Intrinsic Image Decomposition with Step and Drift Shading Separation

Bin Sheng, Ping Li, Yuxi Jin, Ping Tan, and Tong-Yee Lee, *Senior Member, IEEE*

Abstract—Decomposing an image into the shading and reflectance layers remains challenging due to its severely under-constrained nature. We present an approach based on illumination decomposition that recovers the intrinsic images without additional information, e.g., depth or user interaction. Our approach is based on the rationale that the shading component contains the step and drift channels simultaneously. We decompose the illumination into two channels: the step shading, corresponding to the sharp shading changes due to cast shadow or abrupt shape changes; the drift shading, accounting for the smooth shading variations due to gradual illumination changes or slow shape changes. Due to such transformation of turning the conventional assumption that shading has smoothness as reasonable prior, our model has the advantages in handling real images, especially with the cast shadows or strong shape edges. We also apply a much stricter edge classifier along with a reinforcement process to enhance our method. We formulate the problem using a two-parameter energy function and split it into two energy functions corresponding to the reflectance and step shading. Experiments on the MIT dataset, the IIW dataset and the MPI Sintel dataset have shown the success of our approach over the state-of-the-art methods.

Index Terms—Intrinsic images, illumination decomposition, step shading, drift shading, shading edge constraint

1 Introduction

Intrinsic image recovery has been a long-standing problem ever since it was proposed in [1], which aims to separate an observed image into its reflectance (albedo) and shading components. It is a fundamental and significant problem in both computer vision and computer graphics, which has various applications, e.g., shape from shading, retexturing, and material editing [2], [3], [4], [5], [6], [7]. Decomposing an image into two components (the shading image and the reflectance image) is still a challenging problem [8], [9], even though it has been studied for a long time, because of its severely under-constrained nature. Besides, researchers have been looking for an easy-to-use and solid method to handle intrinsic images. Nevertheless, either inefficient or tedious methods can be eventually obtained. Compared with the user-assistant and auxiliary information needed methods, single-image based scheme has the prominent advantage of easy use. However, by far there are rarely effective and reliable single-image based methods being widely used due to lack of valid priors, especially for the smoothness assumption. The single-image based intrinsic images problem has hence become a significant research topic.

In the past few decades, various methods have been developed. Some approaches use an automatic method with only a single image as input, e.g., [10], [11], [12], [13], [14]. Others use image sequences to obtain the intrinsic image, e.g., [15], [16], [17], [18], [19]. Furthermore, learning-based methods are emerging to tackle such recovery problems caused by shading or reflectance [20], [21], [22], [23]. However, considering the difficulty and complexity of obtaining image sequences or user assistance, deep learning based methods need large volume of synthesized datasets to train [21], [22]. In contrast, we concentrate on recovering the intrinsic images from a single image without any other additional information. Previous single-image based methods often adopt a strong but simple assumption that the shading is as smooth as possible over the image, which is insufficient for obtaining accurate estimation. This smoothness assumption breaks down because of the large changes in shading values caused by shadows and abrupt shape changes, which are common in real-world images. Shadows and shape discontinuities increase the ambiguity of shading and reflectance variations, and it is difficult to decompose images that contain them. Thus, we must elaborately build a practical prior to turn the simple smoothness assumption into rigid certainty or determination in the form of illumination decomposition.

To deal with quick shading changes, we further divide the shading component into two channels: the step shading, accounting for quick changes due to shadow or abrupt shape changes; and the drift shading, representing smooth changes on slow variations in illumination or shape. According to these channels, the pixels in a local region of an object have the same step shading values and only differ with respect to their drift shading values, which make the shading

- *B. Sheng is with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail: shengbin@sjtu.edu.cn.*
- *P. Li and Y. Jin are with the Faculty of Information Technology, Macau University of Science and Technology, Macau 999078, China. E-mail: pli@must.edu.mo, yuxijin1808@gmail.com.*
- *P. Tan is with the School of Computing Science, Simon Fraser University, Burnaby, BC V5A 1S6, Canada. E-mail: pingtan@sfu.ca.*
- *T.-Y. Lee is with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan 70101, Taiwan. E-mail: tonylee@mail.ncku.edu.tw.*

Manuscript received 3 Feb. 2018; revised 28 Aug. 2018; accepted 4 Sept. 2018. Date of publication 10 Sept. 2018; date of current version 3 Jan. 2020. (Corresponding author: Bin Sheng.) Recommended for acceptance by K. Zhou. Digital Object Identifier no. 10.1109/TVCG.2018.2869326

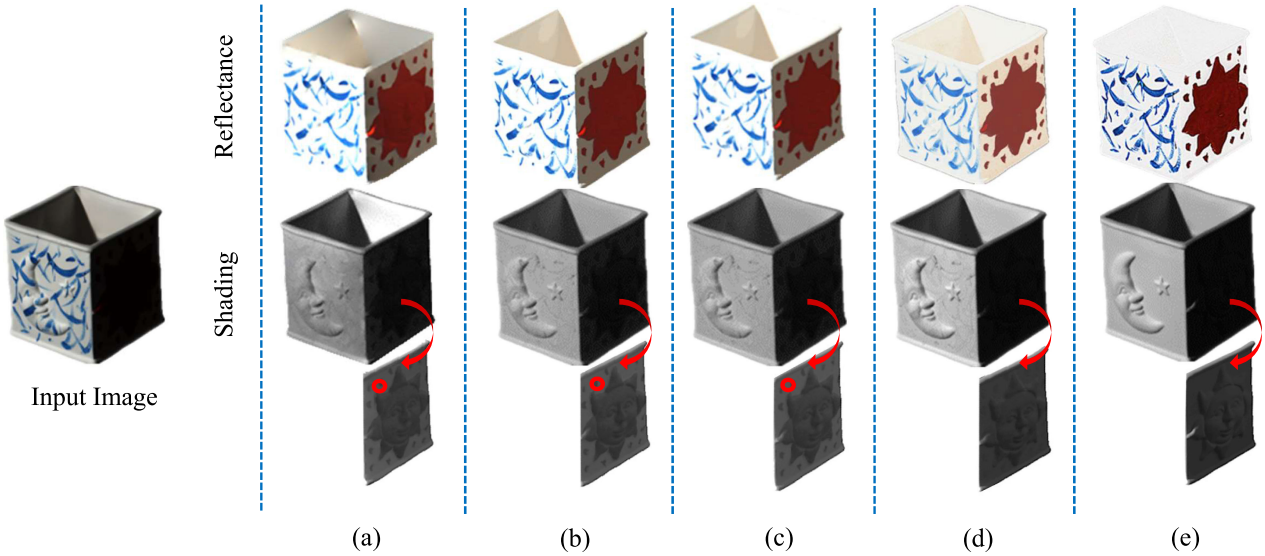


Fig. 1. Comparison of intrinsic image recovery via strong constraints on the reflectance and shading. (a) The Retinex method, (b) the Retinex method combined with a smooth shading prior, (c) the method of [10], which uses a global reflectance prior in addition to the Retinex algorithm and smooth shading prior, (d) our approach, and (e) the ground truth of input image. Desired results support our idea that illumination can be handled by decomposition. (We lightened the dark region of the shading images for visual comparison.)

look smooth. Following this assumption, we can decompose the shading image into two images, namely, a step-shading component image, and a drift-shading component image. Separating the step-shading component from the original shading image, the smoothness assumption holds regardless of shadows or shape discontinuities in image. To formulate our step shading component, we design a shading edge function that detects large lighting derivatives in the image. Different from the Retinex method [24] and other Retinex-based methods, such as [10] and [25], we redefine that large lighting variance coupled with a chromaticity change which is smaller than a certain threshold is considered as an edge in the step shading. By decomposing the shading into two components, our target can be formulated as an energy function optimization problem. We evaluate our method on the MIT dataset [25], the IIW (Intrinsic Images in the Wild) dataset [17], and the MPI Sintel dataset [26]. Experiments verify that our approach outperforms the state-of-the-art intrinsic image recovery methods, both in quantitative and qualitative measurements. Our work on automatic intrinsic image decomposition makes the following three main contributions:

- We decompose shading image into drift shading and step shading channels, which allows appropriate and feasible smoothness priors to different lighting channels, instead of simple and unfaithful assumption by previous methods.
- A function based on a strict classification that distinguishes large shading variation from reflectance formulates the step shading component, which performs better than other Retinex-based methods. Moreover, a reinforcement process has been applied, when dealing with complex scenes.
- We elaborately combine several reasonable priors together to constrain the ill-posed problem, which makes the model itself more effective and convincing.

2 Related Work

Land and McCann [24] studied the human visual system and determined how the light reaches our eyes from objects. They proposed the basic Retinex (Retina and Cortex) algorithm based on a color consistency theory in an ideal Mondrian world. In the Retinex algorithm, large gradients of illumination or chromaticity are considered to be reflectance boundaries. [27] and [28] expanded this algorithm to include two dimensions and enabled the processing of real images. Their method was subsequently applied to color images. For example, [29] applied this approach to remove shadows and obtained a shadow-free image. Other work, such as [30], recovered an intrinsic image from a single image using a strong shading assumption, and classified large chromaticity variances as reflectance changes. Most automatic intrinsic image decomposition methods from a single image are based on Retinex. Some methods used a classifier trained to distinguish reflectance changes from shading changes [12], [13], [20]. Some approaches used additional strong constraints on shading, reflectance, or both (see Fig. 1). Rother et al. [10] used a new prior on reflectance that assumes the reflectance value is sparse to obtain high-quality results without edge information. Shen et al. [11], [31] assumed that similar chromaticity leads to the same reflectance and developed a reflectance sparse prior. Building on [11], Zhao et al. [32] constrained distant pixels with the similar texture to have the same reflectance using texture analysis. Li and Brown [33] built two likelihoods for two layers from the gradient histograms without using a prior on the reflectance. In contrast, we propose a new approach for dealing with the intrinsic image problem of single-image based scheme, which leverages on illumination decomposition. Our method turns the simple assumption used by the previous work that shading is smooth into practical and reasonable prior so that we could achieve better results.

Some researches have relied on image sequences or user assistance to resolve the ambiguities relating to variation

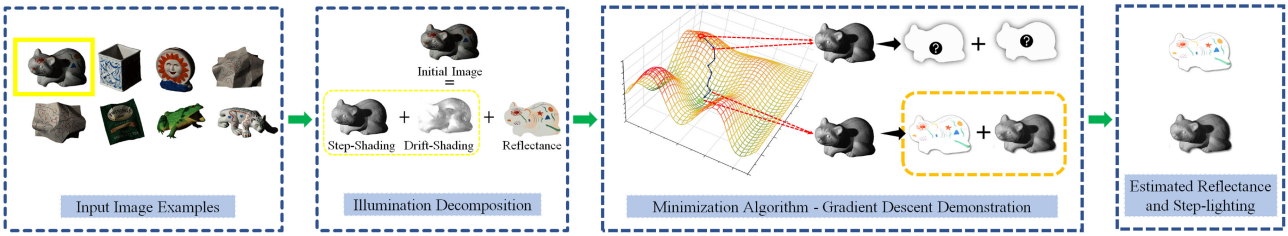


Fig. 2. An overview of our illumination decomposition based method. The input images are restrained by several priors, especially those from our illumination decomposition model. After several iterations by the minimization method, the energy function gets the optimal values, the estimated reflectance and shading images are hence obtained accordingly.

across pixels. Ambient occlusion computed by a stack of images was used to replace a shading smoothness prior in intrinsic decomposition in [34]. Weiss [15] captured the intrinsic image from a sequence of images with a constant reflectance but different illumination conditions. Matsushita et al. [16] derived the intrinsic image from a fixed-viewpoint image sequence by changing the illumination direction. Furthermore, approaches that use multiple viewpoints of a scene have also been developed, e.g., [35], [36]. Bousseau et al. [37] proposed a user-assisted method that focuses on diffuse objects and assumes that local reflectance changes lie on a plane. However, their method is incompatible with multi-colored surfaces [31]. Bell et al. [17] decomposed real-world images in the wild with user-annotated pairs of points that have relative reflectance. Chen and Koltun [38] used depth cues to reconstruct the shading, but the method requires RGB-D images, which are captured by a special camera, such as Microsoft Kinect. Hachama et al. [39] also used depth cues to facilitate intrinsic image recovery by minimizing an energy function consisting of a data term and a regularity term. Compared with the above mentioned methods, our approach only uses a single image and has advantages of usability, simplicity, and the most important thing, feasibility. Because it is unlikely or difficult to obtain extra images to construct the sequence of images or auxiliary information such as depth cues under some conditions. Furthermore, Ye et al. [18] addressed the intrinsic video problem. Their method has two steps: decomposing the initial frame by extending the method of Zhao et al. [32], and propagating the clustered result of the reflectance in the temporal domain. Shen et al. [40] and Lee et al. [19] have also looked at the intrinsic video problem. Given the essence of our method, it is more suitable for handling key frames rather than the whole video, and we will deal with intrinsic video problem by a model of sequence of images scheme in the future.

In recent years, the learning-based methods, especially deep learning based approaches are being proposed to address the intrinsic image problems. Lettry et al. [21] put up with a method based on deep supervised learning, which is composed of the combination of deep convolutional network and several generative adversarial networks. The adversarial loss (trained discriminator networks) gives feedbacks to generative convolutional neural network (CNN) to amend target albedo or shading. Narihira et al. [22] worked out such problems by a similar deep convolutional network, which directly predicts objective reflectance and shading from the input image patches. Zhou et al. [23] proposed a method based on two-stage learning problem, which first produces a data-driven albedo prior, and then integrates this prior to a minimization formulation. Moreover, Fan et al. [41] put forward a deep neural network structure that focuses on finding edge information used to address problems sensitive to them, such as single-image reflection removal and image smoothing. However, learning-based methods are thoroughly dependent on the large volume of training datasets and an abundant amount of computing resources are essential. In addition, because of the complexity of illumination conditions in diverse scenes, overfitting is common.

Unlike the previous methods, we use a refined and enhanced Retinex-based method, which does not require a huge volume of datasets to be trained on or any other auxiliary means. We just focus on recovering intrinsic components from a single image with no extra information by using effective and constructive priors, which demonstrates the previous smoothness assumption with determined illumination decomposition model. Thus, our method has advantages of simplicity, usability and feasibility over previous work. Moreover, we positively apply a concrete model in the form of prior to explain and back up the smoothness assumption, while other researches just rely on a simple assumption. To simplify, we follow the common practice that assumes the light source is white.

3 Approach

Fig. 2 shows the pipeline of our approach. Given an input image, we first formulate an intrinsic image model based on our illumination decomposition assumption, then we get the estimated reflectance and step-shading images of the input image by using gradient descent demonstration.

3.1 Illumination Decomposition Model

Mathematically, a light map is a piecewise continuous function, which can be seen as a combination of step component and smooth component. If we can separate the step part from the illumination map, the assumption on smooth part can be more reasonable. Through observation, we find that the step component of the illumination is usually caused by the dramatic change of lighting shadow and object geometry shape, while the slow change of object geometry shape does not lead to dramatic light gradient. For simplicity purpose, we assume that the illumination model of the objects conforms to the Lambert illumination model. Besides, we propose a new illumination decomposition model to separate the step information and smooth information so that the illumination smooth assumption can be more reasonable.

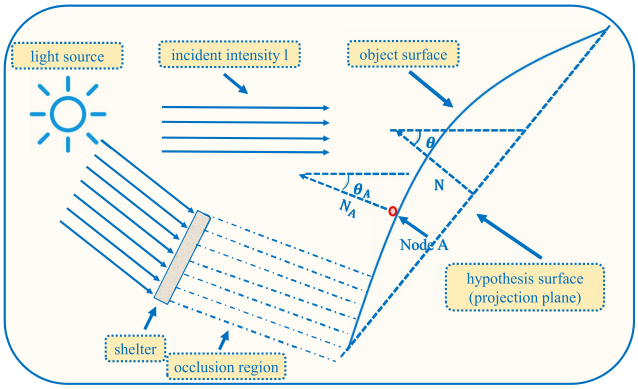


Fig. 3. Our illumination decomposition model. Lights come from the source are classified into the incident light and occluded light. Incident light is further decomposed to the drift lighting and the step lighting, according to our illumination decomposition model.

For an object in the scene, we can divide its surface into different regions according to the normal change by segmentation methods. Here, we take one of those divided regions as an example to explain our illumination decomposition model in detail (see Fig. 3). Since the number of pixels in a region is finite, we can find a pixel which is brightest and select a plane that is perpendicular to the normal direction of the brightest pixel as the main plane of this region. In the way, we assume that light source is infinity, thus the main plane can be one of myriad parallel planes which has no effect on the decomposition result. Randomly choosing a pixel A in the chosen area (see Fig. 3), we can see that the light intensity of pixel A could be computed as follow:

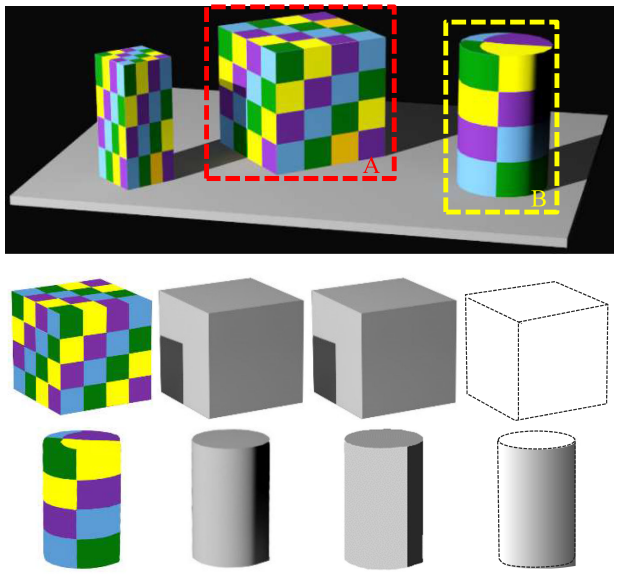
$$L_A = \varepsilon l \cos\theta_A, \tag{1}$$

where, L_A represents illumination intensity of pixel A, l indicates intensity of light source, θ_A is the included angle between the incident light and the normal of pixel A, and ε is the occlusion factor. If pixel A is absolutely sheltered, ε is 0, otherwise, ε is 1.

We conduct the following mathematics transformation to make the concept of step shading and drift shading clearer:

$$L_A = (\varepsilon l \cos\theta)(\cos\theta_A / \cos\theta), \tag{2}$$

where, θ is the included angle between the incident light and the normal of the main plane. $\varepsilon l \cos\theta$ indicates the exposure intensity of the source light on the main plane. $\cos\theta_A / \cos\theta$ is a ratio between 0 and 1, representing the position variations of pixels in the specific area relative to the main plane. Pixels belonging to different regions have different $\varepsilon l \cos\theta$ values, which are discrete and step variables, while pixels belonging to the same region have the same $\varepsilon l \cos\theta$ value. Since the geometric surface of an object changes slowly, $\cos\theta_A / \cos\theta$ is a continuous and smooth variable. Thus, we define $\cos\theta_A / \cos\theta$ as the drift-shading component of illumination to represent the smoothness character of illumination, and $\varepsilon l \cos\theta$ as the step-shading component of illumination to indicate the saltation character of illumination. From the definition of the components, it can be seen that the local lighting of an object is continuous and smoothly changing with the same step component and different drift component, while both the step component and drift component of pixels in different areas are different. If explained in the perspective of visual effects, the step component makes the bright gradation of the object light more prominent, and the drift component makes the object light softer.



(a) Reflectance (b) Shading (c) Step shading (d) Drift shading

Fig. 4. Demonstration of our intrinsic image decomposition method on a simulated simple scene containing one cuboid, one cube, and one cylinder. Our method decomposes the input image into four images: (a) reflectance, (b) shading, (c) step shading, and (d) drift shading. The step-shading and drift-shading results are ideal decompositions according to our shading assumption. This is one of the main contributions of our approach.

To illustrate the two components of illumination much clearer, we simulate a scene of one cuboid, one cube and a cylinder with single light source by Autodesk 3ds Max (see Fig. 4). The scene contains obscured shadows, self-shielding due to dramatic changes in geometry and smooth surface change. Three visible surfaces and shadow on the left side make cube A in Fig. 4 has four different step component values, which can be clearly seen from the second row of Fig. 4c. Since the three surfaces of the cube are parallel to their main plane, the drift component value of cube A is 1, which is white in visual, and we mark the edges of cube A with dotted lines for clarity. The cylinder B has three step component values, namely the top, the left side, and the right side (see Fig. 4c), because its side changes smoothly. We can see from Fig. 4d that the drift component of cylinder B appears as a gradual change process from bright to dark from left to right. One point has to be mentioned is that all changes in the simulated scene are assumed to be ideal for clearly explaining our light decomposition model. According to our definition of step and drift components, our illumination decomposition model could be formulated as:

$$S = S^s \cdot S^d, \tag{3}$$

where, S denotes the illumination image, S^s and S^d indicate the step-shading and drift-shading components, respectively. Since the product of the two components can be rewritten to addition form in the logarithmic domain, we rewrite Eq. (3) as following:

$$\log S = \log\ S^s + \log\ S^d. \tag{4}$$

3.2 Intrinsic Image Decomposition Model

Based on the aforementioned illumination decomposition model, we propose a decomposition approach for intrinsic images, which could well reduce the influence of the shadow on the decomposition result in an image, making the illumination decomposition map more accurate. Besides, our illumination decomposition model can be used as a framework to improve the decomposition accuracy of other single-image decomposition methods. The intrinsic image problem comes down to the problem of finding a illumination decomposition map and an albedo decomposition map for a known image. It is well known that what human see is the comprehensive effect result of factors like illumination and object material. Thus, the image seen by human eyes could be described as the product of the illumination decomposition map and the albedo decomposition map, which is mathematically defined as follow:

$$I = S \cdot R, \tag{5}$$

where, I, S, $R \in \mathbf{R}^3$ are vectors in RGB space. We use I_i, S_i and R_i to indicate the pixel observed at location i of the input image, the shading image and the reflectance image, respectively. For simplicity, we assume the light source is white, i.e., $S = s(1, 1, 1)$, and we define the reflectance component as $R = r\,\vec{R}$. S and R are scalars, indicating the intensity of light and albedo, respectively. $\vec{R}$ can be viewed as the normalized vector for color in RGB space. Thus, we get the final mathematical model for pixel i as:

$$I_i = s_i \cdot r_i \cdot \vec{R_i}\,. \tag{6}$$

For an image with N pixels, we can get $3N$ equations. However, there are $4N$ unknown variables, which is mathematically known as ill-posed problem. According to the illumination decomposition proposed in Section 3.1, Eq. (6) can be rewritten as:

$$I_i = s_i^s \cdot s_i^d \cdot r_i \cdot \vec{R_i}, \tag{7}$$

Eq. (7) is our intrinsic image model. By performing a modulus on both sides, we can get a relation between illumination and albedo:

$$\|I_i\| = s_i^s \cdot s_i^d \cdot r_i. \tag{8}$$

As mentioned in Section 3.1, converting multiplication into addition by logarithmic operation could simplify calculation. Thus, we rewrite Eq. (8) as follow:

$$\log \|I_i\| = \log s_i^s + \log s_i^d + \log r_i. \tag{9}$$

And for the simplicity of notation, we use $\|LI_i\|$, Ls_i^s, Ls_i^d and Lr_i to represent $\log \|I_i\|$, $\log s_i^s$, $\log s_i^d$ and $\log r_i$, respectively.

3.2.1 Edge Detection

To distinguish between the two components of illumination, we design a function φ_{ij} to detect large changes of light gradients as:

$$\varphi_{ij} = \begin{cases} 1, & \text{if}(\|G_{ij}^g\| > \theta^g) \wedge (\|G_{ij}^t\| < \theta^{t0}) \\ 0, & \text{otherwise.} \end{cases} \tag{10}$$

Fig. 5. The real scene images from [17] demonstrate that there are several levels of step shading in a plane because of the complexity of the environment. Therefore, it is innovative and indispensable to consider the decomposition of step shading and drift shading, which benefits trouble-shooting.

Through averaging the sum of the input image's three channel pixel values, and normalizing the color of the input image, respectively, we get the gray-scale image I^g and the tone image I^t. For two adjacent pixels i and j, the brightness gradient G_{ij}^g and the tone gradient G_{ij}^t are defined as $G_{ij}^g = |I_i^g - I_j^g|$ and $G_{ij}^t = \|I_i^t - I_j^t\|_2$. Here, $\|*\|_2$ represents the L2 norm, and $|*|$ indicates the absolute value. If the brightness gradient G_{ij}^g is larger than the threshold θ^g and the tone gradient G_{ij}^t is smaller than the threshold θ^{t0}, it could be inferred that there exists a shadow boundary between pixels i and j, namely $\varphi_{ij} = 1$, otherwise $\varphi_{ij} = 0$ which means the light between pixels i and j varies smoothly and slowly.

For simple scenes with single objects, φ_{ij} works well and shading edge can be easily extracted. However, for complex scenes with multiple objects overlapping, extra shading edge detection method should be brought in to amend φ_{ij}. As shown by the cylinder in Fig. 4, the side facing the light source and the opposite side have different values of step shading, but share the same smooth drift shading. These two components account for the illumination formation. According to the definition of the two shading channels, the local shading regions all have the same step-shading values and only differ in terms of drift-shading values. In other words, the step-shading values of a local shading region should be equal to each other, whereas the drift-shading component should differ. From the visual aspect, step shading produces large visual changes in an object and drift shading makes the shading of an object look smooth. In Fig. 4, three square faces of the cube have different step shading because of its shape discontinuity. The marked regions of Fig. 5 have differences in the step-shading channel because of complex illumination. Region 2 of Fig. 5(left) is the innermost and darkest part of a shadow, where the light source is completely blocked by the occluding body, namely umbra. Observers in region 2 experience a total eclipse. Region 3 of Fig. 5(left) can be seen as penumbra, in which only a portion of the light source is obscured by the occluding body. Observers experience a partial eclipse. While region 1 of Fig. 5(left) is completely exposed to light.

We introduce the method of [42] to find edges in complex scenes. We apply the pre-trained edge detector of [42] to find edges in an image with complex scenes, and use φ_{ij} to detect shading edges. Then, the results from φ_{ij} are checked by those from method of [42], which is a cross-contrast reinforcement process. If an edge detected by φ_{ij} is also seen as an edge by the edge detector of [42], it could be retained as clean shading edges, otherwise, it should be reconsidered as suspensive edge. Since we only use the well-trained edge detector to detect edges, the training procedure of the

Fig. 6. The benefits of amending φ_{ij} using [42]. Top row: the input image. Middle row: the reflectance image and shading image obtained without enhancement of [42]. Bottom row: the reflectance image and shading image obtained with the enhancement of [42].

edge detector will not be further discussed in this paper. Further training detail can be seen from [42]. In Fig. 6, we can experience the benefits by comparing the reference image (reflectance) and the obtained reflectance image by our method. The reference image is achieved by our approach without using edge detector of [42]. Coarse-grained boundary information is largely retained in the reference image, which is undesired. In contrast, our obtained reflectance image enhanced by [42] looks fine and smooth. We can conclude that more detailed and fine-grained edge information can be found and retained by the edge detector of [42], which makes the results more promising.

3.3 Energy of Intrinsic Image Decomposition

Through observation, we know that assumptions and additional constraints for global albedo and texture can make the solution results much closer to the ground truth. Thus, we solve the intrinsic image problem by computing the step shading values $s_1^s, s_2^s, \ldots, s_N^s$ and the reflectance intensities $r_1, r_2, \ldots, r_N$ of all N pixels in the image. All constraints are linearly combined and the energy function is formulated as follow:

$$E = \underset{\substack{Lr_1,Lr_2,\ldots,Lr_N \\ Ls_1^s,Ls_2^s,\ldots,Ls_N^s}}{\arg\min} \; w_1E_s + w_2E_e + w_3E_r + w_4E_g + w_5E_n. \quad (11)$$

The terms of the energy function are described in detail in the following contents. We use the weights $w_1, w_2, \cdots, w_5$ to balance the importance of each term of the energy function.

Smoothness of Drift Shading. The first term of the energy function is the shading smoothness constraint E_s. This term penalizes the large shading derivatives over the entire image. In other previous methods, such as [10], the shading smoothness constraint is formulated as follow:

$$E_s = \sum_{i,j\in p(i)} ((\|LI_i\| - Lr_i) - (\|LI_j\| - Lr_j))^2, \quad (12)$$

where $p(i)$ denotes the 4-neighbor group of pixel i. However, this assumption does not always hold for all conditions, e.g., shadows and surface discontinuities. Thus, we make this assumption more applicable and practical via illumination decomposition.

Fig. 7. Left: the input image. Middle: the shading image by using the traditional shading smooth term. Right: the shading image by using our smooth item E_s.

$$E_s = \sum_{i,j\in p(i)} ((\|LI_i\| - Lr_i - Ls_i^s) - (\|LI_j\| - Lr_j - Ls_j^s))^2. \quad (13)$$

To illustrate the advantages of our smoothness constraint, we compare the shading values with the ground truth, which vary largely at the shadow edges, but the drift-shading component appears smoother than the ground truth. The simple smoothness assumption on the shading image of the other methods clearly breaks down when shadows or other large shading derivatives exist. In contrast, the model is more reasonable in our drift-shading image as the large derivatives are removed from the drift-shading due to decomposition of the shading image into different components. We extract the shading image of the input image by using the traditional shading smooth term and E_s, respectively. Fig. 7 shows the shading results of using the traditional shading smooth term and our smooth term E_s. Since we take the step shading into consideration, the detail of book shelf can be well maintained by our method. While in contrast, the result of using traditional shading smooth term appears badly in the region marked by the red rectangle. One point has to be mentioned is that we make equivalent transformation like $Ls_i^d = \|LI_i\| - Lr_i - Ls_i^s$ over Eq. (9) to ensure the final energy function only has two unknowns.

Smoothness of Step Shading. Inspired by the work of Rother et al. [10], the second term of the energy function is the shading edge constraint E_e as:

$$E_e = \sum_{i,j\in p(i)} (Ls_i^s - Ls_j^s - \varphi_{ij} \times (s_i - s_j))^2. \quad (14)$$

where, s_i and s_j are values of the input image's gray-scale image at pixels i and j, Ls_i^s and Ls_j^s are the values of the step components of our final light resolution map at pixels i and j, and φ_{ij} is the boundary detection function defined in Section 3.2.1. In the extreme case, if $E_e = 0$ or the energy of illumination boundary constraint term is the minimum, we can get $Ls_i^s - Ls_j^s = \varphi_{ij} \times (s_i - s_j)$ from Eq. (14). Assume there is no illumination boundary between pixels i and j (namely $\varphi_{ij} = 0$), then $Ls_i^s - Ls_j^s = 0$, which means there is no light boundary between pixels i and j in the illumination decomposition map. Conversely, if there is a light border between pixels i and j (namely $\varphi_{ij} = 1$), then $Ls_i^s - Ls_j^s = s_i - s_j$, which means there is also a light border between pixels i and j in the light decomposition map. Besides, the gradient of illumination is unchanged. The term E_e is aimed at maintaining the detected light border information in the final light decomposition map. This term formulates the step-shading component and only needs few iterations to reach the minimum. In the minimization in Section 3.4, we use the gray-scale image to initialize Ls^s. The energy of the shading edge term rapidly decreases after the first iteration and only

dozens of iterations are needed in total. This term removes large shading derivations from shading image and makes the smoothness assumption on the drift shading reasonable.

Color Retinex Constraint. The third term of the energy function is the color Retinex constraint E_r. Retinex theory believes that excessive tonal gradient changes and brightness changes are caused by changes in albedo. However, this classification method misclassifies images with shadows. If an image's surface is half shaded and half not shadowed, its brightness gradient may exceed the threshold, but in fact its albedo does not change. Based on [10], we further design a more rigorous and practical classification method to distinguish the light and albedo as:

$$\omega_{ij} = \begin{cases} 1, & \text{if}(\|G^g_{ij}\| > \theta^g) \wedge (\|G^t_{ij}\| > \theta^{t0}) \\ 1, & \text{if} \|G^t_{ij}\| > \theta^{t1} \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where, ω_{ij} is a Retinex operator and represents the result of classification. $\|G^g_{ij}\|$ and $\|G^t_{ij}\|$ are the luminance and tone gradients that have been defined previously. θ^g is the luminance gradient threshold, θ^{t0} and θ^{t1} are the chromaticity gradient threshold.

If the tone gradient value exceeds the threshold θ^{t1}, then we consider the albedo of pixels i and j to be changed, which means $\omega_{ij} = 1$. If the luminance gradient value is larger than the threshold θ^g and at the same time the tone gradient value is larger than the threshold θ^{t0}, it is assumed that the albedo of pixels i and j has changed, which means $\omega_{ij} = 1$; otherwise, the albedo of pixels i and j does not change, namely $\omega_{ij} = 0$. That is to say, there is a color border between pixels i and j when $\omega_{ij} = 1$, and there is no color border when $\omega_{ij} = 0$. From the definition of the two thresholds θ^{t0} and θ^{t1}, it can be concluded that the threshold θ^{t0} is smaller than the threshold θ^{t1}. In order to simplify the selection of the parameters, $\theta^{t0} = \gamma\theta^{t1}$ is used to represent the relation between the two thresholds in general, and γ is a ratio that determines how much chromaticity variance is needed to classify it as a color edge when large illumination variance is detected.

In our experiments, $\gamma \in \{0.1, 0.01\}$. Here, we fixed γ by only using two values to simplify the problem. These thresholds also can be estimated using leave-one-out cross validation. The observed chromaticity gradient between the neighborhood pixels i and j is written as $(\|LI_i\| - \|LI_j\|)$. Therefore, we have:

$$E_r = \sum_{i,j \in p(i)} (Lr_i - Lr_j - \omega_{ij}(\|LI_i\| - \|LI_j\|))^2. \tag{16}$$

In the extreme case, $E_r = 0$ means $Lr_i - Lr_j = \omega_{ij}(\|LI_i\| - \|LI_j\|)$. If there exists a color border between pixels i and j (namely $\omega_{ij} = 1$), $Lr_i - Lr_j = \|LI_i\| - \|LI_j\|$, which also means that the color border should be maintained in the albedo map we eventually solved. But if there exists no color border between pixels i and j (namely $\omega_{ij} = 0$), pixels i and j have same albedo value. Fig. 8 shows the comparison between the reflectance images achieved via the traditional classifier and with the constraint E_r. We can see from the regions marked by the red rectangles that using E_r to detect the reflectance changes could well avoid unnecessary color borders caused by illumination change.

Fig. 8. Left: the input image. Middle: the reflectance image extracted with the traditional classifier to determine the reflectance changes. Right: the reflectance image with the constraint of E_r

Global Sparse Reflectance Constraint. The fourth term of the energy function is the global sparse reflectance constraint E_g:

$$E_g = \sum_i (r_i \vec{R}_i - C(\beta_i))^2. \tag{17}$$

This term is motivated by the finding of [33] that the kinds of color in an image are limited. Many studies, e.g., [31], [36], used a global sparse reflectance prior to constrain the number of reflectance values to a finite set. Our sparse prior on reflectance is modeled using the method presented by Rother et al. [10] as:

$$C(\beta_i) = \frac{1}{|\{i : \beta_i = c\}|} \sum_{i:\beta_i = c} r_i \vec{R}_l, \ c = 1, 2, \cdots, C, \tag{18}$$

where, c denotes the index of different reflectance clusters, C is the number of clusters which is obtained by performing k-means clustering over the input image in RGB color space. β_i is the cluster index indicating the reflectance of pixel i, and $\beta_i \in \{1, 2, \cdots, c\}$. $C(\beta_i)$ is the cluster center that is the average value of all pixels belonging to the cluster indexed by β_i. In the extreme case, $E_g = 0$, and according to Eq. (17), $r_i \vec{R_i} = C(\beta_i)$. In other words, the finally obtained albedo must be one of the clustering results. This constraint E_g means that the difference between our final albedo decomposition results and the clustering results should be as small as possible.

Non-Local Texture Constraint. The fifth term of the energy function is the non-local texture constraint E_n. The texture information an input image contains is limited, which means for a pixel on the surface of an object, its texture information may be consistent with the texture information of other pixels in the image. Then, we can consider the image as a sequence of textures, and pixels with same texture should have same albedo value. According to this global texture constraint, we formulate E_n as:

$$E_n = \sum_{G_k \in \Gamma_n} \sum_{i,j \in G_k} (Lr_i - Lr_j)^2, \tag{19}$$

By comparing the neighborhood, pixels of the entire image can be divided into different texture groups. Γ_n is the set of all texture groups in the image, G_k is the kth texture group of Γ_n, and all pixels in G_k should have same albedo value, which means that the difference among albedo decomposition results of pixels within the same group should be minimal. $i, j \in G_k$ indicates two pixels that belong to the same group G_k. Motivated by the finding of [24], Shen et al. [43] considered the surface of the objects globally, each point on a surface shares the same neighborhood texture with a set of other points. In other words, they have the same local

texture structures, and some local texture structures compose the texture of the object. The groups of pixels generated by this observation could be used to improve the Retinex-based methods. Textured pixel groups are generated by examining all pixels with a 3×3 window in the shading-independent image [36].

3.4 Intrinsic Image Energy Minimization

The energy function of intrinsic image decomposition has been given in Section 3.3. It is obvious that the albedo and light step components are two independent unknowns, they will not affect each other. Thus, the objective function could be divided into two small objective functions according to the unknown quantities as:

$$E_1 = \underset{Lr_1, Lr_2, \ldots, Lr_N}{\arg\min} \; w_1 E_s + w_3 E_r + w_4 E_g + w_5 E_n, \tag{20}$$

$$E_2 = \underset{Ls^s_1, Ls^s_2, \ldots, Ls^s_N}{\arg\min} \; w_1 E_s + w_2 E_e. \tag{21}$$

In E_1, Lr is the only parameter and Ls^s can be considered as a constant. In E_2, Ls^s is the only parameter and Lr can be considered as a constant. For each energy function, we can minimize the energy function respectively by applying the minimization function derived and deduced from [10] and [44]. We can follow the minimization algorithm to iteratively get β and Lr, further to minimize E_1 and E_2 [45]. Finally, the estimated reflectance image and the estimated shading image could be obtained, see Fig. 2 and Algorithm 1 for details. However, the objective function is non-convex for containing the reverse of Lr, and utilizing coordinate descent may lead to a local minimum. Therefore, we need to design an appropriate initialization for Lr and Ls^s.

Algorithm 1. Intrinsic Images Estimation

1: Initialize: $Lr^{(0)}$, $Ls^{s(0)}$;
2: $\beta^{(0)} \leftarrow k$-means clustering;
3: $t \leftarrow 0$;
4: **while** $E^{t+1} - E^t > \theta$ **do**
5: $Lr^{(t+1)} \leftarrow$ optimize Eq. (20) with $\beta^{(t)}$, $Ls^{s(t)}$ fixed;
6: $\beta^{(t+1)} \leftarrow$ cluster $Lr^{(t+1)}$ by k-means;
7: $Ls^{s(t+1)} \leftarrow$ optimize Eq. (21) with $Lr^{(t+1)}$ fixed;
8: $t \leftarrow t + 1$;
9: **end while**

Initialization of Lr and Ls^s. Eq. (20) contains the reciprocal of Lr, which means that we need to find a minimum solution for a non-convex and non-concave function. To avoid finding a local minimum which is not a global minimum in the end, we set four kinds of initial value for Lr, and finally take the decomposition result with the smallest error as the albedo map of the intrinsic image. Obviously, the final output value of Lr is within a certain range, e.g., $0 \leq Lr_i^c \leq 1$, where the superscript c represents one of the RGB channels. Then, we can get $\|LI_i\| \leq Lr_i \leq 3$. Thus, we use a mixed starting point in [10] to calculate the initial values of Lr:

$$Lr_i = \alpha \|LI_i\| + 3(1 - \alpha), \tag{22}$$

where, $\alpha \in \{0.3, 0.5, 0.7\}$. Besides, we also set $Lr_i = 1$ as a case of the initial value for Lr_i. After comparing the four initial values' decomposition results, we take the one with the smallest error as the final result. Since Eq. (21) does not contain the reciprocal of Ls, we set $Ls = 1$ as the initial value. The minimization of the intrinsic image energy could be formulated as Algorithm 1. First, we initialize $Lr^{(0)}$ and $Ls^{s(0)}$ by the initialization strategy mentioned above. Then, we perform the k-means clustering of the intrinsic image to obtain the cluster result and compute the corresponding cluster result index $\beta^{(0)}$ for each pixel. Next, we do the iteration calculation for the energy E till the difference between E_{t+1} and E_t is smaller than a threshold θ. As shown in Algorithm 1, $Lr^{(t+1)}$ is updated by $\beta^{(t)}$ and $Ls^{s(t)}$ with Eq. (20), $\beta^{(t+1)}$ is updated with $Lr^{(t+1)}$, and $Ls^{s(t+1)}$ is updated by $Lr^{(t+1)}$ with Eq. (21). It requires at least 10 iterations to get convergence for our minimization of the intrinsic image energy. Fig. 9 shows the minimization process of our intrinsic image energy. It can be seen from the top sub-figure that the intrinsic image energy rapidly decreases after the first iteration and only 10 iterations are needed. We can also see from the bottom sub-figure that the reflectance image and shading image have similar appearance as the ground truth reflectance and shading images after 10 iterations.

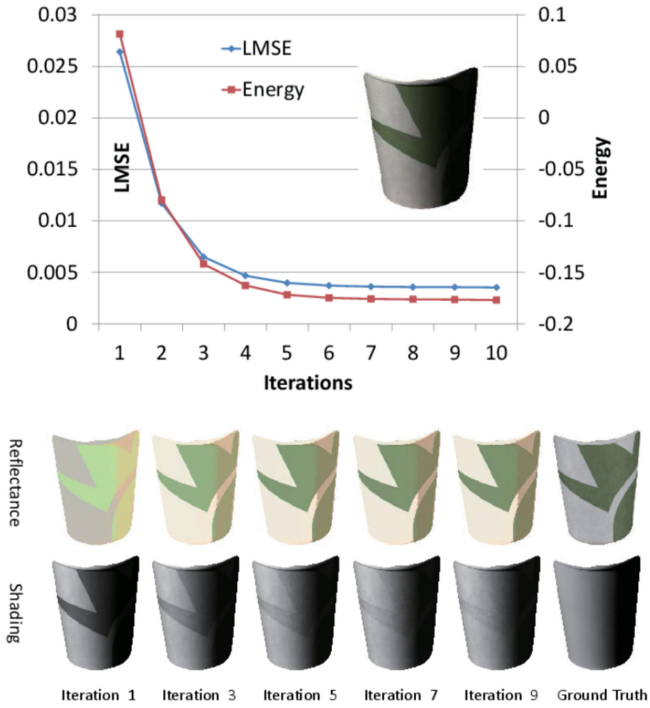


Fig. 9. The energy diagram for intrinsic image of image "cup1".

4 Evaluations

In this section, we evaluate our approach on the MIT dataset with known ground truth [25], the IIW dataset with human annotations [17], and the MPI Sintel dataset [26]. We follow common practice and mainly use the error metrics like LMSE (local mean squared error), MSE (mean squared error), and WHDR (weighted human disagreement rate) to measure the performance on these datasets. The speed of our approach is not a contribution, the performance of the decomposition processing is efficient but not in real time, which is generally at a regular performance level compared with other intrinsic image researches.

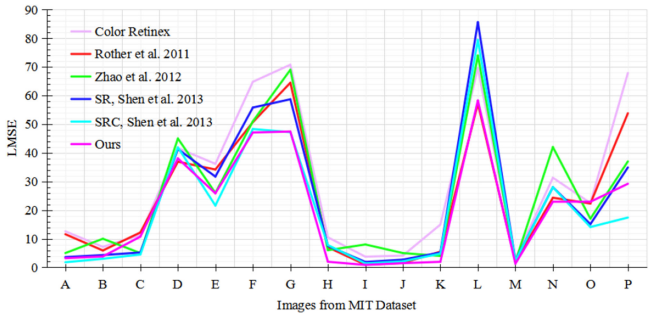


Fig. 10. Intuitive comparison between our method and other approaches on the MIT dataset. The y-axis label "LMSE" denotes "LMSE$\times 10^3$". (A) Box, (B) Cup1, (C) Cup2, (D) Deer, (E) Dinosaur, (F) Frog1, (G) Frog2, (H) Panther, (I) Paper1, (J) Paper2, (K) Raccoon, (L) Squirrel, (M) Sun, (N) Teabag1, (O) Teabag2, and (P) Turtle. We can see that our results have competitive performance, and outperform the others on most examples.

4.1 MIT Dataset

The MIT benchmark dataset with ground truth was presented by [25], and it is very helpful for investigating the intrinsic image recovery problem. We use the LMSE metric in [25] to measure our decomposition quality. The scale-invariant MSE is as follow:

$$\mathrm{MSE}(y, \widetilde{y}) = \|y - \widetilde{\beta}\widetilde{y}\|^2, \tag{23}$$

where, y is a true vector and $\widetilde{y}$ is the estimated one. $\widetilde{\beta}$ satisfies:

$$\widetilde{\beta} = \arg\min_{\beta} \|y - \beta\widetilde{y}\|^2. \tag{24}$$

The local MSE (LMSE) is then presented as follow:

$$\mathrm{LMSE}_k(I, \widetilde{I}) = \sum_{w \in W} \mathrm{MSE}(I_w, \widetilde{I}_w), \tag{25}$$

where, I is the ground truth image of shading or reflectance, $\widetilde{I}$ is the corresponding estimated one. k is the width or height of all local windows, over which MSE is summed by the steps of $k/2$. w is a local window from the local windows set W. LMSE for evaluation is the average LMSE values of the shading and reflectance images. We perform the intrinsic decomposition of 16 images from the MIT dataset, and show in Fig. 10 the LMSE error of our approach in comparison with others, including the Color Retinex, Rother et al. [10], Zhao et al. [32], and Shen et al. [11]. For those non-ideal examples, where our results are better than the others but the assessed value are not perfect, we can figure out that the conditions are really complicated.

Our approach performs better than those of Rother et al. [10] and Zhao et al. [32] on most images. Further, it has a lower average LMSE. Note that by tuning the parameters of the energy function, we could obtain a better LMSE result. The results of the "panther" and "raccoon" images, which contain shadows, are better than those of the other five methods. The results show that our method has achieved the state-of-the-art performance on the MIT benchmark dataset. A single-image based method needs powerful priors to restrict the model itself so that it can effectively solve the under-constrained intrinsic image recovery problem. In other words, the results of our method demonstrate that we have designed and applied applicable and productive priors, which finely supports our idea that the shading should be further decomposed into drift shading and step shading to satisfy shading smoothness assumption. In addition, we have rectified and improved the Retinex method, which our method is based on. Therefore, we also compare our approach with the conventional Retinex algorithm. As shown in Fig. 11, we can see clearly that our method can handle situations with complex illumination better.

4.2 IIW Dataset

Bell et al. [17] released a real-world image dataset IIW in 2014. It contains a large amount of user-annotated data for

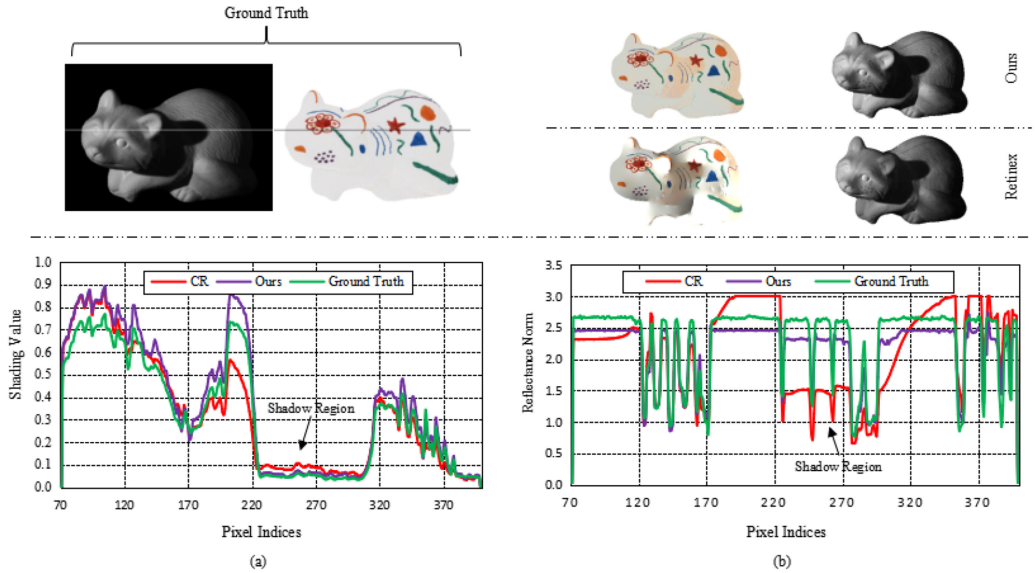


Fig. 11. Comparison between our method and the Retinex algorithm. We focus on the pixels in the shadow region. (a) and (b) are the comparison results. We use three curves (purple, red, and green) to represent our method, the Color Retinex (CR), and the ground truth, respectively. We explore the illumination value of the shadow region in (a), and the reflectance norm (sum of RGB channels) of the shadow region in (b). From the observation of (a), the illumination value of pixel in the shadow region by our method is smaller than by the Retinex, which means the shading component by our approach is more exact. From the observation of (b), the color norm by our method is larger than by the Retinex, which means the color in the shadow region by our method is brighter than by the Retinex.

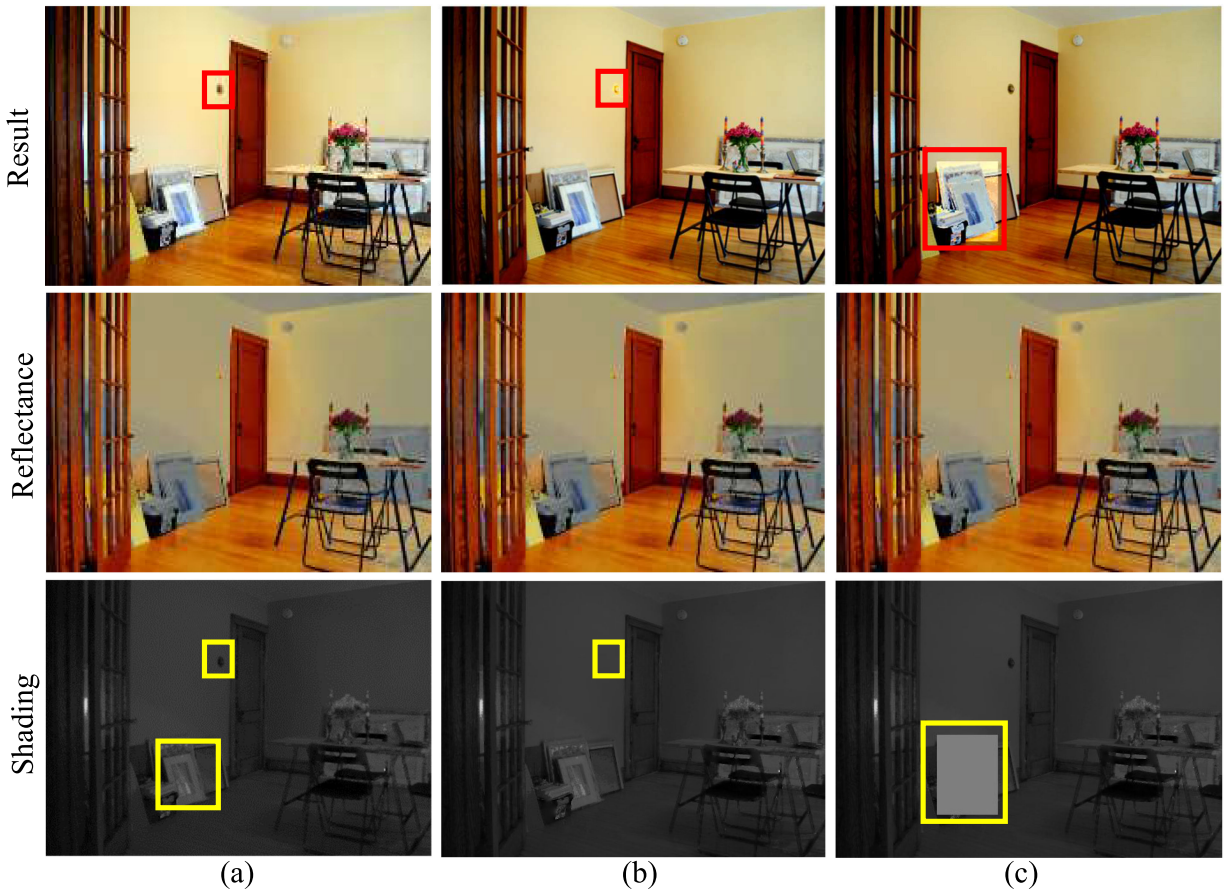


Fig. 12. Limitations of the WHDR error metric. (a) Intrinsic results of real images. (b) and (c) Changing the gray-scale values of the shading image within the yellow rectangle regions. The WHDR error focuses on the reflectance differences and ignores the shading mistakes, it does not change at all in (b) and (c). Therefore, this measure metric is highly biased.

evaluation. The error metric WHDR was proposed to measure the difference between the decomposed results and the human annotations. As [17] noted, the WHDR error focuses on the reflectance differences and ignores the shading mistakes. As shown in Fig. 12, we manipulate some areas of the shading image and compare the decomposed results. The reflectance and shading images in Fig. 12a are intrinsic result of a real image, and we can obtain decomposed result by multiplying intrinsic results. In Fig. 12b, we only change the grayscale values within the yellow rectangle of the shading image, altering the decomposed result. This is also done in Fig. 12c. However, the WHDR error has not been changed in Fig. 12b and 12c, because we have not modified the reflectance image yet. Hence, this measure metric is highly biased. This benchmark was created through millions of crowdsourced annotations of relative comparisons of material properties at pairs of points in each scene [17]. The human annotators are biased to the differences in reflectance. Since this benchmark is based on pairs of points, the WHDR unfairly favors methods that smooth reflectance channel too much. Nevertheless, we compare the WHDR of our method with other state-of-the-art methods, including Bell et al. [17], Zhao et al. [32], Garces et al. [46], Retinex [24], Shen and Yeo [31] on this real image dataset. The results are shown in Fig. 13, which shows that our approach has achieved competitive performance that greatly supports our idea.

Bell et al. [17] proposed a fully-connected CRF (conditional random field) to solve the intrinsic recovery problem. Although it has the lowest mean WHDR error, our illumination decomposition method could be adopted in [17] as a framework to improve its accuracy. As shown in Fig. 10, our approach obtains better LMSE results than other methods, and the LMSE error considers the reflectance as well as the shading, which highly meets our demand. Zhao et al. [32] performed quite well (close to ours) in the real image dataset. This method uses Retinex to decompose the texture while using global constraints to connect the distant parts. Among all the evaluated methods, our approach has the second-lowest mean WHDR error as it only lags behind the method from the creators of the IIW dataset.

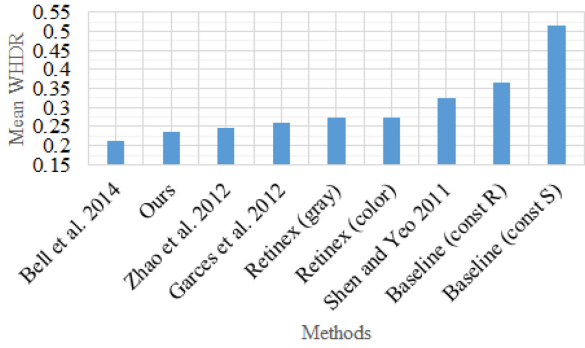


Fig. 13. The mean WHDR over all edges on the real image dataset. (This measure metric is biased to some extent as illustrated in Fig. 12. The reason that our approach lags behind the method in [17] has been further discussed in Section 4.2.)

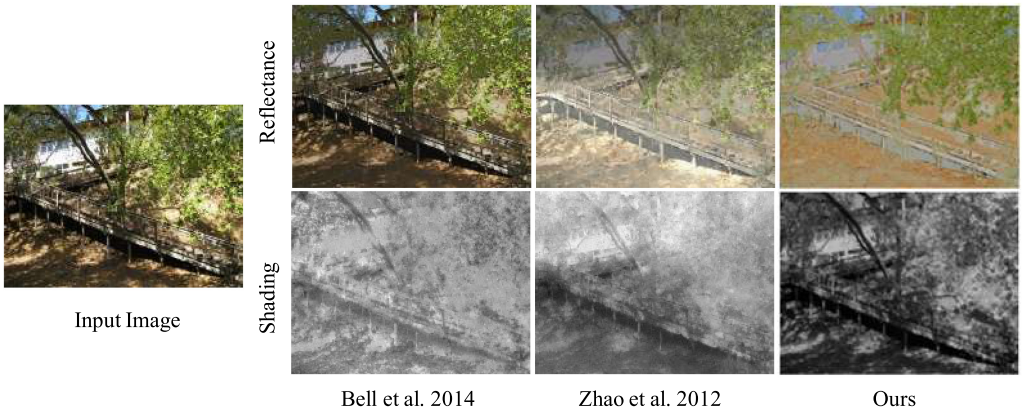


Fig. 14. Comparison with the existing intrinsic image methods (Bell et al. [17] and Zhao et al. [32]) based on a single image. The reflectance (top row) and the shading (bottom row). Our method can reduce the effect of shadow compared with other methods, which keep lots of shading information in the reflectance layers. The results of the other methods are reproduced from http://opensurfaces.cs.cornell.edu/intrinsic/algorithms.

We note that for the subsequent image processing, the accuracy of the shading image is as important as that of the reflectance image. The main idea of our approach is to obtain an exact shading image for this purpose. Garces et al. [46] achieved a decent result. Their method is based on clusters of similar reflectance and requires no user assistance. Retinex [24] performed well with respect to the WHDR measure. Our shading decomposition can be used as a framework for any other Retinex-based methods to improve the decomposition quality. The earlier method by Shen and Yeo [31] performed worse than the Retinex method. The high-frequency textures lead to mistakes in the shading image. In summary, our approach has achieved a promising WHDR performance among the single-image methods. Bell et al. [17] and Zhao et al. [32] kept a lot of shadow information in the reflectance images (see Fig. 14), while our method can reduce the effect of shadow almost completely. However, our method still and only falls behind [17] in these comparisons, which could definitely happen, because of the fact that the WHDR metric highly depends on the reflectance and overlooks the measurement of shading.

At the same time, we have compared our method with the single-image methods of Barron and Malik [47] and Chen and Koltun [38] based on RGB-depth data, and the multi-view method of Duchêne et al. [48] in Fig. 15. The resulting images beside our method come from [48]. As can be clearly seen from Fig. 15 that our approach achieved better performance than the compared methods [38], [47]. We detect the changes in the chromaticity to estimate the changes in the chromaticity. However, our method has certain limitation, as shown in the bright area of the table in Fig. 16c and the dark area under the bridge in Fig. 14, where our approach detects the changes in the chromaticity between the boundaries of these areas. However, there are large areas of light spots or darkness regions existing in these scenes that even human can hardly distinguish some subtle details. Such complicated scenes are very difficult to process.

4.3 MPI Sintel Dataset

We also evaluate our method on the MPI Sintel dataset [26]. We follow [38], and utilize the metrics MSE, LMSE and DSSIM to quantitatively measure the performance of the proposed approach. Results are presented in Table 1 and Fig. 17. From Table 1, we can see that our approach has generally obtained better performance compared to [14], [19], [22], [25], [38]. And in Fig. 17, there is almost no shadow left in our reflectance image, while large amount of remaining shadows exist in the other methods. Our shading image seems to be

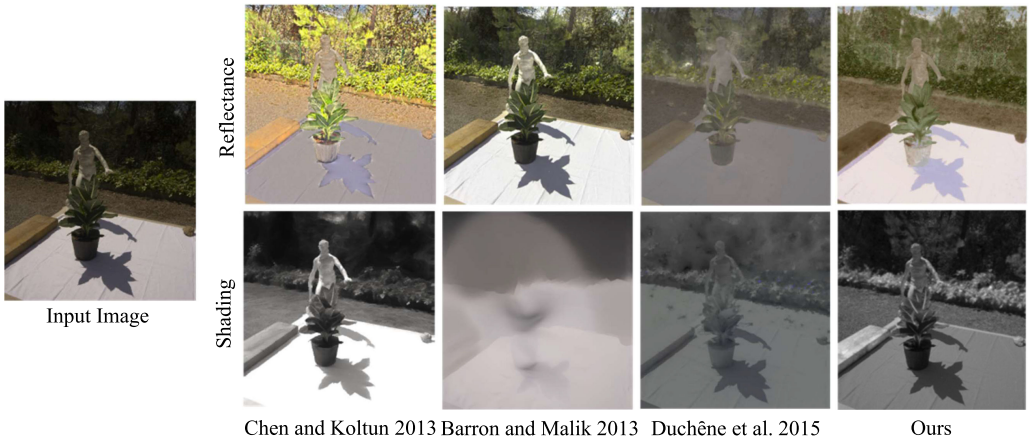


Fig. 15. Comparison with the single-image methods based on RGB-depth data of Chen and Koltun [38], Barron and Malik [47], and the multi-view method of Duchêne et al. [48]. The reflectance (top row) and the shading (bottom row). Although our method could not obtain the performance as the multi-view method [48], which can leverage multi-view image to construct a shading model, yet our approach could get similar or even better result compared with the other methods [38], [47].

Fig. 16. Visual results of intrinsic images recovery on real scenes. We chose three scenes (a), (b), and (c) from a real-world dataset IIW [17]. (a) shows that our method deals with an indoor scene with complex illumination. (b) is an outdoor scene. As shown in (c), in the bright region of the table, there is residual in the reflectance. Overall, our method achieves promising results on real-world images.

TABLE 1
Comparison on the MPI Sintel Benchmarks Between Our Approach and Others

Sintel Image Split	MSE			LMSE			DSSIM		
	A	S	Avg	A	S	Avg	A	S	Avg
Baseline: Shading Constant	5.31	4.88	5.10	3.26	2.84	3.05	21.40	20.60	21.00
Baseline: Albedo Constant	3.69	3.78	3.74	2.40	3.03	2.72	22.80	18.70	20.75
Grosse et al. [25]	6.06	7.27	6.67	3.66	4.19	3.93	22.70	24.00	23.35
Lee et al. [19]	4.63	5.07	4.85	2.24	1.92	2.08	19.90	17.70	18.80
Barron and Malik [14]	4.20	4.36	4.28	2.98	2.64	2.81	21.00	20.60	20.80
Chen and Koltun [38]	3.07	2.77	2.92	1.85	1.90	1.88	19.60	16.50	18.05
Narihira et al. [22]	1.00	0.92	0.96	0.83	0.85	0.84	20.14	15.05	17.60
Ours	0.99	0.95	0.97	0.79	0.75	0.77	18.90	14.96	16.93

Fig. 17. Visual results of intrinsic image recovery on the MPI Sintel dataset. We compare the results of our approach with those by Lee et al. [19], Barron and Malik [14], Chen and Koltun [38], and Narihira et al. [22]. From the reflectance image by our method, we can clearly see that there is almost no shadow left, while large amount of remaining shadows exist in the other methods. Moreover, our shading image also has better visual sense than the others in that with the help of reinforcement process, it looks more natural and has higher resolution and articulation, as shown in the hair and the body part of the character.

more natural and harmonious than the others in the hair and the body part of the character so that it is more like the ground truth, which shows the success of our approach.

4.4 Computation Cost and Complexity

We evaluate all the methods on a 24-core computer with 3.40 GHz CPU and 32 GB memory over an image with the resolution at $3{,}008 \times 2{,}000$ from the IIW Dataset. We have performed intrinsic decomposition of the image for 1000 times to compute for the more meaningful average running time performance for all the methods for stable purpose. Through the testing, the proposed method achieves favorable speed as shown in Table 2, and is significantly faster than that of Bell et al. [17] and Shen and Yeo [31]. Although the proposed method runs slower than Garces et al. [46] and Zhao et al. [32], it gets smaller mean WHDR value than those two methods as shown in Fig. 13, which indicates that our method outperforms these methods. Besides, we also record the average running time of energy optimization in Table 3. As mentioned in Section 3.4, our method requires 10 iterations to achieve the best result. According to Table 3, the whole energy optimization procedure costs about 36.95 seconds. Other operations including image pre-processing and post-processing only take 0.28 seconds, thus our method needs 37.23 seconds in total to get the final intrinsic image and shading image results for the input image.

TABLE 2
Running Time Comparison with Different Methods

Methods	Bell [17]	Shen and Yeo [31]	Garces [46]	Zhao [32]	Ours
Time	210.80 s	235.80 s	4.40 s	24.00 s	37.23 s

TABLE 3
Running Time of Our Energy Optimization

	iter1	iter2	iter3	iter4	iter5
E_1	5.97s	4.82s	3.17s	2.86s	2.23s
E_2	4.26s	3.90s	2.24s	2.03s	1.73s
Total	10.23s	8.72s	5.41s	4.89s	3.96s
	iter6	iter7	iter8	iter9	iter10
E_1	0.97s	0.52s	0.40s	0.17s	0.08s
E_2	0.85s	0.39s	0.23s	0.08s	0.05s
Total	1.82s	0.91s	0.63s	0.25s	0.13s

4.5 Applications and Limitation

Our approach can be applied to several applications, such as relighting, retexturing, and image editing, etc. [5], [6], [7], [49]. Reflectance and shading are significant references to image editing. The extended applications can benefit from the estimated results of our method by its desired performance and high availability under certain conditions. Fig. 18 shows some results on image editing based on our decomposed results. We manipulate images by focusing on the pixels that we are interested in and neglecting irrespective segments to avoid involving unconcerned pixels. Fig. 18a shows an image text replacement according to the estimated effects of shading and albedo. From Fig. 18, we can see that this application has achieved desired results, which demonstrates that our method can be suitably applied as the basic framework for the applications, such as image editing or material editing [49]. Compared with other methods that can also be used as a framework for this kind of application as mentioned in [49], the superiority of our method is that we can provide the availability and usability when users could only access to a single image with no user interference or additional information. The color difference between different objects greatly affects our final result. If the color difference between different objects is very small, the reflectance image will not be that clear to denote the objects in the scene. Fig. 19 shows a failure case under this situation. The color of the floor and the bathroom wall are similar, leading to a poor reflectance image which cannot clearly denote the objects.

5 Conclusion and Future Work

Single-image intrinsic recovery is a very hot research topic as it is easy use with high feasibility. In this paper, we built an illumination decomposition model to design a single-image based method for dealing with images containing shadows. Different from other methods, our approach is based on a solid and reasonable assumption that illumination can be further divided into two components, the step-shading channel and the drift-shading channel. Moreover, an enhanced criterion on Retinex and reinforcement process of edge classification have been well-designed and integrated into our method. We also propose a corresponding solution for the problem as an energy minimization via coordinate descent. Besides, we practically demonstrate a rational prior instead of the conventional simple smoothness assumption so that our approach is much more interpretable and convincing. Through experiments on the MIT intrinsic dataset, the IIW dataset, as well as the MPI Sintel dataset, our method has achieved competitive performance compared with the state-of-the-art methods. In the future, we will work on statistics and learning based intrinsic images. We plan to apply data from diverse points of views to reduce computational complexity. We will also work on L_1 image flattening [50] for further intrinsic image applications like 3D object composition in images/videos as the transform could generate accurate constant images/frames with sparse salient structures eliminating insignificant details.

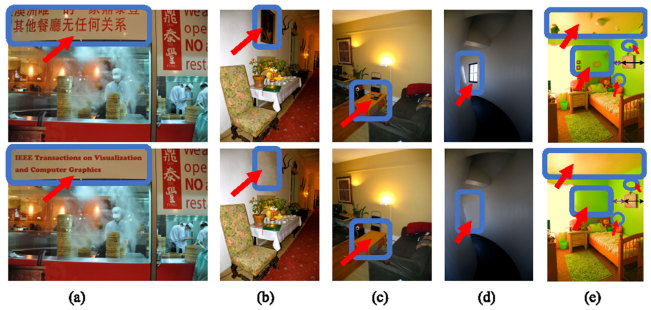


Fig. 18. Our applications on images from the IIW dataset. (a) is text replacement. (b) to (d) present results of removing certain details from the original images. For (e), apart from removing, we also translate the object measured by the purple double-headed arrow and the black double-headed arrow to a new position.

Fig. 19. Failure case of our method. Left to Right: input image, reflectance image, step shading image, and drift shading image.

Acknowledgments

We thank the reviewers for their insightful comments and suggestions. This work was supported in part by the National Natural Science Foundation of China (61572316, 61671290, 61872241), the National Key Research and Development Program of China (2016YFC1300302, 2017YFE0104000), the Science and Technology Commission of Shanghai Municipality (16DZ0501100, 17411952600), and the Ministry of Science and Technology (106-2221-E-006-233-MY2), Taiwan. Bin Sheng, Ping Li, and Yuxi Jin contributed equally to this work.

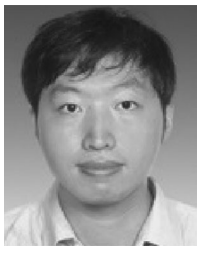
References

[1] H. G. Barrow and J. M. Tenenbaum, "Recovering intrinsic scene characteristics from images," *Comput. Vis. Syst.*, vol. 2, pp. 3–26, 1978.

[2] F.-L. Zhang, J. Wang, E. Shechtman, Z.-Y. Zhou, J.-X. Shi, and S. M. Hu, "PlenoPatch: Patch-based plenoptic image manipulation," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 5, pp. 1561–1573, May 2017.

[3] Z. Zhu, H.-Z. Huang, Z.-P. Tan, K. Xu, and S.-M. Hu, "Faithful completion of images of scenic landmarks using internet images," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 8, pp. 1945–1958, Aug. 2016.

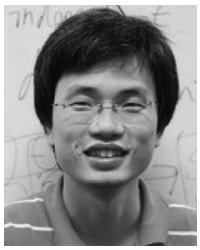
[4] Y. Liang, X. Wang, S.-H. Zhang, S.-M. Hu, and S. Liu, "PhotoRecomposer: Interactive photo recomposition by cropping," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 10, pp. 2728–2742, Oct. 2018.
[5] P. Li, H. Sun, C. Huang, J. Shen, and Y. Nie, "Interactive image/video retexturing using GPU parallelism," *Comput. Graph.*, vol. 36, no. 8, pp. 1048–1059, 2012.
[6] J. Cho, M. Lee, and S. Oh, "Complex non-rigid 3D shape recovery using a procrustean normal distribution mixture model," *Int. J. Comput. Vis.*, vol. 117, no. 3, pp. 226–246, 2016.
[7] S. Han, I. Sato, T. Okabe, and Y. Sato, "Fast spectral reflectance recovery using DLP projector," *Int. J. Comput. Vis.*, vol. 110, no. 2, pp. 172–184, 2014.
[8] A. Morgand, M. Tamaazousti, and A. Bartoli, "A geometric model for specularity prediction on planar surfaces with multiple light sources," *IEEE Trans. Vis. Comput. Graph.*, vol. 24, no. 5, pp. 1691–1704, May 2018.
[9] A. Meka, G. Fox, M. Zollhfer, C. Richardt, and C. Theobalt, "Live user-guided intrinsic video for static scenes," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 11, pp. 2447–2454, Nov. 2017.
[10] C. Rother, M. Kiefel, L. Zhang, B. Schölkopf, and P. V. Gehler, "Recovering intrinsic images with a global sparsity prior on reflectance," in *Proc. 24th Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 765–773.
[11] L. Shen, C. Yeo, and B.-S. Hua, "Intrinsic image decomposition using a sparse representation of reflectance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2904–2915, Dec. 2013.
[12] M. F. Tappen, E. H. Adelson, and W. T. Freeman, "Estimating intrinsic component images using non-linear regression," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis Pattern Recognit.*, 2006, vol. 2, pp. 1992–1999.
[13] M. F. Tappen, W. T. Freeman, and E. H. Adelson, "Recovering intrinsic images from a single image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1459–1472, Sep. 2005.
[14] J. T. Barron and J. Malik, "Shape, illumination, and reflectance from shading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 8, pp. 1670–1687, Aug. 2015.
[15] Y. Weiss, "Deriving intrinsic images from image sequences," in *Proc. 8th IEEE Int. Conf. Comput. Vis.*, 2001, vol. 2, pp. 68–75.
[16] Y. Matsushita, S. Lin, S. B. Kang, and H.-Y. Shum, "Estimating intrinsic images from image sequences with biased illumination," in *Proc. Eur. Conf. Comput. Vis.*, 2004, pp. 274–286.
[17] S. Bell, K. Bala, and N. Snavely, "Intrinsic images in the wild," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 159:1–159:12, 2014.
[18] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez, "Intrinsic video and applications," *ACM Trans. Graph.*, vol. 33, no. 4, pp. 80:1–80:11, 2014.
[19] K. J. Lee, Q. Zhao, X. Tong, M. Gong, S. Izadi, S. U. Lee, P. Tan, and S. Lin, "Estimation of intrinsic image sequences from image +depth video," in *Proc. 12th Eur. Conf. Comput. Vis. - Vol. Part VI*, 2012, pp. 327–340.
[20] M. Bell and E. T. Freeman, in *Learn. Local Evidence Shading Reflectance*, 2001, vol. 1, pp. 670–677.
[21] L. Lettry, K. Vanhoey, and L. van Gool, "DARN: A deep adversarial residual network for intrinsic image decomposition," in *IEEE Winter Conf. Appl. Comput. Vis.*, 2018, pp. 1359–1367.
[22] T. Narihira, M. Maire, and S. X. Yu, "Direct intrinsics: Learning albedo-shading decomposition by convolutional regression," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2992–2992.
[23] T. Zhou, P. Krähenbühl, and A. A. Efros, "Learning data-driven reflectance priors for intrinsic image decomposition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 3469–3477.
[24] E. H. Land and J. J. McCann, "Lightness and retinex theory," *J. Opt. Soc. Amer.*, vol. 61, no. 1, pp. 1–11, 1971.
[25] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, "Ground truth dataset and baseline evaluations for intrinsic image algorithms," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, 2009, pp. 2335–2342.
[26] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 611–625.
[27] A. Blake, "Boundary conditions for lightness computation in Mondrian World," *Comput. Vis. Graph. Image Process.*, vol. 32, no. 3, pp. 314–327, 1985.
[28] B. K. P. Horn, "Determining lightness from an image," *Comput. Graph. Image Process.*, vol. 3, no. 4, pp. 277–299, 1974.
[29] G. D. Finlayson, S. D. Hordley, and M. S. Drew, "Removing shadows from images," in *Proc. 7th Eur. Conf. Comput. Vis.-Part IV*, 2002, pp. 823–836.
[30] B. V. Funt, M. S. Drew, and M. Brockington, "Recovering shading from color images," in *Proc. Eur. Conf. Comput. Vis.*, 1992, pp. 124–132.
[31] L. Shen and C. Yeo, "Intrinsic images decomposition using a local and global sparse representation of reflectance," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2011, pp. 697–704.
[32] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin, "A closed-form solution to retinex with nonlocal texture constraints," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 7, pp. 1437–1444, Jul. 2012.
[33] Y. Li and M. S. Brown, "Single image layer separation using relative smoothness," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 2752–2759.
[34] D. Hauagge, S. Wehrwein, K. Bala, and N. Snavely, "Photometric ambient occlusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 2515–2522.
[35] P.-Y. Laffont, A. Bousseau, and G. Drettakis, "Rich intrinsic image decomposition of outdoor scenes from multiple views," *IEEE Trans. Vis. Comput. Graph.*, vol. 19, no. 2, pp. 210–224, Feb. 2013.
[36] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis, "Coherent intrinsic images from photo collections," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 202:1–202:11, 2012.
[37] A. Bousseau, S. Paris, and F. Durand, "User-assisted intrinsic images," *ACM Trans. Graph.*, vol. 28, no. 5, pp. 130:1–130:10, 2009.
[38] Q. Chen and V. Koltun, "A simple model for intrinsic image decomposition with depth cues," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2013, pp. 241–248.
[39] M. Hachama, B. Ghanem, and P. Wonka, "Intrinsic scene decomposition from RGB-D images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 810–818.
[40] J. Shen, X. Yan, L. Chen, H. Sun, and X. Li, "Re-texturing by intrinsic video," *Inf. Sci.*, vol. 281, pp. 726–735, 2014.
[41] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf, "A generic deep architecture for single image reflection removal and image smoothing," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 3258–3267.
[42] Y. Li, M. Paluri, J. M. Rehg, and P. Dollár, "Unsupervised learning of edges," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1619–1627.
[43] L. Shen, P. Tan, and S. Lin, "Intrinsic image decomposition with non-local texture cues," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2008, pp. 1–7.
[44] C. E. Rasmussen, "Minimization algorithm," 2002. [Online]. Available: www.gatsby.ucl.ac.uk/edward/code/minimize
[45] S. Ding, B. Sheng, Z. Xie, and L. Ma, "Intrinsic image estimation using near-L_0 sparse optimization," *Vis. Comput.*, vol. 33, no. 3, pp. 355–369, 2017.
[46] E. Garces, A. Munoz, J. Lopez-Moreno , and D. Gutierrez, "Intrinsic images by clustering," *Comput. Graph. Forum*, vol. 31, no. 4, pp. 1415–1424, 2012.
[47] J. T. Barron and J. Malik, "Intrinsic scene properties from a single RGB-D image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2013, pp. 17–24.
[48] S. Duchêne, C. Riant, G. Chaurasia, J. L. Moreno, P.-Y. Laffont, S. Popov, A. Bousseau, and G. Drettakis, "Multiview intrinsic images of outdoors scenes with an application to relighting," *ACM Trans. Graph.*, vol. 34, no. 5, pp. 164:1–164:16, 2015.
[49] N. Bonneel, B. Kovacs, S. Paris, and K. Bala, "Intrinsic decompositions for image editing," *Comput. Graph. Forum*, vol. 36, no. 2, pp. 593–609, 2017.
[50] S. Bi, X. Han, and Y. Yu, "An L_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition," *ACM Trans. Graph.*, vol. 34, no. 4, pp. 78:1–78:12, 2015.



Bin Sheng received the PhD degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, China. He is currently an associate professor with the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China. His current research interests include image-based rendering, machine learning, virtual reality, and computer graphics.

Ping Li received the PhD degree from The Chinese University of Hong Kong, Hong Kong, China. He is currently an assistant professor with the Macau University of Science and Technology, Macau, China. His current research interests include image/video stylization, big data visualization, GPU acceleration, and creative media. He has one image/video processing national invention patent, and has excellent research project reported worldwide by *ACM TechNews*.

Yuxi Jin received the BEng degree in software engineering from the Henan University, Kaifeng, China, and the MEng degree from the East China University of Science and Technology, Shanghai, China. She is currently working toward the the PhD degree in computer science in the Faculty of Information Technology, Macau University of Science and Technology, Macau, China. Her current research interests include intrinsic images, machine learning, and computer graphics.



Ping Tan received the PhD degree in computer science and engineering from the Hong Kong University of Science and Technology, Hong Kong, China, in 2007. He is currently an associate professor with the School of Computing Science, Simon Fraser University, Burnaby, BC, Canada. His current research interests include computer vision and computer graphics. He has served as an Editorial Board Member of the *International Journal of Computer Vision*, and the Computer Graphics Forum. He has served as a program committee member of SIGGRAPH, and SIGGRAPH Asia.

Tong-Yee Lee received the PhD degree in computer engineering from Washington State University, Pullman, in May 1995. He is currently a chair professor with the Department of Computer Science and Information Engineering, National Cheng-Kung University, Tainan, Taiwan, ROC. He leads the Computer Graphics Group, Visual System Laboratory, National Cheng-Kung University (http://graphics.csie.ncku.edu.tw). His current research interests include computer graphics, non-photorealistic rendering, medical visualization, virtual reality, and media resizing. He is a senior member of the IEEE and a member of the ACM.