Physics-based Shading Reconstruction for Intrinsic Image Decomposition

Baslamisli, A.S.; Liu, Y.; Karaoglu, S.; Gevers, T.

DOI
10.1016/j.cviu.2021.103183

Publication date
2021

Document Version
Final published version

Published in
Computer Vision and Image Understanding

Computer Vision and Image Understanding 205 (2021) 103183

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Physics-based shading reconstruction for intrinsic image decomposition

Anil S. Baslamisli [a,*], Yang Liu [b], Sezer Karaoglu [b], Theo Gevers [a,b]

[a] University of Amsterdam, Science Park 904, 1098XH Amsterdam, The Netherlands
[b] 3DUniversum, Science Park 400, 1098XH Amsterdam, The Netherlands

ARTICLE INFO

Communicated by Nikos Paragios

MSC:
41A05
41A10
65D05
65D17

Keywords:
Intrinsic image decomposition
Shading
Albedo
Invariant image descriptors

ABSTRACT

We investigate the use of photometric invariance and deep learning to compute intrinsic images (albedo and shading). We propose albedo and shading gradient descriptors which are derived from physics-based models. Using the descriptors, albedo transitions are masked out and an initial sparse shading map is calculated directly from the corresponding *RGB* image gradients in a learning-free unsupervised manner. Then, an optimization method is proposed to reconstruct the full dense shading map. Finally, we integrate the generated shading map into a novel deep learning framework to refine it and also to predict corresponding albedo image to achieve intrinsic image decomposition. By doing so, we are the first to directly address the texture and intensity ambiguity problems of the shading estimations. Large scale experiments show that our approach steered by physics-based invariant descriptors achieve superior results on MIT Intrinsics, NIR-RGB Intrinsics, Multi-Illuminant Intrinsic Images, Spectral Intrinsic Images, As Realistic As Possible, and competitive results on Intrinsic Images in the Wild datasets while achieving state-of-the-art shading estimations.

1. Introduction

Intrinsic image decomposition is the inverse problem of recovering the image formation components, such as reflectance and shading (Barrow and Tenenbaum, 1978). The shading component consists of light effects such as direct illumination, geometry, shadow casts and ambient light. The reflectance component represents the (albedo) color of an object and is free of any lighting effect. Intrinsic images are favorable for various computer vision tasks. For example, albedo images are beneficial for semantic segmentation algorithms because of their illumination invariant representation (Baslamisli et al., 2018a). Similarly, most of the scene editing applications, such as recoloring, rely on albedo images (Ye et al., 2014), whereas shading images are preferred for relighting tasks (Shu et al., 2017).

The pioneering work on intrinsic image computation is the Retinex algorithm by Land and McCann (1971) which uses a heuristic that is based on the rectilinear Mondrian world assumption. In a Mondrian world, where surfaces have piece-wise constant colors, strong gradients correspond to albedo changes, while shading variations are related to weaker ones. Then, using a re-integration algorithm (i.e. Poisson) over the strong (albedo) gradients, the albedo component is computed. However, classifying image gradients into albedo or shading is not a trivial task due to various photometric effects such as strong shadow casts, illuminant color, surface geometry changes or weak albedo transitions. For instance, shadow boundaries or abrupt changes in surface geometry may cause strong intensity shifts and may therefore be interpreted as albedo changes. Moreover, the Mondrian world assumption do not apply to real world scenes. Other traditional approaches usually utilize an optimization process by introducing constraints on the intrinsic components (Gehler et al., 2011; Shen et al., 2011; Barron and Malik, 2015). Most of the priors aim at constraining the albedo component such as global reflectance sparsity, piece-wise constant reflectance or chromaticity reflectance correlation. On the other hand, the shading intrinsic is usually constrained by a smoothness prior.

More recent methods rely on deep learning models, specialized loss functions, and large scale datasets. For example, Baslamisli et al. (2018b) provide an end-to-end solution to the Retinex approach in a deep learning framework, Li and Snavely (2018a) combine four datasets with specialized loss functions to impose constraints, and Lettry et al. (2018a) investigate adversarial learning. With the availability of densely annotated synthetic datasets and multiple constraints on the albedo component, CNN-based methods are capable of estimating high quality albedo maps. However, CNN-based shading estimations regularly suffer from texture and intensity ambiguities (e.g. albedo leakage) introducing (color) artifacts in the shading profiles. See Fig. 1 for an illustration.

In the early days of photometric invariance in computer vision, invariant image descriptors were widely used for different vision tasks. These descriptors are invariant to certain image capturing conditions so that the vision algorithms are not affected by them, such as illumination color, surface geometry or camera position. Successful results were

* Corresponding author.
E-mail address: a.s.baslamisli@uva.nl (A.S. Baslamisli).

https://doi.org/10.1016/j.cviu.2021.103183
Received 5 September 2020; Received in revised form 4 February 2021; Accepted 10 February 2021
Available online 13 February 2021

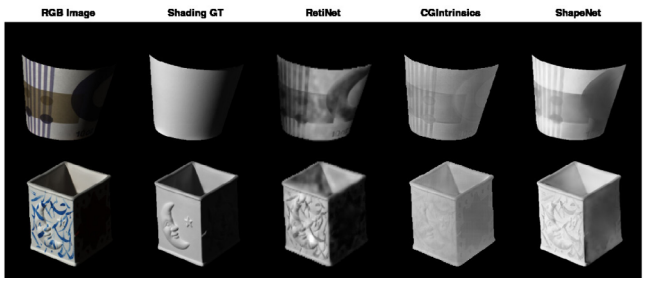

Fig. 1. Color leakage problem in the estimated shading maps. It negatively effects the albedo separation from the shading. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

demonstrated for object recognition (Gevers and Smeulders, 1997), image retrieval (Gevers and Smeulders, 2000), and shadow removal (Finlayson et al., 2006). As CNN-based shading estimations suffer from (color) artifacts, physics-based invariant features may be useful to steer the intrinsic image decomposition process.

Therefore, we investigate the use of photometric invariance and deep learning to compute intrinsic images (albedo and shading). We propose albedo and shading gradient descriptors which are derived from physics-based models. Using the descriptors, albedo transitions are masked out and an initial shading map is calculated directly from the corresponding RGB image gradients in a learning-free manner (unsupervised). Then, an optimization method is proposed to reconstruct the full shading map. Finally, we integrate the shading map into a deep learning model to achieve full intrinsic image decomposition.

Contributions. *1.* We are the first to use photometric invariance and deep learning to address the intrinsic image decomposition task. *2.* We propose albedo and shading gradient descriptors using physics-based models as novel priors. *3.* The shading map is calculated directly from the corresponding RGB image gradients in a learning-free (unsupervised) manner. *4.* We propose a novel deep learning model to leverage the physics-based shading map for the intrinsic image decomposition task. By doing so *5.* we are the first to directly address the color leakage problem in the estimated shading maps. Finally, *6.* we extend the dataset of Baslamisli et al. (2018b) from 15,000 to 50,000 images to train our models, which will be publicly available.

2. Related work

Intrinsic image decomposition is an ill-posed and under-constrained problem. The pioneering work is the Retinex algorithm by Land and McCann (1971) based on the assumption that albedo changes cause large gradients, whereas shading variations result in smaller ones. In general, traditional approaches use different optimization processes to constrain the intrinsic components together with the Retinex heuristic. For example, Gehler et al. (2011) impose constraints on the global albedo sparsity. SIRFS estimates shape, chromatic illumination, albedo, and shading from a single image by applying seven different constraints on the intrinsic components (Barron and Malik, 2015). Intrinsic Images in the Wild (IIW) model combines commonly used priors together with a dense conditional random field (Bell et al., 2014). Shen et al. (2011) use optimization to constraint neighboring pixels having similar intensity values to have similar albedo values. Shen et al. (2008) exploit non-local texture cues by constraining distinct points with the same intensity-normalized textures to have the same albedo values. Furthermore, user interactions are investigated as additional priors to specify albedo values (Bousseau et al., 2009; Shen et al., 2013). Finally, image sequences of the same scene under varying illumination are used to impose constant albedo (Weiss, 2001; Matsushita et al., 2004). Most of the priors mentioned above are related to the albedo intrinsic. It is partially due to color information being more descriptive for robust computer vision algorithms (van de Sande et al., 2009). It is also relatively harder to define priors for the shading intrinsic, because geometry and lighting information are entangled in the representation.

With the introduction of large-scale synthetic datasets, recent research use convolutional neural networks (Shi et al., 2017; Baslamisli et al., 2018a; Li and Snavely, 2018a). Narihira et al. (2015) are the first to use CNNs to learn the task end-to-end in a data-driven manner. Shi et al. (2017) make use of a very large scale dataset along with a specialized network to exploit correlations between the intrinsic components. Baslamisli et al. (2018b) convert the Retinex approach into a deep learning framework together with a physics-based image formation loss. Cheng et al. (2018) use a Laplacian pyramid inspired neural network architecture to exploit scale space properties. Lettry et al. (2018a) explore adversarial residual networks. Fan et al. (2018) apply a domain filter guided by a learned edge map to flatten the albedo estimations. Li and Snavely (2018a) combine four datasets with specialized loss functions. Janner et al. (2017) explore the problem in a self-supervised setting by estimating albedo, shape, and lighting, where shape and lighting estimations are used to train a differentiable shading function. Baslamisli et al. (2019) further decomposes the shading into different photometric effects. Image sequences of the same scene under varying illumination are also explored by deep learning approaches (Lettry et al., 2018b; Li and Snavely, 2018b). Recent work focusing on inverse rendering tasks also achieve superior albedo estimations (Sengupta et al., 2019; Li et al., 2020). Nonetheless, these methods are limited by indoor settings and require additional surface normal and environmental lighting supervision.

CNN-based methods are capable of estimating high quality albedo maps that are mostly free of photometric effects. However, their shading estimations are often negatively affected by albedo transitions causing texture ambiguities and intensity variations, as illustrated in Fig. 1. To mitigate the problem, for example, Zhou et al. (2019) shift the problem of predicting shading to predicting surface normals and lighting properties. Yet, their work is limited by indoor settings and require additional modalities and supervision, similar to inverse rendering works. Another example is CGIntrinsics which over-smooths the shading estimations, yet that in return causes structure loss in the shading maps (Li and Snavely, 2018a). As CNN-based shading estimations suffer from albedo artifacts, invariant image representations may be favorable to steer the process. They were widely used for various image understanding tasks (Drew et al., 1998; Finlayson et al., 1998, 2006; Gevers and Smeulders, 1997, 1998, 2000). One example is the illumination invariant color ratio features used for robust object recognition (Finlayson, 1992). Stricker (1992) combines ratio histograms with boundary histograms for a more robust framework. Nayar and Bolle (1996) utilize color ratios for pose estimation. Matas et al. (1995) embed ratio information into a graph representation also for efficient object recognition. Barnard and Finlayson (2000) identify probable shadow regions using color ratios. Gevers and Smeulders (2001) exploit ratio gradients for image retrieval. As invariant image representations are independent of the certain imaging conditions, they may be useful to improve CNN-based shading estimations as part of intrinsic image decomposition. To this end, in this paper, we investigate the use of photometric invariance and deep learning to compute intrinsic images (albedo and shading).

3. Methodology

3.1. Image formation model

We use the dichromatic reflection model of Shafer (1985) to describe an RGB image. The model defines a surface (image) I as a combination of diffuse I_d and specular I_s reflections as follows:

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

We assume that the diffuse reflection component dominates the imaging conditions and hence the effect of the specular reflection component

is negligible, i.e. $I \approx I_d$. Then, an image I over the visible spectrum ω is modeled by:

$$I_c = m(\vec{n}, \vec{l}) \int_{\omega} f_c(\lambda)\, e(\lambda)\, \rho(\lambda)\, \mathrm{d}\lambda \,, \tag{2}$$

for three color channels $c \in \{R, G, B\}$, where $\vec{n}$ indicates the surface normal, $\vec{l}$ denotes the incoming light source direction, and m is a function of the geometric dependencies (e.g. Lambertian $\vec{n} \cdot \vec{l}$). Furthermore, λ represents the wavelength, f indicates the camera spectral sensitivity, and e describes the spectral power distribution of the light source. Finally, ρ denotes the reflectance i.e. the albedo. Then, assuming a linear sensor response and narrow band filters ($f_c(\lambda_c)$), the equation can be simplified as follows:

$$I_c = m(\vec{n}, \vec{l})\, e(\lambda_c)\, \rho(\lambda_c) \;=\; m(\vec{n}, \vec{l})\, e_c\, \rho_c \,. \tag{3}$$

This equation models an image by the multiplication of its geometry $m(\vec{n}, \vec{l})^x$, albedo ρ_c^x and light source properties e_c^x at pixel x. Then, these characteristics are used to define intrinsic images as follows:

$$I_c^x = S_c^x \times R_c^x \,, \qquad S_c^x = m(\vec{n}, \vec{l})^x\, e_c^x \,, \qquad R_c^x = \rho_c^x \,, \tag{4}$$

where an image I_c at x can be modeled by the element-wise product of its shading S_c and albedo R_c components. If the light source e is colored, then the color information is embedded in the shading component.

3.2. Albedo gradients

Using Eq. (3), the image formation model for the three color channels $c \in \{R, G, B\}$ becomes:

$$\begin{aligned} R^x &= m(\vec{n}, \vec{l})^x\, e_R^x\, \rho_R^x \,, \\ G^x &= m(\vec{n}, \vec{l})^x\, e_G^x\, \rho_G^x \,, \\ B^x &= m(\vec{n}, \vec{l})^x\, e_B^x\, \rho_B^x \,. \end{aligned} \tag{5}$$

Considering only neighboring pixels x_1 and x_2, locally constant illumination can be assumed: $e_c^{x_1} = e_c^{x_2}$ (Land and McCann, 1971). By taking the difference of the logarithmic transformation of each color channel, the albedo descriptors are defined as follows:

$$m_1 = \Delta \log \frac{R}{G} \,, \quad m_2 = \Delta \log \frac{R}{B} \,, \quad m_3 = \Delta \log \frac{G}{B} \,. \tag{6}$$

We illustrate the invariant properties of these albedo descriptors by plugging Eq. (5) into Eq. (6) for m_1 as follows (same also holds for m_2 and m_3):

$$\begin{aligned} m_1 &= \Delta \log \frac{R}{G} = \log \frac{R^{x_1}}{G^{x_1}} - \log \frac{R^{x_2}}{G^{x_2}} \\ &= (\log R^{x_1} - \log G^{x_1}) - (\log R^{x_2} - \log G^{x_2}) \\ &= ((\log m(\vec{n}, \vec{l}))^{x_1} + \log e_R^{x_1} + \log \rho_R^{x_1}) - ((\log m(\vec{n}, \vec{l}))^{x_1} + \log e_G^{x_1} \\ &\quad + \log \rho_G^{x_1}) \;-\; ((\log m(\vec{n}, \vec{l}))^{x_2} + \log e_R^{x_2} + \log \rho_R^{x_2}) - ((\log m(\vec{n}, \vec{l}))^{x_2} \\ &\quad + \log e_G^{x_2} + \log \rho_G^{x_2}) \\ &= \log \frac{\rho_R^{x_1}}{\rho_G^{x_1}} - \log \frac{\rho_R^{x_2}}{\rho_G^{x_2}} = \Delta \log \frac{\rho_R}{\rho_G} \,, \end{aligned} \tag{7}$$

where the remaining factor is only the albedo difference between two channels. The albedo change is a measure that is invariant to surface geometry $\vec{n}$, illumination direction $\vec{l}$, and its intensity and color e. If there is no albedo change (homogeneously colored patch), then the difference is zero. Sensor artifacts or noise may slightly deviate the value from zero. Therefore, the index can be used to identify regions with constant albedo. On the other hand, when the difference deviates significantly from zero, it corresponds to a true albedo change. Hence, this measure encodes spatial information of an image emphasizing on (illumination invariant) albedo edges. Then, we propose the *albedo gradient index* as follows:

$$AGI = \sqrt{(\Delta \log \frac{R}{G})^2 + (\Delta \log \frac{R}{B})^2 + (\Delta \log \frac{G}{B})^2} \,. \tag{8}$$

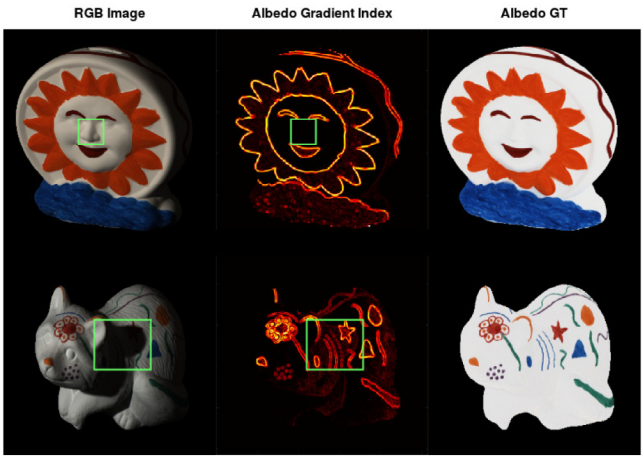

Fig. 2. Finding albedo changes (edges) by the use of the albedo gradient index. Brighter values indicate a higher degree of albedo change. Uniformly colored patches have low scores. Note the similarity of the albedo gradient index and the albedo ground-truth image. The *sun* object shows invariance to geometry and strong shading, and the *raccoon* object demonstrates invariance to shadows. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We calculate the albedo gradients over a local neighborhood (patch) by using derivative filters (e.g. the derivative of a 2D Gaussian or Laplacian) to identify the changes. As a result, the average response of the albedo gradients is calculated. A neighborhood with a higher albedo gradient index value indicates a stronger albedo change, which is also illustrated in Fig. 2. A patch with a constant index yields the homogeneous regions. The albedo gradient index is very intuitive and realized in real time. It is computed for a small threshold to remove possible problems caused by sensor artifacts and noise. The threshold should be set to a small value, because unnecessarily high values may negatively affect the performance by discarding some of the color changes.

3.3. Shading gradients

So far, we have described that the albedo gradient index can be used to identify uniformly colored (homogeneous) patches. In a color image, if the pixel values share the same albedo, then the only source causing those pixel values to change is the shading component. For constant ρ (satisfying $AGI \approx 0$) over an image neighborhood, the shading gradient can be computed by taking the difference of the logarithmic transformation of each color channel. We illustrate it on the red channel as follows (same also holds for green and blue channels):

$$\begin{aligned} \Delta \log R &= ((\log m(\vec{n}, \vec{l}))^{x_1} + \log e_R^{x_1} + \log \rho_R^{x_1}) \\ &\quad - ((\log m(\vec{n}, \vec{l}))^{x_2} + \log e_R^{x_2} + \log \rho_R^{x_2}) \\ &= \log m(\vec{n}, \vec{l})^{x_1} - \log m(\vec{n}, \vec{l})^{x_2} = \Delta \log m(\vec{n}, \vec{l}). \end{aligned} \tag{9}$$

Note that it is only applied on the homogeneous patches. Logarithms are usually preferred to avoid numerical instabilities, yet note that also the derivatives of the RGB channels can be taken to yield the following shading gradient index:

$$SGI = \sqrt{(\Delta R)^2 + (\Delta G)^2 + (\Delta B)^2} \,. \tag{10}$$

Similar to the albedo gradient index, the average response is calculated, which results in representing the gradient field of an RGB image. Note that (non-colored) shadows are included in the shading difference component i.e. when $e_R^{x_1} \neq e_R^{x_2}$.

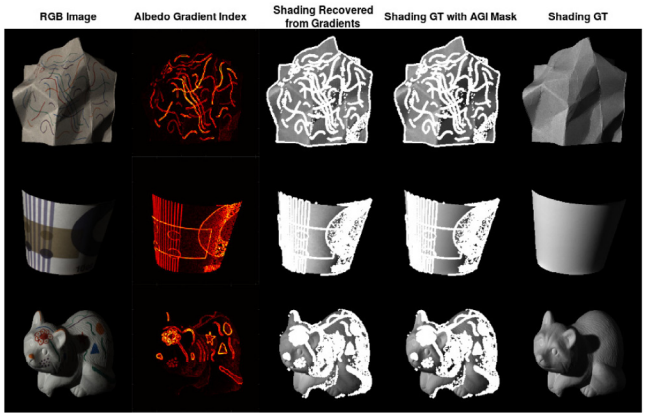

Fig. 3. AGI-assisted physics-based shading gradient index. The albedo gradient index is directly computed from the RGB image. Then, it is used to calculate a shading map by masking out regions that have albedo changes. The same mask is applied to the shading GT to show the resemblance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.4. Shading

After obtaining the shading gradient, we reconstruct the shading map from its shading gradient fields. We use a publicly available algorithm to compute the global least squares reconstruction (Harker and O'Leary, 2008, 2011). Note that the albedo gradient index is used to detect uniformly colored (homogeneous) patches first. Then, the shading gradients are calculated only on the homogeneous patches. As a result, the reconstructed shading map is computed directly from the shading gradient fields of an RGB image in an unsupervised manner. Since it is computed only on the homogeneous image regions (satisfying $AGI \approx 0$), a sparse shading map is obtained. Therefore, the representation is not affected by the albedo changes. The process is illustrated in Fig. 3. In the end, we can generate a sparse shading map that is directly computed from the RGB image that is also very close to the ground-truth representation.

Then, a shading smoothness constraint is used to fill in the gaps based on the neighboring pixel information. To achieve that, we adapt a publicly available optimization framework that is originally designed for the depth completion task (Zhang and Funkhouser, 2018). We modify the model to impose the shading smoothness constraint to achieve a full (dense) shading map. The objective function (E) is defined as the sum of squared errors with two terms $E = E_D + E_S$ as follows:

$$E_D = \sum_{x \in T_{obs}} \|S(x) - S_0(x)\|^2 ,$$
$$E_S = \sum_{p,q \in N} \|S(p) - S(q)\|^2 , \tag{11}$$

where T_{obs} denotes the pixels that are available (not empty) in the initial sparse shading map, which are reconstructed from the RGB gradient fields over the homogeneous regions, and N denotes a neighborhood. E_D measures the distance between the final shading map $S(x)$ and the initial (sparse) shading map $S_0(x)$ at pixel x, i.e. per-pixel reconstruction accuracy. Then, E_S encourages adjacent pixels to have the same shading values, i.e. smoothness.

4. Intrinsic image decomposition

Since the sparse shading map is completed by only a smoothness constraint, the reconstructed dense map may suffer from geometry loss if the initial gaps are too large. It may also suffer from scale problems due to the least squares fitting. Therefore, we integrate the completed dense shading map into a deep learning framework to refine it and also to predict the corresponding albedo image to achieve intrinsic image decomposition. The network is expected to further improve the shading maps by supervised training and also by the differentiation of additional albedo cues. It is also expected to generate better albedo maps as the dense shading map is robust to color leakages and intensity ambiguities. As stated earlier, deep learning based shading estimations are not as good as albedo estimations. They suffer from albedo color leakages mostly due to texture ambiguities and intensity variations (Fig. 1). On the other hand, our physics-based generated shading map is more robust to those leakages as it is computed only on homogeneous regions. As a result, we design a CNN model such that the RGB image only refines the initial shading estimation, and it is not directly involved in the reconstruction phase to avoid any further critical color leakage. The model is illustrated in Fig. 4.

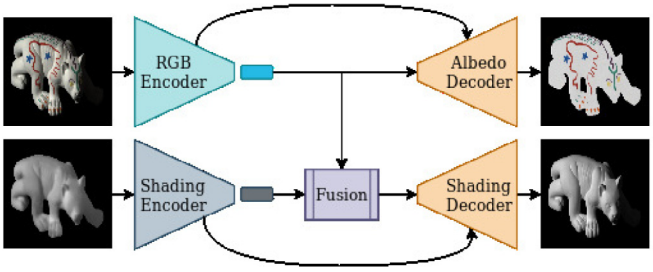

Fig. 4. Proposed model architecture. RGB image guides the shading estimation only during the fusion phase using a 1×1 convolution and a contextual attention module (Yu et al., 2018). Shading decoder only receives shading encoder features through skip connections not to be affected by high resolution RGB color features. Albedo decoder only receives RGB features through skip connections. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Encoders. Encoder blocks use strided convolution layers for downsampling (4 times). Each convolution is followed by residual blocks (He et al., 2016). They are preferred as the deviations from the input are rather small. RGB encoder uses 4 consecutive residual blocks, while the shading encoder uses 1 block with different dilation rates. A residual block is composed of Batch Norm-ReLu-Conv(3x3) sequence, repeated twice. The details are provided in the supplementary material.

Fusion. The final layers of the encoders are fused with a 1×1 convolution and a contextual attention module (Yu et al., 2018) to create a bottleneck such that the related RGB features can properly guide the shading estimation. As a result, the RGB features are fused with the shading features (1) as a (learnable) weighted combination using a $1x1$ convolution, and (2) by the contextual attention module. The contextual attention module learns where to use feature information from known background patches to generate missing patches for the image inpainting task. We adopt their module to our problem such that the shading features use the information from the RGB features. It is expected to help as in a homogeneously colored patch, the only source causing pixel values to change is the shading component, i.e. $\Delta I = \Delta S$. Therefore, in those regions, the shading map and the RGB image are highly correlated. Fusion happens at 16×16 resolution. Preliminary experiments suggested that lower resolutions (i.e. 8×8) cannot reconstruct a decent shading map (too blurry) and higher resolutions (i.e. 32×32) cause further critical color leakages in the shading estimations.

Decoders. The fusion output is fed to the shading decoder, while the albedo decoder takes RGB encoder's final layer as input. Both decoders share the same structure. Encoder features are passed through Conv(3x3)-Batch Norm-LeakyReLu sequence. Then, the feature maps are (bilinearly) up-sampled and concatenated with their encoder counterpart by skip connections. The process is repeated 4 times to reach the final resolution. Shading decoder only receives shading encoder features through skip connections not to be affected by high resolution

color features. Albedo decoder only receives RGB features through skip connections. Therefore, we design a specialized network for the intrinsic image decomposition task for robust shading estimation.

Loss Functions. The loss functions used to train the model are as follows:

$$\mathcal{L}_{Albedo} = \lambda_{A1}\,\mathcal{L}_{pixel} + \lambda_{A2}\,\mathcal{L}_{gradient} + \lambda_{A3}\,\mathcal{L}_{dssim} + \lambda_{A4}\,\mathcal{L}_{perceptual}\,, \quad (12)$$

$$\mathcal{L}_{Shading} = \lambda_{S1}\,\mathcal{L}_{pixel} + \lambda_{S2}\,\mathcal{L}_{gradient} + \lambda_{S3}\,\mathcal{L}_{dssim}\,, \quad (13)$$

$$\mathcal{L}_{Total} = \lambda_{A}\,\mathcal{L}_{Albedo} + \lambda_{S}\,\mathcal{L}_{Shading} + \lambda_{I}\,\mathcal{L}_{Image}\,, \quad (14)$$

where $\mathcal{L}_{pixel}$ is the pixel-wise reconstruction loss, which is a weighted combination of mean-squared-error (MSE) loss and scale-invariant MSE loss, $\mathcal{L}_{gradient}$ denotes the gradient-wise reconstruction loss, $\mathcal{L}_{dssim}$ assesses the structural dissimilarity, $\mathcal{L}_{perceptual}$ measures the reconstruction distance in several feature spaces of a pre-trained VGG16 (Simonyan and Zisserman, 2015), $\mathcal{L}_{Image}$ is the image formation loss to force that the estimated reflectance and shading images should reconstruct the original RGB image (i.e. $I = S \times R$), and the λs are the weights. Note that the loss functions are the standard reconstruction modules and do not impose any intrinsic image characteristics. The implementation details and other training details are provided in the supplementary material.

Dataset. To train our models, we use the ShapeNet dataset of Baslamisli et al. (2018b). The dataset includes around 20,000 (synthetic) images of man-made objects randomly sampled from the original ShapeNet dataset (Chang et al., 2015). Following the setup of Baslamisli et al. (2018b), we render additional images to reach around 50,000 images for training.

5. Experiments and evaluation

We conduct experiments on four datasets of real world objects with ground-truth intrinsics, MIT Intrinsics (Grosse et al., 2009), NIR-RGB Intrinsics (Cheng et al., 2019), Multi-Illuminant Intrinsic Images (Beigpour et al., 2015) and Spectral Intrinsic Images (Chen et al., 2017). In addition, we provide experiments on two scene-level datasets, As Realistic As Possible (Bonneel et al., 2017) a synthetic ground-truth dataset, and Intrinsic Images in the Wild (Bell et al., 2014) a real world complex dataset with relative human annotations. Finally, we provide further qualitative evaluations on real world in-the-wild images. Comparisons are provided against several state-of-the-art intrinsic image decomposition algorithms. We pick three optimization based methods: (i) STAR, a structure and texture aware advanced Retinex model (Xu et al., 2020), (ii) IIW, a framework based on clustering and a dense CRF (Bell et al., 2014), and (iii) SIRFS, a model imposing seven different priors on reflectance, shape and illumination (Barron and Malik, 2015). We include four deep learning based methods: (i) ShapeNet uses specialized decoder links to correlate intrinsics and is trained on 2.5M synthetic objects (Shi et al., 2017), (ii) IntrinsicNet uses deep VGG16 encoder–decoders and an image formation loss, trained on 20K synthetic objects, (iii) RetiNet provides an end-to-end solution to the Color Retinex approach using gradients, trained on 20K synthetic objects, (iv) CGIntrinsics combines two real world scenes (around 3000) and two synthetic scene level datasets (around 20K) for training with additional smoothness constraints to achieve better intrinsics. We use the publicly available models and the original outputs without any fine-tuning or post-processing stages as comparison. To evaluate our proposed method, following the common practice (Grosse et al., 2009), when dense ground-truths are available, we use the mean squared error (MSE), where the absolute brightness of each image is adjusted by least squares as the ground-truth is only defined up to a scale factor and the local mean squared error (LMSE) with window size 20. For Intrinsic Images in the Wild (IIW) dataset's human annotations, we use Weighted Human Disagreement Rate (WHDR) metric as provided by the authors (Bell et al., 2014). All the images are resized to 256 × 256 for fair comparison.

Table 1
Quantitative evaluations on MIT Intrinsic Images dataset. Our proposed model achieves better performance compared against other models on all metrics demonstrating better reconstruction quality. CA module leads to further improvements in performance.

	MSE ↓			LMSE ↓		
	Shading	Albedo	Average	Shading	Albedo	Average
STAR	0.0114	0.0137	0.0126	0.0672	0.0614	0.0643
SIRFS	**0.0066**	0.0129	0.0098	**0.0309**	0.0572	0.0441
IIW	0.0101	0.0210	0.0156	0.0425	0.0720	0.0573
ShapeNet	0.0075	0.0158	0.0117	0.0366	0.0543	0.0455
IntrinsicNet	0.0304	0.0104	0.0204	0.2038	0.0854	0.1446
RetiNet	0.0391	0.0097	0.0244	0.2651	0.0636	0.1644
CGIntrinsics	0.0117	0.0133	0.0125	0.0425	0.0477	0.0451
OURS	0.0069	**0.0060**	**0.0065**	0.0418	**0.0438**	**0.0428**
OURS (w Retinex)	0.0071	0.0060	0.0066	0.0444	0.0438	0.0441
OURS (w/o CA)	0.0075	0.0070	0.0073	0.0454	0.0458	0.0456

5.1. Evaluations on object-level datasets

5.1.1. MIT intrinsic images dataset

The dataset contains 20 real-world objects with ground-truth intrinsic images. Objects are lit by a single directional white light source. We follow the recommendation of the authors and exclude apple, pear, phone and potato objects as they are marked as problematic (Grosse et al., 2009). The quantitative results are provided in Table 1. The table also includes the effect of the contextual attention (CA) module and the quality of our albedo descriptors as ablation studies.

The results show that comparing with the deep learning based estimations, our proposed models achieves better performance at generating albedo and shading maps on the dataset. Optimization based SIRFS results are better than all other learning based models. Its shading estimations yield the best results . It is known that SIRFS achieves superior performance on single and masked objects, yet it generalize poorly to real scenes (Narihira et al., 2015; Li and Snavely, 2018a). Nonetheless, our albedo estimations are superior than SIRFS on all other metrics. On average, we achieve the best results by a substantial margin. Furthermore, the contextual attention module by Yu et al. (2018) leads to further performance boost on all metrics. It emerges as a fundamental building block of our proposed method. Finally, we provide an ablation study to evaluate the quality of our albedo descriptors against the commonly used Color Retinex (Grosse et al., 2009). To this end, we replace our albedo gradients with the gradients of the Color Retinex and keep the rest of the components the same (OURS (w Retinex)), and provide the evaluation. The results further demonstrate that our physics-based albedo gradients achieve better shading reconstructions on both metrics also compared against the heuristic-based Color Retinex gradients.

In addition, we are extremely efficient compared with the optimization-based methods. To process a single image, on average, SIRFS takes 111.38 s, whereas our model takes 1.79 s including the albedo gradient estimation, initial shading recovery from the gradients, filling the initial shading with the smoothness prior, and finally estimating complete intrinsic images. All in all, our model appears 78 times faster than SIRFS. As a side note, IIW model takes 18.09 s, and STAR takes 2.78 s to process a single image on the MIT dataset.[1]

Finally, we provide qualitative evaluations. Fig. 5 demonstrates the effect of the proposed model from the initial step to reach the final shading map with progressive improvement. The results show that our framework first generates an initial shading map where the color transitions are masked out by the physics-based albedo gradient descriptors. Then, the initial shading maps are filled (inpainted/interpolated) with the shading smoothness prior. They are free of color leakages and intensity ambiguities. However, they suffer from scale problems due

[1] The results are provided on Intel Xeon CPU E5-2640 v3 @ 2.60 GHz.

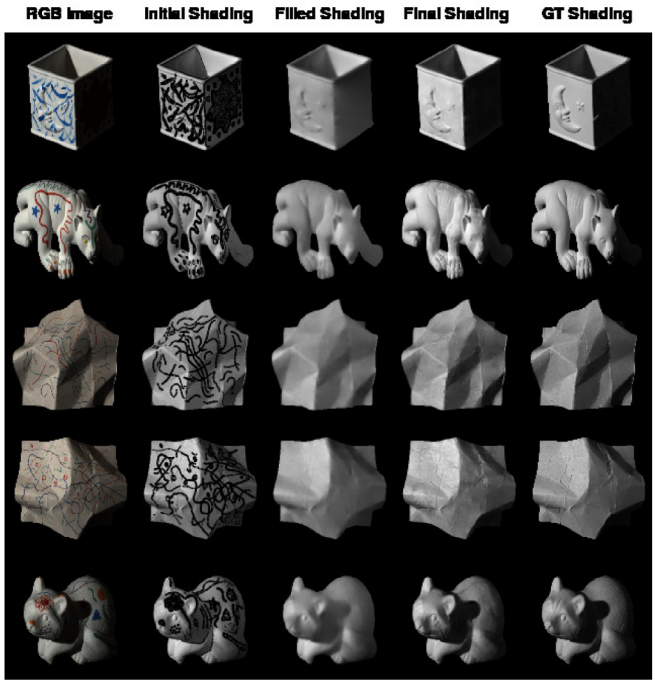

Fig. 5. The effect of the proposed framework. The initial shading maps are free of the color leakage problem. The filled shadings are rather blurry, suffers from scale problems and missing geometric details. The deep model further refines it generating sharper shading maps with proper scale. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to the least squares fitting and they are rather blurry due to the neighborhood smoothness filling. Finally, our deep learning model is able to refine the initially filled shading maps. It makes them sharper, adjusts the scale, and finer geometry details are visible. Fig. 7 provides the qualitative comparison results against the state-of-the-art models. It shows that we achieve better shadow and shading handling in albedo predictions and our albedo estimations are significantly better. We attribute this to our physics-based shading reconstructions as it handles color leakage and intensity ambiguity problems. Thereby, our shading predictions has no or minimum color leakage. Moreover, the shading map estimations by the deep learning methods tend to severely overfit to the *RGB* image producing strong color leakages as texture artifacts and intensity ambiguities.

5.1.2. NIR-RGB intrinsic images dataset

We provide additional cross dataset experiments on NIR-RGB Intrinsic Images dataset, which was mainly generated for near-infrared imagery research (Cheng et al., 2019). It includes seven real-world objects with corresponding ground-truth intrinsics. The quantitative results are provided in Table 2.

The results show that our proposed model achieves better performance compared against other models on all metrics. We especially achieve significantly better albedo estimations. The results further demonstrate the improved generalization ability of our proposed method. In this dataset, deep learning based methods are as good as SIRFS, even more superior in some cases. Finally, Fig. 6 shows qualitative comparisons for a number of images.

The qualitative results further support the quantitative evaluations. Our model predictions are closer to the ground-truth images. The colors of our albedo estimations appear more natural and vivid, and closer to the chromaticity patterns of the input images. Our shading estimations do not include intensity ambiguities or texture artifacts. On the other hand, the intensity ambiguity problem in the shading maps can be observed on ShapeNet and IntrinsicNet estimations on the *candle* and

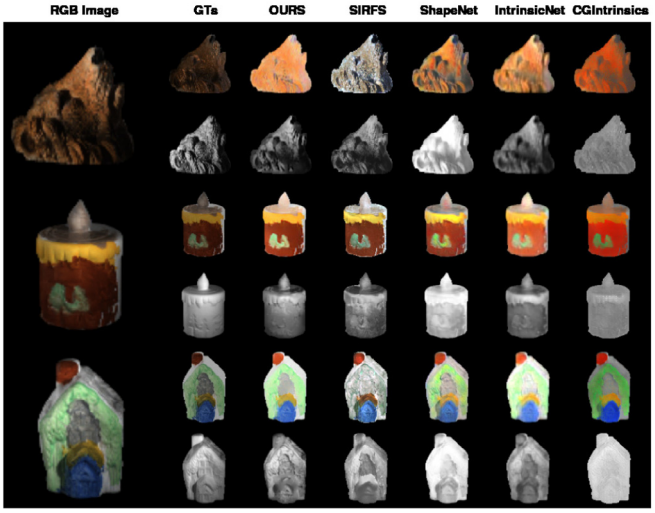

Fig. 6. Qualitative evaluations on NIR-RGB Intrinsic Images dataset. Our albedo maps appear more natural and vivid, and closer to the chromaticity patterns of the input images. Our shading estimations do not include intensity ambiguities or texture artifacts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Quantitative evaluations on NIR-RGB Intrinsic Images dataset. Our proposed model achieves better performance compared against other models on all metrics demonstrating better generalization ability.

	MSE ↓			LMSE ↓		
	Shading	Albedo	Average	Shading	Albedo	Average
STAR	0.0028	0.0017	0.0023	0.0896	0.1131	0.1014
SIRFS	0.0020	0.0009	0.0015	0.0806	0.0950	0.0878
IIW	0.0042	0.0018	0.0030	0.1200	0.1345	0.1273
ShapeNet	0.0019	0.0008	0.0014	0.0701	0.0772	0.0737
IntrinsicNet	0.0021	0.0011	0.0016	0.0748	0.0927	0.0838
RetiNet	0.0028	0.0013	0.0021	0.0959	0.1136	0.1048
CGIntrinsics	0.0027	0.0009	0.0018	0.0862	0.0797	0.0830
OURS	**0.0017**	**0.0006**	**0.0012**	**0.0689**	**0.0609**	**0.0649**

house images. CGIntrinsics's shading smoothness constraint tends to generate over-smoothed estimations and cannot capture fine-grained geometric patterns. For example, the balcony of the *house* object is not visible anymore. SIRFS tends to generate incorrect colors on albedo estimations when a scene is dominated by a single color as in the cases of *lion* and *house* objects. The colors of the CGIntrinsics albedo maps tend to shift towards red.

5.1.3. Multi-Illuminant Intrinsic Images (MIII) dataset

MIT Intrinsic Images and NIR-RGB Intrinsic Images datasets provide images with uniform white illumination. In this experiment, we further test the ability of our proposed method to generalize also to complex multi-illuminant scenarios. The dataset includes five real-world scenes with multi-colored non-uniform lighting, complex geometry, large specularities, and challenging colored shadows (Beigpour et al., 2015). Each scene includes two objects and illuminated with 6 single-illuminant and 9 two-illuminants. The colors of the illuminants vary from orange to blue. In total, there are 75 images with ground-truth intrinsics. The quantitative results are provided in Table 3.

The qualitative results show that our proposed model achieves better performance on almost all metrics. Only the reflectance estimations of CGIntrinsics (Li and Snavely, 2018a) are better on the LMSE metric, but their shading estimations are significantly worse. Thus, compared with other works, on average we achieve the best results by a large margin. Note that optimization based SIRFS (Barron and

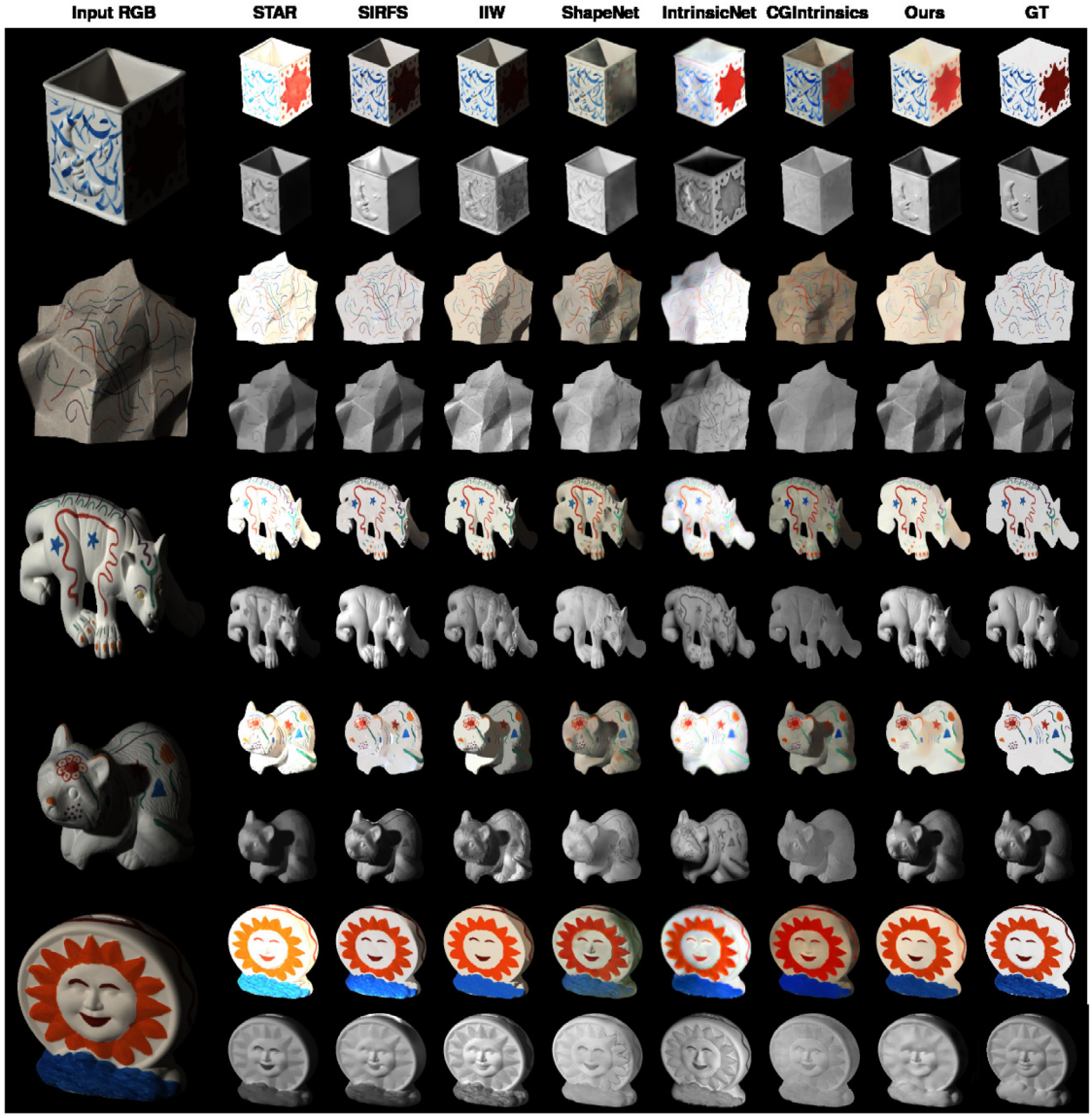

Fig. 7. Comparisons with state-of-the-art models. Our shading predictions are more robust to the color leakage problem, while all other methods tend to overfit to the *RGB* image having severe color leakages in the shading maps. We also achieve significantly better albedo estimations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Malik, 2015) and learning based ShapeNet (Shi et al., 2017) are inherently modeled to estimate multi-colored illumination. Nevertheless, our model emerges more robust to real-world images with multi-colored non-uniform lighting. The results further demonstrate the improved generalization ability of our proposed method.

5.1.4. Spectral Intrinsic Images Dataset (SIID)

The dataset was mainly generated for spectral intrinsic image decomposition research (Chen et al., 2017). It includes nine objects illuminated with two kinds of light sources, one white and one warm-tone white. In total, it has 18 spectral images with corresponding shading ground-truths. The dataset also provides corresponding *RGB* images synthesized from the spectral images that are used as inputs to the models. The quantitative results are provided in Table 4.

The results show that the reconstruction quality of our shading maps are closer to the ground-truths on all metrics. Similar to the MIII dataset experiments with multi-colored non-uniform lighting, our models also achieve more robust results on a different illumination setting of warm-tone white. Finally, Fig. 8 shows qualitative comparisons for a number of images.

Table 3
Quantitative evaluations on MIII dataset with multi-colored non-uniform lighting. Our proposed model achieves better performance and is more robust to multi-colored non-uniform lighting.

	MSE ↓			LMSE ↓		
	Shading	Albedo	Average	Shading	Albedo	Average
STAR	0.0021	0.0023	0.0022	0.0817	0.1350	0.1084
SIRFS	0.0003	0.0003	0.0003	0.1015	0.1417	0.1216
IIW	0.0003	0.0002	0.0003	0.0869	0.1286	0.1078
ShapeNet	0.0002	0.0002	0.0002	0.0846	0.1020	0.0933
IntrinsicNet	0.0002	0.0002	0.0002	0.0597	0.0873	0.0735
RetiNet	0.0002	0.0002	0.0002	0.0590	0.0964	0.0777
CGIntrinsics	0.0004	0.0001	0.0003	0.1172	**0.0707**	0.0940
OURS	**0.0002**	**0.0001**	**0.0002**	**0.0514**	0.0770	**0.0642**

Table 4
Quantitative evaluations on SIID dataset with white and warm-tone white illuminations. Our proposed model achieves better performance and has better generalization ability.

	MSE-s ↓	LMSE-s ↓
STAR	0.0034	0.0192
SIRFS	0.0186	0.0215
IIW	0.0064	0.0164
ShapeNet	0.0129	0.0424
IntrinsicNet	0.0045	0.0189
RetiNet	0.0047	0.0220
CGIntrinsics	0.0142	0.0286
OURS	**0.0027**	**0.0156**

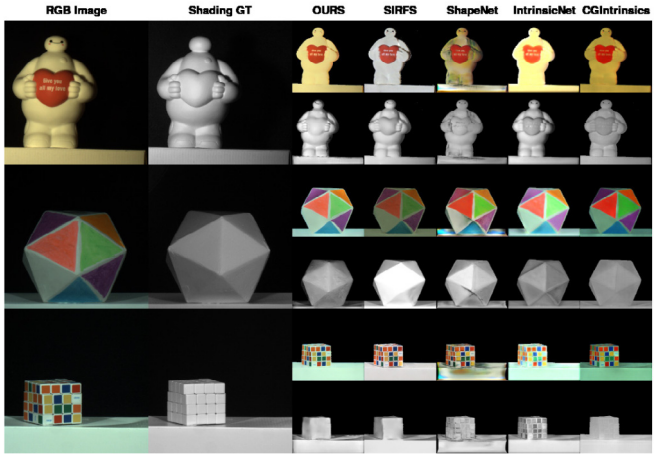

Fig. 8. Qualitative evaluations on SIID. Our albedo maps appear more natural and vivid. Our shading estimations do not include intensity ambiguities or texture artifacts and are closer to the ground-truths. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The qualitative results further support the quantitative evaluations. Our model predictions are closer to the ground-truth images. Our albedo estimations appear more natural and vivid and they are free of geometric effects. Our model is also capable of removing shadow casts on the platforms of the *gypsum* and *cube* objects from the albedo estimations. Since our model is trained only on white light, the color of the light source is also estimated in the albedo. Same behavior is also observed on other models. To overcome this issue, a white balancing algorithm can be applied to the input images as a pre-processing step. Nonetheless, it does not cause significant problems on the reconstruction quality as the ground-truths are not absolute and only defined up to a scale factor (Grosse et al., 2009; Narihira et al., 2015). SIRFS can handle the issue, but it tends to confuse albedo and color of the light source when a scene is dominated by a single color as demonstrated in the previous section. Additional examples can be found in the upcoming sections. Likewise, as mentioned in the previous section, ShapeNet (Shi et al., 2017) is inherently modeled to estimate multi-colored illumination. However, it also fails to differentiate the color of the light source and albedo in this case. It also generates undesired color artifacts on the albedo maps.

As for the shading map generations, our model estimations are free of any texture artifacts and intensity ambiguities. The text on the heart of the *baymax* object is correctly attributed to the albedo map, whereas ShapeNet estimation is contaminated with the texture artifact, and IntrinsicNet and CGIntrinsics estimations both contain texture artifacts and intensity ambiguities. The intensity ambiguity problem is more severe on the shading estimations of the *cube* object. Our model and optimization-based SIRFS can handle those. Nevertheless, our contribution is more significant on the *gypsum* object, where SIRFS tends to generate over-smooth and overly-bright estimations that the geometry is distorted and fine-grained structures are not visible anymore. Our model is also not flawless. For example, we cannot capture the fine geometric details of the *cube* image and our estimation appears more rigid. That is because of the shading smoothness constraint that is used to fill in the gaps of the initial shading map based on the neighboring pixel information. Since the color changes happen near the holes, shading smoothness interpolation also fills in those gaps. Therefore, the shading estimation appears more rigid in those cases.

Fig. 9. Additional real world evaluations on ALOI dataset. Rows (1,2,3) provide examples with textures, and (4,5) with strong shading patterns. Deep learning methods have severe color leakages in the shading maps and cannot handle strong shadings in the albedo maps. Our method is capable of capturing decent albedo and shading maps for also ALOI images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.1.5. Amsterdam Library of Object Images (ALOI) dataset

We provide additional visual comparisons for real world images without ground-truths. For the task, we use Amsterdam Library of Object Images (ALOI) dataset (Geusebroek et al., 2005). Fig. 9 provides a number of examples with different properties to demonstrate the effectiveness of our method. Rows (1,2,3) provide examples with textures and rows (4,5) provides examples with strong shading patterns.

Deep learning based methods have severe color leakages in the shading map estimations for textured objects. CGIntrinsics's shading smoothness constrain negatively effects the shading maps when strong shading patterns are present. It generates homogeneously smooth images such that it cannot properly capture darker regions where the surface normals (geometry) significantly deviate from the incoming light source direction. It can be observed from the *cup* image that the right part of the handle should be covered by the shading pattern and should not be visible. Our proposed work is the only model that can

Table 5
Quantitative evaluations on scene-level ARAP dataset. Our proposed model achieves better performance and generalization ability.

	MSE ↓			LMSE ↓		
	Shading	Albedo	Average	Shading	Albedo	Average
IIW	0.0913	0.0496	0.0705	0.2050	0.0721	0.1386
ShapeNet	0.1218	0.0978	0.1098	0.2400	0.1435	0.1918
IntrinsicNet	0.0889	0.0380	0.0635	0.1867	0.0530	0.1199
RetiNet	0.0874	0.0417	0.0646	0.1875	0.0600	0.1238
OURS	**0.0862**	**0.0337**	**0.0600**	**0.1832**	**0.0482**	**0.1157**

capture that pattern. Similar behavior is also observed for the *wooden cube* in the last row. Likewise, the other models cannot generate a decent albedo map in those cases. ShapeNet generated albedo maps are rather dull colored and blurry. Similarly, CGIntrinsics and IntrinsicNet generated albedo maps tend to be polluted with color artifacts. On the other hand, our model is better at avoiding attributing surface texture to the shading maps, and our albedo estimations are sharper, have better color augmentation and more natural for all cases. SIRFS model is capable of producing decent shading maps for textured objects, as well. However, its albedo predictions are not as decent when an image is dominated by a single color as in the case of 1st and 5th rows. Similarly, it tends to fail to capture decent shading maps when an image has strong shading patterns.

5.2. Evaluations on scene-level datasets

There are several aspects that are challenging for our current setup for the scene level intrinsic image decomposition. Firstly, a scene is composed of multiple objects so that the behavior of the illumination component is more complex. Especially, the ambient light (inter-reflection) effect is way stronger. In addition, our optimization process using the smoothness constraint to fill in the gaps of the initial shading map may be negatively effected if the gaps are filled from different surfaces (e.g. filled with object boundaries). Similarly, cluttered objects may cause way too large gaps to fill. Another thing is that since scene level objects have different scales, one single threshold might not be sufficient to obtain proper gradients. Nonetheless, for the sake of completeness, we also evaluate our model on scene-level images to provide additional insights.

5.2.1. As Realistic As Possible (ARAP) dataset

With the current technology, it is not possible to generate dense ground-truth intrinsic images for any real world scene. Collecting the ground-truth intrinsics happens only on object-level and in a fully-controlled (indoor) laboratory settings, which demands extreme care (Grosse et al., 2009; Chen et al., 2017; Cheng et al., 2019). That is the reason why those datasets are small sampled. Therefore, to evaluate our model on scene-level images, we utilize the synthetic dataset of Bonneel et al. (2017). The dataset provides 53 high quality realistic scene-level renderings with corresponding per-pixel ground-truth intrinsics. Some of the scenes were re-rendered with different illumination settings. Thus, the evaluation is provided for the full dataset of 152 images. The quantitative results are provided in Table 5. The evaluations do not include CGIntrinsics model as it uses ARAP for training (Li and Snavely, 2018a), and also SIRFS model as it is specifically designed for single objects and generalize poorly to real scenes (Narihira et al., 2015; Li and Snavely, 2018a). Compared with other frameworks our proposed model achieves better performance on all metrics also on scene-level images, which further demonstrates our improved generalization ability.

Table 6
Quantitative evaluations on IIW dataset with human annotations. Our proposed model achieve significantly better reflectance predictions among the models trained on object-level ShapeNet dataset.

	Training set	WHDR ↓
STAR	–	32.9%
IIW	–	20.6%
DirectIntrinsics	Sintel	37.3%
CGIntrinsics	SUNCG	26.1%
CGIntrinsics	CGI	18.4%
ShapeNet	ShapeNet (2.5 M)	59.4%
IntrinsicNet	ShapeNet (20 K)	32.1%
RetiNet	ShapeNet (20 K)	37.9%
OURS	ShapeNet (20 K)	28.9%
OURS	ShapeNet (50 K)	28.7%
OURS*	ShapeNet (50 K)	26.8%

*Indicates that the CNN predictions are post-processed with a guided filter.

Fig. 10. Albedo gradient index of scene level images. Brighter values indicate a higher degree of albedo changes. Uniformly colored patches have low scores that can be differentiated from intrinsic color variations. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

5.2.2. Intrinsic Images in the Wild (IIW) dataset

Firstly, we show the albedo gradient index maps for scene-level real-world images in the IIW dataset in Fig. 10. The first row shows that the albedo descriptor does not respond to homogeneously colored regions of the white bedspread, the large blue pillows and the walls. It further ignores the wrinkles of the bedspread and curtains, and diverse color changes are captured. Similarly, the second row demonstrates that the descriptor properly identifies color changes such that the homogeneously colored carpet, pillar, walls and the ceiling is clearly identified by the low response.

For the evaluations, we follow the common practice and utilize the test set used by previous work (Zhou et al., 2015; Li and Snavely, 2018a). The test split includes 1046 images with relative human annotations. The quantitative results are provided in Table 6. We also train our model with less data (20K) to provide a more fair comparison against the models of Baslamisli et al. (2018b).

Comparing with the models trained on object-level ShapeNet dataset, our proposed model achieve significantly better reflectance

predictions. Additional performance boost is achieved by applying a post processing step to enforce piecewise constant reflectance (Nestmeyer and Gehler, 2017). Decreasing the training sample size does not significantly effect the performance for our model's albedo estimations on IIW. Furthermore, our proposed model is significantly better than the structure and texture aware advanced Retinex model, and also DirectIntrinsics model trained on scene-level Sintel dataset. We also achieve on par results with CGIntrinsics model when trained on scene-level SUNCG dataset. The model achieves superior performance by combining the refined and improved renderings of scene-level SUNCG and the integration of ARAP dataset to create their final dataset CGI. It is also worthwhile to note that all the learning based models use data augmentations through random flips, shifts, resizings, and crops, whereas we do not apply any augmentation technique. Finally, Fig. 11 provides qualitative comparisons for shading estimations, and Fig. 12 for albedo estimations.

ShapeNet estimations are contaminated with artifacts and do not appear natural. The shading of the *bed* image includes texture artifacts and the text *AWAI* is directly copied to the shading map in the *girl* image. Similar patterns are also observed in IntrinsicNet estimations. IntrinsicNet generated shading maps also suffer from intensity ambiguities, which can be observed from the *girl* image that the neck of the t-shirt has a darker color. Its albedo estimations are better than ShapeNet's, yet they contain inconvenient brightness artifacts. IIW's albedo estimations appear natural and free of geometry effects. However, its shading generations directly overfit to the *RGB* inputs, and all the texture patterns are clearly visible in the shading maps.

CGIntrinsics trained on scene-level imagery achieves decent albedo predictions with proper smoothing effects, and compared with others, they appear more natural. However, their shading estimations appear way too smooth and hazy and most of the structures are not visible anymore (e.g the stairs or the fine-grained pillars of the church). It also suffers from the same intensity ambiguity problem as IntrinsicNet. On the other hand, our model is also capable of producing scene-level shading maps that are free of texture or intensity ambiguities. The first image shows that our model also works on outdoor scenes capable of handling geometry differences and different light properties. We can also handle the text on the t-shirt of the *girl* image and the text on the *salt box* and correctly attribute them to albedo maps. The windows of the *bed* image are an example where our shading map is negatively effected as our model tries fill in the gaps with insufficient gradient information. Although we did not enforce it as CGIntrinsics, our albedo estimations also appear smooth. However, our method still makes mistakes, such as the face of the *girl* or right side of the *church* appear blurry. Finally, our model is the only one that can handle the strong shadow cast under the bed. Our albedo estimations are free of strong shadow casts in this example, whereas all other models fail to handle it.

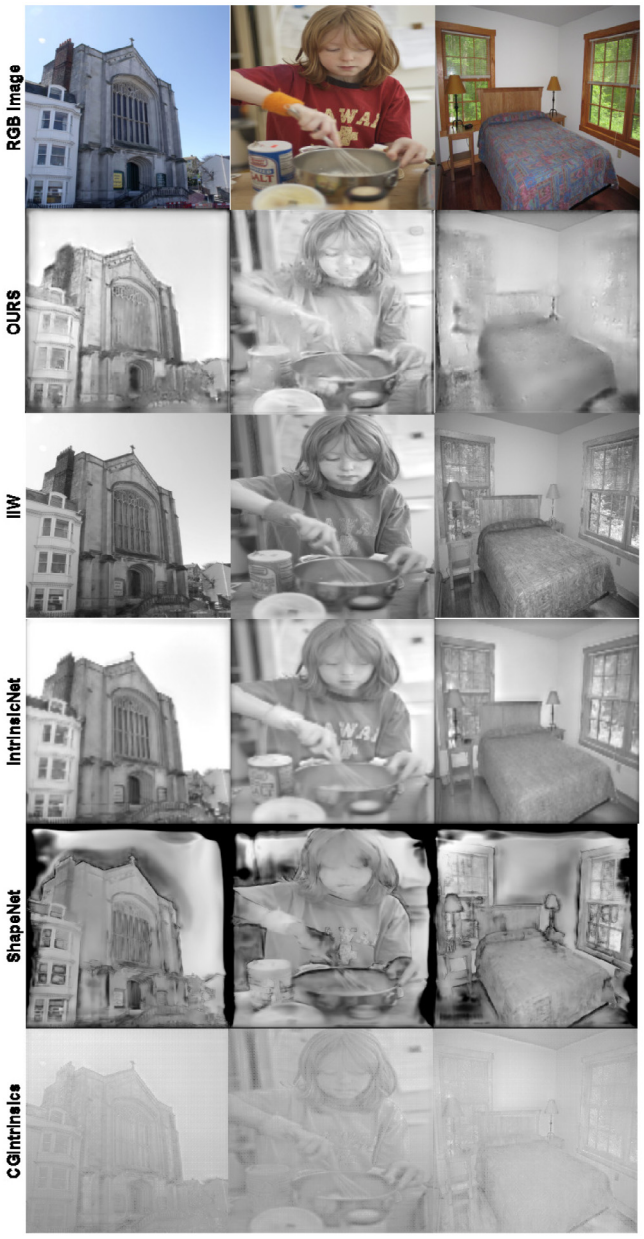

Fig. 11. Shading evaluations on IIW. Our model can produce scene-level shading maps that are free of texture or intensity ambiguities. Other models tend to overfit to the RGB images. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusion

We investigated the use of photometric invariance to steer a deep learning model for intrinsic image decomposition (albedo and shading). We proposed albedo and shading gradient descriptors which are derived from physics-based models as novel priors. Using the descriptors, albedo transitions are masked out and an initial shading map is calculated directly from the corresponding *RGB* image gradients in a learning-free unsupervised manner. Then, an optimization method was proposed to reconstruct the full dense shading map. Finally, we integrated the generated shading map into a novel deep learning framework to refine it and also to predict corresponding albedo image to achieve intrinsic image decomposition. Additionally, to train our model, a large-scale dataset of synthetic images of man-made objects was extended from 20K to 50K.

The evaluations were provided on five different object-level datasets (MIT, NIR-RGB, MIII, SIID, and ALOI), and two scene-level datasets (ARAP and IIW) with comprehensive setups without any fine-tuning or domain adaptation stage. The evaluations proved that our proposed model generated shading maps are more robust to texture artifacts and intensity ambiguities, which has been a long standing problem in the intrinsic image decomposition task. Since our model handles the undesired artifacts in the shading estimations, we also better differentiate albedo changes and achieve superior quantitative results.

Another conclusion is that deep learning based methods tend to overfit to the *RGB* image causing critical color leakages in the shading maps. When quantitatively evaluating, the leakage effect may not be reflected. That suggests that future work should focus on proposing better metrics for evaluation. In addition, the color leakage effect may not be observed when a model is trained and tested (or fine-tuned) on the same dataset (Narihira et al., 2015; Cheng et al., 2018). Therefore, it is important for intrinsic image decomposition methods to provide

Fig. 12. Albedo evaluations on IIW. Our model can generate proper scene-level albedo maps. We can also handle strong shadow casts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

cross-dataset or in-the-wild evaluations. Finally, we also tried to adapt several guided image-to-image translation and feature modulation techniques for our preliminary experiments to refine our initial shading maps with the *RGB* features. In particular, we tried the end-to-end trainable guided filter by Wu et al. (2018), bi-directional guided image-to-image translation by AlBahar and Huang (2019), spatially-adaptive normalization by Park et al. (2019), and deep spatial feature transform by Wang et al. (2018). Unfortunately, none of them were able to address the color leakage problem in the shading maps.

Our model is also not perfect. It might encounter limitations that mainly arise from the physics-based dichromatic reflection model from which the invariant descriptors are derived. Factors causing deviations from the dichromatic reflection model may cause inconsistencies. One example is the type of the surface. Since the model assumes matte surfaces, the descriptors are not expected to properly handle non-matte, glossy surfaces. Another limitation can be caused by image rendering or compression artifacts such as color banding, blur or heavy JPEG compression negatively affecting the physics-based image formation process.

CRediT authorship contribution statement

Anil S. Baslamisli: Methodology, Software, Visualization, Writing - original draft, Formal analysis, Investigation, Data curation. **Yang Liu:** Software, Resources, Visualization. **Sezer Karaoglu:** Project administration, Supervision, Validation, Writing - review & editing. **Theo Gevers:** Conceptualization, Project administration, Validation, Supervision, Writing - review & editing, Resources.

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.

Acknowledgments

This project was funded by the EU Horizon 2020 program No. 688007 (TrimBot2020). We thank Partha Das for his contribution to the experiments.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cviu.2021.103183.

References

AlBahar, B., Huang, J.B., 2019. Guided image-to-image translation with bi-directional feature transformation. In: IEEE International Conference on Computer Vision.
Barnard, K., Finlayson, G.D., 2000. Shadow identification using colour ratios. In: Color and Imaging Conference.
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., Das, P., Le, H.A., Karaoglu, S., Gevers, T., 2019. Shadingnet: Image intrinsics by fine-grained shading decomposition. ArXiv preprint arxiv:1912.04023v2.
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.
Baslamisli, A.S., Le, H.A., Gevers, T., 2018b. CNN based Learning using reflection and retinex models for intrinsic image decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Beigpour, S., Kolb, A., Kunz, S., 2015. A Comprehensive multi-illuminant dataset for benchmarking of intrinsic image algorithms. In: IEEE International Conference on Computer Vision.
Bell, S., Bala, K., Snavely, N., 2014. Intrinsic images in the wild. ACM Trans. Graph.
Bonneel, N., Kovacs, B., Paris, S., Bala, K., 2017. Intrinsic decompositions for image editing. Comput. Graph. Forum (Eurographics State of The Art Report).
Bousseau, A., Paris, S., Durand, F., 2009. User-assisted intrinsic images. ACM Trans. Graph.
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, X., Zhu, W., Zhao, Y., Yu, Y., Zhou, Y., Yue, T., Du, S., Cao, X., 2017. Intrinsic decomposition from a single spectral image. Appl. Opt. 5676–5684.
Cheng, L., Zhang, C., Liao, Z., 2018. Intrinsic image transformation via scale space decomposition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Cheng, Z., Zheng, Y., You, S., Sato, I., 2019. Non-local intrinsic decomposition with near-infrared priors. In: IEEE International Conference on Computer Vision.
Drew, M.S., Wei, J., Li, Z.N., 1998. Illumination-invariant color object recognition via compressed chromaticity histograms of color-channel-normalized images. In: IEEE International Conference on Computer Vision.
Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D., 2018. Revisiting deep intrinsic image decompositions. In: IEEE Conference on Computer Vision and Pattern Recognition.
Finlayson, G.D., 1992. Colour Object Recognition (Master's thesis). Simon Fraser University.

Finlayson, G.D., Chatterjee, S.S., Funt, B.V., 1998. Color angle invariants for object recognition. In: Color and Imaging Conference.
Finlayson, G.D., Hordley, S.D., Lu, C., Drew, M.S., 2006. On the removal of shadows from images. IEEE Trans. Pattern Anal. Mach. Intell. 59–68.
Gehler, P., 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.
Geusebroek, J.M., Burghouts, G.J., Smeulders, A., 2005. The Amsterdam library of object images. Int. J. Comput. Vis. 103–112.
Gevers, T., Smeulders, A., 1997. Object recognition based on photometric color invariants. In: Scandinavian Conference on Image Analysis.
Gevers, T., Smeulders, A., 1998. Image indexing using composite color and shape invariant features. In: IEEE International Conference on Computer Vision.
Gevers, T., Smeulders, A., 2000. Pictoseek: Combining color and shape invariant features for image retrieval. IEEE Trans. Image Process. 102–119.
Gevers, T., Smeulders, A., 2001. Color constant ratio gradients for image segmentation and similarity of texture objects. In: IEEE Conference on Computer Vision and Pattern Recognition.
Grosse, R., 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.
Harker, M., O'Leary, P., 2008. Least squares surface reconstruction from measured gradient fields. In: IEEE Conference on Computer Vision and Pattern Recognition.
Harker, M., O'Leary, P., 2011. Least squares surface reconstruction from gradients: Direct algebraic methods with spectral, Tikhonov, and constrained regularization. In: IEEE Conference on Computer Vision and Pattern Recognition.
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition.
Janner, M., Wu, J., Kulkarni, T.D., Yildirim, I., Tenenbaum, J.B., 2017. Self-supervised intrinsic image decomposition. In: Advances in Neural Information Processing Systems.
Land, E.H., McCann, J.J., 1971. Lightness and retinex theory. J. Opt. Soc. Amer. 1–11.
Lettry, L., Vanhoey, K., van Gool, L., 2018a. DARN: a deep adversarial residual network for intrinsic image decomposition. In: IEEE Winter Conference on Applications of Computer Vision.
Lettry, L., Vanhoey, K., van Gool, L., 2018b. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. In: International Pacific Conference on Computer Graphics and Applications.
Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., Chandraker, M., 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In: IEEE Conference on Computer Vision and Pattern Recognition.
Li, Z., Snavely, N., 2018a. CGIntrinsics: Better intrinsic image decomposition through physically-based rendering. In: European Conference on Computer Vision.
Li, Z., Snavely, N., 2018b. Learning intrinsic image decomposition from watching the world. In: IEEE Conference on Computer Vision and Pattern Recognition.
Matas, J., Marik, R., Kittler, J., 1995. On representation and matching of multi-coloured objects. In: IEEE International Conference on Computer Vision.
Matsushita, Y., Lin, S., Kang, S.B., Shum, H.Y., 2004. Estimating intrinsic images from image sequences with biased illumination. In: European Conference on Computer Vision.
Narihira, T., Maire, M., Yu, S.X., 2015. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: IEEE International Conference on Computer Vision.
Nayar, S.K., Bolle, R.M., 1996. Reflectance based object recognition. Int. J. Comput. Vis. 219–240.
Nestmeyer, T., Gehler, P.V., 2017. Reflectance adaptive filtering improves intrinsic image estimation. In: IEEE Conference on Computer Vision and Pattern Recognition.
Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y., 2019. Semantic image synthesis with spatially-adaptive normalization. In: IEEE Conference on Computer Vision and Pattern Recognition.
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.
Shafer, S., 1985. Using color to separate reflection components. Color Res. Appl. 210–218.
Shen, L., Tan, P., Lin, S., 2008. Intrinsic image decomposition with non-local texture cues. In: IEEE Conference on Computer Vision and Pattern Recognition.
Shen, J., Yang, X., Jia, Y., Li, X., 2011. Intrinsic images using optimization. In: IEEE Conference on Computer Vision and Pattern Recognition.
Shen, L., Yang, X., Li, X., Jia, Y., 2013. Intrinsic image decomposition using optimization and user scribbles. IEEE Trans. Cybern. 425–436.
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.
Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D., 2017. Neural face editing with intrinsic image disentangling. In: IEEE Conference on Computer Vision and Pattern Recognition.
Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations.
Stricker, M.A., 1992. Color and Geometry as Cues for Indexing. Technical Report, University of Chicago, Chicago, IL, USA.
van de Sande, K., Gevers, T., Snoek, C., 2009. Evaluating color descriptors for object and scene recognition. IEEE Trans. Pattern Anal. Mach. Intell. 1582–1596.
Wang, X., Yu, K., Dong, C., Loy, C.C., 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In: IEEE Conference on Computer Vision and Pattern Recognition.
Weiss, Y., 2001. Deriving intrinsic images from image sequences. In: IEEE International Conference on Computer Vision.
Wu, H., Zheng, S., Zhang, J., Huang, K., 2018. Fast end-to-end trainable guided filter. In: IEEE Conference on Computer Vision and Pattern Recognition.
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, J., Lin, Z., Yang, J., Shen, X., Lu, X., Huang, T.S., 2018. Generative image inpainting with contextual attention. In: IEEE Conference on Computer Vision and Pattern Recognition.
Zhang, Y., Funkhouser, T., 2018. Deep depth completion of a single RGB-D image. In: IEEE Conference on Computer Vision and Pattern Recognition.
Zhou, T., Krahenbuhl, P., Efros, A.A., 2015. Learning data-driven reflectance priors for intrinsic image decomposition. In: IEEE International Conference on Computer Vision.
Zhou, H., Yu, X., Jacobs, D.W., 2019. GLoSH: Global-local spherical harmonics for intrinsic image decomposition. In: IEEE International Conference on Computer Vision.