# Intrinsic Scene Decomposition from RGB-D images

Mohammed Hachama
hachamam@gmail.com

Bernard Ghanem
Bernard.Ghanem@kaust.edu.sa

Peter Wonka
pwonka@gmail.com

King Abdullah University of Science and Technology (KAUST)
Thuwal, Saudi Arabia

## Abstract

*In this paper, we address the problem of computing an intrinsic decomposition of the colors of a surface into an albedo and a shading term. The surface is reconstructed from a single or multiple RGB-D images of a static scene obtained from different views. We thereby extend and improve existing works in the area of intrinsic image decomposition. In a variational framework, we formulate the problem as a minimization of an energy composed of two terms: a data term and a regularity term. The first term is related to the image formation process and expresses the relation between the albedo, the surface normals, and the incident illumination. We use an affine shading model, a combination of a Lambertian model, and an ambient lighting term. This model is relevant for Lambertian surfaces. When available, multiple views can be used to handle view-dependent non-Lambertian reflections. The second term contains an efficient combination of $\ell_2$ and $\ell_1$-regularizers on the illumination vector field and albedo respectively. Unlike most previous approaches, especially Retinex-like techniques, these terms do not depend on the image gradient or texture, thus reducing the mixing shading/reflectance artifacts and leading to better results. The obtained non-linear optimization problem is efficiently solved using a cyclic block coordinate descent algorithm. Our method outperforms a range of state-of-the-art algorithms on a popular benchmark dataset.*

## 1. Introduction

Intrinsic image decomposition is the process of decomposing a given image into shading and reflectance components. Such decomposition can be used in multiple applications such as image editing, 3D shape reconstruction, and object recognition.

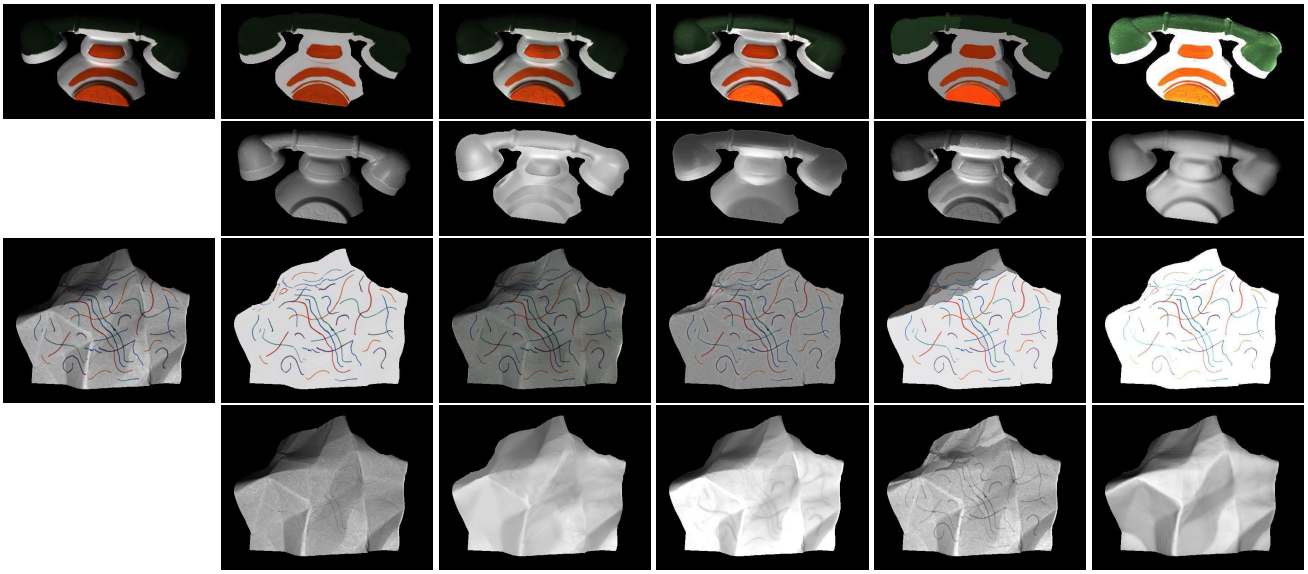
While intrinsic decomposition has been studied since the 1970's, it still remains a challenging problem. It is a highly under-constrained problem in which two unknowns (shading and reflectance) are to be estimated for each observation (image color) that is available. To overcome the fundamental ill-posedness, several approaches have been introduced utilizing additional information such as user interaction[6], priors [10, 2], multiple images [28], videos [30, 14] and depth cues [2, 8].

### 1.1. Related work

One of the earliest works addressing intrinsic image decomposition was the Retinex algorithm [16, 20]. This technique relies on the simple heuristic of associating strong edge gradients with reflectance changes and small gradients with illumination changes. Based on the same assumption, many approaches have been then proposed, using different strategies such as image gradient thresholding [13, 18] or learning gradient variations [27, 26]. A major drawback of these approaches is that the decomposition is analyzed within a small window. To extend beyond local analysis, several works use additional global consistency terms based on textures [24, 31], reflectance clustering [10, 4], and sparsity priors [10, 25]. Nevertheless, the problem remains severely under-constrained. The performance of existing algorithms on complex real-world images remains limited.

Depth cues have also been used in intrinsic decomposition of RGB-D images to further constrain the solution and improve the results. In [3], Barron and Malik present a unified model for a joint estimation of a smoothed depth map, chromatic spatially-varying illumination, and reflectance from a single RGB-D image. They constrain the solution by learning novel priors on albedo local smoothness and global sparsity, and on absolute color. They also introduce generic Gaussian priors on the illumination environment represented using spherical harmonics.

Chen and Koltun [8] showed that high quality decomposition results can be obtained by properly constraining shading components using surface normals. To constrain the albdeo, they use a non-local term that penalizes pairwise albedo differences between image pixels. A set of nearest neighbors for each pixel is defined based on their spatial positions and normals; the shading component of each pixel is



(a) Input (b) Ground truth (c) Barron-Malik [3] (d) Chen et al. [8] (e) Bell et al. [4] (f) Our approach

Figure 1. Intrinsic decomposition of two images from the MIT dataset [11]. (a) Input RGB images. (b) Ground truth albedo and shading images. (c-e) Intrinsic decomposition obtained by state-of-the-art techniques. (f) Intrinsic decomposition obtained by our approach.

constrained to be similar to those of its neighbors.

A related challenge is intrinsic decomposition using several images of the same static scene under different lighting conditions. For time-lapse sequences, Weiss [28] applies a median operator on the log-intensity over all the images as a robust estimator of the log-reflectance derivatives. Laffont et al. [15] used multi-view stereo to automatically reconstruct 3D points and normals, from which they derive relationships between reflectance values at different locations across multiple views. These are later used to robustly estimate reflectance ratios between pairs of points. The reflectance ratios are then taken as constraints to enforce a coherent solution across multiple views and illuminations.

Lee et al. [17] developed a model for intrinsic decomposition of a sequence of RGB-D video frames acquired from a Kinect camera. Their model builds on Retinex with nonlocal constraints, enforcing relationships among the shading components of different surface points according to their similarity in surface orientation. To improve the handling of view-dependent effects, they use temporal constraints that favor consistency in the intrinsic color of a surface point seen in different video frames. The obtained optimization problem depends on a linear system that allows for an efficient solution.

### 1.2. Our approach

In this paper, we propose a novel technique for intrinsic static scene decomposition using one or more RGB-D images obtained from different views. To achieve that, we formulate an energy minimization problem to jointly estimate the albedo and illumination of each point of the scene. The energy is composed of a data term and a regularity term. The first term is related to the image formation process and expresses the relation between the color, the albedo, the surface normals, and the incident illumination. To constrain the problem, we use an efficient combination of priors. We assume that illumination changes smoothly and can be constrained by an $\ell_2$-term, while the albedo tends to be piecewise smooth for which an $\ell_1$-regularization is suited.

The main novelty of our approach is a robust estimation of the intrinsic properties by considering, for each scene surface position, all related points both in the same and in different views. The contribution of each neighbor is quantified using some weights, taking into account spatial distance and view-coherence. This alleviates misalignment errors caused by depth and reconstruction errors that appear in approaches based on image correspondence as in [17] and view-dependent non-Lambertian reflection artifacts encountered in single RGB-D image decomposition [2, 8]. Moreover, unlike previous approaches, albedo and illumination smoothness terms used in this paper do not depend on the image values, gradients, or textures. Therefore, some decomposition artifacts are corrected by avoiding mixing the color with the albedo. In fact, we believe that this is better justified theoretically because priors should not depend on observations. Also, the $\ell_1$-regularization of the albedo constitutes a global coherence sparsity prior. In Figure 1, we show some comparative results of our method.

## 2. Methodology

### 2.1. Overview

The input data of our algorithm is a colored point cloud, representing a static scene. Such a point cloud can be obtained from a single or a set of aligned RGB-D images. In the latter case, each point is labeled by the index of the camera it was taken from.

A surface is extracted from the point cloud using a Poisson reconstruction technique [12]. Based on an implicit representation with a volumetric indicator function, this technique produces surfaces with added parts through hole filling. We apply a post-processing step which removes all surface vertices beyond a chosen distance from the initial point cloud.

Then, we determine intrinsic surface properties by jointly estimating the albedo and illumination of each surface vertex. This is achieved by minimizing an energy composed of a similarity term defined using the data point cloud and a smoothness term. These terms are described in Sections 2.2 and 2.3. The optimization strategy, explained in Section 2.5, is based on cyclic block coordinate descent. The algorithm is initialized by a rough intrinsic image decomposition. We propose in Section 2.4 a technique to efficiently project the initial image decomposition on the surface. More details are given in the supplementary material. Next, we introduce the notation used in the rest of the paper.

#### Notation

- The input point cloud is denoted by

$$\mathcal{Y} = \{\mathbf{u}_p = (\mathbf{x}_p, \mathbf{y}_p, z_p) \in \mathbb{R}^3 \times [0,1]^3 \times \mathbb{N} \\ , p \in \{1, ..., N_p\}\}, \quad (1)$$

  where $N_p$ is the size of the point cloud (number of points). For a given point indexed by $p$, $\mathbf{x}_p$ represents the position, $\mathbf{y}_p$ represents the color, and $z_p$ represents the label (camera index). We denote the vector components with subscripts (for example $\mathbf{y}_p = (y_{p,1}, y_{p,2}, y_{p,3})^T$). Images are supposed to be normalized with color values in the set $[0,1]^3$.

- The reconstructed surface is denoted by

$$\mathbf{S} = \{(\mathbf{s}_i, \mathbf{c}_i, \mathbf{n}_i) \in \mathcal{T} \; ; \; i \in \{1, ..., N_s\}\}, \quad (2)$$

  where $N_s$ represents the size of the surface (the number of vertices) and $\mathcal{T} = \mathbb{R}^3 \times [0,1]^3 \times (\mathbb{S}^2 \times \{1\})$. We denote the location of a surface point by $\mathbf{s}_i$, its color by $\mathbf{c}_i$, while $\mathbf{n}_i$ represents the augmented normal vector obtained by appending a fourth dimension with unity value ($\mathbf{n}_i = (n_i^1, n_i^2, n_i^3, 1)^T$). The normals belong to $\mathbb{S}^2$, the unit sphere of normalized vectors in $\mathbb{R}^3$. For each point $\mathbf{s}_i$, the set of indices of neighboring vertices is denoted by $\mathcal{N}_i \subset \{1, ..., N_s\}$.

- With each surface point $i$, we associate two vectors: its albedo $\mathbf{a}_i \in [0,1]^3$ and incident illumination vector $\mathbf{l}_i \in [0,1]^4$. Thus, we also augment the illumination vector with a fourth component as explained in the next section. For the sake of compact notation, we form a $3 \times N_s$ matrix $\mathbf{A}$ whose columns are the albedo vectors $(\mathbf{a}_i)_{i \in \{1,...,N_s\}}$ of all the vertices. Similarly, we form a $4 \times N_s$ light matrix $\mathbf{L}$. When images are acquired under different lighting conditions, light matrix $\mathbf{L}$ is also labeled by its corresponding camera index.

- Furthermore, we consider weights $w_p^i \in [0,1]$ representing the coherence of a point cloud point $p$ with a surface point $i$. These weights are explained in Section 2.2.1.

- We use the notation $\|.\|$ for the Euclidean norm of $\mathbb{R}^3$.

### 2.2. Data term

The intensity of diffuse lambertian objects can be explained by parametric low dimensional global lighting models such as spherical harmonics or quadratic functions [23, 29]. Here, we use a first order model to represent a local vertex-dependent lighting. We describe the surface color formation model by:

$$\forall i \; : \; \mathbf{c}_i = \mathbf{l}_i^T \mathbf{n}_i \, \mathbf{a}_i, \quad (3)$$

$$= (l_{i,1} n_{i,1} + l_{i,2} n_{i,2} + l_{i,3} n_{i,3} + l_{i,4}) \; \mathbf{a}_i \quad (4)$$

Hence, the color of each surface vertice is the product of a scalar shading value and an albedo vector. We handle non lambertian view-dependent reflections with this model by using multiple views of the surface. Besides, we make a white light assumption for the sake of simplicity. This can be relaxed by estimating colored light which generates a vector-valued shading. Note that the fourth lighting component is usually considered as an ambient lighting term. Since this model is local (vertex-dependent), it can take into account attached and detached shadows.

As the scene surface is reconstructed from a point cloud, a surface vertex can be related to many data points ; especially when multiple RGB-D images are fused. Thus, we define some weights $w_p^i \in [0,1]$ to express the coherence of a surface vertex $i$ with each data point $p$, as detailed in Section 2.2.1. The higher the weight, the more their colors are expected to be similar. Using these weights, we express the data constraint as

$$E_D(\mathbf{A}, \mathbf{L}) = \sum_{i=1}^{N_s} \sum_{p=1}^{N_p} w_p^i \left\| \mathbf{y}_p - \mathbf{c}_i \right\|, \\ = \sum_{i=1}^{N_s} \sum_{p=1}^{N_p} w_p^i \left\| \mathbf{y}_p - \mathbf{a}_i \mathbf{n}_i^T \mathbf{l}_i \right\|. \quad (5)$$

The energy term $E_D$ is the sum of individual data terms $E_{D,i}$, where:

$$E_{D,i}(\mathbf{A},\mathbf{L}) = \sum_{p=1}^{N_p} w_p^i ||\mathbf{y}_p - \mathbf{c}_i||. \quad (6)$$

Hence, the color of each vertex $\mathbf{c}_i$ is a robust weighted spatial median of the point cloud colors. It is worth noting, however, that this color is not estimated directly but implicitly determined from the estimated light and albedo.

#### 2.2.1 Weights

The role of the weights $(w_p^i)_{i,p}$ is twofold. First, they are used to implicitly estimate a surface color for each vertex from the point cloud colors. Basically, this is based on the spatial distance between the vertex and the data points. Second, they impose temporal constraints that favor consistency in the intrinsic properties of a surface vertex seen in different views. This improves the decomposition in cases of non-lambertian reflections with view-dependent effects such as specular highlights. Thus, we use the next formula:

$$w_p^i \;=\; \underbrace{d_s(\mathbf{s}_i,\mathbf{x}_p)}_{\text{Spatial distance}} \;.\; \underbrace{d_v(\mathbf{s}_i,\mathbf{x}_p)}_{\text{view-coherence}}, \quad (7)$$

where the spatial distance is expressed as

$$d_s(\mathbf{s}_i,\mathbf{x}_p) = \exp \frac{||\mathbf{s}_i - \mathbf{x}_p||^2}{\sigma_s^2}, \sigma_s > 0. \quad (8)$$

View-coherence is defined upon the scalar product of the surface normal $n_i = (n_i^1, n_i^2, n_i^3)^T$ and the axis of the camera labeled $z_p$, denoted by $o_{z_p}$:

$$d_v(\mathbf{s}_i,\mathbf{x}_p) = \max(-n_i.o_{z_p}, 0). \quad (9)$$

### 2.3. Smoothness terms

To constrain our ill-posed problem, we add the following terms as regularizers. First, as noted by several authors, the albedo tends to be piece-wise smooth. So, we use a $\ell_1$-norm on the albedo variations:

$$E_A(\mathbf{A}) = \lambda_a \sum_{i=1}^{N_s} \sum_{j\in\mathcal{N}_i} ||\mathbf{a}_i - \mathbf{a}_j||, \quad (10)$$

where $\lambda_a > 0$. The double sum is the $\ell_{1,1}$-sparse norm of the matrix whose coefficients are given by: $\alpha_{i,j} = ||\mathbf{a}_i - \mathbf{a}_j||$. Therefore, this $\ell_1$-regularization of the albedo constitutes a global coherence sparsity prior (sparsity on gradients encourages homogeneous regions and so reduces the total number of colors). On the other hand, assuming that the illumination component changes smoothly over the scene, we add the following constraint that ensures that lighting variation over all surface vertices is small:

$$E_L(\mathbf{L}) = \lambda_l \sum_{i=1}^{N_s} \sum_{j\in\mathcal{N}_i} ||\mathbf{l}_i - \mathbf{l}_j||^2, \quad (11)$$

where $\lambda_l > 0$. Similarly, the double sum here is a $\ell_{2,2}$-norm of the matrix of lighting variation norms. The idea of a spatially-varying model of illumination has also been used in [9]. In this paper, the illumination is modeled as a vector field and regularized by minimizing the $\ell_2-$norm of its gradient.

### 2.4. Initialization

To initialize the surface intrinsic decomposition, we first perform an initial image decomposition using a simple and fast $\ell_2-\ell_p$ image decomposition of the log-intensity derivatives as described in [5]. The obtained shading and albedo images are projected on the surface using depth information and camera parameters. This gives an initial surface albedo and shading $(\mathbf{sh}_i)_{i\in\{1,\dots,N_s\}}$. Then, we estimate an initial illumination field by minimizing the following energy:

$$\begin{aligned} \min_{\mathbf{L}} \quad & J(\mathbf{L}) = \sum_{i}^{N_s} (\mathbf{sh}_i - \mathbf{l}_i^T\mathbf{n}_i)^2 + \lambda_l^0 \sum_{i=1}^{N_s}\sum_{j\in\mathcal{N}_i} ||\mathbf{l}_i - \mathbf{l}_j||^2 \\ s.t. \quad & 0 \le l_{i,0} \le 1\ ,\ 0 \le \mathbf{l}_i^T\mathbf{n}_i \le 1\ ,\ i = 1,\dots,N_s, \end{aligned}$$

where $\lambda_l^0 > 0$ is a smoothness parameter (different from $\lambda_l$) and $\rho > 0$ (set to 0.01 in all experiments). This is a constrained convex optimization problem. We use the Alternating Direction Method of Multipliers to solve it [7]. This method considers the following equivalent problem:

$$\begin{aligned} \min_{\mathbf{L}} \quad & \mathbf{L} + g(\mathbf{Z}) \\ s.t. \quad & \mathbf{L} - \mathbf{Z} = 0, \end{aligned}$$

where $g$ is the indicator function of the constraints set $\mathcal{C}$. This problem is then solved by iterating three steps:

$$\begin{aligned} \mathbf{L}^{k+1} &:= \operatorname{argmin}_{\mathbf{L}} \left( J(\mathbf{L}) + \rho/2 \left\| \mathbf{L} - \mathbf{Z}^k + \mathbf{U}^k \right\|^2 \right), \\ \mathbf{Z}^{k+1} &:= \Pi_{\mathcal{C}}(\mathbf{L}^{k+1} + \mathbf{U}^k), \\ \mathbf{U}^{k+1} &:= \mathbf{U}^k + \mathbf{L}^{k+1} - \mathbf{Z}^{k+1}), \end{aligned}$$

where $\Pi_{\mathcal{C}}$ is the projection operator on the set $\mathcal{C}$. The projection is achieved for each vertex $i$ by double thresholding of $l_{i,0}$ and orthogonal projection on the plane oriented by the normal vector. We found this image-to-surface projection technique very effective to regularize the shading whenever the normal information is available (Figure 2).

Figure 2. Intrinsic image decomposition initialization (middle image) and its image-to-surface projection (right image).

### 2.5. Optimization

Combining the data constraint with the regularization terms, we obtain the global energy:

$$E(\mathbf{A},\mathbf{L}) = \sum_{i=1}^{N_s}\sum_{p=1}^{N_p} w_p^i ||\mathbf{y}_p - \mathbf{a}_i\mathbf{n}_i^T\mathbf{l}_i|| + \lambda_a \sum_{i=1}^{N_s}\sum_{j\in\mathcal{N}_i} ||\mathbf{a}_i - \mathbf{a}_j|| + \lambda_l \sum_{i=1}^{N_s}\sum_{j\in\mathcal{N}_i} ||\mathbf{l}_i - \mathbf{l}_j||^2. \quad (12)$$

The optimization problem is expressed as

$$\begin{aligned} &\min_{\mathbf{A},\mathbf{L}} \quad E(\mathbf{A},\mathbf{L}) \\ &s.t. \quad 0 \le a_{i,k} \le 1\ ,\ k = 1,\dots,3, \\ &\qquad 0 \le l_{i,0} \le 1\ ,\ 0 \le \mathbf{l}_i^T\mathbf{n}_i \le 1\ ,\ i = 1,\dots,N_s. \end{aligned}$$

This is a non linear optimization problem. To solve it, we apply a cyclic block coordinate descent algorithm. This is achieved by minimizing along one direction at a time. We cyclically iterate through each surface vertex index $i$ and minimize $E$ with respect to $(\mathbf{a}_i, \mathbf{l}_i)$. Thus, we define the following $i^{\text{th}}$-subproblem:

$$\begin{aligned} &\min \quad E_i(\mathbf{a}_i,\mathbf{l}_i) \\ &s.t. \quad 0 \le a_{i,k} \le 1\ ,\ 0 \le l_{i,0} \le 1\ ,\ 0 \le \mathbf{l}_i^T\mathbf{n}_i \le 1. \end{aligned} \quad (13)$$

where

$$E_i(\mathbf{a}_i,\mathbf{l}_i) = \sum_{p=1}^{N_p} w_p^i ||\mathbf{y}_p - \mathbf{a}_i\mathbf{n}_i^T\mathbf{l}_i|| + \lambda_a \sum_{j\in\mathcal{N}_i} ||\mathbf{a}_i - \mathbf{a}_j|| + \lambda_l \sum_{j\in\mathcal{N}_i} ||\mathbf{l}_i - \mathbf{l}_j||^2. \quad (14)$$

Inspired by [19], we solve this problem by applying a fixed-point scheme to find the solution of the Euler-Lagrange equation $\nabla E_i = 0$. Such an approach has the advantage of being parameter-free, compared to other descent algorithms requiring the choice or computation of a descent step. The associated iterative scheme for the albedo estimation at vertice $i$ is given by:

$$\mathbf{a}_i^{k+1} = \frac{1}{\alpha_k}\left[\sum_{p=1}^{N_p} \frac{w_p^i\,(\mathbf{n}_i^T\mathbf{l}_i)\,\mathbf{y}_p}{||\mathbf{y}_p - \mathbf{a}_i^k\mathbf{n}_i^T\mathbf{l}_i||} + \lambda_a \sum_{j\in\mathcal{N}_i} \frac{\mathbf{a}_j}{||\mathbf{a}_i^k - \mathbf{a}_j||}\right], \quad (15)$$

where

$$\alpha_k = \sum_{p=1}^{N_p} \frac{w_p^i\,(\mathbf{n}_i^T\mathbf{l}_i)^2}{||\mathbf{y}_p - \mathbf{a}_i^k\mathbf{n}_i^T\mathbf{l}_i||} + \lambda_a \sum_{j\in\mathcal{N}_i} \frac{1}{||\mathbf{a}_i^k - \mathbf{a}_j||}. \quad (16)$$

Similar equations apply for the estimation of $\mathbf{l}_i$. To satisfy the inequality constraints, we adopt a projection approach. For each variable, the obtained value is projected on the interval $[0, 1]$.

### 2.6. Albedo images

In our model, albedo and shading variables are defined on a coarse mesh rather than on a pixel grid. In some cases, the estimated albedo is over-smooth. Thus, when projecting the surface albedo into the image space, the textures are blurred and fine structures are filtered out. To cope with this limitation, an albedo image can be generated by dividing the input RGB image by the estimated shading in such a way that high frequency details are preserved when the shading is smooth. This provides a slight boost in performance when the albedo is estimated in Section 3.2.

## 3. Experimental

We evaluated and compared our method against recent state-of-the-art techniques for intrinsic decomposition of RGB-D images: Barron and Malik [2, 3], Chen and Koltun [8], and Bell et al. [4]. For these competing techniques, results were computed using their software, publicly available on their web pages. In the following, we present quantitative and qualitative evaluations of some simulation results. More results are reported in the supplementary material.

### 3.1. Single RGB-D images

The four algorithms were applied to several images selected from version 2 of the NYU Depth Dataset [21], which consists of RGB images and aligned Kinect depth maps. Comparative results are shown in Figure 3.

To apply our method on a single RGB-D image, a 3D point cloud was generated from the depth image using the intrinsic camera parameters. A surface was estimated using a Poisson reconstruction technique [12]. This technique produces smooth surfaces. Thus, no depth pre-processing (inpaiting and smoothing) was needed. However, large parts have been added through hole filling. We only kept points whose distance to the original point cloud is less than a threshold value. We cross validate the parameters to maximize performance ($\lambda_l^0 = 50$, $\lambda_l = 0.1$, and $\lambda_a = 1$). For [2] and [8], we used parameters provided by their authors as they were optimized for the NYU images.

As shown in Figure 3, our albedo images generally look better than those produced by the other techniques. Thanks to the $\ell_1$-regularization, albedo estimation is more robust to

| | MSE (albedo) | MSE (shading) | MSE (average) | LMSE (albedo) | LMSE (shading) | LMSE (average) |
|---|---|---|---|---|---|---|
| Barron and Malik [2] | 152.5 | 111.9 | 132.2 | 53.6 | 32.0 | 42.8 |
| Chen and Koltun [8] | 130.4 | 106.9 | 118.6 | 51.5 | 31.1 | 41.3 |
| Bell et al. [4] | 145.9 | 152.3 | 149.1 | 40.4 | 42.4 | 41.4 |
| Our approach | 124.5 | 75.0 | 99.7 | 40.5 | **25.2** | 32.9 |
| Our approach (Multiple images) | **95.1** | **71.3** | **83.2** | **36.5** | 26.8 | **31.7** |

Table 1. Quantitative comparison of different intrinsic image decomposition techniques: MSE and LMSE error metrics on the MIT dataset [11] ($\times 10^3$).

illumination variations and tends to be piece-wise smooth. However, our method failed in recovering fine details in the albedo in the second image (e.g., text on the wall and the fire extinguisher) due to the over-smooth albedo estimation. Furthermore, albedo results obtained with the other methods contain many artifacts and are actually very similar to the RGB images. The method of Barron and Malik is based on a hard decomposition constraint which eliminates the reflectance as a free parameter. Therefore, most of the decomposition errors are seen in the reflectance. In Chen and Koltun's method, the reflectance prior is defined on the RGB image intensities leading to a strong correlation between the albedo and the RGB images.

On the other hand, our shading results are globally consistent and less correlated to the image colors. Barron and Malik's shading estimation is strongly related to the estimated depth and a color image soft-segmentation. The results of the two other techniques lack global consistency (e.g., the bed in the first scene and the chair in the second). In addition, we applied our algorithm using multiple images containing various lighting conditions. Performances of the five algorithms are evaluated quantitatively using the standard MSE and LMSE metrics defined in [11]. Results are reported in Table 1 and Figure 4. We can see that our single view method outperforms other methods considering global metrics (MSE) and local metrics (LMSE). As expected, the multi-view method leads to even better results.

### 3.2. Quantitative evaluation on MIT dataset

To quantitatively evaluate our model, we used the MIT Intrinsic Images dataset [11], which provides ground truth albedo and illumination for images of 20 objects along with their depth produced by Barron and Malik [1]. We selected 5 images of each of the 20 objects, obtained with different lighting conditions. Our dataset therefore consists of 100 images. Some images are shown in Figure 4. As in the previous section, we manually set our parameters to $\lambda_l^0 = 50$, $\lambda_l = 1$, $\lambda_a = 0.01$. We kept the same parameters of Barron and Malik as in their implementation (optimized for MIT dataset). For Chen and Koltun [8], as there is no automatic parameter selection, we manually selected the best set of parameters ($w = (1, 0.1, 1, 1, 0.1, 1)$). We mention that for an MIT database image of size $334 \times 334$, the computational time was about 30 seconds for Bell et al. [4], three minutes for our algorithm, and more than one hour for the technique of Barron and Malik [3].

## 4. Conclusion

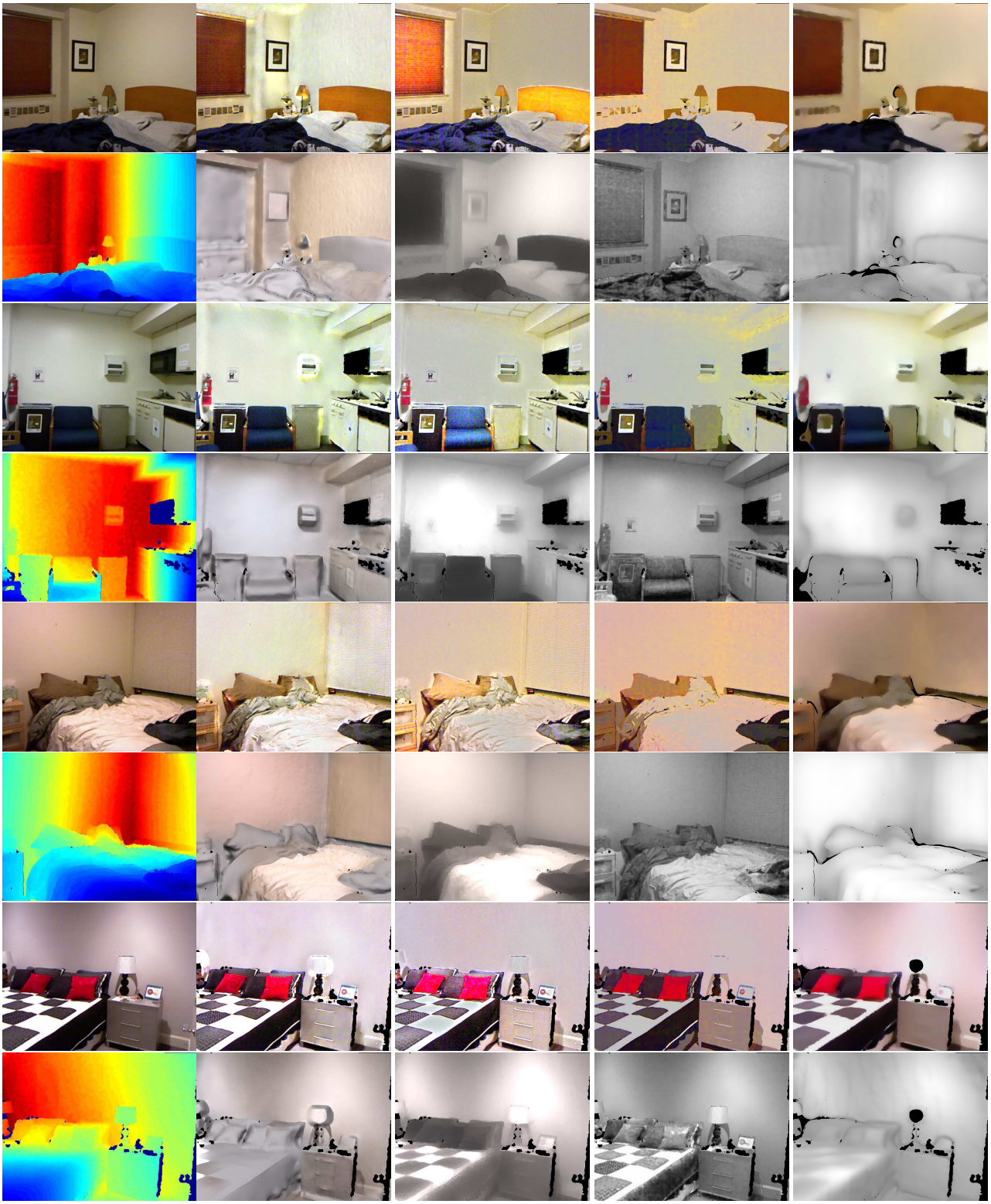
In this paper, we propose a novel approach to intrinsic RGB-D decomposition of a static scene, using one or more RGB-D images. In doing so, we extend and improve existing works dedicated to intrinsic image decomposition. The main contribution of this work is a robust method for the estimation of the intrinsic properties and an improved definition of priors, leading to better results in both albedo and shading estimation. The use of color-independent priors corrects some decomposition artifacts by avoiding mixing the color with the albedo. In addition, our method can optionally make use of multiple views to reduce view-dependent non-lambertian reflection artifacts encountered in single RGB-D image decomposition methods. However, our method relies on a good surface reconstruction and normal estimation. Incorrect and missing depth values limit the performance of our method in some images. Thus, for future work, we would like to investigate learning-based approaches to improve and inpaint depth measurements. Furthermore, we believe that our proposed model can be used in conjunction with or incorporated in popular RGB-D SLAM methods (e.g. Kinect Fusion [22, 32]) to produce improved 3D geometry, along with, color and lighting reconstruction of the scene. In fact, our intrinsic decomposition model can be viewed as another regularization term in the SLAM framework.

## Acknowledgement

Research reported in this publication was supported by competitive research funding from King Abdullah University of Science and Technology (KAUST) with grant number OCRF-2014-CRG3-62140401 and by a post-doctoral fellowship from the Saudi Arabia Basic Industries Corporation (SABIC).

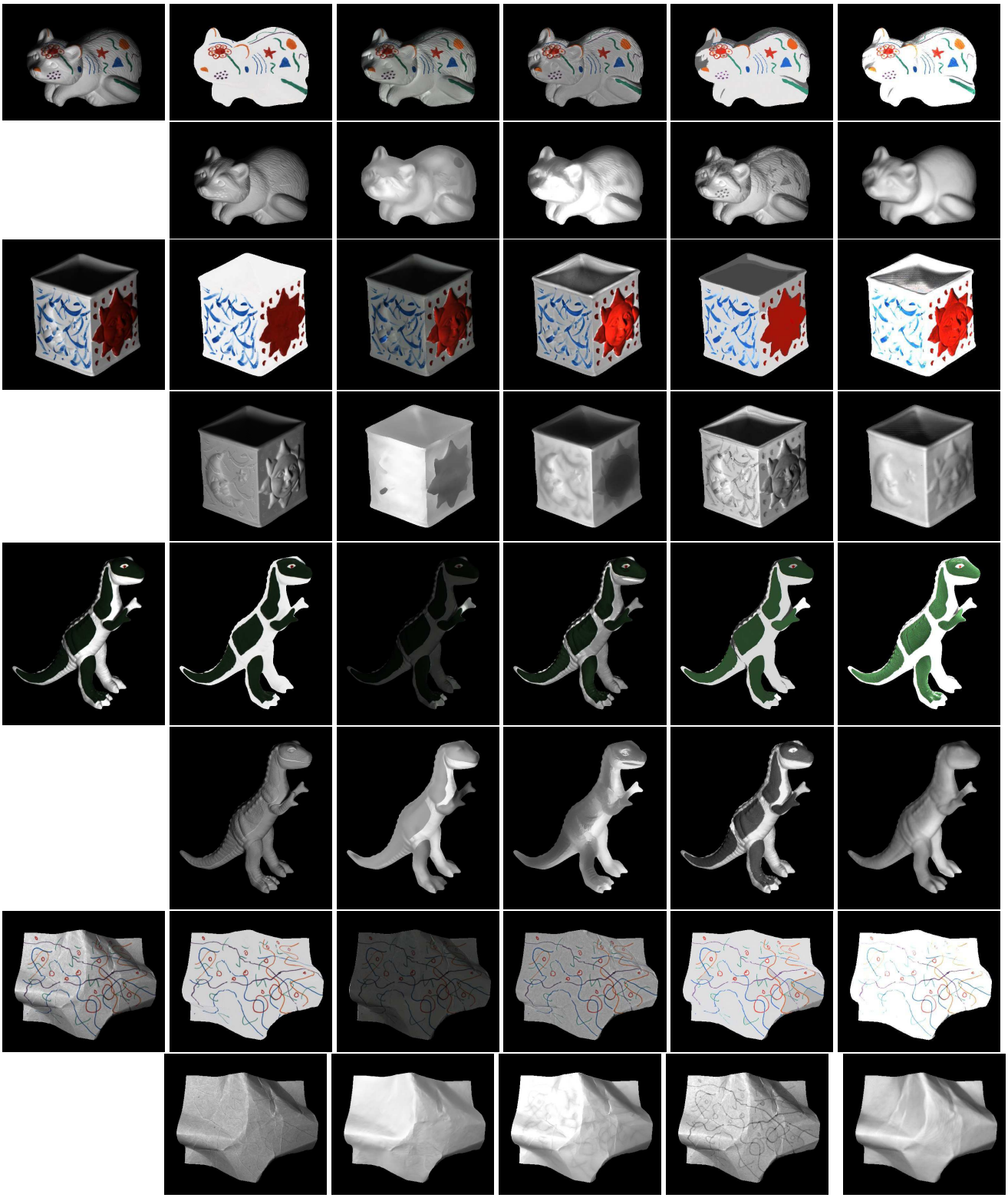
## References

[1] J. T. Barron and J. Malik. Shape, albedo, and illumination from a single image of an unknown object. In *IEEE Con-*



(a) RGB and depth inputs (b) Barron and Malik[2] (c) Chen and Koltun [8] (d) Bell et al. [4] (e) Our approach

Figure 3. Intrinsic decomposition of a single RGB-D image: qualitative comparison of different methods (albedo (top) and shading (bottom)). Results for the competing methods were obtained using their software obtained from their web pages. The RGB-D images were selected from version 2 of the NYU Depth Dataset [21].



(a) Input (b) Ground truth (c) Barron-Malik [3] (d) Chen-Koltun [8] (e) Bell et al. [4] (f) Our approach

Figure 4. Intrinsic decomposition of selected images from the MIT dataset [11]. (a) Input RGB image. (b) Ground truth albedo and shading images. (c-e) Intrinsic decomposition obtained by state-of-the-art techniques ([3], [8], and [4]). (f) Intrinsic decomposition obtained by our approach.

*ference on Computer Vision and Pattern Recognition*, 2012. 6

[2] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013. 1, 2, 5, 6, 7

[3] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. Technical Report UCB/EECS-2013-117, EECS, UC Berkeley, May 2013. 1, 2, 5, 6, 8

[4] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM Trans. Graph.*, 33(4), 2014. 1, 2, 5, 6, 7, 8

[5] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister. Interactive intrinsic video editing. *ACM Transactions on Graphics*, 33(6), 2014. 4

[6] A. Bousseau, S. Paris, and F. Durand. User assisted intrinsic images. *ACM Transactions on Graphics*, 28(5), 2009. 1

[7] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004. 4

[8] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *IEEE International Conference on Computer Vision*, pages 241–248, Dec 2013. 1, 2, 5, 6, 7, 8

[9] D. Forsyth. Variable-source shading analysis. *Int. J. Comput. Vis.*, 91(3):280–302, 2011. 4

[10] P. Gehler, C. Rother, M. Kiefel, Z. Lumin, and B. Schoelkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *Neural Information Processing Systems (NIPS)*, December 2011. 1

[11] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground-truth dataset and baseline evaluations for intrinsic image algorithms. In *International Conference on Computer Vision*, pages 2335–2342, 2009. 2, 6, 8

[12] M. Kazhdan and H. Hoppe. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3):29:1–29:13, July 2013. 3, 5

[13] R. Kimmel, M. Elad, D. Shaked, R. Keshet, and I. Sobel. A variational framework for retinex. *Int. J. Comput. Vision*, 52(1):7–23, Apr. 2003. 1

[14] N. Kong, P. V. Gehler, and M. J. Black. Intrinsic video. In *European Conference on Computer Vision*, volume 8690, pages 360–375, 2014. 1

[15] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis. Coherent intrinsic images from photo collections. *ACM Transactions on Graphics*, 31, 2012. 2

[16] E. H. Land and J. McCann. Lightness and retinex theory. *J. Opt. Soc. Am.*, 61(1):1–11, Jan 1971. 1

[17] K. Lee, Q. Zhao, X. Tong, M. Gong, S. Izadi, S. Lee, P. Tan, and S. Lin. Estimation of intrinsic image sequences from image+depth video. In *European Conference on Computer Vision*, volume 7577, pages 327–340, 2012. 2

[18] Y. Li and M. S. Brown. Single image layer separation using relative smoothness. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2014. 1

[19] Y. Lipman, D. Cohen-Or, D. Levin, and H. Tal-Ezer. Parameterization-free projection for geometry reconstruction. *ACM Trans. Graph.*, 26(3), July 2007. 5

[20] B. F. Mark, M. S. Drew, and M. Brockington. Recovering shading from color images. In *ECCV-92: Second European Conference on Computer Vision*, pages 124–132. Springer-Verlag, 1992. 1

[21] P. K. Nathan Silberman, Derek Hoiem and R. Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5, 7

[22] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *IEEE ISMAR*. IEEE, October 2011. 6

[23] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, pages 497–500, New York, NY, USA, 2001. ACM. 3

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

[25] 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*, Washington, DC, USA, 2011. IEEE Computer Society. 1

[26] M. F. Tappen, E. H. Adelson, and W. T. Freeman. Estimating intrinsic component images using non-linear regression. In *Proceedings of the Computer Vision and Pattern Recognition*, pages 1992–1999, 2006. 1

[27] M. F. Tappen, W. T. Freeman, and E. H. Adelson. Recovering intrinsic images from a single image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(9):1459–1472, Sept. 2005. 1

[28] Y. Weiss. Deriving intrinsic images from image sequences. In *IEEE International Conference on Computer Vision*, volume 2, pages 68–75 vol.2, 2001. 1, 2

[29] T. Weyrich, J. Lawrence, H. Lensch, S. Rusinkiewicz, and T. Zickler. Principles of appearance acquisition and representation. *Foundations and Trends in Computer Graphics and Vision*, 4(2):75–191, 2008. 3

[30] G. Ye, E. Garces, Y. liu, Q. Dai, and D. Gutierrez. Intrinsic Video and Applications. *ACM Trans. Graph. (SIGGRAPH)*, 33(4), 2014. 1

[31] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1437–1444, July 2012. 1

[32] Q.-Y. Zhou and V. Koltun. Dense scene reconstruction with points of interest. *ACM Trans. Graph.*, 32(4):112, 2013. 6