MERLiN: Single-Shot Material Estimation and Relighting for Photometric Stereo

Ashish Tiwari[1], Satoshi Ikehata[2], and Shanmuganathan Raman[1]

[1] Indian Institute of Technology Gandhinagar, Gujarat, India
{ashish.tiwari, shanmuga}@iitgn.ac.in
[2] National Institute of Informatics, Tokyo, Japan
sikehata@nii.ac.jp

Abstract. Photometric stereo typically demands intricate data acquisition setups involving multiple light sources to recover surface normals accurately. In this paper, we propose MERLiN, an attention-based hourglass network that integrates single image-based inverse rendering and relighting within a single unified framework. We evaluate the performance of photometric stereo methods using these relit images and demonstrate how they can circumvent the underlying challenge of complex data acquisition. Our physically-based model is trained on a large synthetic dataset containing complex shapes with spatially varying BRDF and is designed to handle indirect illumination effects to improve material reconstruction and relighting. Through extensive qualitative and quantitative evaluation, we demonstrate that the proposed framework generalizes well to real-world images, achieving high-quality shape, material estimation, and relighting. We assess these synthetically relit images over photometric stereo benchmark methods for their physical correctness and resulting normal estimation accuracy, paving the way towards single-shot photometric stereo through physically-based relighting. This work allows us to address the single image-based inverse rendering problem holistically, applying well to both synthetic and real data and taking a step towards mitigating the challenge of data acquisition in photometric stereo.

Keywords: Intrinsic decomposition · Single-image relighting · Photometric Stereo

1 Introduction

Photometric stereo [43] plays a pivotal role in 3D reconstruction, surface analysis, and material recovery. By analyzing an object's appearance under multiple illumination conditions, it infers per-pixel surface normals. It directly extends to applications such as quality control, industrial inspection, medical imaging, cultural heritage preservation, and robotics, to name a few. However, despite its utility, photometric stereo encounters several challenges that constrain its applicability and accuracy in real-world scenarios.

One significant challenge lies in the complexity of data acquisition, which often demands carefully orchestrated setups involving controlled lighting environments and precise calibration procedures. Due to practical constraints such

as time, cost, and equipment limitations, it is often infeasible to exhaustively sample the entire space of possible lighting configurations. As a result, the acquired dataset may not sufficiently cover all relevant lighting variations, leading to incomplete or inaccurate surface reconstructions.

Key Questions. (a) *Can we leverage the advancements in deep learning research to generate differently illuminated images?* Image relighting has been addressed from various perspectives using deep learning. One stream of works [11, 20, 22, 34, 46, 54] includes the use of convolutional neural networks (CNN), while the other stream of works [23,38,44,45,52] is based on neural radiance fields (NeRF) [29] for relighting and material estimation. The NeRF-based methods have extensively improved the relighting results. However, they rely on multiple calibrated images and lengthy per-scene optimization. Interestingly, CNN-based approaches have achieved relighting in a feed-forward manner from a sparse set of views (as few as one). Initially, works like [11, 46, 54] modeled relighting as an image-to-image translation task where a CNN can be trained with one or more images and novel lighting as input to generate the relit targets.

(b) *Do these synthesized images always guarantee the physical correctness of the relit images?* The image-based relighting methods often produce images that are not physically meaningful because the images may "appear" perceptually realistic even if the underlying shape and material parameters deviate from being physically correct. Physically correct image relighting fundamentally demands an in-depth understanding of geometry, material properties, and illumination. The challenge is more compounded when addressing objects of diverse textures and reflectance properties since these elements interact in complex ways. A stream of works [20, 22, 34] perform relighting through intrinsic parameter estimation. The relighting is performed either via a neural network [34] or through an in-network rendering layer simulating a specific BRDF model [22]. Such an approach offers better controllability and editability of scene parameters. Furthermore, global illumination plays a vital role in physical plausibility. While works like [22] consider global illumination effects due to indirect light bounces, most other CNN-based methods have resorted to direct illumination.

(c) *How can we validate the physical correctness of these relit images?* Interestingly, photometric stereo itself offers a solution. As shown in Figure 1 (c), two sets of images with similar perceptual fidelity can result in widely different normals. Such *"perceptually-correct physically-incorrect"* images fail to generate correct normal estimates through the photometric stereo. Therefore, one could also evaluate the relit images by measuring the performance on the photometric stereo.

Key Ideas and Contributions The following are the key contributions to address the aforementioned observations.

(i) We propose a physically-based global illumination-aware deep network, called MERLiN - <u>M</u>aterial <u>E</u>stimation and <u>ReLighting N</u>etwork, to estimate spatially varying bidirectional reflectance distribution function (svBRDF) parameters such as diffuse albedo, normal, depth, and specular roughness) and jointly perform relighting through a single image. We perform relighting through esti-
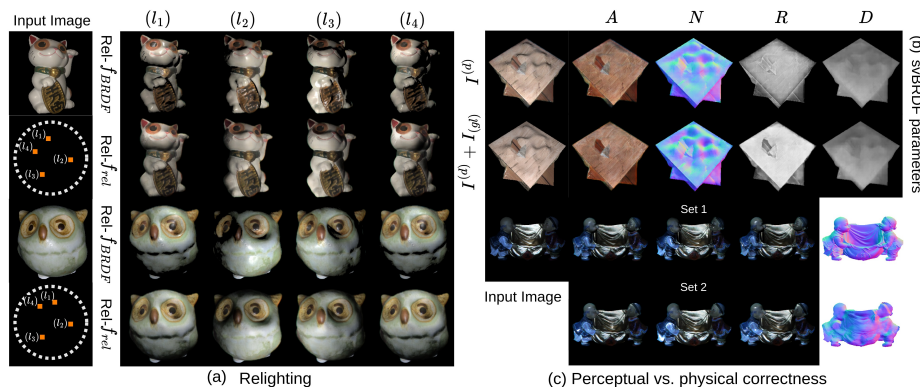
Fig. 1: (a) Effect of relighting (under four light positions l_1, l_2, l_3, l_4) through BRDF rendering layer f_{BRDF} and neural network (f_{rel}), (b) Effect of training with direct (top) vs global illumination (bottom) images. The estimated normals without global illumination are flattened and produce brighter albedo (top). (c) Two different sets of perceptually similar images with different underlying normals maps.

mated image intrinsics and learn the complex relationship between appearance and lighting. The joint learning allows the network to simulate a physically-based rendering process [34] and ensures that the relit images are close to their real counterparts.

(ii) We validate the physical correctness of the relit images through existing photometric stereo benchmarks and compare the accuracy of normal estimation using the relit images and their real counterparts. This way, we take a step towards addressing photometric stereo from a single image via image-based relighting.

2   Related Work

Shape and Material Estimation. Several deep learning frameworks have been designed for the inverse problem over indoor [35] and outdoor [51] scenes for material recognition [2] and estimation [28], reflectance maps extraction [33], surface appearance recovery [19], normal and depth estimation [5]. Others assume a specific class of objects, such as faces [37,39,53] or near planar surfaces [1,19,21] for shape and reflectance recovery. Further, some methods apply to images captured through smartphones [22,34] and a few on in-the-wild images [42]. The ill-posed nature of the problem, especially using a single image, demands more labeled training data for ground supervision. While Li *et al.* [19] leverage the appearance information embedded in unlabeled images of spatially varying materials to self-augment the training process, primarily to reduce the amount of required labeled training data, The authors in [4, 21, 22] train CNNs to regress svBRDF and surface normal using in-network rendering to provide additional supervision during training. However, Sang *et al.* [34] use CNNs to jointly estimate svBRDF

parameters and perform relighting with a single image. While Yi *et al.* [49] use differently lit images during training, they perform single image-based inference during test time. However, they use an off-the-shelf network to remove specularities before performing intrinsic decomposition. In another approach, Wimbauer *et al.* [42] use the priors learned by other networks to aid in the shape and material reflectance. Interestingly, different works have addressed inverse rendering in different flavors, such as intrinsic image decomposition [20, 26] or specularity removal [36, 47] or surface normal estimation [22, 34] or even through photometric stereo [16, 17, 40].

Image Relighting. Image-based relighting has been approached through an image translation perspective [11, 54]. Several other methods [46, 49] have used a sparse set of multiple images for relighting through CNNs. While single image-based relighting is highly ill-posed, methods like [53] have performed relighting over facial images. Others have performed relighting either using an in-network rendering layer [22] or training a relighting network jointly with intrinsic parameter estimation [34] from a single image. In this work, we follow the paradigm of [22, 34] with two important differences - (i) a single-stage network (in contrast to their cascaded network design) with (ii) an in-network global illumination handling for joint material estimation and relighting from a single image allowing us to better model the shape, illumination, and appearance dependencies. While [22] considers global illumination effects through a cascaded CNN, their training is not end-to-end with svBRDF estimation. Specifically, they train the global illumination network separately to estimate second and third light bounce images, given the first bounce image. On the contrary, we use a single-stage global illumination network to perform end-to-end training with material estimation and relighting.

Photometric stereo. Earlier to photometric stereo, Shape from Shading (SfS) [8, 9] methods were proposed to reconstruct shape from single images captured under calibrated illumination, though they usually assume Lambertian reflectance [12]. Later, they were extended to arbitrary shapes and reflectance under known natural illumination [32]. However, due to the severely ill-posed nature of the problem, researchers incorporated multiple images under different lightings for shape estimation and addressed Photometric Stereo. Several methods [6, 10, 16, 17, 40, 43] with the help of meaningfully curated deep learning-based architectures have been used to recover shape, BRDF material, and lighting by generally solving an optimization problem using multiple images of a scene captured under different lighting conditions and/or from multiple viewpoints. However, acquiring these multiple images with controlled lighting for either training or inference is tedious and challenging to apply to objects under arbitrary illumination. Some methods like [41] and [40] have performed photometric stereo through relighting in supervised and self-supervised manner, respectively, using one and two images as input during inference, but multiple images (one or two at a time) for training. Others have performed inverse rendering for photometric stereo in a self-supervised manner [14, 16, 17]. Interestingly, none of the existing works have demonstrated the use of relit images to solve photometric stereo. We take
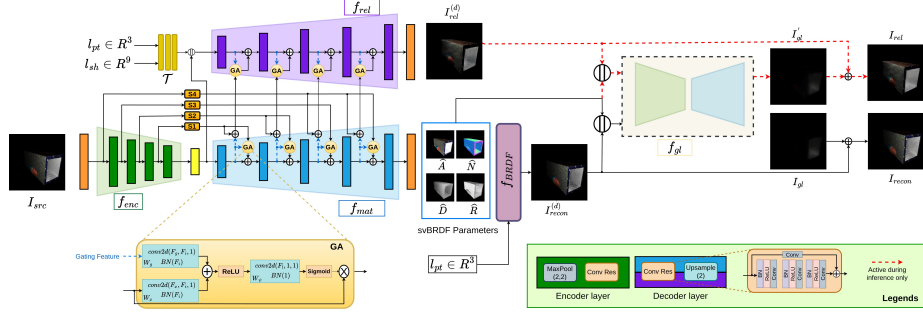
Fig. 2: The proposed framework for single image-based (I_{src}) svBRDF estimation $(\widehat{A}, \widehat{N}, \widehat{D}, \widehat{R})$ and relighting (I_{ref}). The design of the encoder and decoder of the *global illumination network* (f_{gl}) is the same as f_{enc} and f_{inv_dec}, respectively. The superscript (d) and subscript (gl) represent the direct and indirect illumination, respectively. S1-S4 are residual skip connections.

the first step towards photometric stereo through single image-based material estimation and relighting. Our attempt can also be viewed as a bridge between shape from shading [8,9] and photometric stereo [43].

3 Method

Objective: Given a single image of an object under point and/or uncontrolled environment illumination and target lighting, we would like to first estimate four svBRDF parameters - diffuse albedo (A), normal (N), depth (D), and roughness (R) and relight the image under target lighting. Further, we want to use multiple different relit images as input to the two photometric stereo methods [10,24] to evaluate the surface normal estimates and validate the physical correctness of the relit images.

3.1 Material Estimation and ReLighting Network (MERLiN)

We propose a single-stage attention-guided convolutional neural network, called MERLiN, for joint material estimation and relighting from a single image. The network consists of one shared encoder and two decoders, one each for svBRDF estimation and relighting. The skip connections among the encoder and decoders are used for feature sharing across different layers. The design of MERLiN is inspired by the hourglass networks [30, 48] that are well known for hierarchical feature learning and multi-scale processing over different tasks. They have also been applied to inverse rendering [50] and relighting [53] tasks.

As per Figure 2, let us consider an image I_{src} (multiplied with the binary mask) as the input to the encoder f_{enc}.

svBRDF Estimation. Consider a set of features extracted by the encoder as Z_{enc} *i.e.*, $Z_{enc} = f_{enc}(I_{src}, M)$. These features are passed to the *material*

decoder f_{mat} to obtain diffuse albedo $(\widehat{A})$, roughness $(\widehat{R})$, surface normal $(\widehat{N})$, and depth $(\widehat{D})$ jointly rather than independently, such that the following holds.

$$\widehat{A}, \widehat{R}, \widehat{N}, \widehat{D} = f_{mat}(Z_{enc}) \tag{1}$$

The associated skip connections are combined at the respective layer in the decoder. We use only a single decoder to model correlations between the object's shape and material. Such a design has fewer parameters and offers faster runtime speed than the existing cascaded designs [22, 34].

Feature fusion through attention gating. Näively combining the features from the skip connections with the decoder features at respective scales leads to poor results (see Table 1-IDs 1 and 2), primarily due to the underlying redundancies and noise in the skip connections. Therefore, we adopt the attention-gating mechanism of [31] for feature fusion. The information extracted from the coarse scale in the decoder is used as a gating signal to disambiguate irrelevant and noisy responses in skip connections. Such an approach also captures local and global effects such as surface roughness, textures, light intensity fall-off, and specular regions to better model the light interaction with surfaces for joint material estimation and relighting.

Image Reconstruction. To validate the correctness of the estimated intrinsic parameters, we reconstruct the input image through the rendering layer following the microfacet BRDF model [7,13]. Given the diffuse albedo (A), specular roughness (R), normals (N), and depth (D) along with l, v, and h being the light direction, viewing angle, and the half-angle between them, one can easily render the image $I^{(d)}$ as per Equation 2.

$$I^{(d)} = \frac{1}{D^2} \cdot f_{BRDF}(A, R, N, D, l, v) \tag{2}$$

Here, superscript d indicates the image generated under direct illumination, and $1/D^2$ accounts for the light fall-off. The BRDF (f_{BRDF}) can be characterized as per Equation 3.

$$f_{BRDF}(A, N, R, D, l, v) = \frac{A}{\pi} + \frac{\widetilde{M}_f(h, R)\widetilde{F}(v, h)\widetilde{G}(l, v, h, R)}{4(N \cdot l)(N \cdot v)} \tag{3}$$

Here, $\widetilde{M}_f(h, R)$, $\widetilde{F}(v, h)$, and $\widetilde{G}(l, v, h, R)$ are the microfacet distribution, Fresnel, and geometry term. A detailed description of the BRDF model is provided in the supplementary material.

Global Illumination. Incorporating global illumination is crucial for relighting and rendering, especially over intricate shapes with complex material and reflectance such as high specularities or glossiness. Several existing works on material capture [21, 34, 50] and photometric stereo [17, 25, 41] do not explicitly model global illumination effects such as indirect illumination and inter-reflections. While a prior work [22] has considered global illumination, it trained a two-stage cascaded network separately from the BRDF estimation network. However, we train a single-stage global illumination network along with

the BRDF estimation network. The end-to-end learning allows for better co-guidance among global illumination network and material decoder to produce better results. We predict the combined indirect illumination across multiple light bounces instead of modeling individual light bounces, as in [18]. The global illumination network (f_{gl}) design is similar to the encoder-decoder framework described earlier. The output of the rendering layer $I^{(d)}$ (direct illumination) along with estimated intrinsic parameters $(\widehat{A}, \widehat{R}, \widehat{N}, \widehat{D})$ is fed to the network and produces the residual image I_{gl} which when combined with the $I^{(d)}$ yields the final image I with global illumination effects. The residual image is expected to capture the energy contained in higher-order light bounces. Equation 4 describes the image formation with global illumination.

$$I = I^{(d)} + I_{gl} \text{ s.t. } I_{gl} = f_{gl}(I^{(d)}, \widehat{A}, \widehat{R}, \widehat{N}, \widehat{D}) \tag{4}$$

Interestingly, as shown in Figure 1 (b), the network trained with direct lighting only predicts brighter diffuse albedo and flattened normals when evaluated on images with indirect lighting, which aligns with observation in [3,22]. Therefore, we train our network over images with global illumination.

Relighting. We explore two ways of relighting - one through a physically-based rendering based on predicted BRDF (Rel-f_{BRDF}) and the other through a CNN-based decoder (Rel-f_{rel}).

(a) Rel-f_{BRDF}. One way to relight is to use the estimated BRDF parameters and directly render the image under arbitrary target lighting l_{tar} through a BRDF model, as described in Equation 2. Such an approach ensures that relighting explicitly considers the physical plausibility of the intrinsic parameters to better model the global effects, such as specularities and light fall-off. While this approach would generate images under direct illumination, we use the global illumination network (trained on direct illumination images) to obtain the indirect illumination effects and incorporate them for obtaining physically plausible and visually realistic relit images.

(b) Rel-f_{rel}. Another approach is to train a relighting network with material estimation jointly. The features from the encoder Z_{enc} along with the target lighting information l_{tar} is passed through the *relighting decoder* f_{rel} to generate the image I_{rel} relit under target lighting. The skip connections from the encoder and *material decoder* are combined with the *relighting decoder* at the respective scale through the gated-attention mechanism, such that the Equation 5 holds.

$$I_{rel}^{(d)} = f_{rel}(Z_{enc}, Z_{mat}, \mathcal{T}(l_{tar})) \tag{5}$$

Here, $\mathcal{T}$ is a lighting encoder comprising three MLPs that takes the lighting vector as input. The lighting vector could be a 3×1 vector for a point or directional light and a 9×1 vector representing the spherical harmonic coefficients for the environmental illumination. While considering both point and environment lighting, we concatenate both the lighting vectors before feeding to the lighting encoder $\mathcal{T}$. Interestingly, the bidirectional connections across the two decoders allow for better learning of correlated information across two different but related

tasks of inverse rendering and relighting. Additionally, it provides additional supervision for inverse parameter estimation guided by expected consistency in the relit images. The relighting decoder generates images under direct illumination during training. However, during inference, the output of the relighting decoder $I_{rel}^{(d)}$ is passed through the trained global illumination network to infer the higher-order light bounce image I_{gl} and finally combined to obtain the final image $I_{rel} = I_{rel}^{(d)} + I_{gl}$ with global illumination effect.

3.2   Training Details

We train MERLiN over a large synthetic dataset proposed by [22] that contains BRDF parameters and images under point and environment lighting. It contains images under camera-co-located near-field point lighting, *i.e.*, $l_{pt} = [0, 0, 0]$ and object placed at $[0, 0, -1]$) [3] and environment lighting (l_{sh}) represented by 9 SH coefficients per color channel. There are three separate point light images, one for the direct component ($I_{pt}^{(d)}$) and two for subsequent light bounces each (that we combine together to obtain a single image I_{gl}), and one image under environment lighting I_{env} per object. We train MERLiN under two settings: (A) point lighting and (B) point + environment lighting. For setting (A), we consider one point light image such that $I_{src} = I_{pt}^{(d)} + I_{gl}$ with the direct component $I^{(d)} = I_{pt}^{(d)}$, global illumination effects (I_{gl}), and target lighting $l_{tar} = l_{pt}$. For setting (B), we consider $I_{src} = (I_{pt}^d + I_{env}) + I_{gl}$ under point + environment lighting, such that $I_{pt}^{(d)} + I_{env}$ is the direct component and $l_{tar} = [l_{pt}, l_{sh}]$. Moreover, the dataset lacks images under different lighting directions that are needed for relighting. Therefore, we follow [34] to render target images under random point light positions from the frontal hemisphere in an online manner using f_{BRDF} (see Equation 2) and generate ground truth for supervision, allowing the model to learn from more samples under different lights. However, relighting is performed only for the direct component since I_{gl} is only available for $l_{pt} = [0, 0, 0]$.

   Loss function. We use L2 loss to supervise intrinsic components, image reconstruction, and relighting. Consider $\widehat{Y}$ as the estimate of the ground truth Y. The L2 loss $\mathcal{L}$ can be described as per Equation 6.

$$\mathcal{L}_{\star} = \frac{1}{\sum_{i,j} M_{i,j}} ||(Y - \widehat{Y}) \cdot M||_2^2 \tag{6}$$

Here, $Y \in \{A, R, N, D, I_{rec}, I_{rel}\}$ and $\mathcal{L}_a$, $\mathcal{L}_r$, $\mathcal{L}_n$, $\mathcal{L}_d$, $\mathcal{L}_{rec}$, and $\mathcal{L}_{rel}$ are the L2 losses for albedo, roughness, normal, depth, reconstruction, and relighting. M represents the object mask. The final loss function is given as follows.

$$\mathcal{L} = \lambda_a \mathcal{L}_a + \lambda_n \mathcal{L}_n + \lambda_d \mathcal{L}_d + \lambda_r \mathcal{L}_r + \lambda_d \mathcal{L}_d + \lambda_{rec} \mathcal{L}_{rec} + \lambda_{rel} \mathcal{L}_{rel} \tag{7}$$

Here, $\lambda_a = \lambda_r = \lambda_d = \lambda_{rec} = \lambda_{rel} = 1.0$ and $\lambda_n = 2.0$. Additionally, we apply L2 loss over gradients of the roughness map. Mere L2 loss over roughness maps

---

[3] The dataset uses a camera-centric coordinate system with the camera at the origin and x, y, z directions correspond to u, v, d.
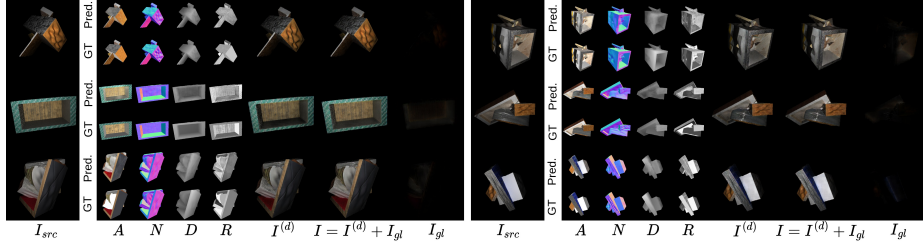
Fig. 3: Qualitative results on the test set of [22] emphasizing global-illumination effects. The superscript (d) and subscript (gl) represent the direct and global illumination components, respectively. Best viewed in PDF with zoom.

produces heavily flattened results and was observed to suppress specularities in the synthesized image. The reconstruction loss includes loss over direct and global illumination components as well.

Training Strategy. We train our network end-to-end on `NVIDIA RTX 5000` with a batch size of 64 using Adam optimizer [15] with the initial learning rate of 1×10^{-4} for image encoder and 2×10^{-4} for decoders and global illumination network and decrease it by half after every five epochs for a total of 25 epochs.

Table 1: Quantitative results over different architectural design choices for images under point light sources from the test set of [22]. Inp Img: whether the input image is under direct illumination (Img-d) or global illumination (Img-g). #InvDec: the number of *material* decoders. FS: Feature Sharing between *relighting* and *material* decoders, GA: Gated Attention, Rel: whether the relighting is through the neural network (f_{rel}) or directly through the BRDF model (f_{BRDF}), and GI: Global Illumination

ID	Design Choices						svBRDF Params (MSE $\times 10^{-2}$)				Relighting (SSIM)	
	Inp Img	#InvDec	FS	GA	Rel	GI	A	R	N	D	Rel-f_{rel}	Rel-f_{BRDF}
1	Im-g	1	✗	✗	f_{rel}	✗	6.154	18.071	4.681	1.958	0.697	0.719
2	Im-g	1	✓	✗	f_{rel}	✗	5.943	17.156	4.617	1.932	0.682	0.724
3	Im-g	1	✗	✓	f_{rel}	✗	5.519	15.277	3.975	1.751	0.701	0.757
4	Im-g	1	✗	✗	f_{rel}	✓	5.614	14.485	3.887	1.713	0.746	0.789
5	Im-g	1	✗	✓	f_{BRDF}	✓	4.723	11.277	3.925	1.632	-	0.844
6	Im-d	1	✓	✓	f_{rel}	✗	4.517	10.113	3.872	1.509	0.764	0.793
7	Im-g	1	✓	✓	f_{rel}	✗	4.162	9.681	3.406	1.462	0.798	0.859
8	Im-g	4	✓	✓	f_{rel}	✓	**3.781**	8.891	3.325	1.012	0.827	0.892
9	Im-g	1	✓	✓	f_{rel}	✓	3.787	**8.267**	**3.311**	**0.975**	0.819	**0.894**

4   Experimental Evaluation

In this section, we compare the performance of MERLiN with benchmark methods [22, 34] on svBRDF estimation and relighting over synthetic and real data. Further, we demonstrate the closeness between the performance of photometric stereo benchmarks [10, 24] evaluated on relit images and their real counterparts.

Table 2: Quantitative comparison of svBRDF estimation (MSE $\times 10^{-2}$) and relighting (SSIM) of MERLiN with Li *et al.* [22] and Sang *et al.* [34] over images under point light global illumination from the test set of [22].

Method	A	R	N	D	Relighting
Li *et al.* [22]	4.868	19.431	3.822	1.505	0.884
Sang *et al.* [34]	3.856	12.781	3.459	1.471	0.872
MERLiN (Ours)	**3.787**	**8.267**	**3.311**	**0.975**	**0.894**

4.1   Ablation Studies

Table 1 demonstrates the effect of several design choices quantitatively by evaluating the mean squared error obtained over the test set of [22]. We observe that feature sharing across the two decoders of intrinsic parameter estimation and relighting performs better than without feature sharing (compare IDs 1 and 2, Table 1). The same applies to components like gated attention (see IDs 1 and 3) and global illumination (see IDs 1 and 4). Interestingly, we observe that joint training helps improve svBRDF estimation, ensuring more physical correctness of inverse parameters (lower MSE for svBRDF params in ID 5 and 9). Note there is no feature sharing in experiment ID 3 since there is no decoder for relighting. However, as shown in Figure 1 (a), the images rendered using f_{BRDF} are still better at capturing the global illumination effects such as specularities (higher SSIM under Rel-f_{BRDF} across all the experiments). Therefore, we show the relighting results generated through f_{BRDF} for all the experiments from hereon. The global illumination and gated attention perform better in tandem than excluding any of them (compare IDs 7 and 9). We also observed that under similar settings, using one decoder for svBRDF parameter estimation offers near-close performance compared to four explicit decoders (one for each parameter) except for the albedo (IDs 8 and 9). This marginal under-performance could be acceptable for reduced network size and higher run-time speed. Moreover, taking images with global illumination as input produces far better results and generalizes well to real images than images under direct illumination (see IDs 6 and 9 and Figure 1: Row 3). Firstly, images under only direct illumination seldom exist in the real world. Moreover, they offer limited information when dealing with arbitrary shapes and materials.

4.2   Quantitative Results on svBRDF Estimation, Reconstruction, and Relighting

We evaluate and compare MERLiN with the two closest benchmark methods [22] and [34] over the test set of [22] in Table 2. We obtained significantly improved svBRDF parameter estimation results over both methods, along with image reconstruction and relighting. It is important to note that while [22] has been trained on global illumination images, [34] considers direct illumination without explicitly considering global illumination effects. However, both frameworks are evaluated on images with global illumination. As a result, we see reduced
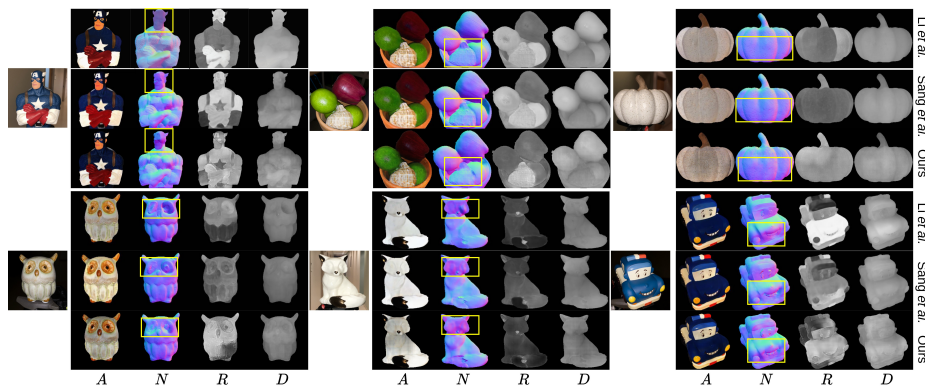
Fig. 4: Qualitative comparison of svBRDF parameters: albedo (A), normal (N), roughness (R), and depth (D) among MERLiN and methods [34] and [22]. Differences can be observed in the marked regions across different svBRDF parameters.

relighting performance of [34] compared to [22]. Interestingly, our in-network end-to-end training of the global illumination layer proves better than the ad-hoc training of the global illumination network in [22] over relighting.

4.3 Qualitative Results

We perform extensive qualitative evaluation over synthetic and real data, especially due to the lack of ground truth for quantitative comparison. Figure 3 shows the results over the test set of [22]. While the estimated svBRDF parameters are close to ground truth, the network also reasonably captures the global illumination effects. The proposed residual design does not explicitly model inter-reflections from surface points that are not visible to the camera. Instead, it only approximates the true global illumination by operating in image space and learning from inter-reflections in the training data. Figure 4 shows the results over real images. We observe that MERLiN produces better inverse parameters than [22] and [34] just from a single-stage architecture. Since physically plausible relighting is the key to our work, we compare the relit images across four objects in Figure 5. While [34] fails to model sharp surface specularities, mainly due to its inability to handle global illumination explicitly. Moreover, this observation aligns with the finding that neural network-based relighting (even with joint training) suffers in handling the global illumination effects but helps in better inverse parameter estimation. The relighting using f_{rel} in MERLiN (see Figure 1 (a)) produces similar results as that of [34]. Moreover, [22] models the specularities a little better, but it spreads the specularities over a larger area. Overall, MERLiN obtains better relighting results exhibiting high photorealism consistent with the underlying material parameters. Figure 6 shows estimated inverse parameters and relighting under environment lighting by varying the
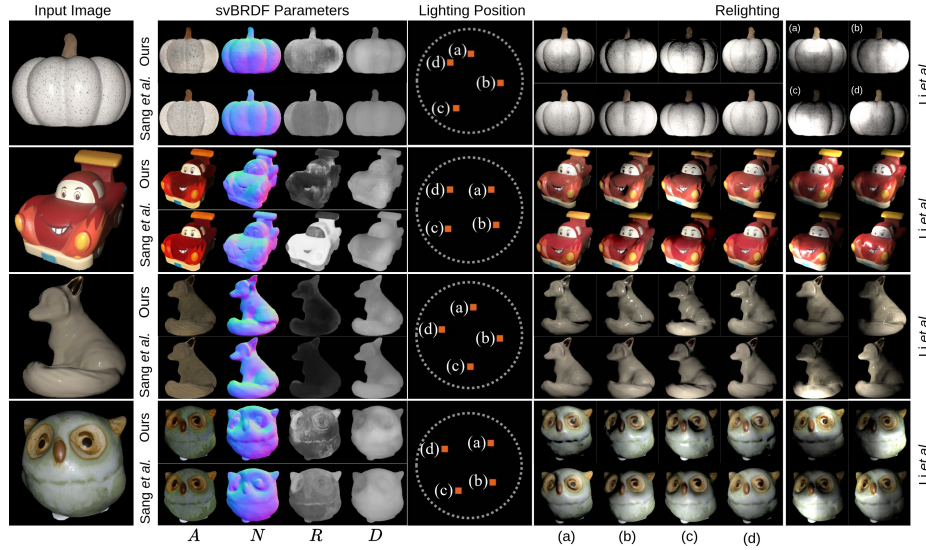
Fig. 5: Qualitative evaluation of the relit images generated through MERLiN, [34], and [22] under point lighting over the test dataset of real images.

point light and the environment map. We observe that the network generates highly realistic images under arbitrary environments.
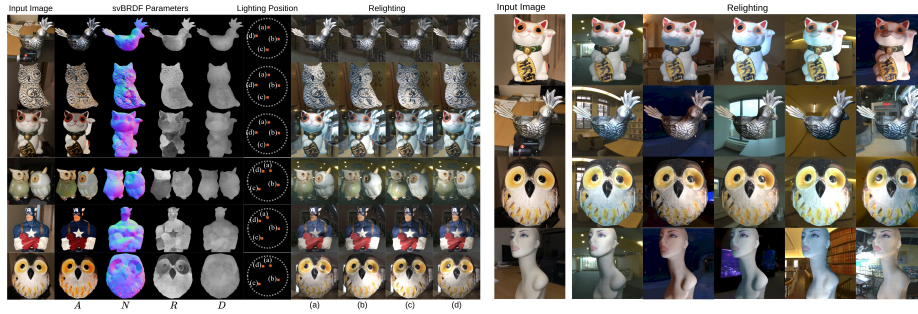
5   Photometric Stereo through Relit Images

Once we have synthesized the images under arbitrary lighting from a single input image, we feed them through the two classic benchmark photometric stereo methods - Fast-NFPS [24] and SDM-UniPS [10] to evaluate the estimated normal maps. We examine their performance over the LUCES dataset [27] - a photometric dataset under near-field lighting. Since MERLiN is trained on images with the perspective camera and near-field lights, evaluation over Fast-NFPS using synthetically relit images from the LUCES dataset is best suited for us.

The rationale behind evaluating photometric stereo over relit images is to answer three questions. (a) Are the relit images physically correct? (b) Are the normal estimates using multiple relit images better than those from a single image? (c) How close are the results when compared to their real counterparts?

In an attempt to answer these questions, we perform a quantitative comparison of the photometric stereo performance using the Fast-NFPS method [24] over relit images generated by MERLiN, Sang *et al.* [34], and Li *et al.* [22]. Table 3 reports the mean angular error (MAE) in degrees over the estimated normals. Specifically, we select an image under near frontal lighting (closest to the camera-light collocated setup) from the LUCES dataset and generate 50 images under lighting from the frontal hemisphere using all the three methods MERLiN, [34], and [22]. For each method, we randomly select 32 images from
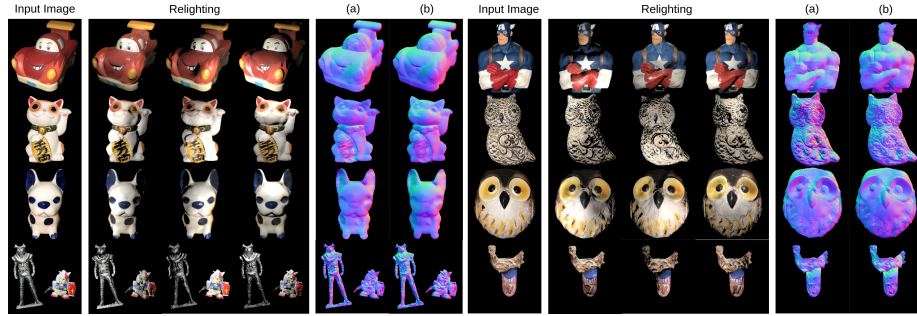
Fig. 6: Relighting through MERLiN under arbitrary environment lightings.
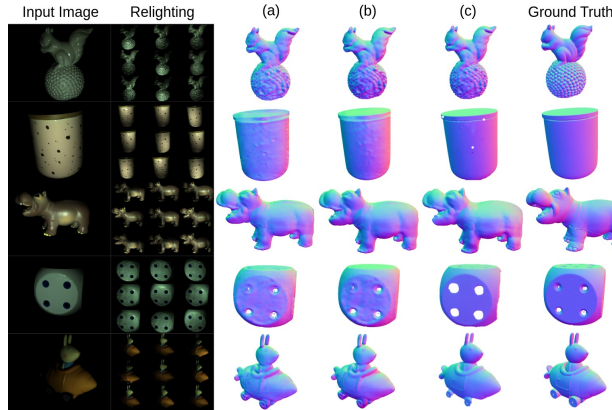
Fig. 7: Qualitative results on photometric stereo over LUCES dataset [27] We compare (a) single image-based normals obtained by the material decoder of MERLiN and the normals estimated by Fast-NFPS [24] through (b) relit images by MERLiN and (c) real 32 images. Note that the relit images are generated by MERLiN using a single image.

the set of 50 images over 5 different runs and pass them through Fast-NFPS under an uncalibrated setting to observe the final MAE. Our quantitative analysis shows that while Fast-NFPS achieves lower MAE over real images compared to relit images generated by MERLiN, the performance surpasses that of normal estimates derived from a single image through MERLiN and the other relighting methods [22,34]. Furthermore, we observe that Sang *et al.* [34] performs well on relatively flat object surfaces but produces higher MAE over complex objects like Buddha, House, and Squirrel, owing to its inability to handle the cast shadows and indirect illumination arising out of underlying surface variations. We also show the qualitative results in Figure 7 comparing the single image normals with normals through relit and real images. We observe that the multiple relit images offer better, if not the same, normal estimates compared to the single image. We also evaluate the SDM-UniPS performance qualitatively on relighted images using the samples from real test data of [22,34]. SDM-UniPS is an interesting choice as it learns global lighting context and is agnostic to any physical lighting model. Figure 8 shows the qualitative comparison of normals estimated by

Fig. 8: Qualitative results on photometric stereo over real test dataset [34]. We compare (a) single image-based normals obtained by the material decoder of MERLiN and the normals estimated by SDM-UniPS [10] through (b) relit images by MERLiN.

Table 3: Mean angular error (MAE) over the normals estimated through a single image, and sets of real and relit images under 32 different point lighting from the LUCES dataset [27] over Fast-NFPS [24] method.

Input	Rel. Method	Bell	Ball	Buddha	Bunny	Die	Hippo	House	Cup	Owl	Jar	Queen	Squirrel	Bowl	Tool	Average
Single Image	-	12.03	10.75	21.26	12.02	9.51	11.23	40.16	19.68	17.62	9.37	20.93	19.94	12.79	21.59	17.06
32 Relit Images	Sang *et al.* [34]	10.09	9.52	19.17	12.69	9.21	10.08	39.42	19.59	17.29	9.79	22.19	19.67	11.96	19.29	16.43
	Li *et al.* [22]	10.33	9.89	18.96	12.03	10.04	10.11	36.88	19.34	16.17	10.51	21.31	19.32	12.23	19.77	16.21
	MERLiN (Ours)	9.51	9.12	18.27	**11.71**	9.12	**10.02**	36.91	19.27	16.97	9.82	20.18	19.05	11.98	19.31	15.80
32 Real Images	-	**7.17**	**6.59**	**14.50**	11.89	**8.63**	10.64	**31.00**	**18.98**	**15.92**	**9.14**	**18.39**	**18.26**	**10.17**	**18.61**	**14.11**

a single image using MERLiN and SDM-UniPS through relit images generated using MERLiN. The photometric stereo through relit images indeed improves the normal estimation, especially evident in Figure 8: rows 2 and 3. However, we find that the method gets distracted by image semantics (picture of cherry on the can, Figure 8, last row) to provide erroneous surface normals.

6   Conclusion

In this work, we took a step towards addressing the data acquisition challenge in photometric stereo through joint material estimation and relighting from a single image. Our single-stage MERLiN network outperformed the baselines with cascaded network architectures over material estimation and relighting, offering faster run-time speed and a low memory footprint. Moreover, explicit global illumination rendering proved effective across all the experiments. Further, we evaluated photometric stereo methods over relit images synthesized from a single input image, shedding light on key questions regarding the physical correctness of relit images, the efficacy of normal estimation using multiple relit images compared to a single image, and the fidelity of results when compared to the real counterparts. It can even be applied to dynamic surface recovery, where a single instance of a dynamic surface can be analyzed under different lighting, allowing photometric stereo for dynamic surfaces.

Acknowledgements

References

1. Aittala, M., Aila, T., Lehtinen, J.: Reflectance modeling by neural texture synthesis. ACM Transactions on Graphics (ToG) **35**(4), 1–13 (2016)
2. Bell, S., Upchurch, P., Snavely, N., Bala, K.: Material recognition in the wild with the materials in context database. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3479–3487 (2015)
3. Chandraker, M.K., Kahl, F., Kriegman, D.J.: Reflections on the generalized bas-relief ambiguity. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). vol. 1, pp. 788–795. IEEE (2005)
4. Deschaintre, V., Aittala, M., Durand, F., Drettakis, G., Bousseau, A.: Single-image svbrdf capture with a rendering-aware deep network. ACM Transactions on Graphics (ToG) **37**(4), 1–15 (2018)
5. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
6. Goldman, D.B., Curless, B., Hertzmann, A., Seitz, S.M.: Shape and spatially-varying brdfs from photometric stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(6), 1060–1071 (2009)
7. Hill, S., McAuley, S., Belcour, L., Earl, W., Harrysson, N., Hillaire, S., Hoffman, N., Kerley, L., Patry, J., Pieké, R., et al.: Physically based shading in theory and practice. In: ACM SIGGRAPH 2020 Courses, pp. 1–12 (2020)
8. Horn, B.K.: Shape from shading: A method for obtaining the shape of a smooth opaque object from one view (1970)
9. Horn, B.K., Brooks, M.J.: Shape from shading. MIT press (1989)
10. Ikehata, S.: Scalable, detailed and mask-free universal photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13198–13207 (2023)
11. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
12. Johnson, M.K., Adelson, E.H.: Shape estimation in natural illumination. In: CVPR 2011. pp. 2553–2560. IEEE (2011)
13. Karis, B., Games, E.: Real shading in unreal engine 4. Proc. Physically Based Shading Theory Practice **4**(3), 1 (2013)
14. Kaya, B., Kumar, S., Oliveira, C., Ferrari, V., Van Gool, L.: Uncalibrated neural inverse rendering for photometric stereo of general surfaces. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3804–3814 (2021)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)

16. Li, J., Li, H.: Neural reflectance for shape recovery with shadow handling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16221–16230 (2022)
17. Li, J., Li, H.: Self-calibrating photometric stereo by neural inverse rendering. In: European Conference on Computer Vision. pp. 166–183. Springer (2022)
18. Li, J., Robles-Kelly, A., You, S., Matsushita, Y.: Learning to minify photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7568–7576 (2019)
19. Li, X., Dong, Y., Peers, P., Tong, X.: Modeling surface appearance from a single photograph using self-augmented convolutional neural networks. ACM Transactions on Graphics (ToG) **36**(4), 1–11 (2017)
20. Li, Z., Snavely, N.: Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In: Proceedings of the European conference on computer vision (ECCV). pp. 371–387 (2018)
21. Li, Z., Sunkavalli, K., Chandraker, M.: Materials for masses: Svbrdf acquisition with a single mobile phone image. In: Proceedings of the European conference on computer vision (ECCV). pp. 72–87 (2018)
22. Li, Z., Xu, Z., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Learning to reconstruct shape and spatially-varying reflectance from a single image. ACM Transactions on Graphics (TOG) **37**(6), 1–11 (2018)
23. Li, Z., Song, L., Chen, Z., Du, X., Chen, L., Yuan, J., Xu, Y.: Relit-neulf: Efficient relighting and novel view synthesis via neural 4d light field. In: Proceedings of the 31st ACM International Conference on Multimedia. pp. 7007–7016 (2023)
24. Lichy, D., Sengupta, S., Jacobs, D.W.: Fast light-weight near-field photometric stereo. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12612–12621 (2022)
25. Lichy, D., Wu, J., Sengupta, S., Jacobs, D.W.: Shape and material capture at home. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6123–6133 (2021)
26. Liu, Y., Li, Y., You, S., Lu, F.: Unsupervised learning for intrinsic image decomposition from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3248–3257 (2020)
27. Mecca, R., Logothetis, F., Budvytis, I., Cipolla, R.: Luces: A dataset for near-field point light source photometric stereo. arXiv preprint arXiv:2104.13135 (2021)
28. Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.P., Richardt, C., Theobalt, C.: Lime: Live intrinsic material estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 6315–6324 (2018)
29. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
30. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14. pp. 483–499. Springer (2016)
31. Oktay, O., Schlemper, J., Folgoc, L.L., Lee, M., Heinrich, M., Misawa, K., Mori, K., McDonagh, S., Hammerla, N.Y., Kainz, B., et al.: Attention u-net: Learning where to look for the pancreas. arXiv preprint arXiv:1804.03999 (2018)
32. Oxholm, G., Nishino, K.: Shape and reflectance estimation in the wild. IEEE transactions on pattern analysis and machine intelligence **38**(2), 376–389 (2015)

33. Rematas, K., Ritschel, T., Fritz, M., Gavves, E., Tuytelaars, T.: Deep reflectance maps. In: Proceedings of the IEEE Conference on computer vision and pattern recognition. pp. 4508–4516 (2016)
34. Sang, S., Chandraker, M.: Single-shot neural relighting and svbrdf estimation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16. pp. 85–101. Springer (2020)
35. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8598–8607 (2019)
36. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1685–1694 (2017)
37. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5541–5550 (2017)
38. Srinivasan, P.P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., Barron, J.T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7495–7504 (2021)
39. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P., Ramamoorthi, R.: Single image portrait relighting. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019)
40. Tiwari, A., Raman, S.: Deepps2: Revisiting photometric stereo using two differently illuminated images. In: European Conference on Computer Vision. pp. 129–145. Springer (2022)
41. Tiwari, A., Raman, S.: Lerps: Lighting estimation and relighting for photometric stereo. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 2060–2064. IEEE (2022)
42. Wimbauer, F., Wu, S., Rupprecht, C.: De-rendering 3d objects in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18490–18499 (2022)
43. Woodham, R.J.: Photometric method for determining surface orientation from multiple images. Optical engineering **19**(1), 139–144 (1980)
44. Wu, H., Hu, Z., Li, L., Zhang, Y., Fan, C., Yu, X.: Nefii: Inverse rendering for reflectance decomposition with near-field indirect illumination. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4295–4304 (2023)
45. Xu, Y., Zoss, G., Chandran, P., Gross, M., Bradley, D., Gotardo, P.: Renerf: Relightable neural radiance fields with nearfield lighting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22581–22591 (2023)
46. Xu, Z., Sunkavalli, K., Hadap, S., Ramamoorthi, R.: Deep image-based relighting from optimal sparse samples. ACM Transactions on Graphics (ToG) **37**(4), 1–13 (2018)
47. Yamamoto, T., Nakazawa, A.: General improvement method of specular component separation using high-emphasis filter and similarity function. ITE Transactions on Media Technology and Applications **7**(2), 92–102 (2019)
48. Yang, J., Liu, Q., Zhang, K.: Stacked hourglass network for robust facial landmark localisation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 79–87 (2017)

49. Yi, R., Zhu, C., Xu, K.: Weakly-supervised single-view image relighting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8402–8411 (2023)
50. Yu, Y., Smith, W.A.: Inverserendernet: Learning single image inverse rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3155–3164 (2019)
51. Yu, Y., Smith, W.A.: Outdoor inverse rendering from a single image using multi-view self-supervision. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3659–3675 (2021)
52. Zhang, X., Fanello, S., Tsai, Y.T., Sun, T., Xue, T., Pandey, R., Orts-Escolano, S., Davidson, P., Rhemann, C., Debevec, P., et al.: Neural light transport for relighting and view synthesis. ACM Transactions on Graphics (TOG) **40**(1), 1–17 (2021)
53. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single-image portrait relighting. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7194–7202 (2019)
54. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)