GLoSH: Global-Local Spherical Harmonics for Intrinsic Image Decomposition
Supplementary Material

Hao Zhou[1] * Xiang Yu[2] David W. Jacobs[1]
[1] University of Maryland, College Park, MD, USA
[2] NEC Laboratories America
[1]{hzhou, djacobs}@cs.umd.edu [2]xiangyu@nec-labs.com

In this supplementary material, we first compare our method to [1] on NYUv2 [7] with more visual results in Section 1. As well, more visual results are compared to [3] and [6] in Section 2 on IIW. Section 3 shows the detail of our network structure. We provide the rendering equation of Spherical Harmonics lighting in Section 4. Limitations of the proposed method are discussed in Section 5.

1. More Results on NYUv2 [7]

Figure 1, Figure 2 and Figure 3 show more comparisons of our results to [1], expanding the results from Figure 5 in the main paper. The red rectangles in reflectance and shading images show that [1] mistakenly decomposes cast shadow into reflectance instead of shading, while the proposed method successfully avoids this problem.

*Hao Zhou is currently at Amazon AWS.

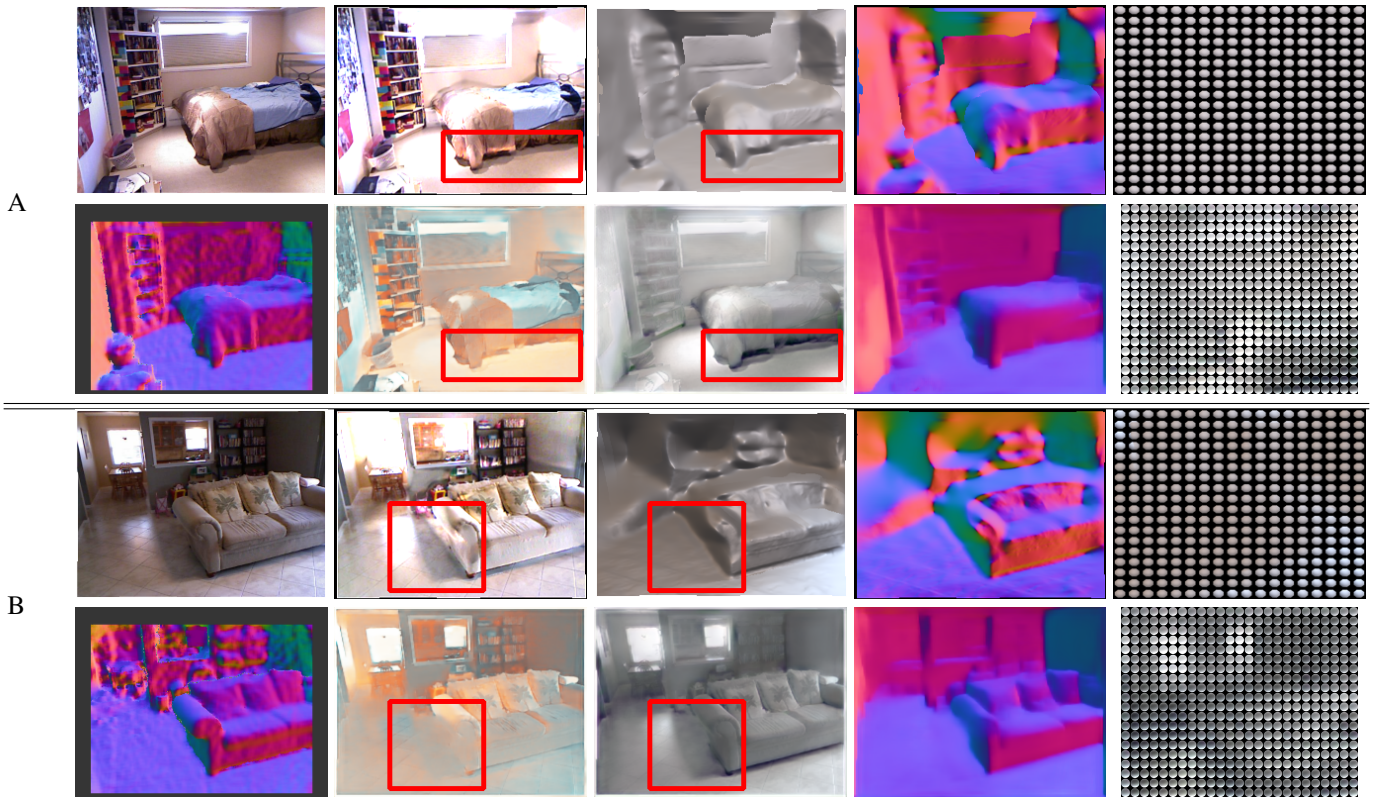

Figure 1. Comparison with [1]. First row of A and B from left to right: input image, reflectance by [1], shading by [1], normal by [1], lighting by [1]. Second row of A and B from left to right: ground truth normal; reflectance, shading, normal, local SH by the proposed method.

Figure 2. Comparison with [1]. First row of A, B, C and D from left to right: input image, reflectance by [1], shading by [1], normal by [1], lighting by [1]. Second row of A, B, C and D from left to right: ground truth normal; reflectance, shading, normal, local SH by the proposed method.

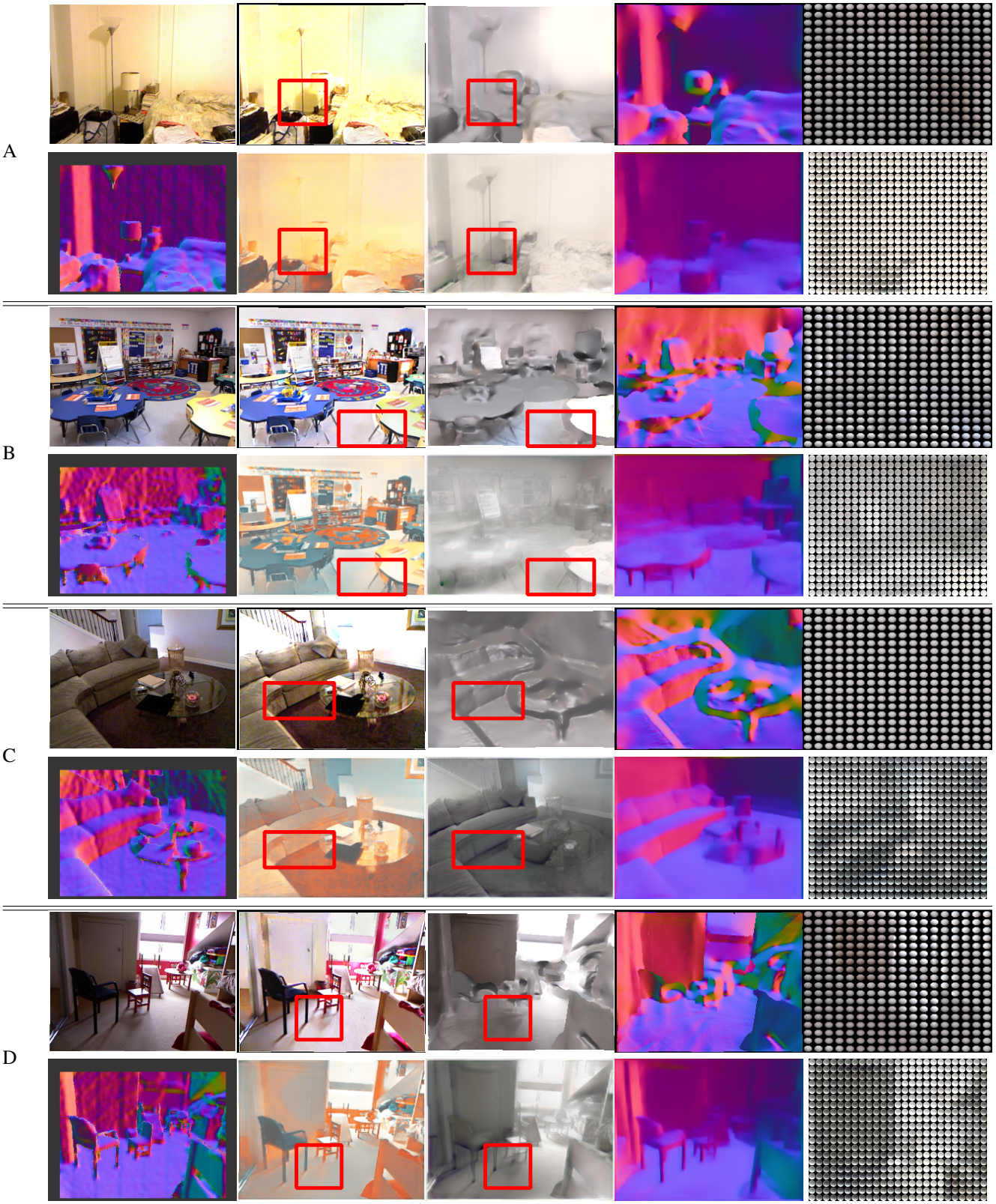

Figure 3. Comparison with [1]. First row of A, B, C and D from left to right: input image, reflectance by [1], shading by [1], normal by [1], lighting by [1]. Second row of A, B, C and D from left to right: ground truth normal; reflectance, shading, normal, local SH by the proposed method.

2. More Results Compared to [3] and [6]

Figure 4, Figure 5 and Figure 6 show more comparison of our results to [3] and [6], expanding results from Figure 6 of the main submission. First of all, the compared methods only decompose the input images into reflectance and shading, while our method decompose it further into reflectance, surface normal and lighting. Results from the compared methods show that the shading images predicted by [3] look like a gray scale version of the input images and that shading images predicted by [6] are of low contrast. Whereas our method predicts more reasonable shading images. This is mainly because we provide a physically reasonable decomposition on shading, *i.e*. decomposing it into surface normal and lighting, where lighting is further regularized with the non-negative constraint to be more physically realistic. We believe this decomposition can implicitly constrain the shading better.

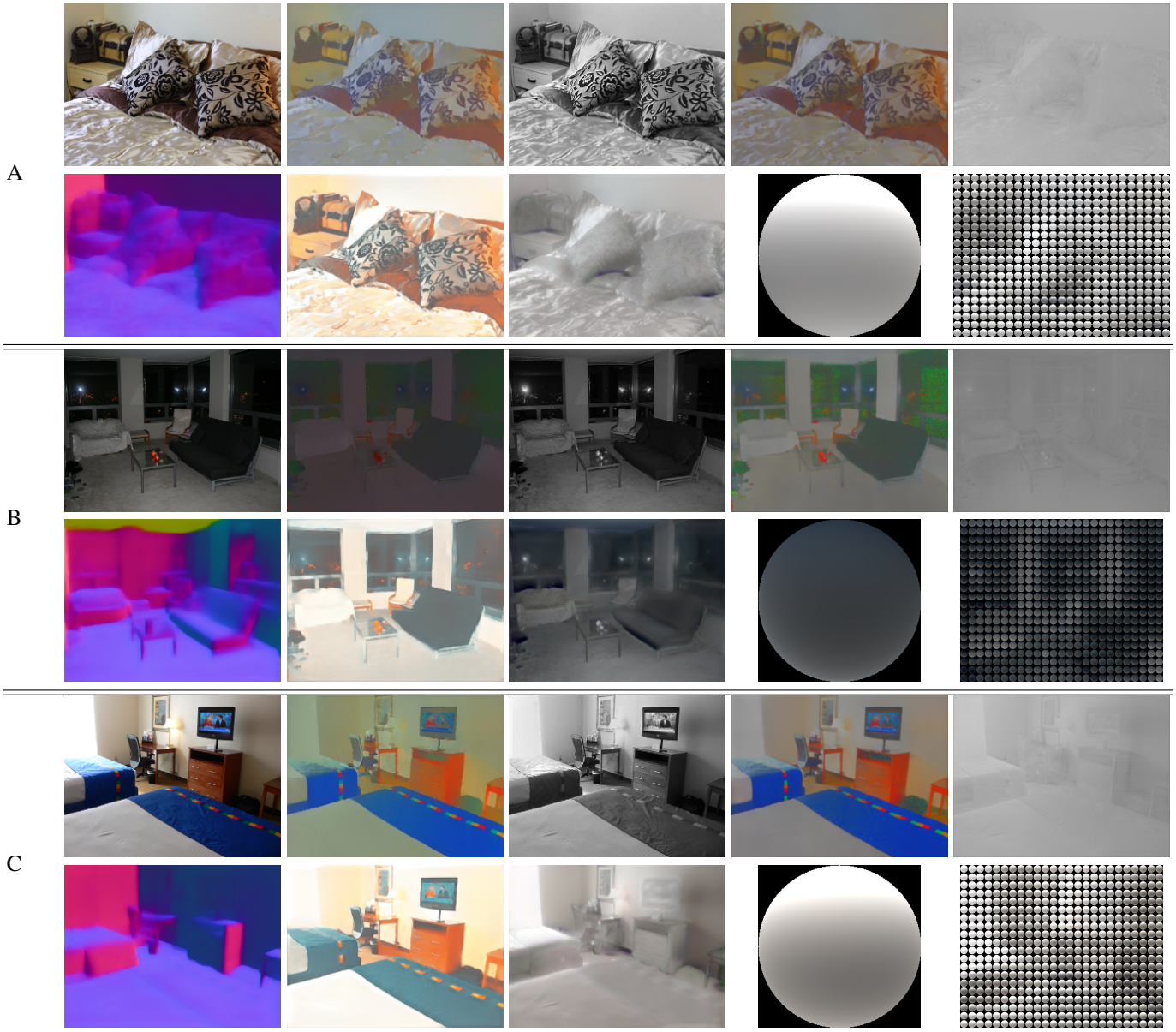

Figure 4. Comparison with state-of-the-art intrinsic image decomposition methods. First row of A, B and C from left to right: input image, reflectance by [3], shading by [3], reflectance by [6], shading by [6]. The second row of A, B and C from left to right: normal, reflectance, shading, global SH and local SH of the proposed method.

Figure 5. Comparison with state-of-the-art intrinsic image decomposition methods. First row of A, B, C and D from left to right: input image, reflectance by [3], shading by [3], reflectance by [6], shading by [6]. The second row of A, B, C and D from left to right: normal, reflectance, shading, global SH and local SH of the proposed method.

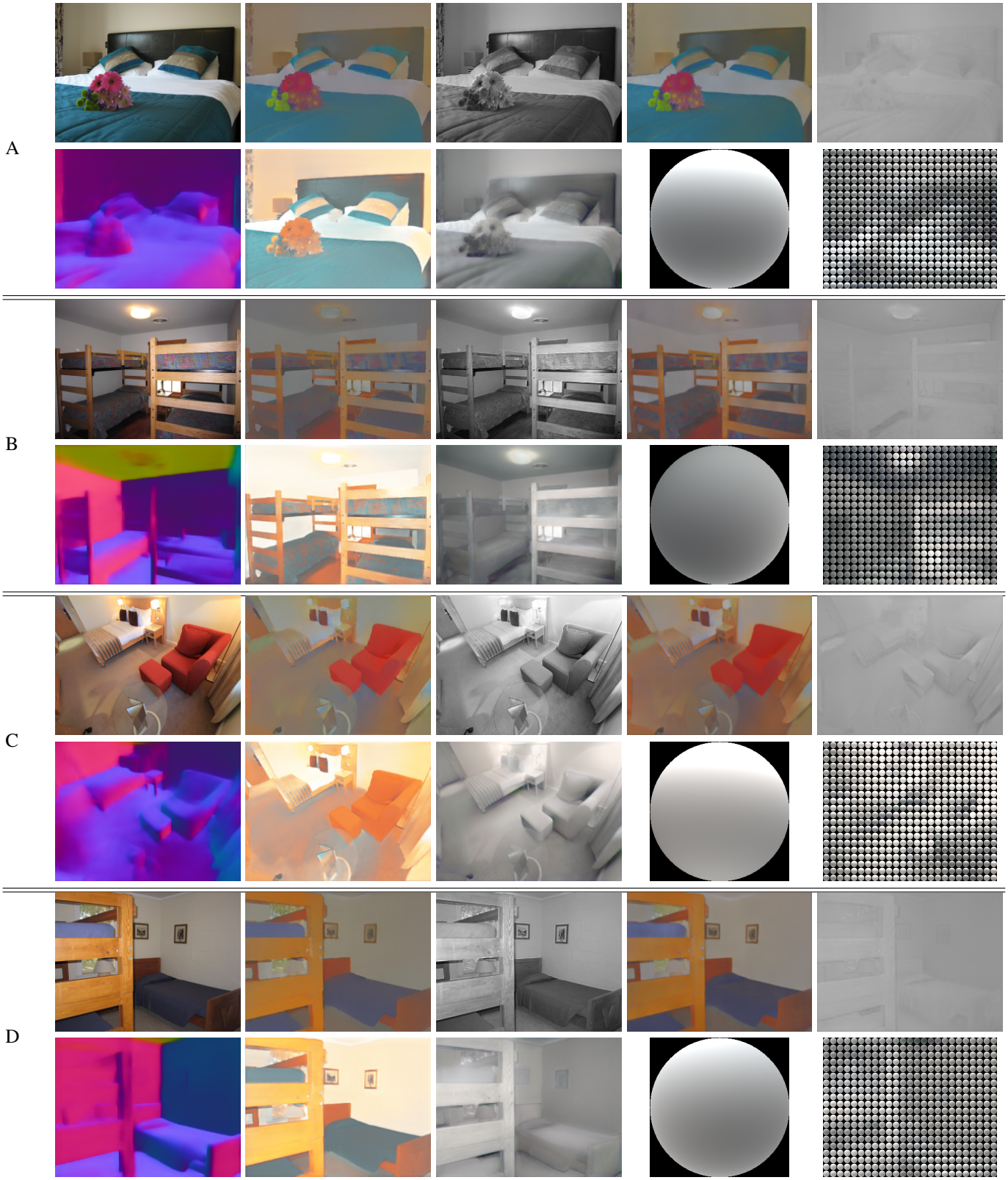

Figure 6. Comparison with state-of-the-art intrinsic image decomposition methods. First row of A, B, C and D from left to right: input image, reflectance by [3], shading by [3], reflectance by [6], shading by [6]. The second row of A, B, C and D from left to right: normal, reflectance, shading, global SH and local SH of the proposed method.

3. Network Structure

As discussed in our paper, we define a three scale coarse-to-fine network structure to naturally incorporate the global and local SH modeling as well as the multi-scale modeling of reflectance and surface normal. Figure 7 shows the framework of our method. Overall our framework can be divided into three scales. We employ a hourglass network [8] for the first scale network and a fully convolutional structure for second and third scale networks. Details are discussed in the following for the three scale network structures.

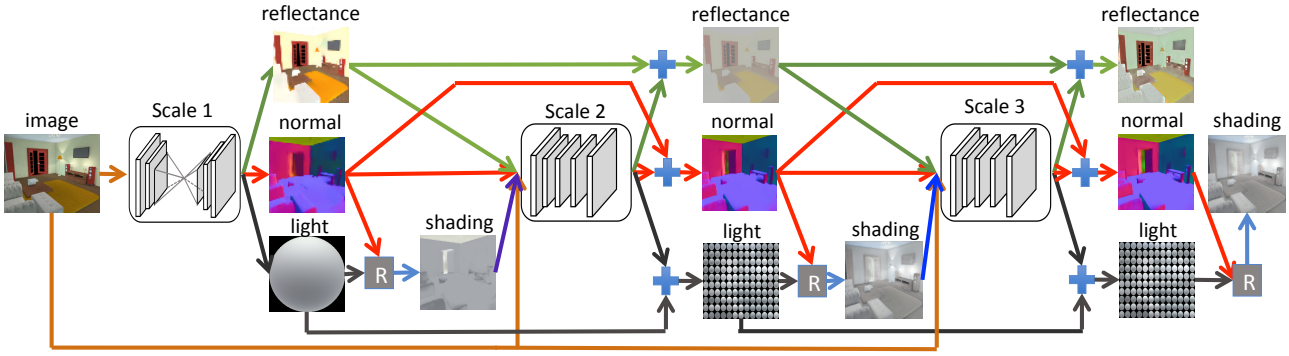

Figure 7. Framework of the proposed method. R means a renderer which takes normal and lighting as input and predicts shading.

Network of First Scale. The first scale network takes a 64×64 image as input and predicts 64×64 reflectance $\mathbf{R}_c$, normal $\mathbf{N}_c$, and a single global Lighting $\mathbf{L}_c$. Shading $\mathbf{S}_c$ can be constructed based on $\mathbf{N}_c$ and $\mathbf{L}_c$. The branches used to predict reflectance and normal have the same hourglass network structure [8], which is illustrated in Figure 8 (a). The branch to predict global SH is shown in Figure 8 (b). The green blocks are shared by all the three branches.

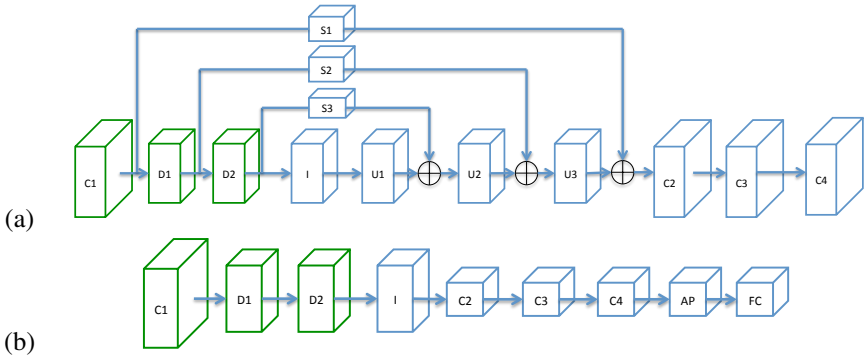

Figure 8. (a) shows the network structure to predict reflectance and normal and (b) shows the network structure for predicting global SH.

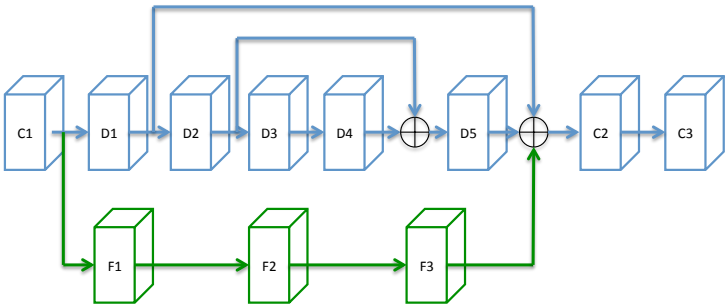

Figure 9. Network structure to predict reflectance, normal and lighting at second and third scale.

In Figure 8, C1, C2, C3, and C4 represent convolutional layers. D1, D2, I, U1, U2, U3, S1, S2, S3 are residual blocks defined in [4]. AP is an average pooling layer and FC represents a fully-connected layer. Each of the convolution layers is

followed by batch normalization [5] and ReLU except for the output layer. Table 3 shows the detailed definition of the block in Figure 8. Since we predict 9 Spherical Harmonics for each channel of the global SH, the number of output channels for global SH is 27.

Table 1. Details of each block in our network. (a) shows the details about each block in Figure 8 (a) and (b) shows the details about each block in Figure 8 (b).

		C1	C2	C3	C4	D1	D2	I	U1	U2	U3	S1	S2	S3	FC
(a)	input channel number	3	64	32	3	64	128	256	248	256	128	64	128	256	-
	output channel number	64	32	3	3	128	256	248	256	128	64	64	128	256	-
	filter size	5	3	3	3	3	3	3	3	3	3	3	3	3	-
	input feature size	64	64	64	64	64	32	16	8	16	32	64	32	16	-
	output feature size	64	64	64	64	32	16	8	16	32	64	64	32	16	-
(a)	input channel number	3	16	64	-	64	128	256	-	-	-	-	-	-	128
	output channel number	64	64	128	-	128	256	16	-	-	-	-	-	-	27
	filter size	5	3	3	-	3	3	3	-	-	-	-	-	-	-
	input feature size	64	8	4	-	64	32	16	-	-	-	-	-	-	-
	output feature size	64	4	2	-	32	16	8	-	-	-	-	-	-	-

Network of Second and Third Scale. Our second and third scale network has the same network structure. Our second network works on images with resolution 128×128 and our third network works on images with resolution 256×256. We define the network to predict residuals of reflectance, normal and lighting using separate networks with no shared layers.

The network structure is illustrated in Figure 9. C1, C2, C3, F1, F2, F3 are convolutional layers, D1, D2, D3, D4, D5 are residual blocks defined in [4]. Each convolutional layer is followed by a Batch Normalization layer [5] and ReLU except for the output layer.

Table 2 (a) shows the detailed definition about each block for networks used to predict residual reflectance and normal. Table 2 (b) shows the detailed definition about each block for networks used to predict local residual SH. For reflectance, we concatenate the image and the upsampled reflectance from the coarse network as input, so the number of input channels is 6. Similarly, we concatenate the image and upsampled normals from the coarse network as input for network for normals. The number of input channels is also 6. For the network used to predict local SH, we concatenate the image, upsampled normal, reflectance and shading as input. As a result, the number of input channels for local SH prediction is 12. Since we predict the color SH for each pixel, the number of output channels is 27.

Table 2. Details about each block in our second and third scale network shown in Figure 9. (a) shows the detailed structure for network used to predict reflectance and normal and (b) shows the detailed structure for the network used to predict lighting.

		C1	C2	C3	D1	D2	D3	D4	D5	F1	F2	F3
(a)	input channel number	6	32	3	32	32	32	64	32	16	32	32
	output channel number	16	3	3	32	32	64	32	32	32	32	32
	filter size	5	3	3	3	3	3	32	32	1	1	1
	dilation size	1	1	1	1	2	4	1	1	1	1	1
(b)	input channel number	12	32	3	32	32	32	64	32	16	32	32
	output channel number	16	3	27	32	32	64	32	32	32	32	32
	filter size	5	3	3	3	3	3	32	32	1	1	1
	dilation size	1	1	1	1	2	4	1	1	1	1	1

4. Rendering equation of Spherical Harmonics

We use Spherical Harmonics (SH) [2, 9] up to the second order, resulting in a 9 dimensional vector to represent lighting for each color channel. More specially, the 9 dimensional spherical harmonics in Cartesian coordinates of the surface normal $\mathbf{N} = (x, y, z)$ are:

$$Y_{00} = \frac{1}{\sqrt{4\pi}}$$

$$Y^o_{11} = \sqrt{\frac{3}{4\pi}}y \qquad Y_{10} = \sqrt{\frac{3}{4\pi}}z \qquad Y^e_{11} = \sqrt{\frac{3}{4\pi}}x$$

$$Y^o_{22} = 3\sqrt{\frac{5}{12\pi}}xy \qquad Y^o_{21} = 3\sqrt{\frac{5}{12\pi}}yz \qquad Y_{20} = \frac{1}{2}\sqrt{\frac{5}{4\pi}}(3z^2 - 1) \qquad Y^e_{21} = 3\sqrt{\frac{5}{12\pi}}xz \qquad Y^e_{22} = \frac{3}{2}\sqrt{\frac{5}{12\pi}}(x^2 - y^2)$$

Given a surface normal $\mathbf{N} = (x, y, z)$ and an SH lighting $\mathbf{L}$, the rendering equation Ψ defined in Equation 1 of the paper is:

$$\begin{aligned} r &= f(\mathbf{N}, \mathbf{L}) \\ &= \sum_{n=0}^{2} \sum_{m=-n}^{n} \alpha_n Y_{nm} l_{nm}, \end{aligned} \tag{1}$$

where $\alpha_n = \sqrt{\frac{4\pi}{2n+1}}$, and l_{nm} is the corresponding element in $\mathbf{L}$.

5. Limitation of the proposed method

The limitation of the proposed method is that Spherical Harmonics cannot model non-diffuse effects of the scene. For example, Figure 10 shows that the specular effect appears in the reflectance image. A more advanced model is needed to deal with non-diffuse effect of the scene, which is beyond the scope of this paper.

Figure 10. From left to right: input image, reflectance, shading, normal, global SH and local SH predicted by the proposed method. Note that the specular effect appears in the predicted reflectance.

References

[1] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, 2013. 1, 2, 3, 4

[2] Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *TPAMI*, 25(2), 2003. 10

[3] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018. 1, 5, 6, 7

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 8, 9

[5] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. 9

[6] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018. 1, 5, 6, 7

[7] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1

[8] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 8

[9] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: Determining the illumination from images of a convex lambertian object. *JOSA*, 2001. 10