Supplementary Material for Domain Randomization-Enhanced Depth Simulation and Restoration for Perceiving and Grasping Specular and Transparent Objects

In the supplementary material, we present the additional sections for this paper, including domain randomization details, network implementation details, additional experiments and results, and additional dataset details.

1 Domain Randomization Details

In this work, we propose the Domain Randomization-Enhanced Depth Simulation (DREDS) approach, leveraging domain randomization and depth sensor simulation to generate photorealistic RGB images and simulated depths with realistic sensor noises. Specifically, during the simulated data generation, we perform domain randomization in the following aspects:

Scene and Object Setting. We focus on hand-scale objects and a table-top setting. We set the scene into the following two types: 1) Category-aware scenes that mainly utilize ShapeNetCore [4] objects from 7 object categories – camera, car, airplane, bowl, bottle, can, and mug. We also have some distractor objects from categories of phone, guitar, cap, *etc.* In total, we leverage 1536 objects for training and 265 objects for evaluation. In our simulated scenes, we load a random number of objects ranging from 6 to 10 with random scales and categories and let them fall freely under gravity onto a ground plane to create random but physically plausible spatial arrangements of objects and prepare cluttered scenes. 2) Category-agnostic scenes. To evaluate the generalization ability to category-novel objects and the performance of grasping, we adopt 60 objects from GraspNet-1Billion [6]. We follow their original poses and arrangements but transfer random types of material as described in the next section.

Material Modeling and Assignment. Few of the existing depth sensor simulators consider modeling a variety of randomized real-world materials, especially specular and transparent materials. In this work, we adopt a bidirectional scattering distribution function (BSDF) [1], a unified representation covering the most common materials. BSDF defines how the light is scattered on a surface to determine the material of each point on the object.

Specifically, we use Disney principled BSDF [2,3] $f_{D\&S}(\phi)$ for diffuse and specular material modeling, where ϕ is the set of scalar parameters or nested functions, including the base color, subsurface, metallic, specular, roughness, anisotropic, *etc.* We use a mix of BSDF $f_T(\psi)$ to represent transparent materials, containing glass BSDF, transparent BSDF, and translucent BSDF to adjust transparency, as well as refraction BSDF to add refraction, and glossy BSDF to add reflection on the surface, *etc*, where ψ means the parameter set from each BSDF function like surface color, index of refraction (IOR), and roughness.

Based on the above BSDF models, we collect an asset of materials with different categories that cover common objects in life, including 1) 27 specular materials including metal, porcelain, clean plastic, paint, *etc.*, 2) 4 transparent materials, 3) 36 diffuse materials including rubber, leather, wood, fabric, coarse plastic, paper, clay, *etc.* We randomize the parameters of the BSDF function for each material within a range, generating a large-scale material collection with wide variations.

We assign one type of material to each object in the scene randomly. For those objects with default colors or texture maps, we mix their colors or textures with the base color of the assigned material in a randomized ratio. It means that we can easily transfer an existing synthetic object dataset to a dataset with a large amount of specular and transparent objects.

Camera Setting. We follow RealSense D415 to set up the projector's parameters (*e.g.*, the IR pattern image, baseline distance) and other cameras' intrinsic parameters. Camera locations and poses are randomized within a range, so that the objects in each scene can be captured from arbitrary directions.

Lighting and Background Setting. We collect 74 HDRI environment maps for training, and 23 for testing, including indoor and outdoor scenes, as well as natural and artificial lighting. An arbitrarily chosen environment map with random intensities is used to simulate realistic ambient illumination. For the background, we pick 81 common indoor materials for training and 23 for evaluation, including wood, marble, tiles, concrete, *etc.* A random selection of these materials is applied to the ground plane to increase variations of the scene.

2 Network Implementation Details

We implement the proposed SwinDRNet and downstream algorithms in PyTorch. We train SwinDRNet for 20 epochs (nearly 146,000 iterations) with batch size 32, using AdamW [7] optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a learning rate of 1e-4, a weight decay of 0.01, as well as a learning scheduler with a linear warmup of 500 iterations and a linear learning rate decay. SwinDRNet takes RGB and raw depth images that are resized to 224*224 as the input, and outputs the restored depth image with the same size for downstream tasks. Note that for SGPA [5], the baseline method of category-level pose estimation, as its performance depends on the number of points of the input point cloud, i.e., the resolution of the depth, the original RGBD images are firstly resized to 224*448, and then sampled at an interval of 1 along the direction of the row to obtain two 224*224 inputs, as well as two 224*224 outputs from the network. We finally interpolate these two outputs in the same sampling way above, to obtain the 224*448 depth as the input to SGPA.

3 Additional Experiments and Results

3.1 Depth Restoration

Qualitative Comparison to State-of-the-art Methods. Figure 1 shows the qualitative comparison of STD dataset, demonstrating that our method can predict a more accurate depth on the area with missing or incorrect values while preserving the depth value of the correct area of the raw depth map.

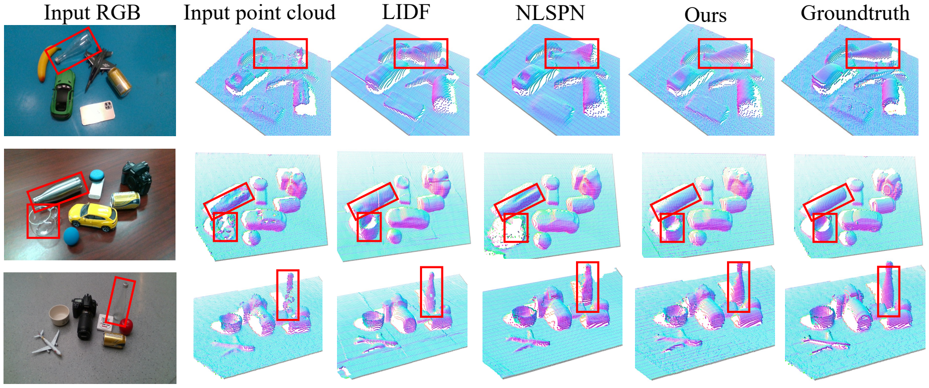

Fig. 1. Qualitative comparison to state-of-the-art methods. For a fair comparison, all the methods are trained on the train split of DREDS-CatKnown. Red boxes highlight the specular or transparent objects.

Cross-Sensor Evaluation. In this work, depth sensor simulation and real-world data capture are both based on Intel RealSense D415. To investigate the robustness of the proposed SwinDRNet on other types of depth sensors, we evaluate the performance on data of two scenes from STD-CatKnown dataset, captured by Intel RealSense D435. Table 1 shows a comparison of the results evaluated on D415 and D435 data after training on DREDS-CatKnown dataset. We observe that SwinDRNet has similar performance on data from these two different depth sensors in each scene, which verifies the good cross-sensor generalization ability of SwinDRNet.

Table 1. Quantitative results for cross-sensor evaluation. The performance of SwinDRNet is evaluated on RGB-D data captured by Intel RealSense D415 and D435 in each of the two scenes.

Scenes	Sensors	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
1	D415	0.017/0.017	0.015/0.016	0.009/0.010	94.62/94.30	98.34/98.60	99.94/99.95
	D435	0.021/0.023	0.022/0.025	0.013/0.015	89.30/86.23	97.95/97.85	99.95/99.98
2	D415	0.013/0.018	0.011/0.014	0.008/0.011	97.93/96.02	99.47/98.94	100.00/100.00
	D435	0.016/0.024	0.015/0.024	0.010/0.017	95.25/89.29	99.16/97.69	100.00/100.00

3.2 Category-level Pose Estimation

Qualitative Comparison to Baseline Methods. Figure 2 shows the qualitative results of different experiments on DREDS and STD datasets. We can see that the qualities of our predictions are generally better than others. The figure also shows that NOCS [9], SGPA [5] and our method all perform better with the help of restoration depth, especially for specular and transparent objects like the mug, bottle and bowl, which indicates that depth restoration does help category-level pose estimation task.

Quantitative Comparison to Restored Depth Inputs. We further evaluate the influence of different restored depths for category-level pose estimation, which is presented in Table 2. The proposed SwinDRNet+NOCSHead network receives the restored depth from SwinDRNet and the competing depth restoration methods for pose fitting. Quantitative results under all metrics demonstrate the superiority of SwinDRNet over other baseline methods in boosting the performance of category-level pose estimation.

Table 2. Quantitative results for category-level pose estimation using different restored depths from SwinDRNet and the competing baseline methods. The left of '/' shows the results evaluated on all objects, and the right of '/' shows the results evaluated on specular and transparent objects.

Methods	IoU25	IoU50	IoU75	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm
	DREDS-CatKnown (Sim)							
NLSPN	**94.7**/98.1	84.6/90.3	65.9/71.2	39.4/39.4	40.3/40.4	65.2/67.8	67.6/70.4	67.6/70.4
LIDF	94.4/97.9	83.3/89.5	59.3/66.4	33.7/37.4	36.3/39.8	57.9/63.7	64.3/69.8	64.6/70.0
Ours	**94.7/98.2**	**84.8/90.8**	**68.0/74.0**	**49.1/51.5**	**50.1/52.9**	**69.8/73.9**	**72.4/77.0**	**72.5/77.1**
	STD-CatKnown (Real)							
NLSPN	92.3/99.5	87.7/94.8	73.5/73.5	45.2/31.5	46.2/33.3	72.5/57.1	75.1/60.9	75.1/60.9
LIDF	92.3/99.1	87.2/93.4	67.0/68.5	34.6/35.4	37.1/40.2	64.7/60.8	70.4/**69.0**	70.5/**69.2**
Ours	**92.4/99.7**	**88.0/95.0**	**75.9/78.8**	**52.9/40.0**	**53.8/41.3**	**77.1/66.3**	**79.1**/68.7	**79.1**/68.7

3.3 Robotic Grasping

The illustration of a real robot experiment for specular and transparent object grasping is shown in Figure 3. We carry out the table-clearing using the Franka Emika Panda robot arm with the parallel-jaw gripper, and RealSense D415 depth sensor for RGBD images capture.

3.4 Ablation Study

To analyze the components of the proposed SwinDRNet, as well as domain randomization and the scale of the proposed DREDS dataset, we conduct the ablation studies with different configurations.

Analysis of the Modules of SwinDRNet. We first evaluate the effect of different modules of SwinDRNet with three configurations: 1) Take the concatenated RGBD images as input without the RGB-D fusion and confidence interpolation module. 2) Have no confidence module compared with SwinDRNet. 3) The complete SwinDRNet. As shown in Table 3, the performance of depth restoration improves when using these two modules. Note that the network with and without the confidence interpolation module obtain the similar depth restoration performance. However, in Table 4, we observe that SwinDRNet with this module achieves higher performance on object pose estimation, because the module keeps the correct geometric features from the original depth input which benefits the downstream task. The results above indicate the effectiveness of the RGB-D fusion and confidence interpolation module of SwinDRNet.

Analysis of Material Randomization. We analyze the effect of material randomization on depth restoration. We create a dataset of the same size as the fully randomized DREDS-CatKnown dataset. The original materials from ShapeNetCore [4] are directly applied to the objects without any transfer or randomization of specular, transparent, diffuse materials. Table 5 shows the results

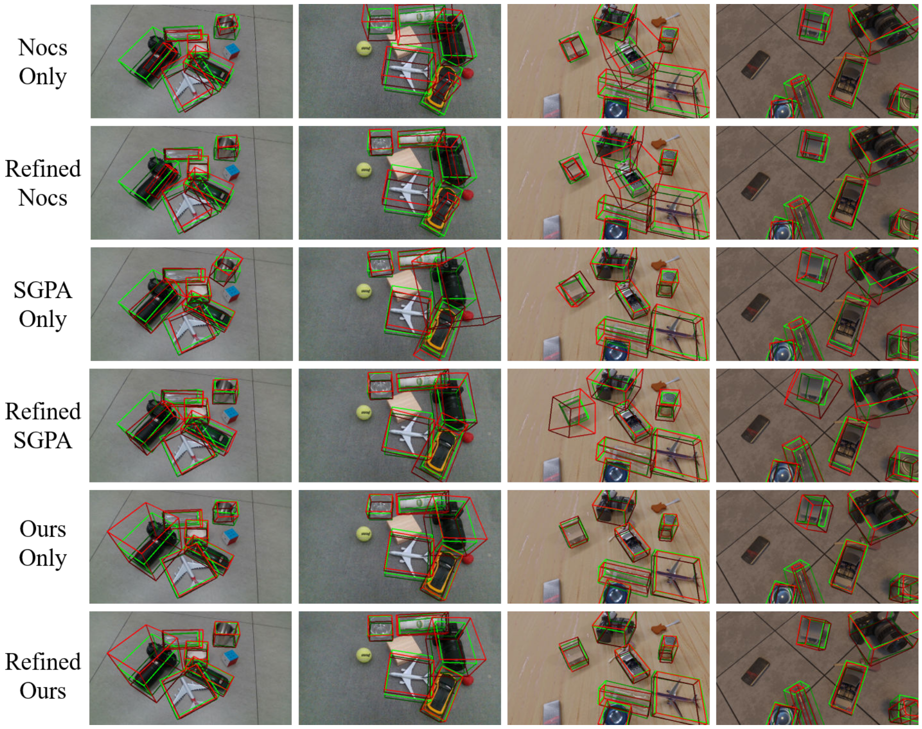

Fig. 2. Qualitative results of pose estimations on DREDS and STD datasets. The ground truths are shown in green while the estimations are shown in red. *only* means using raw depth in the whole experiment, *Refined* means using restored depth for training and inference in SGPA and for pose fitting in NOCS and our method.

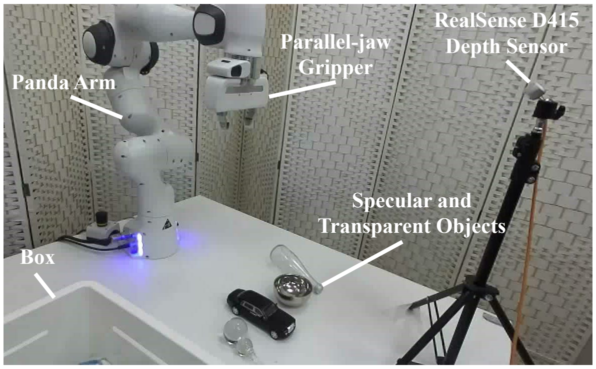

Fig. 3. The setting of real robot experiment for specular and transparent object grasping.

of depth restoration, evaluating on specular and transparent objects. Without material randomization, the performance drops significantly, since the network cannot consider real-world data as the variation of the synthetic training data without seeing sufficient material variation, which demonstrates the significance of material randomization.

Analysis of the Scale of Training Data. In Table 6, we show the performance dependence on the dataset scale. Compared to the full scale, the depth restoration performance of SwinDRNet trained on the half scale also degraded, demonstrating the necessity of the scale of the DREDS dataset for the method.

Table 3. Ablation studies for the effect of different modules on depth restoration. ✓denotes prediction with the module.

Fusion	Confidence	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
		STD-CatKnown					
		0.019/0.027	0.019/0.032	0.0123/0.021	91.09/79.20	98.92/97.73	**99.95/99.91**
✓		**0.014/0.017**	**0.013**/0.017	0.009/0.012	96.33/94.18	**99.36/99.01**	99.92/**99.91**
✓	✓	0.015/0.018	**0.013/0.016**	**0.008/0.011**	**96.66/94.97**	99.03/98.79	99.92/99.85

Table 4. The effect of confidence for category-level pose estimation.

Confidence	IoU25	IoU50	IoU75	5°2cm	5°5cm	10°2cm	10°5cm	10°10cm
	STD-CatKnownl							
	92.4	**88.0**	75.6	51.0	51.9	76.0	78.2	78.3
✓	**92.4**	**88.0**	**75.9**	**52.9**	**53.8**	**77.1**	**79.1**	**79.1**

Table 5. Quantitative results for material randomization on depth restoration task. The left of '/' shows the results evaluated on all objects, and the right of '/' evaluated on specular and transparent objects. Note that only one result is reported on STD-CatNovel, because all the objects are specular or transparent.

Model	RMSE↓	REL↓	MAE↓	$\delta_{1.05}$ ↑	$\delta_{1.10}$ ↑	$\delta_{1.25}$ ↑
	STD-CatKnow (Real)					
Fixed material	0.024/0.038	0.024/0.045	0.015/0.029	86.20/65.63	96.12/90.94	99.87/99.72
Full randomization	**0.015/0.018**	**0.013/0.016**	**0.008/0.011**	**96.66/94.97**	**99.03/98.79**	**99.92/99.85**
	STD-CatNovel (Real)					
Fixed material	0.038	0.051	0.027	67.52	84.86	98.51
Full randomization	**0.025**	**0.033**	**0.017**	81.55	**93.10**	**99.84**

4 Additional Dataset Details

4.1 DREDS Dataset

We present the DREDS-CatKnown dataset, where the category-level objects are from ShapeNetCore [4], and the DREDS-CatNovel dataset, where we transfer random materials to the objects of GraspNet-1Billion [6]. Figure 4 shows the examples and annotations of DREDS dataset. For each virtual scene, we provide the RGB image, stereo IR images, simulated depth, ground truth depth, NOCS map, surface normal, instance mask, *etc.*

Table 6. Ablation study on the scale of training data. SwinDRNet is trained on DREDS-CatKnown and evaluated on the specular and transparent objects of STD.

Scale	RMSE↓	REL↓	MAE↓	$\delta_{1.05}\uparrow$	$\delta_{1.10}\uparrow$	$\delta_{1.25}\uparrow$
	STD-CatKnow (Real)					
Half	0.021	0.020	0.014	92.71	98.54	99.83
Full	**0.018**	**0.016**	**0.011**	**94.97**	**98.79**	**99.84**
	STD-CatNovel (Real)					
Half	0.028	0.037	0.020	80.37	91.16	99.79
Full	**0.025**	**0.033**	**0.017**	**81.55**	**93.10**	**99.84**

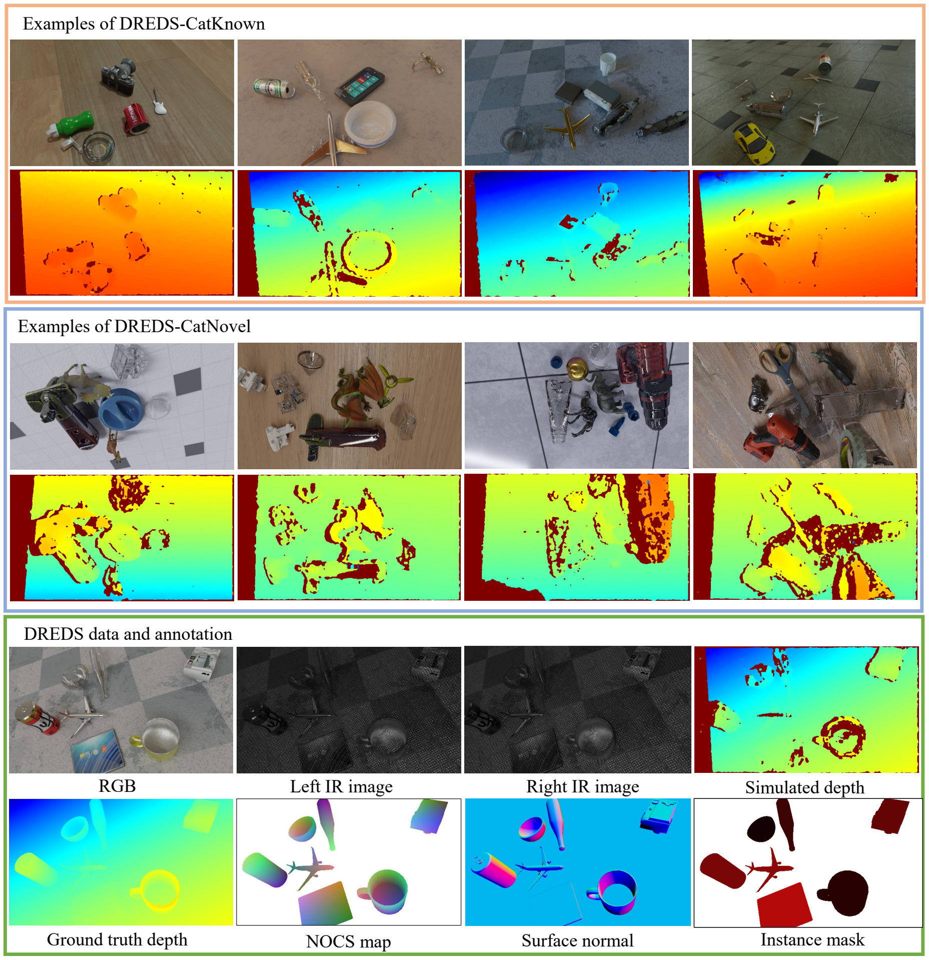

Fig. 4. Paired RGB and simulated depth examples and annotations of DREDS-CatKnown and DREDS-CatNovel datasets.

4.2 STD Dataset

Example of CAD Models. We obtain CAD models of 42 category-level objects and 8 category-novel objects using the 3D reconstruction algorithm. For most of the objects, especially specular and transparent objects, we spray the dye and decorate objects with ink to enhance the reconstruction performance. 50 CAD models are shown in Figure 5.

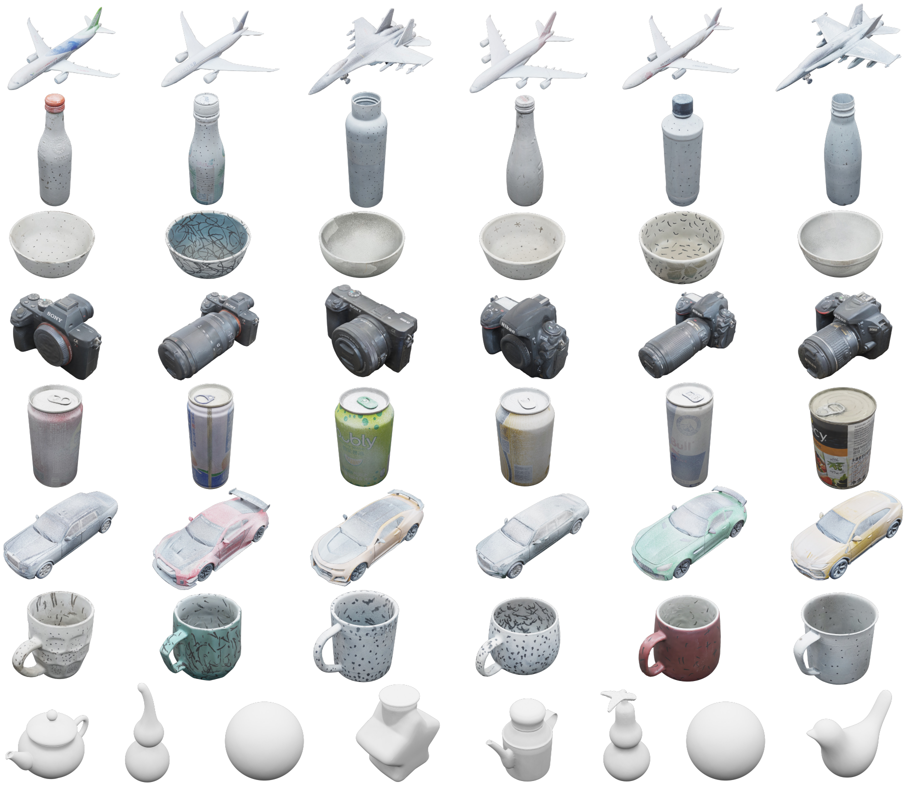

Fig. 5. CAD models of the STD object set. The 1st to 7th rows show 42 objects in 7 categories, and the last row shows 8 objects in novel categories.

Data Annotation. It is quite time-consuming to annotate such a large amount of real data. We propose to annotate the 6D poses of the objects in the first frame of each scene. Then the annotated 6D poses are propagated to the subsequent frames according to the camera poses with respect to the first frame. We calculate the camera poses using COLMAP [8]. In our annotation, we develop a program with GUI, enabling the user to move the CAD model, switching back and forth between the 2D image and 3D point cloud space to determine its pose, which facilitates labeling specular and transparent objects whose point clouds are severely missing or incorrect. After the 6D pose annotation, we can easily render other annotations like the ground truth depth, instance mask, *etc.* Figure 6 shows the examples and annotations of DREDS dataset.

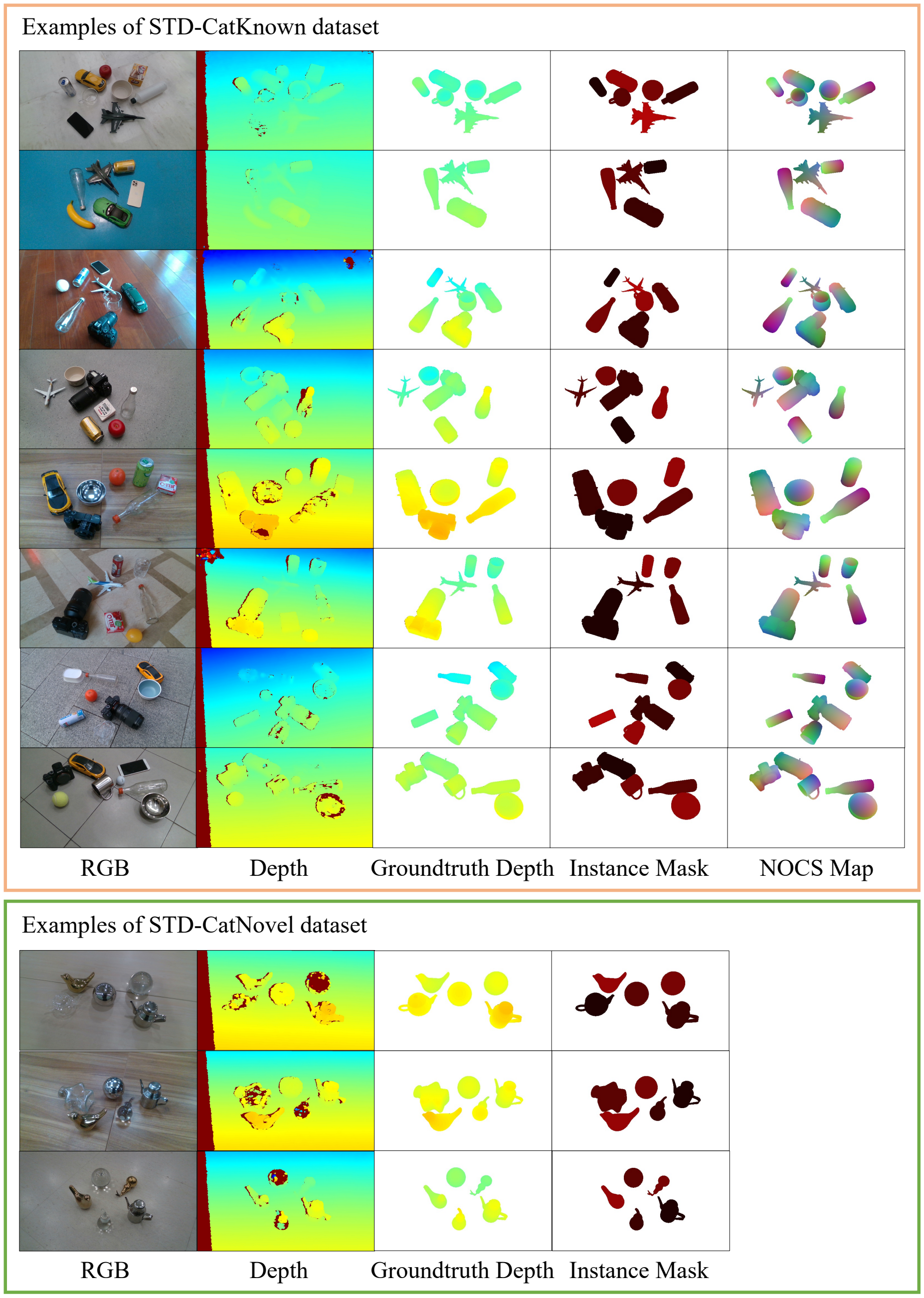

Fig. 6. Examples and annotations of STD-CatKnown and STD-CatNovel datasets. The ground truth depth maps are labeled only in the area of the 42 objects in 7 categories and the 8 objects in novel categories. Moreover, the NOCS maps are not annotated in STD-CatNovel dataset because there does not define the normalized object coordinate space for novel categories.

References

1. Bartell, F.O., Dereniak, E.L., Wolfe, W.L.: The theory and measurement of bidirectional reflectance distribution function (brdf) and bidirectional transmittance distribution function (btdf). In: Radiation scattering in optical systems. vol. 257, pp. 154–160. SPIE (1981)
2. Burley, B.: Extending the disney brdf to a bsdf with integrated subsurface scattering. Physically Based Shading in Theory and Practice'SIGGRAPH Course (2015)
3. Burley, B., Studios, W.D.A.: Physically-based shading at disney. In: ACM SIGGRAPH. vol. 2012, pp. 1–7. vol. 2012 (2012)
4. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
5. Chen, K., Dou, Q.: Sgpa: Structure-guided prior adaptation for category-level 6d object pose estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2773–2782 (2021)
6. Fang, H.S., Wang, C., Gou, M., Lu, C.: Graspnet-1billion: A large-scale benchmark for general object grasping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11444–11453 (2020)
7. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations (2018)
8. Schönberger, J.L., Frahm, J.M.: Structure-from-Motion Revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
9. Wang, H., Sridhar, S., Huang, J., Valentin, J., Song, S., Guibas, L.J.: Normalized object coordinate space for category-level 6d object pose and size estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2642–2651 (2019)