PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposition

Partha Das
University of Amsterdam,
3DUniversum
Amsterdam, The Netherlands
p.das@uva.nl

Sezer Karaoglu
University of Amsterdam,
3DUniversum
Amsterdam, The Netherlands
s.karaoglu@3duniversum.com

Theo Gevers
University of Amsterdam,
3DUniversum
Amsterdam, The Netherlands
Th.Gevers@uva.nl

Abstract

Intrinsic image decomposition is the process of recovering the image formation components (reflectance and shading) from an image. Previous methods employ either explicit priors to constrain the problem or implicit constraints as formulated by their losses (deep learning). These methods can be negatively influenced by strong illumination conditions causing shading-reflectance leakages.

Therefore, in this paper, an end-to-end edge-driven hybrid CNN approach is proposed for intrinsic image decomposition. Edges correspond to illumination invariant gradients. To handle hard negative illumination transitions, a hierarchical approach is taken including global and local refinement layers. We make use of attention layers to further strengthen the learning process.

An extensive ablation study and large scale experiments are conducted showing that it is beneficial for edge-driven hybrid IID networks to make use of illumination invariant descriptors and that separating global and local cues helps in improving the performance of the network. Finally, it is shown that the proposed method obtains state of the art performance and is able to generalise well to real world images. The project page with pretrained and finetuned models and network code can be found here.

1. Introduction

Intrinsic Image Decomposition (IID) is the process of recovering the image formation components such as reflectance (albedo) and shading (illumination) from an image. The reflectance image can be used for albedo texture edits [7, 33, 50], fabric recolouring [48] or semantic segmentation [4]. As for shading, the illumination image can be used for relighting [43] or shape-from-shading i.e. estimating the shape/geometry of objects or scenes [21, 45].

The problem of IID is inherently ill-defined. Therefore, previous IID approaches employ priors to constrain the problem. Retinex [24] is based on gradient information derived from images where shading variations correspond to small (soft) gradients and reflectance transitions to larger (stronger) ones. Other constraints are explored by [2], like piece-wise constancy, parsimony of reflectance, shading smoothness, etc. Other approaches include global sparsity priors on the palette of colours (albedo's) and modelling the problem as latent variable Random Fields by [18]. However, these explicitly imposed constraints (i.e. assumptions about the world) may limit the applicability of these methods. Recently, deep learning based methods are proposed [34, 42]. These methods are based on implicit constraints as formulated by losses, and multiple datasets or image sequences [28, 46]. However, these approaches are purely data-driven and therefore they may be limited in their generalisation abilities (dataset bias). Traditional constraints and deep learning approaches are combined by [16] by means of image edge guidance. However, edge-driven hybrid methods can be influenced by strong illumination conditions. For example, in case of strong shadows, the network may classify shadows as being reflectance edges. This happens when the gradient assumption is violated: illumination changes correspond to soft gradients and reflectance transitions to hard ones. This leads to the well-known problem of shading-reflectance leakage i.e. illumination (strong shadow/shading) transitions which are interpreted/classified as albedo transitions also called hard negative illumination transitions, or simply hard (illumination) negatives.

Therefore, in this paper, an edge-driven hybrid CNN approach is proposed using gradients based on illumination invariant descriptors i.e. Cross Color Ratios (CCR) [19]. CCR are illumination (including shadows and shading patterns) invariant gradients and hence only dependent on albedo changes. To solve for hard negative illumination transitions, a hierarchical CNN is proposed including global and local refinement layers. The global layer ensures a

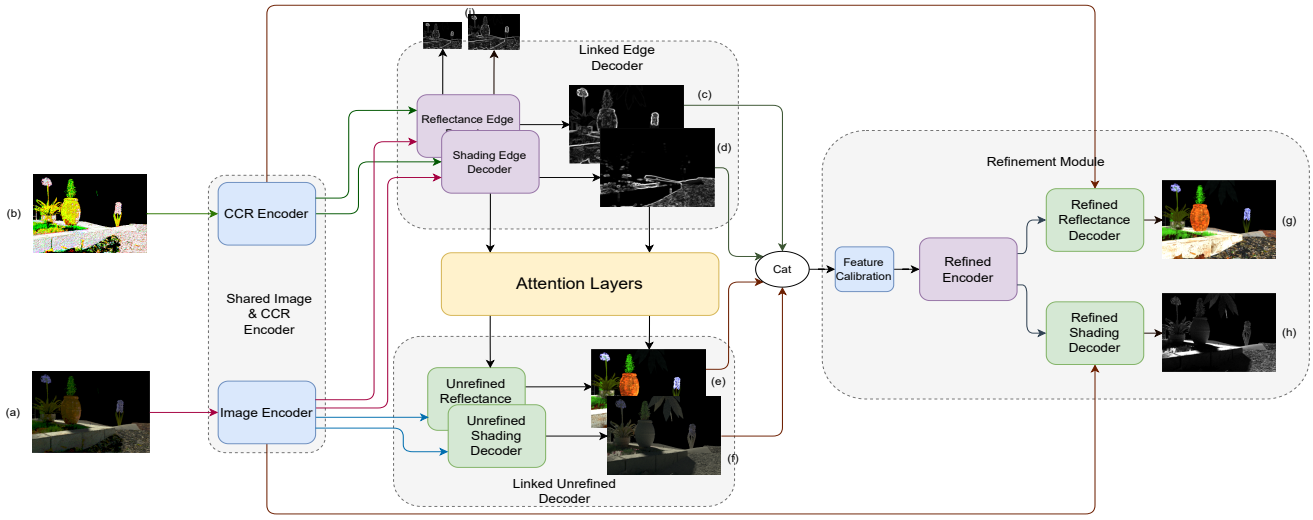

Figure 1. Overview of the proposed Network. The architecture consists of 4 sub-modules denoted by dotted boxes. Inputs to the network are (a) a RGB image and (b) a CCR image. The CCR image is computed from (a). The outputs of the networks are: (c) the reflectance edge, (d) the shading edge, (e) the unrefined reflectance prediction, (f) the unrefined shading prediction, (g) the final refined reflectance, (h) the final refined shading, and (i) the scaled edge outputs @(64, 128).

smooth image decomposition eliminating (soft) negative illumination transitions and therefore minimising shading-reflectance misclassification. The local refinement will eliminate the hard (illumination) negatives. A two-staged approach is beneficial because by adding an incremental parameter space that is conditioned on the previous step, the network is no longer required to model both strong and soft illumination patterns in a single process, but rather a refinement elimination process to remove the hard negatives. The proposed method implicitly encodes the image intrinsics within the network without the need of manual thresholding. Since intrinsic components are spatially dependent, spatial attention layers are included. This allows the network to focus on image areas containing hard negatives. Fig. 1 provides an overview of the proposed network.

In summary, our contributions are as follows:

- An end-to-end edge-driven hybrid approach is proposed for intrinsic image decomposition using gradients based on illumination invariant descriptors.
- To solve for hard negative illumination transitions, a hierarchical approach is taken including global and local refinement layers.
- It is shown that separating the parameter space in a global and local space, rather than a unified parameter space, outperforms single parameter space learning.
- It is shown that the proposed algorithm is able to achieve state of the art performance and is able to generalise well to real world images.

2. Related Work

Earlier methods on intrinsic image decomposition are mainly focused on exploring hand-crafted priors to reduce the solution space. [24] argues that sharp gradient changes belong to reflectance changes, while soft transitions correspond to illumination patterns. [19] proposes the Cross Color Ratios (CCR). CCR are illumination (including shadows and shading patterns) invariant gradients and hence only dependent on albedo transitions. Surprisingly, they have not been employed in the domain of IID.

Other methods, like [40] explore texture cues where image areas having the same reflectance values should also have the same intensity values. [2, 18] adds multiple constraints like piece-wise consistency for reflectance and smoothness priors for shading. Other constraints include textures [18, 41], depth cues [1, 25], infrared priors [15] and surface normals [23]. Shading is also decomposed into geometry and illumination by regularising them individually [13]. Different optimisation frameworks are explored by [8], where Conditional Random Fields are used to optimise pairwise reflectances. More implicit reflectance constraints like multiple frames are explored by [32, 46]. User annotated priors are explored by [8, 10, 11, 35]. For these methods, the application domains are often limited to single objects.

[34] is the first to represent the problem of IID using a CNN. [42] expands this by introducing skip connections and inter-component connections to enforce component inter-dependence. [3, 22, 54] explore decomposing the shading component further into shadows, direct and am-

bient lighting cues. Furthermore, [5] introduces RetiNet by parameterising the Retinex algorithm as an end-to-end learnable framework. Other approaches include Laplacian pyramids based on scale space learning [14], adversarial residual networks [26], inverse rendering [38] and image edge guidance [16]. Finally, [28] combines multiple constraints as loss functions and trains an end-to-end system using 4 different datasets simultaneously. [27] models the problem as differentiable rendering layers and trains it in an end-to-end manner with supervision on reflectance, roughness, normals, depths, etc. Apart from the supervised methods, unsupervised learning approaches for IID are also explored in [28, 30, 49, 51]. Although the results of these approaches are promising, these methods are not always able to fully disentangle (strong) shading and reflectance transitions i.e shading-reflectance leakage problem.

3. Methodology

3.1. Illumination Invariant Gradients

Given an RGB image and two neighbouring pixels p_1 & p_2. Then, the Cross Color Ratios are defined by:

$$M_{RG} = \frac{R_{p_1}\, G_{p_2}}{R_{p_2}\, G_{p_1}}\,, M_{RB} = \frac{R_{p_1}\, B_{p_2}}{R_{p_2}\, B_{p_1}}\,, M_{GB} = \frac{G_{p_1}\, B_{p_2}}{G_{p_2}\, B_{p_1}}\,, \tag{1}$$

where M_{RG}, M_{RB}, M_{GB} are the CCR for the (R, G), (R, B) & (G, B) channel pairs, respectively. Taking the logarithm on both sides of the equation, we obtain:

$$\begin{aligned} log(M_{RG}) &= log(R_{p_1}\, G_{p_2}) - log(R_{p_2}\, G_{p_1})\,, \\ log(M_{RB}) &= log(R_{p_1}\, B_{p_2}) - log(R_{x_2}\, B_{x_1})\,, \\ log(M_{GB}) &= log(G_{p_1}\, B_{p_2}) - log(G_{p_2}\, B_{p_1})\,. \end{aligned} \tag{2}$$

Let the image formation process be modelled by [39]:

$$I = m(\vec{n}, \vec{l}) \int_{\omega} e(\lambda)\, \rho_b(\lambda)\, f(\lambda)\, \mathrm{d}\lambda\,, \tag{3}$$

where, I is the captured image; λ is the incoming light wavelength within the visible spectrum ω; m is a function depending on the object geometry and light sources; $\vec{n}$ denotes the surface normal and $\vec{l}$ corresponds to the light source direction. f indicates the spectral camera sensitivity and e describes the spectral power distribution of the light source. Reflectance is denoted by ρ and is related to the albedo/colour of the object. Discretising the model, we obtain:

$$C_{p_1} = m(\vec{n},\, \vec{l})\, e^{C_{p_1}}(\lambda)\, \rho^{C_{p_1}}(\lambda)\,, \tag{4}$$

where C_{p_1} is colour channel C for pixel p_1 for a RGB image.

For two neighbouring pixels p_1 and p_2, the same illumination conditions can be assumed since they are very close to each other. Hence:

$$e^{C_{p_1}} = e^{C_{p_2}}\,, \tag{5}$$

Combining Eq. (2) and Eq (4) results in:

$$\begin{aligned} log(M_{RG}) = {} & log(R_{p_1}\, G_{p_2}) - log(R_{p_2}\, G_{p_1})\,, \\ log(M_{RG}) = {} & log(R_{p_1}) + log(G_{p_2}) \\ & - log(R_{p_2}) - log(G_{p_1})\,, \\ log(M_{RG}) = {} & log(\rho^{R_{p_1}}(\lambda)) + log(\rho^{G_{p_2}}(\lambda)) \\ & - log(\rho^{R_{p_2}}(\lambda)) - log(\rho^{G_{p_1}}(\lambda))\,. \end{aligned} \tag{6}$$

Hence, CCR are illumination invariant differences and they are only dependent on the reflectance transitions. To reconstruct the intrinsic (shading and reflectance) images from these edges, a CNN is proposed.

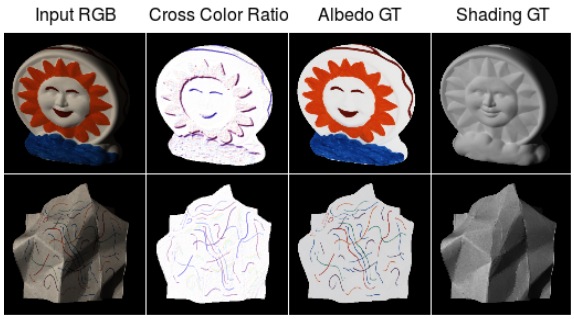

Figure 2. The CCR value becomes 1 where the reflectance is constant. CCR changes correspond to reflectance changes. CCR are illuminantion invariant. Images are gamma corrected for visualisation purposes.

Fig. 2 shows that CCR, computed for two images, correspond to reflectance changes. They are not dependent on illumination changes. However, noisy regions and strong illumination may introduce shading-reflectance leakage transitions. Therefore, we propose a hierarchical CNN including global and local refinement layers. The refinement layers are used to cope with hard negative illumination transitions. To this end, CCR are integrated in the network to steer the learning process through the use of encoded features in: 1) global refinement through edge prediction, and 2) local patch-wise consistent refinement.

3.2. Network Architecture & Details

The network architecture is composed of four subcomponents: 1) A Shared Image & CCR Encoder, 2) Linked Edge Decoder, 3) Unrefined Decoder and 4) Local Refinement Module. The entire network is trained end-to-end.

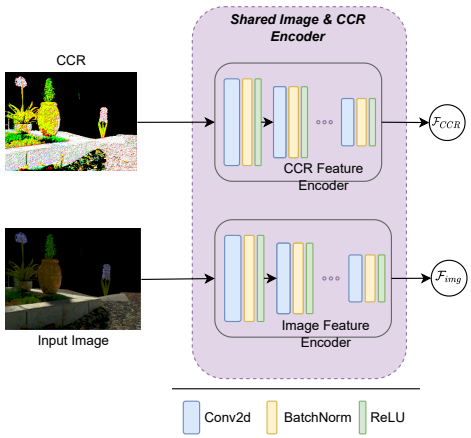

Figure 3. Overview of the shared encoder. The input RGB images and the corresponding CCR are generated by independent encoders. These encoded features are then passed on to the rest of the network as a guidance for the global and local layers.

Shared Image & CCR Encoder: Fig. 3 shows the proposed image and CCR encoders. The input and corresponding CCR image are encoded through two separate encoders. This allows the CCR Encoder to learn illumination invariant reflectance transition feature ($\mathcal{F}_{CCR}$), while the image encoder learns an entangled feature composed of illumination and reflectance cues ($\mathcal{F}_{img}$), independently. These encodings are reused in the later parts of the architecture, (Fig. 1), allowing independent feature usage for both global and local layers.

Linked Edge Decoder: $\mathcal{F}_{CCR}$ & $\mathcal{F}_{img}$, are passed on to the linked edge decoders (Fig. 4). Interconnections within the decoders enable to learn a relational representation of the cues. Thus, the decoder can learn both reflectance edges (d) and illumination edges (e) jointly. Apart from the standard supervision on the output, a feature space scale supervision [47] is also added. To facilitate this, two scales (64×64 and 128×128) are obtained. These scales are transformed directly from the intermediate CNN features through a common convolution. This allows the convolution to learn a transformation from feature to image space. This supervision ensures that the decoder produces edges that are consistent across scales and feature spaces.

Unrefined Decoder: The unrefined decoder consists of a similar set of decoders as the edge decoder set as illustrated in Fig. 4. For every block in the decoder of the Linked edge decoder, the output is fed through an attention layer before being convolved through the respective block in the unrefined decoder. Skip connections (not shown in the figure, for brevity), are also added to the decoders of $\mathcal{F}_{img}$ to provide additional cues. The attention enhanced edge guidance allows the network to focus on global consistencies for the intrinsic images, like segment wise reflectance consistency and smooth gradient changes for illumination. However, these intrinsics are not necessarily locally consistent, but may still contain local imperfections such as shading-reflectance leakage transitions caused by hard negative illumination transitions.

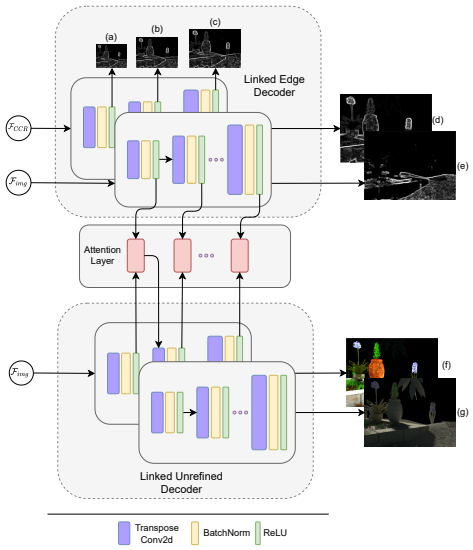

Figure 4. Overview of the linked edge decoder and the linked unrefined decoder. The encoded features from the CCR and the image are used to decode the reflectance (d) and illumination edges (e). Scale space side outputs (a) - (c), used to enforce scale consistency, are also added. The edge features are passed on through the attention layers to the unrefined decoder, which outputs the globally consistent unrefined reflectance (f) and shading (g).

Local Refinement Module: To cope with hard illumination negatives, the output from the previous decoders are passed on to the refinement module. Fig. 5 illustrates this strategy. The reflectance and shading edges, unrefined reflectance and shading pairs are concatenated and convolved through a feature calibration layer. The calibrated features are then passed through an encoder-decoder to obtain the final output. Additional local patch-wise guidance is provided through skip connections from $\mathcal{F}_{CCR}$ & $\mathcal{F}_{img}$, which are also passed through attention layers to selectively focus on hard negative areas (not shown in the figure for brevity).

The shading is computed through a separate decoder. The proposed configuration allows the decoder to use the shading cues as an additional source of information to correct the reflectance and vice versa.

3.3. Loss Functions

To train the network, supervision is added to each of the output channels of the network. These are: 1) the edge loss

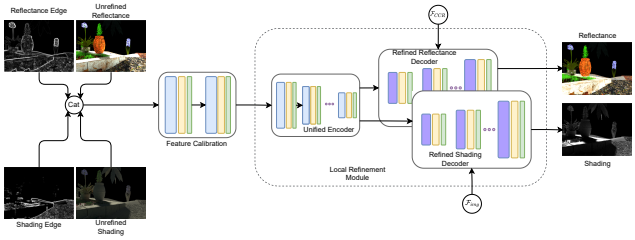

Figure 5. Overview of the local refinement module. The supplementary material provides a higher resolution version.

($\mathcal{L}_e$), 2) the unrefined loss ($\mathcal{L}_u$), and 3) the refined loss ($\mathcal{L}_r$). For each of these outputs, a combination of scale invariant MSE [34] and standard MSE loss is used.

1) Edge Losses: The edge decoder outputs reflectance and shading edges, together with their scaled versions of 64×64 and 128×128. In total, there are 3 outputs for reflectance and shading. The total edge loss is defined by:

$$\begin{aligned} \mathcal{L}_e = \quad & \mathcal{L}_{AE} + \mathcal{L}_{AE64} + \mathcal{L}_{AE128} \\ & + \mathcal{L}_{SE} + \mathcal{L}_{SE64} + \mathcal{L}_{SE128}, \end{aligned} \tag{7}$$

where $\mathcal{L}_{AE}$ & $\mathcal{L}_{SE}$ are the losses on the full scale of reflectance and shading edges; $\mathcal{L}_{AE64}$ & $\mathcal{L}_{SE64}$ are the losses on reflectance and shading feature outputs at scale 64×64; $\mathcal{L}_{AE128}$ & $\mathcal{L}_{SE128}$ are the losses on the 128×128 scale. The ground truth for the edges is calculated using a Canny Edge operator. The reflectance edge is calculated from the reflectance ground truth. The NED [4] dataset (described in more detail in the experimental section) provides fine grained shading decompositions (shadow maps and ambient/inter-reflections). The shadow map is used for the shading edge calculation. For datasets without such ground truth decompositions, the shadow edges are simulated by subtracting the reflectance edges from the shading ground truth edges. Under the Lambertian model, it can be assumed that an image is the multiplication of reflectance and shading. Hence, the subtraction provides an approximation for shading edges.

2) Unrefined Losses: The unrefined decoder is constrained by the following loss:

$$\mathcal{L}_u = \mathcal{L}_{uA} + \mathcal{L}_{uS}, \tag{8}$$

where $\mathcal{L}_{uA}$ is the loss on the unrefined decoder's reflectance output and $\mathcal{L}_{uS}$ on the unrefined decoder's shading output.

3) Refined Loss: Finally, the following loss is applied on the outputs generated by the refined decoder:

$$\mathcal{L}_r = \mathcal{L}_A + \mathcal{L}_S, \tag{9}$$

where $\mathcal{L}_A$ is the loss on the final reflectance output and $\mathcal{L}_S$ on the final shading output.

To enforce component dependence as a supervision, the outputs from the network are recombined and compared with the input image for the reconstruction loss $\mathcal{L}_{rec}$. Since this decoder focuses on localised correction of the outputs, a Structural Dissimilarity (DSSIM) loss is added to regularise the network:

$$\mathcal{L}_{dssim} = \mathcal{L}_{\delta A} + \mathcal{L}_{\delta S}, \tag{10}$$

where $\mathcal{L}_{\delta A}$ and $\mathcal{L}_{\delta S}$ are the losses on the dissimilarity measure of the reflectance and shading respectively. DSSIM measures the divergence of structural changes. This allows the final reflectance and shading, after both global and local corrections, to be closer to the ground truth.

Finally, to make the network explicitly focus on the perceived quality of the decomposition, a perceptual loss is added. The features of a VGG16 [44] network trained on ImageNet are used. This is defined as follows:

$$\mathcal{L}_P(A, \hat{A}) = \sum_i ||\mathcal{F}_{VGG_i}(A) - \mathcal{F}_{VGG_i}(\hat{A})||_1. \tag{11}$$

where $\mathcal{L}_P$ is the perceptual loss; $\mathcal{F}_{VGG}$ is the feature space transform function; A is the predicted reflectance; $\hat{A}$ is the corresponding ground truth; i is the layer index of the VGG16 network (set to the last 4 for all the experiments). Combining all the losses, the total training objective for the network is defined by:

$$\begin{aligned} \mathcal{L} = \mathcal{L}_r + \lambda_u \mathcal{L}_u + \lambda_e \mathcal{L}_e + \lambda_p \mathcal{L}_p \\ + \lambda_d \mathcal{L}_{dssim} + \mathcal{L}_{rec}. \end{aligned} \tag{12}$$

where, the component specific hyper-parameters λ_u, λ_d, λ_e and λ_p are empirically set to be 0.5, 0.4, 0.4 and 0.05 respectively. For more details please see the supplementary material.

4. Experiments and Results

Datasets Quantitative experiments are conducted on four datasets: NED [4], MIT [20], Sintel [12] and IIW [8]. The train/test splits, as provided by the original papers, are used. The proposed network is trained on the NED dataset and finetuned on the other datasets. In addition, qualitative results are provided on the Trimbot dataset [37], which consists of real-world garden dataset. This dataset does not come with ground truth annotations.

Evaluation Metric. Following the literature, the standard MSE error metric, the LMSE metric [20], (with a windows size of 20) and the structural dissimilarity metric (DSSIM) are used. Finally, for IIW, the WHDR metric [8] is used. For all the datasets, the same train and test splits are used for all methods.

4.1. Ablation Study

Influence of Illumination Invariants: In this experiment, the influence of the CCR to steer the global-local process is studied by removing the CCR encoder from the network. In this way, the edge decoder is completely dependent on the RGB image cues given by the input. Tab. 1 shows the results of this experiment.

	Reflectance			Shading		
	MSE	LMSE	DSSIM	MSE	LMSE	DSSIM
w/o Physics Priors	0.0039	0.0449	0.2590	0.0032	0.0812	0.2356
w/o Edge Guidance	0.0033	0.0415	0.2411	0.0029	0.0782	0.2543
w Canny Edge Guidance	0.0061	0.0566	0.2721	0.0034	0.0879	0.2538
w/o Local Refinement	0.0020	0.0361	0.1192	0.0031	0.0768	0.2651
w/o Attention Layers	0.0019	0.0330	0.0776	0.0026	0.0704	0.1301
Proposed	**0.0015**	**0.0289**	**0.0688**	**0.0018**	**0.0489**	**0.1005**

Table 1. An ablation study on the various parts of our network. From the results, the proposed component does indeed have a positive influence on the performance of the network. w - with, w/o - without

From Tab. 1, it is shown that the removal of the CCR (*w/o Physics Priors*) degrades the performance of the proposed network. This is because the modified network now only relies on the RGB edges of the input image. These edges include strong illumination transitions and therefore the network is sensitive to hard negatives. As a result, all the metrics show a decrease in performance. In conclusion, it is beneficial for edge-driven hybrid IID networks to make use of illumination invariant descriptors, rather than learning a data-distribution.

Influence of Reflectance and Shading Edges: (1) The influence of edges as a source of guidance is analysed. The edge decoder part is removed from the unrefined decoder. The CCR features are maintained in the local refinement module. (2) It is tested whether learning the reflectance and shading edges computed directly from the image edges can replace the CCR to the edge translation subnet. Therefore, the input to the CCR encoder in Fig. 1 is replaced by Canny edges calculated directly from the images. This setup corresponds to [16].

For the setup described in (1), results for removing the edge guidance, (*w/o Edge Guidance*) in Tab. 1 shows that the performance decreases. However, it is still an improvement over the previous experiment (*w/o Physics Priors*) in the table. This shows that even if the global edge guidance is removed, the local physics prior is still able to help.

In setup (2), using only Canny edges is shown to degrade the performance. This is because the edge-to-edge transition is lacking any guidance. This subsequently makes the local correction fail. Thus, this experiment shows the importance of consistent global guidance for intrinsic image decomposition.

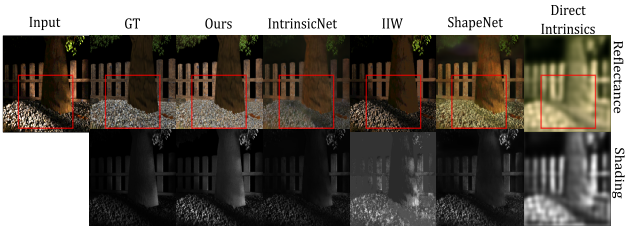

Figure 6. Comparison of the proposed network with sota methods. In the image, the tree trunk has textures and has a hard negative illumination transition on the base. It is shown that the proposed method can both recover from hard negatives and also prevent shadow-reflectance misclassifications in the shading component.

Influence of Local Refinement: The last refinement module is removed and only the edge guided module is kept. This makes the decoder module to handle both global and local consistencies in the same parameter space. The result for this experiment is shown in Tab. 1. The results show that there is an improvement for the explicit global & local parameter disentanglement. This shows that the global and local context separation is an integral part of the proposed network architecture.

Influence of Attention Layers: In this experiment, all the attention layers are removed from the network and replaced by direct connections. The result of this experiment is also shown in Tab. 1.

The results (*w/o Attention Layers* compared to *Proposed*) show that the inclusion of the attention mechanism improves the performance. The metrics (LMSE & DSSIM) demonstrate improvements indicating that the attention layers are beneficial for local corrections. This means that there is no single uniform transformation to be applied to all pixels in an image to recover the intrinsic components. A detailed study supporting this hypothesis can be found in the supplementary material.

4.2. Evaluations & Results

Comparison on NED Dataset: In this experiment, the proposed network is compared to state-of-the-art (sota) methods. All presented methods are re-trained using the NED dataset. For a fair comparison, the same train and test split are used as well as the optimum parameters as mentioned in the respective papers. Recent unsupervised methods of [30], [29] and [51] are also included for comparison. The numerical results are shown in Tab. 2 and the visual results in Fig. 6.

The results show that the proposed method outperforms the baselines for all metrics. As illustrated in Fig. 6, the proposed method recovers the intrinsics more robustly. For example, the proposed method is able to disentangle the shadows at the base of the tree, while other methods suffer from

	Reflectance			Shading		
	MSE	LMSE	DSSIM	MSE	LMSE	DSSIM
	Supervised methods					
Color Retinex [20]	0.0114	0.1204	0.3280	0.0193	0.2334	0.3515
IIW [8]	0.0095	0.1343	0.2098	0.0111	0.1861	0.3511
Direct Intrinsics [34]	0.0073	0.1205	0.3756	0.0065	0.1798	0.3843
IntrinsicNet [5]	0.0035	0.0449	0.2367	0.0037	0.0791	0.2110
ShapeNet [42]	0.0053	0.0597	0.2516	0.0050	0.0910	0.2186
	Unsupervised methods					
USI3D [30]	0.0081	0.0360	0.1886	0.0143	0.0608	0.2140
IIDWW [29]	0.0149	0.0447	0.2229	0.0175	0.0698	0.2346
InverseRenderNet [51]	0.0478	0.0642	0.2751	0.0505	0.2597	0.3382
Ours	**0.0015**	**0.0289**	**0.0688**	**0.0018**	**0.0489**	**0.1005**

Table 2. Numerical evaluation comparison between the proposed architecture and sota baselines on the NED dataset. The first group are supervised methods, second group are unsupervised methods and the final group is the our proposed methods.

hard negatives resulting in discoloured reflectances.

MIT Intrinsics Dataset: The proposed network is finetuned on the MIT Intrinsics dataset. The quantitative numbers are shown in Tab. 3, while visuals are shown in Fig. 7

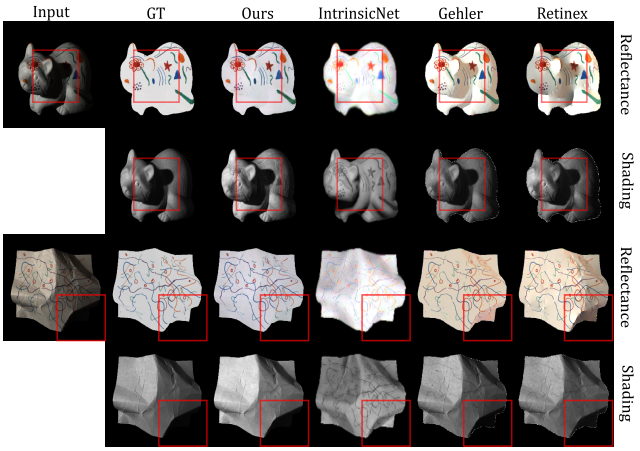

Figure 7. Qualitative evaluation of the proposed network on the MIT dataset. The proposed method is the only method able to disentangle shadows from reflectance cues.

The results show that the proposed method outperforms other baseline methods for all the metrics. The method is able to recover the intrinsic components robustly. For example, the shading map obtained by IntrinsicNet on the raccoon completely misses the shadow. For our method, the reflectance of the paper is much smoother.

MPI Sintel Dataset: The results on the MPI Sintel Dataset are given in Tab. 4.

The proposed method generally outperforms, on average, all other methods except for the LMSE metric. Particularly, the proposed method outperforms other methods for the DSSIM metric for both components and hence robustly

	Reflectance			Shading		
	MSE	LMSE	DSSIM	MSE	LMSE	DSSIM
	Supervised methods					
SIRFS [2]	0.0129	0.0572	-	0.0066	0.0309	-
Gehler *et al.* [18]	0.0065	0.0393	-	0.0051	0.0282	-
Zhou *et al.* [55]	0.0252	-	-	0.0229	-	-
Color Retinex [20]	0.0084	0.0447	-	0.076	0.0343	-
Direct Intrinsics [34]	0.0277	0.0585	0.1526	0.0154	0.0295	0.1328
ShapeNet [42]	0.0278	0.0503	0.1465	0.0126	0.0240	0.1200
CGIntrinsics [28]	0.0221	0.0349	0.1739	0.0186	0.0259	0.1652
CGIntrinsics [28] (MIT Finetuned)	0.0167	0.0319	0.1287	0.0127	0.0211	0.1376
ParCNN [52]	0.0109	0.0462	0.0929	0.0086	0.0537	0.0999
CasQNet [31]	0.0091	0.0212	0.0730	0.0081	0.0192	0.0659
IntrinsicNet [5]	0.0104	0.0854	-	0.0304	0.2038	-
Baslamisli *et al.* [6]	0.0060	0.0438	-	0.0069	0.0418	-
	Unsupervised methods					
STAR [49]	0.0137	0.0614	0.1196	0.0114	0.0672	0.0825
USI3D [30]	0.0156	0.0640	0.1158	0.0102	0.0474	0.1310
IIDWW [28]	0.0126	0.0591	0.1049	0.0105	0.0457	0.1159
InverseRenderNet [51]	0.0234	0.0573	0.1148	0.0186	0.0765	0.1276
Ours	**0.0028**	**0.0136**	**0.0340**	**0.0035**	**0.0183**	**0.0493**

Table 3. Quantitative evaluation comparison of the proposed architecture on the MIT Intrinsic Dataset [20].

	Reflectance			Shading		
	MSE	LMSE	DSSIM	MSE	LMSE	DSSIM
Color Retinex [20]	0.0606	0.0366	0.2270	0.0727	0.0419	0.2400
Lee *et al.* [25]	0.0463	0.0224	0.1990	0.0507	0.0192	0.1770
SIRFS [1]	0.0420	0.0298	0.2100	0.0436	0.0264	0.2060
Chen *et al.* [13]	0.0307	0.0185	0.1960	0.0277	0.0190	0.1650
Direct Intrinsics [34]	0.0100	0.0083	0.2014	0.0092	0.0085	0.1505
Fan *et al.* [16]	0.0069	**0.0044**	0.1194	**0.0059**	**0.0042**	0.0822
Ours	**0.0015**	0.0080	**0.0399**	0.0105	0.0507	**0.0508**

Table 4. Standard numerical evaluation comparison of the proposed method on the MPI Sintel Dataset [12] (scene split).

preserves global and local structural components. From the outputs (included in the supplementary material), it is shown that the shading computed by the proposed method has a lower pixel scale but is structurally consistent. This explains the lower performance on other metrics compared to the DSSIM metric (for shading). The MSE and LMSE metrics are more sensitive to outliers and optimise to a smaller Euclidean distance. Additional details can be found in the supplementary material.

IIW Dataset: For this experiment, the proposed network is finetuned on the IIW dataset. The results for the network are given in Tab. 5. Visuals are shown in Fig. 8.

Due to the nature of the ground truth provided by this dataset, the proposed method can only use the ordinal loss [8] for finetuning on the train set. The original proposed network is trained on the MSE and perceptual losses only. However, it is still able to perform comparatively. GLoSH [54] is the best performing method. However, it needs both the normal and lighting ground truth for supervision. Fig. 8 shows flatter reflectance cues (on the wall and bed), insensitive to illumination patterns (i.e. shadows and shading). [27] needs supervision on normals, depth,

Methods	WHDR (mean)	WHDR (Outdoors only)
Direct Intrinsics [34]	37.3	-
Color Retinex [20]	26.9	-
Garces *et al.* [17]	25.5	-
Zhao *et al.* [53]	23.2	-
IIW [8]	20.6	21.7
Nestmeyer *et al.* [36]	19.5	-
Bi *et al.* [9]	17.7	-
Sengupta *et al.* [38]	16.7	-
Li *et al.* [27]	15.9	21.3
CGIntrinsics [28]	15.5	23.1
GLoSH [54]	15.2	-
Fan *et al.* [16]	15.4	21.6
Ours	21.3	20.8
Fan *et al.* * [16]	**14.45**	20.2
Our*	18.5	**18.4**

Table 5. Performance in terms of the WHDR metric. The proposed method is trained with general image learning losses. When testing it only on outdoor images, the proposed method shows competitive performance. * denotes outputs post-processed with a guided filter.

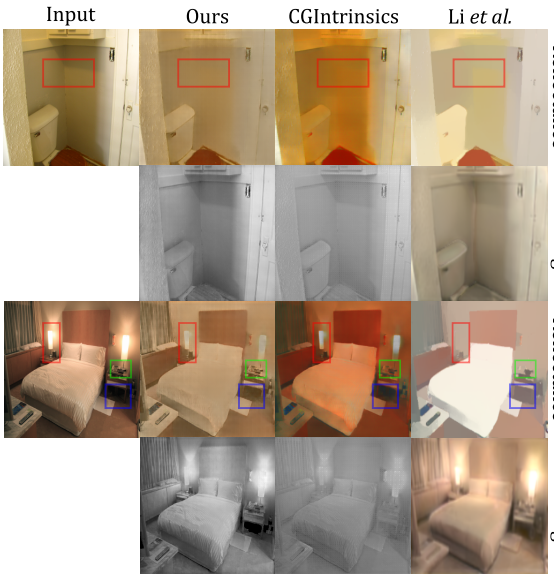

Figure 8. Results of the proposed method compared with CGIntrinsics [28] and [27]. The proposed method shows comparable performance with other methods, despite being trained primarily on outdoor garden images.

roughness and lighting, in addition to the reflectance and shading and has 7 separate training stages. The reflectance cues are missing details, while the shading cues are quite blurry. The proposed network is trained only on outdoor images (domain difference). To test the domain related performance, the WHDR metric is applied to only the outdoor images in the test set. It is shown that the proposed method performs best.

Trimbot Dataset: To test the generalisation of the proposed method to real world scenarios, results on the Trimbot Dataset [37] are shown in Fig. 9.

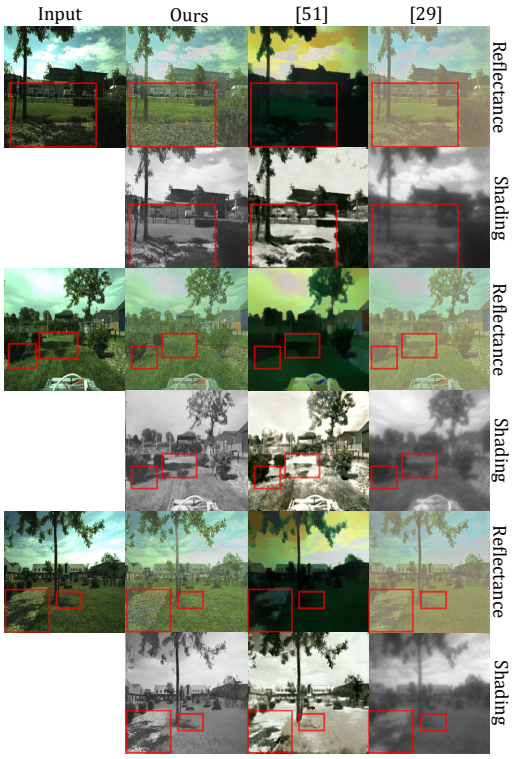

Figure 9. Results of the proposed method on the Trimbot dataset. The proposed method is trained and finetuned on a fully synthetic dataset, yet it can recover proper reflectance by removing both soft and hard illumination patterns. The supplementary materials contain additional results on in-the-wild images.

5. Conclusion

In this paper, an end-to-end edge-driven hybrid approach has been proposed for intrinsic image decomposition. Edges are based on illumination invariant descriptors. To handle hard negative illumination transitions, a hierarchical approach has been taken including global and local refinement layers. The global guidance was integrated through a reflectance and shading edge formulation. For the local guidance, the encoded CCR features were used as a prior to the local refinement module.

Based on extensive ablation study and large scale experiments, is has been shown that (1) it is beneficial for edge-driven hybrid IID networks to make use of illumination invariant descriptors, (2) separating global and local cues into different modules indeed helps in improving the performance of the network both qualitatively and quantitatively, (3) the proposed method obtains sota performance in recovering the intrinsics, and (4) it is able to generalise well to real world images.

References

[1] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, 2013. 2, 7

[2] J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, pages 1670–1687, 2015. 1, 2, 7

[3] Anil S. Baslamisli, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Shadingnet: Image intrinsics by fine-grained shading decomposition. *IJCV*, 129:2445–2473, 2021. 2

[4] A. S. Baslamisli, T. T. Groenestege, P. Das, H. A. Le, S. Karaoglu, and T. Gevers. Joint learning of intrinsic images and semantic segmentation. In *ECCV*, 2018. 1, 5

[5] A. S. Baslamisli, H. A. Le, and T. Gevers. Cnn based learning using reflection and retinex models for intrinsic image decomposition. In *CVPR*, 2018. 3, 7

[6] Anil S. Baslamisli, Yang Liu, Sezer Karaoglu, and Theo Gevers. Physics-based shading reconstruction for intrinsic image decomposition. *Comput. Vis. and Image Understanding*, pages 1–14, 2020. 7

[7] Shida Beigpour and Joost van de Weijer. Object recoloring based on intrinsic image estimation. *ICCV*, pages 327–334, 2011. 1

[8] S. Bell, K. Bala, and N. Snavely. Intrinsic images in the wild. *ACM TOG*, 2014. 2, 5, 7, 8

[9] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *ACM TOG*, 34(4), July 2015. 8

[10] N. Bonneel, K. Sunkavalli, J. Tompkin, D. Sun, S. Paris, and H. Pfister. Interactive intrinsic video editing. *ACM TOG*, pages 197:1–197:10, 2014. 2

[11] A. Bousseau, S. Paris, and F. Durand. User-assisted intrinsic images. *ACM TOG*, pages 130:1–130:10, 2009. 2

[12] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 5, 7

[13] Q. Chen and V. Koltun. A simple model for intrinsic image decomposition with depth cues. In *ICCV*, 2013. 2, 7

[14] L. Cheng, C. Zhang, and Z. Liao. Intrinsic image transformation via scale space decomposition. In *CVPR*, 2018. 3

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

[16] Q. Fan, J. Yang, G. Hua, B. Chen, and D. Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018. 1, 3, 6, 7, 8

[17] Elena Garces, Adolfo Munoz, Jorge Lopez-Moreno, and Diego Gutierrez. Intrinsic images by clustering. *Comput. Graph. Forum (Proceedings of the Eurographics Symposium on Rendering)*, 31(4), 2012. 8

[18] P. V. Gehler, C. Rother, M. Kiefel, L. Zhang, and B. Schölkopf. Recovering intrinsic images with a global sparsity prior on reflectance. In *NeurIPS*, 2011. 1, 2, 7

[19] T. Gevers and A. Smeulders. Color-based object recognition. *PR*, pages 453–464, 1999. 1, 2

[20] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *ICCV*, 2009. 5, 7, 8

[21] Paul Henderson and Vittorio Ferrari. Learning single-image 3d reconstruction by generative modelling of shape, pose and shading. *International Journal of Computer Vision*, 2019. 1

[22] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. B. Tenenbaum. Self-supervised intrinsic image decomposition. In *NeurIPS*, 2017. 2

[23] J. Jeon, S. Cho, X. Tong, and S. Lee. Intrinsic image decomposition using structure-texture separation and surface normals. In *ECCV*, 2014. 2

[24] E. H. Land and J. J. McCann. Lightness and retinex theory. *J. of Optical Society of America*, pages 1–11, 1971. 1, 2

[25] Kyong Joon Lee, Qi Zhao, Xin Tong, Minmin Gong, Shahram Izadi, Sang Uk Lee, Ping Tan, and Stephen Lin. Estimation of intrinsic image sequences from image+depth video. In Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid, editors, *ECCV*, pages 327–340, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 2, 7

[26] L. Lettry, K. Vanhoey, and L. van Gool. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. In *Int. Pacific Conf. on Comput. Graph. and App.*, 2018. 3

[27] Zhengqin Li, Mohammad Shafiei, R. Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. *CVPR*, pages 2472–2481, 2020. 3, 7, 8

[28] Z. Li and N. Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018. 1, 3, 7, 8

[29] Z. Li and N. Snavely. Learning intrinsic image decomposition from watching the world. In *CVPR*, 2018. 6, 7

[30] Y. Liu, Y. Li, Shaodi You, and Feng Lu. Unsupervised learning for intrinsic image decomposition from a single image. *CVPR*, pages 3245–3254, 2020. 3, 6, 7

[31] Y. Ma, X. Jiang, Z. Xia, M. Gabbouj, and X. Feng. Casqnet: Intrinsic image decomposition based on cascaded quotient network. *IEEE TCSVT*, pages 1–1, 2020. 7

[32] Y. Matsushita, K. Nishino, K. Ikeuchi, and M. Sakauchi. Illumination normalization with time-dependent intrinsic images for video surveillance. *IEEE TPAMI*, pages 1336–1347, 2004. 2

[33] A. Meka, M. Zollhöfer, C. Richardt, and C. Theobalt. Live intrinsic video. *ACM TOG*, 2016. 1

[34] T. Narihira, M. Maire, and S. X. Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 1, 2, 5, 7, 8

[35] T. Narihira, M. Maire, and S. X. Yu. Learning lightness from human judgement on relative reflectance. In *CVPR*, pages 2965–2973, June 2015. 2

[36] Thomas Nestmeyer and Peter V. Gehler. Reflectance adaptive filtering improves intrinsic image estimation. *CoRR*, abs/1612.05062, 2016. 8

[37] T. Sattler, R. Tylecek, T. Brox, M. Pollefeys, and R. B. Fisher. 3d reconstruction meets semantics - reconstruction challange 2017. In *Int. Conf. Comput. Vis. Workshop*, 2017. 5, 8

[38] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. *CoRR*, abs/1901.02453, 2019. 3, 8

[39] S. Shafer. Using color to separate reflection components. *Color Research and App.*, pages 210–218, 1985. 3

[40] L. Shen, P. Tan, and S. Lin. Intrinsic image decomposition with non-local texture cues. In *CVPR*, 2008. 2

[41] L. Shen and C. Yeo. Intrinsic images decomposition using a local and global sparse representation of reflectance. In *CVPR*, 2011. 2

[42] J. Shi, Y. Dong, H. Su, and S. X. Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 2017. 1, 2, 7

[43] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. *CoRR*, abs/1704.04131, 2017. 1

[44] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 5

[45] T. Wada, H. Ukida, and T. Matsuyama. Shape from shading with interreflections under proximal light source-3d shape reconstruction of unfolded book surface from a scanner image. In *ICCV*, 1995. 1

[46] Y. Weiss. Deriving intrinsic images from image sequences. In *ICCV*, 2001. 1, 2

[47] Saining "Xie and Zhuowen" Tu. Holistically-nested edge detection. In *ICCV*, 2015. 4

[48] Chen Xu, Yu Han, George Baciu, and Min Li. Fabric image recolorization based on intrinsic image decomposition. *Textile Research J.*, 89(17):3617–3631, 2019. 1

[49] J. Xu, Y. Hou, D. Ren, L. Liu, F. Zhu, M. Yu, H. Wang, and L. Shao. Star: A structure and texture aware retinex model. *IEEE TIP*, pages 5022–5037, 2020. 3, 7

[50] G. Ye, E. Garces, Y. Liu, Q. Dai, and D. Gutierrez. Intrinsic video and applications. *ACM TOG*, 2014. 1

[51] Y. Yu and W. A. P. Smith. Inverserendernet: Learning single image inverse rendering. In *CVPR*, 2019. 3, 6, 7

[52] Y. Yuan, B. Sheng, P. Li, L. Bi, J. Kim, and E. Wu. Deep intrinsic image decomposition using joint parallel learning. In *Comput. Graph. Int. Conf.*, 2019. 7

[53] Q. Zhao, P. Tan, Q. Dai, L. Shen, E. Wu, and S. Lin. A closed-form solution to retinex with nonlocal texture constraints. *IEEE TPAMI*, 34(7):1437–1444, July 2012. 8

[54] Hao Zhou, Xiang Yu, and David W. Jacobs. Glosh: Global-local spherical harmonics for intrinsic image decomposition. In *ICCV*, October 2019. 2, 7, 8

[55] Tinghui Zhou, Philipp Krähenbühl, and Alexei A. Efros. Learning data-driven reflectance priors for intrinsic image decomposition. *CoRR*, abs/1510.02413, 2015. 7