RGB-D Local Implicit Function for Depth Completion of Transparent Objects Supplementary Material

Luyang Zhu[1,2*] Arsalan Mousavian[2] Yu Xiang[2] Hammad Mazhar[2]
Jozef van Eenbergen[2] Shoubhik Debnath[2] Dieter Fox[1,2]
[1]University of Washington [2]NVIDIA

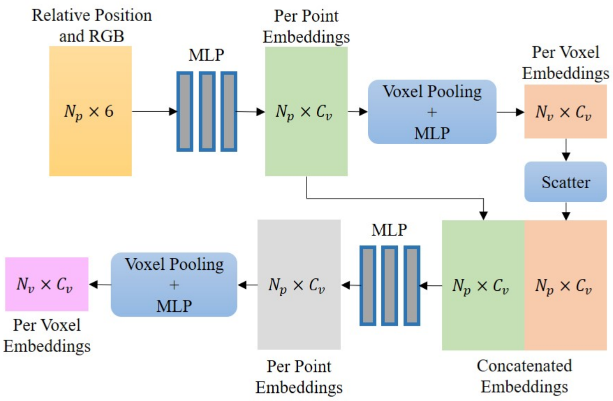

Figure 1. voxel-based PointNet Encoder.

1. Voxel-based PointNet Encoder

In this section, we provide more details about proposed two-stage voxel-based PointNet encoder. As illustrated in Figure 1, We first compute the relative position of valid points P^{valid} with respect to the center of voxels they reside in. The relative position and color of valid points are sent into a shared MLP to produce the initial per-point embedding $\tilde{\mathcal{H}}^{\text{pcl}} \in \mathbb{R}^{N_p \times C_v}$, where N_p is the number of valid points and C_v is the dimension of voxel embedding. Then we apply the max-pooling to embeddings of all points inside each voxel, followed by another MLP to get initial per-voxel embedding $\tilde{\mathcal{H}}^{\text{vox}} \in \mathbb{R}^{N_v \times C_v}$, where N_v is the number of occupied voxels. To generate the second stage input for the i-th valid point P_i^{valid}, we concatenate the point embedding $\tilde{\mathcal{H}}_i^{\text{pcl}}$ and the voxel embedding $\tilde{\mathcal{H}}_k^{\text{vox}}$, satisfying P_i^{valid} reside in v_k^{occ}. Finally, we feed in the new input and repeat the same process as the first stage to get the voxel embedding $\mathcal{H}^{\text{vox}} \in \mathbb{R}^{N_v \times C_v}$.

2. Omniverse Object Dataset

In this section, we provide more details about our Omniverse Object Dataset. To generate the dataset, following categories from ShapeNet [2] are chosen: phone, bowl, camera, laptop, can, bottle. Following objects from ClearGrasp dataset [5] are chosen: cup-with-waves, flower-bath-bomb, heart-bath-bomb, square-plastic-bottle, stemless-plastic-champagne-glass. Note that we only select training objects from ClearGrasp dataset to make sure testing objects are never seen during training. The background textures are randomly selected from the CC0 TEXTURES Dataset [1]. The textures for opaque objects are randomly selected from CC0 TEXTURES Dataset [1] and Describable Textures Dataset [3].

For each image, we provide the following groundtruth data: depth map, instance segmentation, transparent object segmentation, intrinsic and extrinsic camera parameters, 2d/3d bounding box for each object, 6D pose for each object. Since the depth map created from ray-tracing is not accurate for transparent objects, we utilize a two-pass rendering strategy to solve it. Before the rendering, we randomly select some objects and list them as transparent candidates. During the first pass, materials of all objects are set to opaque and we render all groundtruth data including depth map using real time ray-tracing. During the second pass, we set materials of transparent candidates to glass and render the RGB image using path tracing.

3. Additional Results for Ablation Studies

In this section, we provide quantitative results of ablation studies on ClearGrasp [5] Syn-known, Syn-novel and Real-Known dataset. We also provide qualitative comparison of ablation studies on real images.

Depth refinement model. Table 1 shows that depth refinement model can boost the performance of synthetic novel objects while achieving similar results on synthetic known and real known objects. This further proves that depth refinement model can increase the generality of our approach.

[1] Work done while author was an intern at NVIDIA.

Input Modalities. Table 2 shows that both RGB and depth information contribute a lot to the depth accuracy. In Figure 2, we also provide qualitative comparison of input modalities on real images. RGB information can provide visual cues about object shapes. Our approach can only predict flat planes without RGB input. The depth information can help localize the object in metric space. The prediction is far from the table without depth input.

Ray Information. Table 3 (row 1 and 2 in every sub-table) provides quantitative results of the ray information. Figure 3 further visualizes some examples with and without ray information as input. We can see that ray information can help the model reason about the location and orientation of transparent objects.

Positional encoding. Table 3 (row 1 and 3 in every sub-table) shows that positional encoding can improve the performance on both synthetic and real cases. Figure 4 shows that positional encoding helps the model to learn fine details of small objects or under heavy occlusion.

Voxel grid size. Table 3 (row 1,4,5 in every sub-table) shows that the accuracy will drop a lot if the voxel grid size is too large. Figure 5 provides predictions of real images under various voxel grid size. Smaller grid size leads to harder offset regression and the orientation of objects might be wrong (first row). Larger grid size causes objects splitting because of harder classification (second row).

Training Data. Table 4 and Figure 6 provide quantitative and qualitative comparison on different training data respectively. We can see that training the model on both datasets can get best results.

Ray Pooling. Table 5 shows that argmax performs consistently better than weighted sum on all types of testing data. Figure 7 also shows that argmax can better estimate missing depth of transparent objects on real images.

Candidate points selection. Table 6 shows that directly learning offsets of candidate points is better than sampling points heuristically. Figure 8 further provides some examples on real images, showing that learning offset is more robust to strong background textures.

4. Qualitative Results on NYUV2 Dataset

We have done experiments on the NYUV2 dataset [6] to evaluate the performance of our method on general scenes and non-transparent objects. We corrupt the depth map by randomly creating some large holes. Our models are trained to predict the complete depth map given the corrupted depth map and RGB image. As shown in Figure 9, our method can predict reasonable missing depth in general scenes.

5. Failure Cases

Figure 10 provides examples where our approach fails to complete depth of transparent objects from a single RGB-D image. The first limitation (first row) is that pixels of the same object may be classified into different terminating voxels, thus there might be a crack in the reconstructed object. The second limitation (second row) is that there is no explicit constraint in our approach to force objects contacting the table, leading to objects floating in the air.

6. Discussion and Future Works

There are several interesting directions for future works. We can extend our pipeline by treating each pixel's projection as a cone to account for the lateral noise. Generating training data with a realistic depth noise model [4] helps to improve the robustness of our method. We also plan to investigate depth completion of transparent objects in cluttered scenes with heavy occlusion.

Refinement	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
	ClearGrasp Syn-known					
×	0.014	**0.015**	**0.009**	94.36	97.52	99.51
✓	**0.012**	0.017	**0.009**	**94.79**	**98.52**	**99.67**
	ClearGrasp Syn-novel					
×	0.033	0.048	0.026	64.91	87.34	**99.22**
✓	**0.028**	**0.045**	**0.023**	**68.62**	**89.10**	99.20
	ClearGrasp Real-known					
×	**0.027**	**0.032**	**0.019**	**83.50**	**92.71**	98.57
✓	0.028	0.033	0.020	82.37	92.28	**98.63**

Table 1. Depth refinement model. × denotes prediction from first stage while ✓denotes prediction from refinement model.

RGB	Depth	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
		ClearGrasp Syn-known					
✓	✓	**0.014**	**0.015**	**0.009**	**94.36**	**97.52**	**99.51**
	✓	0.061	0.093	0.050	46.53	72.16	92.15
✓		0.031	0.045	0.026	70.73	90.50	98.76
		ClearGrasp Syn-novel					
✓	✓	**0.033**	**0.048**	**0.026**	**64.91**	**87.34**	**99.22**
	✓	0.063	0.102	0.055	35.11	60.42	92.80
✓		0.075	0.119	0.066	34.70	54.81	84.16
		ClearGrasp Real-known					
✓	✓	**0.027**	**0.032**	**0.019**	**83.50**	**92.71**	**98.57**
	✓	0.071	0.098	0.055	38.30	67.51	91.53
✓		0.080	0.124	0.074	30.05	53.47	83.14

Table 2. Ablation studies for effect of different modalities

References

[1] CC0 TEXTURES. https://cc0textures.com/.

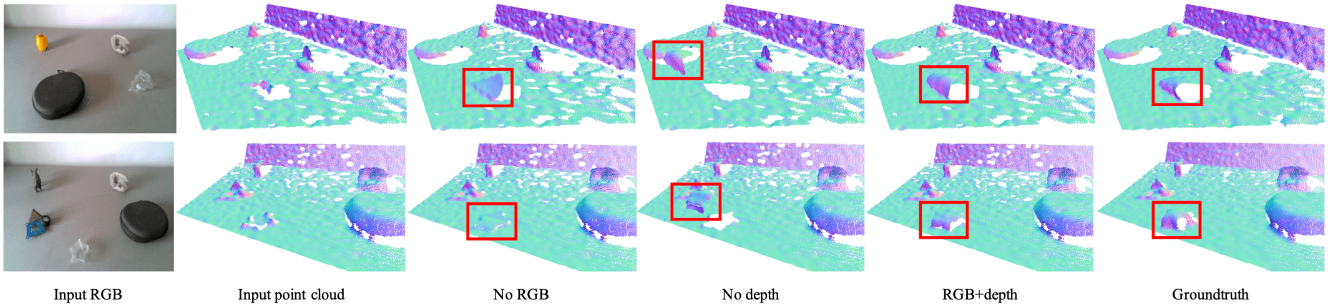

Figure 2. Qualitative results for input modalities. Point clouds are colored by surface normal and rendered in a novel viewpoint to better visualize the 3D shape. The red boxes highlights the interest area. Please zoom in to see details.

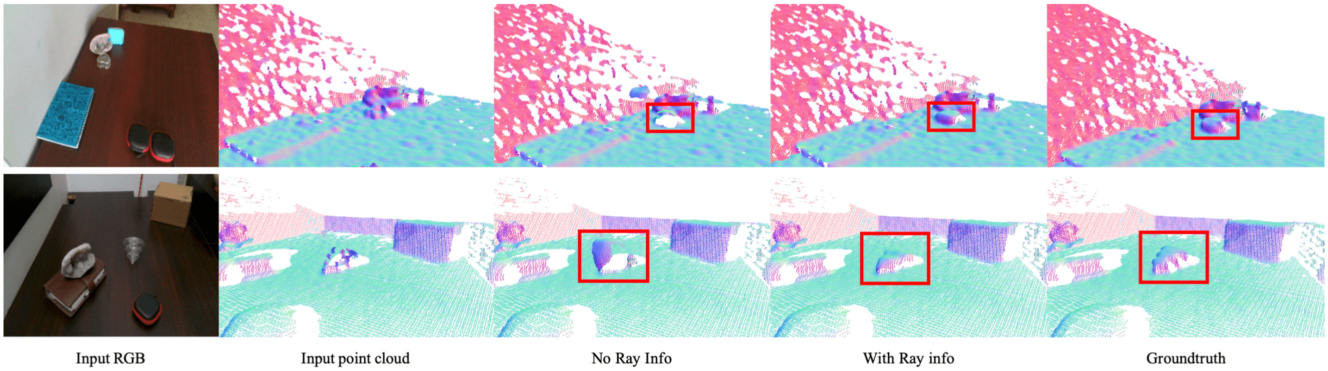

Figure 3. Qualitative results for ray information. Point clouds are colored by surface normal and rendered in a novel viewpoint to better visualize the 3D shape. The red boxes highlights the interest area. Please zoom in to see details.

Ray Info	Pos. Enc	Grid Size	RMSE↓	REL↓	MAE↓	$\delta_{1.05}\uparrow$	$\delta_{1.10}\uparrow$	$\delta_{1.25}\uparrow$
			ClearGrasp Syn-known					
✓	✓	8^3	0.014	**0.015**	**0.009**	**94.36**	97.52	99.51
	N/A	8^3	0.034	0.032	0.019	86.15	93.60	97.92
✓		8^3	0.018	0.024	0.013	89.41	97.33	99.60
✓	✓	4^3	**0.013**	0.017	**0.009**	94.04	**98.00**	**99.70**
✓	✓	16^3	0.017	0.021	0.011	90.62	97.06	99.45
			ClearGrasp Syn-novel					
✓	✓	8^3	0.033	**0.048**	0.026	**64.91**	87.34	**99.22**
	N/A	8^3	0.066	0.089	0.050	49.73	70.88	91.30
✓		8^3	0.041	0.057	0.031	58.88	82.36	97.73
✓	✓	4^3	**0.030**	0.049	**0.025**	64.04	**87.69**	99.09
✓	✓	16^3	0.040	0.057	0.032	61.11	83.85	97.60
			ClearGrasp Real-known					
✓	✓	8^3	**0.027**	**0.032**	**0.019**	**83.50**	**92.71**	**98.57**
	N/A	8^3	0.066	0.072	0.043	61.64	77.98	91.99
✓		8^3	0.032	0.039	0.024	78.07	90.81	96.93
✓	✓	4^3	0.031	0.040	0.024	74.33	90.53	98.47
✓	✓	16^3	0.035	0.044	0.025	71.04	83.90	97.80

Table 3. Ablation Study for different design choices such as including ray information in the embedding, applying positional encoding, and the size of voxels on the accuracy

Data	RMSE↓	REL↓	MAE↓	$\delta_{1.05}\uparrow$	$\delta_{1.10}\uparrow$	$\delta_{1.25}\uparrow$
	ClearGrasp Syn-known					
CG+Omni	**0.014**	**0.015**	**0.009**	**94.36**	**97.52**	**99.51**
Omni	0.063	0.106	0.053	41.80	65.80	89.98
CG	0.023	0.031	0.017	83.56	95.04	99.23
	ClearGrasp Syn-novel					
CG+Omni	**0.033**	**0.048**	**0.026**	**64.91**	**87.34**	**99.22**
Omni	0.062	0.108	0.053	30.11	57.81	91.14
CG	0.041	0.063	0.034	52.69	79.42	98.05
	ClearGrasp Real-known					
CG+Omni	**0.027**	**0.032**	**0.019**	**83.50**	**92.71**	**98.57**
Omni	0.059	0.082	0.048	43.24	70.20	93.74
CG	0.040	0.053	0.031	65.71	84.27	96.60

Table 4. Quantitative effect of training data.

[2] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[3] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[4] Chuong V Nguyen, Shahram Izadi, and David Lovell. Modeling kinect sensor noise for improved 3d reconstruction and tracking. In *2012 second international conference on 3D*

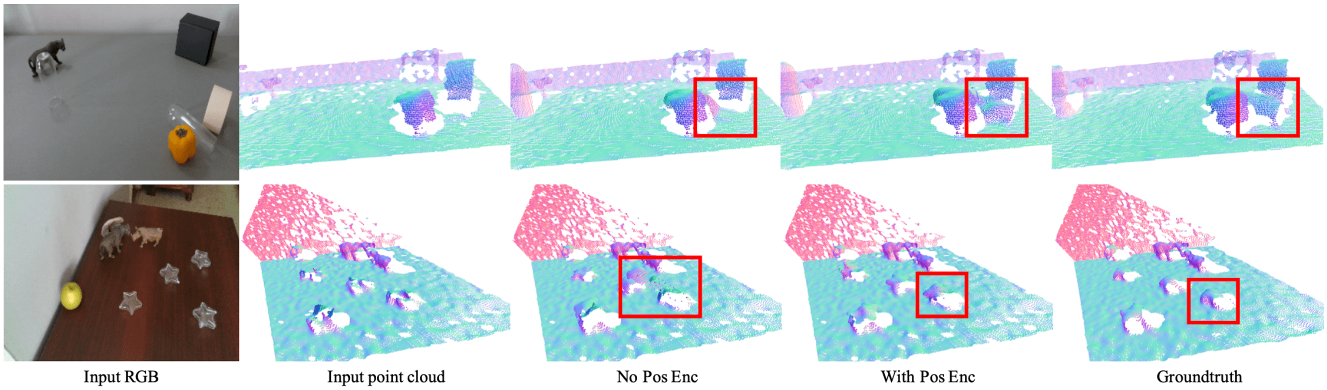

Figure 4. Qualitative results for positional encoding. Point clouds are colored by surface normal and rendered in a novel viewpoint to better visualize the 3D shape. The red boxes highlights the interest area. Please zoom in to see details.

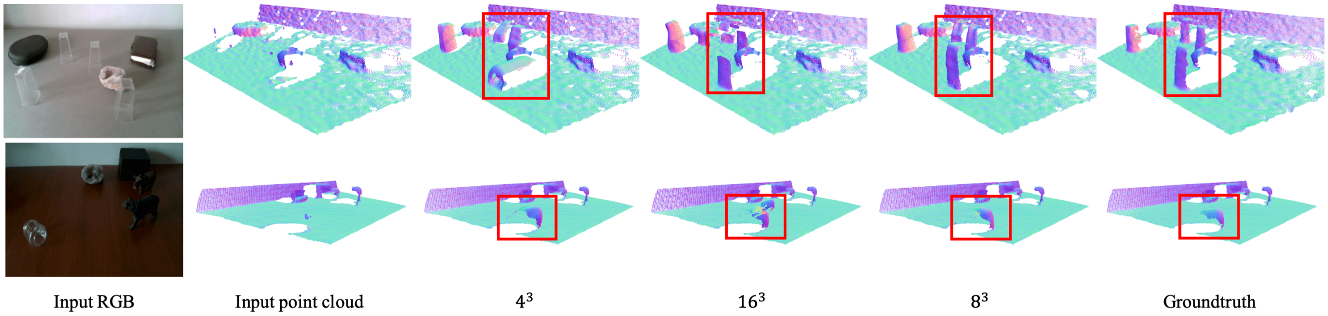

Figure 5. Qualitative results for grid size. Point clouds are colored by surface normal and rendered in a novel viewpoint to better visualize the 3D shape. The red boxes highlights the interest area. Please zoom in to see details.

Ray Pooling	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
	ClearGrasp Syn-known					
Argmax	**0.014**	**0.015**	**0.009**	**94.36**	**97.52**	**99.51**
WeightedSum	0.018	0.028	0.014	85.23	95.78	99.42
	ClearGrasp Syn-novel					
Argmax	**0.033**	**0.048**	**0.026**	**64.91**	**87.34**	**99.22**
WeightedSum	**0.033**	0.051	0.027	63.25	86.14	98.84
	ClearGrasp Real-known					
Argmax	**0.027**	**0.032**	**0.019**	**83.50**	**92.71**	**98.57**
WeightedSum	0.030	0.037	0.022	78.27	91.53	97.83

Table 5. Ablation study for different ray pooling strategies.

imaging, modeling, processing, visualization & transmission, pages 524–530. IEEE, 2012.

[5] Shreeyak Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. Clear grasp: 3d shape estimation of transparent objects for manipulation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 3634–3642, 2020.

[6] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012.

Candidates	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
	ClearGrasp Syn-known					
Learned offset	**0.014**	**0.015**	**0.009**	**94.36**	**97.52**	**99.51**
Sample points	0.019	0.024	0.013	88.84	95.68	98.88
	ClearGrasp Syn-novel					
Learned offset	**0.033**	**0.048**	**0.026**	**64.91**	**87.34**	**99.22**
Sample points	0.035	0.057	0.028	59.17	83.40	97.61
	ClearGrasp Real-known					
Learned offset	**0.027**	**0.032**	**0.019**	**83.50**	**92.71**	98.57
Sample points	0.033	0.041	0.024	73.79	89.22	**98.70**

Table 6. Ablation study for candidate points selection.

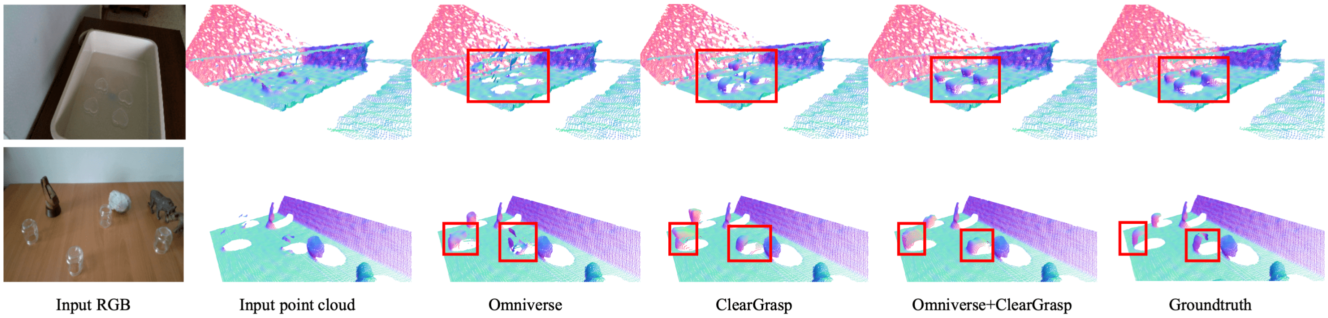

Figure 6. Qualitative results for training data. Point clouds are colored by surface normal and rendered in a novel viewpoint to better visualize the 3D shape. The red boxes highlights the interest area. Please zoom in to see details.

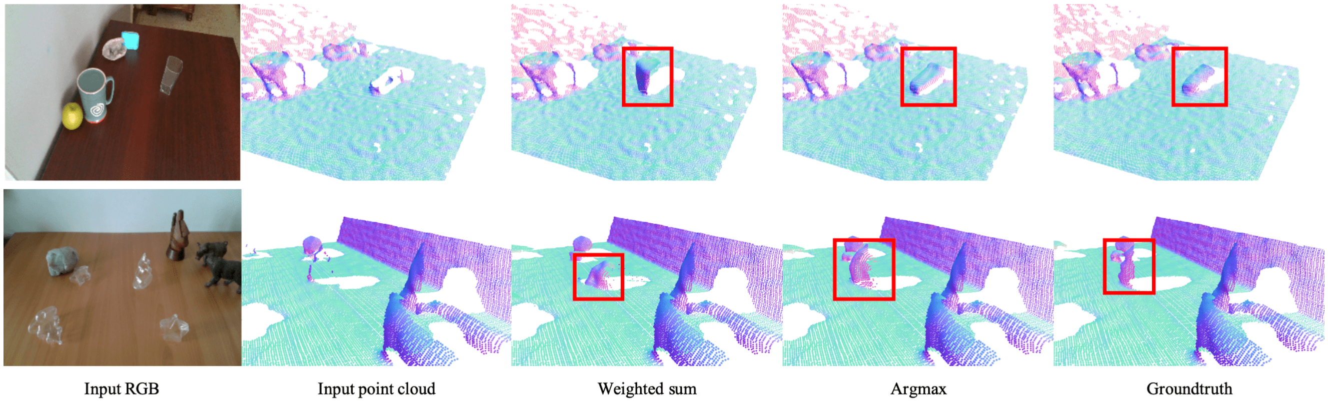

Figure 7. Qualitative results for ray pooling. Point clouds are colored by surface normal and rendered in a novel viewpoint to better visualize the 3D shape. The red boxes highlights the interest area. Please zoom in to see details.

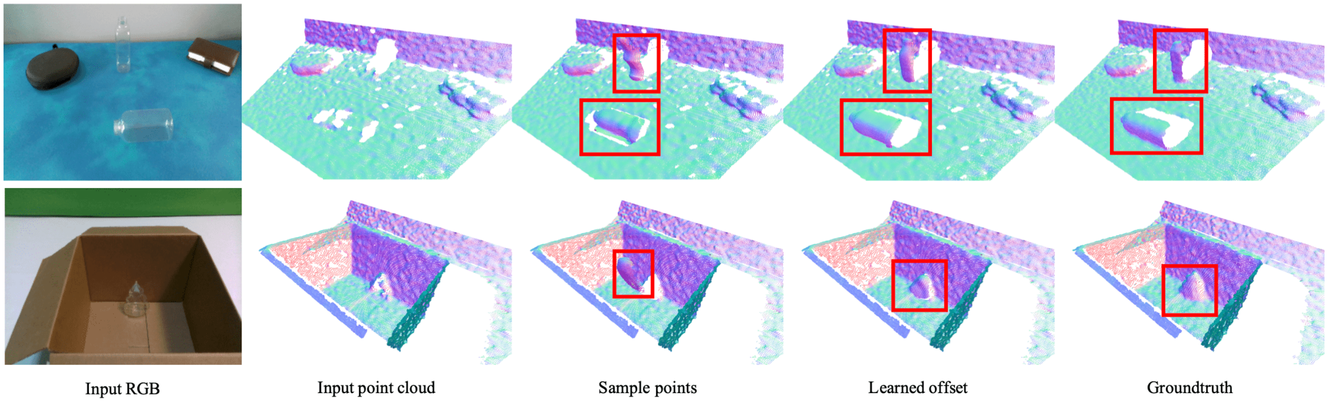

Figure 8. Qualitative results for candidate points selection. Point clouds are colored by surface normal and rendered in a novel viewpoint to better visualize the 3D shape. The red boxes highlights the interest area. Please zoom in to see details.

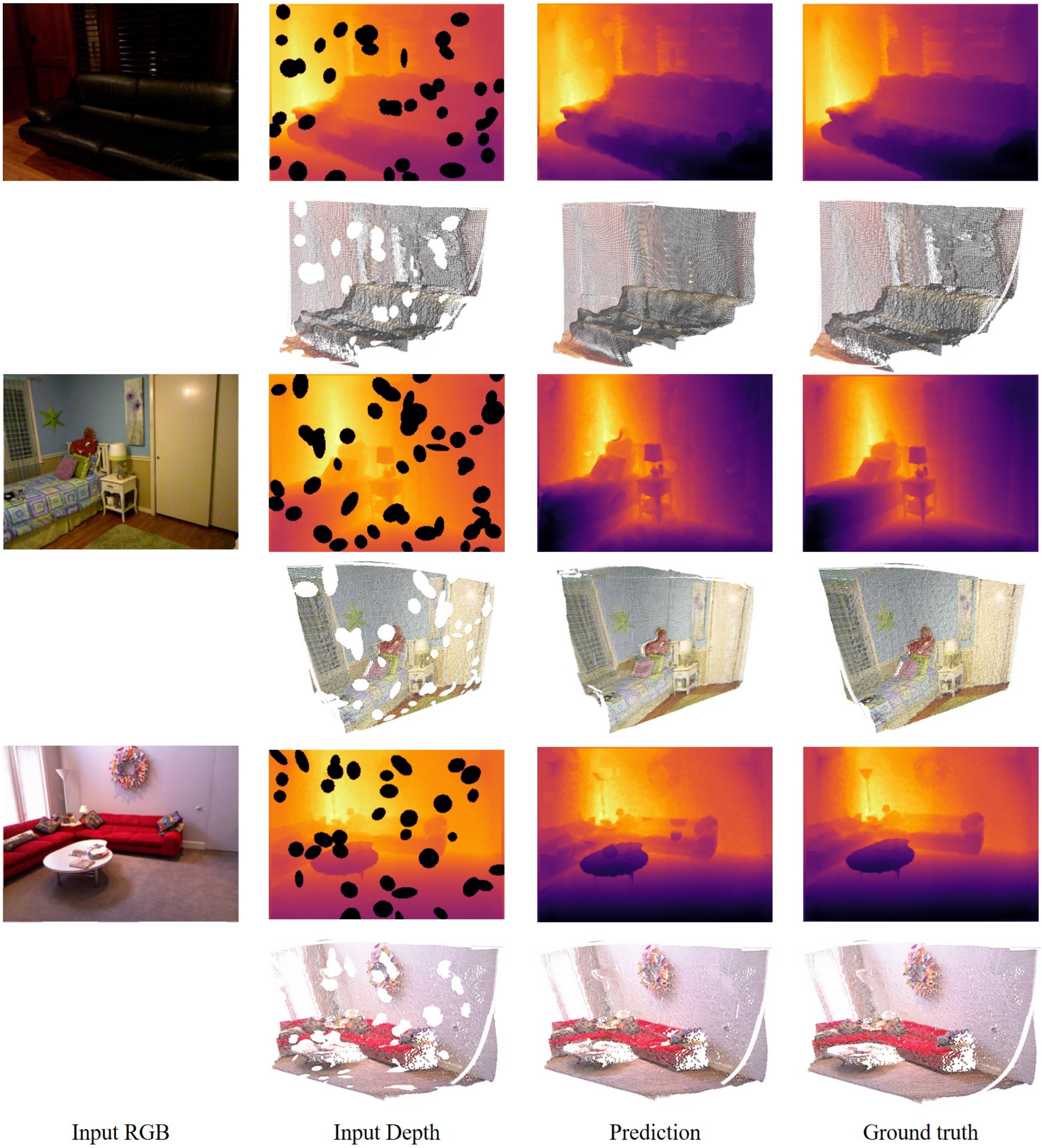

Figure 9. Qualitative results on NYUV2 dataset. For every example, first row from left to right: input RGB, input depth, predicted depth, groundtruth depth; second row from left to right: input point cloud, predicted point cloud, groundtruth point cloud. Point clouds are rendered in a novel viewpoint. Please zoom in to see details.

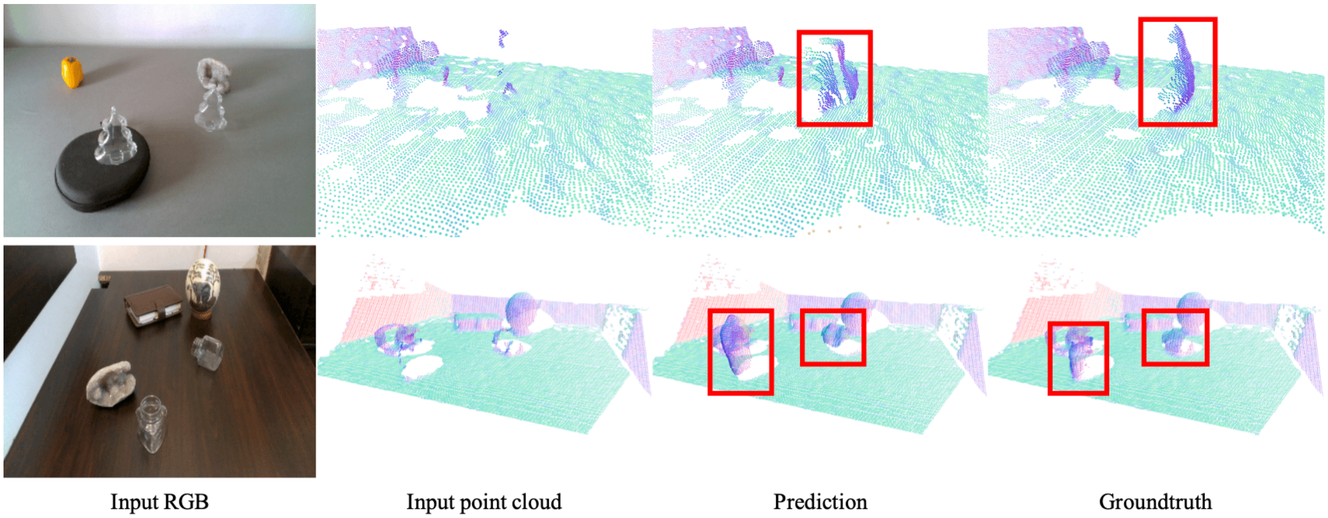

Figure 10. Failure Cases. First row: pixels of the same object are classified into different terminating voxels, leading to a crack in the reconstructed object. Second row: there is no explicit constraint to force objects contacting the table, leading to objects floating in the air. Please zoom in to see details.