Neural Relighting with Subsurface Scattering by Learning the Radiance Transfer Gradient

SHIZHAN ZHU, University of California, Berkeley, USA
SHUNSUKE SAITO, Meta Reality Labs Research, USA
ALJAŽ BOŽIČ, Meta Reality Labs Research, USA
CARLOS ALIAGA, Meta Reality Labs Research, USA
TREVOR DARRELL, University of California, Berkeley, USA
CHRISTOPH LASSNER, Meta Reality Labs Research, USA

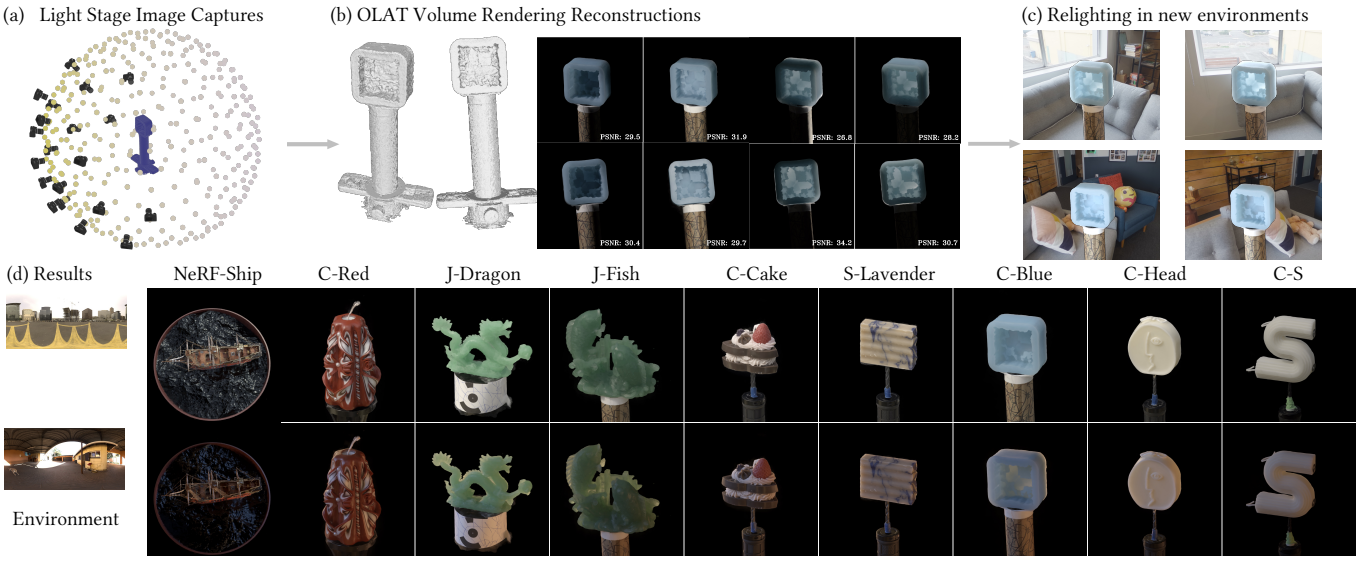


Fig. 1. Our approach reconstructs objects with significant subsurface scattering effects with high fidelity and inserts models into arbitrary environments for relighting. It is fully data-driven and does not assume a particular material representation (such as BRDF or BSSRDF), and can faithfully render high quality appearance under varying lighting conditions and view points. Please see our supplementary video for comprehensive visualizations and comparisons.

Reconstructing and relighting objects and scenes under varying lighting conditions is challenging: existing neural rendering methods often cannot handle the complex interactions between materials and light. Incorporating pre-computed radiance transfer techniques enables global illumination, but still struggles with materials with subsurface scattering effects. We propose a novel framework for learning the radiance transfer field via volume rendering and utilizing various appearance cues to refine geometry end-to-end. This framework extends relighting and reconstruction capabilities to handle a wider range of materials in a data-driven fashion. The resulting models produce plausible rendering results in existing and novel conditions. We will release our code and a novel light stage dataset of objects with subsurface scattering effects publicly available.

1 INTRODUCTION

The ability to relight objects and scenes under varying lighting conditions is crucial in many areas, such as virtual reality, gaming, visual effects, and architecture. It enables artists, designers, and engineers to experiment with many lighting setups without having to physically recreate a scene. It also allows for the creation of more realistic and immersive experiences by accurately simulating the lighting conditions in a virtual environment.

However, relighting remains a challenging task due to the complex interaction between the light and the materials in a scene. Traditional approaches have sought to decompose rendering into geometry, material, and lighting to simplify the problem. For example, opaque materials are represented in Physically Based Rendering (PBR) by the Bidirectional Reflectance Distribution Function (BRDF), which describes how light interacts with a material's surface [Zhang et al. 2021a; Munkberg et al. 2022; Zhang et al. 2021b; Boss et al. 2021b,a]. Similarly, many relighting methods such as [Srinivasan et al. 2021; Chen et al. 2020] rely on decomposing light into its components, such as direct lighting and indirect lighting, to allow for more fine-grained control.

While these approaches have been successful in many cases, they are limited in their ability to handle objects with translucency or subsurface scattering (SSS). This is because these materials are not well approximated by a simple BRDF function and require more complex models, such as the Bidirectional Surface Scattering Reflectance Distribution Function (BSSRDF). However, modeling BSSRDF are

Authors' addresses: Shizhan Zhu, University of California, Berkeley, USA; Shunsuke Saito, Meta Reality Labs Research, USA; Aljaž Božič, Meta Reality Labs Research, USA; Carlos Aliaga, Meta Reality Labs Research, USA; Trevor Darrell, University of California, Berkeley, USA; Christoph Lassner, Meta Reality Labs Research, USA.

computationally expensive and slow to evaluate, neglecting textures beneath the surface (Fig. 3), limiting their practicality for inverse rendering with complex geometry.

Recent works on neural radiance transfer fields [Lyu et al. 2022] have incorporated the idea of pre-computed radiance transfer (PRT) into the neural radiance fields (NeRF) literature, providing promising results for relighting with global illumination effects. However, these approaches rely on a pre-estimated surface, which is nontrivial to reconstruct for objects with SSS or with translucency. Additionally, the separated geometry and appearance optimization is suboptimal, leading to artifacts and unrealistic results.

In this work, we propose a novel framework for relighting that incorporates the optimization of shape and radiance transfer using a volume rendering approach (Fig. 1). Our framework extends the relighting capability to a wider range of materials, including translucent objects with strong SSS effects and textures beneath the surface. Specifically, we use a volume rendering approach to estimate the transfer field and utilize appearance cues to refine the geometry in an end-to-end fashion.

To evaluate our approach, we have recorded real-world objects featuring subsurface scattering effects in a light stage and show that our method produces high quality visual results in recorded and novel lighting conditions. Quantitatively, our approach compares favorably with the current state of the art with a 5 points higher PSNR on average across three datasets.

In summary, we propose a novel framework for neural radiance transfer fields using volume rendering, optimizing appearance and geometry in an end-to-end fashion, which to the best of our knowledge has not been achieved before for optically-thick translucent materials. Additionally, we collected and will release a dataset of objects that exhibit prominent subsurface scattering effects for training and evaluation purposes. These objects have been recorded with high fidelity featuring rich, high frequency spatially-varying details, resulting in 15TiB of data, which is 3000 times larger and notably more detailed than the current highest quality data for research in this area [Deng et al. 2022].

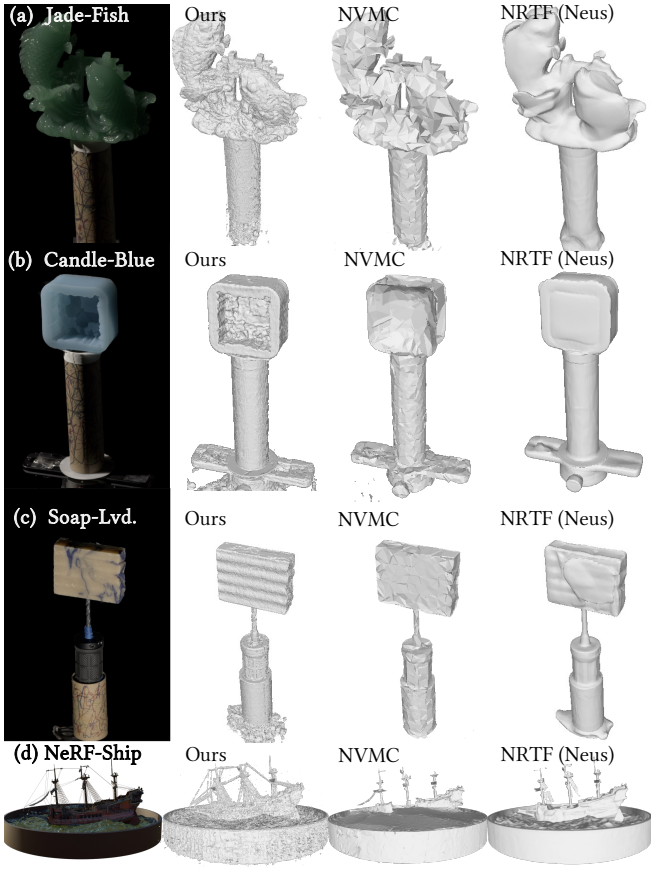


Fig. 2. Despite presented with significant subsurface scattering and translucency in the scene, our approach provides the highest geometric reconstruction quality compared to other approaches (NVMC [Hasselgren et al. 2022]; NRTF [Lyu et al. 2022] via Neus [Wang et al. 2021]). For our approach, we show the extracted mesh using marching cubes from the density in the 512^3 resolution. The high quality geometry is one of the key advantages of our method.

2 RELATED WORK

Relighting and Surface Representations. The problem of relighting an object or a scene under novel lighting conditions has been extensively studied. Usually, the problem is tackled via decomposing the appearance into the lighting and the surface material properties. Early works estimate material given known illumination such as a single light source [Yu et al. 1999; Debevec et al. 2000] or spherical gradient illumination [Fyffe 2009; Guo et al. 2019] with known geometry. [Zhang et al. 2021b] directly model light transports with known illuminations and know geometry. More recently, neural scene representations [Xie et al. 2022] and differentiable rendering [Nimier-David et al. 2019] allow us to jointly optimize BRDF and geometry. Some methods apply inverse rendering using implicit surface to obtain materials [Luan et al. 2021; Zhang et al. 2022; Munkberg et al. 2022]. Other approaches utilize volumetric representations with opacity fields [Bi et al. 2020b,a; Zhang et al. 2021a,b; Boss et al. 2021a]. The required illumination setup can be reduced to a co-located light [Bi et al. 2020b,a], and unknown illuminations [Luan et al. 2021; Zhang et al. 2022, 2021a,b; Boss et al. 2021a]. To reduce the ambiguity in BRDF, the aforementioned methods use parametric BRDFs such as a microfacet model [Walter et al. 2007; Burley and Studios 2012]. However, these parametric models do not support subsurface scattering as they only consider reflectance. In contrast, our approach deals with global light transport effects including subsurface scattering.

Subsurface Scattering. Subsurface scattering refers to light transport inside of a solid substance. It happens with some particular types of materials (such as wax, jade, tiny furs or various fruits), and is quite common in the real world. Since the light might leave the object surface at a different point from where it enters, surface representations (e.g. various BRDFs) cannot represent this type of light transmission. While subsurface scattering can be accurately modeled by volumetric path tracing algorithms [Novák et al. 2018], their run time is typically prohibitive in certain applications, despite efforts to accelerate brute-force computation, e.g. through a

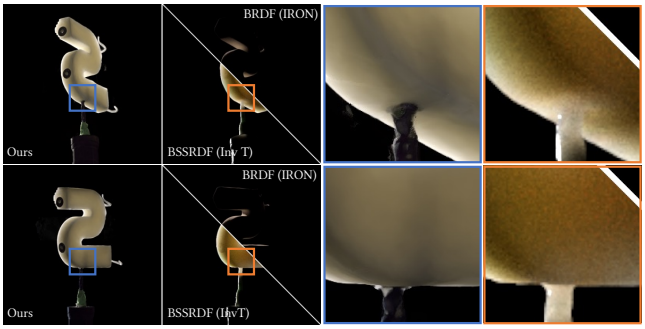


Fig. 3. For relighting objects with subsurface scattering effects (e.g., the translucent soap shown in this figure), the BRDF-based approach [Zhang et al. 2022] renders the object with full opacity when the light comes from the opposite directions, while the BSSRDF-based approach [Deng et al. 2022] cannot capture the texture details and structures beneath the surface (highlighted in the orange squares). In contrast, our approach can faithfully render the right opacity of the object and retain appearance even given the subsurface structure of the drill inside the candle (highlighted in the blue squares).

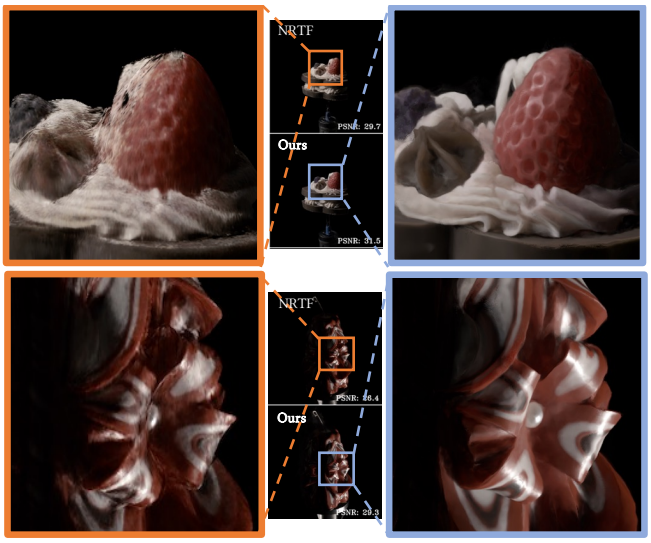


Fig. 4. Volume rendering leads to cleaner surface reconstructions and higher rendering quality compared to NRTF [Lyu et al. 2022].

shape adaptive learned SSS model [Vicini et al. 2019] that relies on a conditional variational auto encoder that learns to sample from a distribution of exit points on the object surface. Some other works have focused on estimating the scattering parameters from images of translucent objects. Inverse Transport Networks [Che et al. 2020] infer the optical properties that control subsurface scattering inside translucent objects of any shape under any illumination. They rely on an encoder decoder where the latter is replaced by a physically-based differentiable path tracer, trained with synthetic images. Prior to that, another approach based on stochastic gradient descent, combined with Monte Carlo rendering and a material dictionary was capable of estimating the scattering materials, inverting the radiative transfer parameters [Gkioulekas et al. 2013]. Nevertheless, since volumetric path tracing can be costly, applying a BSSRDF can be a faster alternative [Deng et al. 2022]. Compared to BRDF-based representations, a higher dimension of inputs (usually 6D for homogeneous materials) is fed to query the outgoing radiance. A relighting algorithm can thus seek to optimize the BSSRDF function with the inverse rendering process so that the resulting material can be relit in conventional rendering engines. Our work follows a different path - we learn our relighting model in a fully data driven fashion, and learn the cached outgoing radiance for each point using a deep neural network, where we bypass the expensive BSSRDF computation in our optimization iteration.

Neural Radiance Fields and Precomputed Radiance Transfer. Neural Radiance Fields (NeRF) [Mildenhall et al. 2020; Barron et al. 2021, 2022] optimize a parameterized volume rendering model from multiple views of the scene so that at test time, novel views can be synthesized from the learned model. Despite its superior rendering quality, NeRF bakes all the lighting and reflective surface information into the RGBs without modeling the interaction of the light and the material. Recent studies [Lyu et al. 2022] have shown promising results for relightable models via incorporating the idea of "precomputed radiance transfer" (PRT) [Sloan et al. 2002] from the real time rendering community. Instead of precomputing and caching the intermediate representation per location, they seek to optimize a cached intermediate representation in the reconstruction process. Notably, [Lyu et al. 2022] relies on a fairly accurate pre-computed surface [Wang et al. 2021], and keeps the lighting appearance optimization separate from the geometry acquisition process. Focused on synthetic images with varying but known illumination, a NeRF extension [Zheng et al. 2021] was presented to reconstruct participating media with full global illumination effects, achieving good results on synthetic data. In contrast, our novel volume rendering framework not only enables optimizing the geometry details with appearance cues, but also works on scenes with partially opaque mass (e.g. thin rope or furs) and demonstrates high quality results on synthetic and real data. It is worth mentioning that a recent concurrent work [Yu et al. 2023] also addressed relighting with translucent objects using scattering functions. In addition to distant point lights, our approach efficiently relights the captured scenes with environment maps with the help of the Median-cut algorithm. Further, we will release high-resolution and large scale light stage dataset with rich lighting effects, such as translucency coupled with specular highlights and translucent shadowing, facilitating future research.

3 METHODOLOGY

3.1 Notation

Our goal is to optimize a relightable neural model from a collection of photos of the object, captured from different camera view points and under varying lighting conditions, that is able to accurately represent strong subsurface scattering effects. Our input includes the set of the input images $\mathcal{I} = \{\mathbf{I}_{c,l}\}$, where $\mathbf{I}_{c,l} \in (\mathbb{R}^+)^{M\times N\times 3}$ are high dynamic range (HDR) images, and c and l represent the camera viewpoint and lighting condition, respectively. We assume the camera poses are known (e.g., computed using photogrammetry software such as Agisoft Metashape), and denote them as $C = \{\mathbf{K}_c, \mathbf{R}_c, \mathbf{t}_c\}$ (camera intrinsic parameters, rotations and translations,

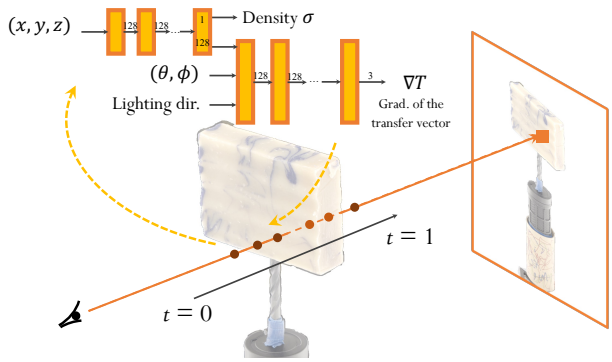


Fig. 5. Illustration of the proposed relighting framework. We devise two MLPs to predict the gradient of the transfer vector for accumulating the HDR value of each ray. See Sec. 3 for details.

respectively). We capture one-light-at-a-time (OLAT) images for training, and denote an OLAT lighting condition as $\mathcal{L} = \{\omega_l\}$, where $\omega_l \in \mathbb{R}^3$ is the ℓ_2 normalized vector representing the incident point light direction relative to the scene center. Since our data capture system uses white light, we parameterize the light using a single channel throughout this paper. During testing, we apply an environment map (*envmap*) $\mathbf{E}_l \in \mathbb{R}^{M_E \times N_E}$, where each pixel of the envmap can be considered as one light source.

We want our relightable model to render the scene under varying *unseen* viewpoints ($\{\mathbf{K}_{\text{query}}, \mathbf{R}_{\text{query}}, \mathbf{t}_{\text{query}}\}$) and lighting conditions (ω_{query} or $\mathbf{E}_{\text{query}}$). Our framework optimizes the geometry as well as the lighting- and viewpoint-varying appearance of the scene in an end-to-end fashion. More precisely, we use the function $f_\Theta(\cdot)$ to denote our model (parameterized by Θ), and denote our model prediction as $\hat{\mathbf{I}}(u, v; \omega \text{ or } \mathbf{E}) = f_\Theta(\mathbf{r}; \omega \text{ or } \mathbf{E}) \in (\mathbb{R}^+)^3$, where $\mathbf{r}$ represents a pixel ray in the space, and (u, v) represents its related pixel coordinates on the image plane under the given camera pose $\{\mathbf{K}, \mathbf{R}, \mathbf{t}\}$. We provide an overview of our approach in Sec. 3.2, and provide details of our volume rendering scheme as well as model details in Sec. 3.3 and Sec. 3.4.

3.2 Method Overview

We devise a volume rendering based neural relightable model that is optimized directly from the image collections of varying camera views and lighting conditions (Fig. 5). The core of our learning framework consists of a volume renderer enabling an end-to-end optimization (Sec. 3.3) and the density-based neural transfer field networks (Sec. 3.4). There are several key differences compared to the existing (neural) relighting approaches. On one hand, unlike [Lyu et al. 2022], our model can be trained from scratch, with no dependency on known estimated surface or other explicit geometry cues whose geometric details are difficult to obtain especially for materials with strong subsurface scattering effects (Fig. 2). Furthermore, training images captured under varying lighting conditions contain rich geometric cues via local micro shadowing or micro reflections, where a direct geometric optimization via an appearance loss is deemed necessary. On the other hand, thanks to our fully data-driven learning scheme, our model does not make any explicit assumptions about material (such as specifying a varying BRDF or BSSRDF) [Zhang et al. 2022; Deng et al. 2022], making it applicable to a wide range of materials, enabling global illuminations and subsurface scattering effects.

3.3 Volume Integration of the Transfer Gradient

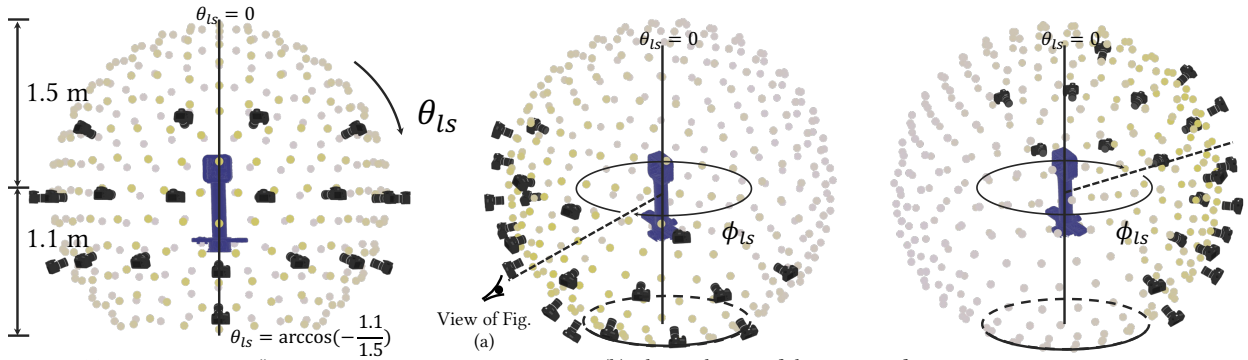
The color of each pixel ray is computed using volume rendering. We denote the points sampled along the ray $\mathbf{r}$ as $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, where N is the total number of points. The model predicts $\hat{\sigma}(\mathbf{x}_i) \in \mathbb{R}^+$ and $\hat{\mathbf{h}}(\mathbf{x}_i; \mathbf{r}; \omega) \in (\mathbb{R}^+)^3$ for every sample point, representing the density and the gradient of the pre-computed transfer vector, respectively. It is worth pointing out that instead of predicting the transfer vector directly as in [Lyu et al. 2022], we predict the transfer vector gradient prediction, which represents the HDR contribution of a particular segment along a particular light transmission direction. It is clear that no HDR delta would incur at a density-free location, and among the non-zero density locations, the HDR contribution at a segment can only be non-negative if a location is visible, i.e. when its volume accumulation weight ($\hat{w}(\mathbf{x})$ in Eq. 1) is positive. This intuition aligns well with our volume accumulation and rendering scheme. We follow the volume integration from [Mildenhall et al. 2020] and obtain the accumulated transfer vector prediction as:

$$\begin{aligned} \hat{\mathbf{I}}(u, v; \omega) &= \sum_{i=1}^{N} \hat{w}(\mathbf{x}_i) \hat{\mathbf{h}}(\mathbf{x}_i; \mathbf{r}; \omega) \\ \text{where } \hat{w}(\mathbf{x}_i) &= \hat{T}_i (1 - \exp(-\hat{\sigma}(\mathbf{x}_i)\delta_i)) \\ \hat{T}_i &= \exp(-\sum_{j=1}^{i-1} \hat{\sigma}(\mathbf{x}_j)\delta_j) \\ \delta_i &= t_{i+1} - t_i. \end{aligned} \tag{1}$$

Our volume rendering scheme demonstrates several key benefits over a surface representation [Lyu et al. 2022] (Fig. 4). First and foremost, obtaining a fairly accurate pre-estimated surface for materials featuring subsurface scattering with detailed geometry is non-trivial. Our model bypasses the difficulties of pre-estimating the surface geometry by applying volume rendering and optimizing the surface density together with appearance. In this case, all local shadowing and reflection effects captured under different lighting conditions are taken into account for geometry estimation, providing stronger cues compared to surface estimation under a single lighting condition. Second, volume rendering enables accurate appearance modeling of semi-opaque materials (e.g. fur) with their subsurface scattering effects, which cannot be trivially achieved using a surface-based rendering framework. Third, similar to other volume rendering-based models, our model is able to optimize the geometry as well as the relightable appearance end-to-end under varying lighting conditions. We do not require model design changes to back propagate the loss gradient back to the geometry prediction [Munkberg et al. 2022]. Our results show that this rendering strongly result in higher fidelity compared to previous surface-based rendering [Lyu et al. 2022].

3.4 End-to-end Learning of Neural Relighting

Architectures. We follow [Mildenhall et al. 2020] and use an MLP to predict the density as well as the transfer vector gradient for



(a) Frontal View of "Group 000" (b) Elevated view of the rotation from group 000 to group 240.

Fig. 6. Illustration of our light stage capture system. A full capture consists of 9 capture groups, with each group labelled as "000", "040", "080", ..., "320", with their number denoting the 40 degree-stepped yaw rotation (see "ϕ_{ls}" in (b)). Lights are visualized as dots and cameras with camera icons. All lights are of the same white color—the visualized dot colors merely refer to the light bulb instance, highlighting that the lights are locked with the cameras when rotating between groups. **(a)** Frontal view of the system (group "000"). The radius of the light stage is 1.5 meters, with its center at 1.1m height—the layout is a bottom-truncated sphere. The light stage illustration in Fig. 1(a) is the elevated view of group "040". **(b)** Rotating from group "000" (b-left) to group "240" (b-right) according to the "ϕ_{ls}" rotation. (b-left) and (b-right) are visualized at an elevated angle. (a) is viewed from the dashed line direction in (b-left).

each sample interval. The MLP consists of 8 fully-connected layers (with a width of 256, and a skip connection in the fourth layer) each for the density as well as the transfer vector gradient prediction respectively. We devise two MLPs tackling the coarse level and fine level of accumulation respectively. To ensure the predicted transfer vector gradient to be non-negative, we use the exponential function as the activation function, following [Mildenhall et al. 2022]. It is worth pointing out that MLPs are just one option for modeling the predictions of each point, we expect that more efficient models [Yu et al. 2021; Fridovich-Keil et al. 2022; Müller et al. 2022; Chen et al. 2022] can be used as well.

Loss functions. We utilize the weighted L2 tonemapped loss [Mildenhall et al. 2022] to supervise the predicted HDR values of each pixel. We also impose an auxiliary mask loss, where we pose L2 regularizers on the density of all points that are sampled on a background ray. Our main focus is the modeling of the foreground objects, and due to the inconsistency of the background appearance (rotation nature of the camera groups, see Sec. 4), we set all the ground truth HDR values in the background to be 0. We sample the rays (from both the foreground as well as the background) with importance sampling, where in each training batch, 1/2 of all rays are from the foreground, 3/8 of all rays are from the near-silhouette area, and 1/8 of total rays are from arbitrary locations in the background. In our real light stage data, since the aspect ratio of the captured images is pretty large, we extend the background area that is outside of the pixel map. We pad them with 0 as their ground truth HDR values. We found this to be useful for clean free-space density predictions.

Using environment map conditioning. We follow [Lyu et al. 2022] to obtain the envmap relighting prediction via accumulating the OLAT HDR prediction. More precisely, we treat each pixel on the envmap as an OLAT point light. Practically, we apply the median cut algorithm [Debevec 2008] to accelerate inference. During accumulation, we reweight the predicted HDR value from each OLAT location by the cosine value of the latitude angle on the envmap to account for the area of lights on the envmap sphere. The aggregated predicted rendering serves as our final prediction of the relighting given the query envmap.

4 LIGHT STAGE DATA ACQUISITION

To facilitate studies on the light-dependent appearance modeling of objects and scenes under significant subsurface scattering effects, it is critical to acquire real-world objects featuring such effects. While existing datasets (e.g., [Deng et al. 2022]) includes captures of two translucent objects, they are often limited by resolution and fidelity of the acquired images, causing local micro geometry details to not be fully captured. To reconstruct a relightable model in a data-driven fashion, we aim to have real-world captures with densely sampled camera viewpoints, complete incident light direction coverage and high-resolution images retaining as much detail as possible. Consequently, we propose a new dataset, consisting of 8 scenes with significant subsurface scattering effects. Our captured data demonstrates high fidelity, preserving rich appearance details, and represents a total of 15TB (3000 times larger than the currently highest quality dataset with similar goals to our knowledge, [Deng et al. 2022]).

As shown in Fig. 6, we place the cameras and the light sources on the spherical light stage cage, while the objects to be captured are placed on a holding table in the center with a height of roughly 1.1 meter. In particular, when capturing the data, our cameras and the light bulbs are fixed on the sphere, while a turntable in the middle can be freely rotated. Ignoring background pixels, this is equivalent to keeping the object scene static to satisfy the consistency of the scene among views, while rotating the cameras and the light bulbs altogether. Throughout the text, we assume that the light stage is configured in the latter case for notational convenience. Our camera/light-bulb sphere radius is roughly 1.5 meter from the surface of the holding table in the middle).* The rotations of the sphere put the whole captured frames into 9 groups, with each group

*Our light bulbs only span roughly between $[0, \frac{3}{2}\pi)$ for θ_{ls}, hence no light bulb has a negative altitude even if the sphere radius is larger than the height of the center—the holding table.

corresponding to one particular rotated setting of the camera-light sphere. On the sphere, we have a total of 20 cameras as well as 331 lighting bulbs (serving as 331 OLAT point lights).[†] Consequently, in each group we captured a $20 \times 331 = 6620$ frames, and for the total 9 groups, we captured a total of 59580 frames for one scene. Our camera captures high dynamic range value for the RGBs, with the cutoff threshold at 4.4019. The original captured frames come with a resolution of 8192×5464. We found a 4 times down-scaling retains most of the texture details and hence we conduct all our experiments on the down-scaled version (2048×1366). Notably, all the captures at the resolution of 2048×1366 still span 15TB of storage. During the capture, the cameras always face toward the objects on the holding table, and we tune the focal length of the camera to best suit the size of the particular objects. We obtain the extrinsic camera poses via an off-the-shelf software with manual corrections. Since the light bulbs shining in the opposite direction of the camera incur significant noise to the reconstruction process (especially considering that the rotation between the group would make the background inconsistent), we introduced several heuristics, including RGB variations and saturation to segment out the background. All the camera poses, light locations as well as the masking information are used by all the approaches in our evaluation sections (Sec. 5), and we shall make all the details about the data publicly available to facilitate future research.

5 EXPERIMENTS

We use both the synthetic data (8 scenes) and the real data we captured (8 scenes) for evaluation and comparisons. All the details on data, training and benchmarking protocols will be released.

Synthetic Data. We use the 8 scenes from the NeRF Blender dataset [Mildenhall et al. 2020] and evaluate them with both their original materials (*Synthetic-Original*) as well as their modified materials with the subsurface scattering shader in Blender [Community 2018] (*Synthetic-SSS*). During training, we use the same 100 camera views given in the training set for each scene as provided by the originally released data [Mildenhall et al. 2020]. To simulate OLAT lighting, we evenly sample 112 incident lighting directions on the upper hemisphere. More precisely, we sample evenly with 7 latitudes in the upper hemisphere, evenly sampled 32 longitudes for each latitude, and left out every other light (to be used during evaluation). The 7×32 OLAT directions exactly correspond to Row 2 through Row 8 of the 16×32 envmap as used in NRTF [Lyu et al. 2022]. We exclude the lower hemisphere for OLAT sampling, mainly due to the fact that most of the scenes in the NeRF blender dataset are rendered as top views, and the OLAT lighting from the bottom produces overall dark renderings. This training setting gives us a total of 11200 training images per scene. To mimic the light stage setting used for real-world data capture, we use only white lights, and use the point light instead of the envmap for rendering the ground truth. More precisely, the point lights are placed roughly 100 units away from the scene center (with about 4 units being the approximate size of each scene). During testing, we use unseen lighting directions as well as unseen camera poses for each test sample. For quantitative evaluation, we stick to the OLAT protocol where there is only one light at a time. For saving evaluation time, we only test 10 out of the unseen 112 lights. We also provide qualitative samples by rendering results with several envmaps downloaded from PolyHaven (e.g., Fig. 10). Since our point lights are single-colored (white), we do the inference with the independent-RGB-channel assumption when relighting under a colored envmap. Following [Lyu et al. 2022] we cast them into a 32×16 envmap to serve as the input. For test time camera poses, we apply the camera views from the test views given in the NeRF blender dataset. For saving evaluation time, we only test 10 out of the unseen 200 test views. This test setup gives us 100 test cases in total for each scene.

Light Stage Data. As introduced in Sec. 4, the proposed light stage data contains 9 groups and 20 cameras per scene (a total of 180 views), with each view consisting of 331 OLAT renderings, thus leading to a total of 59580 HDR images per scene. During training, we use the first 18 cameras in each group, and use 75 out of the 331 OLATs for training, leading to a total of 12150 training images per scene. Testing on real data also only includes samples with both, unseen lighting directions and unseen views. For quantitative evaluation, we use the remaining 2 cameras from each group (a total of 18 views) and 10 unseen OLATs to form our test set (180 images per scene). For qualitative evaluation, we use the same input lighting envmaps as used in the synthetic data benchmark. Since most of our real captures exhibit subsurface scattering, we denote this data with *Real-SSS*.

Evaluation Metrics. We evaluate the predicted pixel map following the standard metric protocol [Lyu et al. 2022; Mildenhall et al. 2020], including PSNR, SSIM and LPIPS [Zhang et al. 2018]. While our main focus is to evaluate the objects of the scene, we follow existing protocols [Mildenhall et al. 2020] to include all the pixels for evaluation. Following most of the recent evaluation conventions (e.g. [Mildenhall et al. 2020]), we evaluate every pixel on the predicted pixel maps (including the background regions). This also includes the areas where the stand holds the captured objects.

Baseline approaches. We compare with several most representative state-of-the-art approaches to highlight the strengths of our neural relightable model. All models are trained with exactly the same data.

- **IRON** [Zhang et al. 2022] is a recent representative BRDF-based relighting approach and achieves state-of-the-art performance with the collocated GGX shader. We underwent major efforts to generalize it to the general setting where the incident light direction, viewing direction and bi-sector direction are no longer identical. Notably, while the GGX shader cannot handle subsurface scattering, optimization in scenarios where lighting is coming from the opposite side of the camera is essential, especially when translucency is present. We used Mitsuba 3 [Jakob et al. 2022] to render the trained textured models and fit the best HDR scaling with the ground truth before computing PSNR.
- **InverseTranslucent** [Deng et al. 2022] is a recent representative state-of-the-art BSSRDF relighting approach. We train the models using spatially varying albedo, sigma (controlling light transmission underneath the surface) and roughness all in the resolution of 256 $\times$ 256. We found [Deng et al. 2022] is sensitive to the geometry

[†]Notably, since the point light locations are locked with the camera during rotation, the OLAT location in different groups are different from each other. In other words, in our whole dataset, there are only up to 20 images that have been recorded with the same lighting.

	Real-SSS	C-Red	J-Dragon	J-Fish	C-Cake	S-Lvd.	C-Blue	C-Head	C-S	Average
PSNR(↑)	IRON [Zhang et al. 2022]	21.6	17.5	20.7	22.2	22.4	19.1	21.4	23.3	21.0
	InverseTranslucent [Deng et al. 2022]	23.3	21.6	23.6	22.9	25.1	21.8	25.0	26.8	23.8
	NRTF [Lyu et al. 2022]	27.5	28.5	28.4	29.7	30.7	29.0	30.7	32.0	29.6
	Ours	**30.9**	**29.0**	**30.3**	**32.3**	**33.2**	**31.2**	**34.1**	**36.3**	**32.2**
SSIM(↑)	IRON [Zhang et al. 2022]	85.5	85.7	82.8	88.6	89.2	82.7	90.0	90.8	86.9
	InverseTranslucent [Deng et al. 2022]	86.2	89.6	84.6	89.7	90.7	86.3	92.3	93.3	89.1
	NRTF [Lyu et al. 2022]	92.3	94.0	92.5	94.1	94.7	92.8	95.8	96.5	94.1
	Ours	**93.4**	**94.7**	**93.3**	**94.8**	**95.7**	**93.8**	**96.9**	**97.6**	**95.0**
LPIPS(↓)	IRON [Zhang et al. 2022]	0.131	0.143	0.173	0.108	0.109	0.179	0.109	0.106	0.132
	InverseTranslucent [Deng et al. 2022]	0.139	0.132	0.165	0.119	0.110	0.186	0.104	0.104	0.132
	NRTF [Lyu et al. 2022]	0.110	0.095	0.125	0.088	0.088	0.139	0.082	0.080	0.101
	Ours	**0.099**	**0.089**	**0.123**	**0.078**	**0.077**	**0.132**	**0.071**	**0.067**	**0.092**

Table 1. Comparison with several state-of-the-art methods on the "Real-SSS" data (8 scenes). Despite optimized on the same data, our results consistently outperform the existing approaches on all scenes and all evaluation metrics. Material abbreviations: "C-" stands for "Candle", "J-" stands for "Jade", and "S-" stands for "Soap".

PSNR (↑)	IRON	Inv. Translucent	NRTF	**Ours**
Synthetic-Original	24.4	23.8	29.0	**33.3**
Synthetic-SSS	23.1	26.9	31.1	**39.3**

Table 2. Comparison on Synthetic-Original and Synthetic-SSS. Please refer to our supplementary materials for further details.

initialization, and thus we provide the baseline with the optimized Neus reconstruction using their original implementation [Wang et al. 2021].

• **NRTF** [Lyu et al. 2022] is a recent state-of-the-art fully data-driven approach that is designed to handle global illumination and potentially subsurface scattering.

For [Zhang et al. 2022; Lyu et al. 2022], we use the provided Neus implementation rather than the original version to obtain object surfaces.

It is worth pointing out that [Zhang et al. 2022] was originally proposed to handle only the PBR based materials with the assumptions that all the objects are fully opaque, and hence it was not proposed to handle our evaluation data of *Synthetic-SSS* and *Real-SSS* (our proposed light stage data). Meanwhile, [Deng et al. 2022] was originally proposed to handle specifically objects with translucency, but not necessarily opaque objects as present in our evaluation data *Synthetic-Original*. We still include all results in the experiments for reference purposes since our approach is able to handle all the types of the materials, further showcasing the wide applicability of the method.

Results. As shown in Tab. 1-2 and Fig. 8-10, our results demonstrate clear advantages compared to all aforementioned methods. Notably, we achieve 5 points overall average PSNR gain (averaging over all the synthetic and real data) over the best-performing existing method thanks to our end-to-end learning framework. We conclude that our relighting approach can not only handle a wider range of material types (in particular objects with subsurface scattering effects) with significantly improved fidelity, but also stays flexible representing vivid and rich geometric structures, such as the thin ropes that are generally not easy to represent using meshes. In contrast to other approaches [Zhang et al. 2022; Deng et al. 2022]

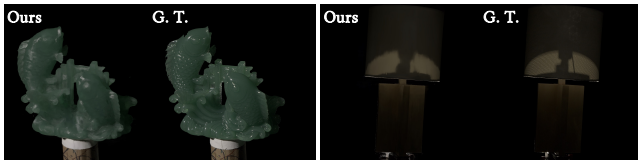


Fig. 7. Failure cases on specular highlights (left) and translucent shadowing (right). The proposed method does not explicitly model specularities and shadowing.

that were designed to handle a relatively narrow range of material types, our approach is able to handle the full variety of materials present in the datasets. This underscores the general applicability of our approach regarding material representations. Please refer to our supplementary materials for additional results.

Limitations. Our approach exhibits two main types of failure modes. First, the proposed method may return blurry results for specular highlights (c.f., Fig. 7—left) since the model does not take specularities into account in a dedicated way. Similarly, our approach does not contain a dedicated model for shadows. In particular, when shadows "penatrate" a thin layer of translucent material (e.g., Fig. 7-right) our model creates blurry boundaries on otherwise hard shadow borders.

Another avenue for future improvement is rendering speed: the proposed model does not yet meet the demand of real-time applications. Further, our relighting algorithm is relying on a light stage capture system and is not yet suited for in-the-wild use.

6 CONCLUSION

We presented a novel volume-rendering based neural relighting approach adept at handling subsurface scattering effects. Thanks to the end-to-end optimization of the radiance transfer gradient on images recorded under various lighting conditions in a light stage, the optimized geometry and appearance reach high quality—even on real data with major subsurface scattering effects. We extensively evaluated the proposed method and established comparisons with several related optimization and modeling approaches and found it to consistently and notably outperform existing work.

REFERENCES

Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5855–5864.

Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5470–5479.

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

Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. 2020b. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*. Springer, 294–311.

Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. 2021a. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12684–12694.

Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. 2021b. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems* 34 (2021), 10691–10704.

Brent Burley and Walt Disney Animation Studios. 2012. Physically-based shading at disney. In *Acm Siggraph*, Vol. 2012. vol. 2012, 1–7.

Chengqian Che, Fujun Luan, Shuang Zhao, Kavita Bala, and Ioannis Gkioulekas. 2020. Towards learning-based inverse subsurface scattering. In *2020 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*. Springer, 333–350.

Zhang Chen, Anpei Chen, Guli Zhang, Chengyuan Wang, Yu Ji, Kiriakos N Kutulakos, and Jingyi Yu. 2020. A neural rendering framework for free-viewpoint relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5599–5610.

Blender Online Community. 2018. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam. http://www.blender.org

Paul Debevec. 2008. A median cut algorithm for light probe sampling. In *ACM SIGGRAPH 2008 classes*. 1–3.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 145–156.

Xi Deng, Fujun Luan, Bruce Walter, Kavita Bala, and Steve Marschner. 2022. Reconstructing Translucent Objects using Differentiable Rendering. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5501–5510.

Graham Fyffe. 2009. Cosine lobe based relighting from gradient illumination photographs. In *SIGGRAPH'09: Posters*. 1–1.

Ioannis Gkioulekas, Shuang Zhao, Kavita Bala, Todd Zickler, and Anat Levin. 2013. Inverse volume rendering with material dictionaries. *ACM Transactions on Graphics (TOG)* 32, 6 (2013), 1–13.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, et al. 2019. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Transactions on Graphics (ToG)* 38, 6 (2019), 1–19.

Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. 2022. Shape, light & material decomposition from images using monte carlo rendering and denoising. *arXiv preprint arXiv:2206.03380* (2022).

Wenzel Jakob, Sébastien Speierer, Nicolas Roussel, and Delio Vicini. 2022. DR. JIT: a just-in-time compiler for differentiable rendering. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–19.

Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. 2021. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *Computer Graphics Forum*, Vol. 40. Wiley Online Library, 101–113.

Linjie Lyu, Ayush Tewari, Thomas Leimkühler, Marc Habermann, and Christian Theobalt. 2022. Neural Radiance Transfer Fields for Relightable Novel-view Synthesis with Global Illumination. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*. Springer, 153–169.

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. 2022. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16190–16199.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. arXiv:2003.08934 [cs.CV]

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15.

Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. 2022. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8280–8290.

Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. 2019. Mitsuba 2: A retargetable forward and inverse renderer. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–17.

Jan Novák, Iliyan Georgiev, Johannes Hanika, Jaroslav Křivánek, and Wojciech Jarosz. 2018. Monte Carlo methods for physically based volume rendering.. In *SIGGRAPH Courses*. 14–1.

Peter-Pike Sloan, Jan Kautz, and John Snyder. 2002. Precomputed radiance transfer for real-time rendering in dynamic, low-frequency lighting environments. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. 527–536.

Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. 2021. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7495–7504.

Delio Vicini, Vladlen Koltun, and Wenzel Jakob. 2019. A learned shape-adaptive subsurface scattering model. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–15.

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

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *arXiv preprint arXiv:2106.10689* (2021).

Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. 2022. Neural fields in visual computing and beyond. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 641–676.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. Plenoctrees for real-time rendering of neural radiance fields. *arXiv preprint arXiv:2103.14024* (2021).

Hong-Xing Yu, Michelle Guo, Alireza Fathi, Yen-Yu Chang, Eric Ryan Chan, Ruohan Gao, Thomas Funkhouser, and Jiajun Wu. 2023. Learning object-centric neural scattering functions for free-viewpoint relighting and scene composition. *arXiv preprint arXiv:2303.06138* (2023).

Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. 1999. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 215–224.

Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. 2022. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5565–5574.

Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. 2021a. 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*. 5453–5462.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 586–595.

Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. 2021b. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–18.

Quan Zheng, Gurprit Singh, and Hans-Peter Seidel. 2021. Neural relightable participating media rendering. *Advances in Neural Information Processing Systems* 34 (2021), 15203–15215.

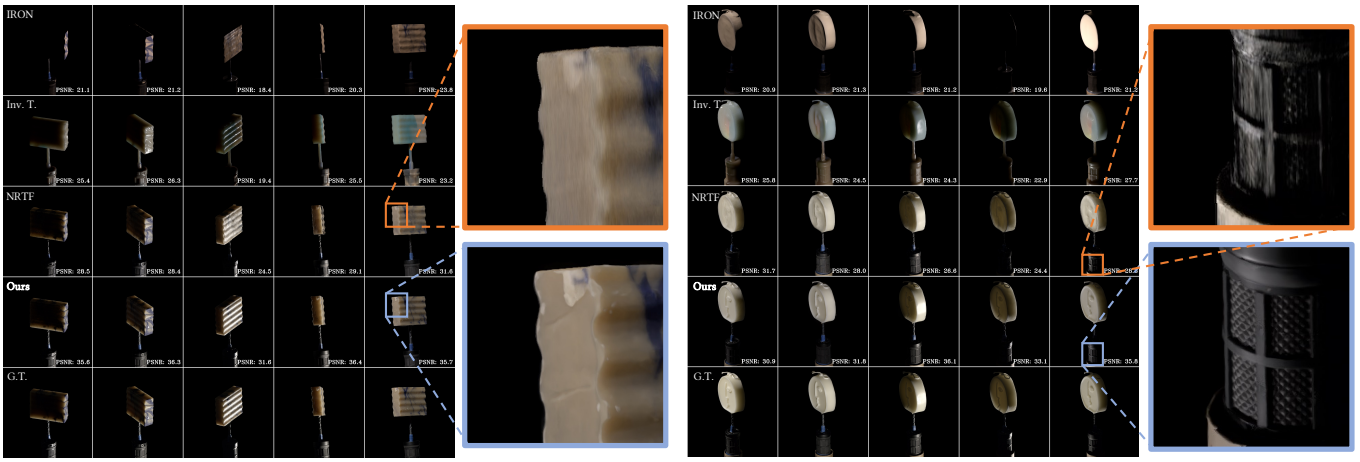


Fig. 8. Detailed comparison for Soap-Lavender (**left**) and Candle-Head (**right**) between our results (Row 4) with other state-of-the-art approaches (IRON [Zhang et al. 2022] in Row 1, InverseTranslucent [Deng et al. 2022] in Row 2, and NRTF [Lyu et al. 2022] in Row 3). Recordings can be found in the last row; all images are held out positions for lights and cameras. Our results show a clear advantage in terms of visual fidelity and geometric accuracy.

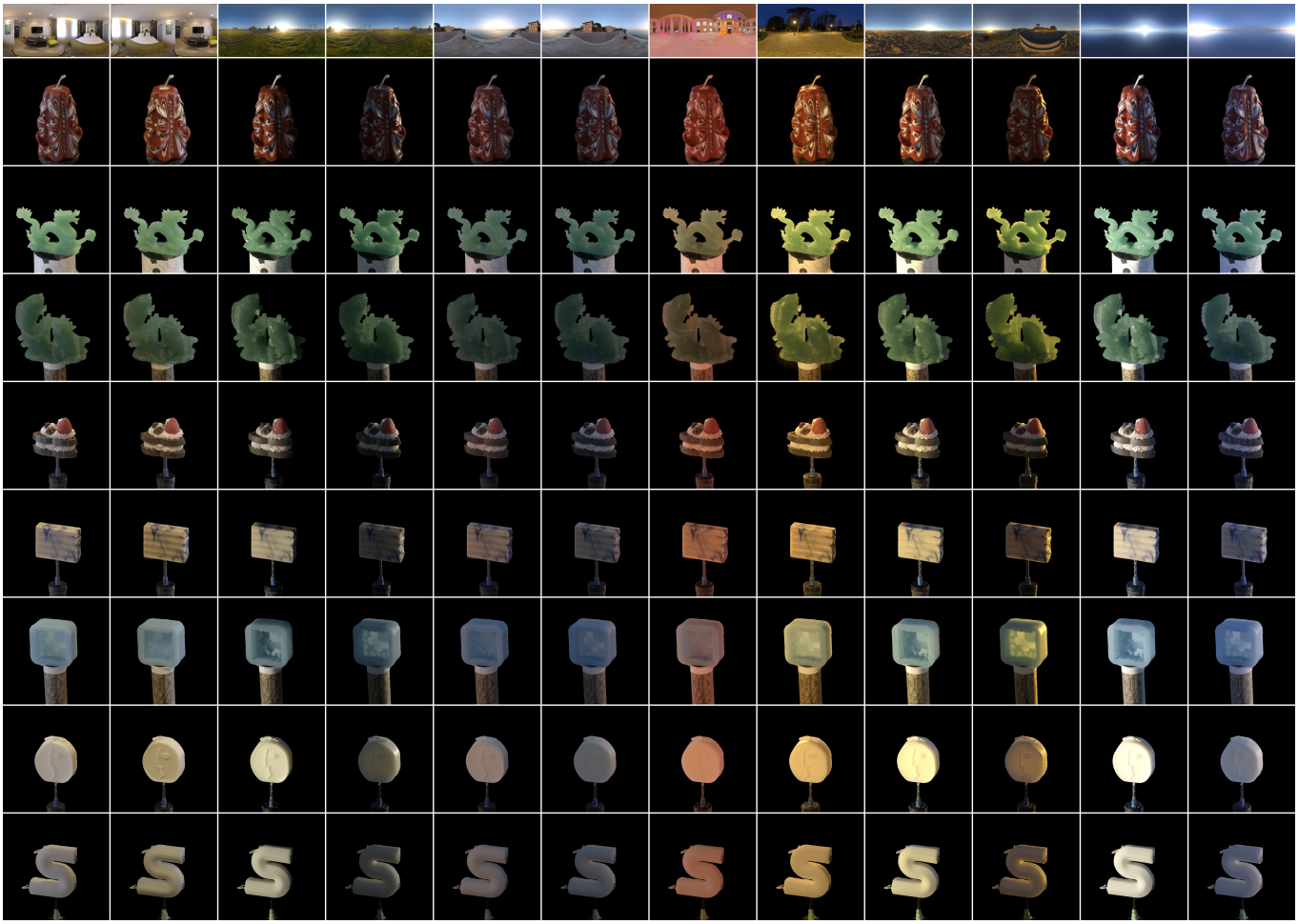


Fig. 9. Envmap relighting results on our real-SSS dataset (light stage captures). The results in each row are from the same scene, while the results in each column are relit using the same environment map.

Fig. 10. Relighting results for various environment maps for the original as well as the translucent version of the synthetic scenes from the Nerf-Blender synthetic datasets (*Synthetic-Original* and *Synthetic-SSS*).