Factorized Inverse Path Tracing
for Efficient and Accurate Material-Lighting Estimation

Liwen Wu, Rui Zhu, Mustafa B. Yaldiz, Yin Hao Zhu, Hong Cai, Janarbek Matai, Fatih Porikli, Tzu-Mao Li, Manmohan Chandraker, Ravi Ramamoorthi

UC San Diego Qualcomm AI Research
{lww026,rzhu,myaldiz,tzli,mkchandraker,ravir}@ucsd.edu
{yinhaoz,hongcai,jmatai,fporikli}@qti.qualcomm.com

Abstract

Inverse path tracing has recently been applied to joint material and lighting estimation, given geometry and multi-view HDR observations of an indoor scene. However, it has two major limitations: path tracing is expensive to compute, and ambiguities exist between reflection and emission. Our Factorized Inverse Path Tracing (FIPT) addresses these challenges by using a factored light transport formulation and finds emitters driven by rendering errors. Our algorithm enables accurate material and lighting optimization faster than previous work, and is more effective at resolving ambiguities. The exhaustive experiments on synthetic scenes show that our method (1) outperforms state-of-the-art indoor inverse rendering and relighting methods particularly in the presence of complex illumination effects; (2) speeds up inverse path tracing optimization to less than an hour. We further demonstrate robustness to noisy inputs through material and lighting estimates that allow plausible relighting in a real scene. The source code is available at: https://github.com/lwwu2/fipt

1. Introduction

We address the task of estimating the materials and lighting of an indoor scene based on image observations (Fig. 1). Recent work has shown that optimizing per-scene material and emission profiles through photometric loss and a differentiable renderer, with geometry reconstructed with the existing 3D reconstruction algorithms [35, 26, 47], can lead to promising results [1, 29, 45]. However, key challenges remain unsolved in these methods: (1) they require expensive Monte Carlo estimation for both the loss and derivative evaluations; (2) inherent ambiguity exists between material and lighting, and this ill-posed inverse problem hinders the optimization. We present an alternative inverse rendering algorithm that outperforms the state-of-the-art in terms of both efficiency and accuracy.

*Equal contribution

Figure 1: Ours vs standard IPT. IPT [1] takes a piecewise constant parameterization of material to reduce Monte Carlo variance and ambiguity for inverse rendering, losing fine spatial details as a result. Directly extending it to complex material representation (e.g. MILO [45]) shows very slow convergence. In contrast, we propose Factorized Inverse Path Tracing (FIPT) to get rid of variance and reduce ambiguities, yielding efficient and high quality BRDF and emission (4th row), appealing relighting (1st row), and object insertion (the bunny on the table). The presented scene is synthetic with the inset showing the input (lower-left sub-figure). We further showcase results on real scenes in Fig. 10 and 11.

Optimizing scene parameters with Monte Carlo differentiable rendering can suffer from high variance and lead to slow convergence. Inspired by classical irradiance caching literature [43], our key idea to address this challenge is to factorize the material term out of the rendering integral and
bake the incoming radiance to significantly speed up inverse rendering. Unlike prior work which also applies a similar factorization (e.g. [32, 22]) but does not consider view-dependent reflections, our method extends to general specular materials and both local and global illumination.

To address the ill-posed nature of joint optimization of material and lighting, we observe that by taking out the emission term in the rendering equation for the first bounce, only emissive surfaces will have high rendering loss. This observation allows us to design an effective way to detect emitters. We incorporate our emitter detection method into a full inverse rendering pipeline and independently estimate the emission after emitter detection.

Overall, our method achieves fast convergence over the material-lighting estimation task thanks to our factorized light transport formulation and emitter extraction strategy (Fig. 1). To demonstrate accurate BRDF-emission comparison, we perform exhaustive experiments on synthetic scenes (Sec. 5.1, 5.2) while also validating on noisy data of captured real scenes (Sec. 5.3). The results show our method is able to obtain high-quality reconstruction for complicated indoor scenes that can easily fail for the state-of-the-art (Tab. 2), yet the training speed is 4-10 times faster (Tab. 3).

2. Related Work

Inverse rendering. Inverse rendering aims to estimate the intrinsic properties of an observed scene, via decoupling material, geometry and lighting which jointly contribute to image appearance. Given the inherent ambiguity between the aforementioned high-dimensional factors, classical methods seek to regularize the solution with a surface rendering objective. Approaches include a low-dimensional surface reflectance representation [46], sparsity priors for intrinsic images [4], and spherical-harmonics-based lighting representation [25]. These methods rely on simplified representation of material or lighting, and their regression-based nature calls for heuristic-based priors which may not be appropriate for a wide variety of scenes.

Earlier work can already photorealistically render synthetic objects in a photograph by estimating lighting and geometry [11, 16, 17]. These methods do not retrieve the materials of the scene, and thus cannot show the reflection of the object on a specular surface in the scene.

Learning-based methods. Learning-based approaches leverage priors learned from datasets. These methods typically take a single image [36, 20, 51, 42, 21] or a pair of stereo images [39], and apply deep learning models to predict spatially-varying materials and lighting. Although learned priors help to regularize individual components, these methods do not explicitly model the physics of global light transport and have to rely on approximated inference [24].

Philip et al. [32] take multiple images and aggregate multiview irradiance and albedo information to a pre-trained network to synthesize the relit image. The network takes physically rendered shading using light sources that are semi-automatically estimated as inputs, and outputs an image after relighting. We show in the results that in our synthetic scenes, their method’s reliance on the network to render the final image can lead to undesired artifacts, while our use of a physically-based renderer delivers more realistic images.

Local or distant lighting. Many recent methods aim to model a specific form of light transport. Some methods focus on a single object or distant illumination (environment map) [12, 49, 28, 7, 8, 50]. Srinivasan et al. [38] model two-bounce volumetric lighting with known light sources, and Yao et al. [44] represent incident radiance as a 5D network. However, optimization of spatially-varying lighting without physically-based constraints is extremely ill-posed especially without abundant observation of light sources. Moreover, object-centric methods do not trivially generalize to indoor settings, where complex lighting effects including occlusion, inter-reflections, and directional highlights call for modeling of long-range interactions of lighting and scene properties.

Global light transport. Most related to our work, to model general global light transport, recent methods [1, 29, 45] build on a per-scene optimization pipeline using a differentiable path tracer [19, 3, 30, 48]. These methods jointly optimize material and lighting along extensively sampled light paths, and thus are subject to incorrect and slow convergence and high variance due to expensive path queries, gradient propagation, and Monte Carlo sampling, as well as the inherent ambiguity between materials and lighting. We propose an inverse rendering pipeline that models the global light transport, but converges significantly faster and more accurately than existing methods. Our variance reduction technique using light baking is inspired by classical rendering methods [43, 18, 37], and we tightly integrate the technique in an inverse rendering pipeline. A concurrent work TexIR [23] adopts similar ideas to ours by using a pre-baked irradiance as HDR texture map to recover scene materials. However, they do not model view-dependent light transport and do not estimate emission.

3. Background

Given posed HDR image captures of an indoor scene, our method builds upon input mesh or existing 3D reconstruction algorithms (e.g. MonoSDF [47]) to further estimate the

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(·)ₜ⁺</td>
<td>dot product clamped to positive value</td>
</tr>
<tr>
<td>ωᵢ</td>
<td>incident (light) direction</td>
</tr>
<tr>
<td>ωₒ</td>
<td>outgoing (viewing) direction</td>
</tr>
<tr>
<td>h</td>
<td>half vector: (ωᵢ + ωₒ)/∥ωᵢ + ωₒ∥₂</td>
</tr>
<tr>
<td>n</td>
<td>surface normal</td>
</tr>
<tr>
<td>a(x)</td>
<td>surface base color</td>
</tr>
<tr>
<td>m(x)</td>
<td>surface metallic</td>
</tr>
<tr>
<td>σ(x)</td>
<td>surface roughness</td>
</tr>
<tr>
<td>kᵣ(x)</td>
<td>diffuse reflectance: a(x)(1 − m(x))</td>
</tr>
<tr>
<td>kₛ(x)</td>
<td>specular reflectance: a(x)m(x)(1 + 0.04(1 − m(x)))</td>
</tr>
<tr>
<td>D(·)</td>
<td>GGX normal distribution [41]</td>
</tr>
<tr>
<td>F(·)</td>
<td>Schlick’s approximation of Fresnel coefficient [34]</td>
</tr>
<tr>
<td>G(·)</td>
<td>Geometry (Shadow-Masking) term [41]</td>
</tr>
</tbody>
</table>

Table 1: Notations
material and lighting of the scene. To ensure the problem is well-constrained, we make similar assumptions about scene acquisition as in previous works [1, 32, 45] that the dominant light sources and most of the scene geometry are observed in input images.

The material is described as a spatially varying BRDF [15] (including the cosine term) with notations specified in Tab. 1:

\[
f(x, \omega_i, \omega_o) = \frac{k_d(x)}{\pi} (n \cdot \omega_i)_+ + \frac{F(\omega_i, h, k_s(x)) D(h, n, \sigma(x)) G(\omega_i, \omega_o, n, \sigma(x))}{4(n \cdot \omega_o)},
\]

where \(k_d = a(1 - m)\) and \(k_s = 0.04(1 - m) + am\) are the diffuse and specular reflectance with base color \(a\) and metallic \(m\) controlling the two coefficients. The emitted light is assumed to be view-independent across the surface: \(L_e(x, \omega_o) = L_e(x)\), which generalizes well to the emission profile of the indoor scene (extension to more complex emitters is possible; see Sec. 5.4).

With the parameterization above, our goal is to find \(a, \sigma, m, L_e\) that minimize the difference of renderings with respect to the ground truth over the training images:

\[
\min_{a, \sigma, m, L_e} \sum_{x, \omega_o} \|L_o(x, \omega_o) - L_{o\|}(x, \omega_o)\|_2\]

\[
L_o(x, \omega_o) = L_e(x, \omega_o) + L_r(x, \omega_o)
\]

\[
L_r(x, \omega_o) = \int \Omega \omega_i L_i(x, \omega_i) f(x, \omega_i, \omega_o) d\omega_i,
\]

where \(L_{o\|}\) is a ground truth RGB pixel obtained from camera ray \((x, \omega_o)\). \(L_o\) denotes the synthesized rendering following the rendering equation [14], where \(L_e\) is the surface emitted radiance and \(L_r\) is the reflected radiance, given by integrating incident radiance \(L_i\) times the BRDF response (Eq. 4). Note, \(L_i\) is defined in a recursive manner with multi-bounce illumination naturally taken into account.

### 4. Factorized Inverse Path Tracing

To optimize re-rendering error (Eq. 2), previous works [1, 29] apply differentiable path tracing to solve Eq. 3 and update BRDF and emission jointly with gradient descent. This approach can be unstable and inefficient: (1) gradient descent optimization is computationally intensive, which limits the number of path tracing samples and therefore increases the estimation variance; (2) fundamental ambiguities exist between BRDF and emission, making emission optimization difficult to regularize or converge. To reduce variance in optimization, we propose a factorized light transport representation (Sec. 4.1) which utilizes pre-baking of diffuse and specular shading maps (Eq. 6) to separate the BRDF coefficients out of the rendering integral.

Our full pipeline is demonstrated in Fig. 2, which optimizes dense BRDF and emission from posed images and scene geometry. The pipeline consists of 3 stages: (1) first,
we find that using a
formulation, we also empirically found this factorized ren-
dering handles mirror-like objects much better (Fig. 9), while
standard Monte Carlo integration can easily obtain unstable
gradients caused by the large BRDF value.

4.2. Image-based shading initialization

As we pre-bake the shadings \( L_d, L_s^0, L_s^1 \), they are fixed
during BRDF-lighting estimation. They need to be properly
initialized; otherwise, the optimization will not converge.
For simplicity, we abstract the integrands in Eq. 6 to the
form: \( L_s = L_s^f \cdot g \), where \( g \) denotes the factorized BRDF
term. In path tracing notation, the shading integral can be
initialized by querying a surface light field approximation:

\[
L_s(x) = L((x_1 \rightarrow x)g(x_1 \rightarrow x) \\
\approx L'(x_1)g(x_1 \rightarrow x),
\]

where \( L \) is the exact surface light field at sampled location \( x_1 \)
towards \( x \), and \( L' \) is its approximation obtained by average
pooling all the input pixels onto a voxel grid spanned on the
scene geometry (Fig. 3 left). Since objects in an indoor scene
are often near-diffuse, and the renderings are essentially low
pass filtering the incident light field [33] that blurs the detail,
we find that using a 256\(^3\) voxel grid with nearest neighbor
radiance query gives good shading approximations (Fig. 3
right).

4.3. Path-traced shading refinement

Eq. 9 gives incorrect shading if the surface light field
is sampled at locations that are mainly specular (\( L \) is view
dependent), which subsequently leads to incorrect BRDF
estimation. Given BRDF-emitter estimations optimized from
Eq. 2 under current shading estimations, we re-estimate the
light transport on specular surfaces by growing the path in
Eq. 9 until the ray either hits an emitter or intersects with a

Figure 3: **Diffuse and specular shadings are initialized** by
tracing a voxel representation of the surface light field \( L' \)
(left), which gives approximations (top row on right) close
to the ground truth (bottom row on the right; obtained by
path tracing).

Figure 4: **Shading refinement**: The cabinet’s diffuse re-
flectance estimation is initially darker than ground truth,
owing to the excessive incident light received from the range
hood that reflects non-diffuse light (2nd column). The arti-
facts are reduced by growing the path for the specular surface
according to the optimized BRDF (1st column), which gives
more accurate shadings that can be used to further refine the
BRDF (3rd column).

Figure 5: **Diffuse radiance cache** from \( L' \) helps reduce vari-
ance and error for shading estimation (2nd image). Without
it, sampling the tiny emitters below the cabinet will be diffi-
cult (1st image), which leads to incorrect shading and albedo
(4th image).

\[
L_s(x) = R(x_n) \prod_{i=1}^{n-1} f(x_{i+1} \rightarrow x_i)g(x_1 \rightarrow x),
\]

s.t. \( \sigma(x_i) \leq 0.6, \forall i < n \)

\[
R(x_n) = \begin{cases} 
L'(x_n), & \sigma(x_n) > 0.6 \\
L_s(x_n), & L_s(x_n) > 0
\end{cases}
\]

where \( n \) is the length of a certain path before it terminates.
The above equation essentially estimates the shadings by
multi-bounced path tracing with \( L' \) being a diffuse radiance
cache, which helps speed up the evaluation and also reduce
error: initial estimations of \( f_s \) may retain large error but
\( L' \) is very close to a diffuse surface light field (as it is also
view-independent). Most of the path hits a diffuse surface
within one to two bounces, such that the errors from the
BRDF will not be magnified (Fig. 5).

Substituting shadings in factorized rendering by their
refinements makes Eq. 5 more closely match the ground
truth light transport, such that BRDF can be re-estimated
with fewer artifacts (Fig. 4: ‘Origin’ VS ‘Refined’). The re-
estimated BRDF in turn is applied to further improve the
shadings, and this BRDF and shading refinement is per-
formed alternatively until convergence.
we make the assumption that an emitter reflects zero light (form solution as the median of RGBs from all input pixels it ill-posed (if the mean of its α sample 100 locations on each triangle and find their corre-

cients, such that the gradient update is balanced between the emission mask value is in the same range as BRDF coeffi-
mization (demonstrated in ablations in Sec. 5.4), because the error. In practice we apply a L1 sparsity loss to

\[ e = \min \sum_{x,\omega_e} \alpha (1 - \alpha) L_r + \alpha L_{gt} - L_{gt} \]  

(12)  
s.t. \( \alpha \rightarrow 0 \).

When a surface is non-emissive, \( \alpha \) will stay small owing to the regularization and the loss is minimized by adjusting \( L_e \) towards \( L_{gt} \); but \( L_r \) cannot model the emission, so \( \alpha \) for an emissive surface has to become large to accommodate the error. In practice we apply a L1 sparsity loss to \( \alpha \). By changing the optimization objective to Eq. 12, we first jointly estimate the BRDF-emission mask, and then threshold the mask to find the emitter (\( \alpha > 0.01 \)). Afterward, each emitter’s emission \( L_e \) is estimated independently from BRDF:

\[ L_e = \begin{cases} \arg \min \sum_{x,\omega_e} \|L_e + L_r - L_{gt}\|_1 & \alpha > 0.01 \\ 0 & \text{otherwise} \end{cases} \]  

(13)

Our formulation is found to be more stable than joint optimization (demonstrated in ablations in Sec. 5.4), because the emission mask value is in the same range as BRDF coefficients, such that the gradient update is balanced between the BRDF and the emission mask. In contrast, surface emission can be much larger than BRDF coefficients, making it more difficult to directly fit or regularize.

Emitter extraction. We assume emission is constant for each mesh triangle. After \( \alpha \) is optimized, we uniformly sample 100 locations on each triangle and find their corre-

casting and the BRDF and the emission mask. In contrast, surface emission

can be ambiguous without any regularization (1st image). By encouraging a surface to be diffuse, specular surfaces still get an incorrect rough-

ness value if no highlights are observed (2nd image). The roughness can be more reasonably estimated with part seg-

mentation for guidance (3rd image). Semantic segmentation (4th image) shows similar results except the roughness for small objects get blurred.

4.5. Optimization

Given either initial or refined shadings, the BRDF and emission mask are optimized using the objective in Eq. 12. We encode BRDF and the emission mask with two MLPs:

\[ (a, m, \sigma) = \text{Sigmoid} \left( \text{MLP}_{\text{brdf}}(x) \right) \]

(14)

where MLP_{brdf} uses hash encoding [27] and MLP_{emin} is a positional encoded MLP [26]. The objective Eq. 12 is converted to a gradient descent loss function as a tone-mapped L2 loss \( l \) plus a L1 regularization term \( l_e \):

\[ l = \sum_{x,\omega} \|\Gamma((1 - \alpha) L_r + \alpha L_{gt}) - \Gamma(L_{gt})\|^2_2 \]

(15)

\[ l_e = \lambda_e \sum_x \|\text{MLP}_{\text{emin}}(x)\|_1, \lambda_e = 1, \]

(16)

where \( \Gamma \) is the tone-mapping function by Munkberg et al. [28] to help suppress noise from high dynamic range values. We prefer neural networks rather than a textured mesh (as in [1, 29]) as scenes with complex geometries can create degenerate UVs, which reduces the BRDF quality.

Roughness-metallic regularization. Surface roughness and metallic can take arbitrary values if there are no high-

lights (\( L_r^e, L_r^m \approx 0 \)), which leads to ambiguity. We prevent this by encouraging surfaces to be diffuse:

\[ l_d = \lambda_d \sum_x (\|1 - \sigma(x)\|_1 + \|m(x)\|_1), \lambda_d = 5e-4, \]

(17)

such that a diffuse surface will not be misinterpreted as a specular surface with weak reflection. To get valid roughness-metallic for input pixels that do not observe highlights, we further assume they stay constant inside each material part, and utilize image-level part segmentation to group input pixels. The roughness-metallic from pixels with highlights are propagated to their corresponding group by
another regularization loss:

\[
l_p = \lambda_p \sum_x \left\| \frac{\sigma(x)}{m(x)} - \frac{\sigma'(x)}{m'(x)} \right\|_1, \quad \lambda_p = 5e-3
\]

\[
\frac{\sigma'(x)}{m'(x)} = \sum_{\text{Seg}(x) = \text{Seg}(x')} \frac{w(x')}{\sum_{x} w(x')} \frac{\sigma(x')}{m(x')}
\]

\[
w(x') = \mathbb{g} \left( \left\| k_d L^0_d + L^1_d \right\|_1 \right), \quad \text{where } \mathbb{g}(\cdot)
\]

where \( \text{Seg}(x) \) gives the segmentation ID for \( x \), \( \mathbb{g}(\cdot) \) denotes stop the gradient, and \( w(x) \) is a propagation kernel that weights the pixel by the amount of highlights. While part segmentation in practice can be hard to obtain, semantic segmentation is readily available from pre-trained model e.g. Mask2Former [10], where multiple material parts may stay inside the same semantic label. To account for such detail loss, we consider two pixels belong to the same material part only if: (1) they share the same semantic ID; (2) have similar albedo value; (3) and are close to each other, which suggests an alternative propagation kernel \( w(x, x') \):

\[
w(x', x) = \mathbb{g} \left( \frac{e^{-\frac{|k(x)-k(x')|^2}{2\sigma^2}} - e^{-\frac{|k(x)-k(x')|^2}{2\sigma^2}}}{\sum_x e^{-\frac{|k(x)-k(x')|^2}{2\sigma^2}}} \right)
\]

By replacing \( w(x) \) with \( w(x, x') \) and changing the regularization weight to \( \lambda_p = 1e-3 \), we can still have reasonable roughness-metallic estimation even with semantic segmentation (Fig. 7).

5. Experiments

We evaluate our method on 4 synthetic and 2 real indoor scenes, where the synthetic scenes are obtained from Bitterli’s rendering resources [5] with large glass objects being removed (as we do not model transmission), and the real scenes are captured by us. Each synthetic scene contains around 200 posed HDR images generated by Mitsuba3 [13], per-camera view BRDF-emission maps generated by Blender [6], and ground truth geometry. For synthetic scenes, we show our method with both part segmentation (FIPT) and semantic segmentation mask (FIPT-sem).

The real scenes (Conference room and Classroom) are captured by a Sony A7M3 camera with around 200 HDR images reconstructed by 5-stop exposure bracketing. The camera poses are estimated from COLMAP [35] and the geometry is reconstructed using MonoSDF [47]. Please refer to the supplementary material for details on real scene capturing and additional results.

5.1. Synthetic: BRDF-emission estimation

While synthetic scenes allow us to directly compare with the ground truth BRDF-emission without noise from geometry or image captures, BRDF parameterizations can vary across different baselines. For fair comparison, we empirically found diffuse reflectance \( k_d \) for diffuse surfaces, roughness \( \sigma \), and the material reflectance defined by

\[
\kappa_a = \int_{0}^{\pi} f d\omega_i
\]

\( f \) are very close across different BRDF models [9, 15]. We therefore measure the PSNR for these metrics in image space for BRDF comparison. The \( k_d \) is compared only for diffuse surfaces, and \( \kappa_a \) is estimated using Monte Carlo integration of 128 samples per pixel. For emission, we estimate the IoU of emission mask and logL2 error of emission map.

### Baselines.

We compare with the original inverse path tracing (IPT) [1] and its extension MILO [45] that also parameterizes spatially varying BRDF with neural networks. IPT assumes BRDF parameters to be constant inside each mesh triangle, and MILO takes manual input of number of emitters. Both IPT and MILO are evaluated by their original authors due to non-public code, and the MILO training is stopped after 10 hours. Meanwhile, we also compare NeILF [44] that models illumination in an unconstrained way and the learning based approach [21] (Li22) for single-view inverse rendering.

### Results.

As is shown in Tab. 2, our method gives the best BRDF and emission estimation with the fastest training speed (Tab. 3) even when only semantic level segmentation is provided (FIPT-sem). The learning-based approach (Li22) fails to generate reasonable reconstruction as it does
Figure 8: **BRDF and emission estimation results on 2 synthetic scenes** shows our method successfully reconstructs material reflectance (1st row), roughness (2nd row), and emission (3rd row) with high frequency details and less ambiguity. Emission estimation is shown as error heatmaps (warmer colors indicate higher emission error; GT emitter boundary is marked in white lines). Input views are shown in upper-right corner.

Figure 9: **Qualitative results of view synthesis (left) and relighting (right) on 2 synthetic scenes** demonstrate accurate light transport can be simulated with our estimated BRDF and emission even for very specular surfaces. The reflection of the chair from the microwave oven can be seen in kitchen scene on top, and mirrors are correctly rendered for bathroom (bottom).

<table>
<thead>
<tr>
<th>Our per-stage profiling</th>
<th>Method</th>
<th>training time</th>
<th>Baselines. Besides IPT, MILO, and Li22, we also consider FVP [32] that performs view synthesis and relighting in a learning-based way. FVP assumes emissions come from saturated regions on the images, which may wrongly classify surfaces with strong reflection as emitters. So we offer ground truth emission to FVP instead as oracle. The renderings for IPT, MILO, and our method are obtained by path tracing with 1024 samples per pixel, which is further denoised by the Optix denoiser [31].</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1 Stage 2 Stage 3</td>
<td>NeILF [44]</td>
<td>1h38min</td>
<td><strong>Results.</strong> As is shown in Tab. 4, our (FIPT and FIPT-sem) estimated BRDF-emission gives the most accurate view synthesis and relighting results. While results from FVP are seemingly visually appealing, the method is not guaranteed to be physically plausible and fails to match the ground truth. As shown in Fig. 9, our method handles specular reflections and even mirror reflection well, which is difficult to be modeled by standard inverse path tracing owing to its high variance.</td>
</tr>
<tr>
<td>Memory</td>
<td>3.2GB</td>
<td>2.8GB</td>
<td>3.4GB</td>
</tr>
<tr>
<td>Time</td>
<td>6min</td>
<td>2min</td>
<td>16min</td>
</tr>
</tbody>
</table>

not utilize multi-view cues, while unconstrained optimization (NeILF) suffers from the ambiguity between material and lighting. While IPT converges, its accuracy is limited by a piece-wise constant constraint to reduce the variance. MILO also fails to reconstruct high frequency details because of the Monte Carlo noise from path tracing, and it requires manual specification of the number of emitters to constrain the emission optimization. In contrast, our method requires no human input during optimization, which allows more stable and faster convergence with results that match the ground truth well (Fig. 8).

### 5.2. Synthetic: view synthesis and relighting

To demonstrate the applications of inverse rendering outputs, we compare the rendered scenes under novel views and novel lighting using estimated BRDF and emission. For quantitative comparison, we tone-map the rendered images with $\gamma = 1/2.2$ then calculate their PSNR with respect to the ground truth.

Table 3: **Averaged training speed comparison** suggests our method is very efficient (right table). The per-stage profiling is shown on the left with Stage 2 and 3 being repeated twice. The comparison is made on a 3090Ti GPU.
5.3. Real-world scenes

Considering that it is not possible to obtain ground truth BRDF and emission from just RGB captures, we only showcase qualitative comparison with MILO, FVP, and Li22 in terms of material-lighting estimation, rerendering, and relighting. Only semantic segmentation is used for the real scenes as ground truth part segmentation is unavailable.

### 5.4. Ablation study

#### Training strategy.

Tab. 5 shows the effect of different training strategies on the kitchen scene. If we jointly optimize the emission and BRDF with the regularization term in IPT [1], the BRDF optimization can still converge, but the emission estimation does not converge given the same amount of time (2 epochs). Since the majority of the scene receives incident light from nearby diffuse surfaces, the reconstruction result is still reasonable without shading refinement (stage 3), but further refining the BRDF estimation helps to correct light transport for specular surfaces. If we simply path-trace the BRDF-emission without using the radiance cache from stage 1, the refined shadings will accumulate too much estimation error causing the subsequent BRDF estimation to deviate from ground truth.

---

**Figure 10**: **BRDF and emission estimation results on 2 real world scenes** demonstrate our method gives reasonable estimation of BRDF and emission. The albedo and roughness preserve details without noticeable artifacts (row 1-2), and the emitters are correctly identified (3rd row).

**Figure 11**: **Rerendering and relighting on 2 real world scenes** show our estimation fits the ground truth well (column 1-5) and gives good rendering under novel light (column 6-9). The inset in lower-right sub-figure shows the reference relighting for the Classroom scene.

<table>
<thead>
<tr>
<th>Method</th>
<th>Bathroom</th>
<th>Bedroom</th>
<th>Livingroom</th>
<th>Kitchen</th>
</tr>
</thead>
<tbody>
<tr>
<td>FVP [32]</td>
<td>23.38</td>
<td>20.49</td>
<td>24.63</td>
<td>20.77</td>
</tr>
<tr>
<td>FIPT</td>
<td>25.42</td>
<td>29.84</td>
<td>30.86</td>
<td>25.38</td>
</tr>
<tr>
<td>FIPT-sem</td>
<td>25.76</td>
<td>29.89</td>
<td>30.84</td>
<td>25.27</td>
</tr>
</tbody>
</table>

Table 4: **Quantitative results (PSNR) of view synthesis and relighting on synthetic scenes** show our estimation yields very consistent rendering under novel views and lighting. View synthesis is unavailable for Li22.

**Results.** As is shown in Fig. 10, our method visually produces more reasonable material reflectance and roughness with emission masks closer to the actual emitters. The rerendering and relighting in Fig. 11 further serve as indirect measurements of the reconstruction quality, where our method is capable of reconstructing the reflection of the emitters on the Conference room wall and the Classroom table, and our relighting is visually closer to reference photos (obtained by switching lights; see supplementary) than the baselines with good reproduction of specular highlights and shadows.
with hollow geometry, we can model the window light as (Fig. 12), which disrupts roughness estimation for regions near windows (see Sec. 6), a majority of the surfaces can still be reconstructed well owing to the diffuse radiance cache.

6. Limitations and Future Work

Our method shares certain limitations with standard inverse path tracing. The framework does not optimize geometry, so that the BRDF and emission estimations can be inaccurate if the input geometry (especially for emitters) is extremely bad (Fig 14, left). Combining differentiable geometry optimization [2, 40] may help improve the robustness. Meanwhile, the BRDF estimation fails if the dominant light source (e.g. the sun) is not directly observed, which can happen very often with environment emitters whose observations are blocked by the windows (Fig. 14, right). Incorporating learning based methods may help. Lastly, the optimization relies on photometric observations, which means it cannot remove ambient occlusion effects out of the BRDF maps (as radiance there is near zero) and our model does not model transparent objects.

Acknowledgements. This work was supported in part by NSF grants 1751365, 2100237, 2105806, 2110409, 2120019, 2127544, ONR grant N000142012529, N000142312526, a Sony research award, gifts from Qualcomm, Adobe and Google, the Ronald L. Graham Chair and the UC San Diego Center for Visual Computing.

Additionally, we would like to thank Bohan Yu, Dejan Azinovic, Julien Philip, and Zhengjin Li for generous assistance in evaluation of their methods, David Forsyth, Shenlong Wang, and Merlin Nimier-David for insightful discussions, as well as Jiaer Zhang for assistance in implementation.
References

[1] Dejan Azinovic, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. Inverse path tracing for joint material and lighting estimation. In CVPR, 2019. 1, 2, 3, 5, 6, 7, 8


[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In ACM Transactions on Graphics (ToG), volume 41, pages 1–15, 2022. 5


3858