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

Zhen Li\(^1\)  Lingli Wang\(^1\)  Mofang Cheng\(^1\)  Cihui Pan\(^{1,*}\)  Jiaqi Yang\(^{2,*}\)

\(^1\)Realsee  \(^2\)Northwestern Polytechnical University

yodlee@mail.nwpu.edu.cn, {wanglingli008, chengmofang001, pancihui001}@realsee.com, jqyang@nwpu.edu.cn

Abstract

We present an efficient multi-view inverse rendering method for large-scale real-world indoor scenes that reconstructs global illumination and physically-reasonable SVBRDFs. Unlike previous representations, where the global illumination of large scenes is simplified as multiple environment maps, we propose a compact representation called Texture-based Lighting (TBL). It consists of 3D mesh and HDR textures, and efficiently models direct and infinite-bounce indirect lighting of the entire large scene. Based on TBL, we further propose a hybrid lighting representation with precomputed irradiance, which significantly improves the efficiency and alleviates the rendering noise in the material optimization. To physically disentangle the ambiguity between materials, we propose a three-stage material optimization strategy based on the priors of semantic segmentation and room segmentation. Extensive experiments show that the proposed method outperforms the state-of-the-art quantitatively and qualitatively, and enables physically-reasonable mixed-reality applications such as material editing, editable novel view synthesis and relighting. The project page is at https://lzleejean.github.io/TexIR.

1. Introduction

Inverse rendering aims to reconstruct geometry, material and illumination of an object or a scene from images. These properties are essential to downstream applications such as scene editing, editable novel view synthesis and relighting. However, decomposing such properties from the images is extremely ill-posed, because different configurations of such properties often lead to similar appearance. With recent advances in differentiable rendering and implicit neural representation, several approaches have achieved significant success on small-scale object-centric scenes with explicit or implicit priors [7,32,43,50,51,55,57,58]. However, inverse rendering of large-scale indoor scenes has not been well solved.

*Co-corresponding authors.

There are two main challenges for large-scale indoor scenes. 1) \textit{Modelling the physically-correct global illumination.} There are far more complex lighting effects, \textit{e.g.}, inter-reflection and cast shadows, in large-scale indoor scenes than object-centric scenes due to complex occlusions, materials and local light sources. Although the widely-used image-based lighting (IBL) is able to efficiently model direct and indirect illumination, it only represents the lighting of a certain position [13, 17, 18, 42]. The spatial consistency of per-pixel or per-voxel IBL representations [26,29,47,59] is difficult to ensure. Moreover, such incompact representations require large memory. Parameterized lights [16,27] such as point light, area light and directional light are naturally globally-consistent, but modeling the expensive global light transport will be in-
evitable [1,37,58]. Thus, simple lighting representations applied in previous works are unsuitable in large-scale scenes.

2) Disentangling the ambiguity between materials. Different configurations of materials often lead to similar appearance, and to add insult to injury, there are an abundance of objects with complex and diverse materials in large-scale scenes. In object-centric scenes, dense views distributed on the hemisphere are helpful for alleviating the ambiguity [14, 22, 32, 35, 55, 56]. However, only sparse views are available in large-scale scenes, which more easily lead to ambiguous predictions [37, 51].

In this work, we present TexIR, an efficient inverse rendering method for large-scale indoor scenes. Aforementioned challenges are tackled individually in the following.

1) We model the infinite-bounce global illumination of the entire scene with a novel compact lighting representation, called TBL. The TBL is able to efficiently represent the infinite-bounce global illumination of any position within the large scene. Such a compact and explicit representation provides more physically-accurate and spatially-varying illumination to guide the material estimation. Directly optimizing materials with TBL leads to expensive computation costs caused by high samples of the monte carlo sampling. Therefore, we precompute the irradiance based on our TBL, which significantly accelerates the expensive computation in the material optimization process. 2) To ameliorate the ambiguity between materials, we introduce a segmentation-based three-stage material optimization strategy. Specifically, we optimize a coarse albedo based on Lambertian-assumption in the first stage. In the second stage, we integrate semantics priors to guide the propagation of physically-correct roughness in regions with same semantics. In the last stage, we fine-tune both albedo and roughness based on the priors of semantic segmentation and room segmentation. By leveraging such priors, physically-reasonable albedo and roughness are disentangled globally.

To summarize, the main contributions of our method are as follows:

1. A compact lighting representation for large-scale scenes, where the infinite-bounce global illumination of the entire large scene can be handled efficiently.

2. A segmentation-based material optimization strategy to globally and physically disentangle the ambiguity between albedo and roughness of the entire scene.

3. A hybrid lighting representation based on the proposed TBL and precomputed irradiance to improve the efficiency in the material optimization process.

2. Related Work

Image-based lighting. Debevec et al. [13] first introduced the high dynamic range (HDR) light probe as a omni-directional lighting representation of a certain position, called IBL, which plays a vital role on inverse rendering and lighting estimation tasks in the computer vision. Barron and Malik [3] fitted the spherical-harmonic (SH) illumination, a parameterized representation of IBL, from a single image. Then they extended this single SH illumination into a set of SH illumination [2]. Zhou et al. [59] predicted the per-pixel SH lighting from a single image. Further, the per-pixel spherical-gaussian (SG) lighting [26] and the per-pixel light probe [29] are applied. Wang et al. [47] proposed a per-voxel SG lighting in the 3D indoor scene. The spatial consistency of illumination is unable to ensure with above lighting representations, especially for 3D large-scale scenes. Although light probes are able to project consistent light probes in others positions with known geometry, the efficiency in the optimization process is limited [25]. Our proposed TBL not only reserves the advantage of IBL, modelling infinite-bounce global illumination efficiently, but also is naturally globally-consistent.

Parametric lighting. Parametric lights, such as point lights, spot lights, area lights and directional lights, are classical lighting representations to define the illumination of a scene in computer graphics. Most methods only use one of the above lighting to represent the illumination of a scene. Nestmeyer et al. [36] predicted a directional light from a face image. Junxuan Li and Hongdong Li [24] used different directional lights as their illumination setting. Several approaches [6, 32, 43, 55] model the illumination of a static object with point lights. Li et al. [27] predicted area lights and SG directional lights for a single indoor image. Zhang et al. [54] optimized a more complex configuration via vertex-based lighting, including point lights, line lights, area lights and a environment map, but this method require user inputs to label where is the light source. Our TBL leverages the HDR texture of the entire scene to represent the illumination at any position within this 3D scene without any user input. Moreover, the TBL is compact and naturally globally-consistent.

Implicit lighting. With great advances in implicit representation, researchers began to use implicit neural networks to represent scenes. In particular, NeRF [33] has shown impressive results on scene representation. It leverages a neural density field and a neural radiance field to model a static small-scale scene. Zhang et al. [58] obtained the incident radiance of a certain position from a pre-trained outgoing radiance field [52]. To go a step further, the neural incident radiance field (NIRF), which represents incident radiance from any direction at any position, have shown great potential for inverse rendering [51] and novel view synthesis [45]. However, without constraints for such powerful implicit representation, the ambiguity between material and lighting is difficult to disentangle. Our explicit TBL is capable of eliminating this ambiguity in a interpretable manner.
The deep neural network has achieved significant success in many vision tasks with large-scale real-world datasets, such as object detection, semantic segmentation and depth estimation. Unfortunately, collecting the annotation of materials in real-world is difficult at scale. IIW [4] only labels sparse pairwise reflectance comparison for a single image. Therefore, most methods [8, 26, 28–30, 42, 47, 60] are alternative to use synthetic dataset to disambiguate the ambiguous properties. These approaches training on synthetic datasets struggle to eliminate the inevitable domain gap though several datasets have achieved photo-realistic results [29, 31, 38]. Optimization-based methods [2, 3, 7, 34, 37, 43, 56, 58] have shown impressive results on real-world multi-view images. It is known that the prior of different materials is essential to be leveraged in the optimization process to alleviate the ambiguity between materials. Barron and Malik [3] applied a smoothness term for albedo. Multiple variants of filters are used to smooth material in follow-up works [5, 26, 29]. Zhang et al. [57] leveraged learning priors from real-world BRDFs and imposed a spatial smoothness prior in the latent space. A similar idea has been adopted by a recent work [58]. Schmitt et al. [40] and Luan et al. [32] assumed similar diffuse albedos to have similar specular ones, and applied a non-local bilateral regularizer for specular albedo. Compared to object-centric scenes [7, 34, 43, 55–58], disentangling the ambiguity between materials is more challenging in the large-scale scene due to sparser observations and wider variety of objects and reflectance properties. Zhang et al. [34] recovered a constant attribute for each floor, wall and ceiling of an empty room. Most similar to ours, Nimier-David et al. [37] and Haefner et al. [20] also assumed similar semantic regions to have similar roughness and specularity. The first difference is we optimize the roughness in a soft manner instead optimizing a constant value [20, 37]. The another difference is we leverage the room segmentation prior, which enables us to recover different roughness even for similar semantic regions. Thanks to our efficient global illumination representation and explicit optimization strategy, our method is 20× faster than the differentiable path tracing-based method [37].

3. Methodology

As shown in Fig. 2, given a set of calibrated HDR images of a large-scale indoor scene, our method aims to accurately recover globally-consistent illumination and Spatially-Varying Bidirectional Reflectance Distribution Functions (SVBRDFs), which can be conveniently integrated into graphics pipelines and downstream applications. To this end, we propose TBL to represent the global illumination of large-scale indoor scenes (Sec. 3.1). In order to improve the optimizing efficiency and quality of re-rendered images in the material estimation stage, we adopt a hybrid lighting representation based on our TBL (Sec. 3.2). To reconstruct physically-reasonable SVBRDFs, we present a segmentation-based three-stage material estimation strategy (Sec. 3.3), which can handle ambiguity of materials in complex large-scale indoor scenes very well.

3.1. Texture-based Lighting

We address the issue of how to represent the illumination of a large-scale indoor scene with TBL. The advantages of TBL respectively are the compactness of neural representation, the global illumination of IBL, and the interpretability and spatial consistency of parametric lights. The proposed TBL, which is a global representation for the entire scene, defines the outgoing radiance for all surface points. We assume that only the diffuse lighting exists in the scene since the diffuse lighting often dominates
the scene, similar to radiosity-based methods [48, 49, 53]. Therefore, the outgoing radiance of one point is typically equal to the value of the HDR texture, i.e., the observed HDR radiance of corresponding pixels in input HDR images.

We initially reconstruct the mesh model of a entire large-scale scene with off-the-shelf classical MVS techniques, e.g., colmap [41]. Finally, we reconstruct the HDR texture based on the input HDR images. Therefore, the global illumination is queried from any direction at any position through the HDR texture, as shown in Fig. 3.

3.2. Hybrid Lighting Representation

According to our proposed TBL, the render equation [21] can be rewritten as follow:

\[ L_o(x, \omega_o) = \int_{H^+} f_r(x, \omega_i, \omega_o) Q(x, \omega_i, G, T_{hdr})(\omega_i, n) d\omega_i \]

where \( H^+ \) denotes hemisphere; \( x \) denotes a surface point; \( \omega_i \) denotes inverse incident light direction; \( \omega_o \) denotes view direction; \( n \) denotes normal; \( f_r \) denotes the BRDF of point \( x \); \( Q \) denotes the HDR lighting of the intersection point between the known geometry \( G \) and the ray \( r(t) = x + t\omega_i \), queried by HDR textures \( T_{hdr} \).

In practice, we calculate the Monte Carlo numerical integration with importance sampling [23]. To decrease the variance, a large sample number seems to be inevitable, which will significantly increase computation cost and memory cost in the optimization process. Inspired by precomputed radiance transfer [19], we precompute irradiance of surface points for the diffuse component. Therefore, the irradiance is queried efficiently without expensive online computation as shown in Fig. 3 (right) [48]. The Eq.1 can be rewritten as:

\[ L_o(x, \omega_o) = L_d(x, \omega_o) + L_s(x, \omega_o) \]

where \( L_d \) denotes the diffuse component and \( L_s \) denotes the specular component. The \( f_r \) of \( L_s \) would be modified as \( f_s \). The formulation can be found in the supplementary material.

We propose two representations to model precomputed irradiance. One is a neural irradiance field (NIrF), a shallow Multi-Layer-Perceptrons (MLP). It takes a surface point \( p \) as input and outputs the irradiance of \( p \). Another one is a irradiance texture (IrT), similar to the light map widely used in computer graphics. Based on such hybrid lighting representation consists of precomputed irradiance for the diffuse component and source TBL for the specular component, the rendering noise significantly decrease and the materials are optimized efficiently. Therefore, the diffuse component of Eq. 2 can be modeled as Eq. 3.

\[ L_d(x, \omega_o) = f_d(x) I_r(x) \]

where \( f_d \) denotes the diffuse BRDF of point \( x \) and \( I_r \) denotes the irradiance of point \( x \).

3.3. Segmentation-based Material Estimation

Instead of optimizing neural material [7, 43, 50, 51, 56–58], which is hard to model a large-scale scene with extremely complex materials and is mismatched to the traditional graphics engines, we directly optimize explicit material textures of the geometry.

We use the simplified Disney BRDF model [9] with Spatially-Varying (SV) albedo and SV roughness as parameters. Optimizing explicit material textures straightforwardly leads to inconsistent and unconverged roughness due to sparse observations [58]. We address this problem by leveraging the priors of semantics and room segmentation. The semantic images are predicted by learning-based models [11] and room segmentation is calculated by the occupancy grid [15]. Our segmentation-based strategy has three phases. Details of each stage are described in the following subsections. 

Stage 1: albedo initialization. We optimize a coarse albedo based on Lambertian assumption instead of initializing the albedo as a constant, which is widely used in object-centric scenes [32]. According to estimated illumination in Sec. 3.2, we can directly calculate the albedo through Eq. 3. However, it recovers over-bright albedo on the highlight regions, which leads to high roughness in the next stage. Therefore, we apply a semantic smoothness constraint to encourage that the albedo/roughness/feature is as close to the mean within the class as possible:

\[ L_{ss} = \sum_c \left| F - \frac{\sum_p F \odot M_{seg}(c)}{\sum_p M_{seg}(c)} + \epsilon \right| \odot M_{seg}(c) \]

where \( L_{ss} \) denotes the semantic smoothness loss; \( c \) denotes one of the classes of semantics; \( F \) denotes the feature image to be smoothed and we use the image-space diffuse albedo \( A \) in this stage; \( p \) denotes the pixel of \( F \); \( \odot \) is an element-wise product; \( M_{seg} \) denotes the mask of semantic segment-
Table 1. Quantitative comparison on our synthetic dataset. Our method significantly outperforms the state-of-the-arts in roughness estimation. NeILF* [51] denotes source method with their implicit lighting representation.

<table>
<thead>
<tr>
<th>Method</th>
<th>Albedo</th>
<th>Roughness</th>
<th>Novel view synthesis</th>
<th>Re-rendering</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>PSNR↑</td>
<td>SSIM↑</td>
<td>MSE↓</td>
<td>PSNR↑</td>
</tr>
<tr>
<td>PhyIR [29]</td>
<td>11.9726</td>
<td>0.6880</td>
<td>0.0635</td>
<td>12.5468</td>
</tr>
<tr>
<td>InvRender [58]</td>
<td>16.9760</td>
<td>0.6305</td>
<td>0.0201</td>
<td>9.1806</td>
</tr>
<tr>
<td>NVDIFFREC [34]</td>
<td>21.2551</td>
<td>0.8100</td>
<td>0.0075</td>
<td>7.6269</td>
</tr>
<tr>
<td>NeILF* [51]</td>
<td>14.2137</td>
<td>0.5184</td>
<td>0.0379</td>
<td>11.5778</td>
</tr>
<tr>
<td>NeILF [51]</td>
<td>17.0707</td>
<td>0.6489</td>
<td>0.0196</td>
<td>11.1654</td>
</tr>
<tr>
<td>Ours</td>
<td>20.4169</td>
<td>0.8514</td>
<td>0.0091</td>
<td>20.2132</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>34.2669</td>
</tr>
</tbody>
</table>

\[ L_{albedo} = |I - L_d| + \beta_{ssa} L_{ss} \]  (5)

where \( \beta_{ssa} \) denotes the weight of semantic smoothness loss for albedo, and \( I \) denotes the input HDR images.

**Stage II: VHL-based sampling and semantics-based propagation.** In multi-view images, only sparse specular cues are observed, which lead to globally inconsistent roughness [58], especially for large-scale scenes. As shown in Baseline of Fig. 9 and NVDIFFREC [34] of Fig. 6, only the roughness of highlight regions is optimized reasonably. Therefore, by leveraging the prior of semantic segmentation, the reasonable roughness of highlight regions would be propagated into the regions with same semantics.

We first render images based on input poses with roughness 0.01 to find virtual highlight (VHL) regions for each semantic class. Then, we optimize the roughness on these VHL regions according to the frozen coarse albedo and illumination. Meanwhile, the reasonable roughness can be propagated into the same semantic segmentation through Eq. 6:

\[ L_{sp} = \sum_c R - \text{quantile}(R \odot M_{shl}(c), q) \odot (M_{seg}(c) - M_{shl}(c)) \]  (6)

where \( L_{sp} \) denotes the semantics-based propagation loss; \( R \) denotes the image-space roughness; \( \text{quantile} \) denotes the \( q \)-th quantiles on VHL regions and we set the \( q \) as 0.4 for robustness; \( M_{shl} \) denotes the mask of VHL regions for each class. The roughness can be optimized by:

\[ L_{roughness} = |I - L_o| + \beta_{sp} L_{sp} \]  (7)

**Stage III: Segmentation-based fine-tuning.** In the last phase, we fine-tune all material textures based on the priors of semantic segmentation and room segmentation. Specifically, we apply a similar smoothness constraint to Eq. 4 and a room smoothness constraint for roughness, which makes the roughness of different rooms smoother in a soft manner. The room smoothness constraint is formulated by Eq. 8:

\[ L_{rs} = \sum_c \left| R - \frac{\sum_p R \odot M_{room}(c)}{\sum_p M_{room}(c)} + \epsilon \right| \odot M_{room}(c) \]  (8)

where \( L_{rs} \) denotes the room segmentation-based smoothness loss; \( M_{room} \) denotes the mask of different room segmentation; \( c \) denotes one of the index of each room.

We do not apply any smoothness constraint for albedo. The total loss is defined as:

\[ L_{all} = |I - L_o| + \beta_{ssr} (L_{ss} + L_{rs}) \]  (9)

where \( \beta_{ssr} \) denotes the weight of segmentation smoothness loss for roughness and we use the image-space roughness \( R \) in \( L_{ss} \) in Stage III.

4. Experiments

4.1. Datasets

**Synthetic dataset.** We create a synthetic scene with diverse material and light sources with a path tracer [26]. We render 24 views for optimization and 14 views as novel views, and render Ground Truth material images for each view. The details of the scene can be found in the supplementary material.

**Real dataset.** Widely used real datasets of large-scale scenes, e.g., ScanNet [12], Matterport3D [10] and Replica [44], lack full-HDR images. Therefore, we collect 10 full-HDR real scenes. For each scene, 10 to 20 full-HDR panoramic images are captured by merging 7 bracketed exposures (from \( \frac{1}{25000} \)s to \( \frac{1}{8} \)s).

4.2. Baselines and Metrics

To our best knowledge, there are only a few methods that recover the SVBRDFs from multi-view images for large-scale scenes. We compare with the following inverse rendering approaches: (1) The state-of-the-art single image learning-based method: PhyIR [29]; (2) The state-of-the-art multi-view object-centric neural rendering methods: InvRender [58], NVDIFFREC [34] and NeILF [51]. Please note that these object-centric approaches are unsuitable to
evaluate on large-scale scenes due to simple illumination representation except for NeILF. For fair comparisons, we integrate their material optimization strategies with our hybrid lighting representation, which is designed specifically for large-scale real-world scenes.¹

We use Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [46] and Mean Squared Error (MSE) to evaluate the material predictions and the re-rendered images for quantitative comparisons. Moreover, we use Mean Absolute Error (MAE) and SSIM to evaluate the relighting images rendered by different lighting representations.

### 4.3. Comparisons

**Evaluation on synthetic dataset.** As shown in Tab. 1 and Fig. 4. Our method significantly outperforms the state-of-the-arts in roughness estimation and our roughness is able to produce physically-reasonable specular reflectance. Moreover, NeILF [51] with our hybrid lighting representation successfully disentangles the ambiguity between material and lighting, compared to their implicit representation.

**Evaluation on real dataset.** More importantly, we conduct the experiment on our challenging real dataset containing complex materials and illumination. The quantitative comparison in Tab. 2 shows our approach outperforms previous methods. Although these methods have approximate re-rendering error, only our proposed approach disentangles globally-consistent and physically-reasonable SVBRDFs while other approaches struggle to produce consistent results and disentangle ambiguity of materials. Note that the low roughness (around 0.15 in ours) leads to the strong highlights, which are similar to GT.

---

¹We tried to compare with optimization-based methods [20, 37], but failed to get available results due to the lack of available source code.

**Table 2. Quantitative comparison of re-rendered images on our real dataset.**

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>MSE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>InvRender [58]</td>
<td>21.9993</td>
<td>0.7668</td>
<td>0.0065</td>
</tr>
<tr>
<td>NVDIFFREC [34]</td>
<td>23.7464</td>
<td>0.8389</td>
<td>0.0044</td>
</tr>
<tr>
<td>NeILF [51]</td>
<td>21.9260</td>
<td>0.7687</td>
<td>0.0066</td>
</tr>
<tr>
<td>Ours</td>
<td><strong>24.6093</strong></td>
<td><strong>0.8623</strong></td>
<td><strong>0.0035</strong></td>
</tr>
</tbody>
</table>

---

**Figure 4. Qualitative comparison on synthetic dataset.** Our method is able to produce realistic specular reflectance. NeILF* [51] denotes source method with their implicit lighting representation.

**Figure 5. Qualitative comparison in the 3D view on challenging real dataset.** This sample is Scene 1. Our method reconstructs globally-consistent and physically-reasonable SVBRDFs while other approaches struggle to produce consistent results and disentangle ambiguity of materials.
Figure 6. **Qualitative comparison in the image view on challenging real dataset.** From left to right: Scene 8 and Scene 9. Red denotes the Ground Truth image. Our physically-reasonable materials are able to render similar appearance to GT. Note that Invrender [58] and NeILF [51] do not produce correct highlights, and NVDIFFREC [34] fails to distinguish the ambiguity between albedo and roughness.

Table 3. **Ablation study of roughness estimation on synthetic dataset.**

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>MSE↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.8012</td>
<td>0.5268</td>
<td>0.1659</td>
</tr>
<tr>
<td>w/o Stage I</td>
<td>10.3561</td>
<td>0.7044</td>
<td>0.0921</td>
</tr>
<tr>
<td>w/o Stage II</td>
<td>7.9627</td>
<td>0.5570</td>
<td>0.1599</td>
</tr>
<tr>
<td>w/o Stage III</td>
<td>17.4177</td>
<td>0.8347</td>
<td>0.0181</td>
</tr>
<tr>
<td>Ours</td>
<td>20.2132</td>
<td>0.9161</td>
<td>0.0095</td>
</tr>
</tbody>
</table>

Table 4. **Quantitative comparison of relighting spheres.** We use 5th order SH coefficients and 12 SG lobes for the comparison.

<table>
<thead>
<tr>
<th>Lighting</th>
<th>Diffuse</th>
<th>Matte Silver</th>
<th>Mirror Silver</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MAE↓</td>
<td>SSIM↑</td>
<td>MAE↓</td>
</tr>
<tr>
<td>SH</td>
<td>0.0602</td>
<td>0.9982</td>
<td>0.0811</td>
</tr>
<tr>
<td>SG</td>
<td>0.0027</td>
<td>0.9995</td>
<td>0.0054</td>
</tr>
<tr>
<td>TBL</td>
<td><strong>0.0021</strong></td>
<td><strong>0.9994</strong></td>
<td><strong>0.0028</strong></td>
</tr>
</tbody>
</table>

segmentation-based fine-tuning in Stage III.

**Effectiveness of TBL.** We compare the proposed TBL to SH lighting and SG lighting widely used in previous methods [18, 26, 47, 59]. As shown in Fig. 7, our TBL exhibits high fidelity in high-frequency features. Moreover, we evaluate the relighting error of three re-rendering virtual spheres rendered by different lighting representations in Tab. 4. Except for accuracy, the TBL only costs around 20 MB storage to represent illumination while the dense grid-based VSG lighting [47] costs around 1 GB storage and the sparse grid-based SH lighting, Plenoxels [39], costs around 750 MB storage. Therefore, our TBL achieves improved accuracy while being compact in storage.

**Effectiveness of Hybrid Lighting Representation.** We compare the hybrid lighting representation described in Sec. 3.2 to source TBL. As shown in Fig. 8, without hybrid lighting representation, the albedo leads to noise and converges slowly. With precomputed irradiance, we can use high resolution inputs to recover detailed materials, and significantly accelerate the optimization process. The IrT produces more detailed and artifacts-free albedo, compared to the NIrF. Furthermore, we also compare to implicit lighting in Tab. 1 and Fig. 4. NeILF [51] with their implicit lighting
fails in disentangling the ambiguity between materials and lighting, e.g., The lighting effects are incorrectly recovered as materials in Fig. 4.

**Effectiveness of the Three Stage Strategy.** The results are shown in Tab. 3 and Fig. 9. The roughness of baseline fails to converge and only the highlight regions are updated. Without albedo initialization in Stage I, albedo in highlight regions is over-bright and leads to incorrect roughness. VHL-based sampling and semantics-based propagation in Stage II is crucial to recover the reasonable roughness of areas where highlights are not observed. Segmentation-based fine-tuning in Stage III produces detailed albedo, makes final roughness smoother and prevent the wrong propagation of roughness between different materials.

**4.5. Applications**

Our final output is a triangle mesh with PBR material textures, which is compatible with standard graphic engines and 3D modeling tools. We demonstrate in Fig. 1 that the proposed approach is able to produce convincing results on material editing, editable novel view synthesis and relighting. Moreover, we show several results of editable novel view synthesis in Fig. 10. Note that the view-dependent specular highlights reasonably change as view changes. See the supplementary material for more results.

**5. Conclusion**

In this paper, we propose a novel inverse rendering framework that recovers globally-consistent lighting and materials from posed sparse-view images and geometry for large-scale scenes. Our texture-based lighting, which not only represents infinite-bounce global illumination but also is compact and globally-consistent, is suitable for modelling illumination of large-scale scenes. Our material optimization strategy leveraging semantics and room segmentation priors is able to reconstruct physically-reasonable and globally-consistent PBR materials. Such a triangle mesh with material textures is compatible with common graphic engines, which benefits several downstream applications such as material editing, editable novel view synthesis and relighting.

**Acknowledgements**

We thank reviewers for their constructive comments, and thank Kunlong Li and Wei Li for generating annotations. This work is supported in part by the National Natural Science Foundation of China (NFSC) (No. 62002295) and Shanxi Provincial Key R&D Program (No. 2021KWZ-03).
References


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


[58] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.*, 2022. 1, 2, 3, 4, 5, 6, 7
