Photometric Inverse Rendering: Shading Cues Modeling and Surface Reflectance Regularization

Jingzhi Bao[1,2] Guanying Chen[3*] Shuguang Cui[4,1]
[1] FNii-Shenzhen [2] SDS, CUHKSZ [3] Sun Yat-sen University [4] SSE, CUHKSZ

Abstract

This paper addresses the problem of inverse rendering from photometric images. Existing approaches for this problem suffer from the effects of self-shadows, inter-reflections, and lack of constraints on the surface reflectance, leading to inaccurate decomposition of reflectance and illumination due to the ill-posed nature of inverse rendering. In this work, we propose a new method for neural inverse rendering. Our method jointly optimizes the light source position to account for the self-shadows in images, and computes indirect illumination using a differentiable rendering layer and an importance sampling strategy. To enhance surface reflectance decomposition, we introduce a new regularization by distilling DINO features to foster accurate and consistent material decomposition. Extensive experiments on synthetic and real datasets demonstrate that our method outperforms the state-of-the-art methods in reflectance decomposition.

1. Introduction

Inverse rendering aims to estimate the shape, materials, and lighting of a scene from 2D images. It finds applications in 3D object digitization, object manipulation, and relighting.

Recently, neural representations have achieved significant success in novel-view synthesis and 3D modeling [44, 46, 52, 75]. Neural radiance fields (NeRF), in particular, model a scene with a Multi-Layer Perceptron (MLP) that maps 3D coordinates and view directions to color and density, resulting in photorealistic rendering [44]. However, NeRF lacks explicit modeling of surface reflectance and lighting, making it unsuitable for relighting tasks. Several methods have been proposed to incorporate physics-based image formation models, enabling the explicit decomposition of reflectance and lighting [53, 83].

Figure 1. Reconstructed 3D assets inserted in a real game scene.

A branch of methods focuses on inverse rendering using images captured under environmental illumination [7, 83, 85]. However, this presents a highly ill-posed problem due to the complex interactions between shape, materials, and lighting. Despite promising results, these methods often suffer from challenges such as the mingling of estimated surface reflectance with illumination effects, especially for real-world objects. In a different approach, IRON [84] has proposed an approach for inverse rendering from photometric images (i.e., multi-view images captured by co-locating a flashlight with a moving camera), yielding impressive results. Compared to environmental illumination, the flashlight (resembling a point light model) simplifies the image formation process, and the captured images contain high-frequency details (e.g., specular highlights), which are beneficial for reflectance estimation [11, 64].

However, IRON [84] presents several shortcomings. Firstly, it assumes an ideal collocated camera-lighting arrangement, neglecting the complications posed by self-shadows that are frequently unfeasible in casual capture contexts, like smartphone use. Secondly, it fails to consider the diverse high-frequency inter-reflections characteristic of multi-view images captured with flashlight illumination. Such oversights can lead to inaccuracies in the estimation of diffuse albedo, as it allows self-shadows to distort the outcomes or unintentionally incorporates specular highlights, particularly within concave regions. Additionally, the absence of effective reflectance regularization in IRON undermines the precision of material decomposition.

[*] Corresponding author
Project page: https://jzbao03.site/projects/PIR/

In this work, we introduce a novel method that leverages rich shading information available in photometric images to achieve robust inverse rendering. Notably, our approach removes the necessity of co-locating the point light source with the camera, instead opting for a joint optimization of the light source's position. This optimization accounts for the intricate interplay between the object's geometry and the light source's position, enabling our algorithm to effectively deduce the presence of self-shadows and significantly diminish their distorting impact on the resultant images. To accurately simulate the effect of inter-reflections, our method integrates an effective importance sampling strategy alongside a differentiable rendering layer. These techniques effectively reducing the unwanted blending of inter-reflections on material properties. To alleviate the ambiguity inherent from reflectance estimation, we incorporate a DINO [9] feature regularization into our inverse rendering framework. The self-supervised DINO method, by learning from extensive unlabeled datasets, captures image features encoding view-consistent contextual information across the scene, providing valuable information to understand the reflectance properties of different image regions and advancing the accuracy of the decomposition process.

In summary, our key contributions are as follows:

- We propose a novel neural inverse rendering framework tailored for photometric images that jointly optimizes object shapes, materials and lighting, achieving accurate reflectance decomposition.
- We harness the shading cues present in photometric images to achieve robust inverse rendering. Our method effectively models self-shadows and employs networks alongside an importance sampling strategy for accurate inference of high-frequency inter-reflections. This approach ensures a detailed and precise rendering by capturing the subtle lighting interactions within the scene.
- We introduce the DINO feature regularization for surface reflectance to group similar materials. Extensive experiments show that our method outperforms existing methods in novel view synthesis and material decomposition.
- We present a new dataset contains 5 scenes captured by a mobile phone in a darkroom. The number of images per scene ranges from 120 to 400.

2. Related Work

Neural Scene Representation Neural scene representations have brought significant advancements to the fields of novel-view synthesis [46, 52, 75]. The neural radiance field (NeRF) [44] adopts a Multi-Layer Perceptron (MLP) to represent a scene by mapping a 3D coordinate and a view direction to color and density, followed by volume rendering for pixel color computation. To address the inherent noise in the surface derived from the density field, various efforts have leveraged the strengths of both volume rendering and surface rendering to enhance surface geometry [47, 63, 76].

Many follow-up methods aim to improve the performance of NeRF on different surfaces types. Ref-NeRF [59] re-parameterizes NeRF's outgoing radiance based on the reflection of the viewing vector with respect to the local normal vector, leading to improved rendering for specular surfaces [17, 90]. Some methods extend NeRF to handle more complex scenes containing mirror surfaces [20, 28, 57, 77, 81] and transparent objects [3, 48, 62].

However, NeRF [44] lacks explicit modeling of surface reflectance and lighting, making it unsuitable for relighting tasks. In this work, we focus on performing inverse rendering from photometric images.

Inverse Rendering with Environment Illumination A subset of methods has emerged to jointly recover the shape, materials, and lighting of objects [15, 24, 29, 32, 33, 39, 42, 54, 55, 78, 93] or entire scene [13, 31, 35, 49, 66, 70, 91] using neural scene representation from multi-view images. These methods consider an unknown distant environment illumination, and adopt diverse representations for shapes (*e.g.*, density [85], SDF [83], and mesh [45]), illuminations (*e.g.*, spherical Gaussian [6] and pre-integrated lighting [7]), and materials [2, 21, 41, 68, 79, 87]. Recently, several studies have emerged focusing on inverse rendering through the application of 3D Gaussian splatting [16, 23, 34, 51]. The primary contributions of these works revolve around acceleration, which is orthogonal to our approach.

Efficiently computing inter-reflections, which involve tracing multiple bounces of rays, is a challenging problem in inverse rendering. Existing methods handle inter-reflections by assuming fixed illumination among multi-view images [14, 53, 67, 73, 74, 82, 86]. For example, InvRender [86] introduces an indirect illumination MLP to map a 3D point to its indirect incoming illumination, directly derived from the outgoing radiance field. However, our setup involves each image being illuminated under a different point light, making the existing approach unsuitable for our scenario.

To regularize the decomposition of reflectance, NeRFactor [85] learns a data prior for BRDFs by training an autoencoder on the *MERL dataset* [43]. Some methods apply low-rank or vector-quantization regularization on the reflectance [86, 88, 89]. In comparison, we introduce a novel regularization without the need for additional training data by distilling the DINO [9] feature into the object's surface. In the context of neural representation, DINO has been adopted in NeRF to scene editing [27] and grouping semantic feature [26].

Inverse Rendering with Point Lights Different from the environment illumination, a point light model simplifies the image formation model and results in images with more high-frequency details, such as specular highlights, which significantly reduce ambiguity in inverse rendering

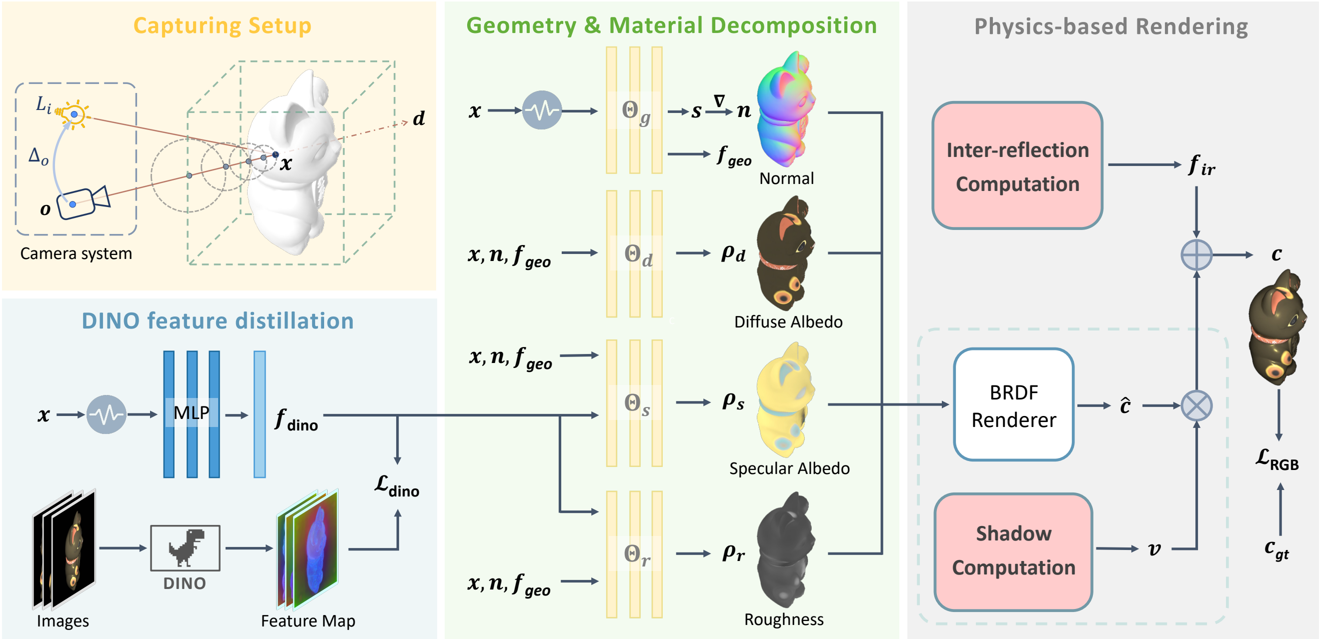


Figure 2. Method overview. Our method optimizes the light source position to account for self-shadows and models inter-reflection. The DINO features are injected into the networks of specular albedo and roughness to regularize the material decomposition.

[37, 58, 69, 80, 92]. Several methods have been proposed to improve the accuracy of inverse rendering by utilizing a point light model [4, 5, 8, 12, 19, 30, 71]. IRON [84] employs the NeuS method [63] to represent the surface and utilizes an edge-aware physics-based surface rendering for geometry refinement and materials estimation. However, IRON assumes an ideal collocated camera-lighting setup and overlooks self-shadows and inter-reflections. In contrast, our method addresses all these issues and explicitly regularizes on surface reflectance to achieve more accurate inverse rendering. While some methods compute shadows during optimization, they typically assume the environment illumination [10, 60, 85] or known point light positions [36, 38, 56, 72]. Differently, our method utilizes self-shadow cues to calibrate the light positions.

3. Method

3.1. Overview

Capturing Setting In this work, we focus on reconstructing object geometries, materials, and illumination conditions from multi-view images lit by a flashlight. Previous methods [12, 84] assume a collocated camera-light setup and optimize the scene with a simplified rendering model. Such an ideal setting is impractical to attain in our daily capture like mobile phone, we consider a more general fixed setup akin to a camera mounted on a mobile phone. Our method (Fig. 2) tackles the complexities introduced by self-shadows, inter-reflections, and ambiguities in reflectance estimation, which are prevalent issues in current techniques.

Rendering Equation In theory, the rendering equation [25] for a surface point $\boldsymbol{x}$ can be written as

$$\hat{I}(\boldsymbol{w}_o;\boldsymbol{x}) = \int_\Omega L_i(\boldsymbol{w}_i;\boldsymbol{x}) f_r(\boldsymbol{w}_o,\boldsymbol{w}_i;\boldsymbol{x})(\boldsymbol{w}_i\cdot\boldsymbol{n})\,\mathrm{d}\boldsymbol{w}_i, \quad (1)$$

where $L_i(\boldsymbol{w}_i;\boldsymbol{x})$ denotes the incoming radiance arriving from direction $\boldsymbol{w}_i$, and $f_r(\boldsymbol{w}_o,\boldsymbol{w}_i;\boldsymbol{x})$ encapsulates the surface's bidirectional reflectance distribution function (BRDF) at $\boldsymbol{x}$. This equation calculates the outgoing radiance $\hat{I}(\boldsymbol{w}_o;\boldsymbol{x})$ of point $\boldsymbol{x}$ in the direction of $\boldsymbol{w}_o$ by integrating all radiance contributions over the upper-hemisphere Ω surrounding the surface normal $\boldsymbol{n}$.

By assuming a point light and considering light visibility and indirect lights, the rendering can be approximated as

$$\begin{aligned}\hat{I}(\boldsymbol{w}_o;\boldsymbol{x}) = L_i(\boldsymbol{w}_i;\boldsymbol{x}) f_r(\boldsymbol{w}_o,\boldsymbol{w}_i;\boldsymbol{x})(\boldsymbol{w}_i\cdot\boldsymbol{n})\\ \times f_v(\boldsymbol{w}_i;\boldsymbol{x}) + f_{ir}(\boldsymbol{w}_o;\boldsymbol{x}),\end{aligned} \quad (2)$$

where $f_v(\boldsymbol{w}_i;\boldsymbol{x})$ indicates the visibility of light along $\boldsymbol{w}_i$ at $\boldsymbol{x}$ that models self-shadows in the rendered image, and f_{ir} accounts for the residual effects attributed to inter-reflections.

Pipeline Our pipeline commences with estimating the object's geometry and surface diffuse albedo using the off-the-shelf neural surface reconstruction framework, NeuS [42]. Subsequently, we utilize physics-based rendering to jointly refine the geometry and materials of the object as well as the position and intensity of the flashlight. Our approach leverages differentiable rendering techniques to accurately model self-shadows and indirect illumination in photometric images, thereby achieving robust material decomposition. Additionally, we reduce the ambiguities of surface reflectance decomposition by integrating self-supervised DINO [9] features from multi-view images. Our method can export the mesh and texture maps of the optimized 3D models, which can be seamlessly integrated into conventional rendering pipelines.

3.2. Neural Scene Representation

Geometry Representation In our approach to representing scene geometry, the geometry is represented by the zero level set of a SDF $\mathcal{S} = \left\{\boldsymbol{x} \in \mathbb{R}^3 \mid s(\boldsymbol{x}) = 0\right\}$ in line with recent advancements in the field [63, 75, 86]. For any point $\boldsymbol{x} \in \mathbb{R}^3$, the signed distance s and a learned local geometric feature descriptor of $\boldsymbol{x}$ are parameterized by a MLP, denoted as $f_{\Theta_g} = (s, \boldsymbol{f}_{\text{geo}}) \in \mathbb{R} \times \mathbb{R}^{256}$.

Color rendering for a pixel is achieved through the projection of a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ from the camera's origin $\mathbf{o}$, extending in the view direction $\mathbf{d}$. In the volumetric field, the color C rendered from a pixel is obtained by integrating along its ray path, with the integral approximated over N discrete points as follows:

$$C(\mathbf{o}, \mathbf{d}) = \sum_{j=1}^{N} T_j \left(1 - \exp\left(-\sigma_j \delta_j\right)\right) \boldsymbol{c}_j, \tag{3}$$

where $T_j = \exp\left(-\sum_{q=1}^{j-1} \sigma_q \delta_q\right)$ denotes the accumulated transmittance at sampled point $\mathbf{r}(t_j)$, and $\boldsymbol{c}_j$ represents the point's color. We incorporate the unbiased density conversion method [42], translating SDF values into density representations for scene's geometry.

Materials Representation To achieve physics-based rendering, our framework decomposes the scene's BRDF into diffuse and specular components, utilizing the roughplastic model for microfacet specular reflection [61]. The materials at a point $\boldsymbol{x}$ include the diffuse albedo ρ_d, specular albedo ρ_s, and roughness ρ_r. These spatially-varying BRDF parameters are encapsulated by MLPs with a positional encoding function and optimizable parameters Θ_d, Θ_s, Θ_r.

3.3. Light Source Optimization

Lighting Model The flashlight is modeled as a white point light source, similar to the configurations in [11, 84]. We assume the photographic equipment is in rigid capture setup, such that the relative positions of the camera and point lights are fixed across images.

Denoting the relative offset as a learnable parameter Δ_o, the incident light direction $\boldsymbol{w}_i(\boldsymbol{x})$ and intensity $L_i(\boldsymbol{w}_i; \boldsymbol{x})$ for a surface point $\boldsymbol{x}$ are defined as

$$\boldsymbol{w}_i(\boldsymbol{x}) = \frac{(\boldsymbol{o} + \Delta_o) - \boldsymbol{x}}{\|(\boldsymbol{o} + \Delta_o) - \boldsymbol{x}\|^2}, L_i(\boldsymbol{w}_i; \boldsymbol{x}) = \frac{L}{\|(\boldsymbol{o} + \Delta_o) - \boldsymbol{x}\|^2}, \tag{4}$$

where $\boldsymbol{o}$ is the camera center, and L represents the learnable scalar intensity of the light.

Visibility Computation By leveraging the object's geometry and the point light's position, we infer self-shadows and mitigate their effects on the captured images. To make the process differentiable, we sample N points along a direction from the point $\boldsymbol{x}$ to the light position, denoted as $\boldsymbol{w}_i$, and calculate the visibility of a point as [85]:

$$f_v(\boldsymbol{w}_i; \boldsymbol{x}) = 1 - \sum_{j=1}^{N} \alpha_j \prod_{k=1}^{j-1} \left(1 - \alpha_k\right), \tag{5}$$

where α_j is the discrete opacity values. We compute the visibility online so that the object geometry and light position can be jointly optimized.

3.4. Differentiable Inter-reflection Computation

Prior methods, like InvRender [86], sample multiple rays and use an MLP to cache the indirect incoming illumination at a surface point as smooth SGs under a static illumination, hindering their application in scenes dynamically captured under directional lighting, such as with a flashlight.

To address the issues of missing indirect illumination details and the high computational load, we propose an online indirect illumination computation strategy based on importance sampling (see supp. for pipeline details).

Importance Sampling Inter-reflection occurs when light reflects from one surface to another. We observe that the specular surfaces exhibit more pronounced inter-reflections, and the main source of indirect illumination for a surface point $\boldsymbol{x}$ comes from its reflective direction $\boldsymbol{w}_r$, which is the reflection of the view angle around the surface normal. To efficiently compute the indirect light, we sample multiple rays near $\boldsymbol{w}_r$ for indirect radiance calculation. Initially, we identify the secondary intersection point $\boldsymbol{x}'$ where the indirect bounce meets the surface, and then we compute the incoming radiance $L_{\text{ind}}(\boldsymbol{w}_i; \boldsymbol{x})$ as the outgoing radiance from $\boldsymbol{x}'$. Radiance rendering of the secondary intersection point only takes into account the intense lighting from the flashlight and $\boldsymbol{x}'$, excluding points occluded from the flashlight. The indirect illumination results from integrating all these incoming radiances over the upper-hemisphere around $\boldsymbol{x}$ surface normal.

To mitigate the artifacts of insufficient sampling, we blend the radiance from incoming direction $\boldsymbol{w}_i$ around the reflective view using a learnable scalar γ. One straightforward parameterization approach involves mapping the scalar from the point coordinates and the implicit geometric feature. However, we discovered that relying on point position and local geometry information often restricts the representation of varying inter-reflections, particularly in concave areas. A more effective strategy involves utilizing common physical properties (distance, view direction and roughness) to deduce dynamic indirect illumination. Consequently, we express the indirect illumination component $f_{ir}(\boldsymbol{x})$ as:

$$f_{ir}(\boldsymbol{x}) = \gamma \cdot \sum_{\boldsymbol{w}_i} L_{\text{ind}}(\boldsymbol{w}_i; \boldsymbol{x}) f_r(\boldsymbol{w}_i, \boldsymbol{w}_o; \boldsymbol{x})(\boldsymbol{w}_i \cdot \boldsymbol{n}). \quad (6)$$

3.5. DINO Regularization

To reduce the inherent ambiguity of reflectance estimation, we introduce a novel reflectance regularization based on the distilled DINO feature field. DINO [9] displays inherent capabilities for object decomposition by training on diverse unlabeled data, and has been successfully distilled into 3D fields for radiance editing [27] and open-vocabulary object grouping [26]. Inspired by these methods, we propose to distill the DINO feature from 2D images to 3D surfaces (object geometry) to learn a fine composition and contextual information of the object, resulting in a more consistent decomposition of surface reflectance and materials. In our implementation, we distill the DINO feature to the initial geometry field by minimizing the loss function:

$$\mathcal{L}_{\text{dino}} = \sum_{\boldsymbol{p}} (\boldsymbol{f}_{\text{dino}}(\boldsymbol{x}(\boldsymbol{p})) - \text{DINO}(\boldsymbol{p}))^2, \quad (7)$$

where $\boldsymbol{p}$ denotes the pixel in the 2D images, $\boldsymbol{x}(\boldsymbol{p})$ indicates the corresponding 3D surface point derived by ray tracing. The distillation process is minimizing the square distance between the learnable DINO feature $\boldsymbol{f}_{\text{dino}}$ on surface and the ViTs pre-trained on 2D images. The distilled DINO features are incorporated into the networks of specular albedo and roughness, providing regularization to enhance the accuracy of material decomposition.

In our experiment, we found the resolution of the DINO feature will also influence the ability of distinguishing the composition of objects. Higher resolution DINO features can assist in achieving finer material decoupling. Empirically, we upsample the image by a scaling factor of two to extract DINO features with a higher resolution.

3.6. Optimization

Differentiable surface point To make the surface point differentiable, we reparameterize the surface intersection equation as previous works [75, 84]:

$$\boldsymbol{x}_{\boldsymbol{\Theta}_g} = \boldsymbol{x} - \frac{\boldsymbol{n}}{\boldsymbol{n}^T \boldsymbol{n}} S_{\boldsymbol{\Theta}_g}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{n} S_{\boldsymbol{\Theta}_g}(\boldsymbol{x}), \quad (8)$$

where $S_{\boldsymbol{\Theta}_g}(\boldsymbol{x})$ denotes the SDF value of point $\boldsymbol{x}$, $\boldsymbol{n}$ is the normal vector at $\boldsymbol{x}$ calculated by $\boldsymbol{n} = \nabla S_{\boldsymbol{\Theta}_g}(\boldsymbol{x})$.

Training Loss The optimization process is formulated as a minimization problem where the total loss $\mathcal{L}$ is a combination of several components, each targeting a specific aspect of the reconstruction:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \mathcal{L}_{\text{ssim}} + \lambda_1 \mathcal{L}_{\text{eik}} + \lambda_2 \mathcal{L}_{\alpha} + \lambda_3 \mathcal{L}_{\text{smooth}} + \lambda_4 \mathcal{L}_{\text{dino}}. \quad (9)$$

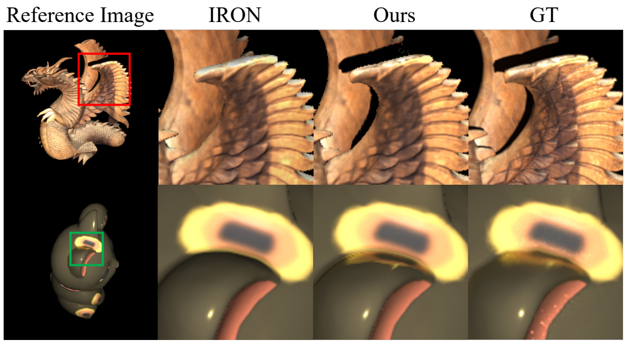


Figure 3. Visual results of self-shadows and inter-reflections.

$\mathcal{L}_{\text{rgb}}$ is the L_2 loss computed on the Gaussian pyramids of the predicted image $\hat{I}$ and the reference image I. $\mathcal{L}_{\text{SSIM}}$ is the SSIM loss [65]. $\mathcal{L}_{\text{eik}}$ is the Eikonal loss [18] to regularize the MLP for a valid SDF. $\mathcal{L}_{\alpha}$ is the roughness range loss, set at 0.5. The first four loss terms are the same as IRON [84]. $\mathcal{L}_{\text{smooth}}$ is the smoothness loss on the specular albedo and roughness as used by [74]. $\mathcal{L}_{\text{dino}}$ denotes the DINO feature alignment loss described in Eq. (7). During the inverse rendering stage, the edge-aware surface rendering proposed by IRON is adopted to refine the geometry [84].

4. Experiments

4.1. Datasets

Synthetic Data The synthetic dataset comprises six objects. Four objects with a variety of shapes and materials, namely *duck*, *maneki*, *horse*, and *dragon*, are used in IRON [84]. To conduct thorough analysis, we adopt another two objects: *marble bowl*, a concave bowl with complex light effects, and *armchair* with self-shadows in multiple views.

We consider a practical non-collocated flashlight and camera setting, similar to a mobile phone setup, with the angle between the camera and flashlight to the object center set to about 3 degrees. We render 200 images from random views under a non-collocated flashlight via Mitsuba [22] for training. We also render 100 images together with their diffuse albedo maps, specular albedo maps and roughness maps for test images to evaluate the quality of novel view synthesis and material decomposition.

Real Data We tested our method on the DRV dataset [5] captured by a nearly collocated camera-flashlight setup, and the Luan dataset [40] captured using a smartphone. We also captured a dataset by an iPhone in a darkroom. Camera poses for real images were derived using COLMAP [50].

4.2. Comparison with Existing Methods

For a fair comparison of material decomposition, we adapted the physical shader in WildLight to roughplastic model [61], as utilized by Mitsuba. We then integrated the

Table 1. Quantitative comparison of novel view rendering results with other inverse rendering methods on the synthetic dataset.

	Duck			*Maneki*			*Horse*			*Marble Bowl*			*Dragon*		
Method	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NeILF++ [82]	31.482	0.9775	0.0591	28.940	0.9498	0.0886	29.657	0.9640	0.0676	28.476	0.9336	0.0856	24.729	0.8986	0.1202
WildLight [12]	-	-	-	29.913	0.9399	0.0787	32.032	0.9669	0.0520	28.219	0.9252	0.0981	26.546	0.9078	0.1155
IRON [84]	31.845	0.9855	0.0320	30.087	0.9550	0.0468	31.713	0.9808	0.0366	27.403	0.9602	**0.0583**	25.516	0.9257	0.0876
Ours	**35.164**	**0.9912**	**0.0273**	**32.979**	**0.9729**	**0.0340**	**33.921**	**0.9851**	**0.0327**	**29.209**	**0.9640**	0.0591	**27.647**	**0.9391**	**0.0767**

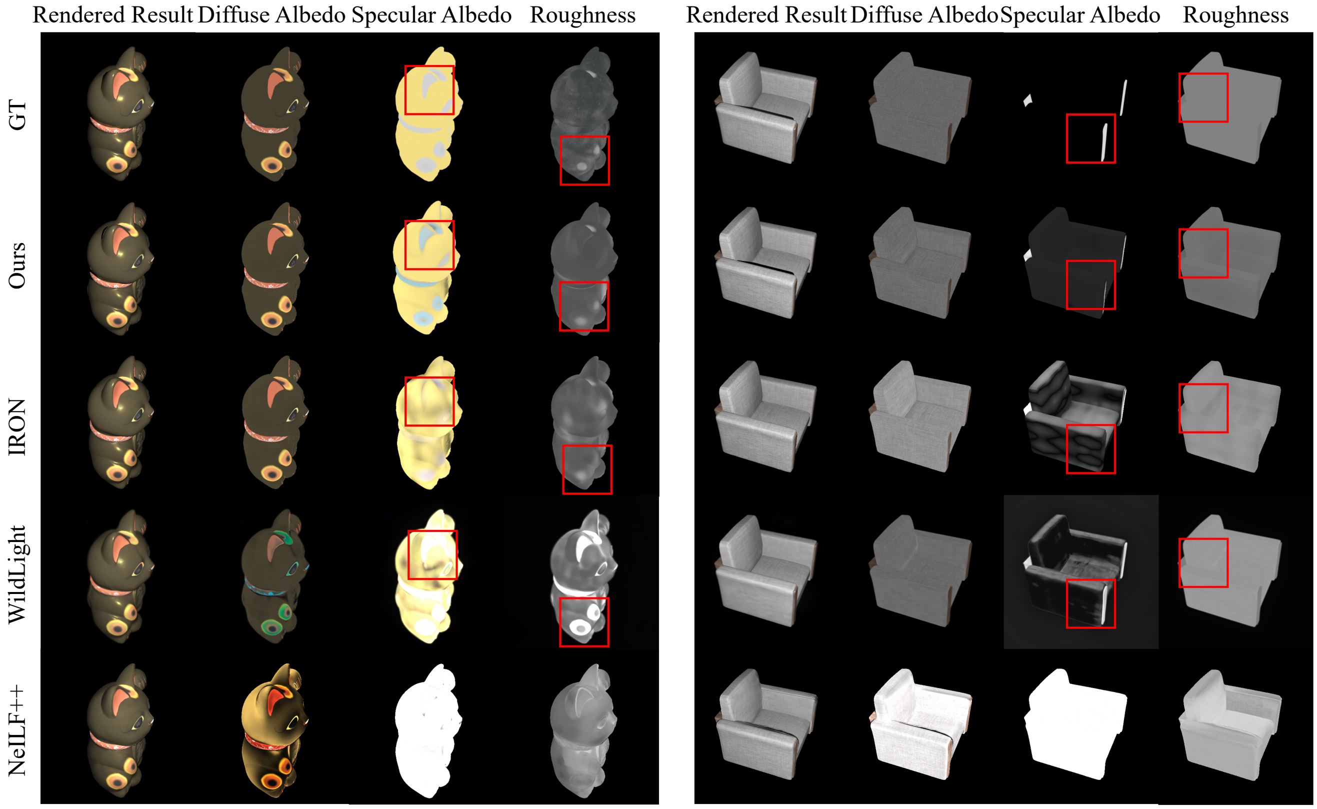


Figure 4. Qualitative comparison of state-of-art methods and our method on the synthetic dataset. The materials of NeILF++[82] are *Base Color*, *Metallic*, *Roughness* defined by simplified Disney principled BRDF and others are using Mitsuba roughplastic model.

decomposed material components along the rays to generate the material maps of the view. Comparison with the methods on volume rendering method DRV [5] and the mesh-based approach PSDR [40] were not conducted in our study, as their codes are not available. We also compared with other implicit methods presented by [82], which recover neural fields while considering inter-reflections.

Our method can recover sharp inter-reflection details and complex self-shadow caused by non-collocated camera and flashlight (see Fig. 3), resulting in more accurate specular albedo and roughness (see Fig. 4). Existing methods for the similar inverse rendering settings (*i.e.*, IRON [84] and WildLight [12]) overlook inter-reflections and self-shadows, leading to inaccuracy in material recovering. Specifically, the diffuse albedo often blends indirect illumination, particularly in concave areas. Self-shadows distort surface reflectance, leading to incorrect specular albedo brightness and noisy roughness. The state-of-the-art implicit method NeILF++ [82] tends to erroneously blend the intensity of moving light sources into material properties.

Table 1 and Table 2 show the quantitative comparison of rendering and material decomposition, respectively. We can see that our method achieves more accurate results, especially in the estimation of specular albedo and roughness.

4.3. Results on Real Data

We compare our method with IRON on the challenging real dataset. Figure 5 showcases the rendered images and material decomposition results. Compared with IRON, our method exhibits fewer shadows and indirect illumination effects baked into the diffuse albedo, given the credit to our modeling of inter-reflection and lighting optimization.

Specifically, in the diffuse albedo of *Xmen* estimated by IRON, the neck region bakes indirect illumination and appears brighter. In IRON's result on *Toy*, self-shadows distort the diffuse albedo, embedding shadows within the material.

Table 2. Quantitative comparison on synthetic data. The predicted albedos are scaled to match the GT light intensity when evaluation.

Method	Roughness MSE $\times 10^{-3}$	Diffuse Albedo PSNR	 SSIM	 LPIPS	Specular Albedo PSNR	 SSIM	 LPIPS	View Synthesis RGB PSNR	 SSIM	 LPIPS
WildLight [12]	106.32	25.631	0.9189	0.1186	17.357	0.8353	0.2016	29.167	0.9300	0.0929
IRON [84]	1.8402	33.175	0.9730	0.0432	25.809	0.8496	0.1645	29.053	0.9604	0.0558
Ours	**0.8808**	**35.777**	**0.9805**	**0.0331**	**29.716**	**0.9136**	**0.1065**	**31.891**	**0.9700**	**0.0488**

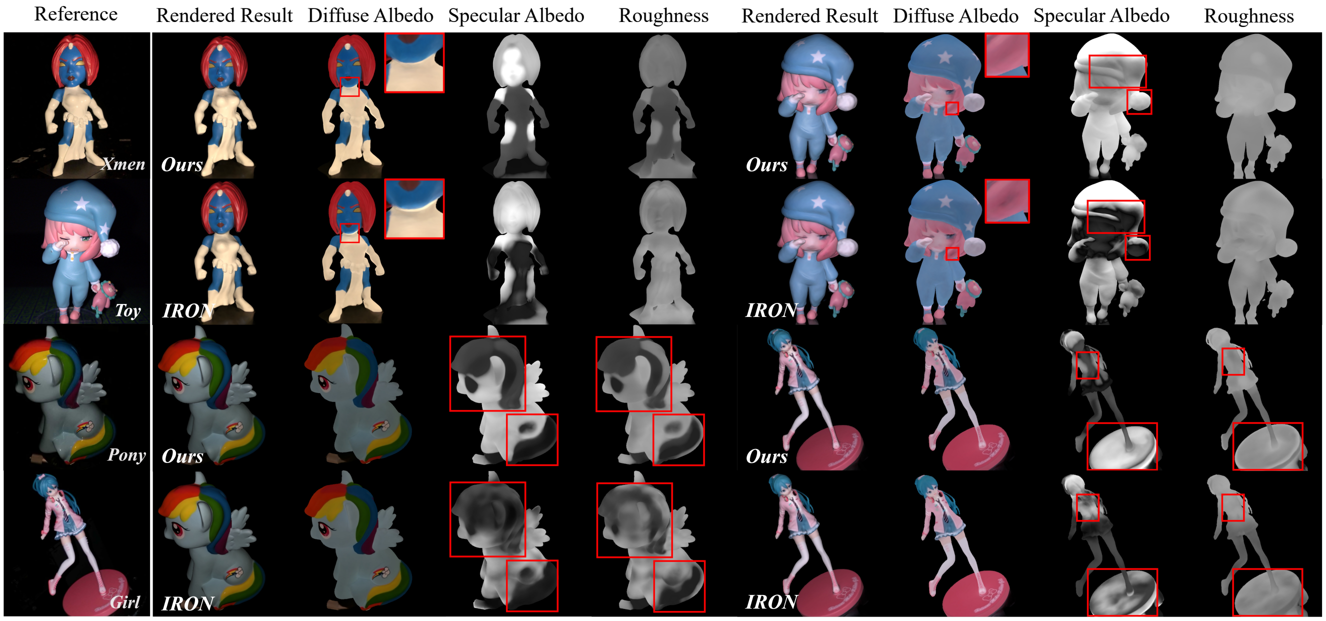


Figure 5. Visual results of material decomposition on real data. For each object, we compare the results of our method and IRON. *Xmen* is from Luan dataset [40], *Toy* is a self-captured dataset, *Pony* and *Girl* are from the DRV dataset [5].

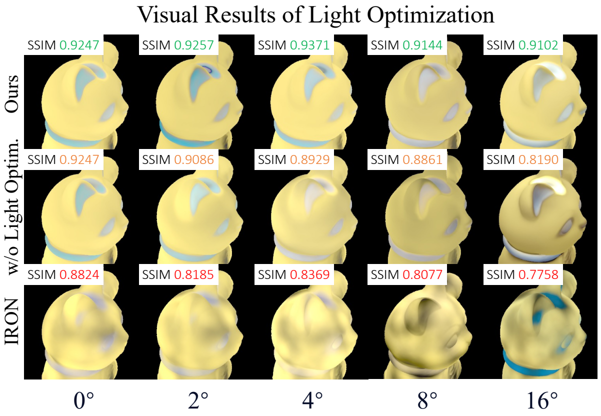


Figure 6. Quantitative and visual Results of Light Optimization.

With the DINO regularization, our method produces more consistent reflectance decomposition (see *Pony* and *Girl*).

4.4. Ablation Studies

To gain a deeper insight into the efficacy of our approach, we have conducted a thorough analysis of our method. We evaluate the inter-reflection modeling, lighting optimization, and DINO regularization, to validate our method.

Evaluation on Lighting Optimization We show the strength of our method in calibrating the camera-lighting offset even in the extreme case. The quantitative and qualitative comparison in Fig. 6 shows that, without the light position optimization, the method fails to accurately estimate materials with large light source deviation.

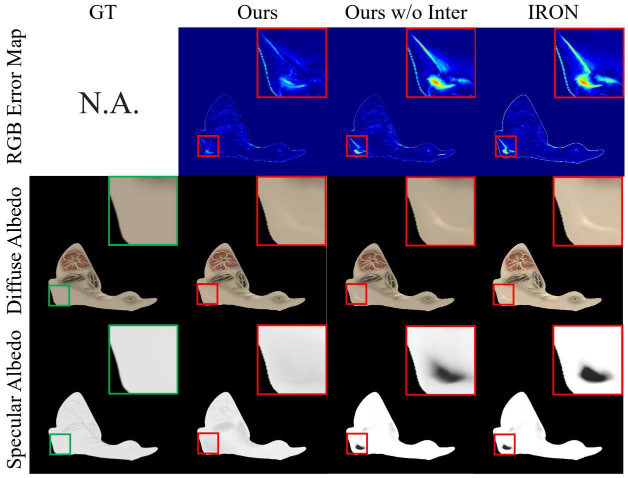
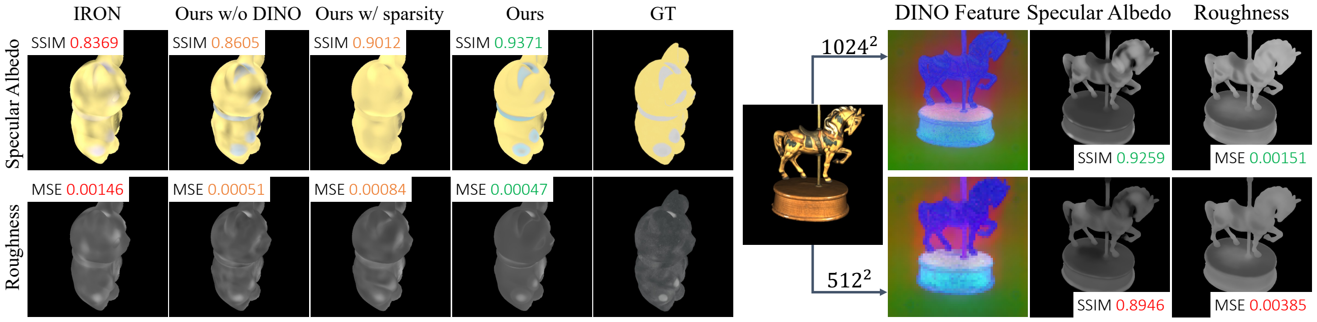


Figure 7. Ablation study on inter-reflection. We render scenes in simplified *collocated setting* without self-shadows.

Evaluation on Inter-reflection Modeling The synthetic data rendered for ablation study on inter-reflection is in col-

Table 3. Quantitative comparison of material estimation on synthetic dataset under collocated camera-lighting (collocated setup).

	Roughness	Diffuse Albedo			Specular Albedo			View Synthesis RGB		
Method	MSE $\times 10^{-3}$	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
WildLight [12]	104.62	27.449	0.7979	0.2320	19.518	0.8420	0.1747	30.520	0.8959	0.0943
IRON [84]	0.8264	33.616	0.9793	0.0419	31.231	0.9159	0.1084	33.827	0.9743	0.0405
Ours w/o Inter-reflect.	0.8112	34.966	0.9807	0.0318	32.500	0.9250	0.0951	34.382	0.9753	0.0393
Ours w/o f_{dino}	0.7638	34.806	0.9812	0.0327	32.526	0.9182	0.1005	34.294	0.9755	0.0393
Ours	**0.6226**	**35.075**	**0.9817**	**0.0316**	**33.128**	**0.9400**	**0.0854**	**34.804**	**0.9762**	**0.0391**



(a) Ablation study on DINO Regularization

(b) The impact of DINO feature resolution on material decoupling

Figure 8. Ablation study on DINO feature regularization (a) and the impact of DINO resolutions on material decoupling (b).

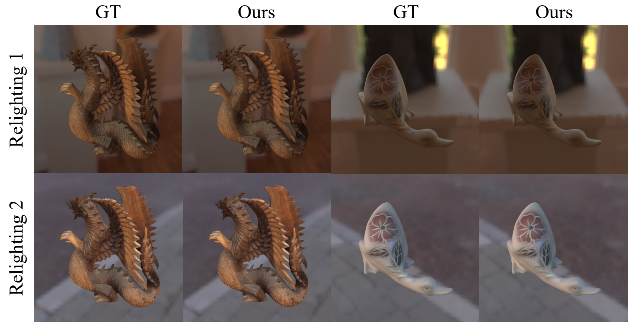


Figure 9. Relighting results with materials estimated our method.

located camera-lighting setting to overview self-shadows and different physical shader settings used in different methods. We ablate the inter-reflection calculation and compare the results in Table 3 and Fig. 7. Our method accurately estimates the materials and without the inter-reflection modeling, the predicted specular albedo has artifacts and diffuse albedo baked the indirect illumination.

Evaluation on DINO regularization Similarly, we evaluate the DINO regularization under the collocated camera-lighting setting for ablation study. Table 3 and Figure 8 (a) shows the quantitative and visual results respectively. Our decomposition results surpass other methods and we also show the DINO regularization is much better than some empirical regularization used in [14, 86]. The DINO feature regularization can ease the inherent difficulty of reflectance decomposition by grouping consistent contextual information across the scene.

In addition, Fig. 8 (b) shows the influence of DINO feature resolution in reflectance decomposition, validating that higher resolution of DINO features can assist in achieving finer material decoupling.

4.5. Relighting Results

We relight the objects with estimated material properties under two environments and show results in Fig. 9. This highlights our method's ability to precisely recover material properties, thereby enabling further relighting applications.

5. Conclusion and Discussion

In this paper, we present a new and effective inverse rendering approach for reconstructing object shapes, materials, and lighting from photometric images. Our method optimizes the light source position to account for self-shadows and employs an online strategy for modeling inter-reflections through a differentiable rendering layer. Additionally, we incorporate the DINO regularization to help the decomposition of surface reflectance. Extensive experiments on synthetic and real datasets demonstrate that our method can address misalignments between camera and light sources and surpass state-of-the-art methods in material decomposition. Due to its ability to finely compose objects, DINO regularization has the potential for application in other inverse rendering settings, such as those involving environment illumination.

Limitations Our method has several constraints. Primarily, we overlooks the consistency of novel views captured by a moving flashlight in the geometry initialization stage. Secondly, the BRDF model we employ is tailored for solid reflective surfaces, thus our approach is not suitable for reflective surfaces. We leave these for our future works.

References

[1] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[2] Hendrik Baatz, Jonathan Granskog, Marios Papas, Fabrice Rousselle, and Jan Novák. Nerf-tex: Neural reflectance field textures. In *CGF*, 2022. 2

[3] Mojtaba Bemana, Karol Myszkowski, Jeppe Revall Frisvad, Hans-Peter Seidel, and Tobias Ritschel. Eikonal fields for refractive novel-view synthesis. In *SIGGRAPH*, 2022. 2

[4] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. 3

[5] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *ECCV*, 2020. 3, 5, 6, 7, 2

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *CVPR*, 2021. 2

[7] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan T. Barron, and Hendrik P.A. Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, 2021. 1, 2

[8] Mohammed Brahimi, Bjoern Haefner, Tarun Yenamandra, Bastian Goldluecke, and Daniel Cremers. Supervol: Super-resolution shape and reflectance estimation in inverse volume rendering. *arXiv preprint arXiv:2212.04968*, 2022. 3

[9] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 4, 5

[10] Ziyu Chen, Chenjing Ding, Jianfei Guo, Dongliang Wang, Yikang Li, Xuan Xiao, Wei Wu, and Li Song. L-tracing: Fast light visibility estimation on neural surfaces by sphere tracing. In *ECCV*, 2022. 3

[11] Ziang Cheng, Hongdong Li, Yuta Asano, Yinqiang Zheng, and Imari Sato. Multi-view 3d reconstruction of a textureless smooth surface of unknown generic reflectance. In *CVPR*, pages 16226–16235, 2021. 1, 4

[12] Ziang Cheng, Junxuan Li, and Hongdong Li. Wildlight: In-the-wild inverse rendering with a flashlight. In *CVPR*, 2023. 3, 6, 7, 8, 2

[13] Changwoon Choi, Juhyeon Kim, and Young Min Kim. Iblnerf: Image-based lighting formulation of neural radiance fields. *arXiv preprint arXiv:2210.08202*, 2022. 2

[14] Youming Deng, Xueting Li, Sifei Liu, and Ming-Hsuan Yang. Dip: Differentiable interreflection-aware physics-based inverse rendering. *arXiv preprint arXiv:2212.04705*, 2022. 2, 8

[15] Yue Fan, Ivan Skorokhodov, Oleg Voynov, Savva Ignatyev, Evgeny Burnaev, Peter Wonka, and Yiqun Wang. Factored-neus: Reconstructing surfaces, illumination, and materials of possibly glossy objects. *arXiv preprint arXiv:2305.17929*, 2023. 2

[16] Jian Gao, Chun Gu, Youtian Lin, Hao Zhu, Xun Cao, Li Zhang, and Yao Yao. Relightable 3d gaussian: Real-time point cloud relighting with brdf decomposition and ray tracing. *arXiv preprint arXiv:2311.16043*, 2023. 2

[17] Wenhang Ge, Tao Hu, Haoyu Zhao, Shu Liu, and Ying-Cong Chen. Ref-neus: Ambiguity-reduced neural implicit surface learning for multi-view reconstruction with reflection. *arXiv preprint arXiv:2303.10840*, 2023. 2

[18] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[19] Heng Guo, Hiroaki Santo, Boxin Shi, and Yasuyuki Matsushita. Edge-preserving near-light photometric stereo with neural surfaces. *arXiv preprint arXiv:2207.04622*, 2022. 3

[20] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *CVPR*, pages 18409–18418, 2022. 2

[21] Yi-Hua Huang, Yan-Pei Cao, Yu-Kun Lai, Ying Shan, and Lin Gao. Nerf-texture: Texture synthesis with neural radiance fields. In *SIGGRAPH*, 2023. 2

[22] Wenzel Jakob. Mitsuba renderer, 2010. 5, 1

[23] Yingwenqi Jiang, Jiadong Tu, Yuan Liu, Xifeng Gao, Xiaoxiao Long, Wenping Wang, and Yuexin Ma. Gaussianshader: 3d gaussian splatting with shading functions for reflective surfaces. *arXiv preprint arXiv:2311.17977*, 2023. 2

[24] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *CVPR*, 2023. 2

[25] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 3

[26] Justin* Kerr, Chung Min* Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *ICCV*, 2023. 2, 5

[27] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, 2022. 2, 5

[28] Georgios Kopanas, Thomas Leimkühler, Gilles Rainer, Clément Jambon, and George Drettakis. Neural point catacaustics for novel-view synthesis of reflections. *TOG*, 2022. 2

[29] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: Neural rendering of objects from online image collections. *TOG*, 2022. 2

[30] Yuandong Li, Qinglei Hu, Zhenchao Ouyang, and Shuhan Shen. Neural reflectance decomposition under dynamic point light. *TCSVT*, 2023. 3

[31] Zhen Li, Lingli Wang, Mofang Cheng, Cihui Pan, and Jiaqi Yang. Multi-view inverse rendering for large-scale real-world indoor scenes. In *CVPR*, 2023. 2

[32] Ruofan Liang, Jiahao Zhang, Haoda Li, Chen Yang, Yushi Guan, and Nandita Vijaykumar. Spidr: Sdf-based neural point fields for illumination and deformation. *arXiv preprint arXiv:2210.08398*, 2022. 2

[33] Ruofan Liang, Huiting Chen, Chunlin Li, Fan Chen, Selvakumar Panneer, and Nandita Vijaykumar. Envidr: Implicit differentiable renderer with neural environment lighting. In *ICCV*, 2023. 2

[34] Zhihao Liang, Qi Zhang, Ying Feng, Ying Shan, and Kui Jia. Gs-ir: 3d gaussian splatting for inverse rendering. *arXiv preprint arXiv:2311.16473*, 2023. 2

[35] Zhi-Hao Lin, Jia-Bin Huang, Zhengqin Li, Zhao Dong, Christian Richardt, Tuotuo Li, Michael Zollhöfer, Johannes Kopf, Shenlong Wang, and Changil Kim. Iris: Inverse rendering of indoor scenes from low dynamic range images. *arXiv preprint arXiv:2401.12977*, 2024. 2

[36] Jingwang Ling, Zhibo Wang, and Feng Xu. Shadowneus: Neural sdf reconstruction by shadow ray supervision. In *CVPR*, pages 175–185, 2023. 3

[37] Isabella Liu, Linghao Chen, Ziyang Fu, Liwen Wu, Haian Jin, Zhong Li, Chin Ming Ryan Wong, Yi Xu, Ravi Ramamoorthi, Zexiang Xu, et al. Openillumination: A multi-illumination dataset for inverse rendering evaluation on real objects. *arXiv preprint arXiv:2309.07921*, 2023. 3

[38] R. Liu, S. Menon, C. Mao, D. Park, S. Stent, and C. Vondrick. What you can reconstruct from a shadow. In *CVPR*, pages 17059–17068, 2023. 3

[39] Yuan Liu, Peng Wang, Cheng Lin, Xiaoxiao Long, Jiepeng Wang, Lingjie Liu, Taku Komura, and Wenping Wang. Nero: Neural geometry and brdf reconstruction of reflective objects from multiview images. *arXiv preprint arXiv:2305.17398*, 2023. 2

[40] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *CGF*, 2021. 5, 6, 7, 2

[41] Alexander Mai, Dor Verbin, Falko Kuester, and Sara Fridovich-Keil. Neural microfacet fields for inverse rendering. In *ICCV*, 2023. 2

[42] Shi Mao, Chenming Wu, Zhelun Shen, and Liangjun Zhang. Neus-pir: Learning relightable neural surface using pre-integrated rendering. *arXiv preprint arXiv:2306.07632*, 2023. 2, 3, 4

[43] Wojciech Matusik, Hanspeter Pfister, Matt Brand, and Leonard McMillan. A data-driven reflectance model. In *TOG*, 2003. 2

[44] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421, 2020. 1, 2

[45] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting Triangular 3D Models, Materials, and Lighting From Images. In *CVPR*, 2022. 2

[46] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, pages 3504–3515, 2020. 1, 2

[47] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, pages 5589–5599, 2021. 2

[48] Jiaxiong Qiu, Peng-Tao Jiang, Yifan Zhu, Ze-Xin Yin, Ming-Ming Cheng, and Bo Ren. Looking through the glass: Neural surface reconstruction against high specular reflections. In *CVPR*, 2023. 2

[49] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *ECCV*, 2022. 2

[50] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, pages 4104–4113, 2016. 5, 2

[51] Yahao Shi, Yanmin Wu, Chenming Wu, Xing Liu, Chen Zhao, Haocheng Feng, Jingtuo Liu, Liangjun Zhang, Jian Zhang, Bin Zhou, et al. Gir: 3d gaussian inverse rendering for relightable scene factorization. *arXiv preprint arXiv:2312.05133*, 2023. 2

[52] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 32, 2019. 1, 2

[53] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021. 1, 2

[54] Cheng Sun, Guangyan Cai, Zhengqin Li, Kai Yan, Cheng Zhang, Carl Marshall, Jia-Bin Huang, Shuang Zhao, and Zhao Dong. Neural-pbir reconstruction of shape, material, and illumination. In *ICCV*, pages 18046–18056, 2023. 2

[55] Jiakai Sun, Zhanjie Zhang, Tianyi Chu, Guangyuan Li, Lei Zhao, and Wei Xing. Joint optimization of triangle mesh, material, and light from neural fields with neural radiance cache. *arXiv preprint arXiv:2305.16800*, 2023. 2

[56] Kushagra Tiwary, Tzofi Klinghoffer, and Ramesh Raskar. Towards learning neural representations from shadows. In *ECCV*, page 300–316, 2022. 3

[57] Kushagra Tiwary, Akshat Dave, Nikhil Behari, Tzofi Klinghoffer, Ashok Veeraraghavan, and Ramesh Raskar. Orca: Glossy objects as radiance-field cameras. In *CVPR*, 2023. 2

[58] Marco Toschi, Riccardo De Matteo, Riccardo Spezialetti, Daniele De Gregorio, Luigi Di Stefano, and Samuele Salti. Relight my nerf: A dataset for novel view synthesis and relighting of real world objects. In *CVPR*, 2023. 3

[59] 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*, 2022. 2

[60] Dor Verbin, Ben Mildenhall, Peter Hedman, Jonathan T Barron, Todd Zickler, and Pratul P Srinivasan. Eclipse: Disambiguating illumination and materials using unintended shadows. *arXiv preprint arXiv:2305.16321*, 2023. 3

[61] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206, 2007. 4, 5

[62] Dongqing Wang, Tong Zhang, and Sabine Süsstrunk. Nemto: Neural environment matting for novel view and

relighting synthesis of transparent objects. *arXiv preprint arXiv:2303.11963*, 2023. 2

[63] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *NeurIPS*, 2021. 2, 3, 4, 1

[64] Xi Wang, Zhenxiong Jian, and Mingjun Ren. Non-Lambertian photometric stereo network based on inverse reflectance model with collocated light. *TIP*, 29:6032–6042, 2020. 1

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: From error visibility to structural similarity. *TIP*, 2004. 5

[66] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representations for inverse rendering of urban scenes. In *CVPR*, 2023. 2

[67] Haoqian Wu, Zhipeng Hu, Lincheng Li, Yongqiang Zhang, Changjie Fan, and Xin Yu. Nefii: Inverse rendering for reflectance decomposition with near-field indirect illumination. In *CVPR*, 2023. 2

[68] Fanbo Xiang, Zexiang Xu, Milos Hasan, Yannick Hold-Geoffroy, Kalyan Sunkavalli, and Hao Su. Neutex: Neural texture mapping for volumetric neural rendering. In *CVPR*, 2021. 2

[69] Yingyan Xu, Gaspard Zoss, Prashanth Chandran, Markus Gross, Derek Bradley, and Paulo Gotardo. Renerf: Relightable neural radiance fields with nearfield lighting. In *ICCV*, 2023. 3

[70] Siqi Yang, Xuanning Cui, Yongjie Zhu, Jiajun Tang, Si Li, Zhaofei Yu, and Boxin Shi. Complementary intrinsics from neural radiance fields and cnns for outdoor scene relighting. In *CVPR*, 2023. 2

[71] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. In *ECCV*, 2022. 3

[72] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. S^3-nerf: Neural reflectance field from shading and shadow under a single viewpoint. In *NeurIPS*, 2022. 3

[73] Ziyi Yang, Yanzhen Chen, Xinyu Gao, Yazhen Yuan, Yu Wu, Xiaowei Zhou, and Xiaogang Jin. Sire-ir: Inverse rendering for brdf reconstruction with shadow and illumination removal in high-illuminance scenes. *arXiv preprint arXiv:2310.13030*, 2023. 2

[74] 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 *ECCV*, 2022. 2, 5

[75] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Ronen Basri, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, pages 2492–2502, 2020. 1, 2, 4, 5

[76] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 34: 4805–4815, 2021. 2

[77] Ze-Xin Yin, Jiaxiong Qiu, Ming-Ming Cheng, and Bo Ren. Multi-space neural radiance fields. In *CVPR*, pages 12407–12416, 2023. 2

[78] Kazuki Yoshiyama and Takuya Narihira. Ndjir: Neural direct and joint inverse rendering for geometry, lights, and materials of real object. *arXiv preprint arXiv:2302.00675*, 2023. 2

[79] Tizian Zeltner, Fabrice Rousselle, Andrea Weidlich, Petrik Clarberg, Jan Novák, Benedikt Bitterli, Alex Evans, Tomáš Davidovič, Simon Kallweit, and Aaron Lefohn. Real-time neural appearance models. *arXiv preprint arXiv:2305.02678*, 2023. 2

[80] Chong Zeng, Guojun Chen, Yue Dong, Pieter Peers, Hongzhi Wu, and Xin Tong. Relighting neural radiance fields with shadow and highlight hints. In *SIGGRAPH*, 2023. 3

[81] Junyi Zeng, Chong Bao, Rui Chen, Zilong Dong, Guofeng Zhang, Hujun Bao, and Zhaopeng Cui. Mirror-nerf: Learning neural radiance fields for mirrors with whitted-style ray tracing. *arXiv preprint arXiv:2308.03280*, 2023. 2

[82] Jingyang Zhang, Yao Yao, Shiwei Li, Jingbo Liu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf++: Inter-reflectable light fields for geometry and material estimation. *ICCV*, 2023. 2, 6, 3

[83] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. PhySG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, 2021. 1, 2

[84] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *CVPR*, 2022. 1, 3, 4, 5, 6, 7, 8, 2

[85] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *TOG*, 2021. 1, 2, 3, 4

[86] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, 2022. 2, 4, 8

[87] Youjia Zhang, Teng Xu, Junqing Yu, Yuteng Ye, Yanqing Jing, Junle Wang, Jingyi Yu, and Wei Yang. Nemf: Inverse volume rendering with neural microflake field. In *ICCV*, 2023. 2

[88] Mingwu Zheng, Haiyu Zhang, Hongyu Yang, and Di Huang. Neuface: Realistic 3d neural face rendering from multi-view images. In *CVPR*, pages 16868–16877, 2023. 2

[89] Hongliang Zhong, Jingbo Zhang, and Jing Liao. Vq-nerf: Neural reflectance decomposition and editing with vector quantization. *arXiv preprint arXiv:2310.11864*, 2023. 2

[90] Licheng Zhong, Lixin Yang, Kailin Li, Haoyu Zhen, Mei Han, and Cewu Lu. Color-neus: Reconstructing neural implicit surfaces with color. *arXiv preprint arXiv:2308.06962*, 2023. 2

[91] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, et al. I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. In *CVPR*, 2023. 2

[92] Shizhan Zhu, Shunsuke Saito, Aljaz Bozic, Carlos Aliaga, Trevor Darrell, and Christop Lassner. Neural relighting with subsurface scattering by learning the radiance transfer gradient. *arXiv preprint arXiv:2306.09322*, 2023. 3

[93] Xiangyang Zhu, Yiling Pan, Bailin Deng, and Bin Wang. Efficient multi-view inverse rendering using a hybrid differentiable rendering method. *arXiv preprint arXiv:2308.10003*, 2023. 2

Photometric Inverse Rendering: Shading Cues Modeling and Surface Reflectance Regularization

Supplementary Material

1. More Details for the Method

1.1. Network Architectures

Neural SDF: $f_{\Theta_g}(\boldsymbol{x}) = (s, \boldsymbol{f}_{\text{geo}})$. We employ an 8-layer MLP featuring a hidden dimension of 256 and incorporate a skip connection at the fourth layer. The network input is the 3D coordinate $\boldsymbol{x}$ encoded with a frequency of 6, to output the SDF value and an implicit local geometric feature. Before optimization, we perform geometric initialization on the network, as described by [1].

Neural diffuse albedo: $f_{\Theta_d}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{n}, \boldsymbol{f}) = \rho_d$. We use an 8-layer MLP featuring a hidden dimension of 256 and a skip connection at the fourth layer. The network inputs include the 3D coordinate $\boldsymbol{x}$ encoded with 10 frequencies, surface normal, and geometric feature. It outputs the diffuse albedo for point $\boldsymbol{x}$.

Neural specular albedo: $f_{\Theta_s}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{f}) = \rho_s$. We employ a 4-layer MLP with a width of 256. The input 3D coordinate $\boldsymbol{x}$ is encoded using 6 frequencies.

Neural roughness: $f_{\Theta_r}(\boldsymbol{x}, \boldsymbol{n}, \boldsymbol{f}_{\text{geo}}) = \rho_r$. We deploy a 4-layer MLP with a width of 256. The input 3D coordinate $\boldsymbol{x}$ is encoded using 6 frequencies.

Blending scalar: $\gamma\left(\|\boldsymbol{x} - \boldsymbol{x}'\|, \boldsymbol{w}_i \cdot \boldsymbol{n}, \rho_r\right) = \gamma$. We use a 4-layer MLP with width 128. The the dot product of normal and view direction uses 6 frequencies.

Neural DINO feature: $f_{\Theta_{\text{dino}}}(\boldsymbol{x}) = \boldsymbol{f}_{\text{dino}}$ We utilize a 4-layer MLP with a width of 256, where the input location $\boldsymbol{x}$ is encoded with 6 frequencies, and the output features a dimension of 384.

1.2. Visibility Computation

For joint optimization of object geometry and light position, we determine the visibility of a surface point $\boldsymbol{x}$ by uniformly sampling $N = 128$ points $\{\boldsymbol{x}_i\}_{i=1}^{N}$ along the path from surface point $\boldsymbol{x}$ to the light source. We obtain the discrete opacity values $\{\alpha\}_{i=1}^{N}$ for these points using the unbiased SDF density conversion method introduced by NeuS [63]:

$$\alpha_i = \max\left(\frac{\Phi_s\left(f\left(\mathbf{p}\left(t_i\right)\right)\right) - \Phi_s\left(f\left(\mathbf{p}\left(t_{i+1}\right)\right)\right)}{\Phi_s\left(f\left(\mathbf{p}\left(t_i\right)\right)\right)}, 0\right). \tag{10}$$

The light visibility of point $\boldsymbol{x}$ in the direction of incident light $\boldsymbol{w}_i$ is represented by the residual transmittance:

$$f_v\left(\boldsymbol{w}_i; \boldsymbol{x}\right) = 1 - \sum_{j=1}^{N} \alpha_j T_j, \tag{11}$$

where α_j is the density value at point $\boldsymbol{x}_j$, and $T_j = \prod_{k=1}^{j-1}\left(1 - \alpha_k\right)$ is the light transmittance at point $\boldsymbol{x}_j$ in the direction $\boldsymbol{w}_i$.

1.3. Inter-reflection Computation

Importance Sampling To model the indirect illumination in scenes dynamically captured with directional lighting, we introduce an online computation approach that combines a differentiable layer and an importance sampling strategy. For a point $\boldsymbol{x}$ and view direction $\boldsymbol{w}_i$, we consider a single light bounce and employ ray marching towards the reflective direction:

$$\boldsymbol{w}_r = 2 \times \boldsymbol{n} - \boldsymbol{w}_i. \tag{12}$$

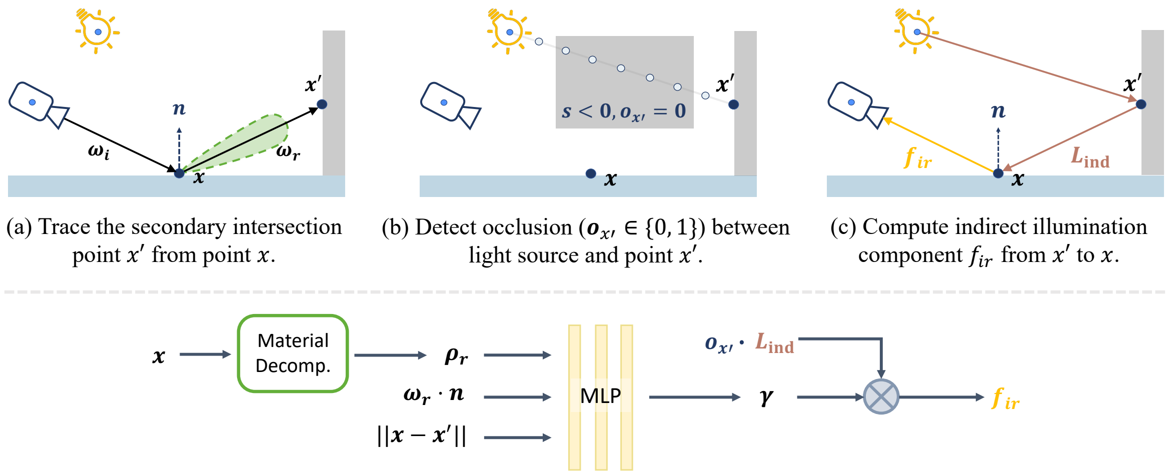
We then identify the secondary intersection point $\boldsymbol{x}'$. To determine if $\boldsymbol{x}'$ is occluded from the light source, we uniformly sample 20 points along the path between the light source and the intersection point. The light is considered occluded by another surface if any of the sampled points exhibit a negative SDF value.

Figure 10 illustrates the process of inter-reflection modeling. If the secondary intersection point $\boldsymbol{x}'$ is unobstructed, we compute the outgoing radiance at $\boldsymbol{x}'$ using the flashlight's incoming radiance. The outgoing result is then combined with the blending coefficient to represent the indirect illumination.

Gradient Backpropogation Given that the blending coefficient is conditioned on the roughness property of the 3D point, there exists a correlation between the roughness property and the blending coefficient, introducing additional ambiguity in material estimation. In our experiments, we found that detaching the roughness of the secondary intersection point $\boldsymbol{x}'$ prior to its input into the blending coefficient network leads to a more precise material decomposition. Moreover, the process of gradient backpropagation starts from the image loss, through the residual component and into the material networks of the secondary intersection point $\boldsymbol{x}'$, fosters the alignment of secondary point radiance with inter-reflection cues. In our experiments, we discovered that disabling the optimization of local geometry at the secondary point reduces the complexity of the optimization process, leading to improved geometric reconstruction, especially in concave areas.

1.4. BRDF Renderer

Our BRDF implementation closely adheres to the Mitsuba roughplastic BRDF model [22], with the distribution pa-



(a) Trace the secondary intersection point x' from point x.

(b) Detect occlusion ($o_{x'} \in \{0, 1\}$) between light source and point x'.

(c) Compute indirect illumination component f_{ir} from x' to x.

(d) Computation details of blending coefficient γ and indirect illumination component f_{ir}.

Figure 10. The illustration of inter-reflection modeling. We take into account rays that are physically rendered at secondary intersection points near the reflective direction of point x, as these rays contribute significantly to the indirect illumination.

rameter specifically set to ‘ggx’. For simplicity, we refer to our configuration as the roughplastic model. Default values are maintained for the internal and external Indices of Refraction and the nonlinear parameter.

Previous methods (IRON [84] and WildLight [12]) employing the renderer relied on an oversimplified BRDF model within an idealized setting where the camera and flashlight are collocated, neglecting the deviations between the camera and light source present in our capture setup. To address this limitation, we have enhanced the simplified roughplastic model to accommodate a broader range of scenarios, allowing for variations in both incident and outgoing light directions.

1.5. Training Details

The training process requires approximately 9 hours on a single RTX3090 GPU with 24GB of memory. We start by training NeuS over 100,000 iterations to initialize the geometry and diffuse albedo networks. For each training iteration, we utilize 512 randomly sampled pixels, employing an ℓ_1 loss along with an eikonal regularization loss. Prior to the rendering phase, we derive the feature maps of images by the pre-trained ViT-S/8 model [9] and executed 10,000 iterations with λ_4 set to 1.0 to extract the DINO feature from 2D feature maps to 3D surfaces. During the physics-based surface rendering stage with total 50,000 iterations, we fixed the geometry and lighting to warmup the BRDFs network for 2,000 iterations to stablize the process, and subsequently, we carried out a joint optimization of the lighting, geometry and BRDFs. The training of blending coefficient network started at the 10,000th iteration. We set the size of rendered image patch as 128×128 and loss weights to $\lambda_1 = 10^{-4}$, $\lambda_2 = 0.1$, $\lambda_3 = 10^{-5}$ and $\lambda_4 = 10^{-5}$. All networks are optimized by corresponding Adam optimizers with learning rate 10^{-4}.

2. More Details for the Dataset

DRV Dataset We acquired the DRV dataset [5] from the authors, comprising five scenes: *Dragon*, *Girl*, *Pony*, *Tree*, and *Cartoon*. Each scene has approximately 400 images, split between training and test sets. The dataset captures images in a darkroom, utilizing a nearly collocated camera-light setup.

Luan Dataset The Luan dataset [40] was captured using a casual smartphone. We noticed that the images exhibit significant noise and motion blur, together with varying exposure times and white balance settings during capture. This inconsistency introduces challenges in maintaining multi-view consistency. We evaluated the scene *Xmen*, which includes 136 images, to compare novel view rendering and material decomposition against the IRON method.

Self-captured Dataset For capturing real-world images, we employed an iPhone 15 to shoot in RAW format, ensuring a linear camera response. Across all photos, we maintained consistent settings for the camera’s exposure time, focus, and white balance. Specifically, the ISO value and shutter speed (exposure time) were fixed at 100 and 1/250s, respectively, with the white balance adjusted to 3,800 Kelvin degrees. Our collection encompasses 5 scenes: *Toy*, *fruit*, *Panda*, *Assassin* and *Bear*, with the number of images per scene varying from 120 to 400. Camera poses were derived using COLMAP [50], and objects were scaled to fit within a unit sphere based on the reconstructed point cloud. The photography sessions took place in a darkroom, positioning the camera 0.15 to 0.3 meters away from

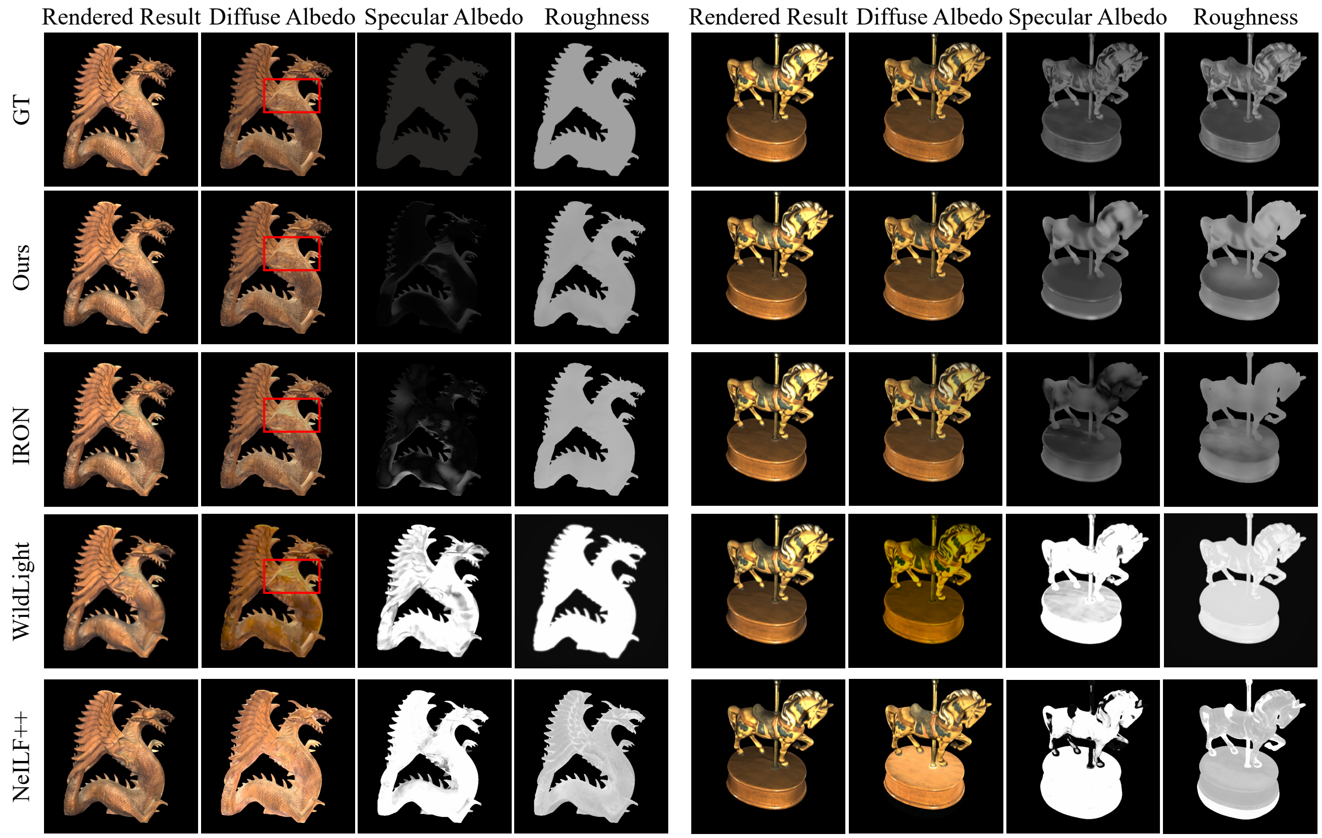


Figure 11. Qualitative comparisons with state-of-art methods on the synthetic dataset (*dragon* and *horse*). The materials of NeILF++[82] are *Base Color*, *Metallic*, *Roughness* defined by simplified Disney principled BRDF.

each object. To achieve comprehensive coverage, we systematically moved the camera in a spiral pattern around the subjects. The separation between the camera lens and the flashlight on the iPhone, roughly 0.015m, results in an approximate 3-degree variation between viewing and lighting angles at a standard distance of 0.25m from the object.

3. More Results and Comparisons

We primarily compare our results with those from IRON [84] and WildLight [12]. Notably, WildLight was unable to reconstruct the synthetic data for *duck*, and as such, its results are not presented in the table within the main paper.

3.1. Results on Synthetic Data

In Table 4, we offer comprehensive results for each synthetic scene captured under casual conditions. Additionally, we provide a qualitative comparison of novel view rendering and material decomposition between our method and earlier methods, as illustrated in Fig. 11. For the *dragon* scene, our method produces a diffuse albedo with less indirect illumination incorporated into the materials. In the *horse* scene, our material decomposition results demonstrate a reduced influence of self-shadows, showing a closer alignment with the ground truth than those obtained with IRON. Even in extreme concave regions, our method is more robust than previous method as shown in Fig. 12.

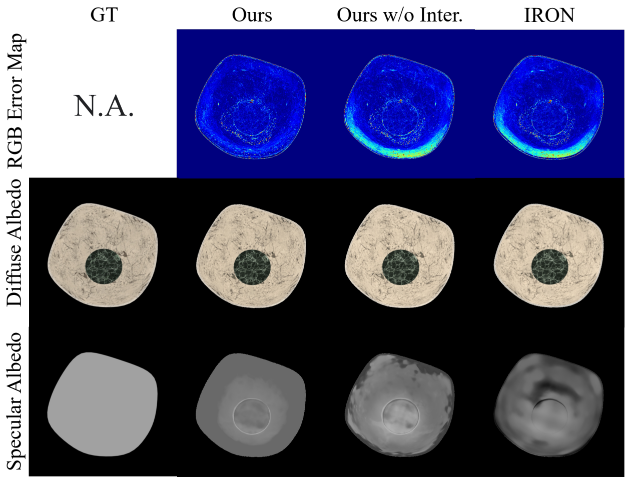


Figure 12. Ablation study on inter-reflection in *marble bowl*.

3.2. Results on Real Data

In Fig. 14, we present our dataset's novel view rendering and material decomposition outcomes. The IRON method

Scene	Method	Roughness	Diffuse Albedo			Specular Albedo			Novel View Synthesis		
		MSE $\times 10^{-3} \downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
duck	WildLight	-	-	-	-	-	-	-	-	-	-
	IRON	2.640	24.693	0.9631	0.0483	18.669	0.9017	0.1075	31.845	0.9855	0.0320
	Ours	**1.059**	**34.871**	**0.9852**	**0.0355**	**23.402**	**0.9474**	**0.0730**	**35.164**	**0.9912**	**0.0273**
maneki	WildLight	93.59	18.151	0.7351	0.4472	12.413	0.8114	0.2497	29.913	0.9400	0.0787
	IRON	1.455	35.367	0.9805	0.0238	18.967	0.8369	0.1732	30.087	0.9550	0.0468
	Ours	**0.467**	**36.098**	**0.9880**	**0.0184**	**22.245**	**0.9371**	**0.0726**	**32.979**	**0.9729**	**0.0340**
horse	WildLight	40.23	24.625	0.9507	0.1032	16.997	0.8401	0.2056	32.032	0.9669	0.0520
	IRON	2.198	31.903	0.9826	0.0363	29.323	0.8701	0.1275	31.713	0.9808	0.0366
	Ours	**1.509**	**33.573**	**0.9880**	**0.0194**	**33.071**	**0.9259**	**0.0917**	**34.206**	**0.9831**	**0.0321**
marble bowl	WildLight	165.2	22.613	0.8862	0.1379	15.600	0.8261	0.2135	28.219	0.9252	0.0981
	IRON	0.321	29.258	0.9623	0.0518	35.035	0.8947	0.1553	27.403	0.9602	**0.0583**
	Ours	**0.172**	**29.881**	**0.9647**	**0.0493**	**39.972**	**0.9729**	**0.0660**	**29.209**	**0.9640**	0.0591
dragon	WildLight	120.8	33.679	0.9208	0.1108	14.432	0.7840	0.2453	26.546	0.9078	0.1155
	IRON	2.815	36.470	0.9675	0.0575	30.902	0.7610	0.2295	25.516	0.9257	0.0876
	Ours	**0.923**	**38.720**	**0.9735**	**0.0390**	**34.894**	**0.8152**	**0.1772**	**27.870**	**0.9406**	**0.0766**
armchair	WildLight	117.9	30.116	0.9242	0.1282	15.748	0.8260	0.2136	29.126	0.9100	0.1200
	IRON	1.612	41.361	0.9818	0.0416	21.960	0.8333	0.1938	27.752	0.9555	0.0734
	Ours	**1.155**	**41.518**	**0.9833**	**0.0370**	**25.442**	**0.8959**	**0.1438**	**31.916**	**0.9655**	**0.0642**

Table 4. Complete results on the synthetic dataset.

	Pony		*Girl*		*Tree*		*Dragon*		*Cartoon*		Average	
Method	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
IRON [84]	29.269	0.9150	27.136	0.9326	31.641	**0.9464**	**32.421**	**0.9317**	30.773	0.9587	30.248	0.9369
Ours	**30.092**	**0.9414**	**27.589**	**0.9365**	**31.765**	**0.9464**	32.251	0.9306	**30.975**	**0.9589**	**30.534**	**0.9428**

Table 5. Quantitative comparison of novel view rendering on DRV dataset.

often incorporates indirect illumination into the diffuse albedo, particularly in concave regions, as observed in the *Fruit* scene. Additionally, specular albedoes produced by IRON method are adversely affected by self-shadows and inter-reflections, as highlighted in specific boxes.

In Table 5, we provide a quantitative comparison that underscores the enhanced performance of our method compared to IRON in terms of novel view rendering within the DRV real dataset. Figure 15 and Fig. 16 complement this with side-by-side qualitative comparisons of our method against IRON regarding to material decomposition. Leveraging DINO regularization for surface decomposition, which effectively clusters similar materials, our approach produces more accurate results for material decomposition, especially in scenarios with a skewed view distribution. We observe that IRON's evaluation metrics for the *Dragon* scene slightly exceed those of our method, this disparity is primarily due to its collocated camera-lighting setup, which inherently minimizes the occurrence of self-shadows within the scene.

3.3. Failure Case

Like many neural surface reconstruction methods, both COLMAP and NeuS presuppose Lambertian observation to guarantee multi-view consistency. Following the same approach as IRON, our method primarily depends on NeuS for geometry initialization but struggles to reconstruct objects with reflective surfaces, as depicted in Fig. 13. The surfaces reconstructed by NeuS and our method exhibit holes within reflective regions.

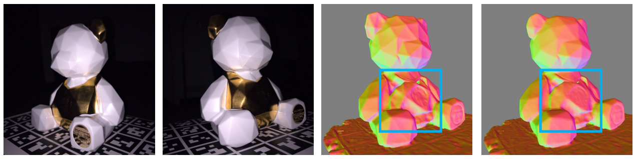

Reference Images NeuS normal Our normal

Figure 13. A failure case on *bear* with reflective surfaces.

4. Video Demos

In the video, we present more comprehensive results to demonstrate the effectiveness of our design, along with additional comparison cases between our method and other inverse rendering methods. Furthermore, we render the reconstructed 3D assets using a traditional graphical pipeline to illustrate their practical applications.

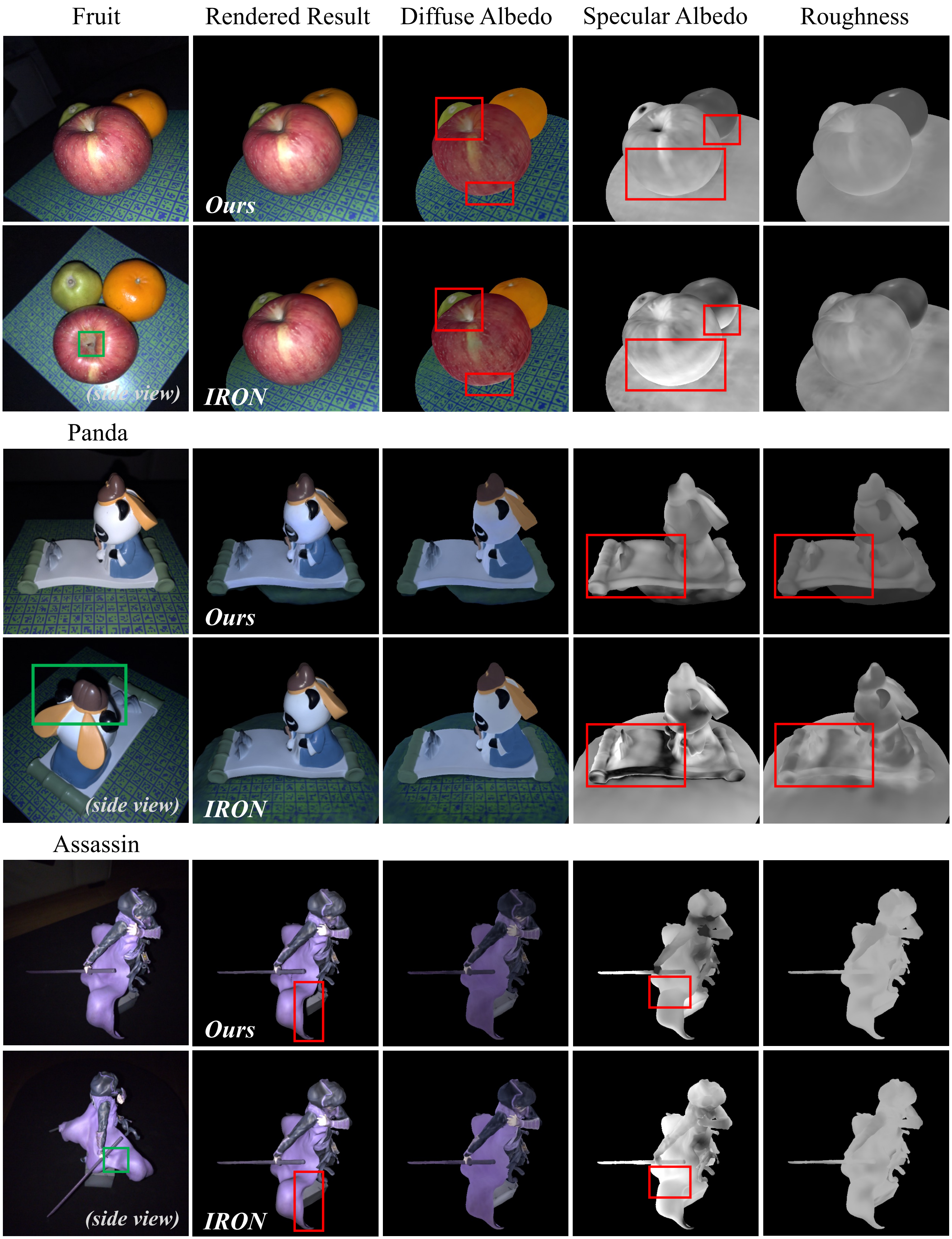


Figure 14. More visual results of material decomposition on our dataset.

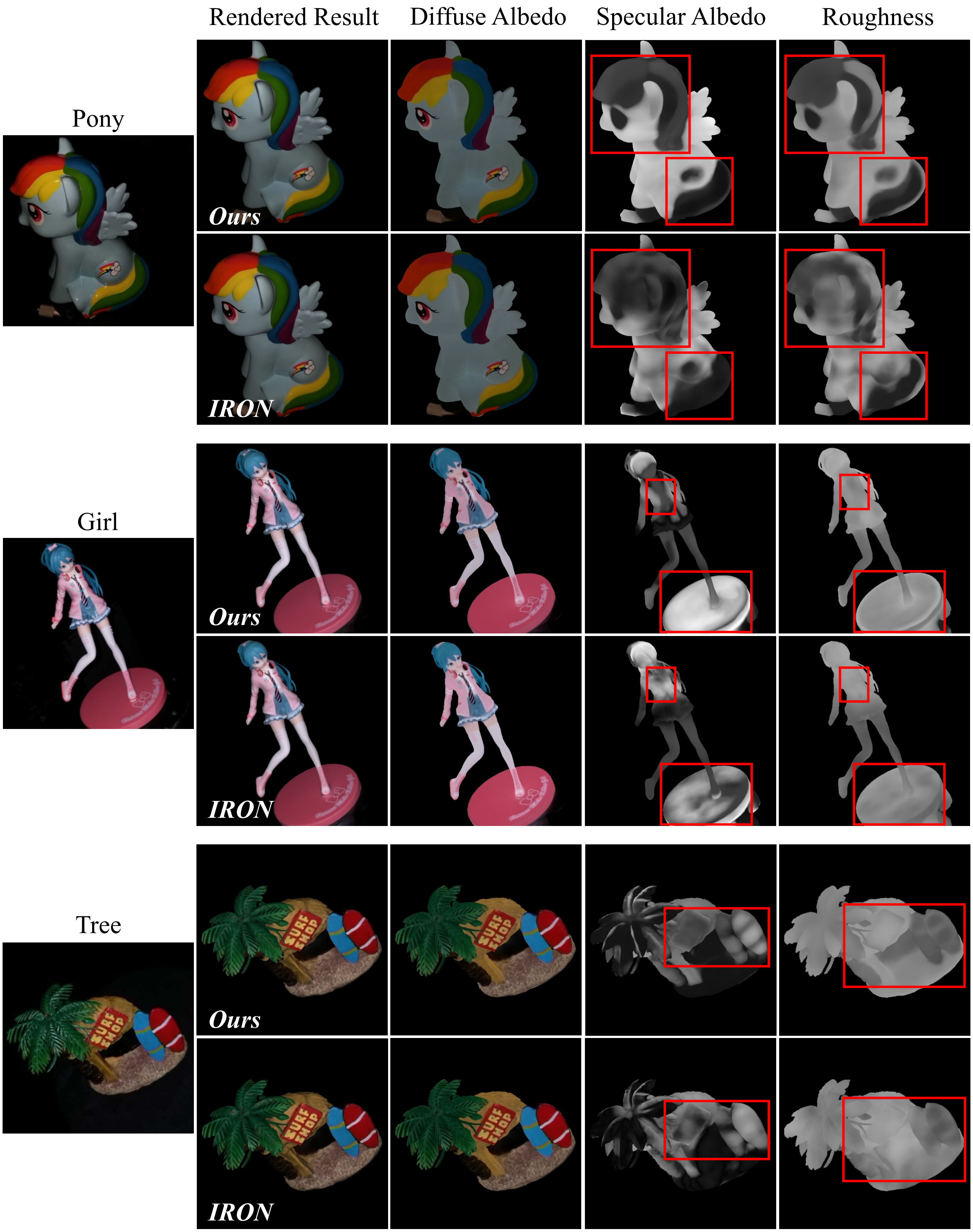


Figure 15. Visual results of material decomposition on DRV dataset.

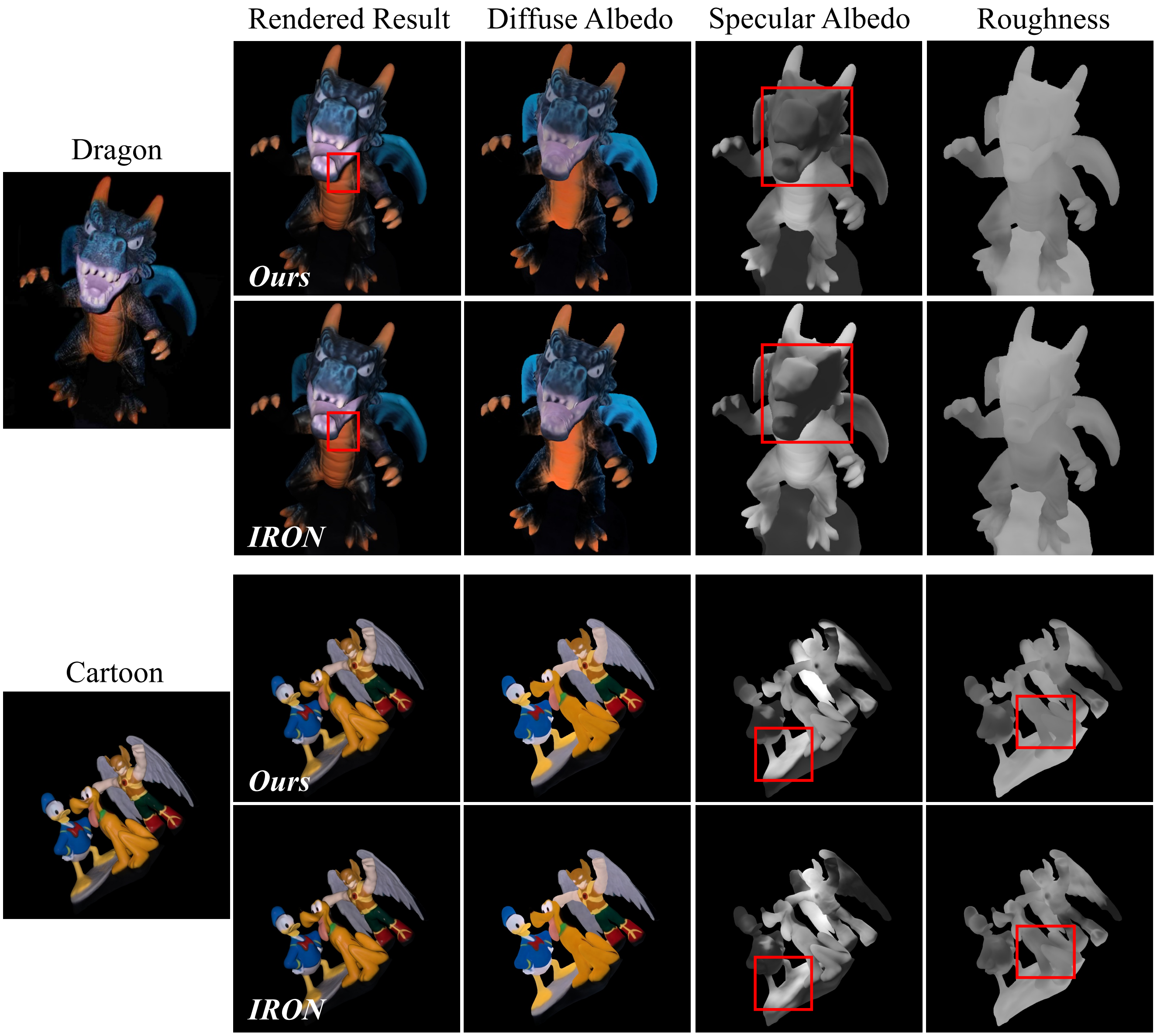


Figure 16. Visual results of material decomposition on DRV dataset (continued).