4 December 2024, Preprint v1 · CC-BY 4.0 Qeios PREPRINT

Research Article

MLI-NeRF: Multi-Light Intrinsic-Aware Neural Radiance Fields

Yixiong Yang[1], Shilin Hu[2], Haoyu Wu[2], Ramon Baldrich[1], Dimitris Samaras[2], Maria Vanrell[1]

1. Universitat Autónoma de Barcelona, Spain; 2. Independent researcher

Current methods for extracting intrinsic image components, such as reflectance and shading, primarily rely on statistical priors. These methods focus mainly on simple synthetic scenes and isolated objects and struggle to perform well on challenging real-world data. To address this issue, we propose MLI-NeRF, which integrates Multiple Light information in Intrinsic-aware Neural Radiance Fields. By leveraging scene information provided by different light source positions complementing the multi-view information, we generate pseudo-label images for reflectance and shading to guide intrinsic image decomposition without the need for ground truth data. Our method introduces straightforward supervision for intrinsic component separation and ensures robustness across diverse scene types. We validate our approach on both synthetic and real-world datasets, outperforming existing state-of-the-art methods. Additionally, we demonstrate its applicability to various image editing tasks. The code and data are publicly available at https://github.com/liulisixin/MLI-NeRF.

Corresponding authors: Yixiong Yang, yixiong@cvc.uab.cat; Shilin Hu, shilhu@cs.stonybrook.edu; Haoyu Wu, haoyuwu@cs.stonybrook.edu; Ramon Baldrich, ramon@cvc.uab.cat; Dimitris Samaras, samaras@cs.stonybrook.edu; Maria Vanrell, maria@cvc.uab.cat

doi.org/10.32388/OQFE20

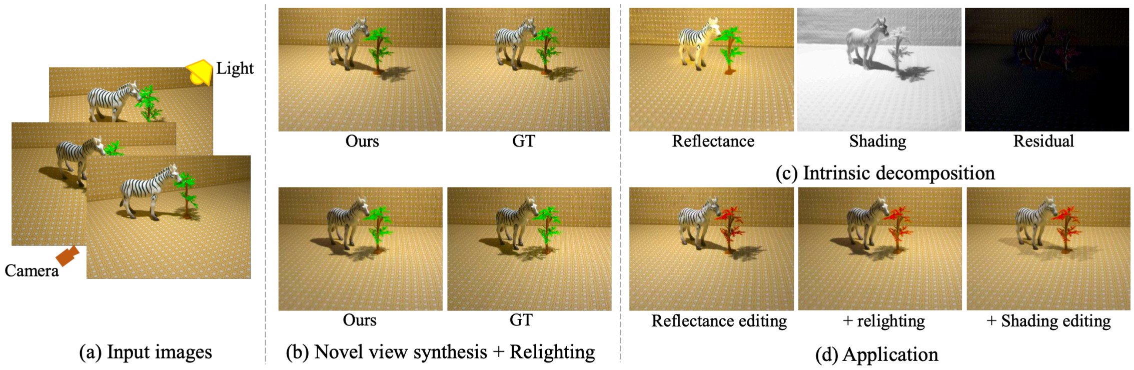

Figure 1. Given real-world images from ReNe dataset[1] (a), our method learns the neural radiance fields that enable novel view synthesis and relighting (b), and intrinsic decomposition (c) simultaneously. Image editing applications (d) can also be employed, such as reflectance editing, reflectance editing plus relighting or shading editing (simulating two lights).

1. Introduction

Neural radiance fields (NeRF) have enabled significant strides in novel view synthesis (NVS) [2][3] including efforts towards scene editing [4], such as recoloring [5] and relighting [6][7]. Scene editing becomes easier when the scene can be decomposed into editable sub-attributes. There are two related approaches to scene editing [8]: inverse rendering and intrinsic image decomposition.

The first approach [9][10][11][12] integrates inverse rendering with neural rendering methods for scene decomposition. They often employ the BRDF [13] to model material properties and jointly optimize geometry, materials, and environmental lighting. However, inverse rendering presents a highly ill-posed challenge: separating material properties and illumination in images often yields ambiguous results, and tracing light within scenes is computationally intensive. These factors limit inverse rendering to object-specific scenarios. The second approach [5][14][15], based on intrinsic image decomposition [16], aims to provide an explainable representation of a scene in terms of components such as reflectance and shading. In general, intrinsic image decomposition is more applicable to a broader range of scenarios, including individual objects and more complex scenes with backgrounds. While IntrinsicNeRF [5] has pioneered the integration of intrinsic decomposition within NeRF, it has not fully leveraged the 3D information available through neural rendering.

Mineralogists illuminate their specimens from different angles to reveal their features. Similarly, varying the light source position is essential for uncovering the intrinsic details of a scene. We aim to enhance intrinsic decomposition quality and expand scene editing capabilities by leveraging multiple light sources to build an intrinsic-aware NeRF. The connection between varying lighting conditions and intrinsic decomposition has been discussed for 2D images[17][18], but not yet in neural rendering, even though there is interest in relighting using neural rendering[7][19].

In this paper, we introduce **MLI-NeRF**, a two-stage method to learn an intrinsic-aware NeRF. In Stage 1, we extend NeRF to incorporate light position information and learn a relightable scene using images captured from various camera angles and light source positions. In the subsequent post-processing, we begin by obtaining normals and light visibility maps for images under multiple lighting conditions using the model from Stage 1 and sphere tracing[20]. We then generate pseudo shadings from the normals, light rays, and light visibility maps. Finally, for each camera pose, pseudo reflectance is generated by combining cues from multiple lighting conditions. In Stage 2, we make our model intrinsic-aware by introducing additional modules for reflectance and shading while restricting light position input to the shading module only, ensuring the independence of the reflectance and light. In this paper, we forego potentially oversimplifying statistical constraints on various illumination-related factors in recent work[5][18][14][21] to instead use the physics-based disentanglement of reflectance and shading and achieve high-quality results.

As illustrated in Fig. 1, our method achieves high-quality intrinsic decomposition results (Fig. 1(c)), as well as NVS and relighting results (Fig. 1(b)). It also enables applications such as reflectance editing, relighting, and shading editing (Fig. 1(d)). Furthermore, our method is applicable across various datasets, including the object-only synthetic NeRF[2] dataset, the real object dataset[22][7], and the ReNe[1] dataset with real-world full scenes. Our contributions are summarized as follows:

- A novel intrinsic-aware NeRF model that integrates multiple light information, enabling applications such as NVS, lighting modification, and scene editing.
- A method that separates intrinsic components by using supervision from generated pseudo intrinsic images. We introduce straightforward physics-based constraints to eliminate the need for statistical priors required by traditional approaches. Our method ensures robustness across various scene types.
- Experimental results across three different datasets demonstrate our method's superior performance in intrinsic decomposition compared to existing state-of-the-art methods, showing advancements

not only in synthetic object-only scenes but also in challenging real scenes with backgrounds and cast shadows.

2. Related Works

Intrinsic decomposition.

Intrinsic decomposition is a classical challenge in computer vision[16], with much of the previous research focused on the 2D image[15][14][23][24]. A key difficulty in this area is the scarcity of real datasets, which need complicated and extensive annotation. This limitation has spurred interest in semi-supervised and unsupervised techniques[17][18][25]. IntrinsicNeRF[5] has been a pioneer in applying intrinsic decomposition to 3D neural rendering. Similar to previous unsupervised methods in 2D, it utilizes statistical priors, including chromaticity and semantic constraints, for guidance. However, these constraints do not accurately reflect physical principles and often fall short in complex scenarios. Our approach leans on 3D information and physical constraints (e.g., variations in illumination) to achieve superior results.

Relighting.

Relighting has recently garnered attention from various perspectives within the field[26]. Data-driven approaches have been explored, with research focusing on portrait scenes[27][28][29][30][31] and extending to more complex scenarios[32][33][34][35][36]. Kocsis et al.[37] have also investigated lighting control within diffusion models, enabling the generation of scenes under varying lighting conditions. Meanwhile, relighting has also received widespread attention within the field of neural rendering[38][22][7][1], achieving impressive relighting outcomes within individual scenes. Among them, Toschi et al. [1] proposed the ReNe dataset, which consists of images captured with various cameras and light poses under controlled lab conditions. Zeng et al.[7] enhanced NeRF relighting with visibility and specular hints. Chang et al.[39] proposes a method of outdoor scene relighting, and they use locations and time to collect the direction of sunlight as input. However, the potential of using information from multiple lights for 3D scene understanding remains unexplored.

Inverse rendering.

Inverse rendering[8] is an alternative approach to recovering the fundamental properties of a scene, which aims at extracting the geometry materials, and lighting of a 3D scene. Recently, the study of inverse rendering methods based on NeRF has become a popular topic. NeRFactor[12] introduced a method to improve the geometric quality of NeRF and incorporated a data-driven BRDF prior. More methods have been developed to address the scenes with different light conditions, including fixed illumination conditions[40], varying light sources[41][42]. Invrender[10] proposed a method for predicting indirect light. L-Tracing[20] introduced an efficient algorithm for estimating visibility without training. SIRe-IR[11] introduced a method for high-illuminance scenes, addressing the issue where previous methods struggled under prominent cast shadows. Liu et al.[43] propose the OpenIllumination dataset with multi-illumination, which focuses on inverse rendering evaluation on real objects. However, inverse rendering methods are primarily based on individual objects and are challenging to extend to large, complex scenes, such as those with backgrounds.

3. Method

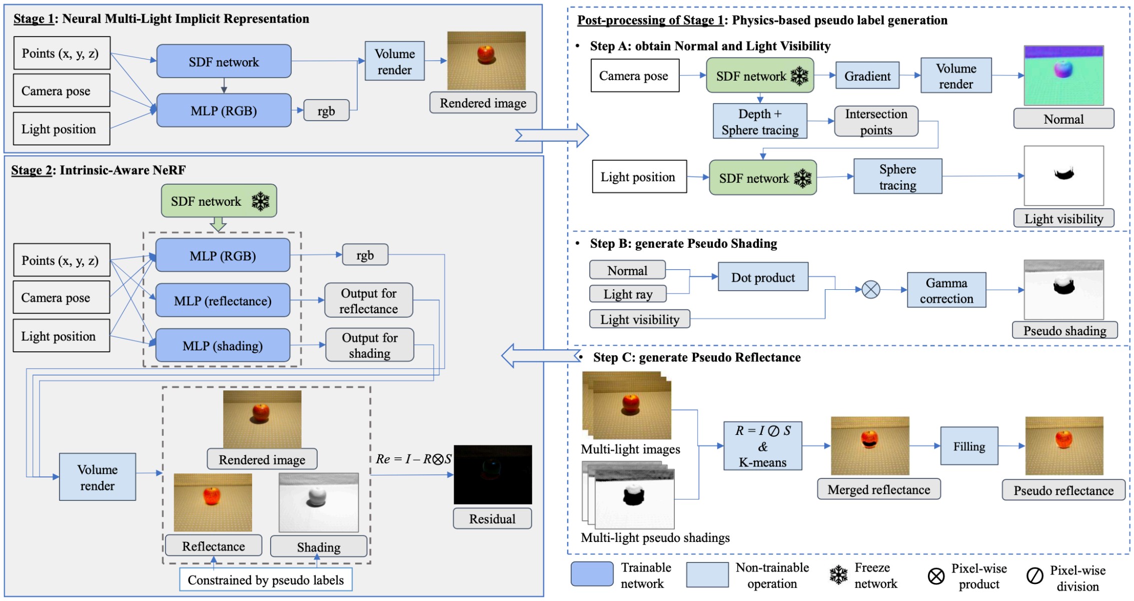

Figure 2. Illustration of the Framework. In Stage 1, we introduce light position as input to extend NeRF for multi-light implicit representation (top left). Following Stage 1, three post-processing steps are applied to generate pseudo labels for reflectance and shading using the proposed physics-based pipeline (right). In Stage 2, we train the intrinsic-aware NeRF based on the model from Stage 1 and the pseudo labels from post-processing (bottom left).

We propose a two-stage method, with the overall framework illustrated in Fig. 2. In previous inverse rendering methods[10], accounting for indirect illumination and shadow has been a critical challenge. Our strategy is to leverage multiple lighting conditions to make the corresponding information more accessible. In the first stage, we train our model to represent scenes under varying camera positions and lighting conditions, enabling NVS and lighting modification. In the subsequent post-processing, we use the model from Stage 1 to generate pseudo intrinsic images of reflectance and shading for each image. Specifically, pseudo shadings are derived from normals and light visibility, while pseudo reflectance is obtained by dividing the image by the shading. Leveraging multiple lighting conditions allows us to effectively gather more detailed information about the scene, thus resulting in high-fidelity pseudo labels. In Stage 2, we retrain the network with added modules for predicting reflectance and shading, restricting light information input solely to the shading module. Pseudo images guide the separation of intrinsic components in this phase.

3.1. Preliminaries

Some traditional methods decompose images into reflectance and shading[16][14][44], primarily targeting diffuse components. Since the real world contains many non-diffuse effects, recent approaches[8][5] add a residual term to account for discrepancies. We follow this setup and model intrinsic decomposition as follows:

$$I(i,j) = R(i,j) \otimes S(i,j) + Re(i,j) \tag{1}$$

where R, S and Re denote Reflectance, Shading and Residual, respectively. Additionally, following the Lambert's cosine law, a shading can be computed using the following formula:

$$S = \vec{N} \cdot \vec{L} \tag{2}$$

where $\vec{N}$ is the normal ray and $\vec{L}$ is the light ray. This physical illumination model guides our following pseudo shading generation and intrinsic-aware NeRF training.

3.2. Stage 1: Neural Multi-Light Implicit Representation

We follow the structure of Neuralangelo[3], which achieves promising results in both small and large scenes. It uses 3D hashing encoding combined with Signed Distance Function (SDF)[45] to represent the implicit geometry, and then employs an MLP to model the color information. The original Neuralangelo does not support light position as an input, so, besides the original MLP input, we incorporate the light position encoded with spherical encoding as an additional input. We illustrate Stage 1 in Fig. 2 (top-left), with formulas as follows:

$$sdf = f(\mathbf{x}),\ \mathbf{c} = \mathrm{MLP}_{color}(\mathbf{x}, \mathbf{n}, \mathbf{feat}, \mathbf{d}, \mathbf{l}) \tag{3}$$

where $f(\cdot)$ is the geometry network that predicts SDF and $\mathrm{MLP}_{color}(\cdot)$ is the color network. $\mathbf{x}$ is the spatial position, $\mathbf{n}$ and $\mathbf{feat}$ are the normal and the features from the SDF network, $\mathbf{d}$ is the view direction, and $\mathbf{l}$ is the light position. Following[3], the loss for Stage 1 is:

$$\mathcal{L}_{S1} = w_{rgb}\mathcal{L}_{rgb} + w_{eik}\mathcal{L}_{eik} + w_{curv}\mathcal{L}_{curv} \tag{4}$$

where $\mathcal{L}_{rgb}$ is the loss of the rendered image, $\mathcal{L}_{\text{eik}}$ represents the Eikonal loss[46], and $\mathcal{L}_{\text{curv}}$ is the curvature loss[3]. The terms w_{eik} and w_{curv} are the corresponding weights.

3.3. Post-processing: Physics-based Pseudo Label Generation

Here we propose a post-processing that generate pseudo labels for reflectance and shading in three steps, as illustrated in Fig. 2 (right). It starts with generating pseudo shading from the normal and the light visibility. We then generate pseudo reflectance using multi-light shadings and images.

Step A. We derive the normal from the gradient of the SDF network. The geometry network also provides depth information which is used to estimate the intersection points in conjunction with sphere tracing[20]. Light visibility, denoted as V, which indicates whether a point is directly illuminated, is obtained by sphere tracing based on the light position and intersection points.

Step B. Since the Eq. (2) does not consider occlusion and other effects, the generation of pseudo-shading in our implementation follows the formula:

$$S^* = (\max(\vec{N} \cdot \vec{L}, 0) \otimes V)^\gamma \tag{5}$$

where $(\cdot)^\gamma$ represents gamma correction, and $\otimes$ denotes pixel-wise product. This gamma correction is essential to adapt to the nonlinear representation of digital images. All image sensors introduce a gamma correction to accommodate the light integration to the nonlinear perception of brightness in the human eye. Since our pseudo-shading is directly created from the 3D world, we need to introduce the sensor representation.

Step C. This step entails inferring high-fidelity pseudo reflectance from multiple pseudo shading. We use the equation $R = I \oslash S$ as a simplified version of Eq. (1) ignoring at this point the residual component that mainly entangles specular and undirected light components.

Our method in this step leverages the trained model from Stage 1 to generate multiple images under different direct light conditions, each accompanied by its corresponding pseudo shading. The flowchart is shown at the right bottom in Fig. 2, and Fig. 3 illustrates the images during the calculation. By calculating $R = I \oslash S$, we can compile the reflectance information for every pixel from multiple light conditions while diminishing the influence of the entangled indirect light. To generate a unique reflectance from multiple pseudo shading, we use the K-means algorithm[47] at the pixel level, incorporating the weights of each pseudo shading to select the most probable pseudo reflectance. This approach allows us to achieve a merged reflectance under varied lighting conditions.

However, some regions within the merged reflectance may appear as holes due to the absence of direct illumination in all lighting conditions. We address these areas with a filling strategy. We compute a weighted distance between hole pixels and non-hole pixels, considering their spatial distance in the image, the angular difference of their normals, and the color difference in the RGB image. Then, we assign the color of the nearest non-hole pixel, based on this weighted distance, to the hole pixel. In fact, our filling strategy is a rough supplementary approach with minimal impact on the overall results. Under the multi-light setting, the hole regions are typically small. Additionally, we reduced the weight of the hole regions during subsequent training.

After our proposed three-step post-processing, we achieve the final pseudo reflectance, as shown in Fig. 3. Our pseudo reflectance offers a straightforward basis for disentanglement and has proven to be more reliable in guiding the following training stage compared to the previous statistical priors. Additionally, we compute weight maps W_R and W_S for both pseudo reflectance and pseudo shading based on the edges of pseudo shading and visibility. Areas with higher pseudo shading values, or those further from visibility edges (where visibility calculations may be prone to errors), exhibit greater credibility in their pseudo labels; conversely, areas closer to visibility edges or with lower pseudo shading values are deemed less reliable.

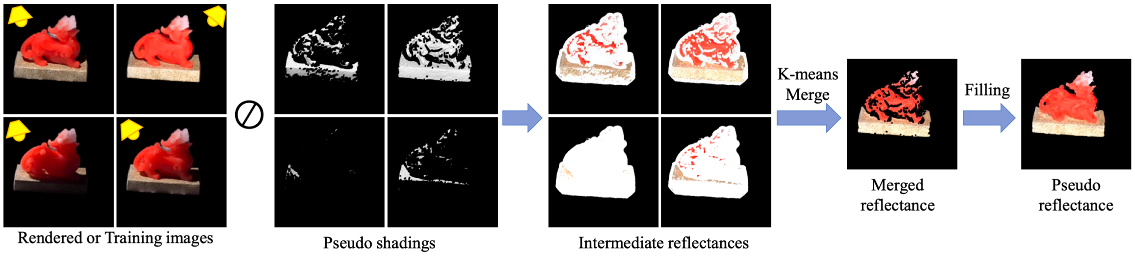

Figure 3. Illustration of the pseudo reflectance generation process in the post-processing.

3.4. Stage 2: Intrinsic-Aware NeRF

As illustrated in Fig. 2 (bottom-left), we use high-fidelity pseudo intrinsic images to guide the intrinsic decomposition learning. Expanding the model from Stage 1, we add two extra MLPs dedicated to generating reflectance and shading outputs, while the geometry network is frozen. Compared to the MLP_{color} in Eq. (3), these two MLPs receive different inputs. Since reflectance is independent of lighting, the $\mathrm{MLP}_{reflectance}$ does not take the light position as input. Additionally, both reflectance and shading are diffuse components, and a diffuse reflecting surface exhibits Lambertian reflection, indicating that it maintains equal luminance when observed from any direction. Therefore, view poses are also excluded from both of their inputs. The formulas are as follows:

$$\begin{aligned}\mathbf{r} &= \mathrm{MLP}_{reflectance}(\mathbf{x}, \mathbf{n}, \mathbf{feat}) \\ \mathbf{s} &= \mathrm{MLP}_{shading}(\mathbf{x}, \mathbf{n}, \mathbf{feat}, \mathbf{l})\end{aligned} \tag{6}$$

where $\mathbf{x}$ is the spatial position, $\mathbf{n}$ and $\mathbf{feat}$ are the normal and the features from the SDF network, and $\mathbf{l}$ is the light position.

After volume rendering, we obtain RGB images, along with reflectance and shading. Subsequently, the residual is derived from Eq. 1. During training, the pseudo labels impose constraints on reflectance and shading:

$$L_{intrinsic} = W_R \cdot \|\hat{R} - R^*\|_1 + W_S \cdot \|\hat{S} - S^*\|_1 \tag{7}$$

where $\hat{R}$ and $\hat{S}$ represent the predicted reflectance and shading, respectively, and R^* and S^* are their corresponding pseudo labels. W_R and W_S represent weight maps for reflectance and shading, derived during pseudo label generation. As demonstrated in[5], the diffuse components dominate the scene, so it is crucial to prevent the training from converging to undesirable local minima (*e.g.* $R = 0, S = 0, Re = I$). Therefore, we introduce a regularization term for Re to ensure that the image is primarily recovered through R and S: $L_{reg} = \|\hat{Re}\|_1$.

Finally, the Stage 2 loss is the weighted sum of:

$$\mathcal{L}_{S2} = w_{\mathrm{rgb}}\mathcal{L}_{rgb} + w_{\mathrm{intrinsic}}L_{intrinsic} + w_{\mathrm{reg}}L_{reg} \tag{8}$$

where w_{rgb}, $w_{\mathrm{intrinsic}}$ and w_{reg} are the corresponding weights.

4. Experiments

Method	Light Setting	Reflectance				Shading			
		PSN↑	SSIM↑	LPIP↓	MSE↓	PSNR↑	SSIM↑	LPIPS↓	MSE↓
InvRender	Single	16.59	0.8228	0.1807	0.0271	—	—	—	—
TensoIR	Single	18.50	0.8518	0.1544	0.0213	—	—	—	—
IntrinsicNeRF	Single	18.02	0.8353	0.2142	0.0226	19.02	0.8660	0.1476	0.0168
Ours	Single	22.44	0.9012	0.0950	0.0072	**24.39**	**0.9188**	**0.0843**	**0.0049**
TensoIR	Multiple	22.70	0.8800	0.1450	0.0087	—	—	—	—
Ours	Multiple	24.40	0.9357	0.0582	0.0051	23.44	0.9225	0.0798	0.0055
PIE-Net	Random	19.59	0.8708	0.1298	0.0153	20.03	0.8868	0.1653	0.0133
Ordinal	Random	18.14	0.8716	0.1251	0.0191	16.80	0.8717	0.1679	0.0293
Ours	Random	**25.48**	**0.9420**	**0.0515**	**0.0041**	**22.90**	**0.9024**	**0.1049**	**0.0072**

Table 1. Quantitative results of the intrinsic decomposition on the **Synthetic Dataset**. We compare different methods under three light settings: single, multiple, and random. Our method outperforms other methods in all settings. **Best** results are marked in **bold**, second best results are underlined.

We quantitatively and qualitatively validate our method and compare it with other approaches, including traditional learning-based intrinsic decomposition methods and neural rendering methods. The comparison encompasses intrinsic decomposition and NVS with relighting. For data-driven methods, we select PIE-Net[15] and Careaga et al.[14] (hereafter referred to as Ordinal). For NeRF-related methods, we choose InvRender[10], TensoIR[9], and IntrinsicNeRF[5]. Among these, InvRender and TensoIR are inverse rendering methods, while IntrinsicNeRF is an intrinsic decomposition method. Additionally, we select NRHints[7], a method focused on relighting, to validate our relighting performance.

4.1. Datasets

Our experiments are conducted on three datasets, each containing four scenes.

Synthetic Dataset.

It contains synthetic data based on Blender, inspired by the synthetic dataset designed by NRHints[7] for relighting. Some scenes in this dataset are derived from NeRF[2]. Since the original dataset does not include ground truth (GT) for intrinsic decomposition, we re-render the data in Blender and export these quantities. Each scene comprises 500 images for training, 100 for validation, and 100 for testing, including intrinsic components for each image. The selection of light positions and camera poses follows the setup described by Zeng et al. [7], with both distributed across a hemispherical space above the scene. The light position and camera pose for each image are randomly and independently selected.

Real Object Dataset.

It includes real objects from the object relighting dataset collected by [22][7], featuring various viewpoints and lighting conditions. The training set size varies from 500 to 2000 images per scene.

ReNe Dataset.

ReNe dataset[1], unlike most datasets used in previous inverse rendering research, contains complete real-world scenes with backgrounds. Notably, the camera poses and light positions in this dataset are concentrated in a specific direction rather than evenly distributed across 360 degrees, which poses challenges for both reconstruction and intrinsic decomposition. This real dataset features 2000 images across scenes, captured from 50 viewpoints under 40 lighting conditions, with lighting and camera poses grid-sampled. Following the dataset split, we use 1628 images (44 camera poses $\times$ 37 light positions) for training. Since the test set is not publicly available, we use the validation set for inference.

4.2. Lighting and Camera Views Settings

In this section, we further clarify the lighting and camera view setups in our experiments.

Grid-sampled or non-grid-sampled.

In the ReNe dataset, camera views and light positions are grid-sampled. In contrast, the Synthetic and Real Object Datasets use independently sampled views and lights. We denote this as non-grid-sampled.

This difference does not affect Stage 1 training but influences the post-processing of Stage 1 (Fig. 2), where we merge results to generate pseudo reflectance.For grid-sampled setups, all light positions are used. For non-grid-sampled setups, because the number of possible light positions is large, four light positions are randomly selected, which our experiments indicate is sufficient.

Additional Lighting Setups for Better Comparison.

Previous methods often struggle with varying lighting conditions. InvRender[10] and IntrinsicNeRF[5] operate under single lighting conditions, while TensoIR[9] supports multiple lights but recommends around four, which cannot handle hundreds of light positions. We introduce two additional settings to better compare intrinsic decomposition methods and examine the impact of different light positions: **single light** (one fixed position) and **multiple lights** (four fixed positions). TensoIR occasionally failed with four lights, so we used three instead in this case.

For the Synthetic Dataset, the original setting is referred to as **random lights**. We re-render scenes in Blender for the single and multiple light tracks, maintaining the same image counts and random seed for camera views. For the ReNe Dataset, the original setting is denoted as **all lights**. We selected 1 light condition as the single light track and 4 light conditions as the multiple lights track, thereby the training sets respectively contain 44×1 and 44×4 images. Notably, the single light setup serves as an extreme condition, diverging from our intent of utilizing multi-light information. This track was designed for fair comparison with other methods and to highlight the importance of multi-light information in intrinsic decomposition.

4.3. Implementation Details

Our model's hyperparameters include a batch size of 2048 and each stage is trained for 500k iterations. We implement the model in PyTorch and use the AdamW[48] optimizer with a learning rate of $1e^{-3}$ for optimization. The experiments can be conducted on a single Nvidia RTX 3090. The training time of Stage 1 follows the training time of Neuralangelo[3], and Stage 2 takes 19 hours. The weights of losses, w_{eik}, w_{curv}, $w_{\text{intrinsic}}$, w_{reg} are set to 0.1, $5e^{-4}$, 1.0, and 1.0, respectively. To evaluate the comparison between predicted images and GT, we employ the following metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM)[49], Learned Perceptual Image Patch Similarity (LPIPS)[50] and Mean Squared Error (MSE).

4.4. Results on the Synthetic Dataset

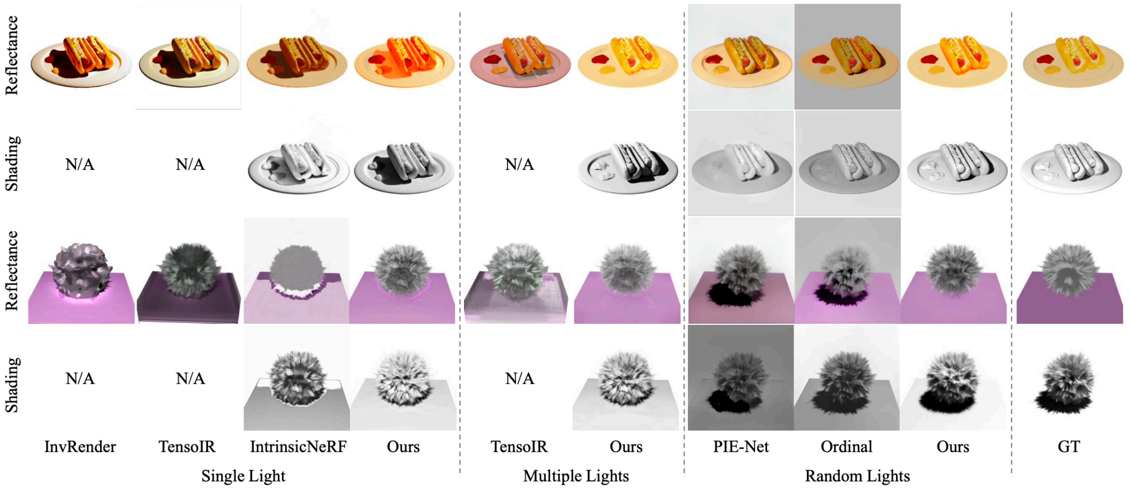

Figure 4. Qualitative Results on the **Synthetic Dataset** with all settings. The same GT reflectance applies across all settings, but GT shading differs due to varying light positions. Here for brevity, only the shading under the Random Lights setting is shown. Compared to other methods, our approach predicts the best reflectance and effectively handles cast shadows.

Comparison with SOTA.

Table 1 presents the quantitative comparison of intrinsic decomposition on all methods and light settings. Since the sequence of camera poses is the same across all settings when generating the dataset, the reflectance GT remains consistent, allowing for direct comparison. Our method outperforms existing methods in predicting reflectance for each light setting, with the best results under the random setting. Due to the different shading GTs, the comparisons are conducted separately. Our method also achieves the best results across all settings. A qualitative comparison is shown in Fig. 4. The single light setting demonstrates that under challenging lighting conditions, such as those with pronounced cast shadows, all methods struggle to achieve consistently good results. This indicates that using multi-light information is a practical approach. However, our method still achieves the most meaningful results in the single light setting of the FurBall scene. In the other two settings, our approach also demonstrates significantly better performance, producing high-quality reflectance and shading results, effectively handling cast shadows.

doi.org/10.32388/OQFE20

Comparison across different light settings.

As shown in Table 1 and Fig. 4, both TensoIR and our method exhibit improved performance as the number of light positions increases. This observation aligns with the findings mentioned in the TensoIR[9]. Additionally, the results in Fig. 4 demonstrate that our method's performance under the multiple light setting (4 lights) is very close to that of the random light setting.

NVS and relighting analysis.

Table 2 shows the quantitative results of NVS and relighting. Since the lights are fixed and known under the Single and Multiple light settings, the results pertain only to NVS, where our method achieves the best performance. Under the Random setting, the results reflect both NVS and relighting, where our method shows comparable performance to NRHints[7].

	Light Setting	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**MSE ↓**
InvRender	Single	23.99	0.8760	0.1109	0.0051
TensoIR	Single	35.00	0.9761	0.0343	0.0019
IntrinsicNeRF	Single	34.53	0.9794	0.0137	0.0005
Ours	Single	**37.51**	**0.9878**	**0.0099**	**0.0003**
TensoIR	Multiple	33.19	0.9645	0.0503	0.0022
Ours	Multiple	**36.12**	**0.9836**	**0.0135**	**0.0003**
NRHints	Random	**32.79**	**0.9674**	0.0369	**0.0007**
Ours	Random	31.20	0.9619	**0.0354**	0.0010

Table 2. Quantitative results of the NVS and relighting on the Synthetic Dataset.

4.5. Results on the Real Object Dataset

Results on the Real Object Dataset allow us to confirm a quite accurate qualitative performance of our intrinsic decomposition. In Fig. 5(a) we show a comparison of our method versus SOTA. Our RGB-rendered images are similar to NRHInts[7] and our estimations for Reflectance and Shading can be compared to PIE-Net[15] and Ordinal[14]. In particular, we want to highlight our estimation of reflectance that does not present any remaining shading effect with respect to the rest of the methods, restoring the vibrant colors of the objects and achieving significantly better results. In Fig. 5(b) we add two more examples where we zoom out two details. At the top is a hole through which our method estimates the texture of the support surface, and at the bottom, it gets a proper reflectance from a strong cast shadow. Our shading presents a better disentangling from reflectance than the rest of the methods. Although we can show a clear overall improvement, some limitations can still be observed from some difficult areas that remain hidden from any camera point of view and light source, as we can see in the unremoved shadow we have under the object's tail of the top zoomed window.

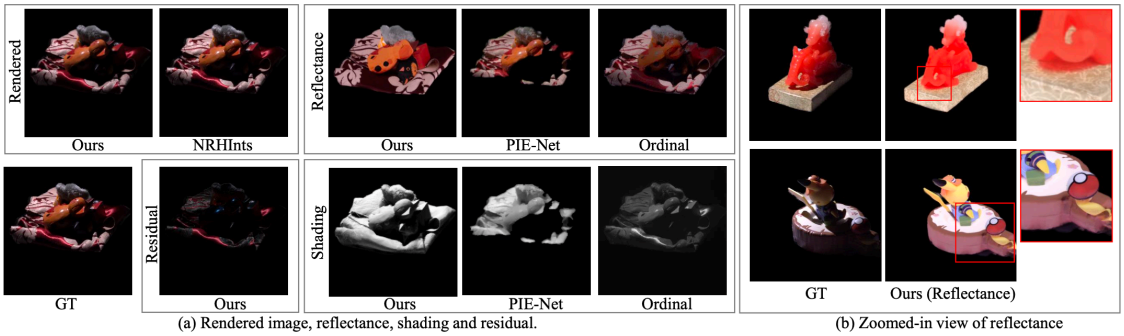

Figure 5. Qualitative Results on the **Real Object Dataset**. (a) Our method compared with NRHints for the rendered image, and PIE-Net and Ordinal for intrinsic decomposition. (b) Our reflectance estimation for two different scenes, with zoomed-in views on the object hole and cast shadow area.

4.6. Results on the ReNe Dataset

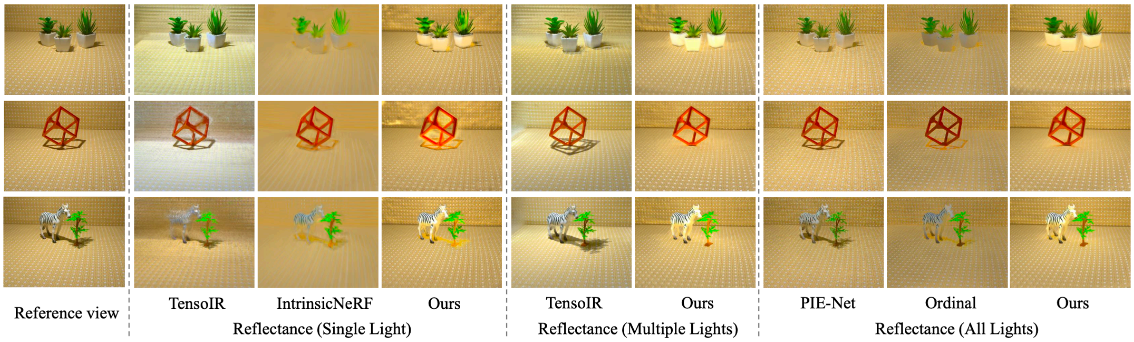

Figure 6. Qualitative Results on the ReNe Dataset. We show the reflectance estimation for a reference view across all settings.

The ReNe dataset presents a significant challenge for previous methods due to its full scenes with backgrounds and camera poses that are mostly concentrated in a specific area rather than being distributed 360 degrees around the scene. The quantitative results of NVS in Table 3 also demonstrate that our method gets the best results in all settings across all metrics.

	Light Setting	PSNR↑	SSIM↑	LPIPS↓	MSE ↓
TensoIR	Single	24.57	0.6063	0.1912	0.0040
IntrinsicNeRF	Single	24.10	0.4009	0.5392	0.0039
Ours	Single	**27.63**	**0.7210**	**0.1136**	**0.0019**
TensoIR	Multiple	26.44	0.6743	0.1572	0.0028
Ours	Multiple	**27.71**	**0.7377**	**0.1078**	**0.0018**
Ours	All	27.48	0.7300	0.1113	0.0019

Table 3. Quantitative results of the NVS and relighting on the ReNe Dataset.

The results from the single light setting (left of Fig. 6) show that all methods struggle to achieve good performance; however, our method produces meaningful results in the first and third scenes. The other neural rendering methods, TensoIR[9] and IntrinsicNeRF[5], fail to achieve correct decomposition, primarily attributed to the failure in distinguishing intrinsic components and also the difficulty in scene reconstruction. In the multiple light setting(middle of Fig. 6), TensoIR incorrectly leaks multiple shading components into the reflectance. In contrast, our method successfully combines multiple lighting conditions to produce clean reflectance. In the all lights setting, our method achieves superior results, outperforming both PIE-Net[15] and Ordinal[14], achieving sharp texture edges, vibrant colors, accurate shadow elimination, and precise background reconstruction. For more results, please refer to the supplementary material.

5. Conclusion

We propose MLI-NeRF, a Multi-light Intrinsic-aware Neural Radiance Field. MLI-NeRF generates pseudo intrinsic images for scenes under different lighting conditions, enabling the learning of intrinsic decomposition without intrinsic GT. Our experiments demonstrate that our method achieves excellent performance across different types of scenes. Our approach relies on fundamental physical principles, highlighting its potential applicability in more complex scenes. Our method enables the simultaneous synthesis of novel views, relighting, and intrinsic decomposition, providing a versatile tool for various editing applications, such as reflectance and shading modification.

Limitations and Future Work

One limitation of our method is the need to know the light positions. As noted in our related works, many methods have proposed ways to obtain light pose information when capturing custom data. These include recording the camera pose fixed with the light source[6][7] and using the time of day to determine the sun's angle in outdoor settings[39]. Another limitation is the computational efficiency: training the Stage 1 model based on Neuralangelo[3] takes around 40 hours, with Stage 2 retraining adding another 19 hours. In fact, our method is applicable to different similar baselines, and we plan to migrate to a more efficient base architecture in the future.

Supplementary Material

In this supplementary material, we present the following:

1. More results on the Synthetic Dataset.

2. More results on the Real Object Dataset.
3. More results on the ReNe Dataset.

We have also submitted a supplementary video to showcase the results of our method on all datasets.

More results on the Synthetic Dataset

From Fig. 7 to Fig. 10, we present additional qualitative results with all settings including Single Light, Multiple Lights, and Random Lights on the Synthetic Dataset.

In the Synthetic Dataset, we conduct experiments on four scenes: Hotdog, FurBall, Drums, and Lego. Across all scenes, we observe that under the original Random Lights setting, our method achieves best results, closely matching the GT. The predicted reflectance and shading are significantly better than those of PIE-Net[15] and Ordinal[14], and the relighting results are comparable to NRHints[7].

Under the Single Light and Multiple Light settings, we compare our method with additional approaches. In the Single Light setting, all methods struggle to consistently produce good intrinsic images, particularly in separating cast shadows. However, our method successfully removes most cast shadows in the predicted reflectance for scenes like FurBall, Drums, and Lego, outperforming other methods. This underscores the importance of multi-light information for intrinsic decomposition.

In the Multiple Lights setting (with four different light positions), our method achieves satisfactory results, very close to the GT and the Random Lights setting, demonstrating that four different light positions are sufficient for our method to perform well. Under the same setting, TensoIR[9] fails to produce satisfactory results, with residual cast shadows mixing into the reflectance.

Overall, since the camera view used in comparisons is the same, the reflectance should have the same GT across all settings. Our predicted reflectance consistently outperforms others in comparisons within the same rows, while also showing improvements as the number of light positions increases.

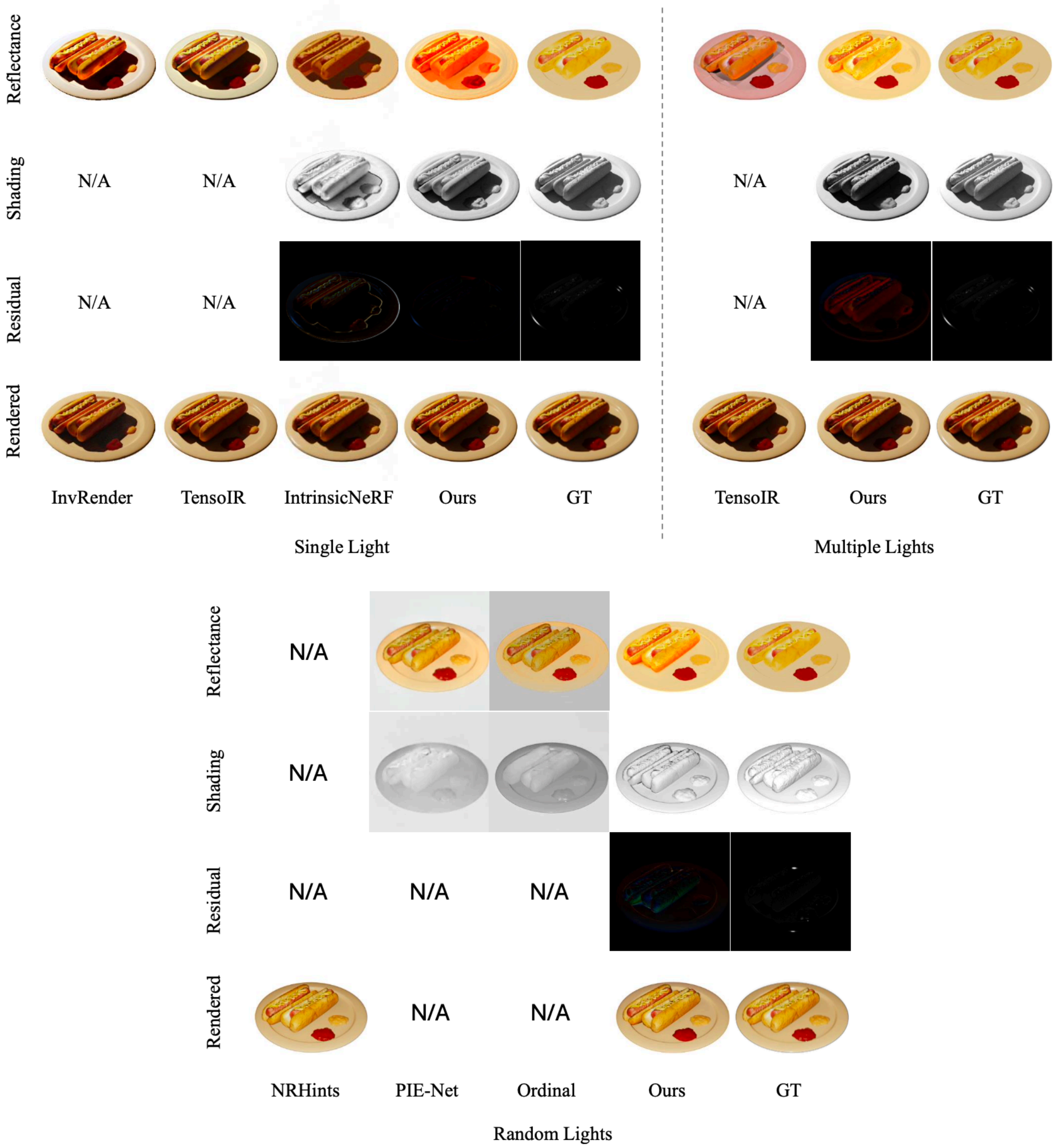

Figure 7. Additional Qualitative Results on the Synthetic Dataset. (Hotdog)

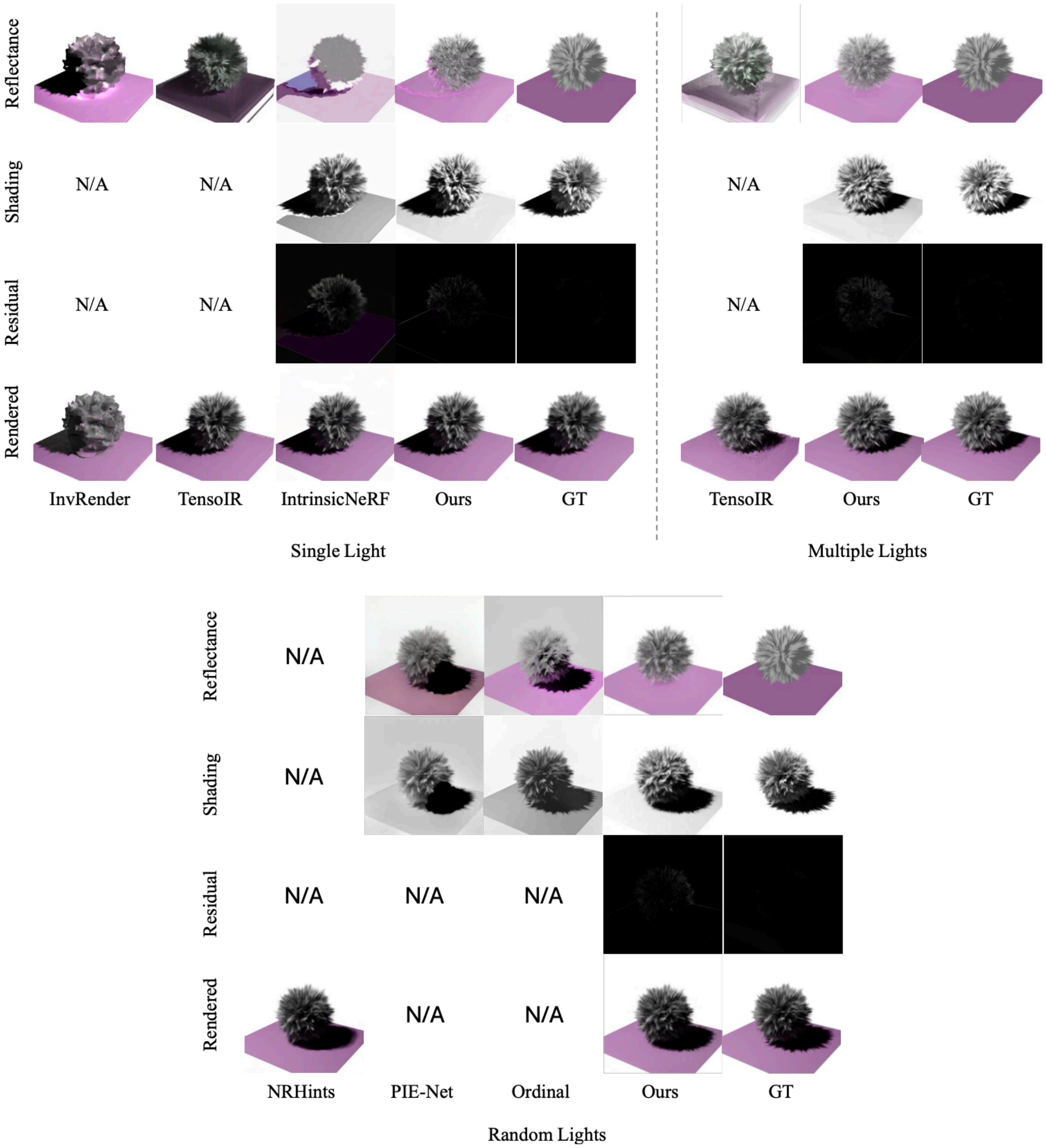

Figure 8. Additional Qualitative Results on the Synthetic Dataset (FurBall). However, it is worth noting that in the GT, the shading of the ground in this model was not correctly rendered in Blender. Our method under the Random Lights setting achieved the most reasonable results.

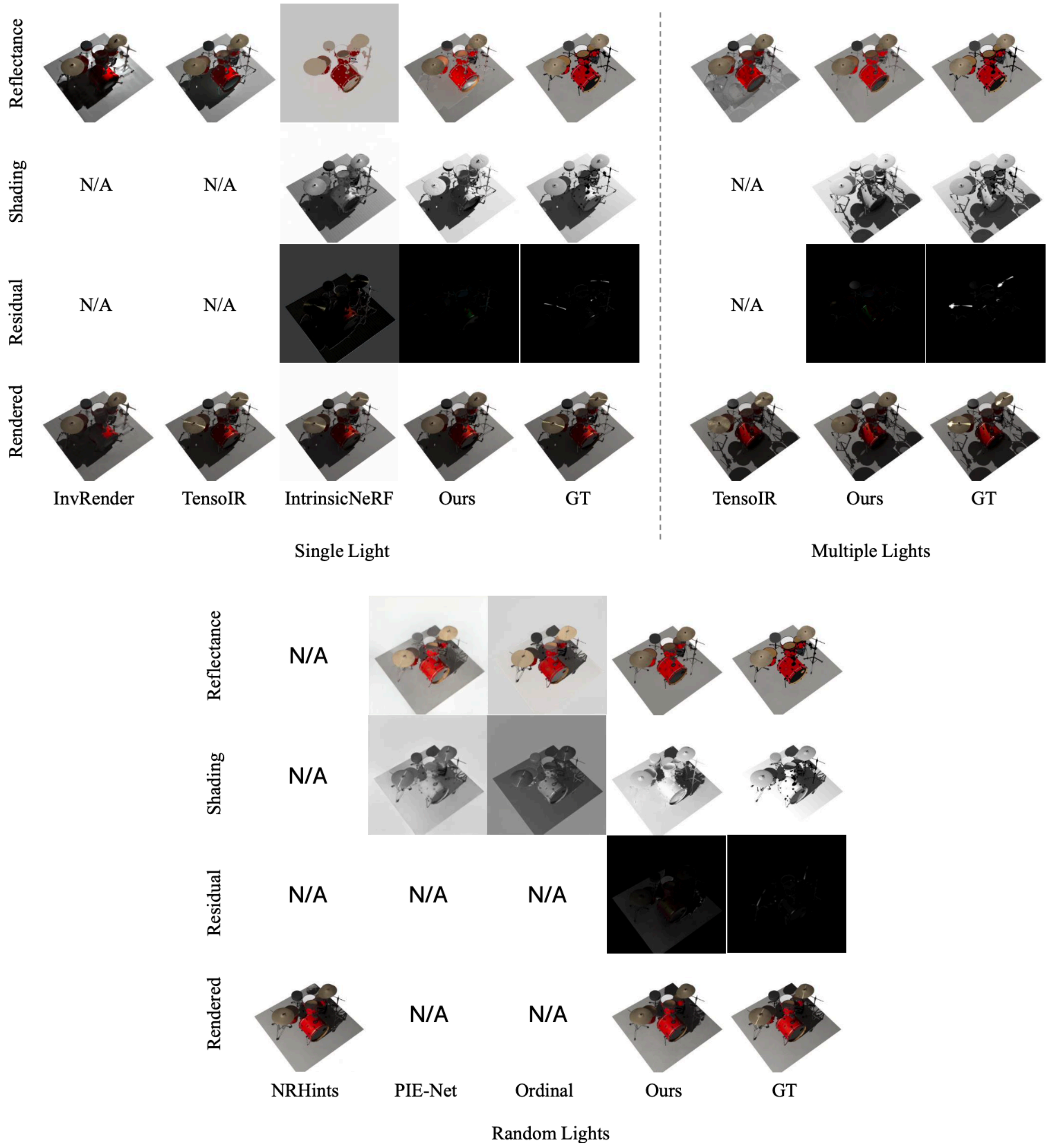

Figure 9. Additional Qualitative Results on the Synthetic Dataset. (Drums)

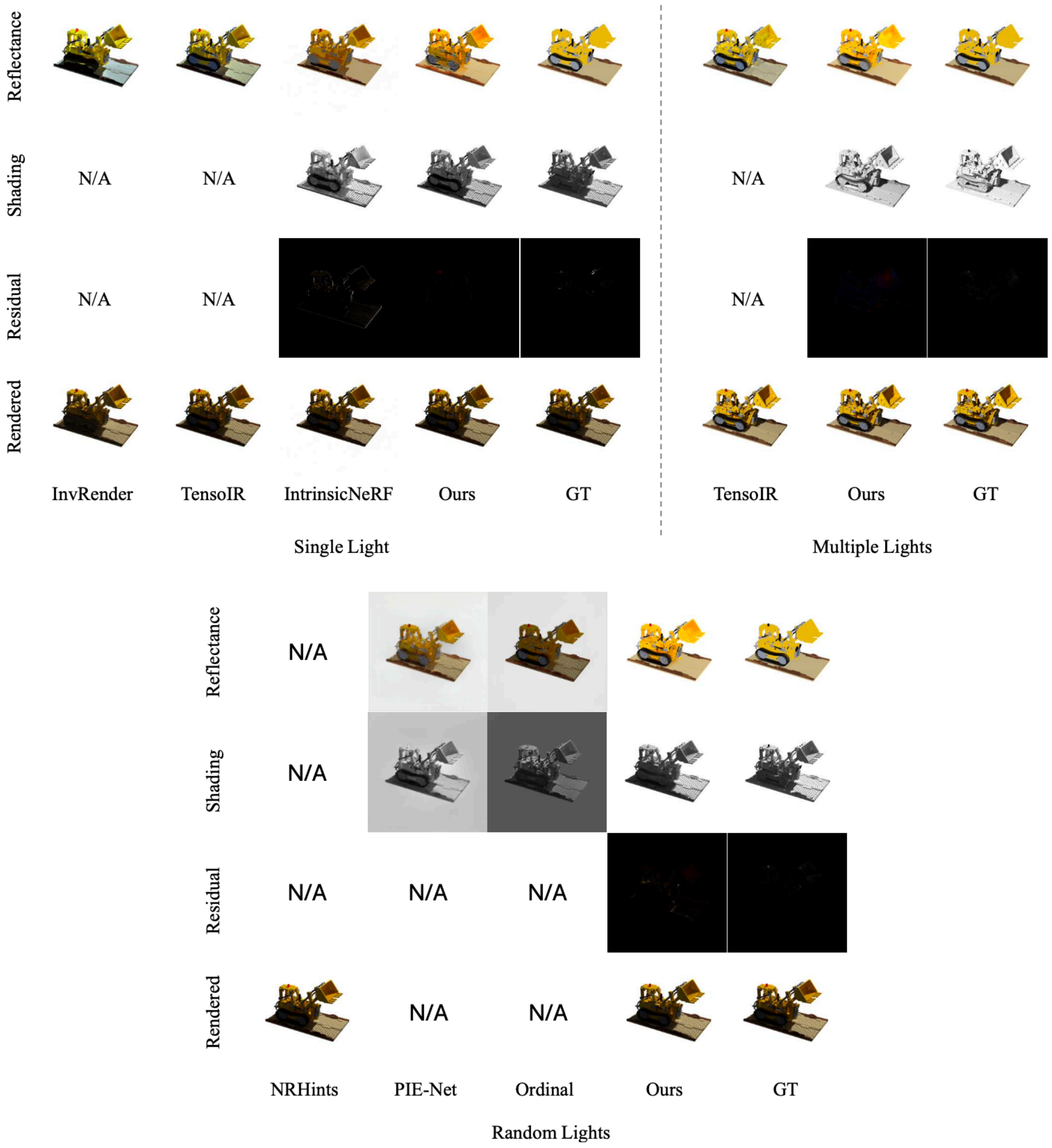

Figure 10. Additional Qualitative Results on the Synthetic Dataset. (Lego)

More results on the Real Object Dataset

We present different results for the 4 scenes of the Real Object Dataset, including Fish, Pikachu, Pixiu, and FurScene[22][7]. In Fig. 11, we show additional results on further scenes. In Fig. 12, we present the results of reflectance and shading from multiple camera views and light positions to demonstrate the coherence of the approach.

doi.org/10.32388/OQFE20

In Fig. 13 and Fig. 14, we zoom in on certain areas to show more details of the results. For reflectance, we want to highlight rows (c) and (d), where our reflectance shows very promising results. In (c), our approach correctly estimates the reflectance, likely due to the robust 3D information considered in the pseudo-shading. In (d), the high quality of the result is probably due to the pseudo-reflectance labels that effectively represent the reflectance in low-light areas. However, a limitation of our approach is evident in the shading of row (a). Although the global surface shading is satisfactory, the edges are not sharp, and there is some confusion in the areas below the object, likely due to none of the viewpoints or light sources providing adequate information for proper reconstruction.

For completeness of the comparison and potential needs, we also compared the relighting results with NRHints[7]. A quantitative evaluation is provided in Table 4. As shown, metrics show our method presents slightly lower PSNR and SSIM but better LPIPS.

	PSNR↑	**SSIM↑**	**LPIPS↓**	**MSE ↓**
NRHints	31.62	**0.9623**	0.0997	—
Ours	30.55	0.9341	**0.0797**	0.0012

Table 4. Quantitative results of the novel view synthesis and relighting on the **Real Object Dataset.**

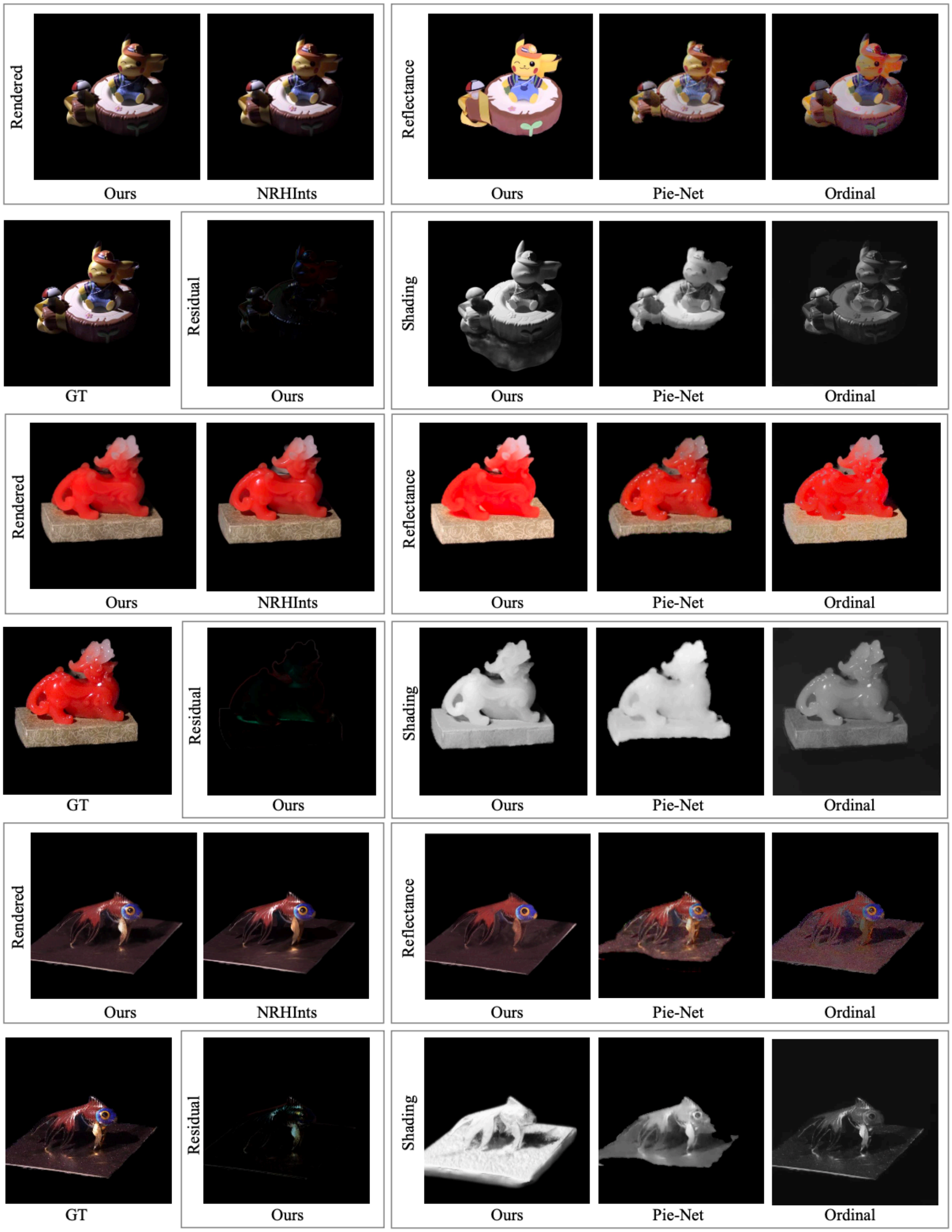

Figure 11. Additional Qualitative Results on the Real Object Dataset.

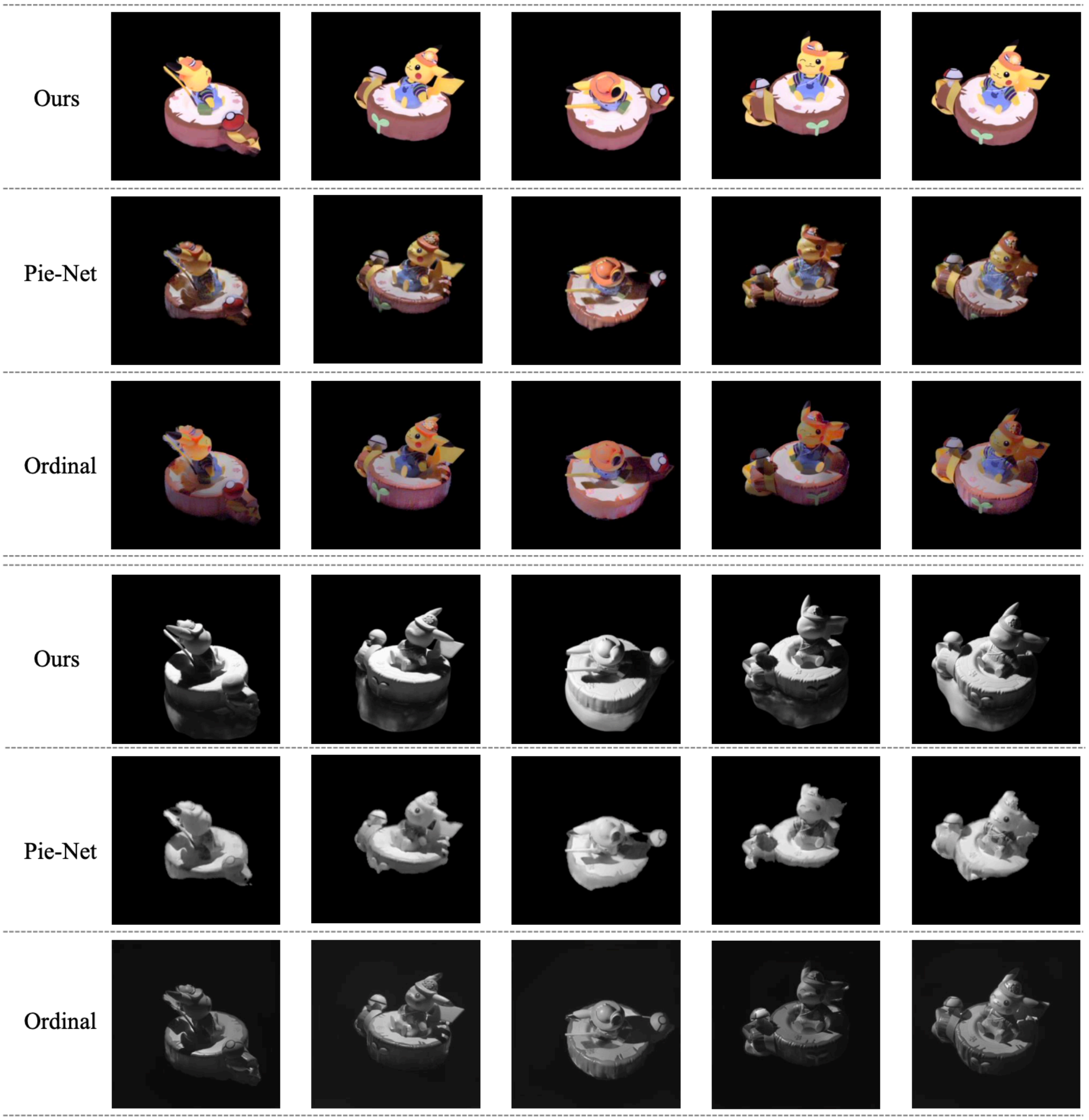

Figure 12. Reflectance and Shading estimation by our method for different points of view of the same scene on the Real Object Dataset.

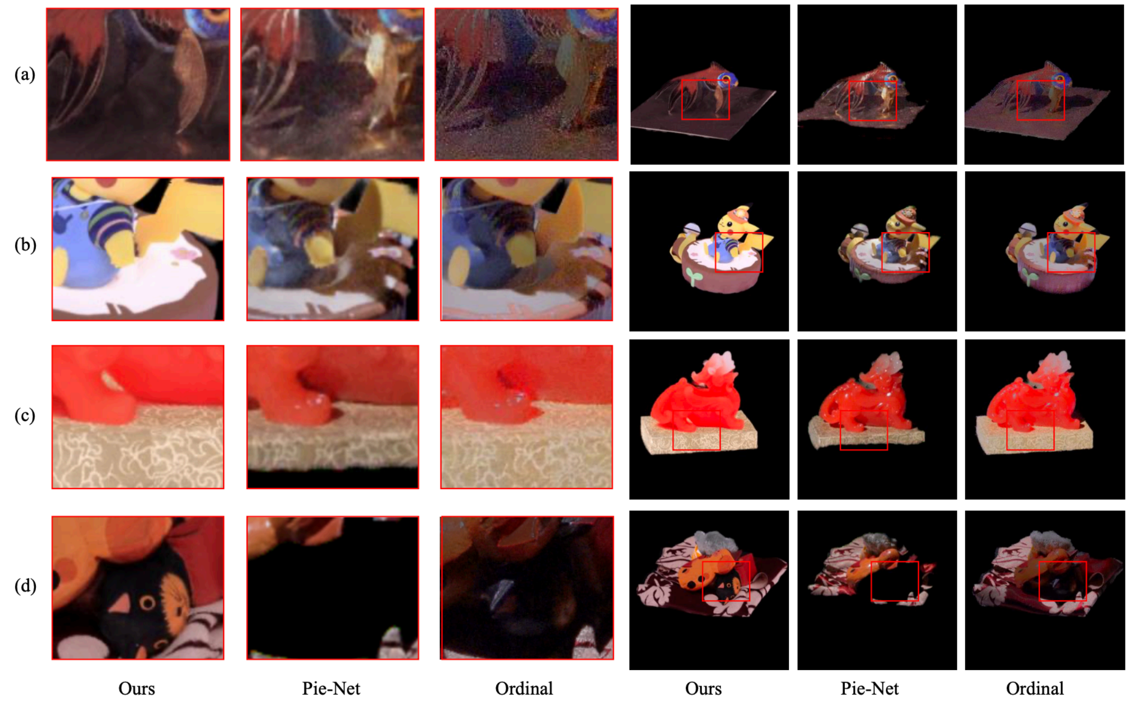

Figure 13. Reflectance Estimation details for different Scenes of the Real Object Dataset.

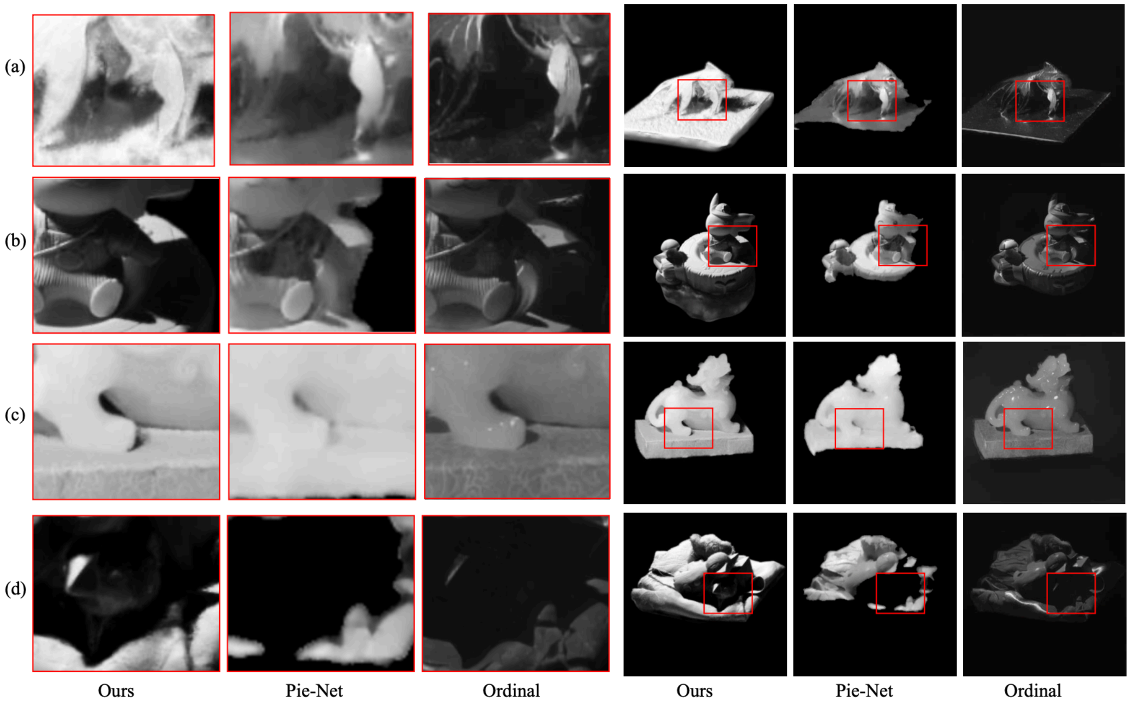

Figure 14. Shading Estimation details for different Scenes of the Real Object Dataset.

8. More results on the ReNe Dataset

Fig. 15 to Fig. 18 present more detailed results on the ReNe dataset, where the four scenes are labeled as apple, cube, garden, and savannah in the original dataset. Similar to our observations on the Synthetic Dataset, our method outperforms previous methods across all settings, including the original ALL Lights, as well as the additional Single Light and Multiple Lights settings. Furthermore, the performance of our method improves as the number of light sources increases. In all scenes under the ALL Lights and Multiple Lights settings, our method consistently produces satisfactory intrinsic images.

The ReNe dataset poses significant challenges for 3D reconstruction and intrinsic decomposition due to the camera views and light views being concentrated within a limited area. As shown in Fig. 15, TensoIR fails to produce valid outputs in the Apple scene under both the Single Light and Multiple Lights settings. Despite these challenges, our method consistently delivers satisfactory results.

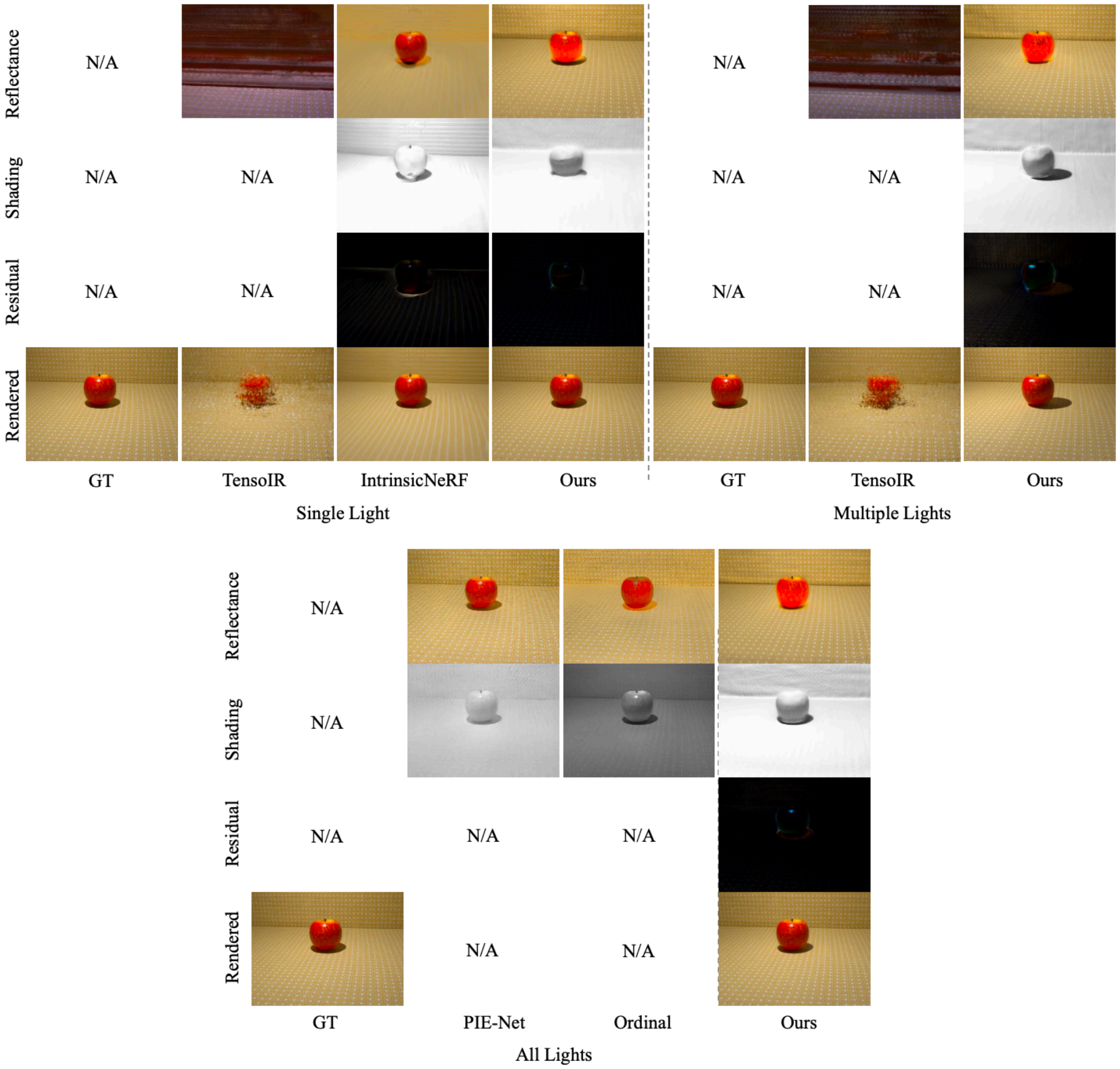

Figure 14. Additional Qualitative Results on the ReNe dataset. (Apple).

doi.org/10.32388/OQFE20

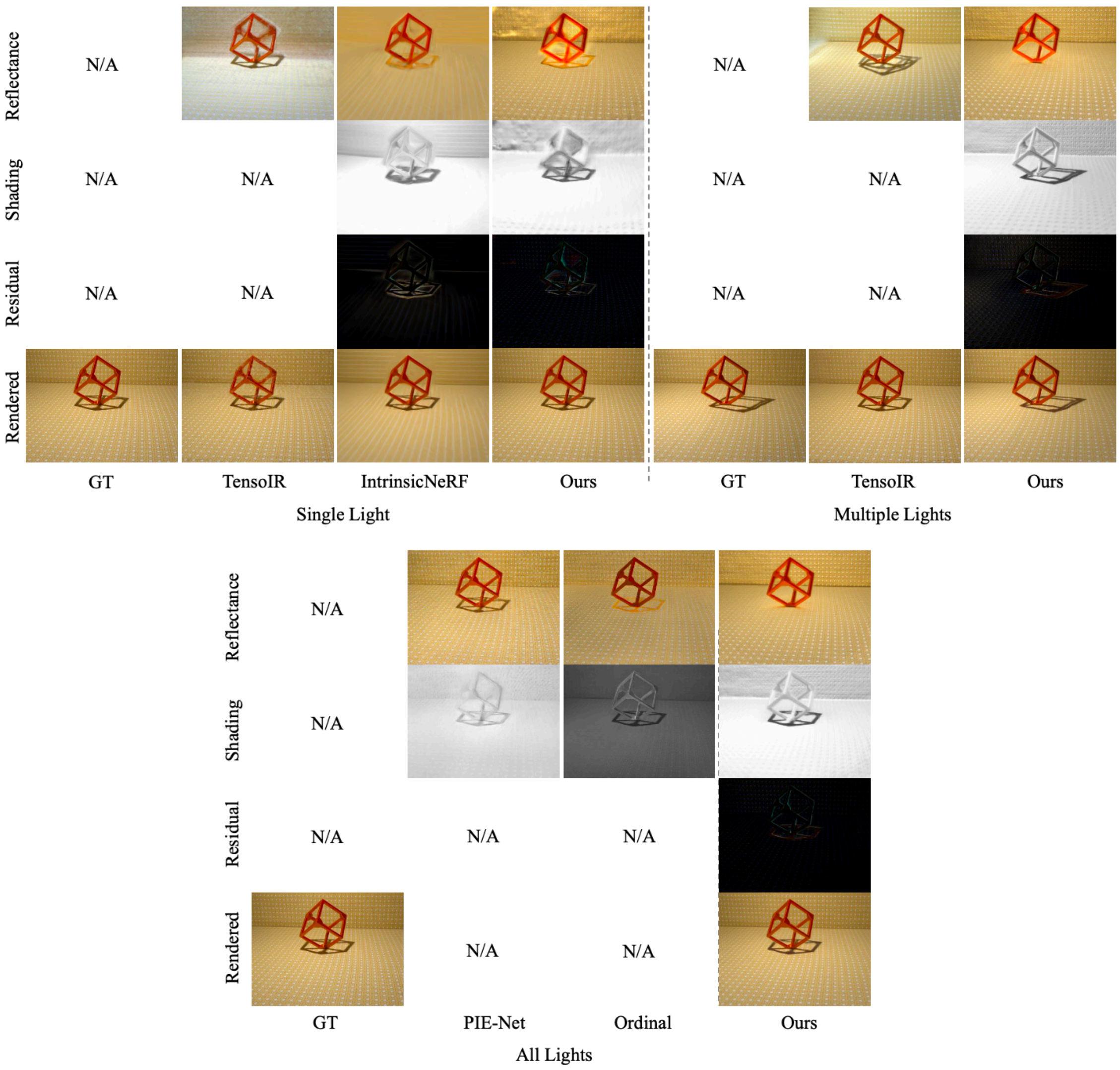

Figure 16. Additional Qualitative Results on the ReNe dataset. (Cube).

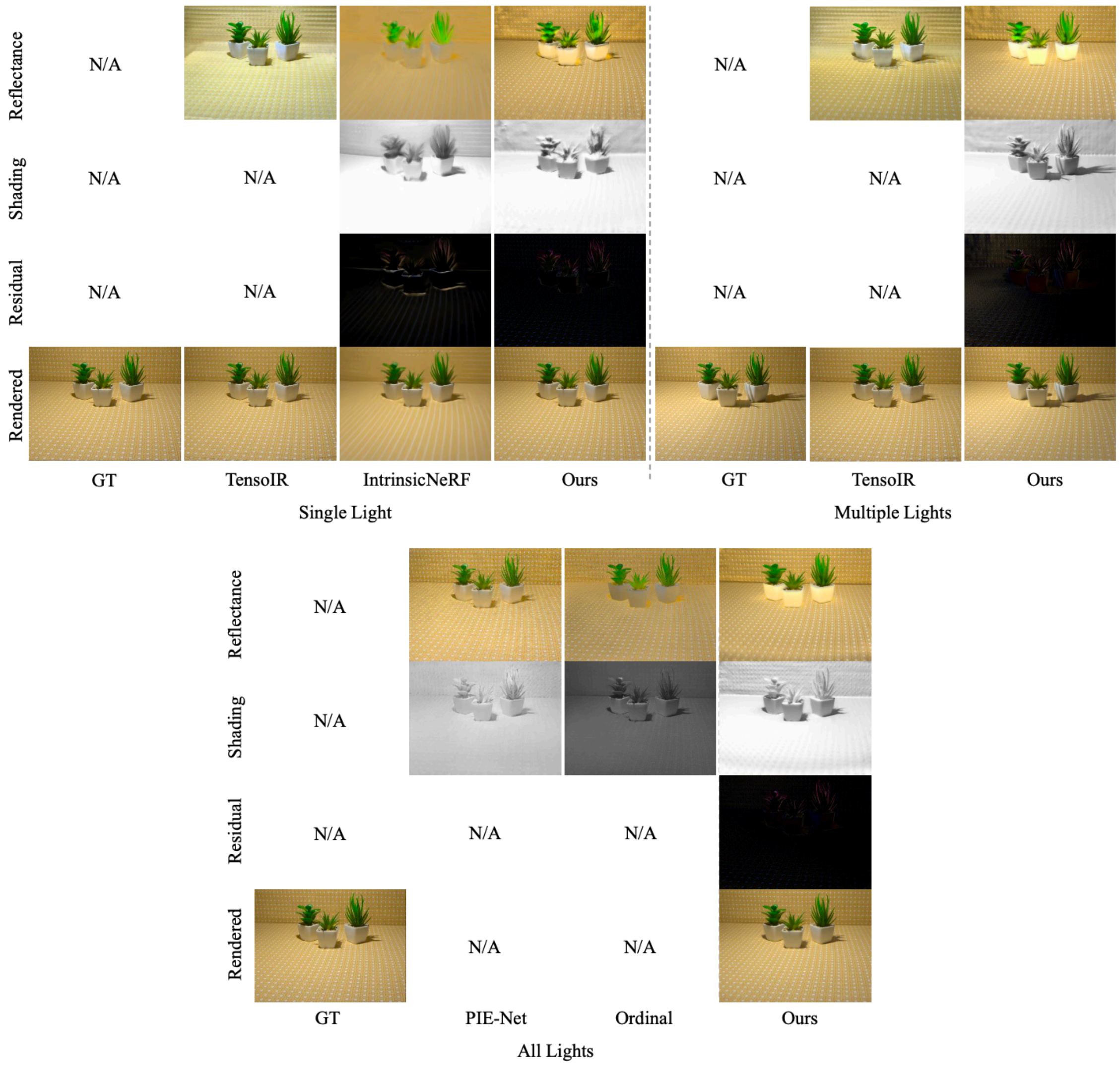

Figure 17. Additional Qualitative Results on the ReNe dataset. (Garden).

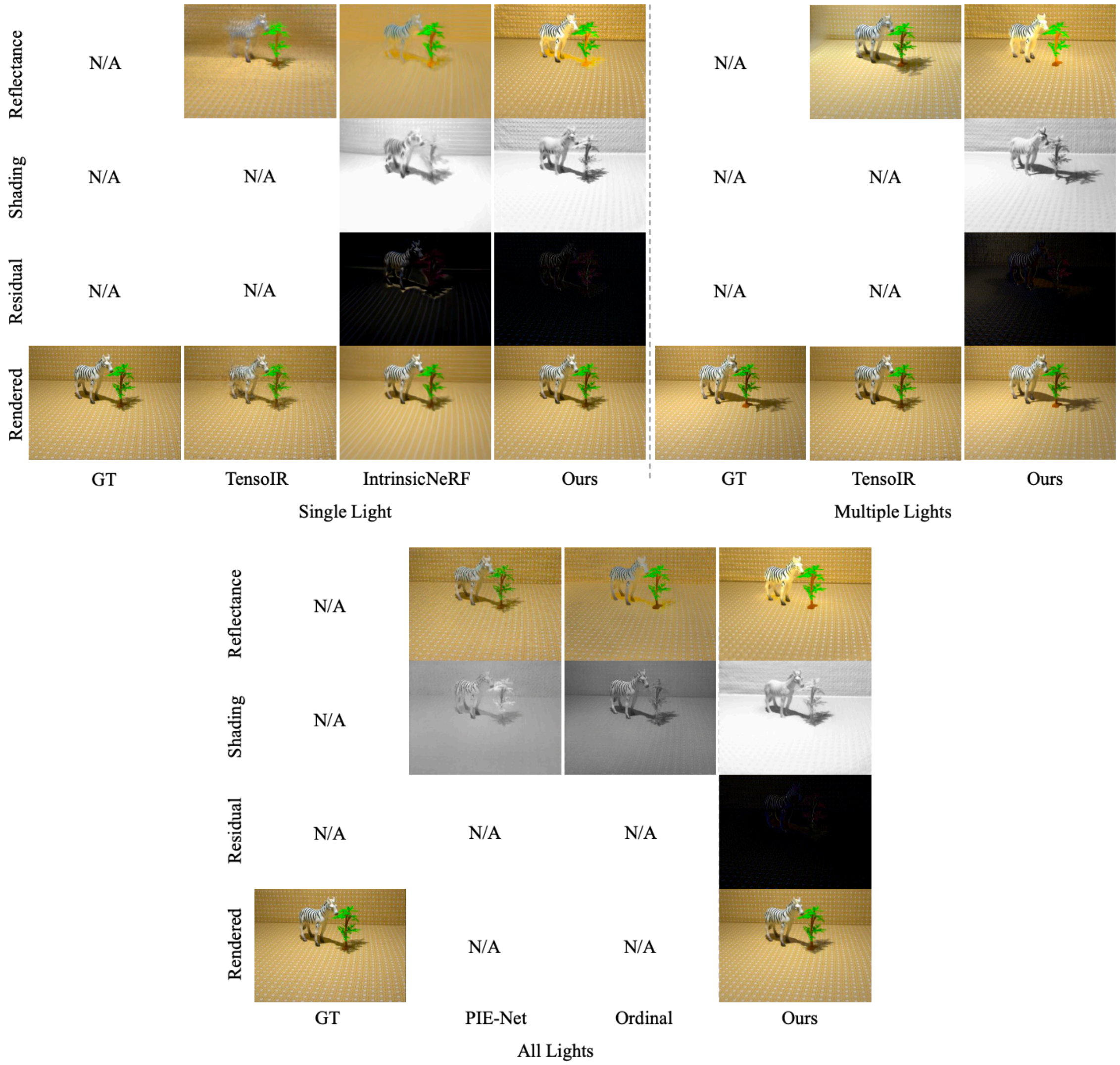

Figure 18. Additional Qualitative Results on the ReNe dataset. (savannah).

Acknowledgments

This work was supported by Grant PID2021-128178OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF ”A way of making Europe”. It was also supported by the Departament de Recerca i Universitats from Generalitat de Catalunya, reference 2021SGR01499. Yixiong Yang is supported by China Scholarship Council.

References

1. a, b, c, d, e*Toschi M, De Matteo R, Spezialetti R, De Gregorio D, Di Stefano L, Salti S. "ReLight My NeRF: A Dataset for Novel View Synthesis and Relighting of Real World Objects." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023. p. 20762-20772.*
2. a, b, c*Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R (2021). "Nerf: Representing scenes as neural radiance fields for view synthesis". Communications of the ACM.* ***65*** *(1): 99–106.*
3. a, b, c, d, e, f*Li Z, M\u00fcller T, Evans A, Taylor RH, Unberath M, Liu M-Y, Lin C-H. Neuralangelo: High-Fidelity Neural Surface Reconstruction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2023.*
4. ^*Wang Y, Wu W, Xu D. "Learning Unified Decompositional and Compositional NeRF for Editable Novel View Synthesis." In: ICCV; 2023.*
5. a, b, c, d, e, f, g, h, i, j*Ye W, Chen S, Bao C, Bao H, Pollefeys M, Cui Z, Zhang G (2023). "IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis". Proceedings of the IEEE/CVF International Conference on Computer Vision.*
6. a, b*Ling J, Wang Z, Xu F (2022). "ShadowNeuS: Neural SDF Reconstruction by Shadow Ray Supervision". arXiv. arXiv:2211.14086.*
7. a, b, c, d, e, f, g, h, i, j, k, l, m, n, o*Zeng C, Chen G, Dong Y, Peers P, Wu H, Tong X (2023). "Relighting Neural Radiance Fields with Shadow and Highlight Hints". In: ACM SIGGRAPH 2023 Conference Proceedings.*
8. a, b, c*Garces E, Rodriguez-Pardo C, Casas D, Lopez-Moreno J (2022). "A survey on intrinsic images: Delving deep into lambert and beyond". International Journal of Computer Vision.* ***130*** *(3): 836–868.*
9. a, b, c, d, e, f*Jin H, Liu I, Xu P, Zhang X, Han S, Bi S, Zhou X, Xu Z, Su H. "TensoIR: Tensorial Inverse Rendering." In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2023.*
10. a, b, c, d, e*Zhang Y, Sun J, He X, Fu H, Jia R, Zhou X (2022). "Modeling Indirect Illumination for Inverse Rendering". In: CVPR, 2022.*
11. a, b*Yang Z, Chen Y, Gao X, Yuan Y, Wu Y, Zhou X, Jin X (2023). "SIRe-IR: Inverse Rendering for BRDF Reconstruction with Shadow and Illumination Removal in High-Illuminance Scenes". arXiv preprint arXiv:2310.13030. arXiv:2310.13030.*
12. a, b*Zhang X, Srinivasan PP, Deng B, Debevec P, Freeman WT, Barron JT (2021). "Nerfactor: Neural factorization of shape and reflectance under an unknown illumination". ACM Transactions on Graphics (ToG).* ***40*** *(6): 1–18.*

13. ^Burley B, Studios WDA. Physically-based shading at disney. In: Acm Siggraph. vol. 2012, 2012. p. 1-7.

14. a, b, c, d, e, f, g, hCareaga C, Aksoy Y. "Intrinsic Image Decomposition via Ordinal Shading". ACM Trans. Graph.. 2023.

15. a, b, c, d, e, fDas P, Karaoglu S, Gevers T. "PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposition". In: IEEE Conference on Computer Vision and Pattern Recognition, (CVPR); 2022.

16. a, b, cBarrow H, Tenenbaum J, Hanson A, Riseman E (1978). "Recovering intrinsic scene characteristics". Comput. Vis. Syst. **2** (3-26): 2.

17. a, bLi Z, Snavely N. "Learning Intrinsic Image Decomposition from Watching the World." In: Computer Vision and Pattern Recognition (CVPR); 2018.

18. a, b, cLettry L, Vanhoey K, Van Gool L (2018). "Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences". Computer Graphics Forum. **37**: 409–419. Wiley Online Library.

19. ^Rudnev V, Elgharib M, Smith W, Liu L, Golyanik V, Theobalt C. NeRF for Outdoor Scene Relighting. In: European Conference on Computer Vision (ECCV); 2022.

20. a, b, cChen Z, Ding C, Guo J, Wang D, Li Y, Xiao X, Wu W, Song L. L-Tracing: Fast Light Visibility Estimation on Neural Surfaces by Sphere Tracing. In: Proceedings of the European Conference on Computer Vision (ECCV); 2022.

21. ^Li Z, Snavely N. "Learning intrinsic image decomposition from watching the world". In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 9039–9048.

22. a, b, c, dGao D, Chen G, Dong Y, Peers P, Xu K, Tong X (2020). "Deferred neural lighting: free-viewpoint relighting from unstructured photographs". ACM Transactions on Graphics (TOG). **39** (6): 258.

23. ^Barron JT, Malik J (2014). "Shape, illumination, and reflectance from shading". IEEE Transactions on Pattern Analysis and Machine Intelligence. **37** (8): 1670–1687.

24. ^Li Z, Snavely N (2018). "CGIntrinsics: Better Intrinsic Image Decomposition through Physically-Based Rendering". In: European Conference on Computer Vision (ECCV).

25. ^Liu Y, Li Y, You S, Lu F (2020). "Unsupervised Learning for Intrinsic Image Decomposition from a Single Image". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 3248--3257.

26. ^Einabadi F, Guillemaut JY, Hilton A (2021). "Deep neural models for illumination estimation and relighting: A survey". Computer Graphics Forum. **40**: 315–331. Wiley Online Library.

27. ^Nestmeyer T, Lalonde JF, Matthews I, Lehrmann A (2020). "Learning Physics-guided Face Relighting under Directional Light". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

2020: 5124–5133.

28. ^*Sun T, Barron JT, Tsai YT, Xu Z, Yu X, Fyffe G, et al. Single image portrait relighting. ACM Trans. Graph.* ***38*** *(4): 79--1, 2019.*

29. ^*Zhou H, Hadap S, Sunkavalli K, Jacobs DW (2019). "Deep single-image portrait relighting". Proceedings of the IEEE International Conference on Computer Vision. 7194--7202.*

30. ^*Pandey R, Orts-Escolano S, Legendre C, Haene C, Bouaziz S, Rhemann C, Debevec PE, Fanello SR (2021). "Total relighting: learning to relight portraits for background replacement." ACM Trans. Graph..* ***40****: 43--1.*

31. ^*Hou A, Sarkis M, Bi N, Tong Y, Liu X (2022). "Face relighting with geometrically consistent shadows". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pages 4217–4226.*

32. ^*Murmann L, Gharbi M, Aittala M, Durand F (2019). "A dataset of multi-illumination images in the wild". Proceedings of the IEEE/CVF International Conference on Computer Vision. 4080--4089.*

33. ^*Helou ME, Zhou R, Süsstrunk S, Timofte R, Afifi M, Brown MS, Xu K, Cai H, Liu Y, Wang LW, et al. AIM 2020: Scene relighting and illumination estimation challenge. arXiv preprint arXiv:2009.12798. 2020.*

34. ^*Puthussery D, Kuriakose M, C V J, et al. WDRN: A wavelet decomposed relightnet for image relighting. arXiv preprint arXiv:2009.06678. 2020.*

35. ^*Wang LW, Siu WC, Liu ZS, Li CT, Lun DPK (2020). "Deep relighting networks for image light source manipulation". arXiv preprint arXiv:2008.08298. Available from: https://arxiv.org/abs/2008.08298.*

36. ^*El Helou M, Zhou R, Susstrunk S, Timofte R (2021). "NTIRE 2021 depth guided image relighting challenge". Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 566–577.*

37. ^*Kocsis P, Philip J, Sunkavalli K, Nie{\ss}ner M, Hold-Geoffroy Y. LightIt: Illumination Modeling and Control for Diffusion Models. In: CVPR; 2024.*

38. ^*Srinivasan PP, Deng B, Zhang X, Tancik M, Mildenhall B, Barron JT (2021). "Nerv: Neural reflectance and visibility fields for relighting and view synthesis". In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 7495–7504.*

39. a, b*Chang Y, Kim Y, Seo S, Yi J, Kwak N. Fast Sun-aligned Outdoor Scene Relighting based on TensoRF. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2024. p. 3626–3636.*

40. ^*Zhang K, Luan F, Wang Q, Bala K, Snavely N. "PhySG: Inverse Rendering with Spherical Gaussians for Physics-based Material Editing and Relighting." In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2021.*

41. ^*Boss M, Braun R, Jampani V, Barron JT, Liu C, Lensch H (2021). "Nerd: Neural reflectance decomposition from image collections". In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 126*

84–12694.

42. ^*Boss M, Engelhardt A, Kar A, Li Y, Sun D, Barron JT, Lensch HP, Jampani V. "SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections." In: Advances in Neural Information Processin g Systems (NeurIPS); 2022.*
43. ^*Liu I, Chen L, Fu Z, Wu L, Jin H, Li Z, Wong CM, Xu Y, Ramamoorthi R, Xu Z, Su H (2023). "OpenIllumination: A Multi-Illumination Dataset for Inverse Rendering Evaluation on Real Objects." In: Oh A, Naumann T, Glob erson A, Saenko K, Hardt M, Levine S, editors. Advances in Neural Information Processing Systems. Curran Associates, Inc.; 2023. p. 36951-36962. Available from: https://proceedings.neurips.cc/paper_files/paper/2023/file/74a67268c5cc5910f64938cac4526a90-Paper-Datasets_and_Benchmarks.pdf.*
44. ^*Fan Q, Yang J, Hua G, Chen B, Wipf D (2018). "Revisiting deep intrinsic image decompositions". Proceedings of the IEEE conference on computer vision and pattern recognition. 8944--8952.*
45. ^*Müller T, Evans A, Schied C, Keller A (2022). "Instant Neural Graphics Primitives with a Multiresolution Ha sh Encoding". ACM Trans. Graph.. **41** (4): 102:1–102:15. doi:10.1145/3528223.3530127.*
46. ^*Gropp A, Yariv L, Haim N, Atzmon M, Lipman Y (2020). "Implicit geometric regularization for learning sha pes". Proceedings of the 37th International Conference on Machine Learning. 2020: 3789--3799.*
47. ^*Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Du bourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011). "Scikit-learn: Mac hine learning in Python". Journal of Machine Learning Research. **12**: 2825–2830.*
48. ^*Loshchilov I, Hutter F (2018). "Decoupled Weight Decay Regularization". In: International Conference on L earning Representations.*
49. ^*Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004). "Image quality assessment: from error visibility to str uctural similarity". IEEE Transactions on Image Processing. **13** (4): 600–612.*
50. ^*Zhang R, Isola P, Efros AA, Shechtman E, Wang O (2018). "The unreasonable effectiveness of deep features as a perceptual metric". In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595.*

Declarations

Funding: This work was supported by Grant PID2021-128178OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF ”A way of making Europe”. It was also supported by the

Departament de Recerca i Universitats from Generalitat de Catalunya, reference 2021SGR01499. Yixiong Yang is supported by China Scholarship Council.

Potential competing interests: No potential competing interests to declare.