Intrinsic Image Decomposition by Pursuing Reflectance Image

Tzu-Heng Lin[1*], **Pengxiao Wang**[1*] and **Yizhou Wang**[1,2]
[1] School of Computer Science, Peking University
[2] Center on Frontiers of Computing Studies, Peking University
lzhbrian@gmail.com, hbxxwpx@gmail.com, yizhou.wang@pku.edu.cn

Abstract

Intrinsic image decomposition is a fundamental problem for many computer vision applications. While recent deep learning based methods have achieved very promising results on the synthetic densely labeled datasets, the results on the real-world dataset are still far from human level performance. This is mostly because collecting dense supervision on a real-world dataset is impossible. Only a sparse set of pairwise judgement from human is often used. It's very difficult for models to learn in such settings.

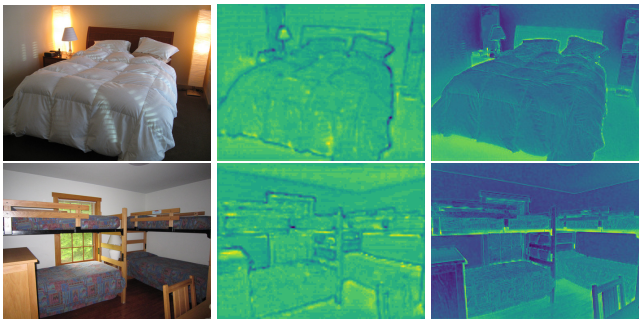
In this paper, we investigate the possibilities of only using reflectance images for supervision during training. In this way, the demand for labeled data is greatly reduced. In order to achieve this goal, we take a deep investigation into the reflectance images. We find that reflectance images are actually comprised of two components: the flat surfaces with low frequency information, and the boundaries with high frequency details. Then, we propose to disentangle the learning process of the two components of the reflectance images. We argue that through this procedure, the reflectance images can be better modeled, and in the meantime, the shading images, though not supervised, can also achieve decent result. Extensive experiments show that our proposed network outperforms current state-of-the-art results by a large margin on the most challenging real-world IIW dataset. We also surprisingly find that on the densely labeled datasets (MIT and MPI-Sintel), our network can also achieve state-of-the-art results on both reflectance and shading images, when we only apply supervision on the reflectance images during training.

1 Introduction

As one of the fundamentals for problems like surface retexturing and object compositing [Bi *et al.*, 2015], intrinsic image decomposition (IID) has been gaining great attention these years. IID aims at decomposing an image into two separate images: the reflectance image and the shading image. Following the lambertian reflection assumption, the image I is modelled as the pixel-wise product of reflectance (or albedo) R and shading S.

$$I \approx R \odot S. \tag{1}$$

While the above assumption is elegant, the acquisition of the groundtruth reflectance and shading images is extremely hard. As a result, MIT Intrinsic [Grosse *et al.*, 2009], MPI-Sintel [Butler *et al.*, 2012], ShapeNet [Shi *et al.*, 2017], CGIntrinsics [Li and Snavely, 2018a] propose to construct datasets with groundtruth reflectance and shading images through rendering objects or scenes with computer softwares. However, these datasets are synthetic and lack realism, thus the algorithms trained on them cannot generalize well to real-world images and limit its usefulness. Subsequently, Intrinsic Image in the Wild (IIW) [Bell *et al.*, 2014], and Shading Annotations in the Wild (SAW) [Kovacs *et al.*, 2017] dataset is proposed to collect supervision from the real-world images on reflectance and shading images, respectively. In these two datasets, humans are asked to make comparisons (darker color or shading) between two points on an images. And eventually, a sparse set of pairwise comparisons is collected



Input | Low res. branch attention mask | High res. branch attention mask

Figure 1: Visualization of the learned attention masks. Greener means larger value, bluer means smaller value. Our method better solves the problem of intrinsic image decomposition in the wild by disentangling the learning process of the two components: the flat surfaces and the boundaries in the reflectance images.

*Equal contribution.

for each dataset. Apparently, compared to the aforementioned synthetic datasets, this kind of data is extremely hard to collect, and thus limits the usage in the practice.

In this paper, we attempt to use only reflectance supervision for training. In this way, we only need to collect label for reflectance images (not both reflectance and shading images). In fact, in the literature, a number of works already tried to solve IID with only reflectance supervision on the IIW dataset [Liu *et al.*, 2019; Wang and Lu, 2019]. However, they usually cannot gain satisfying results. We argue that this is because they did not fully utilize the characteristics of reflectance images. Even with both supervision (reflectance and shading) [Li and Snavely, 2018a; Zhou *et al.*, 2019], the results are also far from satisfactory.

Learning a decomposition model with real-world images with only pairwise supervision on reflectance images is very challenging in the following two aspect. Firstly, it is hard for model to output a reasonable reflectance image with such sparse pairs of comparison groundtruth per image. Secondly, with no supervision on shading, we can only rely on the quality of output reflectance image and use Eqn.(1) to calculate the shading image.

To alleviate the above problems, we decided to pursue reflectance images to solve the problem of IID. We first take a deep investigation on the reflectance images. We observe that there are actually two components in a reflectance image: the flat surfaces with different colors, and the boundaries between these surfaces. The flat surfaces are usually an approximate constant color with low frequency information, and the boundaries are usually sparse yet full of high frequency details.

By leveraging these findings, we then propose **RDNet** (short for **R**eflectance **D**isentanglement **Net**work) to disentangle the learning process of the two components. We argue that through this design, the process of IID can be better modelled. Inspired by OctConv [Chen *et al.*, 2019], we first use a *componential branch separator* module to separate the networks into two branches: a *low resolution branch*, and a *high resolution branch*. We use the *low resolution branch* to capture the flat surfaces for learning the low frequency information, while the *high resolution branch* is used to capture the boundaries for learning the high frequency details. We make this design to reduce the spatial redundancy in extracting low frequency information in the low resolution branch, and to maintain the high frequency details in the high resolution branch. This disentangling process can let the networks focus on more specific tasks and thus gain better performance.

To further make the disentanglement more effective, we propose an *attentional branch merger* module. Two attention masks are learned to attend the spatial location of high or low resolution branches. This makes sure that each branch is actually focusing on learning the component that it is assigned. Eventually, an *attention mask regularization* term on the attention masks are proposed to limit the sparsity for the mask of the *high resolution branch*, and the smoothness for the mask of the *low resolution branch*. This in turn also force each branch to focus on its assigned components. In addition, our model also gains high interpretability through visualization of the attention masks, which have demonstrated the procedure of disentanglement and utilization (*cf.* Figure 1). In summary, our contribution is fourfold:

- We propose RDNet[‡], to our knowledge, the first attempt to disentangle the learning process of the two components in the reflectance images: the flat surfaces, and the boundaries between these surfaces. Also, RDNet only requires reflectance supervision during training.
- Through carefully designed architecture and visualization, RDNet is equipped with high interpretability by demonstrating the procedure of disentanglement and utilization.
- Extensive experiments on the challenging real-world IIW dataset show that RDNet outperforms all current state-of-the-art solutions by a large margin.
- On the densely labeled datasets (MIT and MPI-Sintel), RDNet can also achieve state-of-the-art results on both reflectance and shading images, when only applying supervision on the reflectance image during training.

2 Related Work

Intrinsic image decomposition. The methods of IID can generally be classified into two categories, *optimization based* and *learning based* methods. *Optimization based* methods aim at designing effective priors and attempt to solve this problem by optimizing some energy function for each image [Grosse *et al.*, 2009; Barron and Malik, 2014; Chen and Koltun, 2013; Shen and Yeo, 2011; Zhao *et al.*, 2012; Bi *et al.*, 2015]. *Learning based* methods are becoming the mainstream now with the rapid development of convolutional neural networks [He *et al.*, 2016; Chen *et al.*, 2019; Sun *et al.*, 2019]. They make use of the collected data with supervision, and let the networks learn how to decompose by themselves without the need of human tailored filters and functions [Cheng *et al.*, 2018; Fan *et al.*, 2018; Lettry *et al.*, 2018; Liu *et al.*, 2019; Narihira *et al.*, 2015a; Narihira *et al.*, 2015b; Nestmeyer and Gehler, 2017; Shi *et al.*, 2017; Wang and Lu, 2019; Zhou *et al.*, 2015; Zoran *et al.*, 2015]. More recently, [Zhang *et al.*, 2021; Liu and Lu, 2020; Liu *et al.*, 2020] utilized unsupervised learning in the field of IID.

Reflectance image learning. To solve the IID problem on IIW datatset [Bell *et al.*, 2014], many learning based methods have been proposed. Since the only supervision on the IIW dataset is the pairwise reflectance judgement on reflectance images, most works focus on how to learn and output 'good' reflectance images. A group of works [Narihira *et al.*, 2015b; Zhou *et al.*, 2015; Zoran *et al.*, 2015] start by training a classifier using deep features of the two image patches, then use a globalization method to recover the full reflectance images. [Nestmeyer and Gehler, 2017] develop a CNN approach to output the dense reflectance images directly with the help of a hinge loss function. [Fan *et al.*, 2018] propose to improve [Nestmeyer and Gehler, 2017] by adding a domain filter to the network. More recently, [Wang and Lu,

[‡]Code available in https://github.com/lzhbrian/RDNet

2019] make use of the distribution and divergence of features, and achieve current state-of-the-art results. ERDIN [Liu *et al.*, 2019] modifies an architecture borrowing from the task of super resolution and is able to get even better numbers with a post-processing filter. Researchers also achieve IID by joint training with other tasks [Baslamisli *et al.*, 2018] and auxiliary data [Li and Snavely, 2018a; Li and Snavely, 2018b]. While most learning based methods have provided decent performance, we argue that they didn't make full use of the properties in the reflectance images.

3 Method

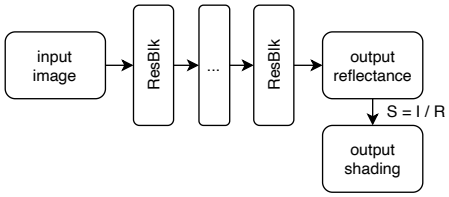
3.1 Preliminaries

Prior to our work, there are mainly two types of methods to solve the problem of IID.

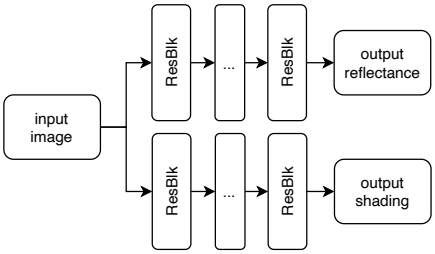
One-headed network (Figure 2a). Many method worked on the sparsely labeled IIW datasets only focus on learning and outputing reflectance images [Liu *et al.*, 2019; Wang and Lu, 2019; Fan *et al.*, 2018]. The network is usually designed to only output the reflectance images. The shading images are subsequently calculated by $S = I/R$.

Two-headed network (Figure 2b). For densely labeled datasets, since the dataset are all synthetic, we are able to obtain the groundtruth reflectance and shading images. This kind of networks are usually designed as two-headed. The input image is sent into two separate networks for obtaining reflectance and shading images, respectively. The two networks are often supervised with separate loss functions.

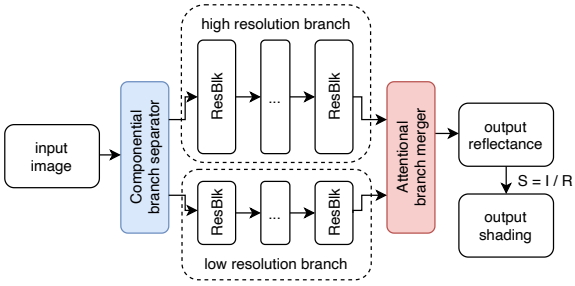
Ours (Figure 2c). In this work, we follow Figure 2a, and only design a dedicated network to output reflectance images. Because the reflectance images are learned well enough, directly calculating shading images by $S = I/R$ can also achieve decent results. In our network, supervision is only applied to the reflectance images during training. No supervision is applied to the shading images.



(a) Existing one-headed networks.



(b) Existing two-headed networks.



(c) Our one-headed network by pursuing reflectance images.

Figure 2: Network architectures.

3.2 RDNet

We now introduce our proposed RDNet (Reflectance Disentanglement Network) for intrinsic image decomposition. The overall architecture is shown in Figure 2c, RDNet composes of three important designs, namely *Componential branch separator*, *Attentional branch merger*, and *Attention mask regularization*.

Componential Branch Separator

As mentioned in Section 1, we observe that there are actually two independent components in an reflectance image: the flat surfaces with different colors, and the boundaries between these surfaces. The flat surfaces component usually consists of constant colors with low frequency information, and the boundaries component consists of sparse edges with high frequency details. In order to disentangle the learning of two components, we propose to separate the input images into two branches through a *componential branch separator* module.

Different from traditional CNN architecture (*cf.* Figure 2a) used for IID, the *componential branch separator* additionally generates a *low resolution branch* from the original feature maps $F_h \in \mathbb{R}^{H \times W \times C_h}$ through a strided 3x3 convolution. This procedure is shown in Figure 3. The new low resolution feature maps possess half the size of the original feature maps, and with a channel number of C_l. The original feature maps with higher resolution is responsible for capturing the boundaries component of the reflectance images, while the new low resolution feature maps are responsible for capturing the flat surfaces component. The design philosophy of this module is that the boundaries component is occupied by high frequency details, and needs a higher resolution feature maps to represent. On the other hand, the flat surfaces component in the reflectance images is occupied by low frequency information, and thus can be embedded in lower resolution feature maps and make the model more effective.

Attentional Branch Merger

After being separated by the *componential branch separator*, the two branches then go through several residual blocks [He *et al.*, 2016] to learn a more meaningful feature representation, respectively. In the end of the two branches, an *attentional branch merger* is used to merge the attention enhanced features of the two branches, and then use several convolutions to output the final reflectance image. Through this pro-

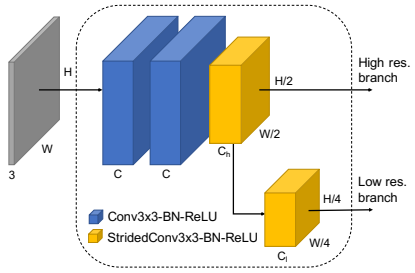


Figure 3: Componential branch separator.

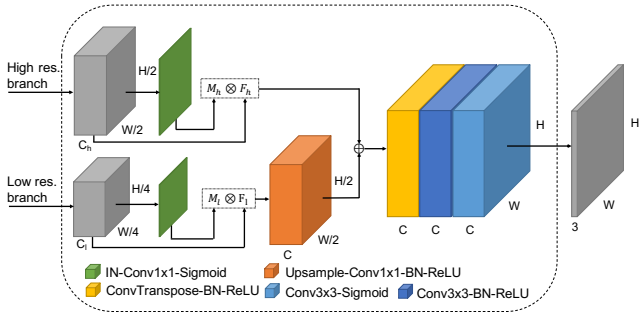


Figure 4: Attentional branch merger.

cess, we are able to ensure that the high and low resolution branches are actually learning the components they are assigned.

In the *attentional branch merger* module (*cf.* Figure 4), two attention masks M^h and M^l are generated with an instance normalization from the feature maps of the two branches, and followed by a 1x1 conv and a sigmoid function. The two attention masks are then multiplied elementwisely with their respective original feature maps, and added together after an upsample procedure apply to the low resolution output. The upsample procedure is realized by a bilinear upsampling, followed by a 1x1 conv, batch normalization and ReLU.

Attention Mask Regularization

In order to learn reasonable attention, we apply two respective regularization terms to the attention masks of high and low resolution branches. For attention mask on the high resolution branch, we use an L_1 regularization term on M^h to ensure the sparse characteristic of boundaries component:

$$L_{mask_h} = \sum |M^h|. \tag{2}$$

Another intuitive choice is to use the gradient maps of images as the supervision signals. However, we argue that the gradient maps shall contain high frequency information brought by shading as well. We only want the ones brought by reflectance. By using L_1 regularization, we are able to let the network learn by itself which high frequency information should be captured with proper constraints.

For attention mask on the low resolution branch, a total variation loss on M^l is used to ensure the smoothness of flat surfaces component:

$$L_{mask_l} = \sum_{i,j} \left|M^l_{i+i,j} - M^l_{i,j}\right|^2 + \left|M^l_{i,j+1} - M^l_{i,j}\right|^2. \tag{3}$$

Reflectance Supervision

For sparsely labeled dataset (*i.e.* IIW), we use a WHDR hinge loss [Nestmeyer and Gehler, 2017] on the output reflectance image:

$$L_{ref} = \ell_{\delta,\xi}(J, R, i) = \begin{cases} \max\left(0, \frac{R_{i_1}}{R_{i_2}} - \frac{1}{1+\delta+\xi}\right) & \text{if } J_i = 1 \\ \max\left(0, \begin{cases} \frac{1}{1+\delta-\xi} - \frac{R_{i_1}}{R_{i_2}}, \\ \frac{R_{i_1}}{R_{i_2}} - (1+\delta-\xi) \end{cases}\right) & \text{if } J_i = E \\ \max\left(0, 1+\delta+\xi - \frac{R_{i_1}}{R_{i_2}}\right) & \text{if } J_i = 2 \end{cases}, \tag{4}$$

where R is the output reflectance image, ξ is the margin, δ is a hyperparameter quantifies the significant level of the relative difference between two points, and $J_i \in \{1, 2, \mathrm{E}\}$ is the i-th groundtruth human judgement indicating that if point i_1 is *darker than* (1), *lighter than* (2), or *equal to* (E) point i_2.

For dense supervision datasets (*i.e.* MIT and MPI-Sintel), we utilize an MSE loss on the output reflectance images. Additionally, the image gradients in the x and y directions are also supervised:

$$\begin{aligned} L_{ref} =& |R - \hat{R}|^2 + \\ & |\nabla_x R - \nabla_x \hat{R}|^2 + |\nabla_y R - \nabla_y \hat{R}|^2, \end{aligned} \tag{5}$$

where R is the output reflectance image, and $\hat{R}$ is the groundtruth reflectance image.

Overall Loss Function

Finally, the overall loss function can be written as:

$$L = L_{ref} + \lambda_{mask_l} * L_{mask_l} + \lambda_{mask_h} * L_{mask_h}, \tag{6}$$

where λ_{mask_l} and λ_{mask_h} is the weight of the regularization terms. Note that our network only outputs the reflectance images and the loss function is only applied on them during training. The shading images are obtained by $S = I/R$, and no supervision is used.

3.3 Discussion

Here, we summarize some desirable properties of RDNet. Firstly, with the *componential branch separator*, we disentangle the learning process of the two components in reflectance images: the flat surfaces and the boundaries. By using feature maps with different resolutions, we successfully disentangle the components with distinct frequency information. Secondly, the *attentional branch merger* can learn the spatial attention of each branch. Through focusing on important spatial location for different components, each of them can be better modelled. Thirdly, the *attention mask regularization* term can limit the sparsity of the high resolution branch attention masks, and the smoothness of the low resolution branch attention masks. This poses a strong prior on the information to be learned, and thus makes the disentanglement and utilization much more effective. Last but not least, we are able to visualize the learned attention masks and reveal the internal mechanism of the model. This makes the model equipped with strong interpretability and alleviate this undesirable property of deep neural networks.

IIW	shd. sup.	WHDR	post filter
Const shading	n/a	51.37	
Const reflectance	n/a	36.54	
Retinex (color)	n/a	26.89	
Retinex (gray)	n/a	26.84	
[Zhou *et al.*, 2015]		19.95	
[Nestmeyer and Gehler, 2017]		19.49	17.69
[Bi *et al.*, 2015]	n/a	17.67	
[Zhou *et al.*, 2019]	√	15.20	
[Li and Snavely, 2018a]	√	14.80	
[Fan *et al.*, 2018]	√	14.45	
[Liu *et al.*, 2019]		14.31	13.76
[Wang and Lu, 2019]		13.90	
Ours		**13.47**	**13.19**

Table 1: Quantitative results (mean WHDR) on the IIW dataset. Most numbers in the table are copied from the paper of [Wang and Lu, 2019]. 'shd.sup.' means the method uses shading supervision during training, note that only learning based methods are applicable. 'post filter' means that the output reflectance image is further post-process by some filter.

MIT Intrinsic	shd. sup.	si-MSE		
		A	S	Avg
[Barron and Malik, 2014]	n/a	0.64	0.98	0.81
[Zhou *et al.*, 2015]	√	2.52	2.29	2.40
[Shi *et al.*, 2017]	√	2.16	1.35	1.75
[Narihira *et al.*, 2015a]	√	2.07	1.24	1.65
[Fan *et al.*, 2018]	√	1.27	0.85	1.06
[Cheng *et al.*, 2018]	√	0.89	0.73	0.81
Ours		**0.66**	**0.57**	**0.61**

Table 2: Quantitative results (×0.01) on the MIT dataset.

4 Experiments

In what follows, we design and conduct extensive experiments to answer the following four research questions.

- **RQ1:** How does our RDNet compare with other state-of-the-art IID methods on the challenging real-world IIW dataset? (Sec.4.2)
- **RQ2:** How does our RDNet perform on the densely labeled MPI-Sintel and MIT datasets? (Sec.4.2)
- **RQ3:** How does each module designed contribute to the improvement of RDNet? (Sec.4.3)
- **RQ4:** How does our RDNet disentangle the two components of reflectance images? (Sec.4.3)

4.1 Datasets and Metrics

Sparsely Labeled Dataset

IIW dataset. The real-world IIW dataset [Bell *et al.*, 2014] contains 872,161 pairwise reflectance comparisons across 5,230 photos. We follow the same settings as [Fan *et al.*, 2018]. For evaluation, the performance is measured on the reflectance images with the Weighted Human Disagreement Rate (WHDR). Since the pairwise judgement of human labels might be subjective, we note that humans actually have a median WHDR of 7.5% across all photos in IIW.

Densely Labeled Datasets

MPI-Sintel dataset. The MPI-Sintel dataset [Butler *et al.*, 2012] contains 8950 images from 18 scene level computer generated images sequences. We follow the same settings as [Cheng *et al.*, 2018]. The performance is measured on si-MSE, si-LMSE, and DSSIM.

MIT dataset. The MIT dataset [Grosse *et al.*, 2009] contains 20 object level images, each with 11 different lighting conditions. We follow the same settings as [Cheng *et al.*, 2018]. The performance is measured on si-MSE.

4.2 Qualitative and Quantitative Results

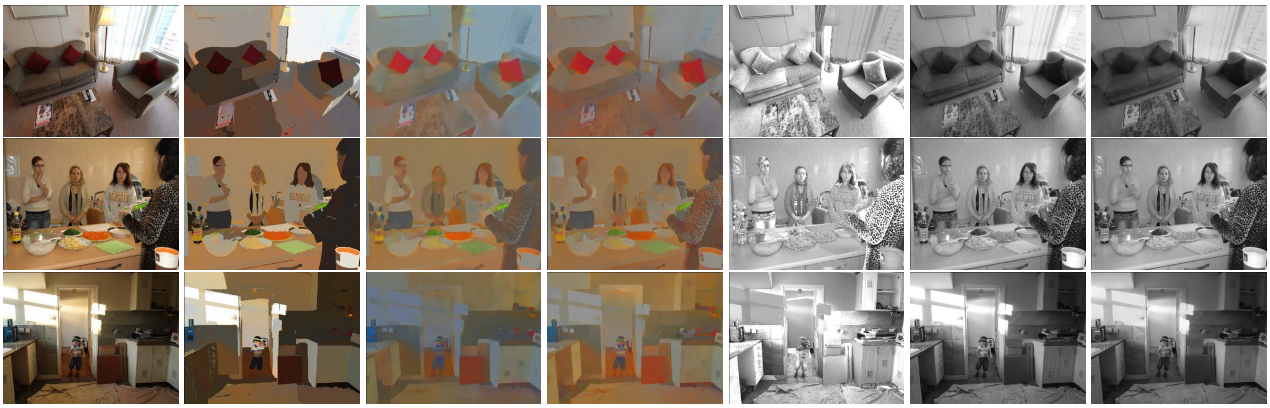
Sparsely Labeled Dataset (RQ1)

Table 1 presents the quantitative results of our proposed method compared with other methods on the IIW dataset. We can see that our method outperforms the current state-of-the-arts of [Wang and Lu, 2019] by a large margin. Note that, without a post-processing filter, we already outperform [Liu *et al.*, 2019], which is the current best number with a post-processing filter. If we further apply a bilateral filter to our output reflectance images, we can further lower our WHDR score to 13.19, which sets the new record for the IIW dataset.

We then compare some qualitative results in Figure 5. [Bi *et al.*, 2015] is the best optimization based method, and [Wang and Lu, 2019] is the current learning based state-of-the-art method. Firstly, we have observed much clearer and sharper reflectance images compared to [Wang and Lu, 2019], and more accurate reflectance images compared to [Bi *et al.*, 2015]. Notice the picture frame on the wall in the first image, the leftmost person's blue top near the waist in the second image, and the sundries on the cabinet of the right side in the third image. Secondly, we are also able to obtain much more accurate shading images. Notice the floor and sofa in the first image, wall and table in the second image, and the ceramic tiles on the right side and floor in the third image. Other methods usually recognize textures in the above objects as shading. Note that we do not have any supervision on shading during training. But with better modelling for reflectance images, the quality of shading images can also be improved considerably.

Densely Labeled Datasets (RQ2)

The performance of previous methods on densely labeled datasets is already very promising. We still conduct experiments on these datasets for two main reasons. Firstly, we want to confirm that our network architectures are suitable for most IID datasets. Secondly, different from existing methods, our network only output and apply supervision on the reflectance images. No supervision on the shading images is used. We want to know whether this setting is sufficient for networks to learn well. From Table 2 and Table 3, we can see that our proposed method outperforms all baselines by a large margin in all metrics on both MIT Intrinsic dataset and MPI-Sintel dataset. Note that all of our baselines use shading images as supervision, while we only use reflectance images as supervision.



Input | Bi *et al.* | Wang and Lu | Ours | Bi *et al.* | Wang and Lu | Ours

Figure 5: Qualitative results on the IIW dataset.

Sintel *scene split*	shd. sup.	si-MSE			si-LMSE			DSSIM		
		A	S	Avg	A	S	Avg	A	S	Avg
[Narihira *et al.*, 2015a]	√	2.01	2.24	2.13	1.31	1.48	1.39	20.73	15.94	18.33
[Fan *et al.*, 2018]	√	1.81	1.75	1.78	1.22	1.18	1.20	16.74	13.82	15.28
[Lettry *et al.*, 2018]	√	1.77	1.84	1.81	0.98	0.95	0.97	14.21	14.05	14.13
[Cheng *et al.*, 2018]	√	1.38	1.38	1.38	0.92	0.93	0.92	8.46	9.26	8.86
Ours		**1.29**	**1.28**	**1.28**	**0.84**	**0.90**	**0.87**	**7.07**	**8.48**	**7.78**

Table 3: Quantitative results (×0.01) on the MPI-Sintel dataset (scene split).

Method	WHDR
Baseline	14.40
+ Componential branch separator	13.84
+ Attentional branch merger	13.74
+ Attention mask regularization	13.65
+ 13 residual blocks	13.47

Table 4: Ablation study.

4.3 Analysis

Ablation Study (RQ3)

To further quantify the contribution of each design of our method, we conduct some ablation study in Table 4. We start by implementing a simple network architecture as in Figure 2a. The network contains 10 residual blocks, and is trained with a WHDR hinge loss (Eqn.4) on the IIW dataset. This architecture is basically the same as the direct intrinsic network in [Fan *et al.*, 2018]. Surprisingly, we find that with appropriate training schedule, this network is able to obtain WHDR=14.40, which is already sufficient to outperform most existing baselines. As a reminder, [Fan *et al.*, 2018] only achieves WHDR=15.40 with the same settings. After adding the *componential branch separator* module, and use a simple sum operation to merge the output feature maps of two branches, we can get WHDR=13.84, which is already the state-of-the-art result. This justifies the effectiveness of the disentanglement. Then, adding the *attentional branch merger*, and the *attention mask regularization* can further get a decrease of 0.1 and 0.09, respectively. By making the network deeper (adding more residual blocks), we continue to lower the score to 13.47.

Disentangling Reflectance Image Components (RQ4)

Finally, for better understanding the internal process of networks, we visualize the attention masks learned in Figure 1. We can clearly observe that the high resolution branch attention mask is paying attention to the boundaries, and the low resolution branch attention mask is focusing on the flat surfaces. This justifies that our network has disentangled the two components as we expected.

5 Conclusion

In this paper, we investigate the possibilities of only using reflectance images for supervision when solving the problem of intrinsic image decomposition. In such way, the demand for labeled data can be greatly reduced. We design RDNet to disentangle the learning process of two components of the reflectance images: the flat surfaces and the boundaries between them. RDNet is equipped with strong interpretability through demonstration of the disentanglement and utilization process. We have shed the light on the internal process of IID through pursuing the characteristics of reflectance images. Extensive experiments have shown that with only reflectance supervision, RDNet is also able to achieve state-of-the-art results on all datasets.

Acknowledgements

This work was supported by MOST-2018AAA0102004, NSFC-62061136001.

References

[Barron and Malik, 2014] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2014.

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

[Bell *et al.*, 2014] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *TOG*, 2014.

[Bi *et al.*, 2015] Sai Bi, Xiaoguang Han, and Yizhou Yu. An l1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. *TOG*, 2015.

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

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

[Chen *et al.*, 2019] Yunpeng Chen, Haoqi Fan, Bing Xu, Zhicheng Yan, Yannis Kalantidis, Marcus Rohrbach, Shuicheng Yan, and Jiashi Feng. Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In *ICCV*, 2019.

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

[Fan *et al.*, 2018] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018.

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

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[Kovacs *et al.*, 2017] Balazs Kovacs, Sean Bell, Noah Snavely, and Kavita Bala. Shading annotations in the wild. In *CVPR*, 2017.

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

[Li and Snavely, 2018a] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018.

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

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

[Liu *et al.*, 2019] Risheng Liu, Cheng Yang, Long Ma, Miao Zhang, Xin Fan, and Zhongxuan Luo. Enhanced residual dense intrinsic network for intrinsic image decomposition. In *ICME*, 2019.

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

[Narihira *et al.*, 2015a] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015.

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

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

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

[Shi *et al.*, 2017] Jian Shi, Yue Dong, Hao Su, and Stella X Yu. Learning non-lambertian object intrinsics across shapenet categories. In *CVPR*, 2017.

[Sun *et al.*, 2019] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019.

[Wang and Lu, 2019] Zongji Wang and Feng Lu. Single image intrinsic decomposition with discriminative feature encoding. In *ICCV Workshop*, 2019.

[Zhang *et al.*, 2021] Qing Zhang, Jin Zhou, Lei Zhu, Wei Sun, Chunxia Xiao, and Wei-Shi Zheng. Unsupervised intrinsic image decomposition using internal self-similarity cues. *PAMI*, 2021.

[Zhao *et al.*, 2012] Qi Zhao, Ping Tan, Qiang Dai, Li Shen, Enhua Wu, and Stephen Lin. A closed-form solution to retinex with nonlocal texture constraints. *TPAMI*, 2012.

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

[Zhou *et al.*, 2019] Hao Zhou, Xiang Yu, and David W. Jacobs. Glosh: Global-local spherical harmonics for intrinsic image decomposition. In *ICCV*, 2019.

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