SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

February 2023, Vol. 66 122105:1–122105:16
https://doi.org/10.1007/s11432-021-3481-3

Reflectance edge guided networks for detail-preserving intrinsic image decomposition

Quewei LI, Jie GUO*, Zhengyi WU, Yang FEI & Yanwen GUO*

State Key Lab for Novel Software Technology, Nanjing University, Nanjing 210023*, China*

Received 24 November 2021/Revised 22 February 2022/Accepted 1 April 2022/Published online 5 January 2023

Abstract Deep learning-based intrinsic image decomposition methods rely heavily on large-scale training data. However, current real-world datasets only contain sparse annotations, leading to textureless reflectance estimation. Although densely-labeled synthetic datasets are available, the large bias between these two categories easily incurs noticeable artifacts (e.g., shading residuals) on reflectance. To address this issue, we introduce reflectance edges that are predicted by a neural network trained on synthetic data with full supervision. Once trained, this network is able to capture high-frequency details of reflectance while greatly reducing the bias stemming from the discrepancy between different data distributions. We design another neural network to remove shading as much as possible from the input image. As this network is trained solely on real-world datasets, little bias will be introduced but the predicted reflectance will be overly smooth due to limited annotations. To recover texture details of the reflectance while still suppressing bias, we leverage a third neural network to progressively fuse feature maps from both reflectance edge maps and coarse-grained reflectance maps. The well-designed fusion strategy makes the best use of features extracted from the real-world data and helps to generate texture-rich reflectance with fewer artifacts. Extensive experiments on multiple benchmark datasets demonstrate the superiority of the proposed method.

Keywords intrinsic image decomposition, detail-preserving, reflectance edges

Citation Li Q W, Guo J, Wu Z Y, et al. Reflectance edge guided networks for detail-preserving intrinsic image decomposition. Sci China Inf Sci, 2023, 66(2): 122105, https://doi.org/10.1007/s11432-021-3481-3

1 Introduction

Intrinsic image decomposition is a classical computer vision problem that aims to extract the illumination-invariant component, i.e., reflectance, and the illumination-variant component, i.e., shading, from a natural image. Intrinsic images can serve a variety of high-level vision tasks and computational photography applications, such as shape from shading [1, 2], re-texturing [3, 4], recoloring [5–7], relighting [8, 9], and face editing [10, 11]. Unfortunately, although being extensively investigated in the past decades, intrinsic image decomposition remains challenging due to its ill-posedness.

With the development of deep learning, a practical and attractive candidate for solving this ill-posed problem is to replace the traditional hand-crafted priors with the data-driven, deep convolutional neural networks (CNN). However, to the best of our knowledge, there are no ground-truth data with dense labels for the complex real-world scenes because of the difficulties in collection. This incurs the detail-losing problem that networks trained on these weakly supervised samples always generate textureless reflectance results [12–15] as highlighted in Figure 1(b). To alleviate the limitation of real-world datasets, several approaches [16–19] suggest utilizing densely-labeled synthetic datasets for training. However, since synthetic data are biased, training networks on these data easily incurs sub-optimal convergence to poor local minima, leading to noticeable shading residuals or inconsistent effects on reflectance when testing for real-world images as highlighted in Figure 1(c).

To address the gap in data distribution of synthetic and real images, we propose a novel framework for detail-preserving intrinsic image decomposition that makes full use of reflectance edges. We first develop a shading removal network (SRNet) trained solely on the real-world dataset (i.e., IIW) to predict the

* Corresponding author (email: guojie@nju.edu.cn, ywguo@nju.edu.cn)

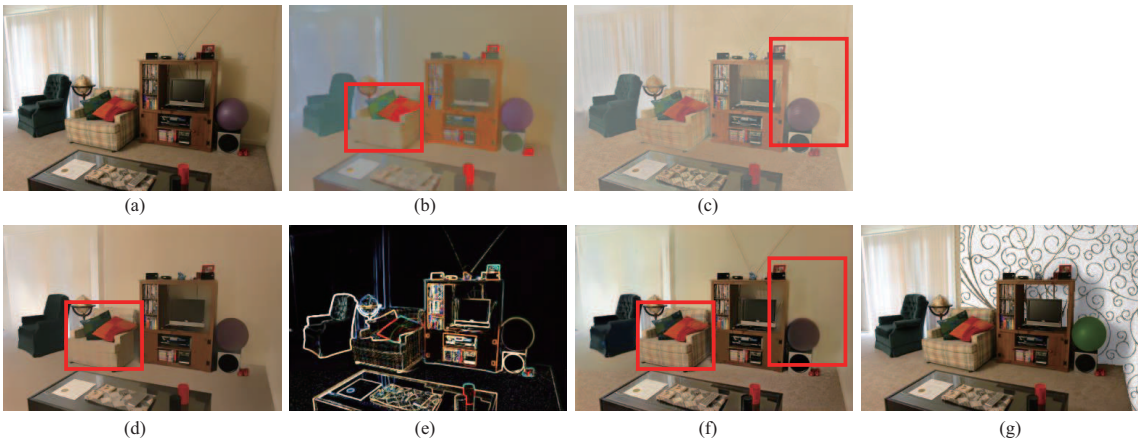


Figure 1 (Color online) We proposed a new framework for intrinsic image decomposition that can generate high-quality reflectance images from coarse-grained (d) to fine-grained (f) with the help of reflectance edges (e). The proposed method is able to preserve sufficient texture details while removing shading effects, and outperforms state-of-the-arts (b) and (c). The decomposed high-quality intrinsic images are beneficial for some image editing tasks, e.g., re-texturing (g). (a) Input; (b) Fan et al. [15]; (c) NIID-Net [19]; (d) ours (coarse); (e) reflectance edges; (f) ours (fine); (g) image editing.

coarse-grained reflectance from the input image (Figure 1(d)). To recover texture details that are missing from the sparsely-labeled real-world data, we resort to synthetic datasets with dense ground truths. In contrast to prior studies [16–19], which directly apply networks trained on synthetic data to predict reflectance images for real-world scenes, we propose a reflectance edge generation network (RegNet) to capture sudden changes along textures. We observe that the data distribution discrepancy between synthetic and realistic datasets in the edge/gradient domain is far less than that in the image domain. As a result, although trained on synthetic datasets, RegNet still performs well on real-world data and is expected to generate a detail-preserving reflectance edge map from the input image (Figure 1(e)). Then, we recover texture details on the coarse-grained reflectance with the help of the estimated reflectance edge map by fusing these two types of features with a well-designed multi-scale fusion network (FuseNet), producing the fine-grained reflectance image with great details and few shading residuals for a real-world scene (Figure 1(f)). The high-quality intrinsic results facilitate some image editing tasks such as material modification in which reflectance is altered while shading effects are preserved as shown in Figure 1(g).

Although the edge/gradient information has been adopted in some previous methods, its usage is significantly different in our pipeline. Previous studies leverage edges/gradients for either loss computation [12, 16, 17, 20] or flattening filters [15] to encourage the predicted reflectance or shading to be piece-wise smooth. Both ignore the texture cues in edge/gradient maps for detail-preserving. In contrast, our reflectance edges are predicted by RegNet and are explicitly used to enhance details in the reflectance map by carefully designing the neural networks, especially the FuseNet which extracts important features from reflectance edges and enhances their influence on the predicted reflectance with a novel multi-scale feature fusion strategy.

To summarize, the main contributions of our work are as follows:

- The introduction of reflectance edges to significantly reduce the bias of synthetic data used in training;
- A multi-scale fusion strategy that is designed to progressively recover details on coarse-grained reflectance with the estimated reflectance edges; and
- A coarse-to-fine intrinsic image decomposition framework that can produce high-quality reflectance images with rich textures.

2 Related work

Optimization-based intrinsic image decomposition. Intrinsic image decomposition has been extensively studied in the past years. Traditional approaches attempted to impose hand-crafted priors to constrain the space of feasible solutions [21–25]. One of the earliest methods is the classical Retinex algorithm [26] which assumes that large gradients of the image are caused by the changes of reflectance, while smaller gradients correspond to illumination changes. This algorithm is later extended by some meth-

ods [27–29] to enable the processing of real images. Since then, many priors have been developed based on physical geometry and illumination understanding, including reflectance sparsity [30, 31], non-local texture cues [32, 33], depth cues [34, 35], and constraints for shading [36–38]. However, the hand-crafted prior assumptions are often violated in complex real-world scenes, especially in the case of cast shadows or highlights, which limits the application of these methods. The use of multi-view stereo information has also been explored by several studies [39, 40]. However, these methods need to reconstruct the 3D points or geometry proxy of the scene for intrinsic image decomposition, which is time-consuming.

Deep intrinsic image decomposition. With the emergence of deep learning, recent methods decompose intrinsic images under the framework of deep neural networks [12, 13, 19, 41–44]. However, real-world datasets with high-quality dense ground truths are extremely hard to acquire and the lack of training data becomes the biggest problem of the data-driven approaches. The most widely used real-world datasets IIW [24] and SAW [45] only provide sparse reflectance and shading annotations for a small collection of pixels, respectively. On the other side, the existing densely-labeled synthetic datasets like CGI [16] and MPI-Sintel [46] often lack realism and have a large bias to the real-world scenes. Other datasets like MIT [21] and ShapeNet [47] only provide samples for single objects, which are far not enough for scene-level predictions. Li and Snavely [16] suggested training on CGI, IIW, and SAW datasets simultaneously to improve the performance on real-world images. However, such a simple joint training strategy may incur unstable training or sub-optimal convergence to local minima due to the discrepancy of distributions between different kinds of datasets, leading to inconsistent effects across large objects. Liu et al. [48] proposed an unsupervised framework for intrinsic image decomposition on a single image but suffered from the problems of noticeable artifacts and shading residuals. Fan et al. [15] proposed a general network architecture for multiple benchmark datasets by utilizing a guidance map to optimize the coarse-estimated reflectance with a domain filter, which achieves excellent flattening effects but is not capable to preserve texture details on the reflectance.

Recently, several approaches [20, 49, 50] explored the use of multiple time-lapse images for deep intrinsic image decomposition. These approaches require a collection of images from a fixed viewpoint of a scene with time-varying environmental parameters. Other approaches tried to incorporate additional information to the model for better performance, such as surface normal or illumination [17–19, 51], semantic segmentation maps [52], and depth maps [53]. Nonetheless, such approaches rely heavily on the quality of the generated auxiliary features to some extent, so the low-quality predictions would seriously affect the intrinsic results. On the contrary, we focus on intrinsic image decomposition for a single image without any additional information.

3 Our method

In this section, we describe the details of the proposed framework. Similar to previous studies [15, 16, 43, 48], we assume that a natural image I is decomposed as the pixel-wise product of the reflectance image R and the shading image S, i.e., $I = R \times S$. As illustrated in Figure 2, our framework consists of three functional components. We first estimate a coarse-grained reflectance image R^c with the SRNet and a reflectance edge map E with the RegNet. Then, R^c, E, and I are passed to a FuseNet, yielding a fine-grained reflectance image R^f and the corresponding shading image S.

3.1 Shading removal

There are a variety of approaches applied to the IIW dataset which only provides sparse annotations of pairwise reflectance comparisons. Although dense reflectance estimation on IIW has been achieved by several recent methods [15, 37, 41], the fine-grained reflectance with great details is still hard to obtain due to the lack of texture information. Considering this, we first estimate a coarse-grained reflectance image that aims at removing the shading components caused by illumination effects. It is practical to do so using the IIW dataset because, on one hand, 2/3 of the annotations are equal judgments, which is beneficial for the pairwise constant reflectance estimation. On the other hand, with some smoothing filters, we are able to generate a densely pseudo ground truth, which is textureless but consistent across large objects like walls and floors, for each example in the IIW dataset. This helps to further remove shading from the input image and achieve better flattening effects [37, 54].

The proposed SRNet adopts a typical U-Net [55] architecture with skip connections. It produces the coarse-grained reflectance image R^c with few shading residuals after trained on the IIW dataset. We

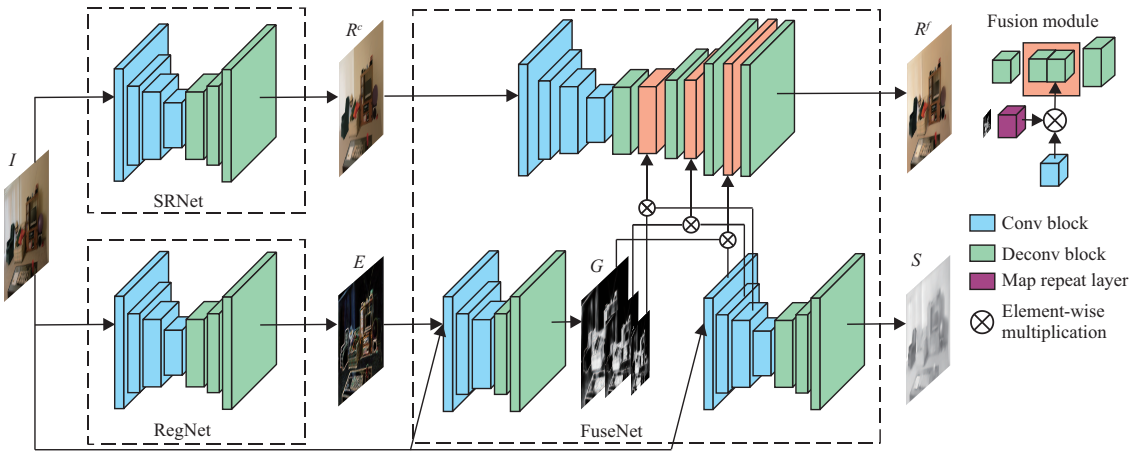


Figure 2 (Color online) Overview of our proposed method. Our shading removal network (SRNet) predicts coarse-grained reflectance images and our reflectance edge generation network (RegNet) predicts reflectance edge maps. The proposed multi-scale fusion network (FuseNet) fuses these two types of features, yielding fine-grained reflectance images and the corresponding shading images.

apply the commonly used weighted human disagreement rate (WHDR) hinge loss [15–17,41] designed on the sparse reflectance judgment $\boldsymbol{J}$ for weakly supervised training. Furthermore, to achieve stable training and flattening effects, we utilize the flattening image filter proposed in [37] to obtain the smooth reference map $\bar{R}^*$ for each IIW image and regard it as a substitute of the unknown dense ground truth. The total loss function for the sparsely-labeled IIW dataset is

$$\begin{aligned}\mathcal{L}_{SR} &= \mathcal{L}_D(R^c, \bar{R}^*) + \lambda_g \mathcal{L}_G(R^c, \bar{R}^*) + \lambda_w \mathcal{L}_W(R^c, \boldsymbol{J}) \\ &= \|R^c - \bar{R}^*\|_1 + \lambda_g \|\nabla R^c - \nabla \bar{R}^*\|_1 + \lambda_w \mathcal{L}_W(R^c, \boldsymbol{J}),\end{aligned} \tag{1}$$

where $\|\cdot\|_1$ denotes the L_1 norm, and λ_g and λ_w are the weights of corresponding loss terms. Similar to prior studies [15–17], in addition to the data loss $\mathcal{L}_D$, we apply the gradient loss $\mathcal{L}_G$ to the reflectance for piecewise smooth predictions. The detailed form of the WHDR hinge loss $\mathcal{L}_W$ and more training details about SRNet are provided in Appendix A.

3.2 Reflectance edge generation

While SRNet succeeds in removing shading from the input image and achieves high numerical performance after training on the IIW dataset, it fails to produce details on the reflectance image. This is because neither the sparse reflectance judgment $\boldsymbol{J}$ evaluates the texture details nor the smooth reference map $\bar{R}^*$ provides enough texture cues. To recover fine-grained details which have been smoothed out by SRNet, we resort to the reflectance edge map, a gradient map for the surface reflectance that captures sudden changes along textures. We achieve this with a RegNet which estimates the reflectance edge map of a natural image, preserving edges of textures while masking out shading edges.

Generating reflectance edge maps by utilizing densely-labeled synthetic datasets is usually straightforward and efficient. As dense ground-truth reflectance images are available, it is possible to predict reflectance edges from the input image in a fully-supervised manner. Furthermore, compared with directly estimating the reflectance image, the proposed RegNet which is trained on synthetic datasets still performs well for the real-world scene on predicting reflectance edges. This is because the bias between the synthetic and realistic data in the edge/gradient domain is far less than that in the image domain. We demonstrate this by respectively projecting 100 randomly selected color images from the synthetic and real-world datasets to the 2D space with principal component analysis (PCA). In the upper-left graph of Figure 3, each blue dot represents a synthetic image while each red dot represents a real-world image. The discrepancy obviously exists between these two categories. We conduct the same experiment on the edges of color image I. Here, the edge map $\mathcal{E}(I)$ is computed as

$$\mathcal{E}(I_i) = \sum_{j \in \mathcal{N}(i)} |I_i - I_j|, \tag{2}$$

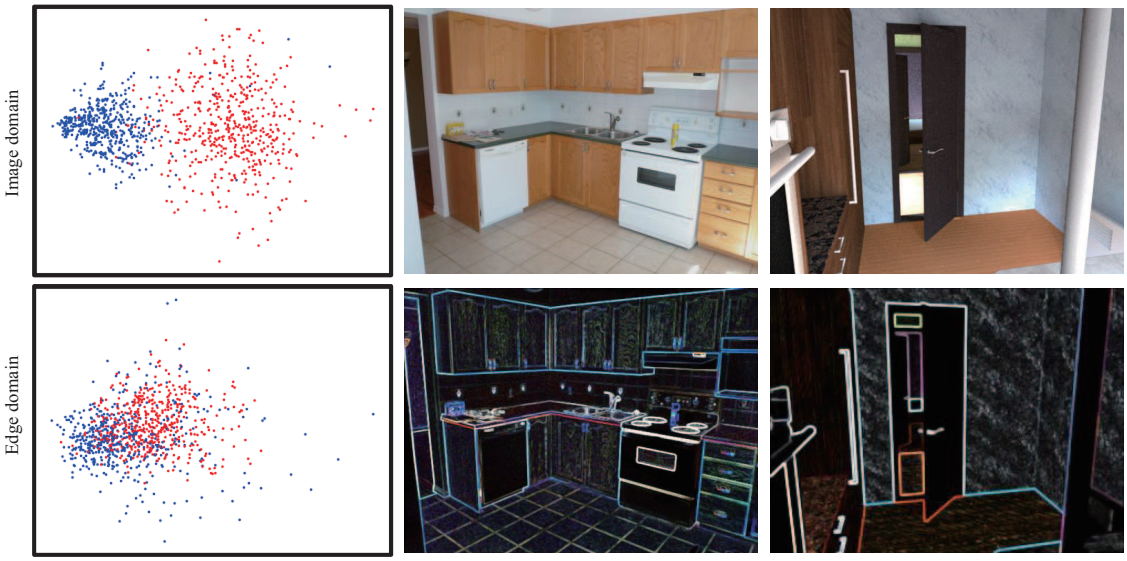


Figure 3 (Color online) Distribution discrepancy between synthetic and realistic datasets in different domains (a). The real-world image/edge (b) is from IIW and the synthetic image/edge (c) is from CGI. (a) Data distribution; (b) real-world data; (c) synthetic data.

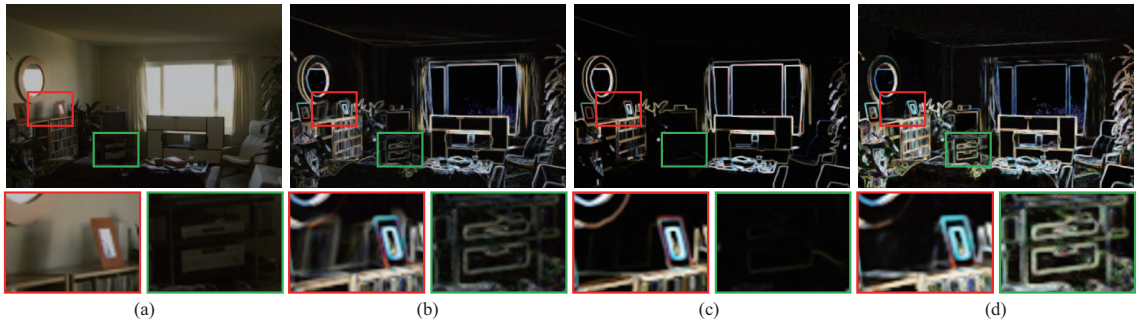


Figure 4 (Color online) Visual quality comparisons among the edge map of the natural image $\mathcal{E}(I)$ (b), the edge map of the coarse-grained reflectance $\mathcal{E}(R^c)$ (c), and the estimated reflectance edge map E (d). (a) is an input image I.

where $\mathcal{N}(i)$ indicates the neighboring pixels in a 3×3 region of pixel i. As shown in the bottom-left graph of Figure 3, the distribution discrepancy of edges between these two categories is much smaller than that in the image domain. The visual comparison in Figures 3(b) and (c) further shows that edge maps are harder to be distinguished (synthetic or real) than color images. This indicates that reflectance edges introduce little bias in our networks.

In practice, we train our RegNet, which has the same network architecture as SRNet, with the CGI dataset [16] since it provides ground-truth intrinsic images with full supervision. We obtain ground-truth reflectance edges $\mathcal{E}(R^*)$ from the dense reflectance map R^* with (2). Then, the loss function for RegNet is defined as

$$\mathcal{L}_{\text{Reg}} = \mathcal{L}_{\mathcal{E}}(E, R^*) = \|E - \mathcal{E}(R^*)\|_1, \tag{3}$$

where $\mathcal{L}_{\mathcal{E}}$ is the L_1 norm between two edge maps. After training, we use RegNet to predict reflectance edge maps for the natural images in IIW. As shown in Figure 4, the estimated reflectance edge map avoids introducing shading edges like shadow boundaries when compared with the edge map of the natural image $\mathcal{E}(I)$ (see the red boxes in Figures 4(b) and (d)), but contains much more texture details than the edge map of the coarse-grained reflectance $\mathcal{E}(R^c)$ (see the green boxes in Figures 4(c) and (d)).

3.3 Multi-scale fusion

Once we have obtained the coarse-grained reflectance image R^c and the reflectance edge map E for an input image I, we fuse them in a FuseNet to generate the fine-grained reflectance image R^f with rich

textures. As shown in Figure 2, the proposed FuseNet contains two autoencoders/branches to generate intrinsic results and a lightweight subnetwork for texture guidance map transformation, all of which are trained on IIW and SAW datasets. We first feed E and I into the transformation network and convert E into a texture guidance map G, which indicates the potential regions of texture details and suppresses shading effects in further treatment. Then, R^c and I are fed into two autoencoders, respectively. One autoencoder maps the input image I to the shading image S, while the other autoencoder progressively recovers texture details when mapping R^c to R^f by using the multi-scale fusion modules. Both autoencoders adopt the U-Net architecture with skip connections.

For texture recovery in R^c without introducing shading effects again, we utilize the texture guidance map G to re-weight the extracted color features at each scale before fusing features from different branches in the multi-scale fusion modules. Assuming the reflectance and color features from the two branches are $\mathcal{F}^r$ and $\mathcal{F}^c$, respectively, the fusion module (FM) at scale l is formulated as

$$\mathrm{FM}(\mathcal{F}_l^r, \mathcal{F}_l^c | G_l) = \mathrm{deconv}(\mathcal{F}_l^r \oplus (\mathcal{F}_l^c \otimes G_l)), \tag{4}$$

where $\oplus$ is the concatenation operation and $\otimes$ denotes element-wise multiplication. By masking out the shading areas in the color features with G, we reduce the risk of re-introducing shading residuals at this stage.

We train FuseNet with a self-supervised loss which assumes that the output fine-grained reflectance image R^f should be consistent with the coarse-grained input R^c, while $\mathcal{E}(R^f)$ should closely match E generated by RegNet. Besides, the WHDR hinge loss is also involved as weak supervision. Therefore, the loss function of the output reflectance is defined as

$$\begin{aligned} \mathcal{L}_{\mathrm{Fuse}}^R &= \mathcal{L}_D(R^f, R^c) + \lambda_\varepsilon \mathcal{L}_\mathcal{E}(R^f, E) + \lambda_w \mathcal{L}_W(R^f, \boldsymbol{J}) \\ &= \|R^f - R^c\|_1 + \lambda_\varepsilon \|\mathcal{E}(R^f) - E\|_1 + \lambda_w \mathcal{L}_W(R^f, \boldsymbol{J}). \end{aligned} \tag{5}$$

For the estimated shading image, we adopt the same loss in [16, 17] for weakly supervised training:

$$\mathcal{L}_{\mathrm{Fuse}}^S = \lambda_{cs} \mathcal{L}_{\text{constant-shading}} + \mathcal{L}_{\text{shadow}}, \tag{6}$$

where λ_{cs} is the balance factor, and $\mathcal{L}_{\text{constant-shading}}$ and $\mathcal{L}_{\text{shadow}}$ correspond to the shading annotations of constant shading regions and shadow boundaries, respectively. The detailed form of the shading loss terms $\mathcal{L}_{\text{constant-shading}}$ and $\mathcal{L}_{\text{shadow}}$ is provided in Appendix B. Furthermore, a reconstruction loss is applied to guarantee the consistency between the input natural image and the pixel-wise product of the reflectance and shading outputs:

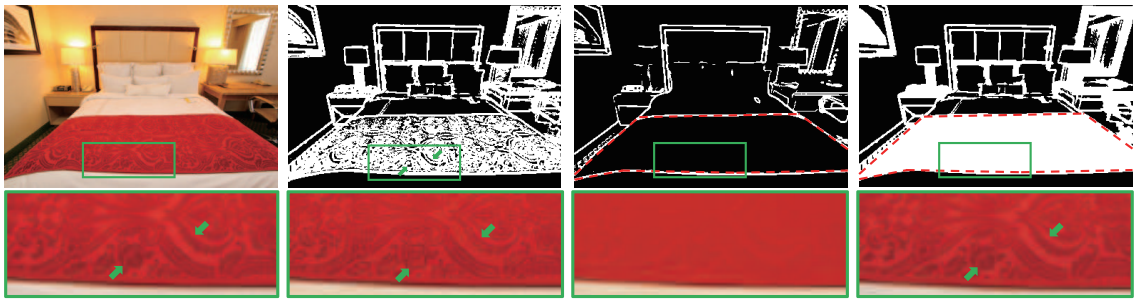
$$\mathcal{L}_{\mathrm{Fuse}}^{\mathrm{Rec}} = \|I - R^f \times S\|_1. \tag{7}$$

Consequently, the total loss function is

$$\mathcal{L}_{\mathrm{Fuse}} = \mathcal{L}_{\mathrm{Fuse}}^R + \mathcal{L}_{\mathrm{Fuse}}^S + \mathcal{L}_{\mathrm{Fuse}}^{\mathrm{Rec}}. \tag{8}$$

To obtain the texture guidance map G which indicates regions of texture details, a simple solution is to directly binarize the reflectance edge map E (denoted as $\Pi(E)$) with a pre-defined threshold. However, since E is essentially a gradient map indicating texture boundaries, $\Pi(E)$ may contain excessive texture details, which affects FuseNet's performance on fine-grained reflectance estimation that the recovered textures on the reflectance are inconsistent with the textures on the input image as pointed out with green arrows in Figures 5(a) and (b). Note that the edge map of such a reflectance image containing inconsistent textures still satisfies the edge loss term in (5). Considering this, we learn to predict G with a lightweight subnetwork. As no ground-truth texture guidance maps are available, we treat the binary mask of E as a reference. That is, we assume that G should be equal to $\Pi(E)$ in non-zero regions. Furthermore, to filter out redundant texture details (fill holes) on $\Pi(E)$, we apply an edge loss between $\mathcal{E}(G)$ and the binary mask of $\mathcal{E}(R^c)$. This is because R^c is textureless and $\Pi(\mathcal{E}(R^c))$ only contains significant reflectance boundaries as highlighted by red dotted lines in Figure 5(c). As a result, G will progressively cover the whole texture regions to (1) minimize its difference with $\Pi(E)$ in non-zero regions, and (2) ensure its edge map $\mathcal{E}(G)$ to be consistent with $\Pi(\mathcal{E}(R^c))$. Therefore, the loss function to train the subnetwork predicting G is given by

$$\mathcal{L}_{\mathrm{TGM}} = \mathcal{L}_D(G, \Pi(E)) + \lambda_\varepsilon \mathcal{L}_\mathcal{E}(G, \Pi(R^c))$$



(a) (b) (c) (d)

Figure 5 (Color online) Different variants of the texture guidance map and their reflectance results (in closeups). The white regions indicate the regions of textures that need to be recovered on reflectance. The binary mask of E (b) contains excessive texture details, generating reflectance with inconsistent textures compared with the input image. The network using the edge map of R^c (c) tends to produce textureless reflectance. In comparison, our complete method guided by G (d) can produce more plausible reflectance consistent to the input image. (a) I; (b) $\Pi(E)$; (c) $\Pi(\mathcal{E}(R^c))$; (d) G.

$$= \|G - \Pi(E, \delta_1)\|_1 + \lambda_\varepsilon \|\mathcal{E}(G) - \Pi(\mathcal{E}(R^c), \delta_2)\|_1, \tag{9}$$

where δ_1 and δ_2 are the pre-defined thresholds, and λ_ε is the balance factor. As shown in Figure 5(d), the high-frequency regions with great texture details are completely covered by G and the network trained with G can produce reflectance with more plausible texture details (consistent with the input image) in visual quality comparisons.

4 Experiments

4.1 Implementation details

Network design. We use the U-Net [55] architecture for all the subnetworks. We adopt the first four convolution blocks of VGG-16 [56] as the back-bone of each encoder (except for the lightweight subnetwork) but reduce the channel number of the last block from 512 to 256. The subnetwork for generating texture guidance maps is designed in a lightweight manner. It has 3 convolutional layers with stride 2, kernel size 3×3, and channel numbers 32, 64, 128 respectively on the encoder side. The decoders are symmetric to the encoders and are equipped with skip connections. Three fusion modules are used to fuse feature maps at different scales in FuseNet. Each (de)convolutional layer is followed by a batch normalization layer and a ReLU activation function except for the last deconvolutional layer, which is equipped with the sigmoid activation function to ensure that the output values fall in the range of [0,1].

Training details. We implement our networks with PyTorch and train them on four NVIDIA RTX 3090 GPUs. We use RMSprop optimizer with the initial learning rate 1E−3 and the power of 0.95 every 10 epochs. The hyper-parameters $\{\lambda_g, \lambda_w, \lambda_\varepsilon, \lambda_{cs}, \delta_1, \delta_2\}$ are set to $\{0.5, 0.3, 1.0, 2.0, 0.4, 0.3\}$ according to validation on a 5% randomly split training data.

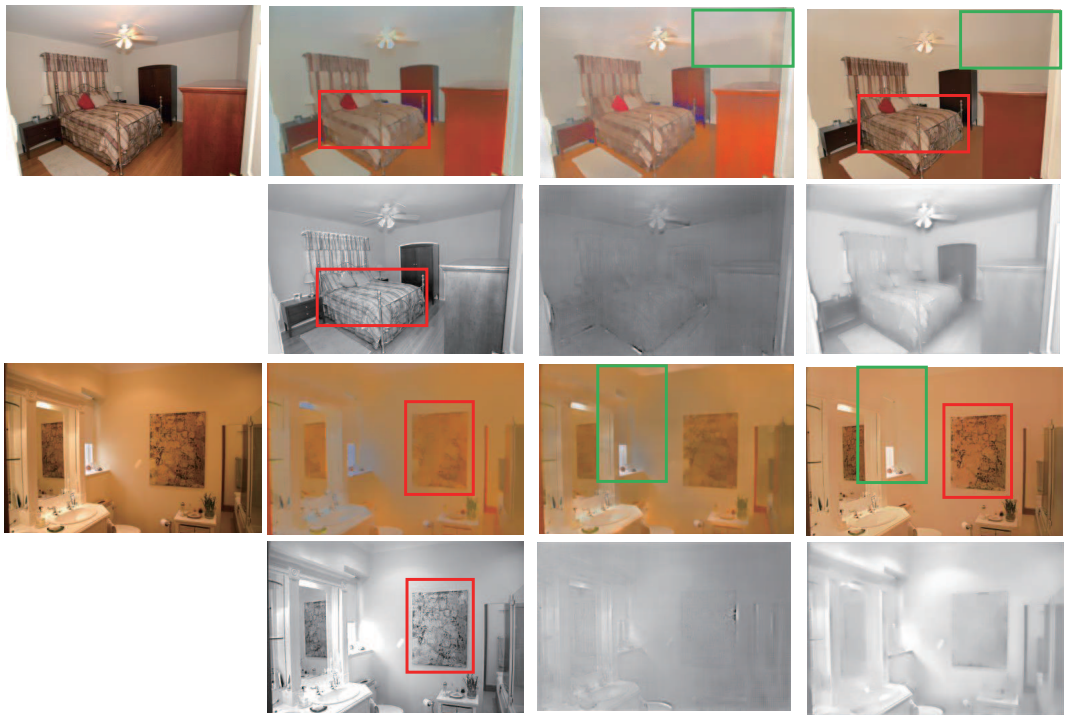
4.2 Evaluation on IIW and SAW

Datasets and evaluation metrics. We evaluate our networks on two widely used real-world datasets, i.e., IIW [24] and SAW [45]. The IIW dataset contains 5230 real images with total 872161 pairs of human-annotated relative reflectance judgments, while the SAW dataset comprises 6677 images in total, providing shading annotations of constant shading regions and shadow boundaries. We follow the train/test split provided by [57] for IIW and [45] for SAW, which is adopted in many recent studies [15–17,19,48]. Since the dense ground truths for both datasets are unavailable, we use the WHDR metric proposed by [24] to evaluate the reflectance on IIW. For the SAW dataset, we adopt the challenge average precision (AP) metric proposed by [16] to evaluate the estimated shading images.

Comparisons. We first compare the proposed method with two state-of-the-art methods, i.e., Fan's network [15] and Li's network [16], which also rely on real-world data for training. Table 1 [58] shows the quantitative analysis for the IIW and SAW datasets. As seen, both methods achieve high numerical performance on IIW, ranking top two among previous studies since they are customized to train on real-world datasets. However, Fan's network tends to predict textureless reflectance images due to the sparse labels adopted in training, leaving many texture residuals on the shading images as

Table 1 Reflectance evaluation (WHDR) on IIW and shading evaluation (AP) on SAW

Method	Dataset	IIW	SAW
		WHDR (%) ↓	AP (%) ↑
Grosse [21]	–	26.9	85.26
Garces [8]	–	24.8	92.39
Zhao [33]	–	23.8	89.72
Bi [37]	–	17.7	–
Bell [24]	–	20.6	92.18
Zhou et al. [13]	IIW	19.9	86.34
Liu et al. [48]	–	18.7	86.48
Nestmeyer [41]	IIW	17.7	88.64
NIID-Net [19]	CGI	16.6	98.40
GLoSH [17]	SUNCG [58]+IIW+SAW	15.2	95.01
Li et al. [16]	CGI+IIW+SAW	14.8	97.93
Fan et al. [15]	IIW	14.5	86.19
Ours (SRNet)	IIW	**14.4**	–
Ours (FuseNet)	CGI+IIW+SAW	14.6	**98.68**



(a) (b) (c) (d)

Figure 6 (Color online) Visual quality comparisons with two popular intrinsic image decomposition methods which also rely on real-world datasets for training. Since sparse annotations are used for supervision, previous methods (Fan et al. and Li et al.) lose many details on the reflectance (highlighted in red boxes), while our method based on reflectance edges retails these important details and avoids shading residuals (shown in green boxes). (a) Input; (b) Fan et al. [15]; (c) Li et al. [16]; (d) ours.

shown in Figure 6(b). By incorporating synthetic training samples, Li's network preserves more texture details. However, due to the bias introduced by synthetic images, such a simple joint training strategy makes the predicted reflectance inconsistent across large objects like walls as shown in Figure 6(c). In comparison, the proposed SRNet achieves the lowest WHDR on IIW, and our complete pipeline achieves the outstanding performance on IIW (only 0.1% higher than the state-of-the-art). Note that WHDR is only computed on sparse positions, ignoring texture details. This may be not reliable for evaluating fine-grained details. As shown in Figure 6(d), the estimated fine-grained reflectance images contain great texture details (highlighted by red boxes) while still preserving consistency across large objects with few shading residuals (highlighted by green boxes). For the testing on SAW, Fan's network fails to achieve a high AP metric (12.49% lower than ours) since the estimated shading images suffer from severe texture

Figure 7 (Color online) More comparisons on reflectance with existing intrinsic image decomposition methods. Note that both NIID-Net and Liu's network suffer from noticeable shading residuals. (a) Input; (b) NIID-Net [19]; (c) Liu et al. [48]; (d) ours (R^c); (e) ours (R^f).

residuals. Though Li's network generates high quantitative results, their shading images are of low contrast. In comparison, our complete pipeline achieves the best performance among all the methods on the AP metric, producing more plausible shading with high contrast and few texture residuals.

In Figure 7, we further compare two additional methods, including NIID-Net [19] which only utilizes the synthetic CGI dataset [16] for training, and Liu's network [48] which is trained in an unsupervised manner. Without utilizing the sparsely-labeled real-world datasets, both NIID-Net and Liu's network preserve much more fine-grained details than Fan's network and Li's network. However, both methods suffer from improper shading residuals, as highlighted in Figures 7(b) and (c). Furthermore, without training on real-world samples, both NIID-Net and Liu's network tend to generate large bias, yielding high WHDR scores on IIW as shown in Table 1. Overall, our method achieves the state-of-the-art performance on both shading removal and texture preserving, and outperforms existing methods both qualitatively and quantitatively.

4.3 Evaluation on other datasets

Datasets and evaluation metrics. For completeness, we further compare our method on the densely-labeled datasets, i.e., the MPI-Sintel and MIT intrinsic datasets. The small-scale MIT intrinsic dataset is also a real-world dataset that only contains 220 images with 20 different objects, each of which has 11 images for different illumination conditions. The MPI-Sintel dataset has a total of 890 synthetic images from 18 virtual scenes. We fine-tune our networks on these two datasets respectively with the provided dense ground truths and then evaluate their performance. For the MPI-Sintel dataset, we adopt the same experiment settings as Fan et al. [15]. Specifically, we use two-fold validation [12] to obtain all 890 MPI-Sintel test results with two kinds of splitting strategies, i.e., scene split and image split, with the train/test list provided by [15]. We also follow their evaluation metrics, employing the mean square error (MSE), local MSE (LMSE), and dissimilarity structural similarity index measure (DSSIM) to evaluate the performance. For the MIT dataset, we use the train/test split provided by [25] and MSE for evaluation.

Comparisons. As shown in Table 2 and Figure 8, our method significantly outperforms previous methods on the MPI-Sintel dataset (achieves the best result for most columns in Table 2) and produces sharper and more accurate results compared with Fan's network [15]. Furthermore, our method achieves comparable performance with state-of-the-art methods on the MIT intrinsic dataset as shown in Figure 9, although this dataset only provides very limited examples for training, which is not sufficient for our

Table 2 Quantitative comparisons on the MPI-Sintel dataset[a)]

		MSE		LMSE		DSSIM	
		Reflectance	Shading	Reflectance	Shading	Reflectance	Shading
Image split	Grosse [21]	0.0606	0.0727	0.0366	0.0419	0.2270	0.2400
	MSCR [12]	0.0100	0.0092	0.0083	0.0085	0.2014	0.1505
	Liu et al. [48]	0.0159	0.0148	0.0087	0.0081	0.1797	0.1474
	Fan et al. [15]	0.0069	0.0059	0.0044	**0.0042**	0.1194	**0.0822**
	Ours	**0.0053**	**0.0057**	**0.0034**	0.0052	**0.0781**	0.0933
Scene split	MSCR [12]	0.0190	0.0213	0.0129	0.0141	0.2056	0.1596
	Fan et al. [15]	0.0189	0.0171	0.0122	**0.0117**	0.1645	0.1450
	Ours	**0.0172**	**0.0167**	**0.0112**	0.0135	**0.1557**	**0.1363**

a) The best results are highlighted in bold.

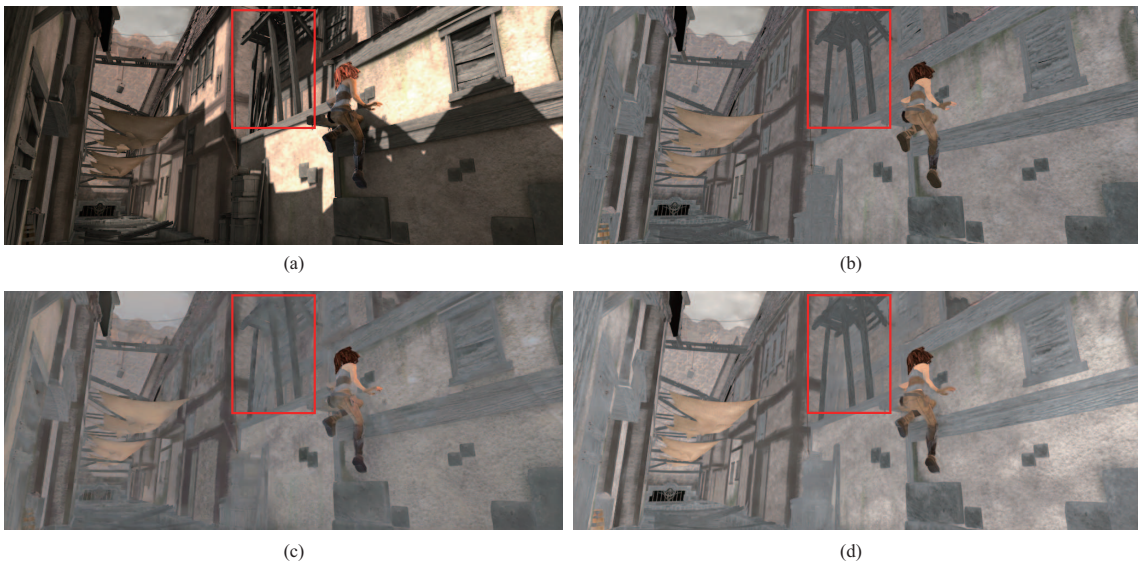


Figure 8 (Color online) Visual quality comparisons of reflectance on the MPI-Sintel dataset. (a) Input; (b) ground truth; (c) Fan et al. [15]; (d) ours.

pipeline. More comparisons on the MPI-Sintel dataset are provided in Appendix D.

4.4 Ablation study

Effectiveness of reflectance edges. The usage of reflectance edge is very important in the proposed method. To verify its effectiveness, we remove it from our complete method. Specifically, we train a model by replacing the texture guidance map (generated from the reflectance edge map) with a mask containing all ones. It means that we directly pass the color features to the multi-scale fusion modules without the re-weighting operation in (4). Furthermore, we remove the edge loss $\mathcal{L}_{\mathcal{E}}$ in (5) from our complete loss function. As shown in Figure 10, without re-weighting the color features, shading residuals like highlights or shadows are left on the reflectance predictions. Although the use of data loss $\mathcal{L}_D$ in (5) can suppress shading effects to some extent, the shading residuals are still noticeable when compared with our complete method. Besides, the performance of texture recovery is also degraded due to the removal of the edge loss $\mathcal{L}_{\mathcal{E}}$.

In Figure 11, we present another experiment by replacing the reflectance edge map E with the image gradient $\mathcal{E}(I)$. As pointed out by red arrows, the shading edges are preserved in $\mathcal{E}(I)$, leading to noticeable shading residuals on the estimated reflectance when serving $\mathcal{E}(I)$ as a guidance for texture recovery. Besides, the texture edges in the dark regions (poor illumination) of the scene are darker than other texture edges, which also raises shading effects on reflectance as highlighted in the green boxes. In comparison, the reflectance edge E generated by the RegNet avoids introducing shading edges and still contains great texture details. Furthermore, the intensity of the reflectance edges is consistent for the whole scene without being affected by illumination variants among different regions. All of these demonstrate that the estimated reflectance edge map E is more suitable than the image gradient $\mathcal{E}(I)$ in our pipeline for

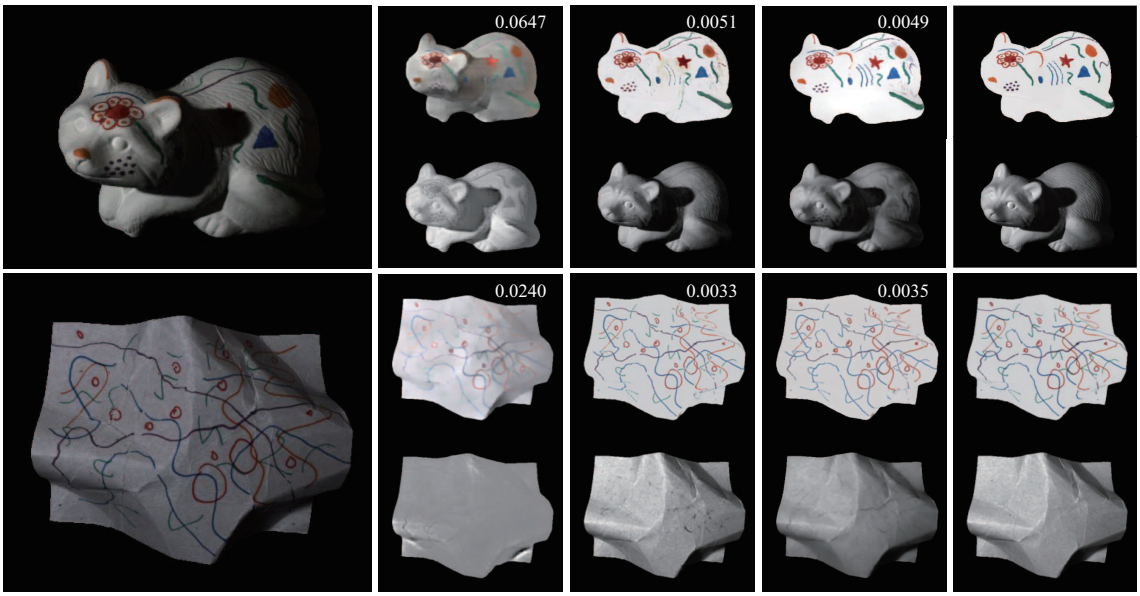


Figure 9 (Color online) Comparisons on the MIT dataset. The MSE value is reported in the top right corner of each image. (a) Input; (b) Li et al. [16]; (c) Fan et al. [15]; (d) ours; (e) ground truth.

Figure 10 (Color online) Visual quality comparisons between the proposed method with (bottom right) and without (bottom left) reflectance edge maps.

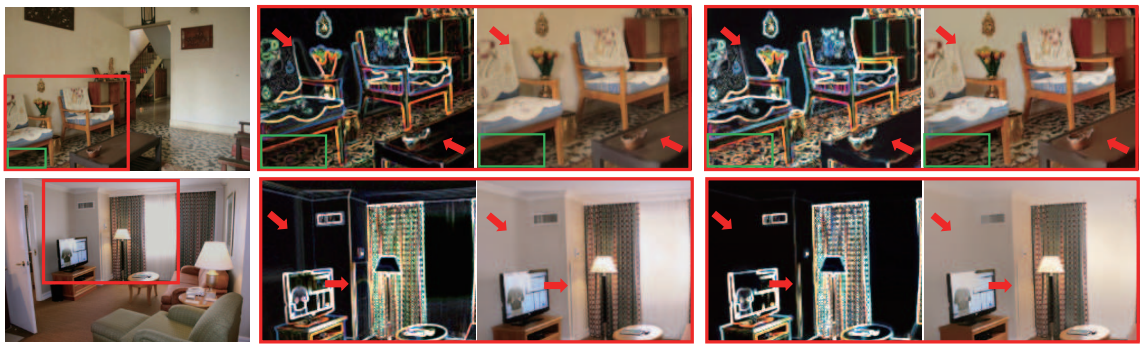


Figure 11 (Color online) Comparisons between image gradient (b) and our estimated reflectance edge map (c). For each closeup group, the left image is the edge map and the right image is the corresponding reflectance. (a) Input; (b) $\mathcal{E}(I)$; (c) E.

texture recovery on reflectance without introducing shading residuals. Table 3 further validates that our complete method outperforms these variant models, i.e., the models trained without reflectance edges or trained with the image gradient. A complete validation of reflectance edges is provided in Appendix C.

Table 3 Ablation study on the variants of our method trained without reflectance edges, with image gradients, without SRNet, and with different losses

Method		IIW	SAW
		WHDR (%) ↓	AP (%) ↑
w/o reflectance edges E		14.86	96.87
Replacing E with $\mathcal{E}(I)$		15.02	96.82
w/o SRNet		18.61	95.24
In SRNet	w/o $\mathcal{L}_W$	16.35	–
	w/o $\mathcal{L}_D$&$\mathcal{L}_G$	15.77	–
In FuseNet	w/o $\mathcal{L}_W$	16.01	93.18
	w/o $\mathcal{L}^S_{\text{Fuse}}$	14.89	93.38
	w/o $\mathcal{L}_D$	15.96	92.96
	w/o $\mathcal{L}_\mathcal{E}$	**14.38**	95.43
	w/o $\mathcal{L}^{\text{Rec}}_{\text{Fuse}}$	14.85	96.00
	w/o $\mathcal{L}_{\text{TGM}}$	14.75	97.53
Our complete method		14.63	**98.68**

Figure 12 (Color online) Comparing our complete model with two variants: model trained without SRNet or edge loss, respectively. (a) Input; (b) variants; (c) our complete model.

Effectiveness of SRNet. We verify the effectiveness of SRNet by removing it from our complete method. Specifically, we directly pass the input natural image to the reflectance branch of FuseNet. As coarse-grained reflectance images are no more available to calculate the data loss in (5), we utilize the smooth reference map generated by the flattening image filter [37] as a candidate. The first row of Figure 12 shows that without SRNet, shading effects cannot be well separated from the reflectance components although texture details are preserved. As expected, our complete method achieves satisfactory results on both shading removal and detail-preserving. Table 3 also validates that the performance of our method shows a significant gap when removing SRNet, especially the WHDR metric. These experiments imply that the use of SRNet plays an important role in removing shading effects.

Validation of the loss functions. To evaluate the effectiveness of the proposed loss functions, we remove each loss term in the complete loss functions respectively. As shown in Table 3, the performance of each variant is decreased compared with our complete method, except for the model trained without edge loss (denoted as 'w/o $\mathcal{L}_\mathcal{E}$' in Table 3) which achieves a better WHDR score on the IIW dataset due to the preference of WHDR. In the second row of Figure 12 we show that both methods perform well on shading removal while our complete method preserves much more texture details.

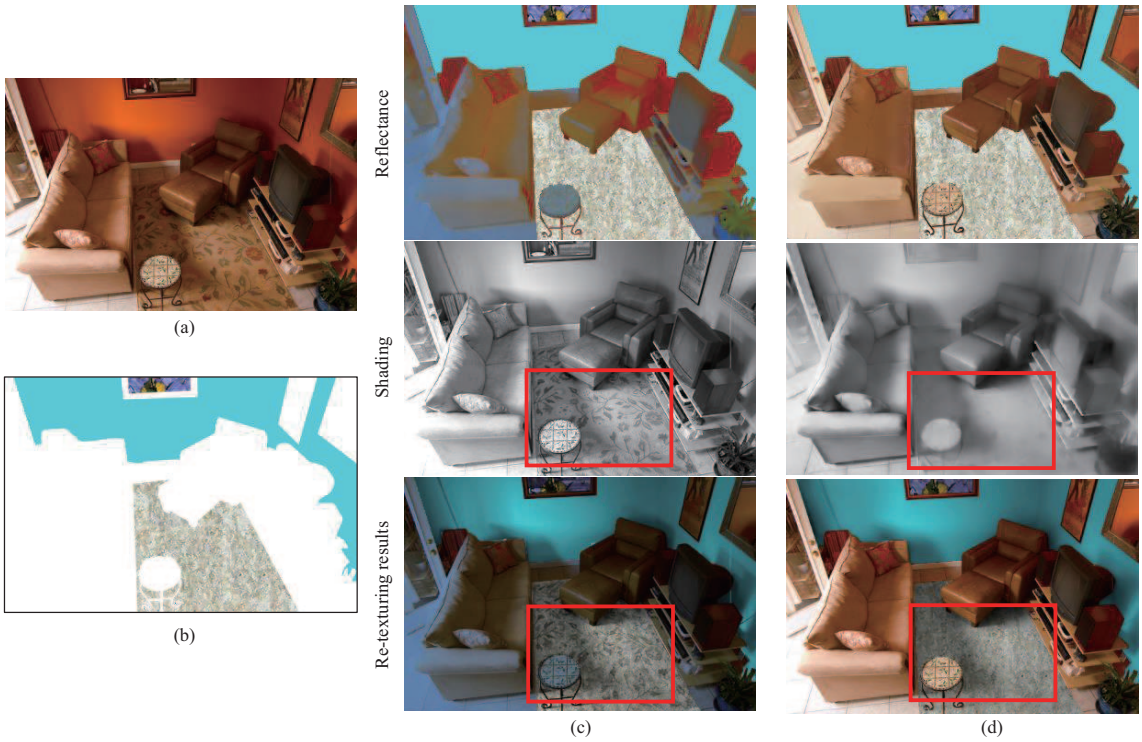


Figure 13 (Color online) Example of re-texturing by altering objects' materials in the input image (a) with the target textures (b). Our method (d) achieves more plausible results than previous method of Fan et al. [15] (c) due to the high quality of estimated intrinsic images. Note that Fan's network still contains old textures on the carpet while our method avoids this.

5 Image editing

To further validate the effectiveness of the estimated high-quality intrinsic images, we apply these intrinsic images to some image editing applications. In Figure 13, we show a re-texturing task in which we alter the texture of the carpet and the color of the wall. Note that the shadows on the carpet and the wall are unaffected. In comparison with Fan's method [15], our method produces more plausible image editing results by achieving better realism without incurring the texture-copy problem. This evidences that our proposed method preserves more texture details on the reflectance such that less will be copied to the shading image. Besides, the image editing results for the illumination-varying image sequence in Figure 14 further show that our method decomposes shading and reflectance quite well and is consistent under different illuminations. We believe our method also facilitates many other image editing or augmented reality applications by providing high-quality reflectance and shading.

6 Limitations

Although achieving state-of-the-art performance, our method suffers from some limitations. Notably, the performance of shading removal for our framework relies heavily on SRNet. Once shading effects cannot be removed on the coarse-grained reflectance, they are likely to be left on the final fine-grained reflectance. Besides, there is a risk of re-introducing shading effects on the reflectance in the process of texture recovery since the estimated reflectance edge map may still contain shading edges such as the strong cast shadow boundaries. We hope these problems would be solved by introducing high-quality real-world datasets or by incorporating shadows removal models [59] to specifically deal with these issues. Furthermore, it is also an interesting future work to generate sharper shading image while still preserving high quality in reflectance.

Figure 14 (Color online) Example of image editing for the illumination-varying image sequence. The image sequence is provided by [49]. (a) Original illumination-varying image sequence; (b) our editing results.

7 Conclusion

In this study, we introduce the reflectance edge and show its superiority on intrinsic image decomposition. The reflectance edge map, which is estimated by RegNet trained on the synthetic dataset (i.e., CGI), is essentially a gradient map capturing sudden changes along textures. This edge map is used to progressively recover texture details on the coarse-grained reflectance image generated by SRNet with multi-scale fusion modules. With the coarse-to-fine strategy, the proposed FuseNet finally produces fine-grained reflectance and the corresponding shading image. Extensive experimental results show that our method outperforms state-of-the-art methods in producing high-quality reflectance with great details and few shading residuals. Several image editing applications further validate the effectiveness of the proposed method. We believe in the future our method can promote practical applications of intrinsic images in augmented reality and be used for other decomposition problems, such as color palette decomposition and recoloring. On one hand, it is easier to change the textures or extract color palettes from the high-quality reflectance than the natural image which contains interference factors such as highlights or shadows. On the other hand, it is efficient to generate high-fidelity results by multiplying the retextured/recolored reflectance and the shading images.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61972194, 62032011) and Natural Science Foundation of Jiangsu Province (Grant No. BK20211147).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Wu C, Zollhöfer M, Nießner M, et al. Real-time shading-based refinement for consumer depth cameras. ACM Trans Graph, 2014, 33: 1–10
2 Zollhöfer M, Dai A, Innmann M, et al. Shading-based refinement on volumetric signed distance functions. ACM Trans Graph, 2015, 34: 1–14
3 Shen J, Yan X, Chen L, et al. Re-texturing by intrinsic video. Inf Sci, 2014, 281: 726–735
4 Meka A, Fox G, Zollhofer M, et al. Live user-guided intrinsic video for static scenes. IEEE Trans Visual Comput Graph, 2017, 23: 2447–2454
5 Tan J, Lien J M, Gingold Y. Decomposing images into layers via RGB-space geometry. ACM Trans Graph, 2017, 36: 1–14
6 Wang Y L, Liu Y F, Xu K. An improved geometric approach for palette-based image decomposition and recoloring. Comput Graph Forum, 2019, 38: 11–22
7 Cui M Y, Zhu Z, Yang Y, et al. Towards natural object-based image recoloring. Comp Visual Media, 2022, 8: 317–328
8 Garces E, Munoz A, Lopez-Moreno J, et al. Intrinsic images by clustering. Comput Graph Forum, 2012, 31: 1415–1424

9 Nestmeyer T, Lalonde J F, Matthews I, et al. Learning physics-guided face relighting under directional light. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020. 5124–5133

10 Li C, Zhou K, Lin S. Simulating makeup through physics-based manipulation of intrinsic image layers. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 4621–4629

11 Shu Z, Yumer E, Hadap S, et al. Neural face editing with intrinsic image disentangling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5541–5550

12 Narihira T, Maire M, Yu S X. Direct intrinsics: learning albedo-shading decomposition by convolutional regression. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 2992–2992

13 Zhou T, Krahenbuhl P, Efros A A. Learning data-driven reflectance priors for intrinsic image decomposition. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 3469–3477

14 Zoran D, Isola P, Krishnan D, et al. Learning ordinal relationships for mid-level vision. In: Proceedings of the IEEE International Conference on Computer Vision, 2015. 388–396

15 Fan Q, Yang J, Hua G, et al. Revisiting deep intrinsic image decompositions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 8944–8952

16 Li Z, Snavely N. CGIntrinsics: better intrinsic image decomposition through physically-based rendering. In: Proceedings of European Conference on Computer Vision (ECCV), 2018

17 Zhou H, Yu X, Jacobs D W. GLoSH: global-local spherical harmonics for intrinsic image decomposition. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 7820–7829

18 Sengupta S, Gu J, Kim K, et al. Neural inverse rendering of an indoor scene from a single image. In: Proceedings of the IEEE International Conference on Computer Vision, 2019. 8598–8607

19 Luo J, Huang Z, Li Y, et al. NIID-Net: adapting surface normal knowledge for intrinsic image decomposition in indoor scenes. IEEE Trans Visual Comput Graph, 2020, 26: 3434–3445

20 Lettry L, Vanhoey K, van Gool L. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. Comput Graph Forum, 2018, 37: 409–419

21 Grosse R, Johnson M K, Adelson E H, et al. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In: Proceedings of IEEE 12th International Conference on Computer Vision, 2009. 2335–2342

22 Tappen M F, Freeman W T, Adelson E H. Recovering intrinsic images from a single image. IEEE Trans Pattern Anal Machine Intell, 2005, 27: 1459–1472

23 Shen L, Yeo C. Intrinsic images decomposition using a local and global sparse representation of reflectance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2011. 697–704

24 Bell S, Bala K, Snavely N. Intrinsic images in the wild. ACM Trans Graph, 2014, 33: 1–12

25 Barron J T, Malik J. Shape, illumination, and reflectance from shading. IEEE Trans Pattern Anal Mach Intell, 2015, 37: 1670–1687

26 Land E H, McCann J J. Lightness and retinex theory. J Opt Soc Am, 1971, 61: 1–11

27 Horn B K P. Determining lightness from an image. Comput Graph Image Process, 1974, 3: 277–299

28 Blake A. Boundary conditions for lightness computation in Mondrian World. Comput Vision Graph Image Process, 1985, 32: 314–327

29 Funt B V, Drew M S, Brockington M. Recovering shading from color images. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 1992. 124–132

30 Omer I, Werman M. Color lines: image specific color representation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004

31 Rother C, Kiefel M, Zhang L, et al. Recovering intrinsic images with a global sparsity prior on reflectance. In: Proceedings of Advances in Neural Information Processing Systems, 2011. 765–773

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

33 Zhao Q, Tan P, Dai Q, et al. A closed-form solution to retinex with nonlocal texture constraints. IEEE Trans Pattern Anal Mach Intell, 2012, 34: 1437–1444

34 Barron J T, Malik J. Intrinsic scene properties from a single RGB-D image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013. 17–24

35 Chen Q, Koltun V. A simple model for intrinsic image decomposition with depth cues. In: Proceedings of the IEEE International Conference on Computer Vision, 2013. 241–248

36 Li Y, Brown M S. Single image layer separation using relative smoothness. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014. 2752–2759

37 Bi S, Han X, Yu Y. An L_1 image transform for edge-preserving smoothing and scene-level intrinsic decomposition. ACM Trans Graph, 2015, 34: 1–12

38 Sheng B, Li P, Jin Y, et al. Intrinsic image decomposition with step and drift shading separation. IEEE Trans Visual Comput Graph, 2020, 26: 1332–1346

39 Laffont P Y, Bousseau A, Drettakis G. Rich intrinsic image decomposition of outdoor scenes from multiple views. IEEE Trans Visual Comput Graph, 2013, 19: 210–224

40 Laffont P Y, Bousseau A, Paris S, et al. Coherent intrinsic images from photo collections. ACM Trans Graph, 2012, 31: 1–11

41 Nestmeyer T, Gehler P V. Reflectance adaptive filtering improves intrinsic image estimation. In: Proceedings of the IEEE

Conference on Computer Vision and Pattern Recognition, 2017. 6789–6798
42 Shi J, Dong Y, Su H, et al. Learning non-lambertian object intrinsics across shapenet categories. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 1685–1694
43 Cheng L, Zhang C, Liao Z. Intrinsic image transformation via scale space decomposition. In: Proceedings of The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018
44 Baslamisli A S, Le H A, Gevers T. CNN based learning using reflection and retinex models for intrinsic image decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 6674–6683
45 Kovacs B, Bell S, Snavely N, et al. Shading annotations in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 6998–7007
46 Butler D J, Wulff J, Stanley G B, et al. A naturalistic open source movie for optical flow evaluation. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2012. 611–625
47 Chang A X, Funkhouser T, Guibas L, et al. ShapeNet: an information-rich 3D model repository. 2015. ArXiv:1512.03012
48 Liu Y, Li Y, You S, et al. Unsupervised learning for intrinsic image decomposition from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020
49 Li Z, Snavely N. Learning intrinsic image decomposition from watching the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 9039–9048
50 Ma W C, Chu H, Zhou B, et al. Single image intrinsic decomposition without a single intrinsic image. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018. 201–217
51 Janner M, Wu J, Kulkarni T D, et al. Self-supervised intrinsic image decomposition. In: Proceedings of Advances in Neural Information Processing Systems, 2017. 5936–5946
52 Baslamisli A S, Groenestege T T, Das P, et al. Joint learning of intrinsic images and semantic segmentation. In: Proceedings of the European Conference on Computer Vision (ECCV), 2018
53 Kim S, Park K, Sohn K, et al. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In: Proceedings of European Conference on Computer Vision. Berlin: Springer, 2016. 143–159
54 Gastal E S, Oliveira M M. Domain transform for edge-aware image and video processing. In: Proceedings of ACM SIGGRAPH 2011, 2011. 1–12
55 Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, 2015. 234–241
56 Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014. ArXiv:1409.1556
57 Narihira T, Maire M, Yu S X. Learning lightness from human judgement on relative reflectance. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015. 2965–2973
58 Zhang Y, Song S, Yumer E, et al. Physically-based rendering for indoor scene understanding using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 5287–5295
59 Wang J, Li X, Yang J. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. 1788–1797