Appendix of UniVoxel: Fast Inverse Rendering by Unified Voxelization of Scene Representation

Shuang Wu*[1], Songlin Tang*[1], Guangming Lu[1], Jianzhuang Liu[2], and Wenjie Pei†[1]

[1] Harbin Institute of Technology, Shenzhen
[2] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences
{wushuang9811,wenjiecoder}@outlook.com
tangsonglin@stu.hit.edu.cn
luguangm@hit.edu.cn
jz.liu@siat.ac.cn

A Overview

We have proposed a novel inverse rendering framework based on the unified voxelization of scene representation. In this appendix, we present more results of our method. We describe the implementation details of the multi-resolution version of our *UniVoxel* in Sec. B and show additional ablation studies in Sec. C. Then we present the quantitative and qualitative results of the Shiny Blender dataset [4] in Sec. D. Furthermore, we show additional results on the MII [7] synthetic dataset and the NeRD [2] real-wold dataset in Sec. E and Sec. F, respectively. Finally, we discuss the limitation of our method in Sec. G.

B Implementation Details

For the multi-resolution hash encoding version of our *UniVoxel*, we employ a similar training paradigm used in our dense voxel grid version. During the first stage, we only optimize the radiance field to accelerate training while using the same resolution setting as the second stage. The total resolution levels of multi-resolution hash grid are set to $L = 16$. The coarsest resolution is 32 and the finest resolution is 2048. The channel of each learnable feature is set to 2 and the hash table size of each resolution level is set to 2^{19}. The tiny MLP network for SDF decoding comprises 1 hidden layer with 64 channels, and the tiny MLP networks for other scene properties comprise 2 hidden layers with 64 channels. The number of Spherical Gaussian lobes is $k = 16$. The weights of the losses are tuned to be $\lambda_{pbr} = 10.0$, $\lambda_{rad} = 10.0$, $\lambda_n = 0.01$, $\lambda_\kappa = 0.1$, $\lambda_\zeta = 0.01$, $\lambda_{sg} = 0.1$ and $\lambda_{white} = 0.01$. We employ additional eikonal loss to regularize SDF value and the weight is tuned to be 0.01. We use the AdamW optimizer with learning rate 0.01, weight decay 0.01, and a batch size of 8192 rays to optimize the scene representation for 10k iterations in both the two stages.

* Equal contribution.
† Corresponding author.

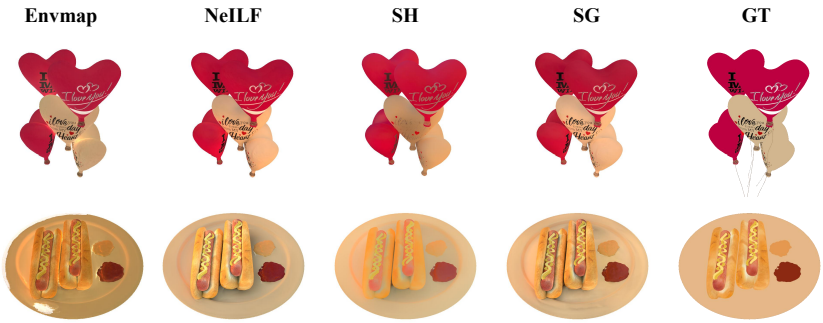

Fig. 1: Visualization of the reconstructed albedo maps by different illumination models.

Fig. 2: Visualization of the albedo maps reconstructed by our method with/without the regularization for Spherical Gaussians.

Table 1: Ray batch size and the required GPU memory for training each method on the MII synthetic dataset.

Method	Batch Size	GPU Memory
TensoIR [3]	4096	≈12GB
MII [7]	1024	≈14GB
UniVoxel(Hash)	8192	≈16GB
UniVoxel	8192	≈19GB

C Additional Ablation Studies

C.1 Comparison of Different Illumination Models

We show the qualitative results of different illumination models in Fig. 1. Using the environment map to model illumination leads to poor albedo maps due to the computational challenges involved in computing light visibility and indirect lighting, making optimization difficult. When employing MLP to predict incident radiance directly, as done by NeILF [6], the lighting tends to be baked into the albedo map without constraints for the illumination. Modeling incident lights using Spherical Harmonics (SH) fails to recover high-frequency illumination, causing color deviations in certain regions of the albedo. The visualization aligns with the quantitative results presented in Tab. 2 of the main paper.

Fig. 3: Comparison of the albedo maps with different SG smoothness loss weight.

C.2 Effectiveness of the Regularization for Spherical Gaussians

We compare the reconstructed albedo optimized with and without the regularization for Spherical Gaussians in Fig. 2. Without L_{sg}, the illumination tends to be baked into the predicted albedo, resulting in poor texture recovery of the pillow, which demonstrates the effectiveness of our proposed regularization in alleviating the material-lighting ambiguity. The visualization aligns with the quantitative results presented in Tab. 3 of the main paper.

C.3 Effect of the SG Smoothness

We compare the estimated albedo maps with different SG smoothness loss weight λ_{sg} of Eq. 16 on *StateOfLiberaty* scene from the NeRD dataset in Fig. 3. It can be observed that using a larger λ_{sg} will result in shadows appearing on the albedo maps. Due to the more complex lighting conditions in outdoor scenes, it is advisable to reduce the constraints on illumination to eliminate these shading components.

C.4 Visualization for Incident Lights

We show the incident light maps in Fig. 4. Our illumination model is able to represent the effect of direct lighting, occlusions and indirect lighting simultaneously. As shown in the *air balloons* scene of Fig. 4, point x_1 locates at the top of the balloons, therefore receiving predominantly ambient lights as its incident lights. On the other hand, point x_2 is located at the saddle point of the balloons, where the surrounding surfaces exhibit low roughness. Consequently, a portion of the incident lights in its incident light map is composed of red light reflected from the neighboring surfaces. In contrast, The environment maps learned by TensoIR [3] only model direct lighting, thus lack the capability to capture such spatially-varying indirect lighting.

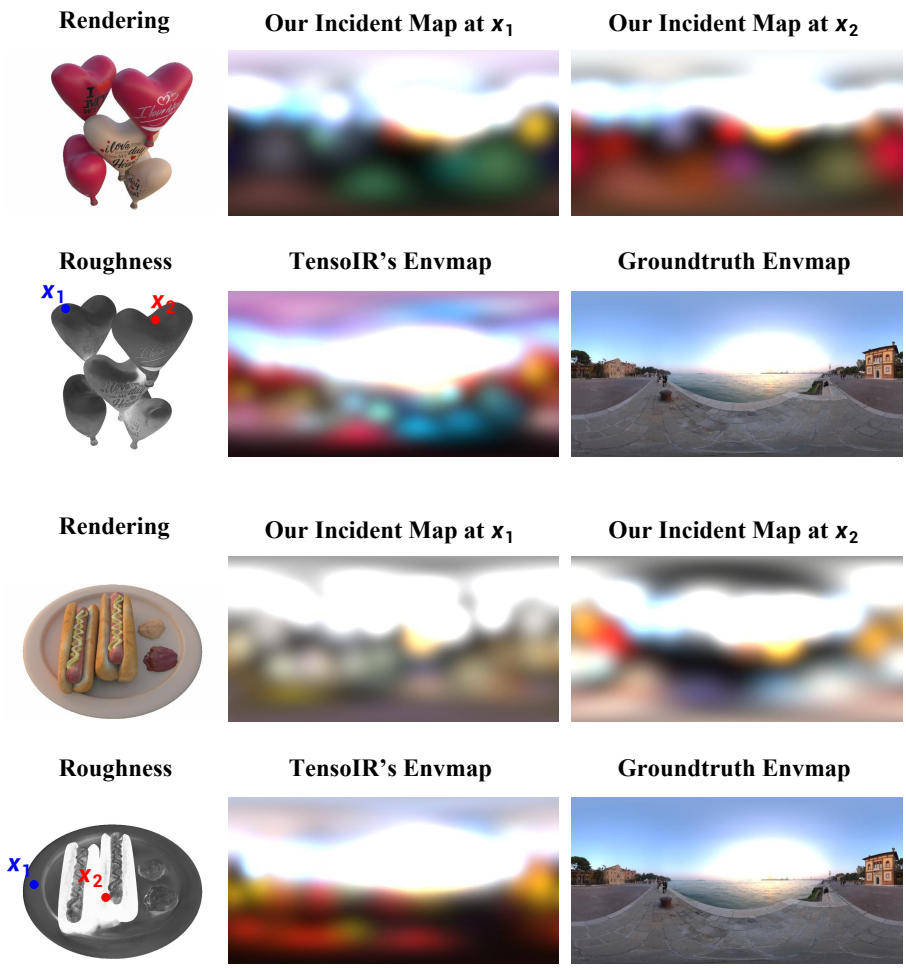

Fig. 4: Visualization of the incident light maps reconstructed by our method. Note that our incident light field is designed for modeling both direct lighting and indirect lighting, while the environment map learned by TensoIR is only designed for modeling direct lighting.

C.5 Comparison of GPU Memory

We present the ray batch size and required GPU memory of each method for training on the MII synthetic dataset in Tab. 1. It can be seen that our method does not significantly exceed the GPU memory of other methods, thanks to our efficient implementation. The GPU memory can be further optimized by the multi-resolution hash encoding version of our *UniVoxel*.

Table 2: Quantitative evaluation on the Shiny Blender dataset. We report the per-scene mean angular error (MAE°) of the normal vectors as well as the mean MAE° over scenes.

MAE° ↓	teapot	toaster	car	ball	coffee	helmet	mean
Mip-NeRF [1]	66.470	42.787	40.954	104.765	29.427	77.904	60.38
Ref-NeRF [4]	9.234	42.870	14.927	1.548	12.240	29.484	18.38
Voxurf [5]	8.197	23.568	17.436	30.395	8.195	20.868	18.110
TensoIR [3]	8.709	60.968	35.483	100.679	15.728	76.915	49.747
Univoxel	6.855	11.515	8.987	1.635	23.654	3.108	**9.292**

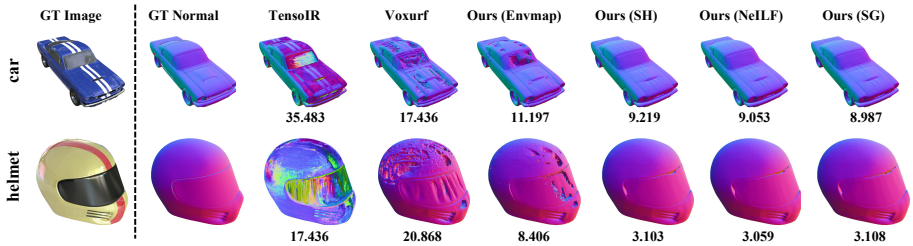

Fig. 5: Qualitative comparison of normal maps on 2 scenes from the Shiny Blender dataset. We report the average MAE° below each image.

D Results on the Shiny Blender Dataset

We conducted experiments on the challenging Shiny Blender dataset [4]. As shown in Tab. 2, our *UniVoxel* achieves better geometric quality compared to other methods. In Fig. 5, we visualize the normal maps produced by different methods, and it can be observed that our *UniVoxel* recovers geometry in the specular regions more accurately than TensoIR and Voxurf. Additionally, we present the recovered geometry, materials and illumination in Fig. 6. It can be seen that TensoIR fails to reconstruct materials in the specular regions and bakes the lighting into the albedo maps, whereas our method predicts realistic materials.

E Additional Results on the MII Synthetic Dataset

From Fig. 7 to Fig. 10, we present complete qualitative results on 4 scenes from the MII synthetic dataset: *air balloons*, *chair*, *hotdog* and *jugs*. Compared to baseline methods, our *UniVoxel* demonstrates superior reconstruction quality in high-frequency details, which is consistent with the quantitative results presented in Tab. 1 of the main paper.

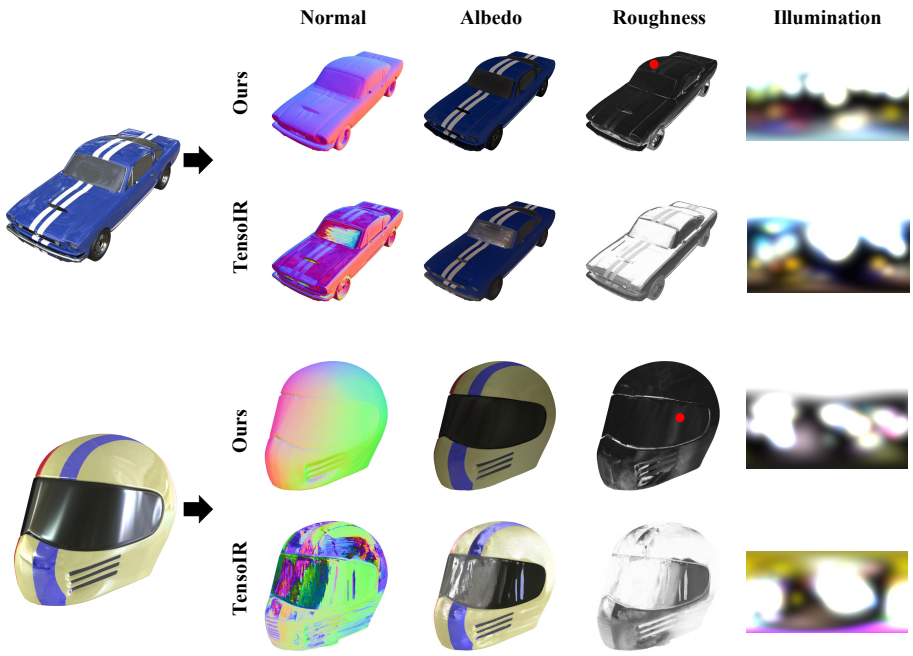

Fig. 6: Qualitative comparison of geometry, materials and illumination on 2 scenes from the Shiny Blender dataset. For our method, we generate the incident light maps at the location of the red points in the roughness maps.

F Additional Results on the NeRD Real-World Dataset

From Fig. 11 to Fig. 13, we show complete qualitative results on the 3 scenes from the NeRD real-word dataset: *StatueOfLiberty*, *Gnome* and *MotherChild*. Although there is no ground truth for reference, we can observe that all baseline methods exhibit poor reconstruction quality in these scenes. The main reason is that the environment maps cannot model the complex lighting conditions in the real world. In contrast, our *UniVoxel* is able to handle various illumination effects, enabling the recovery of geometry and material with relatively superior quality, and the generation of more photo-realistic relighting images.

G Limitation

It is still challenging for our *UniVoxel* to fully decouple lighting from materials, which is also a crucial crux for other inverse rendering methods. For instance, the shadows on the albedo map of the air balloons in Fig. 7 cannot be completely eliminated by our method. This issue could be potentially alleviated by introducing prior knowledge about materials, which we will investigate in future work.

Fig. 7: Qualitative comparison on *air balloons* from the MII synthetic dataset.

Fig. 8: Qualitative comparison on *chair* from the MII synthetic dataset.

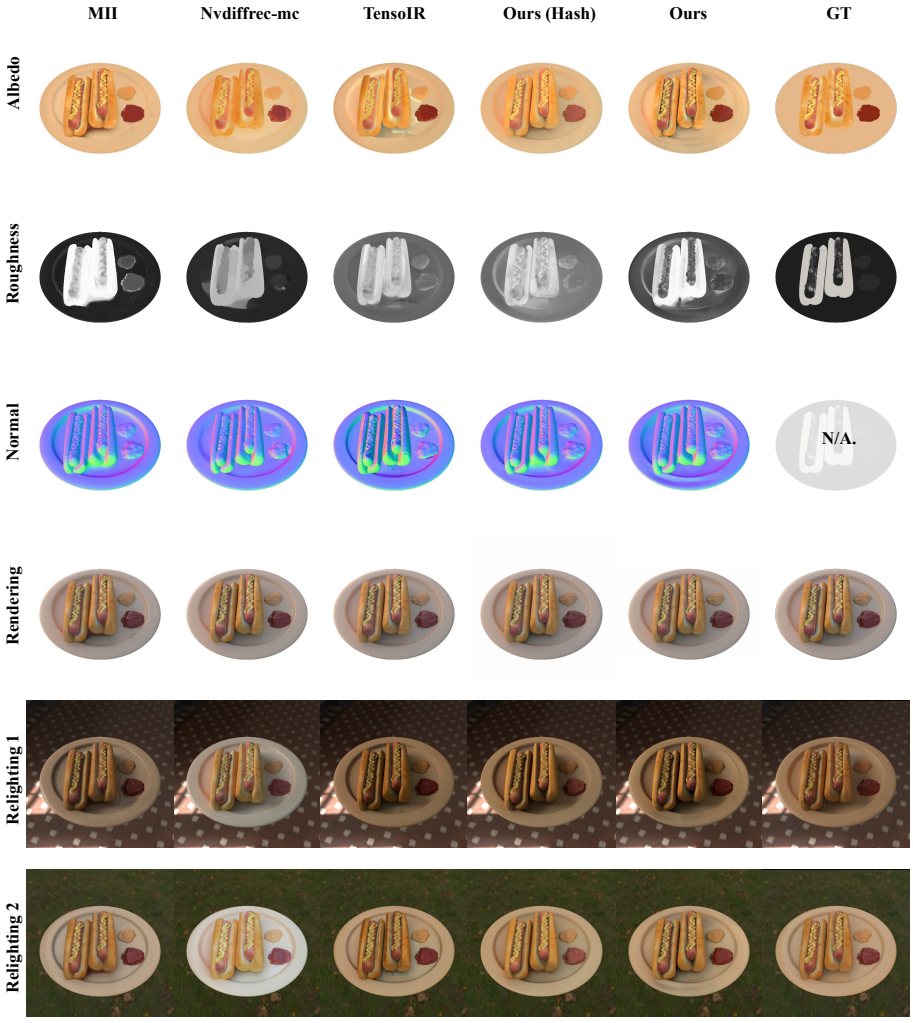

Fig. 9: Qualitative comparison on *hotdog* from the MII synthetic dataset.

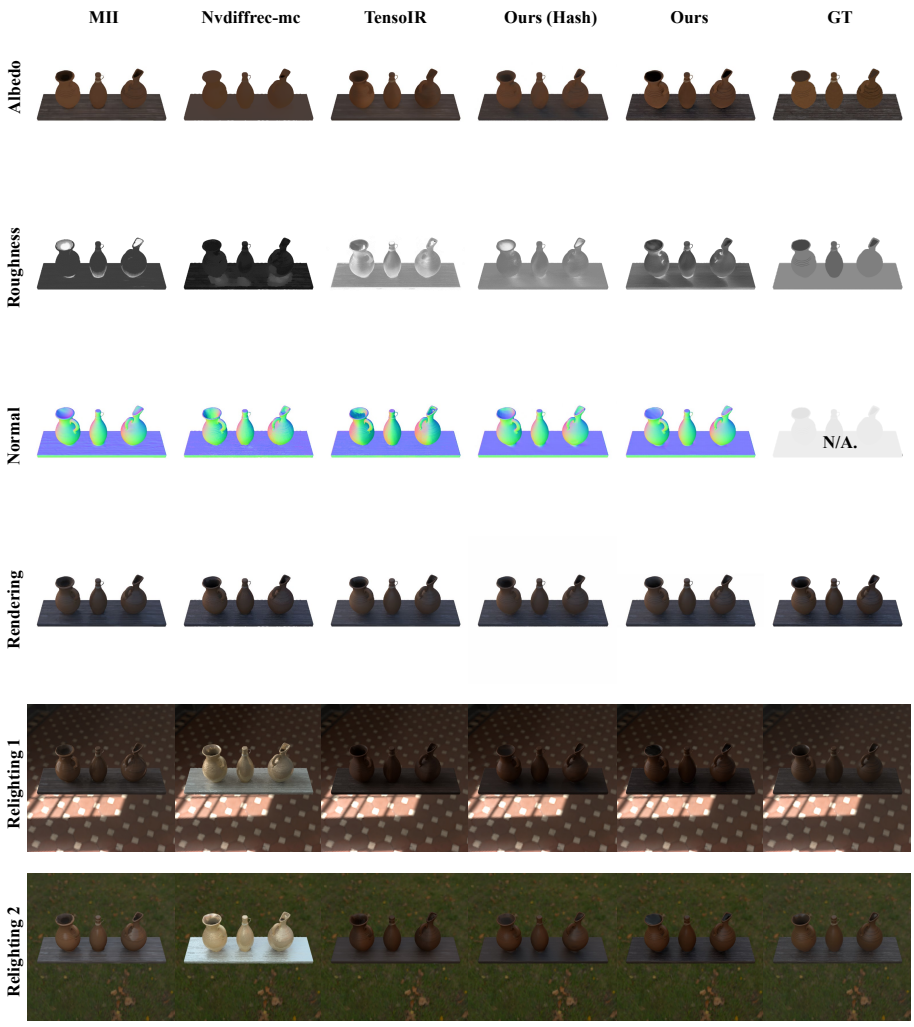

Fig. 10: Qualitative comparison on *jugs* from the MII synthetic dataset.

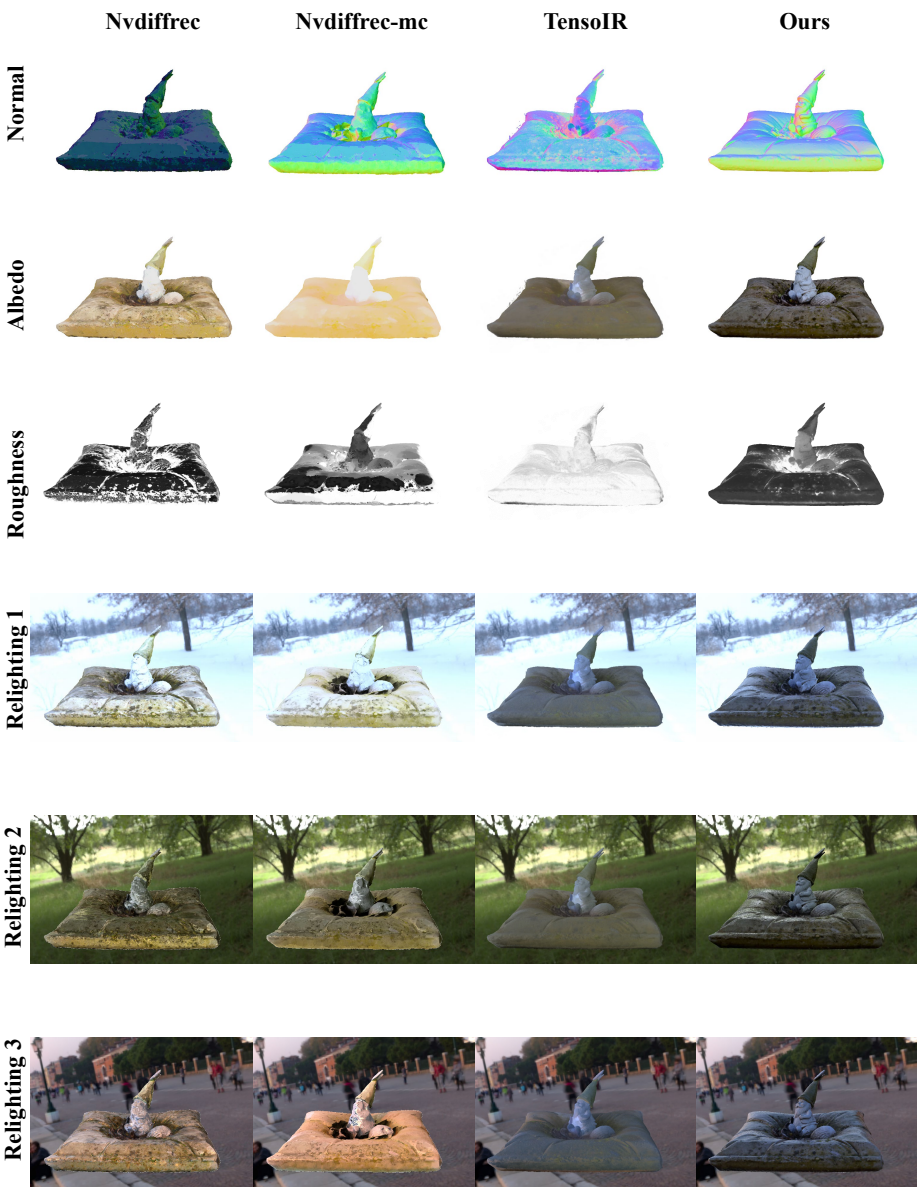

Fig. 11: Qualitative comparison on *Gnome* from the NeRD dataset.

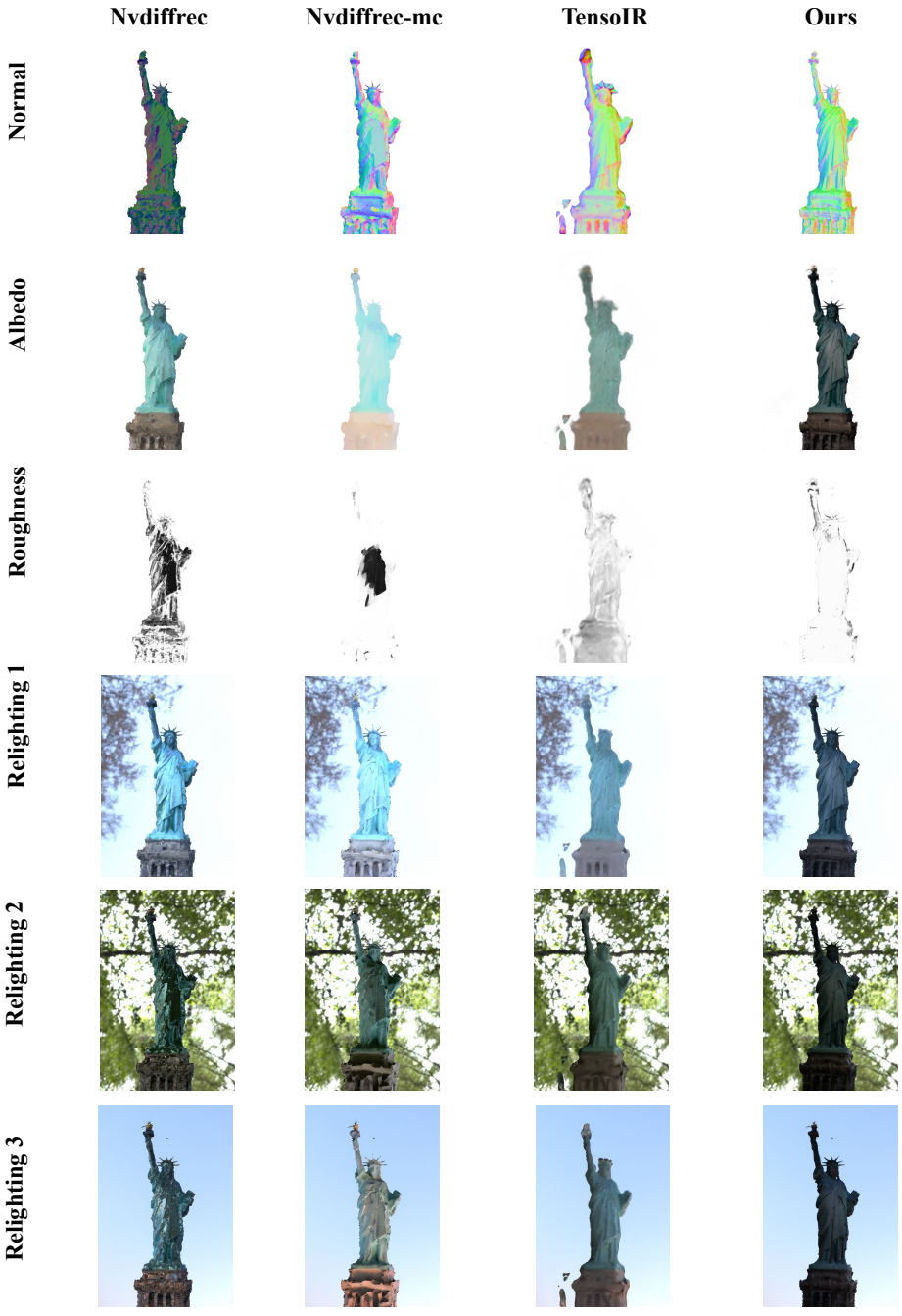

Fig. 12: Qualitative comparisons on *StateOfLiberaty* from the NeRD dataset.

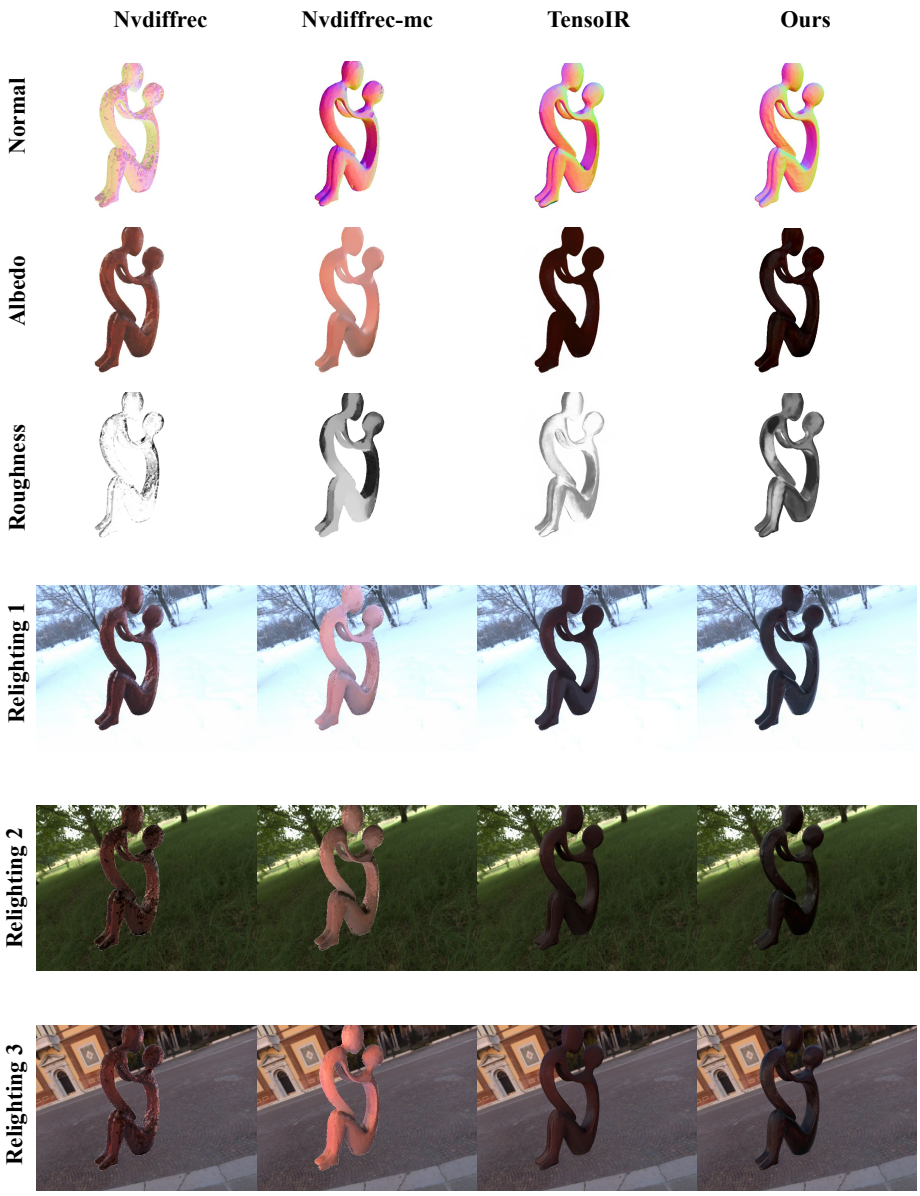

Fig. 13: Qualitative comparisons on *MotherChild* from the NeRD dataset.

References

1. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: ICCV (2021)
2. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.: Nerd: Neural reflectance decomposition from image collections. In: ICCV (2021)
3. Jin, H., Liu, I., Xu, P., Zhang, X., Han, S., Bi, S., Zhou, X., Xu, Z., Su, H.: Tensoir: Tensorial inverse rendering. arXiv preprint arXiv:2304.12461 (2023)
4. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. In: CVPR (2022)
5. Wu, T., Wang, J., Pan, X., Xu, X., Theobalt, C., Liu, Z., Lin, D.: Voxurf: Voxel-based efficient and accurate neural surface reconstruction. arXiv preprint arXiv:2208.12697 (2022)
6. Yao, Y., Zhang, J., Liu, J., Qu, Y., Fang, T., McKinnon, D., Tsin, Y., Quan, L.: Neilf: Neural incident light field for physically-based material estimation. In: ECCV (2022)
7. Zhang, Y., Sun, J., He, X., Fu, H., Jia, R., Zhou, X.: Modeling indirect illumination for inverse rendering. In: CVPR (2022)