HSI-Guided Intrinsic Image Decomposition for Outdoor Scenes

Fan Zhang[1] Shaodi You[2,4] Yu Li[3] Ying Fu[1*]
[1]Beijing Institute of Technology [2]University of Amsterdam
[3]International Digital Economy Academy [4]Jiangsu University of Science and Technology

Abstract

*Intrinisic image decomposition (IID) aims to recover the reflectance and shading components from images and is the prerequisite to many downstream computer vision applications, such as image editing and image relighting. Due to the inherent difficulty in acquiring ground truth reflectance and shading, existing datasets are either synthetic indoor scenes or objects using graphics rendering (*e.g.*, CGIntrinsics and ShapeNet* etc.*) or real photos with very sparse manual annotation (*e.g.*, IIW and SAW* etc.*). Accompanied with the complex nature of outdoor scenes, most IID methods focus on the decomposition of indoor environment. There is still a long way to go before we can handle IID of outdoor scenes. In this paper, we take the attempt to perform intrinsic image decomposition for outdoor scenes when RGB image is not the only thing we can get from the enviroment. With the observation of prior work where nir-infrared (NIR) images are transparent to a range of colourants/dyes, we propose to extend it to more spectra by collecting hyperspectral imaging (HSI) data which are well aligned with RGB images and to perform IID with both of them. We also apply existing mainstream IID methods for comparison to examine current progress and challenges at the road towards IID outdoors. We still make some improvements and find problems when performing IID for outdoor scenes, even though we do not handle it perfectly. The data we collect will be made publicly available for further potential investigation.*

1. Introduction

Intrinsic image decomposition (IID) aims to decompose an image into image formation components with different properties [3], *e.g.*, reflectance and shading. The former describes the surface charateristics of objects like albedo, color and texture, and the latter represents the shape of objects and effects caused by illuminations like shadow cast. This task is under research for several decades and is one of the key problem in the computer vision community. It is helpful to other downstream computer vision tasks like image editing [11], image relighting [48] and so on.

Under the Lambertian assumption, an image is dominated by the diffuse reflection, which is related to albedo of materials, shape of objects and illumination conditions, as formulated in Equations (1) and (2). Thus IID is a highly ill-posed and under-constrained problem, because there are multiple plausible combinations of reflectance and shading to form the same color image. Thus, researchers propose various priors [9, 26, 35, 49, 50] to constrain the problem to find feasible solutions. Retinex [26] assumes that reflectance is piece-wise constant and illumination varies gradually. Later on, traditional method proposed various priors to exploit the essence of images, such as textures cues [9, 44, 49, 50, 54], sparsity of reflectance [39, 45, 46], chromaticity segmentation [20], depth [1, 14, 27], user interaction [12, 43], infrared [17] and *etc*. With the advancement of deep learning, more methods turn to deep models to learn direct decomposition of images from training data [4, 6, 18, 36, 47]. However, the assumption aforementioned do not hold any more because of the complex nature of outdoor scene like varying illumination, shadow casts and specular reflections, which in turn makes these indoor algorithms work abnormally.

Besides, the ground truth of reflectance and shading are difficult to acquire. Existing IID datasets all have their own limitations. According to Equations (1) and (2), collecting ground truth directly requires direct control of the lighting. Therefore, these are either done in a dark room or by graphics rendering. MIT Intrinsics dataset [21] is the first dataset containing carefully constructed ground truth for images of objects. However this dataset only contains 16 objects and only one object in each image. While Intrinsic Images in the Wild (IIW) [10] and Shading Annotations in the Wild (SAW) [24] are two datasets containing sparse, crowd-sourced reflectance and shading annotations on real indoor images. They collect data with the help of users who manually annotate the relative relations between two pixels. These datasets only provide sparse annotations which are not physically reliable and mainly focus on indoor scenes.

*Corresponding author: fuying@bit.edu.cn

Besides, there are also many synthetic datasets generated using rendering, like CGIntrinsics [31], ShapeNet [47], and MPI-Sintel [13]. These datasets are large in amount and provide dense ground truth images of reflectance and shading. The limitation of these datasets is that they also mainly focus on indoor scenes and synthetic data has domain gap with real scenes, although they are proved to be effective in training deep models with good performance. Recently, Cheng *et al*. [17] propose the NIR-RGB dataset utilizing NIR image as shading reference to guide the decomposition based on several gradient priors. This dataset is also small in amount and similar to that of MIT Intrinsics with additional NIR images. There is still no real outdoor dataset for IID, to the best of our knowledge, which means that training a deep model for outdoor scenes is not physically possible.

In this paper, we take the try to perform IID for outdoor scenes when RGB data is not the only thing we can get. With the observation of prior work [17] where NIR image is transparent to a range of colourants/dyes, we propose to collect HSI data together with RGB images of outdoor scenes and to perform IID with both of them. On the basis of this observation, HSI image can capture the shape information of objects and illumination changes of the scene, which is helpful for the IID of outdoor scenes. In addition, we also apply mainstream existing IID methods to outdoor scenes for comparison to examine where we exactly are at the road towards IID outdoors. We investigate their performance on outdoor scenes and find out their drawbacks, which gives us insights on what we should handle when facing outdoor scenes.

The main contributions of this paper can be summerized as follows:

- We take our attempt to perform IID on outdoor scenes with the help of HSI data, which is based on the observation of prior work and makes use of different channels in HSI image within NIR spectra, utilizing their transparency to textures of different materials.
- We propose to collect HSI data together with RGB images for outdoor scenes and guide the decomposition of color images with the help of HSI information.
- We also apply existing mainstream IID methods to outdoor scenes for comparison to investigate their performanceand find out drawbacks for insights on further improvement.

2. Related Work

In this section, we briefly review existing mainstream methods and datasets for IID task.

2.1. IID Methods

Due to the ill-posed nature, traditional methods propose various priors. Retinex theory [26] is one of the earliest methods. It assumes that reflectance is piece-wise constant and illumination changes slowly. Funt *et al*. [19] extend Retinex algorithm to color images using chromacity information due to its invariance to shading component. Bell and Freeman [9], Tappen *et al*. [49, 50], Zhao *et al*. [54] and Shen *et al*. [44] propose to utilize texture cues to guide the IID. While Weiss *et al*. [51] and Matsushita *et al*. [35] estimate intrinsic images based on image sequences assuming that reflectance is constant and the illumination changes. Grosse *et al*. [21] also extend the Retinex theory to color image. Bousseau *et al*. [12] and Shen *et al*. [43] propose to utilize user interaction. [39, 45, 46] propose tp make use of the sparsity of reflectance. There are also methods based on depth cues [1, 14, 27]. Barron *et al*. [2] propose a series of priors to estimate the shape, surface normals, reflectance, shading and illumination from a single image. Xu *et al*. [52] propose a struture and texture aware Retinex mode. Cheng *et al*. [17] propose to uitlize NIR images for IID.

With the advancement of deep learning, many works [23, 25, 37, 38, 55] turn to learning-based models. Narihira *et al*. [36] are the first to learn end-to-end network in a data-driven manner. Zoran *et al*. [56] learn a deep network to classify the pairwise points from both local and global contextual information. Kovacs *et al*. [24] propose to train a CNN to predict per-pixel shading information in an image. Shi *et al*. [47] propose to introduce inter-links between decoders to utilize the correlation between intrinsic components. Janner *et al*. [22] explore the problem in a self-supervised setting and Lettry *et al*. [28] make use of adversarial residual networks. Fan *et al*. [18] apply a flexible loss layer for training a universal model on both fully-labeled and weakly-labeled datasets. Cheng *et al*. [16] use a Laplacian pyramid inspired neural network architecture to exploit scale space properties. Li *et al*. [31] combine four datasets with specialized loss functions for training. There are also learning-based methods based on image sequences [25,29,32]. Baslamisli *et al*. [5] propose to perform IID and semantic segmentation jointly with one network. They also propose RetiNet [6] based on Retinex theory. [7] first get rough shading component with physics-based prior and then get final decomposition with refinement network. There are also unsupervised methods [33, 34] proposed for this task with single image.

While most existing methods focus on indoor IID, in this paper, we take the attempt to perform IID for outdoor images by utilizing HSI, based on the observation that different spectra have varying degrees of transparency to colourants/dyes.

2.2. IID Datasets

Grosse *et al*. [21] collect a dataset called MIT Intrinsics to serve as a benchmark for IID. It is composed of images of 16 real objects which are decomposed into Lambertian shading, Lambertian reflectance, and specularities. Only relative shading and reflectance are provided. Cheng *et al*. [17] propose the NIR-RGB dataset in the way of [21], which contains not noly RGB images with its corresponding reflectance and shading components but also NIR images. Beigpour *et al*. [8] propose the Multi-Illuminant Intrinsic Images (MIII) Dataset containing 75 images with ground-truth intrinsics. Shi *et al*. [47] build a large-scale synthetic dataset based on the ShapeNet dataset, and render millions of synthetic images with specular materials and environment maps. Chen *et al*. [15] present Spectral Intrinsic Images Dataset (SIID) with 18 spectral images. These datasets are all object-level.

Bell *et al*. [10] propose the Intrnsic Images in the Wild (IIW) dataset utilizing crowdsourcing to acquire pair-wise reflectance comparisons for photos. It contains 5,230 photos, includes 875,833 reflectance comparisons. Kovacs *et al*. [24] also make use of crowdsourcing to collect shading annotations and called their dataset Shading Annotations in the Wild (SAW) dataset, which contains 6,677 images including 15K shadow boundary points and 24K constant shading regions. The two datasets both provide no ground truth intrinsic images.

Butler *et al*. [13] propose the MPI-Sintel dataset containing 23 rendered video sequences for assessing optical flow methods. It also offers ground truth reflectance and depth, which thus has been used for IID. Bonneel *et al*. [11] provide a dataset with 53 high quality realistic scene-level renderings under different illumination settings with corresponding per-pixel ground-truth intrinsics. Li *et al*. [31] propose the CGIntrinsics dataset of physically-based rendered images of scenes with full ground truth decompositions. It consists of over 20,000 images of indoor scenes, based on the SUNCG dataset. Baslamisli *et al*. [4, 5] extend a subset of the (synthetic) Natural Environment Dataset (NED) to generate reflectance, direct shading, ambient light and shadow cast ground-truth image, which contains around 25k images of 15 gardens for training and around 5k images of 3 gardens for testing.

In this work, we propose to collect HSI data together with RGB images for outdoor scenes, seeking to perform IID on outdoor scenes with the help of HSI information.

3. Data Preparations

In this section, we first introduce the equipment, collection and details of collecting HSI data together with RGB images for outdoor scenes.

3.1. Motivation

IID is a basic task in computer graphics and vision with a long history and has been studied since 1970s [26]. It aims to decompose color images into image formation components including but not limited to reflectance and shading. It is vital to understand formulation of natural images and those components are benefitial to other tasks because they are no longer intertwined with each other. For example, semantic segmentation can utilize reflectance images for they contain no illumination effects, while shape-from-shading methods can utilize shading images for they describe the shape geometry of objects.

Although it has been studied for decades, ground truth for intrinsic images is very difficult to acquire. It is only possible in strictly controlled laboratory environment to collect object-level intrinsic images [21] and the procedure is complicated, time and labor consuming. It is not applicable to scene-level data collection. Thanks to CG rendering, the problem of lack of data is greatly released by synthetic data. However, existing synthetic datasets mainly contain indoor scenes while outdoor scenes are relatively few. Meanwhile, there also exists domain gap between rendered and real images.

Despite these difficulties, we turn our focus from the decomposition of object-level images or indoor scenes to outdoor scenes. Recently, Cheng *et al*. [17] propose to guide the IID of RGB image with help of NIR image based on the observation that NIR image is transparent to a range of colourants/dyes. Although they only apply their method to object-level images, there is potential for the decomposition of outdoor scenes. It is possible to extend this observation of NIR image to HSI data because different spectra within NIR range of HSI image are diffrently transparent to textures of materials. Thus we make our attemp to perform IID on outdoor images with the help of HSI information.

3.2. Data Collection

To perform IID with both RGB and HSI information, we need to capture the RGB image and corresponding HSI data of the same scene and keep them well aligned with each other. We are equipped with the LightGene Hyperspectral Camera for data collection. This camera is able to capture RGB image and corresponding sparse HSI information at the same time and reconstruct the final dense HSI data based on them. It can also record videos of RGB information together with HSI information, which fits our need.

In particular, the camera can record the spectral information within the range of 449-955nm at the interval of 4nm, resulting in 128 channels in total. The final resolution of RGB image and corresponding HSI data after alignment is 1889×1422 pixels and we crop them by 1600×1200 pixels for final data. On the basis of the observation of NIR image aforementioned, we only take use of the last 60 channels of

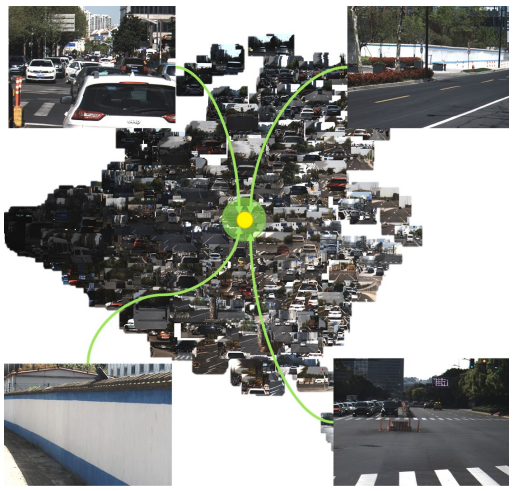

Figure 1. The overview of our collected data for intrinsic image decomposition containing various scenes and illuminations.

the HSI data for following procedures.

We mount our optical system on a car and collect the outdoor data while driving on the road. We collected our data in Shanghai during June. The speed of the vehicle while collecting the data keeps within the range of 20-50km/h. The framerate of RGB camera and HSI camera is 1fps.

3.3. Data Selection

Different from capturing images for static scenes, collecting image data while driving faces the motion blur problem. To obtain as much valid data as we can, we record large amounts of RGB images and corresponding HSI data while driving. After the whole collection work, we get roughly more than 26000 RGB images and corresponding hyperspectral data. These images contain so many motion-blurred or defocused images due to the moving of the car. So we manually filtered out these images with obvious motion-blur and defocusing problem as well as images with bad illumination conditions. The manual filtering results in around 4600 normal images as shown in Figure 1. Because the hyperspectral camera together with RGB camera is working continuously, we collect sequences of images and corresponding HSI data for the same scene. So there are about 260 pairs of RGB image and HSI data covering different scenes.

4. Method

In this section, we first describe the image formation model, then detail how we decompose color images of outdoor scenes into intrinsic images with the help of HSI information.

4.1. Image Formulation Model

IID aims to decompose a color image into its reflectance and shading component. According to the dichromatic reflection model proposed by Shafer [41], an image I^λ can be described as the sum of a diffuse reflection I_d^λ and a specular reflection I_s^λ as follows:

$$I^\lambda = I_d^\lambda + I_s^\lambda, \tag{1}$$

where the diffuse reflection component I_d^λ is commonly assumed to be dominant over the other component under the Lambertian assumption and thus I_s^λ is negligible, *i.e.* $I^\lambda \approx I_d^\lambda$. Furthermore, an image I^λ over the specific spectrum ω is the integration of overall light signal arrived at the camera sensor and can be modelled by as follows:

$$I^\lambda = m(\boldsymbol{n}, \boldsymbol{l}) \int_\omega e(\lambda)\rho(\lambda)f(\lambda)d\lambda, \tag{2}$$

where $\boldsymbol{n}$ and $\boldsymbol{l}$ reprensent the vectors of surface normal and incoming light direction, respectively. $m(\boldsymbol{n}, \boldsymbol{l})$ forms the geometric dependencies and is correlated to the shape of objects. λ stands for the wavelength of light signal and $\rho(\lambda)$ is the reflectance or namely albedo of a surface. It controls the component of reflected light from objects and is only related to the material itself, for which we want to estimate from RGB images. While $e(\lambda)$ denotes the spectral power distribution of light source and is namely the illumination. It together with $m(\boldsymbol{n}, \boldsymbol{l})$ controls the final light signal reflected from the surface of objects and forms the shading we need to estimate. Finally, $f(\lambda)$ is the camera spectral response function and it describes how the light signal is integrated into electric signal after arriving at the sensor. Therefore, following a common assumption of linear response camera model, the above equation can be simplified as:

$$I^\lambda = m(\boldsymbol{n}, \boldsymbol{l})e(\lambda)\rho(\lambda), \tag{3}$$

where I_λ is the captured image of the light signal within the wavelength λ. Thus an image can finally be expressed as:

$$I^\lambda \approx I_d^\lambda = S^\lambda * R^\lambda, \tag{4}$$

where $S^\lambda = m(\boldsymbol{n}, \boldsymbol{l})e(\lambda)$ and $R^\lambda = \rho(\lambda)$, respectively. According to Equation (4), we can know that IID is actually a general task appliable to spectral images [15], *i.e.* all channels within a spectral cube. For the IID task on RGB images, the image is composed of three channels $\{R, G, B\}$ selected from the visible spectrum and the shading component S is single-channel under the white light assumption. If the light source is colored, then the color information should be recorded in S and it becomes a three-channel image instead.

4.2. Energy-based Optimization

In this paper, we target on the IID on color images with the help of HSI information. On the basis of the observation that NIR image is transparent to a variety of textures of materials, Cheng *et al*. [17] propose to minimize the energy function based on smoothness priors of reflectance and shading with the assist of NIR image as guidance. We follow them to solve the same problem making use of HSI information instead of NIR image alone. The overall energy term for the decomposition is formulated as follows:

$$E(S,R) = E_S(S) + E_R(R) + E_I(S,R), \tag{5}$$

where the three elements represent the energy constraints imposed on shading, reflectance and color image, respectively.

For the first energy term $E_S(S)$, Cheng *et al*. [17] make use of the observation that textures are absent in NIR images and impose smoothness constraint on the shading component. Aside from the commonly used constraints on chromacity and intensity of color images according to Colour Retinex [21]:

$$E_S^{rtx}(S) = \sum_{x,y\in\mathcal{N}} \big(\log(S_x) - \log(S_y) - \epsilon_{xy}\big)^2, \tag{6}$$

where ϵ_{xy} is the threshold variable controlling whether the decomposition is based on information from either color image or NIR guidance. In addition, they propose to penalize the local area whose shading variation is greater than that of NIR image:

$$\begin{aligned} E_S^{HSI}(S) = \sum_{x,y\in\mathcal{N}} \big(max(0, |\log(S_x) - \log(S_y)| \\ - |\log(I_x^{nir}) - \log(I_y^{nir})|)\big), \end{aligned} \tag{7}$$

where $\mathcal{N}$ stands for the set of pixels in a local area, x and y denote the pixels in that area.

However, only NIR image is not enough to guide the decomposition on outdoor scenes which are more complicated than objects and indoor scenes for there are various materials and plants which may have fluorescence effect. The smoothness assumption of NIR image may no longer hold any more in outdoor environments. Thus we propose to make use of more information from hyperspectral data to handle such complicated scenes. We choose channels within NIR band from the HSI cube data and last 60 out of 128 channels are used to guide decomposition procedure. Here we adopt the naive way to combine the multi-channel spectral information where we calculate the per-pixel average values of all 60 channels to get the final guidance image. Among different channels, the spectral curves of HSI camera differ from each other and the resulting spectral images looks different, too. However not all areas in the image have obvious difference and we can reduce it by simply averaging among channels, which is more robust and gets affected by textures more difficultly than single NIR image. In a result, we modify the above $E_S^{NIR}(S)$ term into our HSI version:

$$\begin{aligned} E_S^{HSI}(S) = \sum_{x,y\in\mathcal{N}} \big(max(0, |\log(S_x) - \log(S_y)| \\ - |\log(\bar{I}_x^{\lambda}) - \log(\bar{I}_y^{\lambda})|)\big), \end{aligned} \tag{8}$$

where $\bar{I}^{\lambda}$ reprensents the average image of the last 60 channels from the HSI cube data.

With this energy term, we impose a constraint on the shading component that only an area that contains less textures than all 60 channels within the NIR band is recognized to belong to the shading. In this way, we make full use of the diverse information provided by the HSI image and avoid the shading ambiguity caused by textures from reflectance due to the limitation of single NIR image.

In a result, the final energy of shading is the sum of two components in Equations (6) and (8):

$$E_S(S) = E_S^{HSI}(S) + E_S^{rtx}(S). \tag{9}$$

As for the energy term of reflectance and color image, we follow the implementation of Cheng *et al*. [17] to make use of the local and non-local constraints on homogeneity of reflectance and loose constraint on non-Lambertian surfaces. The formulations of $E_R(R)$ and $E_I(S,R)$ are expressed as:

$$E_R(R) = (1-\alpha)E_R^{loc}(R) + \alpha E_R^{non-loc}(R), \tag{10}$$

and

$$\begin{aligned} E_I(S,R) = \beta\big((I - e^{\log(S)+\log(R)}\big)^2 \\ + 0.05\big(\log(I) - \log(S) - \log(R)\big)^2\big). \end{aligned} \tag{11}$$

With energy components all set, we follow Cheng *et al*. [17] to adopt an L-BFGS algorithm to minimize this energy.

5. Experiments

In this section, we first introduce IID methods we apply to the outdoor scenes for comparison. Then we provide both quantitative and qualitative results.

5.1. Compared Methods

Following Section 2, we select 12 methods in total for comparison, including traditional optimization-based methods, supervised and unsupervised methods. For optimization-based methods, we select IIW [10], IID-Optim [42] and STAR [52]. For learning-based methods, we choose several supervised and unsupervised methods trained on synthetic data or real sparse data, including DirectIntrinsic [36], ShapeNet [47], CGIntrinsics [31], IntrinSeg [5], Revisiting [18], InverseRenderNet (IRN for short)

Table 1. Earth Mover's Distance (EMD) between gradient histograms of results from compared methods and ground truth sampled from CGIntrinsics. The lower value the better.

Method	EMD		
	Reflectance	Shading	Avg.
IID-Optim [42]	4.7260	**2.4498**	3.5879
IIW [10]	4.3713	3.6654	4.0184
STAR [52]	6.5376	5.5730	6.0553
IIDWW [32]	4.5869	4.4717	4.5293
UidSequence [29]	5.2263	3.8328	4.5296
USI3D [33]	5.3154	4.5390	4.9272
DirectIntrinsic [36]	4.8424	4.5275	4.6850
ShapeNet [47]	6.0710	3.0560	4.5635
CGIntrinsics [31]	4.9432	5.5385	5.2409
IntrinSeg [5]	4.6229	4.2126	4.4178
Revisiting [18]	4.3952	3.6198	4.0075
IRN [53]	3.4133	3.7276	3.5705
Ours	**3.1754**	3.6579	**3.4167**

[53], UidSequence [29], IIDWW [32] and USI3D [33]. The first six methods belong to the supervised methods and last three methods are unsupervised ones. Results of all methods are produced by the publicly availble code and pretrained models and parameters are kept the same as default.

5.2. Quantitative Results

Due to the lack of ground truth for outdoor scene, we do not adopt commonly used MSE, LMSE and DSSIM metrics for quantitative evalutation. According to [30], reflectance layer holds more sparse distribution for larger gradients compared to shading layer resulting from its piece-wise homogeneity property. Thus, we seek to measure the similarity between gradient histograms of results of all compared methods and ground truth from existing synthetic dataset. Specifically, we utilize the Earth Mover's Distance (EMD) [40] which indicates how close two distributions are to each other. We randomly sample the same amount of ground truth from CGIntrinsics dataset and get gradient maps of both reflectance and shading using Sobel descriptor. With histograms of gradient maps of all results, we can acquire EMD metrics for quantitative comparison as listed in Table 1, with lower values for better performance. We can find that our method gets lowest EMD in terms of reflectance and relatively nice result for shading, as well as lowest value for average performace.

5.3. Qualitative Results

We aim to find out the weaknesses of exsiting methods about their performance on outdoor scenes. Here we provide the results of reflectance and shading component produced by all methods along with input RGB images as well as our HSI-assisted decompositions, as illustrated in Figure 2 and 3.

At first glance, results of all compared methods look diverse and different from each other. But we can still find valuable points among these comparisons. In outdoor scenes, white illumination assumption does not hold anymore and colored illumination changes when images are captured at different time. Methods that assume shading as single-channel layer image suffer from color casts in their reflectances, including our HSI-assisted method. However, several methods that produce colored shading also suffer from color cast problem even in both reflectances and shadings due to domain gap, like IntrinSeg [5] and UniSequence [29].

For optimization-based methods, IID-Optim [42] gets relatvely good reflectances but also suffers from blurring inhomogeneity problems. IIW [10] and STAR [52] also get tidy reflectances, but the former suffers from low contrast and whitening problem while reflectances of latter show light chromacity. For unsupervised methods, the reflectances of IIDWW [32] are overcast, while shadings suffer from checkerboard artifacts due to the usage of deconvolution layers. For UidSequence [29], the reflectances are reddish and the shadings are bluish. Most areas in the reflectances are blurry while shadings are better than reflectances. For USI3D [33], the color of reflectances are different from each other in terms of tone. The shadings look good but also suffer from blurring problem. Among supervised methods, DirectIntrinsic [36] and ShapeNet [47] get relatively worse results compared to other four methods. Reflectances of DirectIntrinsic [36] are not good and only green and red areas like trees, grass are kept. Other areas are not correctly estimated and both shadings and reflectances are blurry. While ShapeNet [47] contains many black areas in the reflectances. As for the last four methods, CGIntrinsics [31] suffers from checkerboard artifacts and over-saturation problem. The reflectance are somewhat over-saturated in terms of the red, orange and blue areas. The roads are less homogeneous. The shading components are over-smoothed that it is hard to identify different areas. Because this model are trained on a mixture of four datasets, it is somewhat robust to real outdoor scenes. For IntrinSeg [5], reflectances look purplish compared to RGB image and shadings are instead yellow. while Revisting [18] looks unnatural and faces inhomogeneity problem. IRN [53] also suffers from artifacts and looks like oil painting.

While our HSI-assisted method produce somewhat reasonable decompositions on outdoor scenes. The reflectances look tidy and keep similar to the scenes captured in the RGB images and the shadings record illumination changes. Focusing on the road surface, reflectance of our method are piece-wise homogenous on areas of roads and changes are kept in shading components. What matters is actually the white roadlines, nearly all methods fail to clear these area in shading component while our method seems to somehow enhance these areas in shadings. The reason is our HSI data capture these roadlines in every single channel

Figure 2. Qualitative results of all compared methods. Gamma correction is applied for better visualization. Compared methods include IIW [10], IID-Optim [42], STAR [52], UidSequence [29], IIDWW [32] and USI3D [33].

of different wavelength and the method cannot discriminate these areas from shading. Our reflectances can also split those shadow casts on the road while strong shadows beneath cars still remain.

6. Conclusion

In this paper, we attempted to perform IID on outdoor scenes with the help of HSI. We first collect large amounts of paired RGB images and corresponding well-aligned HSI

Figure 3. Qualitative results of all compared methods. Compared methods include DirectIntrinsic [36], ShapeNet [47], CGIntrinsics [31], IntrinSeg [5], Revisiting [18], InverseRenderNet (IRN for short) [53].

data and filter out good-quality pairs which suffers no motion blur and defocusing problem. Then we utilized a novel method to perform intrinsic image decomposition on color images with the help of HSI information, inspired by the observation that NIR image is trasparent to a range of colourants/dyes. We also evaluated existing IID methods on our collected data to investigate into their overall performance on outdoor scenes and have found out drawbacks for insights of future improvement. Our collected RGB images together with well-aligned HSI data will be made publicly available for further potential usage. In the future, we will investigate into better strategy of utilizing HSI information to get more reliable intrinsic images regarding the problem of specular reflections, shadow casts and *etc*.

7. Limitation Discussion and Broder Impact

Our method is based on the observation that NIR image is transparent to texture of many materials and it cannot hold perfectly for real outdoor scenes. Although we utilize the last 60 channels of HSI data, textures cannot be totally removed in the resulting guidance image. Better adaptive strategy will be investigated for better utilization of HSI information.

Acknowledgments This work was supported by the National Natural Science Foundation of China under Grants No. 62171038, No. 61827901, No. 62088101 and No. 62006101.

References

[1] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, 2013. 1, 2

[2] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2014. 2

[3] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Computer Vision System*, 2(3-26):2, 1978. 1

[4] Anil S Baslamisli, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Shadingnet: image intrinsics by fine-grained shading decomposition. *IJCV*, pages 1–29, 2021. 1, 3

[5] Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Joint learning of intrinsic images and semantic segmentation. In *ECCV*, 2018. 2, 3, 5, 6, 8

[6] Anil S Baslamisli, Hoang-An Le, and Theo Gevers. Cnn based learning using reflection and retinex models for intrinsic image decomposition. In *CVPR*, 2018. 1, 2

[7] Anil S Baslamisli, Yang Liu, Sezer Karaoglu, and Theo Gevers. Physics-based shading reconstruction for intrinsic image decomposition. *Computer Vision and Image Understanding*, 205:103183, 2021. 2

[8] Shida Beigpour, Andreas Kolb, and Sven Kunz. A comprehensive multi-illuminant dataset for benchmarking of the intrinsic image algorithms. In *ICCV*, 2015. 3

[9] Matt Bell and ET Freeman. Learning local evidence for shading and reflectance. In *ICCV*, 2001. 1, 2

[10] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM TOG*, 33(4):1–12, 2014. 1, 3, 5, 6, 7

[11] Nicolas Bonneel, Balazs Kovacs, Sylvain Paris, and Kavita Bala. Intrinsic decompositions for image editing. *Computer Graphics Forum*, 36(2):593–609, 2017. 1, 3

[12] Adrien Bousseau, Sylvain Paris, and Frédo Durand. User-assisted intrinsic images. In *ACM SIGGRAPH Asia*, pages 1–10. 2009. 1, 2

[13] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 2, 3

[14] Qifeng Chen and Vladlen Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013. 1, 2

[15] Xi Chen, Weixin Zhu, Yang Zhao, Yao Yu, Yu Zhou, Tao Yue, Sidan Du, and Xun Cao. Intrinsic decomposition from a single spectral image. *Applied optics*, 56(20):5676–5684, 2017. 3, 4

[16] Lechao Cheng, Chengyi Zhang, and Zicheng Liao. Intrinsic image transformation via scale space decomposition. In *CVPR*, 2018. 2

[17] Ziang Cheng, Yinqiang Zheng, Shaodi You, and Imari Sato. Non-local intrinsic decomposition with near-infrared priors. In *ICCV*, 2019. 1, 2, 3, 5

[18] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018. 1, 2, 5, 6, 8

[19] Brian V Funt, Mark S Drew, and Michael Brockington. Recovering shading from color images. In *ECCV*, 1992. 2

[20] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. *Computer graphics forum*, 31(4):1415–1424, 2012. 1

[21] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. 1, 2, 3, 5

[22] Michael Janner, Jiajun Wu, Tejas D Kulkarni, Ilker Yildirim, and Joshua B Tenenbaum. Self-supervised intrinsic image decomposition. *arXiv preprint arXiv:1711.03678*, 2017. 2

[23] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *ECCV*, 2016. 2

[24] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. In *CVPR*, 2017. 1, 2, 3

[25] Pierre-Yves Laffont and Jean-Charles Bazin. Intrinsic decomposition of image sequences from local temporal variations. In *ICCV*, 2015. 2

[26] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971. 1, 2, 3

[27] Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin. Estimation of intrinsic image sequences from image+ depth video. In *ECCV*, 2012. 1, 2

[28] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Darn: a deep adversarial residual network for intrinsic image decomposition. In *WACV*, 2018. 2

[29] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. *Computer Graphics Forum*, 37(7):409–419, 2018. 2, 6, 7

[30] Yu Li and Michael S Brown. Single image layer separation using relative smoothness. In *CVPR*, 2014. 6

[31] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018. 2, 3, 5, 6, 8

[32] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, 2018. 2, 6, 7

[33] Yunfei Liu, Yu Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, 2020. 2, 6, 7

[34] Yunfei Liu and Feng Lu. Separate in latent space: Unsupervised single image layer separation. In *AAAI*, 2020. 2

[35] Yasuyuki Matsushita, Stephen Lin, Sing Bing Kang, and Heung-Yeung Shum. Estimating intrinsic images from image sequences with biased illumination. In *ECCV*, 2004. 1, 2

[36] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 1, 2, 5, 6, 8

[37] Takuya Narihira, Michael Maire, and Stella X Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, 2015. 2

[38] Thomas Nestmeyer and Peter V Gehler. Reflectance adaptive filtering improves intrinsic image estimation. In *CVPR*, 2017. 2

[39] Carsten Rother, Martin Kiefel, Lumin Zhang, Bernhard Schölkopf, and Peter Gehler. Recovering intrinsic images with a global sparsity prior on reflectance. *NeurIPS*, 2011. 1, 2

[40] Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *IJCV*, 40(2):99–121, 2000. 6

[41] Steven A Shafer. Using color to separate reflection components. *Color Research & Application*, 10(4):210–218, 1985. 4

[42] Jianbing Shen, Xiaoshan Yang, Yunde Jia, and Xuelong Li. Intrinsic images using optimization. In *CVPR*, 2011. 5, 6, 7

[43] Jianbing Shen, Xiaoshan Yang, Xuelong Li, and Yunde Jia. Intrinsic image decomposition using optimization and user scribbles. *IEEE transactions on cybernetics*, 43(2):425–436, 2013. 1, 2

[44] Li Shen, Ping Tan, and Stephen Lin. Intrinsic image decomposition with non-local texture cues. In *CVPR*, 2008. 1, 2

[45] Li Shen and Chuohao Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR*, 2011. 1, 2

[46] Li Shen, Chuohao Yeo, and Binh-Son Hua. Intrinsic image decomposition using a sparse representation of reflectance. *IEEE TPAMI*, 35(12):2904–2915, 2013. 1, 2

[47] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 2017. 1, 2, 3, 5, 6, 8

[48] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 1

[49] Marshall F Tappen, Edward H Adelson, and William T Freeman. Estimating intrinsic component images using non-linear regression. In *CVPR*, 2006. 1, 2

[50] Marshall F Tappen, William T Freeman, and Edward H Adelson. Recovering intrinsic images from a single image. *IEEE TPAMI*, 27(9):1459–1472, 2005. 1, 2

[51] Yair Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001. 2

[52] Jun Xu, Yingkun Hou, Dongwei Ren, Li Liu, Fan Zhu, Mengyang Yu, Haoqian Wang, and Ling Shao. Star: A structure and texture aware retinex model. *IEEE TIP*, 29:5022–5037, 2020. 2, 5, 6, 7

[53] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *CVPR*, 2019. 6, 8

[54] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with non-local texture constraints. *IEEE TPAMI*, 34(7):1437–1444, 2012. 1, 2

[55] Tinghui Zhou, Philipp Krahenbuhl, and Alexei A Efros. Learning data-driven reflectance priors for intrinsic image decomposition. In *ICCV*, 2015. 2

[56] Daniel Zoran, Phillip Isola, Dilip Krishnan, and William T Freeman. Learning ordinal relationships for mid-level vision. In *ICCV*, 2015. 2