Material acquisition using deep learning

Valentin Deschaintre
Université Côte d'Azur, Inria, Ansys,
valentin.deschaintre@inria.fr

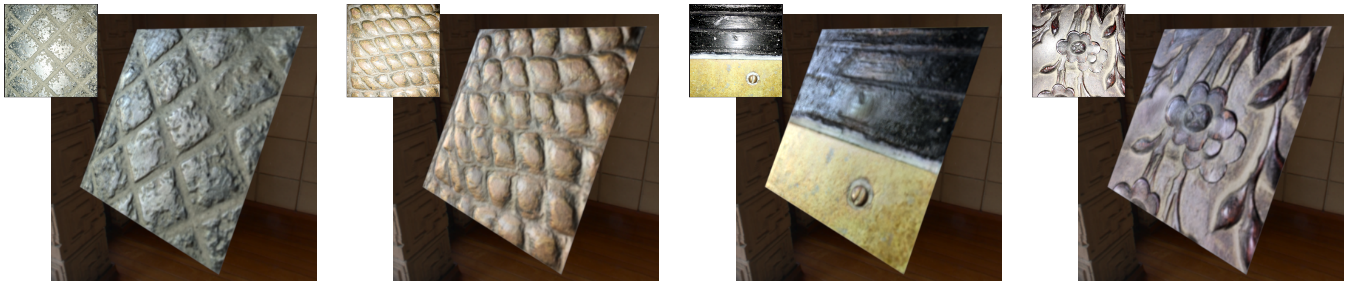

Figure 1: Re-renderings of materials captured using a single flash picture (top-left) and our deep acquisition method.

ABSTRACT

Texture, highlights, and shading are some of many visual cues that allow humans to perceive material appearance in pictures. Designing algorithms able to leverage these cues to recover spatially-varying bi-directional reflectance distribution functions (SVBRDFs) from a few images has challenged computer graphics researchers for decades. I explore the use of deep learning to tackle lightweight appearance capture and make sense of these visual cues. Our networks are capable of recovering per-pixel normals, diffuse albedo, specular albedo and specular roughness from as little as one picture of a flat surface lit by a hand-held flash. We propose a method which improves its prediction with the number of input pictures, and reaches high quality reconstructions with up to 10 images – a sweet spot between existing single-image and complex multi-image approaches. We introduce several innovations on training data acquisition and network design, bringing clear improvement over the state of the art for lightweight material capture.

CCS CONCEPTS

• **Computing methodologies → Reflectance modeling**; *Image processing*.

KEYWORDS

Material capture, Appearance capture, SVBRDF, Deep learning

ACM Reference Format:
Valentin Deschaintre. 2018. Material acquisition using deep learning. *ACM Trans. Graph.* 37, 4, Article 128 (August 2018), 4 pages. https://doi.org/10.1145/3197517.3201378

Author's address: Valentin Deschaintre, Université Côte d'Azur, Inria, Ansys, valentin.deschaintre@inria.fr.

0730-0301/2018/8-ART128 $15.00
https://doi.org/10.1145/3197517.3201378

1 INTRODUCTION

The appearance of real-world objects results from complex interactions between light, reflectance, and geometry. Disentangling these interactions is at the heart of *lightweight appearance capture*, which aims at recovering reflectance functions from one or a few photographs of a surface. This task is inherently ill-posed, since many different reflectances can yield the same observed image. For example, any photograph can be perfectly reproduced by a diffuse albedo map, where highlights are "painted" over the surface. A combination of two strategies is generally employed to deal with this ill-posedness. First, ambiguity can be reduced by collecting additional measurements under different viewing or lighting conditions. While this strategy is currently the most appropriate to achieve high accuracy, it requires precise control of the acquisition process [Xu et al. 2016]. The second strategy is to introduce *a priori* assumptions about the space of plausible solutions. While designing such priors by hand has challenged researchers for decades [Guarnera et al. 2016], Convolutional Neural Networks (CNNs) have emerged as a powerful method to automatically *learn* effective priors from data.

We propose a deep learning approach to lightweight appearance capture, where we use *forward* rendering simulations to train a neural network to solve the ill-posed *inverse* problem of estimating a spatially-varying bi-directional reflectance distribution function (SVBRDF) from one or a few photographs of a flat surface lit by a hand-held flash. While our method shares ingredients with recent work on material capture [Li et al. 2017; Rematas et al. 2017], material editing [Liu et al. 2017], and other image-to-image translation tasks [Isola et al. 2017], achieving high-quality SVBRDF estimation requires several key innovations on training data acquisition and neural network design.

The task of our deep networks is to predict four maps corresponding to *per-pixel* normal, diffuse albedo, specular albedo, and specular roughness of a planar material sample using near-field flash-lit photographs as input. Flash photographs are easy to acquire, and have been shown to contain a lot of information that can be leveraged in inferring the material properties from one [Aittala et al. 2016] or multiple images [Riviere et al. 2016].

This work is a summary of our recent contributions to the light-weight material capture domain [Deschaintre et al. 2018, 2019]. By combining Deep Learning, Computer Vision and Computer Graphics knowledge we design convenient material acquisition systems and improve the state of the art results quality.

2 PROCEDURAL SYNTHESIS OF TRAINING DATA

While several recent papers have shown the potential of synthetic data to train neural networks [Richter et al. 2016; Su et al. 2015], care must be taken to generate data that is representative of the diversity of real-world materials we want to capture. We address this challenge by leveraging Allegorithmic Substance Share [Allegorithmic 2019], a dataset of more than 800 procedural SVBRDFs designed by a community of artists from the movie and video game industry.

We curated a set of 155 high-quality procedural SVBRDFs from 9 material classes from which we generated around 1,850 variants by applying random perturbations to their important parameters.

Instead of pre-rendering a fixed set of training data, we implemented our own SVBRDF renderer in TensorFlow, so that it can be called at each iteration of the training process. At each step, we augment our dataset through random crop and convex combinations of random pairs of SVBRDFs by using α-blending on their maps. The mixing greatly increases the diversity of low-level shading effects in the training data, while staying close to the set of plausible real-world materials, drastically reducing the chance of the network seeing the same material twice.

While rendering our training data on the fly incurs additional computation, we found that this overhead is compensated by the time gained in data loading. In our experiments, training our system with online data generation takes approximately as much time as training it with pre-computed data stored on disk.

3 ONE IMAGE NETWORK ARCHITECTURE

Our problem boils down to translating a photograph of a material into a corresponding SVBRDF map representation, which can be represented as a multi-channel image. The *U-Net* architecture [Ronneberger et al. 2015] has proven to be well suited for a wide range of similar image-to-image translation tasks [Isola et al. 2017]. However, our early experiments revealed that despite its multi-scale design, this architecture remains challenged by tasks requiring the fusion of distant visual information. We address this limitation by complementing the U-Net with a parallel *global features* network tailored to capture and propagate global information.

3.1 U-Net Image-to-Image Network

We adopt the U-Net architecture as the basis of our network design, and follow Isola et al. [2017] for most implementation details. Note however that we do not use their *discriminator* network, as we did not find it to yield a discernible benefit in our problem.

Our base network takes a 3-channel photograph as input and outputs a 9-channel image of SVBRDF parameters – 3 channels for the RGB diffuse albedo, 3 channels for the RGB specular albedo, 2 channels for the x and y components of the normal vector in tangent plane parameterization, and 1 channel for the specular roughness. The input image is processed through a sequence of 8 convolutional layers that perform downsampling (the encoder), followed by a sequence of 8 upsampling and convolutional layers (the decoder).

3.2 Global Features Network

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, such as the recent work of Aittala et al. [2015] where spatial repetitions in the material sample are seen as multiple observations of a similar SVBRDF patch. Taking inspiration from these successful heuristics, we aim for a network architecture capable of leveraging redundancies present in the data.

The hourglass shape of the U-Net results in large footprints of the convolution kernels at coarse spatial scales, which in theory provide long-distance dependencies between output pixels. Unfortunately, we found that this multi-scale design is not sufficient to properly fuse information for our problem. We hypothesize that the ability of the network to compute global information is partly hindered by instance (or batch) normalization, which standardizes the learned features after every convolutional layer by enforcing a mean and standard deviation learned from training data.

We propose a network architecture that simultaneously addresses both of these shortcomings. We add a parallel network track alongside the U-Net, which deals with *global* feature vectors instead of 2D feature maps. The global and convolutional tracks exchange information after every layer.

Each pair of these information exchanges forms a nonlinear dependency between every pixels, providing the network with means to arrive at a consistent solution by repeatedly transmitting local findings between different regions.

3.3 Rendering Loss

The parameterizations of popular BRDF models arise from a combination of mathematical convenience and relative intuitiveness for artists, and the numerical difference between the parameter values of two (SV)BRDFs is only weakly indicative of their visual similarity.

We therefore propose a loss function that is *independent* of the parameterization of either the predicted or the target SVBRDF, and instead compares their *rendered appearance*. Specifically, any time the loss is evaluated, both the ground truth SVBRDF and the predicted SVBRDF are rendered under identical illumination and viewing conditions, and the resulting images are compared pixel-wise. We use the same Cook-Torrance BRDF model [1982] for the ground truth and prediction, but our loss function could equally be used with representations that differ between these two quantities.

We use our in-tensorflow renderer to implement the rendering loss. This strategy has the benefits of seamless integration with the neural network training, automatically-computed derivatives, and automatic GPU acceleration.

Using a fixed finite set of viewing and lighting directions would make the loss blind to much of the angular space. Instead, we formulate the loss as the average error over *all* angles, and follow the common strategy of evaluating it stochastically by choosing the angles at random for every training sample, in the spirit of stochastic gradient descent.

We compare the logarithmic values of the renderings using the l_1 norm. The logarithm is used to control the potentially extreme dynamic range of specular peaks, and because we are more concerned with relative than absolute errors.

While the training a network with only a l_1 loss produces plausible maps when considered in isolation, these maps do not reproduce the appearance of the ground truth once re-rendered. In contrast, the rendering loss yields a more faithful reproduction of the ground truth appearance.

4 MULTI-IMAGE CAPTURE

While we are able to achieve plausible results with a single image in many cases, one image is often not enough to capture the rich appearance of real-world material. We propose an architecture capable of aggregating the information available in any number of pictures, while maintaining a high level of convenience.

4.1 Capture Setup

We designed our method to take as input a variable number of images, captured under uncalibrated light and view directions. We consider two capture setups where we place the material sample within a white paper frame and capture it by holding a smartphone in one hand and a flash in the other, or by using the flash of the smartphone as a co-located light source. Similarly to Hui et al. [2017], we use the four corners of the frame to compute an homography that rectifies the images, and crop the paper pixels away before processing the images with our method.

4.2 Multi-image architecture

Since we are targeting a lightweight capture scenario, we do not provide the network with any explicit knowledge of the light and view position. We rather count on the network to deduce related information from visual cues. The core of our method is a multi-image network composed of several copies of our single-image network -described in 3). The number of copies is dynamically chosen to match the number of inputs provided by the user (or the training sample). All copies are identical in their architecture and weights, meaning that each input receives an identical treatment by its respective network copy. The findings from each single-image network are then fused by a common order-agnostic pooling layer before being processed into a joint estimate of the SVBRDF.

4.3 Multi-image fusion

The second part of our architecture fuses the multiple feature maps produced by the single-image networks to form a single feature map of fixed size.

Specifically, the encoder-decoder track of each single-image network produces a $256 \times 256 \times 64$ intermediate feature map corresponding to the input image it processed. These maps are fused into a single joint feature map of the same size by picking the maximum value reported by any single-image network at each pixel and feature channel. This max-pooling procedure gives every single-image network equal means to contribute to the content of the joint feature map in a perfectly order-independent manner [Aittala and Durand 2018].

The pooled intermediate feature map is finally decoded by 3 layers of convolutions and non-linearities, which provide the network sufficient expressivity to transform the extracted information into four SVBRDF maps. The global features in the fully-connected tracks are max-pooled and decoded in a similar manner.

5 RESULTS

We show results of our one image method on two different materials with strong spatial variations in Fig. 2. We illustrate how the inferred material quality improves over more inputs provided in Fig. 3.

6 CONCLUSION

In my thesis work, I present a number of contributions to the lightweight material acquisition problem. We introduce two network architectures allowing to conveniently capture real world materials with as little as one image, while being able to leverage more information if available.

REFERENCES

Miika Aittala, Timo Aila, and Jaakko Lehtinen. 2016. Reflectance Modeling by Neural Texture Synthesis. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 35, 4 (2016).

Miika Aittala and Fredo Durand. 2018. Burst Image Deblurring Using Permutation Invariant Convolutional Neural Networks. In *The European Conference on Computer Vision (ECCV)*.

Miika Aittala, Tim Weyrich, and Jaakko Lehtinen. 2015. Two-shot SVBRDF Capture for Stationary Materials. *ACM Trans. Graph. (Proc. SIGGRAPH)* 34, 4, Article 110 (July 2015), 13 pages. https://doi.org/10.1145/2766967

Allegorithmic. 2019. Substance Share. (2019). https://share.substance3d.com/

R. L. Cook and K. E. Torrance. 1982. A Reflectance Model for Computer Graphics. *ACM Transactions on Graphics* 1, 1 (1982), 7–24.

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2018. Single-Image SVBRDF Capture with a Rendering-Aware Deep Network. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings)* 37, 128 (aug 2018), 15. http://www-sop.inria.fr/reves/Basilic/2018/DADDB18

Valentin Deschaintre, Miika Aittala, Frédo Durand, George Drettakis, and Adrien Bousseau. 2019. Flexible SVBRDF Capture with a Multi-Image Deep Network. *Computer Graphics Forum (Proceedings of the Eurographics Symposium on Rendering)* 38, 4 (July 2019). http://www-sop.inria.fr/reves/Basilic/2019/DADDB19

Dar'ya Guarnera, Giuseppe Claudio Guarnera, Abhijeet Ghosh, Cornelia Denk, and Mashhuda Glencross. 2016. BRDF Representation and Acquisition. *Computer Graphics Forum* (2016).

Z. Hui, K. Sunkavalli, J. Y. Lee, S. Hadap, J. Wang, and A. C. Sankaranarayanan. 2017. Reflectance Capture Using Univariate Sampling of BRDFs. In *IEEE International Conference on Computer Vision (ICCV)*.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xiao Li, Yue Dong, Pieter Peers, and Xin Tong. 2017. Modeling Surface Appearance from a Single Photograph using Self-augmented Convolutional Neural Networks. *ACM Transactions on Graphics (Proc. SIGGRAPH)* 36, 4 (2017).

Guilin Liu, Duygu Ceylan, Ersin Yumer, Jimei Yang, and Jyh-Ming Lien. 2017. Material Editing Using a Physically Based Rendering Network. In *IEEE International Conference on Computer Vision (ICCV)*. 2261–2269.

K. Rematas, S. Georgoulis, T. Ritschel, E. Gavves, M. Fritz, L. Van Gool, and T. Tuytelaars. 2017. Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2017).

Stephan R. Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for Data: Ground Truth from Computer Games. In *Proc. European Conference on Computer Vision (ECCV)*.

J. Riviere, P. Peers, and A. Ghosh. 2016. Mobile Surface Reflectometry. *Computer Graphics Forum* 35, 1 (2016).

O. Ronneberger, P.Fischer, and T. Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI) (LNCS)*, Vol. 9351. 234–241.

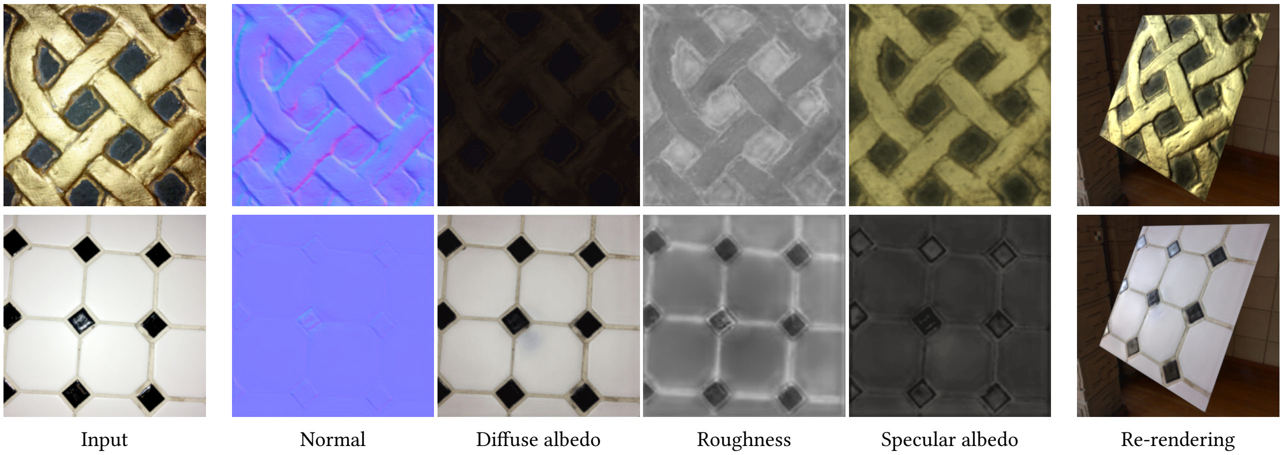

Figure 2: Based on the input photographs (left), our method has recovered a set of SVBRDF maps that exhibit strong spatially varying specular roughness and albedo effects. The gold-colored paint (top) and the highly glossy black tiles (bottom) are clearly visible in the re-renderings of SVBRDF under environment illumination (right).

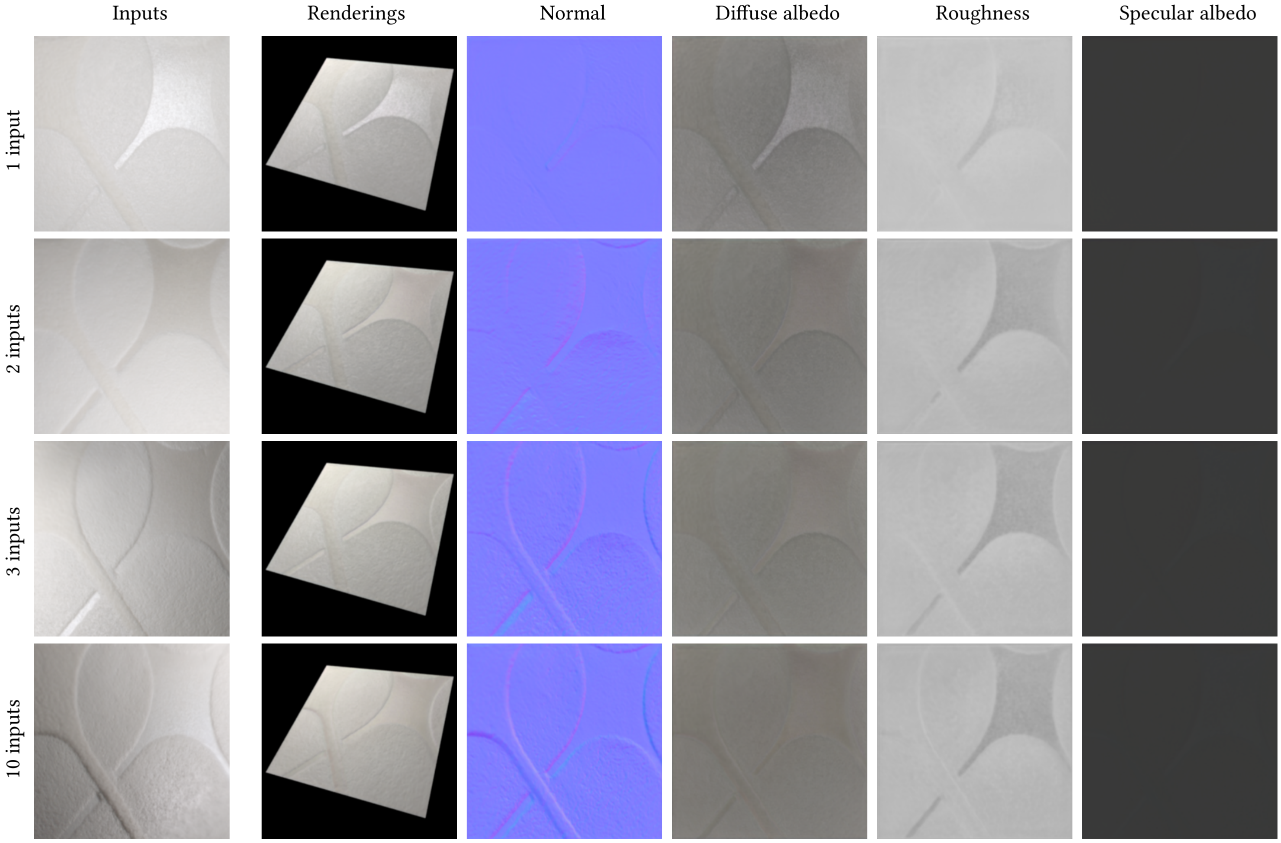

Figure 3: Evaluation on a measured BTF. Three images are enough to capture most of normal and roughness maps. Adding images further improves the result by removing lighting residual from the diffuse albedo, and adding subtle details to the normal and specular maps.

Hao Su, Charles R. Qi, Yangyan Li, and Leonidas J. Guibas. 2015. Render for CNN: Viewpoint Estimation in Images Using CNNs Trained with Rendered 3D Model Views. In *The IEEE International Conference on Computer Vision (ICCV).*

Zexiang Xu, Jannik Boll Nielsen, Jiyang Yu, Henrik Wann Jensen, and Ravi Ramamoorthi. 2016. Minimal BRDF Sampling for Two-shot Near-field Reflectance Acquisition. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)* 35, 6 (2016).