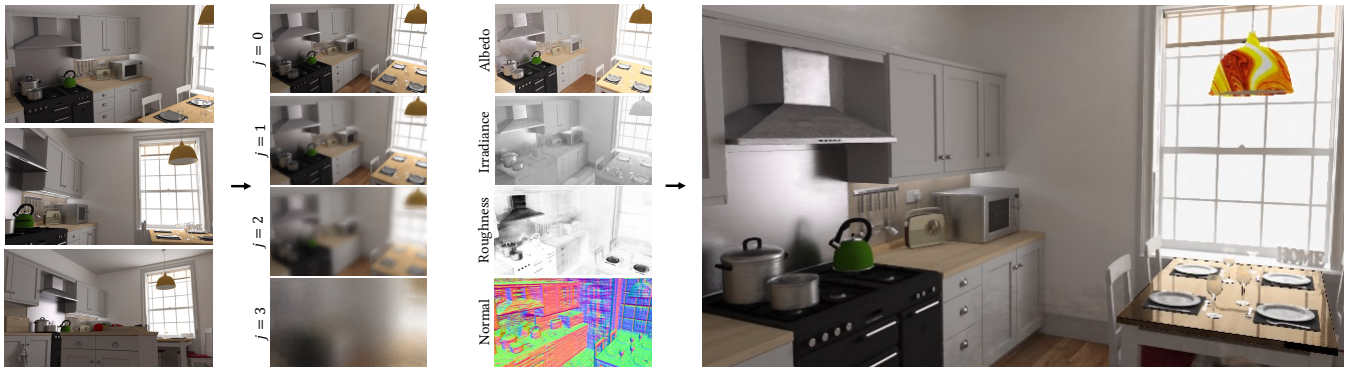
DOI: 10.1111/cgf.14929
Pacific Graphics 2023
R. Chaine, Z. Deng, and M. H. Kim
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 42 (2023), Number 7

IBL-NeRF: Image-Based Lighting Formulation of Neural Radiance Fields

Changwoon Choi*[1] Juhyeon Kim*[1,2] Young Min Kim†[1]

[1]Department of Electrical and Computer Engineering, Seoul National University
[2]Department of Computer Science, Dartmouth College



(a) Input Images (b) Prefiltered Radiance (c) Intrinsic Components (d) Scene Editing

Figure 1: *We propose IBL-NeRF, a neural volume representation with prefiltered radiance field inspired by image-based lighting formulation. (a) Given multi-view images, we optimize the (b) prefiltered radiance field and estimate (c) reflectance properties of the material (albedo, roughness), lighting information (irradiance, prefiltered radiance), and the geometry (normal). (d) One can manipulate the neural scene easily by modifying the decomposed components. Project page: https://changwoon.info/publications/IBL-NeRF*

Abstract

We propose IBL-NeRF, which decomposes the neural radiance fields (NeRF) of large-scale indoor scenes into intrinsic components. Recent approaches further decompose the baked radiance of the implicit volume into intrinsic components such that one can partially approximate the rendering equation. However, they are limited to representing isolated objects with a shared environment lighting, and suffer from computational burden to aggregate rays with Monte Carlo integration. In contrast, our prefiltered radiance field extends the original NeRF formulation to capture the spatial variation of lighting within the scene volume, in addition to surface properties. Specifically, the scenes of diverse materials are decomposed into intrinsic components for rendering, namely, albedo, roughness, surface normal, irradiance, and prefiltered radiance. All of the components are inferred as neural images from MLP, which can model large-scale general scenes. Especially the prefiltered radiance effectively models the volumetric light field, and captures spatial variation beyond a single environment light. The prefiltering aggregates rays in a set of predefined neighborhood sizes such that we can replace the costly Monte Carlo integration of global illumination with a simple query from a neural image. By adopting NeRF, our approach inherits superior visual quality and multi-view consistency for synthesized images as well as the intrinsic components. We demonstrate the performance on scenes with complex object layouts and light configurations, which could not be processed in any of the previous works.

CCS Concepts

• ***Computing methodologies*** → *Computer vision representations; Rendering;*

1. Introduction

Neural radiance field (NeRF) [MST*20] prospers for their superior quality in novel-view synthesis with a simple formulation. A neural

* Authors contributed equally to this work.
† Corresponding author.

network is trained to optimize a colored density volume to directly match multiple posed input images. The formulation is ignorant of any intermediate representations of traditional rendering pipelines, namely surface geometry, light transport, or BRDF. The trained volumetric representation does not trace iterative inter-reflections of rays, or model complex occlusion of the surface geometries. Nonetheless, NeRF can produce detailed subtleties of global illumination and parallax effects.

While NeRF can capture complex effects in general scenes, the implicit formulation limits further analysis or edits of the scenes. Intrinsic decomposition is an attractive choice as it decomposes the captured scene into intrinsic components that can be further manipulated to edit the scene. However, intrinsic decomposition is inherently an ill-posed problem and requires enforcing additional priors or constraints. Prior works often extract an isolated object in a bounding box, selected with exhaustive segmentation masks, for intrinsic decomposition of NeRF. They assume a single low-dimensional environment lighting for the entire scene and incorporate additional knowledge for reflectance properties, such as priors on BRDFs or images captured under different known illuminations. Under the constrained set-up, they sample rays between the approximated environment light and the segmented object with Monte-Carlo integration which can be computationally expensive. Furthermore, such approximation with environment light prohibits viewpoints inside the scene, or a local variation of lights caused from common light fixtures or windows. By relinquishing the flexibility of the original NeRF, existing inverse rendering with NeRF approaches cannot represent everyday environments composed of diverse unsegmented objects.

Instead of extensively simulating multiple bounces of rays with approximated explicit representation, we propose incorporating constraints from the image spaces, extending the NeRF formulation. Specifically, we train a decomposed neural volume, coined IBL-NeRF, to optimize for the implicit light distribution of neural images. This neural representation captures detailed spatial variations of lighting, in contrast to low-dimensional environment mapping. Then we can substitute the illumination integration process into a simple network query for the irradiance. The specular reflection of different surface roughness values is fetched from prefiltered radiance fields of appropriate prefilter levels, similar to texture mipmap. We additionally enforce priors on the intrinsic components for input images, acquired from existing methods for decomposing individual images. By incorporating image-based lighting with implicit intrinsic components, we can efficiently render general scenes without sacrificing the rendering quality of the original NeRF as shown in Fig. 1. We can further edit scenes by changing materials or adding objects, including highly reflective or transparent objects.

In summary, our approach fully leverages the high-quality novel view images of the original NeRF formulation, and yet enables efficient re-generation with approximations inspired from image-based lighting. Our contributions can be listed as following:

- We propose IBL-NeRF, which handles global illuminations with spatially varying lighting and diverse materials given a set of unsegmented images.
- We model the prefiltered radiance of the scene with a neural network of NeRF, and efficiently approximate rendering equations with image-based lighting formulation.
- Our neural representation extracts physically interpretable components of the complex indoor scenes which can be altered to render images with different attributes.

The results are presented with large-scale scenes containing multiple objects, which can not be modeled with previous works employing a single environment lighting with Monte-Carlo integration.

2. Related Works

While NeRF [MST*20] can synthesize photo-realistic novel-view images, one of its limitations is that the radiance information is baked within the implicit neural representation. Several subsequent works propose to distill intrinsic components, such as illumination and reflectance property, and try to achieve inverse rendering with implicit representation, in contrast to reconstructing explicit mesh geometry with multi-view stereo [PMGD21, DRC*15]. They optimize components to match the input images by evaluating the rendering equation with Monte Carlo (MC) method, which requires heavy computation. Neural Reflectance Fields [BXS*20] and NeRV [SDZ*21] adapt ray-marching to account for reflectance, and model the illumination with a single point light and environment light, respectively. Both approaches require multiple images with known lighting configurations as input. NeRFactor [ZSD*21], Hasselgren et al. [HHM22], NeRD [BBJ*21], and PhySG [ZLW*21], on the other hand, factorize radiance fields from unknown light. They concurrently optimize for a single low-dimensional environment light in a coarse resolution (NeRFactor, Hasselgren et al.) or spherical Gaussian (NeRD, PhySG).

In contrast, IBL-NeRF proposes to efficiently synthesize images without explicit Monte-Carlo integration, and utilizes prefiltered radiance which can be evaluated with a single ray sample. Several works [VHM*22, BJB*21] also adapt integrated illumination for efficient rendering. They are either implicitly conditioned on the surface reflectance property, or propose components without physical interpretation. However, previous works using integrated illumination employ a single environment lighting for entire scene and therefore are limited to modeling an isolated object. Concurrent works, such as I^2-SDF [ZHY*23] and TexIR [LWC*23], also exploit spatially-varying light for complex indoor scenes, but they use MC integration or explicit mesh representation, respectively.

Intrinsic decomposition for general scenes requires modeling spatially-varying lighting. With increased degrees of freedom for the already under-constrained problem, scene decomposition requires strong assumptions. Commonly used priors include piecewise constant albedos [CZL18, LS18a, LS18b, MCZ*18, LBP*12], or sparsity of extracted albedo values [MSZ*21, GMLMG12]. A few works exploit data-driven priors instead of hand-crafted priors [BBS14, ZKE15, SGK*19, LSR*20, PEL*21, YTL20], which can be subject to domain discrepancy. IBL-NeRF takes inspiration from the aforementioned prior works using single images, and adds constraints in the image space. Because the neural volume of NeRF is trained with images, the formulation can readily be applied to handle challenging indoor scenes without simplifying the illumination model. Furthermore, IBL-NeRF can naturally find multi-view consistent components, which is not possible with single-image decomposition.

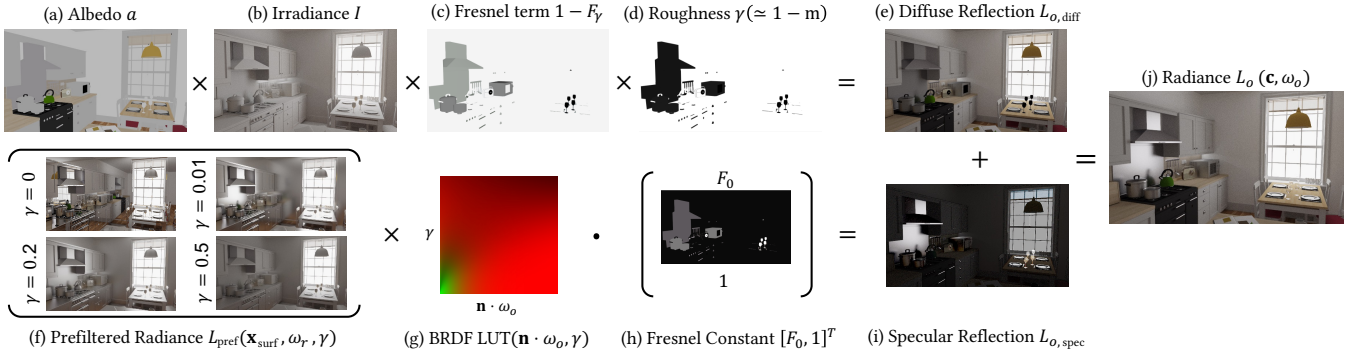


Figure 2: *Overview of the radiance approximation used in IBL-NeRF (Eq. 2). With the combination of inferred (a) albedo, (b) Irradiance, (c) Fresnel term, and (d) roughness, one can obtain diffuse reflection. Also, with (f) multi-level prefiltered radiance, (g) fetched value from LUT and (h) Fresnel constant, we can calculate (i) specular reflection. (j) Final approximated radiance is achieved by the sum of diffuse and specular reflection.*

3. Method

3.1. IBL-NeRF Formulation

3.1.1. Preliminaries

Ray-tracing engines approximate the light transport with samples of rays. The original rendering equation [Kaj86] formulates the outgoing radiance at surface $\mathbf{x}_{\mathrm{surf}}$ as a combination of reflected rays of incoming radiance L_i

$$L_o(\mathbf{x}_{\mathrm{surf}}, \omega_o) = \int_\Omega f_r(\mathbf{x}_{\mathrm{surf}}, \omega_i, \omega_o) L_i(\mathbf{x}_{\mathrm{surf}}, \omega_i)(\mathbf{n} \cdot \omega_i) d\omega_i, \quad (1)$$

where $\mathbf{n}$ and f_r are the surface normal and BRDF at surface $\mathbf{x}_{\mathrm{surf}}$, and ω_i and ω_o are incoming and outgoing direction. Given the scene properties ($\mathbf{n}$ and f_r), the rendered output relies on the diverse distribution of light transport, L_i and L_o, which are 5D functions.

The approximation within game engines [Kar13] replaces the recursive calls of radiances $L_i \rightarrow L_o$ into a single sample of integrated light. L_o is approximated as the sum of two components, namely the diffuse term and the specular term:

$$L_o(\mathbf{x}_{\mathrm{surf}}, \omega_o) = \underbrace{\gamma \times (1 - F_\gamma(\omega_o, \mathbf{n}, \gamma)) \times a \times I}_{L_{o,\mathrm{diff}}} + \underbrace{L_{\mathrm{pref}}(\mathbf{x}_{\mathrm{surf}}, \omega_r, \gamma) \times [F_0, 1]^T \, \mathrm{LUT}(\omega_o \cdot \mathbf{n}, \gamma)}_{L_{o,\mathrm{spec}}}. \quad (2)$$

The diffuse term depends on irradiance $I = \int_\Omega L_i(\mathbf{x}_{\mathrm{surf}}, \omega_i)(\mathbf{n} \cdot \omega_i) d\omega_i$ which integrates all the incoming radiance. Additionally, it is proportional to the surface albedo a, roughness γ, and approximated Fresnel term F_γ. (According to the original paper of [Kar13], the diffuse term is attenuated by (1-metallic) and we approximated it as roughness (γ). More sophisticated approximations could be tried in future works.) Calculating the specular term $L_{o,\mathrm{spec}}$ involves directional components of rays. The split-sum approximation simplifies the specular term into the product of two terms. The first component L_{pref} is the *prefiltered environment map* which summarizes the effects of reflected lights to mimic specular highlights efficiently. It is filtered according to the surface roughness level γ and fetched at the reflected direction. The second component is also precalculated as a 2D lookup texture (LUT). Both diffuse term and specular term are affected by roughness γ. IBL-NeRF allows the decomposition of NeRF by utilizing neural network to represent the pre-computed volumetric light distribution. Detailed descriptions of the approximation are available in the supplementary material.

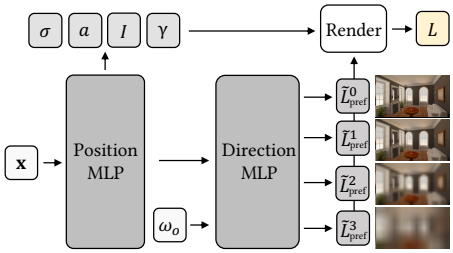


Figure 3: *Architecture of IBL-NeRF. The scene properties dependent on position (volume density σ, albedo a, irradiance I and roughness γ) are extracted from position MLP. Those dependent on the viewing direction (each level of prefiltered radiance $\tilde{L}^j_{pref}$) are obtained from direction MLP.*

3.1.2. Rendering Pipeline of IBL-NeRF

NeRF synthesizes a photo-realistic image applying a volume rendering on a neural volume

$$L_o(\mathbf{c}, \omega_o) = \int_0^\infty V(\mathbf{x}(t), \mathbf{c}) \sigma(\mathbf{x}(t)) L_e(\mathbf{x}(t), \omega_o) dt, \quad (3)$$

where $\mathbf{x}(t) = \mathbf{c} - t\omega_o$ represents points on a ray initiated from the camera position $\mathbf{c}$, and $V(\mathbf{x}(t), \mathbf{c}) = \exp\left(-\int_0^t \sigma(\mathbf{x}(s)) ds\right)$ is the visibility. Given a position $\mathbf{x}$ and an outgoing direction ω_o, the neural volume of NeRF is trained to regress for density σ from the positional MLP and the emitted radiance L_e from the directional MLP. The training objective is to match the results of volume rendering with the pixels in the input images, which enables creating images of the scene only from a set of multiple-view images.

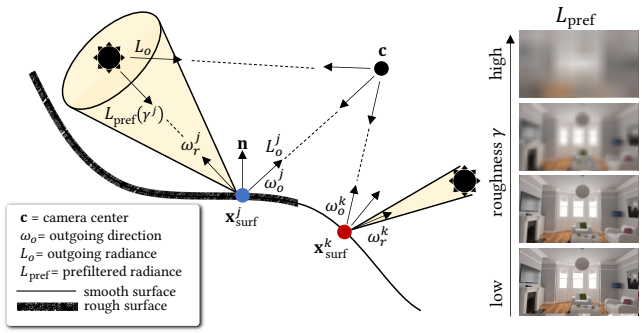


Figure 4: *Specular reflection of IBL-NeRF. The prefiltered radiance field is fetched only from the estimated surface point $\mathbf{x}_{surf}$ with a single reflected ray toward the direction of mirror reflection ω_r. The point on rough surface (intersection of j^{th} ray) fetches the prefiltered radiance convoluted with a wide kernel. In contrast, the point on smooth surface (intersection of k^{th} ray) reads prefiltered radiance field filtered with a narrow kernel.*

	NeRF	NeRFactor	IBL-NeRF (Ours)
Rendering	Volume	Surface	Surface
L_o	Baked	Monte Carlo Integration	Approx w. Eq. 2
L_i	-	Env light w. Visibility Infer	MLP Inference
Time Complexity	$\mathcal{O}(N_s)$	$\mathcal{O}(N_s+N_dN_r)$	$\mathcal{O}(N_s+N_r)$

Table 1: *We compare IBL-NeRF with NeRF and a recent method decomposing NeRF's radiance. The time complexity is measured for the entire training phase. N_s and N_r are the numbers of samples along a camera ray and a reflected ray, and N_d is the number of directional samples over a hemisphere.*

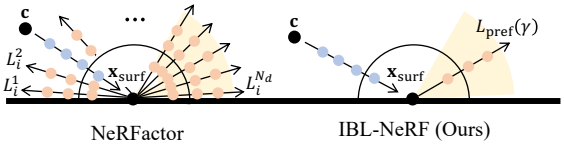


Figure 5: *Illustration for the time complexity analysis in Table 1. Prefiltered radiance fields significantly reduce the sample number during the training phase from $\mathcal{O}(N_s+N_dN_r)$ to $\mathcal{O}(N_s+N_r)$.*

To decompose the radiance of NeRF into physically interpretable components of the scene, we can adapt components as presented in Eq. 2, ignorant of light transport. For each ray, we evaluate albedo a, irradiance I, and roughness γ with the volume density σ. We accumulate the values along the ray using volume rendering following the NeRF formulation in Eq. 3. Also, at the estimated surface point, the network evaluates the prefiltered radiance field L_{pref} of the reflected direction. Due to the computational complexity, the reflected rays are evaluated only at the surface hit position of the ray, which is estimated as $\mathbf{x}_{\text{surf}} = \mathbf{c} - d\omega_o$ [ZSD*21, SDZ*21]. The termination depth $d(\mathbf{c}, -\omega_o)$ of the ray defines the surface point $\mathbf{x}_{\text{surf}}$ and can be obtained with density $\int_0^\infty \exp\left(-\int_0^\infty \sigma(\mathbf{c}-s\omega_o)ds\right) t\sigma(\mathbf{c}-t\omega_o)dt$. We obtain the surface normal from the numerical gradient of the termination depth d: $(\mathbf{n}(\mathbf{x}_{\text{surf}}) = \nabla_{\mathbf{x}} d(\mathbf{x},\omega)/\|\nabla_{\mathbf{x}} d(\mathbf{x},\omega)\|$.) All the values are combined using Eq. 2 to find the output radiance corresponding to the pixel, which is also visualized in Fig. 2.

Fig. 3 shows the modified neural network architecture. The positional MLP infers the components that do not have view dependency, namely, albedo a, irradiance I, and roughness γ, in addition to the volume density σ in the vanilla NeRF. Note that irradiance is inferred from MLP implicitly, instead of explicitly integrating over the hemisphere. The irradiance depends on surface normal, but we assume that it is implicitly handled in the neural network, which takes position $\mathbf{x}$ as input. The directional component is encoded as prefiltered radiance field L_{pref}, and is the output of the subsequent directional MLP. It is modulated by roughness γ and combined to generate the final image. The following subsection further explains the formulation and approximation used for the prefiltered radiance fields.

3.2. Prefiltered Radiance Fields

The prefiltered environment map $L_{\text{pref}}(\mathbf{x}_{\text{surf}}, \omega_r, \gamma)$ in Eq. 2 accounts for the specular reflection with directional components that reside in a high-dimensional space as a single sample. Let us denote the camera observation direction as $-\omega_o$ and its mirror reflection with respect to the surface normal $\mathbf{n}$ at the surface point as ω_r. Unless the surface is a perfect mirror (roughness 0), the reflected rays are evaluated within angular distribution near the reflection direction. As the surface roughness γ increases, prefiltered radiance should be filtered with a wider range kernel. Fig. 4 illustrates the procedure, where the pre-filtered radiance at ω_r is depicted with cones with yellow shade, whose angle indicates the size of convolution kernel for the roughness value.

While there exist several works that approximate specular illumination from a hit point, IBL-NeRF alleviates the need for Monte-Carlo integration and greatly reduces the computational burden. Table 1 summarizes the comparison of IBL-NeRF against NeRFactor [ZSD*21], which is a representative formulation with environment light [ZSD*21, SDZ*21, BXS*20]. Specifically, the Monte-Carlo integration aggregates N_d directional samples of reflected rays from the surface points as shown in shaded cones in Fig. 4. In addition to the N_s samples along the camera ray for the volume rendering of NeRF, each reflected ray is evaluated with N_r samples of towards the surrounding environment lighting. Although NeRFactor directly fetch N_d light samples from environment map according to the visibility MLP output for each direction, they need to query N_r samples along each direction to train visibility MLP which is originally used in NeRV [SDZ*21]. The variants using Monte-Carlo integration therefore require evaluating $\mathcal{O}(N_s + N_dN_r)$ samples. On the other hand, IBL-NeRF proposes fetching a single ray of the prefiltered radiance field $L_{\text{pref}}(\mathbf{x}_{\text{surf}}, \omega_r, \gamma)$ in the place of the environment map, leading to evaluating $\mathcal{O}(N_s + N_r)$ samples, as depicted in Fig. 5.

Additionally, IBL-NeRF can process general scenes with diverse lighting or viewpoints as long as the original NeRF converges. The prefiltered radiance fields is defined for the entire scene volume for

any position $\mathbf{x}_{\text{surf}}$ and or direction ω_r. This is in contrast to the approaches relying on a single environment light which is an infinite-radius spherical image enclosing the entire scene, as they assume an isolated object distant from other scene properties, especially lighting. Therefore it cannot render from viewpoints within the volume, diverse objects spread throughout the scene, or indoor scenes with interior lighting.

Our specular reflection is evaluated as a single ray for the given roughness value within the scene volume since the prefiltered radiance $L_{\text{pref}}(\mathbf{x}_{\text{surf}}, \omega_r, \gamma)$ already aggregates the directional rays. Specifically, IBL-NeRF outputs prefiltered radiance fields L^j_{pref} with different convolution levels j. The prefiltered radiance of the desired roughness γ at a certain point $\mathbf{x}$ with direction ω uses trilinear interpolation as

$$L_{\text{pref}}(\mathbf{x}, \omega, \gamma) = \sum_j w^j(\gamma) L^j_{\text{pref}}(\mathbf{x}, \omega), \tag{4}$$

where $w^j(\gamma)$ is the weight of jth mipmap that depends on the roughness γ as described in Fig. 4. The values stored in the prefiltered radiance fields correspond to specific roughness values, and we linearly interpolate them to adjust to the current γ value. Therefore, we evaluate the prefiltered radiance by fetching a sample of a single ray, similar to texture mipmap.

The prefiltered radiance $\tilde{L}^j_{\text{pref}}$ is inferred from the directional MLPs using the similar volume rendering equation

$$L^j_{\text{pref}}(\mathbf{c}, -\omega_o) = \int_0^\infty V(\mathbf{x}(t), \mathbf{c}) \sigma(\mathbf{x}(t)) \tilde{L}^j_{\text{pref}}(\mathbf{x}(t), -\omega_o) dt. \tag{5}$$

For training, we use a set of images blurred with a discrete set of Gaussian filters from the camera position $-\omega_o$. During the inference of the image, the values of $\tilde{L}^j_{\text{pref}}$ are fetched to render the surface point $\mathbf{x}_{\text{surf}}$ as explained in Sec. 3.1.2 and Eq. 4. Note that the training target is the blurred images observed from the camera $(\mathbf{c}, -\omega_o)$, whereas the inference is evaluated from the reflected direction $(\mathbf{x}_{\text{surf}}, \omega_r)$. The formulation relies on the assumption that training images contain observations of the reflected rays.

3.2.1. Image-Space Approximation

The prefiltered radiance L^j_{pref} of IBL-NeRF incorporates the image-based rendering within the implicit volume of NeRF and achieves computational efficiency. We further analyze the practical considerations with the image-space approximation of Gaussian filters to emulate the specular reflection blobs of different surface roughness. The jth prefiltered radiance L^j_{pref} is approximated for the roughness value γ_j as

$$L^j_{\text{pref}} = \int_\Omega L_i(\mathbf{x}, \omega_i) p(\omega_i | \mathbf{x}, \omega, \gamma_j) d\omega_i \tag{6}$$

$$= \int_S L_i(s_i) p_S(s_i | \mathbf{x}, \omega, \gamma_j) ds. \tag{7}$$

Previous approaches approximate the sampling distribution by inferring radiance multiple times in hemispherical domain Ω (Eq. 6) which is computationally heavy [ZSD*21, SDZ*21]. Our method converts the domain into the image space S of the current view as Eq. 7, where s_i is the screen space coordinate that corresponds to direction ω_i. When rendering for a viewpoint, the viewing direction ω_o can be assumed to be constant, and we can use a globally consistent kernel $L^j_{\text{pref}}(\mathbf{x}, \omega) = K^j(L(s))$, where $K^j(s_i) \propto p_S(s_i | \mathbf{x}, \omega, \gamma_j)$. We include the full derivation of our approximation and plots of $K^j(s_i)$ in the supplementary material. The overall shape of $K^j(s_i)$ is similar to that of the Gaussian function, which is used to approximate $L^j_{\text{pref}}(\mathbf{x}, \omega)$ in our implementation. While Gaussian kernels in image space mostly result in a reasonable approximation, they deviate from direct filtering of the environment map for pixels near the image edge. Additional discussion on our approximation can be found in the supplementary material.

3.3. Training IBL-NeRF

IBL-NeRF imposes constraints on the rendered images to train the neural volume, similar to vanilla NeRF. The objective function is composed of four terms:

$$\mathcal{L} = \mathcal{L}_{\text{render}} + \mathcal{L}_{\text{pref}} + \mathcal{L}_{\text{prior}} + \lambda_{I,\text{reg}} \mathcal{L}_{I,\text{reg}}. \tag{8}$$

The first two components are rendering losses to match the rendered images with the input images. For each pixel of the camera ray $r = (\mathbf{c}, -\omega_o)$, the rendering loss $\mathcal{L}_{\text{render}}$ of approximated radiance is defined as

$$\mathcal{L}_{\text{render}} = \|L_o(r) - \hat{L}_o(r)\|_2^2, \tag{9}$$

where $\hat{L}_o$ is ground truth radiance and L_o is our approximated radiance calculated with Eq. 2. $\mathcal{L}_{\text{pref}}$ is the rendering loss of prefiltered radiance defined as

$$\mathcal{L}_{\text{pref}} = \sum_j \|L^j_{\text{pref}}(r) - L^j_{\text{G}}(r)\|_2^2. \tag{10}$$

L^j_{pref} is inferred prefiltered radiance of j^{th} level and L^j_{G} is the radiance convolved with j^{th} level Gaussian convolution, where $L^0_{\text{G}} = L$.

Inverse rendering is under-constrained in nature, and the remaining two losses incorporate additional prior knowledge to estimate intrinsic components. We obtain the pseudo albedo $\hat{a}$ and irradiance $\hat{I}$ for our input images by applying intrinsic decomposition for single images [BBS14], and use them as data-driven prior. The prior loss $\mathcal{L}_{\text{prior}}$ encourages our inferred albedo a to match the pseudo albedo

$$\mathcal{L}_{\text{prior}} = \|a(r) - \hat{a}(r)\|_2^2. \tag{11}$$

In addition, $\mathcal{L}_{I,\text{reg}}$ is the irradiance regularization loss

$$\mathcal{L}_{I,\text{reg}} = \|I(r) - \mathbb{E}[\hat{I}]\|_2^2, \tag{12}$$

where $\mathbb{E}[\hat{I}]$ is the mean of irradiance (shading) values in training set images. Although the results from single-image decomposition are inconsistent for different viewpoints, our neural volume learns multi-view consistent and smooth results. We provide more detailed comparison between IBL-NeRF and results from single-image decomposition methods in Sec. 4.1.

4. Experiments

Dataset First, we test IBL-NeRF in 12 realistic synthetic indoor scenes [Bit16], which are capable of obtaining ground-truth intrinsic components. We render 100 multi-view images for both training and test set with the OptiX [PBD*10] based path tracer [KK21].

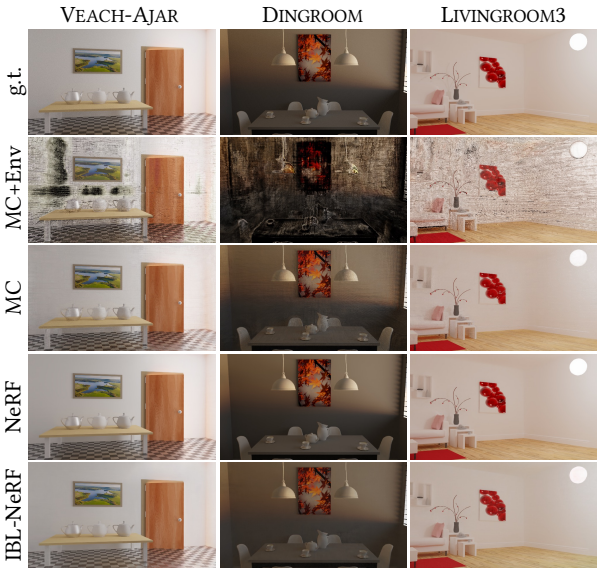


Figure 6: *Qualitative results of novel-view image synthesis.*

Method	MSE ↓	PSNR ↑	SSIM ↑	Time per step (s) Train	Infer
MC + Env [ZSD*21]	0.0369	16.107	0.2763	0.4686	0.1062
MC	0.0016	**30.052**	0.8348	0.4941	0.1084
NeRF	**0.0008**	**34.707**	**0.9253**	**0.0984**	**0.0055**
IBL-NeRF	**0.0014**	29.962	**0.9009**	**0.1559**	**0.0211**

Table 2: *Quantitative results of view synthesis.*

All of the scenes in our dataset exhibit complex lighting with windows or interior lighting and contain multiple objects with challenging material, which cannot be modeled with an environment light. This is in contrast to previous works for decomposing NeRF, which present results with isolated objects [ZSD*21, SDZ*21]. The camera's position and rotation are randomly sampled within a scene bounding box. All the results reported in the manuscript are the novel viewpoints in the testset which are not seen during the training images. We linearly interpolate between the test camera poses to generate results of the supplementary video. Furthermore, we test IBL-NeRF in real-world scenes from ScanNet dataset [DCS*17] and our own captured scene. For ScanNet scenes, we use train/test split from [WLR*21]. The camera poses are estimated with COLMAP [SF16] for real scenes.

Implementation Details The neural network architecture is illustrated in Fig. 3. We use the same MLP configurations with vanilla NeRF [MST*20], except that IBL-NeRF has additional layers to output albedo (a), irradiance (I), and roughness (γ) at position MLP, and 4 parallel layers to emit the prefilterd radiance fields (L^j_{pref}) for each roughness level $0 \leq j \leq 3$. IBL-NeRF is trained for 120k steps with 512 ray samples, and follows the training schedule below to stabilize the process. For the first 10k steps, we only optimize L^j_{pref} and σ with $\mathcal{L}_{\text{pref}}$. Once we obtain stable geometry and prefiltered radiance fields, we additionally optimize for $\mathcal{L}_{\text{pref}} + \mathcal{L}_{\text{render}}$ without

Figure 7: *Qualitative results of intrinsic decomposition and view synthesis on real-world datasets.*

prior. Then we freeze roughness and apply priors $\mathcal{L}_{\text{prior}}, \mathcal{L}_{I,\text{reg}}$ for last 20k steps. We use $\lambda_{I,\text{reg}} = 0.1$ empirically, but we observe that the final result is not very sensitive to the value of $\lambda_{I,\text{reg}}$. Also, we assume monochromatic irradiance for simplicity.

4.1. View Synthesis & Intrinsic Decomposition

Baselines We compare IBL-NeRF with two baselines with Monte Carlo (MC) sampling over a hemisphere of environment light. Since Neural Reflectance Fields [BXS*20] and NeRV [SDZ*21] need known lighting formulation to train models, they cannot be applied to our scenario with unknown lighting conditions. The first baseline (MC) is a variant of IBL-NeRF, which exploits the radiance field ($L^0_{\text{pref}} = L_o$) as incoming light for specular reflection and calculates integration with MC sampling. The second baseline (MC + Env) estimates single environment light for the entire scene as L_i and employs MC integration as in NeRFactor and NeRV. MC + Env is the microfacet BRDF version of NeRFactor [ZSD*21] without visibility inference network, which is the most relevant work to us. We found that the original NeRFactor which exploits learned BRDF prior does not converge in any of our scenes. For all the baselines with Monte Carlo approaches, we use 32×16 resolution environment light following NeRFactor. Also, we use equal-area stratified sampling over hemispheres with 64 samples.

We report the quantitative results of the novel-view synthesis in Table 2 and intrinsic decomposition in Table 3 in terms of MSE, PSNR, and SSIM. IBL-NeRF models outgoing radiance as the

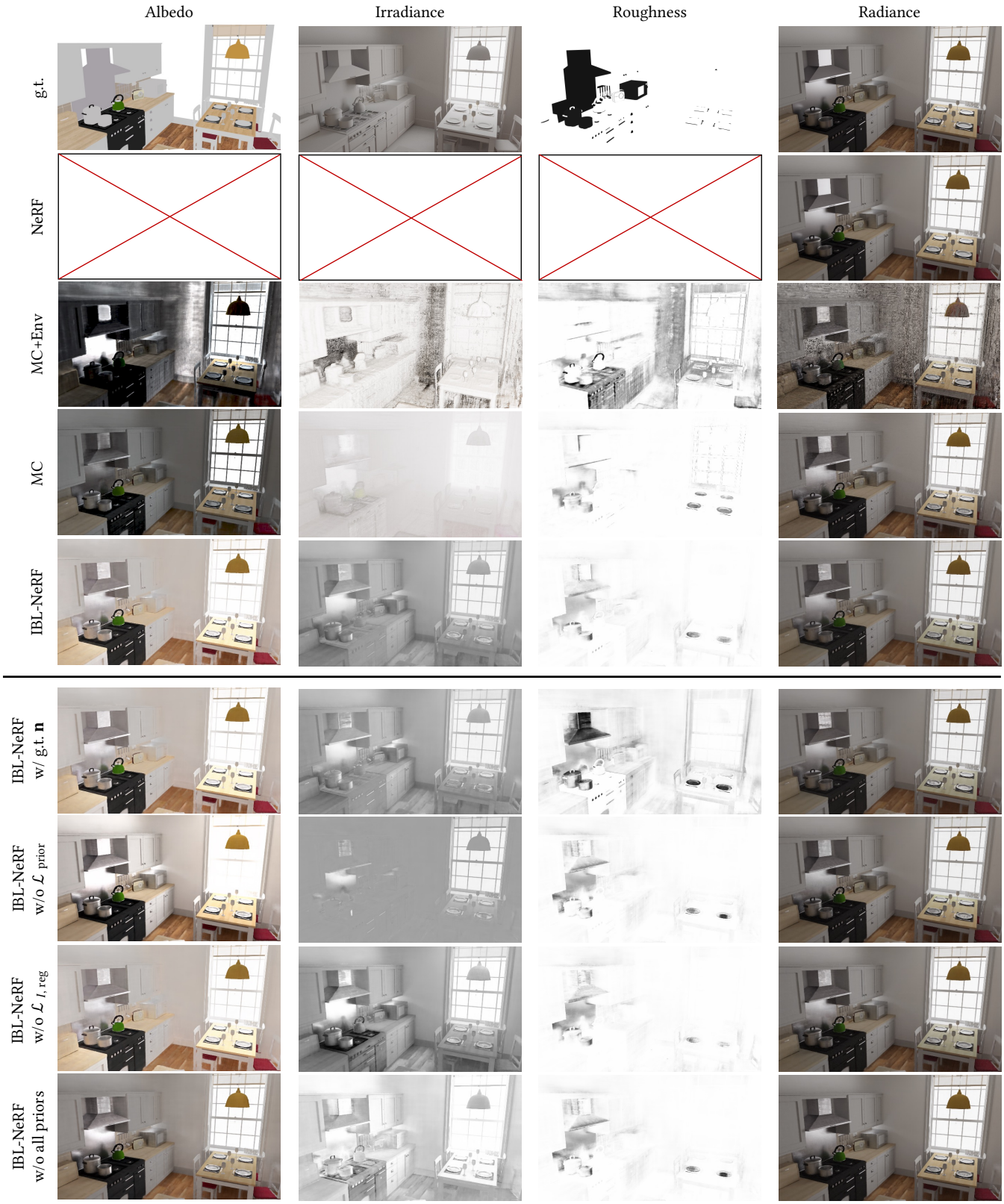


Figure 8: *Qualitative results of intrinsic decomposition and view synthesis on* KITCHEN.

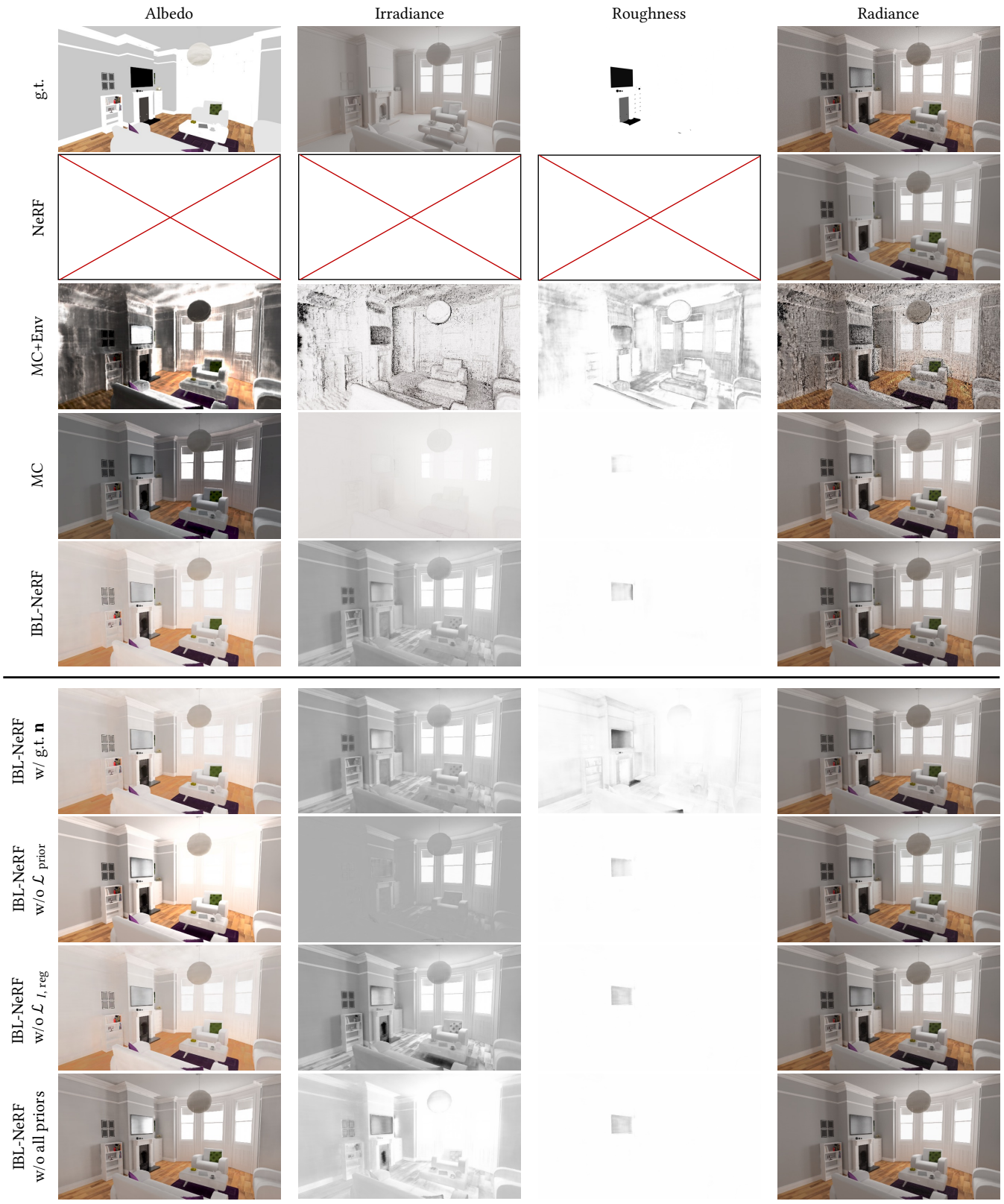


Figure 9: *Qualitative results of intrinsic decomposition and view synthesis on* LIVINGROOM2.

	Albedo			Irradiance			Roughness		
	MSE ↓	PSNR ↑	SSIM ↑	MSE ↓	PSNR ↑	SSIM ↑	MSE ↓	PSNR ↑	SSIM ↑
MC + Env (NeRFactor)	0.1808	8.1273	0.3916	0.1190	10.220	0.2514	0.0910	11.798	0.6217
MC	**0.0543**	14.109	0.7383	**0.0344**	**17.280**	0.7149	0.0722	14.090	0.7474
IBL-NeRF	0.0553	**14.114**	**0.7455**	0.0351	16.435	**0.7778**	**0.0707**	**15.545**	**0.8653**
w/ GT **n**	0.0551	**14.134**	**0.7465**	0.0376	15.986	0.7717	**0.0623**	14.216	0.8220
w/o L_{prior}	0.0664	13.423	0.7107	0.0403	15.609	0.7553	0.0717	15.413	0.8613
w/o $L_{I,\text{reg}}$	0.0551	14.077	0.7362	**0.0337**	16.215	0.7586	0.0710	14.316	0.8588
w/o all priors	0.0775	11.601	0.6911	0.0674	12.147	0.7015	0.0709	15.527	0.8637

Table 3: *Quantitative results of intrinsic decomposition. Compared to MC + Env (recent works on intrinsic decomposition), our result achieves better intrinsic decomposition of the scene. Note that MC shows comparable performance in the expense of substantially longer training time as it performs Monte-Carlo integration (Table 2). Normal information is critical to estimate the shading information (irradiance, roughness), and the additional regularization on albedo or irradiance balances the distribution between different intrinsic components.*

combination of various intrinsic components and concurrently generates images whose quality is comparable to vanilla NeRF. Notably, our approach outperforms the method from NeRFactor (MC + Env) in both intrinsic decomposition and image synthesis results for all error metrics, which supports our claim that using environment lighting with MC sampling is inadequate to express complex indoor scenes. The reconstruction quality is much better by alleviating the environment light and instead adapting our formulation in Eq. 2. Theoretically, the MC baseline should have better results in the expense of computation time, which is almost 3 times slower in the training phase and 5 times slower in the inference phase than IBL-NeRF. However, since there exists a number of invalid samples in the incident radiance that are invisible from training viewpoints, the decomposition of MC is comparable to ours. The results for MC + Env do not incorporate the albedo prior, as it achieves better performance. We report the second baseline method (MC + Env) with $\mathcal{L}_{\text{prior}}$ in supplementary material.

We demonstrate the qualitative results of novel-view synthesis and intrinsic decomposition in synthetic scenes in Fig. 6, 8 and 9, real scenes in Fig. 7. Our approach and MC approach with prefiltered radiance field reconstruct high-quality images in novel viewpoints, which are comparable to vanilla NeRF. On the other hand, objects in large-scale indoor scenes are often occluded by other structures within the scene, and therefore cannot be illuminated appropriately with environment light (MC + Env). The quality of images is significantly worse as it suffers from notable dark and noisy artifacts created from missing viewpoints or ambiguous regions. Fig. 7, 8 and 9 show that IBL-NeRF successfully decomposes the scene attributes in both synthetic and real-world scenes. IBL-NeRF estimates low roughness at metallic surfaces, for example, the ventilator, metallic wall, knobs in the oven, and pots in KITCHEN in Fig. 8, TV in LIVINGROOM2 in Fig. 9. However, our method fails to discover metallic surface that does not have specular variation with respect to viewing direction in the training set. (For example, the fireplace in LIVINGROOM2 has consistent color in the training images.)

Furthermore, IBL-NeRF can easily achieve the inherent multi-view consistency and smoothness of our optimizing process as shown in Fig. 10. While the intrinsic decomposition algorithms for single-view images [BBS14, ZKE15] fail to maintain consistent results, it provides useful guidance for the intrinsic decomposition.

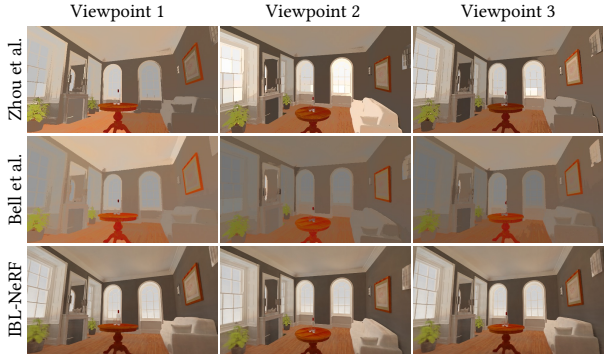


Figure 10: *Visual comparison of albedo estimation between IBL-NeRF and single-image based methods.*

Ablation Studies Fig. 8, 9 and Table 3 also contain results for ablated versions of IBL-NeRF to analyze the important components of the proposed method. The qualitative results with ground-truth normal **n** show cleaner roughness than our original model. The effect of roughness is tightly coupled with the direction of mirror reflection, which is obtained from the surface normal. Recent methods [OPG21, WLL*21] propose to reconstruct high-quality geometry with NeRF formulations, from which IBL-NeRF can learn better decomposition.

Since intrinsic decomposition is an under-constrained problem, prior knowledge on intrinsic components plays a crucial role to disambiguate each component. When we remove $\mathcal{L}_{\text{prior}}$ the albedo contains illumination information which should belong to irradiance, and the irradiance is clipped to the mean value by $\mathcal{L}_{I,\text{reg}}$. Also, without $\mathcal{L}_{I,\text{reg}}$, one cannot estimate correct irradiance, especially on the surface with dark albedo. (e.g., Oven in Fig. 8 should have irradiance similar to nearby furniture, but the dark pixels encourage estimating lower irradiance without $\mathcal{L}_{I,\text{reg}}$.) Removing both priors shows the worst results.

(a) Original Scene (b) Edited Scene

Figure 11: *Example of changing intrinsic components of the scene.*

4.2. Scene Editing

After IBL-NeRF decomposes intrinsic components, one can render realistic novel-view images of altered scenes by modifying the value of each component. For example, we replace roughness of the dining table and albedo of the lamp to edit KITCHEN scene in Fig. 1(d). We demonstrate more results in Fig. 11. In the first row of Fig. 11, we replace albedo of two kettles in VEACH-AJAR to green and red respectively while preserving illumination information. In the second row of Fig. 11, we reduce the roughness of the picture in frame, drawer, and closet door, which results in mirror-like material in BEDROOM scene. We also change the albedo of the middle door of the closet to white and the conference logo is marked on the left door by modifying roughness. In the third row of Fig. 11, we modify the albedo and roughness of the desk pad in our real-world scene to express the marble-like material. Also, one can insert 3D objects inside our trained neural volume with prefiltered radiance field. In Fig. 12, we add 3 objects with different roughness and transparency inside the KITCHEN. The red blobby object is highly reflective and the surrounding scene is clearly reflected on its surface. The green dragon also has a low roughness value but has translucency so the shape of the green kettle behind the object is visible. Finally, the blue teapot has a high roughness value and moderate translucency. The blurry reflection on the teapot accounts for its high roughness value. Note that scene editing could be achieved similarly using Monte Carlo method with NeRF's radiance, but IBL-NeRF outperforms them in terms of speed (Table 2, Infer time).

5. Conclusion

We propose IBL-NeRF, a neural volume representation with prefiltered radiance field. Our approach successfully decomposes the intrinsic components in a large-scale scene with an efficient approximation and prefiltered radiance field, which could not be processed in prior works with Monte Carlo integration of environment light. Furthermore, one can easily edit the scene by modifying each decomposed component or inserting 3D models in our neural volume. Although IBL-NeRF can handle both Lambertian reflection and specular reflection, IBL-NeRF has a limitation in expressing transparent objects or perfect-mirror reflection. One can resolve the ambiguity in a mirror with user interaction as [GKB*21].

Figure 12: *Example of adding new objects to the scene.*

Acknowledgements This work was partly supported by Korea Institute for Advancement of Technology (KIAT) grant funded by the Korea Government (MOTIE) (P0012746, HRD Program for Industrial Innovation), and the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2023.

References

[BBJ*21] BOSS M., BRAUN R., JAMPANI V., BARRON J. T., LIU C., LENSCH H.: Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 12684–12694. 2

[BBS14] BELL S., BALA K., SNAVELY N.: Intrinsic images in the wild. *ACM Transactions on Graphics (TOG) 33*, 4 (2014), 1–12. 2, 5, 9

[Bit16] BITTERLI B.: Rendering resources, 2016. https://benedikt-bitterli.me/resources/. 5

[BJB*21] BOSS M., JAMPANI V., BRAUN R., LIU C., BARRON J. T., LENSCH H. P.: Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *Advances in Neural Information Processing Systems (NeurIPS)* (2021). 2

[BXS*20] BI S., XU Z., SRINIVASAN P., MILDENHALL B., SUNKAVALLI K., HAŠAN M., HOLD-GEOFFROY Y., KRIEGMAN D., RAMAMOORTHI R.: Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824* (2020). 2, 4, 6

[CZL18] CHENG L., ZHANG C., LIAO Z.: Intrinsic image transformation via scale space decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 656–665. 2

[DCS*17] DAI A., CHANG A. X., SAVVA M., HALBER M., FUNKHOUSER T., NIESSNER M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE* (2017). 6

[DRC*15] DUCHÊNE S., RIANT C., CHAURASIA G., LOPEZ-MORENO J., LAFFONT P.-Y., POPOV S., BOUSSEAU A., DRETTAKIS G.: Multi-view intrinsic images of outdoors scenes with an application to relighting. *ACM Transactions on Graphics* (2015), 16. 2

[GKB*21] GUO Y.-C., KANG D., BAO L., HE Y., ZHANG S.-H.: Nerfren: Neural radiance fields with reflections. *arXiv preprint arXiv:2111.15234* (2021). 10

[GMLMG12] GARCES E., MUNOZ A., LOPEZ-MORENO J., GUTIERREZ D.: Intrinsic images by clustering. In *Computer graphics forum* (2012), vol. 31, Wiley Online Library, pp. 1415–1424. 2

[HHM22] HASSELGREN J., HOFMANN N., MUNKBERG J.: Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv:2206.03380* (2022). 2

[Kaj86] KAJIYA J. T.: The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques* (1986), pp. 143–150. 3

[Kar13] KARIS B.: Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice 4*, 3 (2013). 3

[KK21] KIM J., KIM Y. M.: Fast and Lightweight Path Guiding Algorithm on GPU. In *Pacific Graphics Short Papers, Posters, and Work-in-Progress Papers* (2021), Lee S.-H., Zollmann S., Okabe M., Wünsche B., (Eds.), The Eurographics Association. doi:10.2312/pg.20211379. 5

[LBP*12] LAFFONT P.-Y., BOUSSEAU A., PARIS S., DURAND F., DRETTAKIS G.: Coherent intrinsic images from photo collections. *ACM Transactions on Graphics 31*, 6 (2012). 2

[LS18a] LI Z., SNAVELY N.: Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 371–387. 2

[LS18b] LI Z., SNAVELY N.: Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 9039–9048. 2

[LSR*20] LI Z., SHAFIEI M., RAMAMOORTHI R., SUNKAVALLI K., CHANDRAKER M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 2475–2484. 2

[LWC*23] LI Z., WANG L., CHENG M., PAN C., YANG J.: Multi-view inverse rendering for large-scale real-world indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023). 2

[MCZ*18] MA W.-C., CHU H., ZHOU B., URTASUN R., TORRALBA A.: Single image intrinsic decomposition without a single intrinsic image. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 201–217. 2

[MST*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision* (2020), Springer, pp. 405–421. 1, 2, 6

[MSZ*21] MEKA A., SHAFIEI M., ZOLLHÖFER M., RICHARDT C., THEOBALT C.: Real-time global illumination decomposition of videos. *ACM Transactions on Graphics (TOG) 40*, 3 (2021), 1–16. 2

[OPG21] OECHSLE M., PENG S., GEIGER A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 5589–5599. 9

[PBD*10] PARKER S. G., BIGLER J., DIETRICH A., FRIEDRICH H., HOBEROCK J., LUEBKE D., MCALLISTER D., MCGUIRE M., MORLEY K., ROBISON A., ET AL.: Optix: a general purpose ray tracing engine. *Acm transactions on graphics (tog) 29*, 4 (2010), 1–13. 5

[PEL*21] PANDEY R., ESCOLANO S. O., LEGENDRE C., HÄNE C., BOUAZIZ S., RHEMANN C., DEBEVEC P., FANELLO S.: Total relighting: Learning to relight portraits for background replacement. *ACM Trans. Graph. 40*, 4 (jul 2021). URL: https://doi.org/10.1145/3450626.3459872, doi:10.1145/3450626.3459872. 2

[PMGD21] PHILIP J., MORGENTHALER S., GHARBI M., DRETTAKIS G.: Free-viewpoint indoor neural relighting from multi-view stereo. *ACM Trans. Graph. 40*, 5 (sep 2021). URL: https://doi.org/10.1145/3469842, doi:10.1145/3469842. 2

[SDZ*21] SRINIVASAN P. P., DENG B., ZHANG X., TANCIK M., MILDENHALL B., BARRON J. T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 7495–7504. 2, 4, 5, 6

[SF16] SCHONBERGER J. L., FRAHM J.-M.: Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (June 2016). 6

[SGK*19] SENGUPTA S., GU J., KIM K., LIU G., JACOBS D. W., KAUTZ J.: Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8598–8607. 2

[VHM*22] VERBIN D., HEDMAN P., MILDENHALL B., ZICKLER T., BARRON J. T., SRINIVASAN P. P.: Ref-NeRF: Structured view-dependent appearance for neural radiance fields. *CVPR* (2022). 2

[WLL*21] WANG P., LIU L., LIU Y., THEOBALT C., KOMURA T., WANG W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems* (2021), Ranzato M., Beygelzimer A., Dauphin Y., Liang P., Vaughan J. W., (Eds.), vol. 34, Curran Associates, Inc., pp. 27171–27183. URL: https://proceedings.neurips.cc/paper/2021/file/e41e164f7485ec4a28741a2d0ea41c74-Paper.pdf. 9

[WLR*21] WEI Y., LIU S., RAO Y., ZHAO W., LU J., ZHOU J.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* (October 2021), pp. 5610–5619. 6

[YTL20] YI R., TAN P., LIN S.: Leveraging multi-view image sets for unsupervised intrinsic image decomposition and highlight separation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (2020), vol. 34, pp. 12685–12692. 2

[ZHY*23] ZHU J., HUO Y., YE Q., LUAN F., LI J., XI D., WANG L., TANG R., HUA W., BAO H., ET AL.: I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2023), pp. 12489–12498. 2

[ZKE15] ZHOU T., KRAHENBUHL P., EFROS A. A.: Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the IEEE International Conference on Computer Vision* (2015), pp. 3469–3477. 2, 9

[ZLW*21] ZHANG K., LUAN F., WANG Q., BALA K., SNAVELY N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2021), pp. 5453–5462. 2

[ZSD*21] ZHANG X., SRINIVASAN P. P., DENG B., DEBEVEC P., FREEMAN W. T., BARRON J. T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *arXiv:2106.01970* (2021). 2, 4, 5, 6