DOI: 10.1111/cgf.13220

Eurographics Symposium on Rendering 2017
P. Sander and M. Zwicker
(Guest Editors)

Volume 36 (2017), Number 4

Decomposing Single Images for Layered Photo Retouching

Carlo Innamorati Tobias Ritschel Tim Weyrich Niloy J. Mitra

University College London

Figure 1: *Appearance manipulation of a single photograph* (top images) *when using off-the-shelf software like Photoshop directly* (left arrow) *and when using the same in combination with our new layering* (right arrow). *For the car example, the image was decomposed into layers (albedo, irradiance, specular, and ambient occlusion), which were then manipulated individually: specular highlights were strengthened and blurred; irradiance and ambient occlusion were darkened and have added contrast; the albedo color was changed. While the image generated without our decomposition took much more effort (selections, adjustments with curves, and feathered image areas), the result is still inferior. For the statue example, a different decomposition splitting the original image into light directions was used. The light coming from the left was changed to become more blue, while light coming from the right was changed to become more red. A similar effect is hard to achieve in Photoshop even after one order of magnitude more effort. (Please try the edits yourself using the supplementary psd files.)*

Abstract

Photographers routinely compose multiple manipulated photos of the same scene into a single image, producing a fidelity difficult to achieve using any individual photo. Alternately, 3D artists set up rendering systems to produce layered images to isolate individual aspects of the light transport, which are composed into the final result in post-production. Regrettably, these approaches either take considerable time and effort to capture, or remain limited to synthetic scenes. In this paper, we suggest a method to decompose a single image into multiple layers that approximates effects such as shadow, diffuse illumination, albedo, and specular shading. To this end, we extend the idea of intrinsic images along two axes: first, by complementing shading and reflectance with specularity and occlusion, and second, by introducing directional dependence. We do so by training a convolutional neural network (CNN) with synthetic data. Such decompositions can then be manipulated in any off-the-shelf image manipulation software and composited back. We demonstrate the effectiveness of our decomposition on synthetic (i. e., rendered) and real data (i. e., photographs), and use them for photo manipulations, which are otherwise impossible to perform based on single images. We provide comparisons with state-of-the-art methods and also evaluate the quality of our decompositions via a user study measuring the effectiveness of the resultant photo retouching setup. Supplementary material and code are available for research use at `geometry.cs.ucl.ac.uk/projects/2017/layered-retouching`.

1. Introduction

Professional photographers regularly compose multiple photos of the same scene into one image, giving themselves more flexibility and artistic freedom than achievable by capturing a single photo. They do so, by 'decomposing' the scene into individual *layers*, e. g., by changing the scenes physical illumination, manipulating the individual layers (e. g., typically using a software such as Adobe Photoshop), and then composing them into a single image.

A typical manipulation is changing a layer's transparency (or 'weight'): if a layer holds illumination from a specific light direction, this is a direct and easy way to control illumination. Other editing operations include adjustment of hues, blur, sharpening, etc. These operations are applied selectively to some layers, leaving the others unaffected. While the results produced by layered editing could, in principle, also be produced by editing without layers, the separation allows artists, and even novice users, to direct their edits to specific aspects of an image without the need for tediously selecting image regions based on color or shape, resulting in higher efficacy. The key to success is to have a plausible and organized separation into layers available.

Unfortunately, acquiring layers requires either taking multiple photos [CCD03] or to use (layered) rendering [Hec90]. The first option is to capture photos in a studio setup requiring significant setup effort but producing realistic inputs. The other option is to use layered rendering, which is relatively straight-forward and well supported, but the results can be limited in realism.

In this work, we set out to devise a system that combines the strength of both approaches: the ability to directly work on real photos, combined with a separation into layers. Starting from a single photograph, our system produces a decomposition into layers, which can then be individually manipulated and recombined into the desired image using off-the-shelf image manipulation software. Fig. 1 shows two examples, one where specular highlights and albedo were adjusted on the input car image, while on the other directional light-based manipulations were achieved on single input photographs. (Please refer to the supplementary for recorded edit sessions and accompanying PSD files.)

While many decompositions are possible, we suggest a specific layering model that works along two axes: intrinsic features and direction. This is inspired by how many artists as well as practical contemporary rendering systems (e. g., in interactive applications such as computer games) work: first, decomposition into ambient occlusion, diffuse illumination, albedo, and specular shading and second, a decomposition into light directions. Both axes are optional but can be seamlessly combined. Note that this model is not physical. However, it is simple and intuitive for artists and, as we will show, its inverse model is effectively learnable. To invert this model, we employ a deep convolutional neural network (CNN) that is trained using synthetic (rendered) data, for which the ground truth decomposition of a photo into layers is known. While CNNs have recently been used for intrinsic decompositions such as reflectance and shading, we address the novel problem of refined decomposition into ambient occlusion and specular as well as into directions, which is critical for the layered image manipulation workflow. Our contributions are:

1. a workflow, in which a single input photo is automatically decomposed into layered components that are suited for post-capture appearance manipulation within standard image editing software;
2. two plausible appearance decompositions, particularly suited for plausible appearance editing: *i)* advanced intrinsics, including specular and ambient occlusion and *ii)* direction; and
3. a flexible, CNN-based approach to obtain a given type of decomposition, leveraging the state-of-the-art in deep learning.

We evaluate our approach by demonstrating non-trivial appearance edits based on our decompositions and a preliminary user study. We further demonstrate the efficacy of our CNN architecture by applying it to the well-established intrinsic images problem, where it compares favourably to the state-of-the-art methods.

2. Previous Work

Combining multiple photos (also referred to as a "stack" [CCD03]) of a scene where one aspect has changed in each layer is routinely used in computer graphics. For example, NVIDIA IRay actively supports rendered LPE layers (light path expressions [Hec90]) to be individually edited to simplify post-processing towards artistic effects without resorting to solving the inverse rendering problem. One aspect to change is illumination, such as flash-no-flash photography [ED04] or exposure levels [MKVR09]. More advanced effect involve direction of light [ALK*03, RBD06, FAR07], eventually resulting in a more sophisticated user interface [BPB13]. All these approaches require specialized capture to gather multiple images captured by making invasive changes to the scene, limiting their use in practice to change an image post-capture. On-line video and photo communities hold many examples of DIY instructions to setup such studio configurations.

For single images, a more classic approach is to perform intrinsic decomposition into shading (irradiance) and diffuse reflectance (albedo) [BT78, GMLMG12, BBS14], possibly supported by a dedicated UI for images [BPD09, BBPD12], using annotated data [BBS14, ZKE15, ZIKF15], or videos [BST*14, YGL*14]. Recently, CNNs have been successfully applied to this task producing state-of-the-art results [NMY15, SBD15]. For CNNs, a recent idea is to combine estimation of intrinsic properties and depth [SBD15, KPSL16]. We will jointly infer intrinsic properties and normals to allow for a directional illumination decomposition. Also the relation between intrinsic images and filter is receiving considerable attention [BHY15, FWHC17]. We also use a data-driven CNN-based approach to go beyond classic intrinsic image decomposition layers with further separation into occlusion and specular components, as well as directions, that are routinely used in layered image editing (see Sec. 4 and supplementary materials).

In other related efforts, researchers have looked into factorizing components, such as specular [TNI04, MZBK06] from single images, or ambient occlusion (AO) from single [YJL*15] or multiple captures [HWBS13]. We show that our approach can solve this problem at a comparable quality, but requires only a single photo and in combination yields further separation of diffuse shading and albedo without requiring a specialized method.

Despite the advances in recovering reflectance (e. g., with two captures and a stationarity assumption [AWL15], or with dedicated UIs [DTPG11]), illumination (e. g., Lalonde et al. [LEN09] estimate sky environment maps and Rematas et al. [RRF*16] reflectance maps) and CNN-based depth [EPF14] from photographs, no system doing a practical joint decomposition is known. Most relevant to our effort, is SIRFS [BM15] that build data-driven priors for shape, reflectance, illumination, and use them in an optimization setup to recover the most likely shape, reflectance, and illumination under these priors (see Sec. 4 for explicit comparison).

In the context of image manipulations, specialized solutions exist: Oh et al. [OCDD01] represent a scene as a layered collection of color and depth to enable distortion-free copying of parts of a photograph, and allow discounting effect of illumination on uniformly textured areas using bilateral filtering; Khan et al. [KRFB06] enable automatically replacing one material with another (e. g., increase/decrease specularity, transparency, etc.) starting from a single high dynamic range image by exploiting our 'blindness' to certain physical inaccuracies; Caroll et al. [CRA11] achieve consistent manipulation of inter-reflections; or the system of Karsch et al. [KHFH11] that combines many of the above towards compelling image synthesis.

Splitting into light path layers is typical in rendering inspired by the classic light path notation [Hec90]. In this work, different

from Heckbbert's physical $E(S|D)^*L$ formalism, we use a more edit-friendly factorization into ambient occlusion, diffuse light, diffuse albedo, and specular, instead of separating direct and indirect effects. While all the above works on photos, it was acknowledged that rendering beyond the laws of physics can be useful to achieve different artistic goals [TABI07, VPB*09, RTD*10, RLMB*14, DDTP15, SPN*15]. Our approach naturally supports this option, allowing users to freely change layers, using any image-level software of their choice, also beyond what is physically correct. For example, the StyLit system proposed by Fišser et al. [FJL*16] correlates artistic style with light transport expressions, but requires pixels in the image to be labeled with light path information, e. g., by rendering and aligning. Hence, it can take our factorized output to stylize single photographs without being restricted to rendered content.

Figure 2: *The components of our two imaging models. The first row is the intrinsic axis, the second row the directional axis, and the third row shows how one directional element can subsequently be also decomposed into its intrinsics.*

3. Editable Layers From Single Photographs

Our approach has two main parts: an imaging model that describes a decomposition of a single photo into layers for individual editing and a method to perform this decomposition.

Model. The imaging model (Sec. 3.1) is motivated by the requirements of a typical layered workflow (Sec. 3.4): The layers have to be intuitive, they have to be independent, they should only use blend modes available in a (linear) off-the-shelf image editing software and they should be positive and low-dynamic-range (LDR). This motivates a model that can decompose along two axes: intrinsics and directionality. These axes can be combined and use a new directional basis we propose.

Decomposition. The decomposition has two main steps: (i) producing training data (Sec. 3.2) and (ii) a convolutional neural network to decompose single images into editable layers (Sec. 3.3). The training data (Sec. 3.2) is produced by rendering a large number of 3D scenes into image tuples, where the first is the composed image, while the other images are the layers. This step needs to be performed only once and the training data will be made available upon publication. The decomposition (Sec. 3.3) is done using a CNN that consumes a photo and outputs all its layers. This CNN is trained using the (training) data from the previous step. We selected a convolution-deconvolution architecture that is only to be trained once, can be executed efficiently on new input images, and its definition will be made publicly available upon publication.

3.1. Model

We propose an image formation model that can decompose the image along one or two independent axis: intrinsic features or directionality (Fig. 2).

Non-directional model. We model the color C of a pixel as

$$C = O_\mathrm{a}(\rho \cdot E + S), \tag{1}$$

where $O_\mathrm{a} \in [0,1] \in \mathbb{R}$ denotes the *ambient occlusion*, which is the fraction of directions in the upper hemisphere that is blocked from the light; the variable $\rho \in [0,1]^3 \in \mathbb{R}^3$ describes the *diffuse albedo*, i. e., the intrinsic color of the surface itself; the variable $E \in [0,1]^3 \in \mathbb{R}^3$ denotes the *diffuse illumination* (irradiance), i. e., the color of the total light received; and finally, $S \in [0,1]^3 \in \mathbb{R}^3$ is the *specular shading* in units of radiance, where we do not separate between the reflectance and the illumination (see Fig. 2 top row).

This model is a generalization of typical intrinsic images [BKPB17], which only models shading and reflectance, to include specular and occlusion. While in principle occlusion acts differently on diffuse and specular components, we follow Kozlowski and Kautz [KK07], who show that jointly attenuating diffuse and specular reflectance by the same occlusion term is a good approximation under natural lighting, by using O_a as a joint multiplier of $\rho \cdot E$ and S, thus keeping the user-visible reflectance components to a minimum.

In summary, this decomposition produces four layers from each input image that can be combined with simple blending operations in typical image retouching software.

Directional model. The directional model is a generalization of the above. We express pixel color as a generalized intrinsic image as before, but with diffuse illumination depending on the surface normal, and specular shading depending on the reflection vector:

$$C = O_\mathrm{a} \sum_{i=1}^{N} (\rho \cdot b_i(\mathbf{n})E + b_i(\mathbf{r})S)\,, \tag{2}$$

where $b_i \in \mathbb{S}^2 \to \mathbb{R}^3$, $i \in 1 \ldots N$, are basis functions of an N-dimensional lighting basis, parameterized by the surface orientation $\mathbf{n}$, and the reflected orientation $\mathbf{r} \coloneqq 2\langle \mathbf{v}, \mathbf{n} \rangle \mathbf{n} - \mathbf{v}$, respectively. All directions are in view space, so assuming a distant viewer the view direction is $\mathbf{v} \equiv (0,0,1)^\top$ by construction.

Especially for diffuse lighting, a commonly used lighting basis would be first-order spherical harmonics (SH), i. e., $(b_i^{\mathrm{SH}}) = (Y_0^0, Y_1^{-1}, Y_1^0, Y_1^1)$. That basis is shown to capture diffuse reflectance at high accuracy [RH01b]; however, as we aim for a decomposition amenable to be used in traditional photo processing software, which typically quantizes and clamps any layer calculations to $[0,1]$, the negative lobes of SH would be lost when stored in an image layers.

A common positive-only reparameterization would use the six generator functions,

$$\tilde{b}^{\mathrm{SH}}_{1/2} = 1/2 \pm 1/2 Y_1^{-1}\,,$$
$$\tilde{b}^{\mathrm{SH}}_{3/4} = 1/2 \pm 1/2 Y_1^{0}\,,$$
$$\tilde{b}^{\mathrm{SH}}_{5/6} = 1/2 \pm 1/2 Y_1^{1}\,, \tag{3}$$

with $Y_0^0 = \tilde{b}^{\mathrm{SH}}_1 + \tilde{b}^{\mathrm{SH}}_2$, $Y_1^{-1} = \tilde{b}^{\mathrm{SH}}_2 - \tilde{b}^{\mathrm{SH}}_1$, $Y_1^0 = \tilde{b}^{\mathrm{SH}}_4 - \tilde{b}^{\mathrm{SH}}_3$, and $Y_1^1 = \tilde{b}^{\mathrm{SH}}_6 - \tilde{b}^{\mathrm{SH}}_5$. Initial experiments with that lighting basis, however, showed that the necessary blending calculations between the corresponding editing layers lead to excessive quantization in an 8-bit image processing workflow, and even using Photoshop's 16-bit mode mitigated the problem only partially. Moreover, direct editing of the basis function images turned out unintuitive, because

1. the effect of editing pixels corresponding to negative SH contributions is not easily evident to the user;
2. the strong overlap between basis functions makes it difficult to apply desired edits for individual spatial directions only.

This led us to propose a sparser positive-only basis, where spatial directions are mostly decoupled. After experimentation with various bases, we settled on a normalized variant of $\tilde{b}^{\mathrm{SH}}$ as:

$$b_i(\omega) = \tilde{b}^{\mathrm{SH}}_i(\omega)^p \,/\, \sum_{j=1}^{6} \tilde{b}^{\mathrm{SH}}_j(\omega)^p$$
$$= \frac{\big(\langle \omega, \mathbf{c}_i \rangle + 1\big)^p}{\sum_{j=1}^{6} \big(\langle \omega, \mathbf{c}_j \rangle + 1\big)^p}\,, \tag{4}$$

using the dotproduct-based formulation of 1st-order SH, denoting with $\mathbf{c}_i$ the six main spatial directions; the normalization term ensures partition of unity, i. e., $\sum_i^N b_i(\omega) = 1$. Empirically, we found $p = 5$ to offer the best compromise between separation of illumination functions and smoothness. A polar surface plot of the six basis functions overlapping is shown in Figure 3.

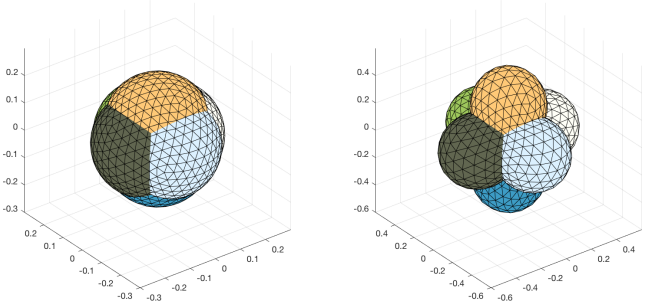

Figure 3: *Directional bases.* Left: $\tilde{b}_i^{SH}$*, a positive-only reparameterization of the 1st-order SH basis exhibits strong overlap between neighboring lobes (drawn as opaque surface plots), and with it strong cross-talk of edits of the associated editing layers;* Right: *our* b_i *(Equation (4)) remains smooth while lobes are separated much more strongly. Note that* $\tilde{b}_i^{SH}$ *has been uniformly rescaled to be partition of unity; the difference in amplitude (see axis labels) further documents the sparser energy distribution in our basis.*

Using this basis, Equation (2) produces 14 layers from an input image – where twelve are directionally-dependent and two are not (ρ and O_a) – that can be combined using any compositing software. As shown in the second and third row of Fig. 2, the 14 output layers can be either collapsed onto 6 directional layers or kept as a combination of both intrinsic and directional decomposition.

Figure 4: *Samples from our set of synthetic training data.*

3.2. Training Data

There are many values of O_a, E, ρ, and S to explain an observed color C, so the decomposition is not unique. In the same way, many normals are possible from a pixel. Inverting this mapping from a single observation is likely to be impossible. At the same time, humans have an intuition how to infer reflectance on familiar objects [KVDCL96]. One explanation can be that they rely on a context $\mathbf{x}$, on the spatial statistics of multiple observations $C(\mathbf{x})$, such that a decomposition into layers becomes possible. In other words, simply not all arrangements of decompositions are equally likely. As described next, we employ a CNN to similarly learn such a decomposition. Training data comprises of synthetic images that show a random shape, with partially random reflectance shaded by random environment map illumination.

Shape. Surface geometry consists of about 2,000 random instances from ShapeNet [C*15] coming from the top-level classes, selected from ShapeNetCore semi-automatically. Specifically, ShapeNetCore has 48 top-level classes among which we use 27. We discarded classes that had either very few models or that were considered uncommon (e. g., birdhouse). We then randomly sampled a tenth of the total models from each class resulting in 1,930 models. These models were also manually filtered to be free of meshing artifacts. Shapes were rendered under random orientation while maintaining the *up* direction intrinsic to each model.

Reflectance. Reflectance using the physically-corrected Phong model [LW94] was sampled as follows: the diffuse colors come directly from ShapeNet models. The specular component k_s is assumed to be a single color. A random decision is made if the material is assumed to be electric or dielectric. If it is electric, we choose the specular color to be the average color of the diffuse texture. Otherwise, we choose it to be a uniform random grey value. Glossiness is set as $n = 3.0^{10\xi}$, where $\xi \in U[0,1]$.

Illumination. Illumination is sampled from a set of 90 high-dynamic-range (HDR) environment maps in resolution 512×256

that have an uncalibrated absolute range of values but are representative for typical lighting settings: indoor, outdoor, as well as studio lights. Illumination were randomly oriented around the vertical axis.

Rendering. After fixing shape, material, and illumination, we synthesize a single image from a random view from a random angle around the vertical axis. To produce C, we compute four individual components, that can be composed into Eq. 1 or further into directions according to Eq. 2 as per-pixel normals are known at render time. Due to the large number of training data required, we use efficient, GPU-based rendering algorithms. The occlusion term O_a is computed using screen-space occlusion [RGS09]. The diffuse shading E is computed using pre-convolved irradiance environment maps [RH01a]. Similarly, specular shading is the product of the specular color k_s selected according to the above protocol, and a pre-convolved illumination map for gloss level n. No indirect illumination or local interactions are rendered.

While this image synthesis is far from being physically accurate, it can be produced easily, systematically and for a very large number of images, making it suitable for learning the layer statistics. Overall we produce 300,000 unique samples in a resolution of 256×256 (ca. 14 GB) in eight hours on a current PC with a higher-end GPU. A fraction of the images totalling to about 30000 were withheld to check for convergence (and detect over-fitting). We also used dropout to prevent over-fitting.

Units. Care has to be taken in what color space learned and training data is to be processed. As the illumination is HDR, the resulting image is an HDR rendering. However, as our input images will be LDR at deployment time, the HDR images need to be tone-mapped to match their range. To this end, automatic exposure control is used to map those values into the LDR range, by selecting the 0.95 luminance percentile of a random subset of the pixels and scale all values such that this value maps to 1. The rendered result C is stored after gamma-correction. All other components are stored in physically linear units ($\gamma = 1.0$) and are processed in physically linear units by the CNN and the end-application using the layers. Doing the final gamma-correction will consequentially be up to the application using the layers later on (as shown in our edit examples).

3.3. Learning a Decomposition

We perform decomposition using a CNN [LBBH98, KSH12] trained using the data produced as described above. Input to the network is a single image such as a photograph. Output for the non-directional variant are five images (occlusion, diffuse illumination, albedo, specular shading, and normals), where occlusion is scalar and the others are three-vector-valued. Note, that normals and intrinsic properties are estimated jointly, such as done before for albedo and depth [SBD15, KPSL16]. The normals are not presented to the user, but only used to perform the directional decomposition.

We have also experimented with letting the CNN directly compute the directional decomposition, but found that having an explicit decomposition using normal and reflected direction to be easier to train and produce better results.

This design follows the convolution-deconvolution idea with cross-links, resulting in a decoder-encoder scheme [RFB15]. The network is fully-convolutional. We start at a resolution of 256×256 that is reduced down to 2×2 through stride-two convolutions. We then perform two stride-one convolutions and increase the number of feature layers in accordance to the required number of output layers (i. e., quadruple for the layers, while the whole step is skipped for normal estimation). The deconvolution part of the network consists of blocks performing a resize-convolution (upsampling followed by a stride-one convolution), cross-linking and a stride-one convolution. Every convolution in the network is followed by a ReLU [NH10] non-linearity except for the last layer, for which a Sigmoid non-linearity is used instead. This is done to normalize the output to the range $[0,1]$. Images with an uneven aspect ratio will be appropriately cropped and/or padded to be square with white pixels. All receptive fields are 3×3 pixels in size except for the first and last two layers that are 5×5. No filter weights are shared between layers. Overall, this network has about 8.5 M trainable parameters.

For the loss function, we combine a per-layer L2 loss with a novel three-fold *recombination* loss, that encourages the network to produce combinations that result in the input image and fulfills the following requirements: (i) the layers have to produce the input, so $C = O_a(E\cdot\rho + S)$; (ii) the components should explain the image without AO, i. e., $C/O_a = E\rho + S$; and (iii) diffuse reflected light should explain the image without AO and specular, so $C/O_a - S = E\rho$. Note that if the network was able to always perform a perfect decomposition, a single L2 loss alone would be sufficient. As it makes errors in practice, the additional loss expressions bias those errors to at least happen in such a way that the combined result does not deviate as much from the input. All losses are in the same RGB-difference range and are weighted equally.

In Tbl. 1, we numerically evaluate the recombination error (i.e., the differences between the original and recombined images) by progressively adding each of the three additional losses to a standard L2 loss. While a positive trend can be observed with the DSSIM metric, these benefits are not as evident on the NRMSE metric.

Overall, the network is a rather standard modern design, but trained to solve a novel task (layers) on novel kind of training data (synthesized, directionally-dependant information). We used TensorFlow [A*15] for our implementation platform and each model requires only several hours on a NVIDIA Titan X GPU with 12 GB on-board RAM to train (both have been trained for 12 hours). We used stochastic gradient descent to solve for the network, which we ran for 6 epochs with batches of size 16. A more detailed description of the network's architecture can be found in the supplementary materials.

Table 1: *Comparing different steps of our recombination loss* (rows) *in terms of two metrics* (columns)*: DSSIM and NRMSE on our validation set.*

Loss	NRMSE	DSSIM
L2	0.2549 ± 0.0807	0.0234 ± 0.0129
L2 + (iii)	0.2598 ± 0.0833	0.0229 ± 0.0126
L2 + (iii) + (ii)	0.2588 ± 0.0799	0.0229 ± 0.0122
L2 + (iii) + (ii) + (i)	**0.2460** ± **0.0787**	**0.0210** ± **0.0119**

Figure 5: *Decomposition of input images into light layers. Please see "Decomposition" in Sec. 4.*

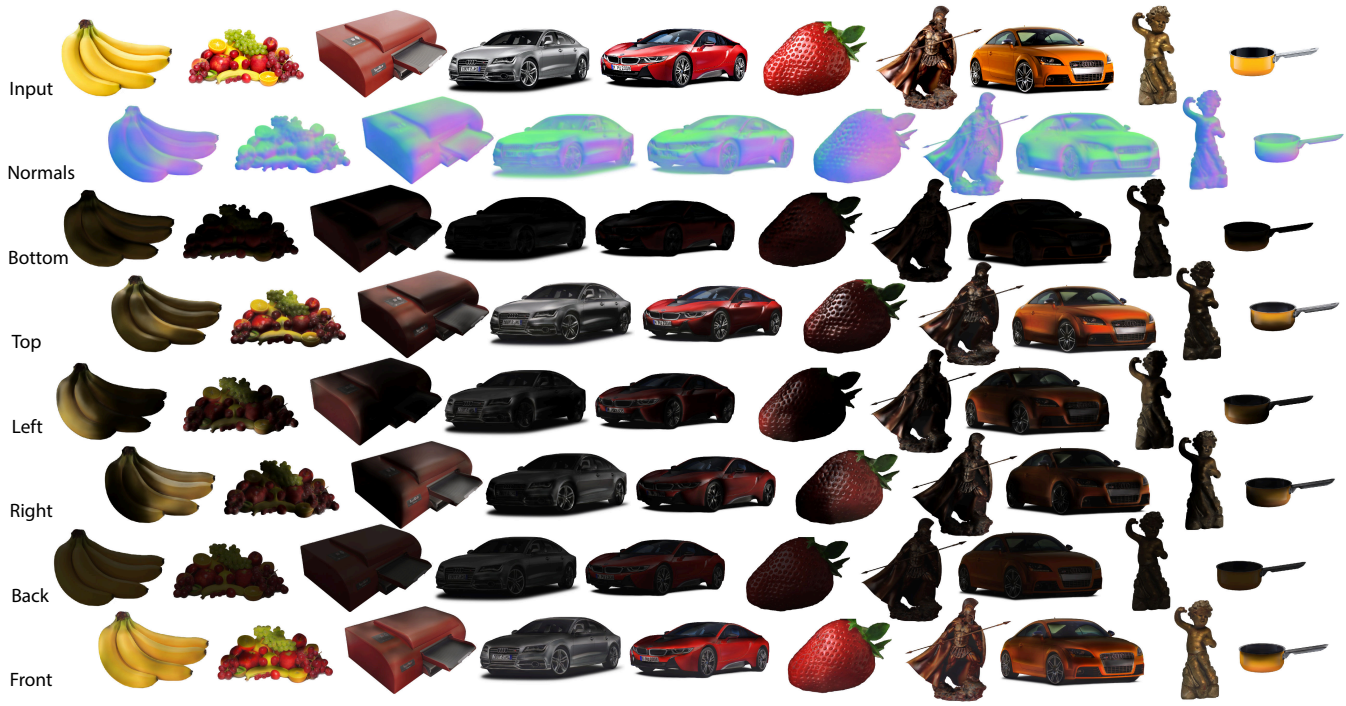

Figure 6: *Decomposition of input images into the six directional layers for different objects. Please see "Decomposition" in Sec. 4.*

3.4. Composition

For composition any arbitrary software that can handle layering, such as GIMP, Adobe Photoshop, or Blender, can be used. Our decomposition is so simple that it can be implemented using a basic set of Photoshop layers of the appropriate additive and multiplicative blend modes, followed by a final gamma mapping. The whole setup process can easily be automated through the use of Photoshop macros. The content is then ready to be manipulated with existing tools with WYSIWYG feedback (please refer to the supplementary for example PSD files).

The artist is free to make any local edits to any of the layers, which usually leads to user-predictable results, owing to the successful decoupling of appearance contributors. Note that we do not limit the manipulation to producing a composition that is physically valid, because this would be unduly limiting artistic expression at this part of the pipeline [TABI07, RGS09, SPN*15].

Figure 7: *$O_a \cdot (\rho \cdot E + S)$ editing.* RED CAR*: specular highlights were strengthened and blurred, while hue was adjusted.* TOY ROBOT*: specular highlights were blurred, while preserving high frequency details.* SOLDIER STATUE*: specular highlights were reduced, and brightness of albedo enhanced. Dark regions were also made more evident. Intuitively, in the single channel edit, the tasks clash against each other.* ORANGE CAR*: specular highlights were boosted and parts of the car directly lit from light sources were emphasized.* BOY STATUE*: the appearance of the statue was adjusted to simulate marble and copper, respectively.* FRUITS*: specular highlights were boosted and given a blue tint. Dark regions were emphasized.*

Figure 8: *Directional editing.* GRAY CAR*: light coming from the right was emphasized.* WOLVERINE STATUE*: blue light sources on the right and left sides of the statue were simulated. Differences with single channel edits are best seen near the tights of the statue.* RED CAR*: light coming mostly from the right was simulated.* TEA SET*: this example uses our 14-way edit. The specular color of the light was changed based on directionality (blue from the left, red from the right). Note how the lidless teapot is fully blue in the single-channel edit, as opposed to the edit using our decomposition, where red can be seen on its right side.* FRUITS*: light coming from the bottom was dimmed and light coming from the top was emphasized.*

4. Results

We report results in form of typical decompositions on images, edits enabled by this decomposition, numeric evaluation, and a preliminary user study. The full material with many more decompositions, high-resolution images, videos, and user study material are found in our supplementary at `geometry.cs.ucl.ac.uk/projects/2017/layered-retouching`.

Note that a segmentation of the object is always provided. Given our context (product photography), where such a segmentation is part of the standard workflow, we assume a user can easily provide a mask, e.g., using GrabCut. In general, the feasibility of automatically obtaining a mask is clear for product images using a monochromatic background.

Decompositions. How well our network performs is best seen when applying it to real images. Naturally, we do not know the reference layer-decomposition or directional decomposition, so their quality can only be judged qualitatively.

Representative results of decomposing images into appearance layers are shown in Fig. 5. The real-world photographs shown represent a mix of natural and man-made objects, whose surface appearance ranges from mostly diffuse, to sheen, to mirror-like gloss, featuring both uniform and structured diffuse albedo.

Overall, the decompositions turn out plausible and provide a clear separation of independent aspects of appearance. That plausibility is critical for successful editing sessions. Existing failure cases are subtle: the tea set exhibits slight cross-talk from diffuse albedo into the specular channel where the porcelain's color blends into pale green; equally, the subtle brown spots on the pear do not end up in the albedo map but show as attenuations in the specular channel; the mirror reflection of a table in the bowl of the food processor (far right), which does not conform with our image formation model, gets evenly distributed across all channels. Interestingly, very dark areas, where an inherent ambiguity between low irradiance, low

diffuse albedo, and high ambient occlusion exists, generally get attributed to a combination of low albedo and irradiance, while ambient occlusion rather credibly remains limited to darkening around strong apparent geometric features. Most importantly, however, we observe that these types of failures are graceful enough to still allow for meaningful edits on a layer-basis.

Directional decompositions are shown in Fig. 6. Their quality largely depends on the quality of the normal map inferred from the input images. We again show a range of materials and objects, and most normal maps look plausible on first sight. Closer scrutiny reveals instances where color or shading variations affect the normal reconstruction, and little can be said about quantitative accuracy of the normals. Nevertheless, given the ill-posedness of the task, we argue that the normal quality is remarkable. This shows in largely plausible directional lighting separations for each object. As we will show below, this still proves sufficient for effective and intuitive editing.

Edits. Typical edits are shown in Fig. 7 and the directional variant in Fig. 8. Note that we support both global manipulations, such as changing the weight of all values in a layer, and local manipulations, such as blurring the highlights or albedo individually. We show a range of edit examples achieved with and without our decomposition. All edits were performed using Photoshop. Generally the effects were much easier to obtain using our decomposition.

Classic intrinsic images [BKPB17] assume S to be zero (no specular) and combine our terms O_a and D, the occlusion and the diffuse illumination, into a single "shading" term that is separated from the reflectance ρ, i. e.,

$$C = O_a(\rho \cdot E + S) \approx O_a(\rho \cdot E + 0) = \underbrace{O_a \cdot E}_{\text{Shading } E'} \cdot \rho. \qquad (5)$$

Similarly, specular separation [TNI04, ABC11], does not identify occlusion and separates into a diffuse term D and a specular term S':

$$C = O_a(\rho \cdot E + S) \approx \underbrace{O_a \cdot \rho \cdot E}_{\text{Diffuse } D} + \underbrace{O_a \cdot S}_{\text{Specular } S'}. \qquad (6)$$

A comparison of our decomposition and typical approaches to generate intrinsic images is shown in Fig. 9. In Table 4 we compare against the same techniques but on our validation dataset. This test dataset was generated using randomly selected ShapeNet models as described before and was not available to our network at training time. We see, that for product images, our approach can perform a better separation into components than other published methods.

At $N = 100$ the std. error of the mean in SSIM units is in the order of, e. g., $0.35/\sqrt{100} = 0.03$, that is, quite low, indicating that even the 100 samples in our test dataset give a good estimate of a methods effect, while still being able to compute the outcome with slower methods.

User study. We have investigated how much the layered representation we suggest can facilitate appearance editing compared to a non-layered representation. To this end, we have given subjects ($N_s = 8$) tasks ($N_t = 2$) they had to complete in Photoshop. Subjects reported had a general technical knowledge of image manipulation

Table 2: *Comparing different methods* (rows) *in terms of two metrics* (columns)*: DSSIM and NRMSE on our validation set.*

Layer	Method	NRMSE	DSSIM
Albedo ρ	IIW [BBS14]	0.484 ± 0.397	0.066 ± 0.040
	SIRFS [BM15]	0.805 ± 0.819	0.066 ± 0.036
	DI [NMY15]	0.598 ± 0.290	0.076 ± 0.038
	Ours	**0.322** ± **0.273**	**0.061** ± **0.035**
Shading E'	IIW [BBS14]	0.298 ± 0.162	0.053 ± 0.034
	SIRFS [BM15]	0.302 ± 0.132	0.057 ± 0.037
	DI [NMY15]	0.389 ± 0.134	0.065 ± 0.036
	Ours	**0.167** ± **0.079**	**0.045** ± **0.033**
Specular S	SRC [TI05]	2.100 ± 1.504	0.084 ± 0.039
	Ours	**0.535** ± **0.299**	**0.060** ± **0.032**
Occlusion O_a	Ours	**0.098** ± **0.034**	**0.050** ± **0.031**
Irradiance E	Ours	**0.158** ± **0.082**	**0.057** ± **0.036**

in Photoshop. Tasks were given in written form. In one condition, the images were split into layers, in the other they were not. The order in which the tasks were performed was randomized. Subjects were asked to perform the ($N_s \times N_t = 16$) trials within 3 minutes and to indicate when they were satisfied with the results. The first task was to "strengthen all the specular highlights" of a car and to "blur the specular highlights across the object to simulate a material with lower glossiness" (Fig. 10-bottom). The second task, operated on a soldier statue instead, was to "brighten the colours of the image but reduce the strength of the specular highlights" (Fig. 10-top).

Two resulting variables were recorded. First, the time until a user would say to be satisfied with the result. Second, the percentage of how often other, external subjects would consider one condition to produce results that achieve the goal more successfully than the other. To this end, random pairs of results produced for one task with and without the condition were shown to Amazon Mechanical (AMT) turk users together with the initial image and the textual description of the goal, asking them in a 2-alternative forced choice (2AFC) "which image achieves the textual goal better" ($N_m = 221$ unique workers). Preference was computed as the AMT confidence-weighted mean, where confidence is the probability of giving the same answer when being asked the same question. Each vote was weighted by the number of consistent answers (same answer to same questions).

The hypothesis is, that the condition has an effect on the outcome. We find this to be the case with statistical significance, that the layered (our) time is 65 seconds (less is better) while it is 151 seconds for no layers ($p < .0001$ paired dependent t-test) and that the layered (our) preference is 70.0 % (more is better) ($p < .0001$ binomial test). While the set of tasks and images is limited and no generally accepted benchmark for layered editing is yet available, this is a first indication that automatically generated layers, such as we suggest, can lead to fast and accurate appearance editing.

Finally, the reader is encouraged to refer to the supplementary material to get a qualitative impression of typical edits produced with and without layered representations. It contains results for the

Figure 9: *Comparison of our approach to three different reflectance and shading estimation techniques SIRFS [BM15], IIW [BBS14], and DI [NMY15]. We run their method on real images and compare their results to ours.*

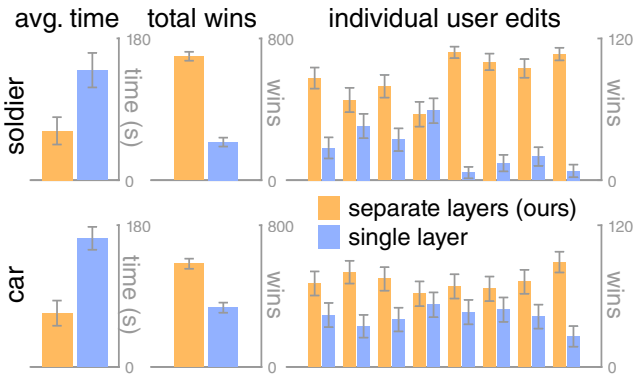

Figure 10: *User study finding. The first row is the first, the bottom row the second task. The right plots shows average time for completion in seconds (less is better, annotated are 0.95 confidence intervals) for both methods. The middle plot shows average preference ratings (more is better). The vertical bar for preference is in units of consistency-weighted votes (see text). The right shows the preference for results produced by each individual user.*

complete set of all inputs we tested; due to space restrictions at 256×256 resolution only. Additionally, we include all the edits results from our user study and the associated original PSD files on which they were performed.

Limitations. Like in many CNN based learning approaches, the shortcoming of our two networks are hard to pin down (for unseen data). We have trained on several different classes of objects, which we found to generalize well to other classes. Images that include phenomena not part of the training, i. e., illumination (individual point lights instead of environment maps), materials (transparency, scattering, anisotropy) or shape (hair, plants) or light transport (soft shadows, indirect light, caustic) remain to be investigated in future work. Also, fundamental limitations of layered image editing remain, such as the inability to completely change the light, change the material or add cast shadows.

5. Conclusion

We have suggested the first decomposition of general images into edit-friendly layers, that were previously only possible either on synthetic images, or when capturing multiple images and manipulating the scene. We have shown that overcoming these limitations allows producing high-quality images, but it also saves capture time and removes the limitation to renderings. Future work could investigate other decompositions such as global and direct illumination, subsurface-scattering or directional illumination or other inputs, such as videos.

6. Acknowledgements

We thank our reviewers for their detailed and insightful comments and the user study participants for their time and feedback. We also thank Paul Guerrero, James Hennessey, Moos Hueting, Aron Monszpart and Tuanfeng Yang Wang for their help, comments and ideas. This work was partially funded by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 642841, by the ERC Starting Grant SmartGeometry (StG-2013-335373), and by the UK Engineering and Physical Sciences Research Council (grant EP/K023578/1).

References

[A*15] ABADI M., ET AL.: TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[ABC11] ARTUSI A., BANTERLE F., CHETVERIKOV D.: A survey of specularity removal methods. In *Comp. Graph. Forum* (2011), vol. 30, pp. 2208–30. 8

[ALK*03] AKERS D., LOSASSO F., KLINGNER J., AGRAWALA M., RICK J., HANRAHAN P.: Conveying shape and features with image-based relighting. In *Proc. IEEE VIS* (2003). 2

[AWL15] AITTALA M., WEYRICH T., LEHTINEN J.: Two-shot SVBRDF capture for stationary materials. *ACM Trans. Graph. (Proc. SIGGRAPH) 34*, 4 (2015). 2

[BBPD12] BOYADZHIEV I., BALA K., PARIS S., DURAND F.: User-guided white balance for mixed lighting conditions. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 31*, 6 (2012). 2

[BBS14] BELL S., BALA K., SNAVELY N.: Intrinsic images in the wild. *ACM Trans. Graph. (Proc. SIGGRAPH) 33*, 4 (2014), 159. 2, 8, 9

[BHY15] BI S., HAN X., YU Y.: An l 1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM Trans. Graph. (Proc. SIGGRAPH) 34*, 4 (2015), 78. 2

[BKPB17] BONNEEL N., KOVACS B., PARIS S., BALA K.: Intrinsic decompositions for image editing. *Comp. Graph. Forum (Eurographics State of the Art Reports) 36*, 2 (2017). 3, 8

[BM15] BARRON J. T., MALIK J.: Shape, illumination, and reflectance from shading. *IEEE Tr. Pat. An. & Mach. Intel. (PAMI)* (2015). 2, 8, 9

[BPB13] BOYADZHIEV I., PARIS S., BALA K.: User-assisted image compositing for photographic lighting. *ACM Trans. Graph. (Proc. SIGGRAPH) 32*, 4 (2013). 2

[BPD09] BOUSSEAU A., PARIS S., DURAND F.: User-assisted intrinsic images. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 28*, 5 (2009). 2

[BST*14] BONNEEL N., SUNKAVALLI K., TOMPKIN J., SUN D., PARIS S., PFISTER H.: Interactive intrinsic video editing. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 33*, 6 (2014). 2

[BT78] BARROW H., TENENBAUM J.: Recovering intrinsic scene characteristics. *Comput. Vis. Syst.* (1978). 2

[C*15] CHANG A. X., ET AL.: Shapenet: An information-rich 3d model repository. *CoRR abs/1512.03012* (2015). 4

[CCD03] COHEN M. F., COLBURN A., DRUCKER S.: *Image stacks*. Tech. Rep. MSR-TR-2003-40, Microsoft Research, July 2003. 2

[CRA11] CARROLL R., RAMAMOORTHI R., AGRAWALA M.: Illumination decomposition for material recoloring with consistent interreflections. *ACM Trans. Graph. (Proc. SIGGRAPH) 30*, 4 (2011). 2

[DDTP15] DONG B., DONG Y., TONG X., PEERS P.: Measurement-based editing of diffuse albedo with consistent interreflections. *ACM Trans. Graph. (Proc. SIGGRAPH) 34*, 4 (2015). 3

[DTPG11] DONG Y., TONG X., PELLACINI F., GUO B.: AppGen: interactive material modeling from a single image. *ACM Trans. Graph. (Proc. SIGGRAPH Asia) 30*, 6 (2011). 2

[ED04] EISEMANN E., DURAND F.: Flash photography enhancement via intrinsic relighting. *ACM Trans. Graph. (Proc. SIGGRAPH) 23*, 3 (2004). 2

[EPF14] EIGEN D., PUHRSCH C., FERGUS R.: Depth map prediction from a single image using a multi-scale deep network. In *Proc. Neur. Inf. Proc. Sys. (NIPS)* (2014). 2

[FAR07] FATTAL R., AGRAWALA M., RUSINKIEWICZ S.: Multiscale shape and detail enhancement from multi-light image collections. *ACM Trans. Graph. (Proc. SIGGRAPH) 26*, 3 (2007). 2

[FJL*16] FIŠER J., JAMRIŠKA O., LUKÁČ M., SHECHTMAN E., ASENTE P., LU J., SÝKORA D.: StyLit: Illumination-guided example-based stylization of 3D renderings. *ACM Trans. Graph. (Proc. SIGGRAPH) 35*, 4 (2016). 3

[FWHC17] FAN Q., WIPF D. P., HUA G., CHEN B.: Revisiting deep image smoothing and intrinsic image decomposition. *CoRR abs/1701.02965* (2017). 2

[GMLMG12] GARCES E., MUNOZ A., LOPEZ-MORENO J., GUTIERREZ D.: Intrinsic images by clustering. *Comp. Graph. Forum (Proc. Eurogr. Symp. Rendering) 31*, 4 (2012). 2

[Hec90] HECKBERT P. S.: Adaptive radiosity textures for bidirectional ray tracing. *ACM SIGGRAPH Computer Graphics 24*, 4 (1990). 2

[HWBS13] HAUAGGE D., WEHRWEIN S., BALA K., SNAVELY N.: Photometric ambient occlusion. In *Proc. IEEE Conf. Comp. Vision & Pat. Rec. (CVPR)* (2013). 2

[KHFH11] KARSCH K., HEDAU V., FORSYTH D., HOIEM D.: Rendering synthetic objects into legacy photographs. In *ACM Trans. Graph. (Proc. SIGGRAPH Asia)* (2011), vol. 30. 2

[KK07] KOZLOWSKI O., KAUTZ J.: Is accurate occlusion of glossy reflections necessary? In *Proc. Appl. Percept. in Gr. & Vis. (APGV)* (2007), pp. 91–98. 3

[KPSL16] KIM S., PARK K., SOHN K., LIN S.: Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *Proc. Eur. Conf. Comp. Vision (ECCV)* (2016), pp. 143–59. 2, 5

[KRFB06] KHAN E. A., REINHARD E., FLEMING R. W., BÜLTHOFF H. H.: Image-based material editing. *ACM Trans. Graph. (Proc. SIGGRAPH) 25*, 3 (2006). 2

[KSH12] KRIZHEVSKY A., SUTSKEVER I., HINTON G. E.: Imagenet classification with deep convolutional neural networks. In *Proc. Neur. Inf. Proc. Sys. (NIPS)* (2012). 5

[KVDCL96] KOENDERINK J., VAN DOORN A., CHRISTOU C., LAPPIN J.: Perturbation study of shading in pictures. *Perception 25*, 1009-26 (1996). 4

[LBBH98] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based learning applied to document recognition. *Proc. IEEE 86*, 11 (1998), 2278–324. 5

[LEN09] LALONDE J.-F., EFROS A. A., NARASIMHAN S. G.: Estimating natural illumination from a single outdoor image. In *Proc. IEEE Intl. Conf. on Comp. Vision (ICCV)* (2009). 2

[LW94] LAFORTUNE E. P., WILLEMS Y. D.: *Using the modified phong reflectance model for physically based rendering*. Tech. Rep. CW 197, Dept. Computerwetenschappen, KU Leuven, Nov. 1994. 4

[MKVR09] MERTENS T., KAUTZ J., VAN REETH F.: Exposure fusion: A simple and practical alternative to high dynamic range photography. *Comp. Graph. Forum (Proc. Pacific Graphics) 28*, 1 (2009). 2

[MZBK06] MALLICK S. P., ZICKLER T., BELHUMEUR P. N., KRIEGMAN D. J.: Specularity removal in images and videos: A PDE approach. In *Proc. Eur. Conf. Comp. Vision (ECCV)* (2006). 2

[NH10] NAIR V., HINTON G. E.: Rectified linear units improve restricted boltzmann machines. In *Proc. Intl. Conf. Mach. Learn. (ICML)* (2010), pp. 807–14. 5

[NMY15] NARIHIRA T., MAIRE M., YU S. X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proc. IEEE Intl. Conf. on Comp. Vision (ICCV)* (2015). 2, 8, 9

[OCDD01] OH B. M., CHEN M., DORSEY J., DURAND F.: Image-based modeling and photo editing. In *Proc. SIGGRAPH* (2001). 2

[RBD06] RUSINKIEWICZ S., BURNS M., DECARLO D.: Exaggerated shading for depicting shape and detail. *ACM Trans. Graph. (Proc. SIGGRAPH) 25*, 3 (2006). 2

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *Proc. Med. Image Comp. and Comp.-Assisted Int.* (2015). 5

[RGS09] RITSCHEL T., GROSCH T., SEIDEL H.-P.: Approximating dynamic global illumination in image space. In *ACM SIGGRAPH Symp. Interact. 3D Gr. & Games (i3D)* (Feb. 2009). 5, 6

[RH01a] RAMAMOORTHI R., HANRAHAN P.: An efficient representation for irradiance environment maps. In *Proc. SIGGRAPH* (2001). 5

[RH01b] RAMAMOORTHI R., HANRAHAN P.: A signal-processing framework for inverse rendering. In *Proc. SIGGRAPH* (New York, NY, USA, 2001), ACM, pp. 117–128. 3

[RLMB*14] RICHARDT C., LOPEZ-MORENO J., BOUSSEAU A., AGRAWALA M., DRETTAKIS G.: Vectorising bitmaps into semi-transparent gradient layers. *Comp. Graph. Forum (Proc. Eurogr. Symp. Rendering) 33*, 4 (2014), 11–19. 3

[RRF*16] REMATAS K., RITSCHEL T., FRITZ M., GAVVES E., TUYTELAARS T.: Deep reflectance maps. In *Proc. IEEE Conf. Comp. Vision & Pat. Rec. (CVPR)* (2016). 2

[RTD*10] RITSCHEL T., THORMÄHLEN T., DACHSBACHER C., KAUTZ J., SEIDEL H.-P.: Interactive on-surface signal deformation. In *ACM Trans. Graph. (Proc. SIGGRAPH)* (2010), vol. 29. 3

[SBD15] SHELHAMER E., BARRON J. T., DARRELL T.: Scene intrinsics and depth from a single image. In *CVPR Workshops* (2015), pp. 37–44. 2, 5

[SPN*15] SCHMIDT T.-W., PELLACINI F., NOWROUZEZAHRAI D., JAROSZ W., DACHSBACHER C.: State of the art in artistic editing of appearance, lighting and material. In *Comp. Graph. Forum* (2015). 3, 6

[TABI07] TODO H., ANJYO K.-I., BAXTER W., IGARASHI T.: Locally controllable stylized shading. *ACM Trans. Graph. (Proc. SIGGRAPH) 26*, 3 (2007), 17. 3, 6

[TI05] TAN R. T., IKEUCHI K.: Separating reflection components of textured surfaces using a single image. *IEEE Tr. Pat. An. & Mach. Intel. (PAMI) 27*, 2 (2005). 8

[TNI04] TAN R. T., NISHINO K., IKEUCHI K.: Separating reflection components based on chromaticity and noise analysis. *IEEE Tr. Pat. An. & Mach. Intel. (PAMI) 26*, 10 (2004). 2, 8

[VPB*09] VERGNE R., PACANOWSKI R., BARLA P., GRANIER X., SCHLICK C.: Light warping for enhanced surface depiction. In *ACM Trans. Graph. (Proc. SIGGRAPH)* (2009), vol. 28, ACM. 3

[YGL*14] YE G., GARCES E., LIU Y., DAI Q., GUTIERREZ D.: Intrinsic video and applications. *ACM Trans. Graph. (Proc. SIGGRAPH) 33*, 4 (2014), 80. 2

[YJL*15] YANG W., JI Y., LIN H., YANG Y., BING KANG S., YU J.: Ambient occlusion via compressive visibility estimation. In *Proc. IEEE Conf. Comp. Vision & Pat. Rec. (CVPR)* (2015). 2

[ZIKF15] ZORAN D., ISOLA P., KRISHNAN D., FREEMAN W. T.: Learning ordinal relationships for mid-level vision. In *Proc. IEEE Conf. Comp. Vision & Pat. Rec. (CVPR)* (2015), pp. 388–96. 2

[ZKE15] ZHOU T., KRAHENBUHL P., EFROS A. A.: Learning data-driven reflectance priors for intrinsic image decomposition. In *Proc. IEEE Conf. Comp. Vision & Pat. Rec. (CVPR)* (2015), pp. 3469–3477. 2