Multi-view Inverse Rendering for Large-scale Real-world Indoor Scenes Supplemental Material

Zhen Li[1] Lingli Wang[1] Mofang Cheng[1] Cihui Pan[1,*] Jiaqi Yang[2,*]
[1]Realsee [2]Northwestern Polytechnical University
yodlee@mail.nwpu.edu.cn,{wanglingli008,chengmofang001,pancihui001}@realsee.com,jqyang@nwpu.edu.cn

In this supplementary material, we provide more details of implementation (Sec. A), proposed datasets (Sec. B), additional experimental results (Sec. C) and discussions (Sec. D).

A. Details of Implementation

A.1. BRDF Model

In Sec. 3.2 in the main paper, f_d and f_s are defined as:

$$f_d = \frac{A}{\pi}, f_s = \frac{DFG}{4(n \cdot v)(n \cdot l)} \tag{1}$$

where A is albedo; l denotes light direction; n denotes normal; v denotes view direction; D denotes Normal Distribution Function (NDF); F denotes Fresnel function and G is the Geometry Factor. We adopt a simplified D, F and G [6, 12].

The specular D:

$$\begin{aligned} D &= \frac{\alpha^2}{\pi((n \cdot h)^2(\alpha^2 - 1) + 1)^2}, \\ h &= bisector(v, l), \\ \alpha &= R^2. \end{aligned} \tag{2}$$

The specular F:

$$F = 0.04 + (1 - 0.04)2^{(-5.55473(v \cdot h) - 6.98316)(v \cdot h)} \tag{3}$$

The specular G:

$$\begin{aligned} G &= G_1(l)G_1(v), \\ G_1(v) &= \frac{n \cdot v}{(n \cdot v)(1 - k) + k}, \\ k &= \frac{(R+1)^2}{8}. \end{aligned} \tag{4}$$

*Co-corresponding authors.

Figure 1. **Overview of our synthetic dataset.** It contains diverse materials and objects.

Table 1. **Comparison of costs.** N denotes the number of images. Our method achieves competitive performance on costs compared to the highly efficient method, NVDIFFREC [14]. The performance of TSDR∗ [15] is reported by their paper.

Method	Time (s)	Memory (MB)
TSDR∗ [15]	43200	-
InvRender [18]	50×N	5547
NVDIFFREC [14]	42×N	**2159**
NeILF∗ [17]	144×N	$>$ 32510
NeILF [17]	80×N	9783
Ours	**41**×N	2543

A.2. Implementation

We use neural networks to predict the depth image [5] and semantic segmentation [2] for each input image. The 3D mesh of whole scene is reconstructed with depth images and poisson surface reconstruction algorithm [7]. The room segmentation is calculated by occupancy grid [3].

Figure 2. **Overview of our challenging real dataset.** Our dataset consists of 10 Full-HDR indoor scenes with extremely complex lighting, geometry and materials.

Table 2. **Detailed quantitative comparison on our challenging real dataset. Although NVDIFFREC [14] reaches similar performance to our method, it fails to distinguish the ambiguity between albedo and roughness.**

Method	InvRender [18]			NVDIFFREC [14]			NeILF [17]			Ours		
	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓	PSNR↑	SSIM↑	MSE↓
Scene 1	23.4773	0.8367	0.0045	24.6780	0.8776	0.0034	23.5793	0.8405	0.0044	**25.5872**	**0.8984**	**0.0028**
Scene 2	22.3096	0.7603	0.0059	23.6182	0.8092	0.0043	22.5556	0.7691	0.0056	**24.1521**	**0.8450**	**0.0038**
Scene 3	21.8565	0.7959	0.0065	22.9661	0.8582	0.0050	21.8175	0.7994	0.0066	**25.3452**	**0.8820**	**0.0029**
Scene 4	21.0931	0.7443	0.0078	22.3015	0.8150	0.0059	21.0957	0.7464	0.0078	**23.0425**	**0.8451**	**0.0050**
Scene 5	23.0713	0.7764	0.0049	23.8165	0.8012	0.0042	23.3284	0.7897	0.0046	**24.2985**	**0.8367**	**0.0037**
Scene 6	23.0081	0.7885	0.0050	25.0760	0.8682	0.0031	22.7081	0.7860	0.0054	**26.1958**	**0.8943**	**0.0024**
Scene 7	20.5928	0.7395	0.0087	22.0116	0.8149	0.0063	20.5794	0.7512	0.0088	**23.1939**	**0.8481**	**0.0048**
Scene 8	20.8998	0.7083	0.0081	**25.8481**	**0.8816**	**0.0026**	20.4024	0.6965	0.0091	25.3344	0.8542	0.0029
Scene 9	21.2149	0.7474	0.0076	24.0453	0.8615	0.0039	20.7916	0.7331	0.0083	**24.3945**	**0.8732**	**0.0036**
Scene 11	22.4695	0.7710	0.0057	23.1026	0.8015	0.0049	22.4023	0.7747	0.0058	**24.5486**	**0.8461**	**0.0035**
Mean	21.9993	0.7668	0.0065	23.7464	0.8389	0.0044	21.9260	0.7687	0.0066	**24.6093**	**0.8622**	**0.0035**

We use 2048 samples to precompute the irradiance of sampled surface points. The NIrF is trained for 2000 epochs with the batch size of 16 and the total size of 1024 and we use the Adam optimizer [8] with a learning rate of 1e-4. The resolution of IrT is 1024×1024.

In material estimation, we use the Adam optimizer [8] with a learning rate of 3e-2 for 40 epochs in all three stages. We set β_{ssa} as 10 in stage 1, set β_{sp} as 1 in stage 2 and set β_{ssr} as 0.1 in stage 3. The resolution of albedo texture to be optimized is 2048×2048 and the resolution of roughness texture to be optimized is 4096×4096. We use 16 samples to re-render the specular component in material estimation. Considering the efficiency of optimization and the natural global illumination of proposed TBL, we apply nvdiffrast [9] with deferred shading to backward the gradient of image-space materials into corresponding textures. We note that nvdiffrast is orthogonal to our pipeline, which can be replaced by other differentiable renderers [4, 10, 13]. The pre-computed IrT takes around 10 minutes and the optimization process of material takes around 20 minutes.

B. Details of Proposed Datasets

B.1. Synthetic Dataset

As described in Sec. 4.1 in the main paper, to enable more comprehensive analysis, we create a synthetic scene with diverse material and light sources with a path tracer [11]. As shown in Fig. 1, the virtual scene is consists of three rooms and several objects with different materials. We generate 40 HDR panoramas, and corresponding poses, semantic segmentation, depth, albedo and roughness annotations, and the entire geometry. We use 24 views as input and others as novel views for the novel view synthesis.

B.2. Full-HDR Real Dataset

As described in Sec. 4.1 in the main paper, we capture 10 Full-HDR real indoor scenes due to the lack of Full-HDR real dataset. We first use neural networks to predict the corresponding depth images, and leverage SFM and MVS [16] to reconstruct the 3D mesh with the RGB texture. As shown in Fig. 2, 3D indoor scenes are reconstructed. Note that each indoor scene only contains 10 to 20 images. Therefore, the

Figure 3. **Additional samples of applications.** We edit the roughness of floors in Scene 1, Scene 4 and Scene 6, and the albedo for all scenes. Compared to source images, our method still reproduces realistic and consistent lighting effects after editing.

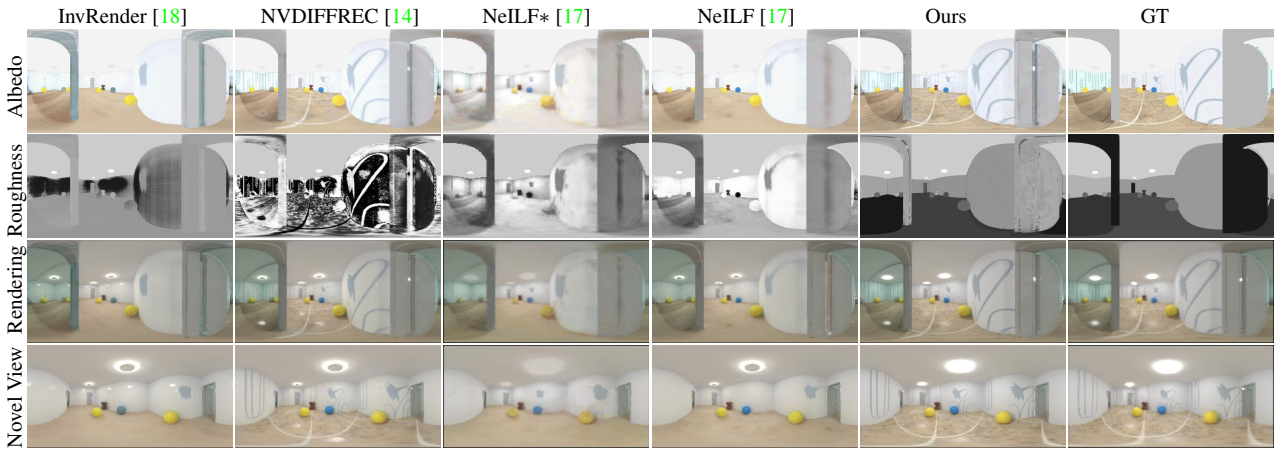

Figure 4. **Additional samples of qualitative comparison on synthetic dataset.** Our method reconstructs globally-consistent and physically-reasonable SVBRDFs while other approaches struggle to reduce ambiguity of materials.

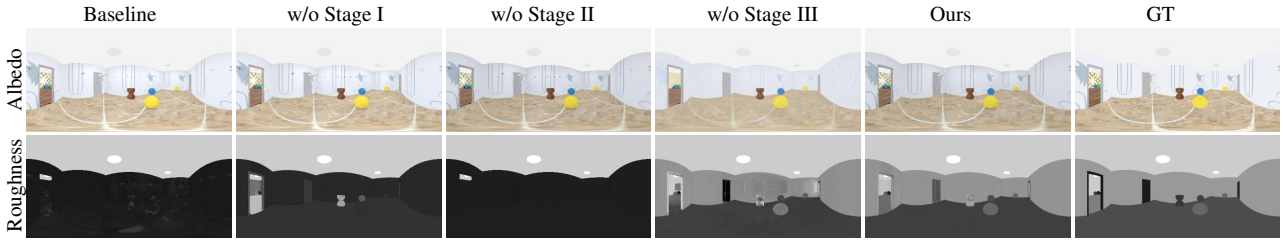

Figure 5. **Ablation study of three-stage material optimization on synthetic dataset.**

inverse rendering on these real scenes is extreme challenging.

C. Details of Experiments

C.1. Postprocessing

We change the albedo and roughness of ceiling and lamps as a postprocessing on synthetic dataset. We empirically found that the predictions of each approach on these regions are easily prone to local minimal. Based on the observation that the roughness and albedo of ceiling is high in most scenes, we set the roughness of ceiling as 0.8 and the albedo as 0.9. Please note that we update the results for each method on synthetic dataset.

C.2. Results on Costs

We compare the time cost and memory cost of material optimization to the multi-view inverse rendering methods in Tab. 1. Please note that all methods apply our efficient hybrid lighting representation except for NeILF∗ [17]. With our hybrid lighting representation, the efficiency of NeILF [17] is significantly improved. In material optimization, our approach achieves the comparable performance on costs to the previous highly efficient method, NVIDFFREC [14]. The calculation of IrT with a resolution of 1024× 1024 takes 10 minutes and costs around 2 GB GPU memory. The optimization process of material takes 20 minutes and also costs around 2 GB GPU memory. Note that the differentiable path tracing-based method [15] takes 12 hours per scene with a significant amounts of GPU memory [15].

C.3. Additional Results for Applications

In Sec.4.5 in the main paper, we demonstrate the capability of our method on several mixed-reality applications, such as material editing, editable novel view synthesis and relighting. We show more results on these applications in Fig. 3. Benefiting from our triangle mesh and PBR materials output, which is compatible with standard engines, we can easily edit the properties in a physical manner. We change the albedo or roughness according to the semantic segmentation, *e.g.*, the wooden floors become ceramic floors by changing the albedo of floors. Furthermore, we are able to render physically-reasonable novel views based on our 3D geometry and material textures, which is orthogonal to material editing, as shown in third column in Fig. 3. Last but not least, the entire scene can be rendered under new different illumination, as shown in last column in Fig. 3. Please refer to supplementary videos for more animations.

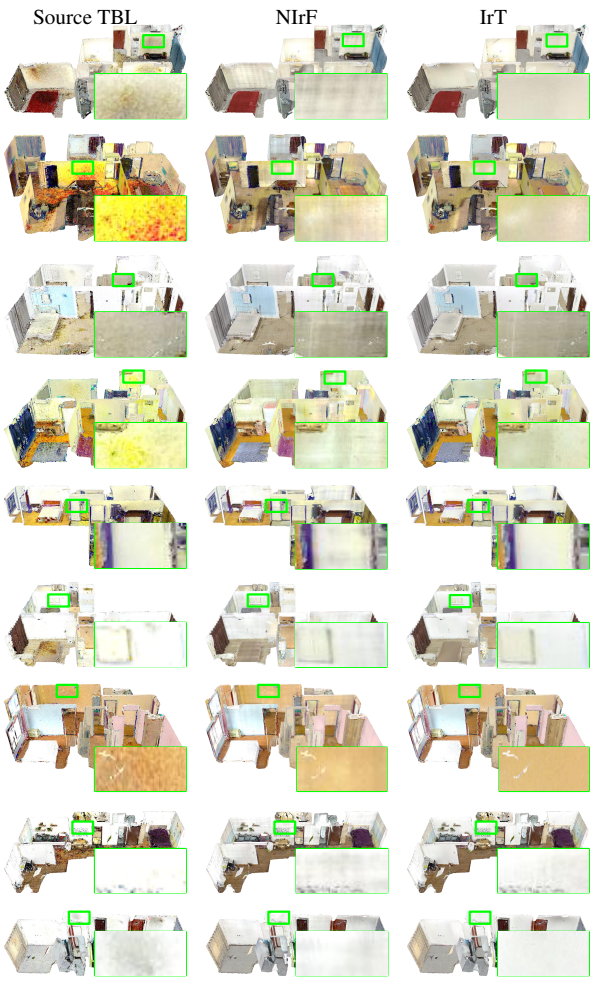

Figure 6. **Ablation study of hybrid lighting representation.** From top to down: Scene 1, Scene 2, Scene 3, Scene 4, Scene 5, Scene 6, Scene 7, Scene 8 and Scene 9. IrT recovers detailed albedo with less artifacts.

C.4. Additional Results on Synthetic Dataset

We provide more qualitative comparisons on synthetic dataset in Fig. 4. Our approach is superior than other inverse rendering methods on roughness estimation. And our physically-reasonable and globally-consistent SVBRDFs are able to produce realistic novel views. Note that NeILF [17] with our hybrid lighting representation more successfully disentangles the ambiguity between materials and lighting than NeILF$*$ [17] with their implicit lighting representation.

C.5. Additional Results on Real Dataset

As shown in Tab. 2 in the main paper, our approach outperforms previous neural rendering methods. The detailed results of each real scene are shown in Tab. 2. Note that we do not compare to PhyIR [12] with rendering error because it uses LDR panoramas as input. Although NVDIFFREC [14] reaches competitive performance to our method, it fails to distinguish the ambiguity between albedo and roughness in Fig. 7, Fig. 8 and Fig. 9. Our approach is able to reconstruct physically-reasonable and globally-consistent SVBRDF. Such properties re-render similar specular reflectance to GT with less wrong highlights in albedo, which proves we disentangle the ambiguity of materials successfully.

Table 3. **Ablation study of the quality of semantics.**

Property (PSNR)	0*0	16*16	32*32	64*64	128*128	256*256
Albedo	20.4169	20.7858	20.7353	21.0199	20.8364	19.7991
Roughness	20.2132	19.8076	19.9088	19.8964	17.8038	13.5650

C.6. Additional Results for Ablation studies

We showcase the effectiveness of our three-stage material optimization on synthetic dataset in Fig. 5. As described in Sec. 4.4 in the main paper, the Baseline only update the highlight regions of roughness. Without Stage I, the roughness leads to incorrect result. Without Stage II, the performance of roughness estimation will decrease dramatically. Without Stage III, the ableo is over-blur and the roughness is unsmooth.

As shown in Fig. 7 in the main paper, we show one sample for ablating the effectiveness of hybrid lighting representation. We show more results in Fig. 6. The proposed IrT recovers detailed albedo with less noise.

Additionally, we show more ablation studies of our material optimization strategy on real dataset in Fig. 10 and Fig. 11.

Finally, we show the performance of our method as the semantic segmentation mask becomes less accurate. We randomly change a cube region with wrong semantic labels for each input image. As shown in Tab. 3, our method is surprisingly robust as the length of cube increases.

C.7. Bad Cases

As described in Sec. 4.6 in the main paper, our method lead to recover bright albedo and low roughness when the light source is not captured. In Scene 8 in the Fig. 8, we reconstruct over-high albedo and over-low roughness nearby the window because the sun is not captured. The learning prior will be helpful for disentangling the ambiguity between materials in such cases.

D. More Discussions

D.1. Limitations and Future works

There are some limitations of our method. First, we rely on the HDR images to recover the proposed lighting representation for large-scale scene. To lift this limitation,

Figure 7. **Additional samples of qualitative comparison in the 3D mesh view on challenging real dataset.** From top to down: Scene 2, Scene 3, Scene 4 and Scene 5. Our method reconstructs globally-consistent and physically-reasonable SVBRDFs while other approaches struggle to produce inconsistent results and reduce ambiguity of materials.

the joint optimization of lighting and material will be explored. Second, our VHL-based sampling and semantics-based propagation requires that light sources are visible in the scene. If light sources are not captured, our method leads to recover bright albedo and low roughness. In such cases, we have to leverage the learning prior to alleviate the ambiguity of materials. Finally, although the geometry reconstructed by MVS is enough for our method, a more accurate geometry would lead to more accurate predictions.

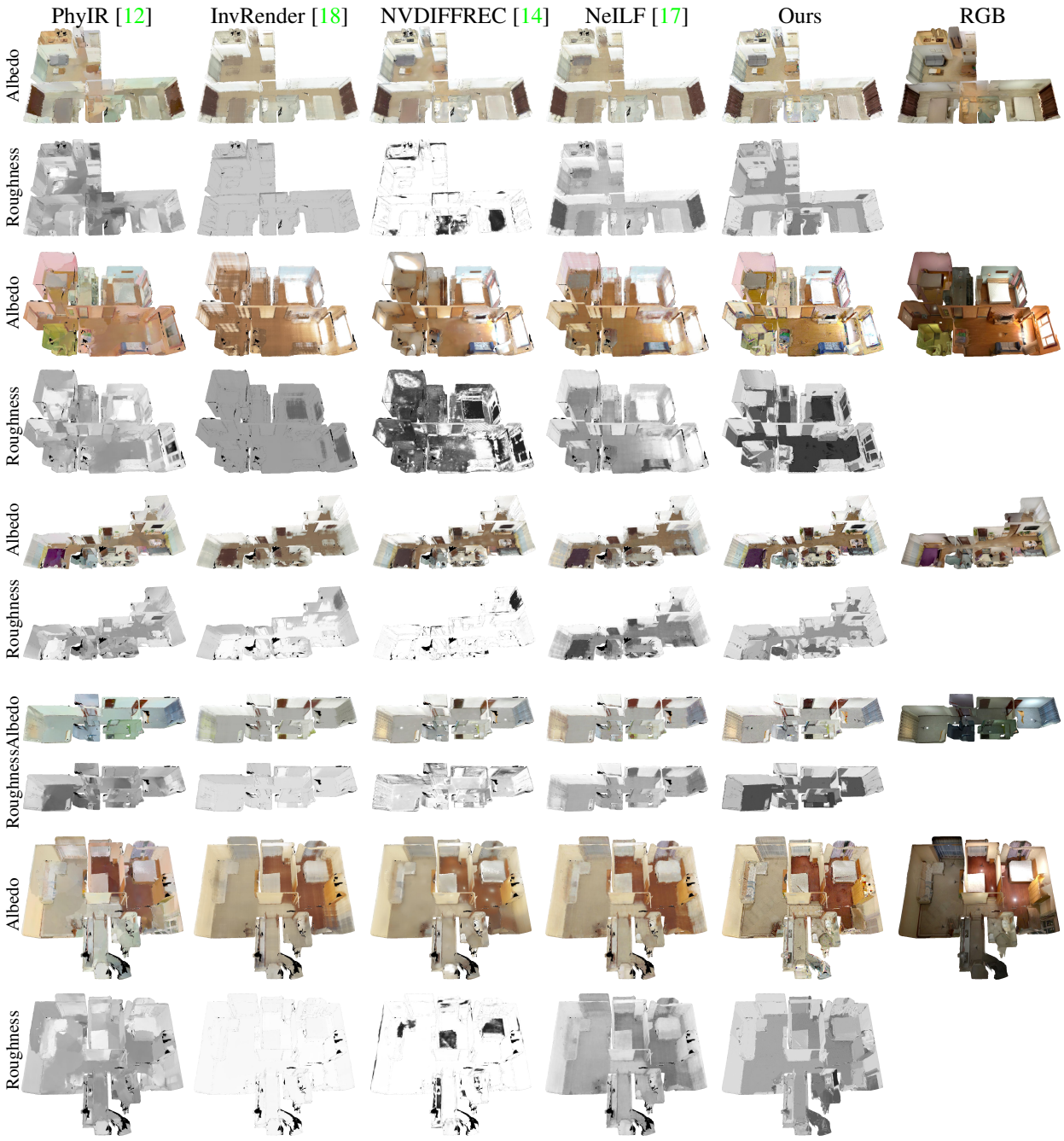

Figure 8. **Additional samples of qualitative comparison in the 3D mesh view on challenging real dataset.** From top to down: Scene 7, Scene 8, Scene 9 and Scene 11. Our method reconstructs globally-consistent and physically-reasonable SVBRDFs while other approaches struggle to produce inconsistent results and reduce ambiguity of materials.

D.2. TBL and Path tracing

The main pros of TBL is much less time and memory costs, compared to the path tracer [1, 15]. Our method only takes 30 minutes while [15] takes 12 hours per scene, reported in their paper. Moreover, the accuracy and robustness of TBL also is higher than the path tracer. If the recursive rendering equation can be computed instantly, the high gradient caused by the recursion and low samples in path sampling still do not ensure steady convergence [15]. On the one hand, our TBL models the complex light transport as a relatively simple local shading, which ensures more robust optimization. On the other hand, the global illumi-

nation of path tracing is finite-bounce while the TBL represents infinite-bounce global illumination, corresponding to real world. Therefore, the global illumination of TBL is more accurate.

In some cases, both our TBL and the path tracer do not work well, *e.g*., some important light sources or regions are missing, transparent/translucent objects, participating media and caustics. The differentiable volume rendering and neural rendering will be nice choices for such hard cases. I agree that some effects, *e.g*., a chain of specular reflections and retroreflections could be solved well using a path tracer while our TBL fails to model such effects. However, such effects are rare in most indoor scenes.

D.3. Broader Impacts

As described in the main paper, our method is able to produce realistic and physically-reasonable images with modified materials or illumination. Therefore, creating deepfake is a major potential negative impact. We can limit the target scenarios to prevent malicious use cases.

References

[1] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2019. 7

[2] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. https://github.com/open-mmlab/mmsegmentation, 2020. 1

[3] Alberto Elfes. Using occupancy grids for mobile robot perception and navigation. *Computer*, 22:46–57, 1989. 1

[4] Shi-Min Hu, Dun Liang, Guo-Ye Yang, Guo-Wei Yang, and Wen-Yang Zhou. Jittor: a novel deep learning framework with meta-operators and unified graph execution. *Science China Information Sciences*, 63(222103):1–222103, 2020. 2

[5] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters.*, 2021. 1

[6] Brian Karis and Epic Games. Real shading in unreal engine 4, 2013. 1

[7] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the Fourth Eurographics Symposium on Geometry Processing*, page 61–70, 2006. 1

[8] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2015. 2

[9] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. *ACM Transactions on Graphics.*, 39(6), 2020. 2

[10] Tzu-Mao Li, Miika Aittala, Frédo Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Transactions on Graphics. (Proceedings of SIGGRAPH Asia.)*, 37(6):222:1–222:11, 2018. 2

[11] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, pages 2475–2484, 2020. 2

[12] Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi. Yang. Phyir: Physics-based inverse rendering for panoramic indoor images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022. 1, 5, 6, 7

[13] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *Proceedings of the IEEE International Conference on Computer Vision.*, Oct 2019. 2

[14] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Mueller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022. 1, 2, 4, 5, 6, 7, 9

[15] Merlin Nimier-David, Zhao Dong, Wenzel Jakob, and Anton Kaplanyan. Material and Lighting Reconstruction for Complex Indoor Scenes with Texture-space Differentiable Rendering. In *Eurographics Symposium on Rendering.*, 2021. 1, 4, 7

[16] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *Proceedings of the European Conference on Computer Vision.*, 2016. 2

[17] Yao Yao, Jingyang Zhang, Jingbo Liu, Yihang Qu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf: Neural incident light field for physically-based material estimation. In *Proceedings of the European Conference on Computer Vision.*, 2022. 1, 2, 4, 5, 6, 7, 9

[18] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022. 1, 2, 4, 6, 7, 9

Figure 9. **Additional samples of qualitative comparison in the 2D image view on challenging real dataset.** From left to right and from top to down: Scene1, Scene2, Scene3, Scene4, Scene5, Scene6, Scene7 and Scene 11. Red denotes the Ground Truth image.

Figure 10. **Additional samples of ablation study of material optimization on challenging real dataset.** From top to down: Scene 1, Scene 2, Scene 3 and Scene 4.

Figure 11. **Additional samples of ablation study of material optimization on challenging real dataset.** From top to down: Scene 1, Scene 2, Scene 3 and Scene 4.