

Appendix

A ROTATION OF SPHERICAL HARMONICS

Spherical harmonics $Y_{lm}(\theta, \phi)$ are functions defined on the surface of a sphere. To rotate spherical harmonics, we use the Wigner D-matrix (Wigner, 1931). The Wigner D-matrix represents rotation operators for angular momentum states and facilitates rotations in the spherical harmonics space. Algorithm A1 outlines the process for rotating spherical harmonics given a rotation transformation R .

Algorithm 1: Rotate Spherical Harmonics with Wigner D-matrix

Input: Spherical harmonics Y_{lm} , Rotation matrix R Euler angles (α, β, γ)

Output: Rotated spherical harmonics \tilde{Y}_{lm}

```

begin
   $(\alpha, \beta, \gamma) \leftarrow \text{RotationMatrix2EulerAngles}(R)$ 
  foreach  $l$  in degrees of spherical harmonics do
     $D^l \leftarrow \text{ComputeWignerDMatrix}(l, \alpha, \beta, \gamma)$ 
    for  $m = -l$  to  $l$  do
       $\tilde{Y}_{lm} \leftarrow \sum_{m'=-l}^l D_{mm'}^l Y_{lm'}$ 
    end
  end
  return  $\tilde{Y}_{lm}$ 
end

```

B DATASETS

B.1 REAL DATASET

As shown in Fig. A1a, we capture the dataset in the lab environment using a Zivid 2 camera which offers a very high nominal depth precision of $0.3mm$. The object was placed at the center of a turntable, with an additional ChArUco board used for camera pose estimation. Fig. A1b displays an example of the camera pose estimation results. We positioned the cameras at 5 or 7 different elevation angles, aiming roughly at the center of the object. Both color and depth images were captured at a resolution of 1944×1200 . All images were cropped according to the camera intrinsic parameters to ensure compatibility with a simple pinhole camera model, aligning with the camera model used in 2DGS. The images were further downsampled by half, resulting in a final image resolution of 944×560 . To obtain the ground truth mesh, we performed TSDF fusion (Newcombe et al., 2011) using the Open3D (Zhou et al., 2018) library. The voxel size was set to $0.3mm$ and the truncated threshold to $1.5mm$. Examples of the real dataset are shown in Fig. A4.

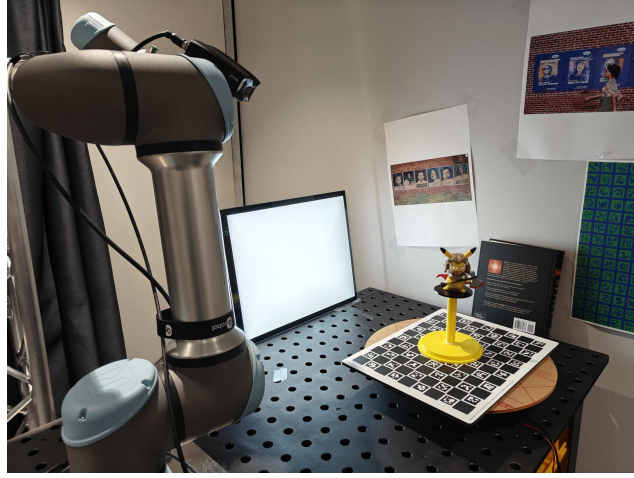
B.2 SYNTHETIC DATASET

The synthetic dataset is rendered using Blender’s (Community, 2018) Cycles engine. The object is placed within a typical indoor scene. Cameras are positioned to view the object from 5 elevation angles uniformly distributed within $(-30, 30)$ degrees. For each camera position, the object is rotated by 10 degrees per step, resulting in 180 images at a resolution of 800×800 for each object. Of these images, 60% (108) are used for training and 20% are allocated for evaluation and testing respectively. Examples of the synthetic dataset are shown in Fig. A5.

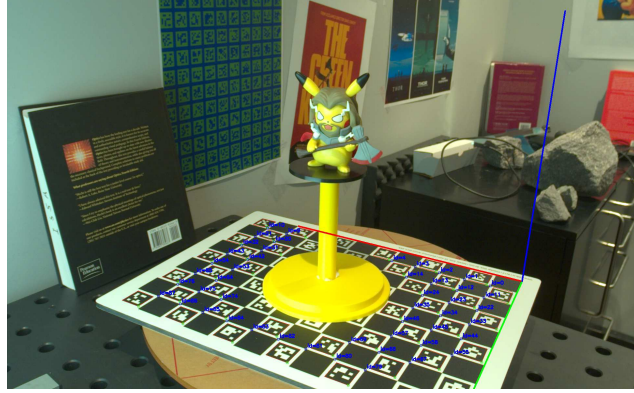
C ADDITIONAL EXPERIMENTS

C.1 ADDITIONAL EXPERIMENTS ON MORE COMPREHENSIVE SCENARIOS

To further validate the robustness and generalizability of our proposed method, we conducted additional experiments on more complex scenarios. We synthesize an additional dataset where the

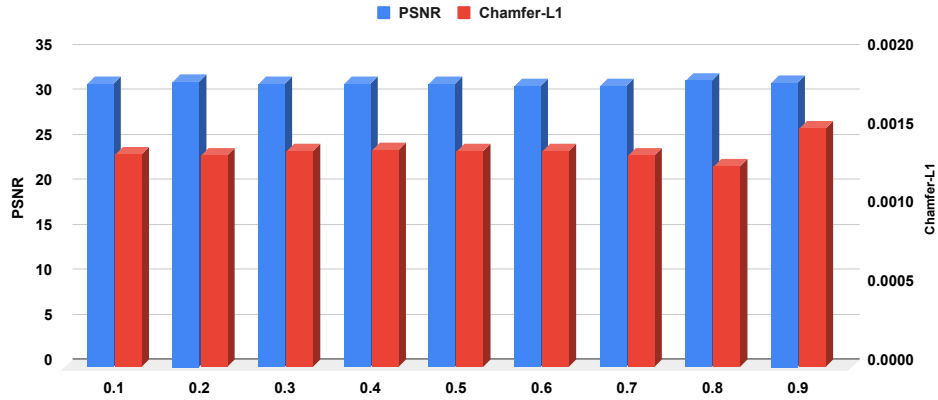


(a) Our image capture setup.



(b) Visualization of camera pose estimation result.

Figure A1: Dataset capture setup.

Figure A2: Performance of our method vs. probability threshold τ .

object not only rotates but also traverses an elliptical path with random radii and rotations, similar to the motion of a planet around a star. As reported in Tab. A2, the increased movement complexity leads to a slight decrease in performance, our method remains competitive with 2DGS in terms of

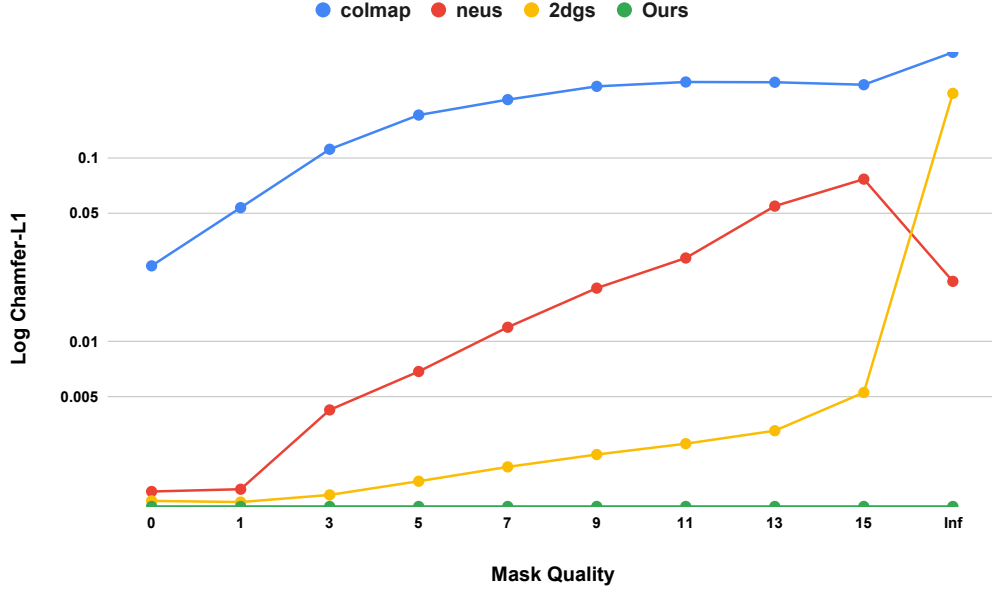


Figure A3: Reconstruction metric (Chamfer- \mathcal{L}_1) w.r.t. different levels of mask quality. A higher quality value means a lower quality mask, and Inf means no mask is applied.

	Methods	Crab	Insect	Leaves	Marci	Cockchafer	Miyuki	Pigeon	Plant1	Plant2	Avg.
w/ mask	COLMAP	0.4420	0.3745	6.0029	1.1857	0.4001	0.8536	0.4014	2.1398	1.7228	1.5025
	NeuS	0.5033	0.3114	0.9274	0.8205	0.4622	0.6948	0.7476	1.8733	1.0726	0.8237
	2DGS	0.4587	0.3551	0.6156	0.7757	0.3194	0.5548	0.5288	0.8849	0.8910	0.5982
w/o mask	NeuS	13.4557	2.7616	9.3365	30.1131	33.9812	29.8241	9.6987	13.4225	24.3946	18.5542
	2DGS	39.2561	43.8407	38.0376	89.3802	60.4679	53.3285	44.2398	-	31.5418	50.0116
	D-3DGS	13.2224	177.8401	68.6867	24.1406	22.0735	13.4585	63.2310	51.7127	31.3708	51.7485
	S2GS (ours)	0.4361	0.3297	0.5154	0.6481	0.3378	0.5592	0.5529	0.9829	0.5378	0.5444

Table A1: All experiment results on the synthetic dataset. We report the Chamfer- $\mathcal{L}_1 \downarrow$ and color each cell as **best**, **second**, and **third**. The results are scaled up by $100\times$ for better comparison. Masks are perfectly obtained.

novel-view synthesis. However, the reconstruction quality of 2DGS drops dramatically due to worse mask quality while our method performs only slightly worse than the rotation-only case.

Method	PSNR	Chamfer- \mathcal{L}_1	IoU
2DGS	22.50	4.7072	0.5269
Ours	22.37	0.8019	0.6432

Table A2: Experiment results on the new synthetic dataset.

C.2 ANALYSIS OF HYPERPARAMETER SENSITIVITY.

We perform experiments on the sensitivity of hyper-parameters. To be specific, we evaluate the influence of the probability threshold τ which we use to segment the object and the background. Figure A2 shows our method performs robustly against the hyperparameter.

C.3 INFLUENCE OF MASK QUALITY

To demonstrate the effectiveness of the proposed method, we conducted additional experiments to evaluate the influence of mask quality. Specifically, we simulated image masks with varying levels of quality by applying image dilation to the original masks. The experimental results are presented

Method	0	1	3	5	7	9	11	13	15	Inf
COLMAP	2.5895	5.3659	11.1479	17.1102	20.7523	24.5067	25.8846	25.8010	25.0094	37.5152
NeuS	0.1535	0.1579	0.4261	0.6896	1.2004	1.9590	2.8565	5.4713	7.6668	2.1323
2DGS	0.1365	0.1345	0.1471	0.1745	0.2089	0.2439	0.2793	0.3286	0.5301	22.4533
Ours	0.1272	0.1272	0.1272	0.1272	0.1272	0.1272	0.1272	0.1272	0.1272	0.1272

Table A3: Reconstruction evaluation metric (Chamfer- \mathcal{L}_1) on the real dataset with different mask quality. A higher quality value means a lower quality mask, and Inf means no mask is applied. The results show that the performance of other methods drops dramatically with worse image masks, while our method remains consistent since we don't rely on image masks. The results are scaled up by $100\times$ for better comparison.

in Tab. A3 and Fig. A3. It is evident that other methods rely heavily on the quality of the mask, whereas our method maintains consistent performance as it does not utilize image masks at all. Notably, NeuS (Wang et al., 2021) fails in 7 out of 9 cases when the mask quality deteriorates beyond a certain threshold ≥ 11 .

C.4 ADDITIONAL QUALITATIVE RESULTS

We show more qualitative results on the real dataset in Fig. A6 and the synthetic dataset in Fig. A7.



Figure A4: Examples of the real dataset and SAM masks.

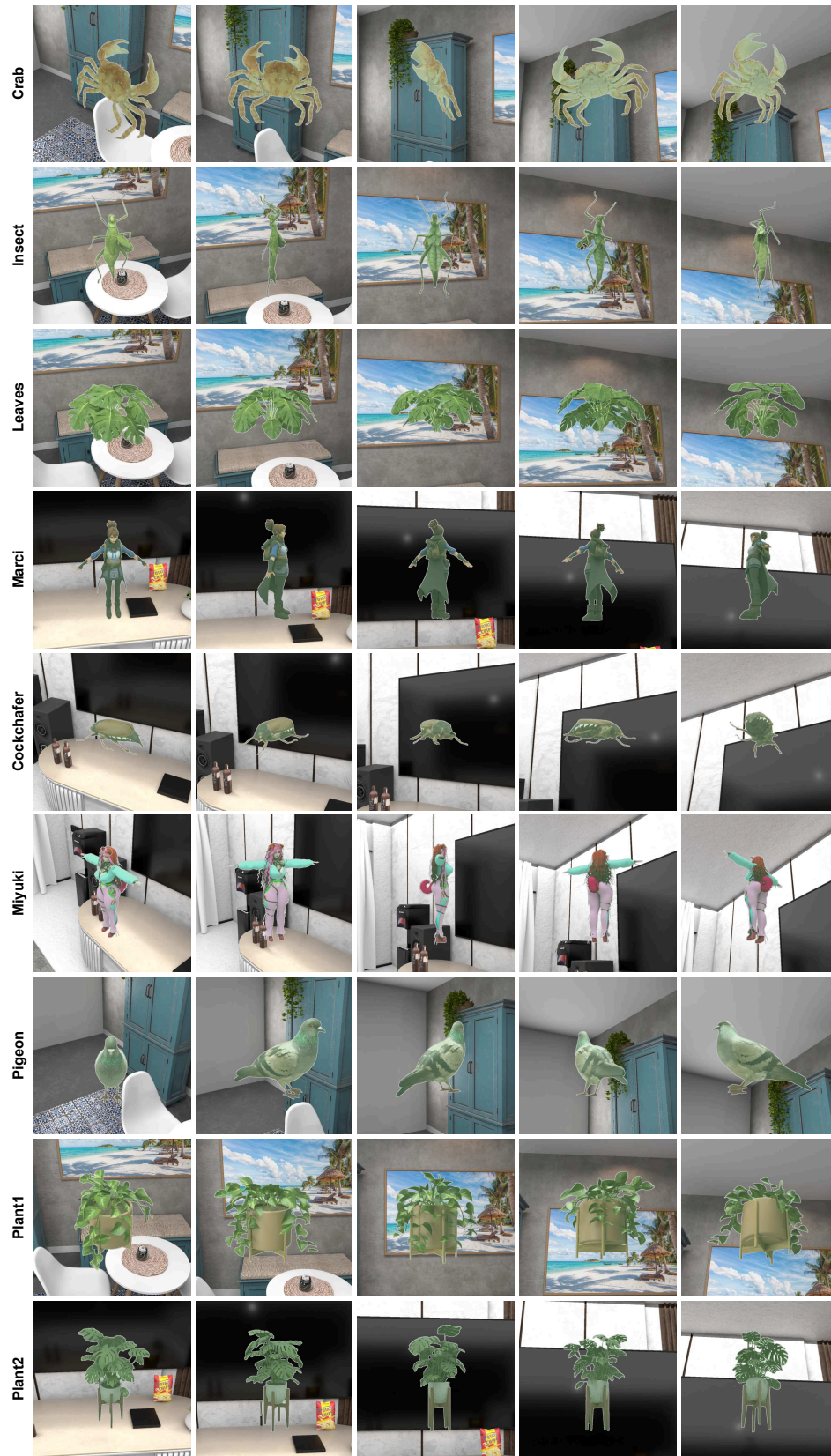


Figure A5: Examples of the synthetic dataset and SAM masks.

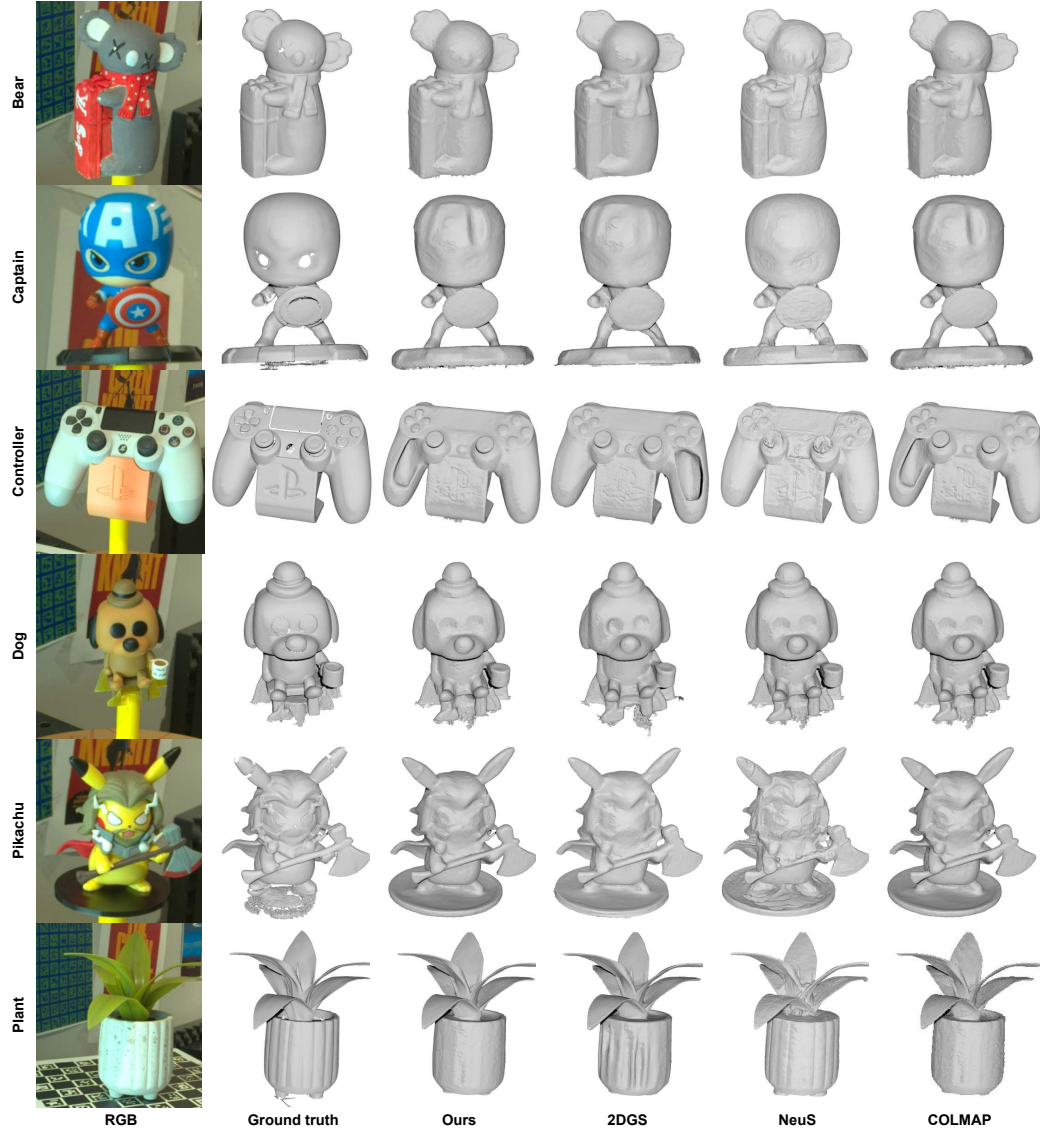


Figure A6: Qualitative results on the real dataset. Note that our method works without an image mask while others work with masks from Track Anything (Yang et al., 2023a).

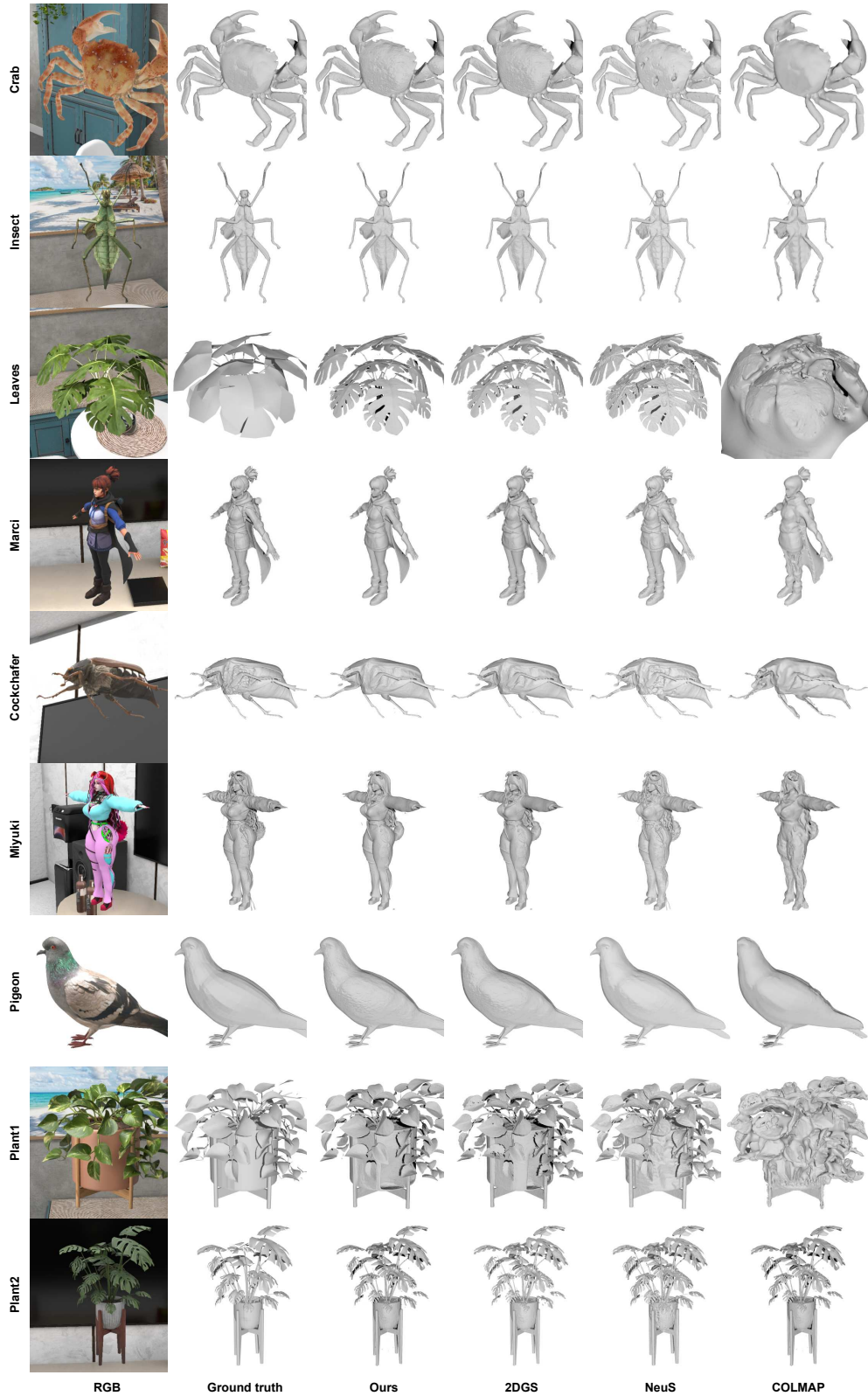


Figure A7: Qualitative results on the synthetic dataset. Note that our method works without an image mask while others work with ground truth masks.