Learning Implicit Surface Light Fields

Michael Oechsle[1,2] Michael Niemeyer[1] Christian Reiser[1]
Lars Mescheder[1,3] Thilo Strauss[2†] Andreas Geiger[1]
[1]Max Planck Institute for Intelligent Systems and University of Tübingen
[2]ETAS GmbH, Bosch Group, Stuttgart [3]Amazon, Tübingen
[1]{firstname.lastname}@tue.mpg.de [†]{firstname.lastname}@etas.com

Abstract

Implicit representations of 3D objects have recently achieved impressive results on learning-based 3D reconstruction tasks. While existing works use simple texture models to represent object appearance, photo-realistic image synthesis requires reasoning about the complex interplay of light, geometry and surface properties. In this work, we propose a novel implicit representation for capturing the visual appearance of an object in terms of its surface light field. In contrast to existing representations, our implicit model represents surface light fields in a continuous fashion and independent of the geometry. Moreover, we condition the surface light field with respect to the location and color of a small light source. Compared to traditional surface light field models, this allows us to manipulate the light source and relight the object using environment maps. We further demonstrate the capabilities of our model to predict the visual appearance of an unseen object from a single real RGB image and corresponding 3D shape information. As evidenced by our experiments, our model is able to infer rich visual appearance including shadows and specular reflections. Finally, we show that the proposed representation can be embedded into a variational auto-encoder for generating novel appearances that conform to the specified illumination conditions.

1. Introduction

Recently, neural implicit representations of 3D objects have emerged as a powerful paradigm for reasoning about the geometry and texture of objects from a single image as input [36, 41–43, 50, 53]. The main advantage of implicit models is their ability to represent 3D structure continuously while handling arbitrary geometric topologies. Unfortunately, however, existing implicit methods are not able to model the full visual appearance of 3D objects which requires reasoning about the complex interplay of light, geometry and surface properties. Consequently, the resulting 3D reconstructions appear "lifeless", missing view- and light-dependent illumination effects such as shadows or specular reflections.

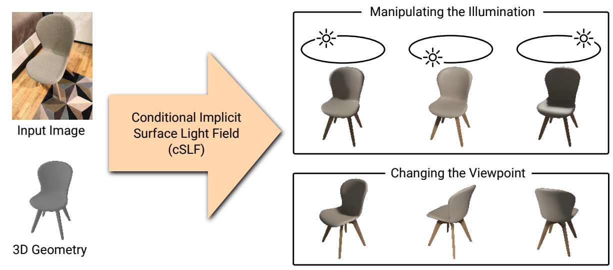

Figure 1. **Overview.** Given an RGB image and corresponding (predicted or true) 3D geometry as input, our conditional implicit surface light field model allows for manipulating the light configuration and synthesizing novel viewpoints. Note how shadows and specular reflections are faithfully captured by our model.

In the graphics and vision community, synthesizing rich visual appearance of objects using learning-based methods has emerged as a popular research direction with the promise to replace the currently dominating, but slow and work-intensive 3D modeling and physically-based rendering [16] paradigm. Some methods learn neural rendering of voxel features [38, 52] or neural texture maps [56] for synthesizing novel views of an object. Similarly, Chen et al. [8] learn the surface light field of a single object. Even though these methods achieve high realism, they must be retrained for each new object and they do not allow modification of the illumination. In contrast, in this work, we are interested in the following, more challenging scenario: Our goal is to infer *conditional* surface light fields which allow for manipulating illumination conditions and which can be applied to novel, previously unseen objects.

More specifically, we propose an implicit representation for parameterizing conditional surface light fields. As illustrated in Fig. 1, our model takes an RGB image and

[1]This work was done prior to joining Amazon.

corresponding 3D shape information of an object as input and allows for generating photo-realistic images of the object from arbitrary viewpoints and light configurations. Towards this goal, we train our model to regress color values given a 3D location on the object's surface, the camera viewpoint and the location and color of a point light source. We empirically demonstrate that our network represents high-frequency material properties while generalizing across object shapes, light settings and viewpoints. While trained on synthetic renderings and ground truth geometry, we find that our method generalizes to real images and captures complex physical illumination phenomena including shadows and specular highlights. Exploiting the well-known rendering equation [17], we demonstrate that our model extends beyond point light sources, enabling implicit rendering using realistic environment maps. We also extend our representation to the generative setting which allows for synthesizing novel appearances conforming to given illumination conditions and for transferring appearances from one model to another. Our contributions are summarized as follows:

- We propose **Conditional Implicit Surface Light Fields (cSLF)**, a novel appearance representation of 3D objects which allows for rendering novel views and varying the light color and the light location.
- We experimentally verify that our method is able to represent textures, diffuse and specular reflections as well as shadows of a 3D object.
- We apply our model for predicting the appearance of a novel object from a single RGB image given ground truth and inferred 3D shape information.
- We demonstrate the generative modeling capabilities of our representation by transferring and synthesizing novel physically plausible appearances.

Code is available at `https://github.com/autonomousvision/cslf`

2. Related Work

We now discuss the most related works on neural and differentiable rendering, material and surface light field estimation and implicit shape and appearance modeling.

Neural Rendering: The recent success of Generative Adversarial Networks (GANs) [6,11,20,21] has enabled novel image synthesis approaches which employ 2D or 3D convolutional neural networks to generate image content directly from a latent code. These methods mainly differ in terms of the input representation. Examples include voxel-based representations [2,38–40,47,52], primitive-based representations [25], depth-based representations [1,63] and texture-based representations [34, 56]. In contrast to CNN based rendering approaches, our method is inherently 3D consistent as we directly predict the object appearance in 3D. Moreover, our model allows for fine grained viewpoint and lighting control.

Image-based relighting approaches [18,32,45,51,54,59, 61] infer images from the same view in other light settings. In contrast, our method can reason about novel views and light configurations from a single image.

Differentiable Rendering: Recently, several methods proposed to backpropagate gradients through the rendering pass [9,22,27,30]. This allows for predicting explicit 3D appearance representations including texture maps [9, 22, 27] and voxel colors [13]. Chen et al. [9] learn to predict the geometry and texture of a 3D mesh as well as the light position from a single image. In contrast to these methods, we are interested in inferring a conditional surface light field which also captures more complex physical light-transport phenomena including both shadows and specular highlights. Furthermore, our implicit representation is continuous and independent of the underlying geometric representation.

Materials and Surface Light Fields: Existing methods on reflectance capture estimate (sv)BRDFs from one or multiple measurements [3, 5, 19, 24, 26, 28, 29, 33, 62]. In contrast to these works which use explicit material models and illumination maps, we learn an implicit model which directly represents a surface light field for given lighting and geometry. A surface light field represents the outgoing light at a given location on a 3D surface as a function of the viewing direction [12, 58]. Maximov et al. [31] propose a method for representing material-light interaction with a neural network which maps normals and the viewing direction to color values. Chen et al. [8] model the appearance of a triangular mesh based on the UV coordinates and the viewing direction. In contrast to traditional surface light fields we do not assume fixed illumination but instead condition on the light source. This allows our model to manipulate illumination at inference time. Furthermore, while all aforementioned models consider a single object, our approach is able to predict light fields for unseen objects.

Implicit 3D Representations: Recently, implicit 3D geometry models have gained popularity [10, 36, 43] as they circumvent the discretization inherent to traditional explicit representations. Implicit appearance models represent texture information by learning the mapping from 3D location to color [41, 42, 50, 53]. However, they are restricted to simple textures which represent only diffuse material properties. In contrast, our model represents both diffuse and specular material properties as well as shadows. Furthermore, modeling surface properties allows for capturing high-frequency details which cannot be captured with texture models (see Fig. 5 in our experiments). [37], proposes neural radiance fields as an implicit representation of visual appearance that allows for photo-realistic novel view synthesis. Very recent, concurrent work [60] introduces a view-dependent implicit representation for modeling appearance and geometry from multi-view images. While [37] and [60] can only represent single scenes with the original light setting, our model allows for predicting the appearance of a novel scene from a single image and for rendering unseen light configurations.

3. Method

An overview of our model is provided in Fig. 2. We take a 2D image and the corresponding 3D shape as input, and encode both into latent representations ($\mathbf{z}$ and $\mathbf{s}$) which are fed into our conditional implicit Surface Light Field (cSLF). Given a 3D point $\mathbf{p}$, its viewing direction $\mathbf{v}$ and the light configuration $\mathbf{l}$, our cSLF model then outputs the predicted color value for this point.

We first describe the physical image formation process and traditional surface light field representations. Next, we introduce our cSLF model. Finally, we provide details about the network architectures and the loss functions.

3.1. Background

Physically-based rendering techniques [16] utilize the rendering equation to calculate the amount of light that radiates from a specific surface location into the direction of the camera. Let $\mathbf{p} \in \mathbb{R}^3$ denote a 3D surface point and $\mathbf{v} \in \mathbb{R}^3$ the viewpoint of the camera, i.e., the vector from $\mathbf{p}$ to the camera center. Let further $\mathbf{r} \in \mathbb{R}^3$ denote the incoming light direction. The rendering equation describes how much of the light arriving at $\mathbf{p}$ is reflected into the camera direction $\mathbf{v}$:

$$L(\mathbf{p},\mathbf{v}) = \int_{\Omega} \text{svBRDF}(\mathbf{p},\mathbf{r},\mathbf{v}) \cdot L_{\text{in}}(\mathbf{r}) \cdot (\mathbf{n}^T\mathbf{r})\, d\mathbf{r} \quad (1)$$

Here, $\text{svBRDF}(\mathbf{p},\mathbf{r},\mathbf{v})$ denotes the spatially varying bidirectional reflectance distribution function which models the interaction between light and the surface of a 3D object by determining the proportion of light reflected in outgoing direction $\mathbf{v}$ from incoming light direction $\mathbf{r}$. Furthermore, L_{in} is the incoming radiance and $\mathbf{n} \in \mathbb{R}^3$ describes the surface normal at point $\mathbf{p}$.

Surface light fields [12, 58] instead directly represent the radiated light as a function that associates a color value $\mathbf{c} \in \mathbb{R}^3$ with every surface location $\mathbf{p}$ and view direction $\mathbf{v}$. Formally, they are described by the following mapping:

$$L_{\text{SLF}}(\mathbf{p},\mathbf{v}) : \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}^3 \quad (2)$$

While this mapping can be learned using neural networks [8, 31], surface light fields do not allow for varying the light configuration. Instead, the illumination properties are "hard-coded" in this representation. In contrast, the model proposed in the following is conditioned on the light configuration, thus allowing the modification of illumination parameters at inference time.

3.2. Conditional Implicit Surface Light Fields

In this work, we aim at representing illumination-dependent visual appearance by learning a surface light field conditioned on a single light source. We refer to this representation as conditional implicit Surface Light Field (cSLF)

$$L_{\text{cSLF}}(\mathbf{p},\mathbf{v},\mathbf{l},\mathbf{s},\mathbf{z}) : \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R}^M \times \mathcal{S} \times \mathcal{Z} \rightarrow \mathbb{R}^3 \quad (3)$$

where $\mathbf{l} \in \mathbb{R}^M$ denotes the parameters of a point light source[1] (e.g., its location in 3D space and color). $\mathbf{s} \in \mathcal{S}$ and $\mathbf{z} \in \mathcal{Z}$ encode the object shape and image content, respectively. We use encodings $\mathbf{s}$ and $\mathbf{z}$ only for tasks involving unseen objects, such as single view appearance prediction.

We represent L_{cSLF} implicitly using a neural network. In contrast to traditional surface light fields (2), our implicit representation allows for querying the light field for any light configuration $\mathbf{l}$ and object encoding $(\mathbf{s},\mathbf{z})$. This allows our model to generalize to novel, previously unseen 3D objects.

While our cSLF model considers a single point light source during training, more complex light settings, including spatial varying illumination, can be achieved by combining multiple evaluations of Eq. (3) at inference time. Discretizing the domain Ω in Eq. (1) and letting $\mathcal{L} = \{\mathbf{l}_1, \ldots, \mathbf{l}_K\}$ denote a sampling-based approximation of the environment map (i.e., $\mathbf{l}_k$ in $\mathcal{L}$ corresponds to one pixel in the environment map), we obtain the aggregated surface light field as:

$$L(\mathbf{p},\mathbf{v},\mathcal{L},\mathbf{s},\mathbf{z}) = \frac{1}{|\mathcal{L}|} \sum_{\mathbf{l} \in \mathcal{L}} L_{\text{cSLF}}(\mathbf{p},\mathbf{v},\mathbf{l},\mathbf{s},\mathbf{z}) \quad (4)$$

We split the parameterization of Eq. (3) into a two-step process. We first map the 3D location $\mathbf{p}$ conditioned on global shape ($\mathbf{s}$) and image ($\mathbf{z}$) information into an D-dimensional appearance feature $\mathbf{f}$. The resulting *appearance field* is described by:

$$a_\theta(\mathbf{p},\mathbf{s},\mathbf{z}) : \mathbb{R}^3 \times \mathcal{S} \times \mathcal{Z} \rightarrow \mathbb{R}^D \quad (5)$$

Note that in contrast to the global image feature, $\mathbf{f} = a_\theta(\mathbf{p},\mathbf{s},\mathbf{z})$ is a localized appearance representation of the 3D object. As $\mathbf{f}$ is independent of the viewpoint $\mathbf{v}$ and lighting $\mathbf{l}$, it captures mostly material properties. The interaction between the object surface and light is modeled using a *lighting model* which maps the appearance $\mathbf{f}$, the viewpoint $\mathbf{v}$, the light $\mathbf{l}$ and the shape $\mathbf{s}$ into an RGB value:

$$l_\theta(\mathbf{f},\mathbf{v},\mathbf{l},\mathbf{s}) : \mathbb{R}^D \times \mathbb{R}^3 \times \mathbb{R}^M \times \mathcal{S} \rightarrow \mathbb{R}^3 \quad (6)$$

[1] In practice, we use a small area light source to avoid hard non-differentiable shadows.

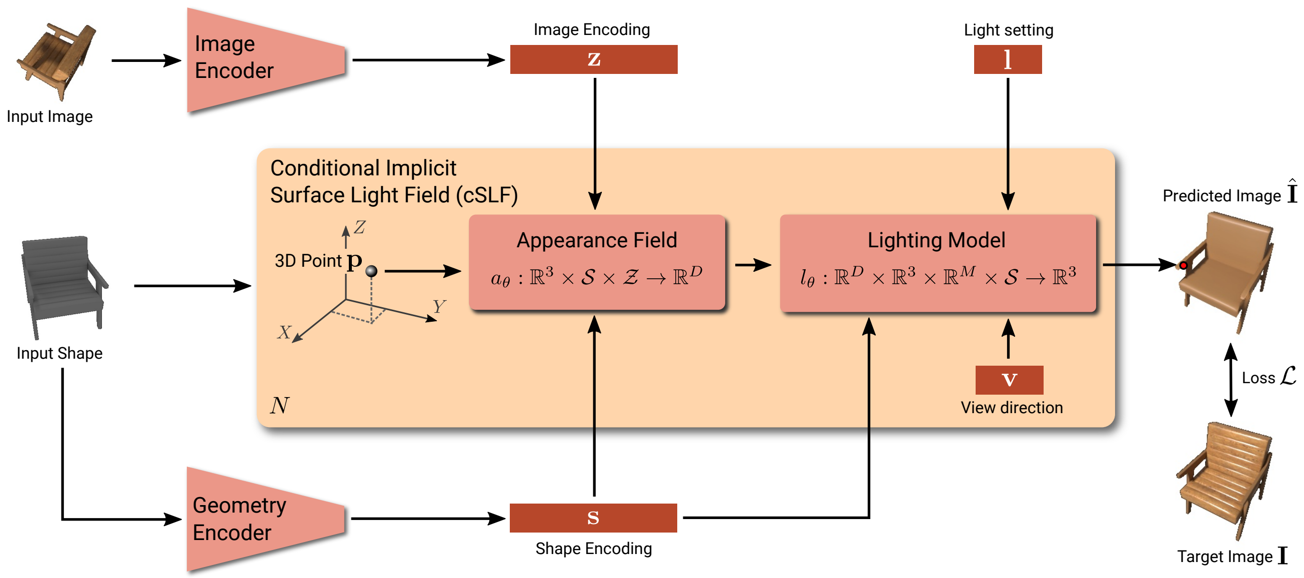

Figure 2. **Model Overview.** Our Conditional Implicit Surface Light Field (cSLF) takes an encoding of a real image ($\mathbf{z}$) and corresponding 3D shape ($\mathbf{s}$) as well as a light configuration ($\mathbf{l}$) as input and outputs a color value for every 3D location $\mathbf{p} \in \mathbb{R}^3$.

Splitting the inference process into two steps allows for precomputing the appearance features. Thus, when manipulating the light setting, only the lighting model $l_\theta(\mathbf{v}, \mathbf{f}, \mathbf{l}, \mathbf{s})$ needs to be evaluated. This is particularly important when considering more complex light settings (e.g., environment lighting) as in Eq. (4).

3.3. Network Architectures

We now briefly discuss the network architectures. See supplementary for details.

Image Encoder: We use the image encoder of [36] which consists of a ResNet-18 that outputs a 512-dimensional image encoding $\mathbf{z}$.

Geometry Encoder: To provide global shape information to our model, we sample 2048 points from the object surface and use the PointNet-based residual network of [36] which encodes the 3D point cloud into a shape vector $\mathbf{s}$.

Surface Light Field: We parameterize the networks $a_\theta(\cdot)$ and $l_\theta(\cdot)$ of our surface light field model using fully-connected residual networks. To circumvent the intrinsic low-frequency bias of fully-connected neural networks, we encode all spatial inputs $(\mathbf{p}, \mathbf{v}, \mathbf{l})$ with a Fourier encoding γ using k octaves [37, 46]:

$$\gamma(\mathbf{x}) = \left[\sin(2^0\pi\mathbf{x}), \cos(2^0\pi\mathbf{x}) ... \sin(2^k\pi\mathbf{x}), \cos(2^k\pi\mathbf{x})\right] \tag{7}$$

We found this partiularly useful for single object experiments with high-frequency details, see supp. material.

3.4. Loss Functions

We train our model in a supervised fashion on photo-realistic renderings of 3D models. More specifically, we randomly sample a batch of 3D models along with one random input and one random target view. We also randomly sample the point light source location and color ($\mathbf{l}$). We randomly select N pixels from the target view and project them into 3D, resulting in a set of 3D points $\mathbf{p}$. We also render photo-realistic RGB images for both the input and the target view using the Blender Cycles renderer. During training, we apply a photometric loss between the predicted color values $\hat{\mathbf{I}}$ and the ground truth color values $\mathbf{I}$:

$$\mathcal{L}(\hat{\mathbf{I}}, \mathbf{I}) = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \|\hat{\mathbf{I}}_b - \mathbf{I}_b\|_1 \tag{8}$$

where $\hat{\mathbf{I}}_b$ represents the predicted color values for all N points of batch $\mathcal{B}$.

As our generative model, we use a Variational Autoencoder (VAE) [23, 48]. More specifically, we use a ResNet-based encoder $q_\phi(\mathbf{z}|\mathbf{I}, \mathbf{s})$ which maps an image $\mathbf{I}$ and a shape representation $\mathbf{s}$ to the mean and variance of a normally-distributed latent variable $\mathbf{z}$. The decoder is given by the surface light field model described in Section 3.2 which predicts a color value for every pixel, given its 3D location $\mathbf{p}$, the viewpoint $\mathbf{v}$, the light configuration $\mathbf{l}$, the object shape $\mathbf{s}$ and the latent code $\mathbf{z}$. The loss function for training our VAE model is thus given by

$$\mathcal{L}_{\text{VAE}} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} [\beta \, \text{KL}(q_\phi(\mathbf{z}_b|\mathbf{I}_b, \mathbf{s}_b) \parallel \mathcal{N}(\mathbf{z}_b|0, \mathbf{I})) + \|\hat{\mathbf{I}}_b - \mathbf{I}_b\|_1] \tag{9}$$

where $\text{KL}(\cdot)$ represents the Kullback-Leibler divergence, $\mathbf{z}_b$ follows the posterior distribution $q_\phi(\mathbf{z}|\mathbf{I}_b, \mathbf{s}_b)$ and $\beta = 1$ is a trade-off parameter [15] between the KL-divergence and the reconstruction loss.

4. Experimental Evaluation

We conduct three different types of experiments. First, we investigate the ability of our method to capture the surface light field of a single object illuminated by a point light source. Next, we use our model to predict the surface light field of previously unseen objects from a single image. Furthermore, we provide results for geometry reconstruction and appearance prediction from a single image by applying our method to the output of Occupancy Network [36] reconstructions. Finally, we show qualitative results from our generative model which learns the conditional distribution of surface light fields given the object geometry.

Dataset: In our single object experiments, we show results on 3D models with high-quality materials, texture and geometry from free3d.com and turbosquid.com. To further investigate high-frequency appearance details, we use a planar high-quality svBRDF scan from [35]. For our single view prediction task, we choose ShapeNet categories (cars, lamps, sofas) with rich appearance information [7]. While ShapeNet provides a large variety of 3D models, it is limited in terms of complex materials. We therefore additionally use the Photoshapes dataset [44] which contains 5452 meshes of chairs with a large variety of realistic materials. We split the datasets into a training set, a validation set and a test set with ratio $7:1:2$. For testing our models on real-world data we use the Pix3D dataset [55].

Rendering: We use the Blender Cycles Renderer [4] to generate ground truth RGB images and depth maps from random viewpoints and light locations, sampled from the upper hemisphere. More details on the rendering process are provided in the supplementary.

Baselines: As many existing neural rendering techniques use 2D CNNs for image synthesis, we consider a CNN-based image-to-image translation network as a baseline for our method. Similar to [42, 56], we employ a U-Net architecture [49] which maps a depth image from the same viewpoint to an RGB image, conditioned on the light location and the geometry of the object. We refer to this baseline as Img2Img, see supplementary for details. As a second baseline, we use Texture Fields [42], a state-of-the-art implicit texture model.

4.1. Single Object Experiment

We first investigate the representation power of our model wrt. high-frequency textures, shadows and reflections by training separate models for all objects and test our model's ability to generalize to novel unseen viewpoints and light configurations. In Fig. 3, we provide predictions of our model for a diverse set of 3D models, varying views and light locations. We use renderings from 100 views with 40 light configurations. We apply our method for representing planar svBRDF samples and show a zoom-in of the barrel model in Fig. 4. To analyze shadows, we choose a chair with thin armrests in Fig. 5. For analyzing reflections, we use a car model which contains highly specular materials in Fig. 6. For training, we render 50 views with 30 lighting conditions each. For evaluation, we choose views that are substantially different from the training views wrt. viewpoint and light location. We also show the nearest neighbor from the training set (Fig. 5 left).

High-frequency Details: Our model is able to accurately infer textural details as well as complex appearances in different light settings, see Fig. 3. For a more in-depth investigation, we also train our model to represent high-frequency planar svBRDFs. Our method performs on par with the Img2Img baseline, even though this architecture is known to be highly effective for representing 2D images. In the 3D domain our method performs superior to the baseline in capturing high-frequency details (see results on 3D barrel).

Shadow Analysis: Fig. 5 shows a comparison of our model to Texture Fields [42], a state-of-the-art implicit model for texture representation and the Im2Im baseline. As illustrated in Fig. 5, our method successfully infers shadows for novel view and lighting configurations. Note that synthesizing accurate shadows requires reasoning about the underlying geometry and is hence very challenging. In contrast, the texture-based model [42] "averages" shadows and texture details due to its inability to capture illumination effects. The Im2Im baseline is not able to properly represent shadows, that may result from missing 3D reasoning.

Reflection Analysis: To further investigate the capability of our model to accurately represent surface light fields, we apply our method to a car model which consists of highly specular materials in Fig. 6. By representing complex illumination as a combination of light sources, our method is able to reason about the appearance of 3D objects in complex illumination settings. We represent environment maps as the superposition of point light sources using Eq. (4) and show examples of the car model rendered using different environment maps in Fig. 6. We find that our model is able to accurately capture specular reflections, which allows for rendering 3D objects under complex illumination.

4.2. Single View Appearance Prediction

While reasoning about the inter-dependence between light, surface, and viewpoint of a single object is a challenging task, we now address the even more challenging problem of inferring surface light fields for novel, previously unseen objects. To successfully solve this task, the network must implicitly infer appearance properties and geometric properties (e.g., surface normals, shadowing effects, etc.) of objects from the input image and shape encoding. For this experiment, we render images and depth maps from 10 dif-

Figure 3. **Generalization to unseen viewpoints and light configurations.** Predicted novel views of high-quality 3D models for different light locations. Our method predicts photo-realistic visual appearances with high-frequency details as well as plausible shadows, reflections and specularities.

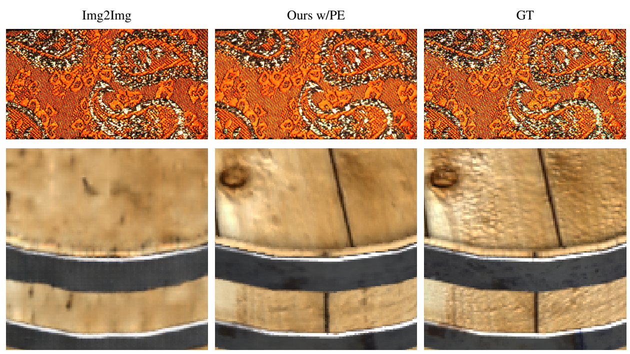

Figure 4. **High-frequency Details.** Qualitative comparison to the Img2Img baseline. Our method performs similarly for planar svBRDFs, while showing more textural details for 3D objects.

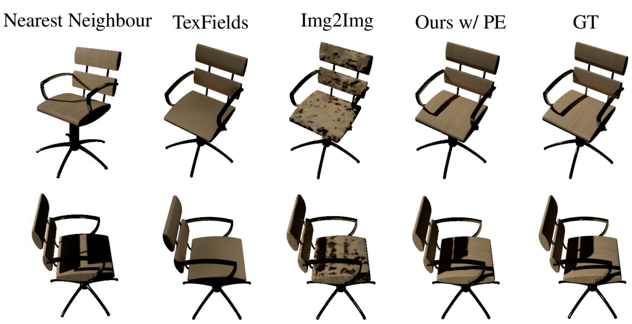

Figure 5. **Shadow Analysis.** Comparison of TextureFields [42], the Im2Im baseline and our model with respect to the nearest neighbor and ground truth for two different views.

Figure 6. **Reflection Analysis.** We show results for a highly specular car with two different environment maps.

ferent viewpoints per object with 4 randomly sampled light locations per view. For each object class, we train a separate model using an image resolution of 256^2 pixels.

Metrics: We evaluate all methods on three different metrics. First, we compute the FID score between a set of predicted images and ground truth images [14]. We further report the Structure Similarity Image Metric (SSIM) [57] and a l_1 distance measure in the Inception feature space between the predicted image and the ground truth as in [42].

Surface Light Field Prediction: In our first experiment, we predict the appearance of a novel object from a single RGB image and the corresponding ground truth geometry embedding. Fig. 7 shows the predictions for different light locations using our method, the Img2Img baseline and Texture Fields [42].

Does the model predict accurate textures? Fig. 7 shows that our method is able to reconstruct details from the input view like the brown armrests and chair legs. While our methods reconstruct plausible colors, the Img2Img baseline produces artifacts at borders between two color regions.

Does the model predict plausible shadows for different light configurations? Fig. 7 shows that our method is able to reason about shadows. Note that even for the geometry of the last example, our model predicts a rough silhouette. However, a general limiting factor is the presence of fine-structured geometry that complicates predicting shadows and reflections. The Img2Img baseline leads to very noisy predictions. The Texture Field baseline lacks a physical understanding of image formation and thus “averages” shadows in the training data, leading to blurry results.

How well does the model represent specular reflections? The examples in Fig. 7 contain specularities as well as more diffuse reflection properties. In contrast to the Img2Img Baseline and Texture Fields, our method is able to predict plausible specular reflections (brown chair) as well as diffuse materials (red couch). Note that our approach can't model transparencies.

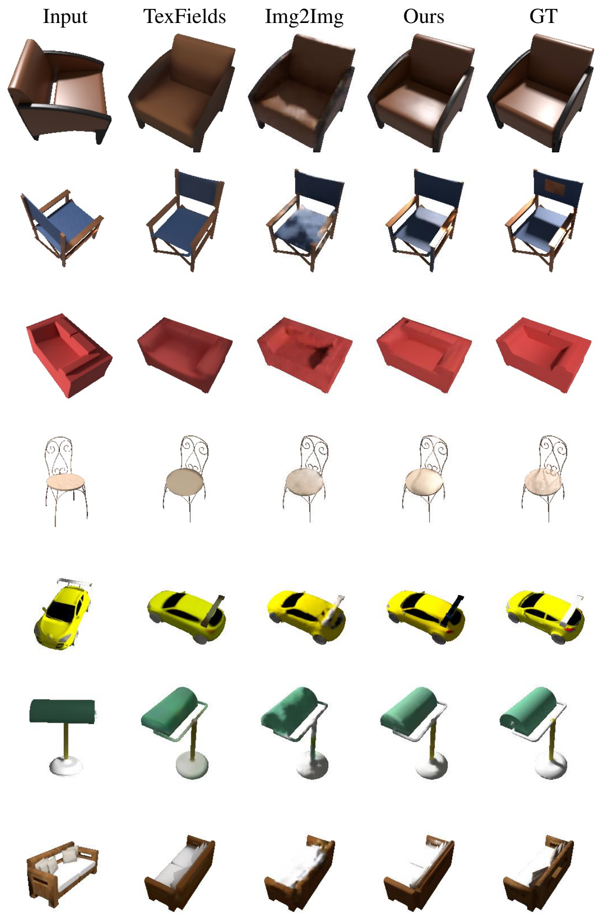

Figure 7. **Single View Appearance Prediction.** Results of our method and the provided baselines for predicting the appearance of a novel, previously unseen object.

Quantitative Comparison: In Table 1, we quantitatively compare our method to the baselines. While our method and the Img2Img baseline compare similarly in SSIM, our method performs favorably in terms of FID and Feature-l_1 distance. As expected, the Texture Field baseline is significantly worse compared to the other methods, since it does not capture the full visual appearance. However, we remark that none of the metrics focuses explicitly on shadows or surface properties, highlighting the importance of our qualitative comparisons.

Ablation Study: We hypothesized that our model relies on geometric information (shape encoding $\mathbf{s}$ in Fig. 2) to model shadows and surface properties. We therefore compare our approach against a version without shape encoding (Ours w/o shape $\mathbf{s}$) in Table 2. Indeed, removing the shape encoding leads to a degradation in all metrics. We further investigate if our method can also be trained on only two views per object with different light settings (2 views). While two input views per object are sufficient for learning to infer plausible appearance from a single image, the complex physical light formation process cannot be modeled well. We provide qualitative results in the supplementary.

Figure 8. **Image-based Reconstruction.** Reconstruction of geometry and appearance based on a single RGB image. We show the input image in the first column and predictions of our 2-step model for different light configurations in the following columns.

Image-based Reconstruction: While we use ground truth geometry as input in all previous experiments, such information is often not available in real-world settings. We therefore investigate whether our approach allows for inferring the appearance when using inaccurate image-based 3D reconstructions as input. Towards this goal, we train the model proposed in [41] on our renderings and combine it with our surface light field to reconstruct both geometry and appearance from a single RGB image. As evidenced by Fig. 8, our surface light field model allows for inferring accurate textures, reflections and shadow effects even when using less accurate, inferred 3D geometry as input.

Real Images: We now investigate if our approach trained on synthetic data generalizes to real images as input. We use the Pix3D dataset [55] for this purpose, as it offers real input images with aligned 3D geometry and apply the provided instance masks to the input images. As illustrated in Fig. 9, our approach successfully reasons about the appearance of real objects despite being trained on synthetic data only (both in terms of input and ground truth supervision).

	FID			SSIM			Feature-ℓ_1		
	TexFields	Img2Img	Ours	TexFields	Img2Img	Ours	TexFields	Img2Img	Ours
photoshape	33.467	27.019	**20.602**	0.949	**0.964**	0.961	0.160	0.142	**0.131**
cars	75.411	89.401	**39.170**	0.927	**0.941**	0.935	0.215	0.223	**0.187**
lamps	48.297	46.617	**42.464**	0.960	**0.970**	0.969	0.165	0.153	**0.150**
sofas	43.593	40.424	**32.172**	0.937	**0.963**	0.961	0.163	0.145	**0.136**
mean	50.192	50.865	**33.602**	0.943	**0.959**	0.957	0.176	0.166	**0.151**

Table 1. **Single View Appearance Prediction.** Quantitative comparison of our approach with the Img2Img baseline. Our method performs better on the FID and Feature-l_1 metric. In the less accurate SSIM metric all methods perform similarly.

	FID	SSIM	Feature-ℓ_1
Ours	**20.602**	**0.961**	**0.131**
Ours (2 V)	25.922	0.952	0.149
Ours w/o $\mathbf{s}$	24.771	0.956	0.143

Table 2. **Ablation Study.** Quantitative results when removing the input shape encoding ($\mathbf{s}$) and reducing the number of training views per object to 2.

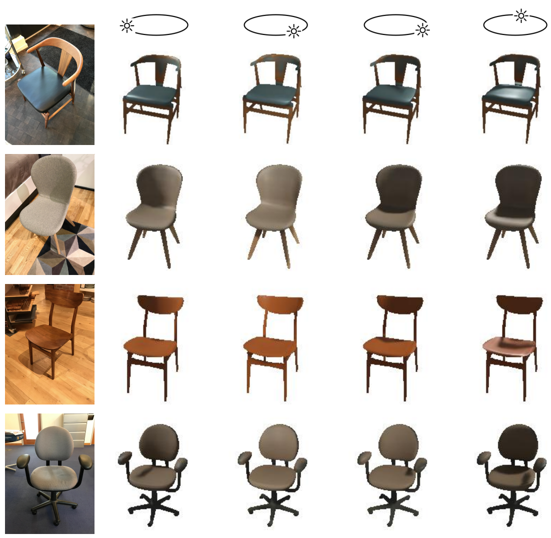

Figure 9. **Real Images.** Results of our method when using real images of unseen objects as input. Even though our model is trained on synthetic images, it predicts plausible appearance for real images. Note that we used masked images as input.

4.3. Generative Model

In addition to predicting appearance from a single image, we use our approach as a generative model for learning the conditional distribution of surface light fields given a 3D model. In Fig. 10, we show the output of our model for randomly sampled appearance codes $\mathbf{z}$ as well as latent space interpolations using our VAE-based generative model. Our model learns a meaningful latent space, which also allows for transferring appearance from one model to another. We observe that the predictions contain plausible specular reflections and shadows.

5. Conclusion

We proposed conditional implicit surface light fields, a novel representation of visual appearance of 3D objects. Our model reasons about complex visual effects including texture, diffuse and specular reflection as well as shadows. Our representation allows for inferring arbitrary viewpoints with novel light configurations. We find that conditional implicit surface light fields are an effective representation of visual appearance. In future work, we plan to further increase the representation power of this model by exploiting convolutional features.

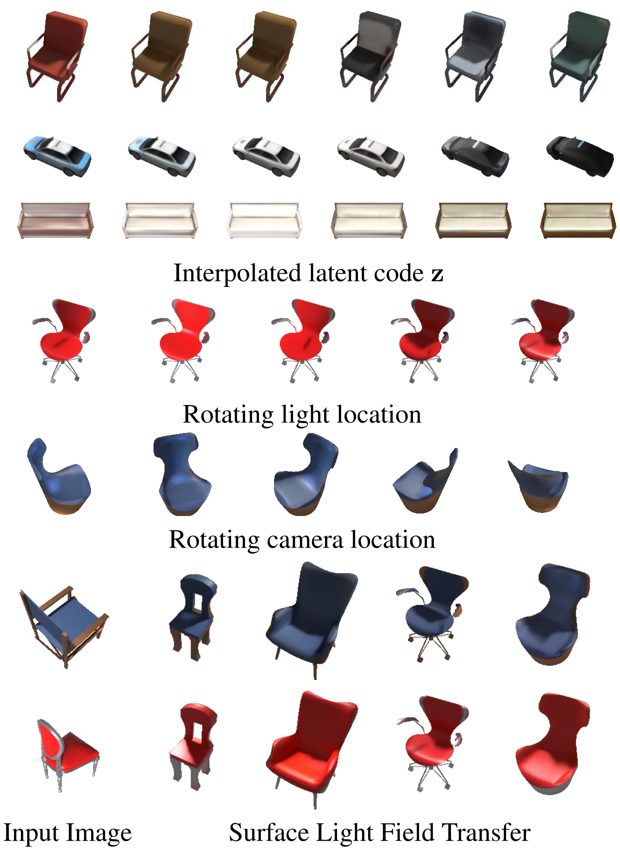

Figure 10. **Generative Model.** We show latent space interpolations (top), samples with varying light and camera locations (middle) and surface light field transfer to different shapes (bottom).

Acknowledgements This work was supported by an NVIDIA research gift. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Michael Niemeyer.

References

[1] H. Alhaija, S. Mustikovela, A. Geiger, and C. Rother. Geometric image synthesis. In *Proc. of the Asian Conf. on Computer Vision (ACCV)*, 2018. 2

[2] S. Bi, Z. Xu, K. Sunkavalli, M. Hašan, Y. Hold-Geoffroy, D. Kriegman, and R. Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2

[3] S. Bi, Z. Xu, K. Sunkavalli, D. J. Kriegman, and R. Ramamoorthi. Deep 3d capture: Geometry and reflectance from sparse multi-view images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[4] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Blender Institute, Amsterdam, 5

[5] M. Boss, V. Jampani, K. Kim, H. P. Lensch, and J. Kautz. Two-shot spatially-varying brdf and shape estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[6] A. Brock, J. Donahue, and K. Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2019. 2

[7] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *arXiv.org*, 1512.03012, 2015. 5

[8] A. Chen, M. Wu, Y. Zhang, N. Li, J. Lu, S. Gao, and J. Yu. Deep surface light fields. In *ACM Trans. on Graphics*, 2018. 1, 2, 3

[9] W. Chen, H. Ling, J. Gao, E. Smith, J. Lehtinen, A. Jacobson, and S. Fidler. Learning to predict 3d objects with an interpolation-based differentiable renderer. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 2

[10] Z. Chen and H. Zhang. Learning implicit fields for generative shape modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[11] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2

[12] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen. The lumigraph. In *ACM Trans. on Graphics*, 1996. 3

[13] P. Henzler, N. J. Mitra, , and T. Ritschel. Escaping plato's cave: 3d shape from adversarial rendering. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[14] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NIPS)*, 2017. 6

[15] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2017. 4

[16] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 1, 3

[17] J. T. Kajiya. The rendering equation. In *ACM Trans. on Graphics*, 1986. 2

[18] Y. Kanamori and Y. Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Trans. on Graphics*, 2018. 2

[19] K. Kang, C. Xie, C. He, M. Yi, M. Gu, Z. Chen, K. Zhou, and H. Wu. Learning efficient illumination multiplexing for joint capture of reflectance and shape. *ACM Trans. on Graphics*, 2019. 2

[20] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *Proc. of the International Conf. on Learning Representations (ICLR)*, 2018. 2

[21] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[22] H. Kato, Y. Ushiku, and T. Harada. Neural 3d mesh renderer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[23] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2014. 4

[24] Z. Li, K. Sunkavalli, and M. Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Proc. of the European Conf. on Computer Vision (ECCV)*, 2018. 2

[25] Y. Liao, K. Schwarz, L. Mescheder, and A. Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. *arXiv.org*, 1912.05237, 2019. 2

[26] L. Liu and I. Stamos. A systematic approach for 2d-image to 3d-range registration in urban environments. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2007. 2

[27] S. Liu, W. Chen, T. Li, and H. Li. Soft rasterizer: Differentiable rendering for unsupervised single-view mesh reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[28] S. Lombardi and K. Nishino. Reflectance and natural illumination from a single image. In A. W. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, editors, *Proc. of the European Conf. on Computer Vision (ECCV)*, 2012. 2

[29] S. Lombardi and K. Nishino. Reflectance and illumination recovery in the wild. *IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI)*, 38(1):129–141, 2016. 2

[30] M. M. Loper and M. J. Black. Opendr: An approximate differentiable renderer. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2014. 2

[31] M. Maximov, L. Leal-Taixé, M. Fritz, and T. Ritschel. Deep appearance maps. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2, 3

[32] A. Meka, C. Häne, R. Pandey, M. Zollhöfer, S. R. Fanello, G. Fyffe, A. Kowdle, X. Yu, J. Busch, J. Dourgarian,

P. Denny, S. Bouaziz, P. Lincoln, M. Whalen, G. Harvey, J. Taylor, S. Izadi, A. Tagliasacchi, P. E. Debevec, C. Theobalt, J. P. C. Valentin, and C. Rhemann. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. 2019. 2

[33] A. Meka, M. Maximov, M. Zollhöfer, A. Chatterjee, H. Seidel, C. Richardt, and C. Theobalt. LIME: live intrinsic material estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[34] A. Meka, R. Pandey, C. Haene, S. Orts-Escolano, P. Barnum, P. Davidson, D. Erickson, Y. Zhang, J. Taylor, S. Bouaziz, C. Legendre, W.-C. Ma, R. Overbeck, T. Beeler, P. Debevec, S. Izadi, C. Theobalt, C. Rhemann, and S. Fanello. Deep relightable textures - volumetric performance capture with neural rendering. In *ACM Trans. on Graphics*, 2020. 2

[35] S. Merzbach, M. Hermann, M. Rump, and R. Klein. Learned fitting of spatially varying brdfs. *Computer Graphics Forum*, 38(4), 2019. 5

[36] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 4, 5

[37] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 2, 3, 4

[38] T. Nguyen-Phuoc, C. Li, S. Balaban, and Y. Yang. Rendernet: A deep convolutional network for differentiable rendering from 3d shapes. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 1, 2

[39] T. Nguyen-Phuoc, C. Li, L. Theis, C. Richardt, and Y. Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[40] T. Nguyen-Phuoc, C. Richardt, L. Mai, Y. Yang, and N. Mitra. Blockgan: Learning 3d object-aware scene representations from unlabelled images. *arXiv.org*, 2020. 2

[41] M. Niemeyer, L. M. Mescheder, M. Oechsle, and A. Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 7

[42] M. Oechsle, L. Mescheder, M. Niemeyer, T. Strauss, and A. Geiger. Texture fields: Learning texture representations in function space. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 1, 2, 5, 6

[43] J. J. Park, P. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[44] K. Park, K. Rematas, A. Farhadi, and S. M. Seitz. Photoshape: Photorealistic materials for large-scale shape collections. In *ACM Trans. on Graphics*, 2018. 5

[45] J. Philip, M. Gharbi, T. Zhou, A. A. Efros, and G. Drettakis. Multi-view relighting using a geometry-aware network. 2019. 2

[46] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. A. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks, 2019. 4

[47] K. Rematas and V. Ferrari. Neural voxel renderer: Learning an accurate and controllable rendering tool. *arXiv.org*, 1912.04591, 2019. 2

[48] D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *Proc. of the International Conf. on Machine learning (ICML)*, 2014. 4

[49] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015. 5

[50] S. Saito, Z. Huang, R. Natsume, S. Morishima, A. Kanazawa, and H. Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv.org*, 1905.05172, 2019. 1, 2

[51] S. Sengupta, J. Gu, K. Kim, G. Liu, D. Jacobs, and J. Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[52] V. Sitzmann, J. Thies, F. Heide, M. Nießner, G. Wetzstein, and M. Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2

[53] V. Sitzmann, M. Zollhöfer, and G. Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems (NIPS)*, 2019. 1, 2

[54] T. Sun, J. T. Barron, Y. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi. Single image portrait relighting. *ACM Trans. on Graphics*, 2019. 2

[55] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5, 7

[56] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: image synthesis using neural textures. In *ACM Trans. on Graphics*, 2019. 1, 2, 5

[57] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. on Image Processing (TIP)*, 13(4):600–612, 2004. 6

[58] D. N. Wood, D. I. Azuma, K. Aldinger, B. Curless, T. Duchamp, D. H. Salesin, and W. Stuetzle. Surface light fields for 3d photography. In *ACM Trans. on Graphics*, 2019. 2, 3

[59] Z. Xu, K. Sunkavalli, S. Hadap, and R. Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Trans. on Graphics*, 2018. 2

[60] L. Yariv, M. Atzmon, and Y. Lipman. Universal differentiable renderer for implicit neural representations. *arXiv.org*, 2003.09852, 2020. 3

[61] H. Zhou, S. Hadap, K. Sunkavalli, and D. Jacobs. Deep single-image portrait relighting. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2019. 2

[62] Z. Zhou, G. Chen, Y. Dong, D. P. Wipf, Y. Yu, J. Snyder, and X. Tong. Sparse-as-possible SVBRDF acquisition. *ACM Trans. on Graphics*, 2016. 2

[63] J. Zhu, Z. Zhang, C. Zhang, J. Wu, A. Torralba, J. Tenenbaum, and B. Freeman. Visual object networks: Image generation with disentangled 3d representations. In *Advances in Neural Information Processing Systems (NIPS)*, 2018. 2