SIGNet: Intrinsic Image Decomposition by a Semantic and Invariant Gradient Driven Network for Indoor Scenes

Partha Das[1,3](✉), Sezer Karaoğlu[1,3], Arjan Gijsenij[2], and Theo Gevers[1,3]

[1] CV Lab, University of Amsterdam, Amsterdam, The Netherlands
{p.das,th.gevers}@uva.nl
[2] AkzoNobel, Amsterdam, The Netherlands
arjan.gijsenij@akzonobel.com
[3] 3DUniversum, Amsterdam, The Netherlands
s.karaoglu@3duniversum.com

Abstract. Intrinsic image decomposition (IID) is an under-constrained problem. Therefore, traditional approaches use hand crafted priors to constrain the problem. However, these constraints are limited when coping with complex scenes. Deep learning-based approaches learn these constraints implicitly through the data, but they often suffer from dataset biases (due to not being able to include all possible imaging conditions).

In this paper, a combination of the two is proposed. Component specific priors like semantics and invariant features are exploited to obtain semantically and physically plausible reflectance transitions. These transitions are used to steer a progressive CNN with implicit homogeneity constraints to decompose reflectance and shading maps.

An ablation study is conducted showing that the use of the proposed priors and progressive CNN increase the IID performance. State of the art performance on both our proposed dataset and the standard real-world IIW dataset shows the effectiveness of the proposed method. Code is made available here.

Keywords: Priors · Semantic segmentation · Intrinsic image decomposition · CNN · Indoor dataset

1 Introduction

An image can be defined as the combination of an object's colour and the incident light on it projected on a plane. Inverting the process of image formation is useful for many downstream computer vision tasks such as geometry estimation [18], relighting [34], colour edits [5] and Augmented Reality (AR) insertion and interactions for applications like the Metaverse. The process of recovering the object colour (reflectance or albedo) and the incident light (shading) is known

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-25066-8_35.

L. Karlinsky et al. (Eds.): ECCV 2022 Workshops, LNCS 13803, pp. 605–620, 2023.
https://doi.org/10.1007/978-3-031-25066-8_35

as Intrinsic Image Decomposition (IID). As the problem is ill-defined (with only one known), constraint-based approaches are explored to limit the solution space. For example, as an explicit gradient assumption, softer (or smoother) gradient transitions are attributed to shading transitions, while stronger (or abrupt) ones are related to reflectance transitions [21]. Colour palette constraints in the form of sparsity priors and piece-wise consistency are also employed for reflectance estimation [1,15]. However, these approaches are based on strong assumptions of the imaging process and hence are limited in their applicability.

Implicit constraints, by means of deep learning-based methods, are proposed to expand previous approaches [26]. For these methods, the losses implicitly formulate the constraints and are dependent on the training data. These methods learn a flexible representation based on training data which may lead to dataset biases. [23] integrates multiple datasets to manage the dataset bias problem. However, introducing more datasets only acts as an expansion of the imaging distribution. Additionally, multiple purpose-built losses are needed to train the network. An alternative approach of combining constraints and deep learning is explored in [12] where edges are used as an additional constraint to guide the network. However, edges at image locations with strong illumination effects, like pronounced cast shadows, may lead to edge misclassification resulting in undesirable effects like shading-reflectance leakages.

On the other hand, [2] forgoes priors and specialised losses to leverage joint learning of related modalities. They explore semantic segmentation as a closely related task to IID, arguing that jointly learning the semantic maps provides the network information to jointly correct for reflectance-shading transitions. However, no explicit guidance or constraint between the semantics and reflectance

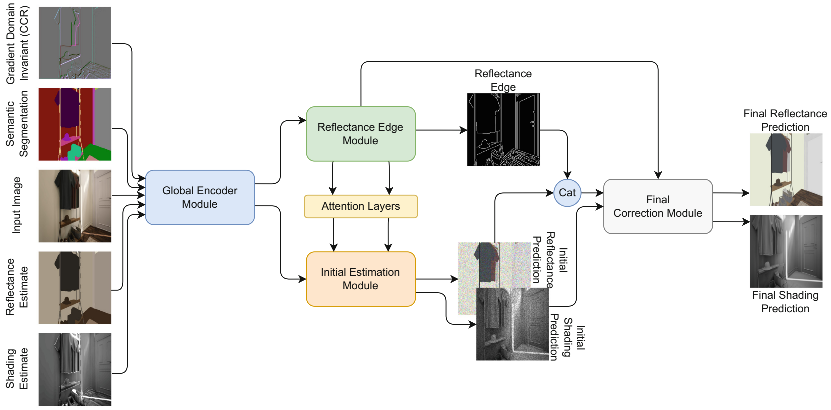

Fig. 1. The proposed network overviews. The network consists of i) the global encoder module, ii) the reflectance edge module, iii) the initial estimation module, and iv) the final correction module. The final reflectance and shading outputs are used for all the evaluations. Please refer to the supplementary for more details. Images shown here are ground truth images, for illustrative purposes.

are imposed. The network thus relies on learning the constraints from the ground truth semantic, reflectance and shading jointly. Moreover, only outdoor gardens are considered, where most natural classes (e.g., bushes, trees, and roses) contain similar colours (i.e., constrained colour distributions).

This paper exploits physical and statistical image properties for IID of indoor scenes. Illumination and geometry invariant descriptors [16] yield physics-based cues to detect reflectance transitions, while statistical grouping of pixels in an image provides initial starting estimates for IID components. To this end, a combination of semantic and invariant transition constraints is proposed. Semantic transitions provide valuable information about reflectance transitions i.e., a change in semantics most likely matches a reflectance transition but not always the other way around (objects may consist of different colours). Illumination invariant gradients provide useful information about reflectance transitions but can be unstable (noisy) due to low intensity. Exploiting reflectance transition information on these two levels compensates each other and ensures a stronger guidance for IID. In addition, indoor structures, like walls and ceilings, are often homogeneously coloured. To this end, the semantic map can be used as an explicit homogeneous prior. This allows for integrating an explicit sparsity/piecewise consistency (homogeneity) prior in the form of constant reflectance colour.

In this paper, a progressive CNN is employed, consisting of two stages. The first stage of the network exploits the prior information to arrive at an initial estimation. This estimation is based on the semantics, the invariant guided boundaries, and sparsity constraints. The second stage of the network takes the initial estimation and fine-tunes it using the original image cues to disentangle the reflectance and shading maps while being semantically correct. This allows the network to separate the problem into two distinct solution spaces that build progressively on each other. In addition, it allows the network to learn a continuous representation that can extrapolate even when the priors contain errors. An overview of the proposed network is shown in the Fig. 1.

While deep learning networks have shown very good performance, they require high quality datasets. Traditional physical-based rendering methods are often time and resource intensive. Recently, these methods are more efficient i.e., real time on consumer hardware. Hence, a dataset of physical-based and photo-realistic rendered indoor images is provided. The synthetic dataset is used to train the proposed method.

In summary, our contributions are as follows:

- **Algorithm**: An end-to-end semantic and physically invariant edge transition driven hybrid network is proposed for intrinsic image decomposition of indoor scenes.
- **Insight**: The use of component specific priors outperforms learning from a single image.
- **Performance**: The proposed algorithm is able to achieve state-of-the-art performance on both synthetic and real-world datasets.
- **Dataset**: A new ray-traced and photo-realistic indoor dataset is provided.

2 Related Works

A considerable amount of effort has been put in exploring hand-crafted prior constraints for the problem of IID. [21] pioneered the field by assuming reflectance changes to be related to sharper gradient changes, while smoother gradients correspond to shading changes. Other priors have been explored like piece-wise constancy for the reflectance, and smoothness priors for shading [1], textures [15]. Constraints in the form of additional inputs have also been explored. [22] explores the use of depth as an additional input, while [19] explores surface normals. Near infrared priors are used by [10] to decompose non-local intrinsics. Humans in the loop is also studied by [8] and [27]. However, these works mostly focus on single objects and do not generalise well to complete scenes.

In contrast to the use of explicit (hand-crafted) constraints, deep learning methods that implicitly learn (data-driven) specific constraints are also explored [26]. [3] explores disentangling the shading component into direct and indirect shading. [39] differentiates shading into illumination and surface normals in addition to reflectance. [6] uses a piece-wise constancy property of reflectances and employs Conditional Random Fields to perform IID. [12] shows that image edges contain information about reflectance edges and uses them as a guidance for the IID problem. [23] reduces the solution space by using multiple task specific losses. [31] directly learns the inverse of the rendering function. Finally, [2] forgoes losses and jointly learns semantic segmentation to implicitly learn a posterior on the IID, while [30] uses estimated semantic features as a support for an iterative competing formulation for IID. However, the above approaches do not explicitly integrate the physics-based image formation information and rely on the datasets containing a large set of imaging conditions. Hence, they may fall short for images containing extreme imaging conditions such as strong shadows or reflectance transitions. Large datasets [23,25,29] are proposed to train networks. Unfortunately, they are limited in their photo-realistic appearance.

Unlike IID, physics-based image formation priors have been explored in other tasks. [13] introduces Colour Ratios which are illumination invariant descriptors for objects. [16] then introduces Cross Colour Ratios which are both geometric and illumination invariant reflectance descriptors. [4] shows the applicability of the descriptors to the problem of IID. In contrast to previous methods, in this paper, a combination of explicit image formation-based priors and implicit intrinsic component property losses are explored.

3 Methodology

3.1 Priors

Semantic Segmentation: [2] shows that semantic segmentation provides useful information for the IID problem. However, components are jointly learned and hence their method lacks any explicit influence of the component's property. Since object boundaries correspond to reflectance changes such boundary information can serve as a useful global reflectance transition guidance for the network. Furthermore, homogeneous colour (i.e., reflectance) constraints (e.g., a wall has a uniform colour) can be imposed on the segmentation explicitly. To this end, in this paper, an off-the-self segmentation algorithm Mask2Former [9] is used to obtain segmentation maps.

Invariant Gradient Domain: Solely using semantic regions as priors may cause the network to be biased to the regions generated by the segmentation method. To prevent such a bias, an invariant (edge) map is included as an additional prior to the network. In this work, Cross Colour Ratios (CCR) [16] are employed. These are illumination invariants i.e., reflectance descriptors. Given an image I with channels Red (R), Green (G) and Blue (B) and neighbouring pixels p_1 and p_2, CCR is defined by $M_{RG} = \frac{R_{p_1} G_{p_2}}{R_{p_2} G_{p_1}}$, $M_{RB} = \frac{R_{p_1} B_{p_2}}{R_{p_2} B_{p_1}}$ and $M_{GB} = \frac{G_{p_1} B_{p_2}}{G_{p_2} B_{p_1}}$ where, R_{p_1}, G_{p_1} and B_{p_1} are the red, green, and blue channel for pixel p_1. Descriptors M_{RG}, M_{RB} and M_{GB} are illumination free and therefore solely depending on reflectance transitions. Using the reflectance gradient as an additional prior allows the network to be steered by reflectance transitions.

Reflectance and Shading Estimates: Consider the simplified Lambertian [32] image formation model: $I = R \times S$, where shading (S) is the scaling term on the reflectance component (R). Hence, for a given constant reflectance region, all the pixels are different shades of the same colour. In this way, the reflectance colour becomes a scale optimisation for which the pixel mean of a segment can be used: $\mathcal{M}_c = \sum_{n=1}^{N} I_n^c$ where, $\mathcal{M}_c$ is the channel-specific mean of the pixels. $\mathcal{M}_R$, $\mathcal{M}_G$ and $\mathcal{M}_B$ values are then spread within the region to obtain an initial starting point for reflectance colour based on the homogeneity constraint. Conversely, these values can be inverted using the image formation to obtain the corresponding scaled shading estimates. A CNN is then employed to implicitly learn the scaling for both priors. Additionally, since the mean of the segment does not consider textures, a deep learning method is proposed to compensate it by means of a dedicated correction module, see Sect. 3.2. The supplementary material provides more visuals for these priors.

3.2 Network Architecture

The network consists of 4 components: i) Global encoder blocks, ii) Reflectance edge Decoder, iii) Initial estimation decoder and iv) Final correction module. The network is trained end-to-end. The input to the network is an image and its corresponding segmentation obtained by Mask2Former [9]. The CCR, Reflectance and Shading estimates are computed from the input image for the respective encoder blocks. Additional details and visuals for the modules can be found in the supplementary materials.

Global Encoder Module: The input image, the segmentation image, the average reflectance estimate, inverse shading estimate and the CCR images are encoded through their respective encoders. The encoders share the same configuration, but the intermediate features are independent of each other. The semantic features ($\mathcal{F}_S$) provide guidance for the general outlines of object boundaries, while the CCR features ($\mathcal{F}_G$) focus on local reflectance transitions, possibly

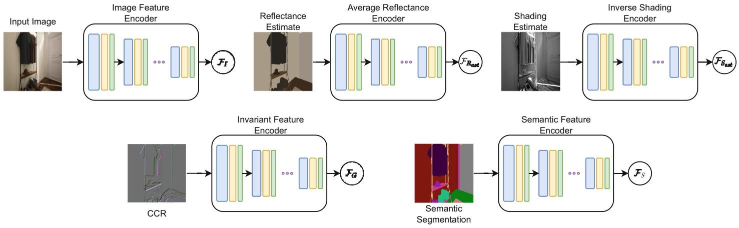

Fig. 2. Overview of the global encoder module. Each of the inputs are provided with their independent encoders to enable modality specific feature learning. The respective features are used in the downstream decoders to provide component specific information for the network.

including textures. Correspondingly, the average reflectance estimate features ($\mathcal{F}_{R_{est}}$) and the inverse shading estimate features ($\mathcal{F}_{S_{est}}$) provide a starting point for the reflectance and the shading estimation, respectively. Finally, the image features ($\mathcal{F}_I$) provide the network a common conditioning to learn the scaling and boundary transitions for the intrinsic components. Figure 2 shows the overview of the module.

Reflectance Edge Module: This sub-network decodes the reflectance edges of the given input. The decoded reflectance and edges are used as an attention mechanism to the initial estimation module to provide (global) region consistency. The features, $\mathcal{F}_S$ and $\mathcal{F}_G$ are concatenated with the image features $\mathcal{F}_I$ and passed on to the edge decoder. The semantic and CCR features provide object and reflectance transitions, respectively. The image features allow the network to disentangle reflectance from illumination edges. Corresponding skip connections from $\mathcal{F}_I$, $\mathcal{F}_{R_{est}}$ and $\mathcal{F}_G$ encoders are used to generate high frequency details. Scale space supervision, following [36], is provided by a common deconvolution layer for the last 2 layers, for scales of 64×64 and 128×128, yielding a scale consistent reflectance edge prediction. The ground truth edges are calculated by using a Canny Edge operation on the ground truth reflectance. Figure 3 shows an overview of the module.

Initial Estimation Module: The initial estimation decoder block focuses on learning the IID from the respective initial estimates of the intrinsic (Fig. 3). It consists of two parallel decoders. The Reflectance decoder learns to predict the first estimation from $\mathcal{F}_I$ and $\mathcal{F}_{R_{est}}$. The features are further augmented with the learned boundaries from the reflectance edge decoder passed through an attention layer [35]. $\mathcal{F}_S$ is also passed to the decoder to guide global object transitions and acts as an additional attention. Similarly, the Shading decoder only receives $\mathcal{F}_I$ and $\mathcal{F}_{S_{est}}$, focusing on properties like smoother (shading) gradient changes. The reflectance and shading decoders are interconnected to provide an additional cue to learn an inverse of each other. Skip connections from the respective encoders to the decoders are also given. This allows the network to learn an implicit scaling on top of the average reflectance and the inverse shading estimation. The output at this stage is guided by transition and reflectance boundaries and may suffer from local inconsistencies like shading-reflectance leakages.

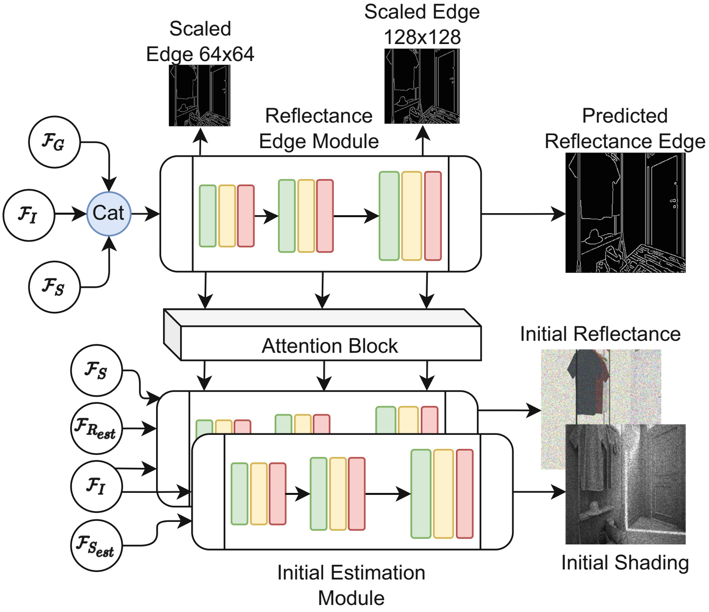

Fig. 3. Overview of the reflectance edge and the attention guided initial estimation module. The edge module takes the image encoder, semantic encoder, and the invariant encoder feature to learn a semantically and physically guided reflectance transition. The edge features are then transferred through an attention block to the initial estimation decoder module. The reflectance decoder in this module takes the semantic encoder, image encoder and the average reflectance estimation features and input. The shading decoder correspondingly takes the image encoder along with the average shading estimation feature. Interconnections in the decoder allows the network to use reflectance cues for shading and vice versa.

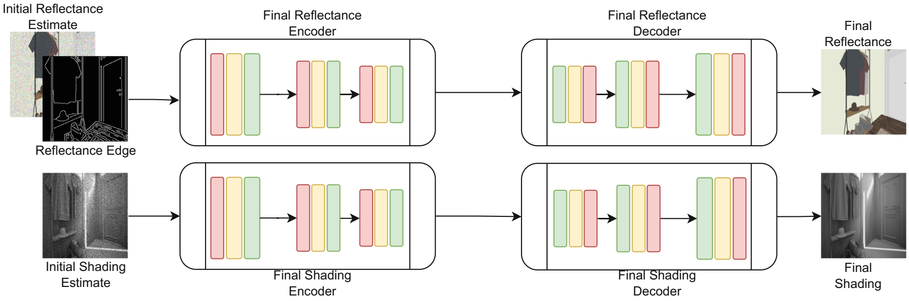

Fig. 4. The final decoder module. The initial reflectance and shading estimates from the previous step are further corrected to obtain the final reflectance and shading. The encoder consists of an independent parallel reflectance and shading encoder. The reflectance encoder takes receives the initial reflectance and the reflectance edge as an input, while the shading encoder receives the initial shading. Two parallel decoders are used for reflectance and shading to obtain the final IID outputs.

Final Correction Module: To deal with local inconsistencies, a final correction module is proposed. First, the reflectance edge from the edge decoder and the reflectance from the previous decoder is concatenated and passed through a feature calibration layer. This allows the network to focus on local inconsistencies guided by global boundaries. The output is then passed through a final reflectance encoder. The shading from the previous module is similarly passed through another encoder block. The output of these two encoders is then passed through another set of parallel decoders for the final reflectance and shading output. Since the reflectance and shading from the previous block is already globally consistent, this decoder acts as a localised correction. To constrain the corrections to local homogeneous regions, skip connections (through attention layers) of encoded reflectance edge features are provided to the decoders. In this way, the network limits the corrections to the local homogeneous regions and recover local structures like textures. Skip connections from the respective reflectance and shading encoders are provided to include high frequency information transfer. The reflectance and shading features in the decoder are shared within each other to enforce an implicit image formation model. Figure 4 shows the overview of the module.

3.3 Dataset

Unreal Engine [11] is used to generate a dataset suited for the task. The rendering engine supports physically based rendering, with real-time raytracing (RTX) support. The engine first calculates the intrinsic components from the various material and geometry property of the objects making up the scene. Then, the illumination is physically simulated through ray tracing and lighting is calculated. Finally, all these results are combined to render the final image. Since the engine calculates the intrinsic components, ground truth intrinsic is recovered using the respective buffer. The dataset consists of dense reflectance and shading ground-truths. The network learns the inversion of this process.

Assets from the unreal marketplace are used to generate the dataset. These assets are professionally created to be photo realistic. 5000 images are generated of which 4000 images are used for training, and 1000 are used for validation and testing. To evaluate the generalisability of the network, Intrinsic Images in the Wild (IIW) [6] is used as a real-world test. Figure 5 shows a number of samples from the dataset. The dataset generated is comparatively small. However, the purpose of the dataset is that the network learns an efficient physics guided representation, rather than a dataset dependent one. The pretrained model and the dataset will be made available.

3.4 Loss Functions and Training Details

MSE loss is applied for each output of the network: (i) Initial estimation loss ($\mathcal{L}_e$ & $\mathcal{L}_i$) and (ii) Final correction loss ($\mathcal{L}_f$). $\mathcal{L}_e$ is the loss applied on the scale space reflectance edge. $\mathcal{L}_i$ is the loss on the reflectance and shading output from the initial estimation module. Additional losses are also applied on the reflectance and

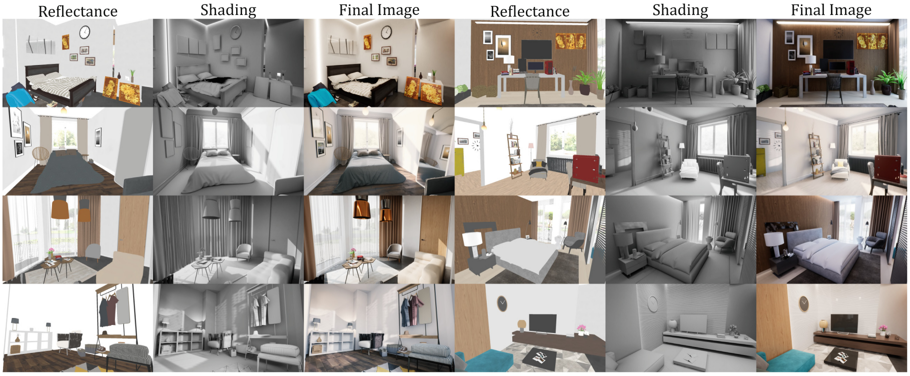

Fig. 5. Samples from the proposed dataset. The dataset comes with the corresponding dense reflectance and shading maps. The dataset consists of various everyday objects and lighting, containing both near local light sources, like lamps, and more global light sources like sunlight and windows.

shading output from the final correction module. This reflectance and shading are also combined and compared with the input image for a reconstruction loss. These 3 losses are collected in the term $\mathcal{L}_f$. An invariance loss $\mathcal{L}_{Norm}$ is added between the normalised RGB and the prediction of the network for each segment. A Total Variation (TV) loss ($\mathcal{L}_{TV}$) is included to deal with the assumption that large indoor classes like walls and ceilings are homogeneously coloured. This loss is only applied to ceilings and wall pixels and minimises the TV between the prediction and the ground truth reflectance. Finally, to encourage perceptually consistent and sharper textures, a perceptual and dssim loss are included and grouped as $\mathcal{L}_\delta$. The final loss term to minimise for the network thus becomes:

$$\begin{aligned} \mathcal{L} = \lambda_e \, \mathcal{L}_e + \lambda_i \, \mathcal{L}_i + \mathcal{L}_f \\ + \mathcal{L}_{Norm} + \mathcal{L}_{TV} + \mathcal{L}_\delta \end{aligned} \tag{1}$$

where λ_e and λ_i are weighting terms for the edge and initial estimation losses. They are empirically set to 0.4 and 0.5, respectively. The network is trained for 60 epochs, with a learning rate of $2e - 4$ and the Adam [20] optimiser. Please refer to the supplementary materials for more details.

4 Experiments

4.1 Ablation Study

To study the influence of different architecture components and losses, an ablation study is conducted. For a fair evaluation, the ablation study is performed on the test-set of the rendered dataset. For all the ablations, all hyper-parameters are kept constant. The results of the ablation study are presented in Table 1.

Table 1. Ablation study for the proposed network. For each experiment, the respective parts of the network are modified. All the experiments are conducted on the same test and train split of the proposed dataset. All the applicable hyper-parameters are kept constant.

	Reflectance			Shading		
	MSE	LMSE	DSSIM	MSE	LMSE	DSSIM
w/o final correction	0.0029	0.0020	0.0225	0.0044	0.0035	0.0276
w/o priors	0.0105	0.0047	0.0444	0.0054	0.0034	0.0399
w canny edges	0.0032	0.0037	0.0229	0.0031	0.0049	0.0293
w/o average estimates	0.0030	0.0023	0.0232	0.0041	0.0043	0.0267
w/o reflectance edge module	0.0097	0.0156	0.3254	0.0033	0.0061	0.0270
No DSSIM loss	0.0131	0.0240	0.3704	0.0041	0.0055	0.1488
No perceptual loss	0.0032	0.0022	0.0289	0.0032	0.0038	0.0285
No invariant & homogeneity loss	0.0032	0.0027	0.0288	**0.0024**	**0.0024**	0.0318
Proposed	**0.0026**	**0.0018**	**0.0219**	0.0030	0.0033	**0.0252**

Influence of Final Correction Module: In this experiment, the influence of the final correction module is studied. The output from the initial estimation decoder is taken as the final output.

From the results, it is shown that the final correction module helps in improving the outputs. The improvement in the DSSIM metric for both components shows that the final correction module is able to deal with structural artefacts.

Influence of Priors: The influence of all the priors is studied in this ablation. The additional priors are removed, and the network only receives the image as an input. All network structures are kept the same. This setup studies if the network can disentangle the additional information from the input image without any specific priors.

Removing all priors makes the network to perform worse for all metrics. In this setting, the network only uses the image to derive both the reflectance and shading changes. This is challenging for strong illumination effects. This shows that the priors are an important source of information enabling a better disentanglement between intrinsic components.

Influence of Specialised Edges: This experiment studies the need of specialised edges obtained from the semantic transition boundaries and invariant features. The edges obtained from the input image are provided to the network. The study focuses on whether the network can distinguish between reflectance, geometry, and shadow edges directly from the image.

From the results, it is shown that using image edges is not sufficient. Image edges can be ambiguous due to the presence of shadow edges. However, the performance is still better than using the image as the only input, showing that edges yield, to a certain extent, useful transition information.

Influence of Reflectance and Shading Estimate Priors: In this experiment, the efficacy of the statistic-based homogeneous reflectance and the inverted shading estimate is studied.

Removing the average reflectance and shading estimates degrades the performance. With the priors of the estimates, the network can use its learning capacity to deal with the scaling of the initial estimation to obtain the correct IID. The network needs to learn the colour as well as the scaling within the same learning capacity.

Influence of Reflectance Edge Guidance Module: For this experiment, the edge guidance module is removed. As such, the network is then forced to learn the attention and the reflectance transition boundaries implicitly as part of the solution space.

Removing the reflectance edge module results in the second worse result. This shows that, apart from the priors, the ability to use those features to learn a reflectance transition, is useful. It is shown that without such a transition guidance, the network is susceptible of misclassifying shadow edges as reflectance transitions. Furthermore, it is shown that without this module, the reflectance performance suffers more than the shading performance. Hence, using a learned edge guidance allows the network to be more flexible and better able to distinguish between true reflectance transitions.

Influence of Different Losses: The influence of the different losses is studied in this experiment. For each sub-experiment, the same proposed structure is used, and the respective losses are selectively turned off.

From the results, it is shown that the DSSIM loss contributes to a large extend, to the performance, because this loss penalises perceptual variations like contrast, luminance, and structure. As such, by removing the supervision, the network learns an absolute difference which is not expressive to smaller spatial changes. Similar trend of performance decrease is shown when removing the perceptual and homogeneity losses. This is expected since both losses contribute to region consistency. With the addition of the losses on the reflectance, the shading values suffer slightly. However, structurally they perform better when including the losses, as shown by the DSSIM metric. This indicates that applying such a loss helps not only to achieve a better reflectance value, but it also jointly improves shading, resulting in sharper outputs.

4.2 Comparison to State of the Art

On the Proposed Dataset: To study the influence of the dataset, the proposed network is compared to baseline algorithm's performance. For these experiments, the standard, MSE, LMSE and the DSSIM metric are used. The baselines are chosen based on their performance of the Weighted Human Disagreement Rate (WHDR), widely used in the literature. Hence, [23] is chosen as a baseline. [39] does not provide any publicly available code, hence is not included. Although [12] is the state of the art, their provided code generates errors when trying to run on custom datasets and hence is not used for comparison. For completeness, [37] and [33] is also compared. [37] uses an optimization-based method based on the pioneering Retinex model. Since it is a purely physical constraint-based model, it is included for comparison. For a fair comparison, methods focusing on indoors are used. [2] assumes outdoor settings and requires semantic ground truths to train and hence is not included. For all the networks, they are retrained on the dataset that is proposed in this paper, using the optimum hyperparamters as mentioned in the respective publication. The results are shown in Table 2 and Fig. 6

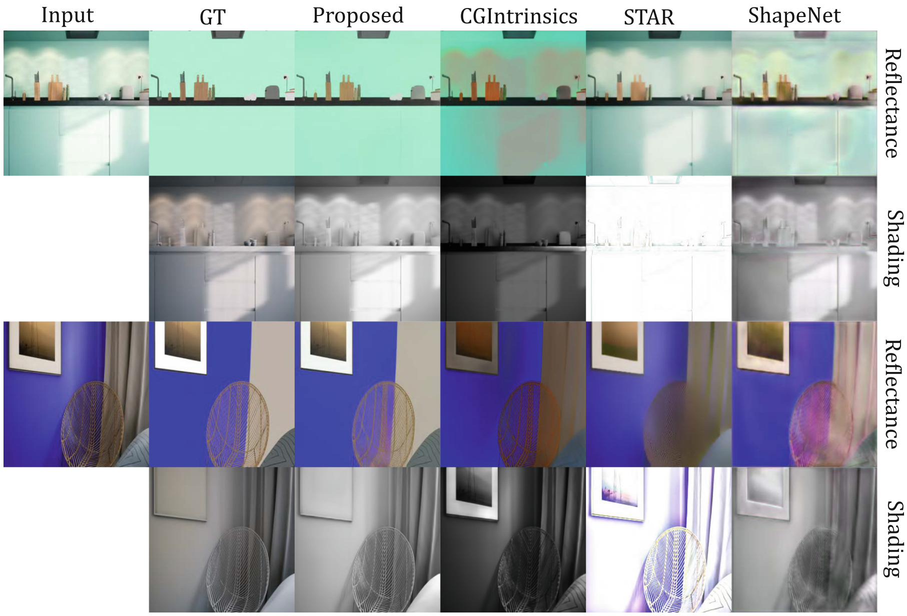

Fig. 6. Comparison of the proposed to baseline methods. It is shown that the proposed method is able to better disentangle the illumination effect. In comparison, CGIntrinsics, which has comparable results on the WHDR SoTA, suffers from discolouration. STAR misses the illumination while ShapeNet suffers from artefacts.

From the table it is shown that our proposed model is able to provide the highest scores. From the figure, the baselines suffer from strong illumination effects. CGIntrinsics discolours the regions while STAR mostly fails. ShapeNet, suffers from artefacts and colour variations around the illumination regions. In comparison, the proposed network is able to recover from such effects.

On IIW [6]**:** The proposed network is finetuned on the IIW dataset and compared to the baselines. The training and testing splits are used as specified in the original publication. For the baselines, the numbers are obtained from the respective original publications. The results are shown in Table 3 and visuals in Fig. 7.

Table 2. Comparison to the baseline methods on the proposed dataset. It is shown that the proposed method outperforms all other methods.

	Reflectance			Shading		
	MSE	LMSE	DSSIM	MSE	LMSE	DSSIM
ShapeNet [33]	0.0084	0.0133	0.1052	0.0065	0.0129	0.1862
STAR [37]	0.0304	0.0166	0.1180	0.0290	0.0128	0.1572
CGIntrinsics [23]	0.0211	0.0156	0.0976	0.0848	0.0577	0.2180
Proposed	**0.0026**	**0.0018**	**0.0219**	**0.0030**	**0.0033**	**0.0252**

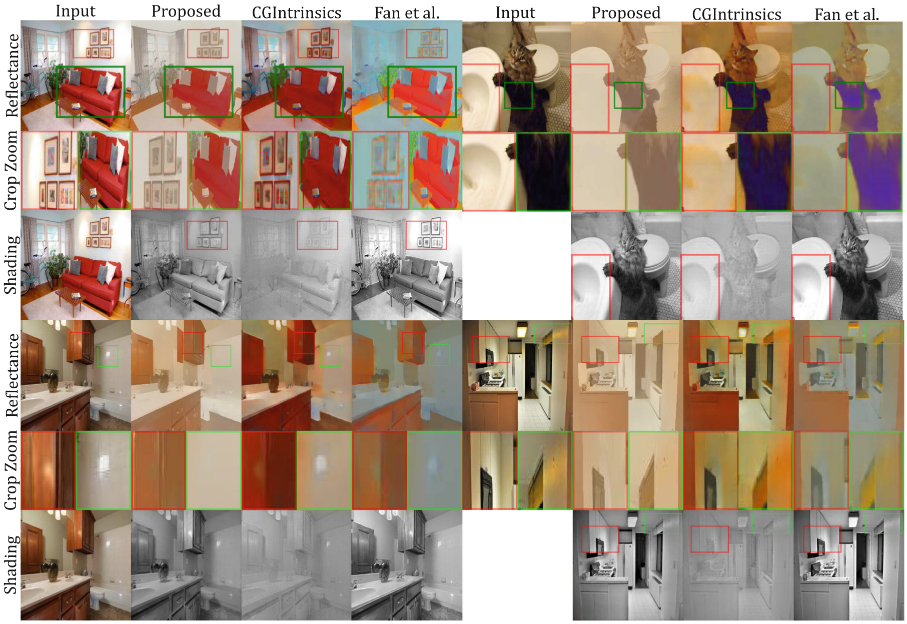

Fig. 7. Visual results on the IIW test set. Compared to CGIntrinsics [23] and Fan *et al.* [12], the proposed method disentangles better the shading and highlights (highlighted in red boxes), showing a smoother reflectance. Both CGIntrinsics and [12] are unable to remove the highlights from the reflectance, resulting in discolouration. They are also susceptible to reflectance colour change as be seen on the cat and furniture (highlighted green boxes). The proposed method is able to better retain the original colour in the reflectance. (Color figure online)

Table 3. Baseline comparison for the IIW dataset. Results marked with * are post-processed with a guided filter [28]

Methods	WHDR (mean)
Direct intrinsics [26]	37.3
Color retinex [17]	26.9
Garces *et al.* [14]	25.5
Zhao *et al.* [38]	23.2
IIW [6]	20.6
Nestmeyer *et al.* [28]	19.5
Bi *et al.* [7]	17.7
Sengupta *et al.* [31]	16.7
Li *et al.* [24]	15.9
CGIntrinsics [23]	15.5
GLoSH [39]	15.2
Fan *et al.* [12]	15.4
Proposed	15.2
CGIntrinsics [23]*	14.8
GLoSH [39]*	14.6
Fan *et al.* [12]*	14.5
Proposed*	**13.9**

The IIW dataset does not contain dense ground truth and hence is only finetuned with the ordinal loss. A guided filter [28] is used to further improve the results. Overall, our proposed method is on par with GLoSH [39] which is the best performing method without any post filtering. However, they need both lighting and normal information as supervision, while the proposed method is trained with just reflectance and shading, along with a smaller dataset ($58,949$ images of [39] vs. 5000 of the proposed method). For the filtered results, the proposed method is able to achieve a comfortable lead compared to the current best of 14.5 obtained by [12], showing the efficiency of the current model.

5 Conclusions

In this paper, an end-to-end prior driven approach for indoor scenes has been proposed for the task of intrinsic image decomposition. Reflectance transitions and invariant illuminant descriptors has been used to guide the reflectance decomposition. Image statistics-based priors have been used to provide the network a starting point for learning. To integrate explicit homogeneous constraints, a progressive CNN was used. To train the network, a custom physically rendered dataset was proposed.

An extensive ablation was performed to validate the proposed network showing that: i) using explicit reflectance transition priors helps the network to achieve an improved intrinsic image decomposition, ii) image statistics-based priors are helpful for simplifying the problem and, iii) the proposed method attains sota performance for the standardised real-world dataset IIW.

References

1. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE TPAMI **37**, 1670–1687 (2015)
2. Baslamisli, A.S., Groenestege, T.T., Das, P., Le, H.A., Karaoglu, S., Gevers, T.: Joint learning of intrinsic images and semantic segmentation. In: ECCV (2018)
3. Baslamisli, A.S., Das, P., Le, H., Karaoglu, S., Gevers, T.: Shadingnet: image intrinsics by fine-grained shading decomposition. IJCV **129**, 2445–2473 (2021)
4. Baslamisli, A.S., Liu, Y., Karaoglu, S., Gevers, T.: Physics-based shading reconstruction for intrinsic image decomposition. Comput. Vis. Image Understanding, 1–14 (2020)
5. Beigpour, S., van de Weijer, J.: Object recoloring based on intrinsic image estimation. In: ICCV, pp. 327–334 (2011)
6. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM TOG **33**, 1–12 (2014)
7. Bi, S., Han, X., Yu, Y.: An l1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. ACM TOG **34**(4) (2015). https://doi.org/10.1145/2766946
8. Bonneel, N., Sunkavalli, K., Tompkin, J., Sun, D., Paris, S., Pfister, H.: Interactive intrinsic video editing. ACM TOG **33**, 197:1–197:10 (2014)
9. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. arXiv (2021)
10. Cheng, Z., Zheng, Y., You, S., Sato, I.: Non-local intrinsic decomposition with near-infrared priors. In: ICCV (2019)
11. Epic Games: Unreal engine. https://www.unrealengine.com
12. Fan, Q., Yang, J., Hua, G., Chen, B., Wipf, D.: Revisiting deep intrinsic image decompositions. In: CVPR (2018)
13. Finlayson, G.D.: Colour Object Recognition. Master's thesis, Simon Fraser University (1992)
14. Garces, E., Munoz, A., Lopez-Moreno, J., Gutierrez, D.: Intrinsic images by clustering. Comput. Graph. Forum **31**(4) (2012). https://www-sop.inria.fr/reves/Basilic/2012/GMLG12
15. Gehler, P.V., Rother, C., Kiefel, M., Zhang, L., Schölkopf, B.: Recovering intrinsic images with a global sparsity prior on reflectance. In: NeurIPS (2011)
16. Gevers, T., Smeulders, A.: Color-based object recognition. PR **32**, 453–464 (1999)
17. Grosse, R., Johnson, M.K., Adelson, E.H., Freeman, W.T.: Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: ICCV (2009)
18. Henderson, P., Ferrari, V.: Learning single-image 3D reconstruction by generative modelling of shape, pose and shading. Int. J. Comput. Vis. **128**, 835–854 (2019)
19. Jeon, J., Cho, S., Tong, X., Lee, S.: Intrinsic image decomposition using structure-texture separation and surface normals. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8695, pp. 218–233. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10584-0_15

20. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014). https://arxiv.org/abs/1412.6980, arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015
21. Land, E.H., McCann, J.J.: Lightness and retinex theory. J. Opt. Soc. Am. **61**, 1–11 (1971)
22. Lee, K.J., et al.: Estimation of intrinsic image sequences from image+depth video. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7577, pp. 327–340. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33783-3_24
23. Li, Z., Snavely, N.: Cgintrinsics: better intrinsic image decomposition through physically-based rendering. In: ECCV (2018)
24. Li, Z., Shafiei, M., Ramamoorthi, R., Sunkavalli, K., Chandraker, M.: Inverse rendering for complex indoor scenes: shape, spatially-varying lighting and svbrdf from a single image. In: CVPR, pp. 2472–2481 (2020)
25. Li, Z., et al.: Openrooms: an end-to-end open framework for photorealistic indoor scene datasets. CoRR abs/2007.12868 (2020). https://arxiv.org/abs/2007.12868
26. Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: learning albedo-shading decomposition by convolutional regression. In: ICCV (2015)
27. Narihira, T., Maire, M., Yu, S.X.: Learning lightness from human judgement on relative reflectance. In: CVPR, pp. 2965–2973 (2015). https://doi.org/10.1109/CVPR.2015.7298915
28. Nestmeyer, T., Gehler, P.V.: Reflectance adaptive filtering improves intrinsic image estimation. CoRR abs/1612.05062 (2016). https://arxiv.org/abs/1612.05062
29. Roberts, M., et al.: Hypersim: a photorealistic synthetic dataset for holistic indoor scene understanding. In: International Conference on Computer Vision (ICCV) 2021 (2021)
30. Saini, S., Narayanan, P.J.: Semantic hierarchical priors for intrinsic image decomposition. CoRR abs/1902.03830 (2019). https://arxiv.org/abs/1902.03830
31. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. CoRR abs/1901.02453 (2019). https://arxiv.org/abs/1901.02453
32. Shafer, S.: Using color to separate reflection components. Color Res. App. **10**, 210–218 (1985)
33. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: CVPR (2017)
34. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. CoRR abs/1704.04131 (2017). https://arxiv.org/abs/1704.04131
35. Tang, H., Qi, X., Xu, D., Torr, P.H.S., Sebe, N.: Edge guided gans with semantic preserving for semantic image synthesis. CoRR (2020)
36. Xie, S., Tu, Z.: Holistically-nested edge detection. In: ICCV (2015)
37. Xu, J., et al.: Star: a structure and texture aware retinex model. IEEE TIP **29**, 5022–5037 (2020)
38. Zhao, Q., Tan, P., Dai, Q., Shen, L., Wu, E., Lin, S.: A closed-form solution to retinex with nonlocal texture constraints. IEEE TPAMI **34**(7), 1437–1444 (2012). https://doi.org/10.1109/TPAMI.2012.77
39. Zhou, H., Yu, X., Jacobs, D.W.: Glosh: global-local spherical harmonics for intrinsic image decomposition. In: ICCV (2019)