Multi-view Multi-illuminant Intrinsic Dataset

Shida Beigpour
shida@mpi-inf.mpg.de
Mai Lan Ha
hamailan@informatik.uni-siegen.de
Sven Kunz
sven.kunz83@gmail.com
Andreas Kolb
andreas.kolb@uni-siegen.de
Volker Blanz
blanz@informatik.uni-siegen.de

Institute for Vision and Graphics
University of Siegen
Siegen, Germany

Abstract

This paper proposes a novel high-resolution multi-view dataset of complex multi-illuminant scenes with precise reflectance and shading ground-truth as well as raw depth and 3D point cloud. Our dataset challenges the intrinsic image methods by providing complex coloured cast shadows, highly textured and colourful surfaces, and specularity. This is the first publicly available multi-view real-photo dataset at such complexity with pixel-wise intrinsic ground-truth. In the effort to help evaluating different intrinsic image methods, we propose a new perception-inspired metric based on the reflectance consistency. We provide the evaluation of three intrinsic image methods using our dataset and metric.

1 Introduction

Decomposing an image into its intrinsic components (e.g. reflectance and shading) has always been a fundamental concept in computer vision research. During the last decades, intrinsic image research has seen great improvement. While in the early days intrinsic image was limited to grayscale, in the recent years, performing a joint optimization of colour and 3D surface is proving to produce superior results [2].

New trends in computer vision such as fusion of colour and depth (RGB-D) strongly benefit from a correct reflectance estimation in multi-view scenes. This task requires an illumination invariant estimation of reflectance to reproduce the true colour of the surface despite illumination variation and specularities (Fig 1). Hao Li *et al.*, for example, use illumination-colour invariant reflectance estimation to produce the correct colour of their 3D model [26]. Furthermore, 3D reconstruction using photo collections deals with internet photographs captured under varying illumination conditions ([19] [28] [13]). Here producing the true colour and texture of the surface is highly challenging due to the presence of shadows and illumination colours which result in strong artefacts unless a correct reflectance estimation is performed. Therefore, multi-view intrinsic image estimation is desired.

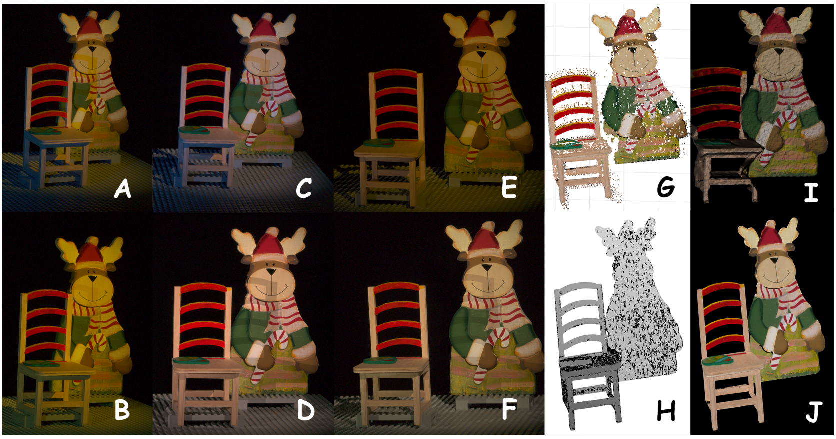


Figure 1: An example of the multi-view multi-illuminant setup (A-F), rough point cloud (G), raw depth (H), 3D surface (I), and ground-truth reflectance (J).

The contribution of our work comprises two main parts. Firstly, we introduce a new stereo and multi-view photo dataset for intrinsic image under multiple coloured (non-uniform) lights with full precise ground-truth which improves over the state-of-the-art in the following aspects: a) pixel-wise depth value and dense point cloud. b) highly textured and colourful objects and background suitable for intrinsic image and colour constancy research. c) multi-view setup which is especially interesting regarding specularities. Secondly, we propose a new perception-inspired intrinsic image evaluation metric to better judge the precision of the recovered reflectance for multi-illuminant scenarios. We also provide the evaluation of three intrinsic image methods using our dataset and metric. Fig. 1 demonstrates an example of a scene captured with six different cameras under different illumination conditions along with its raw depth, point cloud, and a rough surface reconstruction. Here the advantage of using reflectance instead of the captured pixel colour for the 3D surface colour is evident.

2 Related work

2.1 Intrinsic Image Decomposition

The majority of intrinsic image methods are based on the study of Retinex theory [25]. They can either use single RGB image or multiple images and video as the input.

Single RGB image: Barrow and Tenenbaum published a method to recover intrinsic image through edge classification [3]. A similar idea that classifies derivatives of reflectance changes from shading changes was proposed by Tappen *et al.* [35]. To strengthen the constraints from Retinex theory, different cues are used such as non-local texture [33] or integration of luminance amplitude, hue and texture [21]. A shadow-free decomposition method was proposed by Finlayson *et al.* [12]. In order to improve the performance, sparse reflectance prior is used [32], [39]. Some methods use clustering regions of similar reflectance as constraint automatically [15] or using user's assistance [9]. A combination of multiple priors has been proven to greatly improve the results [16], [1]. Serra *et al.* [31] proposed a method using surface descriptors on names and shades of colours. Recently, there is an increasing work of intrinsic image using convolutional neural network [29], [40], [41].

Multiple RGB images and video: Instead of only using a single RGB image, multiple RGB images of the same scene under different lighting conditions and/or different viewpoints are used in the decomposition [37], [24]. The coherence of the result is reinforced by the reflectance constancy of the same points in different input images [24]. Intrinsic decomposition for videos is a challenging task. Besides the ill-posed nature of intrinsic image, it requires the temporal coherence. Few methods for intrinsic video have been developed in recent years [38], [7], [23]. The common constraints used in these methods are the spatial sparse reflectance prior and temporal smoothness.

RGB-D: Another approach to solve the ill-posed problem of intrinsic image decomposition and improve the results is to make use of a depth map in different ways [2], [11], [18], [34]. Extended from single RGB-D image, Hachama *et al.* proposed to use single or multiple RGB-D images from different views [18]. Furthermore, the optimizations for intrinsic image techniques also improve the depth map [2] and can be used in depth transfer [22]. Shi *et al.* also take advantage of intrinsic image estimation with RGB data to recolour, re-texture and compose objects in a scene [34].

2.2 Datasets

Datasets have been created to evaluate intrinsic image results. They can be synthetic or created from real-world photos with single or multiple illuminants and other enriched features.

Synthetic Data: Butler *et al.* introduced the MPI Sintel dataset [10] originally for evaluating optical flow. Besides the ground-truth flow fields, it provides full intrinsics ground-truth. Barron *et al.* [2] composed a pseudo-synthetic dataset by enhancing the MIT Intrinsic Image dataset with depth map and rendering colored illumination. Beigpour *et al.* [4] provided full intrinsic ground-truth for rendered scenes with multiple objects lit by multiple distinctly coloured lights.

Real-world Photographs: One popular real-world dataset is MIT dataset created by Grosse *et al.* [17]. It provides full and precise physics-based intrinsic ground-truth. The scenes are constructed with a single object and white illuminant. Tappen *et al.* [36] created a dataset using different types of papers with green marker scrabbles. Therefore, the red channel was used as the shading ground-truth. A more comprehensive dataset is provided by Beigpour *et al.* [5] with full precise intrinsic ground-truth. The dataset captures multiple illuminants with some specularities and low resolution depth map from a Kinect. For more natural images, Bell *et al.* published a large-scale dataset that is annotated by crowd-source, thus the ground-truth is not available for every pixel [6] and illumination colour is ignored.

2.3 Evaluation Metrics

Many intrinsic image methods use mean squared error (MSE) or local mean squared error (LMSE) metric to evaluate the results. As Grosse *et al.* stated in [17], the MSE gives heavy penalty for a small error in the decomposition results. On the other hand, the LMSE tries to average out the errors across the whole image by computing the MSE and estimating the scale factor for individual patches [17]. Therefore, the global consistency on the evaluation is not enforced. Bell *et al.* introduced a weighted human disagreement rate (WHDR) metric [6]. The evaluation is based on human judgement on individual pairs of points without ground-truth. Bell *et al.* stated that the WHDR has a margin of error of 7.5% . This could potentially result in the metric being less discriminative for methods with similar performance.

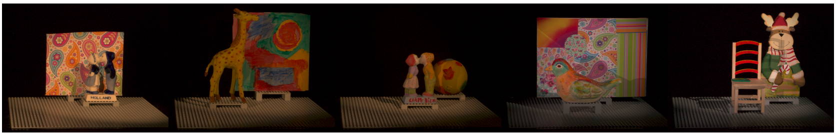


Figure 2: The five scenes (1 to 5 from left to right) captured by one of the six cameras.

3 Multi-view Dataset and Ground-truth

Inspired by the work of Beigpour *et al.* [5], we create "multi-illuminant" scenes and compute their shading and reflectance ground-truth. Here we propose a number of improvements which we believe extensively boost the applicability of our dataset over existing real-world intrinsic image datasets: high resolution dense depth image and point cloud, multi-view setup using 6 cameras, and textured and colourful objects. Fig. 2 shows the 5 scenes from one of the cameras. In total the dataset consists of 20 illumination conditions, 5 scenes, and 6 cameras. Our complete dataset consists of 600 high-resolution (5208×3476) images along with their ground-truth and is publicly available online[1]. In the following, we provide technical information regarding our dataset and well as acquisition of the ground-truth data.

3.1 Scene characteristics and lighting

While the work of Beigpour *et al.* captured the essence of multi-illuminant condition, the scenes presented limited colours and virtually no texture. A common shortcoming of physics-based ground-truth datasets in this research is a lack of colourful and textured surfaces. Many intrinsic estimation methods therefore based their optimization on sparse reflectance which displays a limited number of distinct colours. Only few methods have addressed colour texture [20]. Our dataset contains challenging textures which makes shadow removal more difficult. We also included objects which present strong colour variations, smooth change of intensity in reflectance colour, and/or blending of different colours. All of these further challenge the common assumptions on reflectance.

Furthermore, to challenge the "smooth shading" assumption, we have created complex multi-illuminant scenarios which results in coloured shadow edges. Moreover, to take advantage of multi-view setup, we have added more specular illumination conditions since specularity is view dependent. In total, for each scene, we provide 20 lighting conditions: 4 single-illuminant, 11 multi-illuminant, and 5 specular multi-illuminant[2]. In order to maintain a high Colour Rendering Index (CRI), we use a halogen bulb with a stabilized DC power supply to avoid changes in the intensity during the capturing. Using a set of broadband filters we produce a set of distinctly coloured lights.

3.2 Setup

All the 6 cameras are equipped with zoom lens and focus is locked on the middle point of the scene. A polarizing filter is placed in front of each of the cameras. The cameras in our setup are operated in Manual mode. The automatic white balance is turned off. In order to keep the image sharp for the whole scene, we chose a small aperture to increase the depth of field. We calculate this value based on the distance of the camera to the objects, focal length,

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

[2] Beigpour *et al.* have provided 17 illumination condition with only 9 multi-illuminant and 2 specular condition.

sensor size, and the dimensions of the scene's volume. To reduce noise, we set the ISO to the base value which, together with using a small aperture, leads to the need for long exposure times. As all of our scenes are static and the lighting is stable, this will not lead to artefacts. We further empirically find appropriate exposure time for each of the lighting conditions in order to maintain the most number of valid pixels (not too dark or clipped). As camera noise is unavoidable and very small fluctuations in the light intensity is not impossible we average several images for each lighting condition and remove the dark current image.

3.3 Intrinsic Ground-Truth Calculation

Real-photo datasets with precise ground-truth are crucial in many computer vision and machine learning topics especially the field of intrinsic image. Yet there are only few datasets of such available in this research. The main reason for that is the precise and cumbersome procedure which is require to create a reliable intrinsic image dataset. This procedure have first been introduced by Grosse *et al.* [17] on the basis of the compact form of the image formation model which is widely accepted in this field[3]:

$$I(\mathbf{x}) = S(\mathbf{x})R(\mathbf{x}) + C(\mathbf{x}) \quad , \tag{1}$$

where S, R, and C are the Shading, Reflectance, and Specular components of the image I when $\mathbf{x}$ indicates pixel coordinates. In the following, we drop the $\mathbf{x}$ notation without loss of generality. This definition has been extended by Beigpour *et al.* [5] where the shading component S also represents the illumination colour. In other words, the brightness level (grayscale) of each pixel in this image represents shading and the chromatic value is the illumination chroma as defined in the field of colour constancy. So far, only few intrinsic image methods provide colour values for S (e.g, [1]).

We calculate the intrinsic ground-truth for our images by following the standard procedure developed by Grosse *et al*. Here we briefly discuss the main steps and refer the readers to their paper for details and proof.

Specular Component: We use cross-polarization technique for obtaining a diffuse image I_D by using polarizing filters in front of the light sources and the cameras. Images obtained without these filters (I) contains specularity. Specular component can then be calculated by subtracting the diffuse image from the original: $C = I - I_D$

Reflectance: After capturing the scene under the 20 different lighting conditions, we carefully cover the surface of each object with a very thin and uniform diffuse gray spray paint. We then recapture each scene and lighting condition with the uniformly gray object. The image of the gray painted object is proportional to the shading image and is refereed to as relative shading $\widetilde{S}$. Division of the original image by the relative shading image results in relative reflectance $\widetilde{R}$. Therefore, we have $\widetilde{R} = I_D/\widetilde{S}$, where $\widetilde{S} \propto S$ and $\widetilde{R} \propto R$. Intrinsic image methods estimate the intrinsic components up to a constant magnitude, i.e. relative shading and reflectance. Real-capture intrinsic datasets therefore need only to provide the relative values of shading and reflectance [17]. By definition, the reflectance formula only generates valid results for pixels whose shading is positive. As low-brightness areas are more prone to noise, we discard pixels whose brightness is less than a threshold.

[3] Please refer to [5] and [17] for the full format of the image formation model.

Alignment: It is crucial to assure that all images of the scene are pixel-wise aligned. Even a slightest displacement which would result in a sub-pixel shift could falsify the ground-truth. There's no straightforward way to correct these errors in post-processing since a small displacement in object location can possibly result in changes in shadow boundaries. We use LEGO plates to place the object back on its exact position in the scene after gray-painting. To assure accuracy, we remove and re-shoot any scene in which there's a displacement. This in practice resulted in removing two-third of the captured images.

3.4 Depth and 3D

In the recent years and thanks to the advancements in 3D imaging, a number of intrinsic image methods have been proposed which take advantage of depth information about the image to improve the estimation accuracy. Such methods have presented results on datasets that are synthetic [10], their depth information is synthetically estimated [17], or do not provide reflectance/shading ground-truth [30]. Beigpour *et al.* [5] only provides very low resolution coarse depth information from Kinect Time-of-flight due to the differences between the RGB and depth camera in both resolution and Field-of-View. This results in their depth image to be 16 times lower resolution according to their paper.

To solve this issue, we take advantage of stereo setup with active illumination to produce full resolution dense depth values for each scene. This enables us to provide a dense point-cloud as well as a 3D mesh. Our camera system setup consists of 3 pairs of cameras (6 in total) whose positions are fixed with relation to each other and the scene. The cameras' intrinsic and extrinsic parameters are calibrated using a checker board [8]. Each of these pairs is used to calculate the scene geometry (i.e., depth map and 3D point cloud) in the stereo manner. To achieve better accuracy for parts of the scene with less prominent texture, we use active illumination, i.e., a noise pattern projected on the scene using a projector.

3.5 Discussion

The current dataset, as the only real-photo multi-illuminant and multi-view dataset with accurate intrinsic ground-truth and full resolution depth, offers several important improvements over the existing work. We plan to further extend this in the future. It is worth noting here that incorporating each of such features will be highly challenging. For example, outdoor scenery and illumination is by nature very dynamic as a moving cloud or even a leaf on a tree would falsify the results. While a crowd sourced labelling approach can be used for multi-illuminant and outdoor images, the human judgement is subjective and prone to error. We would also like to create a dataset which provides accurate inter-reflection ground-truth.

4 Point-wise Consistency Metric (PCM)

We propose a new perception-inspired error metric, which is based on *CIE Lab* colour space and the standard visual colour difference measurement CIE DE2000 [27], [14], to evaluate reflectance results generated from intrinsic image methods against the ground-truth. The proposed metric measures the perceptual error of the estimated reflectance without any human subject input and can hence be easily and automatically calculated for any new dataset.

Our main inspiration in designing the point-wise consistency metric (PCM) is rooted in the observation that the quality of an intrinsic image method's result is proportional to its *reflectance consistency* with respect to the ground-truth throughout different illuminations, strong shadows, and specularity. Here, our notion of reflectance consistency conforms to the following principles. Firstly, if two points p and q are perceptually similar in the ground-truth reflectance then they should also be similar in the estimated reflectance. Secondly, the brightness difference between a pair of points in the estimated reflectance should be similar to the ground-truth for the same pair.

4.1 Overview

Given the ground-truth reflectance image G and the estimated reflectance image T, our *point-wise consistency metric* works as follows:

1. Select a set of point pairs $\mathcal{S} = \{(p,q)\}$ in G that are *perceptually similar*.

2. Compute the *point-wise consistency error* $PCE(G,T)$ as follows:

$$PCE(G,T) = \frac{1}{|\mathcal{S}|} \sum_{(p,q)\in\mathcal{S}} f(pce_{G,T}(p,q)), \tag{2}$$

$$f(pce_{G,T}(p,q)) = \begin{cases} 1, & pce_{G,T}(p,q) > \sigma \\ \frac{pce_{G,T}(p,q)}{\sigma}, & otherwise \end{cases}, \tag{3}$$

$$pce_{G,T}(p,q) = \triangle E_{00}(G(p)-G(q), T(p)-T(q)), \tag{4}$$

where $\triangle E_{00}$ denotes CIE DE2000 colour distance [27], [14]. $pce_{G,T}(p,q)$ is the difference between the similarity of the points in the ground-truth and the estimated reflectance for a pair of points (p,q). f is a linear function to normalize $pce_{G,T}(p,q)$ to $[0,1]$ with a cut-off threshold σ. The value of σ is set such that there are about 10% of point pairs (p,q) or less which have $pce_{G,T}(p,q) > \sigma$ for all the evaluated methods. We do not want to have many error values that are normalized to 1 because it will make the evaluation between different methods less distinctive and less accurate.

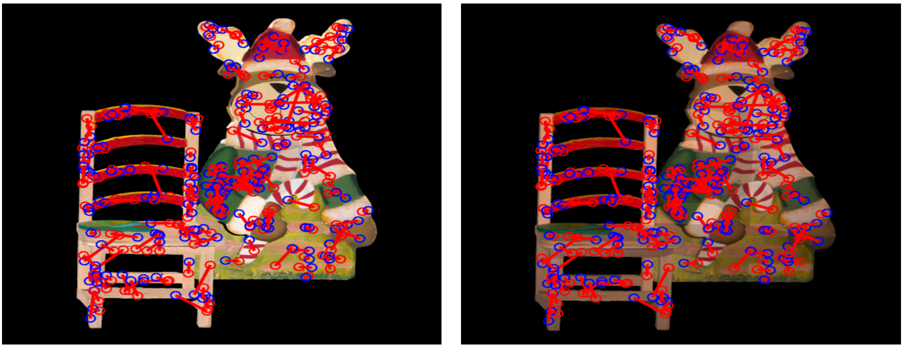
4.2 Point Selection Strategy

We randomly select pairs of perceptually similar points in the ground-truth image G. For every pair, the points are selected in the region of interest and need to be perceptually similar in colours, taking both chroma and luminance into account. Furthermore, the distance between two points in a pair also follows a normal distribution.

We define the region of interest to be on object surfaces, at properly lit pixels and not on objects' contours or edges. We build the mask M forming the region of interest as below:

$$M = \mathrm{ero}\big(M_{\text{scene}} \cap \overline{M_{\text{under-exp}}}\big) \cap \mathrm{ero}(\overline{M_{\text{edges}}}) \ , \tag{5}$$

where $M_{\text{scene}}, M_{\text{under-exp}}$, and M_{edges} are the masks for the scene, the under exposed pixels, and the edges inside the object, respectively. M_{scene} is created manually to mask out the background of the scene. $M_{\text{under-exp}}$ is produced automatically by setting the threshold for under-exposed pixels. M_{edges} is the edge map computed using Canny edge detection method. The erosions make sure that selected points are not close to the contours or edges.



(a) Selected pair points on ground-truth reflectance (b) Selected pair points on estimated reflectance under white illumination

Figure 3: 200 pairs of points selection with the colour difference threshold $\varepsilon = 10$ and the mean distance between point pairs $\mu_d = 20$

We select point pairs (p,q) by first randomly select p within the region of interest M. Point q is selected also inside M with a distance from p that follows a normal distribution, i.e. $\|p,q\|$ is $N(\mu_d,\sigma_d)$. Furthermore, we only accept perceptually similar point pairs, i.e.

$$\triangle E_{00}(G(p),G(q)) < \varepsilon \quad , \tag{6}$$

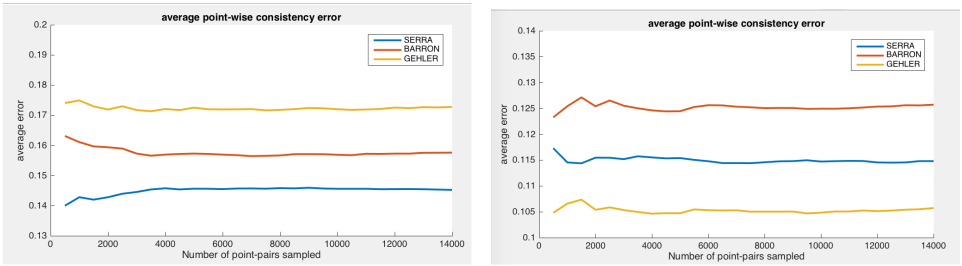
where ε is set to 10 as we observe that two colours are perceptually similar if their $\triangle E_{00}$ is less than 10.

A note on CIE colour measures

- We use $\triangle E_{00}$ to measure the similarity between the two selected points in Eq. (6) because $\triangle E_{00}$ offers the perceptual uniformity by correcting problems with blue colours and also improves the performance on gray colours [27]. This point selection criteria sets the constraint for the first principle of the reflectance consistency to be fulfilled.
- To compute the pair-wise consistency error, we utilize $\triangle E_{00}$ creatively in Eq. (4). One can think of using $\triangle E(G(p),G(q))$, $\triangle E(T(p),T(q))$ and compute the difference between them. However, this straight forward solution results 0 when the colours of points p and q are swapped between G and T. When combining $pce_{G,T}(p,q)$ in Eq. (4) with the point selection criteria in Eq. (6), our reflectance consistency principles hold.

4.3 Sampling and Results

In our experiment, we scale the images in the dataset to 1042 × 696 pixels to reduce the reflectance reconstruction time for different evaluated methods. We first start by sampling 500 pairs of points and increase 500 pairs for each iteration. The maximum number of sampling is 16,000 pairs for an average of 120,000 pixels in the regions of interest M in our dataset. For p percentile inliers selection, we choose $p = 95$. In other words, we eliminate 5% of the highest errors. When the number of samplings is large, this 5% elimination guarantees there's no outliers due to noises, but we still have enough number of samplings to fairly evaluate the reflectance between different methods. The results analysis in the Fig. 4 shows that



(a) Scene 5 with green and orange illuminant

(b) Scene 2 with white and yellow illuminant

Figure 4: The average point-wise consistency error of scene 5 (a) and scene 2 (b) with the colour difference threshold $\varepsilon = 10$ and the mean distance between pair points $\mu_d = 20$

the metric is able to distinct well the performance of different methods on different scenes with different illumination and is stable when the number of pairs increases to 8.000 and beyond.

Our point-wise consistency metric (PCM) evaluates the reflectance reconstruction of different methods based on the reflectance consistency principles. The PCM can be extended to evaluate reflectance at pairs of points that are perceptually different. It is also possible to develop the error estimation for shading based on this perceptual and statistical approach.

5 Evaluation

Table1 presents evaluation of three intrinsic image methods, namely: Serra *et al.* [31], Barron *et al.* [1], and Gehler *et al.* [16] using their publicly available codes and default parameters on a subset of our dataset (i.e. images captured by one of the 6 cameras for all the scenes and illuminations)[4]. The goal of this paper is not to compare the performance of all the existing methods, but rather to demonstrate the application of our dataset and metric. We further group our illumination conditions to four categories, i.e. easy (whitish), moderately coloured, hard (strongly coloured), and specular based on their complexity level. Evaluation is performed using PMC and LMSE. As PCM only evaluates reflectance data, we restrict LMSE in the same way in order to get comparable results.

To give further insight into the two measures, we pick a sequence of 20 different illuminations for a fixed camera pose for scene 5 (see Fig. 2, right). Fig. 5 compares the evaluation based on PCM and LMSE and shows the results of all three methods. PCM delivers consistent results ranking the methods' performance over all 20 illumination conditions. LMSE, on the other hand, shows large variations across different illuminations. Consulting LMSE for the 4th illumination , Barron and Gehler are similar and Serra is worse, while for PCM Serra is better than Barron and both better than Gehler. Visually on Fig. 6, all three methods deliver imperfect results for this example, but Gehler's methods appear to yield stronger color artefacts, e.g., on the bear's sleeve which PCM accounts for.

[4] Due to high computation time of the evaluated methods, we scale our images to 20% of their original resolution.

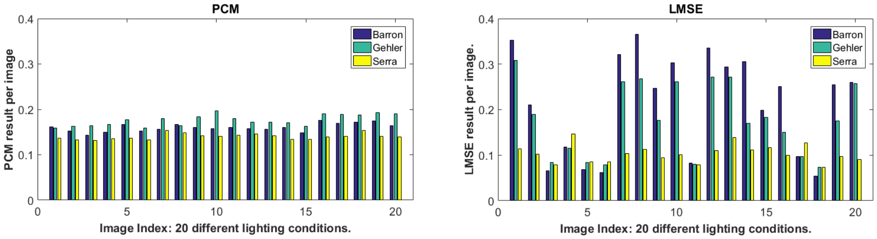


Figure 5: PCM and LMSE results on scene 5 under 20 different illuminations.

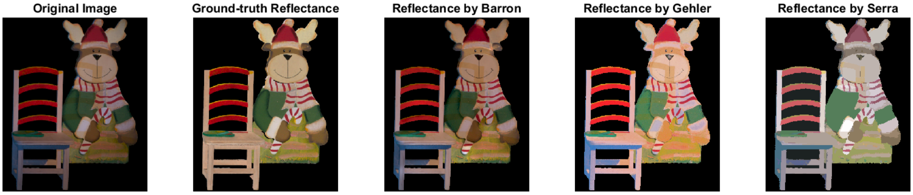


Figure 6: Reflectance results for the three methods on scene 5, 4th illumination.

Evaluation	Method	easy	moderate	hard	specular	Total
PCM	Barron *et al.*	0.149	0.151	0.157	0.157	0.154
	Gehler *et al.*	0.151	0.154	0.158	0.164	0.157
	Serra *et al.*	0.123	0.125	0.125	0.126	0.125
LMSE (reflectance)	Barron *et al.*	0.305	0.383	0.485	0.387	0.403
	Gehler *et al.*	0.277	0.341	0.441	0.336	0.360
	Serra *et al.*	0.253	0.300	0.319	0.289	0.296

Table 1: Evaluation of the methods using PCM and LMSE.

6 Conclusions

We contribute to the intrinsic image research a new high resolution multi-view real-photo dataset with precise pixel-wise reflectance and shading ground-truth. Our dataset presents challenging surface colour texture and multi-coloured illumination. To the best of our knowledge, the current work is the only real-photo dataset that provides all these features as well as high resolution depth. We also propose a new perceptually inspired metric to evaluate the intrinsic methods' results based on the reflectance consistency principle. We evaluate three state-of-the-art intrinsic methods on our dataset using LMSE and our proposed PCM metrics. We believe that our dataset and the PCM metric can help in improving the quality of intrinsic image methods in complex scenes.

Acknowledgment

This research was partially funded by the German Research Foundation (DFG) as part of the research training group GRK 1564 “Imaging New Modalities”. We would also like to thank Daniel Schneider and Dr. Rainer Bornemann from University of Siegen for their invaluable support regarding our experimental setup.

References

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

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

[3] Harry G. Barrow and Jay M. Tenenbaum. Recovering intrinsic scene characteristics from images. *Computer Vision Systems*, pages 3–26, 1978.

[4] Shida Beigpour, Marc Serra, Joost van de Weijer, Robert Benavente, María Vanrell, Olivier Penacchio, and Dimitris Samaras. Intrinsic image evaluation on synthetic complex scenes. In *IEEE International Conference on Image Processing (ICIP)*, pages 285–289, 2013.

[5] Shida Beigpour, Andreas Kolb, and Sven Kunz. A comprehensive multi-illuminant dataset for benchmarking of the intrinsic image algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 172–180, 2015.

[6] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014.

[7] Nicolas Bonneel, Kalyan Sunkavalli, James Tompkin, Deqing Sun, Sylvain Paris, and Hanspeter Pfister. Interactive intrinsic video editing. *ACM Transactions on Graphics (TOG)*, 33(6):197, 2014.

[8] Jean-Yves Bouguet. Camera calibration toolbox for matlab. 2004.

[9] Adrien Bousseau, Sylvain Paris, and Frédo Durand. User-assisted intrinsic images. In *ACM Transactions on Graphics (TOG)*, volume 28, page 130. ACM, 2009.

[10] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012*, pages 611–625. Springer, 2012.

[11] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 241–248, 2013.

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

[13] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building rome on a cloudless day. In *Computer Vision–ECCV 2010*, pages 368–381. Springer, 2010.

[14] E.N. Dalal G. Sharma, W. Wu. The CIEDE2000 color-difference formula: Implementation notes, supplementary test data, and mathematical observations. *Color Research and Application*, 30(1), February 2005.

[15] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. In *Computer Graphics Forum*, volume 31, pages 1415–1424. Wiley Online Library, 2012.

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

[17] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009.

[18] Mohammed Hachama, Bernard Ghanem, and Peter Wonka. Intrinsic scene decomposition from rgb-d images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 810–818, 2015.

[19] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the world* in six days*(as captured by the yahoo 100 million image dataset). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3287–3295, 2015.

[20] Junho Jeon, Sunghyun Cho, Xin Tong, and Seungyong Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *Computer Vision–ECCV 2014*, pages 218–233. Springer, 2014.

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

[22] Naejin Kong and Michael J Black. Intrinsic depth: Improving depth transfer with intrinsic images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3514–3522, 2015.

[23] Pierre-Yves Laffont and Jean-Charles Bazin. Intrinsic decomposition of image sequences from local temporal variations. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 433–441, 2015.

[24] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frédo Durand, and George Drettakis. Coherent intrinsic images from photo collections. *ACM Transactions on Graphics*, 31(6), 2012.

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

[26] Hao Li, Etienne Vouga, Anton Gudym, Linjie Luo, Jonathan T. Barron, and Gleb Gusev. 3d self-portraits. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2013)*, 32(6), November 2013.

[27] M Ronnier Luo, Guihua Cui, and B Rigg. The development of the CIE 2000 colour-difference formula: CIEDE2000. *Color Research & Application*, 26(5), 2001.

[28] Ricardo Martin-Brualla, David Gallup, and Steven M Seitz. 3d time-lapse reconstruction from internet photos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1332–1340, 2015.

[29] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2992–2992, 2015.

[30] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *European Conference on Computer Vision (ECCV)*, pages 746–760, 2012.

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

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

[33] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–7. IEEE, 2008.

[34] Jian Shi, Yue Dong, Xin Tong, and Yanyun Chen. Efficient intrinsic image decomposition for rgbd images. In *Proceedings of the 21st ACM Symposium on Virtual Reality Software and Technology*, pages 17–25. ACM, 2015.

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

[36] Marshall F Tappen, Edward H Adelson, and William T Freeman. Estimating intrinsic component images using non-linear regression. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1992–1999. IEEE, 2006.

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

[38] Genzhi Ye, Elena Garces, Yebin Liu, Qionghai Dai, and Diego Gutierrez. Intrinsic video and applications. *ACM Transactions on Graphics (TOG)*, 33(4):80, 2014.

[39] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1437–1444, 2012.

[40] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3469–3477, 2015.

[41] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. Learning ordinal relationships for mid-level vision. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 388–396, 2015.