EUROGRAPHICS 2022 / R. Chaine and M. H. Kim
(Guest Editors)

Volume 41 (2022), Number 2

MaterIA: Single Image High-Resolution Material Capture in the Wild

Rosalie Martin, Arthur Roullier, Romain Rouffet, Adrien Kaiser and Tamy Boubekeur

Adobe Research

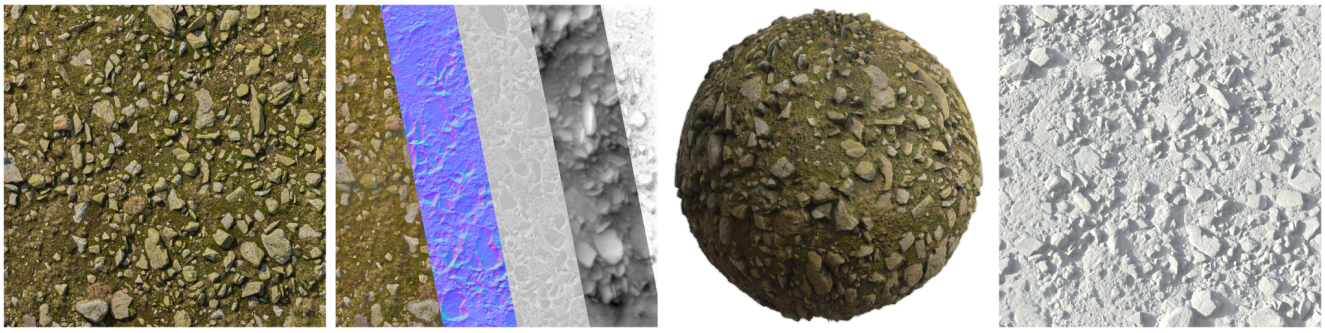

Figure 1: *From left to right: input photo (*4096 × 4096 *pixels), SVBRDF maps (albedo, normal, roughness, height, ambient occlusion) reconstructed with our method, rendering of the reconstructed material and rendering with a constant albedo to visualize the extracted geometry.*

Abstract

*We propose a hybrid method to reconstruct a physically-based spatially varying BRDF from a single high resolution picture of an outdoor surface captured under natural lighting conditions with any kind of camera device. Relying on both deep learning and explicit processing, our PBR material acquisition handles the removal of shades, projected shadows and specular highlights present when capturing a highly irregular surface and enables to properly retrieve the underlying geometry. To achieve this, we train two cascaded U-Nets on physically-based materials, rendered under various lighting conditions, to infer the spatially-varying albedo and normal maps. Our network processes relatively small image tiles (*512 × 512 *pixels) and we propose a solution to handle larger image resolutions by solving a Poisson system across these tiles. We complete this pipeline with analytical solutions to reconstruct height, roughness and ambient occlusion.*

Keywords: Material Capture, SVBRDF, Shadow Removal, Deep Learning, Dataset Synthesis, Delighting

1. Introduction

Digital materials provide 3D objects with a realistic appearance by describing how light interacts with them. One popular way to model such materials is through *spatially varying bidirectional reflectance distribution functions* (SVBRDF) which supply, at each point of the 3D surface domain, the parameters of an underlying pointwise reflectance model, as well as information about the geometric mesostructures. In our work, we target standard models for *Physically-Based Rendering* (PBR material) that rely on a microfacet BRDF model based on the GGX normal distribution function. Such models parameterize appearance with a collection of 2D texture maps that associate each texel to reflectance (diffuse albedo, roughness) and mesogeometric (normal, height, ambient occlusion) values.

Previously, capturing such materials from real samples implied complex hardware setup, with for instance multiple calibrated photos taken under controlled lighting [PF14]. Recently, a new trend emerged thanks to advances in machine learning (see Section 2) where a single, uncalibrated e.g., smartphone photo of the sample is used to reverse-engineer the individual components of its SVBRDF. Indeed, the vast majority of such methods require the photo to be taken with a flash to help a neural architecture drawing cues from the reflectance and mesogeometric properties of the sample. This turns out to be extremely efficient for indoor scenarios, for SVBRDF that exhibit moderate midscale geometric features and a certain degree of stationarity which copes acceptably with a low resolution output. However, when it comes to outdoor scenarios, some assumptions fall short. First, the camera flash may have no

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd. Published by John Wiley & Sons Ltd.

effect, with lighting being dominated by the sky and/or the sun. Second, geometric mesostructures may express a greater variety of relative scales and features. Third, practical use cases in the industry require fairly high resolution captures since a small sample, e.g. a forest ground or a cliff wall, may simply tile poorly and unrealistically. In this paper, we propose a single image method to address such outdoor material capture scenarios.

Aiming for high resolution material capture quickly faces the neural dimensionality problem: even with extremely large compute resources, the neural capacity should focus on the most ill-posed aspects of the outdoor scenario if high resolution, e.g. 1024^2 to 4096^2 pixels is expected in the SVBRDF maps. The prominent presence of strong geometric mesostructures as well as outdoor lighting conditions has two consequences that dominate the problem: first the material sample may exhibit strong self-shadowing and second the geometric structure often exists at multiple scales in the same sample. For this reason, we propose a hybrid method, where a neural architecture is employed on the most challenging aspects of the problem, and scalable numerical methods are used on easier tasks. More precisely, on the reflectance side, we focus on delighting the input image to bootstrap the material extraction while, on the mesogeometric side, we focus on reconstructing the normal field of the material sample. In essence, we create a *cascaded* neural architecture where we start by designing a neural network to separate the irradiance and the specular contribution while recovering the material's albedo map, before running a second neural net, conditioned on the output of the first, devised to reconstruct the normal map in a multiscale fashion. The last step of our method recovers the remaining height, roughness and ambient occlusion maps using standard numerical methods.

We deployed our approach in a commercial application designed for users without extensive technical background nor advanced equipment in capture or photography, where ease of use and productivity are fundamental. This context influences our final implementation, designed towards better performance (see Section 4), and our design choices, to ensure that minimal assumptions are taken on the input image in terms of capturing device, lighting and quality. Such "wild" capture use cases include forest grounds, stone walls and rock cliffs, taken at daylight without constraint on the weather, with an ordinary camera device and without flash lighting. Still, our method can also be used on any similar picture, e.g. from the internet.

Contributions

We propose a SVBRDF acquisition method from a single image that targets highly-irregular materials captured in a natural lighting environment, where self-cast shadows have to be removed. Our main contributions are:

Hybrid reconstruction: we designed a deep neural architecture to resolve the most ambiguous tasks – namely *delighting* the input picture and *extracting* the geometric gradients, while relying on explicit numerical methods to deduce the remaining SVBRDF channels. As such, the available neural capacity is dedicated to the most ill-posed problems and, combined with our tiling approach, helps scaling to high-resolution inputs.

Synthetic dataset generation: we describe a method to generate large collections of SVBRDF materials starting from a smaller set of procedural material graphs that provide realistic variations. We do so by varying their hyper-parameters following a Gaussian distribution centered on their presets and amplifying this data set by piling distributions of objects fetched from large, material-specific *atlases*.

Illumination decomposition: we decompose the delighting problem into the prediction of the irradiance and specular contribution to compute the albedo. In addition, an enhancer map is predicted to improve the delighting of saturated areas (darkest shadows and specular spots).

Cascaded neural architecture: we propose a *cascaded* U-Nets approach that exploits the result of the delighting stage as a clue for the geometry estimation. More precisely, the geometry U-Net uses the output irradiance map of the delighting U-Net to predict the normal map.

Seamless high-resolution outputs: we propose a seamless reconstruction from tiles of the albedo and normal maps in the gradient domain by solving a Poisson system, combined with a multiscale blending to reduce ambiguities at the scale of a tile.

We report on important implementation details, based on NVidia CuDNN and Tensor Cores, which allowed us to ship our method in a large-audience creative product.

2. Related Work

Image Translation. The task of converting one image into another has benefited greatly from progress in deep-learning based image segmentation, and the advent of the U-Net architecture [RFB15] specifically. Based on a sequential encoder-decoder, it was the first deep network to recompute a full sized output from a latent code while preserving the high frequencies of the input, thanks to *skip connections* mapping the encoded input features to the decoded features at each scale. This architecture can be used to generate a modified version of the input image, by simply adapting the loss function. Hence, most work presented below, including our approach, relies on the U-Net to predict the attributes of the object or material to capture. In particular, in our work, we use a pixel-level loss for all tasks as we are interested in preserving as much fidelity to the input as possible.

Shadow Removal and Intrinsic Image Decomposition. As our method is tailored to capture a material under natural lighting conditions, one of the tasks we aim at solving is to remove the shadows from the picture, which is a long standing problem in computer vision and computer graphics [SSL12]. While early results of shadow removal from a single picture were promising for single object scenes [FHLD05; LEN10], Qu et al. [QTH*17] and Wang et al. [WLY18] recently introduced learning-based methods to perform this task on more complex scenes. However, in their work the shadows are cast by large objects outside of the picture's field of view and the target is not an albedo free of any lighting information, which is what we seek. Wang et al. [WLY18] use cascaded UNets to perform the shadow removal task conditionally to the shadow mask detection, which we take inspiration from and show that it helps recovering an accurate representation of the material geometry.

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

Intrinsic image decomposition, namely the separation of natural images into reflectance and shading components, is close to our proposal of predicting the irradiance and specular contribution maps in order to compute the final albedo map. This task was tackled by many researchers, starting with the work of Weiss [Wei01] based on image sequences, up to recent unsupervised learning based methods [JWK*17; MCZ*18; LYYL20]. In the single image context, some approaches propose to leverage a prior geometry estimation of the scene to predict the reflectance and shading [LHL*20]. Yu et al. [**yu_2019**] are going further by introducing an inverse renderer and tackling outdoor scenes, while most work in the area focuses on indoor scenes, or even on a single object. However, these methods assume a Lambertian scene and do not handle the specular highlights. The extracted reflectance often lacks the contrast, sharpness, and fine grain detail of a material albedo.

Single-View Shape Capture. Methods to capture objects from a single image often extract the appearance properties of the object in addition to its shape, represented by a depth or normal map.

An early line of work focused on the estimation of the shape and the reflectance map of an object in a picture, while assuming knowledge of either shape or lighting conditions [Woo80; Mar98] or Lambertian reflectance [JA11]. Recently, Rematas et al. [RRF*16] use two U-Nets to respectively infer the shape map and the reflectance map, with supervision on both. The reflectance map is further factored out into BRDF parameters and illumination map in a follow-up paper [GRR*18]. Liu et al. [LCY*17] use a differentiable rendering layer to add further supervision over renders obtained with the predicted parameters, compared to ground truth materials. Although, these approaches share by design the limitation of only modeling constant, *non*-spatially varying BRDFs.

Recent work introduces the estimation of *spatially varying* BRDF parameters in addition to the object shape. On top of using a rendering loss and estimating lighting parameters, Li et al. [LXR*18] propose to train one U-Net per material channel, where encoders are shared. Similarly to our work, this is made possible by the use of a large scale synthetic dataset. In the last two years, a number of improvements have been proposed, including training jointly with relighting [SC20], two shot capture [BJK*20], multiple polarized inputs [DLG21] and multi-scale architecture [LWSJ21].

While able to capture variations of color and roughness, these methods remain focused on small objects photographed under flash lighting and are not designed to handle the large scope of details seen in high quality digital materials. In addition, the use of flashlight, while being better suited to capturing specular properties, can deteriorate portions of the input by saturating highly specular areas and preventing a total recovery of the underlying material. In our unconstrained scenario, we want to avoid both the need for flash lighting and its destructive effect on the input, although the uncontrolled natural light makes the recovery of material properties a harder task.

Single-View Material Capture. Methods specifically tailored for the acquisition of SVBRDF maps consider the observed object as a flat surface. This task is closer to image translation, as there is no ambiguity between the object's shape (a flat surface) and the local geometry of the material at the surface of the shape. See [GGG*16] for an overview of early results of SVBRDF capture. Recent works on acquisition from a single picture have shown good results using deep learning approaches, often by making assumptions on the material or imposing constraints on the capture process, such as the distance to the sample or the type of light source.

In particular, a first line of work focuses on the translation of one input image into multiple material channels. A single model is used to capture a variety of materials in a single feed-forward pass, which is fast and well suited to real-time use cases. Li et al. [LDPT17] estimate diffuse albedo and normal maps using a U-Net like architecture, as well as a single value for the specular level, estimated with a convolutional network. They alleviate the lack of labeled data with a self-augmentation scheme where predictions are used to create new input renders. Deschaintre et al. [DAD*18] predict SVBRDF maps from a single flash-lit image. They augment the U-Net architecture with a global features track to propagate information from non-burnt areas and mitigate the remaining specular spot artifacts from the flash. They match differentiable renders, although their large high quality dataset, produced with a new on-the-fly material blending method, would be well suited to direct supervision. *Materials for Masses* [LSC18] uses a U-Net with one common encoder but multiple decoders, and supervises jointly on the ground truth PBR maps and differentiable renders. It uses a specific map with radial coordinates to compensate for the degradation of the input by the required flashlight. Guo et al. [GLT*21] mitigate such saturated pixels at the level of network features, by handling the specular highlights within the convolution operation. Tini [Tin20] presents competitive results on the estimation of the normal map, including a Gaussian blending scheme to stitch overlapping tile predictions. More recently, Zhou et al. [ZK21] propose an adversarial loss, in the spirit of *Pix2Pix* [IZZE17]. They require flash-lit inputs but use their light position estimation to generate differentiable renders, supervised with another adversarial loss. Martin et al. [MMP19] propose a single-input high-resolution delighter that produces an albedo map for outdoor scenarios, that we take inspiration from.

Other approaches use iterative optimization to estimate material maps, such as the seminal work from Aittala et al. [AAL16] which assumes that the material is stationary and requires the use of the flashlight during the acquisition. Their work only matches the statistics of the input and does not provide a pixel-wise translation. Also featuring an iterative approach, Gao et al. [GLD*19] obtain the material parameters by optimizing the latent space of an encoder to match differentiable renders, starting from multiple input images. *MaterialGAN* [GSH*20] pushes the idea further by using a *StyleGAN* [KLA21] architecture instead of the usual encoder-decoder. Deschaintre et al. [DDB20] process tiles of large resolution inputs by fine-tuning a version of their original network, pretrained on natural non-flash lit examples, with multiple close-up images. Although refinement iterations lead to high fidelity capture, they add a strong time constraint and are slower than direct, feed-forward approaches.

Still starting from a photo input, some methods extend the capture process by generating new samples of the input material. Zhao et al. [ZWX*20] train a generator to estimate the material parame-

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

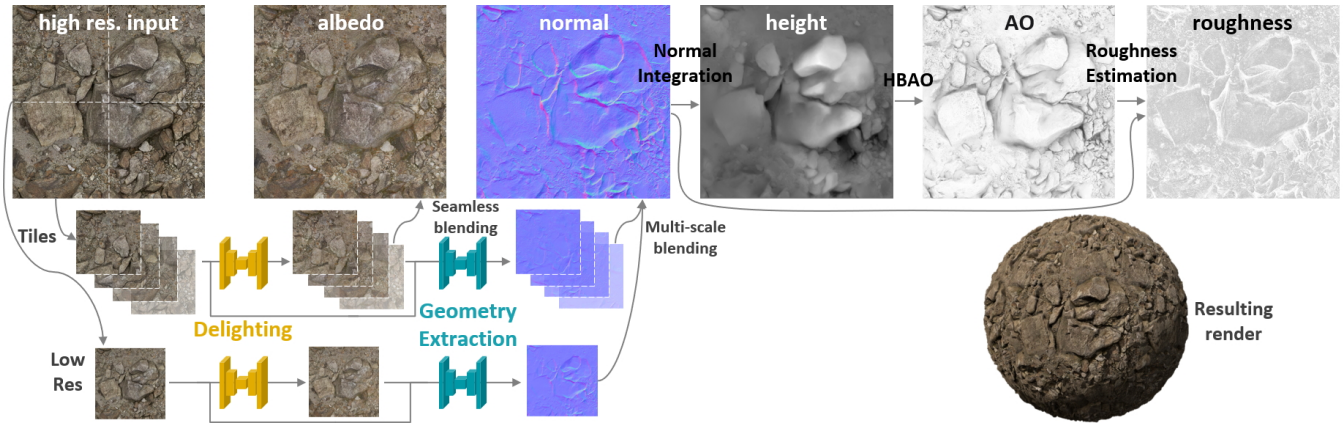

Figure 2: *Method overview. High-resolution input from a camera is split into* 512×512 *tiles, from which the albedo and normal maps are inferred by two cascaded U-Nets before being stitched back together and seamlessly reconstructed. A downscaled version of the input is also processed through the same network to get the low frequencies of the normal map. High-resolution and low-resolution normals are then blended, before deducing the height, ambient occlusion and roughness maps.*

ters but also expand them spatially, although their method requires training the model for each input. Henzler et al. [HDMR21] propose a unique model to handle the capture and the generation over an infinite spatial extent for a variety of materials. They only match texture statistics, which limits their approach to capturing stochastic textures. Last, inverse procedural modeling methods [HDR19; SLH*20] have recently shown impressive results in generating the material graph itself from a single picture, but do not match the input pixel-wise, nor address the geometric mesostructures, and assume flashlight capture as well as access to a large database of material graphs at runtime.

The methods described above focus on retrieving the SVBRDF of a material. As such, they are not considering self-projected shadows and, while providing good results on flat materials, may not perform as well on irregular material geometries, frequent in outdoor scenarios. Additionally, the use of the flashlight is often a constraint of the capture, while we consider natural, unconstrained outdoor lighting, where flash lighting has no effect because of either the environment light or the distance to the material. Besides, most existing methods are trained on 256×256 pixels image inputs, and do not address the question of larger input sizes, up to 4K.

3. High-resolution Material Acquisition

3.1. Overview

Our method reconstructs a digital material from a single photo in the form of an SVBRDF modelled as a collection of five 2D texture maps. These maps define, at each location on a surface, the reflectance attributes of a microfacet model [CT82; Kar13] based on the GGX normal distribution function [WMLT07], as well as the geometric components of the material. More precisely, they include the diffuse albedo and the roughness of the BRDF as well as the normal and the displacement (or height) value of the material mesostructure. Last, an ambient occlusion map is computed from the height map. Our method (Fig. 2) starts by inferring the albedo map through a *Delighter* U-Net (or D-UNet), before inferring the normal map through a *Geometry* U-Net (or G-UNet) using the output of the Delighter U-Net. We run it on a per-tile basis where 512×512 pixels tiles are processed before recovering the high-resolution output by interpolating in the gradient domain. Height, roughness and ambient occlusion maps are analytically deduced in the final stage, providing a complete PBR material.

3.2. Dataset

Although our goal is to reconstruct a PBR material from a photograph, creating training pairs using real photographs as input and accurate corresponding SVBRDFs is a very tedious process. Indeed, in the case of highly irregular materials, getting an albedo map with no remaining shade in the occluded regions requires a manual cleaning with no guarantees on its correctness. This would also require many captures to get enough variations in the data and lighting conditions in order to obtain a representative dataset. Instead, using synthetic data to train a neural network has proven to be effective and reliable, as shown by Li et al. [LSC18] and Deschaintre et al. [DAD*18]. Nonetheless, care must be taken to ensure the realism of the synthesized data and the representativeness of our target domain – outdoor materials. To that end, we propose a dedicated data synthesis pipeline based on parametric materials.

Initial Dataset We initiate our dataset with a collection of parametric material graphs in the Substance 3D format [Ado21]. Each such parametric material is described by a *directed acyclic graph* that outputs the different channels of an SVBRDF, and exposes a set of high-level parameters to control procedural variations. We selected the graph set to cover outdoor categories (grounds, stones, terracottas, plasters, concrete/asphalt). Across most material databases – including ours – the geometric information is inconsistently balanced between the height map and the normal map. To mitigate this issue, we compute a reliable version of the normal map, to use as ground truth, by blending the normal derived from the height map and the raw normal map produced by the material graph.

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

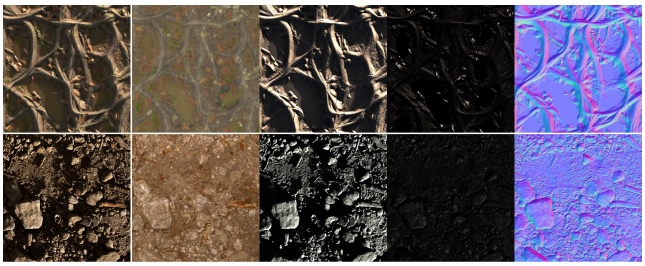

Figure 3: *Two examples of synthetic training sets. Left to right: Input (material rendered under specific lighting), target albedo, irradiance map, specular contribution map, normal map.*

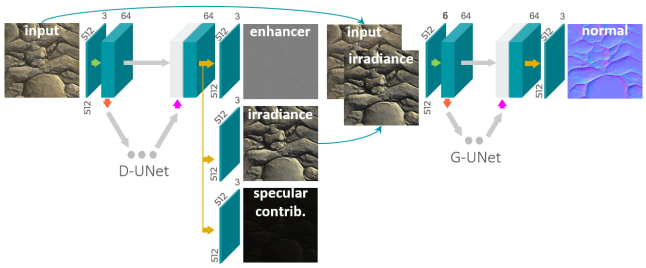

Figure 5: *Complete network architecture: two UNets are cascaded, G-UNet being conditional to the output of D-UNet.*

Dataset augmentation Tweaking parameters of procedural materials in their allowed ranges can actually produce highly unrealistic materials, which require a time-consuming manual cleaning pass, making iterations on the dataset very inefficient. Consequently, to generate a wide variety of realistic materials, we adopt a conservative strategy by sampling parameter variations following a Gaussian distribution centered on the values of *presets* defined in the material graph by its creator, with a variance equal to 5% of the parameter range. This provides enough variability while preserving the overall aspect of the material, as originally designed.

Furthermore, real life outdoor scenes rarely offer "clean" materials to capture, and often exhibit a certain complexity due to the presence of multiple objects that collectively form the appearance of a ground or a cliff – and this is precisely this realistic appearance that we aim at capturing. Such layouts often cause sharp edges and strong cast shadows that we want to show to our neural architecture during training. Therefore, we amplify our data set by splatting small assets such as stones, leaves, or wood sticks using a blue-noise distribution over each base material. These assets are fetched from atlases and designed on a per-category basis to preserve semantics of the synthetic output (see Fig. 4).

We produce 14 renders of each material in 2048×2048 pixels using a set of seven HDR environment maps, while varying their horizontal rotation. These lighting condition samples are drawn from a parametric HDR environment map which is itself modelled as a parametric graph and simulates natural lighting conditions with a sun light at different times of day and a diffuse sky. The color temperature of the sun and sky are also slightly varied during the rendering to increase the robustness of the delighting to color shifting in illuminated and shadowed parts. Our real-time renderer uses image-based lighting and ray tracing to generate, in addition to the final image, the irradiance and specular contribution maps. These are used as ground truth maps during training to learn to decouple the contribution of light and material properties, as presented in Section 3.3.

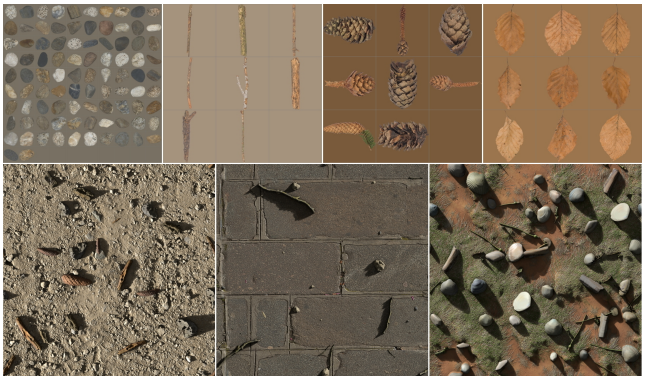

Figure 4: *Atlas splatting on synthetic materials. Top: examples of atlases (albedo map). Bottom: Final renders of atlas splatting on grounds, provoking strong shadows.*

Last, we proceed with a final data augmentation step by rotating, flipping, scaling, and finally cropping our images to a 512×512 resolution – the input resolution of our neural nets. Ultimately, starting from roughly 350 parametric material graphs, we generate 2100 materials, resulting in 30000 renders and finally around 240000 training tuples of 512×512 pixels (see one example in Fig. 3), with for each tuple the rendered material used as input and a set of ground truth maps (albedo, irradiance, specular contribution and homogenized normal) used for supervision.

3.3. Network architecture

Our network is a deep convolutional neural network made of two cascaded UNets, the Delighter (*D-UNet*), focused on recovering the albedo, and the Geometry extractor (*G-UNet*), focused on retrieving the normal map, leveraging the Delighter's output (see Fig. 5).

The delighter uses the illumination decomposition formula in or-

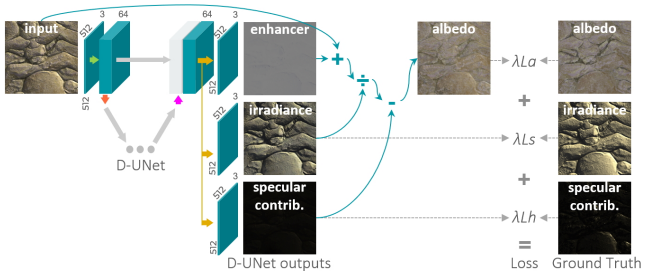

Figure 6: *D-UNet architecture.*

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

der to recover the albedo, starting from the input:

$$input = albedo * irradiance + specular \quad (1)$$

This decomposition helps to disambiguate the captured color, by splitting it into the intrinsic reflectance of the captured material (albedo) and the irradiance and specular highlights produced by the incident lighting. We reverse this formula to deduce the albedo given the input, irradiance and specular contribution, and introduce the enhancer map as follows:

$$albedo = \frac{input + enhancer - specular}{irradiance} \quad (2)$$

The role of our enhancer map is to enhance the input, that is possibly quantized or clamped and may lack information in the saturated and darkest areas. Ground truth irradiance and specular contribution maps are available in our dataset, but there is no ground truth for the enhancer, which brings slights modifications to the input in order to rectify the damaged areas, thus mitigating the instability of the division by the irradiance map. D-UNet (see Fig. 6) takes as input a 512×512 lit RGB image and is composed of five levels of convolutional blocks on each side of the latent space. It predicts three maps: *irradiance*, *specular contribution*, and *enhancer*. Fig. 12 shows the evolution of the Delighter, starting with a direct albedo prediction, introducing the irradiance and specular maps, and finally adding the enhancer map.

The input of G-UNet is the concatenation of the input image and the predicted irradiance map, leading to a six-dimensional input (see Fig. 5). It outputs the X and Y components of the normal vector, the Z component being deduced. G-UNet is shallower than D-UNet, with four levels of convolutional blocks and fewer feature maps in the deepest layers. The cascaded UNets approach allows to use the semantic information provided by the irradiance map as a clue to recover the geometry. The complete architecture of these two networks is given in Section 1 of the supplemental material.

We use a combination of all the output maps to compute the loss, each one being based on the L1 norm between the prediction and the ground truth map. We first train the Delighter UNet with the following loss function:

$$loss = \lambda_a loss_{albedo} + \lambda_s loss_{irradiance} + \lambda_h loss_{specular} \quad (3)$$

where λ_a is set to 1.0, λ_s and λ_h are set to 0.7 in our experiments, to give slightly more importance to the error in the albedo map as it corresponds to the final target. Once the training has converged, we freeze the weights of the first UNet and train the second one to retrieve the normal map. To encourage the network to infer a normal map with sharp edges, we multiply the L1 error image by a weight map which is inversely proportional to the Z component of the ground truth normal, in order to give more importance to normals pointing further from the inverse view direction.

We found that using a *Bounded ReLU* as activation function with a threshold of 6 helps the network to converge, by preventing perturbations in the input signal to accumulate across the layers. We use nearest neighbor upsampling in the decoding part, and mirror padding at all stages to reduce artifacts at the boundaries of each 512×512 tile.



Figure 7: *Seamless albedo reconstruction. Left: input picture from camera (*1024×1024 *pixels). Middle: output albedo tiles stitched together, with visible seams. Right: seamless Poisson solving.*

3.4. Reaching high-resolution

While existing deep learning methods usually operate on small input images, often limited to 256×256 pixels due to memory and speed limitations, an artist would typically need at least 2048×2048 pixels to produce a valuable material. To tackle the ability to extract a material from a high-resolution picture, we propose a method that processes small tiles separately, associated to a merging strategy to produce high-quality results.

Seamless blending in the gradient domain. We first split the input image into 512×512 pixels tiles without overlap, and predict the albedo and normal maps for each tile using our network. We then stitch all the predicted tiles together to recover a full resolution albedo and normal, and solve the *Poisson* equation to remove the seams between tiles on each resulting image, as an application of *Poisson Image Editing* [PGB03]. For the albedo map, we use the color gradients as the guidance field, and provide the entire image borders as the boundary values for the Poisson equation, which is solved on each color component separately. The normal map carries unit vectors, for which XYZ components cannot be be interpolated separately. Therefore, we convert the normal vector into particular angular coordinates, θ being the angle from the Z axis along XZ plane, between $-\frac{\pi}{2}$ and $\frac{\pi}{2}$, and φ being the angle formed by Y axis and OP (see Fig. 8). This representation leverages the fact that normals are pointing up, and avoids any modulo 2π in the gradient computation:

$$\begin{cases} x = sin(\theta)sin(\varphi) \\ y = cos(\varphi) \\ z = cos(\theta)sin(\varphi) \end{cases} \leftrightarrow \begin{cases} \theta = atan2(x,z) \\ \varphi = acos(y) \end{cases} \quad (4)$$

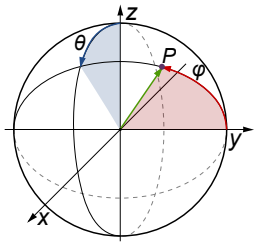

Figure 8: *Angular coordinates used for the Poisson reconstruction of the normal vectors.*

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

We compute the gradient of each angular coordinate separately and solve the Poisson equation in this space.

This process removes the seams appearing when all output tiles are stitched together by smoothing the gradients, and allows the acquisition of a high-resolution material. This method works particularly well for the albedo because the tiles are from the same material and present a color consistency at their boundaries (see Fig. 7).

The result for the normal map is more dependent on the size of the portrayed features and the positions of seams. The results are satisfactory for small features, but when features exceed the tile size or are located across two tiles, the high ambiguity at the tile scale can make the network fail at recovering the geometry. In addition, Poisson solving can introduce low-frequency gradient artifacts that lead to unwanted geometry once converted into a height map through integration.

Multi-scale blending of the normal map. To mitigate this issue, we adopt a multi-scale approach to first obtain low frequency geometry information and then blend it with the high frequency details. We downscale the input image into a unique 512×512 image, that we process with the same network as the tiles, to get a low-resolution normal map. The large scale components of the material's geometry are entirely contained in this low-resolution normal map, that we blend with the seamless high-resolution normal as follows:

$$normal = \alpha \left(normal_{HrHf} - normal_{HrLf}\right) + (1-\alpha)\, normal_{LrUp} \tag{5}$$

where α is the frequency balance parameter, $normal_{HrHf}$ is the seamless output at the resolution of the input image, $normal_{HrLf}$ is obtained from $normal_{HrHf}$ after removing high frequencies, i.e. by downscaling to 512×512 pixels then upscaling to the input resolution, and $normal_{LrUp}$ is the 512×512 predicted normal upscaled to the input resolution (see Fig. 9).

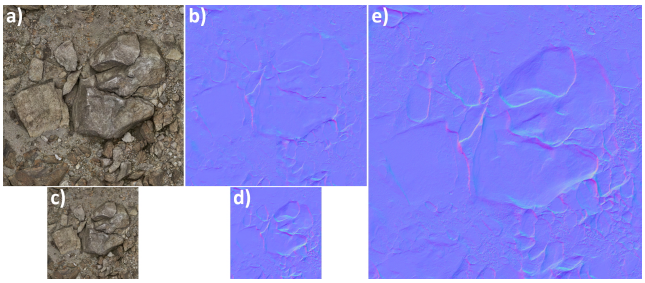

Figure 9: *Multi-scale normal blending. a) Input high-resolution picture from camera (*2048×2048 *pixels); b) Normal inferred from high-resolution input, where large features have not been recovered; c) Low resolution input, after downscale at* 512×512 *pixels; d) Low-resolution normal inferred from low resolution input; e) Resulting high-resolution normal (*2048×2048 *pixels) after multi-scale blending (*α*=0.5), containing both low-frequency geometry and high-frequency details.*

3.5. Full SVBRDF recovery

Our two cascaded U-Nets, coupled with seamless reconstruction, predict full resolution albedo and normal maps, which are the most ambiguous ones in outdoor scenarios, consuming the full neural capacity of the system. Then, the height map is computed from the normal map, through a normal integration following Durou et al. [DC07]. This method implies solving a Poisson system using normal divergence as the guidance field. The ambient occlusion map is computed using *HBAO* (Image-Space Horizon-Based Ambient Occlusion) [BSD08]. The displacement factor to consider for this computation is exposed to the user.

We estimate the roughness map using an approximation that models the relationship between the normal distribution function and the actual surface exposition to outdoor conditions. Essentially, we link the local geometry of the surface to the roughness value through the normalized curvature estimate $\mathcal{H}$ and the ambient occlusion $\mathcal{A}$:

$$roughness = \max(1, (1-\mathcal{H}) + (1-\mathcal{A})) \tag{6}$$

Our insight is that salient edges are more likely to be exposed, and thus polished, so the lower their roughness becomes. Similarly, accessibility (AO) is correlated to surface exposure, hence roughness. This formulation allows recovering fine grain roughness information from the middle range variations of curvature and ambient occlusion. This is clearly a crude approximation, but it is commonly used by 3D artists in material creation workflows. This formulation is illustrated in Fig. 1 of the supplemental material.

Our method allows estimating the roughness variation induced by the high and low frequency geometry of the captured material, considering how much a surface is exposed, however, as we do not constrain the incident lighting at capture time, we cannot be sure to have enough specular information in the input to deduce the overall roughness value. Consequently, we adjust the global roughness value by moving the mid-grey value of the roughness map histogram, which is exposed to the user, as well as the roughness variation range.

Overall, the user controls the process using the ambient occlusion factor (displacement), the mid roughness value and roughness variation amount, as well as the balance α between the low and high frequencies of the normals.

4. Results

4.1. Visual quality

Figure 10 presents an inference result run on synthetic data. Our model successfully recovers albedo and material geometry and allows, after computing the remaining PBR channels, for a realistic render of the extracted virtual material that is very close to the input picture. We can observe that our estimated roughness brings the final material appearance closer to ground truth in the rendering, although there is clearly room for improvement. More results of prediction on synthetic renderings can be found in Section 3.1 of the supplemental material.

In addition, we have created a test set dedicated to checking the

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

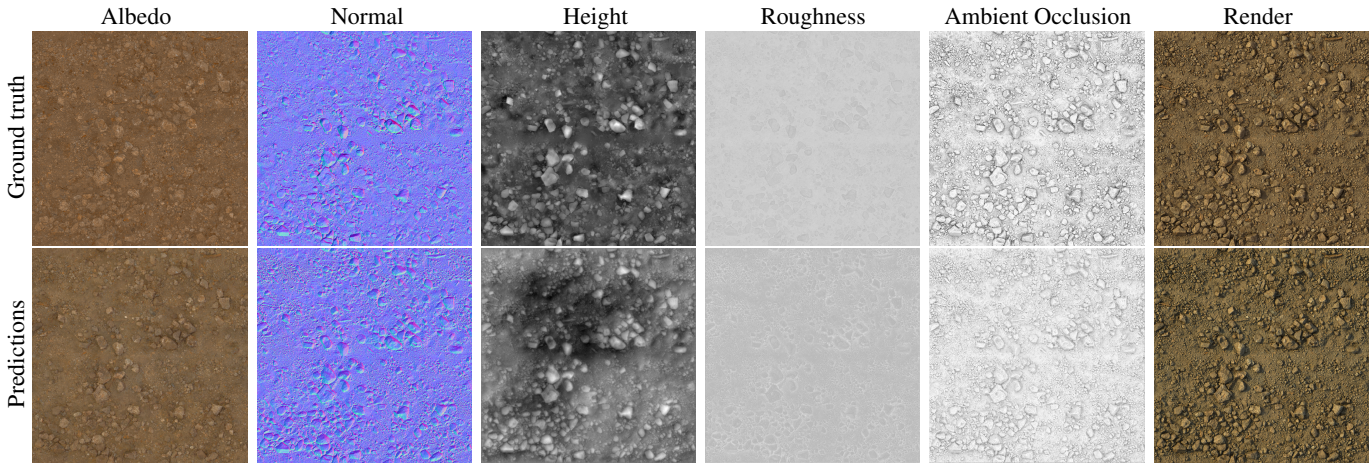

Figure 10: *SVBRDF reconstruction on the synthetic test set (*2048×2048 *pixels), predicted from the render in top-right. The resulting render (bottom-right) allows ensuring that the whole SVBRDF channels are close enough to the ground truth. Our model successfully removes the shadows and the specular highlights from the input image, and provides a result close to the ground truth albedo, with details and contrast preserved. The material geometry is also well recovered in the normal map including sharp edges. While the local effect of light is well removed, we observe some global color drift, where the predicted albedo has an average color closer to the input than to the ground truth. This shows that our network can not always decouple the intrinsic material color from that of the environment, which is a complex semantic task, the global environment color being highly dependent on the location of the capture and type of material.*

visual quality of the network output on a selection of real captures. For this purpose, we have aggregated 120 pictures of grounds and stones taken in the context of photometry and photogrammetry, using several camera devices, under various lighting conditions and scales. We show four of these results in Fig. 11, additional samples in Figure 3 of the supplemental material, and the whole test set is also available as supplemental material.

4.2. Evaluating predictions

We evaluate the quality of the network prediction using three metrics: Mean Squared Error, Mean Absolute Error, and SSIM Error, comparing the predicted maps to the ground truth ones (see Tab. 1).

4.3. Ablation study

We started our work on the Delighter UNet by predicting directly the albedo map without the intermediary irradiance and specular maps. The dataset was first made of procedural materials, rendered with greyscale environment maps to avoid color shifting due to the color temperature of the environment map light. Results are satisfying on inputs with small shadows and with a neutral color temperature. Fig. 12 shows two failure cases of this version, presented in column b).

Metric	Input/GT albedo	Albedo pred/GT	Normal pred/GT
MAE	0.07671	0.0507	0.0406
MSE	0.0150	0.00631	0.0049
SSIM	0.3834	0.240	0.263

Table 1: *Average metrics computed on the entire test set: Mean Average Error (MAE), Mean Square Error (MSE) and Structural Similarity (SSIM). The* Input/GT albedo *metric represents how different the input image is from the target ground truth albedo.*

We make the hypothesis that the network needs more capacity to interpret the effect of light on the material color. Hence, we train it to infer an irradiance map and a specular map, by designing the predicted albedo as the result of the illumination decomposition formula presented in Section 3.3:

$$albedo = \frac{input - specular}{irradiance} \tag{7}$$

However, with no ground truth associated to these illumination maps, the loss can only be computed on the target albedo, leading to unexpected information in the predicted irradiance and specular maps, even though the resulting albedo has reasonable quality on most examples. In addition, this decomposition does not solve the failure cases (see column c) of Fig. 12) and this version tends to shift towards red colors in the darkest shadows.

These results highlight several issues with this dataset. First, using grayscale environment maps prevents the network from learning the effects of color temperature of the sun and sky on lit and occluded parts. Second, this dataset lacks examples of large shadows cast by small objects such as stones and wood sticks. We then reworked this dataset to include colored environment maps, material augmentation using atlases (see Section 3.2) and generation of the ground truth irradiance and specular maps associated to each render. We changed the loss to a combination of the error on the predicted irradiance, specular, and albedo maps. Column d) of Fig. 12 shows a better delighting on three examples. The shadows are better removed and the overall contrast of the albedo is increased. However, as the illumination decomposition involves a division by

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

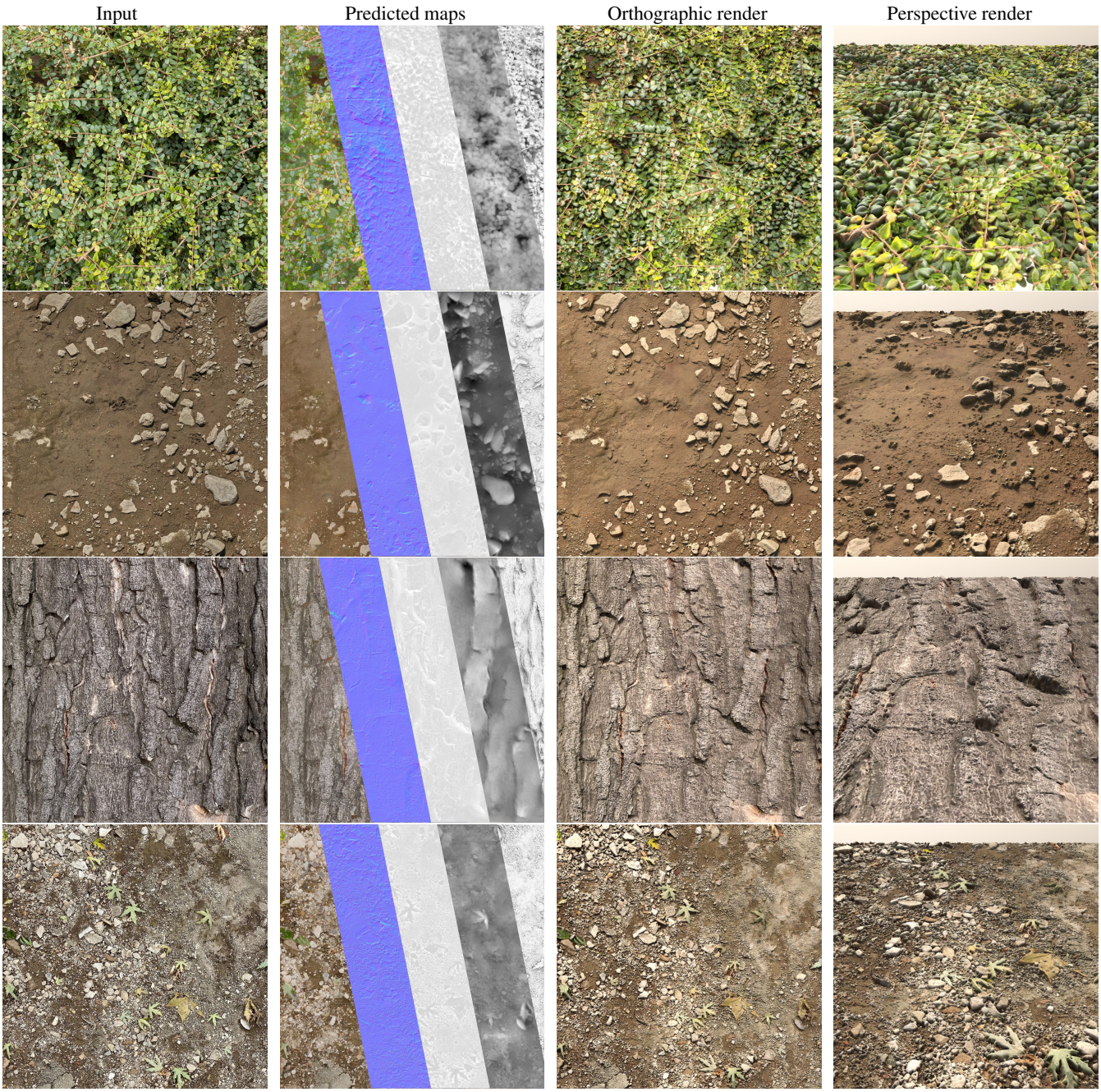

Figure 11: *MaterIA results on four real use cases. Left: Input picture in* 4096×4096 *pixels, taken with a smartphone in natural lighting condition. Second column: Predicted albedo, normal map, roughness, height and ambient occlusion. Last two columns: Resulting material rendered with an environment map that reproduces closely the lighting environment of the capture. Our model generalizes well on these real pictures (left) and allows capturing a complete high-resolution PBR material (second column) that can be synthetically lit with no shadows artifacts and a convincing recovery of the geometry, as can be assessed from the new generated shadows (right).*

the irradiance map, the results can be unstable in the darkest areas, where saturated pixels can appear. This is also caused by the missing information in the input in saturated areas (darkest shadows and specular spots).

The final Delighter architecture adds a third output called the enhancer map, whose role is to help recover information in these saturated areas. The enhancer map has no ground truth and is added to the input before applying the illumination decomposition. Column e) presents the results of this last Delighter version and shows

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

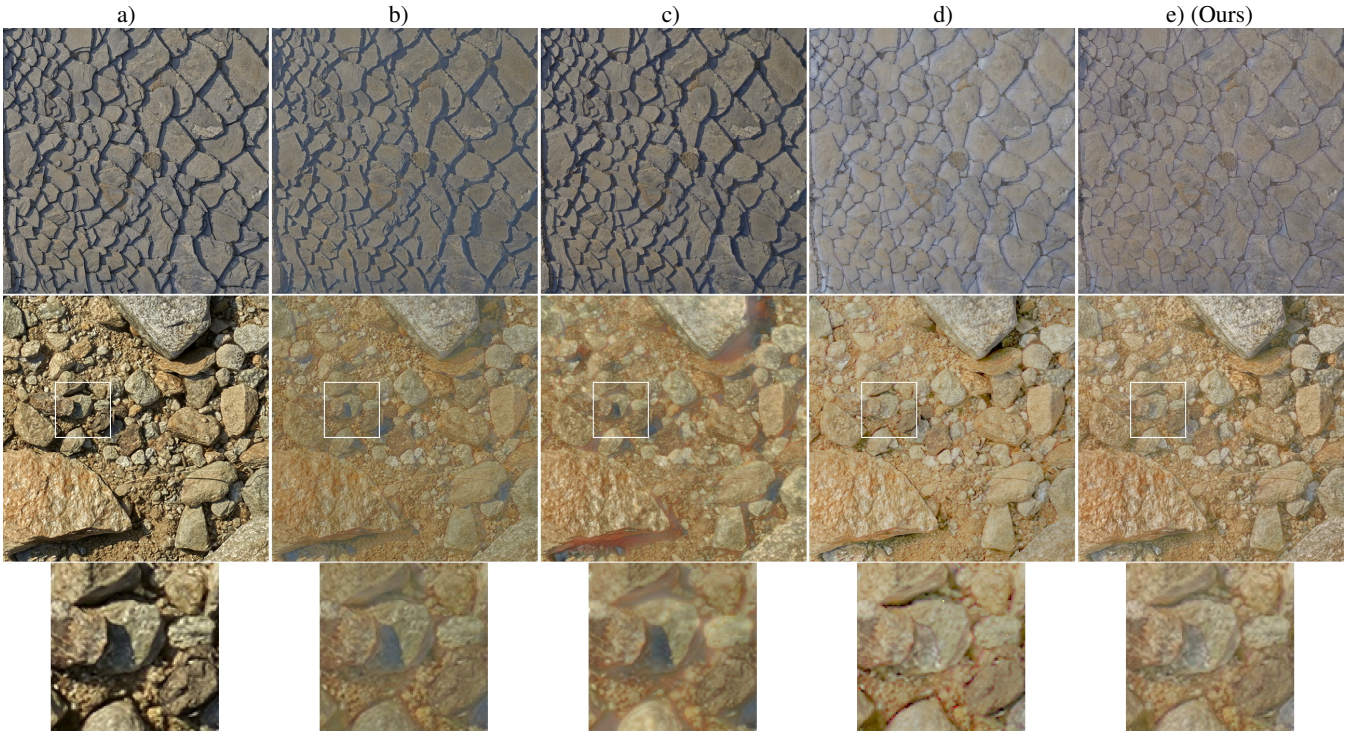

Figure 12: *Ablation study on the delighting. a) Input picture from camera (512 × 512 pixels). b) Delighter UNet [MMP19] trained at predicting directly the albedo map, using the initial dataset made of materials rendered with greyscaled environment map, without atlas augmentation. c) Modification of b) to predict an irradiance and a specular output, with no supervision on these maps. The loss is computed on the resulting albedo only, which leads to a lack of constraint on the irradiance and specular outputs, and unsatisfying delighting results. d) Reworked dataset with colored environment maps and atlas augmentation, and ground truth irradiance and specular maps generated. The Delighter is trained at predicting the irradiance and specular maps, with a loss involving these two maps only. The albedo is deduced using the light decomposition formula (detailed in 3.3). The last row is a zoom of the second row and allows to observe some saturated pixels in the darkest areas, in the form of chromatic aberrations. e) Adding the enhancer map as third output map of the Delighter network mitigates the issues in the darkest areas. The loss becomes a combination of irradiance, specular and resulting albedo. This is our final Delighter network.*

that the aberrations visible in column d) have disappeared, while producing comparable delighting results.

4.4. Comparison to existing work

Method	Albedo pred/GT		Normal pred/GT	
	RMSE	SSIM	RMSE	SSIM
[YLD*18]	0.166	0.613	0.119	0.484
[DAD*18] (no flash)	0.223	0.598	0.135	0.507
[DAD*18] (our data)	0.166	0.589	0.135	0.496
[Tin20]	–	–	0.106	0.592
Ours	**0.126**	**0.722**	**0.104**	**0.630**

Table 2: *Average metrics computed on 28 ray-traced synthetic renders at 1024 × 1024 pixels with ground truth maps: Root Mean Square Error (RMSE, lower is better) and Structural Similarity with a kernel of 7 (SSIM, higher is better).* Deepbump *[Tin20] only extracts a normal map, hence the missing Albedo metric. Examples of input images used to compute these metrics are given in Figure 2 of the supplemental material.*

In order to evaluate the quality and robustness of our method against state of the art, we compare our SVBRDF extraction with three existing methods from Ye et al. [YLD*18], Deschaintre et al. [DAD*18] and Hugo Tini [Tin20]. For each method, we provide a quantitative evaluation of the predicted albedo and normal against the ground truth maps (Table 2). We also show qualitative results for predicted albedo and normal (Fig. 13) as well as renders (Fig. 14).

For the method of Deschaintre et al. [DAD*18], we use the network trained with directional lighting and sky dome [DDB20], that we call "no flash", so the results are closer to our work. In addition, we trained this same network and renderer on the material maps generated using our method, that we call "our data", to highlight the added quality brought about by our data augmentation approach. However, we can notice that the lack of displacement in the renderings produced by this method during the training prevents removing the strong shadows, even when trained on our atlas-augmented material dataset. Overall, we can observe that our method provides a significant improvement in the visual quality of the resulting SVBRDF.

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

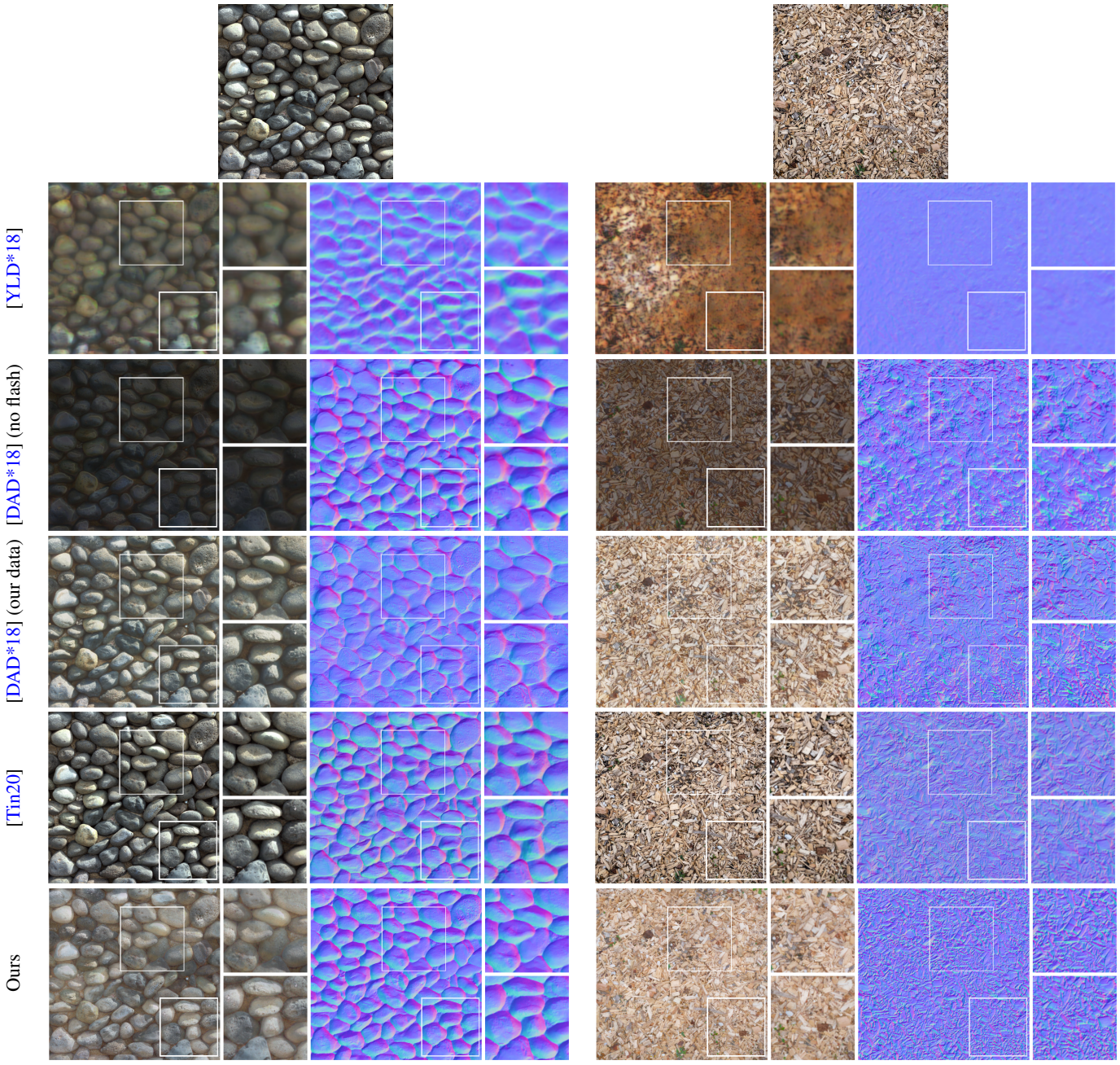

Figure 13: *Comparison of predicted albedo and normal maps for different methods. For both of these real examples, we can see that our method recovers well delighted albedo maps, and provides a normal map with both large scale features and small details. On the left, the very strong sun light leads to some remaining soft shadows in our albedo, which might be visible when rendering it with novel views or lights. While some other methods achieve a result similar to ours for one map or the other, none of them reach a high quality for the two maps. More results of predicted maps are given in Figure 4 of the supplemental material.*

The normal maps of Tini [Tin20] are comparable to our reconstructed geometry in both high and low frequency, but we also provide basecolor and additional SVBRDF parameters. In some examples, the normal maps of Tini have slightly stronger gradients, in particular when mixing flat and bumpy areas, such as pebbles in grass (Fig. 14, left) or in mud (Fig. 5 of the supplemental material, left).

As our albedo extraction using illumination decomposition can relate to the intrinsic image decomposition domain, we compare our method to three intrinsic image decomposition approaches, Yu

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

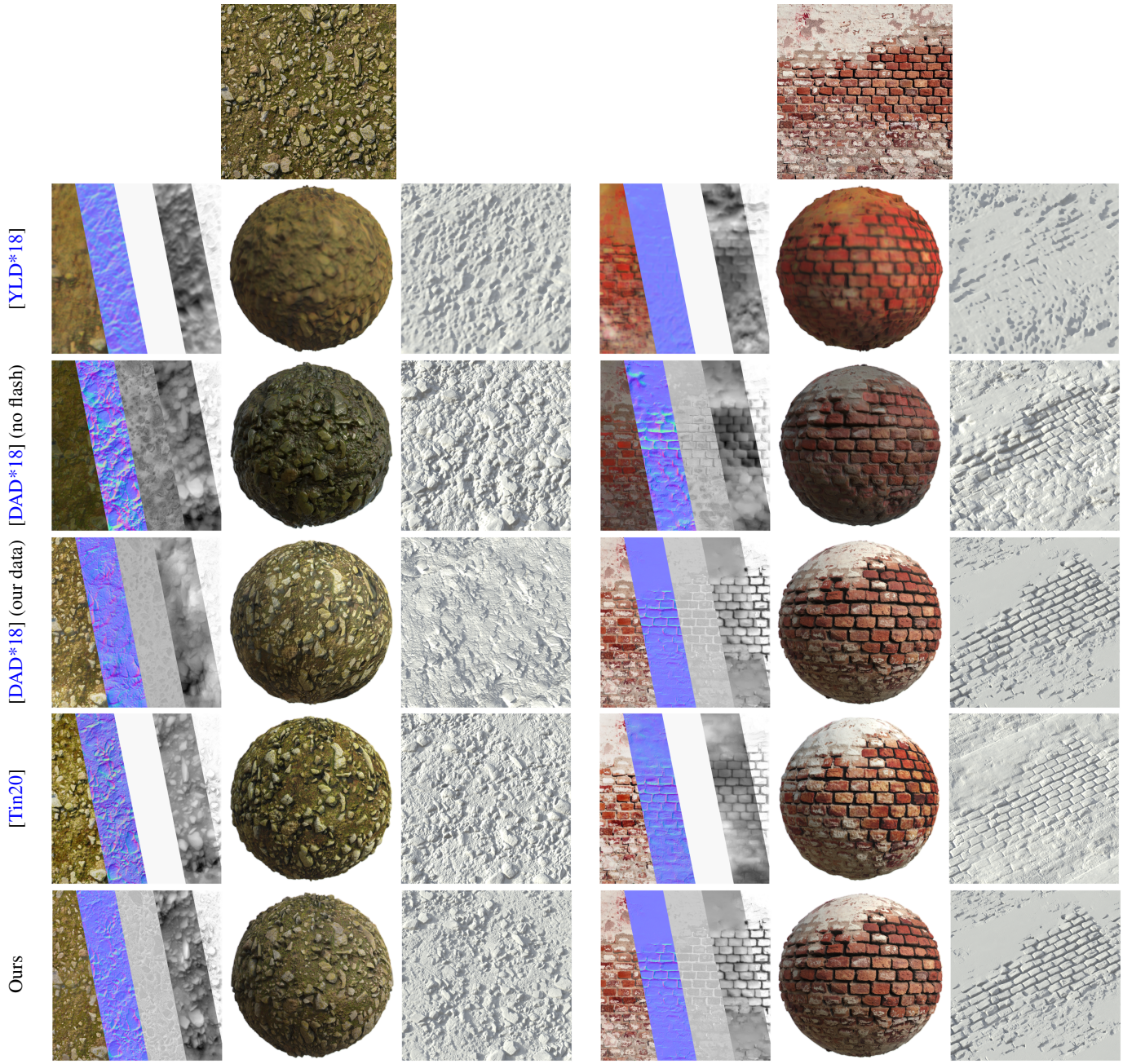

Figure 14: *Comparison of SVBRDF acquisition and resulting renders for different methods on two real samples. Top row: input photo. Then, for each method, from left to right: extracted SVBRDF maps (albedo, normal, roughness) as well as height and ambient occlusion maps computed with the same process as our method, resulting material rendered on a sphere, clay render on a plane. Thanks to the well-delighted albedo and accurate normal map, our method allows correct rendering of the input material in new lighting conditions. The white clay renders highlight the high quality geometry recovered by our method. More renders are given in Fig. 5 of the supplemental material.*

et al. [**yu_2019**], Liu et al. [LYYL20] and Luo et al. [LHL*20]. Fig. 15 shows that these methods have a tendency to produce flat solid colors and do not preserve the high frequency details in the resulting reflectance. This is suitable for the *image manipulation* domain which benefits from large flat regions, but not for the *material reconstruction* domain, where the albedo map is expected to contain higher frequency content for realistic rendering. Fig. 16 shows that our albedo and normal recovery results in a better geometry interpretation than the reflectance and normal extraction of

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

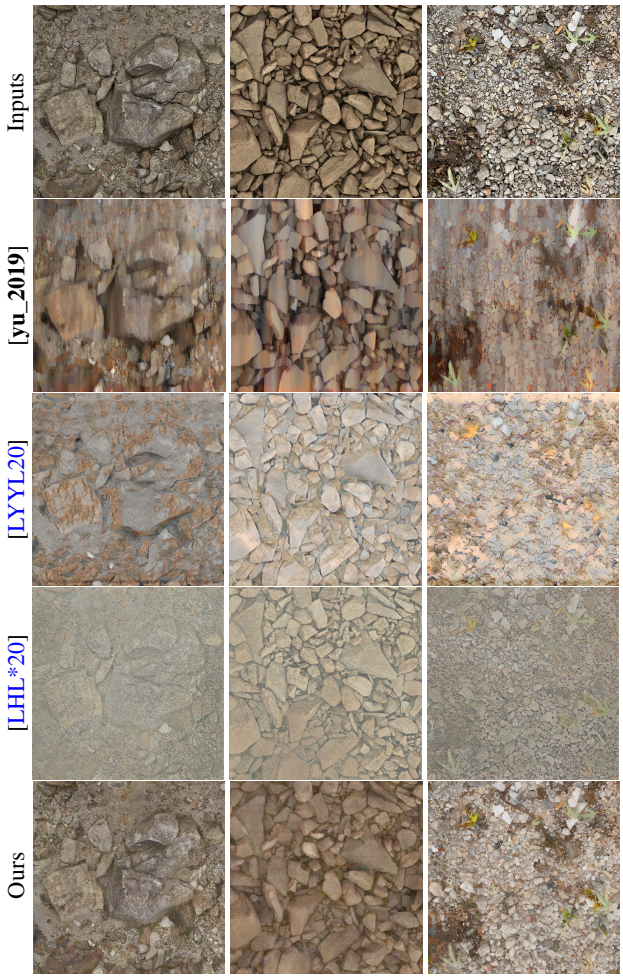

Figure 15: *Comparison of our delighting method with three intrinsic image decomposition methods, on three real examples. From top to bottom: input photos, results of* [**yu_2019**]*,* [LYYL20]*,* [LHL*20] *and ours. For each method, we ran the pretrained network available online and show the resulting reflectance. We can observe that* [**yu_2019**] *and* [LYYL20] *methods produce somehow blurry results with unexpected color artifacts, while* [LHL*20] *prediction is closer to the expected output. However, it lacks contrast and fine grain details that are required for the albedo of a SVBRDF material, and that our method preserves well.*

Luo et al. [LHL*20]. As such, our material-specific data set could be applied to such intrinsic decomposition methods.

4.5. User study

In order to evaluate the quality of our material extraction from the point of view of 3D artists, we ran a user study by asking 21 artists to rank the predictions and subsequent renderings given by four compared methods. To compute normalized scores, we assign a score of 1.0 to the first ranked method, 0.6 for the second, 0.3 for the third, and 0.0 for the last one. Then, we average these scores over all users and examples, to get the final scores in Table 3. In addition, screenshots of the survey sent to users are given in Figure 6 of the supplemental material.

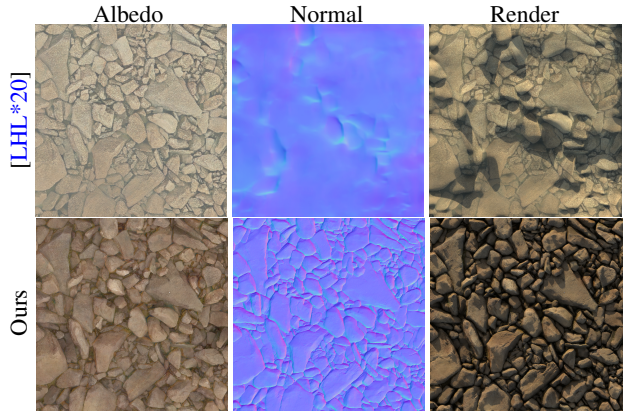

Figure 16: *Comparison to* [LHL*20] *for the geometry components extraction, on the second example of Fig.15. Only the albedo and normal predictions are used for our method, the height map is computed from the normal map the same way for both methods, and a constant roughness has been set.* [LHL*20] *does not predict a proper geometry of the captured material.*

4.6. Implementation details and performance

We use the Tensorflow framework in Python [MAP*15] to train our model. Our network requires around 20GB of memory during the training stage. We use an NVidia Quadro GV100 GPU for the training, and it takes around five days to get a fully converged model. To deploy our model into the commercial application, we implemented our network using a C++ API based on native CUDA and CuDNN libraries. We export the weights of the network from Tensorflow as a binary file, that is read by the C++ API when the network is built. We also developed optimized CUDA kernels for the operations that are not natively available in CuDNN .

The Poisson equation is solved using the Intel®MKL Poisson library, using the color gradients as vector fields for the albedo, and the spherical coordinates gradients for the normal map. We run the following process for both albedo and normal maps. The gradients

Method	Albedo	Normal
[YLD*18]	0.38	0.03
[DAD*18] (no flash)	0.48	0.38
[Tin20]	–	**0.77**
Ours	**0.75**	0.73

Table 3: *User study scores for different methods, averaged over 10 material extraction examples and 21 survey responses.* Deepbump *[Tin20] only extracts a normal map, hence the missing Albedo score. While we can see that the albedo predicted by our method is largely preferred by artists, our normal map has a score similar to* Deepbump*, which shows that both method extract an accurate geometry from the input.*

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

Step	Duration (ms)
Inference high-resolution ($16 \times 512 \times 512$ tiles)	0.75
Inference low-resolution (single 512×512 tile)	0.7
Poisson solving on albedo and normal	1.0
Multi-scale normal blending	0.2
Height computation	1.0
AO and Roughness computation	0.5

Table 4: *MaterIA performance from a* 2048×2048 *pixels input picture on a NVidia Quadro GV100 GPU.*

of the prediction in the horizontal and vertical axes are computed for each tile and channel in a dedicated CUDA kernel, then stitched together similarly to the predicted tiles. We define the boundary of the Poisson equation as the one-pixel border of the entire image, and set the Dirichlet boundary condition using the predicted values (RGB values for the albedo and angular coordinates for the normal). We set the gradient to zero at the boundary of all tiles in the considered direction; e.g. we set a zero gradient for all pixels of the right border in the horizontal gradient, and respectively for the bottom border in the vertical gradient. Finally, acquiring the full SVBRDF from a 2048×2048 pixels picture takes around 4 seconds, with the time breakdown shown in Tab. 4.

5. Limitations and future work

Our solution shows limitations in the predicted normal map if the low-resolution and high-resolution inferences are opposite. In that case, blending them cancels the relief and leads to unsatisfactory results. We plan to rework the network architecture to condition the high-resolution inference to the low-resolution result. Adopting a multi-scale approach directly within the network architecture should also remove the need for post-process blending. Also, using a synthetic dataset made of material renders introduces a bias in the data learned, and even if we found out that our model generalizes pretty well on real pictures, we know that some realistic effects cannot be represented in 2.5 dimensions, in particular the shadows cast by small objects that are not directly on the floor, like twigs for instance. Last, we would like to investigate network compression to reduce the memory footprint of our network once deployed.

Capturing a material in the wild implies having the most flexible setup, in particular for the lighting conditions. This would make a roughness map prediction very unstable as the input may not contain any meaningful information allowing to determine the material roughness. This is why we adopted this roughness approximation based on the curvature and ambient occlusion, combined with a user input to adjust the global range. Our method, although not accurate and requiring user input, provides satisfactory results in the targeted domain of outdoor materials.

Our method is not designed to extract SVBRDF from metallic objects. As we work in the roughness/metallic workflow (see introduction), the albedo used to train our delighter network is not split between diffuse and specular maps, hence cannot be used to estimate a metallic map. Predicting the latter could be the object of further research, as our data creation pipeline allows to generate a metallic map from procedural materials.

6. Conclusion

We proposed a method which evolves state-of-the-art in high-resolution SVBRDF capture from a single outdoor image, interleaving deep learning and explicit non-learned formulations. Our method successfully removes the shadows coming from the lighting while preserving the contrast of the input, resulting in a high quality albedo map. As such, our method leverages domain learning to solve the under-constrained problem of delighting, while using the full capacity of single-task specialized networks. The decomposition of the albedo map following the illumination, i.e. into irradiance and specular contribution maps, helps the network to understand the semantics of the image. Exploiting this semantic information by concatenating the predicted irradiance map to the input of the normal map inference shows good results at recovering the geometry. Moreover, our multi-scale post-processing of the normal map leads to well-preserved low and high frequencies, that reflect in a height map reproducing faithfully the captured geometry. Eventually, the fast computation of the roughness, even though not fully accurate, shows good results allowing the addition of small details in the light reflection when rendering the captured PBR material. Additionally, our procedural data generation and augmentation strategy effectively produced a huge and realistic dataset of synthetic materials, with in particular the atlas splatting strategy overcoming the lack of highly irregular materials. Last, our tiling approach reduces the memory footprint while processing high resolution inputs, with a seamless reconstruction for high quality resulting materials. Last, we optimized our model using NVidia CuDNN and deployed it within a widely used software product.

Acknowledgements. We thank Nicolas Wirrmann for the Adobe Substance Designer renderer, used for rendering our materials. We thank all the technical artists who created the Substance 3D Assets library. We thank Jean-François El Hajjar for his implementation of the normal integration algorithm, used to recover the height map. We also thank Arthur Meyer for his early contribution to the delighter. We thank Baptiste Manteau and Maxime Morel for their help on the product deployement of our method. Finally we thank Valentin Deschaintre for his feedback on the manuscript.

References

[AAL16] AITTALA, MIIKA, AILA, TIMO, and LEHTINEN, JAAKKO. "Reflectance modeling by neural texture synthesis". *ACM ToG* 35.4 (2016), 1–13 3.

[Ado21] ADOBE. *Substance Source*. https://substance3d.adobe.com/assets/. 2021 4.

[BJK*20] BOSS, MARK, JAMPANI, VARUN, KIM, KIHWAN, et al. "Two-Shot Spatially-Varying BRDF and Shape Estimation". *Proc. CVPR*. 2020, 3981–3990 3.

[BSD08] BAVOIL, LOUIS, SAINZ, MIGUEL, and DIMITROV, ROUSLAN. "Image-Space Horizon-Based Ambient Occlusion". *ACM SIGGRAPH 2008 Talks*. 2008, 22:1–22:1 7.

[CT82] COOK, R. L. and TORRANCE, K. E. "A Reflectance Model for Computer Graphics". *ACM ToG* 1.1 (1982), 7–24 4.

[DAD*18] DESCHAINTRE, VALENTIN, AITTALA, MIIKA, DURAND, FREDO, et al. "Single-image SVBRDF capture with a rendering-aware deep network". *ACM ToG* 37.128 (2018), 1–15 3, 4, 10–13.

[DC07] DUROU, JEAN-DENIS and COURTEILLE, FREDERIC. "Integration of a Normal Field without Boundary Condition". *Proc. PACV*. 2007, 8 p. 7.

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.

[DDB20] DESCHAINTRE, VALENTIN, DRETTAKIS, GEORGE, and BOUSSEAU, ADRIEN. "Guided Fine-Tuning for Large-Scale Material Transfer". *Computer Graphics Forum* 39.4 (2020), 91–105 3, 10.

[DLG21] DESCHAINTRE, VALENTIN, LIN, YIMING, and GHOSH, ABHIJEET. "Deep polarization imaging for 3D shape and SVBRDF acquisition". *Proc. CVPR*. 2021, 15567–15576 3.

[FHLD05] FINLAYSON, GRAHAM D, HORDLEY, STEVEN D, LU, CHENG, and DREW, MARK S. "On the removal of shadows from images". *IEEE TPAMI* 28.1 (2005), 59–68 2.

[GGG*16] GUARNERA, DARYA, GUARNERA, GIUSEPPE CLAUDIO, GHOSH, ABHIJEET, et al. "BRDF representation and acquisition". *Computer Graphics Forum* 35.2 (2016), 625–650 3.

[GLD*19] GAO, DUAN, LI, XIAO, DONG, YUE, et al. "Deep inverse rendering for high-resolution SVBRDF estimation from an arbitrary number of images." *ACM ToG* 38.4 (2019), 134–1 3.

[GLT*21] GUO, JIE, LAI, SHUICHANG, TAO, CHENGZHI, et al. "Highlight-aware two-stream network for single-image SVBRDF acquisition". *ACM ToG* 40.4 (2021), 1–14 3.

[GRR*18] GEORGOULIS, STAMATIOS, REMATAS, KONSTANTINOS, RITSCHEL, TOBIAS, et al. "Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning". *IEEE TPAM* 40.8 (2018), 1932–1947 3.

[GSH*20] GUO, YU, SMITH, CAMERON, HAŠAN, MILOŠ, et al. "MaterialGAN: Reflectance Capture using a Generative SVBRDF Model". *ACM ToG* 39.6 (2020), 254:1–254:13 3.

[HDMR21] HENZLER, PHILIPP, DESCHAINTRE, VALENTIN, MITRA, NILOY J., and RITSCHEL, TOBIAS. "Generative Modelling of BRDF Textures from Flash Images". *ACM ToG* 40.6 (2021), 284:1–284:13 4.

[HDR19] HU, YIWEI, DORSEY, JULIE, and RUSHMEIER, HOLLY. "A Novel Framework for Inverse Procedural Texture Modeling". *ACM ToG* 38.6 (2019), 1–14 4.

[IZZE17] ISOLA, PHILLIP, ZHU, JUN-YAN, ZHOU, TINGHUI, and EFROS, ALEXEI. "Image-to-Image Translation with Conditional Adversarial Networks". *Proc. CVPR*. 2017, 5967–5976 3.

[JA11] JOHNSON, MICAH K and ADELSON, EDWARD H. "Shape estimation in natural illumination". *Proc. CVPR*. 2011, 2553–2560 3.

[JWK*17] JANNER, MICHAEL, WU, JIAJUN, KULKARNI, TEJAS D, et al. "Self-Supervised Intrinsic Image Decomposition". *Proc. NIPS*. 2017, 5938–5948 3.

[Kar13] KARIS, BRIAN. *Real Shading in Unreal Engine 4*. Tech. rep. Epic Games, 2013 4.

[KLA21] KARRAS, TERO, LAINE, SAMULI, and AILA, TIMO. "A Style-Based Generator Architecture for Generative Adversarial Networks". *IEEE TPAMI* 43.12 (2021), 4217–4228 3.

[LCY*17] LIU, GUILIN, CEYLAN, DUYGU, YUMER, ERSIN, et al. "Material Editing Using a Physically Based Rendering Network". *Proc. ICCV*. 2017, 2280–2288 3.

[LDPT17] LI, XIAO, DONG, YUE, PEERS, PIETER, and TONG, XIN. "Modeling Surface Appearance from a Single Photograph using Self-augmented Convolutional Neural Networks". *ACM ToG* 36.4 (2017), 1–11 3.

[LEN10] LALONDE, JEAN-FRANÇOIS, EFROS, ALEXEI A, and NARASIMHAN, SRINIVASA G. "Detecting ground shadows in outdoor consumer photographs". *Proc. ECCV*. 2010, 322–335 2.

[LHL*20] LUO, JUNDAN, HUANG, ZHAOYANG, LI, YIJIN, et al. "NIID-Net: Adapting Surface Normal Knowledge for Intrinsic Image Decomposition in Indoor Scenes". *TVCG* 26.12 (2020), 3434–3445 3, 12, 13.

[LSC18] LI, ZHENGQIN, SUNKAVALLI, KALYAN, and CHANDRAKER, MANMOHAN. "Materials for Masses: SVBRDF Acquisition with a Single Mobile Phone Image". *Proc. ECCV*. Vol. 11207. 2018, 74–90 3, 4.

[LWSJ21] LICHY, DANIEL, WU, JIAYE, SENGUPTA, SOUMYADIP, and JACOBS, DAVID W. "Shape and Material Capture at Home". *Proc. CVPR*. 2021, 6119–6129 3.

[LXR*18] LI, ZHENGQIN, XU, ZEXIANG, RAMAMOORTHI, RAVI, et al. "Learning to reconstruct shape and spatially-varying reflectance from a single image". *ACM ToG* 37.6 (2018), 1–11 3.

[LYYL20] LIU, YUNFEI, YU, LI, YOU, SHAODI, and LU, FENG. "Unsupervised Learning for Intrinsic Image Decomposition From a Single Image". *Proc. CVPR*. 2020, 3245–3254 3, 12, 13.

[MAP*15] MARTIN ABADI, ASHISH AGARWAL, PAUL BARHAM, et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015 13.

[Mar98] MARSCHNER, STEPHEN ROBERT. "Inverse rendering for computer graphics". PhD thesis. Cornell University, 1998 3.

[MCZ*18] MA, WEI-CHIU, CHU, HANG, ZHOU, BOLEI, et al. "Single image intrinsic decomposition without a single intrinsic image". *Proc. ECCV*. 2018, 201–217 3.

[MMP19] MARTIN, ROSALIE, MEYER, ARTHUR, and PESARE, DAVIDE. "De-lighting a High-resolution Picture for Material Acquisition". *EGSR / Industry Track*. 2019, 69–72 3, 10.

[PF14] PAPADHIMITRI, THOMA and FAVARO, PAOLO. "Uncalibrated Near-Light Photometric Stereo". *Proc. BMVC*. 2014 1.

[PGB03] PEREZ, PATRICK, GANGNET, MICHEL, and BLAKE, ANDREW. "Poisson Image Editing". *ACM ToG* 22.3 (2003), 313–318 6.

[QTH*17] QU, LIANGQIONG, TIAN, JIANDONG, HE, SHENGFENG, et al. "DeshadowNet: A Multi-context Embedding Deep Network for Shadow Removal". *Proc. CVPR*. 2017, 2308–2316 2.

[RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. "U-Net: Convolutional Networks for Biomedical Image Segmentation". *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. 2015, 234–241 2.

[RRF*16] REMATAS, KONSTANTINOS, RITSCHEL, TOBIAS, FRITZ, MARIO, et al. "Deep Reflectance Maps". *CVPR*. 2016, 4508–4516 3.

[SC20] SANG, SHEN and CHANDRAKER, MANMOHAN. "Single-Shot Neural Relighting and SVBRDF Estimation". *Proc. ECCV*. 2020, 85–101 3.

[SLH*20] SHI, LIANG, LI, BEICHEN, HAŠAN, MILOŠ, et al. "MATch: Differentiable Material Graphs for Procedural Material Capture". *ACM ToG* 39.6 (2020), 1–15 4.

[SSL12] SANIN, ANDRES, SANDERSON, CONRAD, and LOVELL, BRIAN C. "Shadow detection: A survey and comparative evaluation of recent methods". *Pattern recognition* 45.4 (2012), 1684–1695 2.

[Tin20] TINI, HUGO. *DeepBump*. May 2020. URL: https://hugotini.github.io/deepbump 3, 10–13.

[Wei01] WEISS, YAIR. "Deriving intrinsic images from image sequences". *Proc. ICCV*. Vol. 2. 2001, 68–75 3.

[WLY18] WANG, JIFENG, LI, XIANG, and YANG, JIAN. "Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal". *Proc. CVPR*. 2018, 1788–1797 2.

[WMLT07] WALTER, BRUCE, MARSCHNER, STEPHEN R., LI, HONGSONG, and TORRANCE, KENNETH E. "Microfacet Models for Refraction through Rough Surfaces". *Proc. EGSR*. 2007, 195–206 4.

[Woo80] WOODHAM, ROBERT J. "Photometric method for determining surface orientation from multiple images". *Optical engineering* 19.1 (1980), 191139 3.

[YLD*18] YE, WENJIE, LI, XIAO, DONG, YUE, et al. "Single Image Surface Appearance Modeling with Self-augmented CNNs and Inexact Supervision". *Computer Graphics Forum* 37.7 (2018), 201–211 10–13.

[ZK21] ZHOU, XILONG and KALANTARI, NIMA KHADEMI. "Adversarial Single-Image SVBRDF Estimation with Hybrid Training". *Computer Graphics Forum* 40.2 (2021), 315–325 3.

[ZWX*20] ZHAO, YEZI, WANG, BEIBEI, XU, YANNING, et al. "Joint SVBRDF Recovery and Synthesis From a Single Image using an Unsupervised Generative Adversarial Network". *EGSR- DL-only Track*. 2020, 53–66 3.

© 2022 The Author(s)
Computer Graphics Forum © 2022 The Eurographics Association and John Wiley & Sons Ltd.