Flexible SVBRDF Capture with a Multi#Image Deep Network

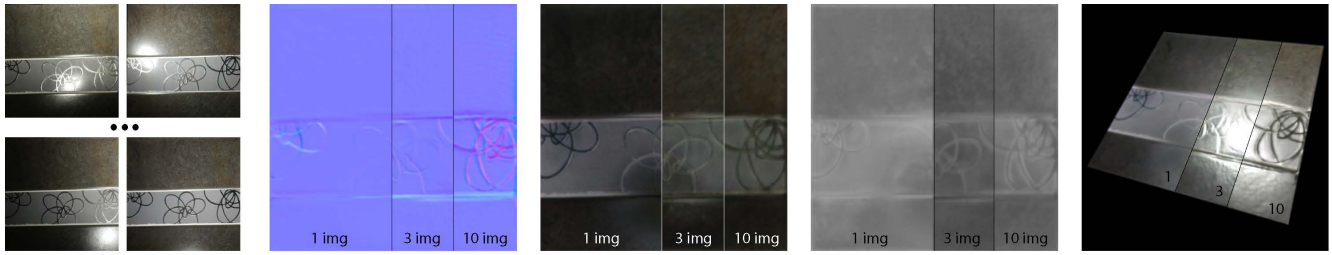
Eurographics Symposium on Rendering 2019
T. Boubekeur and P. Sen
(Guest Editors)

Volume 38 (2019), Number 4

Flexible SVBRDF Capture with a Multi-Image Deep Network

Valentin Deschaintre[1,3], Miika Aittala[2], Fredo Durand[2], George Drettakis[1] and Adrien Bousseau[1]

[1] Université Côté d'Azur, Inria
[2] MIT CSAIL
[3] Optis for Ansys



(a) Input images (b) Normals (c) Diffuse albedo (d) Roughness (e) Re-rendering

Figure 1: *Our deep learning method for SVBRDF capture supports a variable number of input photographs taken with uncalibrated light-view directions (a, rectified). While a single image is enough to obtain a first plausible estimate of the SVBRDF maps, more images provide new cues to our method, improving its prediction. In this example, adding images reveals fine normal variations (b), removes highlight residuals in the diffuse albedo (c), and reveals the difference of roughness between the stone, the stripe, and the thin pattern (d). Please see supplemental materials for animated re-renderings.*

Abstract

Empowered by deep learning, recent methods for material capture can estimate a spatially-varying reflectance from a single photograph. Such lightweight capture is in stark contrast with the tens or hundreds of pictures required by traditional optimization-based approaches. However, a single image is often simply not enough to observe the rich appearance of real-world materials. We present a deep-learning method capable of estimating material appearance from a variable number of uncalibrated and unordered pictures captured with a handheld camera and flash. Thanks to an order-independent fusing layer, this architecture extracts the most useful information from each picture, while benefiting from strong priors learned from data. The method can handle both view and light direction variation without calibration. We show how our method improves its prediction with the number of input pictures, and reaches high quality reconstructions with as little as 1 *to* 10 *images – a sweet spot between existing single-image and complex multi-image approaches.*

CCS Concepts

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

Keywords: Material capture, Appearance capture, SVBRDF, Deep learning

This paper is a low resolution version of our full paper, available here : https://www-sop.inria.fr/reves/Basilic/2019/DADDB19/.

1. Introduction

The appearance of most real-world materials depends on both viewing and lighting directions, which makes their capture a challenging task. While early methods achieved faithful capture by densely sampling the view-light conditions [Mca02, DVGNK99], this exhaustive strategy requires expensive and time-consuming hardware setups. In contrast, lightweight methods attempt to only perform a few measurements, but require strong prior knowledge on the solution to fill the gaps. In particular, recent methods produce convincing spatially-varying material appearances from a single flash photograph thanks to deep neural networks trained from large quantities of synthetic material renderings [DAD*18,LSC18].

However, in many cases a single photograph simply does not contain enough information to make a good inference for a given material. Figure 1(b-d) illustrates typical failure cases of single-image methods, where the flash lighting provides insufficient cues of the relief of the surface, and leaves highlight residuals in the diffuse albedo and specular maps. Only additional pictures with side views or lights reveal fine geometry and reflectance details.

We propose a method that leverages the information provided by additional pictures, while retaining a lightweight capture procedure. When few images are provided, our method harnesses the power of learned priors to make an educated guess, while when additional images are available, our method improves its prediction to best explain all observations. We achieve this flexibility thanks to a deep network architecture capable of processing an arbitrary number of input images with uncalibrated light-view directions. The key observation is that such image sets are fundamentally unstructured. They do not have a meaningful ordering, nor a predetermined type of content for any given input. Following this reasoning, we adopt a pooling-based network architecture that treats the inputs in a perfectly order-invariant manner, giving it powerful means to extract and combine subtle joint appearance cues scattered across the inputs.

Our flexible approach allows us to capture spatially-varying materials with 1 to 10 images, providing a significant improvement over single-image methods while requiring much fewer images and less constrained capture than traditional multi-image methods.

2. Related Work

We first review prior work on appearance capture, focusing on methods working with few images. We then discuss deep learning methods capable of processing multiple images.

Appearance capture. The problem of acquiring real-world appearance has been extensively studied in computer graphics and computer vision, as surveyed by Guarnera et al. [GGG*16]. Early efforts focused on capturing appearance under controlled view and lighting conditions, first using motorized point lights and cameras [Mca02, DVGNK99] and later using complex light patterns such as linear light sources [GTHD03], spherical gradients [GCP*09], Fourier basis [AWL13], or deep-learned patterns [KCW*18]. While these methods provide high-quality capture of complex material effects – including anisotropy, they require tens to hundreds of measurements acquired using dedicated hardware. In contrast, recent work manages to recover plausible spatially-varying appearance (SVBRDF) from very few pictures by leveraging strong priors on natural materials [WSM11, AWL15, AAL16, RWS*11, DWT*10, HSL*17] and lighting [LN16, DCP*14, RRFG17]. In particular, deep learning is nowadays the method of choice to automatically build priors from data, which allows the most recent methods to only use one picture to recover a plausible estimate of the spatially-varying appearance of flat samples [LDPT17, YLD*18, DAD*18, LSC18], and even the geometry of isolated objects [LXR*18]. However, while impressive in many cases, the solutions produced by these single-image methods are largely driven by the learned priors, and often fail to reproduce important material effects simply because they are not observed in the image provided as input, or are too ambiguous to be accurately identified without additional observations. We address this limitation by designing an architecture that supports an arbitrary number of input images. Compared to existing single-image methods [LDPT17, YLD*18, DAD*18, LSC18], our multi-image approach produces results of increasing quality as more images are provided. Compared to optimization-based multi-image methods [RPG16, HSL*17], our deep-learning approach requires much fewer images to produce high-quality solutions – 1 to 10 instead of around a hundred, while retaining much of the convenience of handheld capture. Nevertheless, the lightweight nature of our method makes it hard to reach the accuracy of solutions based on calibrated view and light conditions.

Multi-image deep networks. Many computer vision tasks become better posed as the number of observations increases, which calls for methods capable of handling a variable number of input images. For example, classical optimization approaches assign a data fitting error to each observation and minimize their sum. However, implementing an analogous strategy in a deep learning context remains a challenge because most neural network architectures, such as the popular U-Net used in prior work [LDPT17, DAD*18, LSC18], require inputs of a fixed size and treat these inputs in an asymmetric manner. These architectures thus cannot simultaneously benefit from powerful learned priors as well as multiple unstructured observations.

Choy et al. [CXG*16] faced this challenge in the context of multi-view 3D reconstruction and proposed a recurrent architecture that processes a sequence of images to progressively refine its prediction. However, the drawback of such an approach is that the solution still depends on the order in which the images are provided to the method – the first image has a great impact on the overall solution, while subsequent images tend to only modify details. This observation motivated Wiles et al. [WZ17] to process each image of a multi-view set through separate encoders before combining their features through max-pooling, an order-agnostic operation. Aittala et al. [AD18] and Chen et al. [CHW18] apply a similar strategy to the problems of burst image deblurring and photometric stereo, respectively. In the field of geometry processing, Qi et al. [QSMG17] also apply a pooling scheme for deep learning on point sets, and show that such an architecture is an universal approximator for functions whose inputs are set-valued. Zaheer et al. [ZKR*17] further analyze the theoretical properties of pooling architectures and demonstrate superior performance over recurrent architectures on multiple tasks involving loosely-structured set-valued input data. We build on this family of work to offer a method that processes images captured in an arbitrary order, and that can handle uncalibrated viewing and lighting conditions.

3. Capture Setup

We designed our method to take as input a variable number of images, captured under uncalibrated light and view directions. Figure 2 shows the capture setup we experimented with, 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 Paterson et al. [PCF05] and Hui et al. [HSL*17], 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. We capture pictures of 3456×3456 pixels and resize them to 256×256 pixels after cropping.

4. Multi-Image Material Inference

Our goal is to estimate the spatially-varying bi-directional reflectance distribution function (SVBRDF) of a flat material sample given a few aligned pictures of that sample. We adopt a parametric representation of the SVBRDF in the form of four maps representing the per-pixel surface normal and diffuse albedo, specular albedo and specular roughness of a Cook-Torrance [CT82] BRDF model.

The core of our method is a multi-image network composed of several copies of a single-image network , as illustrated in Figure 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 subsequently processed into a joint estimate of the SVBRDF.

We now detail the single-image network and the fusion mechanism, before describing the loss we use to compare the network prediction against a ground-truth SVBRDF. We detail our generation of synthetic training data in Section 5.

The source code of our network architecture along with pre-trained weights is available at https://team.inria.fr/graphdeco/projects/multi-materials/

4.1. Single-image network

We base our architecture on the single-image network of Deschaintre et al. [DAD*18], which was designed for a similar material acquisition task. The network follows the popular U-Net encoder-decoder architecture [RPB15], to which it adds a fully-connected track responsible for processing and transmitting global information across distant pixels. While the original architecture outputs four SVBRDF maps, we modify its last layer to instead output a 64-channel feature map, which retains more information to be processed by the later stages of our architecture. We also provide pixel coordinates as extra channels to the input to help the convolutional network reason about spatial information [LLM*18, LSC18].

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.

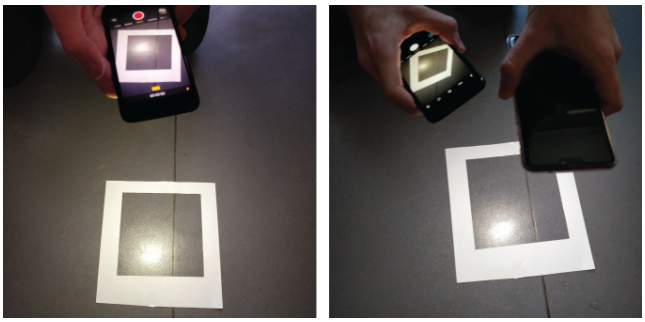


Figure 2: *We use a simple paper frame to help register pictures taken from different viewpoints. We use either a single smartphone and its flash, or two smartphones to cover a larger set of view/light configurations.*

4.2. 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 [AD18, CHW18].

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. Through end-to-end training, the single-image networks learn to produce features which are meaningful with respect to the pooling operation and useful for reconstructing the final estimate.

While we vary the number of copies of the single-view network between 1 and 5 during training, an important property of this architecture is that it can process an arbitrarily large number of images during testing because all copies share the same weights, and are ultimately fused by the pooling layer to form a fixed-size feature map. In our experiments, we vary the number of input images from 1 to 10 at testing time.

4.3. Loss

We evaluate the quality of the network prediction with a differentiable *rendering loss* [LSC18, LXR*18, DAD*18]. We adopt the loss of Deschaintre et al. [DAD*18], which renders the predicted SVBRDF under multiple light and view directions, and compare these renderings with renderings of the ground-truth SVBRDF under the same conditions. The comparison is performed using an l_1 norm on the logarithmic values of the renderings to compress the high dynamic range of specular peaks.

Following Li et al. [LSC18], we complement this rendering loss with four l_1 losses, each measuring the difference between one of the predicted maps and its ground-truth counterpart. We found this direct supervision to stabilize training. Our final loss is a weighted mixture of all losses, $L = L_{\text{Render}} + 0.1\big(L_{\text{Normal}} + L_{\text{Diffuse}} + L_{\text{Specular}} + L_{\text{Roughness}}\big)$.

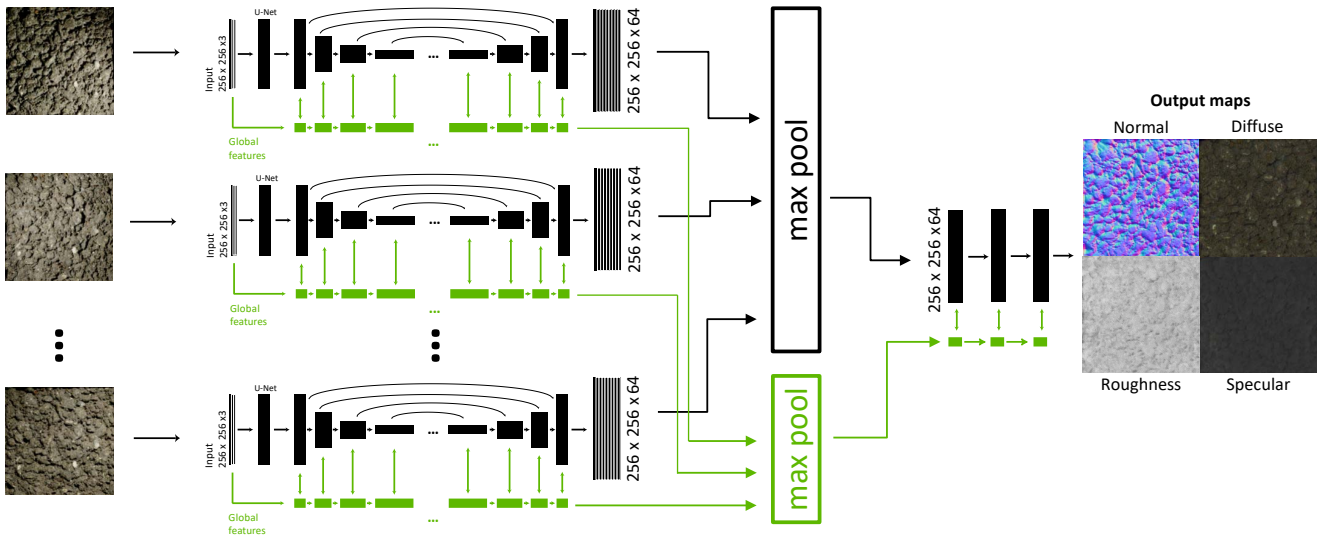


Figure 3: *Overview of our deep network architecture. Each input image is processed by its copy of the encoder-decoder to produce a feature map. While the number of images and network copies can vary, a pooling layer fuses the output maps to obtain a fixed-size representation of the material, which is then processed by a few convolutional layers to produce the SVBRDF maps.*

4.4. Training

We train our network for 7 days on a Nvidia GTX 1080 TI. We let the training run for 1 million iterations with a batch size of 2 and input sizes of 256×256 pixels. We use the Adam optimizer [KB15] with a learning rate set to 0.0002 and $\beta = 0.5$.

5. Online Generation of Training Data

Following prior work on deep-learning for inverse rendering [RGR*17, LDPT17, DAD*18, LSC18, LXR*18, LCY*17], we rely on synthetic data to train our network. While in theory image synthesis offers the means to generate an arbitrary large amount of training data, the cost of image rendering, storage and transfer limits the size of the datasets used in practice. For example, Li et al. [LSC18] and Deschaintre et al. [DAD*18] report training datasets of 150,000 and 200,000 images respectively. This practical challenge motivated us to implement an online renderer that generates a new SVBRDF and its multiple renderings at each iteration of the training, yielding up to 2 million training images in practice.

We first explain how we generate numerous ground-truth SVBRDFs, before describing the main features of our SVBRDF renderer.

5.1. SVBRDF synthesis

We rely on procedural, artist-designed SVBRDFs to obtain our training data. Starting from a small set of such SVBRDF maps, Deschaintre et al. [DAD*18] perform data augmentation by computing 20,000 convex combinations of random pairs of SVBRDFs. We follow the same strategy, although we implemented this material mixing within TensorFlow [AAB*15], which allows us to generate a unique SVBRDF for each training iteration while only loading a small set of base SVBRDFs at the beginning of the training process. We use the dataset proposed by Deschaintre et al., which contains 1,850 SVBRDFs covering common material classes such as plastic, metal, wood, leather, *etc*, all obtained from Allegorithmic Substance Share [All18].

5.2. SVBRDF rendering

We implemented our SVBRDF renderer in TensorFlow, so that it can be called at each iteration of the training process. Since our network takes rectified images as input, we do not need to simulate perspective projection of the material sample. Instead, our renderer simply takes as input four SVBRDF maps along with a light and view position, and evaluates the resulting rendering equation at each pixel. We augment this basic renderer with several features that simulate common effects encountered in real-world captures:

Viewing conditions. We distribute the camera positions over an hemisphere centered on the material sample, and vary its distance by a random amount to allow a casual capture scenario where users may not be able to maintain an exact distance from the target. We also perform random perturbations of the field-of-view (set to 40° by default) to simulate different types of cameras. Finally, we apply a random rotation and scaling to the SVBRDF maps before cropping them to 256×256 pixels, which simulates materials of different orientations and scales.

Lighting conditions. We simulate a flash light as a point light with angular fall-off. We again distribute the light positions over an hemisphere at a random distance to simulate a handheld flash. Other random perturbations include the angular fall-off to simulate different types of flash, the light intensity to simulate varying exposure, and the light color to simulate varying white-balance. Finally,

we also include the simulation of a surrounding lighting environment in the form of a second light with random position, intensity and color, which is kept fixed for a given input SVBRDF.

Image post-processing. We have implemented several common image degradations – additive Gaussian noise, clipping of radiance values to 1 to simulate low-dynamic range images, gamma correction and quantization over 8 bits per channel.

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, making the actual rendering virtually free.

6. Results and Evaluation

We evaluate our method using a test dataset of 32 ground truth SVBRDFs not present in the set used for training data generation. We also use measured Bidirectional Texture Functions (BTFs) [WGK14] to compare the re-renderings of our predictions to real-world appearances. Finally, we used our method to acquire a set of around 80 real-world materials. Since our method does not assume a controlled lighting, we used either the camera flash or a separate smartphone as the light source for those acquisitions. All results in the figures of the main paper were taken with two phones; please see supplemental for all results and examples acquired with a single phone. Resulting quality is similar in both cases.

6.1. Number of input images

A strength of our method is its ability to cope with a variable number of photographs. We first evaluate whether additional images improve the result using synthetic SVBRDFs, for which we have ground truth maps. We measure the error of our prediction by re-rendering our predicted maps under many views and lights, as done by the rendering loss used for training. Figure 4 plots the SSIM similarity metric of these re-renderings averaged over the test set for an increasing number of images, along with the SSIM of the individual SVBRDF maps. While most improvements happen with the first five images, the similarity continues to increase with subsequent inputs, stabilizing at around 10 images. The diffuse albedo is the fastest to stabilize, consistent with the intuition that few measurements suffice to recover low-frequency signals. Surprisingly, the quality of the roughness prediction seems on average independent of the number of images, suggesting that the method struggles to exploit additional information for this quantity. In contrast, the normal prediction improves with each additional input, as also observed in our experiments with real-world data detailed next. We provide RMSE plots of the same experiment as supplemental materials.

Using the same procedure, in Figure 5 we perform an ablation study to evaluate the impact of including random perturbations of the viewing and lighting conditions in the training data. As expected, the network trained without perturbation does not perform as well as our complete method on our test dataset that includes

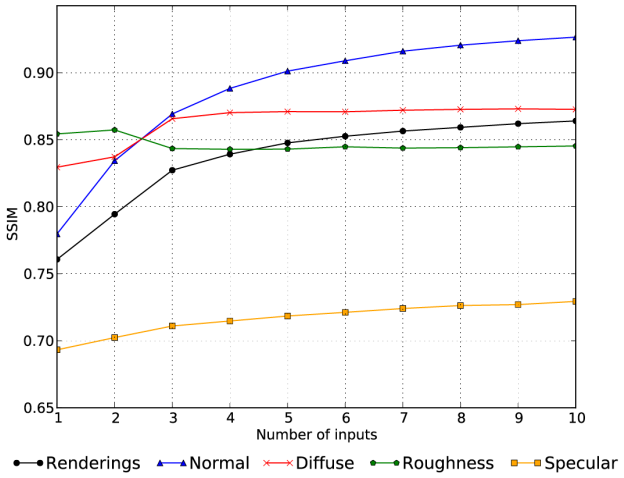


Figure 4: *SSIM of our predictions with respect to the number of input images, averaged over our synthetic test dataset. The SSIM of re-renderings increases quickly for the first images, before stabilizing at around 10 images. The normal maps strongly benefit from new images. Diffuse and specular albedos also improve with additional inputs, which is not the case of the roughness that remains stable overall. We provide similar RMSE plots as supplemental materials.*

view and light variations similar to those in casual real world capture. We trained both networks for 750,000 iterations for this experiment.

Figure 6 shows our predictions on a measured BTF material from the Bonn database [WGK14], using 1, 2, 3 and 10 inputs. For this material, normals, diffuse albedo and roughness estimations improve with more inputs. In particular, the normal map progressively captures more relief, the diffuse albedo map becomes almost uniform, and the embossed part on the upper right is quickly recognized as shinier than the remaining of the sample.

For a real material capture we performed (Figure 7), we see similar effects: normals are improved with more inputs, and the difference of roughness between different parts is progressively recovered. However, we do not have access to ground truth maps for these real-world captures.

Overall, our results in Fig. 4-10 and in supplemental material illustrate that our method achieves our goals: adding more pictures greatly improves the results, notably removing artifacts in the diffuse albedo while improving normal estimation. Our method enhances the quality of recovered materials while maintaining a casual capture.

6.2. Comparison to multi-image optimization

We compare our data-driven approach to a traditional optimization that takes as input multiple images captured under the assumption of known and precisely calibrated light and viewing conditions. Given these conditions we solve for the SVBRDF maps that minimize re-rendering error of the input images, as measured by our

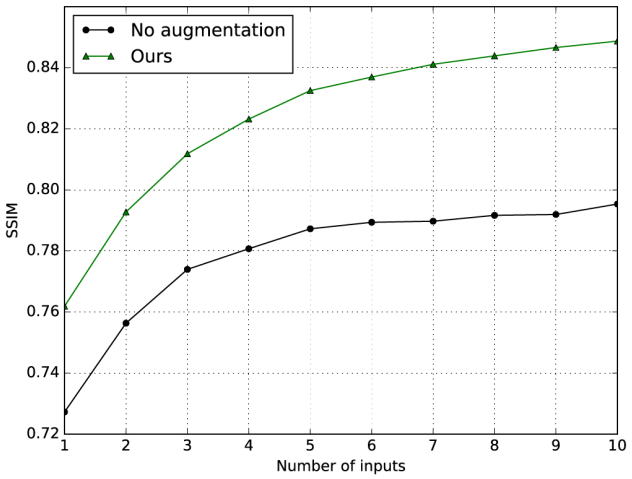


Figure 5: *Ablation study. Comparison of SSIM between our method (green) and a restricted version (black) where the network is trained with lighting and viewing directions chosen on a perfect hemisphere, and with all lighting parameters constant (falloff exponent, power, etc.). Our complete method achieves higher SSIM when tested on a dataset with small variations of these parameters, showing that it is robust to such perturbations that are frequent in casual real world capture.*

rendering loss. We further regularize this optimization by augmenting the loss with a total-variation term that favors piecewise-smooth maps. We solve the optimization with the Adam algorithm [KB15]. While the optimization stabilizes after 900K iterations, we let it run for a total of 2M iterations to ensure full convergence, which takes approximately 3.5 hours on an NVIDIA GTX 1080 TI. Given the non-convex nature of the optimization, we initialize the solution to a plausible estimate obtained by setting the diffuse albedo map to the most fronto-parallel input, the normal map to a constant vector pointing upward, the roughness to zero and the specular albedo to gray. We use synthetic data for this experiment, which provides us with full control and knowledge of the viewing and lighting conditions needed by the optimization, as well as with ground truth maps to evaluate the quality of the outcome.

Figure 8 compares the number of input images required to achieve similar quality between the classical optimization and our method, using view and light directions uniformly distributed over the hemisphere. On rather diffuse materials (stones, tiles), the optimization needs a few dozen calibrated images to achieve a result of similar quality to the one produced by our method using only 5, uncalibrated images. A similar number of images is necessary for a material with uniform shininess (scales). However more than 900 images were necessary for our optimization to reach the quality obtained by our method on a material with significant normal and roughness variations (wood). Overall, our method achieves plausible results with much fewer inputs captured under unknown lighting, although classical optimization can recover more precise SVBRDFs if provided with enough carefully-calibrated images.

6.3. Comparison to alternative deep learning methods

We first compare our architecture to a simple baseline composed of the network by Deschaintre et al. [DAD*18] augmented to take 5 images instead of one. This baseline achieves an average SSIM of 0.826, similar to the SSIM of 0.847 produced by our method for the same number of inputs. This evaluation demonstrates that our multi-image network performs as well as a fixed network while providing the freedom to vary the number of input images.

We next compare to the recent single-image methods of Deschaintre et al. [DAD*18] and Li et al. [LSC18], which both take as input a fronto-parallel flash photo. Figure 9 provides a visual comparison on synthetic SVBRDFs with ground truth maps, Figure 12 provides a similar comparison on BTFs measured from 81x81 pictures, which allow ground-truth re-renderings, and Figure 10 and 11 provide a comparison on real pictures. While developed concurrently, both single-image approaches suffer from the same limitations. The co-located lighting tends to produce low-contrast shading, reducing the cues available for the network to fully retrieve normals. Adding side-lit pictures of the material helps our approach retrieve these missing details. The fronto-parallel flash also often produces a saturated highlight in the middle of the image, which both single-image methods struggle to in-paint convincingly in the different maps. While the strength of the highlight could be reduced by careful tuning of exposure, saturated pixels are difficult to avoid in real-world capture. In contrast, our method benefits from additional pictures to recover information about those pixels.

Another limitation of these two single-image methods is that the flash highlight cannot cover all parts of the material sample. This lack of information can cause erroneous estimations, especially when the sample is composed of multiple materials with different shininess. Providing more pictures gives a chance to our method to observe highlights over all parts of the sample, as is the case in Figure 7, where the difference in roughness in the upper right only becomes apparent with the 4th input.

6.4. Limitations

Since our method builds on the single-image network of Deschaintre et al. [DAD*18], it inherits some of its limitations. First, the method is limited to materials that can be well represented by an isotropic Cook-Torrance BRDF. We also observe that the method tends to produce correlated maps and interpret dark materials as shiny, as shown in Figure 13(top) where despite several pictures, albedo variations of the cardboard get interpreted as normal variations, and the black letters get assigned a low roughness. This behavior reflects the content of our training data, since most artist-designed SVBRDFs have correlated maps.

Since we rectify the multi-view inputs with a simple homography, we do not correct for parallax effects produced by surfaces with high relief. This approximation may yield misalignment in the input images, which in turn reduces the sharpness of the predicted maps. In addition, our SVBRDF representation, training data, and rendering loss do not model cast shadows. While shadows are mostly absent in pictures taken with a co-located flash, they can appear when using a handheld flash and remain visible in some of our results, as shown in Figure 13 (bottom).

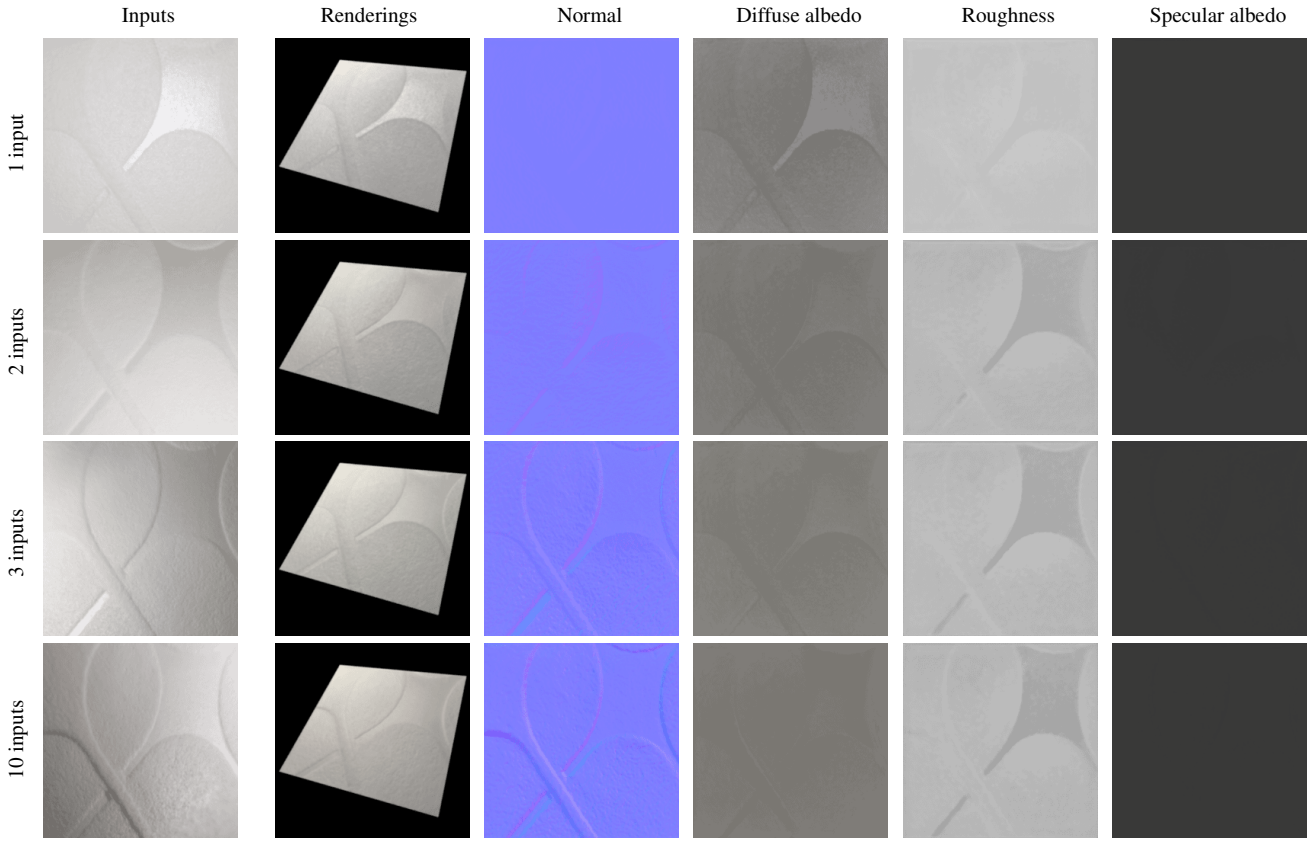


Figure 6: *Evaluation on a measured BTF. Three images are enough to capture most of normal and roughness maps. Adding images further improves the result by removing lighting residual from the diffuse albedo, and adding subtle details to the normal and specular maps.*

7. Conclusion

With the advance of deep learning, the holy grail of single-image SVBRDF capture recently became a reality. Yet, despite impressive results, single-image methods offer little margin to users to correct for erroneous predictions. We address this fundamental limitation with a deep network architecture that accepts a variable number of input images, allowing users to capture as many images as needed to exhibit all the visual effects they want to capture of a material. Our method bridges the gap between single-image and many-image methods, allowing faithful material capture with a handful of images captured from uncalibrated light-view directions.

Acknowledgments

We thank Yulia Gryaditskaya, Simon Rodriguez and Stavros Diolatzis for their support during the deadline as well as Anthony Jouanin and Vincent Hourdin for regular feedback. We also thank Zhengqin Li and Kalyan Sunkavalli for their help with evaluation. This work was partially funded by an ANRT (http://www.anrt.asso.fr/en) CIFRE scholarship between Inria and Optis, the ERC Advanced Grant FUNGRAPH (No. 788065, http://fungraph.inria.fr), and by software and hardware donations from Adobe and Nvidia.

References

[AAB*15] ABADI M., AGARWAL A., BARHAM P., BREVDO E., CHEN Z., CITRO C., CORRADO G. S., DAVIS A., DEAN J., DEVIN M., GHEMAWAT S., GOODFELLOW I., HARP A., IRVING G., ISARD M., JIA Y., JOZEFOWICZ R., KAISER L., KUDLUR M., LEVENBERG J., MANÉ D., MONGA R., MOORE S., MURRAY D., OLAH C., SCHUSTER M., SHLENS J., STEINER B., SUTSKEVER I., TALWAR K., TUCKER P., VANHOUCKE V., VASUDEVAN V., VIÉGAS F., VINYALS O., WARDEN P., WATTENBERG M., WICKE M., YU Y., ZHENG X.: TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. URL: https://www.tensorflow.org/. 4

[AAL16] AITTALA M., AILA T., LEHTINEN J.: Reflectance modeling by neural texture synthesis. *ACM Transactions on Graphics (Proc. SIGGRAPH) 35*, 4 (2016). 2

[AD18] AITTALA M., DURAND F.: Burst image deblurring using permutation invariant convolutional neural networks. In *The European Conference on Computer Vision (ECCV)* (2018). 2, 3

[All18] ALLEGORITHMIC: Substance share, 2018. URL: https://share.allegorithmic.com/. 4

[AWL13] AITTALA M., WEYRICH T., LEHTINEN J.: Practical SVBRDF capture in the frequency domain. 2

[AWL15] AITTALA M., WEYRICH T., LEHTINEN J.: Two-shot SVBRDF capture for stationary materials. *ACM Trans. Graph. (Proc. SIGGRAPH) 34*, 4 (July 2015), 110:1–110:13. URL: http://doi.acm.org/10.1145/2766967, doi:10.1145/2766967. 2

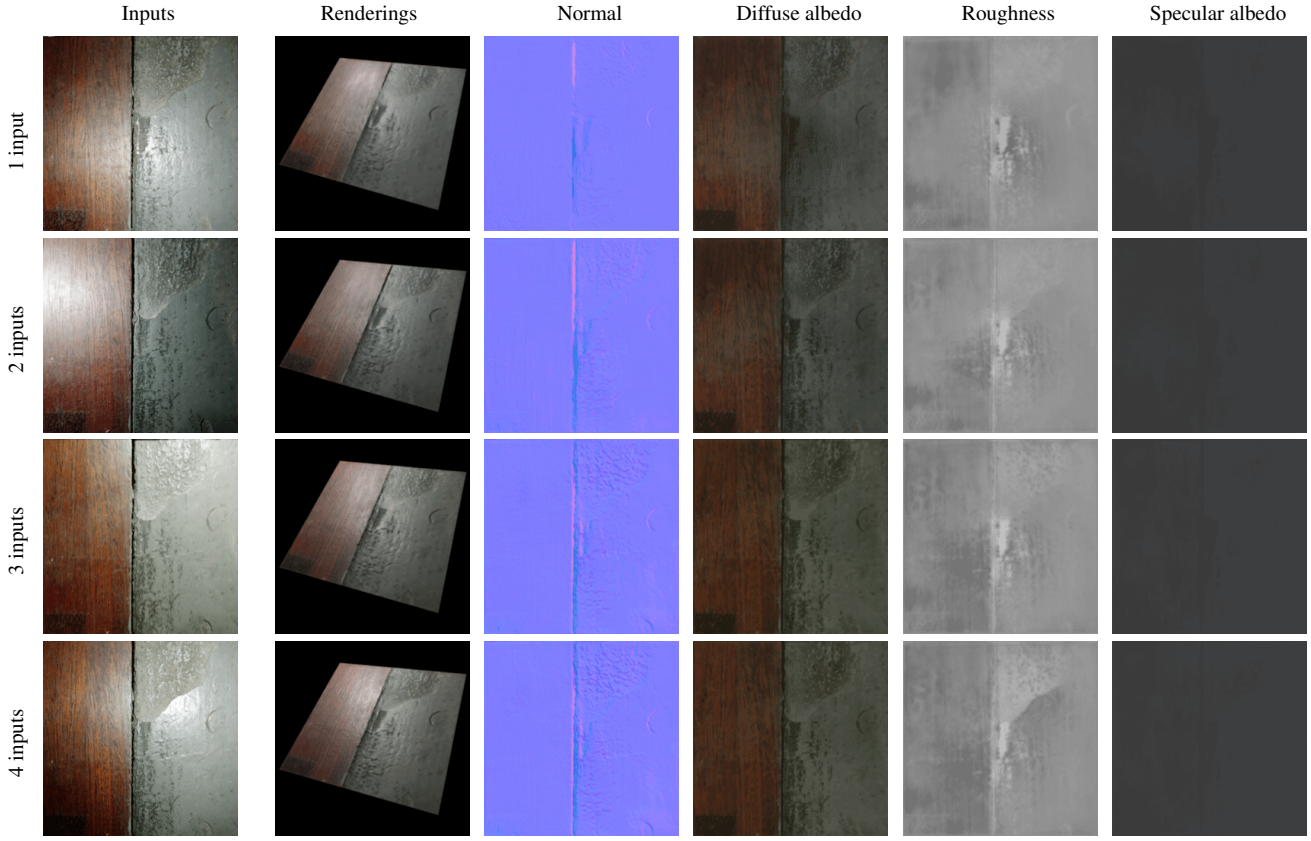


Figure 7: *A single flash picture hardly provides enough information for surfaces composed of several materials. In this example, adding images allows the recovery of normal details, and the capture of different roughness values in different parts of the image. Note in particular how the 4th image helps capturing a discontinuity of the roughness on the right part.*

[CHW18] CHEN G., HAN K., WONG K.-Y. K.: Ps-fcn: A flexible learning framework for photometric stereo. In *The European Conference on Computer Vision (ECCV)* (2018). 2, 3

[CT82] COOK R. L., TORRANCE K. E.: A reflectance model for computer graphics. *ACM Transactions on Graphics 1*, 1 (1982), 7–24. 3

[CXG*16] CHOY C. B., XU D., GWAK J., CHEN K., SAVARESE S.: 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *IEEE European Conference on Computer Vision (ECCV)* (2016), pp. 628–644. 2

[DAD*18] DESCHAINTRE V., AITTALA M., DURAND F., DRETTAKIS G., BOUSSEAU A.: Single-image svbrdf capture with a rendering-aware deep network. *ACM Transactions on Graphics (SIGGRAPH Conference Proceedings) 37*, 128 (aug 2018), 15. URL: http://www-sop.inria.fr/reves/Basilic/2018/DADDB18. 1, 2, 3, 4, 6, 14

[DCP*14] 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 (Proc. SIGGRAPH Asia) 33*, 6 (2014). 2

[DVGNK99] DANA K. J., VAN GINNEKEN B., NAYAR S. K., KOENDERINK J. J.: Reflectance and texture of real-world surfaces. *ACM Transactions On Graphics (TOG) 18*, 1 (1999), 1–34. 1, 2

[DWT*10] DONG Y., WANG J., TONG X., SNYDER J., BEN-EZRA M., LAN Y., GUO B.: Manifold bootstrapping for svbrdf capture. *ACM Transactions on Graphics (Proc. SIGGRAPH) 29*, 4 (2010). 2

[GCP*09] GHOSH A., CHEN T., PEERS P., WILSON C. A., DEBEVEC P.: Estimating specular roughness and anisotropy from second order spherical gradient illumination. In *Computer Graphics Forum* (June 2009), vol. 28, p. 4. 2

[GGG*16] GUARNERA D., GUARNERA G. C., GHOSH A., DENK C., GLENCROSS M.: BRDF Representation and Acquisition. *Computer Graphics Forum* (2016). 2

[GTHD03] GARDNER A., TCHOU C., HAWKINS T., DEBEVEC P.: Linear light source reflectometry. *ACM Trans. Graph. 22*, 3 (July 2003), 749–758. URL: http://doi.acm.org/10.1145/882262.882342, doi:10.1145/882262.882342. 2

[HSL*17] HUI Z., SUNKAVALLI K., LEE J. Y., HADAP S., WANG J., SANKARANARAYANAN A. C.: Reflectance capture using univariate sampling of brdfs. In *IEEE International Conference on Computer Vision (ICCV)* (2017). 2

[KB15] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)* (2015). 4, 6

[KCW*18] KANG K., CHEN Z., WANG J., ZHOU K., WU H.: Efficient reflectance capture using an autoencoder. *ACM Transactions on Graphics (Proc. SIGGRAPH) 37*, 4 (July 2018). 2

[LCY*17] LIU G., CEYLAN D., YUMER E., YANG J., LIEN J.-M.: Material editing using a physically based rendering network. In *IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2261–2269. 4

[LDPT17] LI X., DONG Y., PEERS P., TONG X.: Modeling surface ap-

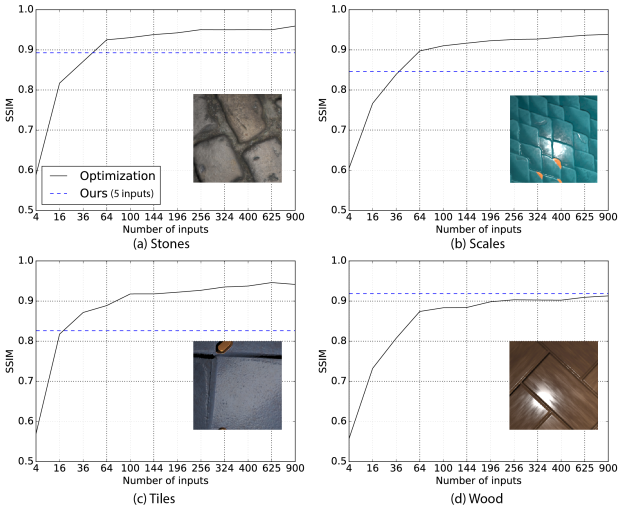


Figure 8: *SSIM on re-renderings for the maps obtained by our method with 5 images (dotted blue) and by a classical optimization method with an increasing number of input images (black). The classical optimization requires several dozens of calibrated pictures to outperform our method on rather diffuse or uniform materials (stones, tiles, scales), while requiring many more for a more complex material (wood).*

pearance from a single photograph using self-augmented convolutional neural networks. *ACM Transactions on Graphics (Proc. SIGGRAPH) 36*, 4 (2017). 2, 4

[LLM*18] LIU R., LEHMAN J., MOLINO P., SUCH F. P., FRANK E., SERGEEV A., YOSINSKI J.: An intriguing failing of convolutional neural networks and the coordconv solution. *CoRR abs/1807.03247* (2018). 3

[LN16] LOMBARDI S., NISHINO K.: Reflectance and illumination recovery in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI) 38* (2016), 129–141. 2

[LSC18] LI Z., SUNKAVALLI K., CHANDRAKER M.: Materials for masses: SVBRDF acquisition with a single mobile phone image. *Proceedings of ECCV* (2018). 1, 2, 3, 4, 6

[LXR*18] 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 (Proc. SIGGRAPH Asia)* (2018). 2, 3, 4

[Mca02] MCALLISTER D. K.: *A Generalized Surface Appearance Representation for Computer Graphics*. PhD thesis, 2002. 1, 2

[PCF05] PATERSON J. A., CLAUS D., FITZGIBBON A. W.: Brdf and geometry capture from extended inhomogeneous samples using flash photography. *Computer Graphics Forum (Proc. Eurographics) 24*, 3 (Sept. 2005), 383–391. 2

[QSMG17] QI C. R., SU H., MO K., GUIBAS L. J.: Pointnet: Deep learning on point sets for 3d classification and segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017). 2

[RGR*17] REMATAS K., GEORGOULIS S., RITSCHEL T., GAVVES E., FRITZ M., GOOL L. V., TUYTELAARS T.: Reflectance and natural illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)* (2017). 4

[RPB15] RONNEBERGER O., P.FISCHER, BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (2015), vol. 9351 of *LNCS*, pp. 234–241. 3

[RPG16] RIVIERE J., PEERS P., GHOSH A.: Mobile surface reflectometry. *Computer Graphics Forum 35*, 1 (2016). 2

[RRFG17] RIVIERE J., RESHETOUSKI I., FILIPI L., GHOSH A.: Polarization imaging reflectometry in the wild. *ACM Transactions on Graphics (Proc. SIGGRAPH)* (2017). 2

[RWS*11] REN P., WANG J., SNYDER J., TONG X., GUO B.: Pocket reflectometry. *ACM Transactions on Graphics (Proc. SIGGRAPH) 30*, 4 (2011). 2

[WGK14] WEINMANN M., GALL J., KLEIN R.: Material classification based on training data synthesized using a btf database. In *European Conference on Computer Vision (ECCV)* (2014), pp. 156–171. 5

[WSM11] WANG C.-P., SNAVELY N., MARSCHNER S.: Estimating dual-scale properties of glossy surfaces from step-edge lighting. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia) 30*, 6 (2011). 2

[WZ17] WILES O., ZISSERMAN A.: Silnet : Single- and multi-view reconstruction by learning from silhouettes. *British Machine Vision Conference (BMVC)* (2017). 2

[YLD*18] YE W., LI X., DONG Y., PEERS P., TONG X.: Single image surface appearance modeling with self-augmented cnns and inexact supervision. *Computer Graphics Forum 37*, 7 (2018), 201–211. 2

[ZKR*17] ZAHEER M., KOTTUR S., RAVANBAKHSH S., POCZOS B., SALAKHUTDINOV R. R., SMOLA A. J.: Deep sets. In *Advances in Neural Information Processing Systems (NIPS)*. 2017. 2

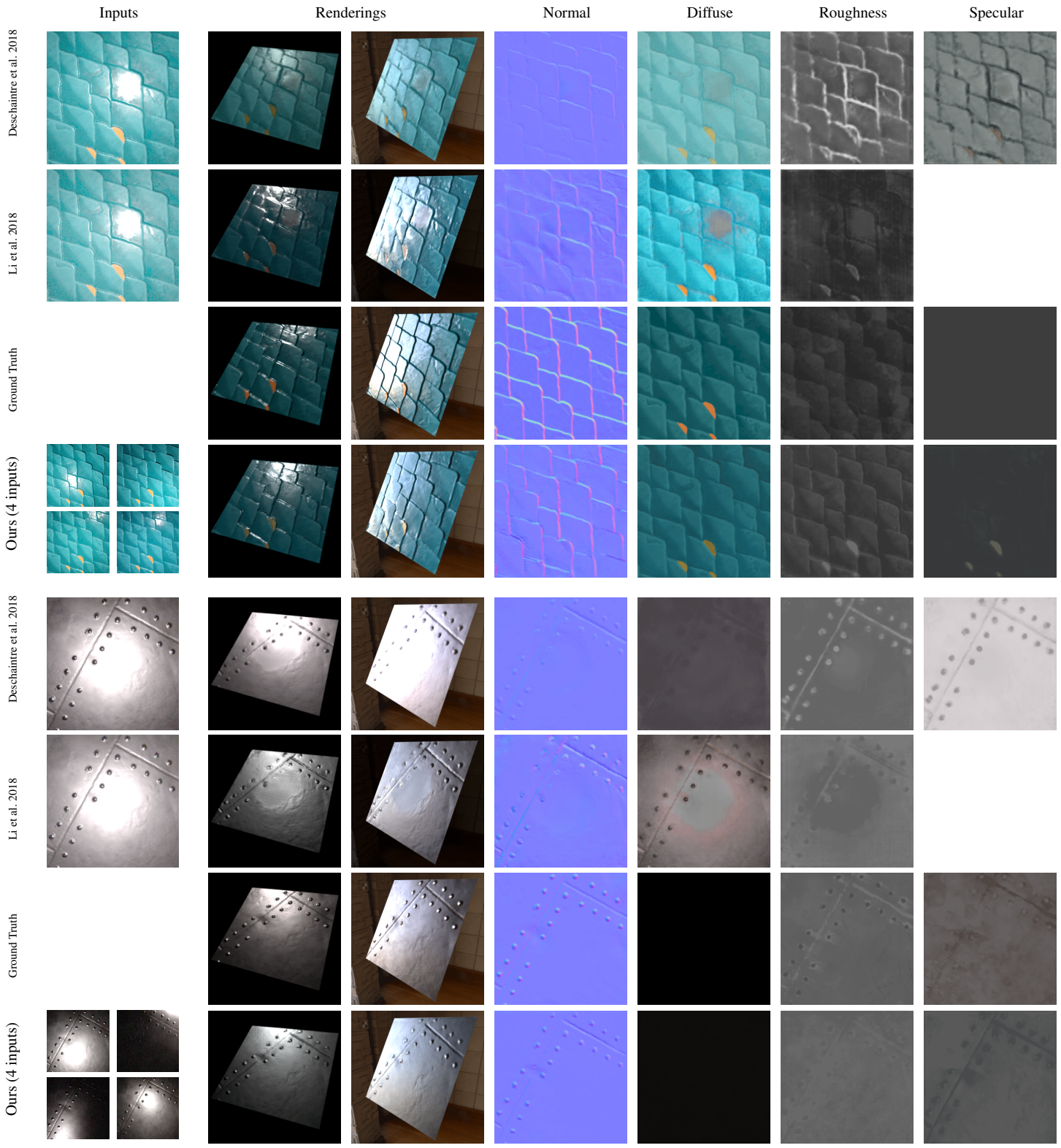


Figure 9: *Comparison against single-image methods on synthetic SVBRDFs. Our method leverages additional input images to obtain SVBRDF maps closer to ground truth. In particular, single-image methods under-estimate normal variations and fail to remove the saturated highlight on shiny materials. See supplemental materials for more comparisons and results.*

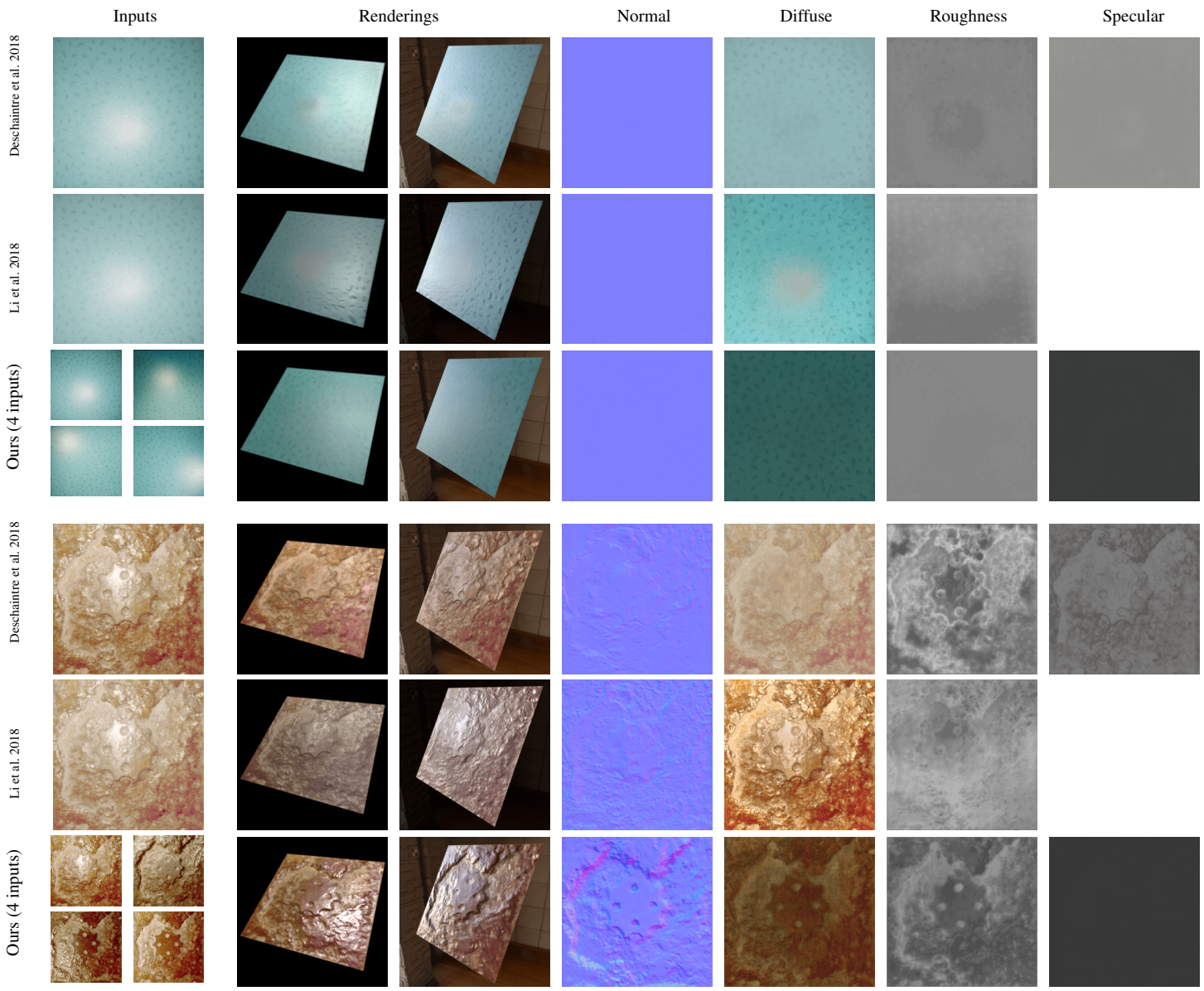


Figure 10: *Comparison against single-image methods on real-world pictures. Our method recovers more normal details, and better removes highlight and shading residuals from the diffuse albedo. See supplemental materials for more comparisons and results.*

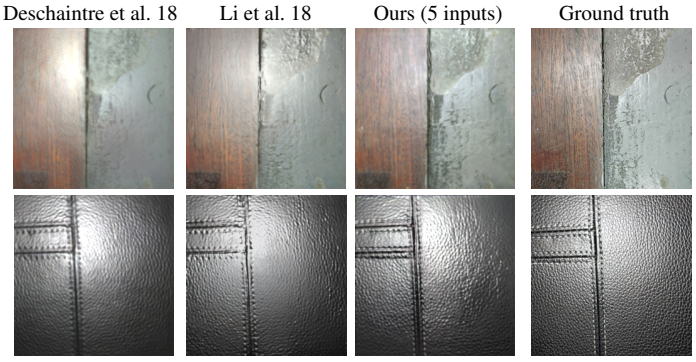


Figure 11: *Comparison to real-world relighting. Each column shows re-renderings of a captured material, except the last column which shows a picture of that material under a similar lighting condition (not used as input). We manually adjusted the position of the virtual light to best match the ground truth. Similarly, we adjusted the light power for each method separately since each has its own arbitrary scale factor. Overall, our method better reproduces the normal and gloss variations of the materials. In particular, single-image methods tend to flatten the bumps of the leather and orient them towards the center of the picture, where the flash highlight appeared in the input. For individual result maps, see supplemental materials.*

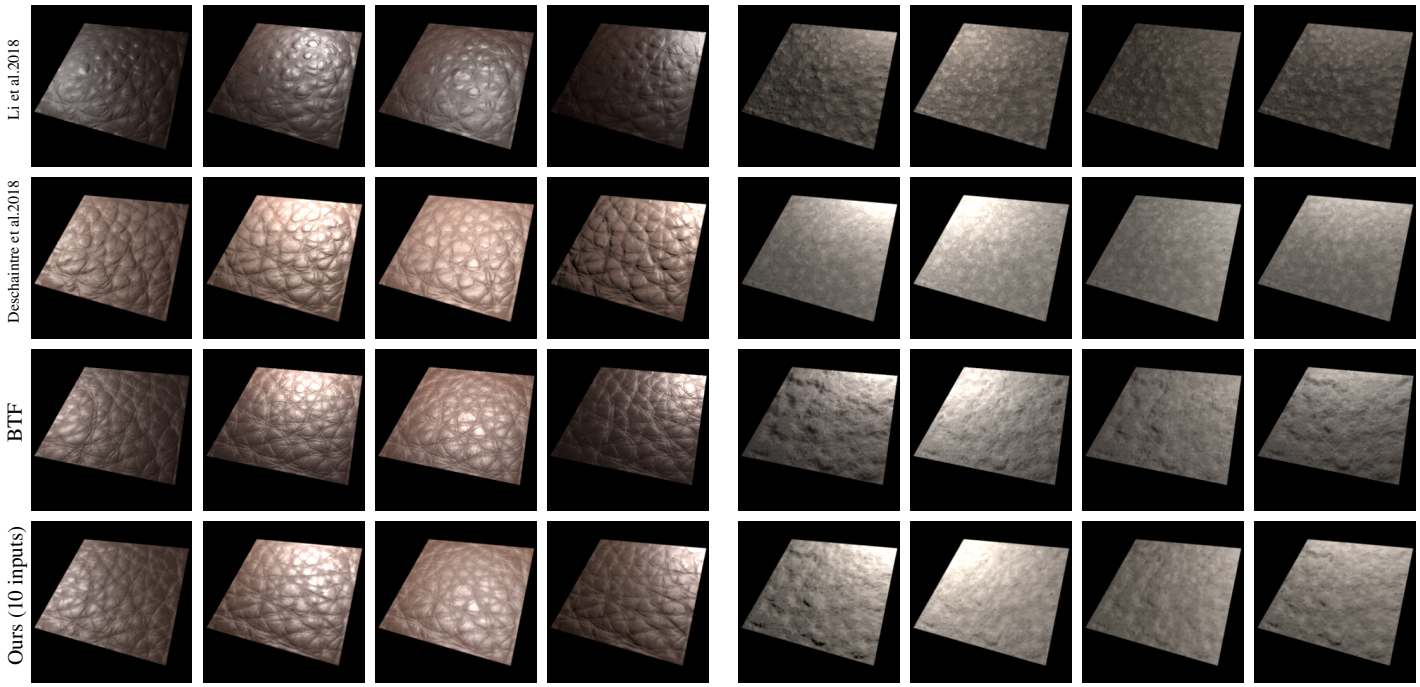


Figure 12: *Comparison against single-image methods on a measured BTF with ground truth re-renderings. Our method globally captures the material features better.*

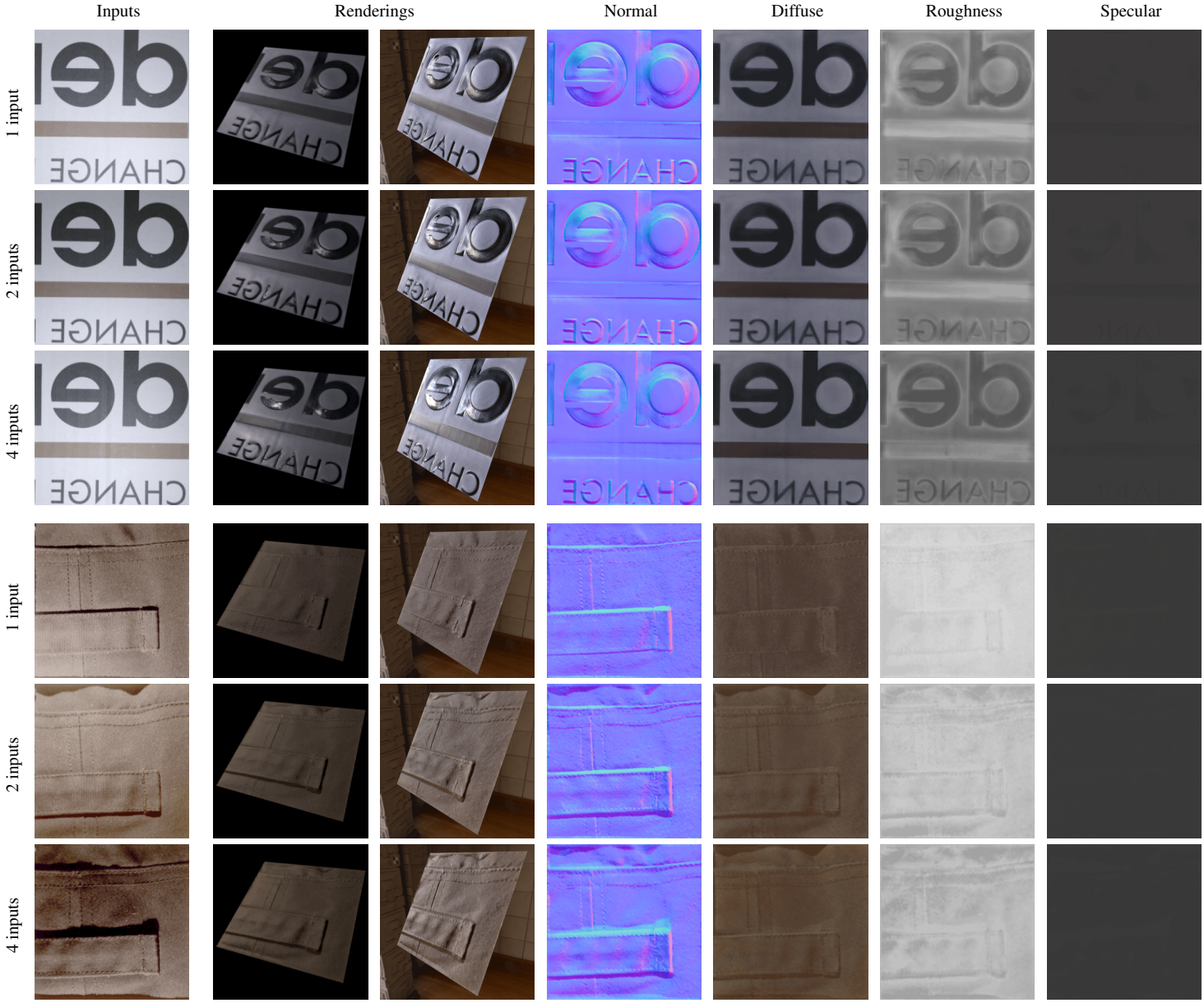


Figure 13: *Limitations. We inherits some of the limitations of the method by Deschaintre et al. [DAD*18], such as the tendency to produce correlated maps and to interpret dark pixels as shiny (top). Our SVBRDF representation, training data and loss do not model cast shadows. As a result, shadows in the input pollute some of the maps (bottom).*