Single Image Neural Material Relighting

James Bieron
College of William & Mary
Williamsburg, USA
jcbieron@wm.edu

Xin Tong
Microsoft Research Asia
Beijing, China
xtong@microsoft.com

Pieter Peers
College of William & Mary
Williamsburg, USA
ppeers@siggraph.org

ABSTRACT

This paper presents a novel *neural material relighting* method for revisualizing a photograph of a planar spatially-varying material under novel viewing and lighting conditions. Our approach is motivated by the observation that the plausibility of a spatially varying material is judged purely on the visual appearance, not on the underlying distribution of appearance parameters. Therefore, instead of using an intermediate parametric representation (e.g., SVBRDF) that requires a rendering stage to visualize the spatially-varying material for novel viewing and lighting conditions, neural material relighting directly generates the target visual appearance. We explore and evaluate two different use cases where the relit results are either used directly, or where the relit images are used to enhance the input in existing multi-image spatially varying reflectance estimation methods. We demonstrate the robustness and efficacy for both use cases on a wide variety of spatially varying materials.

CCS CONCEPTS

• **Computing methodologies** → **Image-based rendering**; **Reflectance modeling**.

KEYWORDS

Spatially-varying material, Neural Relighting

ACM Reference Format:

James Bieron, Xin Tong, and Pieter Peers. 2023. Single Image Neural Material Relighting. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings (SIGGRAPH '23 Conference Proceedings), August 6–10, 2023, Los Angeles, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3588432.3591515

1 INTRODUCTION

Recovering the spatially-varying appearance of a material from a limited number of measurements is a challenging problem in computer graphics that has received significant attention in the past decade. The application of machine learning to appearance modeling [Dong 2019] enabled the recovery of plausible spatially-varying bidirectional reflectance functions (SVBRDFs) from a planar sample from as little as a single photograph. These recent advances in machine learning-driven appearance modeling can be categorized in two classes: *direct inference methods* and *neural inverse rendering methods*. Direct inference methods rely on a neural network to directly convert the input image in the desired SVBRDF parameter maps without the need for additional processing steps. Alternatively, neural inverse rendering methods perform an online optimization that matches a rendering of the recovered SVBRDF parameters to the input photograph, with one or more steps in the optimization process replaced by a learned component (e.g., using a learned optimization domain).

Estimating the spatially varying appearance from a single photograph is highly underconstrained. Both direct inference and neural inverse rendering methods learn a non-linear mapping from the space of visual material appearance to the higher dimensional space of SVBRDF parameter maps. Typically, specular reflections are not observed at every surface point, and hence both direct inference and neural inverse methods must somehow decide how to implement the non-linear mapping despite incomplete observations. Given the inherent richness of spatially varying materials, this process is ambiguous and direct inference and neural inverse rendering methods therefore aim to recover the most *plausible* SVBRDF parameter maps. Once the SVBRDF parameter maps are estimated, new visualizations of the material can be generated by effectively performing another non-linear mapping from the SVBRDF parameter space back to the visual material appearance space. Our key observation is that the plausibility of the resulting spatially varying material is judged purely on its visual appearance, and not on the distribution of the underlying SVBRDF parameter maps.

In this paper, we take a different approach to appearance modeling. Instead of learning a mapping between two different spaces, we learn how to navigate the visual appearance space directly. This has two major advantages compared to going through an intermediate SVBRDF parameter space. First, learning to navigating the visual appearance space only requires a loss defined in the same space (namely visual material appearance). In contrast, prior SVBRDF estimation methods need to balance possibly conflicting losses defined in different spaces: the SVBRDF parameter space and the visual material appearance space (i.e., parameter loss versus render loss). Second, when leveraging skip connections, prior SVBRDF methods need to translate image features to SVBRDF features. While correlated, this is not a one-to-one mapping. In contrast, our method leverages skip connections between two identical domains, avoiding the need for additional translations, resulting in a more effective propagation of information from the input photograph to the relit output image. To control the navigation, we specify the destination by providing the target view direction and point light position, resulting in a revisualization of the material present in the input photograph. This process is conceptually akin to relighting, hence we name our method *"neural material relighting"*.

SIGGRAPH '23 Conference Proceedings, August 6–10, 2023, Los Angeles, CA, USA

ACM ISBN 979-8-4007-0159-7/23/08…$15.00
https://doi.org/10.1145/3588432.3591515

We show that neural material relighting from a single photograph is able to reproduce equally or more plausible visual material appearance and that it generalizes better compared to prior SVBRDF estimation methods. Our solution builds on a poweful encoder-decoder architecture with passthrough connections, residual blocks [He et al. 2016] and highlight aware convolutions [Guo et al. 2021], as well as a novel multi-resolution injection strategy for specifying the target lighting during decoding. Our neural material relighting is trained on the INRIA-SVBRDF dataset [Deschaintre et al. 2018] using three different losses: an image similarity loss to ensure visual similarity to reference relit images, a conditional discriminator loss that promotes similarity of the material appearance between the input and relit image, and a perceptual loss to ensure that inevitable differences are perceptually plausible.

We present two use cases of our neural material relighting network. First, we demonstrate that neural material relighting can produce, given a single input photograph, plausible relit images under a novel view and point light for a wide range of spatially varying materials. These relit images can then be directly used in existing rendering systems. Second, we can also use neural material relighting to produce a set of intermediate synthetic input images for any existing multi-image SVBRDF parameter map estimation method, thereby improving reconstruction quality and extending the capabilities of SVBRDF estimation methods that require multiple input photographs to operate on a single input photograph.

2 RELATED WORK

We focus our discussion of related work on selected learning based appearance modeling approaches in relighting and SVBRDF estimation. We refer the reader to the excellent surveys by Dong [2019] on neural appearance modeling, and by Einabadi et al. [2021] on relighting.

Relighting. Relighting directly infers changes in an object's visual appearance under varying incident lighting from a set of photographs of a subject under controlled lighting conditions [Debevec et al. 2000]. Recently, with rise in popularity of machine learning methods, relighting has seen renewed interest. At a high level, learning based relighting methods can be categorized based on whether they are specifically (over)trained for relighting a single object [Bemana et al. 2020; Chen et al. 2020; Gao et al. 2020; Guo et al. 2019; Ren et al. 2015; Srinivasan et al. 2021; Zhang et al. 2021], or whether they rely on a pretrained model to relight an object from a small set of photographs, e.g., for face relighting [Meka et al. 2019; Sun et al. 2019; Yeh et al. 2022; Zhou et al. 2019], human body relighting [Kanamori and Endo 2018], and general outdoor [Griffiths et al. 2022; Philip et al. 2019] and indoor scene relighting [Philip et al. 2021; Xu et al. 2018]. Our method is most similar to the second class of methods that rely on a pretrained model to relight, in our case, a planar material sample. However, unlike the majority of the relighting methods in second class, our method is not limited to a fixed viewpoint (albeit in texture space) and features more complex variations in surface normal and surface reflectance.

SVBRDF Estimation. A popular representation of surface appearance is by means of the spatially-varying bidirectional reflectance function (SVBRDF), a collection of 2D maps that serve as the per-surface point parameters of an analytical BRDF model such as the Cook-Torrance BRDF model [1982] or GGX BRDF model [Walter et al. 2007] and a local shading frame in the form of a local surface normal. A common strategy for creating an SVBRDF is by an inverse rendering process that searches for the 2D property maps that, when rendered, best matches a series of reference photographs of a physical material exemplar.

Estimating an SVBRDF from a single photograph is an ill-conditioned problem as it has more unknowns (9 or more BRDF parameters) than knowns (3 observations) per surface point. Before the use of machine learning, robust estimation of an SVBRDF from a single photograph was only possible for a restricted class of texture-like materials [Aittala et al. 2016]. Machine learning, and in particular convolutional neural networks, made it practical to estimate plausible SVBRDFs from a single photograph. Many variants have been introduced that estimate SVBRDF-based representations under uncontrolled lighting for planar surfaces [Li et al. 2017; Martin et al. 2022; Ye et al. 2018] and complex indoor scenes [Li et al. 2020], and from a flash-photograph of a planar sample [Deschaintre et al. 2018; Guo et al. 2021; Henzler et al. 2021; Li et al. 2018a; Vecchio et al. 2021; Wen et al. 2022; Zhou and Kalantari 2021] and general objects [Li et al. 2018b; Sang and Chandraker 2020]. Our neural appearance relighting also infers surface appearance from a single flash photograph of a planar material exemplar. However, our work differs from these SVBRDF estimation methods in that we bypass the SVBRDF estimation step and directly produce a revisualization of the material for a new view and light condition. As a consequence, our training loss does not need to balance differences between property maps and the visual appearance of the material (expressed in prior work with an additional rendering loss [Deschaintre et al. 2018]), and it can better leverage information sharing via skip connections.

Of special note is the work by Sang and Chandraker [2020] who learn both SVBRDF estimation and (fixed viewpoint) relighting at the same time. However, unlike our work, Sang and Chandraker's relighting network is limited to a single fixed view, for an arbitrary object, and requires estimated SVBRDF maps (recovered jointly) as an input, and is therefore more similar to Deep Shading [Nalbach et al. 2017]. Our neural material relighting network directly operates on the input photograph and can relight for any viewpoint, albeit limited to a planar surface.

An alternative strategy to promote visual similarity of a recovered SVBRDF is to provide multiple input photographs of the material sample. Following the success of learning-based approaches in single-image SVBRDF estimation, several multi-image methods have been introduced ranging from direct inference methods [Deschaintre et al. 2019] to neural inverse rendering methods where one or more components in the optimization pipeline are replaced by a learned component [Fischer and Ritschel 2022; Gao et al. 2019; Guo et al. 2020; Ye et al. 2021; Zhou et al. 2022; Zhou and Kalantari 2022], to differentiable rendering approaches [Azinović et al. 2019; Bi et al. 2020]. Neural material relighting is complementary to these multi-image SVBRDF estimation methods, by allowing us to augment a single input photograph to a small collection of relit images that can subsequently be used as synthetic input to a multi-image method, thereby extending the range (of number of input images) on which these methods can operate.

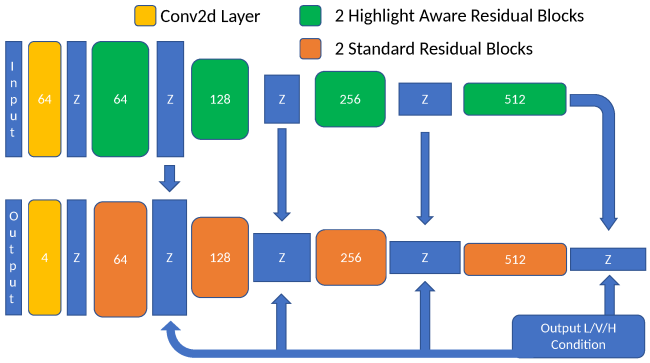

Figure 1: Summary of the neural material relighting network architecture.

3 METHOD OVERVIEW

Neural material relighting takes as input a single photograph of a planar material sample viewed from straight above and lit with a colocated point light (i.e., camera flash). We assume the input photograph is captured by a camera with a FOV of 28° and resampled to a 256×256 resolution. We deliberately train for such a narrow FOV so that images captured with a larger FOV can be easily cropped to mimic the correct FOV before resampling; the reverse (going from narrow to wide) would be more difficult. We concatenate the per-pixel corresponding z coordinate of the view/light direction to the captured input photograph (i.e., in total: 3 + 1 input channels). In addition, a 9 channel decoder-condition 'image' containing per-pixel output view, lighting, and halfway vectors is provided to control the appearance of the output. The resulting output is a *rectified* photograph where each pixel's appearance is relit based on the corresponding view and light directions in the output condition. The rectification ensures that each surface point on the material sample is mapped to the same surface location in the photograph irrespective of the view direction, thereby alleviating the network from learning the projective mapping and avoiding foreshortening issues (i.e., spatial low pass filtering) at grazing view angles. Note that while we provide a light source direction, we do so per-surface point, hence we can specify point lighting (i.e., converging directions) at different distances without actually encoding the distance, consequently neural material relighting models point lighting but without the distance-squared fall-off; this can be easily added afterwards by scaling each pixel appropriately.

4 NETWORK ARCHITECTURE

Our network follows an encoder-decoder architecture with residual blocks [He et al. 2016] as the core processing units. The encoding stage consists of a regular 2D convolution layer with a 7×7 kernel that expands the 4 channel input image to a 64 channel latent vector. This latent vector is then processed by two residual blocks (with ReLU activations and batch normalization) in which the 2D convolution layers have been replaced with *highlight aware* convolution layers [Guo et al. 2021] to reduce burn-in. This highlight aware double residual block is repeated 4 times, preceeded by a downsample layer for all but the first double residual block.

The decoder network also follows a similar structure of four double residual blocks, but with *regular convolutions* since highlight aware convolution are designed for encoding only. We also include skip connections for efficient information sharing between each latent vector in the encoder to the corresponding decoder latent vectors. In addition, we provide the output condition (9 channels containing the output view, lighting, and halfway vector) to the decoder by downsampling the output condition to the appropriate size and concatenating it to *each* latent vector in the decoder. Finally, we add a tanh activation to the 2D convolution layer at the end that reduces the final latent vector from 64 channels to a 3 channel relit output image.

While each of the components that comprise our network architecture are known (i.e., residual blocks, highlight aware convolutions, and skip connections), the combination is novel with respect to appearance modeling, as is the manner in which the output condition is injected in the decoding process. Figure 1 summarizes the network architecture.

Discussion: Neural description versus SVBRDF parameter maps. At first glance, one could argue that the encoded neural description is the equivalent of a neural encoding of the SVBRDF parameter maps. For classic SVBRDF parameter maps, one could consider the renderer to be equivalent to a (fixed) decoder. However, our neural material description is a not a generative description due to the inclusion of skip connections between the encoder and the decoder/neural renderer for efficient information sharing. Consequently, unlike an SVBRDF representation, the latent encoding does not need to contain all the small scale details needed to exactly reproduce the appearance as these are injected by the skip connections. In addition, an SVBRDF representation is limited by the expressiveness of the model and the relation between the different properties can be ambiguous for a given observation (i.e., different SVBRDF property maps can result in the same appearance for a given view and light condition). In contrast, our neural description is agnostic to the underlying physical interpretation and only encodes information relevant for producing a relit appearance from the input photograph.

5 LOSS & TRAINING

We train our neural material relighting network on the INRIA SVBRDF dataset [Deschaintre et al. 2018]. In order to avoid bias towards the particularities inherent to the INRIA SVBRDF dataset, we compose our test set of 40 unique spatially-varying materials as an even mixture of test materials from the INRIA SVBRDF test set and SVBRDFs from other sources. Note we only use renderings of the materials during training, and the network never sees any of the SVBRDF property maps. We use a combination of three losses: a data loss $\mathcal{L}_d$, a perceptual loss $\mathcal{L}_p$, and a conditional discriminator loss $\mathcal{L}_c$. The data loss $\mathcal{L}_d$ is the resolution-normalized L_1 error between the reference and relit image. $\mathcal{L}_d$ guides the network training towards reproducing the appearance of the training set. However, this loss only considers per-pixel losses, and it tends to bias the solution towards blurred highlights (a small shift in highlight edge produces a large error, hence blurring on average minimizes the error due to misalignments). We address this issue by including a VGG perceptual loss [Johnson et al. 2016] to drive the solution

towards a plausible relit image and a discriminator loss $\mathcal{L}_c$ *conditioned* on the input image that judges whether the relit image is the same material as the input image. The discriminator consists of a resolution dependent number of 2D convolution layers (kernel size 4, stride 2, and 8 output channels) with a leaky ReLU activation function and a batch-norm layer. An adaptive max-pooling followed by a fully connected layer completes the discriminator network. We use 5 layers for 128×128 and 6 for 256×256 resolution inputs. The condition images are fed together with the input image to the network. The final loss is then:

$$\mathcal{L} = \lambda_d \mathcal{L}_d + \lambda_p \mathcal{L}_p + \lambda_c \mathcal{L}_c, \tag{1}$$

with $\lambda_d = 1$, $\lambda_p = 0.01$, and $\lambda_c = 0.025$.

We found that the sampling of the training exemplars in each training batch is critical for obtaining a well behaved neural relighting network. Unlike SVBRDF-based methods, appearance relighting cannot rely on the extrapolation capabilities of the underlying model. We use a batch size of 16, and each batch consists of 4 different materials. Each material is relit and viewed from 4 different view/light combinations with none of the directions shared between the materials. Hence, in each training batch the network sees 4 materials and 16 different view/light combinations. However, the sampling of view and lighting also matters. A key challenge for the network is to learn how to 'move' the highlight. Hence, a majority of training samples should feature a highlight. However, the network also needs to learn how to relight diffuse surface reflectance, thus some portion of training samples should should be highlight free. Equally important is that the light source varies in distance, so that the network learns to take in account the relative difference in light directions between neighboring surface points. To address these concerns, we follow the procedure outlined below, assuming that the material sample forms a square with corners at -1 and $+1$ in x and y coordinates:

(1) We select a camera position p_{cam} by uniformly sampling a point on a hemisphere with radius 4 surrounding the sample.
(2) Next, we pick where the center of a highlight p_h should be (if the surface was perfectly flat) by sampling a normal distribution with a standard deviation of two and centered at a uniformly sampled position c_d on the material surface:

$$c_d = \mathcal{U}(-1, +1) \tag{2}$$
$$p_h = \mathcal{N}(c_d, 2) \tag{3}$$

This ensures that a significant portion of the highlights will appear on the sample (due to the mean always lying on the surface) with a non-negligible chance that it falls outside (due to the standard deviation being the size of the sample).
(3) We compute the main (non-normalized) light direction by reflecting the vector from the camera to the ideal highlight center around the z-axis: $l = reflect(p_h - p_{cam}, z)$. Note: $|l| = |p_h - p_{cam}| \approx 4$.
(4) Finally, we compute the point light source *position* p_l by scaling the resulting main lighting vector l by 1 plus the absolute value of a normal distributed random value with mean zero and standard deviation of two: $p_l = p_h + l(1 + |\mathcal{N}(0, 2)|)$. This ensures that the network generalizes to different light source distances.

We train the conditional discriminator network, using a regular mean square loss, simultaneously with the neural material relighting network by providing, for each batch, a positive exemplar (a relit image from the same material) and a negative exemplar (a relit image from another material). Including a negative sample is important as we want the discriminator to learn to decide whether the relit image depicts the same material as the input. Including only positive training examples would result in a network that essentially ignores the input image. In our implementation, we use the same set of reference relit images as used in the batch for training the neural material relighting network since it contains both positive as well as negative exemplars (given a reference material).

We exploit the full convolutional architecture of our network to improve robustness and to speed up the training process. We first train our network on 128×128 crops from the INRIA SVBRDF dataset, after which we refine the network weights on 256×256 crops. Because our network is fully convolutional, we can use the same weights when doubling the resolution without needing to add extra layers. However, the number of convolution layers in the discriminator varies with resolution, and therefore, when changing training resolution, we train the discriminator again from scratch.

6 RESULTS

We implemented and trained our network in PyTorch with the following hyperparameters: learning rate of 10^{-4}, variational beta of 0.5, and a learning rate decay of 1% every 10,000 batches. We train for 500,000 batches at 128×128 resolution on a single Nvidia RTX A40, followed by a refinement for an additional 150,000 batches at 256×256 resolution distributed over four Nvidia RTX A40. Once trained, material relighting takes 15ms on an Nvidia RTX 2070ti. We validate our neural appearance relighting network for two use cases: *direct relighting* and as an *input augmentation* step for existing SVBRDF estimation methods. All results in this section are at 256×256 resolution.

6.1 Direct Relighting

For the first use-case the output of the neural relighting is directly used. Hence, the quality of the relit images is of primary concern. Figure 2 and 3 show visual comparisons with respect to the reference and with respect to prior single image SVBRDF methods (we include comparisons to [Zhou and Kalantari 2021] and [Gao et al. 2019] (using [Zhou and Kalantari 2021] as starting point) for both figures plus [Deschaintre et al. 2018] for Figure 2 and [Guo et al. 2021] for Figure 3) for a selection of materials not used in training. The results for most prior methods [Deschaintre et al. 2018; Gao et al. 2019; Zhou and Kalantari 2021] are generated with the authors' provided trained networks; there does not exist a publicly available solution for the two-stream highlight aware network [Guo et al. 2021] and the corresponding results in Figure 3 are computed from the SVBRDF property maps from Gou et al.'s supplemental material. Note, all results are rectified (i.e., shown in texture space); the azimuthal view angle used for relighting is listed in the first column. Observe how neural appearance relighting is able to reproduce challenging spatially varying specular reflections (e.g., discontinuous highlights in the 2nd material, and the specular reflections on ridges

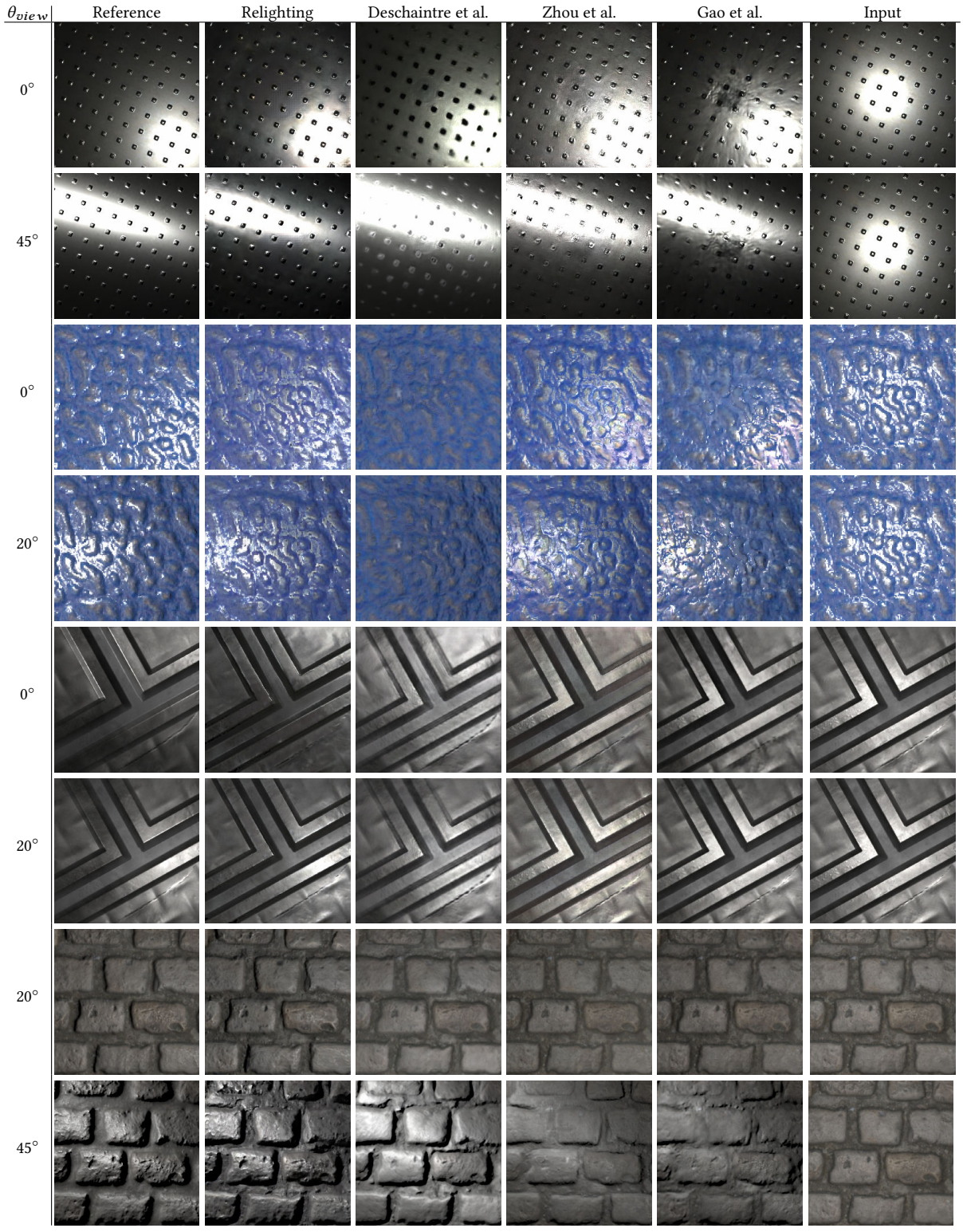

Figure 2: Qualitative comparison of neural relit results (at 256×256 resolution) against three prior SVBRDF estimation methods for a variety of materials. The first column list the azimutal angle of the view angle for relighting.

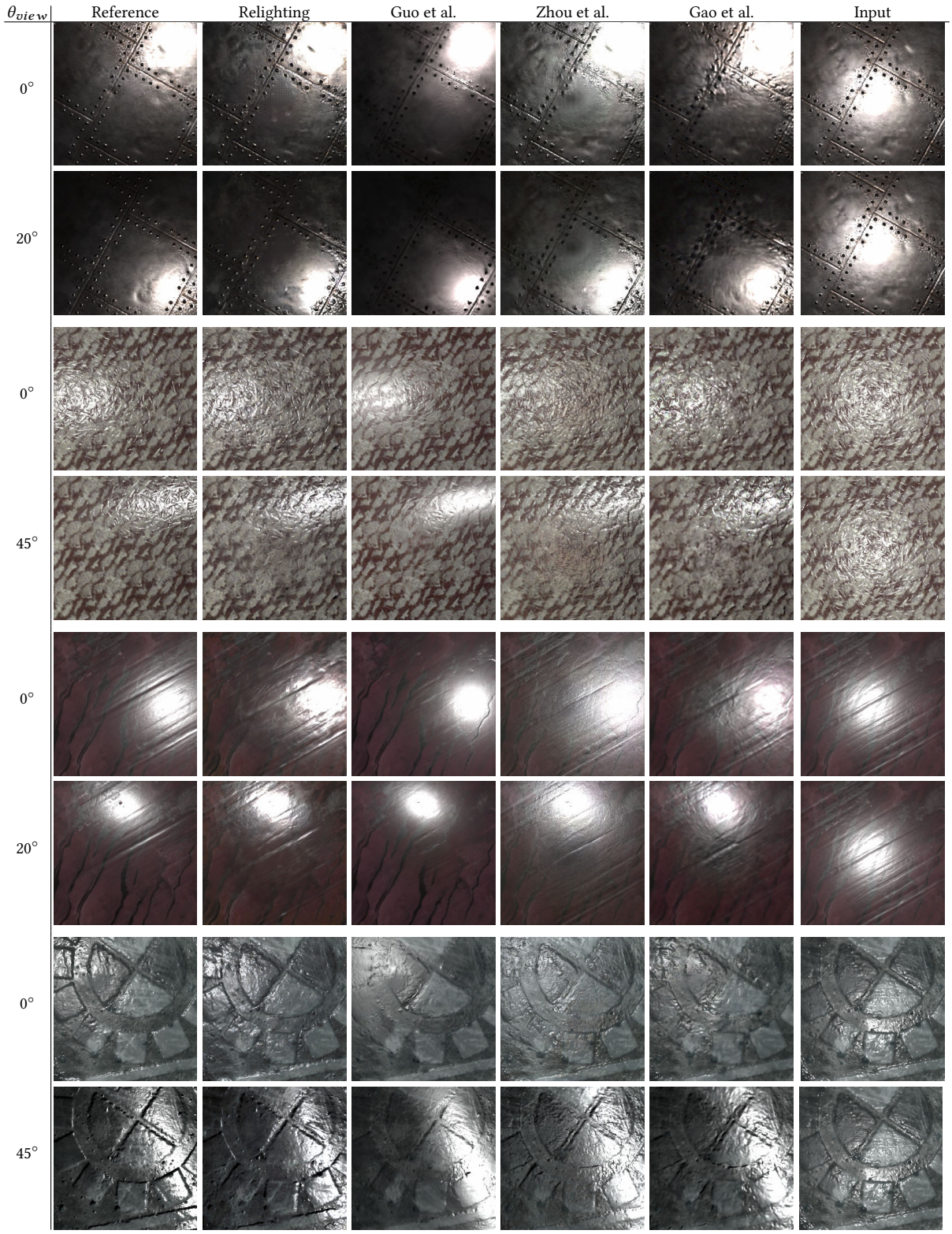

Figure 3: Qualitative comparison of neural relit results (at 256×256 resolution) against three prior SVBRDF estimation methods for a variety of materials. The first column list the azimutal angle of the view angle for relighting.

Table 1: Quantitative comparison of LPIPS errors on renderings at 256×256 resolution obtained with neural material relighting and representative prior single-image SVBRDF estimation methods.

Method	Material Relighting	Our SVBRDF	Deschaintre et al. [2018]	Zhou et al. [2021]	Gao et al. [2019]
LPIPS	**0.1902**	0.1966	0.2713	0.2375	0.2323

in the 3rd example), handle specular reflections for small geometrical details (e.g., the highlights on the "nubs" in the first example are correctly oriented), and model large scale normal variations and foreshortening (e.g., the bricks and tiles in the last examples for both result figures). While there are clear differences with the reference relit image due to the highly ambiguous nature of single image relighting, neural material relighting produces overall visually more plausible results with less artifacts than the four prior methods. In addition, Table 1 summarizes LPIPS [Zhang et al. 2018] errors averaged over a test set of 40 materials rerendered for 3 view directions (0°, 20°, and 45°) and 49 light directions chosen such that for each view the highlights are regularly distributed in the texture space. These errors also confirm that neural material relighting produces more plausible relit images than prior work.

Despite not explicitely enforcing similarity between neighboring views or light directions, neural material relighting produces relit images that change smoothly with varying view and light. We refer to the supplementary video for a demonstration.

Finally, the robustness of neural appearance relighting outside the training set is further demonstrated in Figure 4 on photographs captured by a cellphone. While no reference photographs under novel lighting are available, the results show plausible relit images. To better gauge the generalization capabilities our neural material relighting outside the training set, we also include a comparison to the SVBRDF method of Zhou and Kalantari [2021]. Note how material relighting is able to produce more plausible results, ranging from retaining the fine-scale details without over or underestimating the specular highlight (1st row and 4th row), more plausible recreating of appearance effects due to normal variations (2nd and 3rd row), and reproducing complex highlights (5th row). We refer to the supplementary video for a comparison under varying view and lighting that further reinforces the plausibility differences, as well as a comparison to deep inverse rendering [Gao et al. 2019].

6.2 Input Augmentation

A second use case of neural material relighting is to augment a single input photograph to a set of relit synthetic photographs that are subsequently used as an input to a multi-image SVBRDF estimation methods. We demonstrate this use case with deep inverse rendering [Gao et al. 2019] (using the estimate of the adversarial SVBRDF estimation method [Zhou and Kalantari 2021] as starting point) for which we synthesize 5 new relit input images with varying light positions ensuring that each images' highlight is contained within the camera view from a single captured photograph (Figure 5). The SVBRDFs recovered from the augmented input shows more plausible rerenderings with less artifacts than from a single input photograph. Over all 40 test materials, the augmented results yield a significantly lower LPIPS error (0.2021) compared to without augmentation (0.2323).

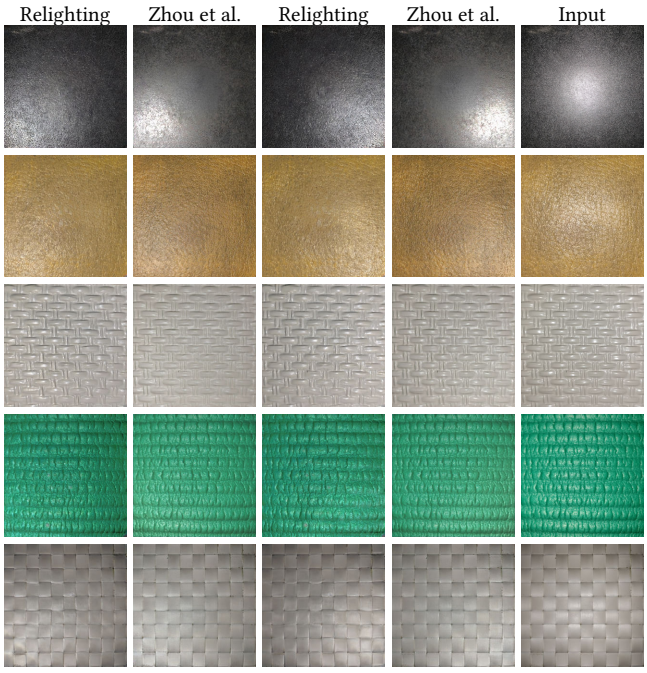

Figure 4: Neural material relighting for two light source positions (1st and 3rd column) on photographs captured with a handheld camera (last column) and compared to [Zhou and Kalantari 2021] (2nd and 4th column).

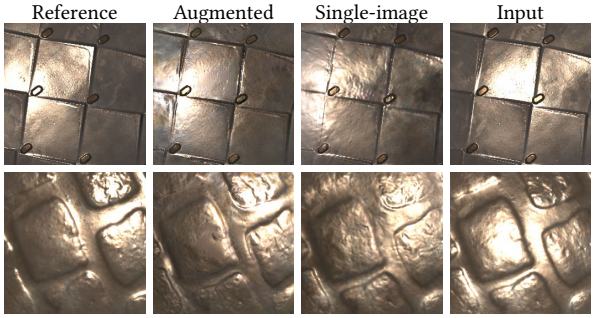

Figure 5: Neural material relighting can be used to generate synthetic inputs to multi image SVBRDF estimation methods. In this example we generated 5 synthetic input images for Deep Inverse Rendering [Gao et al. 2019] yielding more plausible revisualizations than from a single input image.

7 ABLATION STUDY

We perform a number of ablation and sensitivity experiments to provide further insight and to validate the design of the network architecture. All ablation experiments are performed on images and networks trained at 128×128 resolution.

Table 2: Quantitative comparison of LPIPS errors on renderings at 128×128 resolution for variations in the model: residual vs. regular convolution blocks, and highlight aware (HA) vs. standard convolutions.

Backbone Convolution	Residual HA	Residual Standard	Regular HA	Regular Standard
LPIPS Err.	**0.1735**	0.1792	0.1774	0.1871

Table 3: Quantitative comparison of LPIPS errors on renderings at 128×128 resolution for models trained with variations in input and out specification.

Variant	Our Method	Remove Input Z	Remove Output H	Pass Output at Encoder
LPIPS Err.	**0.1735**	0.1785	0.1828	0.1941

Network Architecture. We validate the importance of using both residual blocks and highlight aware convolutions by comparing results from a network where the residual blocks are replaced by regular convolution blocks, with and without highlight aware convolutions (Table 2). Numerically, the combined residual blocks and highlight aware convolutions provide the lowest average LPIPS error. In general, we find that the residual blocks are able to better reproduce the shape of the specular highlights, while the highlight aware convolutions reduce burn-in artifacts.

Loss Terms. Figure 6 shows the impact of each loss term on the relighting quality. Using only the data loss results in blurred highlights. Adding the perceptual loss, sharpens the highlights but it fails to capture the correct highlight details (e.g., the ridge highlight on the right). Finally, adding the discriminator yields the highest quality highlights.

Input/output Specification. Our neural material relighting network takes as additional input (besides the photograph) the z coordinate of the lighting direction. Since the input lighting is always the same, one could argue that this extra input is unnecessary. The errors in Table 3 show that without this information the network does not perform as well. Inclusion of the z component serves a similar role as the so-called "coord-conv" trick [Liu et al. 2018] to help the network learn the location and statistics of specular highlights, due to the strong correlation with the z coordinate; only providing the (x, y) coordinate (cf. coord-conv trick) fails to capture this correlation.

The output conditions passed to the relighting network do not only contain the view and lighting directions per surface point, but also the halfway vector. The corresponding error in Table 2 confirms that including the halfway vector improves the result quality.

Currently, neural material relighting concatenates the (spatially scaled) output conditions to each feature vector in the decoder. However, a more common strategy is to concatenate the conditions to the input. The corresponding error in Table 3 shows that this does not yield a good result due to two reasons. First, during training the encoder might learn erroneous correlations between the output conditions and the input photograph; injecting the output conditions only in the decoder cleanly separates material encoding from neural rendering. Second, the network would need to learn to correctly downsample the output direction images (including renormalization) for effective use at each level in the decoder.

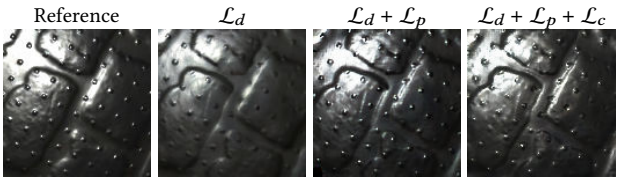

Figure 6: Impact of the different loss terms: data loss $\mathcal{L}_d$, VGG perceptual loss $\mathcal{L}_p$, and the conditional discriminator loss $\mathcal{L}_c$.

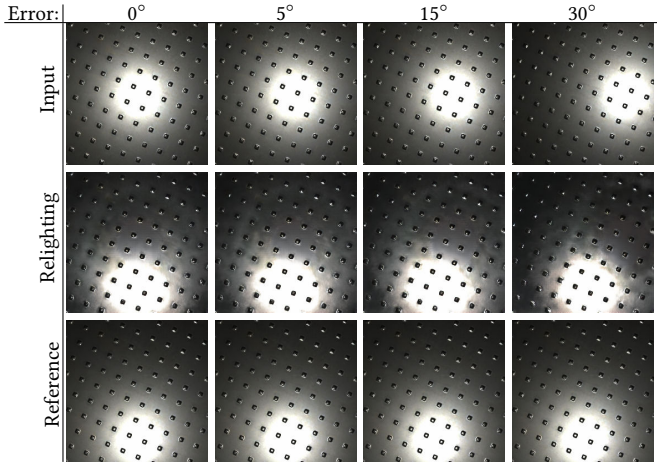

Figure 7: Neural material relighting is robust to deviations of the light source position for up to 5° from a colocated configuration.

Input Robustness. The easiest way to obtain a photograph of a material lit by a colocated light source is by handheld capture with a cell phone. However, the camera flash light is not exactly colocated due to physical constraints. Figure 7 shows that our method is robust for deviations of up to 5° between the light source and sensor. While still plausible, at larger deviations the quality degrades gracefully.

Output Lighting Generalization. While our method is trained for a limited range of variations in light source distance, neural material relighting generalizes well to light source positions outside this range. Figure 8 shows that neural material relighting produces plausible relit results when placing the light source at 0.5, 2, 10, and 100 units distance from the material sample.

Relighting vs. SVBRDF Estimation. The results in Figure 2 and Table 1 show that neural material relighting can produce more plausible relit results than existing single-image SVBRDF estimation methods. To better understand the difference between neural material relighting and SVBRDF estimation, we perform an additional ablation experiment where we use the same architecture, loss terms,

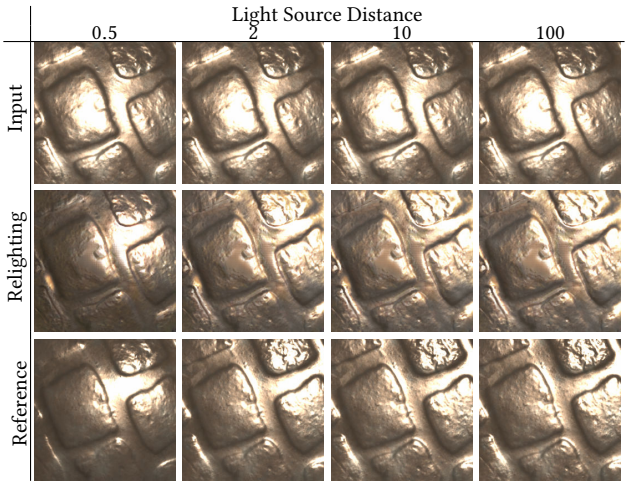

Figure 8: Neural material relighting is robust to moving the point light away from the positions seen during training.

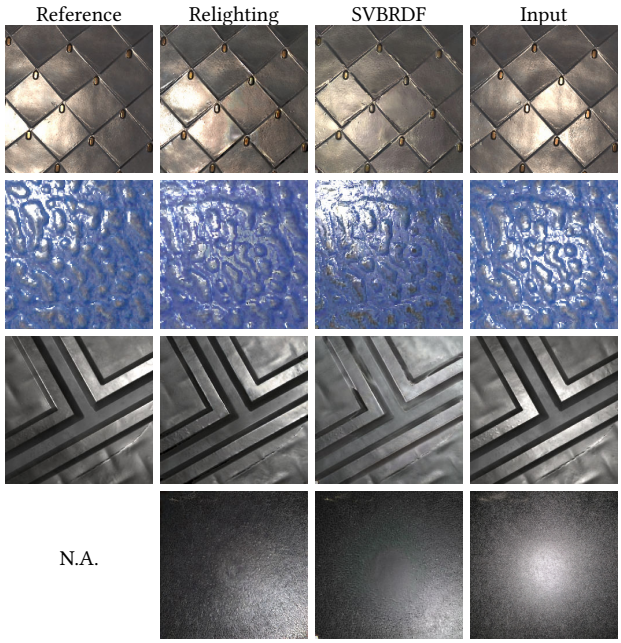

Figure 9: SVBRDF estimation using a similar architecture tends to generalize less well to materials that are challenging or that are from outside the training set.

and training procedure, but output SVBRDF property maps instead of relit images, and include an additional L_1 SVBRDF property map loss. Figure 9 shows a comparison between both methods on a variety of synthetic materials, as well as a captured material (without a reference); the LPIPS errors over the test set are also listed in Table 1. While the LPIPS errors over the test set are closer to neural material relighting than those from prior SVBRDF estimation methods, we find that the SVBRDF estimation version of our network is less stable to train due to the conflicting loss functions, and that it produces an artifact (i.e., 'stuck' pixels) in the lower left corner; this artifact is very noticeable but not captured by LPIPS errors. Furthermore, we observe that for materials from the INRIA-SVBRDF *test* set both methods perform well. However, for challenging materials more dissimilar from the training set, we observe that material relighting tends to generalize better. This is also confirmed by comparing the average LPIPS error on the 20 non-INRIA test materials (0.2431 for the SVBRDF estimation versus 0.2248 for neural material relighting).

8 LIMITATIONS

Our neural material relighting is not without limitations. Due to the convolutional nature of the decoding network, our material relighting network is limited by the view/light combinations seen during training. Currently, our relighting network is only trained for planar materials and it cannot handle mapping the material over a curved surface. Possibly including such cases during training could help extend the capabilities, although we expect a more powerful or deeper network architecture might be needed. Alternatively, we could also subdivide the curved surface in patches and relight the material for each patch. Similarly, our neural material relighting network does not support relighting a single selected pixel without relighting the whole material. This makes our method less suited for ray tracing based rendering systems. An interesting avenue for future research would be to replace the decoder by an MLP that takes a pixel coordinate as additional input, and that outputs the relit pixel value. Furthermore, our lighting network currently is only able to relight from a single point light, making relighting with environment maps expensive (cf. classic image-based relighting [Debevec et al. 2000]). Extending our method to direct relighting with environment lighting is another avenue for future research. Finally, neural material relighting can fail to reproduce correct highlights if the input does not contain many specular highlights. Furthermore, despite the highlight aware convolutions, severe oversaturation can still lead to burn-in (Figure 10). However, existing SVBRDF methods typically also fail on these challenging materials.

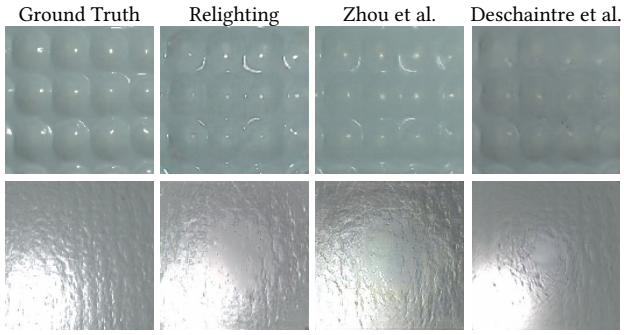

Figure 10: Top: Neural material relighting requires a sufficient number of pixels featuring a specular highlight in the input photograph to correctly reproduce highlights. Bottom: Despite the highlight aware convolutions, severe oversaturation still causes burn-in.

9 CONCLUSION

In this paper we presented neural material relighting, a novel strategy for appearance modeling that from a single photograph produces a relit image of the material without going through an intermediate SVBRDF estimation and rendering step. Our learning based method features an encoder-decoder network architecture with residual blocks and highlight aware convolutions trained with a combination of three loss terms: a data loss, a perceptual loss, and a conditional loss. Besides directly using the relit materials as is, neural material relighting can also be used to create synthetic input images to drive multi-image SVBRDF estimation methods thereby extending the conditions under which these methods can operate.

ACKNOWLEDGMENTS

This research was supported in part by NSF grant IIS-1909028.

REFERENCES

Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance modeling by neural texture synthesis. *ACM Trans. Graph.* 35, 4 (2016).

Dejan Azinović, Tzu-Mao Li, Anton Kaplanyan, and Matthias Nießner. 2019. Inverse Path Tracing for Joint Material and Lighting Estimation. In *CVPR*.

Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. 2020. X-Fields: Implicit Neural View-, Light- and Time-Image Interpolation. *ACM Trans. Graph.* 39, 6, Article 257 (nov 2020).

Sai Bi, Z. Xu, K. Sunkavalli, David Kriegman, and Ravi Ramamoorthi. 2020. Deep 3D Capture: Geometry and Reflectance from Sparse Multi-View Images. In *CVPR*.

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 *CVRP*. 5598–5609.

Robert L. Cook and Kenneth E. Torrance. 1982. A Reflectance Model for Computer Graphics. *ACM Trans. Graph.* 1, 1 (1982), 7–24.

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 (SIGGRAPH '00)*. 145–156.

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-image SVBRDF capture with a rendering-aware deep network. *ACM Trans. Graph.* 37, 4 (2018).

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible SVBRDF Capture with a Multi-Image Deep Network. *Comp. Graph. Forum* 38, 4 (2019).

Yue Dong. 2019. Deep appearance modeling: A survey. *Visual Informatics* 3, 2 (2019), 59–68.

Farshad Einabadi, Jean-Yves Guillemaut, and Adrian Hilton. 2021. Deep Neural Models for Illumination Estimation and Relighting: A Survey. *Comp. Graph. Forum* 40, 6 (2021), 315–331.

Michael Fischer and Tobias Ritschel. 2022. Metappearance: Meta-Learning for Visual Appearance Reproduction. *ACM Trans. Graph.* 41, 6, Article 245 (nov 2022).

Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2020. Deferred Neural Lighting: Free-viewpoint Relighting from Unstructured Photographs. *ACM Trans. Graph.* 39, 6 (2020).

Duan Gao, Xiao Li, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. 2019. Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images. *ACM Trans. Graph.* 38, 4 (2019).

David Griffiths, Tobias Ritschel, and Julien Philip. 2022. OutCast: Single Image Relighting with Cast Shadows. *Comp. Graph. Forum* 43 (2022).

Jie Guo, Shuichang Lai, Chengzhi Tao, Yuelong Cai, Lei Wang, Yanwen Guo, and Ling-Qi Yan. 2021. Highlight-Aware Two-Stream Network for Single-Image SVBRDF Acquisition. *ACM Trans. Graph.* 40, 4, Article 123 (2021).

Kaiwen Guo, Peter Lincoln, Philip L. Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Ryan Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul E. Debevec, and Shahram Izadi. 2019. The relightables: volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.* 38, 6, Article 217 (2019).

Yu Guo, Cameron Smith, Miloš Hašan, Kalyan Sunkavalli, and Shuang Zhao. 2020. MaterialGAN: Reflectance Capture Using a Generative SVBRDF Model. *ACM Trans. Graph.* 39, 6, Article 254 (2020).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*. 770–778.

Philipp Henzler, Valentin Deschaintre, Niloy J. Mitra, and Tobias Ritschel. 2021. Generative Modelling of BRDF Textures from Flash Images. *ACM Trans. Graph.* 40, 6, Article 284 (2021).

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *ECCV*. 694–711.

Yoshihiro Kanamori and Yuki Endo. 2018. Relighting Humans: Occlusion-Aware Inverse Rendering for Full-Body Human Images. *ACM Trans. Graph.* 37, 6, Article 270 (Dec. 2018).

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. *ACM Trans. Graph.* 36, 4 (2017).

Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *CVPR*. 2475–2484.

Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. 2018a. Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image. In *ECCV*. 74–90.

Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. 2018b. Learning to reconstruct shape and spatially-varying reflectance from a single image. *ACM Trans. Graph.* 37, 6 (2018).

Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. 2018. An Intriguing Failing of Convolutional Neural Networks and the CoordConv Solution. In *NeurIPS*. 9628–9639.

Rosalie Martin, Arthur Roullier, Romain Rouffet, Adrien Kaiser, and Tamy Boubekeur. 2022. MaterIA: Single Image High-Resolution Material Capture in the Wild. *Comp. Graph. Forum* 41, 2 (2022), 163–177.

Abhimitra Meka, Christian Häne, Rohit Pandey, Michael Zollhöfer, Sean Ryan Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul E. Debevec, Christian Theobalt, Julien P. C. Valentin, and Christoph Rhemann. 2019. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Trans. Graph.* 38, 4, Article 77 (2019).

Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, Hans-Peter Seidel, and Tobias Ritschel. 2017. Deep Shading: Convolutional Neural Networks for Screen Space Shading. *Comp. Graph. Forum* (2017).

Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. 2019. Multi-view Relighting Using a Geometry-aware Network. *ACM Trans. Graph.* 38, 4, Article 78 (July 2019).

Julien Philip, Sébastien Morgenthaler, Michaël Gharbi, and George Drettakis. 2021. Free-Viewpoint Indoor Neural Relighting from Multi-View Stereo. *ACM Trans. Graph.* 40, 5, Article 194 (2021).

Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. 2015. Image based relighting using neural networks. *ACM Trans. Graph.* 34, 4, Article 111 (2015).

Shen Sang and M. Chandraker. 2020. Single-Shot Neural Relighting and SVBRDF Estimation. In *ECCV*.

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 *CVPR*. 7495–7504.

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E. Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Trans. Graph.* 38, 4, Article 79 (2019).

Giuseppe Vecchio, Simone Palazzo, and Concetto Spampinato. 2021. SurfaceNet: Adversarial SVBRDF Estimation From a Single Image. In *ICCV*.

Bruce Walter, Stephen R. Marschner, Hongsong Li, and Kenneth E. Torrance. 2007. Microfacet Models for Refraction through Rough Surfaces. In *EGSR*. 195–206.

Tao Wen, Beibei Wang, Lei Zhang, Jie Guo, and Nicolas Holzschuch. 2022. SVBRDF Recovery from a Single Image with Highlights Using a Pre-trained Generative Adversarial Network. *Comp. Graph. Forum* 41, 6 (2022).

Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep Image-based Relighting from Optimal Sparse Samples. *ACM Trans. Graph.* 37, 4, Article 126 (July 2018).

Wenjie Ye, Yue Dong, Pieter Peers, and Baining Guo. 2021. Deep Reflectance Scanning: Recovering Spatially-varying Material Appearance from a Flash-lit Video Sequence. *Comp. Graph. Forum* 40, 6 (2021), 409–427.

Wenjie Ye, Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2018. Single Image Surface Appearance Modeling with Self-augmented CNNs and Inexact Supervision. *Comp. Graph. Forum* 37, 7 (2018), 201–211.

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to Relight Portrait Images via a Virtual Light Stage and Synthetic-to-Real Adaptation. *ACM Trans. Graph.* 41, 6, Article 231 (nov 2022).

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T Barron, Ravi Ramamoorthi, and William T Freeman. 2021. Neural Light Transport for Relighting and View Synthesis. *ACM Trans. Graph.* 40, 1 (2021), 1–17.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. 2019. Deep Single-Image Portrait Relighting. In *ICCV*. 7193–7201.

Xilong Zhou, Milos Hasan, Valentin Deschaintre, Paul Guerrero, Kalyan Sunkavalli, and Nima Khademi Kalantari. 2022. TileGen: Tileable, Controllable Material Generation and Capture. In *SIGGRAPH Asia 2022 Conference Papers*. Article 34.

Xilong Zhou and Nima Khademi Kalantari. 2021. Adversarial Single-Image SVBRDF Estimation with Hybrid Training. *Comp. Graph. Forum* (2021).

Xilong Zhou and Nima Khademi Kalantari. 2022. Look-Ahead Training with Learned Reflectance Loss for Single-Image SVBRDF Estimation. *ACM Trans. Graph.* 41, 6, Article 266 (nov 2022).