

# MultimodalStudio: A Heterogeneous Sensor Dataset and Framework for Neural Rendering across Multiple Imaging Modalities

## Supplementary Material

In this document, we provide some additional details on *MultimodalStudio*. Firstly, we show all the scenes included in *MultimodalStudio* (*MMS*)-*DATA* and we add a further description of them (Sec. A). Then, we present more in-depth results of the geometrical calibration for the multimodal setup (Sec. B). In Sec. C, we show an extensive set of comparisons between the renderings produced by *MMS-FW* and the ground truth for all the involved modalities, and we provide an exhaustive overview of the results achieved on every scene. Finally, we present some results and considerations on the few shot scenario (Sec. D).

### A. Dataset

We acquired 32 scenes from 50 different viewpoints. The viewpoints are arranged as shown in Fig. 1. The scene subject was placed on a table and enlightened by 3 halogen light bulbs equipped with diffusers. The sunlight could not penetrate the acquisition environment. Halogen lights were chosen because their emission spectrum is more suited to multispectral acquisition than that of LED light bulbs. LEDs are generally optimized to emit only in the visible band, but we are also interested in infrared emission given that the setup has a Near-Infrared (NIR) camera. Instead, halogen light bulbs have an emission spectrum that also covers bands beyond the visible range. The three lights were placed as the three vertices of an equilateral triangle centered at the object, but higher than it. On the table, a couple of ChAruco boards were placed to ease the rig pose estimation when running the Structure-from-Motion (SfM) algorithm, and to estimate the arbitrary scale factor that the SfM introduces. In Fig. 2 a complete overview of the acquired scenes is shown. It is possible to see that they capture a wide variety of common objects made of materials of different nature. In Tab. 1 it is possible to see which type of materials is present in each scene, for better classification. Hereunder, we report some additional details concerning specific scenes:

- **Clock and Glass Clock:** the clock in these scenes (Figs. 2g and 2n) is almost entirely made of plastic.
- **Laurel Wreath:** the wreath is made of real laurel but the leaves have been dried out by time (Fig. 2q).
- **Orchid:** the orchid in this scene is not a real organic flower and it is entirely made of plastic (Fig. 2t). It is also included in the scene Bouquet (Fig. 2e) along with real living plants.
- **Teddy Bear:** behind the teddy bear (Fig. 2x), a X-Rite colorchecker is placed for color calibration purposes.
- **Trophies:** the trophies upper part is metal, as well as the medals, while the middle and lower parts are made of plastic (Fig. 2ab).
- **Vases:** in the vases scene (Fig. 2ad), the horse statue is plastic, the vase with lid is ceramic, and the remaining two vases are made of glass.

In all the other scenes, the material appearance reflects the actual material nature.

We defined a fixed test and training split for the viewpoints (it is the same for all the scenes): we used viewpoints number 9, 19, 29, 39, and 49 as test views, while all the remaining ones are used as training views. Consider that the pictures have been acquired by moving the camera rig around the object in 2 circular patterns, a lower one (views 0-24) and an upper one (views 25-49) as shown in Fig. 1. However, recall that the camera rig is moved manually, as explained in Sec. 3.2 of the main paper, thus the actual viewpoint for each view can slightly change across different scenes.

### B. Geometrical Calibration

As anticipated in Sec. 3.3, the sensor calibration is performed employing five different ChAruco boards. We displaced the ChAruco boards in a column to ensure that the ChAruco patterns always span the whole vertical field-of-view of every sensor. Capturing the calibration patterns on every region of the image plane enables a better intrinsics and distortion parameter estimation, thus it is important that the matched features lay close to the frame border too. Due to this displacement, we only needed to shift the rig horizontally while capturing images to achieve the complete vertical and horizontal field-of-view coverage. Moreover, the patterns were acquired also at different distances from the cameras, in order to achieve a more robust geometrical calibration. In Tab. 2, the reprojection error plots for each sensor are shown. It is possible to observe that the Silios CMS-C1 multispectral camera calibration is the least accurate: this is explainable considering that the calibration was performed using the demosaicked frames. Indeed, demosaicking a  $3 \times 3$  multispectral pattern (Fig. 3) is not trivial: the employed bilinear interpolation introduces some block artifacts that may reduce the calibration pattern localization accuracy.

Scene	Diff.	Glossy	Reflect.	Transp.	Plastic	Metal	Wood	Organic	Paper	Cloth	Glass
African Art	x	x			x		x				
Aloe	x	x			x			x			
Bird House		x					x				
Book	x								x		
Bouquet		x			x			x			
Chess		x					x				
Clock		x	x		x						
Easter Egg			x		x						
Fan		x			x	x					
Forest Gang 1	x									x	
Forest Gang 2	x									x	
Fruits	x	x			x			x			
Gamepads	x	x			x						
Glass Clock			x	x	x						x
Globe		x			x						
Laptop		x			x						
Laurel Wreath	x		x		x			x		x	
Lego Ship		x			x						
Makeup	x	x	x	x	x	x				x	x
Orchid		x			x						
Pillow	x		x		x					x	
Plant		x			x			x			
Steel Pot			x			x					
Teddy Bear	x				x					x	
Tin Box 1			x			x					
Tin Box 2			x			x			x		
Toys		x		x	x						
Trophies		x	x		x	x				x	
Truck	x				x						
Vases	x	x		x	x						x
Watering Can 1	x				x						
Watering Can 2	x				x						

Table 1. Table showing which types of material are present in each scene of *MMS-DATA*.

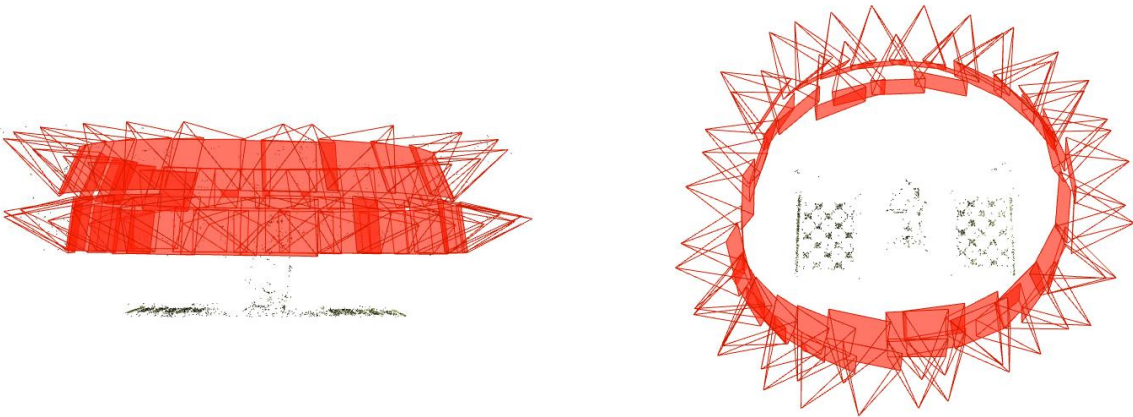


Figure 1. Overview of the RGB camera poses in a sample scene.

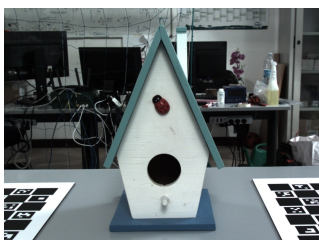




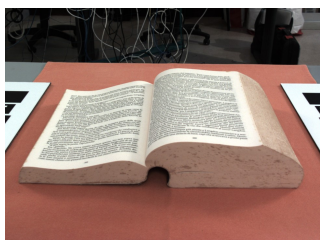
(a) African Art



(b) Aloe



(c) Bird House



(d) Book



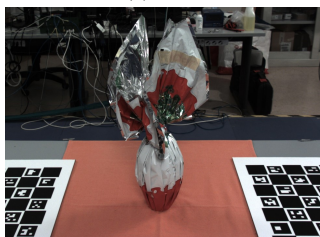
(e) Bouquet



(f) Chess



(g) Clock



(h) Easter Egg



(i) Fan



(j) Forest Gang 1



(k) Forest Gang 2



(l) Fruits



(m) Gamepads



(n) Glass Clock



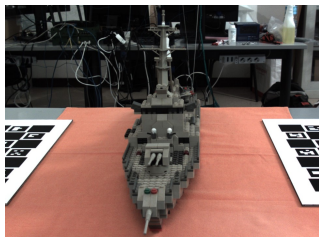
(o) Globe



(p) Laptop



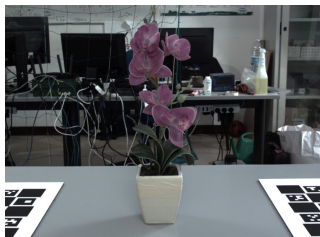
(q) Laurel Wreath



(r) Lego Ship



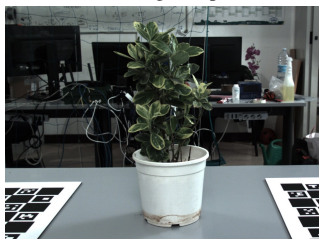
(s) Makeup



(t) Orchid



(u) Pillow



(v) Plant



(w) Steel Pot



(x) Teddy Bear

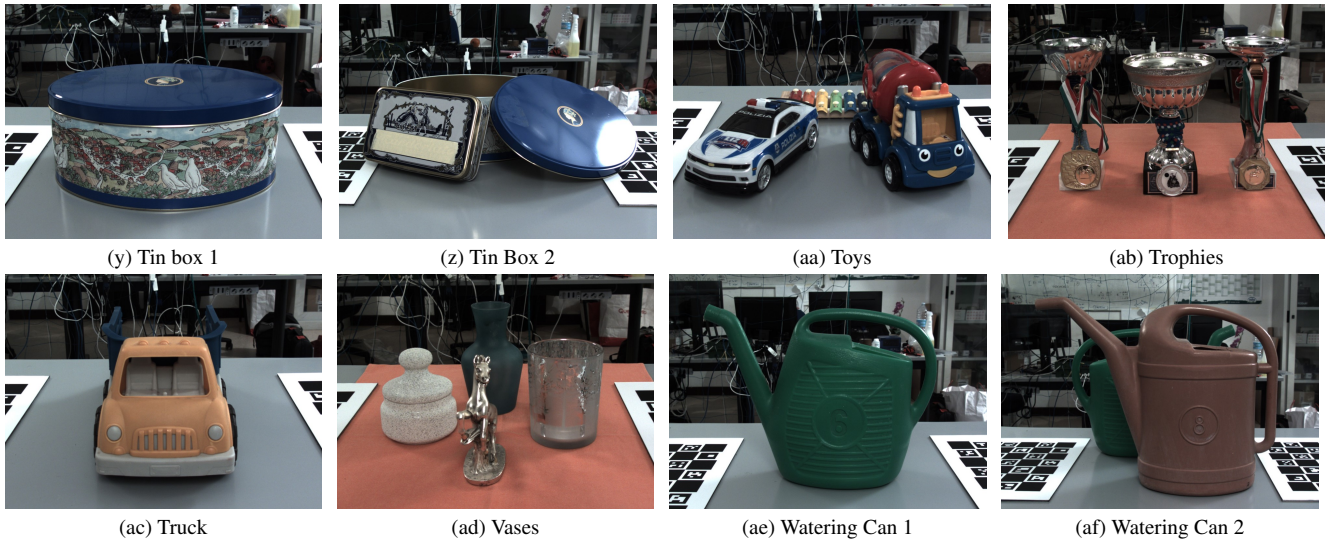


Figure 2. The complete set of scenes of *MMS-DATA*.

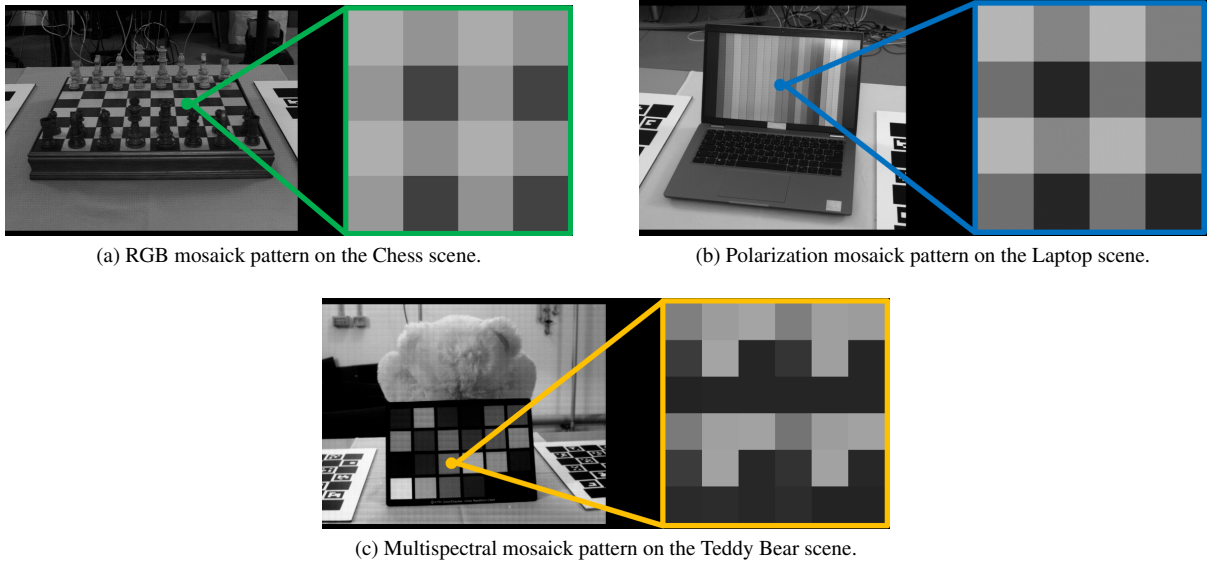


Figure 3. RGB, Polarization and Multispectral mosaick patterns. The RGB and Pol patterns are composed by 4 pixel arranged in a  $2 \times 2$  square. The MS pattern includes 9 pixels arranged in a  $3 \times 3$  square. In the figures, four patterns per modality are shown.



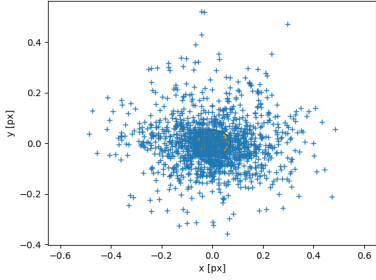
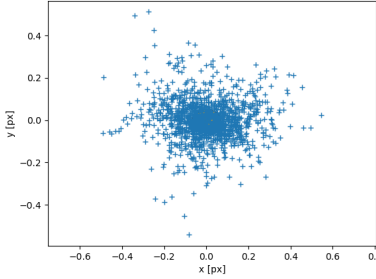
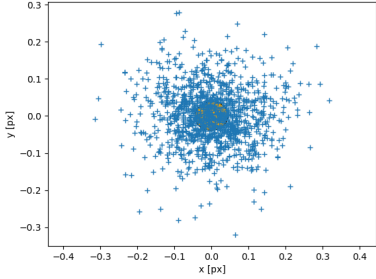
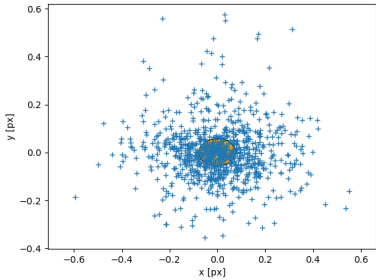
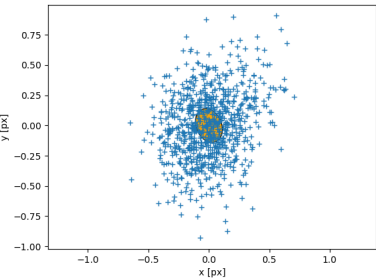
Modality	Sensor	Error Plot	RMSE (px)
RGB	Basler acA2500-14gm		0.18
Mono	Basler acA2500-14gc		0.17
NIR	Basler acA1300-60gmNIR		0.11
Pol	FLIR Blackfly S BFS-U3-51S5P		0.20
MS	Silios CMS-C1		0.36
Modality	Sensor	Error Plot	RMSE (px)

Table 2. Sensors used for the dataset acquisition and corresponding calibration accuracies. The error plots show the reprojection error distribution in terms of pixels. The yellow blob represents the error standard deviation.

Macroframes	Training Mod.	Test Mod.	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
5	RGB	RGB	15.73	0.63	0.30
	RGB-NIR	RGB	16.75	0.66	0.27
	RGB-NIR Mono-Pol-MS	RGB	19.46	0.73	0.20
10	RGB	RGB	16.97	0.68	0.25
	RGB-NIR	RGB	22.17	0.81	0.15
	RGB-NIR Mono-Pol-MS	RGB	25.84	0.89	0.09

Table 3. Few-shot results averaged on *MMS-DATA*. Metrics computed on the demosaicked frames rendered by the model.

### C. Additional Results

In this section, we provide an extensive overview of the rendering results obtained by the five-modality training described in Sec. 5.2 and in Tab. 4 of the main paper. In Tab. 4 it is possible to see the PSNR achieved for each scene, averaged on the five test views. Analyzing the results, we observe that the model struggles with scenes mostly containing reflective or transparent materials. For instance, these include the Pillow, Steel Pot, Tin Box 1, Tin Box 2, Trophies, and Glass Clock scenes. This mainly happens because of specular reflections, as confirmed by Figs. 4 to 8 where the reflection on the desk is challenging to estimate for all the modalities. The reflected radiance is a high frequency function because it is highly dependent on the viewpoint, thus it is difficult to predict for arbitrary viewpoints far from the training views. Even if our model employs Spherical Harmonics (SH) encoding [2] to ease the estimation of view-dependent high-frequency details, it is still not enough to always accurately estimate reflections from novel viewpoints. One possible solution is to introduce the estimation of the Bidirectional Reflectance Distribution Function (BRDF), as proposed in [1, 2], which is a parametric model that describes how light is reflected according to the specific properties of the surface materials; we leave it for future work.

In Figs. 9 and 10, we show some qualitative renderings of the normal and depth maps estimated by our model and some examples of perfectly aligned renderings of different modalities. Finally, in Figs. 11 to 13 we present some Polarization and Multispectral renderings, respectively. For the Polarization (Pol) renderings, we show both the Angle of Linear Polarization (AoLP) and the Degree of Linear Polarization (DoLP): it is possible to observe the limited difference in the measured and estimated polarization for mostly reflective or diffusive scenes. The Multispectral (MS) renderings show how the different multispectral channels capture different bands of the visible spectrum.

### D. Few-Shot Experiments

Given the number of advantages introduced by the use of multiple modalities, we also investigated whether their use could help to obtain better results in the few-shot scenario. For this reason, we performed some additional tests to evaluate the impact of additional modalities on the RGB rendering quality. We conducted two few-shot experiments involving the single-, two-, and five-modality training, by training the model with 5 and 10 macroframes, respectively. The results are reported in Tab. 3. Even if we did not explicitly develop our model to address the few-shot task, it is possible to see that introducing additional modalities is beneficial in terms of RGB rendering PSNR. With the introduction of a single additional modality (namely, the NIR in this case), we obtain an increase in PSNR of 1 dB and 5 dB for the 5 and 10 macroframes cases, respectively. This trend is also confirmed considering the single- and the five-modality training: we achieve a gain up to 3.5 dB and 9 dB, respectively. Analogously, the SSIM and the LPIPS also improve.

These preliminary but encouraging results, obtained with a model not specifically developed to handle the few-shot task, open new possibilities of investigation of the few-shot neural rendering and 3D reconstruction tasks in the multimodal scenario.

RGB

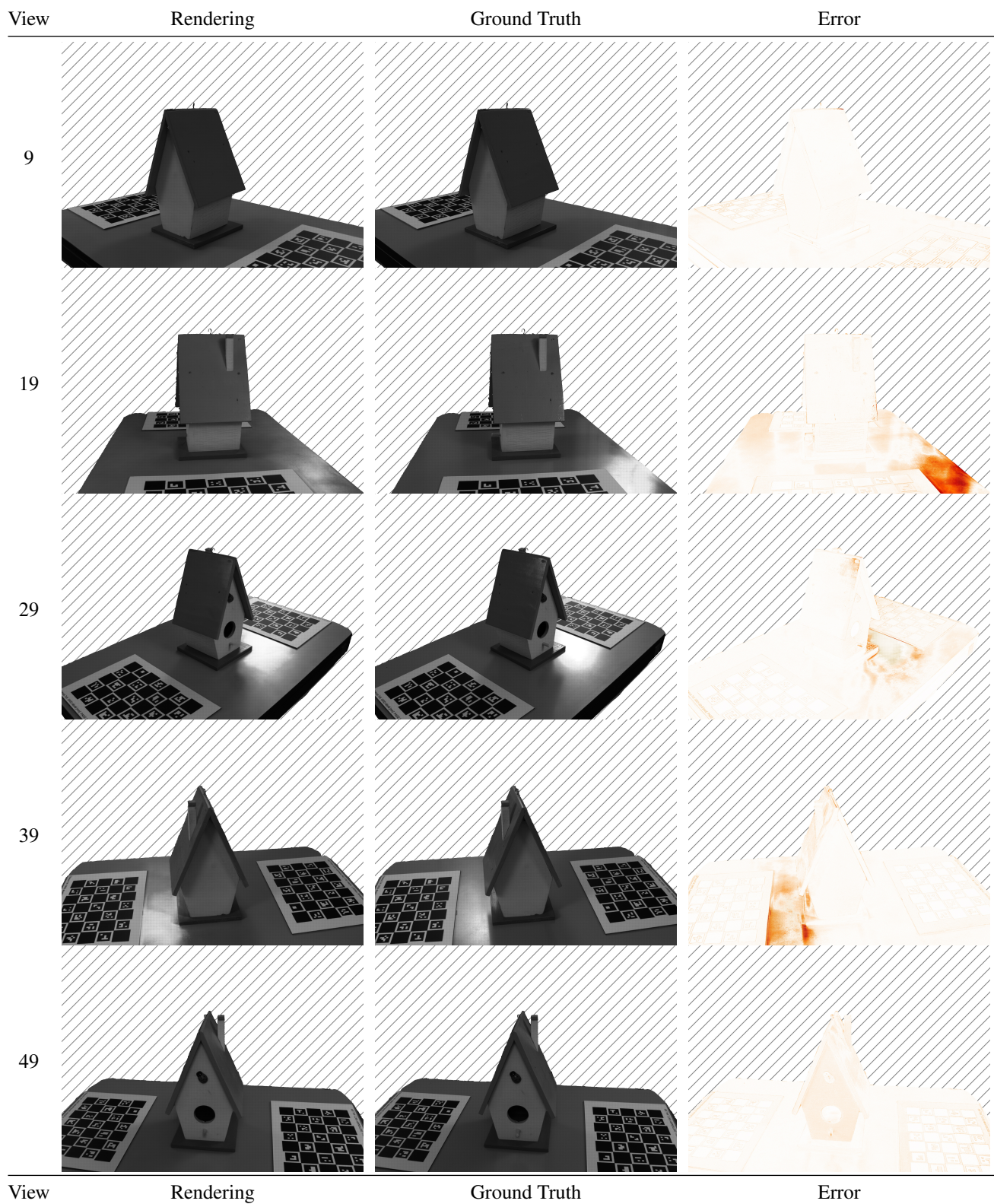


Figure 4. Mosaicked RGB renderings, ground truth and error maps of the Bird House scene from the five different test views.

Mono

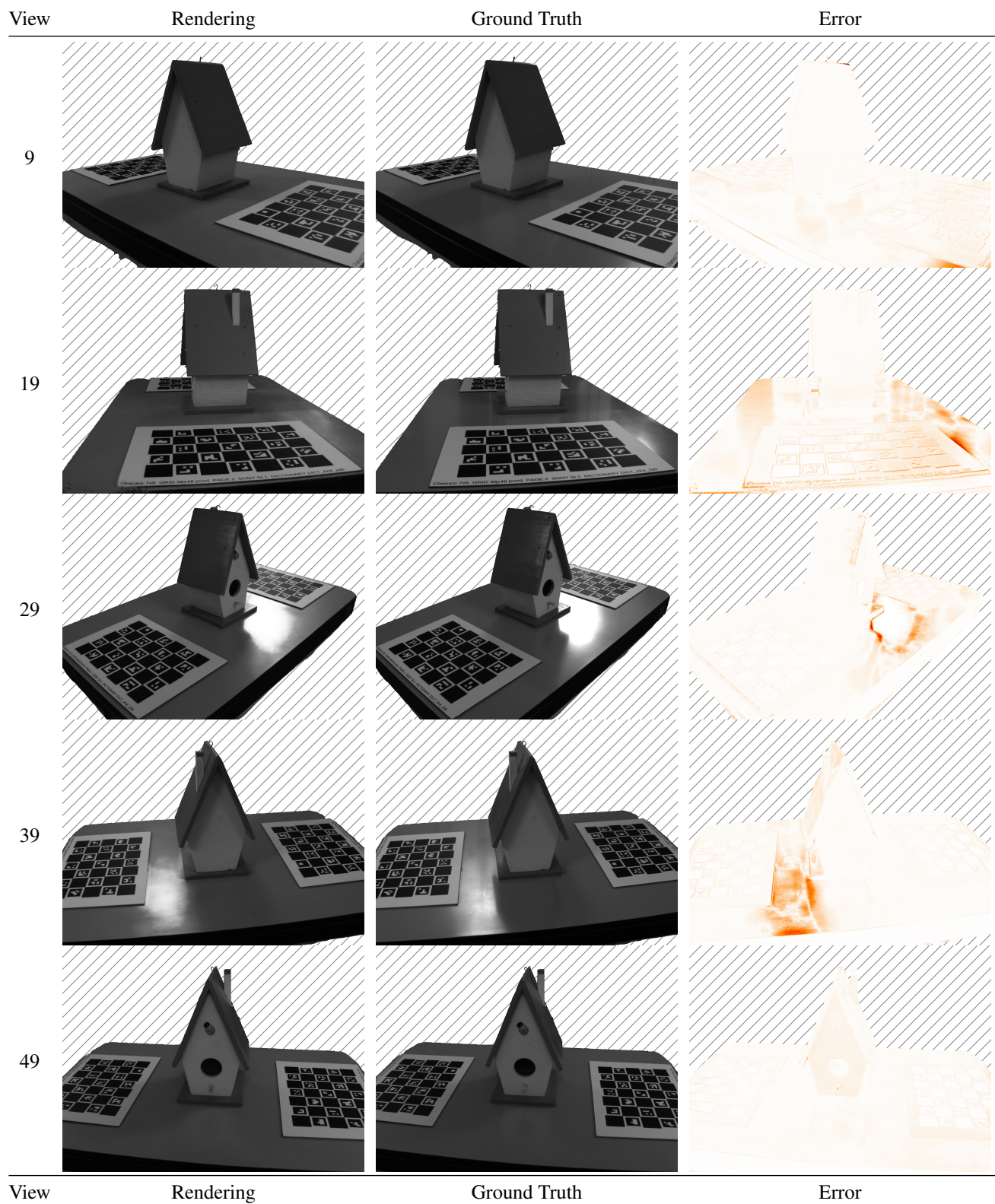


Figure 5. Mono renderings, ground truth and error maps of the Bird House scene from the five different test views.



# Near-Infrared (NIR)




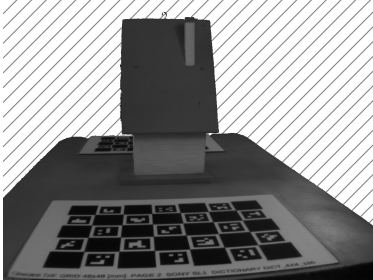
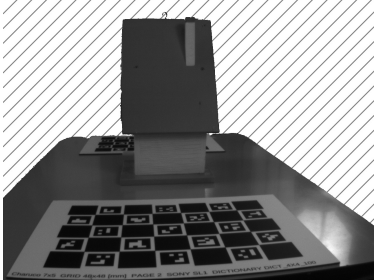

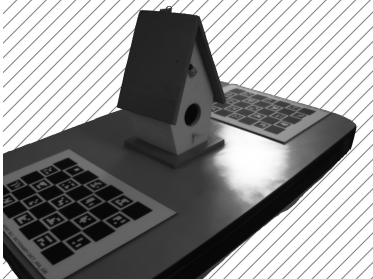
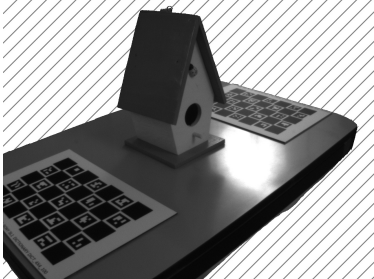
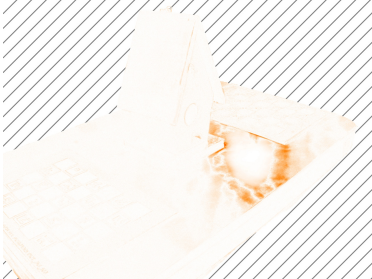
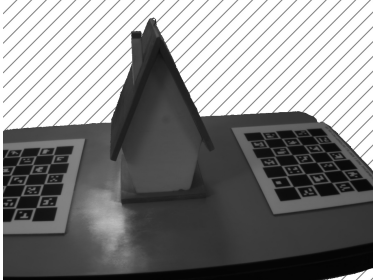
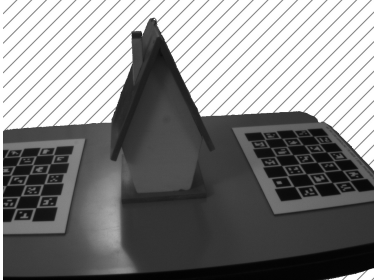

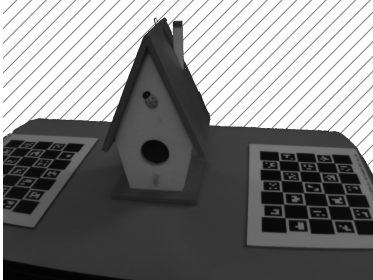
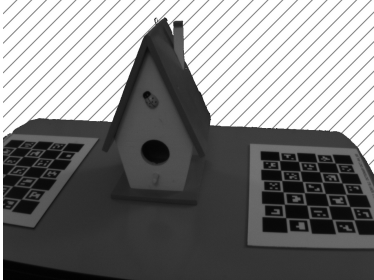

View	Rendering	Ground Truth	Error
9			
19			
29			
39			
49			
View	Rendering	Ground Truth	Error

Figure 6. Near-Infrared renderings, ground truth and error maps of the Bird House scene from the five different test views.

# Polarization (Pol)

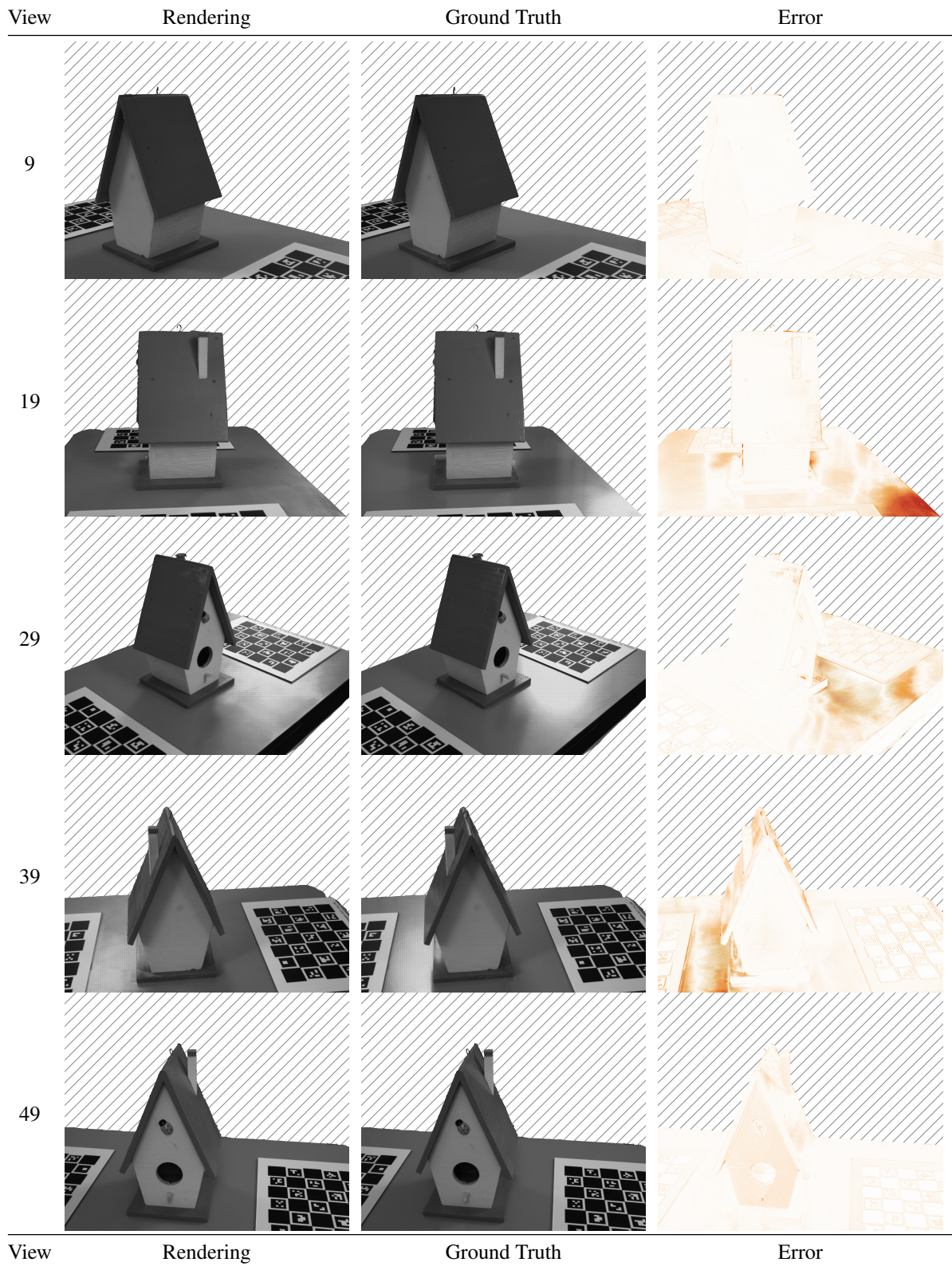


Figure 7. Mosaicked Polarization renderings, ground truth and error maps of the Bird House scene from the five different test views.

# Multispectral (MS)

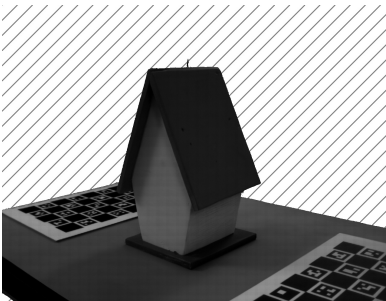
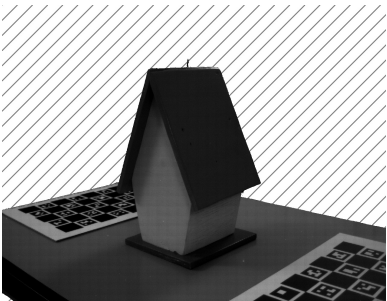
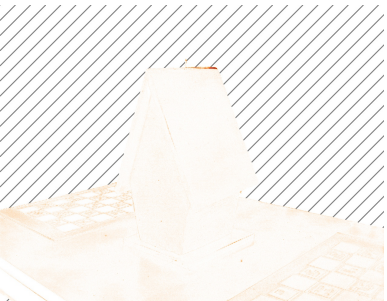
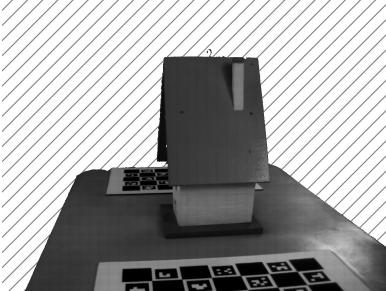
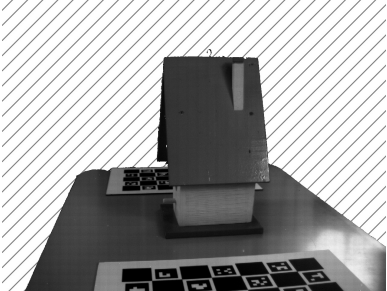
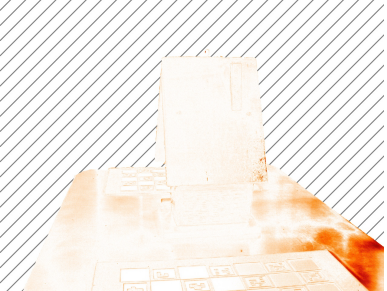
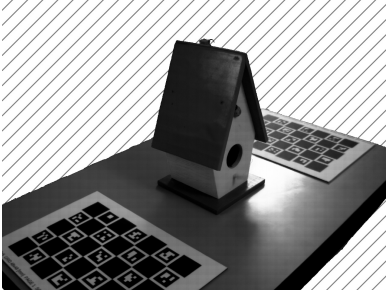
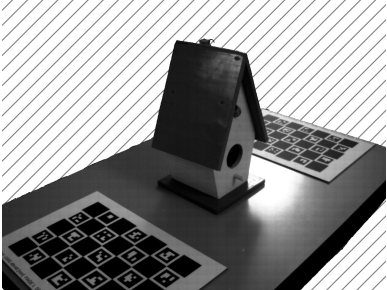

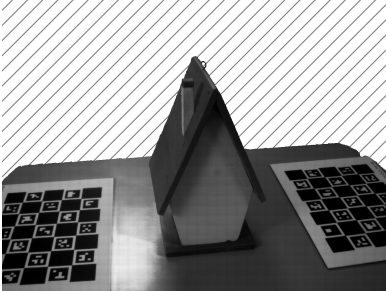
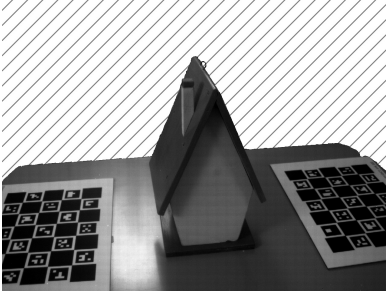

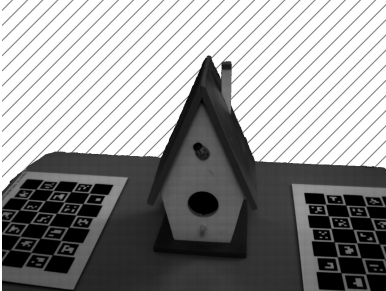

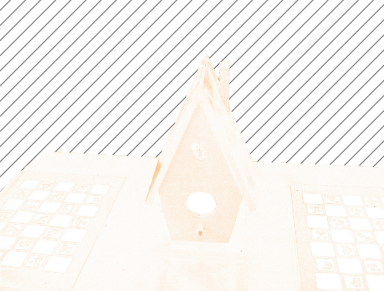
View	Rendering	Ground Truth	Error
9			
19			
29			
39			
49			
View	Rendering	Ground Truth	Error

Figure 8. Mosaicked Multispectral renderings, ground truth and error maps of the Bird House scene from the five different test views.



Training Mod.	Test Mod.	African Art	Aloe	Bird House	Book	Bouquet
RGB - Mono NIR - Pol - MS	RGB	34.87	31.47	30.40	34.84	29.73
	Mono	35.19	32.00	31.83	35.15	30.16
	NIR	35.94	33.78	32.18	37.27	33.24
	Pol	31.85	28.31	29.20	32.50	27.06
	MS	32.91	28.94	30.38	33.96	28.07
Training Mod.	Test Mod.	Chess	Clock	Easter Egg	Fan	Forest Gang 1
RGB - Mono NIR - Pol - MS	RGB	33.59	30.90	29.24	31.08	33.56
	Mono	33.62	32.22	30.75	31.88	34.23
	NIR	35.84	33.68	31.14	32.70	34.90
	Pol	31.53	29.04	26.71	28.04	33.09
	MS	31.46	30.16	28.56	30.59	32.77
Training Mod.	Test Mod.	Forest Gang 2	Fruits	Gamepads	Glass Clock	Globe
RGB - Mono NIR - Pol - MS	RGB	32.39	34.55	36.19	28.05	35.95
	Mono	32.82	34.56	34.56	29.75	35.29
	NIR	33.22	34.19	36.76	31.34	37.52
	Pol	32.49	32.99	33.85	25.94	32.70
	MS	31.21	33.37	34.90	27.10	34.04
Training Mod.	Test Mod.	Laptop	Laurel Wreath	Lego Ship	Makeup	Orchid
RGB - Mono NIR - Pol - MS	RGB	36.88	33.37	33.51	33.68	30.50
	Mono	36.02	34.28	34.46	31.78	31.03
	NIR	37.32	35.07	36.36	34.68	33.34
	Pol	34.04	30.61	31.80	31.28	28.21
	MS	34.45	30.81	32.11	33.09	29.00
Training Mod.	Test Mod.	Pillow	Plant	Steel Pot	Teddy Bear	Tin Box 1
RGB - Mono NIR - Pol - MS	RGB	27.67	29.89	27.52	35.83	28.20
	Mono	28.24	31.49	28.70	36.38	29.19
	NIR	29.18	32.97	29.10	37.16	31.34
	Pol	25.66	28.47	24.89	33.72	26.11
	MS	26.28	29.66	26.17	33.74	27.37
Training Mod.	Test Mod.	Tin Box 2	Toys	Trophies	Truck	Vases
RGB - Mono NIR - Pol - MS	RGB	29.34	33.33	25.88	31.45	34.83
	Mono	27.96	31.96	26.84	30.76	34.11
	NIR	28.51	34.48	27.99	31.84	34.96
	Pol	27.31	30.31	24.34	29.72	32.48
	MS	28.17	31.56	25.36	29.10	34.37
Training Mod.	Test Mod.	Watering Can 1	Watering Can 2	Mean		Std.
RGB - Mono NIR - Pol - MS	RGB	33.31	32.01	32.00		2.83
	Mono	34.93	31.64	32.31		2.49
	NIR	35.23	33.03	33.63		2.59
	Pol	29.74	29.59	29.80		2.79
	MS	31.75	30.76	30.69		2.64

Table 4. Five-modality mosaicked training results of all the scenes of *MMS-DATA* in terms of PSNR (dB).

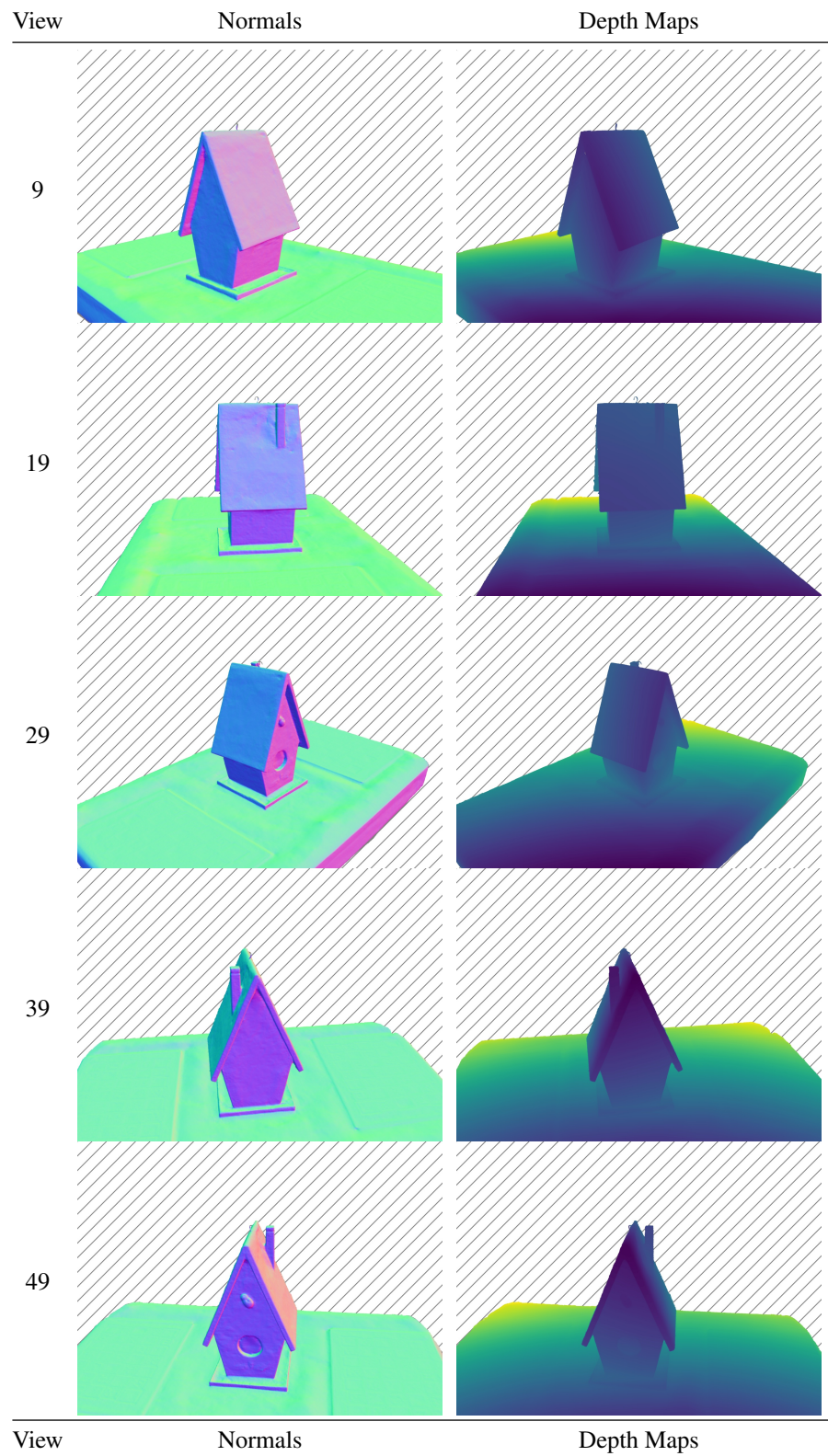


Figure 9. Normal and depth maps of the Bird House scene for the five different test views.



Figure 10. Aligned renderings of the Fruits, Aloe and Laurel Wreath scenes from the viewpoint number 49. Colors are for visualization purposes only.



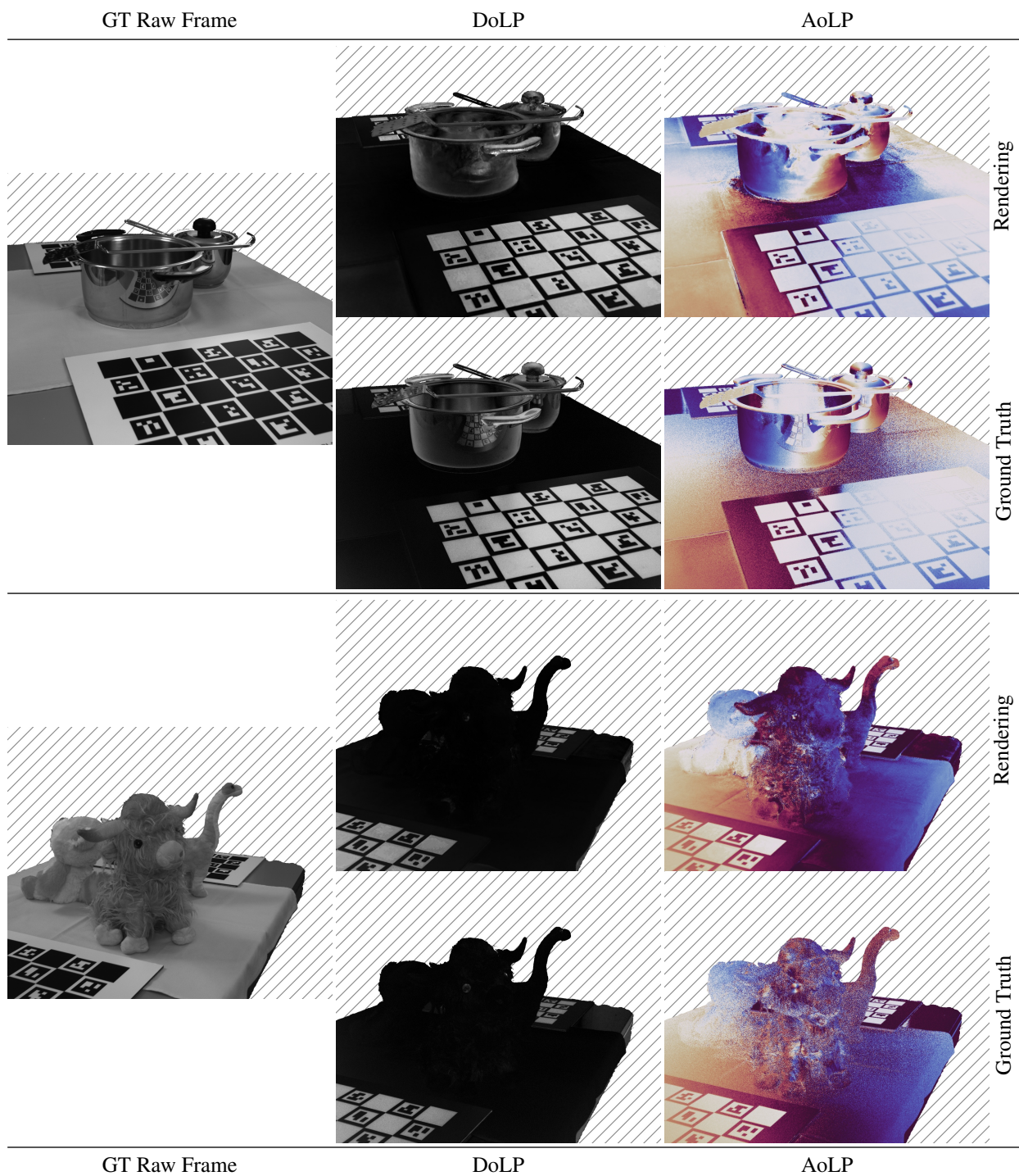


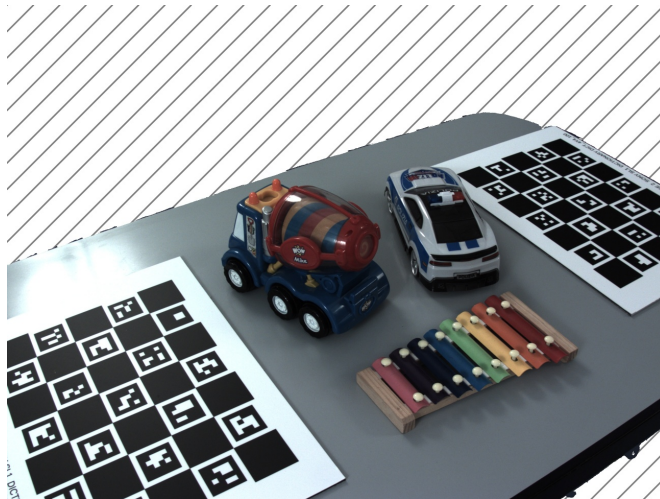
Figure 11. Degree of Linear Polarization (DoLP) and Angle of Linear Polarization (AoLP) of two scenes with different properties: on the upper side, Steel Pot, with mainly reflective materials; on the lower side, Forest Gang 1, with mainly diffusive materials.



Ch.	Band	Rendering	Ground Truth	Ch.	Band	Rendering	Ground Truth
0	692 nm			1	653 nm		
2	611 nm			3	572 nm		
4	541 nm			5	503 nm		
6	464 nm			7	431 nm		

Figure 12. Teddy Bear scene. On the top, the RGB ground truth view of the scene, for color reference purposes. On the bottom, the Multispectral individual channel renderings and ground truths.

RGB Ground Truth



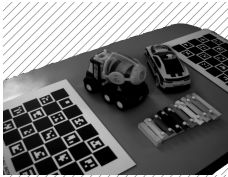
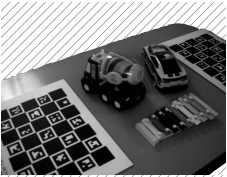
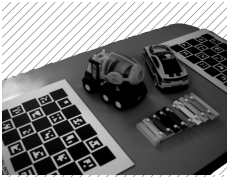
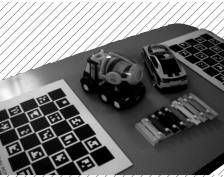
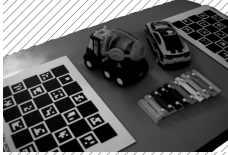
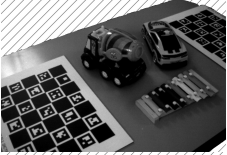
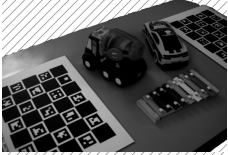
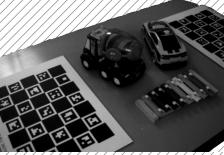

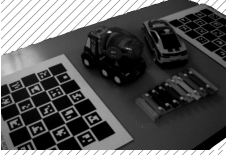
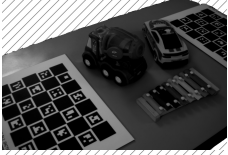
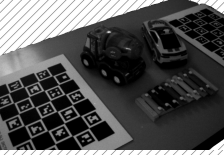
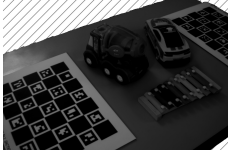
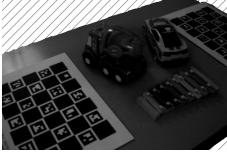
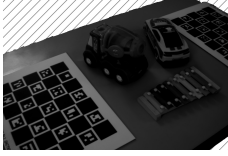
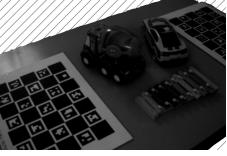
Ch.	Band	Rendering	Ground Truth	Ch.	Band	Rendering	Ground Truth
0	692 nm			1	653 nm		
2	611 nm			3	572 nm		
4	541 nm			5	503 nm		
6	464 nm			7	431 nm		

Figure 13. Toys scene. On the top, the RGB ground truth view of the scene, for color reference purposes. On the bottom, the Multispectral individual channel renderings and ground truths.

## References

- [1] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, pages 12684–12694, 2021.
- [2] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *CVPR*, pages 5481–5490. IEEE, 2022.