MatTrans: Material Reflectance Property Estimation of Complex Objects with Transformer

Liping Wu[1], Bin Cheng[1], Wentao Chao[1], Juli Zhao[2], and Fuqing Duan[1]

[1] School of Artificial Intelligence, Beijing Normal University, 100875, Beijing, CHINA
202121081045@mail.bnu.edu.cn 1401050696@qq.com chaowentao@mail.bnu.edu.cn
fqduan@bnu.edu.cn
[2] School of Data Science and Software Engineering, Qingdao University 266071, Qingdao, CHINA
zhaojl@yeah.net

Abstract. Material Reflectance Property Estimation of an object is challenging and it can be used in realistic rendering to make the appearance of objects realistic. Current research focuses primarily on the near-planar objects, with little attention paid to complex-shaped objects. In this paper, we propose a method called MatTrans to estimate geometry and material reflectance properties with Transformer. Specifically, a Transformer Encoder module is designed to fuse local and global information for each material property respectively, and then a cascaded network with residual learning is introduced to estimate the geometry and reflectance properties of any 3D object surface from a single image. Extensive experiments validate that our method brings a clear improvement over previous methods for single-shot capture of spatially varying BRDFs.

Keywords: Reflectance property estimation · Transformer · Complex-shaped objects.

1 Introduction

Surface appearance modeling of an object has always been a research hot spot in computer graphics, and it is widely used in 3D animation, games, virtual reality and so on. The appearance of a real-world object is the result of interactions between the light and the object. The reflectance of light on object surface varies with different materials, and it directly affects the appearance of the object. For example, under the same lighting conditions, metal is brighter than wood because the specular reflection effects the material reflectance property of opaque objects is typically represented by the Spatially Varying Bidirectional Reflectance Distribution Function (SVBRDF) [1, 2] which describes how the incident irradiance at different points on the objects surface affects the emissivity in a given reflected direction. SVBRDF is primarily determined by the object's

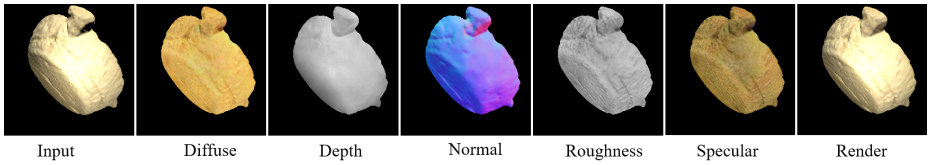

Fig. 1. We propose a cascaded residual network based on the Transformer for recovering arbitrary shapes and spatially-varying BRDF from a single mobile phone image. Our approach produces high-quality material property maps. The rendering pictures with these properties are also very close to the inputs.

shape and properties such as diffuse reflectance, roughness, and specular reflectance. As shown in Figure 1, surface appearance modeling of an object is to estimate the properties such as diffuse albedo, roughness, surface normal and specular albedo at the 3D object surface point corresponding to each pixel from the collected texture images. Based on these material properties and the geometric information of the object's surface, the appearance of the object under any light illumination can be rendered. However, in addition to these geometric and material properties, illumination and viewpoint also influence the appearance of an object, and different combinations of these factors may lead to the same appearance. This makes the material reflectance property estimation challenging, especially for complex objects.

Most of previous methods rely on well-designed and calibrated light field acquisition systems to densely collect the texture photos of the object under different illuminations and viewpoints and estimate the reflectance properties by inverse rendering of these texture photos. It is commonly a time-consuming and laborious work to collect the data of an object. In recent years, reflectance property estimation based on deep learning has gradually gained popularity. Generally, deep learning methods [3, 7, 10] estimate the reflectance property from sparse multi-view images, and they reduce the complexity of data acquisition and reflectance estimation. However, the available material datasets that are requisite are very limited, and most of them are about near-planar samples, which is unfit for the material reflectance property modeling of complex shape objects. Aside from more descriptive dataset, disambiguating shape and spatially-varying material necessitates novel network architectures that can reason about appearance at multiple scales, such as understanding both local shading and non-local shadowing and lighting variations, especially in the case of unknown complex geometry. Mainstream methods use encoder-decoder architectures based on convolutional neural layers due to their good feature extraction abilities. However, convolutional layers tend to focus more on local information, and they struggle to aggregate distant information and propagate it to fine-scale details. To solve this limitation, we design a novel coarse-to-fine cascaded network with the Transformer to estimate shape and SVBRDF parameters.

There are three main contributions in our paper: (1) We generate a material dataset by rendering a set of 3D models of cultural relics with complex sur-

face shapes. (2) We improve the encoder-decoder architecture using the Transformer module which can extract information over long distances. Therefore, our method has both local and global modeling capabilities over previous methods. (3) We add residual blocks to the network to learn residual information in the refined networks, which can ease the network learning while improving the estimation effect.

Experiments on different datasets demonstrate that our method brings a clear improvement over state-of-the-art methods for single-shot capture of spatially varying BRDFs.

2 Related works

Non-Deep learning-based methods. In recent years, research on reflectance property modeling mainly focuses on the reflectance estimation using sparse light field data captured by convenient lightweight acquisition devices. The literatures [28, 41] adopt a near-field camera and a directional light source to take photos of near-planar samples, and fit SVBRDF from the collected sparse data with a prior hypothesis about the system configuration or material. The works [40, 8] estimate SVBRDF from a video of rotating objects under unknown natural illumination with the sparsity prior in gradient domain of natural illumination. Nam et al. [27] design a multi-stage iterative inverse rendering framework to jointly reconstruct SVBRDF, normal and 3D shape of an object surface from a set of photos taken by a camera with built-in flash. Baek et al. [2] leverage the physical relationship between the polarized appearance and geometric characteristics of the object to estimate the material appearance and normal from a set of polarization images. Michael et al. [17] use a high-frequency spatially modulated light source and a camera that is precisely aligned with the light source to capture modulated images of the object and recover the geometry and reflectance from these images through light modulation and demodulation and stereo vision reconstruction. Borom et al. [33] use a computational illumination device to capture a set of images of objects under continuous spherical harmonic illumination to recover the geometric structure and SVBRDF. Wu et al. [39] jointly optimize the camera pose, reflectance, ambient illumination and normal from a set of RGB and depth images captured by a RGB-D camera under an unknown ambient illumination. Most of these methods use priors on illumination mode or motion to optimize the reflectance properties, and the computational complexity is very high. Compared to these methods, our deep learning-based approach is simpler and more efficient.

Deep learning-based methods. With the ever-increasing successful applications of deep learning on many visual tasks, it has attracted wide attention in research on object appearance modeling for its simplicity, high efficiency and easy training. However, existing works [21, 7, 23] are mainly about reflectance estimation of near-planar objects, and few work is for arbitrary three-dimensional objects. Kang et al. [18] propose a framework based on the asymmetric deep autoencoder to automatically learn the effective illumination modes and reconstruct

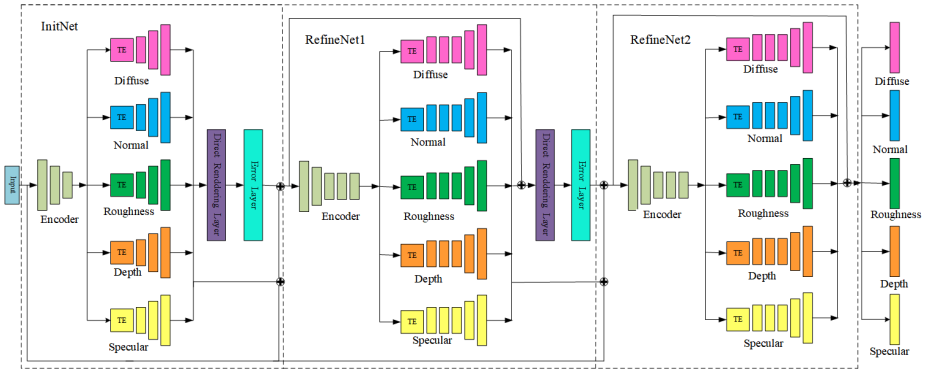

Fig. 2. Our network framework is mainly based on the encoder-decoder structure. Different colors represent different material properties. Before each decoder structure, in order to extract global features, we add a Transformer Encoder TE. We use the L2 distance to constrain the generated material properties. Our method adopts a coarse-to-fine manner, we use a cascaded network RefineNet to learn the residuals, which is also based on the encoder-decoder structure. Unlike the InitNet, the input of the network is composed of the input of the InitNet and the material properties generated in the previous stage. The figure on the left shows the InitNet whose encoder is composed of six convolutional layers, and each decoder is also composed of six deconvolutional layers. Besides, we use the skip-connections in each decoder to help the network recover detailed information when upsampling. The figure on the right shows RefineNet, the network is also an encoder-decoder structure, in addition to being different from the initial network in the input, residual blocks are also used in the encoding and decoding process.

SVBRDF from the images captured in these illumination modes. However, this method relies on precisely designed acquisition equipment. Li et al. [24] generate a large scale material dataset of 3D objects, and trained a deep neural network using this dataset to recover the SVBRDF and geometry of a 3D object of arbitrary shape from a single RGB image. The network physically incorporates the illumination representation and the differentiate rendering of the scene appearance, and adopts a cascaded structure to iteratively refine the prediction results. Based on the dataset in [24], Bi et al. [3] propose a deep multi-view reflectance estimation network architecture to predict per-view SVBRDF, and geometry and SVBRDF are obtained by fusing these per-view estimations. The material dataset they used is generated by rendering a set of three-dimensional data models and the data modes are manually generated by combining a few regular primitive shapes like cone, cylinder, and so on. Therefore, the generated material data always has a certain mismatch with the real cases. Cheng et al. [4] propose an end-to-end network based on attention mechanism to estimate the reflectance properties of any 3D object surface from a single image. However, they do not estimate the geometric properties. On the basis of this work, we

use the Transformer to improve the global modeling capability and estimate the geometry and reflectance properties simultaneously.

There have been many studies that try to avoid the need for a large amount of self-supervised data by using a GAN [11] to estimate material properties now. Zhao et al. [43] present an unsupervised generative adversarial neural network that addresses both SVBRDF capture and synthesis of high-resolution SVBRDF maps from a single image at the same time. Guo et al. [12] train the HA-convolution to "guess" the saturated pixels (specular highlight area) by the unsaturated area surrounded, making the extracted features more uniform. Guo et al. [13] present MaterialGAN, a deep generative convolutional network based on StyleGAN2 [19], trained to synthesize realistic SVBRDF parameter maps. They show that MaterialGAN could be used as a powerful material prior for an inverse rendering framework. Most of these works are aimed at near-planar samples. Compared to these works, our method is more suitable for estimating surface material properties of complex 3D objects. In recent years, the diffusion model [16] has achieved great success in the community of image synthesis. Vecchio et al. [35] present ControlMat, a generative method, based on a Latent Diffusion Model (LDM) [29] to produce SVBRDFs. They map SVBRDFs to a latent space and train a diffusion model to sample from this latent space. To generate high-quality images Vecchio et al. [36] also adopted a structure named MatFuse similar to controlnet [42] by adding conditional information to guide image generation. However, GAN is unstable and prone to collapse during training and the inference of diffusion model is computationally slow, which is what we don't want. Our method can generate high-quality SVBRDFs maps both stably and quickly.

An alternative to data-driven relighting is inverse rendering, which involves optimizing a set of trial model parameters based on the discrepancy between rendered and reference photographs. Inverse rendering poses a complex non-linear optimization problem at its core. Recent advancements in differentiable rendering have facilitated more robust inverse rendering for intricate scenes and capture conditions. Munkberg et al. [26] employ an alternating optimization approach, refining an implicit shape representation (i.e., signed distance field) as well as reflectance and lighting defined on a triangle mesh. Hasselgren et al. [14] extend the work of Munkberg et al. [26] by incorporating a differentiable Monte Carlo renderer to handle area light sources and integrating a denoiser to mitigate gradient computation issues caused by Monte Carlo noise during non-linear optimization. Similarly, Fujun et al. [25] utilize a differentiable Monte Carlo renderer to estimate shape and spatially-varying reflectance from a limited number of colocated view/light photographs. All these methods primarily focus on direct lighting effects, potentially yielding suboptimal outcomes for objects or scenes with pronounced interreflections. However, all aforementioned approaches ultimately represent shapes using triangle meshes, limiting their applicability to objects with well-defined surfaces; moreover, the accuracy of these methods inherently depends on the representational capacity of underlying BRDF and lighting models.

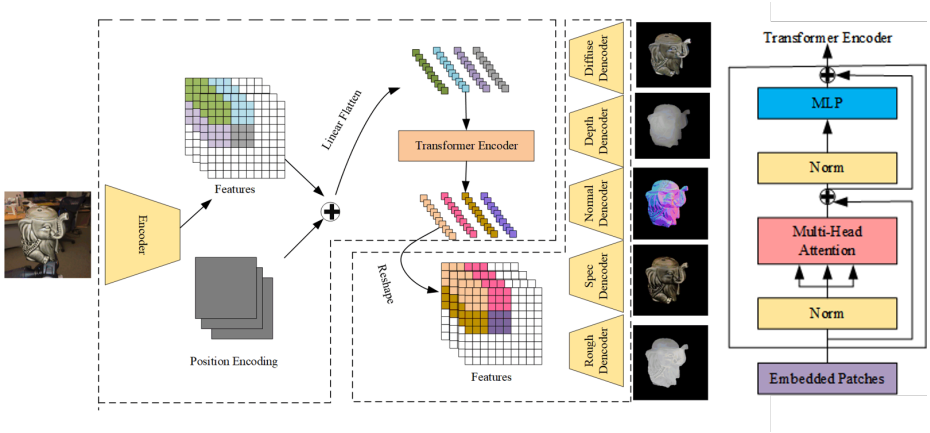

Fig. 3. The framework of the Transformer Encoder module. We divide the feature maps into small patches of the same size in the last layer of the encoder, and we supplement the position information between different tokens by position embedding and use them as inputs for the Transformer Encoder, where the encoder structure is shown on the right. After Transformer Encoder, we reshape the outputs to the same size as the feature maps output by the encoder. Then we input the newly generated feature maps into different decoder branches to get different material properties.

3 Method

Given a single image of a 3D object captured under a flashlight, we estimate the shape and spatially varying BRDF from the image. The overall framework of our method is shown in Figure 2. Simliar with the prior works [24, 4], we also adopt a cascaded network architecture, which includes an initial estimation network and several refined estimation networks. We adopt the U-Net [30] as the basic network architecture of the initial and refined estimation networks. The U-Net has proven to be suited for a wide range of similar image-to-image translation tasks. However, previous works have shown that this network does not perform well enough when dealing with tasks that require fusing long-distance visual information. Considering the transformer has an excellent global modeling effect, we add the Transformer [9] module to extract better global features.

3.1 Initial estimation network

Initial estimation network consists of a single encoder and five decoders for different shape and SVBRDF parameters: diffuse albedo(A), roughness (R), surface normal(N), specular(S) and depth(D). Since the object appearance is decided by the object surface shape and material reflectance properties, the five shape and SVBRDF parameters are closely correlated. Therefore, the network can be considered as a multi-task learning network and the five parameter estimation tasks share a single encoder, which can boost the generalization performance

of the material reflectance estimation task by shared information representation learning of the multiple tasks. In order to extract global features, we add a Transformer module named as TE before each decoding structure.

The input image is processed through a sequence of 6 convolutional layers that perform downsampling (the encoder), followed by a sequence of 6 upsampling and convolutional layers (the decoder). The resolution of the input image is halved after each convolutional layer. Such an hourglass-shaped network can obtain more accurate feature maps. Assume A_0, N_0, R_0, D_0, and S_0 denote the initial estimations of diffuse, normal, roughness, depth, and specular reflectance properties respectively, and $InitNet$ is the initial estimation network.

$$A_0, N_0, R_0, D_0, S_0 = InitNet(I, M) \tag{1}$$

where I is the input image, and M is the mask of the object in the image, which is used to extract the object in the image.

3.2 Refined estimation network

Following the initial estimation network, two refined estimation networks are connected to enhance the learning ability of the network and optimize the estimation, where the structure of each refined estimation network is the same to the initial estimation network. It is composed of two convolutional layers and three residual blocks. The inputs of each cascaded stage include the input image, the shape SVBRDF from the previous stage as well as the rendering error associated with these previous predictions in relation to the input image. This enables each cascaded stage to iteratively improve the predictions by taking into account the rendering error observed in the previous stage. Like the initial estimation network, a Transformer module is also added before each decoding structure to learn the global features. In order to ease the network learning and improve the estimation effect, a residual block [15] is added to each cascaded network. We name the optimization network as RefineNet. Take the diffuse properties as an example, Let A_0 represent the diffuse albedo of the $Initnet$, $RefineNet_n$ represent the n_{th} optimization module and A_n represent the diffuse albedo of the n_{th} optimization module.

$$\begin{aligned} A_n, N_n, R_n, D_n, S_n = & RefineNet_n(I_{n-1}, M, A_{n-1}, N_{n-1}, \\ & R_{n-1}, D_{n-1}, S_{n-1}, Err_{n-1}) \end{aligned} \tag{2}$$

where I_{n-1} represents the rendered image in the previous network and Err_{n-1} represents the L2 loss between the previous-stage rendering result and the ground truth.

3.3 Transformer Encoder

Distant regions of a material sample often offer complementary information to each other for SVBRDF recovery. This observation is at the heart of many past

methods for material capture. Convolution alone cannot capture information over long distances. As we all know, the Transformer [34] has a good extraction function for long-distance information due to the unique advantages of self-attention. We use the Transformer to estimate the shape and surface reflectance properties from a single image. The Transformer structure is shown in Figure 3.

We use the multi-head attention mechanism during the Transformer Encoder and finally the module will output feature maps which have both local information and long-distance supplementary information, which is beneficial for reconstructing SVBRDFs. Extensive experiment results illustrate it.

3.4 Dataset

Generally, methods based on deep learning require a large training dataset. However, most works are based on near-planar datasets, and the only available material dataset of three-dimensional objects is the one in the literature [24], where they used complex 3D data models generated by a random combination of several artificial 3D shapes like spheres, cubes and so on, and the generated material data are the reflectance and geometry property maps rendered from multiple projection viewpoints. Since the surfaces of these 3D models are simple smooth surfaces, this dataset is not well-suitable for the reflectance estimation of some real 3D objects with complex surfaces (like some complex cultural antiques). We downloaded 82 cultural relic models with relatively complex shapes or materials from Sketchfab. To keep original material distribution and realistic appearance of these antique models, we directly utilized the original texture maps of the models instead of reassigning an alternative material, where some missing reflectance properties are completed manually. We linearly transformed diffuse, roughness, and specular to generate new material parameters for data augmentation. As for the rendering pictures, we use the physically motivated microfacet BRDF model in [37, 6, 5, 20, 32]. Let c_{diff}, n, $Roughness$, and F_0 represent the diffuse albedo, normal, roughness and specular albedo respectively, and v, l represent the view and light direction respectively. Then, the BRDF model is expressed as:

$$f(l,v) = \frac{c_{diff}}{\pi} + \frac{D(h)F(v,h)G(l,v,h)}{4(n\cdot l)(n\cdot v)} \tag{3}$$

Where $D(h)$ $F(v,h)$ and $G(l,v,h)$ are the normal distribution, fresnel, and geometric terms respectively. These terms are defined as follows:

$$D(h) = \frac{\alpha^2}{\pi\left[(n\cdot h)^2(\alpha^2-1)+1\right]^2} \tag{4}$$

$$\alpha = Roughness^2 \tag{5}$$

$$k = \frac{(Roughness+1)^2}{8} \tag{6}$$

$$F(v,h) = F_0 + (1-F_0)2^{-[5.55473(v\cdot h)+6.8316](v\cdot h)} \tag{7}$$

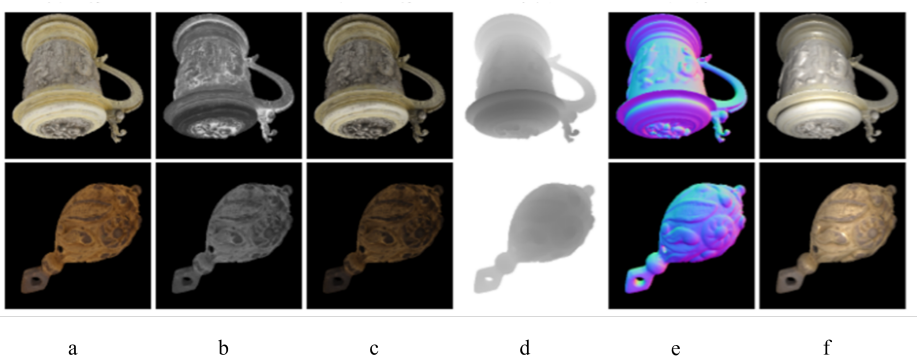

Fig. 4. Samples from the dataset. (a)-(e) Diffuse albedo, roughness, specular albedo, depth and surface normals. (f) rendered images by a dominant point light source collocated with the camera and the reflectance properties.

$$G_1(v) = \frac{n \cdot v}{(n \cdot v)(1-k)+k} \tag{8}$$

$$G(l, v, h) = G_1(l)G_1(v) \tag{9}$$

In the end, we generated 37720 samples from 20 different viewpoints. Each sample is composed of a rendered appearance image, three reflectance property maps of diffuse albedo, roughness specular albedo and two geometric property maps of depth and normal. The resolution of each map is 256×256. Samples are shown in Figure 4.

3.5 Training

Loss Function We have the same loss function for both $InitNet$ and each $RefineNet$ stage. For diffuse albedo, normal, roughness, specular and rending results, we use L2 loss to constrain the network so that the predicted material properties are closer to the ground truth. Since the range of depths is larger than that of other BRDF parameters, we use an inverse transformation to project the depth map into a fixed range. Let $\tilde{d}_i$ be the initial output of depth prediction network of pixel i; the real depth d_i is given by

$$d_i = \frac{1}{\sigma \cdot \left(\tilde{d}_i + 1\right) + \epsilon} \tag{10}$$

where σ=0.4, ϵ=0.2. We also calculate L2 loss for depth. Considering the depth information learned by the initial network is already good so in the Refine network, we no longer convert depth, and directly learn the residual information of depth so the final loss can be written as :

$$\mathcal{L} = \lambda_a \mathcal{L}_a + \lambda_n \mathcal{L}_n + \lambda_r \mathcal{L}_r + \lambda_d \mathcal{L}_d + \lambda_{\mathrm{s}}\, \mathcal{L}_{\mathrm{s}}\ + \lambda_{\mathrm{render}}\, \mathcal{L}_{\mathrm{render}} \tag{11}$$

where λ_a=λ_n=λ_{render}=1 and λ_r=λ_d=0.5 represent the different weights corresponding to the losses.

Training Strategies Considering that designing an end-to-end network may cause the number of parameters at one time to be too large, which could result in overflow of memory. So we train the network in three phases according to the cascaded network. We set the batchsize as 5 during training. We use Adam optimizer, with a learning rate of 10^{-4} for the encoder and 4×10^{-4} for the decoders. The learning rate is halved every two epochs during training. As for the epoch at each stage, we set it as 15, 8, 6 respectively. We trained the network for approximately two days using an NVIDIA 2080 GPU.

4 Experiments

We validate the effectiveness of our method with evaluations of synthetic and real data, firstly, we conduct ablation experiments to analyzes the each module. Then, we compare the results with state-of-the-art works and use the average L2 loss of estimation on test data to evaluate the performance. We use Dataset1 and Dataset2 to represent our dataset and Li's dataset. For Dataset1, we choose 66 models with a total of 30360 samples as the training data and the rest are testing data. The metric shows that our method is superiority of state-of-the-art works.

4.1 Ablation Experiment

In this section, we use our dataset to train and test our network to validate the effect of our method.

Coarse-to-Fine Network We first verify the effect of the cascaded network module. We calculate the initial network estimation results and the refined network estimation results respectively. The average estimation errors of the test dataset are shown in Table 1 with InitNet and RefineNet. It can be seen that the reflectance property estimation accuracy of the cascaded network is improved compared to the initial network. Especially, the rendering result is boosted signifivantly by the RefineNet. This validates the effectiveness of the cascaded network module. The material reflection property estimation results of the RefineNet are shown in Figure 5.

	InitNet	RefineNet
diffuse(10^{-2})	**6.6801**	6.8951
roughness(10^{-1})	**1.76660**	1.76697
depth(10^{-2})	3.9389	**3.7384**
normal(10^{-2})	9.1590	**8.3738**
specular(10^{-1})	1.1115	**1.0773**
render(10^{-3})	8.116	**1.335**

Table 1. Quantitative comparisons L2 errors of InitNet and RefineNet.

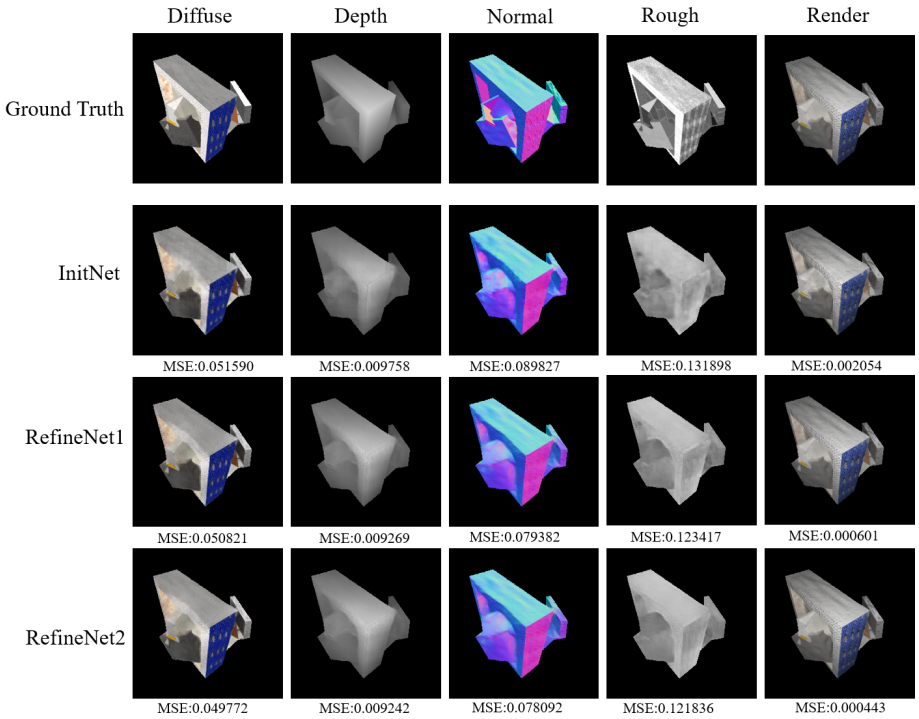

Fig. 5. Effect of our cascaded design, we annotated the mse loss under corresponding photo, which proves that cascaded networks are effective.

Residual Learning. To test the effectiveness of the residual learning in the cascaded network, we train two variants of our basic network with or without residual learning. The final test results are shown in Table 2. As we can see, most evaluated results for the network with residual learning are better than results without residual learning, which demonstrate the ability of residual learning.

Transformer Encoder. Then, we analyze the effect of the Transformer Encoder module. We take the network with residual and specular reflection properties as the baseline, which does not include a module with global feature extraction. As a comparison, we add Transformer Encoder on the baseline and retrain the network. The final quantitative results are shown in Table 3. The results show that after adding the Transformer module, the L2 loss of related attribute estimation is significantly reduced, which proves that the improvement of our work is very effective. In order to prove the unique advantages of the Transformer Encoder module, we use the non-local [38] module with the same global feature extraction, and the final result is shown in Table 3. As we can see, using the non-local module also improves the results of material properties, which also implicitly proves the necessity of adding global information, while the improvement is limited compared with the Transformer. Table 4 shows our

	with residual	without residual
diffuse(10^{-2})	**6.8951**	7.2965
roughness(10^{-1})	1.76697	**1.765836**
depth(10^{-2})	3.7384	**3.6874**
normal(10^{-2})	**8.3738**	8.7979
specular(10^{-1})	**1.07731**	1.10250
render(10^{-3})	**1.335**	4.346

Table 2. Quantitative comparisons L2 errors of with or without residual learning during the RefineNet.

	baseline	with Transformer	with non-local
diffuse(10^{-2})	8.7629	**6.8951**	8.4230
roughness(10^{-1})	1.77358	**1.76697**	1.78906
depth(10^{-2})	4.8313	**3.7384**	4.5657
normal(10^{-2})	8.6889	**8.3738**	8.6593
specular(10^{-1})	1.25774	**1.07731**	1.23236
render(10^{-3})	1.354	**1.335**	1.364

Table 3. Quantitative comparisons L2 errors illustrating the influence of the Transformer Encoder. To demonstrate the unique advantages of the Transformer Encoder, we compare the result with adding a non-local module that also have global feature extraction.

full version of ablation experiments. The network model we used, in the end, achieves the best results on all indicators, which illustrates that our design is reasonable.

4.2 Generalization to real data

In order to prove that our method can have a relatively accurate estimate of the material of complex objects, We demonstrate our method on several real objects. The material reflection property estimation results of several real images are shown in Figure 7. As we can see, the images we rendered are very close to the inputs, which indicates that the material property parameters have strong reliability. In order to verify the effectiveness of the proposed method we test the data under different lighting directions. The results are in shown figure 8, as we can see, the proposed method also have good results under novel lighting directions. To make our experimental results more convincing, we retrain our method on Dataset2 and the final results for real data are shown in Figure 9. The final results show that our method achieves very good results on both datasets.

	diffuse(10^{-2})	normal(10^{-2})	roughness(10^{-1})	depth(10^{-2})	specular(10^{-1})	render(10^{-3})
baseline	9.3375	9.0161	1.78722	4.5985	-	2.908
+spec	8.8422	9.2881	1.77868	4.8189	1.21809	3.374
+spec res	8.7629	8.6889	1.77358	4.8313	1.25774	1.354
+spec res non-local	8.4230	8.6593	1.78906	4.5657	1.23236	1.364
+spec res Transformer	**6.8951**	**8.3738**	**1.76697**	**3.7384**	**1.07731**	**1.335**

Table 4. Our full version of ablation experiments. We use Li's network as the baseline, and we add specular reflection properties, a residual learning in cascaded networks, a non-local module or Transformer Encoder.

	diffuse(10^{-2})	normal(10^{-2})	roughness(10^{-1})	depth(10^{-2})	specular(10^{-1})	render(10^{-3})
Li's model& Dataset1	9.3375	9.0161	1.78722	4.5985	-	2.908
Sang's model& Dataset1	8.2564	8.7905	1.79537	4.2596	-	2.083
Zheng's model& Dataset1	9.1106	9.1854	1.81733	3.9065	-	2.573
Cheng's model &Dataset1	9.9011	-	1.97100	-	1.24516	-
Our's model & Dataset1	**6.8951**	**8.3738**	**1.76697**	**3.7384**	**1.07731**	**1.335**
Li's model & Dataset2	4.868	3.822	1.943	1.505	-	1.637
Our's model & Dataset2	**4.5042**	**3.33811**	**1.77122**	**1.3456**	-	**0.379**

Table 5. To demonstrate the effectiveness of our method, we test it on two datasets, where Dataset1 represents the complex artifact dataset and Dataset2 represents Li generated synthetic dataset. We select Li's as the comparison methods on both datasets, respectively. We use L2 loss as the evaluation metric.

4.3 Comparison Experiment

In related works, the one in Li et al. [24] is the work to restore the material reflectance properties of objects with arbitrary geometry from a single image. They adopt a cascaded network structure and estimate the material and geometric properties at the same time in each cascaded module. A rendering layer is also included in each cascaded module to render the estimation results, and the rendering error along with the current estimation results are taken as the input into the next cascaded module. This method trains each cascaded network module separately and then assembles them together in a cascaded way. At the same time, we also compare Sang's [31] and Zheng's [22] methods on our dataset. There is also another work in Cheng et al. [4] for complex objects. However, they only estimated the reflection properties of the object but not the

	diffuse	normal	roughness	specular	render
Zhao[43]	0.062	0.065	0.153	0.079	0.080
Guo[13]	0.065	0.070	0.159	0.094	0.081
Kalantari[44]	0.067	0.058	0.124	0.089	0.061
Our	**0.058**	**0.054**	**0.115**	**0.069**	**0.058**

Table 6. The RMSE error on Deschaintr's near-planar dataset.

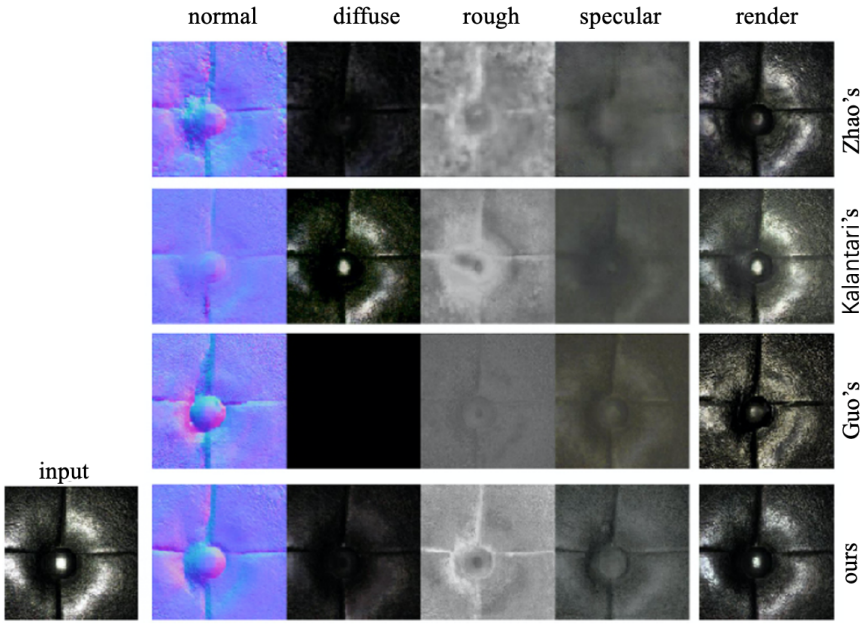

Fig. 6. Comparison on near-planar dataset. All the re-rendered results were produced under identical lighting conditions and viewing direction.

geometric properties. In the comparisons, we also validate the generated dataset is necessary for reflectance estimation of arbitrary complex shape objects. We train and test both network models with Dataset1 and Dataset2 respectively. Since the sample in Dataset2 does not include specular albedo, we remove the specular estimation in our network structure when training with Dataset2 and take the specular albedo as a constant when rendering. The estimation errors are shown in Table 5. We can see that the errors of all material properties of our method are minimal. Some results for real captured images are shown in Figure 10. Since we have no geometric models of these images, we can not compare with the Cheng's method. From the figure, we can see that the rendering results of our method are closer to the input images.

In order to prove the effectiveness of our method, we also trained the network using Deschaintre's [7] near-planar dataset and compared it with Zhao's [43], Guo's [13], Kalantari's [44] methods. We use root mean squared error (RMSE) to evaluate the quality of the reflectance parameters. The final results are shown in the Table 6. As we can see, our method also achieves the best results on near-planar datasets, which verifies the robustness of our method. Visual comparisons are shown in figure 6. It is evident that our results exhibit the closest resemblance to the input pictures, effectively reconstructing the SVBRDF maps.

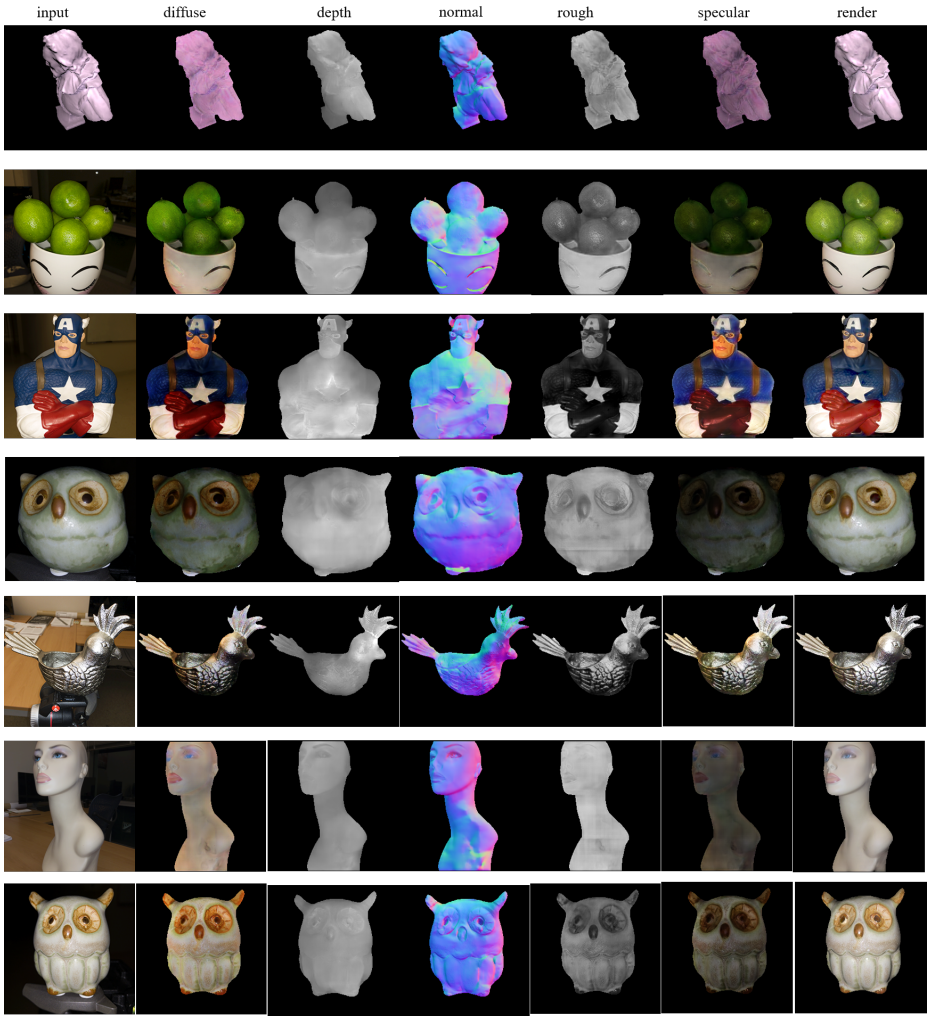

Fig. 7. Results on real objects. As we can see, we achieve high-quality recovery of shape and spatially-varying BRDF.

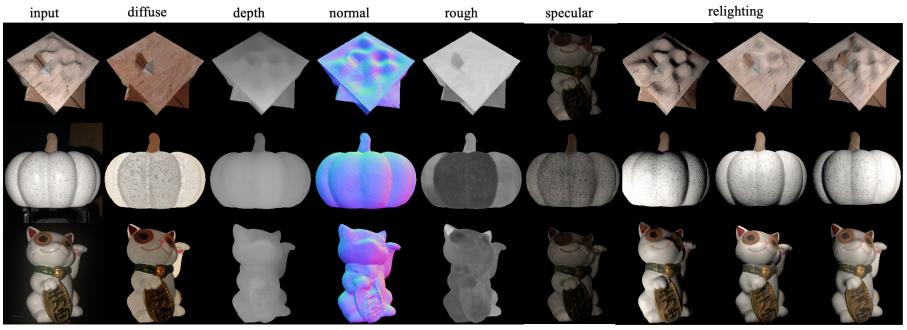

Fig. 8. Results rendered from novel lighting directions. We show the input image, the estimated shape and BRDF parameters and the rendered output under different lighting directions.

5 Conclusion

Surface appearance modeling of an object has always been a research hot spot in computer graphics. Recently, deep learning-based methods have gradually become the mainstream. However, current research focuses primarily on the near-planar objects, with little attention paid to complex-shaped objects. In this paper, we propose a method called MatTrans to estimate geometry and material reflectance properties with Transformer. Specifically, a Transformer Encoder module is designed to fuse local and global information for each property respectively, and then a cascaded network with residual learning is introduced to estimate the geometry and reflectance properties of any 3D object surface from a single image. Extensive experiments validate that our method which includes an initial estimation network and several refined estimation networks brings a clear improvement over state-of-the-art methods for single-shot capture of spatially varying BRDFs.

References

1. Aittala, M., Weyrich, T., Lehtinen, J., et al.: Two-shot svbrdf capture for stationary materials. ACM Trans. Graph. **34**(4), 110–1 (2015)
2. Baek, S.H., Jeon, D.S., Tong, X., Kim, M.H.: Simultaneous acquisition of polarimetric svbrdf and normals. ACM Trans. Graph. **37**(6), 268–1 (2018)
3. Bi, S., Xu, Z., Sunkavalli, K., Kriegman, D., Ramamoorthi, R.: Deep 3d capture: Geometry and reflectance from sparse multi-view images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5960–5969 (2020)
4. Cheng, B., Zhao, J., Duan, F.: Material reflectance property estimation of complex objects using an attention network. In: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW). pp. 632–633. IEEE (2022)

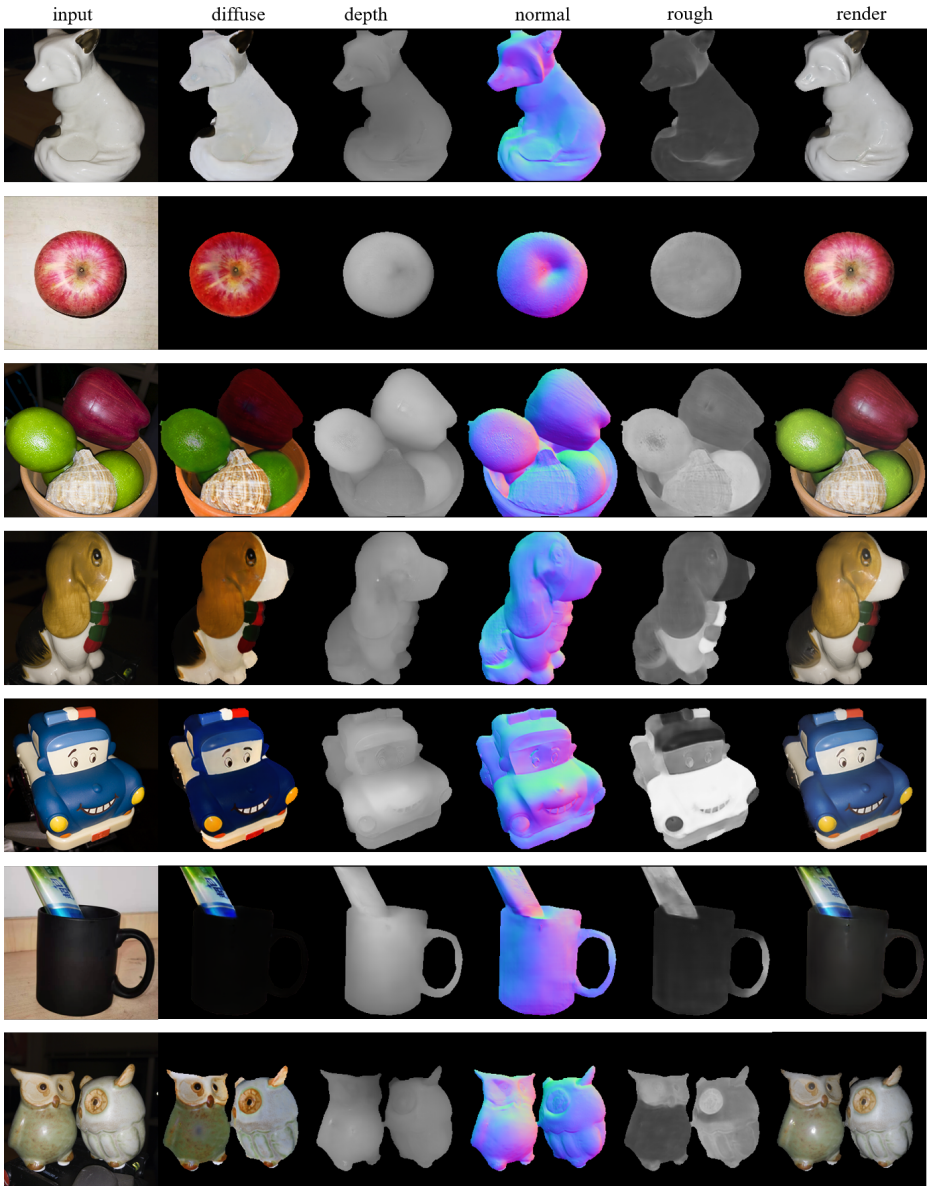

Fig. 9. we train our network on Li's dataset. We also get the high-quality recovery of shape and spatially-varying BRDF on real objects.

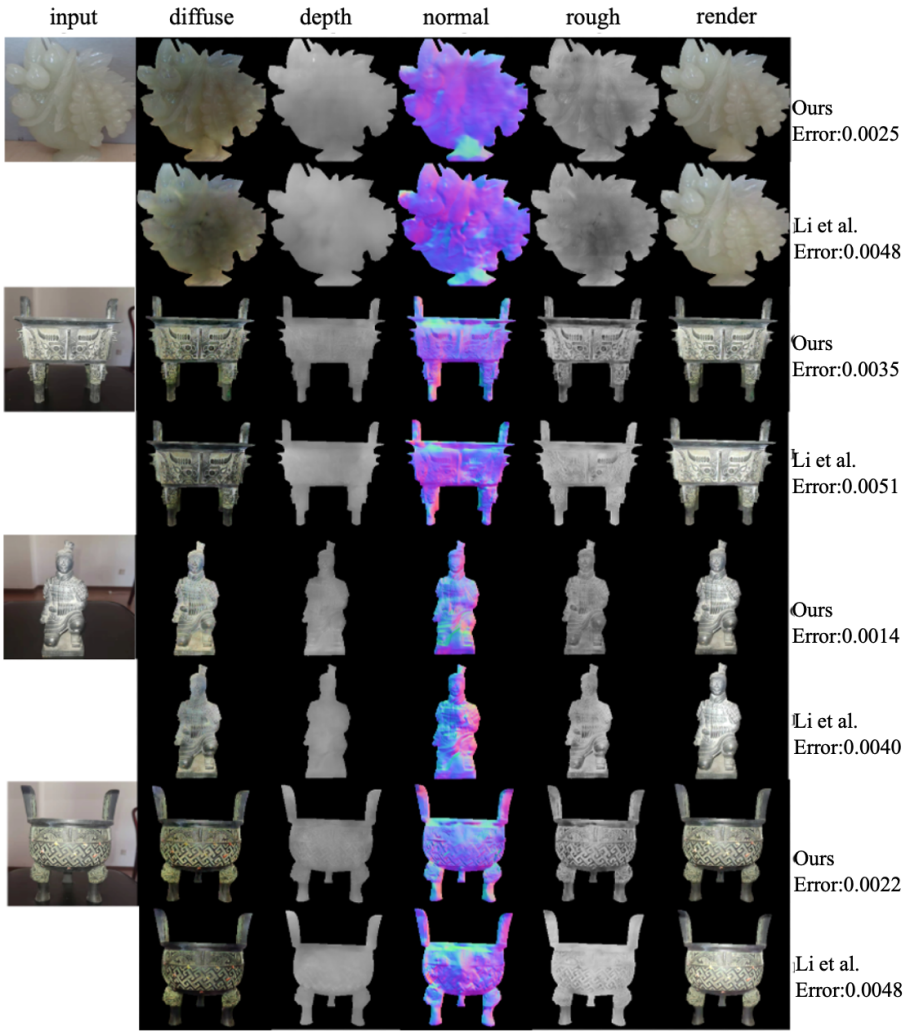

Fig. 10. Comparison with Li et al. [24] on SVBRDF estimation on real images, our results are better in comparison.

5. Cook, R.L., Torrance, K.E.: A reflectance model for computer graphics. ACM Siggraph Computer Graphics **15**(3), 307–316 (1981)
6. Cook, R.L., Torrance, K.E.: A reflectance model for computer graphics. ACM Transactions on Graphics (ToG) **1**(1), 7–24 (1982)
7. 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)
8. Dong, Y., Chen, G., Peers, P., Zhang, J., Tong, X.: Appearance-from-motion: Recovering spatially varying surface reflectance under unknown lighting. ACM Transactions on Graphics (TOG) **33**(6), 1–12 (2014)
9. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
10. Gao, D., Li, X., Dong, Y., Peers, P., Xu, K., Tong, X.: Deep inverse rendering for high-resolution svbrdf estimation from an arbitrary number of images. ACM Trans. Graph. **38**(4), 134–1 (2019)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets, in 'advances in neural information processing systems 27', curran associates (2014)
12. Guo, J., Lai, S., Tao, C., Cai, Y., Wang, L., Guo, Y., Yan, L.Q.: Highlight-aware two-stream network for single-image svbrdf acquisition. ACM Transactions on Graphics (TOG) **40**(4), 1–14 (2021)
13. Guo, Y., Smith, C., Hašan, M., Sunkavalli, K., Zhao, S.: Materialgan: reflectance capture using a generative svbrdf model. arXiv preprint arXiv:2010.00114 (2020)
14. Hasselgren, J., Hofmann, N., Munkberg, J.: Shape, light, and material decomposition from images using monte carlo rendering and denoising. Advances in Neural Information Processing Systems **35**, 22856–22869 (2022)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
17. Holroyd, M., Lawrence, J., Zickler, T.: A coaxial optical scanner for synchronous acquisition of 3d geometry and surface reflectance. ACM Transactions on Graphics (TOG) **29**(4), 1–12 (2010)
18. Kang, K., Chen, Z., Wang, J., Zhou, K., Wu, H.: Efficient reflectance capture using an autoencoder. ACM Trans. Graph. **37**(4), 127–1 (2018)
19. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
20. Lagarde, S.: Spherical gaussian approximation for blinn-phong, phong and fresnel. Random Thoughts about Graphics in Games blog, June **3** (2012)
21. 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)
22. Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2475–2484 (2020)

23. 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)
24. 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)
25. Luan, F., Zhao, S., Bala, K., Dong, Z.: Unified shape and svbrdf recovery using differentiable monte carlo rendering. In: Computer Graphics Forum. vol. 40, pp. 101–113. Wiley Online Library (2021)
26. Munkberg, J., Hasselgren, J., Shen, T., Gao, J., Chen, W., Evans, A., Müller, T., Fidler, S.: Extracting triangular 3d models, materials, and lighting from images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8280–8290 (2022)
27. Nam, G., Lee, J.H., Gutierrez, D., Kim, M.H.: Practical svbrdf acquisition of 3d objects with unstructured flash photography. ACM Transactions on Graphics (TOG) **37**(6), 1–12 (2018)
28. Riviere, J., Peers, P., Ghosh, A.: Mobile surface reflectometry. In: ACM SIGGRAPH 2014 Posters, pp. 1–1 (2014)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
31. Sang, S., Chandraker, M.: Single-shot neural relighting and svbrdf estimation. In: European Conference on Computer Vision. pp. 85–101. Springer (2020)
32. Schlick, C.: An inexpensive brdf model for physically-based rendering. In: Computer graphics forum. vol. 13, pp. 233–246. Wiley Online Library (1994)
33. Tunwattanapong, B., Fyffe, G., Graham, P., Busch, J., Yu, X., Ghosh, A., Debevec, P.: Acquiring reflectance and shape from continuous spherical harmonic illumination. ACM Transactions on graphics (TOG) **32**(4), 1–12 (2013)
34. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
35. Vecchio, G., Martin, R., Roullier, A., Kaiser, A., Rouffet, R., Deschaintre, V., Boubekeur, T.: Controlmat: A controlled generative approach to material capture. arXiv preprint arXiv:2309.01700 (2023)
36. Vecchio, G., Sortino, R., Palazzo, S., Spampinato, C.: Matfuse: Controllable material generation with diffusion models. arXiv preprint arXiv:2308.11408 (2023)
37. Walter, B., Marschner, S.R., Li, H., Torrance, K.E.: Microfacet models for refraction through rough surfaces. In: Proceedings of the 18th Eurographics conference on Rendering Techniques. pp. 195–206 (2007)
38. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7794–7803 (2018)
39. Wu, H., Wang, Z., Zhou, K.: Simultaneous localization and appearance estimation with a consumer rgb-d camera. IEEE transactions on visualization and computer graphics **22**(8), 2012–2023 (2015)
40. Xia, R., Dong, Y., Peers, P., Tong, X.: Recovering shape and spatially-varying surface reflectance under unknown illumination. ACM Transactions on Graphics (TOG) **35**(6), 1–12 (2016)

41. Xu, Z., Nielsen, J.B., Yu, J., Jensen, H.W., Ramamoorthi, R.: Minimal brdf sampling for two-shot near-field reflectance acquisition. ACM Transactions on Graphics (TOG) **35**(6), 1–12 (2016)
42. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
43. Zhao, Y., Wang, B., Xu, Y., Zeng, Z., Wang, L., Holzschuch, N.: Joint svbrdf recovery and synthesis from a single image using an unsupervised generative adversarial network. In: EGSR (DL). pp. 53–66 (2020)
44. Zhou, X., Kalantari, N.K.: Adversarial single-image svbrdf estimation with hybrid training. In: Computer Graphics Forum. vol. 40, pp. 315–325. Wiley Online Library (2021)