

Reflectance and Natural Illumination from Single-Material Specular Objects Using Deep Learning

Stamatios Georgoulis[†], Konstantinos Rematas[†], Tobias Ritschel, Efstratios Gavves,
Mario Fritz, Luc Van Gool, and Tinne Tuytelaars

Abstract—In this paper, we present a method that estimates reflectance and illumination information from a single image depicting a single-material specular object from a given class under natural illumination. We follow a data-driven, learning-based approach trained on a very large dataset, but in contrast to earlier work we do not assume one or more components (shape, reflectance, or illumination) to be known. We propose a two-step approach, where we first estimate the object's reflectance map, and then further decompose it into reflectance and illumination. For the first step, we introduce a Convolutional Neural Network (CNN) that directly predicts a reflectance map from the input image itself, as well as an indirect scheme that uses additional supervision, first estimating surface orientation and afterwards inferring the reflectance map using a learning-based sparse data interpolation technique. For the second step, we suggest a CNN architecture to reconstruct both Phong reflectance parameters and high-resolution spherical illumination maps from the reflectance map. We also propose new datasets to train these CNNs. We demonstrate the effectiveness of our approach for both steps by extensive quantitative and qualitative evaluation in both synthetic and real data as well as through numerous applications, that show improvements over the state-of-the-art.

Index Terms—Reflectance maps; Intrinsic images; Reflectance; Natural illumination; Specular shading; Convolutional Neural Networks

1 INTRODUCTION

A classic computer vision task is the decomposition of an image into its intrinsic properties, i. e. its shape, reflectance, and illumination. The physics of image formation is based on the complex interplay of these properties; the light (i. e. illumination) hits a surface with specific orientation (i. e. shape) and material properties (i. e. reflectance) and is reflected to the camera. Factoring an image into its intrinsic components, however, is a very difficult and under-constrained task, as the same visual result might be due to many different combinations of intrinsic object properties.

For the estimation of those properties a common practice in literature is to assume one or more components to be known or simplified and try to estimate the others. On the one hand, traditional approaches to intrinsic images or shape-from-shading try to constrain either reflectance, by assuming Lambertian materials [1], [2], [3], or illumination, by having a controlled lighting environment such as point light sources [4], [5], [6]. On the other hand, recent approaches allow for less constrained reflectance and illumination. Yet, in this case shape is either assumed to be known, given in the form of a scanned 3D model [7], [8], or it is restricted to having trivial geometry (e. g. spheres) [9].

We go beyond these simplifying assumptions and estimate reflectance and illumination in a more general setting where the shape of the object is not given but instead it comes from a known class (e. g. cars). This is motivated by the observation that humans may exploit prior knowledge and expectations (e. g. car bodies have similar local structures) to deal with ambiguities in perception

[10]. As such, we hope that focusing on objects from a known class would allow us to exploit similar cues in a learning-based scheme. Furthermore, we observe that there are strong priors about illumination and photo content (e. g. the sky is blue and always on top), that are harder to capture in parametric models [11], [12] or carefully designed physics formulas [7], [8]. Instead, going for a data-driven, learning-based approach trained on a very large dataset, allows us to leverage such priors in the learning process. The latter is essential for dealing with the ambiguous cases one encounters in these decomposition problems.

To keep the complexity of the problem under control, we propose a two-step approach. First, we estimate a shape-independent representation of the appearance, in the form of a reflectance map [13]. Second, we decompose it into material and illumination. To carry out these tasks we employ Convolutional Neural Networks (CNNs) as they have shown unprecedented performance in other tasks with similar requirements [2], [14], [15]. The input to our method is a single 2D image, depicting a single-material object from a known class, and its segmentation mask. The latter is just used for background removal. The output is the reflectance of the object, expressed as Bidirectional Reflectance Distribution Function (BRDF) [16] parameters, and the illumination, expressed as a high dynamic range (HDR) spherical illumination map. An overview of our two-step pipeline can be seen in Fig. 1.

Besides allowing for a better understanding and analysis of 2D imagery, the ability to estimate reflectance maps lends itself to a broad spectrum of applications, including appearance transfer, inpainting, and augmented reality, while its further decomposition into reflectance and illumination enables powerful image editing applications, such as material transfer and illumination editing.

As mentioned earlier, we opted for a two-step approach where: (1) we estimate a reflectance map from a 2D image (Fig. 1, Step 1), and (2) we decompose it into reflectance and illumination (Fig. 1, Step 2). There are several reasons behind this choice: (1)

- [†] S. Georgoulis and K. Rematas contributed equally to this work.
- S. Georgoulis, L. Van Gool and T. Tuytelaars are with KU Leuven, ESAT-PSI, iMinds, Belgium. L. Van Gool is also with ETH Zurich, Switzerland.
- K. Rematas is with University of Washington, USA.
- T. Ritschel is with University College London, UK.
- E. Gavves is with University of Amsterdam, Netherlands.
- M. Fritz is with Max Planck Institute for Informatics, Germany.

Manuscript received December 21 2016.

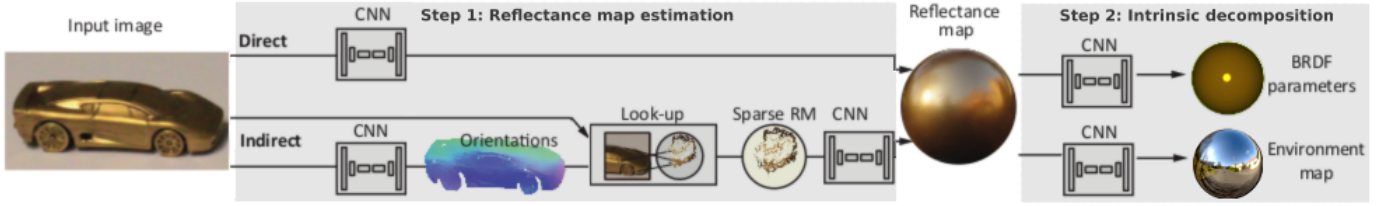


Fig. 1. Overview of our approach. From the input image, in a first step we estimate a reflectance map either directly from the input image itself or indirectly with additional supervision, and in a second step we decompose the reflectance map into reflectance parameters and an illumination map.

A connected framework trained end-to-end (i. e. from images to reflectance and illumination) would require a prohibitively large amount of GPU memory. (2) Even if (1) was possible, there is a lack of large scale databases, especially for reflectance and illumination, needed for training such a connected system. (3) Previous approaches in similar tasks [7], [8] have shown that optimizing in discrete steps (some parameters are kept fixed while estimating the rest) helps in keeping the training process stable.

For the reflectance map estimation, we propose two different approaches: The first approach (Fig. 1, Direct) directly estimates a reflectance map from the input image using an end-to-end learning framework based on a CNN with de-convolutions. The second approach (Fig. 1, Indirect) leverages additional supervision at training time, to first predict per-pixel surface normals, which are then used to compute sparse reflectance maps from the visible normals of the object. Given the sparse reflectance map, a learning-based sparse data interpolation scheme is introduced to arrive at the final reflectance map.

For the decomposition of the reflectance map into material and illumination, we investigate three different approaches: The first approach independently estimates BRDF parameters and illumination using two different CNN architectures. The second approach jointly estimates both by employing a single CNN that shares the first convolutional layers. Finally, the third approach combines the use of CNNs with classic inverse rendering techniques.

Our key contributions can be summarized as:

- We propose the first deep learning formulation to infer reflectance maps from a 2D image and to further decompose them into material parameters and natural illumination.
- We show new capabilities of CNN architectures, mapping from the image to the directional domain, performing learning-based sparse data interpolation as well as mapping from low dynamic range to high dynamic range data.
- In order to train and evaluate our two-step approach, we provide new datasets that include large scale synthetic data to facilitate the training of deep learning models as well as real data to provide a realistic testing regime.

The paper is organized as follows: Related work is presented in Sec. 2. Next, Sec. 3 introduces some basic definitions used throughout the paper. In Sec. 4 we present our CNN framework for estimating reflectance maps that we further decompose into reflectance and illumination in Sec. 5. Following this, Sec. 6 describes the new datasets used for training. Experimental results are reported in Sec. 7. Finally, Sec. 8 concludes the paper.

2 RELATED WORK

Intrinsics are the individual physical properties that yield a scene’s appearance through their interaction [1]. As an example, incoming

light reflected on a material’s surface in the direction of the observer yields an appearance influenced by the surface’s reflectance as well as the scene’s illumination. 3D shape is another intrinsic property of the objects in a scene, that also influences appearance through the surface’s orientation (normals). Ideally, one can retrieve all these pieces of the appearance jigsaw puzzle separately. In practice, even if one fixes a single component (shape, reflectance, or illumination) by assuming it to be known or by just simplifying it, what one is left with is still a hard decomposition problem for the remaining two components. Sometimes one also keeps two of the three intertwined, only retrieving the third as a separate entity.

As making assumptions about one or more of the intrinsics is important to get a handle on the decomposition problem, it is also relevant to better understand their natural statistics. Databases of reflectance [17], [18] or illumination [19], [20] samples have allowed to acquire such statistics, but exploiting them in computation remains challenging. Recent databases focus on images captured in the wild, e.g. annotated for reflectance using crowd-sourcing [3]. We built upon these recent advances and propose a new dataset that captures reflectance maps and normals for the specular case, which are not well represented in prior recordings (e. g. the intrinsic image decomposition tasks [2], [21] assume diffuse surfaces).

Next, we describe related work, that we mainly found in the three core research strands listed below. The following discussion gradually homes in on work that gets closer and closer to ours.

Reflectance Maps It is not always required to separate reflectance and illumination. *Reflectance maps* [13] - that assign an appearance (i. e. RGB color) to a surface orientation, thereby combining reflectance and illumination - suffice for many important applications. Examples are novel view synthesis (if the 3D shape is available) [22], [23] or material exchanges [24]. Such reflectance maps can be obtained in multiple ways, e. g. using Internet photo collections of diffuse objects to produce a rough 3D shape and then extracting reflectance maps in a second step [25].

In computer graphics, reflectance maps are popular to capture, transfer, and manipulate the orientation-dependent appearance of photo-realistic or artistic shading. They are also known as “lit spheres” [26] or “MatCaps” [27]. A special user interface is typically required to map surface orientation to appearance at sparse points in an image, from which orientations are interpolated for in-between pixels to fill the lit sphere (e. g. Rematas et al. [22] manually aligned a 3D model with an image to generate reflectance maps). Khan et al. [24] made small diffuse objects in a single cluttered image to appear specular or transparent, but they rely on manual interventions and mainly aim for plausible photo-realistic results. Instead, our results do not just look plausible, but stay closer to desired ground-truth even when scene parameters are changed significantly. Surface reflectance and scene illumination

are naturally separated in our case.

Factoring Images Classic *intrinsic images* factor an image into reflectance and illumination [1]. Similarly, *shape-from-shading* decomposes into reflectance and shading, eventually leading to an orientation (normal) map or even a full 3D shape.

Recently, factoring images has received renewed interest. Lombardi and Nishino [8] as well as Johnson and Adelson [28] have studied the relation of shape, reflectance, and natural illumination. A key idea in these works is, that under natural illumination, appearance and orientation are in a much more specific relation, as used in photometric stereo [29]), than for a single point light, where many similar appearances for totally different orientations can be present. They present different optimization approaches that allow for estimation of one component if at least one other component is known. In this work, we assume that the object is made of a single material (multi-material objects as in [30] have to be ruled out), and its object class and segmentation mask are known. The latter is only used to segment the object from the background. We then aim at factoring out reflectance and illumination, in a two-step approach where first we estimate the reflectance map and then we factor the produced reflectance map into material and illumination. Overall, compared to approaches such as [8] or [28], our method solves a more general problem in one axis (i. e. the number of unknown intrinsics) but is more constrained in another (i. e. the class of the object should be known a priori).

Barron and Malik [31] factor shaded images into shape, reflectance, and lighting, but only for scalar reflectance, i. e. diffuse albedo, and for limited illumination frequencies. In a very different vein, a recent approach by Richter and Roth [32] first estimates a diffuse reflectance map using approximate normals and then refines the normal map using the reflectance map as a guide. Different from our approach, they assume diffuse surfaces to be approximated using 2nd-order spherical harmonics (SH) and learn to refine the normals from the reflectance map using a regression forest. We compare the reflectance maps produced by our more general approach to reflectance maps using an SH basis, which are limited to diffuse materials only, in our experimental results.

Having estimated the reflectance map from the input image, in the second step of our pipeline we address a problem similar to Lombardi and Nishino [8]: an object with a single, unknown material on the surface (homogeneous surface reflectance) is observed under some unknown natural illumination. Hence, in their case the shape is known, and the reflectance and illumination remain to be separately retrieved. Although we address a similar problem as these previous works, our solution is fundamentally different: instead of seeking to invert the physical process under the guidance of manually designed - thus limiting - priors, our work entirely relies on data to learn the backward mapping from a reflectance map to its intrinsics. Our results indicate this inverse mapping can be learned, leading to high-quality, detailed, yet naturalistic illumination maps. The underlying network has learned cues such that the fact that windows are bright or that it is the sky that is blue and not so much the object. Moreover, our approach is the first to perform a slightly altered task, that is much closer to practice, where the image to decompose is captured using a low dynamic range (LDR) sensor, yet the resulting illumination map has high dynamic range (HDR) as required in re-synthesis tasks. Instead, previous works have typically considered either HDR input [8], which implies the capture of multiple exposures per image making the capturing process rather impractical, or produced only LDR illumination maps [11], [12].

Our approach is inspired by the ideas presented by Dror et al. [33]. In their paper, they classify materials from grayscale, HDR images of spheres under unknown, complex illumination by training SVM classifiers on image features (i. e. histogram statistics computed on the original image and its wavelet transform). Although both methods rely on machine learning techniques, we tackle a less constrained problem and solve for more unknowns. In particular, we start from a single RGB, LDR image of an object (e. g. car) with unknown shape and regress both reflectance parameters and illumination maps.

Deep Learning In recent years *Convolutional Neural Networks* (CNNs) have shown strong performance across different domains. In particular, the models for object recognition by Krizhevsky et al. [34] and detection by Girshick et al. [35] can be seen as a layer-wise encoder of successively improved features. Based on ideas of encoding-decoding strategies similar to auto-encoders, convolutional decoders have been developed [36], [37] to decode condensed representations back to images. This has led to fully convolutional or de-convolutional techniques that have seen wide applicability for tasks where there is a per-pixel prediction target. In [38], [39], this paradigm has been applied to semantic image segmentation, whereas in [40], image synthesis was proposed given object class, view, and view transformations as input and synthesizing segmented new object instances as output. Similarly, Kulkarni et al. [41] proposed the *deep convolutional inverse graphics networks* with an encoder-decoder architecture, that, given an image, can synthesize novel views. In contrast, our approach achieves a new mapping to intrinsic properties - the reflectance map, reflectance, and illumination.

Deep lambertian networks [42] apply deep belief networks to the joint estimation of a surface's reflectance, an orientation map, and the direction of a single point light source. They rely on Gaussian Restricted Boltzmann Machines to model the prior of the albedo and the surface normals for inference from a single image. In contrast, we address specular materials under general illumination, and further factor into reflectance and illumination.

Another branch of research proposes to use neural networks for depth estimation [14], [43], [44], normal estimation [15], [43], [45], intrinsic image decomposition [2], [21], and lightness [46]. Wang et al. [45] show that a careful mixture of deep architectures with hand-engineered models allows for accurate surface normal estimation. Observing that normals, depth, and segmentations are related tasks, Eigen et al. [15] propose a coarse-to-fine, multi-scale, and multi-purpose deep network that optimizes depth, normal estimation, and semantic segmentation. Likewise, Li et al. [43] apply deep regression using CNNs for depth and normal estimation, whose output is further refined by a conditional random field. Going one step further, Liu et al. [44] propose to embed both the unary and the pairwise potentials of a conditional random field in a unified deep network. In contrast to these approaches, our goal is not normals, but rather reflectance and illumination estimation - although our "indirect approach" estimates normals as a by-product.

3 DEFINITIONS

Before presenting our two-step pipeline in detail we introduce some basic definitions that will be used throughout the paper. We begin with the *reflectance map* $L(\omega) \in \mathcal{S}^+ \rightarrow \mathbb{R}^3$ [13], which is a map from orientations ω in the positive half-sphere \mathcal{S}^+ to the

RGB radiance value L leaving that surface to a distant viewer. It combines the effect of *reflectance* and *illumination*.¹

There are multiple ways to parametrize orientation ω . Horn and Sjöberg [13] used positional gradients which are suitable for an analytic derivation but less attractive for computation as they are defined on the infinite real line. We instead parametrize the orientation simply by s, t the normalized surface normal's x and y components. Dropping the z coordinate is equivalent to drawing a sphere under orthographic projection with exactly this reflectance map. Note, that orientations of surfaces in an image only cover the upper half-sphere \mathcal{S}^+ , so we only need to parametrize a half-sphere, avoiding to deal with spherical functions.

To arrive at the notion of reflectance maps, as well as the surface reflectance model that we will use, we recall the definition of the rendering equation [47] (RE) which states that for one wavelength

$$L_o(\mathbf{x}, \omega_o) = L_e(\mathbf{x}, \omega_o) + \int_{\Omega^+} f_r(\mathbf{x}, \omega_i, \omega_o) L_i(\mathbf{x}, \omega_i) \langle \omega_i, \mathbf{n}(\mathbf{x}) \rangle^+ d\omega_i, \quad (1)$$

where L_o is the outgoing radiance, L_e the emitted radiance, L_i the incoming radiance, f_r the BRDF, and $\mathbf{n}(\mathbf{x})$ the surface orientation. The radiances are both functions of position \mathbf{x} and direction ω . The reflected part is the integral over the upper hemisphere \mathcal{S}^+ of the product of incoming light L_i , BRDF f_r , and the dot product of surface normal $\mathbf{n}(\mathbf{x})$ and integration direction ω_i . In this work, it is assumed that *i)* there is no emission, *ii)* the positions of light entry and exit do not differ (translucent objects are excluded), *iii)* there is only a single material (one surface reflectance model to be considered), *iv)* the object is seen under orthographic projection from an infinitely far-away observer, *v)* that the incoming light comes from a distant scene and as such only depends on direction (environment map illumination), and *vi)* there are no shadows. These simplify the RE to the following function

$$L_o(\omega_o) = \int_{\Omega^+} f_r(\omega_i, \omega_o) L_i(\omega_i) \langle \omega_i, \mathbf{n} \rangle^+ d\omega_i, \quad (2)$$

which refers to the reflectance map L_o of the illumination L_i and the surface reflectance f_r . Henceforth, for simplification we refer to the surface reflectance model f_r as the *material*. A data-driven BRDF would be an ideal such reflectance model, but here it is simplified to the seven-parameter Phong model [48]

$$f_r(\omega_i, \omega_o) = k_d + k_s \cdot \langle r(\omega_i, \mathbf{n}), \omega_o \rangle^{k_g}, \quad (3)$$

where k_d is called the *diffuse color*, k_s the *specular color*, k_g the *glossiness*, and $r(\cdot, \cdot)$ the mirror reflection of L_i .

As both the illumination L_i and the reflectance map L_o are two-dimensional functions of direction ω , we represent them as images using the described s, t parametrization. Nevertheless, other parametrizations could also be used, such as the Lambert, latitude-longitude or the mirror-ball mappings [49].

4 STEP 1: FROM IMAGES TO REFLECTANCE MAPS

In this section, we present a solution for the first step of our pipeline, which is the estimation of the reflectance map when a single 2D image depicting a single-material object from a known class (e. g. cars) and its segmentation mask are given as input.

1. For the case of a mirror sphere, as it is here, it captures illumination [19] but is not limited to it. It also does not only capture surface reflectance [17], which would be independent of illumination, but rather joins the two.

Motivation We address a challenging inverse rendering problem that is highly under-constrained. Therefore, any solution needs to mediate between evidence from the data and prior expectations. In the general setting of specular materials and unknown natural illuminations, modeling prior expectations over reflectance maps - let alone obtaining a parametric representation - seems problematic. This motivated us to follow a data-driven approach in an end-to-end learning framework, where the dependence of reflectance maps on object appearances is learnt from a substantial number of synthesized images for a given object class.

Overview We want to estimate the reflectance map of a single-material object depicted in a single RGB image (see Fig. 1, Step 1). This is equivalent to estimating how a sphere [26] with the same material as the object would look like from the same camera position and under the same illumination. We propose two approaches to estimate reflectance maps: a *Direct* (Sec. 4.1) and an *Indirect* one (Sec. 4.2). Both have a general RGB image as input and a reflectance map as an output. The *Indirect* method also produces a conjoint per-pixel normal map. Both variants are trained from and evaluated on the new SMASHING dataset introduced in detail in Sec. 6.1. For now, we can assume the training data to consist of pairs of 2D RGB images (domain) and reflectance maps (range) with the latter in the parametrization explained in Sec. 3.

4.1 Direct approach: An end-to-end learning-based model for inferring reflectance maps

In the *Direct* approach (Fig. 1, Step 1, Direct), we learn a mapping between the object's segmented image and its reflectance map, following a convolutional-deconvolutional architecture.

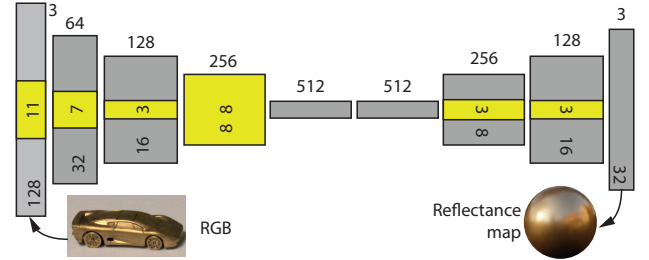


Fig. 2. Architecture of the *Direct* approach for the reflectance map estimation (see also Fig. 1, Step 1, Direct). The bottom numbers represent the spatial resolution and the ones on top the size of the feature channels for the corresponding convolutional layer. Finally, the yellow boxes in the middle indicate the filters' size.

Fig. 2 shows the proposed architecture. Starting from a series of convolutional layers, each followed by batch normalization, ReLU, and pooling layers, the size of the input feature maps is reduced to 1×1 . After continuing with two fully-connected layers, the feature maps are up-sampled until the output size is 32×32 pixels. In all convolutional layers a stride of 1 and zero padding are used such that the output has the same size as the input. The final layer uses an Euclidean loss between the RGB values for the predicted and the ground-truth reflectance map.

In short, for the *Direct* approach the network needs to learn how to “encode” the input image to a reflectance map. Note that, this is a particularly challenging task, as the model has to learn not only how to map the image pixels to locations on the reflectance map (change from image to directional domain), but also to impute and interpolate appearance for unobserved normals.

4.2 Indirect approach: Estimating reflectance maps from inferred normals using sparse data interpolation

As an alternative for the *Direct* approach described above, we also explored an *Indirect* approach that explicitly incorporates domain knowledge about the RE and the relation between the input image and corresponding reflectance map.

The *Indirect* approach (Fig. 1, Step 1, Indirect) proceeds in four steps: *1a*) estimating per-pixel orientation maps from the RGB image, *1b*) up-sampling the orientation map to the full available input image resolution, *1c*) changing from the image domain into the directional domain, producing a sparse reflectance map, and *1d*) predicting a dense reflectance map from the sparse one.

The steps *1a* and *1d* are modeled by CNN architectures, while the steps *1b* and *1c* are prescribed transformations, related to the parametrization of the reflectance map. Next, we detail each step.

(1a) Normals estimation Our goal is to predict a surface orientation (normal) map from the RGB image. To this end, we use our simplified parametrization of the directional domain to coordinates in a flat 2D image of a lit sphere (see Sec. 3). Specifically, we seek to find the s, t parameters according to our reflectance map parametrization.

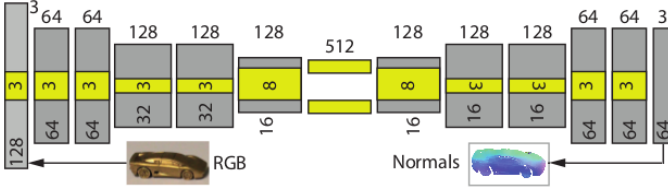


Fig. 3. Architecture of the normals estimation sub-step of our *Indirect* approach for estimating the reflectance map (see also Fig. 1, Step 1, Indirect). The notation is the same as in Fig. 2. Note that, the middle elements here correspond to fully convolutional filters.

For this task we train a CNN, whose architecture is shown in Fig. 3. Inspired by recent works in normals estimation from a single image [15], [43], [45], we opted for a deeper architecture which has proven more efficient in this task. Specifically, the network is fully convolutional as in [38] and it consists of a series of convolutional layers, each followed by ReLU and pooling layers, that reduce the spatial extent of the feature maps. After the fully convolutional layers, there is a series of de-convolutional layers that up-scale the feature representation to half of the original image’s size. Finally, we use two Euclidean losses between the predicted and the L_2 normalized ground-truth normals. The first one takes into account only the s, t coordinates of our simplified parametrization, while the second uses the original x, y, z coordinates of the normals (also explaining why the features channel in the last layer of Fig. 3 has a size of 3 instead of 2). We have experimentally found that this design helps in improving the quality of predicted normals.

(1b) Normals up-sampling In the above network the orientations are estimated at a decimated resolution of 64×64 , so the number of orientation samples is in the order of thousands. The input images however are of resolution 128×128 with ten-thousands of pixels. Note that, a full-resolution orientation map is useful for resolving all appearance details in the orientation domain. Also, the appearance of one orientation in the reflectance map can be related to all high-resolution image pixels with that orientation. As such, intended applications performing shape manipulation in the 2D image (cf. Sec. 7.3) will benefit from a refined map. To

produce this high-resolution orientation map, we use joint bilateral upsampling [50] as also done in range images [51].

(1c) Change-of-domain Next, we want to reconstruct a sparse reflectance map from the high-resolution orientation map of the previous step and the input image. This is a prescribed mapping transformation: The pairs of appearance L_p and orientation ω_p in every pixel p are unstructured samples of the continuous reflectance map function $L(\omega)$ ($= L_o(\omega_o)$) we seek to recover. Our goal now is to map these samples from the image to the directional domain, constituting the reflectance map. The most straightforward solution is to perform scattered data interpolation

$$L(\omega) = \left(\sum_{p=1}^n w(\langle \omega, \omega_p \rangle) \right)^{-1} \sum_{p=1}^n w(\langle \omega, \omega_p \rangle) L_p, \quad (4)$$

where $w(x) = \exp(-(\sigma \cos^{-1}(x))^2)$ is an RBF kernel.

In practice however, the orientation estimates are noisy and the requirements of a global reflectance map (directional illumination, orthographic projection, no shadows) are never fully met, asking for a more robust estimate. We found darkening due to shadows to be the largest issue in practice. Therefore, we instead perform a max operation over all samples closer than a threshold $\epsilon = \cos(5^\circ)$,

$$L(\omega) = \max\{w(\langle \omega, \omega_p \rangle) L_p\}, \quad w(x) = \begin{cases} 1 & \text{if } x > \epsilon \\ 0 & \text{otherwise.} \end{cases} \quad (5)$$

If one orientation is observed under different amounts of shadow, only the one that is not in shadow will contribute - which is the intended effect. Still, the map resulting from this step is sparse due to normals that were not observed in the image, as seen in the example of Fig. 4 (Sparse RM). This requires imputing and interpolating the sparse data in order to arrive at a dense estimate.

(1d) A learning-based approach for sparse data interpolation The result of the previous step is a sparse reflectance map, that is noisy due to errors from incorrectly estimated normals and has missing information at orientations that were not observed in the image. Note that, the latter is not a limitation of the normal estimation, but even occurs for ground-truth surface orientations: if an orientation is not present, its appearance remains unknown.

One solution is to directly use Eq. 4 to get a dense output. As will be shown in Sec. 7.1 though, this leads to poor performance. Instead, we propose a learning-based approach to predict a dense reflectance map from a sparse and noisy one. Accordingly, the network is trained on pairs of sparse and dense reflectance maps. The sparse ones are created using the steps *1a-1c* on synthetic data where the target reflectance map is known by rendering a sphere.

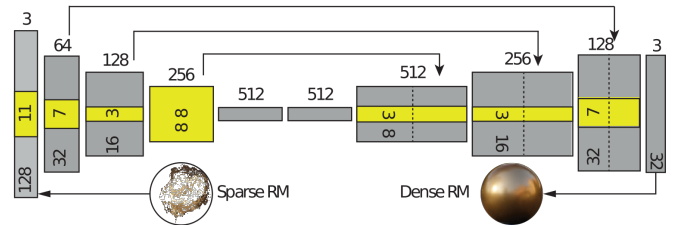


Fig. 4. Architecture of the last sub-step of our *Indirect* approach for the reflectance map estimation. The notation is the same as in Fig. 2.

The employed CNN architecture is shown in Fig. 4. Note, that it is very similar to the architecture of our *Direct* approach (see Fig. 2). Input is the sparse reflectance map and output the dense

one. Since both are in the same domain, we use the output of the convolutional layers as additional cues. Specifically, after each de-convolution layer we concatenate its output with the feature map from the respective convolution layer. This is a common practice in CNN architectures with similar tasks, as it helps preserving the local structure of the predicted image. Finally, an L_2 loss between the predicted and the ground-truth dense reflectance map is used.

5 STEP 2: FROM REFLECTANCE MAPS TO MATERIAL PARAMETERS AND NATURAL ILLUMINATION

In the previous section we presented our approach for estimating the reflectance map of an object from a single image. In what follows, we show how to further decompose the estimated reflectance map into its intrinsics: material (i. e. reflectance) and illumination.

Overview The input to our material and illumination decomposition (Fig. 1, Step 2) is the LDR reflectance map estimated from the first stage of our pipeline (Fig. 1, Step 1). In general, a reflectance map can be obtained in several other ways. For example, when a spherical sample of the desired material is available, it can directly be put under the desired illumination and photographed. In practice, this is usually not the case though - the sample has a different shape. If the shape is known, i. e. its normals are known, its reflectance map can be retrieved, at least for all observed surface orientations. In this latter case, only the last step of our *Indirect* reflectance map estimation needs to be applied. If the shape is unknown, several options have been explored to acquire it, including 3D scanning, structure-from-motion, depth sensors, CNN-based depth extraction [14], [43], [44] or directly estimating the normals using deep learning [15], [43], [45]. Although in this paper we assume that the reflectance map is given from the first stage of our pipeline (Fig. 1, Step 1), for the sake of generality it is useful keeping these other options in mind.

The outputs of Step 2 are: (1) the Phong reflectance parameters (see Eq. 3), and (2) an HDR illumination (environment) map in the parametrization of Sec. 3. The illumination map is an HDR spherical image, expressing illumination’s directional dependency. Remember that, HDR is a critical property to have for illumination [20], [49], as without it re-illumination is likely to fail in many real-world cases. Note that, our estimated illumination is still HDR even when the input is only LDR, which is a generalization over previous approaches that require HDR inputs [7], [8].

We enable this mapping by proposing *DeLight-Net*, a framework of CNNs, trained on synthetic data. All CNNs take as input a dense reflectance map. *Material CNN* outputs a parameter vector, which is 7-dimensional in the case of the Phong reflectance model: one color for the diffuse, one for the specular component, and a glossiness value, defining how shiny the material is (see Eq. 3). *Illumination CNN* outputs the HDR illumination map. These two independent CNNs comprise our INDEP. approach. We also propose two variants: JOINT shares intermediate representations to perform the estimation jointly, and SEQUEN. combines the *Illumination CNN* with classic inverse rendering techniques to estimate the Phong parameters.

5.1 Independent material and illumination estimation

Our INDEP. approach builds on *Material CNN* and *Illumination CNN* to independently estimate material parameters and natural illumination from a dense reflectance map. For both networks we used Huber loss for regression. We have experimentally found that

this choice nicely balances between learning the dynamic range and the color distribution of the illumination map.

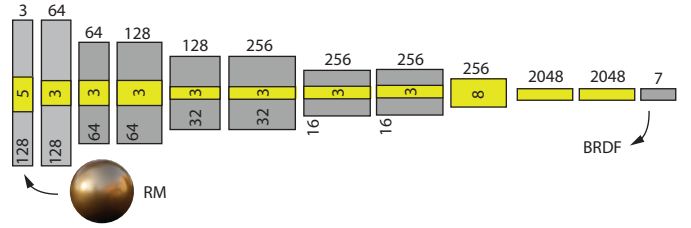


Fig. 5. The *Material CNN* for estimating Phong reflectance parameters. The notation is the same as in Fig. 2.

Material CNN As already mentioned, the input to this network is a 2D image of the dense reflectance map, while the output is a 7-parameter Phong vector. The design of the network is shown in Fig. 5. Overall, the network consists of multiple convolutional layers reducing the resolution, followed by several fully-connected layers. Note that, each convolutional unit is always followed by batch normalization and ReLU.

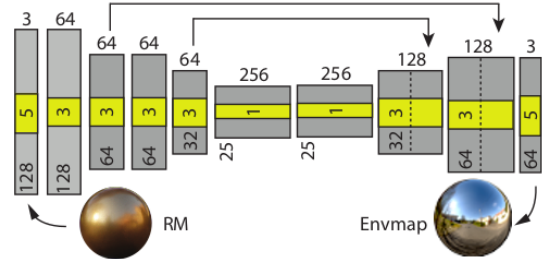


Fig. 6. The *Illumination CNN* for estimating natural illumination. The notation is the same as in Fig. 2.

Illumination CNN As already mentioned, the input to this network is the same dense reflectance map as in *Material CNN*, while its output is an HDR illumination map of half the spatial resolution (Fig. 6). The feature spatial resolution is gradually reduced by about one order of magnitude, from 128 to 25, with the middle layers applied in a fully convolutional fashion. Also, two layers of de-convolution are added, that take intermediate results from the previous same-resolution convolutional layers into account (similar design as in Fig. 4). We remind that by doing so, fine spatial details can be preserved. As before, each convolutional unit is always followed by batch normalization and ReLU.

5.2 Joint material and illumination estimation

Besides the independent estimation of material and illumination, as discussed in Sec. 5.1, we also experimented with a network estimating both somewhat more jointly. In this JOINT approach, the network shares the first two layers of *Material CNN* and *Illumination CNN*, as seen in Fig. 5 and Fig. 6 respectively, and is consequently split, preserving the individual architectures that result in two outputs with their independent losses.

5.3 Sequential material and illumination estimation

While the two approaches explained above can estimate material parameters and natural illumination separately or jointly, we also investigate an alternative that combines CNNs with classic inverse

rendering. For this SEQUEN. approach, we use the output of the *Illumination CNN* as an input to a classic inverse rendering solution for material estimation. To this end, we show how Phong reflectance parameters can be estimated from a reflectance map and known illumination in a closed form solution. Going back to our simplified reflectance map from Eq. 2, when BRDF f_r is Phong,

$$L_o(\omega_o) = \underbrace{k_d \int L_i(\omega_i) \langle \omega_i, \mathbf{n} \rangle^+ d\omega_i}_{\text{Diffuse}} + \underbrace{k_s \int L_i(\omega_i) \langle (r(\omega_i, \mathbf{n}), \omega_o) \rangle^{k_g} \langle \omega_i, \mathbf{n} \rangle^+ d\omega_i}_{\text{Specular}} \quad (6)$$

it can be written as a linear combination of a diffuse reflectance map L_d and a gloss-dependent specular reflectance map L_s :

$$L_o(\omega_o) = k_d \underbrace{L_d(\omega_o)}_{\text{Diffuse RM}} + k_s \underbrace{L_{s,k_g}(\omega_o)}_{\text{Specular RM}}.$$

Having observed many pixel samples of L_o , and having estimated L_i using *Illumination CNN*, L_d and L_{s,k_g} can be computed for all values of k_g . Furthermore, if we hold k_g fixed, estimating k_d and k_s is a linear least-squares problem: Let $\mathbf{l}_o, \mathbf{l}_d, \mathbf{l}_{s,k_g}$ be vectors of those pixels for a gloss level k_g . So $\mathbf{l}_o = k_d \mathbf{l}_d + k_s \mathbf{l}_{s,k_g}$ or $\mathbf{A}\mathbf{x} = \mathbf{b}$, where $\mathbf{A} = [\mathbf{l}_d | \mathbf{l}_{s,k_g}]$, $\mathbf{x} = (k_d, k_s)$, and $\mathbf{b} = \mathbf{l}_o$. This can efficiently be solved for \mathbf{x} for every gloss level k_g by inverting a 2×2 matrix. In order to find the optimal gloss level k_g , a line search for discrete gloss levels is performed, in our case on 100 levels, logarithmically spaced.

This procedure is only applicable because the number of non-linear parameters is low in the Phong model and would not scale to more complex material models. Still, as we show later, estimating Phong parameters analytically and illumination using *Illumination CNN* outperforms more complex material models.

6 DATASETS

To train the two steps of our pipeline a large number of images is required. Since it is very difficult to acquire many real images - at least in the order of ten-thousands - together with their ground-truth 3D shape, material (i. e. reflectance), and HDR illumination, we opted for synthetically rendered images for the training process. Unfortunately, there is also a lack of large scale databases of scanned material samples and HDR illumination maps. As such, we generated two datasets for training each step of our pipeline with emphasis given on different aspects every time.

6.1 The SMASHING challenge dataset



Fig. 7. Our dataset for the reflectance map estimation consists of synthetic images with random view, 3D shape, material, illumination, and exposure.

For the reflectance map estimation (Fig. 1, Step 1), we propose the Specular MATerials on SHapes with complex IllumiNation

(SMASHING) challenge. It includes a dataset of real as well as synthetic images, ground-truth reflectance maps, and normals (where available), results from different methods for baseline comparisons, and a set of metrics that we propose to evaluate and compare performance. The data, baselines, methods as well as performance metrics are publicly available².

Our dataset combines synthetic images, photographs, and images from the web, all depicting cars. We have manually segmented foreground and background for every image.

Synthetic images Synthetic images are produced with random *i)* views, *ii)* 3D shapes, *iii)* materials, *iv)* illumination, and *v)* exposure. A preview can be seen in Fig. 7. The view is sampled from a random position around the object, looking at the center of the object with a FOV of 40° . The 140 3D shapes come from the free 3D Warehouse repository, indexed by Shapenet [53]. For each sample the object orientation around the y axis is randomized. Illumination is provided by 40 free HDR illumination maps collected from the Internet (for more details visit the project’s webpage [52]). The exposure is sampled over the “key” parameter of Reinhard et al.’s photographic tone mapper [54], between 0.4 and 0.6. For materials, the MERL BRDF database [18] containing 100 materials is used. Overall 60k sample images from that space are generated. We define a training-test split so that no shape, material or illumination is shared between the training and test set.

Photographs As real test images, we have recorded photos of six toy cars that were completely painted with a single car lacquer, placed in four different lighting conditions and photographed from five different views, resulting in a total of 120 images. For the corresponding ground-truth reflectance maps, we placed in the same locations spheres painted with the same material. Again, those real images were manually segmented from the background.

Internet images In order to provide an even more challenging test set, we collect an additional 32 car images from the Internet. Here we do not have access to ground-truth normals or reflectance maps, but this setting provides a realistic test case for image-based editing methods. Again, we have manually segmented out the car body. This allows the qualitative evaluation of single-material normal and reflectance map prediction. Note that, for the Internet images we used networks that were trained on synthetic data from segmented meshes to contain only the car body.

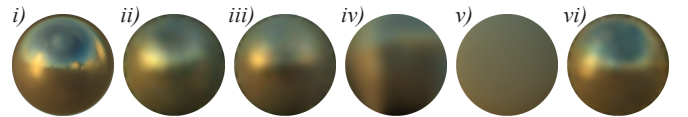


Fig. 8. Different methods to reconstruct reflectance maps.

Methods and metrics We include six different methods to reconstruct reflectance maps, visualized in Fig. 8: *i)* ground-truth, *ii)* our *Direct* approach, *iii)* our *Indirect* approach, *iv)* an approach that follows our *Indirect* one, but instead of using a CNN for sparse interpolation, it relies on an RBF reconstruction as described in Eq. 4, *v)* spherical harmonics (SH) where we project the ground-truth reflectance map to the SH domain, and *vi)* an *Indirect* approach where the estimated normals are replaced by ground-truth normals.

To assess the quality of the reflectance map estimation step we employ two different metrics. The first is the traditional L_2

² The two datasets, MATLAB code (CNN models, loss functions, etc), as well as extensive visualizations of the following results can be found on the project’s webpage [52].

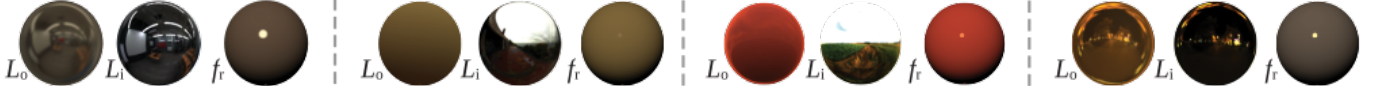


Fig. 9. Four examples of training data for decomposing reflectance maps into material parameters and natural illumination: Each triplet is a sample of the training set. From left to right: the input reflectance map L_o , the output illumination map L_i , and material f_r .

error between all defined pixels of the reflectance map in RGB and the second is the DSSIM structural difference [55] that excels in measuring the similarity between two images.

6.2 DeLight-Net dataset

For the reflectance map decomposition (Fig. 1, Step 2), our training data consist of a set of synthetically rendered images of reflectance maps with random materials under random HDR illuminations from random views. Fig. 9 shows such examples of training data.

Training materials were again taken from the MERL BRDF database [18], in particular, the Phong fit made therein. There are 100 materials overall - in our case 67 were used for training and 33 for testing - including diffuse, glossy, and mirror-like appearances.

For illumination we used 70 free HDR illumination maps in total - 60 for training and 10 for testing - from the commercial content supplier HDR Maps [56]. These images are radiometrically calibrated, i.e. they differ from the true physical RGB radiance units by only a factor. We found this not to be the case for other HDR illumination maps found on the Internet, which is crucial for re-lighting. All illumination maps were also rendered as mirror spheres and consequently re-sampled to 128×128 pixels, which is the resolution of the maps we will later infer.

View positions are sampled from a random direction in the xz -plane with a random declination of $\pm 10^\circ$; an orthographic projection is used. The shape is always a highly-tessellated sphere with analytic normals. For rendering we use the full convolution of the illumination map with the Phong parametric model (see Eq. 2). This convolution is computationally demanding and to keep it tractable when producing massive training data, it was implemented using GPUs. The rendering result is a 128×128 image. Overall, we produced approximately 50k sample images of synthetically rendered reflectance maps. Note that for the testing set both the material as well as the illumination map are never seen before. The training and benchmark data as well as the CNN architectures used are publicly available [52].

Two variants of the resulting images are kept, with slightly different purposes: an HDR and an LDR variant. For the HDR variant we apply the natural logarithm to the RGB data, stored as a 32-bit float image file, as also done in [18] to avoid bias towards differences in the higher intensity ranges during training. For the LDR variant we simulate the exposure process, as follows: First we automatically choose an exposure level using the (5,95)-percentiles. Second, linear radiance values are mapped into the (0,1)-range and quantized uniformly into 256 values (8-bit). Finally, the values are mapped back to absolute radiance and stored in a 32-bit float format. This procedure simulates the information available to a contemporary capturing device with EXIF information (aperture, exposure time, ISO): radiance quantized to 8-bit in an appropriately chosen exposure, allowing to re-scale it to absolute radiance, but with quantization and clipping.

7 EXPERIMENTS

In this section, we perform the experimental analysis of the proposed pipeline. Since we rely on a two-step approach, we find it fit to first evaluate each step individually and compare with their corresponding baselines before assessing the system as a whole. As such, in Sec. 7.1 we first evaluate the proposed end-to-end *Direct* approach for the reflectance map estimation (Fig. 1, Step 1) on the new SMASHING challenge and compare it to the *Indirect* approach in its different variants (see Sec. 6.1). Second, in Sec. 7.2 we perform extensive evaluations for the results produced by the reflectance map decomposition framework (Fig. 1, Step 2) and compare it with state-of-the-art approaches [7], [8]. Finally, in Sec. 7.3 we analyze the qualitative performance of our combined pipeline through various applications including a wide range of image-based editing tasks.

7.1 Evaluation of reflectance map estimation

Setup Here, we provide results for our *Direct* method that learns to predict reflectance maps directly from the input image in an end-to-end scheme, as well as several variants of our *Indirect* approach that utilize intermediate results facilitated by additional supervision through normals at training time (cf. Fig. 1, Step 1). The variants of the *Indirect* scheme are based on our estimated normals, but differ in their second stage that has to perform a type of data interpolation to arrive at a dense reflectance map, given the intermediate sparse estimate. For the interpolation, we investigate the proposed learning-based approach *Indirect (CNN)* as well as using Radial Basis Function interpolation *Indirect (RBF)*. Furthermore, we provide best case analysis by using ground-truth normals in the *Indirect* approach *Indirect (GT Normals)* (only possible for synthetic data) and computing a diffuse version of the ground-truth by means of spherical harmonics *SH (GT Normals)*. The latter gives an upper bound on the result that could be achieved by methods relying on a diffuse material assumption. Quantitative results for the different approaches are summarized in Table 1.

TABLE 1
Quantitative results for the reflectance map estimation (cf. Fig. 1, Step 1) using the different methods defined in Sec. 6.1.

Method	Synthetic		Real	
	MSE	DSSIM	MSE	DSSIM
<i>Direct</i>	.0019	.0209	.0120	.0976
<i>Indirect (RBF)</i>	.0038	.0250	.0116	.0814
<i>Indirect (CNN)</i>	.0018	.0180	.0143	.0991
SH (GT Normals)	.0044	.0301	.0114	.0914
Indirect (GT Normals)	.0008	.0111	—	—

Reflectance map analysis Overall, we observe consistency among the two investigated metrics, MSE and DSSIM (as defined in Sec. 6.1), in how they rank approaches. We obtain accurate estimations for the synthetic set of the SMASHING challenge dataset for our *Direct* as well as the best *Indirect* methods. The

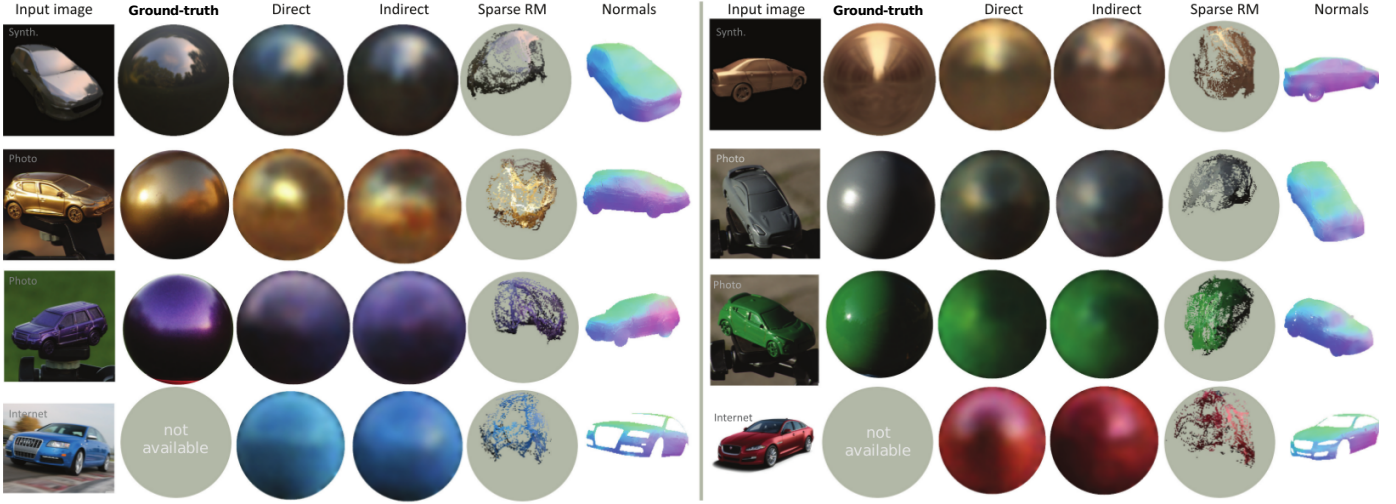


Fig. 10. Results of different variants and steps for our reflectance map estimation (Fig. 1, Step 1). From left to right: input image, ground-truth reflectance map (RM), RM result of the *Direct* approach, RM result of the *Indirect* approach, the intermediate sparse RM produced in the *Indirect* variant, normals produced by the *Indirect* variant as well. Each result is annotated to come from the synthetic, photographed or Internet part of our database. For the Internet-based part, no ground-truth RM is available. Please visit the project’s webpage [52] for exhaustive results in this format.

quantitative findings are underpinned by the visual results, e.g. showing the predicted reflectance maps in Fig. 10. The performance on the real images is generally lower with the error roughly increasing by one order of magnitude. Yet, the reconstruction still preserves rich specular structures and gives a truthful reconstruction of the represented material.

In more detail, we observe that the best *Direct* and *Indirect* approach perform similar on the synthetic data, although *Direct* did not use the normal information during training. For the real examples, this form of additional supervision seems to pay off more and the RBF interpolation scheme achieves best results in the considered metrics. A closer inspection to the results though, clearly shows the limitations of the image-based metrics. While the RBF-based technique yields a low error, it frequently fails to generate well-localized highlight features on the reflectance map (see also an indicative illustration in Fig. 8). We encourage the reader to visit the project’s webpage [52], where a detailed visual comparison for all methods is provided.

The ground-truth baselines give further insights into improvements over prior diffuse material assumptions and the future potential of the method. The *SH (GT Normals)* baseline shows that our best methods improve over a best case diffuse estimate with a large margin for the DSSIM metric - highlighting the importance of considering more general reflectance maps. The *Indirect (GT Normals)* illustrates a best case analysis of the *Indirect* approach where we provide ground-truth normals. The results show a potential performance leap by having better estimated normals.

Normals analysis Table 2 quantifies the error in the normals estimation by the first stage of our *Indirect* approach. This experiment is facilitated by the synthetic data where normals are available. L_2 corresponds to a network using the Euclidean loss on the x, y, z components of the normals, while *Dual* uses the two losses described in Sec. 4.2. *Up* refers to a network trained on up-sampled normals. Both dual loss and joint up-sampling improve the estimated normals. Despite the fact that this analysis is conducted on synthetic data, we observe that our models predict very convincing normals even in the most challenging scenario that we consider (see Fig. 10 and Fig. 12).

TABLE 2
Normals estimation of *Indirect* approach on synthetic data.

	Mean	Median	RMSE
L_2	14.3	9.1	20.6
<i>Dual</i>	13.4	8.2	19.8
<i>Dual & Up</i>	13.3	8.2	19.9

7.2 Evaluation of material and illumination estimation

Here, we evaluate our approach for decomposing the reflectance map (cf. Fig. 1, Step 2), by first using it in a synthetic re-synthesis benchmark, where images are re-synthesized for original or novel illuminations and materials starting from the estimated components, and second, on real photographs of reflectance maps.

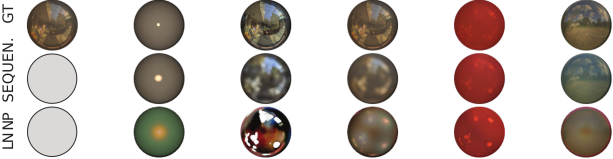
Synthetic re-synthesis benchmark Evaluating a successful decomposition is not trivial due to the complex interaction of material, illumination, shape, and viewpoint. The established evaluation protocol [7], [8], [18] is to measure the L_2 error between renderings using the estimated and ground-truth components respectively. Indeed, the reflectance map decomposition into material parameters and natural illumination allows direct evaluation of a) the estimated material parameters by rendering them under a point light source (*Point light*), b) the estimated natural illumination by rendering on a mirror sphere (*Mirror Mat.*), c) both estimated material parameters and natural illumination by re-rendering them together as a reflectance map (*Re-synthesis*).

We enhance this protocol, by also including two extensions inspired by real-world applications: d) we evaluate how well the estimated material parameters perform under different illumination (*Nat. Illum.*). To do so, we compute the reflectance map of the estimated material illuminated by a new illumination map, not included in the training set. And finally e) we measure how well the estimated natural illumination generalizes to new materials (*MERL Mat.*). This is performed by selecting a random MERL material, not contained in our training data, and rendering it under the estimated illumination.

The different approaches, represented as rows in Table 3, are:

TABLE 3

Synthetic evaluation for our material and illumination estimation (Fig. 1, Step 2). Rows represent the different approaches and columns the different tasks. Results are reported for two error metrics, LRMSE and DSSIM (lower is better). The images on top are samples from selected rows and columns. The best method for a task is shown in bold.



	Point light		Mirror Mat.		Re-synthesis		MERL Mat.		Nat. Illum.	
	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM
<i>HDR input</i>										
INDEP.	.0055	.0677	.0603	.1821	.0118	.0685	.0232	.0341	.0006	.0466
JOINT	.0082	.0753	.0590	.1782	.0117	.0770	.0200	.0339	.0006	.0529
SEQUEN.	.0062	.0326	.0603	.1821	.0016	.0175	.0232	.0341	.0008	.0209
LN DP [7]	.0245	.1450	.2537	.3299	.0002	.1485	.0288	.0854	.0019	.1423
LN NP [7]	.0263	.1664	.2862	.3124	.0001	.0243	.0292	.0433	.0018	.0605
<i>LDR input</i>										
INDEP.	.0082	.0691	.0626	.1901	.0011	.0624	.0270	.0354	.0006	.0472

“INDEP.” is our approach with independent CNNs for estimating the material parameters and natural illumination (see Sec. 5.1). “JOINT” is our joint material and illumination estimation (see Sec. 5.2). “SEQUEN.” refers to sequentially estimating natural illumination and material parameters (see Sec. 5.3). “LN” refers to the method of Lombardi and Nishino [7]. A comparison with their work is made, both when using the default values for their priors (“LN DP”) and when using no priors (“LN NP”), which might depend on the types of materials and illuminations used [7].

All the above are evaluated on the DeLight-Net dataset (cf. Sec. 6.2). Since the method of Lombardi and Nishino [7] require HDR inputs, we perform the comparison on the HDR variant of the dataset (upper part of Table 3). We also compare to our INDEP. approach trained on the LDR variant (last row of Table 3), which better relates to the LDR reflectance map estimated in the first step of our pipeline. To the best of our knowledge, our method is the first to learn this LDR to HDR mapping.

The final quantitative measure is the difference between the re-synthesized image produced using the estimated decomposition and the re-synthesized image produced using the ground-truth decomposition. The root mean square-error of the logarithm of HDR radiance (LRMSE) and a colored, multi-resolution structured similarity index [55] ran on the tone-mapped LDR result (DSSIM) are used to compare the re-synthesized image to the ground-truth.

Overall, we find that our methods outperform the method of Lombardi and Nishino [7], according to all metrics, with one exception, which is discussed below. When using our estimated material parameters and re-rendering with a point light (*Point light*), our CNNs outperform competitors by a large margin in LRMSE (three-fold improvement) and our SEQUEN. approach by a similar factor according to DSSIM. Using the estimated natural illumination and re-rendering on a mirror sphere (*Mirror Mat.*), is best done using our JOINT approach, again outperforming competitors by a substantial factor according to both metrics. According to LRMSE, for the task of re-synthesizing the input image with both the estimated material parameters and natural illumination (*Re-synthesis*), the approach of Lombardi and Nishino [7] comes out best. This is to be expected, as their approach specifically seeks

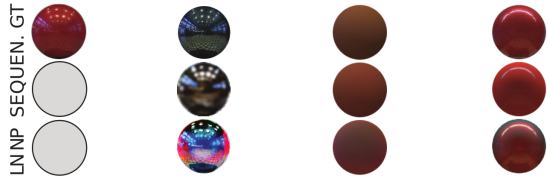
to minimize in LRMSE the pair of material and illumination that if re-synthesized give the original input. According to DSSIM however, which likely is a better measure, our SEQUEN. approach works best also for this case. When using the estimated components and re-rendering with a new material (*MERL Mat.*) or illumination (*Nat. Illum.*) from the corpus our JOINT approach performs best for both metrics with one exception. According to DSSIM, for the *Nat. Illum.* task our SEQUEN. approach comes out first. Again, the difference to competitors is the strongest in terms of the *Nat. Illum.* task, where a three-fold improvement is achieved, while for the *MERL Mat.* task the difference is almost twice as much. Remarkably, the decomposition performance from LDR inputs is on par with the HDR case, although the problem is more difficult.

In general, our SEQUEN. approach excels in estimating the material parameters whereas our JOINT approach comes marginally first when estimating the natural illumination. We found the latter marginal improvement to be less important in practice, so our SEQUEN. approach is generally the preferred choice.

Real reflectance maps The synthetic re-synthesis benchmark has been evaluated on the basis of a large choice of variations on a large number of reflectance maps, illuminations, and materials. Capturing a similar amount of reflectance maps ourselves is in practice not possible, so we opted for a smaller set of pairs of materials and illumination maps where the ground-truth illumination was also acquired. In particular we use a set of 25 materials under 4 different natural illuminations that we have acquired specifically for this task (see also the project’s webpage [52]).

TABLE 4

Evaluation on real reflectance maps for our material and illumination estimation (Fig. 1, Step 2). The notation is the same as in Table 3



	Mirror Mat.		MERL Mat.		Nat. Illum.	
	LRMSE	DSSIM	LRMSE	DSSIM	LRMSE	DSSIM
<i>HDR input</i>						
INDEP.	0.929	0.376	0.099	0.062	1.111	0.183
JOINT	0.933	0.365	0.052	0.043	1.110	0.186
SEQUEN.	0.929	0.376	0.099	0.062	1.223	0.106
LN NP [7]	5.402	0.662	1.722	0.071	3.938	0.187
<i>LDR input</i>						
INDEP.	0.950	0.376	0.092	0.059	1.155	0.214

The results are summarized in Table 4. The tasks are similar to the ones in our synthetic re-synthesis benchmark, but in a more restricted way, as we do not have the ground-truth material available; such a task would require a gonioreflectometer. As the ground-truth HDR illumination is available however (i.e. we scanned it using a chrome sphere), we can compute the difference between the ground-truth illumination and the estimated illumination rendered in a mirror (*Mirror Mat.*). Furthermore, we can re-synthesize, using not just a mirror, but instead a new material from a database, here again MERL (*MERL Mat.*). Finally, we can predict how the estimated material would look under a different illumination, as the same reflectance maps were captured under this different illumination too (*Nat. Illum.*). Note that, without ground-truth for the material, re-rendering under point light illumination (*Point light*) and re-synthesis using the estimated components (*Re-*

synthesis) are not possible. For brevity, the LN NP [7] is compared to our approaches: INDEP., JOINT. and SEQUEN..

We find that the results are consistent with the synthetic evaluation, and our SEQUEN. approach outperforms LN NP [7] for both LRMSE and DSSIM metrics.

7.3 Qualitative results and applications

Automatically extracting reflectance maps from images - together with the normal information we get as a by-product - and decomposing them into material and illumination facilitates a range of image-based editing applications, such as material acquisition and transfer and shape manipulation. In what follows, we evaluate the performance of our two-step pipeline through numerous example applications. The project's webpage [52] contains all the images and videos that complement our following presentation.

Estimating reflectance maps and normals from images

Results of estimated reflectance maps are presented in Fig. 10, also showing the quality of the predicted normals. The first row shows two examples on synthetic images, the second and third row on real images and the last row on web images (no reference reflectance map is available here). Notice how the overall appearance, reflecting the interplay between material and the complex illumination, is captured by our estimates. In most examples, highlights are reproduced and even a schematic structure of the environment can be seen in the case of very specular materials.

Inserting virtual objects in a scene Fig. 11 shows synthesized images (column 2-5) that we have rendered from 3D models using the reflectance map automatically acquired from the images in column 1. Here, we use ambient occlusion [57] to produce virtual shadows. This application shows how material representations can be acquired from real objects and transferred to a virtual object. Notice, how the virtual objects match in material, specularity, and illumination to the source image on the left.

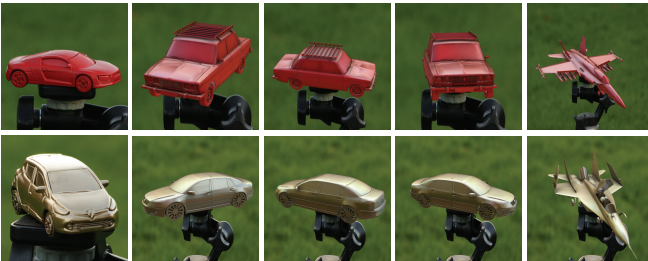


Fig. 11. Transfer of reflectance maps estimated from real photographs (1st column) to virtual objects (other columns) of different shape. The project's webpage [52] shows animations of those figures.

Transferring appearance between images A useful application of our approach is the appearance transfer between different objects in different scenes. To do so, we first estimate reflectance maps for each object independently, swap the estimated reflectance maps, and then use the estimated normals to re-render the objects using a normal look-up table from the new reflectance map. To preserve details, such as shadows and textures, we first re-synthesize each object with its original reflectance map, save the per-pixel difference in LAB color space, then re-synthesize with the swapped reflectance map and add the saved difference in LAB color space. An example is shown in Fig. 12. Despite the uncontrolled illumination conditions, we achieve photo-realistic transfer of the appearance, making it hard to distinguish source from target.



Fig. 12. Appearance transfer application. Images on the diagonal are the original input. Off-diagonal images have the appearance of the input in their column transferred to the input shape of their row.

Manipulating shape Since we estimate reflectance maps and surface normals, this enables various manipulation applications that work in the directional or normal domain. Fig. 13 shows such an application, where the surface orientation is changed, e. g. using a special painting interface, and new appearance for the new orientations is sampled from the reflectance map. As before, we save and restore the delta between the original and re-synthesized reflectance map to keep details and shadows. The final result gives a strong sense of 3D structure while maintaining an overall consistent appearance w.r.t. material and scene illumination.



Fig. 13. Shape manipulation application. A user has drawn to manipulate the normal map extracted from our *Indirect* approach. The reflectance map and the new normal map can be used to simulate the new shape's appearance. For a live demo visit the project's webpage [52].

Estimating material and illumination from reflectance maps Starting from a reflectance map we can decompose it into its intrinsic material and illumination. The estimated material parameters and natural illumination can then be used to re-render the object of interest in different scenes, change its material or even replace the object itself with another. Some of the many editing possibilities that our method enables are summarized in Fig. 15. Our pipeline allows re-rendering objects with different materials (*horizontal variation*, Fig. 15), under different illuminations (*intra-block vertical variation*, Fig. 15), or for different shapes (*inter-block vertical variation*, Fig. 15). Results are visualized in pairs, where the left half shows re-synthesis using our estimated decomposition and the right half the same re-synthesis using reference material and illumination. The input reflectance maps are marked with a



Fig. 14. Illumination (*top row*) and material (*bottom row*) change (*3rd-5th columns*), originally estimated from real photos (*1st column*). For more details see *Manipulating material and illumination from real photos* in Sec. 7.3. The project’s webpage [52] contains a video with more such examples.

dotted circle. We clearly see that our approach can reconstruct plausible materials and illumination maps with fine details.

Manipulating material and illumination from real photos

Perhaps the most interesting and practical application is interactive material and illumination manipulation from real photos. We begin from a segmented image of the object of interest, which is the car’s body in our case. Using the CNNs of Sec. 4 we first estimate the normal orientations and consequently the reflectance map (*Indirect* approach). From the estimated reflectance map we then decompose into material and illumination using the CNNs of Sec. 5 (*SEQUEN.* approach). Finally, we re-render (Fig. 14) the imaged object (*1st column*) under different illumination (*1st row*) and different material (*2nd row*). The results for two car models are shown in Fig. 14. For more car models you can visit the project’s webpage [52]. Note that we have explicitly modeled only the car’s body and not the lights, mirrors, windows, etc (same as in Fig. 12). The recovered results look nevertheless realistic and convincing.

8 CONCLUSION

We have presented a deep learning approach to estimate natural illumination information and surface reflectance characteristics from a single 2D image that facilitates new image-based rendering applications. We showed that our technique works with complex 3D shapes, specular materials, and under complex natural illumination. In order to achieve our goal, we have developed new deep learning architectures that for the first time achieve sparse data interpolation, mapping from the image to the directional domain, and inferring high dynamic range data from low dynamic range input. The application of deep learning techniques to this domain is facilitated by our novel large scale synthetically rendered dataset that is accompanied by real-world testing data in order to evaluate our approach. Our proposed methods outperform prior work in this area, which highlights the potential of deep learning approaches in inverse rendering tasks and computer graphics in general. In the future, we would like to include additional effects, such as indirect illumination and shadows, in our reflectance model and experiment with more complicated parametric models for reflectance, that will both bring this work even closer to real-world applications.

ACKNOWLEDGMENTS

This work was supported by Toyota Research Institute and FWO project “Structure from Semantics” (#G086617N).

REFERENCES

- [1] H. Barrow and J. Tenenbaum, “Computer vision systems,” *Computer Vision Systems*, vol. 2, 1978.
- [2] T. Zhou, P. Krahenbuhl, and A. A. Efros, “Learning data-driven reflectance priors for intrinsic image decomposition,” in *ICCV*, 2015, pp. 3469–3477.
- [3] S. Bell, K. Bala, and N. Snavely, “Intrinsic images in the wild,” *TOG*, vol. 33, no. 4, p. 159, 2014.
- [4] B. K. Horn and M. J. Brooks, *Shape from shading*. MIT press, 1989.
- [5] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, “Shape-from-shading: a survey,” *TPAMI*, vol. 21, no. 8, pp. 690–706, 1999.
- [6] S. Georgoulis, M. Proesmans, and L. Van Gool, “Tackling shapes and brdfs head-on,” in *3DV*, vol. 1. IEEE, 2014, pp. 267–274.
- [7] S. Lombardi and K. Nishino, “Reflectance and natural illumination from a single image,” *ECCV*, pp. 582–595, 2012.
- [8] —, “Reflectance and illumination recovery in the wild,” *TPAMI*, vol. 38, no. 1, pp. 129–141, 2016.
- [9] S. Georgoulis, V. Vanweddigen, M. Proesmans, and L. Van Gool, “A gaussian process latent variable model for brdf inference,” in *ICCV*, 2015, pp. 3559–3567.
- [10] D. Kersten, P. Mamassian, and A. Yuille, “Object perception as bayesian inference,” *Psychol.*, vol. 55, pp. 271–304, 2004.
- [11] F. Romeiro, Y. Vasilyev, and T. Zickler, “Passive reflectometry,” *ECCV*, pp. 859–872, 2008.
- [12] F. Romeiro and T. Zickler, “Blind reflectometry,” in *ECCV*. Springer, 2010, pp. 45–58.
- [13] B. K. Horn and R. W. Sjöberg, “Calculating the reflectance map,” *Applied optics*, vol. 18, no. 11, pp. 1770–1779, 1979.
- [14] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” in *NIPS*, 2014, pp. 2366–2374.
- [15] D. Eigen and R. Fergus, “Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture,” in *ICCV*, 2015, pp. 2650–2658.
- [16] F. E. Nicodemus, “Directional reflectance and emissivity of an opaque surface,” *Applied optics*, vol. 4, no. 7, pp. 767–775, 1965.
- [17] K. J. Dana, B. Van Ginneken, S. K. Nayar, and J. J. Koenderink, “Reflectance and texture of real-world surfaces,” *TOG*, vol. 18, no. 1, pp. 1–34, 1999.
- [18] W. Matusik, “A data-driven reflectance model,” Ph.D. dissertation, MIT, 2003.
- [19] P. Debevec, “Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography,” in *SIGGRAPH classes*. ACM, 2008, p. 32.
- [20] R. O. Dror, T. K. Leung, E. H. Adelson, and A. S. Willsky, “Statistics of real-world illumination,” in *CVPR*, vol. 2. IEEE, 2001, pp. II–II.
- [21] T. Narihira, M. Maire, and S. X. Yu, “Direct intrinsics: Learning albedo-shading decomposition by convolutional regression,” in *ICCV*, 2015, pp. 2992–2992.
- [22] K. Rematas, T. Ritschel, M. Fritz, and T. Tuytelaars, “Image-based synthesis and re-synthesis of viewpoints guided by 3d models,” in *CVPR*, 2014, pp. 3898–3905.
- [23] K. Rematas, C. H. Nguyen, T. Ritschel, M. Fritz, and T. Tuytelaars, “Novel views of objects from a single image,” *TPAMI*, vol. 39, no. 8, pp. 1576–1590, 2017.
- [24] E. A. Khan, E. Reinhard, R. W. Fleming, and H. H. Bühlhoff, “Image-based material editing,” *TOG*, vol. 25, no. 3, pp. 654–663, 2006.

- [25] T. Haber, C. Fuchs, P. Bekaer, H.-P. Seidel, M. Goesele, and H. P. Lensch, "Relighting objects from image collections," in *CVPR*. IEEE, 2009, pp. 627–634.
- [26] P.-P. J. Sloan, W. Martin, A. Gooch, and B. Gooch, "The lit sphere: A model for capturing npr shading from art," in *Graphics interface*, vol. 2001, 2001, pp. 143–150.
- [27] S. Spencer, *ZBrush Character Creation: Advanced Digital Sculpting*. John Wiley & Sons, 2011.
- [28] M. K. Johnson and E. H. Adelson, "Shape estimation in natural illumination," in *CVPR*. IEEE, 2011, pp. 2553–2560.
- [29] A. Hertzmann and S. M. Seitz, "Example-based photometric stereo: Shape reconstruction with general, varying brdfs," *TPAMI*, vol. 27, no. 8, pp. 1254–1264, 2005.
- [30] S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, T. Tuytelaars, and L. Van Gool, "Natural illumination from multiple materials using deep learning," *arXiv preprint arXiv:1611.09325*, 2016.
- [31] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *TPAMI*, vol. 37, no. 8, pp. 1670–1687, 2015.
- [32] S. R. Richter and S. Roth, "Discriminative shape from shading in uncalibrated illumination," in *CVPR*, 2015, pp. 1128–1136.
- [33] R. O. Dror, E. H. Adelson, and A. S. Willsky, "Estimating surface reflectance properties from images under unknown illumination," in *Human Vision and Electronic Imaging*, 2001, pp. 231–242.
- [34] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.
- [35] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.
- [36] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, "Deconvolutional networks," in *CVPR*. IEEE, 2010, pp. 2528–2535.
- [37] H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng, "Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations," in *ICML*. ACM, 2009, pp. 609–616.
- [38] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015, pp. 3431–3440.
- [39] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015, pp. 447–456.
- [40] A. Dosovitskiy, J. Tobias Springenberg, and T. Brox, "Learning to generate chairs with convolutional neural networks," in *CVPR*, 2015, pp. 1538–1546.
- [41] T. D. Kulkarni, W. F. Whitney, P. Kohli, and J. Tenenbaum, "Deep convolutional inverse graphics network," in *NIPS*, 2015, pp. 2539–2547.
- [42] Y. Tang, R. Salakhutdinov, and G. Hinton, "Deep lambertian networks," *arXiv preprint arXiv:1206.6445*, 2012.
- [43] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs," in *CVPR*, 2015, pp. 1119–1127.
- [44] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *CVPR*, 2015, pp. 5162–5170.
- [45] X. Wang, D. Fouhey, and A. Gupta, "Designing deep networks for surface normal estimation," in *CVPR*, 2015, pp. 539–547.
- [46] T. Narihira, M. Maire, and S. X. Yu, "Learning lightness from human judgement on relative reflectance," in *CVPR*, 2015, pp. 2965–2973.
- [47] J. T. Kajiya, "The rendering equation," in *SIGGRAPH Computer Graphics*, vol. 20, no. 4. ACM, 1986, pp. 143–150.
- [48] B. T. Phong, "Illumination for computer generated pictures," *Communications of the ACM*, vol. 18, no. 6, pp. 311–317, 1975.
- [49] P. E. Debevec, C. J. Taylor, and J. Malik, "Modeling and rendering architecture from photographs: A hybrid geometry and image-based approach," in *Computer graphics and interactive techniques*. ACM, 1996, pp. 11–20.
- [50] J. Kopf, M. F. Cohen, D. Lischinski, and M. Uyttendaele, "Joint bilateral upsampling," in *TOG*, vol. 26, no. 3. ACM, 2007, p. 96.
- [51] D. Chan, H. Buisman, C. Theobalt, and S. Thrun, "A noise-aware filter for real-time depth upsampling," in *M2SFA2*, 2008.
- [52] "Project's webpage," http://homes.esat.kuleuven.be/~sgeorgou/DRM_DeLight/.
- [53] Y. Li, H. Su, C. R. Qi, N. Fish, D. Cohen-Or, and L. J. Guibas, "Joint embeddings of shapes and images via cnn image purification," *TOG*, vol. 34, no. 6, pp. 234–1, 2015.
- [54] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda, "Photographic tone reproduction for digital images," *TOG*, vol. 21, no. 3, pp. 267–276, 2002.
- [55] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Signals, Systems and Computers*, vol. 2. IEEE, 2003, pp. 1398–1402.
- [56] "Hdrmaps," <https://www.hdrmaps.com/>.
- [57] S. Zhukov, A. Iones, and G. Kronin, "An ambient light illumination model," in *Rendering Techniques*. Springer, 1998, pp. 45–55.



Stamatios Georgoulis is a PhD student at the Department of Electrical Engineering of KU Leuven. He received the Diploma degree in Electrical and Computer Engineering from the Aristotle University of Thessaloniki, Greece, in 2011, where he also worked as a student research assistant and received the National Award at Microsoft's Imagine Cup. His research focuses on extracting surface and lighting characteristics from images.



Konstantinos Rematas is a post-doctoral researcher at the University of Washington. He received his BSc in Computer Science from Aristotle University, Greece, in 2008. He continued his MSc studies in Media Informatics at RWTH, Germany. In March 2016 he received his PhD from KU Leuven, Belgium. His research focuses on computer vision methods for virtual reality.



Tobias Ritschel is a Senior Lecture at the University College London. After receiving a PhD from Saarland University in 2009, he was a post-doctoral researcher at Télécom ParisTech and a Senior Researcher at MPI Informatik, Saarbrücken. His interests include interactive and non-photorealistic rendering, human perception, and data-driven graphics. He received the Eurographics Young Researcher (2014) and PhD dissertation (2011) award.



Efstratios Gavves is an Assistant Professor with the QUVA Lab at University of Amsterdam in the Netherlands. He received his Ph.D. in 2014 at the University of Amsterdam. He was a post-doctoral researcher at the KU Leuven from 2014 - 2015. His research interests include statistical and deep learning with applications on computer vision tasks, like object recognition, image captioning, action recognition, tracking, memory networks, and recurrent networks.



Mario Fritz is senior researcher at the Max Planck Institute for Informatics and Junior Faculty at the Saarland University. He is heading a group on Scalable Learning and Perception. His research interest are centred around computer vision and machine learning but extend to natural language processing and robotics. From 2008 to 2011, he did his postdoc at the International Computer Science Institute as well as UC Berkeley. He received his PhD in 2008 from TU Darmstadt.



Luc Van Gool is full professor at KU Leuven and ETH Zurich. He has authored over 200 papers in this field. He was the Program Chair of ICCV 2005 and the General Chair of ICCV 2011 and ECCV 2014. His main interests include 3D reconstruction and modeling, tracking and gesture analysis, and object recognition. He has received several best paper awards. He is the Co-Founder of the Eyetracks, GeoAutomation, Kooaba, eSaturnus, and Procedural companies.



Tinne Tuytelaars is a professor at the Electrotechnical department of KU Leuven. Her research focuses on image understanding, including object, scene, and action recognition. In 2009, Tinne Tuytelaars received an ERC starting independent researcher grant. She has been one of the Program Chairs of the European Conference on Computer Vision 2014 and one of the General Chairs of IEEE Conference on Computer Vision and Pattern Recognition 2016.

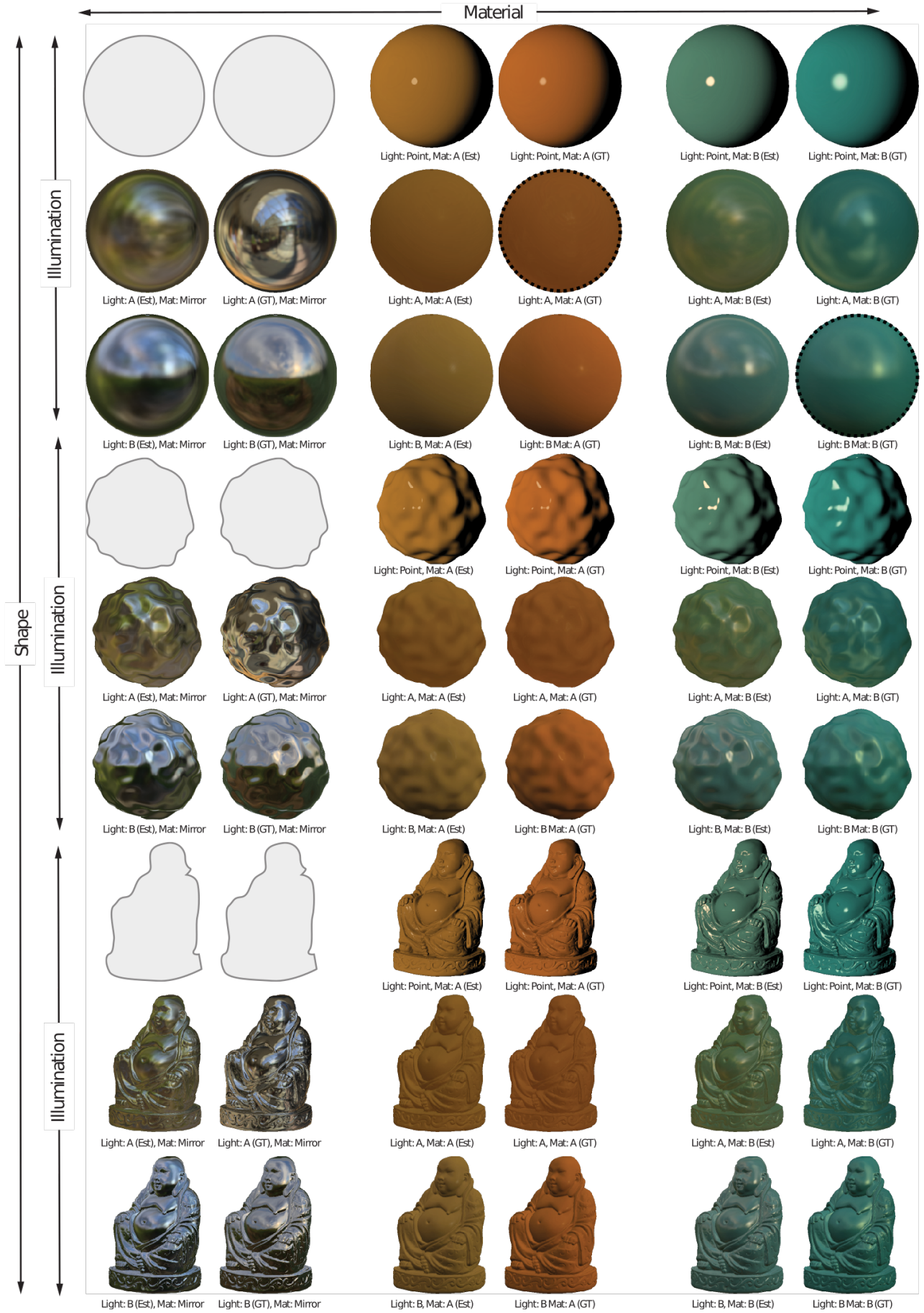


Fig. 15. Some of the many manipulations enabled by our approach. Please see the text in Sec. 7.3 for a more thorough explanation.