Neural Radiance Transfer Fields for Relightable Novel-view Synthesis with Global Illumination

Linjie Lyu[1], Ayush Tewari[2], Thomas Leimkühler[1], Marc Habermann[1], and Christian Theobalt[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus
[2]MIT

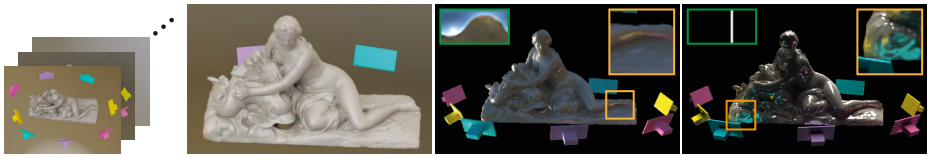

Fig. 1. Our method takes multiple views of a scene under one unknown illumination condition as input and allows novel-view synthesis and relighting (corresponding environment maps in green insets) with intricate multi-bounce illumination (orange insets).

Abstract. Given a set of images of a scene, the re-rendering of this scene from novel views and lighting conditions is an important and challenging problem in Computer Vision and Graphics. On the one hand, most existing works in Computer Vision usually impose many assumptions regarding the image formation process, e.g. direct illumination and predefined materials, to make scene parameter estimation tractable. On the other hand, mature Computer Graphics tools allow modeling of complex photo-realistic light transport given all the scene parameters. Combining these approaches, we propose a method for scene relighting under novel views by learning a neural precomputed radiance transfer function, which implicitly handles global illumination effects using novel environment maps. Our method can be solely supervised on a set of real images of the scene under a single unknown lighting condition. To disambiguate the task during training, we tightly integrate a differentiable path tracer in the training process and propose a combination of a synthesized OLAT and a real image loss. Results show that the recovered disentanglement of scene parameters improves significantly over the current state of the art and, thus, also our re-rendering results are more realistic and accurate.

1 Introduction

The image formation process is influenced by many factors such as the scene geometry, the object materials, the lighting, and the properties of the recording camera. Recovering these properties solely from the final images of the scene is an important inverse problem in Computer Vision and enables several applications such as scene understanding, virtual reality, and controllable image synthesis.

Since this is an ill-posed and challenging inverse problem, existing methods make several assumptions about the 3D scene. Common assumptions are that scenes are diffuse [47], or can be described by some predefined material models [54, 51]. Importantly, most methods only consider the direct scene illumination [30, 54, 3, 4, 38, 51, 25]. These assumptions limit existing methods from recovering accurate and rich scene properties, resulting in limited re-rendering results as well, e.g. global illumination effects cannot be modeled.

In parallel, the field of Computer Graphics has extensively researched the problem of photorealistic image synthesis. These methods take a well-defined 3D scene and render a realistic image. Methods have explored different ways of modeling indirect illumination using path tracing. Since most path tracing methods are inefficient, precomputed radiance transfer (PRT) was introduced as an efficient approximation of global illumination [36, 35, 37, 43, 46, 16] . However, these approaches usually do not consider *recovering the PRT solely from images*.

In this paper, we combine the learning of precomputed radiance transfer function with inverse rendering, thus combining the best of Computer Vision and Computer Graphics. The precomputed radiance transfer function is parameterized as a neural network. Thus, it does not require any predefined approximation function, e.g. spherical harmonics. As we model the material using a learned PRT, our method does not share common limitations with existing inverse rendering methods – our method is capable of dealing with complex light paths such as indirect reflections and shadows, and is also not limited to any predefined BRDF model. Our method is learned on multi-view observations of a scene under a single unknown light condition. In addition to the PRT, it also recovers the scene illumination as an environment map, and the scene geometry defined as a neural signed distance field. Thus, our method enables applications such as novel-view synthesis and global relighting using environment maps. Existing methods, which enable these applications while taking global illumination into account, rely on light-stage datasets, where the object is captured from multiple views under different light conditions. We show that such a setup is not essential. In contrast to real light-stage data, we generate synthetic light-stage data of the scene using a high-quality renderer. This enables correct disentanglement of the material and illumination properties in the scene, while the real multi-view data allows us to capture photorealistic details and to overcome the common assumptions made by the renderer. In summary, our contributions are:

- A method for recovering the radiance transfer field from images of objects under an unknown light condition, hence enabling free-viewpoint relighting with realistic global illumination.
- A *neural* precomputed radiance transfer (PRT) field for multi-bounce global illumination computation and neural implicit surface rendering.
- A novel supervision strategy leveraging a differentiable ray-tracer for physically based scene reconstruction, multiple light bounce rendering, and a new synthetic OLAT supervision.

Our qualitative and quantitative results demonstrate a clear step forward in terms of the recovery of scene parameters as well as the synthesis quality of

our approach under novel views and lighting conditions when comparing to the previous state of the art. We will make the code and the new dataset publicly available.

2 Related Work

In the following, we focus on previous work concerning radiance transfer and inverse rendering. Although our method also recovers the scene geometry using an off-the-shelf implicit geometry reconstruction approach [44], it is not the main focus of this work and, thus, we do not review related work in this area.

Precomputed Radiance Transfer. Precomputed Radiance Transfer (PRT) [36] is a powerful approach for efficient rendering of global illumination [35]. Typically, static geometry and reflectance in combination with distant illumination are assumed, which allows to partially precompute light transport for free-viewpoint synthesis and dynamic lighting. Extensions to e.g., dynamic objects [37] or near-field illumination [43, 46] exist. The generic formalization of PRT [16] enables the incorporation of arbitrarily complex light transport, including multi-bounce light paths. While these works improve the runtime of the forward rendering pipeline, they do not consider recovering the PRT solely from a set of real world images of the object. In contrast, we employ this concept to efficiently decompose illumination and reflectance for view synthesis and relighting, taking into account full global illumination, but apply these concepts in an inverse setting where we aim to recover the PRT from images by means of training a neural PRT network.

The versatility of PRT has encouraged the exploration of different angular basis functions, such as spherical harmonics [36, 13], Haar wavelets [26], spherical isotropic [42] and anisotropic [49] radial basis functions. While the inherent prior of such basis functions can be beneficial for inverse problems, they also limit the range of illumination effects that can be explained by such a basis. As a remedy, we use the primal directional basis [9] in combination with a neural network, to overcome the limitations of classical basis functions. We encode the full radiance transfer into a neural field [40, 48]. Recently, Rainer et al. [32] have explored PRT-inspired neural field-based forward rendering of synthetic scenes – with full knowledge of all scene parameters, as in most works discussed above. In contrast, our framework is concerned with global illumination-aware novel-view synthesis and relighting from multi-view data under one unknown illumination condition. Further, the transfer from distant illumination to local lighting is traditionally concerned with the *incoming* radiance at a surface point [12], i.e., the convolution with reflectance is excluded from precomputation to increase efficiency and to reduce storage requirements. In contrast, we follow ideas from PRT-based relighting [26, 41] in that we directly predict *outgoing* radiance.

Inverse Rendering and Relighting. Inverse rendering [22, 33] aims at estimating scene properties such as geometry, lighting, and materials from image observations. In this work, we are particularly interested in decoupling lighting using

multi-view data, and therefore focus our literature review on corresponding related work in illumination decomposition and relighting.

Controllable illumination in a multi-view light stage [7] is conceptually the most straightforward way of obtaining a light-reflectance decomposition in the presence of global illumination, via one-light-at-a-time (OLAT) captures. Even though not trivial, novel-view synthesis and relighting boils down to clever interpolation [39, 53, 29]. In contrast, input to our method is casually-captured multi-view data under unknown illumination, while embedding *synthetic* OLAT data generation into the training process to aid disentanglement.

Techniques for inverse rendering from multi-view data typically impose strong assumptions on lighting and material, with shading models commonly only considering *direct* illumination [30, 54, 3, 4, 38, 51, 25]. Different scene representations have been explored in this context, including meshes [30, 25], signed distance functions (SDFs) [51], or neural radiance or reflectance fields [23, 54, 3, 4, 38]. A common paradigm is the explicit reconstruction of a material representation, e.g., an albedo and roughness map, limiting them to recover appearance effects within the range of these predefined representations. In contrast, our approach seeks to decompose observed color into illumination and a radiance transfer function in a surface-based scene representation, enabling relighting with intricate *indirect* illumination, while reconstructing materials only for supervision.

Incorporating multiple light bounces into inverse rendering and relighting can be done by using heuristic lighting models [14, 21], by assuming known illumination [8], or by employing physically-based rendering to approximate irradiance [31]. Chen et al. [5] approximate PRT in neural rendering, given geometry, without physically-based modeling of multiple light bounces. Thul et al. [41] utilize PRT in a custom optimization to perform global illumination-aware decomposition of lighting and materials, approximating the required gradients with a single-bounce model. In contrast, differentiable path tracing [28, 18, 2] can be used to obtain full gradients for global illumination-aware inverse rendering [1, 27]. We also leverage the concept of differentiable path tracing [28] during training as a means for achieving disentanglement. Different from path tracing, our performance at inference time is independent of light transport complexity and by design produces noise-free renderings of multi-bounce illumination.

3 Method

Our method takes $m \approx 64$ posed multi-view images under an unknown illumination condition as input and allows efficient novel-view synthesis and relighting for the object depicted in these images. To this end, our approach leverages the concept of precomputed radiance transfer (PRT) to factor multi-view observations into illumination and reflectance. Thus, at inference, novel illumination conditions in the form of environment maps can be multiplied with our learned reflectance field given a user-defined camera viewpoint.

In more detail, the rendering equation and its equivalent formulation in the PRT framework forms the theoretical foundation of our approach, and we provide

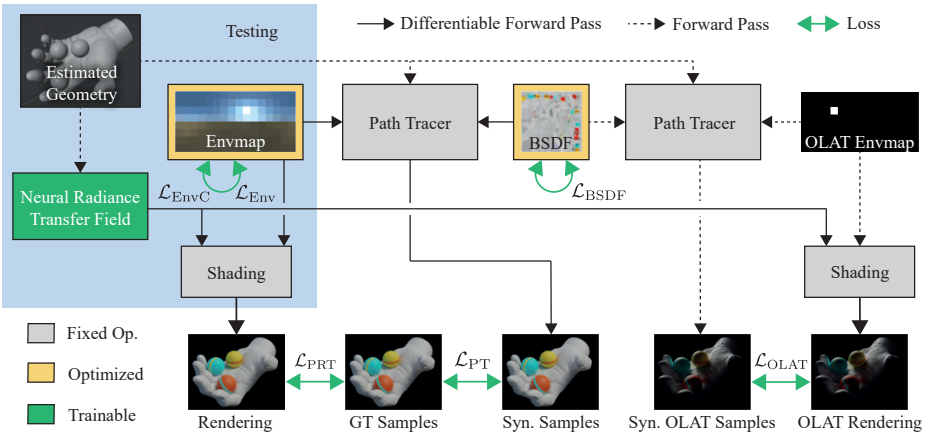

Fig. 2. Overview of our pipeline. The Shading blocks evaluate Eq. 2. The directional inputs to the radiance transfer field are omitted to avoid clutter. The blue area marks the parts run at test time, when any environment map can be used to light the scene with full global illumination.

a brief overview of it (Sec. 3.1). Then, we introduce our neural radiance transfer field (NRTF), a neural network that takes as input a point on the surface and its normal, as well as the incoming and outgoing light directions and predicts the radiance transfer of the scene. This radiance can then be multiplied with an arbitrary environment map enabling global illumination relighting (Sec. 3.2). To achieve this, we first estimate scene geometry from the available multi-view observations using implicit surface reconstruction [44] (Sec. 3.2). In order to train the NRTF, an approximate disentanglement between the observed material and lighting in the multi-view training images is required. Our solution for this is to leverage a differentiable path tracer [28]. It allows the joint optimization of a spatially-varying BSDF and the environment map (Sec. 3.3). Once the BSDF is obtained, the path tracer can be used to synthesize one-light-at-a-time (OLAT) renderings of the scene (Sec. 3.4). The NRTF is trained using a combined loss, consisting of a real image loss that helps to recover photoreal material effects beyond the effects possible with the BSDF model, as well as a synthetic OLAT loss that acts as a prior improving generalization to novel lighting conditions (Sec. 3.4). An overview of our pipeline is shown in Fig. 2.

3.1 Background

We are interested in estimating radiance L arriving from scene point $\mathbf{x} \in \mathbb{R}^3$ in direction $\boldsymbol{\omega}_o \in \Omega$, where Ω denotes the space of 3D directions, i.e., points on the unit sphere. The rendering equation [11] describing global light transport can be formulated as

$$L\left(\mathbf{x}, \boldsymbol{\omega}_o\right) = \int_{\Omega_+} L\left(x, \boldsymbol{\omega}_i\right) \rho\left(\mathbf{x}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o\right) \left(\boldsymbol{\omega}_i \cdot \mathbf{n}\right) d\boldsymbol{\omega}_i \ , \tag{1}$$

where Ω_+ is a hemisphere centered at the surface normal $\mathbf{n}$ of $\mathbf{x}$, $\boldsymbol{\omega}_i$ is an incoming direction, and ρ is the bidirectional scattering distribution function (BSDF) encoding spatially-varying surface material reflectance. Solving this integral equation including global illumination, i.e., multiple light bounces with potentially complex inter-reflections, lends itself to a recursive algorithm like path tracing [11], which stochastically samples light paths to obtain a Monte Carlo estimate of the solution. While modern differentiable variants of path tracing for inverse rendering [27] show promising results, they suffer from high computational costs, especially in the presence of complex light paths. To gain efficiency, we consider distant but otherwise arbitrary illumination, i.e., lighting that can be modeled using an environment map. Therefore, inspired by PRT, we rewrite the rendering equations as

$$L\left(\mathbf{x}, \boldsymbol{\omega}_o\right) = \int_{\Omega_+} L_{\mathrm{e}}\left(\boldsymbol{\omega}_i\right) T\left(\mathbf{x}, \mathbf{n}, \boldsymbol{\omega}_i, \boldsymbol{\omega}_o\right) d\boldsymbol{\omega}_i \ , \tag{2}$$

where $L_e\left(\boldsymbol{\omega}_i\right)$ is the incoming environment light from direction $\boldsymbol{\omega}_i$, which is notably independent of $\mathbf{x}$. The crucial ingredient of this formulation is the collapsed radiance transfer function T, which transforms the global distant illumination L_e from direction $\boldsymbol{\omega}_i$ into local reflected radiance at position $\mathbf{x}$ into direction $\boldsymbol{\omega}_o$. Given an environment map and the scene-specific transfer function T, all that is needed to compute global illumination for a pixel is to obtain the primary intersection point $\mathbf{x}$, evaluate T for all environment map texels, multiply with the respective illumination, and sum all contributions. If T is compact and easy to evaluate, arbitrarily complex global illumination can be efficiently computed on a GPU in a map-reduce fashion.

3.2 Neural Radiance Transfer Field (NRTF)

We model the radiance transfer function T using our neural radiance transfer field

$$T_\theta\left(\mathcal{H}(\mathbf{x}), \mathbf{n}, \mathcal{F}(\boldsymbol{\omega}_i), \mathcal{F}(\boldsymbol{\omega}_o)\right) = \mathbf{c}_t, \tag{3}$$

where $\mathbf{c}_t \in \mathbb{R}^3$ denotes transferred RGB color and θ indicates the trainable parameters. We parameterize T_θ using a multi-layer perceptron (MLP). Here, we apply a multi-resolution hash encoding $\mathcal{H}(\cdot)$ [24] to the 3D position $\mathbf{x}$, and a spherical harmonics encoding $\mathcal{F}(\cdot)$ [50] to light directions $\boldsymbol{\omega}_i$ and $\boldsymbol{\omega}_o$. The hash encoding enables faster training and evaluation of our networks. Details about the network architecture and the encoding strategy can be found in the supplemental document. When rendering an image from an arbitrary camera view centered at $\mathbf{o}$, we shoot a ray $\mathbf{r}(t) = \mathbf{o} + t\boldsymbol{\omega}_o$ through a pixel with 2D coordinate $\mathbf{u}$, and compute the intersection point $\mathbf{x}$ with respect to the scene geometry. At $\mathbf{x}$, we now evaluate a discretized version of Eq. 2 using our learned T_θ: With $\hat{\boldsymbol{\omega}}$ denoting discrete incoming directions, corresponding to the pixels of a discretized environment map $\hat{L}_{\mathrm{e}}$, we write

$$L_\theta(\mathbf{u}) = L_\theta\left(\mathbf{x}, \boldsymbol{\omega}_o\right) = \sum_{\hat{\boldsymbol{\omega}}_i} \hat{L}_{\mathrm{e}}(\hat{\boldsymbol{\omega}}_i) \cdot T_\theta\left(\mathcal{H}(\mathbf{x}), \mathbf{n}, \mathcal{F}(\hat{\boldsymbol{\omega}}_i), \mathcal{F}(\boldsymbol{\omega}_o)\right). \tag{4}$$

Note that this process is repeated for each pixel of the output image. It is worth emphasizing again that this formulation can capture multi-bounce lighting effects and complex material reflectance. Importantly, $\mathbf{x}$, $\boldsymbol{\omega}_o$, and $\hat{L}_\mathrm{e}$, i.e., the camera and the environment map can be modified at test time, enabling free-viewpoint rendering and scene relighting. In the following, we explain how we first obtain the scene geometry from the set of multi-view images and then provide details on how the NRTF can be trained without ground truth scene lighting and material.

Geometry Estimation. In general, our approach is agnostic to the type of surface-based geometry representation. Recent neural rendering methods [40] have demonstrated state-of-the-art shape reconstruction results using implicit neural SDF representations. We leverage the recently proposed NeuS [44] for computing the SDF geometry of the object. NeuS takes multi-view images and camera poses as input and reconstructs the geometry, represented as a neural field. Since rendering an explicit mesh is significantly more efficient than rendering an SDF, we extract a mesh from the implicit surface using Marching Cubes [20] and use this mesh in our method. We utilize Blender's "Smart UV Project" operator [6] to automatically generate the texture map for the mesh extracted from the SDF.

3.3 Path Tracing for Initial Light and Material Estimation

As an initial step, we leverage the state-of-art differentiable path tracer Mitsuba 2 [28] to optimize material properties and scene illumination. We choose a blended BSDF type, where a rough conductor BSDF with roughness α and a diffuse BSDF with a 512×512 texture A is combined using a convex combination with weight w. Illumination is represented as a 32×16 environment map $\hat{L}_\mathrm{e}$. Jointly optimizing light and material properties is difficult due to the ambiguities in the image formation process. In order to make our optimization stable, we assume the object to have a specular material that does not vary spatially. However, we use a spatially-varying diffuse texture A for capturing details. While these assumptions are often incorrect for many complex scenes, we show that our neural radiance transfer function is capable of reconstructions beyond these assumptions. Using the reconstructed geometry, it is straightforward to obtain foreground masks for each input view I_i, and we define the set of all foreground pixels as $\mathcal{M}_i$.

We jointly optimize $w, \alpha, A, \hat{L}_\mathrm{e}$ from the input multi-view images and the precomputed geometry, using the following loss term:

$$\mathcal{L}(w, \alpha, A, \hat{L}_\mathrm{e}) = \mathcal{L}_\mathrm{PT}(w, \alpha, A, \hat{L}_\mathrm{e}) + \lambda_\mathrm{reg}\mathcal{L}_\mathrm{reg}(A, \hat{L}_\mathrm{e})\,. \tag{5}$$

It consists of a data term and a regularizer that is weighted by λ_reg. The data term is defined as

$$\mathcal{L}_\mathrm{PT}(w, \alpha, A, \hat{L}_\mathrm{e}) = \sum_{i=1}^{m} \sum_{\mathbf{u} \in \mathcal{M}_i} ||\hat{I}_i(\mathbf{u}) - I_i(\mathbf{u})||^2\,. \tag{6}$$

Here, $\hat{I}_i$ is the path-traced reconstruction using up to five light bounces from the estimated scene parameters rendered from the ith viewpoint, $\mathbf{u}$ denotes 2D pixel coordinates, and $||\cdot||$ is the Euclidean norm. We use two regularizers to better constrain the problem:

$$\mathcal{L}_{\text{reg}}(A, \hat{L}_{\text{e}}) = \sum_{\hat{\boldsymbol{\omega}}_i} |\nabla \hat{L}_{\text{e}}(\hat{\boldsymbol{\omega}}_i)| + \lambda_{\text{BDSF}} \sum_{\mathbf{u}' \in \mathcal{M}_{\text{tex}}} |\nabla A(\mathbf{u}')| \tag{7}$$

where $\nabla(\cdot)$ denotes the image gradient, λ_{BDSF} is a weighting factor, and $|\cdot|$ denotes the L1 norm. $\mathbf{u}'$ are 2D uv-coordinates in the texture map and $\mathcal{M}_{\text{tex}}$ is the set of texels that is covered by the unwrapped geometry. The first term, $\mathcal{L}_{\text{Env}}$, is a regularizer on the environment map reconstruction, while the second term, $\mathcal{L}_{\text{BSDF}}$, regularizes the texture reconstruction. Both encourage image gradient sparsity. We refer to the supplemental document for more details.

3.4 Training the Neural Radiance Transfer Field

OLAT Synthesis. Our goal is to train the neural transfer field for the input scene. If we train the neural network only with the input illumination condition, the network can easily overfit, thus, not being able to disentangle illumination and material. Traditionally, learning-based methods, which disentangle material and illumination properties, rely on light-stage capture setups [7, 39, 53, 29]. In contrast to these approaches, we only rely on a single illumination condition. We show that it is possible to train for disentanglement even in this more challenging setup, by simulating a virtual light stage. Using the reconstruction obtained with the differentiable path tracer, we render synthetic images of the scene under novel one-light-at-a-time (OLAT) illumination conditions. Here, only one pixel on the environment map is active at a time. We sample OLAT images for training and novel camera views from every incoming light direction. We use Blender [6] to render the OLAT images with the reconstructed geometry and material as input. In total, we synthesize $N_{\text{c}} * N_{\text{e}}$ OLAT images as extra supervision, where N_{c} is the number of sampled camera views and N_{e} is the number of texels in the environment map. Note that the OLAT representation forms a complete basis for illumination conditions, i.e., any environment map can be computed as a linear combination of OLAT environment maps. Using these OLATs for our network supervision enables generalization to unseen illumination conditions and camera views.

NRTF Training. We train our NRTF in two stages. First, we train on the OLAT dataset using the following loss:

$$\mathcal{L}_{\text{OLAT}}(\theta) = \sum_{i=1}^{N_{\text{c}}} \sum_{\mathbf{u} \in \mathcal{M}_{\text{i}}} \left|\left| \frac{L_{\theta,i}(\mathbf{u}) - O_i(\mathbf{u})}{\text{sg}\left(L_{\theta,i}(\mathbf{u})\right) + \epsilon} \right|\right|^2, \tag{8}$$

Here, O_i is the ith OLAT image from Blender and $L_{\theta,i}$ is the corresponding estimate from our NRTF using Eq. 4. Stop gradient is denoted by $\text{sg}(\cdot)$. We set

$\epsilon = 1e-3$ to avoid division by zero and optimize for the network parameters θ. As shown in Noise2Noise [17], this loss works better for high-dymanic range images in the presence of path-tracing noise. Training on the OLAT images enables relighting and novel-view synthesis using the learned transfer function.

However, the method so far is heavily influenced by the lighting-reflectance ambiguity, and by the assumption of a global specularity parameter. Thus, in a second step, we further finetune the network on the input multi-view images. Here, we sample images from the real images as well as the synthetic OLAT images in a minibatch for training. The loss for this stage is defined as

$$\mathcal{L}(\theta, \tilde{L}_\mathrm{e}) = \mathcal{L}_\mathrm{OLAT}(\theta, \tilde{L}_\mathrm{e}) + \lambda_\mathrm{PRT}\mathcal{L}_\mathrm{PRT}(\theta, \tilde{L}_\mathrm{e}) + \lambda_\mathrm{EnvC}\mathcal{L}_\mathrm{EnvC}(\tilde{L}_\mathrm{e})\,. \tag{9}$$

$\mathcal{L}_\mathrm{OLAT}$ is used for the OLAT images in the batch. It is defined as in Eq. 8, however, we also finetune the environment map $\tilde{L}_\mathrm{e}$ in this stage using the previously obtained environment map $\hat{L}_\mathrm{e}$ for initialization. We further use a masked L2 loss for real images as:

$$\mathcal{L}_\mathrm{PRT}(\theta, \tilde{L}_\mathrm{e}) = \sum_{i=1}^{m} \sum_{\mathbf{u}\in\mathcal{M}_i} ||L_{\theta,i}(\mathbf{u}) - I_i(\mathbf{u})||^2\,, \tag{10}$$

Training on real images allows us to update the environment map. We add a regularizer, which penalizes the output to be too far from the initial environment map.

$$\mathcal{L}_\mathrm{EnvC}(\tilde{L}_\mathrm{e}) = \sum_{\hat{\boldsymbol{\omega}}_i} ||\tilde{L}_\mathrm{e}(\hat{\boldsymbol{\omega}}_i)) - \hat{L}_\mathrm{e}(\hat{\boldsymbol{\omega}}_i)||^2\,, \tag{11}$$

where $\hat{L}_e(\hat{\boldsymbol{\omega}}_i)$ denotes the initial environment map estimate.

4 Results

Next, we report results of the experiments we conducted to evaluate our method. We construct five synthetic scenes to showcase global illumination effects and further utilize four real scenes from the DTU dataset [10]. For each scene, we take 32-64 input views with a resolution of 800×600 pixels. During training, all our environment maps have a resolution of 32×16 pixels in latlong format, but this resolution can be different at test time due to our continuous neural-field formulation. On a single Quadro RTX 8000 GPU, training takes half an hour for initial light and material estimation, eight hours for OLAT training, and an additional 16 hours for the final joint optimization to reach highest-quality results. Factorizing lighting and reflectance is fundamentally ambiguous [15] and cannot be resolved from image observations in general [34, 19], especially when allowing for spatially-varying materials [33]. To aid meaningful comparisons nevertheless, we follow the procedure of Zhang et al. [54] and other works for all qualitative and quantitative results on synthetic scenes: We compute the mean RGB value of the ground truth environment map and normalize our estimated lighting by its inverse.

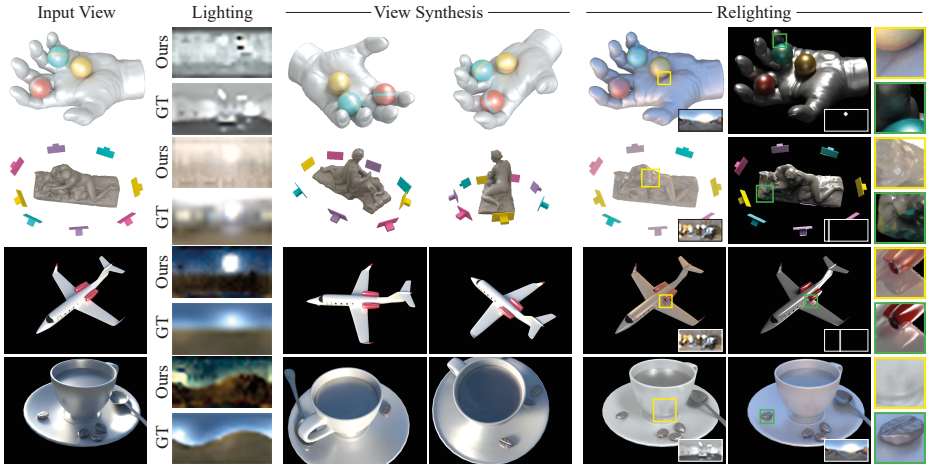

Fig. 3. Qualitative results on synthetic data. First column: Example input view for training. Second column: Our estimated environment map and the corresponding ground truth. Third and fourth column: Novel view synthesis results of our approach. Note that our method achieves sharp and accurate novel views that are almost indistinguishable from the input views in terms of quality. Last two columns: Relighting results of our method using novel environment maps (see insets). Also for novel lighting conditions our approach achieves convincing results with sharp specular reflections and secondary light bounce effects, e.g. indirect reflections on the wing of the airplane.

4.1 Qualitative Results

In Fig. 3, we demonstrate qualitative results of our method. Next to a representative input view (first column), we show the estimated and ground truth lighting (second column), followed by two exemplary novel views created with our method (third and fourth column). Finally, we show relighting results (remaining two columns) using different environment maps (insets). We see that our method produces high-quality relightable novel views, while successfully incorporating global-illumination effects like higher-order specular reflections and subtle color bleeding (see also Fig. 1). In our supplemental video, we further show that our method is also temporally stable when smoothly changing the camera view or rotating an environment map.

Real Data. In Fig. 4, we show results of our method on real scenes of the DTU dataset [10]. We can successfully synthesize high-quality novel views and plausible relighting. This shows that our method is robust to such real world captures, which are very challenging due to the lack of very precise camera calibration and foreground segmentation, camera noise, and other effects that are typically not present in synthetic datasets.

Beyond the Model Assumptions. In Fig. 5, we show that our approach can learn spatially-varying material effects beyond the ones that can be explained by the

Fig. 4. Qualitative results on real data [10]. First column: Example input views used for training our method. Second and third column: Novel view results. Note that even for real data our method achieves realistic novel view renderings. Last three columns: Relighting results using the environment maps depicted in the insets. Also here, note that our method can achieve plausible relighting effects.

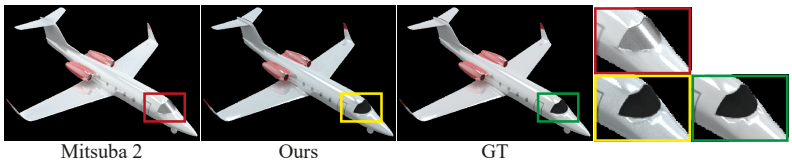

Fig. 5. Supervision with real input images lets our NRTF (center) learn appearance effects beyond the material model used during initialization with Mitsuba 2 (left), producing images close to the ground truth (right). Notice that our approach captures spatially-varying material roughness, see close-ups. Images show a relit novel view.

light and material models of the differentiable path tracer [28]. This is due to the real image loss, which lets the network learn appearance effects from real observations.

4.2 Comparisons

We compare our approach against several alternatives on the task of novel-view synthesis with relighting: On the one hand, we analyze the capabilities of stand-alone differentiable path-tracing using Mitsuba2 [28], which can be used to perform inverse rendering in the presence of global illumination. On the other hand, we consider three recent neural field-based inverse-rendering approaches PhySG [51], Neural-PIL [4], and NeRFactor [54], all employing only a direct illumination model. We omit a comparison to NeRD [3] as Neural-PIL can be considered as the follow-up. We also compare with RNR [5], providing it with the same geometry as ours. Further, we provide more results on NeRFactor [54] dataset in the supplemental document.

Table 1. Numerical comparisons for novel-view synthesis and relighting. We compare to the recent state of the art Mitsuba2 [28], PhySG [51], Neural-PIL [4], NeRFactor[54] and RNR [5] in terms of image-based metrics, i.e. PSNR and SSIM, and perceptual metrics, i.e. LPIPS. For both tasks, novel view synthesis and relighting, we achieve the best performance.

	Novel View Synthesis			Novel View Synthesis & Relighting		
Method	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**	**PSNR ↑**	**SSIM ↑**	**LPIPS ↓**
Mitsuba2 [28]	23.50	0.7567	0.0763	21.69	0.5722	0.0812
PhySG [51]	20.52	0.8563	0.2577	17.30	0.6252	0.2736
Neural-PIL [4]	17.07	0.5563	0.1159	14.76	0.4895	0.1328
NeRFactor [54]	21.97	0.6394	0.1691	15.83	0.6470	0.2033
RNR [5]	22.54	0.8122	0.0960	18.06	0.7009	0.1081
Ours	**28.73**	**0.9151**	**0.0454**	**23.06**	**0.8247**	**0.0692**

A qualitative comparison is shown in Fig. 6. Despite the fact that our method provides the sharpest and most realistic results, it is worth noting that our method is the only one that can recover accurate indirect lighting effects, e.g. the self-reflection on the wing of the airplane and the color spill of the squares onto the statue. This is further confirmed by the quantitative analysis in Table 4.2. We compute image errors for four scenes, each with five views and three lighting conditions according to three metrics on the tasks of novel-view synthesis and novel-view synthesis with relighting. In particular, we evaluate the Peak Signal-to-noise Ratio (PSNR), the Structural Similarity Index Measure (SSIM) [45], and the learned perceptual image patch similarity (LPIPS) [52]. PSNR and SSIM are purely image-based metrics and, thus, sometimes do not reflect the *perceived* image quality. For this reason, we also provide the perceptual LPIPS metric. We observe that our approach again delivers the highest-quality results for both tasks across all metrics.

4.3 Ablation & Extension

Here, we study ablations and extensions in order to gain further insights into our system. All results are compiled into Table 4.3, where the evaluation protocol is the same as in Sec. 4.2. First, we consider omitting the OLAT loss (Sec. 3.4). We observe that result quality reduces significantly for the relighting task compared to our full method. This is due to the poor disentanglement of lighting and reflectance and the fact that the network can overfit to the lighting condition in the training data, which also explains why the novel view synthesis without the OLAT loss is slightly more accurate than our method.

Second, we investigate the behavior of our approach when input views are captured under *multiple* unknown illumination conditions. In this experiment, we use three different environment maps. When reconstructing geometry (Sec. 3.2), we select only a subset of multi-view images with the same illumination condition, while during initial light and material estimation (Sec. 3.3), we optimize

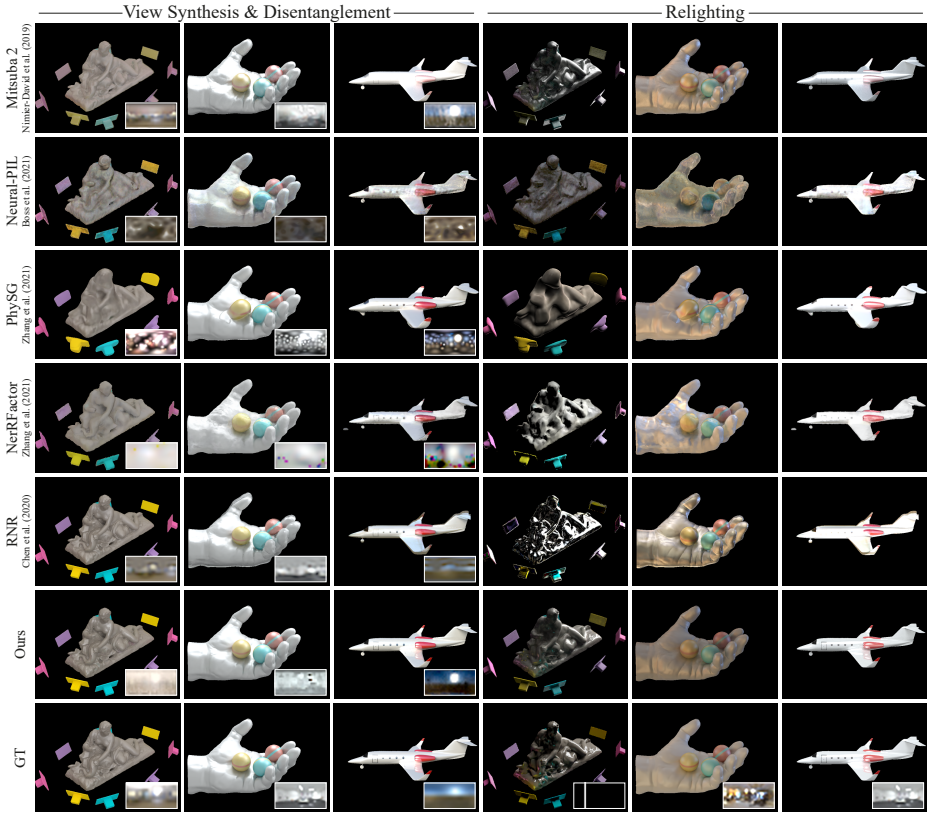

Fig. 6. Comparisons to related works [28, 51, 4, 5, 54] for novel-view synthesis and disentanglement of lighting (left three columns), and relighting (right three columns). Note that for both tasks we achieve the best results in terms of rendering quality. It is also worth noting that we are the only method, which can accurately reproduce the indirect illumination effects such as the self-reflection on the wing of the airplane.

for three individual environment maps. Not surprisingly, we observe that this extended setup increases result quality even more compared to our full single-lighting approach. Yet, it has a significantly less pronounced effect compared to the omission of the OLAT training stage, indicating that our pipeline achieves a solid disentanglement for the single-illumination condition.

5 Limitations and Future Work

Although our method improves the state of the art in terms of image quality and global illumination handling, it still has some limitations, which open up future work in this direction. In particular, our method relies on an accurate geometry estimate of the scene and we are not jointly optimizing the scene geometry along with the material and lighting of the scene. Future work could involve a

Table 2. Ablations and extensions. First, we evaluate the effect of the proposed synthetic OLAT loss. One can clearly see that without the OLAT loss the performance of our method drastically drops for the relighting task. This can be explained by the fact that the OLAT loss acts as a regularizer and prevents overfitting to the single environment map that is recovered during training. Moreover, we evaluate how our method performs when the object was observed under *multiple* lighting conditions. Interestingly, with this additional input, our method can achieve even better results, especially for the relighting task.

	Novel View Synthesis			Novel View Synthesis & Relighting		
Method	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓	**PSNR** ↑	**SSIM** ↑	**LPIPS** ↓
w/o OLAT Loss	31.35	0.9668	0.063	10.03	0.4547	0.4487
Full	29.56	0.9418	0.066	24.76	0.8665	0.069
Multiple Envmaps	30.62	0.9166	0.043	26.67	0.9071	0.047

joint reasoning of all these aspects in a differentiable manner such that optimizing all scene aspects jointly can be achieved. Further, our relighting results are only correct up to a global scale due to the inherent ambiguity between scene illumination and the object material. Here, future research could explore a minimal setup required to disentangle such ambiguities, e.g. it may be that a single measurement on the surface can resolve the ambiguity. Last, our method takes several seconds per frame. Ideally, it would run at real time enabling interactive scene relighting with global illumination. Thus, exploring more efficient scene representations could be an interesting research branch for the future.

6 Conclusion

We presented neural radiance transfer fields, which enable global illumination scene relighting and view synthesis given multi-view images of the object. At the technical core, our method implements the concept of precomputed radiance transfer that disentangles illumination from appearance. To this end, we propose a neural radiance transfer field represented as an MLP and show how at train time differentiable path tracing and a dedicated OLAT loss can be used to let the network accurately learn such a disentanglement. Once trained, our rendering formulation allows novel-view synthesis and relighting, which is aware of global-illumination effects. Our results demonstrate a clear improvement over the current state of the art while future work could involve further improving the runtime and a joint reasoning of geometry, material, and scene lighting.

Acknowledgements. We would like to thank Xiuming Zhang for his help with the NeRFactor comparisons. Authors from MPII were supported by the ERC Consolidator Grant 4DRepLy (770784).

References

1. Azinović, D., Li, T.M., Kaplanyan, A., Nießner, M.: Inverse path tracing for joint material and lighting estimation. In: Proc. Computer Vision and Pattern Recognition (CVPR), IEEE (2019)
2. Bangaru, S., Michel, J., Mu, K., Bernstein, G., Li, T.M., Ragan-Kelley, J.: Systematically differentiating parametric discontinuities. ACM Trans. Graph. **40**(107), 107:1–107:17 (2021)
3. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.P.: Nerd: Neural reflectance decomposition from image collections. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 12684–12694 (October 2021)
4. Boss, M., Jampani, V., Braun, R., Liu, C., Barron, J., Lensch, H.: Neural-pil: Neural pre-integrated lighting for reflectance decomposition. Advances in Neural Information Processing Systems **34** (2021)
5. Chen, Z., Chen, A., Zhang, G., Wang, C., Ji, Y., Kutulakos, K.N., Yu, J.: A neural rendering framework for free-viewpoint relighting. In: CVPR (2020)
6. Community, B.O.: Blender - a 3d modelling and rendering package (2018), http://www.blender.org
7. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 145–156 (2000)
8. Goel, P., Cohen, L., Guesman, J., Thamizharasan, V., Tompkin, J., Ritchie, D.: Shape from tracing: Towards reconstructing 3d object geometry and svbrdf material from images via differentiable path tracing. In: 2020 International Conference on 3D Vision (3DV). pp. 1186–1195. IEEE (2020)
9. Hao, X., Baby, T., Varshney, A.: Interactive subsurface scattering for translucent meshes. In: Proceedings of the 2003 symposium on Interactive 3D graphics. pp. 75–82 (2003)
10. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413. IEEE (2014)
11. Kajiya, J.T.: The rendering equation. In: Proceedings of the 13th annual conference on Computer graphics and interactive techniques. pp. 143–150 (1986)
12. Kautz, J., Sloan, P.P., Lehtinen, J.: Precomputed radiance transfer: theory and practice. In: ACM SIGGRAPH 2005 Courses, pp. 1–es (2005)
13. Kautz, J., Snyder, J., Sloan, P.P.J.: Fast arbitrary brdf shading for low-frequency lighting using spherical harmonics. Rendering Techniques **2**(291-296), 1 (2002)
14. Laffont, P.Y., Bousseau, A., Drettakis, G.: Rich intrinsic image decomposition of outdoor scenes from multiple views. IEEE transactions on visualization and computer graphics **19**(2), 210–224 (2012)
15. Land, E.H., McCann, J.J.: Lightness and retinex theory. Josa **61**(1), 1–11 (1971)
16. Lehtinen, J.: A framework for precomputed and captured light transport. ACM Transactions on Graphics (TOG) **26**(4), 13–es (2007)
17. Lehtinen, J., Munkberg, J., Hasselgren, J., Laine, S., Karras, T., Aittala, M., Aila, T.: Noise2noise: Learning image restoration without clean data. arXiv preprint arXiv:1803.04189 (2018)
18. Li, T.M., Aittala, M., Durand, F., Lehtinen, J.: Differentiable monte carlo ray tracing through edge sampling. ACM Transactions on Graphics (TOG) **37**(6), 1–11 (2018)

19. Lombardi, S., Nishino, K.: Reflectance and illumination recovery in the wild. IEEE transactions on pattern analysis and machine intelligence **38**(1), 129–141 (2015)
20. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
21. Lyu, L., Habermann, M., Liu, L., Tewari, A., Theobalt, C., et al.: Efficient and differentiable shadow computation for inverse problems. In: ICCV. pp. 13107–13116 (2021)
22. Marschner, S.R.: Inverse rendering for computer graphics. Cornell University (1998)
23. 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: The European Conference on Computer Vision (ECCV) (2020)
24. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. arXiv:2201.05989 (Jan 2022)
25. Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S.: Extracting triangular 3d models, materials, and lighting from images. arXiv preprint arXiv:2111.12503 (2021)
26. Ng, R., Ramamoorthi, R., Hanrahan, P.: All-frequency shadows using non-linear wavelet lighting approximation. In: ACM SIGGRAPH 2003 Papers, pp. 376–381 (2003)
27. Nimier-David, M., Dong, Z., Jakob, W., Kaplanyan, A.: Material and lighting reconstruction for complex indoor scenes with texture-space differentiable rendering (2021)
28. Nimier-David, M., Vicini, D., Zeltner, T., Jakob, W.: Mitsuba 2: A retargetable forward and inverse renderer. ACM Transactions on Graphics (TOG) **38**(6), 1–17 (2019)
29. Pandey, R., Orts-Escolano, S., LeGendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., Fanello, S.: Total relighting: Learning to relight portraits for background replacement. vol. 40 (August 2021). https://doi.org/10.1145/3450626.3459872
30. Philip, J., Gharbi, M., Zhou, T., Efros, A.A., Drettakis, G.: Multi-view relighting using a geometry-aware network. ACM Trans. Graph. **38**(4), 78–1 (2019)
31. Philip, J., Morgenthaler, S., Gharbi, M., Drettakis, G.: Free-viewpoint indoor neural relighting from multi-view stereo. ACM Transactions on Graphics (TOG) **40**(5), 1–18 (2021)
32. Rainer, G., Bousseau, A., Ritschel, T., Drettakis, G.: Neural precomputed radiance transfer. Computer Graphics Forum (Proc. Eurographics) **41**(2) (2022), http://www-sop.inria.fr/reves/Basilic/2022/RBRD22
33. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 117–128 (2001)
34. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for reflection. ACM Transactions on Graphics (TOG) **23**(4), 1004–1042 (2004)
35. Ritschel, T., Dachsbacher, C., Grosch, T., Kautz, J.: The state of the art in interactive global illumination. In: Computer graphics forum. vol. 31, pp. 160–188. Wiley Online Library (2012)
36. Sloan, P.P., Kautz, J., Snyder, J.: Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In: Proc. SIGGRAPH. pp. 527–536 (2002)
37. Sloan, P.P., Luna, B., Snyder, J.: Local, deformable precomputed radiance transfer. ACM Transactions on Graphics (TOG) **24**(3), 1216–1224 (2005)

38. 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. pp. 7495–7504 (2021)
39. Sun, T., Xu, Z., Zhang, X., Fanello, S., Rhemann, C., Debevec, P., Tsai, Y.T., Barron, J.T., Ramamoorthi, R.: Light stage super-resolution: continuous high-frequency relighting. ACM Transactions on Graphics (TOG) **39**(6), 1–12 (2020)
40. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Wang, Y., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., et al.: Advances in neural rendering. arXiv preprint arXiv:2111.05849 (2021)
41. Thul, D., Tsiminaki, V., Ladickỳ, L., Pollefeys, M.: Precomputed radiance transfer for reflectance and lighting estimation. In: 2020 International Conference on 3D Vision (3DV). pp. 1147–1156. IEEE (2020)
42. Tsai, Y.T., Shih, Z.C.: All-frequency precomputed radiance transfer using spherical radial basis functions and clustered tensor approximation. ACM Transactions on graphics (TOG) **25**(3), 967–976 (2006)
43. Wang, J., Ramamoorthi, R.: Analytic spherical harmonic coefficients for polygonal area lights. ACM Transactions on Graphics (Proceedings of SIGGRAPH 2018) **37**(4) (2018)
44. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
45. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004). https://doi.org/10.1109/TIP.2003.819861
46. Wu, L., Cai, G., Zhao, S., Ramamoorthi, R.: Analytic spherical harmonic gradients for real-time rendering with many polygonal area lights. ACM Transactions on Graphics (TOG) **39**(4), 134–1 (2020)
47. Wu, S., Rupprecht, C., Vedaldi, A.: Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1–10 (2020)
48. Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., Sridhar, S.: Neural fields in visual computing and beyond. arXiv preprint arXiv:2111.11426 (2021)
49. Xu, K., Sun, W.L., Dong, Z., Zhao, D.Y., Wu, R.D., Hu, S.M.: Anisotropic spherical gaussians. ACM Transactions on Graphics (TOG) **32**(6), 1–11 (2013)
50. Yu, A., Fridovich-Keil, S., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. arXiv preprint arXiv:2112.05131 (2021)
51. Zhang, K., Luan, F., Wang, Q., Bala, K., Snavely, N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In: CVPR (2021)
52. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 586–595. IEEE Computer Society, Los Alamitos, CA, USA (jun 2018). https://doi.org/10.1109/CVPR.2018.00068, https://doi.ieeecomputersociety.org/10.1109/CVPR.2018.00068
53. Zhang, X., Fanello, S., Tsai, Y.T., Sun, T., Xue, T., Pandey, R., Orts-Escolano, S., Davidson, P., Rhemann, C., Debevec, P., et al.: Neural light transport for relighting and view synthesis. ACM Transactions on Graphics (TOG) **40**(1), 1–17 (2021)

54. 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. ACM Transactions on Graphics (TOG) **40**(6), 1–18 (2021)