

Joint Material and Illumination Estimation from Photo Sets in the Wild

Tuanfeng Y. Wang Tobias Ritschel Niloy J. Mitra

University College London, UK



Figure 1: We factor a set of images (left) showing objects with different materials (red, yellow, black, white plastic) under different illumination into per-image illumination and per-object material (top right) that allows for novel-x applications such as changing view, illumination, material, or mixed illumination/material (red chair in the left-bottom imaged environment) (bottom right).

Abstract

Faithful manipulation of shape, material, and illumination in 2D Internet images would greatly benefit from a reliable factorization of appearance into material (i.e., diffuse and specular) and illumination (i.e., environment maps). On the one hand, current methods that produce very high fidelity results, typically require controlled settings, expensive devices, or significant manual effort. To the other hand, methods that are automatic and work on ‘in the wild’ Internet images, often extract only low-frequency lighting or diffuse materials. In this work, we propose to make use of a set of photographs in order to jointly estimate the non-diffuse materials and sharp lighting in an uncontrolled setting. Our key observation is that seeing multiple instances of the same material under different illumination (i.e., environment), and different materials under the same illumination provide valuable constraints that can be exploited to yield a high-quality solution (i.e., specular materials and environment illumination) for all the observed materials and environments. Similar constraints also arise when observing multiple materials in a single environment, or a single material across multiple environments. Technically, we enable this by a novel scalable formulation using parametric mixture models that allows for simultaneous estimation of all materials and illumination directly from a set of (uncontrolled) Internet images. The core of this approach is an optimization procedure that uses two neural networks that are trained on synthetic images to predict good gradients in parametric space given observation of reflected light. We evaluate our method on a range of synthetic and real examples to generate high-quality estimates, qualitatively compare our results against state-of-the-art alternatives via a user study, and demonstrate photo-consistent image manipulation that is otherwise very challenging to achieve.

1. Introduction

Estimating realistic material (i.e., reflectance) and illumination along with object geometry remains a holy grail of shape analysis. While significant advances have been made in the recent years in predicting

object geometry and pose from ‘in the wild’ Internet images, estimation of plausible material and illumination has remained elusive in uncontrolled settings and at a large scale.

Successful material and illumination estimation, however, will



Figure 2: Comparison to alternatives (projective texturing, average RGB of intrinsic images [BM15b]). We see that only a proper separation into specular materials and natural illumination can predict appearance in novel views. Other approaches miss the highlight, even in the original view (average of intrinsic), or does not move under view changes (projective texturing). Please refer to the accompanying video to judge the importance of moving highlights under view changes.

enable unprecedented quality of AR and VR applications like allowing realistic ‘transfer’ of objects across multiple photographs, or inserting high-quality replicas of virtual objects into Internet images. For example, in Figure 1, imagine transferring the red chair from one image to another. Currently, this task is challenging as we neither have access to the (red) chair’s material, nor the illumination in the target scene.

The naive solution of simply copying and pasting a 2D cutout is unsatisfactory as it easily leads to low fidelity results (e.g., unrealistic highlights), and more importantly, does not allow for pose adjustments (see Figure 2) or relighting.

In this paper, we investigate the problem of material and illumination estimation *directly* from ‘in the wild’ Internet images. The key challenge is that material and illumination are never observed independently, but only as the result of the convolving reflection operation with (estimated) normal direction and view direction (assuming access to rough geometry and pose estimates). Thus, in absence of further assumptions, we *cannot* uniquely recover material or illumination from single observations (i.e., images). Instead we rely on *linked* observations. We observe that often Internet images record the same objects in different environments (i.e., illuminations), or multiple objects in the same environments. Such *linked* observations among all the materials and illuminations forms a (sparse) *observation matrix* providing critical constraints among the observed materials and illumination parameters. We demonstrate that such a special structure can be utilized to robustly and accurately estimate all the material and illumination parameters through a global optimization.

We choose a formulation based on the basic rendering equation in combination with available per-pixel geometry estimation. However, there are multiple challenges: (i) access to only approximate proxy geometry for the scene objects with rough pose estimates leads to inaccurate normal estimates; (ii) partial observations due to view bias (e.g., chair backs are photographed less often) and sparsely observed normal directions (e.g., flat regions in man-made objects); (iii) working with the rendering equation when updating material and illumination parameters in an inverse problem setup is inefficient in a standard physically-based rendering pipeline; and finally, (iv) access to limited data due to sparsely observed joint material-illumination pairs.

In order to overcome the above challenges, we propose a novel formulation using parametric mixture models. We propose to approximate the reflection operator and its derivative with respect to material and illumination in terms of Anisotropic Spherical Gaus-

sians (see [XSD*13]) that can be efficiently utilized to jointly optimize for the materials and illumination at a large scale (i.e., involving multiple materials and illuminations). This optimization is driven by two neural networks that were trained on a large set of materials and illuminations to predict the gradient the optimization will follow. For example, in Figure 1, we observe 4 different colored (i.e., material) chairs under 8 different illuminations (only 6 images shown in the teaser figure) with linked observations. Only using these limited observations, our algorithm extracts high-quality material and illumination estimates, which can then be used for non-trivial image manipulation.

We extensively evaluate our method on both synthetic and real data, both quantitatively and qualitatively (using a user study). We demonstrate that increasing the amount of linked material-illumination observations improves the quality of both the material and illumination estimates. This, in turn, enables novel image manipulations previously considered to be very challenging. In summary, our main contributions are: (i) proposing the problem of coupled material and illumination estimation from a set of Internet images; (ii) formulating an efficient and scalable algorithm that allows high-quality material and illumination estimation from a set of images; (iii) using a neural network to approximate the complicated gradient of reflected light with respect to material and illumination parameters; and (iv) utilizing the estimations to enable realistic photo-realistic image manipulations.

2. Related Work

Our goal is to perform an advanced *intrinsic image* decomposition (as a factorization into materials and illumination) using an *image collection* with application in *photo-realistic image manipulation*. In the following, we review the related works.

Materials and illumination estimation from images. The classic intrinsic image decomposition problem [BT78] is highly ambiguous as many shapes, illuminations, and reflectances can explain one observation made. When geometry and material for the objects in an image are known, finding the illumination is a problem linear in a set of basis images [MG97]. Reflectance maps [HB89] can also be used to map surface orientation to appearance, allowing for a limited range of applications, such as novel views [RNR*16]. In absence of such information, alternatives regularize the problem using statistics of each component such as texture [STL08], or exploit user annotations on Internet images [BBS14] to develop a CRF-based decomposition approach. The latter method is widely considered to be the state-of-the-art for uncontrolled observations.

Haber et al. [HFB^{*}09] used observations of a single known geometry observed in a small set of images to estimate a linear combination of basis BRDFs and pixel-basis lighting. Aittala et al. [AWL15] capture texture-like materials by fitting SVBRDF using texture statistics to regularize a non-linear optimization on single image capture. An alternate recent trend is to use machine learning to solve inverse rendering problems. Deep learning of convolutional neural networks CNNs (cf., [KSH12]) has been used to decompose Lambertian shading [TSH12, BM15a], albedo in combination with other factors [NMY15], intrinsics from rendered data [SDSY16], decompose images into rendering layers [IRWM17], or multiple materials under the same illumination [GRR^{*}16].

A complementary but related problem is shape-from-shading, where again many shapes can explain a given image [HB89]. Alternately, illumination estimation has made use of shadows on diffuse receivers with known geometry [GHH01, SSI03, PSP09]. Another option is to assume access to outdoor illumination following a parametric sky model [LEN09]. Approaches to jointly solve for shape, reflectance and illumination in a single image or multiple materials under the same illumination using optimization with priors were suggested [ON15, LN16]. Our method, being applicable to direct observation of specular materials under uncontrolled illumination, is more general and works on a large variety of materials in many illuminations.

Image-based rendering (see monograph [SCK08]) can be used to re-create view-dependent appearance when a sufficiently dense set of images is provided. It then can produce novel views but fails to transfer to novel shapes or novel illuminations.

Image and shape collections. Visual computing has made increasing use of data, particularly image and/or shape collections with the aim to exploit cross observations. Starting from illumination [Deb98] and its statistics [DLAW01], measurements of BRDFs [MPBM03], we have seen models of shape [OLGM11], appearance [NNRS15], object pose estimate [AME^{*}14], object texture [WSH^{*}16], object attributes [HOM15] made possible by discovering correlation across observations in image and/or 3D model collections. In the context of shape analysis, mutual constraints of instances found across images or 3D scenes in the collection have been used to propose room layouts [YYT^{*}11], material assignments [JTRS12], or scene color and texture assignments [CXY^{*}15]. Instead, we directly estimate materials and illumination, rather than solving an assignment problem.

Photo-realistic image manipulation. While the rendering equation [Kaj86] explains image formation to the largest part with advanced signal processing perspective for the forward theory of light transport [RH01], practical solutions are still lacking for the inverse problem, i.e., estimating object materials and illumination directly from (uncontrolled) photographs. Instead, specialized user interfaces can assist this process [OCDD01], or multiple images of the same (static) scene can be used to improve relighting and material estimation [HFB^{*}09]. Several manipulations of images are possible, even without knowing the decomposition into shape, illumination and material due to perceptual effects [KRFB06].

In terms of state-of-the-art image manipulations, 3-Sweep [CZS^{*}13] propose a generalized cylinder-based interactive

tool to efficiently generate part-level models of man-made objects along with inter-part relations, and use them to enable a diverse variety of non-trivial object-level interactions; Khogade et al. [KSES14] align stock 3D models with input images to estimate illumination and appearance, and use them for impressive object-level image manipulations; while Karsch et al. [KSH^{*}14] estimates a comprehensive 3D scene model from a single, low dynamic range photograph and uses the information to insert virtual objects into the scene.

3. Overview

Starting from a set of linked photographs (i.e., multiple objects observed in different shared environments), our goal is to retrieve object geometry with pose predictions and estimate per-object materials and per-environment illuminations. The estimated information can then be used to faithfully re-synthesize original appearance and more importantly, obtain plausible view-dependent appearance. Figure 2 shows baseline comparisons to alternative approaches to assign materials to photographed objects. We observe that even if the geometry and light is known (we give all the approaches access to our estimated environment maps, if required), the highlights would either be missing (using intrinsic image [BM15b] for estimating average albedo), or not move faithfully (e.g., with projective texturing) under view changes.

As input, we require a set of photographs of shared objects with their respective masks (see Figure 3). In particular, we assume the materials segmentation to be consistent across images. As output, our algorithm produces a parametric mixture model (PMM) representation of illumination (that can be converted into a common environment map image) for each photograph and the reflectance parameters for every segmented material. We proceed in three steps.

First, we estimate object geometry and pose, and convert all the input images into an unstructured reflectance map for each occurrence of one material in one illumination in Section 4.1. Since we work with very few images collected from the wild, our challenge is that this information is very sparse, incomplete, and often contradict each other.

Second, we solve for illumination for each image and reflectance model parameters for each material in Section 4.4. This requires combining a very large number of degrees of freedom, as fine directional lighting details as well as accurate material parameters to be estimated. The challenge is that a direct optimization can easily involve many variables non-linearly coupled and lead to a cost function that is highly expensive even to evaluate as it involves solving the forward rendering equation, e.g., [LN16]. For example, representing images and illumination in the pixel basis leads to an order of $10^4\text{-}10^5$ variables (e.g., $128 \times 256 \times \text{number-of-environment-maps}$). At the same time, evaluating the cost function for every observation pixel would amount to gathering illumination by iterating all pixels in the environment map, i.e., an inner loop over all 128×256 environment map pixels inside an outer loop across all the $640 \times 480 \times \text{number-of-images-in-the-collection}$ observations. This quickly becomes computationally intractable.

Instead, we introduce a solution based on parametric mixture-model (PMM) representation of illumination to *inverse* render-

ing, which has been successfully applied to forward rendering [GKMD06, WRG*09, WDR11, XSD*13, VKŠ*14]. Our core contribution is to take PMM a step further by introducing the parametric mixture reflection operator and an approximation of its gradient, allowing to solve the optimization in a scalable fashion involving many materials and environments. The gradient approximation uses a neural network to map from observed reflected light, light and material parameters to changes of light and material parameters. It is trained on a set of synthetic images rendered from many illuminations and many materials.

Third, the estimated material and illumination information can directly be used in standard renderers. The challenge in such applications is to capture view-dependent effects such as moving highlights. In Section 5.3, we show applications to manipulating images, changing the illumination and/or material and/or view, transferring materials to objects in other images, or inserting objects into new illumination (see also supplementary materials).

4. Algorithm

We now explain our approach in details.

4.1. Acquiring Geometry and Reflectance Maps

We start from a set of images with the relevant materials segmented consistently across the image collection. Designer websites (e.g., Houzz) and product catalogs (e.g., Ikea) regularly provide such links. Here we assume that the links are explicitly available as input. First, we establish a mapping between illumination material-pairs and observed appearance.



Figure 3: RGB, normal, and segmentation of a typical input image.

Per-pixel labels. For the input images, we used per-pixel orientation (screen-space normals) (Figure 3) obtained using render-for-CNN [SQLG15] trained on the ShapeNet to retrieve object geometry and pose estimates. We found this to provide better quality normal predictions than those obtained via per-pixel depth [EF15] and normal [WFG15] estimation.

Reflectance maps. The rendering equation [Kaj86] states that

$$L_o(\mathbf{x}, \mathbf{n}, \omega_o) = \underbrace{L_e(\mathbf{x}, \omega_o)}_{\text{Emit}} + \int_{\Omega} \underbrace{f_r(\mathbf{x}, \omega_i, \omega_o)}_{\text{BRDF}} \underbrace{L_i(\mathbf{x}, \omega_i)}_{\text{Incom.}} \underbrace{\langle \mathbf{n}, \omega_i \rangle^+}_{\text{Geometry}} d\omega_i, \quad (1)$$

where \mathbf{x} is the position, \mathbf{n} the surface normal at location \mathbf{x} , ω_o the observer direction, L_o is the observed radiance, L_e is light emission, L_i is the incoming illumination, and f_r the bi-directional reflectance distribution function (BRDF) [Nic65].

We assume a simplified image formation model that allows for

using a slightly generalized variant of reflectance maps [HB89]: (i) distant illumination, (ii) convex objects, i.e., no shadows or inter-reflections, (iii) *spatially invariant* BRDFs, and (iv) no emission. Note that we do *not* assume a distant viewer as typical reflectance map does. This simplifies Eq. 1 to

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

A classic reflectance map is parameterized either by normal \mathbf{n} or by the observer direction ω_o . Instead of making such a split, we take a less structured approach tailored to our problem: an *unstructured reflectance map* (URM) denoted by \mathcal{O} that uses a list that holds in each entry a tuple of (i) normal \mathbf{o}_n , (ii) half-angle vector \mathbf{h} , (iii) observed radiance o_L (cf. Figure 4), and (iv) indices o_m and o_i of the material and illumination, respectively. We denote \mathbf{h} as the half-angle vector for front ($-z$) and observer direction, $\mathbf{h} := (\langle 2\mathbf{n}, \omega_o \rangle - \mathbf{n} \cdot \omega_o + (0, 0, -1)) / 2$. This parametrization will provide a more convenient way to index information. An example visualization of the URM by projecting the \mathbf{n} as well as the \mathbf{h} coordinate using latitude-longitude is seen in Figure 4.

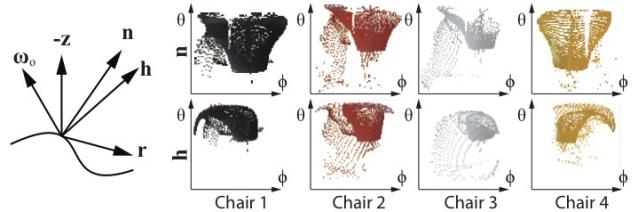


Figure 4: Schema and actual Unstructured Reflectance Maps of the chairs in the first column of Figure 1. Each point is an observed color for a specific surface orientation \mathbf{n} and half-angle vector \mathbf{h} .

To acquire the URM from an image with given per-pixel position and orientation, we apply inverse gamma correction such that o_L is in physically linear units. Note, that although we do not know the absolute scale inside each photo, we do not need it for most applications. Further, we do not differentiate between objects and consider only their materials (i.e., an object with two material parts are essentially treated as two materials).

4.2. Representation

Illumination. We use Parametric Mixture Models (PMMs) to represent illumination. PMMs have been used for pre-computed light transport [GKMD06, TS06, WRG*09], BTF compression [WDR11], interactive rendering [Tok15], importance sampling [VKŠ*14], or even in caustic design [PJJ*11]. A PMM encoded as

$$g(\omega|\Theta) := \sum_{l=1}^{n_p} p(\omega|\Theta_l) \approx L(\omega) \quad (3)$$

is a sum of n_p *lobe* functions $p(\omega|\Theta_l)$ that depend on a parameter vector Θ_l to approximate, in our setting, the incoming or outgoing light function $L(\omega)$. All parameter vectors Θ_l of one PMM are combined in a parameter matrix Θ . In our case, the domain of g is the sphere Ω parameterized using latitude-longitude representation $\omega = (\theta, \phi) \in [0, 2\pi) \times [0, \pi]$.

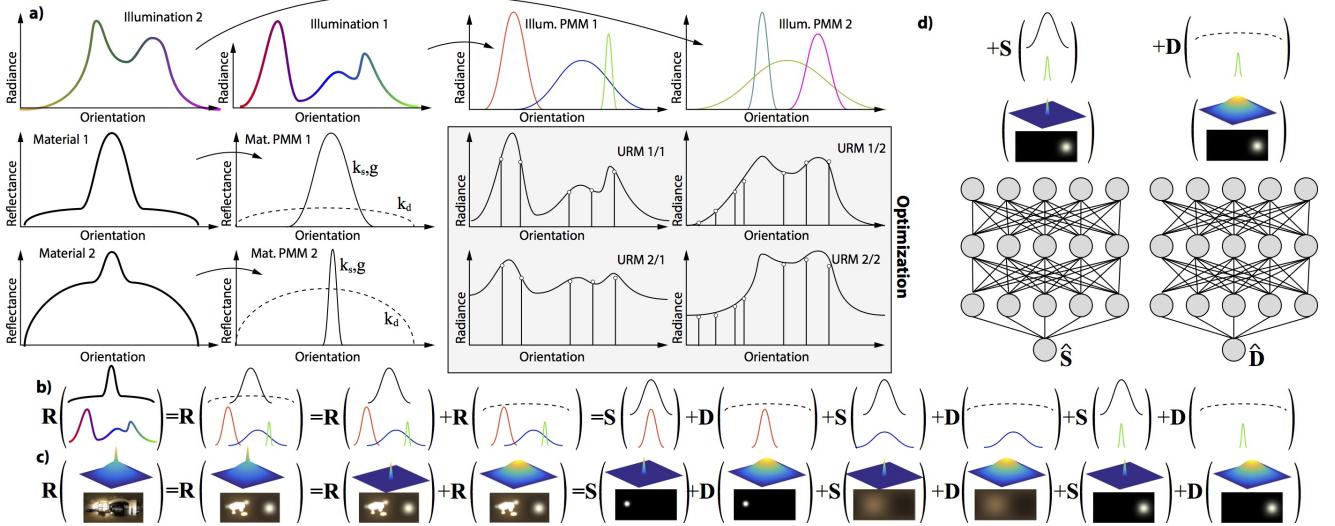


Figure 5: The three main ideas to enable large-scale optimization: (a) approximating illumination as parametric mixture models and the BRDF as a sum of a diffuse and a specular component; (b, c) expressing reflection as a sum of diffuse and specular reflections of individual lobes; and (d) approximating derivative of diffuse and specular reflection of ISGs using corresponding neural nets.

As mode functions, we employ Isotropic Spherical Gaussians (ISGs) [GKMD06, TS06, VKŠ*14]. An ISG lobe has the form

$$p(\omega|\Theta) := w \cdot \exp(-\sigma(\omega - \mathbf{z})^2),$$

where $w \in \mathbb{R}^+$ is the weight of the lobe, σ is its variance and \mathbf{z} the *mean* direction. Consequently, a lobe is described by parameter vector $\Theta = (w, \sigma, \mathbf{z})$. To work with RGB values all weight components w in this paper are vector-valued, but the variance parameter σ is scalar. For each image, we use an ISG PMMs with $n_p = 32$ components to represent unknown illuminations.

Material. We assume the material to be of the form

$$f_r(\omega_i, \omega_o | \rho) = \underbrace{k_d f_d(\omega_i, \omega_o)}_{\text{Diffuse}} + \underbrace{k_s f_s(\omega_i, \omega_o | r)}_{\text{Specular}}, \quad (4)$$

a parametric model that can be split into the weighted sum of a diffuse and a specular component f_d and f_s with weights k_d and k_s , respectively. We choose Lambertian as the diffuse model and GGX [WMLT07] that has a single roughness parameter r as the specular model. The material parameters are therefore a tuple $\rho = (k_d, k_s, r) \in \mathbb{R}^7$ of RGB diffuse and specular reflectance and a scalar roughness parameter. We denote the BRDF parameter vector of material j as $\rho^{(j)}$. Note that we do not need to represent f_r using a PMM, which would introduce unnecessary approximation error.

4.3. Reflection

Using standard notation [ATS94] for light transport, we express reflection as an operator \mathbf{R} , mapping the function of incoming light L_i to a function of reflected outgoing light L_o :

$$L_o(\omega_o) = \mathbf{R}(L_i | \rho)(\omega_o) = \int_{\Omega} \underbrace{L_i(\omega_i)}_{\text{Illumination}} \underbrace{f_r(\omega_i, \omega_o | \rho)}_{\text{BRDF}} d\omega_i. \quad (5)$$

When using an ISG to represent the illumination, we suggest to use a *parametric reflection operator* $\mathbf{R}(\Theta | \rho)$ that maps from a single illumination ISG lobe Θ and a material ρ to a reflected light. As we assume the BRDF to be a sum of a diffuse and a specular part, we can similarly define \mathbf{D} and \mathbf{S} that are respectively the diffuse and the specular-only reflection and $\mathbf{R} = \mathbf{D} + \mathbf{S}$. So, finally we have

$$L_o(\omega_o) = \sum_{l=1}^{n_l} \mathbf{D}(\Theta_l | \rho) + \mathbf{S}(\Theta_l | \rho). \quad (6)$$

4.4. Formulation

Our task is to find a set of illuminations and a set of materials that explain the acquired observations (see the previous section). Next, we describe how to represent reflectance and illumination as well as introduce the parametric reflectance operator, its derivative with respect to material and illumination, and an approximation method for efficient joint optimization for material and illumination given the observations (see Figure 5).

Cost function. Our main objective function quantifies how well a set of materials and illuminations explain the input observation. It should be fast to evaluate and allow for an effective computation of its gradient with respect to illuminations and materials in order to be useful in an optimization. We formulate the objective as:

$$c(\Theta, \rho | \mathcal{O}) := \sum_{o \in \mathcal{O}} \underbrace{\left\| o_L - \sum_{l=1}^{n_p} \mathbf{R}(\Theta_l | \rho^{(o_m)}) (o_\omega) \right\|^2}_{\text{Data}} + \lambda p(\Theta). \quad (7)$$

The gradient of this function with respect to the illumination and material comprises of evaluating \mathbf{R} , which involves convolving an

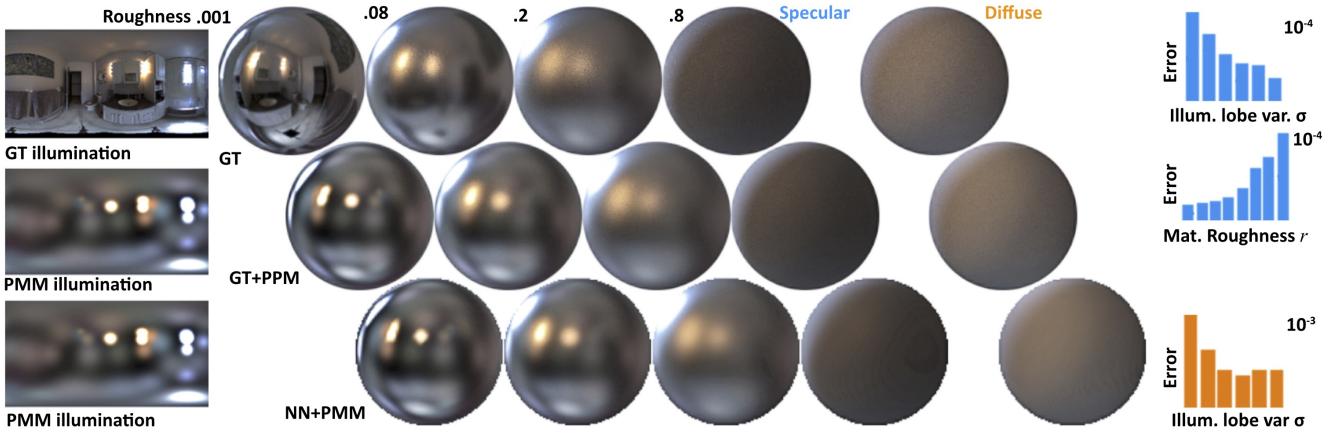


Figure 6: Evaluation of the neural network. The first row shows GT renderings with a GT envmap. The second row shows again GT rendering, but using the GMM fit to the envmap. This is an upper bound on the NN quality, as it works on the GMM representation. The third row shows the NN result. In the horizontal direction, specular results of increasing roughness are followed by the diffuse result in the rightmost column. The plots on the right below show the error distribution as a function of different parameters.

illumination lobe with the BRDF. This is both costly to compute and we need to find its derivative. To this end, we will employ a learning-based approach, as described next.

Neural network. The input to this neural network (NN) is the parameters of a single illumination lobe, the material parameters, and the observation direction ω . We call this approximation $\hat{\mathbf{R}}$. The output is an RGB value. We keep the NNs for the diffuse and specular components to be separate and independently process the RGB channels. The corresponding approximations using NNs are denoted as $\hat{\mathbf{D}}$ and $\hat{\mathbf{S}}$, respectively. The network architecture is shown in Figure 7. The input to the network is a 12-dimensional vector and differs between diffuse and specular NNs. Both consume the parameters of a single illumination lobe (direction and variance). However, the diffuse net consumes the normal while the specular net consumes the half-angle. All layers are full convolutional with 288 units in each layer. The networks are trained from 200k samples from renderings of spheres produced using Blender. An evaluation of this architecture is found in Figure 6.

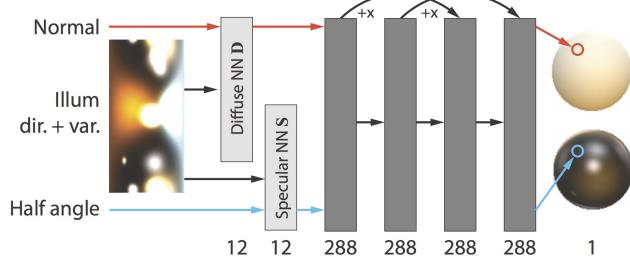


Figure 7: Our diffuse (orange) and specular (blue) neural network architecture, that consumes either normal and a single illumination lobe, or half-angle (left) and a lobe to produce a color (right).

Prior. As reflectance is typically more chromatic than illumination is, our prior penalizes the variance of the illumination lobe colors i.e., their RGB weights, as in $p(\Theta) = V(q(\Theta)) = E(q(\Theta)^2) - E(q(\Theta))^2$, where $q(\Theta) = \sum_{\Theta \in \Theta} \sum_{i=1}^{n_p} \Theta_{w,i}$. In other words by, first computing the average color of all lobes $q(\Theta)$ and second the variance $V(q(\Theta))$ of those three channels.

Optimization. Armed with a fast method (see above) to evaluate the cost function, we employ LBFGS [CBLN97] in combination with randomization. As the full vector \mathcal{O} does not fit into memory, we use a randomized subset that fits GPU memory in each iteration and dynamically change the randomization set across iterations. We stop our optimization when each observation on average has been sampled 5 times.

4.5. Rendering

For rendering, the result of the optimization is simply converted into an HDR environment map image by evaluating the estimated PMM for each pixel. Such an environment map along with estimated diffuse/specular parameters are then used with standard offline and online rendering applications as shown in our results.

5. Results

In this section, we evaluate our approach on synthetic and real image collection data (Section 5.1), compare to alternatives (Section 5.2) and give examples of potential applications (Section 5.3). We use L-BFGS solver for all the experiments. The complexity of our optimization in terms of the number of variables is $(7m + 6n_p n)$ and hence is linear in terms of the number of input entries in the material \times environment observation matrix. For example, for a five-photo, five-material matrix dataset, it costs about 30 minutes using a NVIDIA Titan X GPU. Pre-training the reflection operator \mathbf{R} , both diffuse and specular components, takes about three hours on the same specification.



Figure 8: Results on the INTERNET-LAPD dataset of four images of police cars with two materials. The first row shows the input images. The second row the reflectance maps. The observed ones are marked with black circles. In this example, all are observed. When an RM is not observed, it is re-synthesized. The third row shows our estimated illumination. Recall, that it is defined in camera space. The fourth row contains a re-synthesis using our material and illumination. Please note, that such a re-synthesis is not possible using a trivial factorization as all images have to share a common material that sufficiently explains the images. The last row shows a re-synthesis from a novel view, as well as a rendering of the materials in a new (Uffizi) illumination.

5.1. Evaluation

5.1.1. Datasets

We evaluated our method using three types of data sets, each holding multiple image collections acquired in different ways. The full resolution images and result images/video are included in the supplementary material.

The first comprises of SYNTHETIC images rendered using Mitsuba [Jak10] and a collection of HDR environment maps. Note that here we have access to ground-truth per-pixel normals and material labels. Here, we have rendered 3 objects in 3 different scenes with both spheres and real-world shapes that allow synthetic re-combination in an arbitrary fashion. This allows us to evaluate the proposed approach under different input variants, validating its scalability.

The second data set consists of real images collected from the INTERNET. We have manually selected the images (using iconic object name-based Google search) and masked the image content. This dataset has three sets of photographs: the LAPD car (INTERNET-LAPD), the Docksta table (INTERNET-DOCKSTA), and the Eames DSW chair (INTERNET-EAMES). For geometry, we estimated used coarse quality meshes available from ShapeNet. Images are good

for qualitative evaluation but do not allow to quantify the error, especially in novel views.

The third dataset contains PHOTOS we have taken from designer objects we choose under illumination conditions (in our labs and offices). We have 3D-scanned these objects (using Kinect) to acquire their (rough) geometry. The photos are taken in five different environments and 7 materials are considered.

5.1.2. Qualitative evaluation

Visual quality. We show results of our approach in Figures 8, 9, 10, and 17. We evaluate the main objective of our approach, i.e., estimating illumination and reflectance from a photo collection. In each figure, we show the input images, rendering of all objects' materials from original view (with the background from input images) and a novel view as well as visualizations of the material and illumination alone. Input images are shown on the top with the outputs we produce on the bottom (see supplementary for full images). Observed reflectance maps are shown encircled in black. The objects are rendered from an identical novel view, which is more challenging than rendering from the original view. The material is shown by re-rendering it under a new illumination. The exposure between all images is identical, indicating that the reflectance is directly in

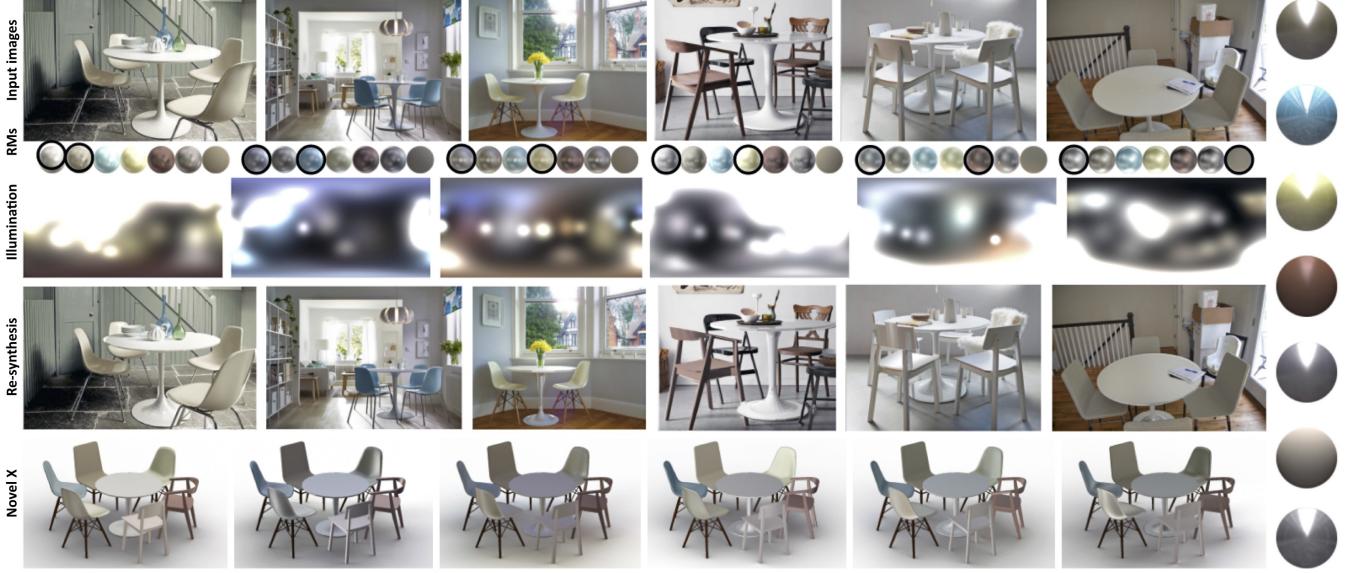


Figure 9: Results on the INTERNET-DOCKSTA dataset of six images of IKEA furniture with seven materials in the protocol of Figure 8.



Figure 10: Results on INTERNET-EAMES dataset of six images of a celebrated Eames chair with four materials in the protocol of Figure 8. The last row shows a pair of novel views, under input illumination where the first image has unobserved, the second an observed material.

the right order of magnitude and can transfer to new illuminations. While the illumination alone does not appear natural, shadow and shading from it produce plausible images, even of unknown materials or new objects. Recall that large parts of the objects are not seen in any of the input images and hence large parts of the environment maps are only estimated from indirect observation. Recall that our method does not use any data-driven prior to regularize the results.

The INTERNET-LAPD in Figure 8 shows a single object made from multiple materials. Figure 9 shows many chairs in many photos with one common ‘linking’ object INTERNET-DOCKSTA: the chair with material M_1 , that is used to calibrate all the other objects, which are only observed sometimes. Figure 10 shows multiple Eames chairs from a set of photos INTERNET-EAMES. All show

plausible highlights and colors, albeit only observing a fraction of the combinations. Please see the supplemental video for an animation of the highlights under changing view or object rotations. Figure 17 shows photos taken for this work, with all objects in all illumination conditions. Note, that both vases are made of multiple materials. The geometry of all objects in this part (except the chairs) is very approximate and acquired by a depth sensor. Still a good result quality is possible.

Prediction. Using the INTERNET-EAMES data sets, we are able to test the predictive ability of our approach by *leaving out* one image from the collection and compare it to the acquired ground truth. This is seen in Figure 12, where starting from the first image, we predict

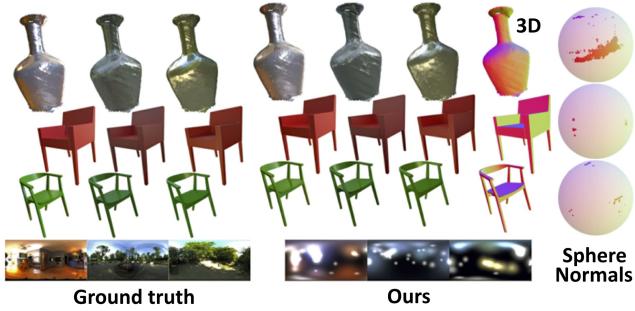


Figure 11: Here, we show the effect of Gauss-sphere coverage: Even for non-round objects with flat areas that have a bad coverage of the Gauss sphere the reconstruction (left) is similar to the reference (right).

the red chair in the second image and the yellow chair in the third image.



Figure 12: Estimating materials and illumination using all chairs except the rendered one. Left: reference image; Middle: the red chair is rendered; Right: the yellow chair is rendered.

Progressive estimation. A key property of our approach is to consolidate information from multiple images to disambiguate material and illumination. This characteristic implies that adding more images to the photo set should reduce the error. Figure 13 confirms the rise in performance as more images are added to the linked set.

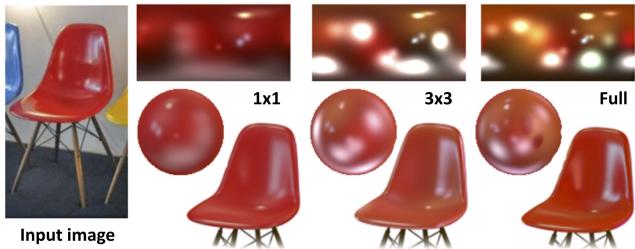


Figure 13: Progression of quality from left to right. Every row shows, for a selected material what the additional images can add to the quality in terms of re-rendering, material, and illumination.

5.1.3. Quantitative evaluation

We evaluate the effect of certain aspects of our method on the end-result (Figure 14). The error is quantified as DSSIM [WBSS04] structural image distance (smaller distance indicates better match) between a reference image rendered with known illumination and material compared to another rendering using our estimated material

and illumination. Images were gamma-corrected and linearly tone-mapped before comparison.

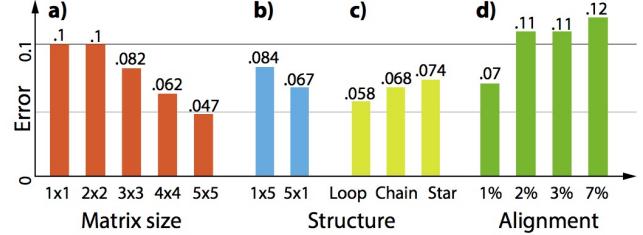


Figure 14: Effect of different input properties (horizontal) on the quality of our approach in terms of the DSSIM error (vertical, less is better): matrix size, structure, and alignment error.

Matrix size. In Figure 14a, we show the effect of increasing matrix size on the error of predicting the entries for the SYNTHETIC data set. Here, the matrix \mathcal{O} is complete, i.e., all material-illumination pairs are observed. We see, that with increasing size, the estimation for all entries gets more correct while the task being solved in some sense is also bigger (more different illuminations). Note that the total error residue can go up, but the estimation gets more accurate (compared to the ground truth).

When the matrix is reduced to a single row or column (Figure 14b) our approach can still estimate illumination and material. For a 1×5 matrix, which estimates a single material form multiple illuminations, the approach does well; but slightly degrades for a 5×1 setting, where multiple objects are seen under the same illumination.

Label quality. We assume the input images to have per-pixel normal and material labels. In Figure 14d, we study the effect of incorrect normal estimates by adding a different label of uniform noise to the normal. We see that good normal fair better with the error in the order of one percent, while larger errors produce an error that saturates still at a low total value. Also, note that for the INTERNET and the PHOTOS datasets, the geometry models are coarse and/or noisy. But in absence of ground truth, we could not measure estimation error.

5.2. Comparison

We compare possible alternatives to our approach as shown in Figure 2 and supplementary material. A simple approach could be using image-based rendering based approaches [SCK08], however, these approaches require either flat geometry or a high-enough number of images to reproduce specular appearance, neither of which is available in our input images that show a single image of one condition only. Effectively, IBR would amount to projective texturing in our setting, that is compared to our approach in Figure 2a. An alternative could be to run a general intrinsic image approach [BM15b] on the input images and use the average pixel color of the albedo k_d image as the diffuse albedo. The specular could then be the color that remains. While this would provide a specular value k_s , it is not clear how to get a glossiness value g (see Figure 2b).



Figure 15: Various photo-realistic image manipulations (e.g., object insertion) made possible using our estimated material and environment parameters (see Section 5.3 for details).

5.3. Application

A typical application of our approach is photo-realistic manipulation of objects in Internet images as shown in Figure 15. Having estimated the material and illumination parameters from all the images, we can insert virtual replica into the image (Figure 15b, 15d), transfer reflectance estimated from other Internet images to new scenes (Figure 15a and Figure 15c), or introduce new object with material under the estimated illumination. Please note that the estimated environment maps were used to render object shadows on (manually added) ground planes (Figure 15a, 15c, and 15d).

5.4. User Study

We have compared our approach to the similar approach (SIRFS) that extract intrinsic images and lighting in a user study. When asking $N = 250$ subjects if one of five animated turn-table re-renders using our material information or the model of SIRFS is preferred when showing both in a space-randomized 2AFC the mean preference was 86.5 % in our favor (std. error of the mean is 2.1 %). The chart of the user response, their mean, the exact sample counts and standard errors for individual images are presented in Figure 16.

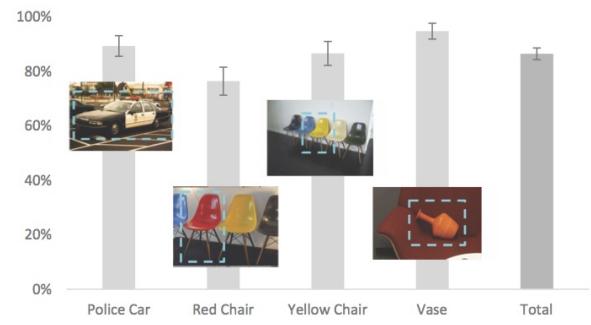


Figure 16: User study results. the vertical axis is the preference for ours, so more is better. Kinks are standard error of the mean, where small means certainty about the outcome.

5.5. Limitations

Our approach has three main limitations: First, we use a distant light model, which results in estimation errors in large rooms with interior lights. Second, although our environment map estimates lead to photo-realistic back projections and predictions, in absence of any data-driven regularization the illumination estimates themselves may look unnatural. This is primarily due to the limited samplings we have in our input measurements. Further, when objects are in close proximity, they may ‘show up’ in environment map estimations. For example, in Figure 10, we see the chair’s present red shading. This is because the incoming light for the white chair is distorted by the reflection of the red chair as shown in P4 (please refer to the video in the supplementary material).

6. Conclusion

We presented a novel optimization formulation for joint material and illumination estimation from collections of Internet images when different objects are observed in varying illumination conditions sharing coupled material and/or illumination observations. We demonstrated that such a linked material-illumination observation structure can be effectively exploited in a scalable optimization setup to recover robust estimates of material (both diffuse and specular) and effective environment maps. The estimations can then be used for a variety of compelling and photo-realistic image manipulation and object insertion tasks.

This work opens up several interesting future directions to pursue. In this work, we manually curated the photo sets to test our approach. While collecting datasets with star structure for the observation matrix is not so difficult – one can search with single keywords, especially for iconic designs – gathering more data linked with tighter connections (e.g., loops, full matrix, etc.) is more challenging. One option would be to harvest user annotations and existing links (like in Houzz or Pinterest websites) to collect such data. Such data sets can open up new material-illumination estimation pipelines – this is particularly exciting as we can update our estimates in an incremental way leading to simultaneous improvement of all the associated estimates. Another interesting direction would be to improve the



Figure 17: Results on the PHOTOS of five images of furniture and objects with seven materials in the protocol of Figure 8.

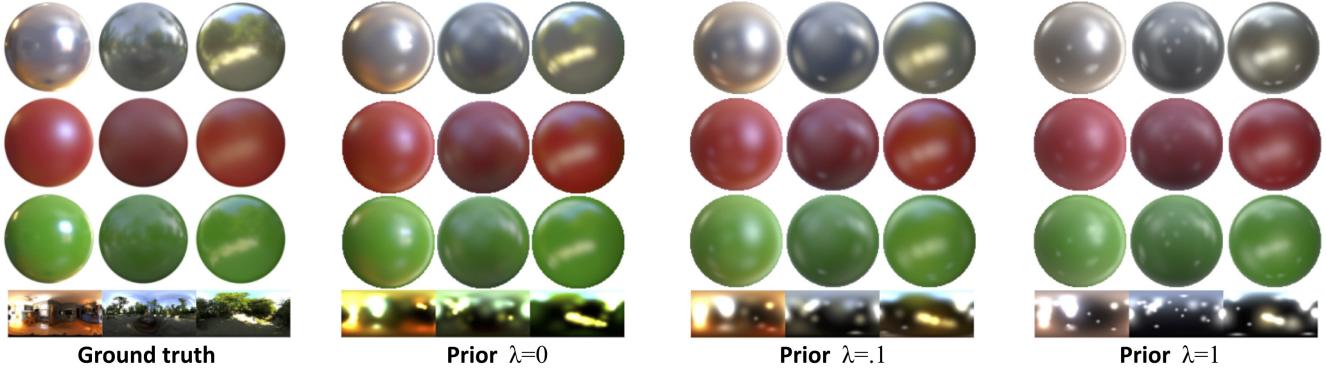


Figure 18: Effect of increasing prior. The top three rows show a 3×3 input. The next row shows illumination. Following the GT left, we increase the prior weight left to right. We note that illumination chromaticity decreases with increasing weight and that a good trade-off is likely at 0.1.

environment map estimates – one option would be to additionally use data priors to project the estimates environment maps to some data-driven manifold of environment maps measurements. Finally, an interesting future direction is to use the material and illumination estimates to improve geometry and pose estimates by refining correspondences by (partially) factoring our shading and illumination effects.

References

- [AME*14] AUBRY M., MATURANA D., EFROS A. A., RUSSELL B. C., SIVIC J.: Seeing 3d chairs: exemplar part-based 2D-3D alignment using a large dataset of cad models. In *CVPR* (2014), pp. 3762–69.
- [ATS94] ARVO J., TORRANCE K., SMITS B.: A framework for the analysis of error in global illumination algorithms. In *Proc. SIGGRAPH* (1994), pp. 75–84.
- [AWL15] AITTALA M., WEYRICH T., LEHTINEN J.: Two-shot SVBRDF capture for stationary materials. *ACM Trans. Graph. (Proc. SIGGRAPH)* 34, 4 (2015), 110:1–110:13.
- [BBS14] BELL S., BALA K., SNAVELY N.: Intrinsic images in the wild. *ACM Trans. Graph.* 33, 4 (2014), 159.
- [BM15a] BARRON J. T., MALIK J.: Intrinsic scene properties from a single rgb-d image. *PAMI* (2015).
- [BM15b] BARRON J. T., MALIK J.: Shape, illumination, and reflectance from shading. *PAMI* (2015).

- [BT78] BARROW H. G., TENENBAUM J. M.: Recovering intrinsic scene characteristics from images. *Comp. Vis. Sys.* (1978).
- [CBLN97] CZHU C., BYRD R. H., LU P., NOCEDAL J.: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Transactions on Mathematical Software* 23, 4 (1997), 550–560.
- [CXY*15] CHEN K., XU K., YU Y., WANG T.-Y., HU S.-M.: Magic decorator: Automatic material suggestion for indoor digital scenes. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 34, 6 (2015).
- [CZS*13] CHEN T., ZHU Z., SHAMIR A., HU S.-M., COHEN-OR D.: 3sweep: Extracting editable objects from a single photo. *ACM Trans. Graph.* 32, 6 (2013).
- [Deb98] DEBEVEC P.: Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *SIGGRAPH* (1998).
- [DLAW01] DROR R. O., LEUNG T. K., ADELSON E. H., WILLSKY A. S.: Statistics of real-world illumination. In *CVPR* (2001).
- [EF15] EIGEN D., FERGUS R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV* (2015).
- [GHH01] GIBSON S., HOWARD T., HUBBOLD R.: Flexible image-based photometric reconstruction using virtual light sources. *Comp. Graph. Forum* 20, 3 (2001), 203–214.
- [GKMD06] GREEN P., KAUTZ J., MATUSIK W., DURAND F.: View-dependent precomputed light transport using nonlinear gaussian function approximations. In *Proc. i3D* (2006).
- [GRR*16] GEORGULIS S., REMATAS K., RITSCHEL T., FRITZ M., TUYTELAARS T., VAN GOOL L.: Natural illumination from multiple materials using deep learning. *arXiv:1611.09325* (2016).
- [HB89] HORN B. K. P., BROOKS M. J. (Eds.): *Shape from Shading*. MIT Press, 1989.
- [HFB*09] HABER T., FUCHS C., BEKAER P., SEIDEL H.-P., GOESELE M., LENSCHE H. P., ET AL.: Relighting objects from image collections. In *CVPR* (2009).
- [HOM15] HUETING M., OVSJANIKOV M., MITRA N.: Crosslink: Joint understanding of image and 3d model collections through shape and camera pose variations. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 34, 6 (2015).
- [IRWM17] INNAMORATI C., RITSCHEL T., WEYRICH T., MITRA N. J.: Decomposing single images for layered photo retouching. *Computer Graphics Forum (Proc. Eurogr. Symp. on Rendering)* 36, 4 (2017), 15–25.
- [Jak10] JAKOB W.: Mitsuba renderer, 2010. <http://www.mitsuba-renderer.org>.
- [JTRS12] JAIN A., THORMÄHLEN T., RITSCHEL T., SEIDEL H.-P.: Material memex: Automatic material suggestions for 3d objects. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 31, 5 (2012).
- [Kaj86] KAJIYA J. T.: The rendering equation. In *ACM SIGGRAPH* (1986).
- [KRFB06] KHAN E. A., REINHARD E., FLEMING R. W., BÜLTHOFF H. H.: Image-based material editing. *ACM Trans. Graph.* (2006).
- [KSES14] KHOLGADE N., SIMON T., EFROS A., SHEIKH Y.: 3d object manipulation in a single photograph using stock 3d models. *ACM Trans. Graph. (Proc. SIGGRAPH)* 33, 4 (2014).
- [KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *NIPS* (2012).
- [KSH*14] KARSCH K., SUNKAVALLI K., HADAP S., CARR N., JIN H., FONTE R., SITTIG M., FORSYTH D.: Automatic scene inference for 3d object compositing. *ACM Trans. Graph.* 33, 3 (June 2014).
- [LEN09] LALONDE J.-F., EFROS A. A., NARASIMHAN S. G.: Estimating natural illumination from a single outdoor image. In *CVPR* (2009), pp. 183–190.
- [LN16] LOMBARDI S., NISHINO: Reflectance and illumination recovery in the wild. *PAMI* (2016).
- [MG97] MARSCHNER S. R., GREENBERG D. P.: Inverse lighting for photography. In *Color and Imaging Conf.* (1997), pp. 262–265.
- [MPBM03] MATUSIK W., PFISTER H., BRAND M., McMILLAN L.: A data-driven reflectance model. *ACM Trans. Graph.* (2003).
- [Nic65] NICODEMUS F. E.: Directional reflectance and emissivity of an opaque surface. *Applied optics* (1965).
- [NMY15] NARIHIRA T., MAIRE M., YU S. X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV* (2015).
- [NNRS15] NGUYEN C. H., NALBACH O., RITSCHEL T., SEIDEL H.-P.: Guiding image manipulations using shape-appearance subspaces from co-alignment of image collections. *Computer Graphics Forum (Proc. Eurographics 2015)* 34, 2 (2015).
- [OCDD01] OH B. M., CHEN M., DORSEY J., DURAND F.: Image-based modeling and photo editing. In *Proc. SIGGRAPH* (2001).
- [OLGM11] OVSJANIKOV M., LI W., GUIBAS L., MITRA N. J.: Exploration of continuous variability in collections of 3d shapes. *ACM Trans. Graph.* 30, 4 (2011).
- [ON15] OXHOLM G., NISHINO K.: Shape and reflectance estimation in the wild. *PAMI* (2015).
- [PJ*11] PAPAS M., JAROSZ W., JAKOB W., RUSINKIEWICZ S., MATUSIK W., WEYRICH T.: Goal-based caustics. *Comp. Graph. Forum (Proc. Eurographics)* 30, 2 (2011), 503–511.
- [PSP09] PANAGOPOULOS A., SAMARAS D., PARAGIOS N.: Robust shadow and illumination estimation using a mixture model. In *CVPR* (2009), pp. 651–658.
- [RH01] RAMAMOORTHI R., HANRAHAN P.: A signal-processing framework for inverse rendering. In *SIGGRAPH* (2001).
- [RNR*16] REMATAS K., NGUYEN C., RITSCHEL T., FRITZ M., TUYTELAARS T.: Novel views of objects from a single image. *TPAMI* (2016).
- [SCK08] SHUM H.-Y., CHAN S.-C., KANG S. B.: *Image-based rendering*. Springer Science & Business Media, 2008.
- [SDSY16] SHI J., DONG Y., SU H., YU S. X.: Learning non-lambertian object intrinsics across shapenet categories. *arXiv:1612.08510* (2016).
- [SQLG15] SU H., QI C. R., LI Y., GUIBAS L. J.: Render for CNN: Viewpoint estimation in images using cnns trained with rendered 3d model views. In *The IEEE International Conference on Computer Vision (ICCV)* (December 2015).
- [SSI03] SATO I., SATO Y., IKEUCHI K.: Illumination from shadows. *PAMI* 25, 3 (2003), 290–300.
- [STL08] SHEN L., TAN P., LIN S.: Intrinsic image decomposition with non-local texture cues. In *CVPR* (2008).
- [Tok15] TOKUYOSHI Y.: Virtual spherical gaussian lights for real-time glossy indirect illumination. *Comp. Graph. Forum* 34, 7 (2015), 89–98.
- [TS06] TSAI Y.-T., SHIH Z.-C.: All-frequency precomputed radiance transfer using spherical radial basis functions and clustered tensor approximation. *ACM Trans. Graph.* 25, 3 (2006), 967–976.
- [TSH12] TANG Y., SALAKHUTDINOV R., HINTON G.: Deep lambertian networks. In *ICML* (2012).
- [VKŠ*14] VORBA J., KARLÍK O., ŠIK M., RITSCHEL T., KŘIVÁNEK J.: On-line learning of parametric mixture models for light transport simulation. *ACM Trans. Graph. (Proc. SIGGRAPH)* 33, 4 (2014), 101.
- [WBSS04] WANG Z., BOVIK A. C., SHEIKH H. R., SIMONCELLI E. P.: Image quality assessment: from error visibility to structural similarity. *IEEE TIP* 13, 4 (2004), 600–612.
- [WDR11] WU H., DORSEY J., RUSHMEIER H.: A sparse parametric mixture model for BTF compression, editing and rendering. *Comp. Graph. Forum* 30, 2 (2011), 465–473.
- [WFG15] WANG X., FOHEY D. F., GUPTA A.: Designing deep networks for surface normal estimation. In *CVPR* (2015).

[WMLT07] WALTER B., MARSCHNER S. R., LI H., TORRANCE K. E.: Microfacet models for refraction through rough surfaces. In *Proc. EGSR* (2007), pp. 195–206.

[WRG*09] WANG J., REN P., GONG M., SNYDER J., GUO B.: All-frequency rendering of dynamic, spatially-varying reflectance. *ACM Trans. Graph. (Proc. SIGGRAPH)* 28, 5 (2009), 133.

[WSH*16] WANG T. Y., SU H., HUANG Q., HUANG J., GUIBAS L., MITRA N. J.: Unsupervised texture transfer from images to model collections. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 35, 6 (2016).

[XSD*13] XU K., SUN W.-L., DONG Z., ZHAO D.-Y., WU R.-D., HU S.-M.: Anisotropic spherical gaussians. *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* 32, 6 (2013).

[YYT*11] YU L.-F., YEUNG S. K., TANG C.-K., TERZOPoulos D., CHAN T. F., OSHER S.: Make it home: automatic optimization of furniture arrangement. *ACM Trans. Graph.* 30, 4 (2011), 86.