Workshop on Material Appearance Modeling (2018)
H. Rushmeier and R. Klein (Editors)

Deep Dual Loss BRDF Parameter Estimation

Mark Boss[1], Fabian Groh[1], Sebastian Herholz[1], Hendrik P. A. Lensch[1]

[1]University of Tübingen

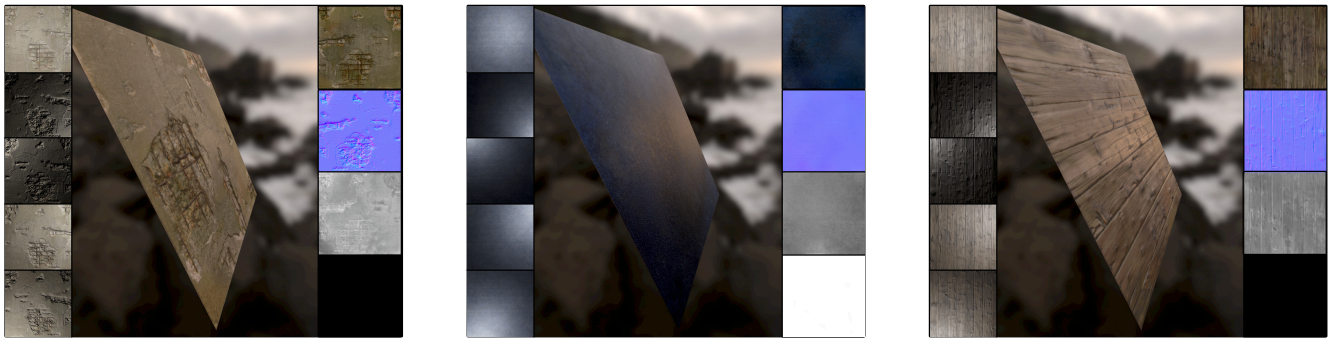

Figure 1: *The proposed neural network predicts spatially-varying BRDF from five input photographs with fixed viewing position and varying light positions. From left to right: input positions, rendered predictions and predicted parameters for three examples.*

Abstract

Surface parameter estimation is an essential field in computer games and movies. An exact representation of a real-world surface allows for a higher degree of realism. Capturing or artistically creating these materials is a time-consuming process. We propose a method which utilizes an encoder-decoder Convolutional Neural Network (CNN) to extract parameters for the Bidirectional Reflectance Distribution Function (BRDF) automatically from a sparse sample set. This is done by implementing a differentiable renderer, which allows for a loss backpropagation of rendered images. This photometric loss is essential because defining a numerical BRDF distance metric is difficult. A second loss is added, which compares the parameters maps directly. Therefore, the statistical properties of the BRDF model are learned, which reduces artifacts in the predicted parameters. This dual loss principal improves the result of the network significantly. Opposed to previous means this method retrieves information of the whole surface as spatially varying BRDF (SVBRDF) parameters with a sufficiently high resolution for intended real-world usage. The capture process for materials only requires five known light positions with a fixed camera position. This reduces the scanning time drastically, and a material sample can be obtained in seconds with an automated system.

Categories and Subject Descriptors (according to ACM CCS): I.3.3 [Computer Graphics]: Picture/Image Generation—Line and curve generation

1. Introduction

With the advance of processing power and improvements in rendering algorithms, movies and video games approach photorealism. The 3D models of characters and scenes are highly detailed, and the behavior of light is recreated realistically in rendering algorithms. To achieve this high level of realism, the correct behavior of surfaces is critical. These materials are often captured using photogrammetry or with a Bidirectional Texturing Function (BTF) measurement device. Either way is a time-consuming process.

However, artists are capable of reproducing information about reflective behavior from a few images under different light conditions by leveraging their previous knowledge about similar objects. At the same time, neural networks are reaching human performance in many areas such as speech or image recognition in the recent years [HZRS15, SSS*17]. We present a neural network

DOI: 10.2312/mam.20181199

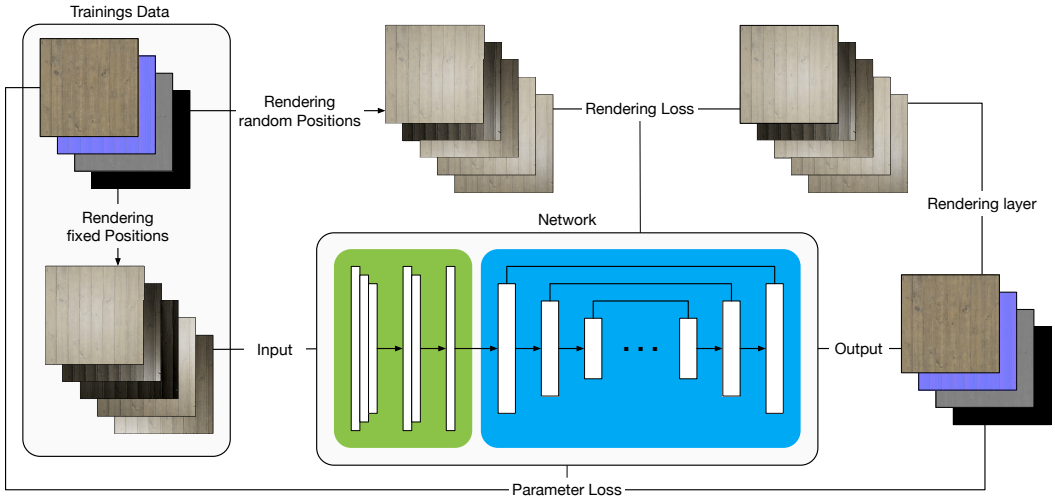

Figure 2: *Overview of the network architecture. The BRDF parameters from the dataset are rendered with the five fixed view and light positions. Additionally, the parameters are rendered during the training with ten random view and light positions. The five images with fixed positions are passed to the network. Here, the images are first processed in three 3D convolution layers (green) reducing the depth dimension from the five stacked images, and the result is then passed to the U-Net (blue). Finally, the predictions are rendered with the light and view positions from the random images in a rendering layer. An error is calculated between the ground truth and re-rendered predicted loss images. Additionally, the predicted BRDF parameters are compared to the ground truth maps. The loss from both errors is combined.*

which predicts SVBRDF parameters for isotropic materials from multiple input images. The Cook-Torrance model [CT82] with influences from the Disney BRDF [Bur12] is selected as the underlying BRDF model. The key contributions of this work are:

- BRDF estimation from five input images with a fixed viewing angle but varying light positions with an encoder-decoder CNN
- A network architecture with a dual loss defined as a:
  - Loss against rendered images from random view and light positions using a differentiable renderer, which is used to define a Cook-Torrance specific loss term
  - Loss against the ground truth parameters, which learns statistical properties of commonly used BRDF parameter sets and thus reduces artifacts in the predictions

2. Related Work

The concept of decreasing measurement time for BRDFs is an active area of research. Several approaches achieving a high speedup are developed recently [AWL15,LDPT17,NJR15]. As materials are often related to each other, several methods which leverage this property are developed. Nielsen et al. [NJR15] perform a Principal Component Analysis (PCA) on the MERL dataset [MPBM03] to search for optimal sampling positions and use this knowledge to reconstruct the BRDF from the sparse sampling positions. The results of this work are homogeneous BRDF parameters.

Aittala et al. [AWL15] use the property that materials are often stationary, which means they consist of repeating patterns. An image with a flash and one without are taken and divided into a grid. A single tile is fitted against the other tiles, and thus multiple half vectors between view and light are calculated. The result is an SVBRDF from a small low-resolution area of the material.

Li et al. [LDPT17] are the first, who explore the possibility of using CNNs for this problem. They use an encoder-decoder CNN with skip connections to estimate BRDF parameters from a single flash image. The result of this method is spatially varying information about the diffuse and normal parameters and homogeneous parameters for the roughness and specular information.

3. Network Architecture

The general task in our framework is to extract an SVBRDF from five images of a surface with different lighting conditions. The used BRDF is an analytical Cook-Torrance microfacet model [CT82] in combination with the diffuse term and the metallic property of the Disney BRDF [Bur12]. In this case, the BRDF is represented by its parameters, which are stored in a multichannel image. In general, this is a transformation between different styles of images. The U-net architecture from Ronneberger et al. [RFB15] is a well-suited architecture for these kinds of problems [IZZE,ZLW18].

3.1. Dual Loss Formulation

When designing a network architecture, the error metric is an important part. Only when done correctly, the predictions fulfill the desired purpose. Since we want to use the predicted BRDF parameters under arbitrary light conditions and viewing angles, we have two main constraints: First, the rendered prediction should be plausible and match the input data. Second, The predicted parameters should be similar to ones generated by an artist. Therefore, we propose a dual loss, which encompasses these constraints.

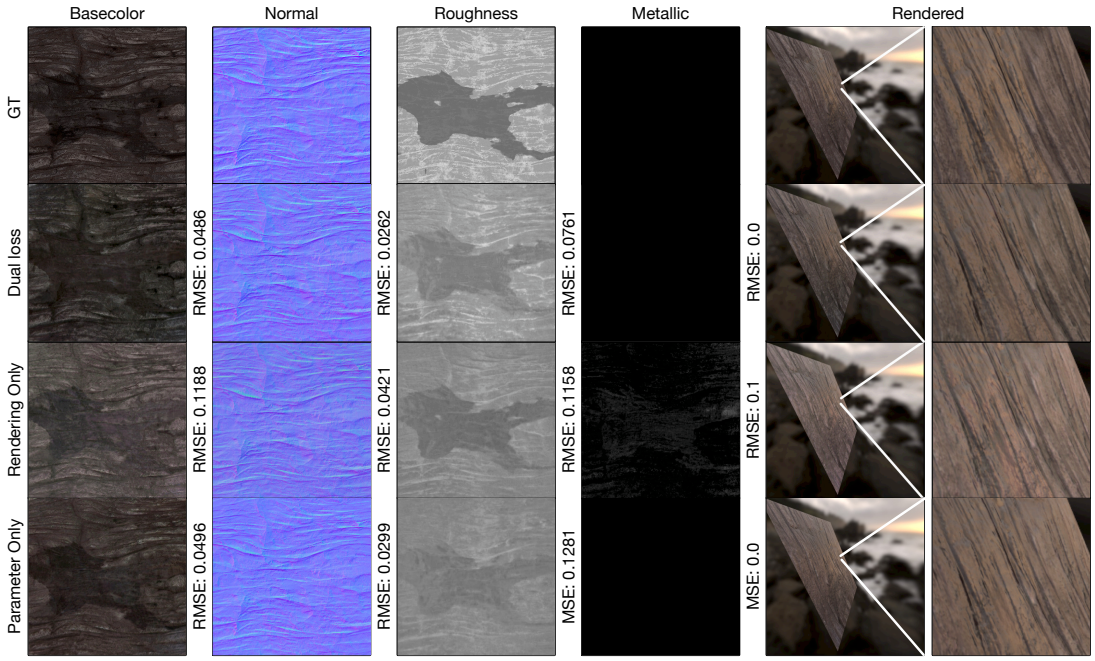

Figure 3: *Comparison for a predicted rock material. The "Parameter Only"-loss is calculated only on the parameter maps and the "Rendering Only"-loss is calculated on the rendered loss images. The dual loss is a joined loss between a photometric rendering loss and the direct parameter loss.*

Rendering Loss This loss term is based on the Mean Absolute Error (MAE) between a ground truth and predicted rendered image. To backpropagate the error to the network weights, we use a differentiable render, which is directly implemented with Tensorflow operations. The renderer generates images using the predicted parameters from ten random view and light positions. For five of these positions, a sharp highlight is enforced by first selecting the light position and then mirroring the light ray at the surface normal from a random point on the surface. The ten ground truth and re-rendered images are transformed with the formula: $\log(x+1)$. Thereby high specular peak values of the renderings are reduced.

Parameter Loss The rendering loss can introduce artifacts from specular highlights, due to lost and hard to recover information in these areas. This violates both constraints by neither matching artists generated parameter, nor matching the input and being plausible. By calculating the MAE between the predicted BRDF parameters and the ground truth, the network learns the statistical properties of commonly used BRDFs.

3.2. Encoder-Decoder Network

The U-Net architecture is the basis of the network design with inspiration from the specific implementation details of Isola et al. [IZZE]. As seen in Figure 2 the input of the network consists of five low dynamic range images. The resolution of the input and output is 512×512. The output produces an image with eight channels corresponding to the SVBRDF parameters: three for the RGB basecolor, three for the normal vector, one for the roughness and one for the metallic mask. The metallic mask is used to split the base color in diffuse and specular color.

To extend the original implementation of Ronneberger et al. [RFB15] and Isola et al. [IZZE] to multiple inputs, we stack our five input images in a depth dimension. Three 3D convolution layers with a kernel size of three are used to merge the inputs images with a stride of two in the depth dimension. The feature counts for each of these convolutions are 16, 32, 32.

The merged input is now passed through nine convolution layers, which perform the downsampling, followed by nine transposed convolution layers for upsampling. Each of the encoder layers decreases the spatial resolution but increases the feature size, and each decoder layer the other way around. This hourglass shape forces the encoder to compress the information into a global feature vector. Additional skip connections are added between encoder and decoders layers of matching size to reconstruct spatial details [RFB15].

In detail, the encoder outputs 64, 128, 256, and 512 features in the following layers. The downsampling is implemented with a stride of two in the convolutions. The decoder uses the same feature outputs but in reversed order. A stride of two in the transposed convolutions is used for upsampling. In the four coarsest layer of the decoder, a dropout is applied. A kernel size of four, batch normalization and a leaky ReLu with a 0.2 weight are used throughout the encoder-decoder network. To output, the BRDF parameter the feature size of the last transposed convolution layer is set to eight and the sigmoid nonlinearity is used to achieve the [0, 1] range.

4. Training and Dataset

The required training data is gathered from the following libraries: textures.com, cgbookcase.com, cc0textures.com, freepbr.com, poliigon.com, and 3d-wolf.com. These libraries provide 815 high-resolution samples. Each sample is randomly rotated, scaled, and cropped to 512×512 pixels. Additionally, the hue, saturation, brightness, and contrast is adjusted randomly. Afterward, every material is blended with another material. The total number of samples in the dataset is 40750 samples afterward. The network is trained on this dataset with a batch size of four in 249.000 steps, and the Adam optimizer [KB14] is used with a learning rate of 0.001 for the first 100 epochs and 0.0001 afterward. The training takes three days on a single Nvidia 1080 TI.

5. Evaluation

For evaluation, a small dataset of 29 handpicked, challenging materials is created. Over the whole test set, the RMSE (Root-Mean-Square-Error) is 0.0808 for the basecolor, 0.0201 for the normal, 0.0343 for the roughness, and 0.0629 for the metallic parameter map. As seen the network is capable of estimating normal maps with near-perfect accuracy. In Figure 3 a rock material is compared to the ground truth with both single losses and the dual loss. Here it is visible, that the "Rendering Only"-loss is introducing errors due to the difficulty of attributing specular behavior for a specific map. This is especially noticeable in the metallic parameter map. The "Parameter Only"-loss is not displaying this problem. The combination of both loss formulations produces the best visual and error metric results. The Mean Squared Error for each map and loss are annotated in Figure 3.

6. Conclusion

We propose a framework for automated BRDF estimation which only depends on five input images with a fixed viewing position. An error metric which takes the rendering context into account is added with a differentiable renderer. The resulting SVBRDF parameter maps are estimated at a reasonably high resolution and can be used in video games and movie productions. Due to the low number of input images and the fixed viewing position, the capture time is low, allowing simple setups without motorized parts. This allows designing different scanning devices in the future.

References

[AWL15] AITTALA M., WEYRICH T., LEHTINEN J.: Two-shot svbrdf capture for stationary materials. *ACM Trans. Graph. 34*, 4 (July 2015), 110:1–110:13. 2

[Bur12] BURLEY B.: Physically based shading at disney. In *SIGGRAPH 2012* (2012), vol. Course Notes. 2

[CT82] COOK R. L., TORRANCE K. E.: A reflectance model for computer graphics. *ACM Transactions on Graphics 1*, 1 (1982), 7–24. 2

[HZRS15] HE K., ZHANG X., REN S., SUN J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR abs/1502.01852* (2015). 1

[IZZE] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. 2, 3

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *CoRR abs/1412.6980* (2014). 4

[LDPT17] 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 36*, 4 (jul 2017), 1–11. 2

[MPBM03] MATUSIK W., PFISTER H., BRAND M., MCMILLAN L.: A data-driven reflectance model. *ACM Transactions on Graphics 22*, 3 (July 2003), 759–769. 2

[NJR15] NIELSEN J. B., JENSEN H. W., RAMAMOORTHI R.: On optimal, minimal BRDF sampling for reflectance acquisition. *ACM Transactions on Graphics 34*, 6 (oct 2015), 1–11. 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. *CoRR* (2015). 2, 3

[SSS*17] SILVER D., SCHRITTWIESER J., SIMONYAN K., ANTONOGLOU I., HUANG A., GUEZ A., HUBERT T., BAKER L., LAI M., BOLTON A., CHEN Y., LILLICRAP T., HUI F., SIFRE L., VAN DEN DRIESSCHE G., GRAEPEL T., HASSABIS D.: Mastering the game of go without human knowledge. *Nature 550* (Oct. 2017), 354–. 1

[ZLW18] ZHANG Z., LIU Q., WANG Y.: Road extraction by deep residual u-net. *IEEE Geoscience and Remote Sensing Letters 15*, 5 (May 2018), 749–753. 2