# A Comprehensive Multi-Illuminant Dataset for Benchmarking of Intrinsic Image Algorithms

Shida Beigpour, Andreas Kolb
Chair of Computer Graphics and Multimedia Systems, University of Siegen
Hölderlinstrae 3, 57076 Siegen, Germany
{ shida.beigpour, andreas.kolb} @uni-siegen.de

Sven Kunz
sven.kunz83@gmail.com

## Abstract

*In this paper, we provide a new, real photo dataset with precise ground-truth for intrinsic image research. Prior ground-truth datasets have been restricted to rather simple illumination conditions and scene geometries, or have been enhanced using image synthesis methods. The dataset provided in this paper is based on complex multi-illuminant scenarios under multi-colored illumination conditions and challenging cast shadows. We provide full per-pixel intrinsic ground-truth data for these scenarios,* i.e. *reflectance, specularity, shading, and illumination for scenes as well as preliminary depth information. Furthermore, we evaluate 3 state-of-the-art intrinsic image recovery methods, using our dataset.*

## 1. Introduction

One of the main goals in the image processing research, from the early days until now, is the ability to describe the scene in terms of "intrinsic" characteristics like depth, shape, surface orientation (normals), incident light, and reflectance at each visible surface point [4]. Each of these intrinsic characteristics provides us with valuable cues for scene understanding. Several methods have been proposed which tackle different aspects of this problem in an image-based manner, i.e. utilizing image-based scene representations with a fixed camera pose. While color constancy methods are more interested in the chromatic value of the incident light on the object surface, intrinsic image estimation methods try to separate the effects of the lighting and scene geometry from the object's reflectance, *i.e.* its color and texture.

A major difficulty for intrinsic image research is getting access to datasets of images which depict realistic scenarios, and provide precise ground-truth information regarding all intrinsic scene parameters, i.e. reflectance, specularity, shading, and illumination. This kind of datasets allows the evaluation of the performance of any intrinsic image method, as well as providing training data for these algorithms. The first reliable dataset is published by Grosse *et al*. [15], known as the MIT dataset. While the MIT dataset has paved the way for many recent intrinsic image methods, it suffers from limitations like: simple lighting condition (a single white light source with no ambient lighting, no colored interreflections, no colored lighting edges, no colored shadows, and very little specularities), simple toy objects with limited color and texture, and no geometry data.

Recently Barron *et al*. [2] have synthetically rendered the MIT images to provide more complex lighting. In their later work [3], they have also featured synthesized range (depth) data for the MIT dataset with synthesized Kinect-like noise, and shown the value of such complimentary information about the scene in improving the performance of intrinsic image recovery. Furthermore, they have provided results on the NYU dataset [21] for which range data, but no real intrinsic image ground-truth is available.

In the current work, we provide a new, real-photo ground-truth dataset applicable to intrinsic image research, which strongly improves over the existing ones in terms of illumination and scene complexity, and which includes full intrinsic ground-truth data. Inspired by Grosse *et al*. [15] and Beigpour *et al*. [6], our approaches to scene setup, data acquisition, and ground-truth determination comprise the following contributions, addressing new challenges in the field which were previously not considered:

- Creation of the first reliable multi-illuminant real-photo dataset for the purpose of intrinsic image research presenting complex geometry, multi-colored non-uniform lighting, large specularities, and chal-

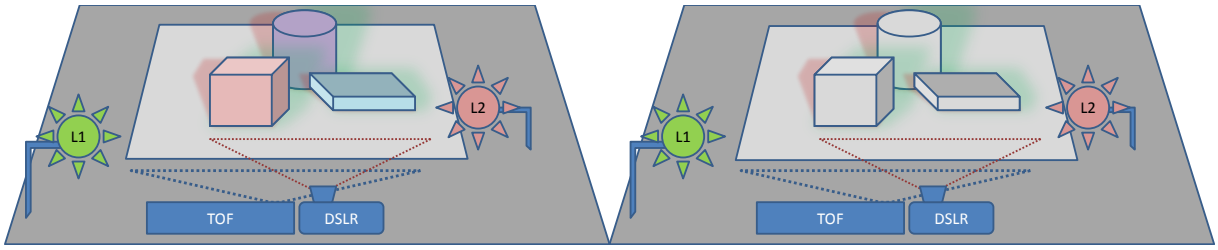

Figure 1. Schematic example of the setup: Scene with colored objects (left) and geometrically identical with objects painted in gray (right).

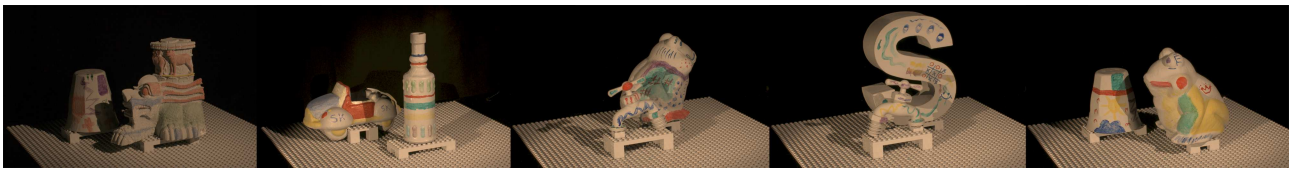
Figure 2. The scenes: We have designed five scenes with complex geometry and shading effects. Here we show the colored scene versions.

lenging colored shadows.

- Precise per-pixel ground-truth of reflectance, specularity, shading, and illumination for scenes under multi-colored illumination conditions.

- Preliminary depth information of the scene on a coarse resolution level.

- Evaluation of 3 state-of-the-art intrinsic image recovery method using our dataset.

The main contributions of the proposed dataset over the MIT dataset are that it features: colored light, non-uniform multi-colored illumination (e.g., challenging colored cast shadows, color variations of the incident light dependent on the surface normal) with pixel-wise accurate ground-truth, large specularities, a preliminary depth map, as well as adding more complexity to the scene geometry (e.g., scenes contain multiple objects with very different geometries). Thus, our dataset enables advanced evaluations of existing and future intrinsic image algorithms.

The reminder of this paper is organized as follows: Section 2 summarizes the related work; Section 3 provides details on our dataset acquisition and ground-truth calculation; a benchmarking of state-of-the-art using our dataset is given in Section 4; Section 5 discusses challenges and future work followed by conclusions in Section 6.

## 2. Related work

In this section we briefly explain the image formation model, which is fundamental for the intrinsic images concept (Sec. 2.1). Afterwards, we summarize the related work in intrinsic image dataset creation (Sec. 2.2) as well as the intrinsic image recovery methods (Sec. 2.3).

### 2.1. Image Formation Model

Recovering the intrinsic characteristics of the scene based on acquired imagery is a highly complex problem, which requires a model of the image formation as such. These models often use an image space representation, i.e. all characteristic parameters are given in image coordinates denoted by $\mathbf{x}$.

One of the early models for image formation is the Dichromatic Reflection Model (DRM) [24], which decomposes the image $I$ into its diffuse and specular components:

$$I^c(\mathbf{x}) = m_d(\mathbf{x}) \int_\omega b(\lambda, \mathbf{x})\, e(\lambda, \mathbf{x}) \rho^c(\lambda)\, d\lambda \qquad (1)$$
$$+ m_s(\mathbf{x}) \int_\omega i(\lambda)\, e(\lambda, \mathbf{x}) \rho^c(\lambda)\, d\lambda,$$

where $e(\lambda, \mathbf{x})$ represents the energy of the incident light, and $b(\lambda, \mathbf{x})$ is the surface albedo. We integrate over the wavelength $\lambda$ of the visible spectrum $\omega$. $\rho^c(\lambda), c \in \{R, G, B\}$ denotes the spectral sensitivity of the camera (sensor) for the R, G, and B channels. We will omit the channel superscript $c$ without loss of generality. The achromatic terms $m_d(\mathbf{x})$ and $m_s(\mathbf{x})$ summarize the geometric dependence of the reflected diffuse and specular light, respectively, to the view angle, direction of the incident light and the surface normal for the observed scene location related to the image location $\mathbf{x}$.

A common assumption is the neutral interface reflection, *i.e.* the Fresnel reflectance $i$ is independent of $\lambda$ and can be omitted from Eq. 1 as it is a constant factor. Note that intrinsic image decomposition methods only provide shading and reflectance values up to a constant scalar multiple of the real values [15].

Later work has extended this model to account for non-uniform illumination, *e.g.* by considering shadows.

Maxwell *et al*. [20] formulate a multi-illuminant image as a linear combination of images under different illumination conditions. In their work, they focus on ambient illumination as they are mainly interested in chromatic shadows. However, considering Eq. 1, variations in the color of the incident light can be accounted for directly, as $e(\lambda, \mathbf{x})$ can encode the spatially varying information depending on the image coordinate.

In the case of the MIT dataset [15], the illumination is homogeneous and white, leading to $e(\lambda, \mathbf{x}) = \text{const}$ in Eq. 1, i.e. it is independent of the wavelength and the position. Using the assumption that the image of the object can be modeled as a linear combination of shading, reflectance, and specularity, Grosse *et al*. [15] simplify the representation of the model to its main components:

$$I(\mathbf{x}) = S(\mathbf{x})R(\mathbf{x}) + C(\mathbf{x}) \ , \tag{2}$$

where $S$, $R$, and $C$ represent the shading, reflectance, and specular components respectively. We refer readers to the work of Grosse *et al*. [15] for further explanation.

### 2.2. Datasets

As mentioned in the previous section, the MIT dataset by Grosse *et al*. [15] is the main resource for intrinsic image research. It is constructed from single-object scenes and white illuminant scenario with precise reflectance, and shading ground-truth. Further work by Barron *et al*. [3] used synthetic rendering on top of the MIT dataset in order to achieve more complex multi-illuminant ground-truth data, as well as depth information. Beigpour *et al*. [6] also extends the MIT concept to account for non-uniform illumination, but their approach is solely based on synthetic image generation for scene geometry and image formation, whereas the illumination data is synthesized from real light spectra. Furthermore, they provided a thorough study of state-of-the-art in intrinsic image recovery, but the data does not account for noise, scenes are entirely diffuse, and no depth information has been provided. Some color constancy datasets have targetd multi-illuminant scenes [14, 5, 31] but do not provide any shape and reflectance ground-truth.

Recent advances in the field of computer graphics based on physically motivated models produce realistic-looking images and, thus, have encouraged researches to use synthetic data as ground truth. In general, replacing real-world measurements by synthetic, computer generated imagery (CGI) is problematic in two ways. First, the photometric consistency of CGI compared to equivalent real-world imagery has yet to be validated. Considering image forensics, several methods exist that can classify photo-realistic CGI from real photos with more than $90\%$ accuracy [28]. Second, there is a strong overlap between the models used for rendering CGI and physics-based models used by intrinsic image methods. Thus the evaluation of different methods using synthesized data is potentially biased in favor of methods which rely on similar models as to which the synthesized data has been created.

The creation of a real multi-illuminant intrinsic image dataset with complete ground truth involves laborious, time consuming, and very precise procedure as well as expertise in the field. This paper takes on the challenge of providing such a dataset which extends over the MIT dataset in terms of complexity of the illumination and scene geometry, along with its full intrinsic ground-truth data.

### 2.3. Intrinsic Image Estimation Methods

Many works in intrinsic image recovery have been proposed in the literature. The earlier approaches relied on the assumption that the reflectance changes produce sharp edges, while shading changes are smooth [18]. Considering that the object reflectance is independent of the illumination, Weiss [29] has used several images of a scene under different lighting conditions to solve this problem. Entropy minimization, e.g. Finlayson *et al*. [10], and learning-based approaches, e.g. Tappen *et al*. [27], have proven to be advantageous. Non-local texture constraint [25], integration of local luminance amplitude with hue and texture [16], and image clustering [12] further improved the results.

Use of sparse reflectance priors has resulted in strong improvements in performance and results in the works of Shen *et al*. [26] Gehler *et al*. [13] and Serra *et al*. [23]. Barron *et al*. [2] combine priors on the local smoothness and global sparsity of the reflectance with priors on shape (flatness, outward-facing orientation at the occluding contour, and local smoothness).

In section 4 we evaluate some of the most recent methods in this field by Gehler *et al*. [13], Serra *et al*. [23] and Barron *et al*. [2] on our multi-illuminant real-photo ground-truth dataset.

## 3. Dataset and Ground-Truth Acquisition

The main contribution of this work is providing a dataset of images of real objects lit in complex, non-uniform, multi-colored lighting conditions with precise ground-truth, i.e. the Lambertian reflectance and shading as well as specularity. Furthermore we provide pixel-wise accurate color of the incident light on the objects. Our complete dataset is publicly available online[1].

Our work is inspired by works of Grosse *et al*. [15] and Beigpour *et al*. [6]. But it differs from the former mainly in that it present complex multi-colored lighting in the scenes with precise pixel-wise ground-truth as well as depth information. The latter has also provided the multi-colored lighting, but only on synthetic data. In this section we ex-

[1] http://www.cg.informatik.uni-siegen.de/data/iccv2015/intrinsic

plain in detail our framework for capturing the images and extracting the ground-truth information.

### 3.1. Scene Setup

Our dataset consists of 5 distinct scenes, each of which contains two colorful objects with complex geometry; see Fig. 2. A schematic example of a scene is given in Fig. 1. Lighting in each scene is provided using a DC-950 regulated illuminator with a 150W quartz halogen light source fitted with an IR filter. The intensity of the light can be controlled. Using a dual branch fiber optics light guide, the light is directed through two filter-wheels fitted with color filters.

One wheel is mounted on the scene's ground level at the front-left side and is equipped with a set of orange ('O'), yellow ('Y'), and white ('$W_1$') filters, and the other wheel is mounted on the top-front-right side of the scene and is equipped with a set of blue ('B'), green ('G'), and white ('$W_2$') filters. $W_1$ and $W_2$ are results of two neutral density filters with different levels - in order to regulate the scene's brightness - but have the same chromaticity. Each of these wheels act as a separate light source. The spectra of each of the individual illuminants, i.e. light source plus filter, is given in Fig. 3.

Using the combination of filters, we produce a set of 15 lighting conditions $\mathbb{L} = \{\ell_1, ..., \ell_{15}\}$ consisting of 9 two-illuminant and 6 single-illuminant (by blocking one of the sources) lighting conditions. In the following, we denote an illumination condition as

$$\text{LR with } \begin{cases} \text{L} \in \{\text{O, Y, W}_1\text{, N}\} & \text{(left filter color)} \\ \text{R} \in \{\text{B, G, W}_2\text{, N}\} & \text{(right filter color).} \end{cases}$$

Thus, OB stands for orange from left and blue from right. We use 'N' to indicate absence of an illuminant, e.g. NB means blue illuminant from right while the left source is blocked. The position of the two light sources are rigidly fixed in relation to the scene and the camera.

In order to represent the complexity of natural scenes, our scenes feature different types of surfaces and geometries, *e.g.* flat, curved, sharp or smooth edges, as well as various types of reflectance, *e.g.* uniformly colored and sparse scribbles. The illumination in the scenes resulted in both soft shadows and shadows with sharp edges, as well as smooth blending of differently colored illuminants over a curved surface. Fig. 4 presents an example of the lighting conditions of an acquired scene.

### 3.2. Data Acquisition and Calibration

The scene is captured using a Nikon D300 camera coupled with a Kinect version 2 Time-of-Flight (ToF) depth camera. The two cameras are rigidly attached together at about 8 cm distance, so the relative transformation is not altered during the experiments. The extrinsic and intrinsic parameters of this stereo-like setup are calibrated beforehand using a checkerboard and the standard camera calibration model from Bouguet's Matlab toolbox [9]. This way, we estimate a mapping between the images of the two cameras in order to register the color and depth images.

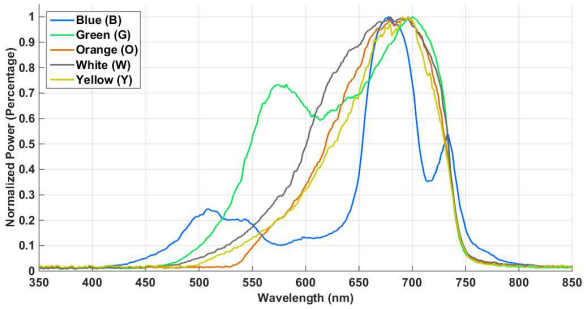


Figure 3. The illumination spectra: By placing color filters in front of the halogen lamp we create 5 distinctly colored light. Spectral distribution of each light is presented in this plot.

For the Nikon camera we use a 30mm Sigma zoom lens to acquire our scenes. The Kinect camera, on the other hand, has a wide field of view optics, which can unfortunately not be exchanged. Alternatively using the Kinect's built-in color camera is not possible, as its colors are not reliable.

In order to acquire reliable data, we disable the automatic white balance of the Nikon camera, set all the parameters of the camera to manual, and capture RAW images. Since Eq 1 is based on the linearity of the image, it is crucial to use the uncompressed raw output of the camera without any manipulation which is usually done by the software embedded in the camera. We use raw images for our evaluations and calculation of the ground-truth. But since linear raw images often appear dull and dark on the screen or when printed, we apply the color transform of Akkaynak *et al.* [1] (created using a Macbeth color checker) along with a gamma encoding as a post processing step for visualization of our images in this paper.

Regarding the depth acquisition, several problems arise due to the operation principle of ToF cameras in general, as well as to the restricted functionality and flexibility of the specific ToF Kinect camera. Most of the error sources related to ToF cameras, such as motion artefacts or interference with background illumination are less prominent due to our specific acquisition setup (see Kolb *et al.* [17] for a survey on ToF cameras and their error sources). Two aspects are more relevant in our setup, i.e. the range measurements of ToF are less reliable for darker object regions and highly reflective objects may lead to so-called multi-path interference, where different light paths from the camera's illumination unit to the sensor chip superimpose and lead to wrong depth values. To circumvent both effects, we capture

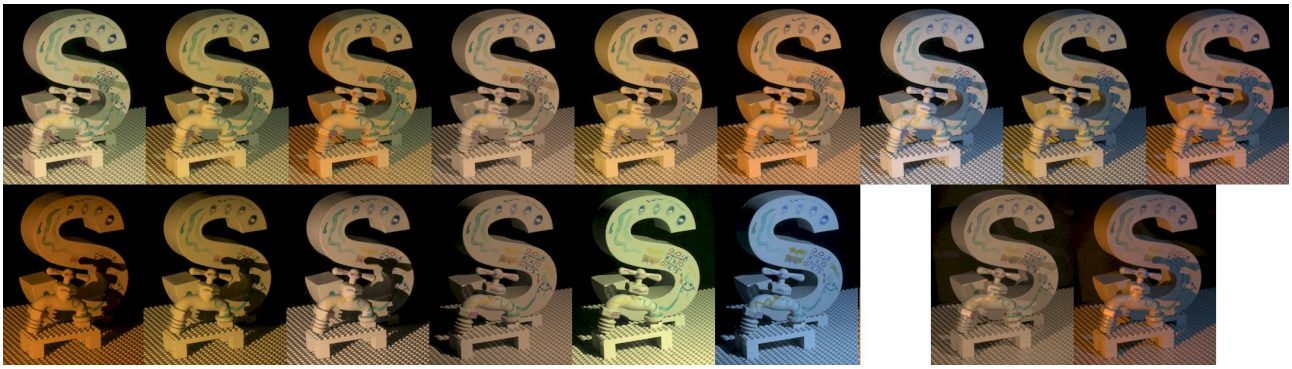

Figure 4. An example of the 17 different illumination conditions: Diffuse Multi-illuminant (Top-row), Diffuse Single-illuminant (bottom-row left), Specular Multi-illuminant (bottom-row right).

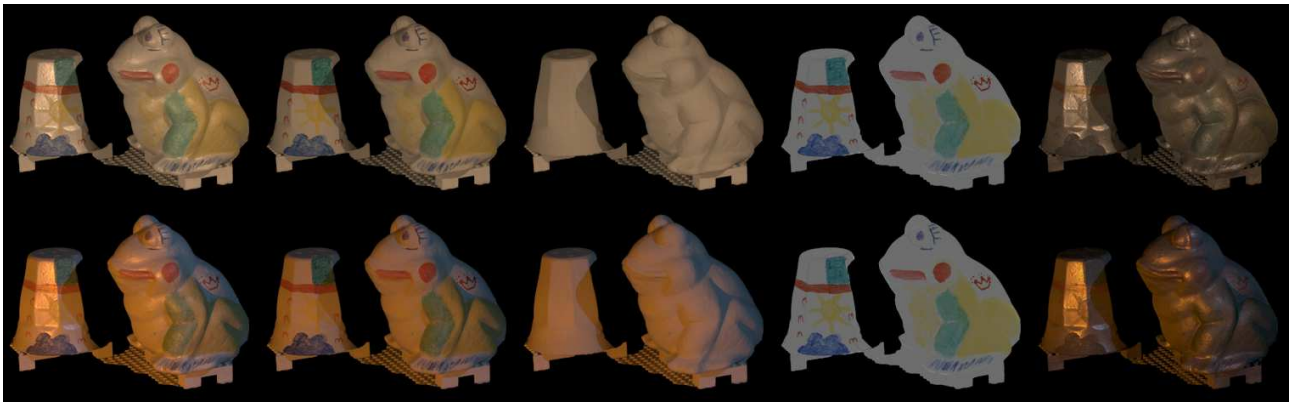

Figure 5. Determined Ground-Truth: The original image $I_{\text{org}}$ (left), the diffuse image $I_{\text{diff}}$ (second from left), the relative shading image $\widetilde{S}$ (center), the relative reflectance image $\widetilde{R}$ (second from right), and the specular image $I_{\text{spec}}$ (right) for the $W_1W_2$ (top row) and the OB (bottom row) lighting condition. Note that these images are enhanced for visualization using the gamma curve.

the depth images from the diffuse gray version of the scene (see Sec. 3.3). Due to the baseline between the Nikon and the Kinect, we further have the problem of false color assignments due to occlusions, *i.e*. scene points observed by the range sensor, but not by the color camera may get assigned to the color of the occluded object portion. We use the *shadow-maps* like approach from Lindner *et al*. [19] in order to identify these points and mask them as black.

As far as the specific ToF Kinect camera is concerned, we already mentioned the wide field of view optics, which can unfortunately not be exchanged. Furthermore, Kinect's ToF can only provide images in a VGA resolution, while the Nikon has 12 Mega-pixel resolution. Therefore, any of the two options for image based fusion of the range and the depth information implies difficult challenges:

- A depth-to-color transformation results in a very sparse depth representation of the scene with an up-scaling factor of about 16. Interpolating depth at this level of sparsity is an unsolved problem. We apply an inpainting method to illustrate the sub-optimality of the result; see Fig. 6, left.
- A color-to-depth transformation yields a strongly down-sampled color image, which would discard a vast amount of color information and may even be subject to aliasing; see Fig. 6, right.

Even though, one might argue that using lower resolution input data is feasible due to intrinsic image decomposition methods being often computationally exhaustive, we opted for not using the depth information at this stage, as most existing intrinsic image decomposition methods do not feature depth information so far. However, we provide the preliminary, sparse depth information, as it is still an additional and valuable cue. We will address the challange of providing full range, high resolution ground-truth depth data and depth upsampling in our future research.

### 3.3. Determination of Ground-Truth

The main goal of the paper is to provide ground-truth data for several scenes under varying illumination conditions. As the shading and the reflectance values can only be determined up to a constant scalar multiple of the real values, we cannot calculate the true shading image $S$ and the true reflectance image $R$ from Eq. 2, but only *relative shading image* $\widetilde{S}$ and the *relative reflectance image* $\widetilde{R}$.

Condering Eq. 2 the following relations hold:

$$I_{\text{org}} = S \cdot R + I_{\text{spec}} \text{ and } I_{\text{diff}} = S \cdot R, \qquad (3)$$
$$S \propto \widetilde{S} \text{ and } R \propto \widetilde{R}.$$

where $I_{\text{org}}$ denotes the *original image*, $I_{\text{diff}}$ the *diffuse image*, and $I_{\text{spec}}$ the *specularities*.

**Specularities:** Fig. 5 demonstrates an example from our dataset with specularity, shading, and reflectance ground-truth. The first step in creating the ground-truth is to separate the specular image $I_{\text{spec}}$ from the diffuse component $I_{\text{diff}}$. Following the work of Grosse *et al.* [15], we use a cross-polarization approach by placing linear polarizing filters in front of each of the light sources. First we block one of the light sources and rotate the polarizing filter placed in front of the camera lens until there are no specularities visible in the scene. We place a chrome sphere in the scene to judge the quality of the cross-polarization. Then we uncover the second light source and rotate its polarizing filter until the specularities caused by the second source also disappear from the chrome sphere. This yields $I_{\text{org}}$ and $I_{\text{diff}}$, and from Eq. 3 we get: $I_{\text{spec}} = I_{\text{org}} - I_{\text{diff}}$.

Note that even though the polarizing filters' nominal performance is less than 100% the captured images demonstrate no visible specularities on the objects.

**Reflectance:** The second and more difficult step is to separate reflectance from shading and illumination. Similarly to Grosse *et al.* [15], we create exact replica of the scene, one colored version for capturing reflectance and shading and one gray version for capturing shading only. We need to fulfill two main requirements in setting up the two versions of each scene:

- The geometric accuracy in the acquisition needs to be at pixel precision, and
- any inter-reflection must equal for both scene variants, as variations in the inter-reflection lead to erroneous reflectance.

In order to cope with the first requirements, we use only one instance of each object, where the initial object version is painted in a diffuse grayish reference color (RAL7042 standard paint; RGB=(142,146,145)); our alignment approach is described later in the text. The colored object versions are generated by sparsely applying various chromatic paint. In principle, one could also use initially colorful objects and uniformly paint them gray, but we found it easier to meet the first requirement using the prior approach. The second requirement is handled by placing the scene and the camera in a black tent, thus removing all ambient inter-reflections, and painting the back side of the objects in diffuse gray in order not to cause colored reflections on the objects. As we apply chromatic paint sparsely, the variation in the inter-object reflection is minimized.

The diffuse acquisition of the initial grayish and the colored scene versions yields the *diffuse gray image* $I_{\text{diff}}^{\text{gray}}$ and the *diffuse colored image* $I_{\text{diff}}^{\text{col}}$. From Eq. 3 we determine

$$\widetilde{S} = I_{\text{diff}}^{\text{gray}}, \quad \widetilde{R} = I_{\text{diff}}^{\text{col}} / I_{\text{diff}}^{\text{gray}}. \qquad (4)$$

As some intrinsic color reconstruction methods can handle specularities, or even use them to optimize their results, we additionally provide the captured *specular colored image* $I_{\text{spec}}^{\text{col}}$ for each scene, i.e. without polarizing filters, for the $\text{W}_1\text{W}_2$ and the OB illumination conditions.

**Alignment:** It is crucial to keep the position of the objects, lights, and the camera fixed between the acquisition of the colored versions $I_{\text{diff}}^{\text{col}}$ and the gray version $I_{\text{diff}}^{\text{gray}}$. These images must be aligned at pixel accuracy. It is not enough to align them in a post-processing step as even slight displacements can drastically alter the cast shadows and other features of the scene. Painting the objects on the set could result in some movements. Grosse *et al.* [15] have constructed a platform balanced on a group of spheres to be able to position the objects on their exact position. We have found that in our case, as our scenes consist of multiple objects with different geometries, Grosse *et al.*'s approach is difficult to apply. Therefore, we have fixed a large LEGO plate on the ground level of the scene and glued each of the objects on a platform made of a small LEGO plate with 4 legs on the corners; see Fig. 2. This structure provides us with more freedom as well as a very precise way of repositioning the objects in the scene. We have examined the images of the scene with great scrutiny and discarded the scenes with any slightest visible displacement or rotation.

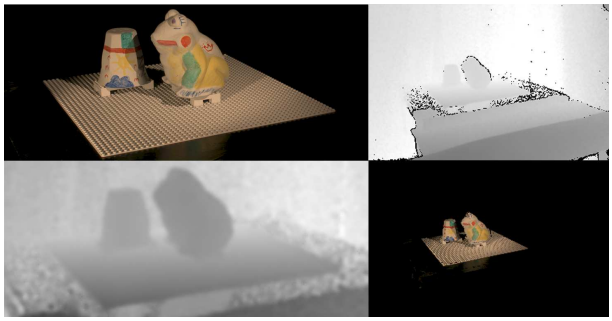

Figure 6. Top-row: Original RGB (left) and Depth (right) images; Bottom-row: Depth-to-color transformation using an inpainting method (left) and color-to-depth transformation (right). The resolution ratio is about 16:1.

### 3.4. Noise Analysis and Reduction

Since we calibrate the Nikon and Kinect cameras together, we need to keep the focal length constant for all

scenes. To avoid blurring artifacts, we have chosen a small aperture for which the depth of field would give us enough room to place each scene [11].

Dark pixels in the image tend to be much more prone to noise. In our setup we are interested especially in the effect of shadows. To avoid artifacts in the shadow areas we have tested each lighting condition with several different exposure times and choose the one which produces reliable images, *i.e.* in which the object's pixels' RGB values are not being over or under exposed. For each lighting condition, we capture the gray and the colored version of the scene with the same exposure time. As our scenes are stationary, we can allow for longer exposure, while keeping the ISO low enough to reduce noise.

As any camera-based image acquisition is affected by noises, we evaluate the noise-level of the Nikon in order to determine an appropriate number of images that need to be taken and averaged for noise reduction. Therefore, we run a sequence of acquisitions of the Macbeth chart, manually segment one of its gray-level patches, and compute the standard deviation for R, G, and B using all pixel values within the segmentation. Fig. 7 depicts the resulting standard deviation as function of the number of averaged images.

In the current work, we use the average of 10 images to construct one shot, which is a feasible compromise between the amount of data to be acquired and the noise level achieved.

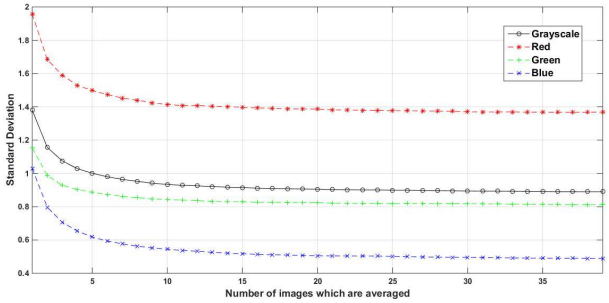


Figure 7. The standard deviation of a gray cell inside a Macbeth chart as function of the number of averaged images.

## 4. Benchmarking

In this section we evaluate Barron *et al*. [2], Gehler *et al*. [13], and Serra et. al. [23], three of the state-of-the-art intrinsic image recovery methods using the proposed dataset. We use the publicly available code for each of these methods from the authors' own web pages. All the parameters are set to the default provided with these codes and no further training on our data has been done. Tab. 1 provides quantitative evaluation of each method on our dataset [2]. To better explain the results we have grouped our lighting conditions from Sec. 3.1 in six categories:

$$\begin{aligned}
\text{White} &\in \{\text{W}_1\text{W}_2\ ,\ \text{W}_1\text{N}\ ,\ \text{NW}_2\} \\
\text{MonoColor} &\in \{\text{ON}\ ,\ \text{YN}\ ,\ \text{NB}\ ,\ \text{NG}\} \\
\text{ColorWhite} &\in \{\text{W}_1\text{B}\ ,\ \text{W}_1\text{G}\ ,\ \text{OW}_2\ ,\ \text{YW}_2\} \\
\text{MultiColor} &\in \{\text{OB}\ ,\ \text{OG}\ ,\ \text{YB}\ ,\ \text{YG}\} \\
\text{SpecWhite} &\in \left\{\{W_1W_2\}_{Spec}\right\} \\
\text{SpecMultiColor} &\in \left\{\{OB\}_{Spec}\right\}
\end{aligned}$$

where the $Spec$ subscript denotes that the polarizing filter has been removed to allow for specularities. Here we use two different metrics, namely *Local Mean Square Error (LMSE)* and *Angular Error* ($E_a$).

As discussed before, the ground truth shading and reflectances can only be provided up to a scalar multiple of the true values. To tackle this issue, Grosse *et al*. [15] define LMSE as the scale-invariant mean square error (MSE):

$$MSE(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\alpha}\hat{\mathbf{x}}\|^2 \ , \tag{5}$$

where $\mathbf{x}$ and $\hat{\mathbf{x}}$ are the vector-valued true and estimated values respectively with $\hat{\alpha} = argmin_\alpha \|\mathbf{x} - \hat{\alpha}\hat{\mathbf{x}}\|^2$ . Here $\alpha$ is fitted over each local square sized window $w$ of size $k$. The LMSE is given by:

$$LMSE_k(\mathbf{x}, \hat{\mathbf{x}}) = \sum_{w \in W} MSE(\mathbf{x}_w, \hat{\mathbf{x}}_w) \ . \tag{6}$$

We refer the readers to Grosse *et al*. [15] for more details. We provide LMSE values for each of the methods regarding the reflectance and shading in Tab. 1. While the shading ground-truth of Grosse *et al*. is a 2D grayscale image, our shading ground-truth contains an extra dimension, i.e. the color of the illuminant at each pixel. Here we only use the gray values of our ground truth and the estimated shadings by the 3 methods.

Barron *et al*. [2] also provides color values for shading which integrates the illumination estimate. To evaluate the per-pixel illumination estimates, we use the Angular Error which is a common metric in the field of color constancy [5]. For a pair of vectors $\mathbf{i}_{gt}$ and $\mathbf{i}_{est}$ which denote *ground truth* and *estimated* illuminant color respectively, we define:

$$E_a = \arccos\left((\mathbf{i}_{gt})^T(\mathbf{i}_{est})\right) \ . \tag{7}$$

In Tab. 1, the illumination evaluation is produced by calculating the mean value of the per-pixel $E_a$ over the entire estimated and ground truth shading. As [13, 23] only provide grayscale shading, we compare Barron *et al*. [2] against the baseline that is to set the estimation to a gray

[2]Examples of results are available for qualitative comparison in the supplementary document.

| Evaluation | Method | Diffuse Acquisition | | | | Specular Acquisition | |
|---|---|---|---|---|---|---|---|
| | | White | MonoColor | ColorWhite | MultiColor | SpecWhite | SpecMultiColor |
| Reflectance | Barron *et al*. | 0.045 | 0.233 | 0.158 | 0.264 | 0.051 | 0.410 |
| | Gehler *et al*. | 0.155 | 0.263 | 0.194 | 0.257 | 0.154 | 0.280 |
| | Serra *et al*. | 0.049 | 0.117 | 0.096 | 0.133 | 0.044 | 0.102 |
| Shading | Barron *et al*. | 0.018 | 0.024 | 0.024 | 0.029 | 0.015 | 0.028 |
| | Gehler *et al*. | 0.013 | 0.014 | 0.011 | 0.015 | 0.013 | 0.023 |
| | Serra *et al*. | 0.014 | 0.019 | 0.020 | 0.024 | 0.019 | 0.030 |
| Illumination | Barron *et al*. | 2.0° | 23.0° | 17.6° | 29.8° | 4.7° | 48.9° |
| | Baseline | 21.9° | 29.3° | 23.9° | 28.0° | 21.9° | 28.5° |

Table 1. Evaluation results.

value (R=G=G). Note that $E_a$ is invariant to the scale factor which exists between the true value of the illuminant at each pixel and our ground truth.

According to Tab. 1, the methods are more challenged when the illumination color strongly deviates from white as it is very difficult to separate the effect illumination color from the natural intrinsic color of the object. The lowest error is achieved when the illumination is all white. The presence of specularities also increases the error. Over all, Barron *et al*. perform better when estimating the illumination since they consider colored lighting in the optimizations, Gehler *et al*. are more accurate in estimation of shading, and Serra *et al*. perform better in recovering reflectance.

## 5. Discussion and Future Work

Here we briefly discuss challenges and avenues for future work:

**Lighting:** Even though using the gray-painted version of the scene, a reliable ground-truth can be calculated, this strongly relies on all the other conditions in the scene, especially the lighting, to be unchanged. Outdoor lighting and natural scenes are widely dynamic, e.g., moving clouds, shaking leaves in a light breeze. Bell *et al*. [7] tackle this using a crowd-sourced ground-truth labeling performed by human subjects. But the authors admit that this creates errors since the human judgment is subjective and does not always hold up to the reality due to perceptual errors. We will further investigate the incorporation of ambient light and complex backgrounds in the future.

**Interreflections:** Global illumination in real-world results in colored interreflections which though interesting to have in a dataset, could falsify the ground-truth because the colored interreflections caused by the diffuse-diffuse reflection will not be present on the gray-painted scene; therefore, the ground-truth calculation wrongly considers them as part of the reflectance. Despite our best effort, some minor interreflections are still present in our data, e.g. car fenders in scene 2. We argue that these effects can be over-looked as in our data they cause only an average 1.69° alteration in recovered ground-truth reflectance color in terms of angular error for an area of about .4% of total object pixels. This is far below affecting the numbers in Table 1. In the future, we will explore more effective ways of avoiding such artifacts while featuring scenes with interreflections.

**Texture:** The MIT dataset does not contain color texture and has minimal 3D texture. Our dataset features some fine 3D textures (e.g., the brush strokes, the texture of the 3D-printed dragon in Scene 1, and the pattern on the LEGO plate). We would like to extend our dataset to include more color variations and 3D textures like fur and woven textile.

**Depth upsampling:** The current real depth information acquired with a ToF-Kinect is sparse in resolution compared to the color imagery. Some works in the literature have so far attempted to tackle range data upsampling using corresponding RGB data [22].

## 6. Conclusion

In this paper, we have presented the methodology to generate real-photo ground-truth dataset for intrinsic image research with complex, multi-illuminant scenarios under non-uniform lighting and challenging colored cast shadows. We provide full per-pixel intrinsic ground-truth data for these scenarios, including reflectance, specularity, shading, and illumination. Furthermore, we have evaluated 3 state-of-the-art intrinsic image recovery methods, using our dataset. Our versatile, ground-truth dataset can be used for developing and evaluating any current and future intrinsic image reconstruction technique. We hope that as the MIT dataset, despite its shortcomings, has sparked the creation of several intrinsic image methods in the past, our dataset can contribute to the new developments in this field.

Intrinsic video decomposition methods are becoming a recent trend in the community [30, 8]. It would be interesting to extend our dataset to dynamic scenes.

## Acknowledgment

This research was partially funded by the German Research Foundation (DFG) as part of the research training group GRK 1564 Imaging New Modalities

## References


[1] D. Akkaynak, T. Treibitz, B. Xiao, U. A. Gürkan, J. J. Allen, U. Demirci, and R. T. Hanlon. Use of commercial off-the-shelf digital cameras for scientific data acquisition and scene-specific color calibration. *Journal of the Optical Society of America A, Optics, image science, and vision*, 31(2):312–321, 2014. 4

[2] J. Barron and J. Malik. Color constancy, intrinsic images, and shape estimation. In *European Conference on Computer Vision (ECCV)*, pages 55–70, 2012. 1, 3, 7

[3] J. T. Barron and J. Malik. Intrinsic scene properties from a single RGB-D image. *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 17–24, 2013. 1, 3

[4] H. Barrow and J. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, pages 3–26, 1978. 1

[5] S. Beigpour, C. Riess, J. van de Weijer, and E. Angelopoulou. Multi-illuminant estimation with conditional random fields. *IEEE Transactions on Image Processing*, 23(1):83–95, jan 2014. 3, 7

[6] S. Beigpour, M. Serra, J. van de Weijer, R. Benavente, M. Vanrell, O. Penacchio, and D. Samaras. Intrinsic image evaluation on synthetic complex scenes. In *IEEE International Conference on Image Processing (ICIP)*, pages 285–289, 2013. 1, 3

[7] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014. 8

[8] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister. Interactive intrinsic video editing. *ACM Transactions on Graphics (TOG)*, 33(6):197, 2014. 8

[9] J.-Y. Bouguet. Camera calibration toolbox for matlab. 2004. 4

[10] G. Finlayson, M. Drew, and C. Lu. Intrinsic images by entropy minimization. In *European Conference on Computer Vision (ECCV)*, pages 582–595, 2004. 3

[11] D. Flemin. Depth of Field Calculator. http://www.dofmaster.com/dofjs.html, 2005. [Online; accessed 20-April-2015]. 7

[12] E. Garces, A. Munoz, J. Lopez-Moreno, and D. Gutierrez. Intrinsic images by clustering. In *Computer Graphics Forum*, volume 31, pages 1415–1424. Wiley Online Library, 2012. 3

[13] P. Gehler, C. Rother, M. Kiefel, L. Zhang, and B. Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *Advances in Neural Information Processing Systems (NIPS)*, pages 765–773, 2011. 3, 7

[14] A. Gijsenij, R. Lu, and T. Gevers. Color Constancy for Multiple Light Sources. *IEEE Transactions on Image Processingn*, 2011. 3

[15] R. Grosse, M. Johnson, E. Adelson, and W. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *IEEE International Conference on Computer Vision*, pages 2335–2342, 2009. 1, 2, 3, 6, 7

[16] X. Jiang, A. J. Schofield, and J. L. Wyatt. Correlation-based intrinsic image extraction from a single image. In *European Conference on Computer Vision (ECCV)*, pages 58–71. Springer, 2010. 3

[17] A. Kolb, E. Barth, R. Koch, and R. Larsen. Time-of-flight cameras in computer graphics. *Computer Graphics Forum*, 29(1):141–159, 2010. 4

[18] E. Land. The retinex theory of colour vision. *Scientific American*, 237(6):108–129, 1977. 3

[19] M. Lindner, M. Lambers, and A. Kolb. Sub-pixel data fusion and edge-enhanced distance refinement for 2D/3D images. *International Journal of Intelligent Systems Technologies and Applications*, 5(3):344–354, 2008. 5

[20] B. Maxwell, R. Friedhoff, and C. Smith. A Bi-Illuminant Dichromatic Reflection Model for Understanding Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, June 2008. 3

[21] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, pages 746–760, 2012. 1

[22] J. Park, H. Kim, Y.-W. Tai, M. Brown, and I. Kweon. High quality depth map upsampling for 3D-TOF cameras. In *IEEE International Conference on Computer Vision (ICCV)*, pages 1623–1630, Nov 2011. 8

[23] M. Serra, O. Penacchio, R. Benavente, and M. Vanrell. Names and shades of color for intrinsic image estimation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 278–285, 2012. 3, 7

[24] S. Shafer. Using color to separate reflection components. *Color Research and Application*, 10(4):210–218, 1985. 2

[25] L. Shen, P. Tan, and S. Lin. Intrinsic image decomposition with non-local texture cues. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008. 3

[26] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 697–704. IEEE, 2011. 3

[27] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(9):1459–1472, 2005. 3

[28] X. Wang, Y. Liu, B. Xu, L. Li, and J. Xue. A statistical feature based approach to distinguish PRCG from photographs. *Computer Vision and Image Understanding*, 128(0):84 – 93, 2014. 3

[29] Y. Weiss. Deriving intrinsic images from image sequences. In *International Conference on Computer Vision*, pages 68–75, 2001. 3

[30] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez. Intrinsic video and applications. *ACM Transactions on Graphics (TOG)*, 33(4):80, 2014. 8

[31] I. M. Ziko, S. Beigpour, and J. Y. Hardeberg. Design and creation of a multi-illuminant scene image dataset. *Image and Signal Processing, Lecture Notes in Computer Science*, 8509:531–538, June 2014. 3