Supplementary Material:
Lighting, Reflectance and Geometry Estimation from 360° Panoramic Stereo

In this Supplementary Material, we present videos, additional implementation details, and further results of our method. Section 1 describes the contents in the videos. The detailed architecture of our RN-Net is illustrated in Section 2 and Table 1. More comparisons on the reflectance and normal estimation are shown in Section 3. We also present more visualizations for ablation studies in Section 4.

1. Video for Illumination Estimation

Two videos are provided for showing our estimated illumination map and inserted mirror-objects. Some example screenshots of the videos are shown in Fig. 1. **We strongly encourage the reader to watch the attached video for a better appreciation of the performance of our method.**

In the video of scene 'barbershop', we compare our estimated lighting with Li *et al*. [2], Lighthouse [3], and the ground truth. Li *et al*. [2] proposed a method to per-pixel-independently estimate the lighting in the scene. Hence, it is easy to notice the inconsistency of their lighting between each frame. Li *et al*. [2] utilized a spherical Gaussian to represent the lighting in low-frequency. This representation of the lighting is not suitable for mirror-object insertion, as the mirror-object looks diffuse in their results. Lighthouse [3] took the scene geometry into consideration, to generate the spatially-coherent lighting at each location. However, they simplified the scene geometry to a coarse-to-fine model to save computational resources. Hence, their estimated lighting also lacks high-frequency structures. Besides, due to the limited field of view in their perspective input, their imagination of the unseen scene relies on the past data. In the results, they fail to recover the unseen region. Hence, the reflections on their mirror-object are far from satisfactory. Our method takes the 360° stereo input to fully observe the lighting and geometry of the entire scene. Note that our estimated illumination map changes smoothly under the constraints of scene geometry. We are also able to recover the high-frequency lighting in high quality. The 3D spatially-coherent and high-definition lighting both enable us to insert the moving mirror-object in the scene with realistic reflection effects.

In the video of scene 'hall', we demonstrate a mirror-sphere, relighted by our estimated illumination map, moving around the entire scene.

Figure 1. The above are screenshots from our two supplementary videos. The top compares our method to other methods on a synthetic scene 'barbershop' along with the ground truth. The bottom demonstrates insertion of a moving mirror-sphere in the real scene 'hall.' Videos are found in the attached `.mp4` files.

2. Details of RN-Net

The detailed architecture of RN-Net is illustrated in Table 1. The input is first processed by the Encoder. After four residual blocks in the Encoder, the output feature map at 4 is then fed to the two decoders for reflectance and normal estimation separately. When training the network, we found that the reflectance requires more layers to learn as

Encoder		
0	7×7 conv, k_0 features , stride 2	$H/2 \times W/2 \times k_0$
	3×3 max pooling, stride 2	$H/4 \times W/4 \times k_0$
	$(3 \times 3$ conv, k_1 features$) \times 2$, residual	$H/4 \times W/4 \times k_1$
1	$(3 \times 3$ conv, k_1 features$) \times 2$, residual	$H/4 \times W/4 \times k_1$
	$(3 \times 3$ conv, k_2 features$) \times 2$, stride 2, residual	$H/8 \times W/8 \times k_2$
2	$(3 \times 3$ conv, k_2 features$) \times 2$, residual	$H/8 \times W/8 \times k_2$
	$(3 \times 3$ conv, k_3 features$) \times 2$, stride 2, residual	$H/16 \times W/16 \times k_3$
3	$(3 \times 3$ conv, k_3 features$) \times 2$, residual	$H/16 \times W/16 \times k_3$
	$(3 \times 3$ conv, k_4 features$) \times 2$, stride 2, residual	$H/32 \times W/32 \times k_4$
4	$(3 \times 3$ conv, k_4 features$) \times 2$, residual	$H/32 \times W/32 \times k_4$

Reflectance Decoder			Normal Decoder		
	$2\times$ bilinear upsample	$H/16 \times W/16 \times k_4$		$2\times$ bilinear upsample	$H/16 \times W/16 \times k_4$
	$(3 \times 3$ conv, k_3 features$) \times 2$, residual	$H/16 \times W/16 \times k_3$			
r_1	$(3 \times 3$ conv, k_3 features$) \times 2$, residual	$H/16 \times W/16 \times k_3$	n_1	$(3 \times 3$ conv, k_3 features$) \times 2$, residual	$H/16 \times W/16 \times k_3$
	Add 3 and r_1	$H/16 \times W/16 \times k_3$		Add 3 and n_1	$H/16 \times W/16 \times k_3$
	$2\times$ bilinear upsample	$H/8 \times W/8 \times k_3$		$2\times$ bilinear upsample	$H/8 \times W/8 \times k_3$
	$(3 \times 3$ conv, k_2 features$) \times 2$, residual	$H/8 \times W/8 \times k_2$			
r_2	$(3 \times 3$ conv, k_2 features$) \times 2$, residual	$H/8 \times W/8 \times k_2$	n_2	$(3 \times 3$ conv, k_2 features$) \times 2$, residual	$H/8 \times W/8 \times k_2$
	Add 2 and r_2	$H/8 \times W/8 \times k_2$		Add 2 and n_2	$H/8 \times W/8 \times k_2$
	$2\times$ bilinear upsample	$H/4 \times W/4 \times k_2$		$2\times$ bilinear upsample	$H/4 \times W/4 \times k_2$
	$(3 \times 3$ conv, k_1 features$) \times 2$, residual	$H/4 \times W/4 \times k_1$			
r_3	$(3 \times 3$ conv, k_1 features$) \times 2$, residual	$H/4 \times W/4 \times k_1$	n_3	$(3 \times 3$ conv, k_1 features$) \times 2$, residual	$H/4 \times W/4 \times k_1$
	Add 1 and r_3	$H/4 \times W/4 \times k_1$		Add 1 and n_3	$H/4 \times W/4 \times k_1$
	$2\times$ bilinear upsample	$H/2 \times W/2 \times k_1$		$2\times$ bilinear upsample	$H/2 \times W/2 \times k_1$
	$(3 \times 3$ conv, k_0 features$) \times 2$, residual	$H/2 \times W/2 \times k_0$			
r_4	$(3 \times 3$ conv, k_0 features$) \times 2$, residual	$H/2 \times W/2 \times k_0$	n_4	$(3 \times 3$ conv, k_0 features$) \times 2$, residual	$H/2 \times W/2 \times k_0$
	Add 0 and r_4	$H/2 \times W/2 \times k_0$		Add 0 and n_4	$H/2 \times W/2 \times k_0$
	$2\times$ bilinear upsample	$H \times W \times k_0$		$2\times$ bilinear upsample	$H \times W \times k_0$
	$(3 \times 3$ conv, k_5 features$) \times 2$, residual	$H \times W \times k_5$			
r_5	$(3 \times 3$ conv, k_5 features$) \times 2$, residual	$H \times W \times k_5$	n_5	$(3 \times 3$ conv, k_5 features$) \times 2$, residual	$H \times W \times k_5$
Reflectance	3×3 conv, 3 features	$H \times W \times 3$	Normal	3×3 conv, 3 features	$H \times W \times 3$

Table 1. Network architecture of RN-Net. All convolutional layers use a ReLu activation and Batch normalization, except for the prediction layer. For RN-Net in original scale Φ_1, we set $H = 512, W = 1024$ and $k_{0,\dots,5} = [10, 16, 32, 64, 128, 10]$. For RN-Net in small scale $\Phi_{\frac{1}{4}}$, we set $H = 128, W = 256$ and $k_{0,\dots,5} = [64, 64, 128, 256, 512, 32]$.

Figure 2. The 360° stereo camera setup we use for capturing the real scene.

it contains more high-level features than normal estimation. Hence, we reduce the layers of normal decoder by half to speed up the training and reduce overfitting.

3. Reflectance and Normal Estimation

As shown in Fig. 3, our results are quantitatively better than all the competing methods. Besides, the errors are only computed based on the cropped region of the image for fair comparison. It is worth mentioning that our method also performs well outside the cropped regions.

We test our method on public real scene 'room' and 'hall', provided by 360SD-Net [4] [1], as shown in Fig. 4.

4. Ablation Study

We provide more visualizations for the comparison on the ablated versions of our method in Figs. 5 and 6. From top to bottom, the row denotes the 360° input, origin RN-Net, pyramid RN-Net, and our full method with the rendering and total variation refinement, respectively. The scale-invariant mean-square-error (sMSE) and mean angular error

[1] 360SD-Net [4] used a similar 360° stereo setup to capture the real data. Their data can be acquired at https://github.com/albert100121/360SD-Net.

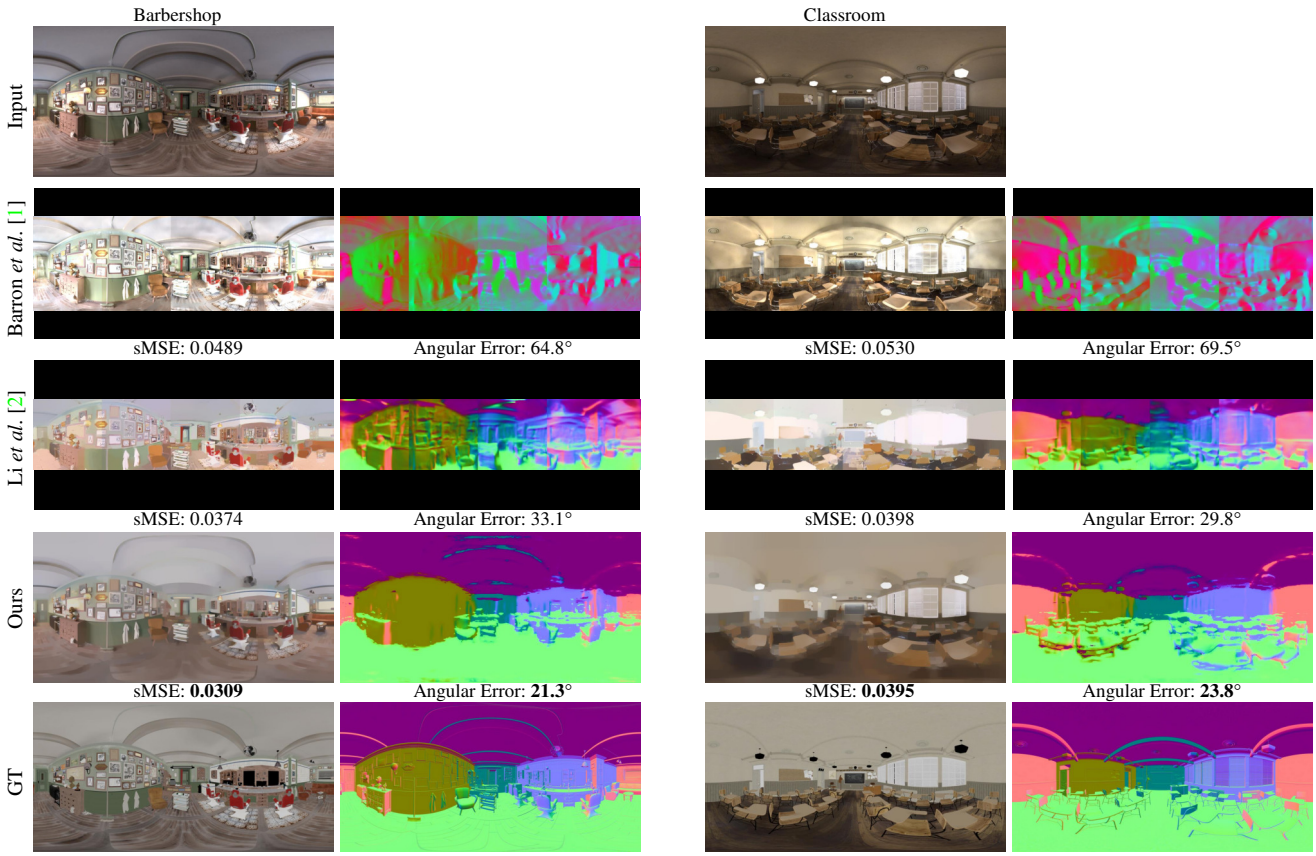

Figure 3. Comparison to other methods: estimated reflectance and normal on synthetic scene 'barbershop' and 'classroom'. The scale-invariant mean-square-error (sMSE) and mean angular error shown in the bottom is evaluated on the cropped regions for all the methods. **Better view on screen with zoom-in.**

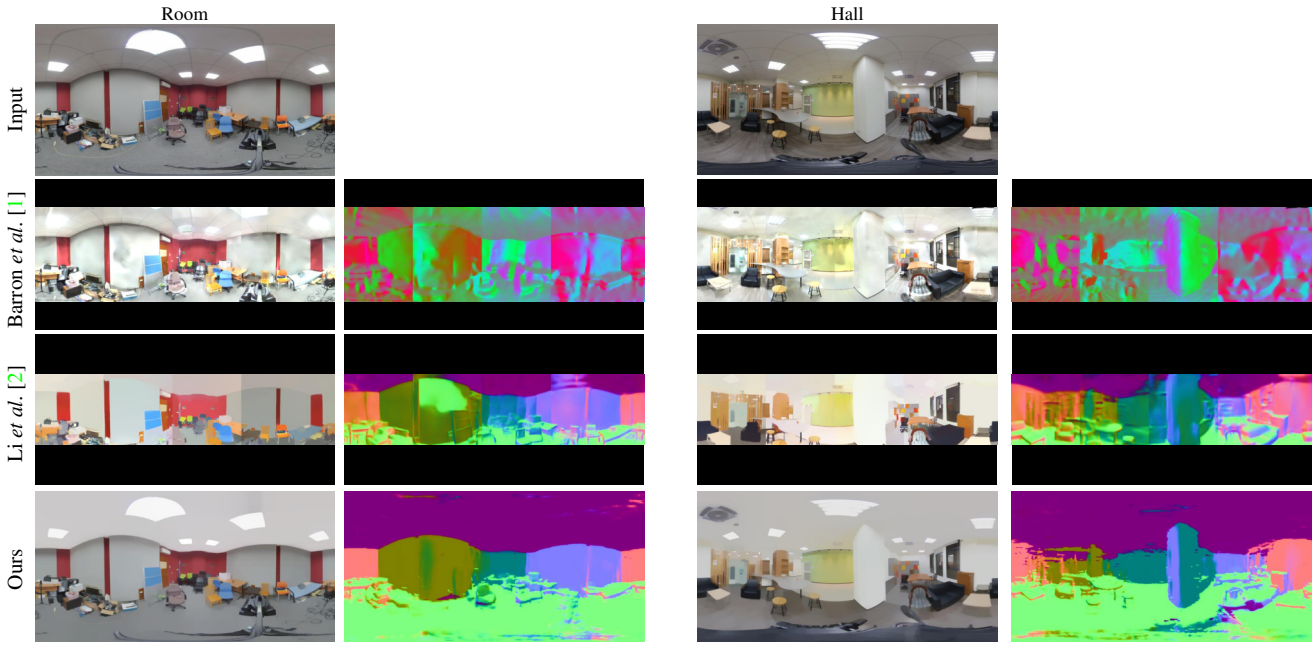

Figure 4. Comparison to other methods: estimated reflectance and normal on real scene 'room' and 'hall'. **Better view on screen with zoom-in.**

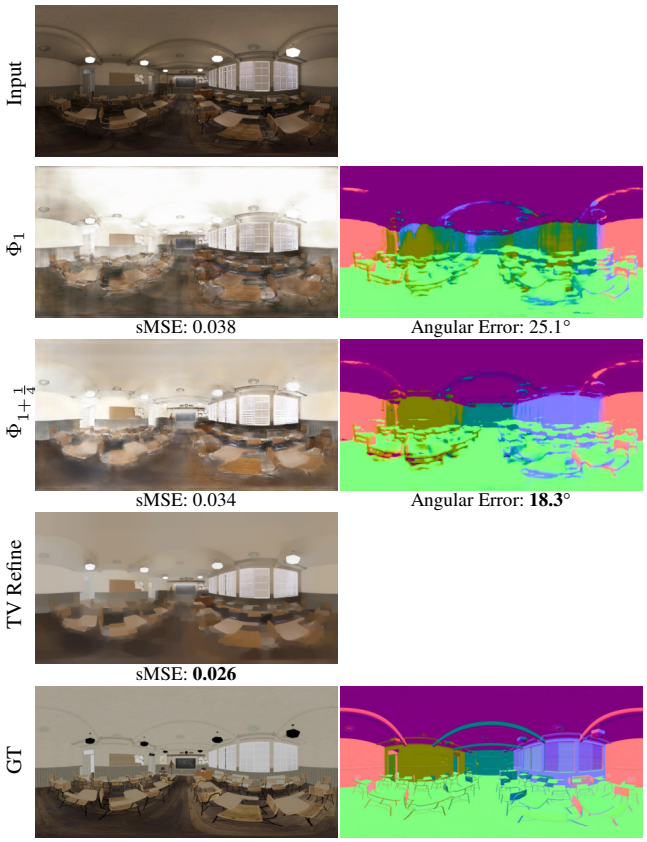

Figure 5. Ablation study: estimated reflectance and normal of the ablated versions of our method on synthetic scene 'classroom'.

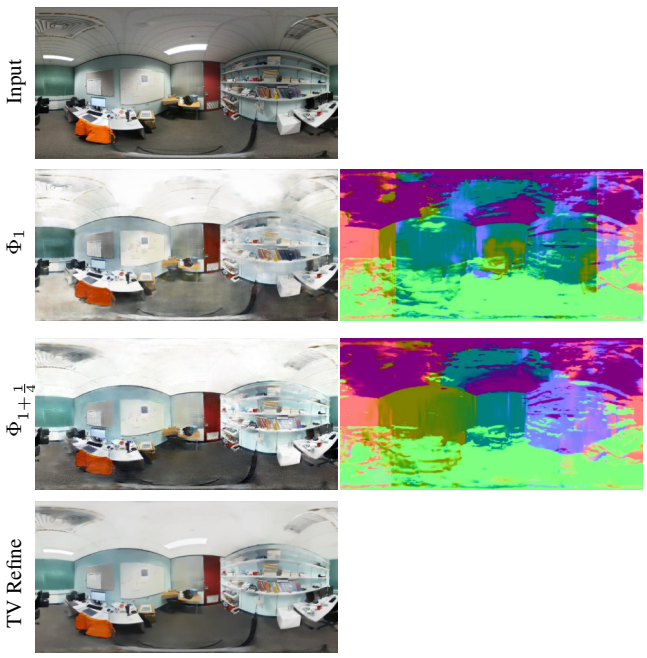

Figure 6. Ablation study: estimated reflectance and normal of the ablated versions of our method on the real scene 'office'.

is computed on the full 360° images in ablation study. Both figures demonstrate that the pyramid structure improves reflectance and normal. It is also clear that the rendering and refinement module can effectively reduce the noise and outliers of the reflectance map.

References

[1] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 3

[2] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 1, 3

[3] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 1

[4] Ning-Hsu Wang, Bolivar Solarte, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 360sd-net: 360 stereo depth estimation with learnable cost volume. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 582–588. IEEE, 2020. 2