On the Importance of Accurate Geometry Data for Dense 3D Vision Tasks – Supplementary Material

HyunJun Jung*[1], Patrick Ruhkamp*[1,2], Guangyao Zhai[1], Nikolas Brasch[1], Yitong Li[1], Yannick Verdie[1,3], Jifei Song[3], Yiren Zhou[3], Anil Armagan[3], Slobodan Ilic[1,4], Ales Leonardis[3], Nassir Navab[1], Benjamin Busam[1,2]

[1] Technical University of Munich, [2] 3Dwe.ai, [3] Huawei Noah's Ark Lab, [4] Siemens AG, * Equal Contribution

hyunjun.jung@tum.de, p.ruhkamp@tum.de, guangyao.zhai@tum.de, b.busam@tum.de

1. Dense 3D Vision Tasks

1.1. Monocular Depth Estimation

Following the results on monocular depth estimation in the main paper, we describe the implementation details of the training, show additional results on different scenes and provide additional metrics on different test scenes.

Implementation Details For all our depth estimation experiments, we use PyTorch [6] and train for 20 epochs for comparability using Adam [3]. Monocular approaches are trained with a batch size of 12 on one NVIDIA RTX-3090 GPU. We chose $\lambda_s = 10^{-3}$ and sample S with $T = 10$ frames offset due to small relative camera movement between frames and the high frame rate. The RGB inputs are scaled to 480×320 for supervised training and to 320×160 for self-supervised training, respectively. The depth network regresses dense depth predictions on four pyramid levels, each with half the resolution of the previous. Pose network and augmentations follow [2]. We choose an initial learning rate of 1×10^{-4} for 15 epochs, which we decrease to 1×10^{-5} after 15 epochs in the self-supervised setting. For the supervised case, we start with a learning rate of 1×10^{-3}, which we decrease every five epochs by a factor of ten.

1.1.1 Quantitative evaluation

Test scenes. Table 1 summarizes the extensive quantitative evaluation of the supervised training with different depth modalities as supervision signal for different test scenes. Test scene 1 has a similar background compared to the training scenes and includes additional unseen objects. The scene is also observed from viewing angles that differ significantly from the training data. The background in test scene 2 is only partly observed in the training data and it includes mostly unseen objects. Test scene 3 is similar to test scene 2, but with a modified object layout and difficult lighting in the background from an additional bright light source above the scene. The additional test set with (partly) seen scenes is an additional test split which includes the first 10 frames of each training sequence. Please note that these frames have not been used during training. Here, we first test all predictions against the rendered ground truth (Top) and additionally on each individual respective modality (Bottom) to highlight the overfitting issue of invalid ground truth from each modality. The results suggest that overall the supervision with accurate rendered ground truth achieves to generalize best for (mostly) unknown scenes. It is noticeable, that the active stereo achieves to produce good predictions for transparent objects and also performs well for reflective ones. The I-ToF and D-ToF predictions suffer from incorrect ground truth values for such objects.

Overfitting on (partly) seen scenes. The (partly) seen scene shows generally lower overall errors for all modalities as compared to the (mostly) unseen test scenes 1,2, and 3. Again, the active stereo can provide decent depth supervision for reflective and transparent objects, where the ToF sensors cannot provide valid depth. The prediction of the background of the scene performs worst for the active stereo, as the textureless wall is still problematic for the sensor.

When testing on the respective modality itself, the overfitting issue due to incorrect depth values of the sensor becomes apparent. It can be noticed, that for objects where the respective sensor cannot yield accurate depth values (e.g. transparent objects for I-ToF or reflective objects for D-ToF), the errors are significantly lower, indicating overfitting to the specific sensor modality.

Table 1. **Depth prediction comparison when training with different modalities and tested on different unseen scenes and seen scenes.** (Top) Evaluation against GT of depth predictions on the test set with dense supervision from different depth modalities. (Bottom) Predictions evaluated on respective modality. Error is reported as Sq.Rel. and RMSE in mm.

	Mask	Full Scene		Background		All Objects		Textured		Reflective		Transparent	
	Metric	Sq.Rel.	RMSE	Sq.Rel.	RMSE	Sq.Rel.	RMSE	Sq.Rel.	RMSE	Sq.Rel.	RMSE	Sq.Rel.	RMSE
Test 1	I-ToF	24.78	148.09	22.25	151.07	29.62	123.19	16.47	99.08	102.79	214.60	44.29	134.44
	D-ToF	24.23	151.72	23.74	159.28	22.85	110.88	16.22	101.12	57.14	148.61	30.23	107.23
	Active Stereo	32.15	173.72	33.84	184.16	22.23	116.57	19.55	114.07	64.27	167.71	12.92	69.49
Test 2	I-ToF	27.42	123.79	22.66	116.86	39.85	139.67	48.66	144.92	16.15	99.44	25.15	122.25
	D-ToF	23.00	115.40	21.18	113.27	27.89	119.59	30.00	112.92	15.81	90.89	23.73	117.72
	Active Stereo	25.94	124.17	25.50	126.28	27.18	117.04	32.81	121.24	16.40	101.86	15.73	95.27
Test 3	I-ToF	36.82	152.51	35.92	153.26	38.75	147.14	34.09	127.51	20.21	110.85	55.09	183.14
	D-ToF	32.99	145.50	35.64	153.07	25.90	120.35	19.92	96.01	21.59	105.41	37.26	149.66
	Active Stereo	31.63	141.77	35.24	151.37	22.44	110.42	23.47	106.63	14.49	94.51	21.21	109.53
T. Seen	I-ToF	9.87	77.99	4.62	57.10	33.91	133.46	6.18	60.48	35.65	119.76	91.30	224.27
	D-ToF	15.43	93.31	11.62	79.89	31.12	123.97	4.40	51.91	17.42	82.29	89.19	212.55
	Active Stereo	9.43	88.30	9.28	88.24	9.11	75.21	6.32	65.54	12.98	65.73	16.62	98.75
Tested on Modality:													
Test Seen	I-ToF	8.34	52.29	8.57	50.00	7.01	58.85	3.80	43.44	23.28	95.38	13.69	65.41
	D-ToF	8.05	50.43	6.82	45.50	13.52	66.34	9.00	54.15	30.91	87.71	27.92	87.32
	Active Stereo	39.25	101.76	40.87	102.29	30.32	90.00	32.24	90.49	23.36	72.21	37.25	101.23
	GT	1.12	28.81	0.71	24.41	2.65	40.41	1.83	34.89	2.16	29.55	5.02	52.43

1.1.2 Qualitative predictions

Figures 1, 2 and 3 show predictions on exemplary frames of the test scenes 1, 2 and 3, together with the different sensor modalities and the error plot of the prediction compared against the ground truth. The training with rendered ground truth generally performs best. Both ToF sensors show incorrect depth values for reflective or transparent objects which also translates to incorrect predictions in these areas (compare Fig. 1. The predictions when training with active stereo as supervision are more blurry and show less distinct edges at depth boundaries when compared to other modalities, which may arise from many depth pixels being invalidated by the sensors around such boundaries (compare Fig. 2). The very challenging test scene 3 with bright lighting and many unseen objects is difficult to predict for all training setups (compare Fig. 3. We can see similar artifacts as described above. Additionally, the unseen trophy object with partly reflective and partly transparent material shows large errors for the sensor inputs as well as for its predictions. The desk surface is also incorrectly captured by the D-ToF sensors due to large reflections and MPI from the background.

1.2. Implicit Reconstruction

Implementation Details As mentioned in the main paper, we follow NeRF [4] and build upon the work of [7] without a depth completion network, but leverage the respective sensor depth with a scale-invariant depth loss $\mathcal{L}_{\mathrm{D}}$. We use images with a resolution of 640×480 and process batches of 1024 rays. We set λ_{D} to 0.1 and the learning rate to 0.0005 and optimize for 100k iterations with Adam optimizer [3].

Table 2. Relative Pose Error of SLAM and SfM.

Error	Direct (DSO)	Dense dToF	Dense iToF	Dense AS	SfM
rot [deg]	0.22	0.18	0.51	0.56	10.76
trans [cm]	0.27	0.31	0.68	0.62	2.86

1.3. Camera Pose Estimation

The analysis above focuses on dense monocular depth estimation and novel view synthesis as recent and important approaches - for which pixelwise prediction and evaluation are crucial. We add results for direct SLAM (DSO) [1], KinectFusion [5] with different depth modalities, and COLMAP SfM [8] in Fig. 4.

Tab. 2 summarizes the relative pose error for different approaches (cf. Fig 4). Note the pose accuracy results for KinectFusion [5] align with the depth results from Tab.2 in the main paper.

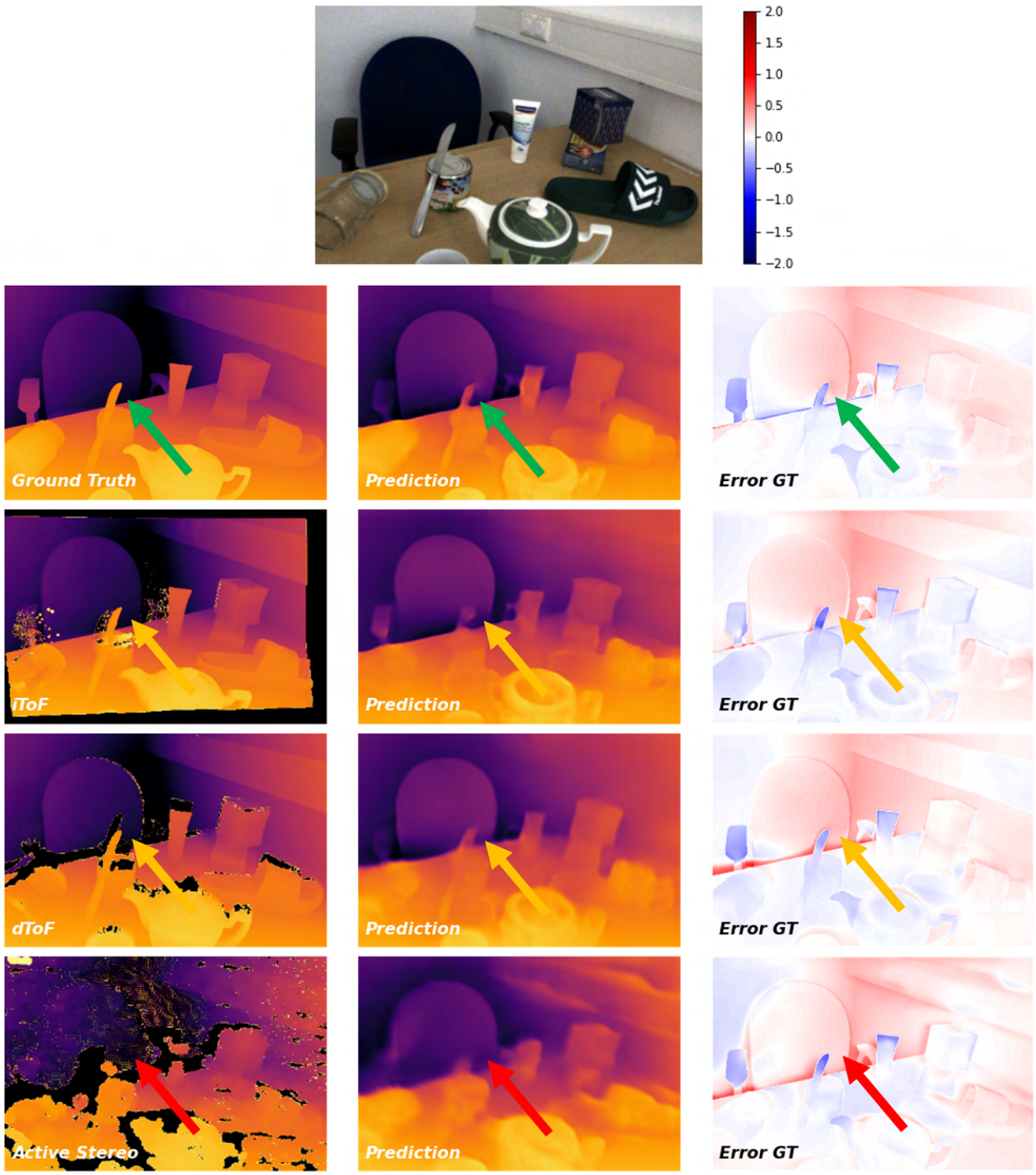

Figure 1. Qualitative evaluation on test scene 1. Each depth modality, the network prediction when trained with supervision of each modality, and the error, are shown as qualitative evaluation.

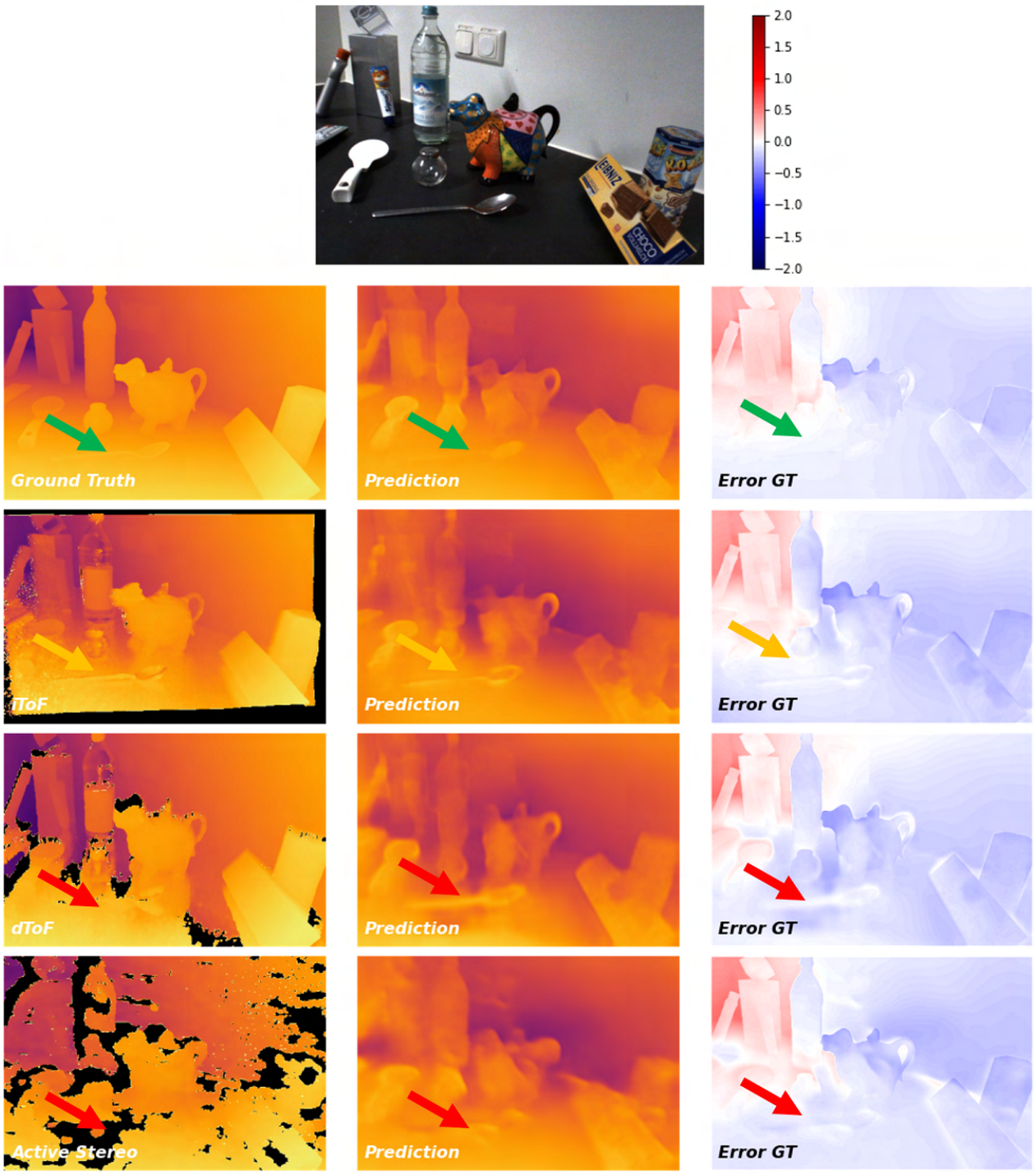

Figure 2. Qualitative evaluation on test scene 2. Each depth modality, the network prediction when trained with supervision of each modality, and the error, are shown as qualitative evaluation.

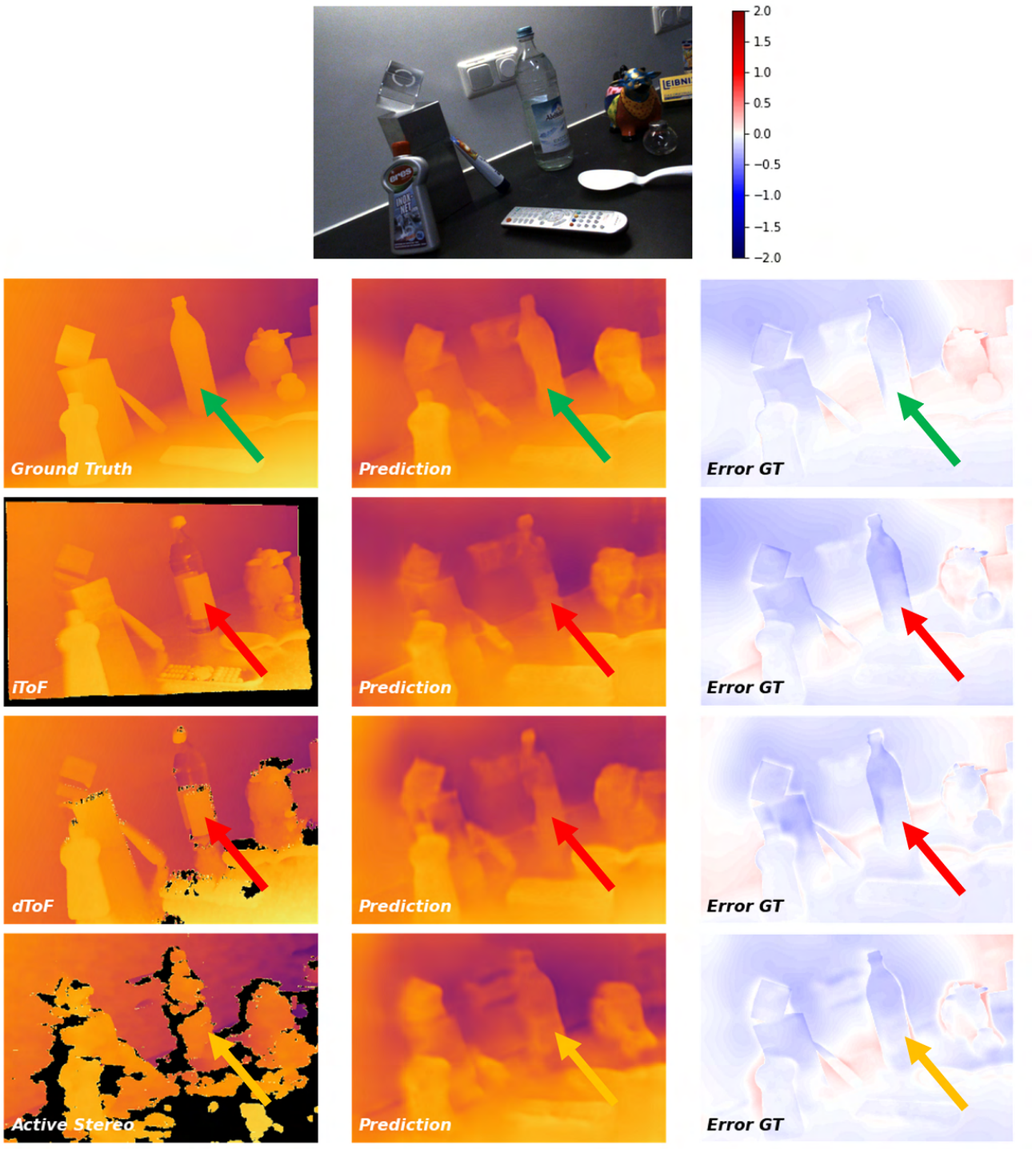

Figure 3. Qualitative evaluation on test scene 3. Each depth modality, the network prediction when trained with supervision of each modality, and the error, are shown as qualitative evaluation.

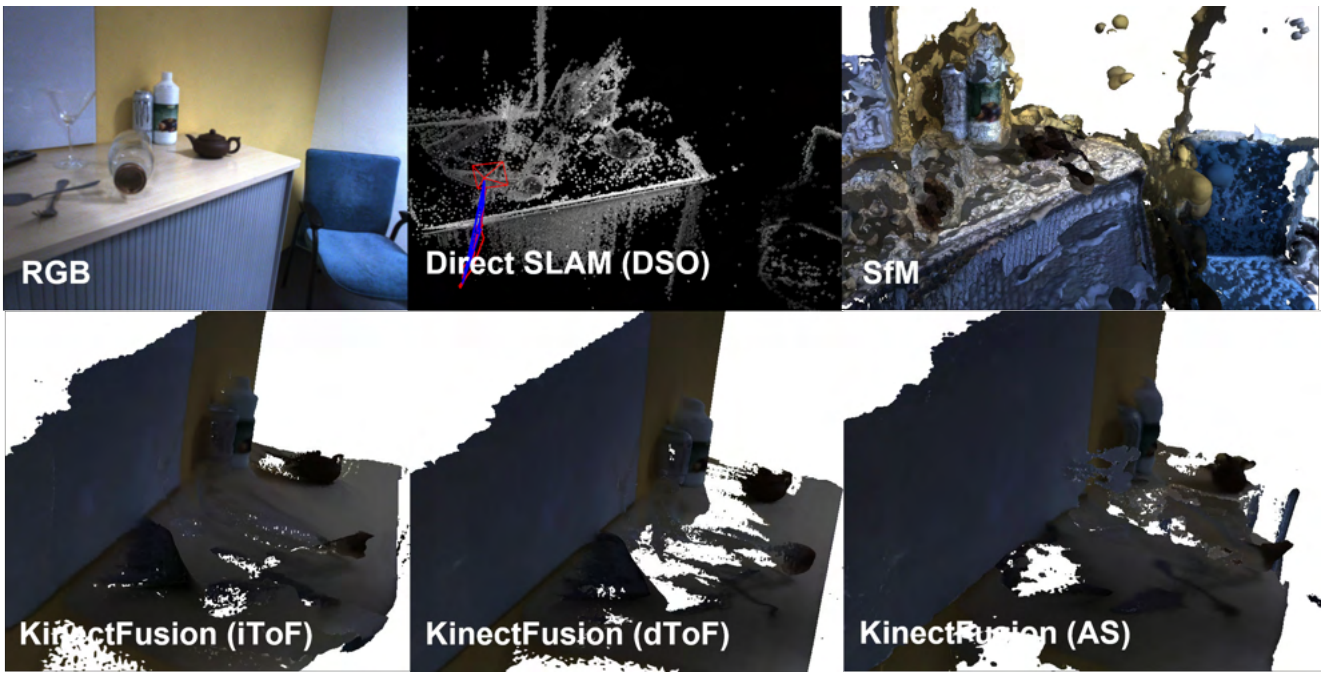

Figure 4. Qualitative reconstruction results from SLAM and SfM.

2. Dataset

2.1. Detailed Dataset Description

Sec. 3 of the main paper mentioned that our dataset uses multiple images/depth sensors to collect the dataset with highly accurate annotations of the scene using the robotic arm in a synchronized manner. This section shows the detailed description of data we include in our dataset.

2.1.1 Polarization Camera

Fig. 5 shows examples of images included for the polarization camera. As mentioned in Sec. 2 of the main paper, a polarization camera provides images with different polarization angles, which can extract cues like the surface normal by using the physical property of object material in the scene. The polarization camera we used in our dataset (See Sec. 3 in the main paper) provides polarized images at 4 different angles (0, 90, 180 270 degrees) which are saved in a single 2x2 image (Fig. 5, (a)). A regular RGB image is obtained by averaging the 4 images (Fig. 5, (b)). To showcase the results of the depth map trained with different depth cameras, we include warped depth images from each depth camera into the polarization camera coordinates using the extrinsic between the two cameras and its depth image (Fig. 5, (d-g)). These can be additionally used for RGBD-based depth completion research. On top of that, we include extra information, such as instance map (Fig. 5, (c)) to help train or validate pipelines for categorical level tasks, accurate 6d pose of the camera as the 4x4 matrix obtained from the robotic arm, extrinsic transformation between cameras as 4x4 matrices and camera intrinsics as 3x3 matrix.

2.1.2 D-ToF Camera

Fig. 6 shows an example of images included for the D-ToF camera. Direct ToF (D-ToF) camera senses the depth information of its surrounding by emitting an infrared signal and measuring the difference in time between the emitted and received signal. The quality of this modality highly depends on the reflection of the signal. It often suffers from specific physical noise such as Multi-Path-Interference (MPI) or strong material dependent artefacts (Fig. 9). For the D-ToF camera, we provide the depth map from the camera (Fig. 6, (a)) as well as its rendered ground truth depth map (Fig. 6, (b)) such that one can also research on D-ToF refinement pipelines to reduce such errors. As in the polarization camera, we include extra information such as instance label map (Fig. 6, (c)), camera pose, intrinsic and extrinsics of the camera as well.

2.1.3 I-ToF Camera

Fig. 7 shows image examples for the I-ToF camera. Indirect ToF (I-ToF) cameras sense the depth information of their surrounding by emitting a frequency modulated signal and measuring the return signal. Unlike Direct ToF (D-ToF), I-ToF cameras do not calculate the time difference to infer the depth. Instead, the camera correlates the returning signal with phase-shifted emitting signals to generate 4 different measurements, called correlation images. These are measured as sinus functions of distance $((\sin(d), \cos(d), -\sin(d), -\cos(d)) = (c_1, c_2, c_3, c_4)$ in Fig. 7, (a)). Either arc-tangent formula or convolutional neural networks can be used to extract depth information from the correlation images. As I-ToF modality also relies on the reflection of the signal like in D-ToF, it suffers from similar artefacts, such as MPI and material dependent artefacts (compare qualitative results of the test scenes in Figs. 1, 2 and 3). Here, we provide raw correlation images and depth map from the camera (see Fig. 7, (a,b)) as well as its rendered ground truth depth (Fig. 7, (c)) such that one can train I-ToF depth improvement pipelines either from raw signal or from I-ToF depth itself. As the other cameras, extras such as instance map (Fig. 7, (d)), camera pose, intrinsic and extrinsics are included.

2.1.4 Active Stereo Camera

Fig. 8 shows the examples of images included for the Active Stereo camera. Stereo depth estimation infers depth using and photometric consistency and geometrical constraints from epipolar geometry and triangulates the depth map from the disparity between left and right cameras. As the disparity is calculated via matching on the image itself, the stereo based depth estimation methods suffers less from the specific material, but they suffer from other aspects such as stereo occlusion and large texture-less regions. Active projection (Active Stereo) is used to overcome this issue. We provide both, active and passive stereo left / right images (Fig. 8, (a),(b)) and raw depth from the camera (active, Fig. 8, (c)) as well as the rendered ground truth (Fig. 8, (d)). This allows to use our dataset to improve stereo methods from either passive or active stereo and also depth refinement pipelines. Similar to the other cameras, extras such as instance map (Fig. 8, (e)), camera pose, intrinsic and extrinsics are included.

2.2. Error Analysis on Different Modality

In this section, we show specific errors on each depth modality to illustrate the implication of the depth quality when the given modality is used as the ground truth, as well as advantage of using our rendered depth as the ground truth.

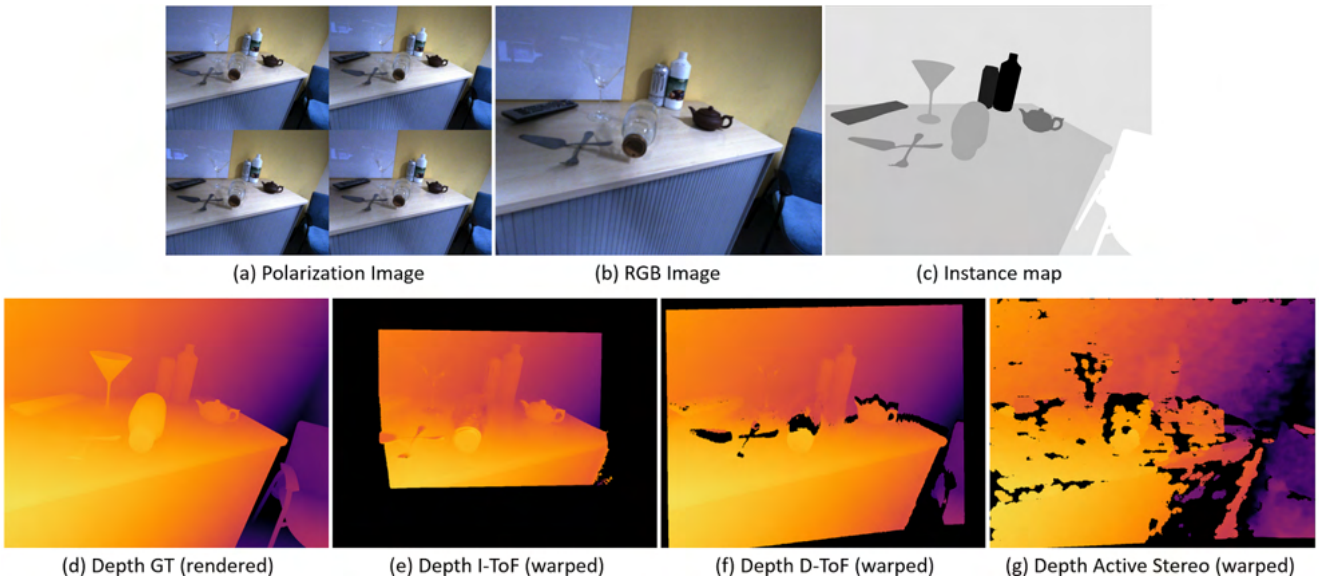

Figure 5. Example of the images included for the polarization camera input (top) together with instance label map and depth estimates warped onto the same coordinate reference frame.

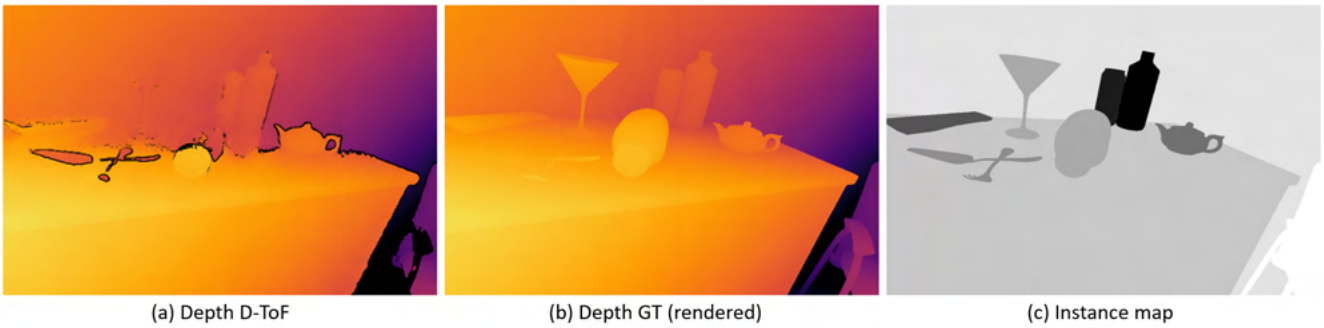

Figure 6. Example of the images included for the D-ToF camera: its depth map (left), ground truth depth (centre) and an object instance label map (right).

2.2.1 D-ToF Camera

As mentioned in Subsec. 2.1.2, D-ToF modality suffers from its own reflection-based nature, such as MPI and material dependent artefacts. When the angle of the surface normal of the scene is close to the incident angle of the infrared signal, the strength of the reflected signal becomes weak due to scattering effects (Fig. 9, (a) blue arrow) while multiple scattered signals from the other surfaces which has more traveling distance are received and with stronger strength (Fig. 9, (a) red arrow) and interfere with the original signal (MPI), producing a wrong measurement of the depth on the area with further distance which looks like a reflection or shadow of the object to the surface (Fig. 9, (b) red marker). This effect can be intensified when the surface material is reflective, which gives even stronger artefact as its reflective surface bounces even weaker and noisier signal with less attenuation (Fig. 9, (a,b) yellow arrow&marker). On the other hands, when the surface material is transparent, the emitted infrared signal rather goes through the object in the both ways (Fig. 9, (a) green arrow) which at the end ignores the object and the sensor produce the depth value as similar level as its background (Fig. 9, (b) green marker - material dependent artefact). Quality of the depth map degrades slightly around some boundaries after warping into the RGB frame (Fig. 10, (b), red), while the invalid regions actually helps to invalidate more area on wrong depth especially on the reflective objects (Fig. 10, (b), green) , which might become beneficial when it is used in the training.

2.2.2 I-ToF Camera

As mentioned in Subsec. 2.1.3, I-ToF modality suffers by its own reflection based nature as well similar to D-ToF,

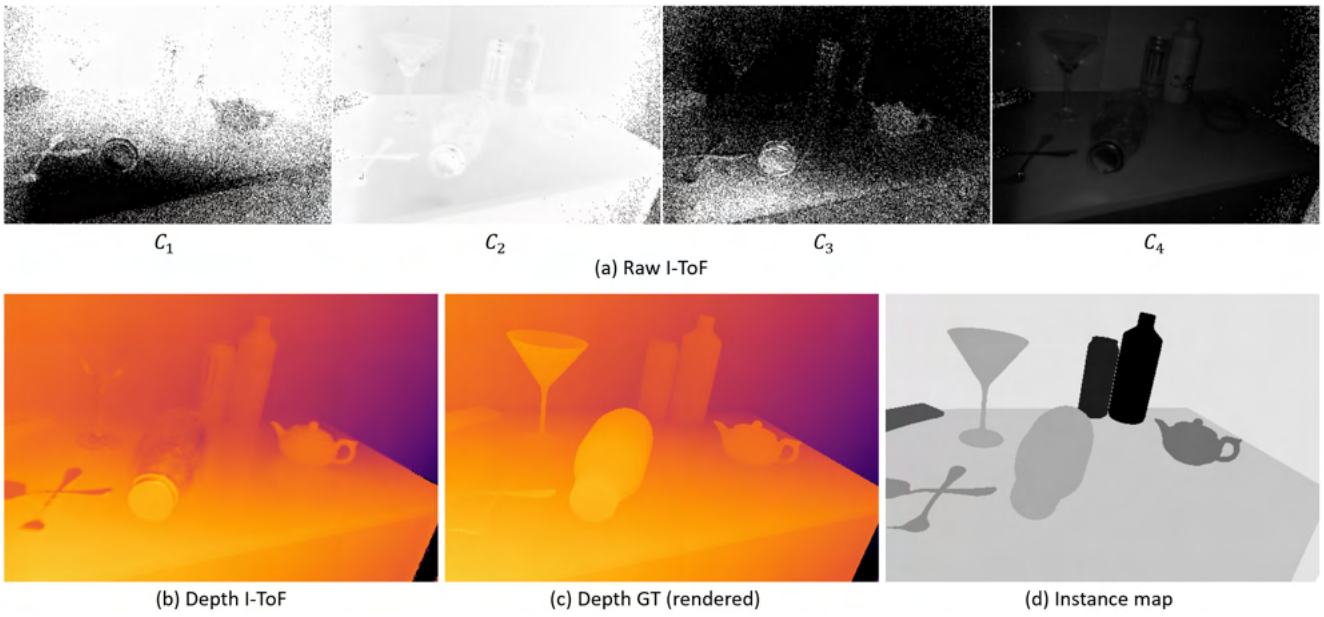

Figure 7. Example of the images included for the I-ToF camera.

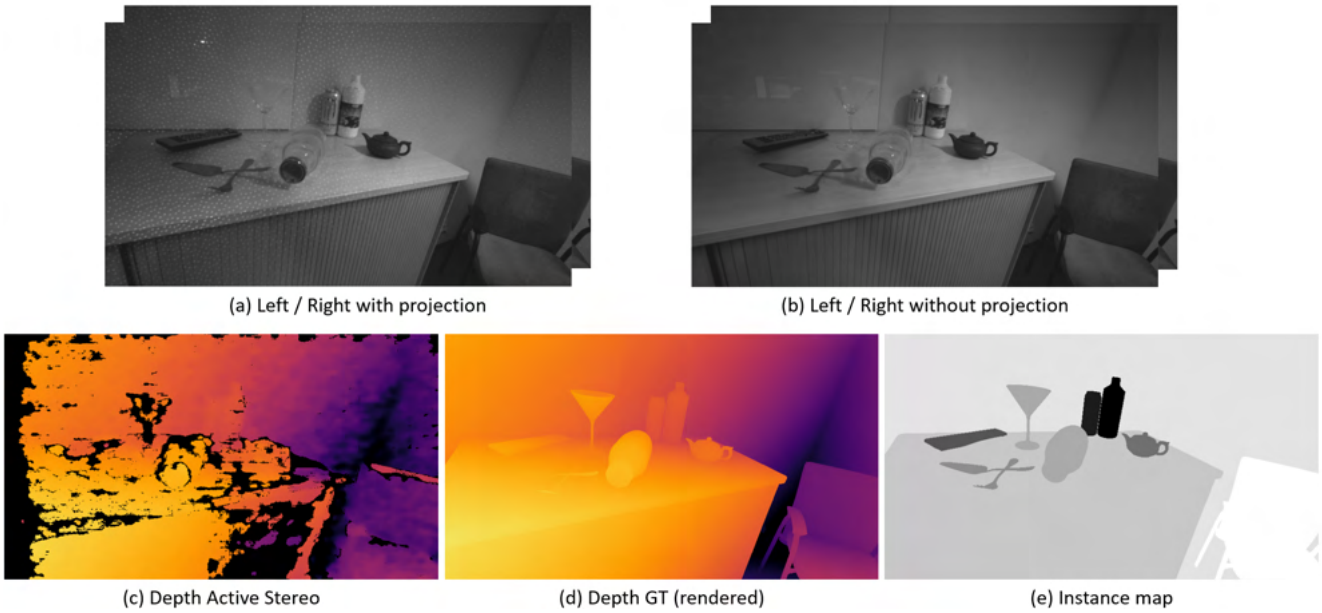

Figure 8. Example of the images included for the Active Stereo camera.

such as MPI and material dependent artefact (Fig. 11. Although the quality of depth itself seems better as the depth itself is more dense (with less invalid region) and amount of the artefacts are less, it is hard to say I-ToF modality is better than D-ToF as these two camera are in different price range and power level. Also less invalid area but rather with wrong depth didn't help invalidating depth (Fig. 12) not like in D-ToF case, which could result in artefact in the prediction when it is used as GT during the training.

2.2.3 Active Stereo Camera

As the stereo camera uses left and right matching with photoelectric cue, depth map suffers less on the challenging material as the projection can be visible on the surface as well as left-right check can be performed to invalidate region with the wrong depth. For this reason, depth on glass or the reflective object is significantly more accurate compared to either of ToF modality (Fig. 13, green arrow). On

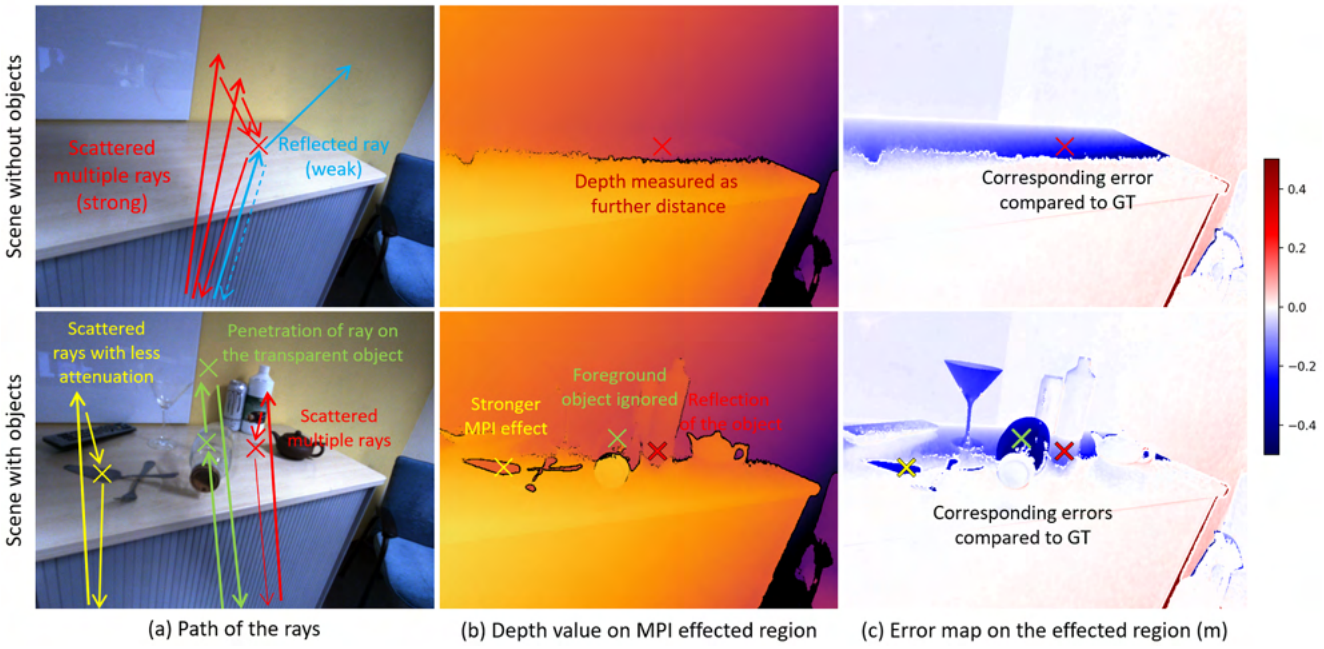

Figure 9. Detailed ray paths with MPI and surface material induced error on D-ToF modality. While D-ToF produces dense and sharp depth, its quality is highly dependent on the surface material and the incident angle.

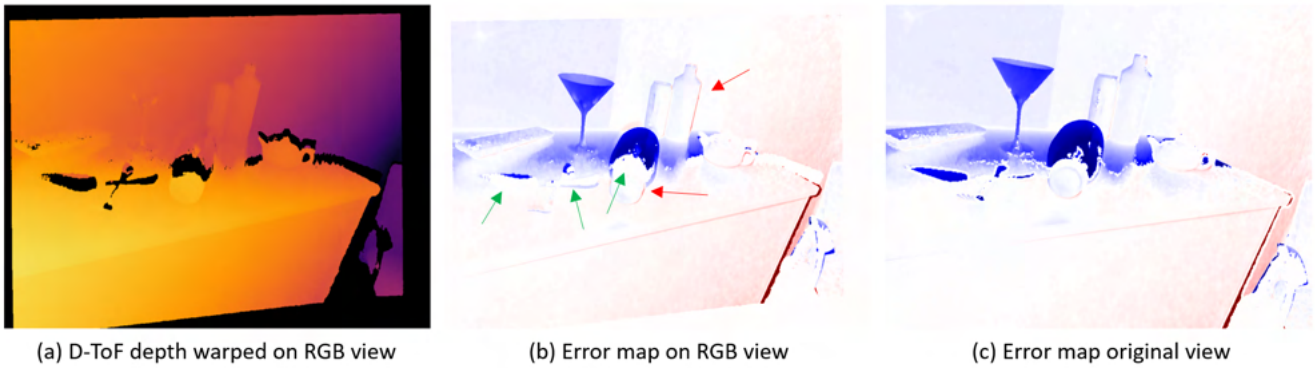

Figure 10. Error after warping D-ToF into RGB view. Slight errors are introduced on some edges (red) while expansion of the invalid area helps to invalidate on the reflective objects (green).

the other hands, due to its nature of pattern projection far distance that depth quality gets worsen as the scene gets further (Fig. 13, red arrow) the projection pattern gets attenuated and spread in the far distance. Moreover, the depth map in general is more blurry, jittery, sparse and has wrong values on some regions without being invalidated (Fig. 13, orange arrow) which can introduce negative influence when it is used as GT, such as blurriness and depth jittering. Error introduced by warping is trivial (Fig. 14) as the original depth map is already blurry and sparse.

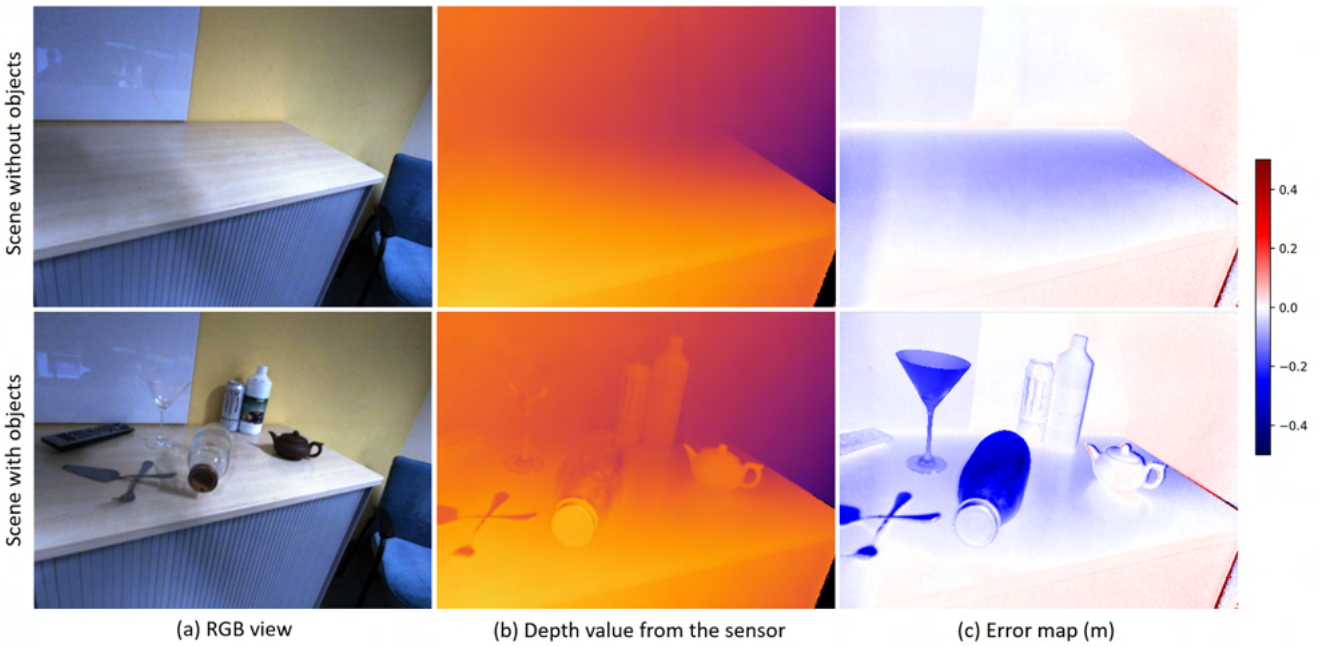

Figure 11. Depth quality from I-ToF camera. I-ToF modality suffers from same type of artefect as D-ToF. While depth map itself is more sense and suffers less from MPI artefact on the table.

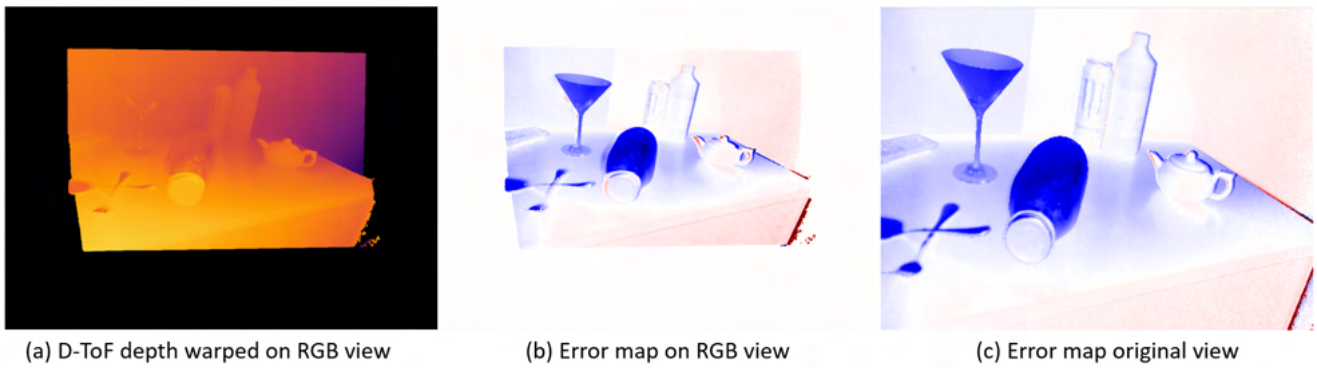

Figure 12. Error after warping I-ToF into RGB view. Not like D-ToF, most of depth error exists without being invalidated, which might introduce more error when it used as GT during the training.

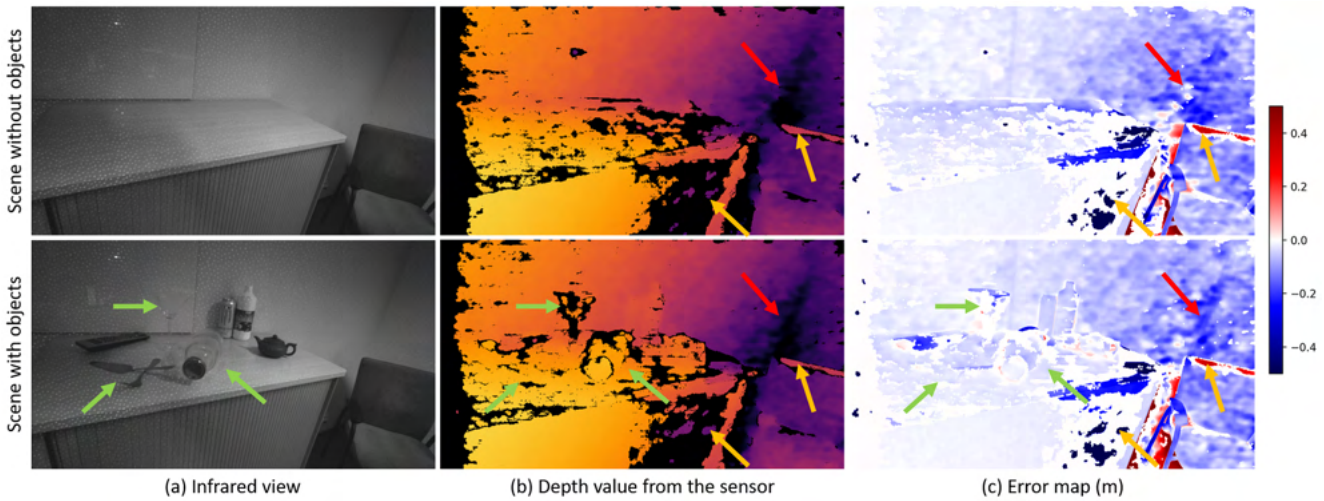

Figure 13. Depth quality from Active Stereo camera. While depth map suffers less on the challenging material, quality of depth itself is far behind either of ToF modality in multiple aspects, such as sharpness, variance, sparsity.

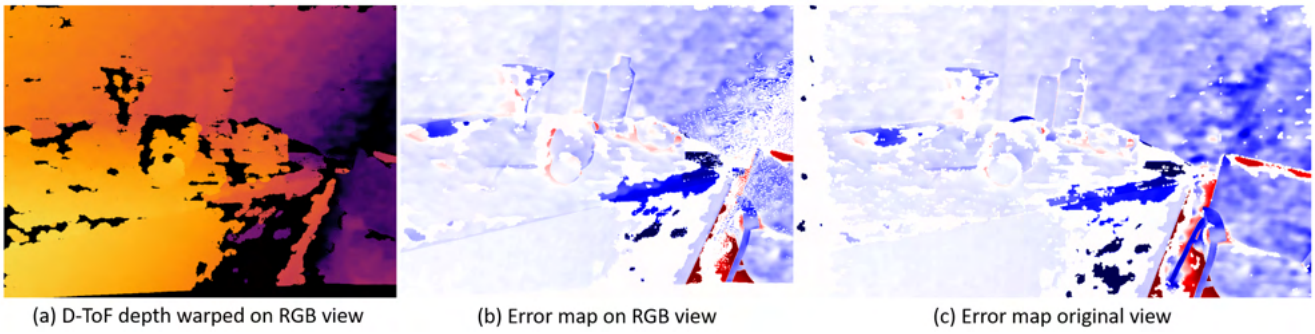

Figure 14. Error after warping Active Stereo into RGB view. Note that there isn't significant change in the depth quality after the warping.

2.3. Detailed Background and Objects Description

As described in Sec. 4 in the main paper, our dataset comprises a total of 13 scenes divided into 10 scenes for training and 3 testing scenes composed of a mixture of 4 different chairs, 6 different tables, 64 household objects from 8 plus 4 different categories (i.e. cup, teapot, bottle, remote, boxes, can, glass, cutlery and tube, shoe, plastic kitchenware, trophy) and and 7 different indoor areas. Test sets have 1 unseen background and 2 seen backgrounds with and without different lighting and contain a mixture of seen/unseen objects from seen/unseen categories. In this section, we show detailed images of backgrounds, chairs, tables, and other objects. Fig. 15 and 17 respectively show images of 3 chairs and 6 tables used in the dataset and their corresponding meshes. Fig. 18 and 19 show a collection of household objects used in training and test set. Fig. 16 shows 9 backgrounds used in the dataset and their corresponding meshes.

2.3.1 Detailed Scene Description

As described, our training set is composed of 10 scenes, and the test set is composed of 3 scenes. For each scene, we include 2 different trajectories. Each trajectory covers 2 setups with and without objects (naked scene). This sums up to 800-1200 frames per scene and a total of ca. 10k frames. In this section, we show several sample images of the scenes in Fig. 20, 21, and 22, 23. Each of them consists of an annotated mesh and RGB images with different types of rendering, which show the diversity and quality of our dataset.

2.3.2 Partial Scanning of the Scene and Mesh Fitting

As mentioned in Sec. 3 in the main paper, we use partial scanning and mesh fitting to annotate background, large objects, and objects outside the robotic workspace. This section shows images of partial scanning and the mesh fitting from one of the scenes as an example. The green box in Fig. 24, (a) shows annotated meshes of the objects by the robotic arm. Once the objects are annotated, the scene is partially scanned with multiple viewpoints to make the scanning dense and cover multiple facets of the background. Note that the center of the scanning is not yet in the robot base coordinates (Fig. 24, (a) blue box). Once the partial scanning is done, the scanned mesh is then fit onto the annotated objects, such that the partially scanned mesh origin concides with the robot base (Fig. 24, (b)). Once the scanned mesh is put to robot base coordinates, we fit background, large objects, and distant objects meshes also in robot base coordinates to annotate them (Fig. 25, (a)). Fig. 25, (b-c) shows the result of the annotated mesh. For the mesh fitting, we used Artec Studio 10 Professional (Artec 3D, Luxembourg) which runs a point correspondence and ICP-based method to fit the meshes.

Figure 15. Chairs used in the dataset. Chairs in group (a) are used for the training set and the chair in (b) is used for the test set.

(b) Backgrounds used in the test set

(a) Backgrounds used in the training set

Scanned Mesh | Real Image | Scanned Mesh | Real Image | Scanned Mesh | Real Image

Figure 17. Tables used in the dataset. Tables in group (a) are used for the training set and the table in (b) is used for the test set. Note that, unlike small objects or chairs, we decide not to scan some parts of the large tables (e.g. end of their legs) as the cameras cannot see the part in their trajectories.

Figure 18. Collection of small household objects used in the training set. Objects from 8 household categories are used in the training set, 3 of which have photometrically challenging surface material - partially reflective (can), transparent (glass/plastic), reflective (cutlery).

Figure 19. Collection of small household objects used in the test set. The test set comprises a mixture of seen (left column) and unseen (mid column) objects from 8 seen categories and a few objects from unseen categories (right column - tube, slipper, plastic kitchenware, trophy) are used.

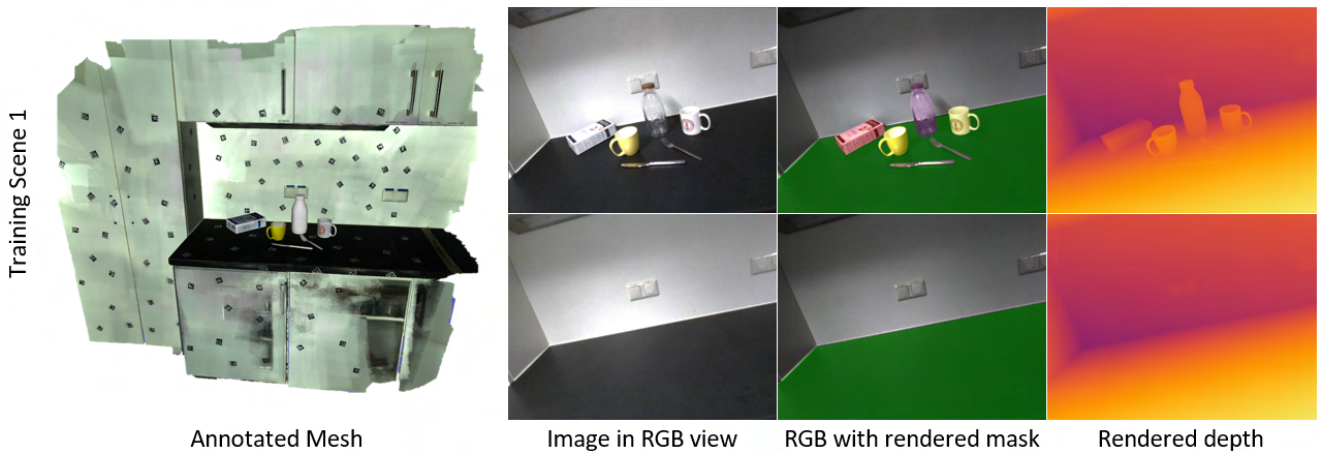

Figure 20. Example images from Training Scene 1. The annotated mesh is shown on the left together with an RGB view from the scene (second from left) with and without objects. The overlayed masks (second from right) and the rendered depth (right) illustrate the annotation quality of our data.

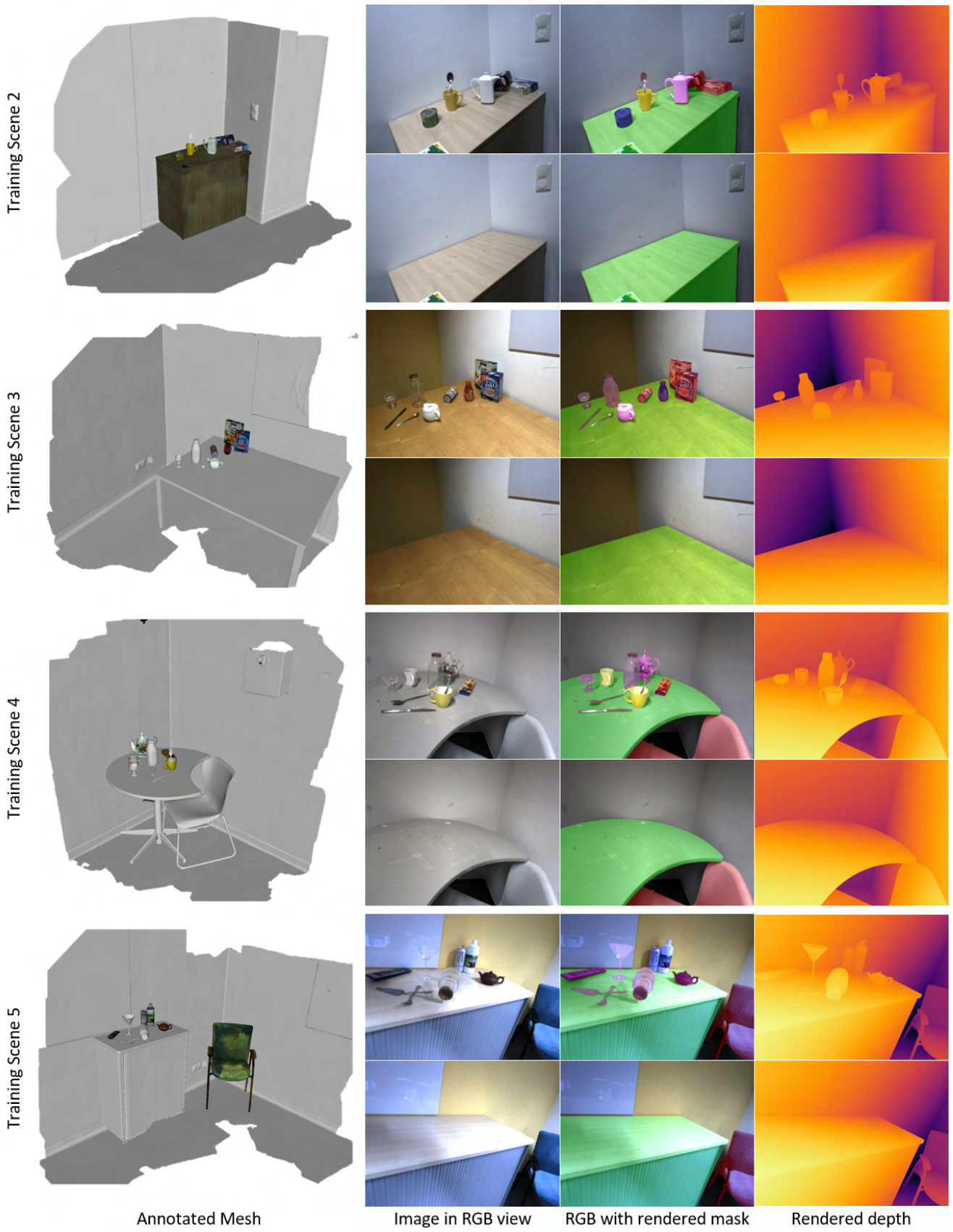

Figure 21. Example images from Training Scene 2-5. The annotated mesh for 4 different scenes is shown on the left together with an RGB view from the scene (second from left) with and without objects. The overlayed masks (second from right) and the rendered depth (right) illustrate the annotation quality of our data.

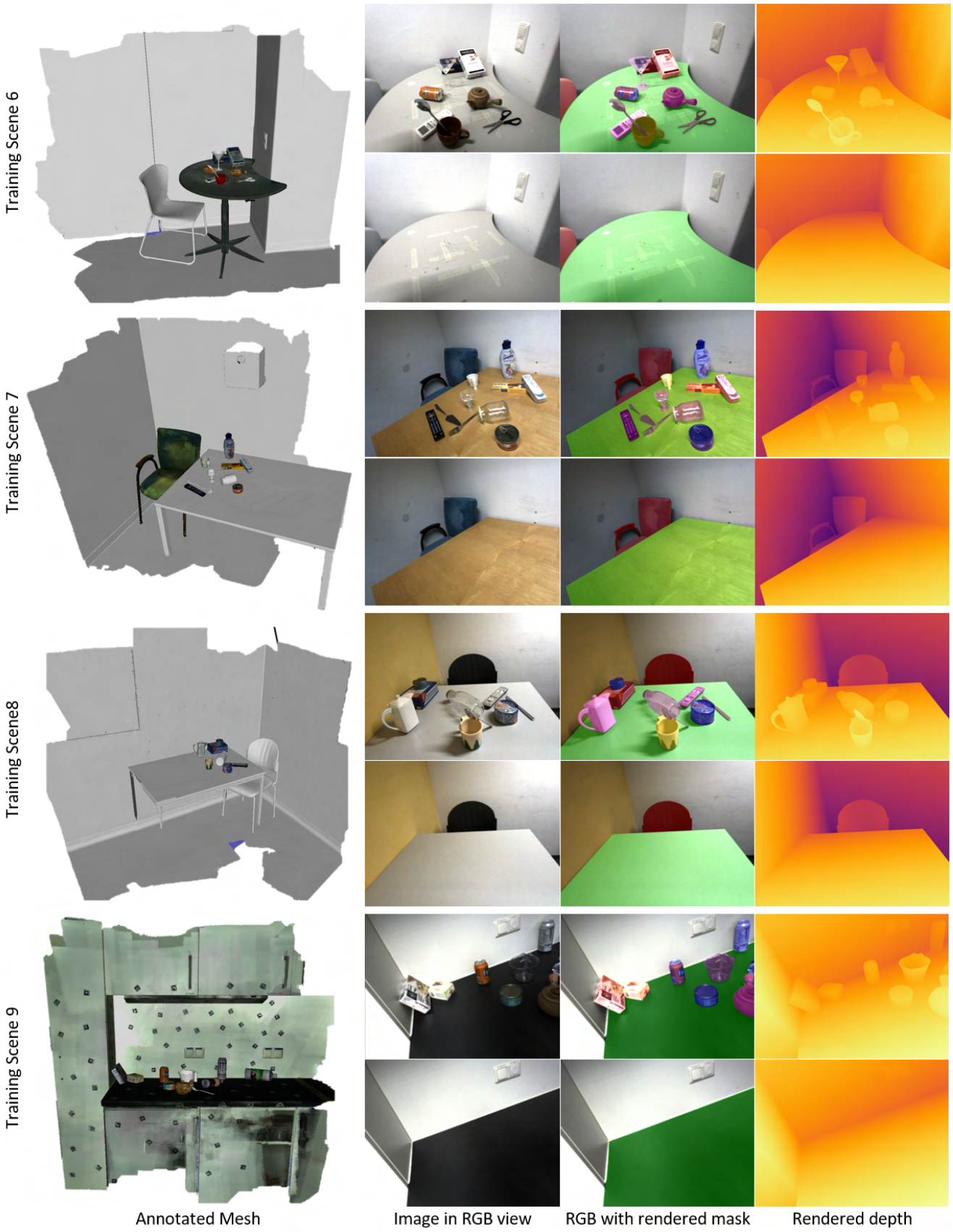

Figure 22. Example images from Training Scene 6-9. The annotated mesh for four different scenes is shown on the left together with an RGB view from the scene (second from left) with and without objects. The overlayed masks (second from right) and the rendered depth (right) illustrate the annotation quality of our data.

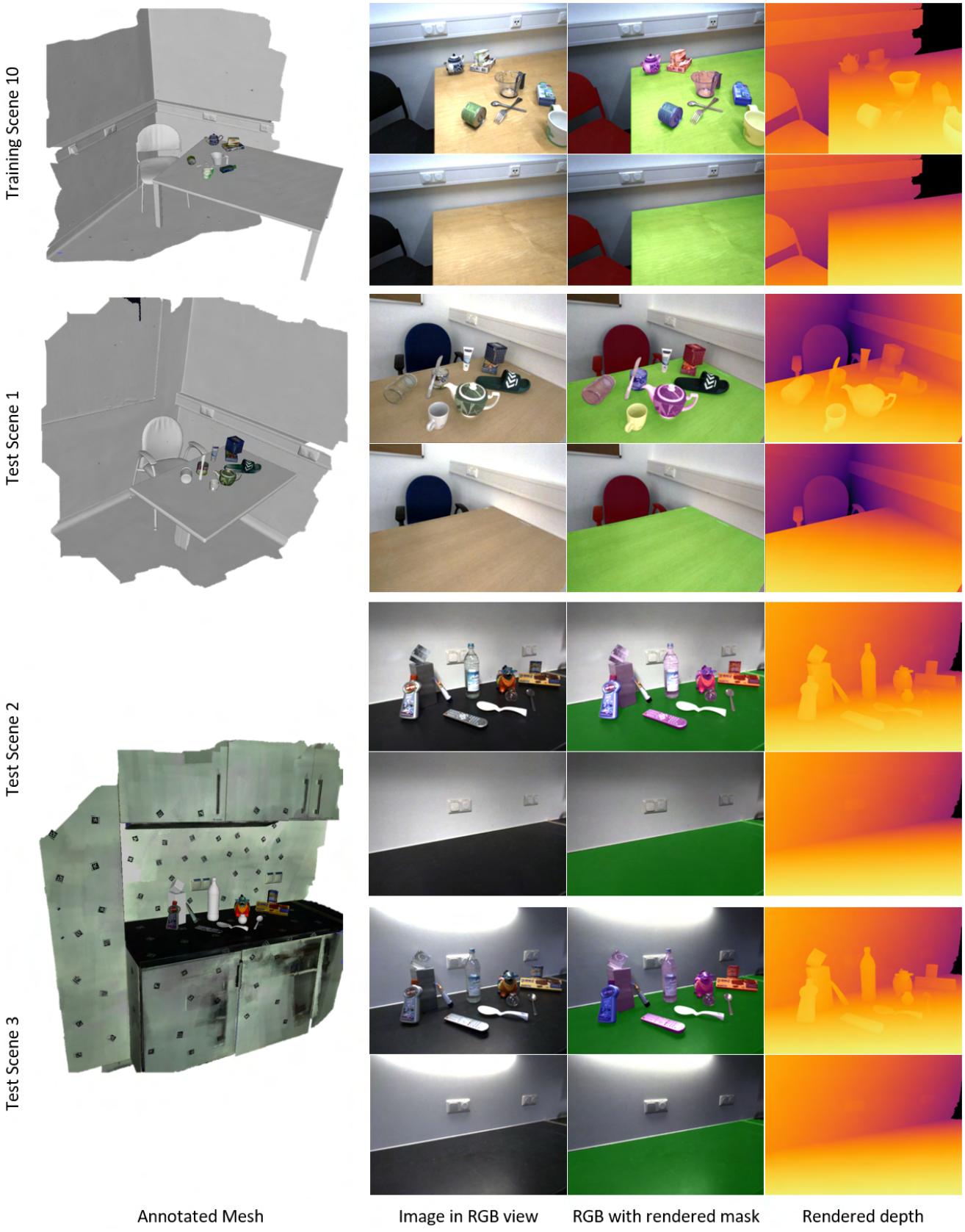

Figure 23. Example images from Training Scene 10 and Test scene 1-3. The annotated mesh is shown on the left together with an RGB view from the scene (second from left) with and without objects. The overlayed masks (second from right) and the rendered depth (right) illustrate the annotation quality of our data. Note that the test scene 2,3 are recorded in the exactly same pose and trajectory but with the different lighting.

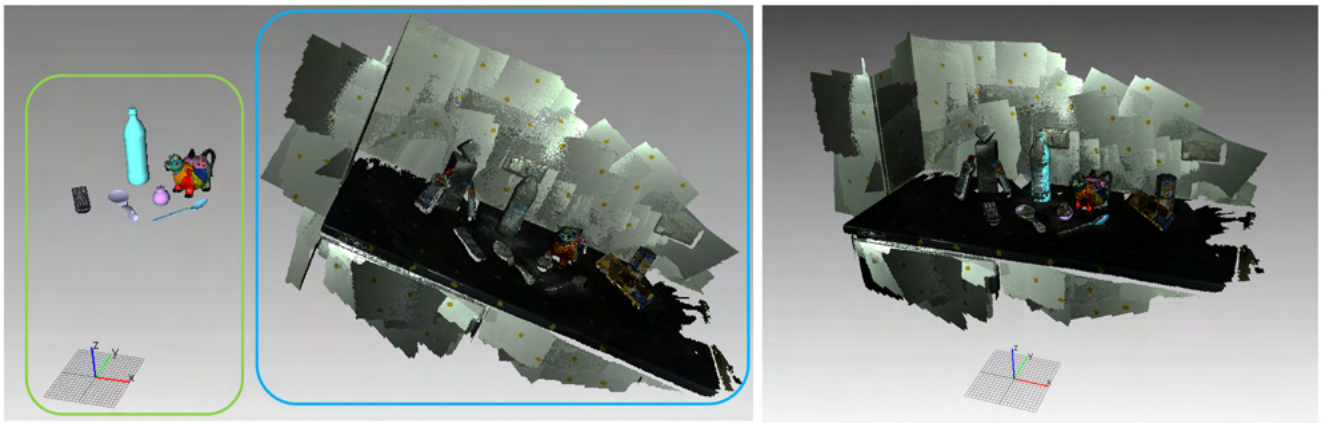

Figure 24. Example of partial scanning of the scene before and after the fitting on scene 13. Note that the center of the partial scanned mesh is aligned to robot base (xyz coordinate marker) after fitting it onto the mesh of the annotated objects.

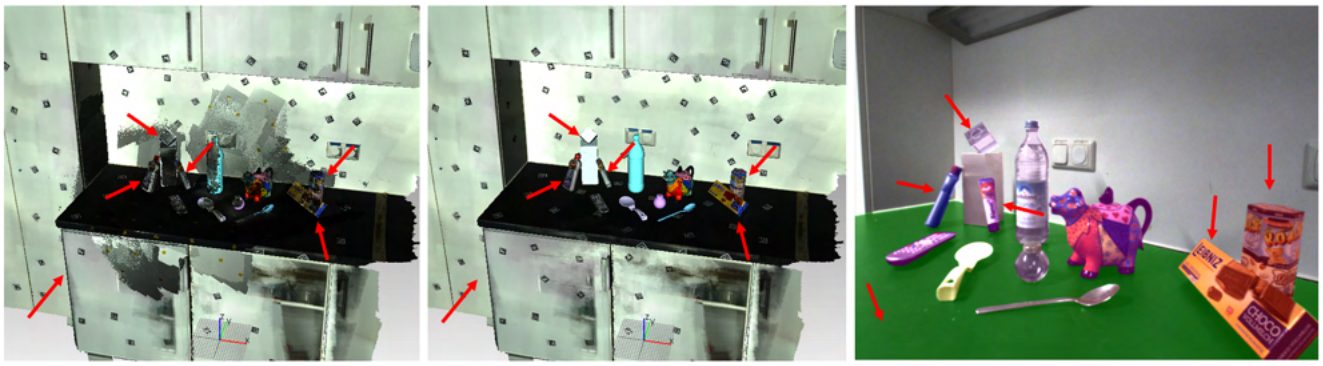

Figure 25. Example of far objects and background fitting onto partially scanned mesh. Left: Background and objects are fit to partial scans. Centre: All annotated meshes are shown without partial scans. Right: Corresponding scene from the camera viewpoint with augmented object masks. Note that the annotation quality of meshes with partial scans and robot arm is similar. The annotated meshes via partial scanning are marked with red arrows.

References

[1] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 2

[2] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. In *The International Conference on Computer Vision (ICCV)*, 2019. 1

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1, 2

[4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2

[5] Richard A. Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J. Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, pages 127–136, 2011. 2

[6] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS-W*, 2017. 1

[7] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 2

[8] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2