Intrinsic image decomposition using physics-based cues and CNNs

Das, P.; Karaoglu, S.; Gevers, T.

DOI
10.1016/j.cviu.2022.103538
Publication date
2022
Document Version
Final published version
Published in
Computer Vision and Image Understanding

Computer Vision and Image Understanding 223 (2022) 103538

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Intrinsic image decomposition using physics-based cues and CNNs

Partha Das *, Sezer Karaoglu, Theo Gevers

University of Amsterdam, Science Park 904, Amsterdam, 1098XH, The Netherlands
3D Universum, Science Park 400, Amsterdam, 1098XH, The Netherlands

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
41A05
41A10
65D05
65D17

Keywords:
Computer vision
Physics based vision
Intrinsics image decomposition
Deep learning

ABSTRACT

Intrinsic image decomposition is the decomposition of an image into its reflectance and shading components. The intrinsic image decomposition problem is inherently ill-posed, since there can be multiple solutions to compute the intrinsic components forming the same image. In this paper, we explore the use of physics-based priors. We also propose a new architecture that separates the learning components in a stacked manner. We explore various ways of integrating such priors into a deep learning system. Our method is trained and tested on a large synthetic garden dataset to assess its performance. It is evaluated and compared to state-of-the-art methods using two standard intrinsic datasets. Finally, the pre-trained network is tested on real world images to show the generalisation capabilities of the network.

1. Introduction

Intrinsic Image Decomposition (IID), under the Lambertian assumption, is the decomposition of an image (I) into its constituent Reflectance (R) and Shading (S) images (Barrow and Tenenbaum, 1978). Reflectance is independent of the lighting conditions and corresponds to the colour of the object. Shading is a function of the object's geometry and lighting conditions. IID is beneficial for different computer vision applications, for example, shading cues helps to recover the shape of an object (Wada et al., 1995), while reflectance cues allow for object material editing (Ye et al., 2014; Meka et al., 2016; Beigpour and van de Weijer, 2011) and semantic segmentation (Baslamisli et al., 2018a).

For a given I, there can be different combinations of R and S. Therefore, IID is considered as an under-constrained problem. Early IID methods use various constraints like depth and geometry cues (Barron and Malik, 2015; Gehler et al., 2011; Shen et al., 2008). These constraints are usually derived from the image formation model. For example, the Retinex model (Land and McCann, 1971) defines IID components in terms of gradients. Reflectance is attributed to stronger gradient changes and shading to weaker gradient changes. Similarly, Barron and Malik (2015) defines reflectance as a piece-wise constant image containing a limited colour palette. Based on a general reflection model, these methods are quite robust at generalisation and are not dataset biased. However, these physics-based methods are less effective for scenes containing complex illumination conditions.

Recently, deep learning-based methods (Narihira et al., 2015; Li and Snavely, 2018; Fan et al., 2018; Bell et al., 2014) are proposed to compute IID in data-driven way. For example, Li and Snavely (2018) use multiple datasets to solve the IID problem in an end-to-end deep learning way. These methods show robust performance in decomposing scenes containing complex illumination cues. However, a drawback is that these methods are purely data-driven, and dataset biased. As a result, they do not generalise well to scenes that are different than the scenes used in the training set.

The two classes of approaches to compute the IID have different sources of weaknesses and may not necessarily offer a complete framework by themselves. However, their combination could provide a more robust and complete strategy. Therefore, we propose a new integrated approach to compute IID by leveraging the best of two worlds. We propose (generalised) physics-based cues to steer a (data-driven) CNN model by an integrated IID processing pipeline.

To steer our data-driven CNN pipeline, the physics-based cues are derived from illumination and geometry invariants namely (1) Colour Ratios (Finlayson, 1992) (CR), and (2) Cross Colour Ratios (Gevers and Smeulders, 1999) (CCR). The aim is to compute an invariant representation as an early stage of the pipeline using CCR and CR defining the surface albedo and shading cues respectively.

An overview of the proposed network is given in Fig. 1. The model uses an invariant representation to steer the (data-driven) CNN model by an integrated IID processing pipeline. The invariant representation

* Corresponding author at: University of Amsterdam, Science Park 904, Amsterdam, 1098XH, The Netherlands.
E-mail address: p.das@uva.nl (P. Das).

https://doi.org/10.1016/j.cviu.2022.103538
Received 11 August 2021; Received in revised form 8 June 2022; Accepted 16 August 2022
Available online 23 August 2022

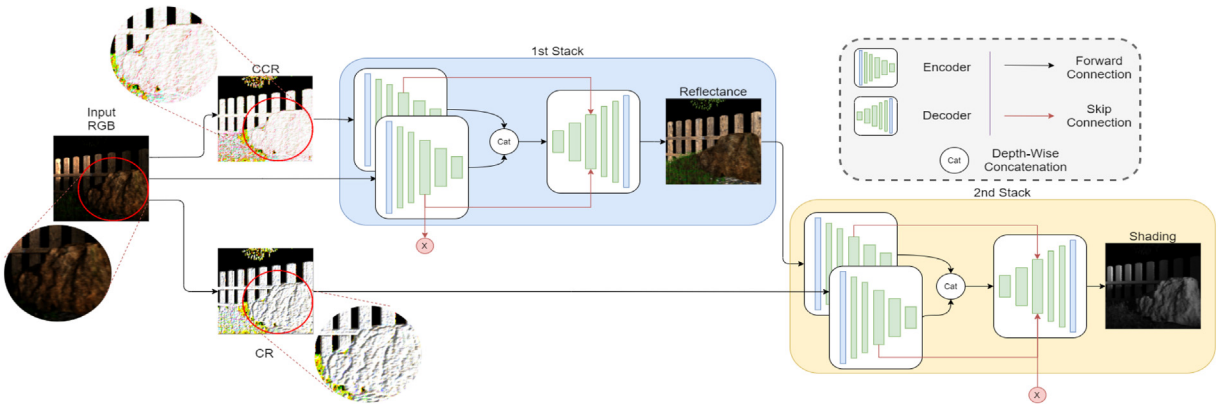

Fig. 1. An overview of the proposed network. *RGB* images are used to calculate the pixel-wise *CCR* and *CR* images. *CCR* gives a uniform reflectance, while *CR* contains geometric information, as can be observed on the highlighted rock (corresponding zoom-crops are shown). The *CCR* and input images are then passed on to separate encoders. The resulting features are concatenated and decoded into a reflectance image. The reflectance and *CR* images are used by two new encoders. The features are then decoded into a shading image. Skip connections are added between the encoder and decoders. Only one skip connection is shown for brevity. The image encoder features are shared by both stacks. The separated stacks model the interdependence of the components. More details about the encoder and decoder structure can be found in the supplementary material. Best viewed on a screen.

includes both CR and CCR invariants. The computation of these physics-based invariants may become erroneous for complex scenes where the imaging process does not follow the reflection of a general model. To address this shortcoming, a large learning network capacity is exploited to cope with possible sources of erroneous invariant values introduced by complex imaging conditions such as inter-reflections, coloured shadows, and specularities.

An important characteristic of the IID problem is the inter-dependency between the reflectance and shading components. For deep learning-based models this by simultaneously learning the components (Narihira et al., 2015; Shi et al., 2017), thus relying on the simultaneous forward and backward propagation to learn the inter-dependency. However, this may lead to artefacts, like shading leakages in the reflectance and vice versa. Therefore, a stacked approach is proposed to model the reflectance–shading inter-dependency where the shading stack exploits reflectance cues as priors.

The proposed method is trained on the NED (Baslamisli et al., 2018a) synthetic dataset. The trained model is then tested on different IID datasets including the MIT (Grosse et al., 2009) and Sintel (Butler et al., 2012) datasets. In addition, we provide qualitative evaluations on real-world datasets such as NYU (Silberman et al., 2012) and Trimbot outdoor garden (Sattler et al., 2017) dataset. Our approach is compared to state-of-the-art methods in terms of usual IID metrics and generalisation capabilities.

The contributions of this paper are as follows:

- We propose (generalised) physics-based cues to steer a (data-driven) CNN model by an integrated IID processing pipeline.
- A stacked approach is used to exploit reflectance cues for shading computation based on the imaging formation process, i.e. separating the intrinsic components for the learning phase is beneficial. The proposed method is able to cope with the dataset bias problem and has generalisation capacity to unseen domains.

2. Related work

Physics based: The image formation model is used to address the problem of IID. The landmark work by Land and McCann (1971) proposes the Retinex algorithm where stronger gradients are used to identify reflectance changes. Funt et al. (1992) extends this principle to colour images. Barron and Malik (2015) explores component specific properties by introducing physics-based constraints. A smoothness prior is added to the shading image to compute a rough depth map. Reflectance specific constraints like piece-wise constancy and parsimony, are then used to optimise the reflectance. Sheng et al. (2020) extends the shading constraints to compensate for shading changes. Chen and Koltun (2013) directly breaks down the shading component into the physical interaction of object geometry and light. Li et al. (2021a) uses RGB-D images coupled with a sparsity regularisation term to obtain IID components. However, the above methods are often designed for single objects or need specialised hardware for the additional input. Moreover, they may fail for scenes with complex imaging conditions (e.g. scenes containing unusual illumination and reflectance properties).

Learning based: With the introduction of large datasets for the IID problem (Chang et al., 2015; Li and Snavely, 2018; Butler et al., 2012; Li et al., 2021b), various learning based methods are proposed. Narihira et al. (2015) employs an end-to-end CNN to parameterise the IID problem. Shi et al. (2017) extends it by enforcing inter-dependence between the intrinsic components. Baslamisli et al. (2018b) uses a deep learning approximation of the Retinex algorithm, by learning in the gradients domain with a CNN model. The above-mentioned datasets are synthetic and do not always model the image formation complexities of real-world scenes. Bonneel et al. (2014) introduces a human-in-the-loop approach to address this. This is further explored to generate sparse crowd sourced reflectance and shading annotations (Bell et al., 2014; Kovacs et al., 2017; Narihira et al., 2015). Zhou et al. (2015) uses these sparse annotations to steer the proposed network. Li and Snavely (2018) makes use of both synthetic and real-world datasets and multiple losses to learn the IID problem. On the other hand, edge maps are used as guidance priors to enforce piece-wise smooth reflectance (Fan et al., 2018), while Zhu et al. (2021) uses images rendered by game engines for training on outdoor scenes. Cheng et al. (2018) explores different scale space properties. Shelhamer et al. (2015) and Kim et al. (2016) include depth in their formulation in a joint learning manner. Luo et al. (2020) uses surface normal information to cope with illumination for indoor images. Baslamisli et al. (2018a) uses semantic segmentation to jointly learn intrinsic image decomposition. Sengupta et al. (2019) also introduces an inverse rendering network that decomposes an input into its albedo, normals and illumination. However, these fully data-driven methods are dataset biased and hence will come up short to generalise well to unseen data.

The two classes of IID approaches both have their weaknesses: (1) physics-based methods may fall short in performing well in realistic environments and (2) (data-driven) CNN models may suffer from the dataset bias problem. Therefore, we propose a combined approach to compute the IID. Our method uses the strength of one approach to cope with the weakness of the other approach. To this end, our method uses a (generalised) invariant representation derived from general physics modelling to steer the (data-driven) CNN model which can cope with erroneous invariant values caused by complex scenes.

3. Methodology

3.1. Image formation

Under the Lambertian assumption of diffuse image formation (Shafer, 1985), we have:

$$I = m(\vec{n}, \vec{l}) \int_{\omega} \rho_b(\lambda)\, e(\lambda)\, f(\lambda)\, \mathrm{d}\lambda\,, \tag{1}$$

where, I is the captured image. λ is the incoming light wavelength within the visible spectrum ω. m is a function of geometry of the object and lighting. $\vec{n}$ denotes surface normal and $\vec{l}$ denotes the light source direction. f indicates the spectral camera sensitivity and e describes the spectral power distribution of the light source. ρ denotes reflectance and is related to the colour of the object. Assuming a linear sensor response, a narrow band filter and a single (white) light source, Eq. (1) can be simplified to:

$$I = S \times R\,, \tag{2}$$

where R is the reflectance image, the (albedo) colour of the object, and S is the shading component associated with geometry and illumination. Obviously, Eq. (2) can have multiple solutions for the same I value.

3.2. Physics-based invariants

In this paper, (generalised) invariant representations are used to condition the IID process.

Colour Ratios: Finlayson (1992) introduces *Colour Ratios* (CR), which are illuminant invariant edge descriptors. Consider two neighbouring pixels, x_1 and x_2, in a RGB image coming from a flat surface with a locally constant light source:

$$e^c(\vec{x}_1) = e^c(\vec{x}_2) \tag{3}$$

where, $e^c(x_i)$ is the spectral power distribution function of the illumination for colour channel c and pixel x_i. Assuming the same light source is justifiable for (adjacent) neighbouring pixels. Then, CR is defined by:

$$MC_R = \frac{R_{x_1}}{R_{x_2}},\; MC_G = \frac{G_{x_1}}{G_{x_2}},\; MC_B = \frac{B_{x_1}}{B_{x_2}}\,. \tag{4}$$

where MC_R, MC_G, MC_B are Colour Ratios for R, G and B channels respectively. $[R_{x_1}, R_{x_2}]$, $[G_{x_1}, G_{x_2}]$ and $[B_{x_1}, B_{x_2}]$ are neighbouring pixels. Taking the logarithm of Eq. (4) results in:

$$\begin{aligned} log(MC_R) &= log(R_{x_1}) - log(R_{x_2})\,,\\ log(MC_G) &= log(G_{x_1}) - log(G_{x_2})\,,\\ log(MC_B) &= log(B_{x_1}) - log(B_{x_2}). \end{aligned} \tag{5}$$

Expanding R_{x_1} with Eq. (1), gives:

$$R_{x_1} = m(\vec{n}\,\vec{l})\; e^{R_{x_1}}(\lambda)\; \rho^{R_{x_1}}(\lambda) \tag{6}$$

Substituting Eq. (6) in Eq. (5), results in:

$$\begin{aligned} log(MC_R) &= log(R_{x_1}) - log(R_{x_2}),\\ &= log(m(\vec{n_{x_1}}\,\vec{l_{x_1}}) + log(e^{R_{x_1}}(\lambda)) + log(\rho^{R_{x_1}}(\lambda))\\ &\quad - log(m(\vec{n_{x_2}}\,\vec{l_{x_2}}) - log(e^{R_{x_2}}(\lambda)) - log(\rho^{R_{x_2}}(\lambda))\\ &= log(\rho^{R_{x_1}}(\lambda)) - log(\rho^{R_{x_2}}(\lambda)) \end{aligned} \tag{7}$$

where $\vec{l_{x_1}} = \vec{l_{x_2}}$ and $\vec{n_{x_1}} = \vec{n_{x_2}}$ assuming a white light source and a flat surface, respectively. Finally, $e^{R_{x_1}} = e^{R_{x_2}}$ from Eq. (3). The same argument holds for the other channels.

Cross Colour Ratios: The assumption of flat surfaces limits the applicability of the descriptors. Gevers and Smeulders (1999) introduces *Cross Colour Ratios* (CCR) that are independent of both illumination and object geometry:

$$M_1 = \frac{R_{x_1}\, G_{x_2}}{R_{x_2}\, G_{x_1}}\,, M_2 = \frac{R_{x_1}\, B_{x_2}}{R_{x_2}\, B_{x_1}}\,, M_3 = \frac{G_{x_1}\, B_{x_2}}{G_{x_2}\, B_{x_1}}\,. \tag{8}$$

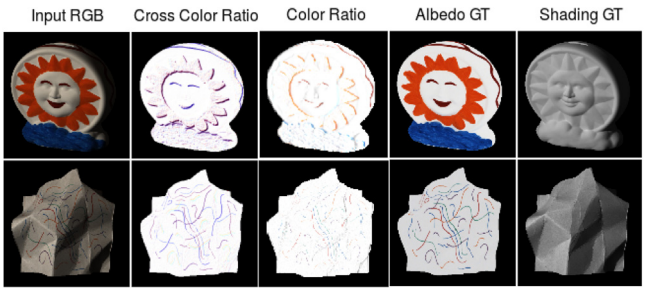

Fig. 2. Colour Ratio (CR) and Cross Colour Ratio (CCR) outputs. Both CR & CCR are calculated from the input RGB image. Note that the influence of geometry is discarded by the CCR computation and that edges resemble reflectance ground-truths. Conversely, the CR image preserves shading information. The mouth and nose are visible for the Sun (CR) image, while fold lines are visible for the paper (CR) image. (MC_R, MC_G, MC_B) and (M_1, M_2, M_3) are combined (3 channels) for CR and CCR visualisations, respectively. Images are gamma corrected.

where M_1, M_2, M_3 are CCR descriptors for (R, G), (R, B) & (G, B) channel pairs respectively. Taking the logarithm on both sides results in:

$$\begin{aligned} log(M_1) &= log(R_{x_1}\, G_{x_2}) - log(R_{x_2}\, G_{x_1})\,,\\ log(M_2) &= log(R_{x_1}\, B_{x_2}) - log(R_{x_2}\, B_{x_1})\,,\\ log(M_3) &= log(G_{x_1}\, B_{x_2}) - log(G_{x_2}\, B_{x_1})\,. \end{aligned} \tag{9}$$

Expanding Eq. (9) by Eq. (6) we get:

$$\begin{aligned} log(M_1) &= log(R_{x_1}\, G_{x_2}) - log(R_{x_2}\, G_{x_1})\\ log(M_1) &= log(R_{x_1}) + log(G_{x_2})\\ &\quad - log(R_{x_2}) - log(G_{x_1})\\ log(M_1) &= log(m(\vec{n_{x_1}}\,\vec{l_{x_1}}) + log(e^{R_{x_1}}(\lambda)) + log(\rho^{R_{x_1}}(\lambda))\\ &\quad + log(m(\vec{n_{x_2}}\,\vec{l_{x_2}}) + log(e^{G_{x_2}}(\lambda)) + log(\rho^{G_{x_2}}(\lambda))\\ &\quad - log(m(\vec{n_{x_2}}\,\vec{l_{x_2}}) - log(e^{R_{x_2}}(\lambda)) - log(\rho^{R_{x_2}}(\lambda))\\ &\quad - log(m(\vec{n_{x_1}}\,\vec{l_{x_1}}) - log(e^{G_{x_1}}(\lambda)) - log(\rho^{G_{x_1}}(\lambda))\\ log(M_1) &= log(\rho^{R_{x_1}}(\lambda)) + log(\rho^{G_{x_2}}(\lambda))\\ &\quad - log(\rho^{R_{x_2}}(\lambda)) - log(\rho^{G_{x_1}}(\lambda)) \end{aligned} \tag{10}$$

Hence, for curved surfaces (i.e., $\vec{n_{x_1}} \neq \vec{n_{x_2}}$), the geometric and the illumination term are cancelled out. The resulting descriptor corresponds to a reflectance (albedo) indicator. The terms $M2$ and $M3$ can be similarly derived and are not shown for brevity.

For constant reflectance, the value of CCR is 1. The value changes for albedo changes i.e., reflectance transitions. Fig. 2 shows CR & CCR examples taken from the MIT Dataset (Grosse et al., 2009). From the figure, it is shown that the CCR ignores shadow and shading effects. However, the CCR computation becomes erroneous when the imaging process does not follow the reflection model including noise, sensor artefacts and non-Lambertian cues.

In conclusion, CCR corresponds to (albedo) reflectance changes only and CR encodes both shading and reflectance transitions. In the following section, the integration of the invariant representations into an end-to-end trainable network is discussed.

3.3. Network architectures

The network takes a RGB image (I) as input. Using Eqs. (7) and (10), the CR invariants (CR_i) and CCR invariants (CCR_i) are obtained. I, CR_i and CCR_i are passed through separate encoders to obtain the encoded features for image ($\mathcal{F}_{img}$), CCR ($\mathcal{F}_{cr}$) and CR ($\mathcal{F}_{ccr}$) respectively. This allows the network to learn a rich feature representation for each of the individual descriptors. For example, $\mathcal{F}_{ccr}$ corresponds to reflectance features, like illumination invariance and reflectance change

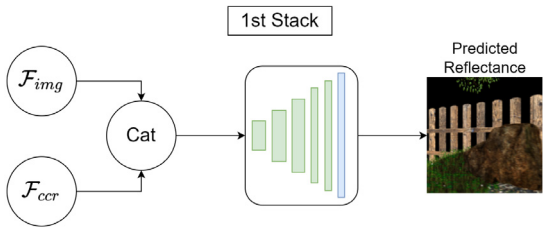

Fig. 3. The features from the image and CCR encoders are concatenated depth wise and passed through a decoder. The output is the predicted reflectance. This makes up the first stack.

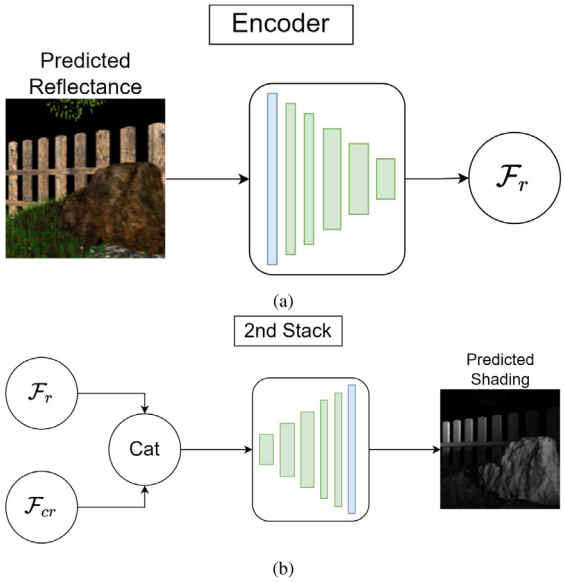

Fig. 4. (a) The predicted reflectance is passed through an encoder to obtain the reflectance features. (b) The encoded reflectance features along with the CR encoded features are concatenated depth-wise and passed to the final decoder. The output of this decoder is the predicted shading. This is the second stack of the network.

boundaries, while $\mathcal{F}_{cr}$ corresponds to both illumination and reflectance transitions.

The encoded features $\mathcal{F}_{img}$ and $\mathcal{F}_{ccr}$ are concatenated and passed on to the reflectance decoder. Skip connections from both the two respective encoders are used in the reflectance decoder to preserve corresponding scale space feature consistency. This allows the network to exploit $\mathcal{F}_{ccr}$, which are illumination and geometry invariant features, to disentangle the reflectance from the image. This makes up the first stack illustrated in Fig. 3.

According to Eq. (2), reflectance is closely linked to shading. Reflectance features are exploited to over-constrain the shading decomposition. The predicted reflectance is passed through an encoder to obtain encoded reflectance features ($\mathcal{F}_r$). To further guide the shading decomposition, $\mathcal{F}_{cr}$ is concatenated to $\mathcal{F}_r$ before passing them on to the decoder. The $\mathcal{F}_{cr}$ encodes reflectance and shading transition information, while $\mathcal{F}_r$ provides only reflectance transition information. Since these transitions are exclusive, the network is able to disentangle the shading transitions. Thus, the shading decoder uses $\mathcal{F}_{cr}$ and $\mathcal{F}_r$ to learn the physics based shading transition, while $\mathcal{F}_{img}$ provides the perceptual shading from the image directly. Skip connections from $\mathcal{F}_{img}$, $\mathcal{F}_{cr}$ and $\mathcal{F}_r$ are added to the decoder to provide corresponding scale space encoder features. This makes up the second stack, as illustrated in Fig. 4.

To further enforce inter-component dependency, the reflectance stack is pre-trained first, on the same dataset, keeping the shading stack disabled. Once the reflectance stack gains sufficient convergence (around 50 epochs), the shading stack is enabled. This allows the shading stack to focus on useful reflectance cues, along with the CR invariant. Without the delayed start on the shading stack, the network will learn reflectance and shading simultaneously. Since the reflectance cues will not be sufficiently converged for the network at the beginning, the reflectance prior for shading is not guaranteed to be correct.

An overview of the network is given in Fig. 1. Scale Invariant MSE losses (Shi et al., 2017) are used for reflectance, shading and reconstructed images to train the network. To further integrate the physical constraints, the CCR loss is used to train the reflectance. The reflectance prediction of the network is used to compute the CCR invariants, using Eq. (10). This is then compared with the CCR invariants obtained from the input RGB image. Since CCR is an illumination and geometry invariant descriptors, the CCR invariants obtained from the reflectance should match the CCR calculated from the input RGB, enforcing an explicit physics constraint.

Finally, the mutual exclusion of reflectance and shading edges is further exploited in the form of an edge divergence regularisation. Reflectance and shading edges are obtained from the network predictions using a Canny edge operator. The edges are thresholded to obtain a binary mask. These binary edges are used as a regularisation to the network, where the overlap between the reflectance and shading edges are minimised. This allows the network to learn a physical model which prevents shading leakages in the reflectance prediction and vice-versa. The edge regularisation is defined by:

$$\mathcal{E}_R = \frac{1}{N}\sum \|(r_c + s_c) - 1\| \tag{11}$$

where, R_c and S_c are the reflectance and shading edges, respectively and N is the number of edge pixels. $\mathcal{E}_R$ is the edge divergence regularisation, which penalises when reflectance and shading edges overlaps.

The network is optimised with Adam optimiser with a learning rate of $2e{-}4$. The reflectance stack is trained separately for $250k$ iterations. Then, both the reflectance and shading stacks are trained for a total of $750k$ iterations. More details about the network architecture and losses can be found in the supplementary material.

3.4. Datasets

The proposed method is trained and assessed on the Natural Environment Dataset (NED) (Baslamisli et al., 2018a). The dataset consists of synthetic garden scenes. All scenes are rendered using the physics-based Blender's Cycles Renderer. Ambient light is simulated using real HDR sky images. The dataset consists of around $32K$ images of which $25K$ images are used for training. The dataset contains around 40 different parks/gardens under 5 lighting conditions. Our network is also fine-tuned and evaluated on the MIT (Grosse et al., 2009) and Sintel (Butler et al., 2012) datasets. To test real-world performance, visual results on the NYU Dataset (Silberman et al., 2012) and the Trimbot dataset (Sattler et al., 2017) are provided. An ablation study on the influence of the different parts of the proposed method is provided in the supplementary material.

4. Experiments and results

4.1. Influence of physics-based cues for IID

In this experiment, the influence of CR and CCR for IID is assessed for two off-the-shelf networks. ShapeNet (Shi et al., 2017) is used to test an IID network and VGG19 (Simonyan and Zisserman, 2015) is taken to test a higher learning capacity network. VGG19 is configured as an encoder–decoder architecture for image-to-image translation. ShapeNet is based on a non-Lambertian assumption. Therefore, the network is modified to output only two IID components for the Lambertian assumption. Extra encoder paths are created for CR and CCR inputs for both architectures. Skip connections are added. Four experiments are conducted per network, (1) without any modification, (2) with CR as the only prior for both components, (3) with CCR as the only prior

Table 1
Influence of physics-based cues for two off-the-shelf architectures. Adding CCR to the reflectance stack and CR to the shading stack yields better results. Using CCR and CR as loss, however degrades the performance.

	Conditioning	MSE		LMSE		DSSIM	
		Reflectance	Shading	Reflectance	Shading	Reflectance	Shading
ShapeNet	–	0.0053	0.0050	0.0597	**0.0910**	0.2516	0.2186
	CR	0.0049	0.0051	0.0800	0.1161	0.1520	0.1792
	CCR	0.0048	0.0049	0.0792	0.1100	0.1510	0.1703
	CCR & CR	**0.0045**	**0.0046**	**0.0748**	0.1065	**0.1438**	**0.1689**
VGG	–	0.0119	0.0094	0.1463	0.1693	0.1987	0.2108
	CR	0.0034	0.0032	0.0579	0.0827	0.1372	0.1640
	CCR	0.0027	**0.0027**	0.0431	**0.0616**	0.1010	0.1113
	CCR & CR	**0.0024**	**0.0027**	**0.0404**	0.0625	**0.0953**	**0.1105**
	CR Loss	0.0082	0.0056	0.0842	0.1219	0.3200	0.3419
	CCR Loss	0.0028	0.0030	0.0491	0.0691	0.1839	0.2456
	Proposed	**0.0019**	**0.0021**	**0.0343**	**0.0509**	**0.0805**	**0.0873**

for both the components, and (4) with CCR as the prior for reflectance and CR for shading. Both networks are trained on the NED dataset using the same data split. Stacks are not included for the ShapeNet and VGG networks.

The proposed network, with the stacked learning, uses the VGG backbone. The quantitative results are presented in Table 1.

It is shown that adding physics-based invariant priors to existing architectures improves the performance, regardless of the underlying architecture. The additional modifications of the priors are simple copies of the encoder blocks. However, using CR and CCR only as a loss instead of a prior, degrades the performance, even when the proposed stacked learning is used. Subsequently, enabling the stack with explicit physics constraints in the form of an invariant descriptor consistency loss and edge regularisation improves the performance. This shows not only the flexibility of the use of physics-based cues, but also the physics-based learning capability of the network. Additional details can be found in the supplementary material.

4.2. Comparison to state of the art

In this experiment, performance of the proposed method to different state-of-the-art methods is studied. The proposed method is trained on the NED (synthetic garden scenes) (Baslamisli et al., 2018a) dataset. Each epoch takes about 20 min for the first stack and 30 min for the full pipeline with the NED dataset, on a Nvidia RTX A6000 GPU.

MIT Intrinsics(real objects) (Grosse et al., 2009) and Sintel(synthetic animated scenes) (Butler et al., 2012) datasets are used to test the generalisation of the proposed method. For all methods, the train and test splits provided by the authors are used. All experimental settings are kept the same according to the respective publications.

4.2.1. Comparison on the NED dataset

Results are shown in Table 2. Qualitative results are shown in Fig. 5.

Despite being trained on a single type of loss, the proposed method outperforms all baselines. Fig. 5 shows a better disentanglement of IIDs by the proposed method, while preserving fine details (e.g. the bush in the third row). The baselines miss fine details and generate blurry outputs. Additionally, shadow leakages are shown to be minimised (row 5, shadow at the base of the tree), which are misclassified as reflectance cues by the baselines.

4.2.2. Comparison on MIT and sintel

Numerical results are presented in Tables 3 and 4. MIT visuals are provided in Fig. 6. The results for Grosse et al. (2009), Barron and Malik (2015) and Gehler et al. (2011) are computed by the code provided by the authors. The results for Narihira et al. (2015), Shi et al. (2017), and Li and Snavely (2018) are obtained from Li and Snavely (2018).

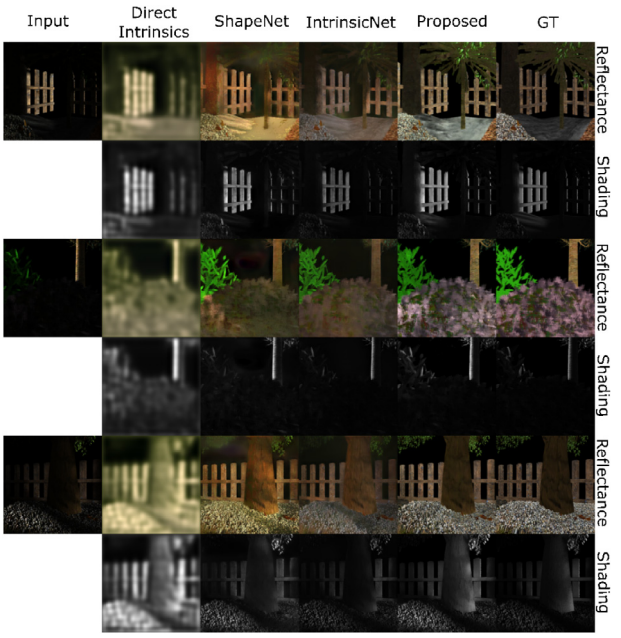

Fig. 5. Comparison between outputs of the proposed method and baselines. It can be observed that the proposed method is able to better separate shading from reflectance (e.g. shading leakages for the baselines).

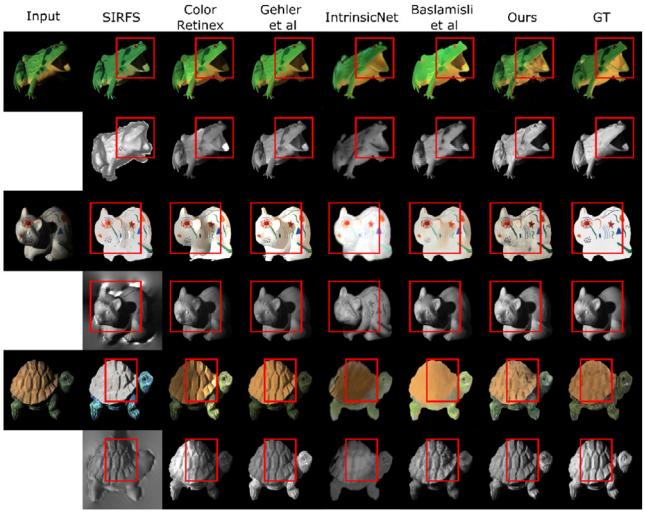

Fig. 6. Qualitative results on MIT dataset. The proposed method is better in separating reflectance and shading cues for objects containing complex lighting effects.

Baslamisli et al. (2018b, 2020) are provided by the author. The missing values are due to the missing results in the original work.

The proposed method with finetuning is shown to outperform other baselines for almost all metrics. Fig. 6 shows better recovery of reflectance and shading patterns by the proposed method. The baselines may incorrectly classify shadows by reflectance cues. In the second row, the shadow on the frog's mouth is missed by the baselines, while the proposed method can disentangle the IID components correctly. The fifth row demonstrates a challenging case for the turtle. The proposed method preserves texture while removing shadows. The baselines methods employ specialised loss functions and specific datasets. Our method is only trained on a synthetic garden dataset and a simple loss, showing that the (generalised) invariant representation and component separation is beneficial in modelling the image formation process.

Table 2
Comparison of the proposed method with state-of-the-art baselines for the NED dataset. The proposed method shows an improvement across all evaluation metrics.

	MSE		LMSE		DSSIM	
	Reflectance	Shading	Reflectance	Shading	Reflectance	Shading
Grosse et al. (2009)	0.0114	0.0193	0.1204	0.2334	0.3280	0.3515
Bell et al. (2014)	0.0095	0.0111	0.1343	0.1861	0.2098	0.3511
Narihira et al. (2015)	0.0073	0.0065	0.1205	0.1798	0.3756	0.3843
Shi et al. (2017)	0.0053	0.0050	0.0597	0.0910	0.2516	0.2186
Li and Snavely (2018)	0.0149	0.0175	0.0447	0.0698	0.2229	0.2346
Baslamisli et al. (2018b)	0.0035	0.0037	0.0449	0.0791	0.2367	0.2110
Yu and Smith (2019)	0.0478	0.0505	0.0642	0.2597	0.2751	0.3382
Liu et al. (2020)	0.0081	0.0143	0.0360	0.0608	0.1886	0.2140
Proposed	**0.0019**	**0.0021**	**0.0343**	**0.0509**	**0.0805**	**0.0873**

Table 3
Comparison of performance on the MIT dataset with current baselines.

	MSE		LMSE		DSSIM	
	Reflectance	Shading	Reflectance	Shading	Reflectance	Shading
Gehler et al. (2011)	0.0065	0.0051	0.0393	0.0282	–	–
Barron and Malik (2015)	0.0129	0.0066	0.0572	0.0309	–	–
Baslamisli et al. (2018b)	0.0104	0.0304	0.0854	0.2038	–	–
Li and Snavely (2018)	0.0167	0.0127	0.0319	0.0211	0.1287	0.1376
Yu and Smith (2019)	0.0234	0.0186	0.0573	0.0765	0.1148	0.1276
Yuan et al. (2019)	0.0109	0.0086	0.0462	0.0537	0.0929	0.0999
Xu et al. (2020)	0.0137	0.0114	0.0614	0.0672	0.1196	0.0825
Liu et al. (2020)	0.0156	0.0102	0.0640	0.0474	0.1158	0.1310
Ma et al. (2020)	0.0091	0.0081	0.0212	**0.0192**	0.0730	0.0659
Baslamisli et al. (2020)	0.0060	0.0069	0.0438	0.0418	–	–
Proposed (MIT Finetuned)	**0.0047**	**0.0045**	**0.0210**	0.0220	**0.0647**	**0.0608**

Table 4
Numerical results for the Sintel dataset. The proposed method is finetuned.

	MSE		LMSE		DSSIM	
	Reflectance	Shading	Reflectance	Shading	Reflectance	Shading
Retinex	0.0606	0.0727	0.0366	0.0419	0.2270	0.2400
Lee et al. (2012)	0.0463	0.0507	0.0224	0.0192	0.1990	0.1770
Barron and Malik (2013)	0.0420	0.0436	0.0298	0.0264	0.2100	0.2060
Chen and Koltun (2013)	0.0307	0.0277	0.0185	0.0190	0.1960	0.1650
Narihira et al. (2015)	0.0100	0.0092	0.0083	0.0085	0.2014	0.1505
Fan et al. (2018)	0.0069	0.0059	**0.0044**	**0.0042**	0.1194	0.0822
Proposed	**0.0010**	**0.0010**	0.0046	0.0047	**0.0450**	**0.0400**

4.3. Real world dataset evaluation

In this experiment, the generalisation capability of the network is tested. The proposed method is trained only on synthetic images. Since there are no dense ground truth annotations for these datasets, only the qualitative evaluations are shown. The datasets are real world gardens (Sattler et al., 2017) and indoor (Silberman et al., 2012) images. The results are shown in Figs. 7 & 8, respectively.

The proposed method can separate reflectance and shading for complex light and object interactions (Fig. 7). The cast shadows around bushes and trees are removed from reflectance images. The bushes, grass and the trees are shown to maintain more uniformity in reflectance predictions.

The proposed method trained and fine-tuned on only the synthetic dataset before it is applied to indoor images (NYU Dataset). Fig. 8 shows that it can recover reflectance and shadings. Predicted reflectance for the wall behind the toilet is flat (third row, fifth column). Similarly, the shadow under the sink is mostly removed (second row, fifth column). This shows that the proposed method has generalisation capabilities.

Some failure cases are shown in Fig. 9. For extreme cases like nearby light sources or specularities, the proposed method falls short to obtain the correct decomposition.

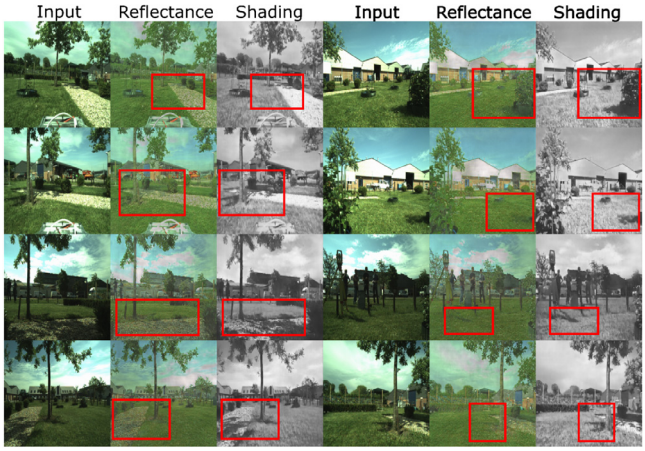

Fig. 7. Qualitative results for Trimbot dataset images. The proposed method generalises well to unseen real world images. It is able to distinguish reflectance and shading cues for complex light and object interactions.

5. Conclusion

In this paper, the use of physics-based descriptors to constrain the problem of intrinsic image decomposition has been investigated.

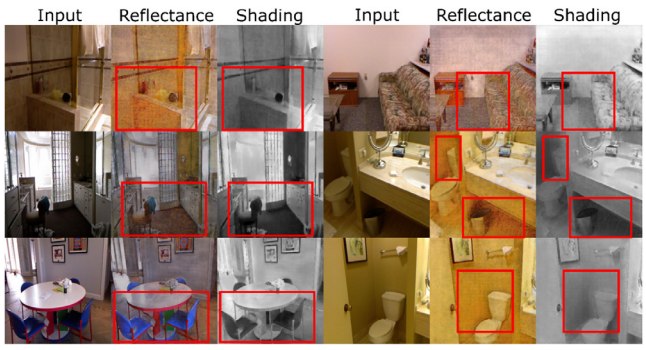

Fig. 8. Qualitative results for the NYU dataset. The proposed method removes shadows from reflectance (first row, fifth column, base of the sofa). The shadows around the toilet are removed, while the wall behind it has a much flatter reflectance (third row, fifth column). This shows the generalisation capability of the proposed method.

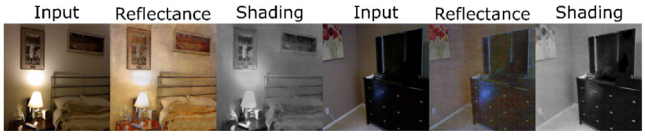

Fig. 9. Some failure cases for the NYU dataset. Scenes with strong local light sources (first column) violates the assumption of locally consistent illumination. Thus, the light is classified as a separate texture. The fourth column contain glossy surfaces violating the Lambertian assumption. This results in erroneous reflectance separation.

Physics-based cues in the form of Colour Ratios and Cross Ratios have been explored. Inter-dependency of reflectance and shading have been exploited through a stacked approach. Component specific physics-based priors and stacked learning have shown to improve IID performance. Qualitative and quantitative improvement on two standard benchmark datasets were discussed. The network has shown to be able to cope well with dataset bias and generalise well by training on synthetic data and testing on real world, unseen images.

For future directions, stricter guidance from the illumination invariant descriptors could be integrated in the form of edges. Further, explicit edge guidance and attention could be used.

CRediT authorship contribution statement

Partha Das: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Visualization. **Sezer Karaoglu:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review and editing, Supervision, Project administration, Funding acquisition. **Theo Gevers:** Conceptualization, Methodology, Formal analysis, Investigation, Writing – review and editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This project was funded by the NWO EDL Project No. P16-25 P2.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2022.103538.

References

Barron, J.T., Malik, J., 2013. Intrinsic scene properties from a single RGB-d image. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 17–24.
Barron, J.T., Malik, J., 2015. Shape, illumination, and reflectance from shading. IEEE Trans. Pattern Anal. Mach. Intell. 1670–1687.
Barrow, H.G., Tenenbaum, J.M., 1978. Recovering intrinsic scene characteristics from images. Comput. Vis. Syst. 3–26.
Baslamisli, A.S., Groenestege, T.T., Das, P., Le, H.A., Karaoglu, S., Gevers, T., 2018a. Joint learning of intrinsic images and semantic segmentation. In: European Conference on Computer Vision. pp. 1–17.
Baslamisli, A.S., Le, H., Gevers, T., 2018b. CNN based learning using reflection and retinex models for intrinsic image decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 6674–6683.
Baslamisli, A.S., Liu, Y., Karaoglu, S., Gevers, T., 2020. Physics-based shading reconstruction for intrinsic image decomposition. In: 2020 Computer Vision and Image Understanding (CVIU). pp. 1–14.
Beigpour, S., van de Weijer, J., 2011. Object recoloring based on intrinsic image estimation. In: IEEE International Conference on Computer Vision. pp. 327–334.
Bell, S., Bala, K., Snavely, N., 2014. Intrinsic images in the wild. ACM Trans. Graph. (TOG).
Bonneel, N., Sunkavalli, K., Tompkin, J., Sun, D., Paris, S., Pfister, H., 2014. Interactive intrinsic video editing. ACM Trans. Graph. 197:1–197:10.
Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J., 2012. A naturalistic open source movie for optical flow evaluation. In: European Conference on Computer Vision. pp. 611–625.
Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F., 2015. ShapeNet: An Information-Rich 3D Model Repository. Technical Report, Stanford University — Princeton University — Toyota Technological Institute at Chicago.
Chen, Q., Koltun, V., 2013. A simple model for intrinsic image decomposition with depth cues. In: IEEE International Conference on Computer Vision. pp. 241–248.
Cheng, L., Zhang, C., Liao, Z., 2018. Intrinsic image transformation via scale space decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 656–665.
Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.P., 2018. Revisiting deep intrinsic image decompositions. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 8944–8952.
Finlayson, G.D., 1992. Colour Object Recognition (Masters thesis). Simon Fraser University.
Funt, B., Drew, M., Brockington, M., 1992. Recovering shading from color images. In: Computer Vision — ECCV'92. ECCV 1992.
Gehler, P.V., Rother, C., Kiefel, M., Zhang, L., Schölkopf, B., 2011. Recovering intrinsic images with a global sparsity prior on reflectance. In: Advances in Neural Information Processing Systems. pp. 765–773.
Gevers, T., Smeulders, A., 1999. Color-based object recognition. Pattern Recognit. 453–464.
Grosse, R.B., Johnson, M.K., Adelson, E.H., Freeman, W.T., 2009. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: IEEE International Conference on Computer Vision. pp. 2335–2342.
Kim, S., Park, K., Sohn, K., Lin, S., 2016. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: European Conference on Computer Vision. pp. 143–159.
Kovacs, B., Bell, S., Snavely, N., Bala, K., 2017. Shading annotations in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 850–859.
Land, E.H., McCann, J.J., 1971. Lightness and retinex theory. J. Opt. Soc. Am. 1–11.
Lee, K.J., Zhao, Q., Tong, X., Gong, M., Izadi, S., Lee, S.U., Tan, P., Lin, S., 2012. Estimation of intrinsic image sequences from image+depth video. In: European Conference on Computer Vision. pp. 327–340.
Li, Z., Snavely, N., 2018. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In: European Conference on Computer Vision. pp. 381–399.
Li, K., Wang, Y., Ye, X., Yan, C., Yang, J., 2021a. Sparse intrinsic decomposition and applications. Signal Process., Image Commun. 95, 116281. http://dx.doi.org/10.1016/j.image.2021.116281, URL https://www.sciencedirect.com/science/article/pii/S0923596521001132.
Li, Z., Yu, T., Sang, S., Wang, S., Song, M., Liu, Y., Yeh, Y.-Y., Zhu, R., Gundavarapu, N.B., Shi, J., Bi, S., Yu, H.-X., Xu, Z., Sunkavalli, K., Haan, M., Ramamoorthi, R., Chandraker, M., 2021b. OpenRooms: An open framework for photorealistic indoor scene datasets. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 7186–7195.
Liu, Y., Li, Y., You, S., Lu, F., 2020. Unsupervised learning for intrinsic image decomposition from a single image. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3245–3254.
Luo, J., Huang, Z., Li, Y., Zhou, X., Zhang, G., Bao, H., 2020. NIID-net: Adapting surface normal knowledge for intrinsic image decomposition in indoor scenes. IEEE Trans. Vis. Comput. Graphics 26 (12), 3434–3445.
Ma, Y., Jiang, X., Xia, Z., Gabbouj, M., Feng, X., 2020. CasQNet: Intrinsic image decomposition based on cascaded quotient network. IEEE Trans. Circuits Syst. Video Technol. 1.

Meka, A., Zollhöfer, M., Richardt, C., Theobalt, C., 2016. Live intrinsic video. ACM Trans. Graph. (SIGGRAPH).

Narihira, T., Maire, M., Yu, S.X., 2015. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: IEEE International Conference on Computer Vision. p. 2992.

Narihira, T., Maire, M., Yu, S.X., 2015. Learning lightness from human judgement on relative reflectance. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 2965–2973.

Sattler, T., Tylecek, R., Brox, T., Pollefeys, M., Fisher, R.B., 2017. 3D reconstruction meets semantics - reconstruction challenge 2017. In: IEEE International Conference on Computer Vision Workshop. pp. 1–7.

Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J., 2019. Neural inverse rendering of an indoor scene from a single image. In: IEEE International Conference on Computer Vision. pp. 1–21.

Shafer, S., 1985. Using color to separate reflection components. In: Color research and applications. pp. 210–218.

Shelhamer, E., Barron, J.T., Darrell, T., 2015. Scene intrinsics and depth from a single image. In: IEEE International Conference on Computer Vision Workshop. pp. 235–242.

Shen, L., Tan, P., Lin, S., 2008. Intrinsic image decomposition with non-local texture cues. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–7.

Sheng, B., Li, P., Jin, Y., Tan, P., Lee, T.-Y., 2020. Intrinsic image decomposition with step and drift shading separation. IEEE Trans. Vis. Comput. Graphics 26, 1332–1346.

Shi, J., Dong, Y., Su, H., Yu, S.X., 2017. Learning non-lambertian object intrinsics across ShapeNet categories. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 5844–5853.

Silberman, N., Hoiem, D., Kohli, P., Fergus, R., 2012. Indoor segmentation and support inference from RGBD images. In: European Conference on Computer Vision. pp. 746–760.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. pp. 1–14.

Wada, T., Ukida, H., Matsuyama, T., 1995. Shape from shading with interreflections under proximal light source-3D shape reconstruction of unfolded book surface from a scanner image. In: IEEE International Conference on Computer Vision. pp. 66–71.

Xu, J., Hou, Y., Ren, D., Liu, L., Zhu, F., Yu, M., Wang, H., Shao, L., 2020. STAR: A structure and texture aware retinex model. IEEE Trans. Image Process. 5022–5037.

Ye, G., Garces, E., Liu, Y., Dai, Q., Gutierrez, D., 2014. Intrinsic video and applications. ACM Trans. Graph. (SIGGRAPH).

Yu, Y., Smith, W.A.P., 2019. InverseRenderNet: Learning single image inverse rendering. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3155–3164.

Yuan, Y., Sheng, B., Li, P., Bi, L., Kim, J., Wu, E., 2019. Deep intrinsic image decomposition using joint parallel learning. In: Computer Graphics International Conference. pp. 336–341.

Zhou, T., Krähenbühl, P., Efros, A.A., 2015. Learning data-driven reflectance priors for intrinsic image decomposition. CoRR.

Zhu, Y., Tang, J., Li, S., Shi, B., 2021. DeRenderNet: Intrinsic image decomposition of urban scenes with shape-(in)dependent shading rendering. CoRR abs:2104.13602. arXiv:2104.13602. URL https://arxiv.org/abs/2104.13602.