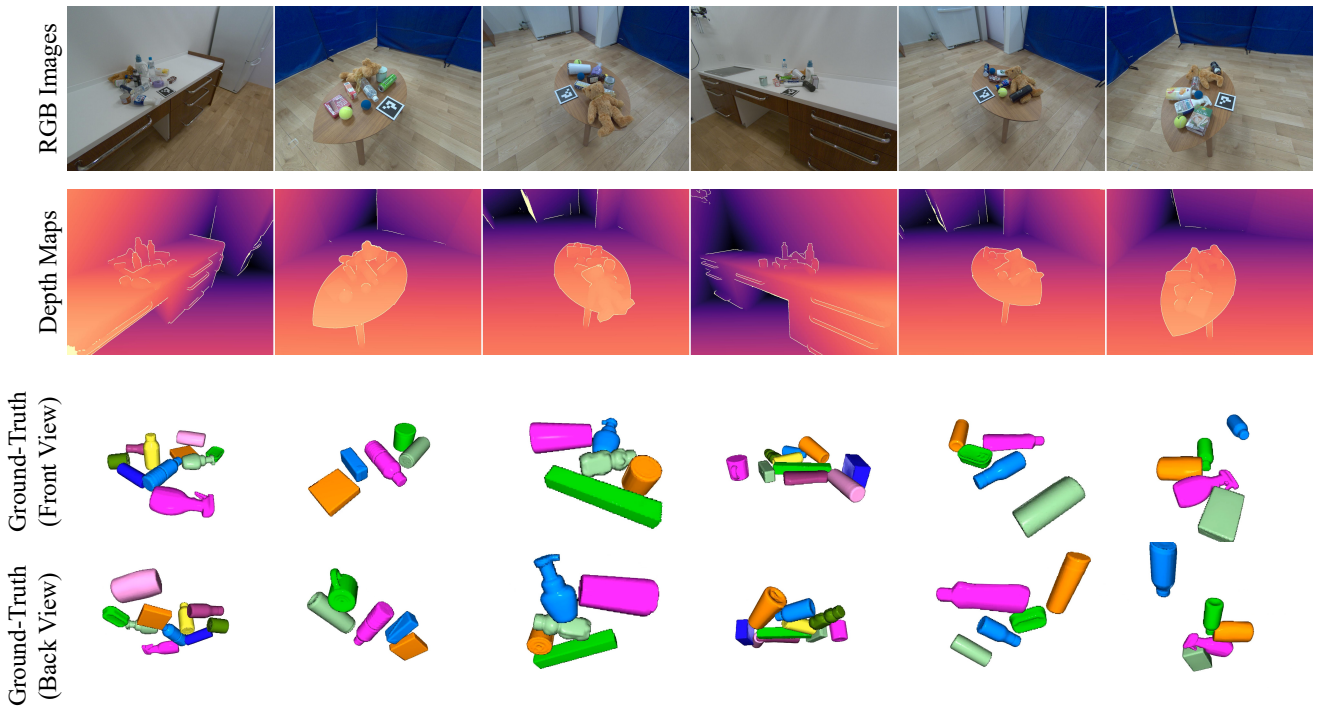

Figure 3. Examples from the hard split of the ReOcS dataset, where ground-truth 3D models are represented as octrees derived from their corresponding meshes.

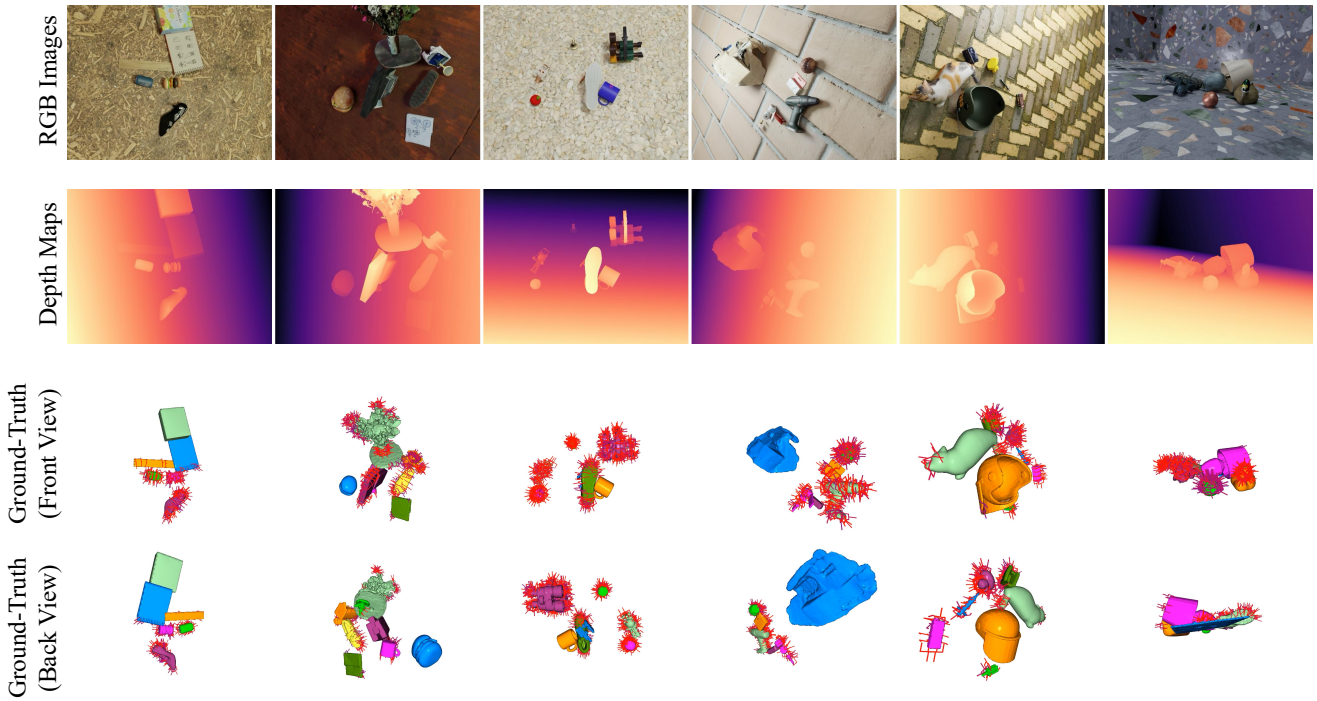

Figure 4. Examples from the ZeroGrasp-11B dataset, where ground-truth 3D models and grasp poses are represented as octrees derived from their corresponding meshes and two-finger parallel grippers.

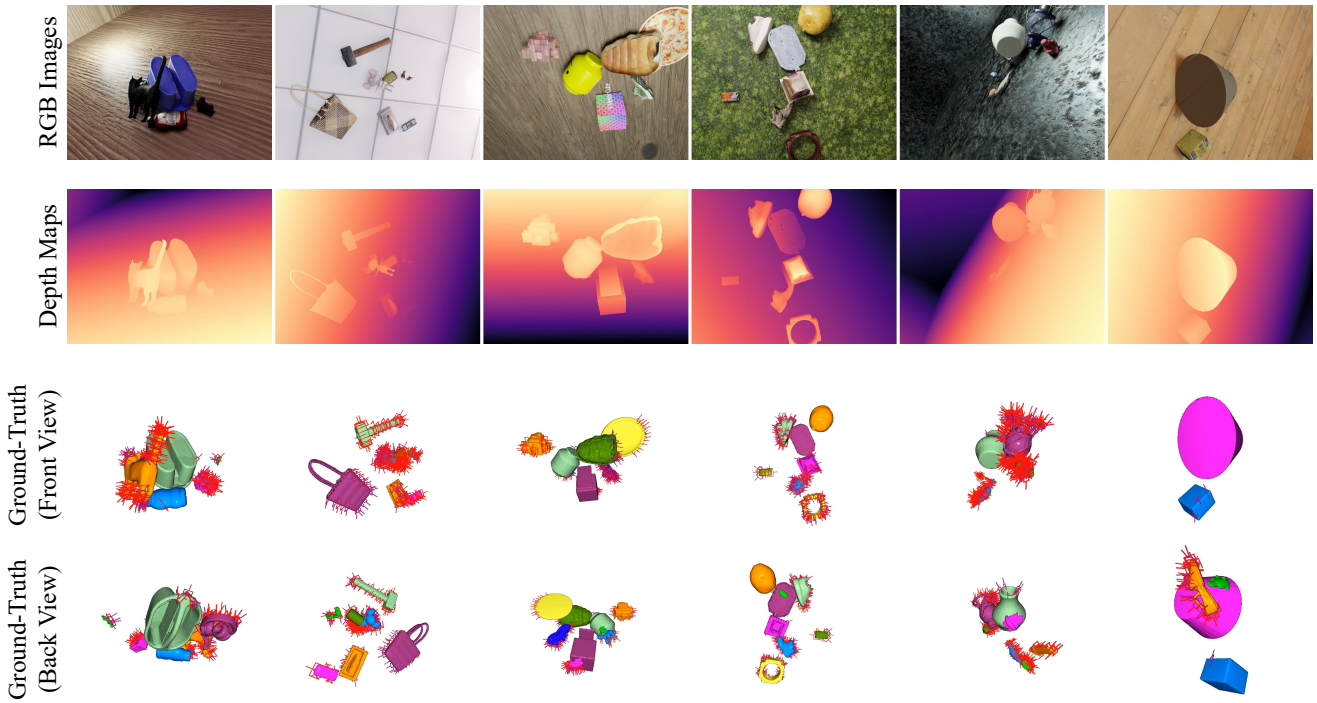

Figure 5. Examples from the ZeroGrasp-11B dataset, where ground-truth 3D models and grasp poses are represented as octrees derived from their corresponding meshes and two-finger parallel grippers.

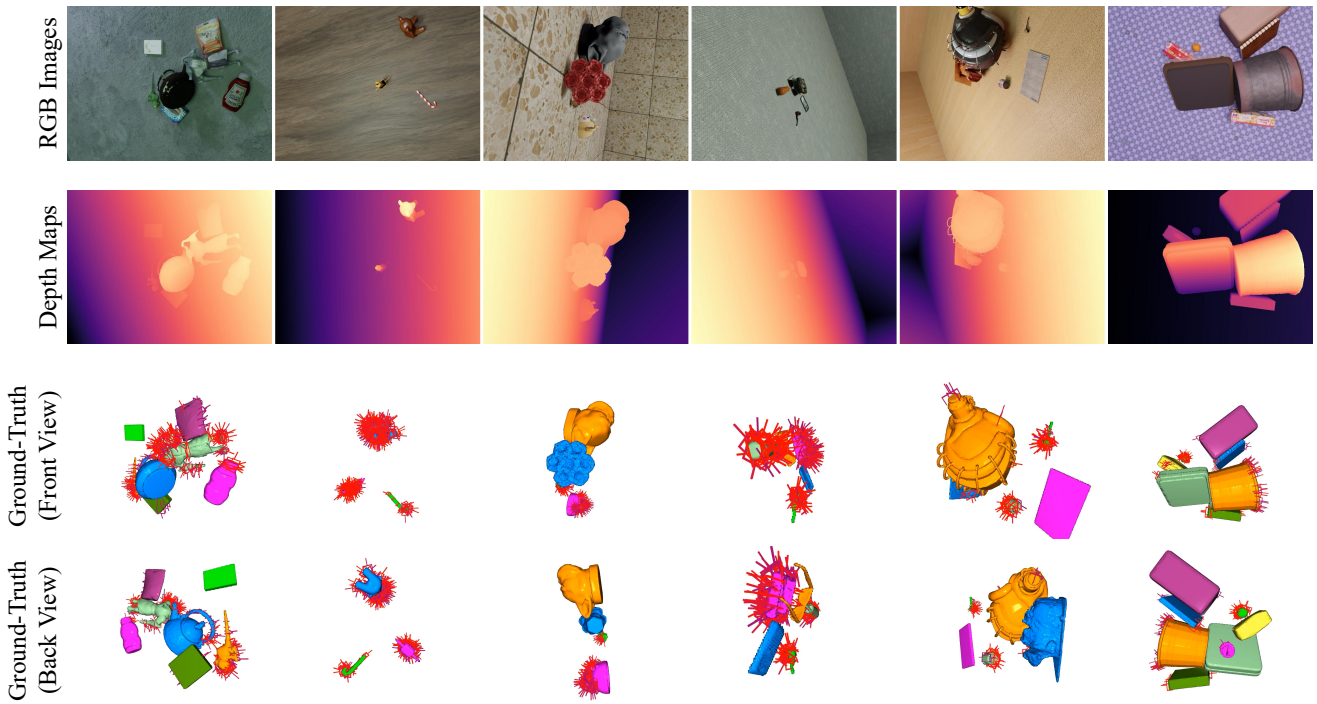

Figure 6. Examples from the ZeroGrasp-11B dataset, where ground-truth 3D models and grasp poses are represented as octrees derived from their corresponding meshes and two-finger parallel grippers.

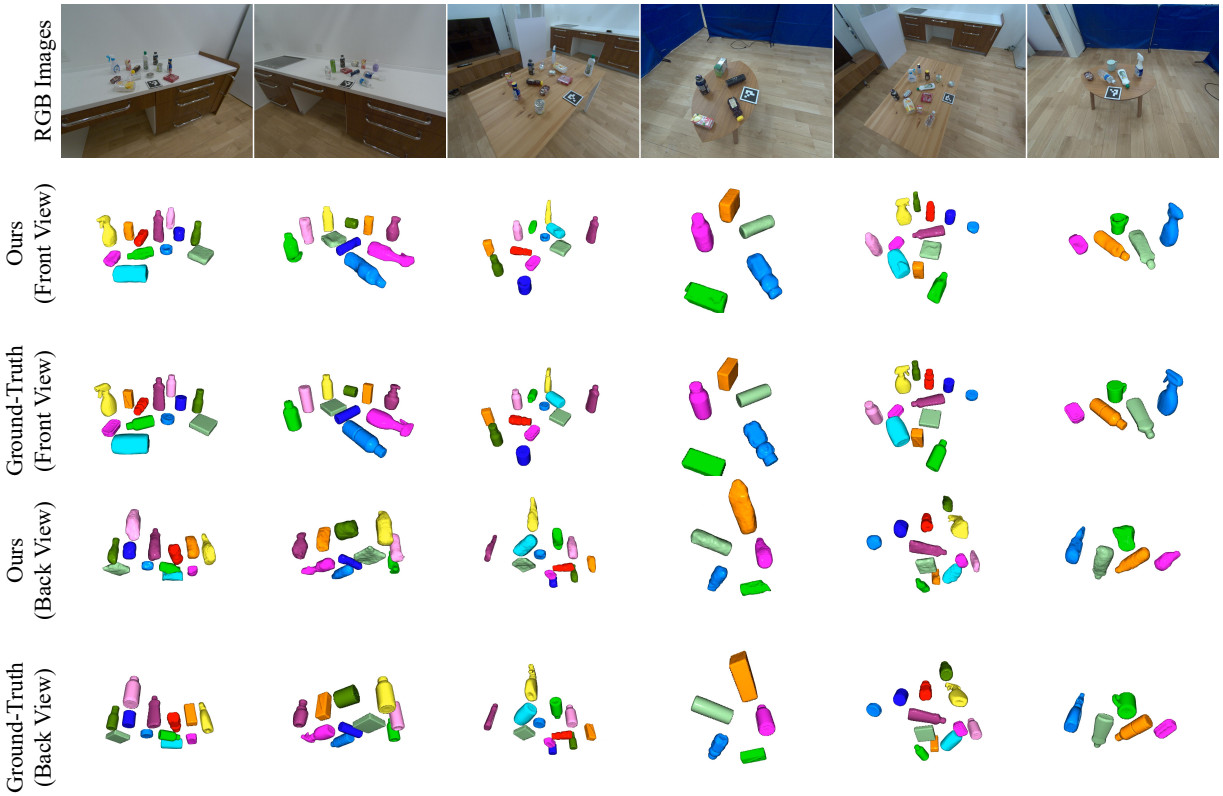

Figure 7. Results of 3D reconstruction on the easy split of the ReOcS dataset.

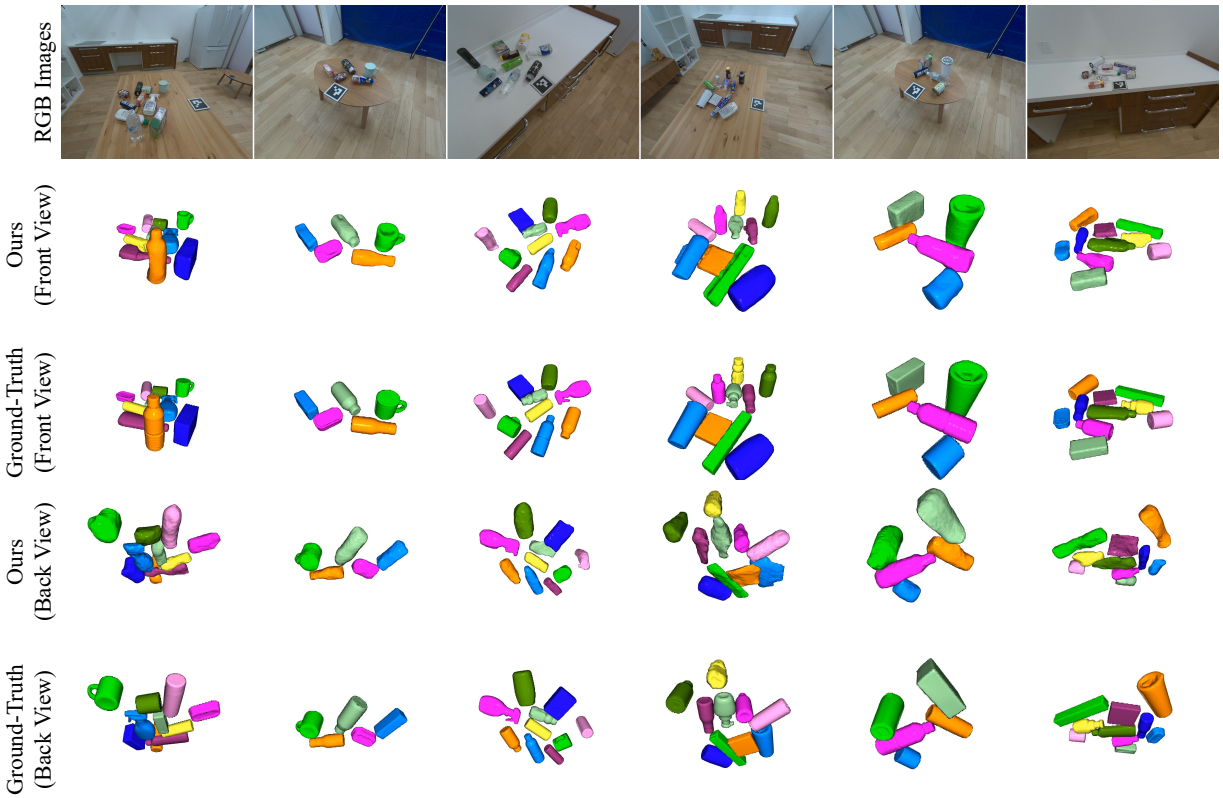

Figure 8. Results of 3D reconstruction on the normal split of the ReOcS dataset.

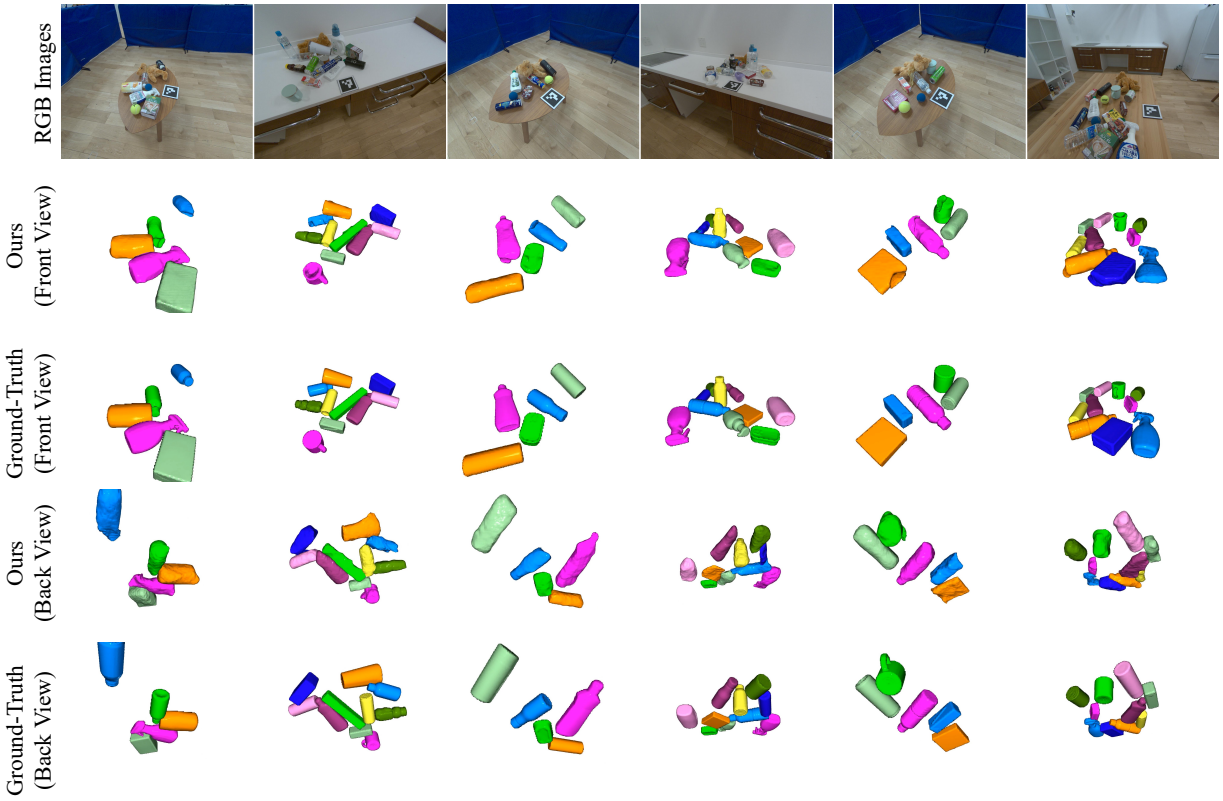

Figure 9. Results of 3D reconstruction on the hard split of the ReOcS dataset.

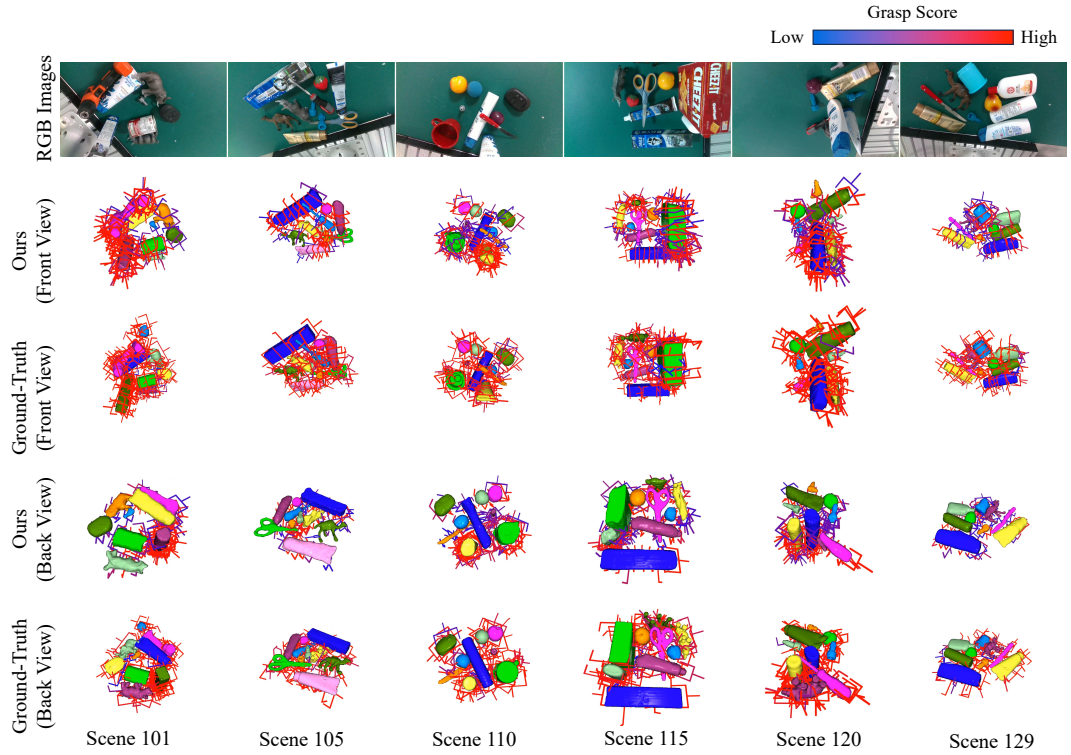

Figure 10. Results of grasp pose prediction on the seen split of the GraspNet-1B dataset. Grasp-NMS [3] is applied to discard redundant grasps for better visibility of 3D reconstructions and grasp poses.

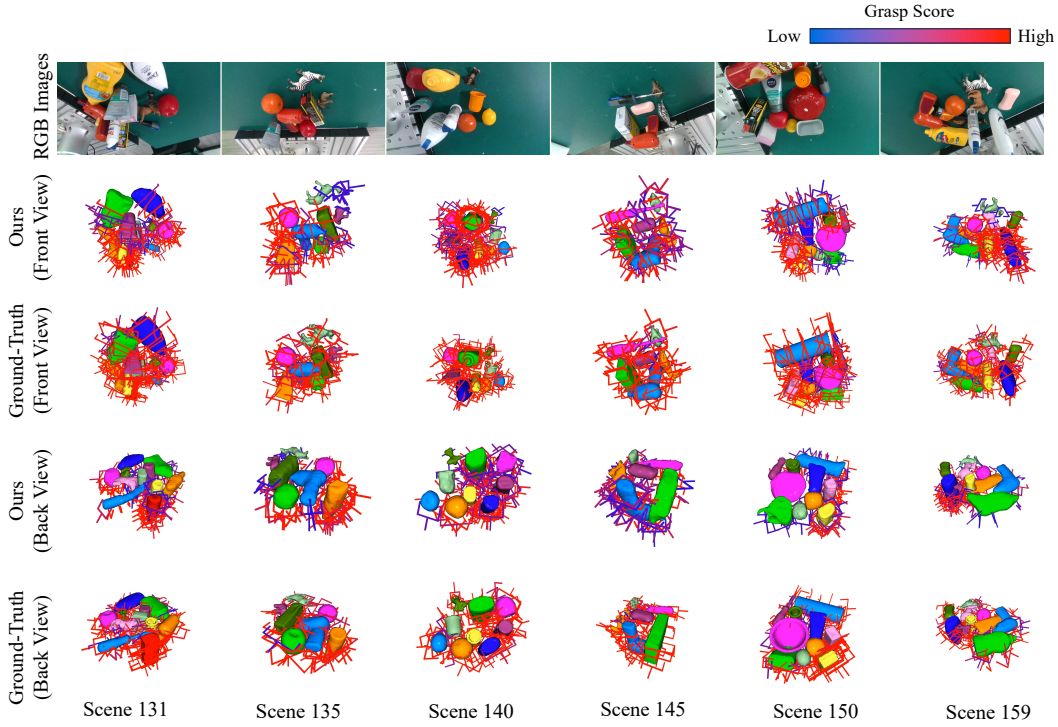

Figure 11. Results of grasp pose prediction on the similar split of the GraspNet-1B dataset. Grasp-NMS [3] is applied to discard redundant grasps for better visibility of 3D reconstructions and grasp poses.

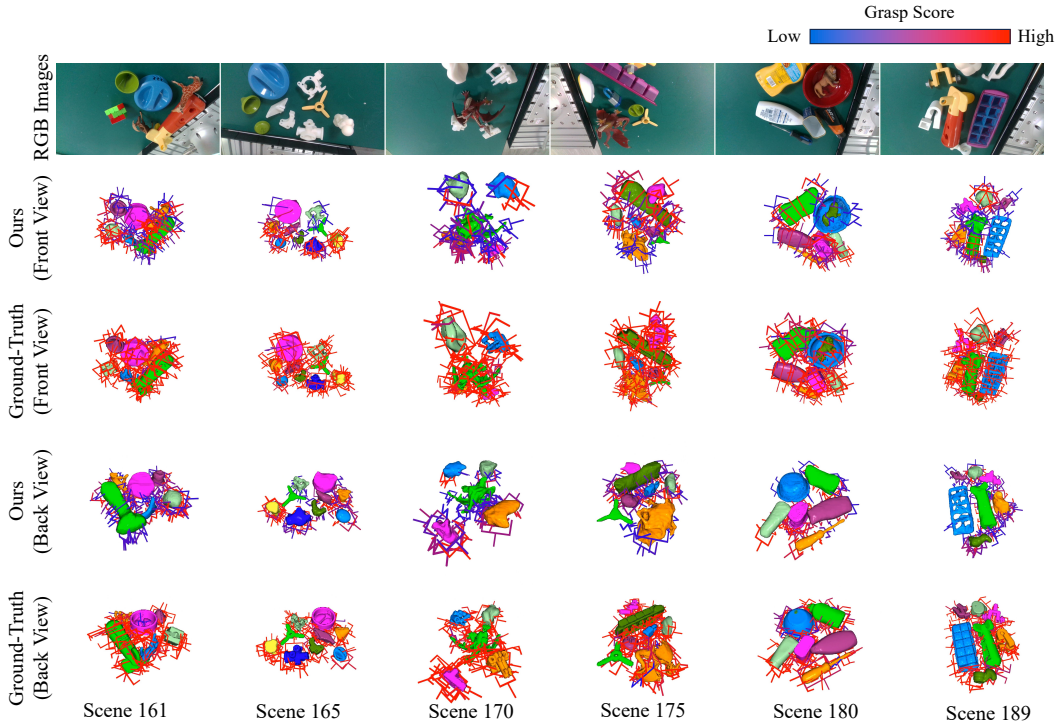

Figure 12. Results of grasp pose prediction on the novel split of the GraspNet-1B dataset. Grasp-NMS [3] is applied to discard redundant grasps for better visibility of 3D reconstructions and grasp poses.