

What Is Around The Camera?

Stamatios Georgoulis
KU Leuven

Mario Fritz
MPI for Informatics

Konstantinos Rematas
University of Washington

Tinne Tuytelaars
KU Leuven

Tobias Ritschel
University College London

Luc Van Gool
KU Leuven & ETH Zurich

Abstract

How much does a single image reveal about the environment it was taken in? In this paper, we investigate how much of that information can be retrieved from a foreground object, combined with the background (i.e. the visible part of the environment). Assuming it is not perfectly diffuse, the foreground object acts as a complexly shaped and far-from-perfect mirror. An additional challenge is that its appearance confounds the light coming from the environment with the unknown materials it is made of. We propose a learning-based approach to predict the environment from multiple reflectance maps that are computed from approximate surface normals. The proposed method allows us to jointly model the statistics of environments and material properties. We train our system from synthesized training data, but demonstrate its applicability to real-world data. Interestingly, our analysis shows that the information obtained from objects made out of multiple materials often is complementary and leads to better performance.

1. Introduction

Images are ubiquitous on the web. Users of social media quite liberally contribute to this fast growing pool, but also companies capture large amounts of imagery (e.g. street views).

Simultaneously, users have grown aware of the risks involved, forcing companies to obfuscate certain aspects of images, like faces and license plates.

We argue that such images contain more information than people may expect. The reflection in a foreground object combined with the background - the part of the environment revealed directly by the image - can be quite telling about the entire environment, i.e. also the part behind the photographer. Apart from privacy issues, several other applications can be based on the extraction of such environment maps. Examples include the relighting of the foreground object in a modified environment, the elimination of the photographer

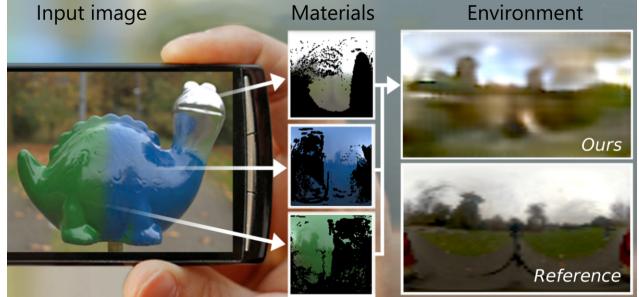


Figure 1. Given approximate surface normals, our approach predicts a truthful reconstruction of the environment from a single low dynamic range image of a (potentially multi-material) object.

in the reflection by the foreground object, or figuring out that a photo might have been nicely framed, but that the actual surroundings may not be quite as attractive as one might have been lead to believe (e.g. browsing through real estate).

By combining large amount of synthesized data and deep learning techniques, we present an approach that utilizes partial and ambiguous data encoded in surface reflectance in order to reconstruct a truthful appearance of the surrounding environment the image was taken in. While we assume a rough scene geometry to be given, we do not make any strong assumptions on the involved materials or the surrounding environment. Yet, we obtain a high level of detail in terms of color and structure of that environment. Beyond the new level of reconstruction quality obtained by our method, we show that our learned representations and final predictions lend to a range of application scenarios. Experiments are conducted on both synthetic and real images from the web.

Prior work on estimating the illuminating environment has either focused on retrieving a template map from a database (e.g. [18]) or recovering key elements to approximate it (e.g. [46]), but both directions are limited in terms of the achievable fidelity. Equally, prior work on inverse rendering [27, 24, 23] has the potential to derive arbitrary maps with radiometric accuracy, but to date are not able to do so with accurate appearance. In this spirit, we seek a different tradeoff. By not constraining our solution to be radiomet-

rically accurate, we investigate how far we can push the quality of overall appearance of the map. We hope that these two approaches can complement each other in the future.

Our work can be seen as an extension of the ideas presented in [13] and is the first to recover such highly detailed appearance of the environment. We achieve this in a learning-based approach that enables joint learning of environment and material statistics. We propose a novel deep learning architecture that combines multiple, partially observed reflectance maps together with a partial background observation to perform a pixel-wise prediction of the environment.

2. Previous work

Object appearance is the result of an intriguing jigsaw puzzle of unknown shape, material reflectance, and illumination. Decomposing it back into these intrinsic properties is far from trivial [3]. Typically, one or two of the intrinsic properties are assumed to be known and the remaining one is estimated. In this work, we focus on separating materials and illumination when the partial reflectance maps of multiple materials seen under the same illumination plus a background image are known. Such an input is very typical in images, yet not so often studied in the literature.

Key to this decomposition into intrinsic properties is to have a good understanding of their natural statistics. Databases of material reflectance [7, 28, 4] and environmental illumination [8, 11] allow the community to make some first attempts. Yet, exploiting them in practical decompositions remains challenging.

Beyond image boundaries Revealing "hidden" information about the environment in which an image was taken, has been used for dramatic purposes in movies and tv shows, but it has also attracted research interest. In [41], Torralba *et al.* are using a room as a camera obscura to recover the scene outside the window. Nishino *et al.* [30] estimate panoramas from the environment reflected in a human eye. On the other hand, [48] estimate a plausible panorama from a cropped image by looking at images with similar structure, while [43] extend the content of the image by high level graph matching with an image database. Moreover, a single image can contain hidden "metadata" information, such as GPS coordinates [16] or the photographer identity [40].

Reflectance maps Reflectance maps [17] assign appearance to a surface orientation for a given scene/material, thus combining surface reflectance and illumination. Reflectance maps can be extracted from image collections [15], from a known class [34, 33], or using a CNN [35]. In computer graphics, reflectance maps are used to transfer and manipulate appearance of photo-realistic or artistic "lit spheres" [39] or "MatCaps" [37]. Khan [19] made diffuse objects in a photo appear specular or transparent using manipulations of the image background that require manual intervention.

Factoring illumination Classic intrinsic images factor

an image into shading and reflectance [3]. Larger-scale acquisition of reflectance [28] and illumination [8] have allowed to compute their statistics [11] helping to better solve inverse and synthesis problems. Nevertheless, intrinsic images typically assume diffuse reflectance. Barron and Malik [2] decompose shaded images into shape, reflectance and illumination, but only for scalar reflectance, *i.e.* diffuse albedo, and for limited illumination frequencies. Richter *et al.* [36] estimated a diffuse reflectance map represented in spherical harmonics using approximate normals and refined the normal map using the reflectance map as a guide, but their approach is suitable to represent low-frequency illumination, while our environment maps reproduce fine details.

Separating material reflectance (henceforth simply referred to as 'material') and illumination was studied by Lombardi and Nishino [24, 23, 26]. In [24, 23], they present different optimization approaches that allow for high-quality radiometric estimation of material and illumination from known 3D shape and a single HDR RGB image, whereas in [26], they use multiple HDR RGBZ images to acquire material and illumination and refine shape using a similar optimization-based framework also handling indirect lighting. Here, we address a more general problem than these approaches: they consider one or more objects with a single, unknown material on their surface (homogeneous surface reflectance) observed under some unknown natural illumination. However, most real-life objects are made of multiple materials and as noted in [21, 49, 25] multiple materials help to estimate surface reflectance under a single point light source. Furthermore, they throw away the image background that is naturally captured in their images anyway.

In this paper, we assume that the objects consist of any number of materials (we investigate up to 5), that they are segmented into their different materials as well as from the background, and that the reflectance maps of all materials are extracted (conditions that are met by *e.g.* using a Google Tango phone as in [47]). We then ask how these multiple materials under the same non-point light illumination plus the background can help a deep architecture to predict the environment. Note that, unlike [24, 23] our primary goal is to estimate a "textured" environment map that is perceptually close to ground truth and can provide semantic information (*e.g.* the object is placed in a forest); later in Sec. 6 we examine if we can recover the dynamic range too as a secondary goal. Most importantly, we aim to do this starting from standard LDR images, as the HDR ones used in [24, 23, 26] imply the capture of multiple exposures per image making the capturing process impractical for non-expert users.

Barron *et al.* [1] made use of similar data to resolve spatially-varying, local illumination. While ours is spatially invariant (distant), we can extract it both with more details, in HDR and from non-diffuse surfaces.

Earlier work has also made use of cues that we did not

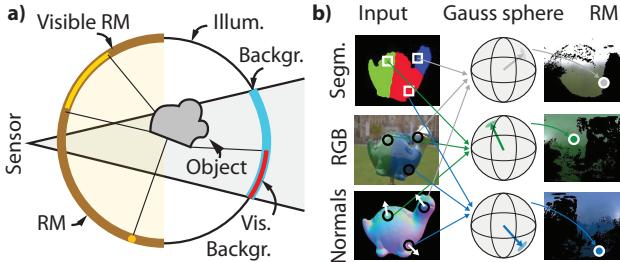


Figure 2. **a)** Illustration of Sec. 3. **b)** Converting 3 example input pixels into reflectance map pixels using normals and segmentation.

consider, such as shadows [38], as they may only be available in some scenes. The work of [20, 46] has shown how to fit a parametric sky model from a 2D image, but cannot reproduce details such as buildings and trees and excludes non-sky, *i.e.* indoor settings. Karsch *et al.* [18] automatically inferred environment maps by selecting a mix of nearest neighbors (NN) from a database of environment maps that can best explain the image assuming diffuse reflectance and normals have been estimated. They demonstrate diffuse relighting but specular materials, that reveal details of a reflection, hardly agree with the input image as seen in our results section.

3. Overview

We formulate our problem as learning a mapping from n_{mat} partial LDR reflectance maps [17] and a background LDR image to a single consensual HDR environment map. In particular, we never extract illumination directly from images, but indirectly from reflectance maps. We assume the reflectance maps were extracted using previous work [12, 44, 22, 35]. In our datasets, we analyze the limits of what reflectance map decomposition can do and we do not consider the error introduced by the reflectance map itself.

A reflectance map $L_o(\omega)$ represents the appearance of an object of a homogeneous material under a specific illumination. Under the assumptions of (i) a distant viewer, (ii) distant illumination, (iii) in the absence of inter-reflections or shadows (convex object) and (iv) a homogeneous material, the appearance depends only on the surface orientation ω in camera space and can be approximated as a convolution of illumination and material (*i.e.* BRDF) [32].

The full set of orientations in \mathbb{R}^3 is called the 3D *Gauss* sphere Ω (the full circle in Fig. 2 a and b). Note, that only at most half of the orientations in \mathbb{R}^3 are visible in camera space, *i.e.* the ones facing into the direction of the camera. This defines the positive Gauss sphere Ω^+ (the brown half-circle in Fig. 2 a). Also note, that due to the laws of reflections, surfaces oriented towards the viewer also expose illumination coming from behind the camera. The ideal case is a one-material spherical object, that completely contains all observable normals. When its surface behaves like a perfect mirror, that is even better. Then a direct (but partial)

environment map is directly observable. In practice, we only observe some orientations for some materials and other orientations for other materials. Sometimes, multiple materials are observed for one orientation, but it also happens that for some orientations, no material might be observed at all. Moreover, the materials tend to come with a substantially diffuse component in their reflectance, thus smearing out information about the environment map. In Fig. 2 a, the brown part shows the half-sphere of the reflectance map and the yellow part within shows the object normals actually observed in the image, for the example object in the figure.

A second piece of input comes from the background. The visible part of the background in the image shows another part of the illumination, this time from the negative half sphere. In Fig. 2 a, the visible part of the background is shown in red, the rest - occluded by the foreground - in blue.

The illumination $L_i(\omega)$ we will infer from both these inputs covers the full sphere of orientations Ω (the full circle in Fig. 2 a). Other than the reflectance map, it typically is defined in world space as it does not change when the viewer's pose changes. For the actual computations, both the input (partial reflectance maps and partial background) and the output (illumination) are represented as two-dimensional images using a latitude-longitude parameterization.

The mapping $f := L_o \rightarrow L_i$ we seek to find is represented using a deep CNN. We propose a network that combines multiple convolutional stages - one for each reflectance map, that share weights, and another one for the background - with a joint de-convolutional stage that consolidates the information into a detailed estimate of the illumination.

The training data consists of tuples of reflectance maps l_o with a single background image, *i.e.* inputs, and a corresponding illumination l_i , *i.e.* output.

4. Dataset

Our dataset consists of synthetic training and testing data (Sec. 4.1) and a manually-acquired set of test images of real objects captured under real illumination (Sec. 4.2). Upon publication, the dataset and code will be made available.

4.1. Synthetic data

We now explain how to synthesize train and test data.

Rendering Images are rendered at a resolution of 512×512 using variations of geometry, illumination, materials and views. The geometry is a random object from the ShapeNet [5] class “car”. Later, we show results of our pipeline for both cars and other shapes though (*e.g.* Fig. 1). As large 3D shape datasets from the Internet do not come with a consistent segmentation into materials, we perform a simple image segmentation after rasterization. To this end, we perform k -means clustering ($k = n_{\text{mat}}$) based on positions and normals, both weighted equally and scaled to the range $(-1, 1)$, to divide the shapes into three regions, to be

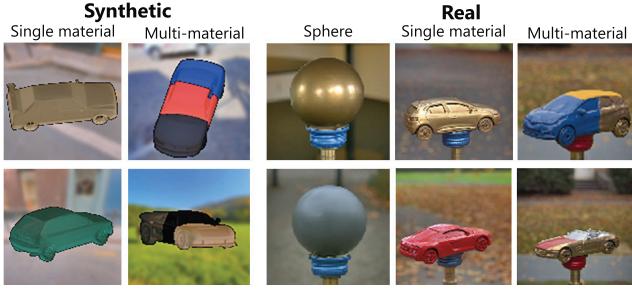


Figure 3. Example images from our dataset. **1st col:** Synthetic images of cars with a single material. **2nd col:** Synthetic images of cars with multiple materials. **3rd col:** Photographs of spheres with a single material. **4th col:** Photographs of toy cars with a single material. **5th col:** Photographs of toy cars with multiple materials.

covered with three different ‘materials’. Per-pixel colors are computed using direct (no global illumination and shadows) image-based illumination [10]. We also store per-pixel ground-truth positions and normals. As materials we used the 100 BRDF samples from the MERL database [28]. The illumination is randomly selected from a set of 105 publicly available HDR environment maps that we have collected. The views are sampled randomly over the sphere, with a fixed field-of-view of 30 degrees. Synthetic examples can be seen in the first two columns of Fig. 3.

Extracting reflectance maps The pixel j in the reflectance map of material i is produced by averaging all pixels with material i and orientation ω_j . This is shown for three pixels from three different materials in Fig. 2 b. The final reflectance maps contain 128×128 pixels. These are typically partial with sometimes as little as 10% of all normals observed (see some examples in Fig. 4). Even sparser inputs have not been studied (*e.g.* hallucination works [14]).

Background extraction The background is easily identified for these synthetic cases, by detecting all pixels where the geometry did not project to. To make the network aware of depth-of-field found in practice, the masked background is filtered with a 2D Gaussian smoothing kernel ($\sigma = 2$).

Building tuples To test our approach for an arbitrary number of materials, n_{mat} , we build tuples by combining n_{mat} random reflectance maps extracted from images with a single material. For each tuple we make sure to use mutually exclusive materials observed under the same illumination.

Splitting For the single-material case, from the 60 k synthetic images generated, 54 k are used for training and 6 k for testing. Note that, no environment map is shared between the two sets - 94 for training and 11 for testing, with the split generated randomly once. For the multi-material case, we used the same protocol (identical split) for two different sub-cases: multiple single-material objects (*i.e.* the tuples), and single multi-material objects (*e.g.* Fig. 3, 2nd column). In both cases, the environment maps are augmented by rotations around the y axis, which is essential for preventing the

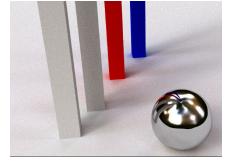
networks to memorize the training samples.

4.2. Real data

While training can be done on massive synthetic data, the network ultimately is to be tested on real images. To this end, we took photographs of naturally illuminated single-material and multi-material objects with known geometry. All images in this set - 112 in total - were used for testing and never for training. Moreover, all 3D models, materials and illuminations in this set are unknown to the train set.

Capture The images are recorded in LDR with a common DSLR sensor at a resolution of 20 M pixels and subsequently re-scaled to match the training data. To compare with reference, we also acquired the environment map in HDR using a chrome sphere. We recorded 7 f -stops [9], which

is a compromise between shadow fidelity and capture time, but as the inset figure shows - an example rendering with an estimated environment map (Fig. 8, row 2) - it is enough to produce clearly directed shadows. Three variants were acquired: spheres, single-material objects and multi-material objects (see Fig. 3). For the single-material case, 84 images were taken, showing 6 spheres and 6 toy cars with different materials each and placed under 7 different illuminations. The multi-material data comprises of 30 images, showing 6 different objects (4 cars and 2 non-cars), each painted with 3 materials, captured under 9 different illuminations (6 and 3 respectively). Some materials repeat, as overall 12 different materials were used.



Extracting reflectance maps and background From all images, reflectance maps are extracted in the same way as for the synthetic images. Per-pixel normals are produced using virtual replica geometry from online repositories or scanned using a structured-light scanner. These models were manually aligned to the 2D images. Material and background segmentation was also done manually for all images.

5. Network Architecture

Our network consists of three parts (Fig. 4) - some of them identical in structure and some sharing weights. First, there is a convolutional *background* network. Second, n_{mat} convolutional *de-reflection* networks that share parameters but run on the reflectance maps of different materials. Third, a final de-convolutional *fusion* network takes as input intermediate stages as well as end results from all reflectance nets, together with the result of the background net, to produce the HDR environment map as an output. All parts are trained jointly end-to-end using an L1 loss on the illumination samples, after applying the natural logarithm and converting them to CIE Lab space. We experimentally found that these

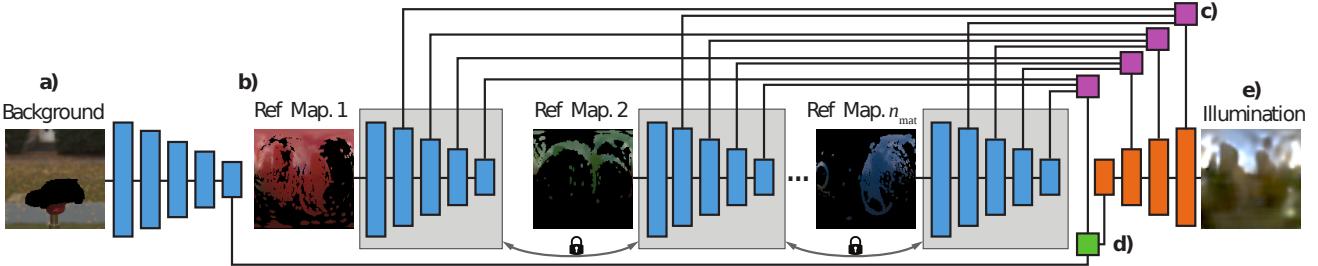


Figure 4. Our CNN architecture (*from left to right*). First, the background image is encoded using one independent sub-network (*blue*). Next, each partial reflectance map is encoded using n_{mat} de-reflection sub-networks that share parameters (*blue*). Finally, these two sources of information are fused in a de-convolution network (*orange*). Here, information from all levels of the partial reflectance maps is included (*violet*) as well as the global encoding of the background (*green*). Sub-network details are found in the text and supplementary material.

choices nicely balance between learning the dynamic range, structure and color distribution of the environment map.

Background network Input to the background network (blue part in Fig. 4, a) is an LDR background image in full resolution *i.e.* 128×128 converted to CIE Lab space. The output is a single, spatially coarse encoding of resolution 4×4 . The reduction in spatial resolution is performed as detailed in supplementary material. Only the final output of the encoding step will contribute to the fusion (Fig. 4, d).

De-reflection network The de-reflection network (blue parts in Fig. 4, b) consumes partial, LDR environment maps also converted to CIE Lab space, where undefined pixels are set to black. It has the same structure as the background network. It starts with the full, initial reflectance map at a resolution of 128×128 and reduces to a spatial resolution of 4×4 . We can support an arbitrary number of materials n_{mat} ; however the network needs to be trained for a specific number of materials. In any case, the de-reflection networks are trained with shared parameters (siamese architecture; locks in Fig. 4). We want each of these networks to perform the same operations so the reflectance maps do not need to come in a particular order.

Fusion network The fusion network (Fig. 4, e) combines the information from the background and the de-reflection network. The first source of information are the intermediate representations from the reflectance maps (*violet*, Fig. 4, c). They are combined using plain averaging with equal weights. This is done at each scale of the de-reflection, respectively, at each level of the fusion. The second source of information is the background (*green* in Fig. 4, d). Here, only a single, spatial level is considered, *i.e.* that of its output. This encoding is concatenated with the average of the encodings from all reflectance maps on the coarsest level (*i.e.* their spatial resolution matches). Result of this sub-network is the final 64×64 HDR environment map.

Missing pixels Although the input reflectance maps are partial, the output environment map is complete. This is an advantage of the proposed CNN architecture, as in the absence of information the fusion network learns how to

perform sparse data interpolation for the missing pixels, as in [35]. Our learning-based approach naturally embeds priors from the statistics of reflectance and illumination in the training data (*e.g.* the sky is blue and always on top) which is hard to model with physics-based analytic formulas [24, 23].

Training details We use mini-batch (size 4) gradient descent, a log-space learning rate of (-3, -5, 50), a weight decay of .0005, a momentum of .95 and train for 50 iterations.

6. Results

6.1. Quantitative evaluation

We quantify to which extent our approach can infer the environment from an LDR input. Since we are interested in estimating an environment map that is closer to the truthful illumination and does not only have HDR content, we use the perceptualized DSSIM [45] (less is better) as our evaluation metric. This metric captures the structural similarity between images [29, 35, 31, 6], that is of particular importance when the environment’s reflection is visible in a specular surface, such as the ones we target in this paper. The evaluation protocol includes the next steps: The .90-percentile is used to find a reference exposure value for the ground truth HDR environment map. We then apply the same tone-mapper with this authoritative exposure to all HDR alternatives. This way we ensure that we achieve a fair comparison of the different alternatives also w.r.t. HDR.

Model variants and baselines The results of different variants of our approach and selected baseline methods are presented in terms of performance in Table 1 and visual quality in Fig. 5. Below, we describe the different approaches:

- **SINGLET** uses only a single reflectance map, *i.e.* our de-reflection network with $n_{\text{mat}} = 1$, but without background.
- **SINGLET+BG** also uses a single reflectance map, as before, but includes the background network too.
- **BEST-OF-SINGLETS** executes the $n_{\text{mat}} = 1$ de-reflection-plus-background network for each singlet of a triplet individually and then chooses the result closest to the reference by an oracle (we mark all oracle methods in gray).

- NEAREST NEIGHBOR is equivalent - an upper bound - to [18], but as their dataset of illuminations is not publicly available, we pick the nearest neighbor to ground-truth from our training dataset by an oracle so that DSSIM is minimized.¹
- MASK-AWARE MEAN executes $n_{\text{mat}} = 1$ de-reflection-plus-background network for each singlet of a triplet individually and then averages the predicted environment maps based on the sparsity masks of the input reflectance maps.
- TRIPLET combines the information from three reflectance maps via our de-reflection ($n_{\text{mat}} = 3$) and fusion networks, without using background information.
- TRIPLET+BG represents our full model that combines the de-reflection ($n_{\text{mat}} = 3$), fusion and background networks.

Numerical comparison All variants are run on all subsets of our test set: synthetic and real, both single and multi-material, for all objects. Results are summarized in Table 1. Note that, as the table columns refer to different cases, cross-column comparisons are generally not advised, but below we try to interpret the recovered results. For the synthetic cars, we see a consistent improvement by adding background information already for the SINGLET - even outperforming BEST-OF-SINGLETS. Across all experiments, there is consistent improvement from SINGLET to TRIPLET to TRIPLET+BG. TRIPLET+BG has consistently the best results - in particular outperforming the NEAREST NEIGHBOR, which indicates generalization beyond the training set environment maps as well as the hand-crafted fusion scheme MASK-AWARE MEAN. Overall, it is striking that performance for the multi-material case is very strong. This is appealing as it is closer to real scenarios. But it might also be counter-intuitive, as it seems to be the more challenging scenario involving multiple unknown materials with less observed orientations. In order to analyze this, we first observe that for SINGLET, moving from the single to the multi-material scenario does not affect performance much. We conclude that our method is robust to such sparser observation of normals. More interestingly, our best performance in multi-material scenario is only partially explained by exploiting the “easiest” material, which we see from BEST-OF-SINGLETS. The remaining margin to TRIPLET indicates that our model indeed exploits all 3 observations and that they contain complementary information.

Visual comparison Example outcomes of these experiments, are qualitatively shown in Fig. 5. Horizontally, we see that individual reflectance maps can indeed estimate illumination, but contradicting each other and somewhat far from the reference (columns labeled SINGLET in Fig. 5). Adding the BG information can improve color sometimes (columns +BG in Fig. 5). We also see that a nearest neighbor oracle approach (column NN in Fig. 5) does not perform well. Proceeding with triplets (column TRIPLET in Fig. 5)

¹Note that, [18] is the only other published work with the same input (a single LDR image) and output (a HDR non-parametric environment map).

we get closer to the true solution. Further adding the background (OURS in Fig. 5) results in the best prediction. We see that as the difficulty increases from spheres over single- and multi-material to complex shapes, the quality decreases while a plausible illumination is produced in all cases. Most importantly, the illumination can also be predicted from complex, non-car multi-material objects such as the dinosaur geometry as seen in the last row. Supplementary material visualizes all the alternatives for the test dataset.

Varying the number of materials In another line of experiments we look into variation of n_{mat} in Table 2. Here, the number of input reflectance maps increases from 1 up to 5. In each case we include the background and run both on spheres and single-material cars, for which these data are available for $n_{\text{mat}} > 3$. Specifically, we use the real singlets, that we combine into tuples of reflectance maps according to the protocol defined in Sec. 4. We see that, although we have not re-trained our network but rather copy the shared weights that were learned using $n_{\text{mat}} = 3$ materials, our architecture does not only retain efficiency across an increasing number of materials in both cases, but in fact uses the mutual information to produce even an increase in quality. This is in agreement with observations that humans are better in factoring illumination, shape and reflectance from complex aggregates than for simple ones [42].

Further analysis To assess the magnitude of the recovered dynamic range in Fig. 6 a we plot the distribution of luminance over the test dataset (yellow color) and compare it to the distribution of estimated illuminations (red color). Despite the fact that our method operates using only LDR inputs, we observe that in the lower range the graphs overlap (orange color), but in the higher we do not reproduce some brighter values found in the reference. This indicates that our results are both favorable in structure as seen from Table 1 and Table 2 as well as according to more traditional measures such as log L1 or L2 norms [24, 23].

We also evaluate the spatiality of the recovered illumination, *i.e.* dominant light source direction. In 97.5% of the test dataset environment maps, the estimated brightest pixel (dominant light) is less than 1 pixel away from ground-truth, which indicates a fairly accurate prediction.

From the numerical and visual analysis presented above and in supplementary material we can extract useful insights. First, we plot the DSSIM w.r.t. the sparsity of the input reflectance maps on the test set (see Fig. 6 c). We observe that for sparser reflectance maps, the DSSIM becomes higher (the error increases). For sparser reflectance maps the network has more unobserved normals (pixels in the reflectance map) to hallucinate, making inference relatively harder. Second, we study the visual quality of the results w.r.t. material attributes like specularity. Fig. 6 b visualizes from left to right the recovered visual details for an increasing specularity. The visual quality increases as the more specular a

Table 1. DSSIM error (less is better) for different variants (*rows*) when applied to different subsets of our test set (*columns*). The best alternative is shown in **bold**. Oracle analysis using ground-truth information are shown in gray. Variant images are seen in Fig. 5.

	Synthetic			Real		
	Cars (Single)	Cars (Multi)	Spheres	Cars (Single)	Cars (Multi)	Non-cars
SINGLET	.311±.011	.316±.011	.324±.002	.337±.002	.335±.005	.315±.002
SINGLET + BG	.281±.010	.277±.008	.360±.003	.360±.002	.366±.005	.341±.002
BEST-OF-SINGLETS	.304±.011	.307±.011	.314±.001	.330±.002	.324±.004	.312±.004
NEAR. NEIGH.	.277±.009	.277±.009	.360±.002	.360±.002	.332±.007	.313±.004
MASK-AWARE MEAN	.290±.012	.293±.012	.306±.002	.324±.002	.305±.004	.285±.002
TRIPLETS	.268±.011	.277±.011	.313±.001	.332±.002	.284±.002	.288±.001
TRIPLETS + BG	.210±.007	.226±.007	.305±.001	.315±.001	.272±.004	.279±.001



Figure 5. Alternative approaches (*left to right*): **1**) input(s). **2, 4 and 6**) our approach for $n_{\text{mat}} = 1$. **3, 5 and 7**) the same, including a background. **8**) the nearest neighbor approach. **9**) the mask-aware mean approach. **10**) our approach for $n_{\text{mat}} = 3$. **11**) the same, including a background, *i.e.* full approach. **12**) reference. For a quantitative version of this figure see Table 1. For all images see supplementary material.

Table 2. Reconstruction on different number of materials n_{mat} .
Spheres Cars (Single)

	Spheres	Cars (Single)
SINGLET + BG	.360±.003	.360±.002
DOUBLETS + BG	.320±.002	.327±.002
TRIPLETS + BG	.305±.001	.315±.001
QUADRUPLETS + BG	.309±.001	.306±.001
QUINTUPLETS + BG	.292±.001	.295±.001

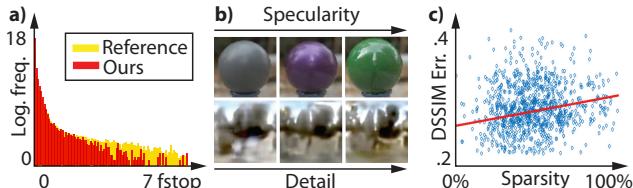


Figure 6. Result analysis: **a**) Predicted vs. ground-truth luminance distribution, **b**) Visual quality w.r.t. material specularity, **c**) DSSIM error w.r.t. reflectance map sparsity.

material is, the more it reveals about the environment.

Comparison with related work It is important to explain why existing approaches are not directly applicable in our case. [24, 23] do not handle multiple materials. One might argue that they can still be applied to the segmented

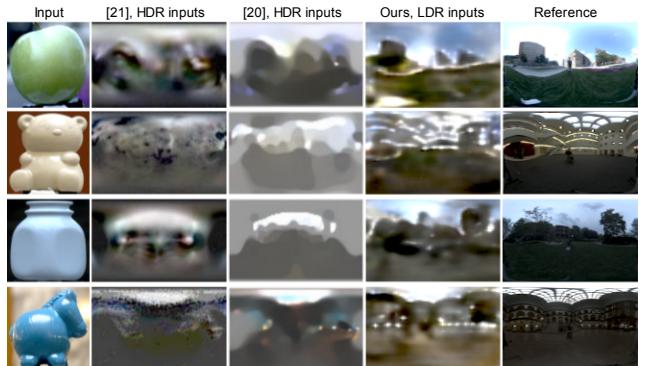


Figure 7. Visual comparison with [24, 23]. Images were taken from the respective papers. [24, 23] use HDR inputs, we use LDR.

subregions, but then it is unclear how to merge the different generated outputs; the papers do not provide a solution. In this case, one would still need techniques like **BEST-OF-SINGLETS** or **MASK-AWARE MEAN** to proceed. Instead, our method naturally fuses the features from the segmented materials and background producing a single output. [25] uses multiple materials but works for point light sources and can not be assumed to extend to natural illumination. [26] works with multiple single-material objects under natural illumina-

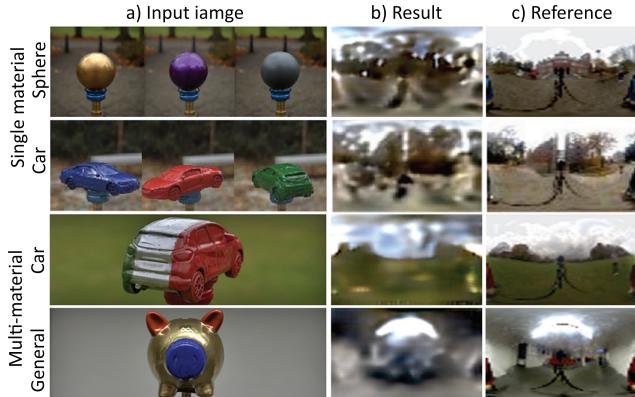


Figure 8. Results for our approach on real objects: spheres, single-material cars, multi-material cars, multi-material non-cars. **a)** input LDR images. **b)** our predicted HDR environment. **c)** the ground-truth. Exhaustive results are found in supplementary material.

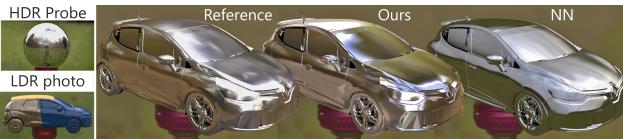


Figure 9. Comparison of re-rendering using the reference, ours, and nearest neighbor for a specular material. Ours is more similar to the reference, while not requiring to acquire an HDR light probe.

tion but requires multiple images as input (≥ 3 according to the authors). Most importantly, all of [25, 24, 23, 26] require HDR images as input (*i.e.* taking multiple pictures under different exposures) making them unworkable for our single-shot LDR input, and do not leverage the image background that is naturally captured in their images anyway.

Nevertheless, we provide an indicative visual comparison between our method and [24, 23] for the dataset of [24] in Fig. 7. While [24, 23] are able to capture the major light components of the scene, our approach recovers detailed structures not only from the lights but also from the surrounding elements (*e.g.* buildings and trees).

6.2. Qualitative results and applications

The visual quality is best assessed from Fig. 8, that shows, from left to right, the input(s) (a), our estimated (b) and ground-truth environment map (c). The difficulty ranges: starting from spheres, we proceed to scenes that combine three single-material objects over single objects with multiple materials to non-car shapes with multiple materials. This shows how non-car shapes at test time can predict illumination, despite training on cars and car parts. In each case, a reasonable estimate of illumination is generated as seen from the two last columns in Fig. 8 and supplementary material.

Relighting To verify the effectiveness in a real relighting application, we show how re-rendering with a new material looks like when illumination is captured using our method vs.



Figure 10. Material/shape edits on web images (see the text below).

a light probe. In the traditional setup (which we also used to acquire the reference for our test data) one encounters multiple exposures, (semi-automatic) image alignment, and a mirror ball with known reflectance and geometry. Instead, we have an unknown object with unknown material and a single LDR image. Note how similar the two rendered results are in Fig. 9. This is only possible when the HDR is also correctly acquired. Instead, a nearest-neighbor oracle approach already performs worse; the reflection alone is plausible, but far from the reference. For more relighting examples see the supplemental video.

Images from the web Ultimately, our method is to be used on images from the web. Fig. 10 shows examples of scenes (top row), on which we run our method and then re-render editing either the material (middle row) or the shape (bottom row). The rendered results look nevertheless convincing. The supplementary material contains more results that indicate the method’s performance on everyday images.

7. Conclusion

We have shown an approach to estimate natural illumination in HDR when observing a shape with multiple, unknown materials captured using an LDR sensor. We phrase the problem as a mapping from reflectance maps to environment maps that can be learned by a suitable novel deep convolution-de-convolution architecture we propose. Training and evaluation are both made feasible thanks to a new dataset combining both synthetic and acquired information. Due to its learning-based nature, we believe that our approach can be possibly extended to compensate for known material segmentation and geometry, and achieve higher fidelity results if trained on a larger “in the wild” dataset.

Acknowledgements This work was supported by the

FWO (G086617N) and the DFG (CRC 1223).

References

- [1] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgbd image. *PAMI*, 2015.
- [2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *PAMI*, 2015.
- [3] H. G. Barrow and J. M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Comp. Vis. Sys.*, 1978.
- [4] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Trans. Graph.*, 2014.
- [5] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.
- [6] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013.
- [7] K. J. Dana, B. Van Ginneken, S. K. Nayar, and J. J. Koenderink. Reflectance and texture of real-world surfaces. *ACM Trans. Graph.*, 1999.
- [8] P. Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. *SIGGRAPH*, 1998.
- [9] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH*, 2008.
- [10] P. E. Debevec, C. J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. *Proc. SIGGRAPH*, 1996.
- [11] R. O. Dror, T. K. Leung, E. H. Adelson, and A. S. Willsky. Statistics of real-world illumination. In *CVPR*, 2001.
- [12] D. Eigen and R. Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV*, 2015.
- [13] S. Georgoulis, K. Rematas, T. Ritschel, M. Fritz, L. Van Gool, and T. Tuytelaars. Delight-net: Decomposing reflectance maps into specular materials and natural illumination. *arXiv preprint arXiv:1603.08240*, 2016.
- [14] S. Georgoulis, V. Vanwellingen, M. Proesmans, and L. Van Gool. A gaussian process latent variable model for brdf inference. In *ICCV*, 2015.
- [15] T. Haber, C. Fuchs, P. Bekaer, H.-P. Seidel, M. Goesele, H. P. Lensch, et al. Relighting objects from image collections. In *CVPR*, 2009.
- [16] J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *CVPR*, 2008.
- [17] B. K. Horn and R. W. Sjoberg. Calculating the reflectance map. *App. Opt.*, 1979.
- [18] K. Karsch, K. Sunkavalli, S. Hadap, N. Carr, H. Jin, R. Fonte, M. Sitig, and D. Forsyth. Automatic scene inference for 3d object compositing. *ACM Trans. Graph.*, 2014.
- [19] E. A. Khan, E. Reinhard, R. W. Fleming, and H. H. Bültmann. Image-based material editing. *ACM Trans. Graph.*, 2006.
- [20] J.-F. Lalonde, A. A. Efros, and S. G. Narasimhan. Estimating the natural illumination conditions from a single outdoor image. *IJCV*, 2012.
- [21] H. Lensch, J. Kautz, M. Goesele, W. Heidrich, and H.-P. Seidel. Image-based reconstruction of spatial appearance and geometric detail. *ACM Trans. Graph.*, 2003.
- [22] 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.
- [23] S. Lombardi and K. Nishino. Reflectance and illumination recovery in the wild. *PAMI*, 2016.
- [24] S. Lombardi and K. Nishino. Reflectance and natural illumination from a single image. In *ECCV*, 2012.
- [25] S. Lombardi and K. Nishino. Single image multimaterial estimation. In *CVPR*, 2012.
- [26] S. Lombardi and K. Nishino. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from RGB-D images. *3DV*, 2016.
- [27] D. Mahajan, R. Ramamoorthi, and B. Curless. A theory of frequency domain invariants: Spherical harmonic identities for brdf/lighting transfer and image consistency. *PAMI*, 2008.
- [28] W. Matusik, H. Pfister, M. Brand, and L. McMillan. A data-driven reflectance model. *ACM Trans. Graph.*, 2003.
- [29] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015.
- [30] K. Nishino and S. K. Nayar. Corneal imaging system: Environment from eyes. *IJCV*, 2006.
- [31] D. Pathak, P. Krähenbühl, S. X. Yu, and T. Darrell. Constrained structured regression with convolutional neural networks. In *arXiv:1511.07497*, 2015.
- [32] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *SIGGRAPH*, 2001.
- [33] K. Rematas, C. Nguyen, T. Ritschel, M. Fritz, and T. Tuytelaars. Novel views of objects from a single image. *TPAMI*, 2017.
- [34] K. Rematas, T. Ritschel, M. Fritz, and T. Tuytelaars. Image-based synthesis and re-synthesis of viewpoints guided by 3d models. In *CVPR*, 2014.
- [35] K. Rematas, T. Ritschel, E. Gavves, M. Fritz, and T. Tuytelaars. Deep reflectance maps. In *CVPR*, 2016.
- [36] S. Richter and S. Roth. Discriminative shape from shading in uncalibrated illumination. In *CVPR*, 2015.
- [37] Right Hemisphere. ZBruhs MatCap, 2015.
- [38] I. Sato, Y. Sato, and K. Ikeuchi. Illumination from shadows. *PAMI*, 2003.
- [39] 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*, 2001.
- [40] C. Thomas and A. Kovashka. Seeing behind the camera: Identifying the authorship of a photograph. In *CVPR*, 2016.
- [41] A. Torralba and W. T. Freeman. Accidental pinhole and pinspeak cameras. *IJCV*, 2014.
- [42] P. Vangorp, J. Laurijssen, and P. Dutré. The influence of shape on the perception of material reflectance. 2007.
- [43] M. Wang, Y.-K. Lai, Y. Liang, R. R. Martin, and S.-M. Hu. Bigger-picture: Data-driven image extrapolation using graph matching. *ACM Trans. Graph.*, 2014.
- [44] X. Wang, D. F. Fouhey, and A. Gupta. Designing deep networks for surface normal estimation. In *CVPR*, 2015.
- [45] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Proc.*, 2004.
- [46] S. H. E. G. Yannick Hold-Geoffroy, Kalyan Sunkavalli and J.-F. Lalonde. Deep outdoor illumination estimation. In *CVPR*, 2017.
- [47] E. Zhang, M. F. Cohen, and B. Curless. Emptying, refurbishing, and relighting indoor spaces. *ACM Trans. Graph.*, 2016.
- [48] Y. Zhang, J. Xiao, J. Hays, and P. Tan. Framebreak: Dramatic image extrapolation by guided shift-maps. *CVPR*, 2013.
- [49] T. Zickler, R. Ramamoorthi, S. Enrique, and P. N. Belhumeur. Reflectance sharing: Predicting appearance from a sparse set of images of a known shape. *PAMI*, 2006.