Learning to Separate Multiple Illuminants in a Single Image
(Supplementary Material)

Zhuo Hui,[1] Ayan Chakrabarti,[2] Kalyan Sunkavalli,[3] and Aswin C. Sankaranarayanan[1]
[1]Carnegie Mellon University [2]Washington University in St. Louis [3]Adobe Research

(a) Input photograph (b) Light editing results

Figure 1: Our technique is able to separate a single photograph (a) into individual light components. Given that, we recolor each separated image to create photo-realistic results with novel lighting conditions (b). Our technique does not require any additional scene information, and can realistically retain all shading effects while changing the color and brightness of each light source.

List of Figures

1. Light Editing

Our method is able to separate a single photograph into separated images, each of which illuminated by a single light source. This enables us to adjust the brightness as well as the color chromaticity of a particular light simply by operating on the corresponding separated results. In Figure 1, we showcase the results by editing on the light colors and brightness. Note that our method is able to retain the shading of the illuminant while changing its color and brightness. In comparison to the past techniques utilizing global color transformation, our method enables the color adjustments for a local light source. As can be seen in Figure 1 (bottom), we can adjust the highlight on the floor with light yellow color while making indoor light being light blue.

2. Network Architectures

We provide details on the architecture of networks used in the paper.

ChromNet. We adopt an architecture similar to that of Johnson et al. [5] to map the input image to a three channel reflectance chromaticity map. Specifically, the input first goes through the block, which consists of

$$\texttt{Reflection} - (\texttt{C-f7s1-o64}) - \texttt{Norm} - \texttt{ReLu},$$

where `C-f7s1-o64` denotes a 7×7 stride-1 convolution layer with 64 output channels, `Reflection` denotes the layer that pads the boundary of the input with reflection (`ReflectionPad2d` in PyTorch) and `Norm` denotes the batch normalization layer.

Next, we feed the stream into a series of two convolutional blocks, each of which consists of

$$\begin{aligned}&(\texttt{C-f3s2-o128}) - \texttt{Norm} - \texttt{ReLu}-\\&(\texttt{C-f3s2-o256}) - \texttt{Norm} - \texttt{ReLu}.\end{aligned}$$

After that, the stream goes into a series of 9 residual blocks, each with

$$(\texttt{C-f3s1-o256}) - \texttt{Norm} - (\texttt{C-f3s1-o256}) - \texttt{Norm},$$

We then utilize two up-sampling blocks, each of which performs nearest neighbor up-sampling by a factor of 2 followed by a 3×3 convolution

$$\texttt{UpSample(2)} - (\texttt{C-f3s1-o128}) - (\texttt{Norm}) - (\texttt{ReLu})-$$
$$-\texttt{UpSample(2)} - (\texttt{C-f3s1-o64}) - (\texttt{Norm}) - (\texttt{ReLu}),$$

where `UpSample(2)` is the up-sampling layer with the factor of 2.

Finally, we produce the reflectance chromaticity estimates via

$$\texttt{Reflection} - (\texttt{C-f7s1-o3}) - \texttt{Tanh}.$$

ShadingNet. For the ShadingNet, we expect it to take reflectance chromaticity estimates as inputs, and solve for the two illuminant shadings. As shown in the derivations in Section 3 in the main paper, we need to involve deriving γ from the chromaticities and the original input on a purely per-pixel basis. To facilitate this in a pixel-wise manner, we first concatenate the input image and reflectance chromaticity in the color channel and then process the input with

$$\begin{aligned}&(\texttt{C-f1s1-o16}) - \texttt{LeakyReLu}-\\&(\texttt{C-f1s1-o16}) - \texttt{Norm} - \texttt{LeakyReLu}.\end{aligned}$$

After that, we adopt a variant of the U-Net like architecture, similar to the network of Isola et al. [4], with a single encoder while two decoders producing the illuminant shading maps. That is, the stream first goes into the encoder, which is

$$\texttt{CB64}-\texttt{CB128}-\texttt{CB256}-\texttt{CB512}-\texttt{CB512}-\texttt{CB512}-\texttt{CB512},$$

where `CBk` denotes the block with output channel k, which consists of a 4×4 stride-2 convolution layer followed by `Norm` and `LeakyReLu` layer. Note that the number of input channel of `CBk` is equal to the number of output channels in the previous block, and the input channel of the first block is set as 16.

The two decoders with identical architecture receive skip-connections from intermediate layers of the encoder, each of which consists of

$$\texttt{CB512}-\texttt{CB512}-\texttt{CB512}-\texttt{CB512}-\texttt{CB256}-\texttt{CB128}-\texttt{CB64},$$

where `CBk` denotes the block, which consists of an `UpSample(2)` followed by a 3×3 stride-1 convolution with `Norm` and `LeakyReLu` layer.

Finally, we append `C-f1s1-o3` to the end of each decoder to output the illuminate shadings.

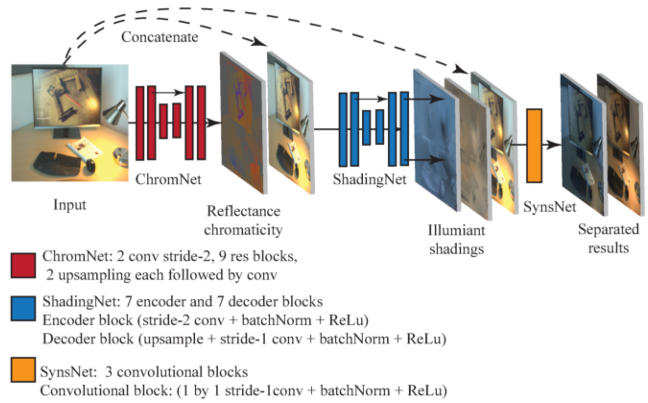


Figure 2: Our proposed three sub-networks.

SeparateNet. For the SeparateNet, we use several pixel-wise layers to express the computation in (7) in the main paper. In particular, our third sub-network concatenates the two predicted shading maps and the input RGB photograph into a nine-channel input, and uses

$$\begin{aligned}&(\texttt{C-f1s1-o64}) - \texttt{LeakyReLU}-\\&(\texttt{C-f1s1-o64}) - \texttt{Norm} - \texttt{LeakyReLu}.\end{aligned}$$

Finally, we append `C-f1s1-o6` to the end to produce the separated images.

3. Results on Synthetic Benchmark Dataset

In this section, we showcase the performance of variants of networks on the synthetic benchmark dataset. In particular, we evaluate against the following baselines:

- (**SingleNet**) that directly predicting the separated images,
- (**Chrom-only**), the first sub-network, i.e., ChromNet, that only supervises chromaticity,
- (**Final-Only**), the full network where we train with supervision on only the final output,
- (**Full-Direct**) where we supervise chromaticity, shading and the separated images,
- (**Full+[3]**). where we take only the reflectance chromaticity estimates of the Full-Direct trained model and use Hui et al.'s algorithm [3] for the actual separation.

We also compare the proposed method against the state-of-the-art intrinsic image decomposition techniques [7, 1, 6]. To produce the separated images, we first compute the reflectance by using these methods and then utilize the reflectance chromaticity to separate illuminants with [3]. We also compare against two physical-based approaches, which require additional information beyond a single image. For these methods, we provide the ground truth light colors for Hsu et al. [2] and ground truth reflectance chromaticity for Hui et al. [3]. Note that all the scenes in Figures 3, 4 and 5, are realistic rendered in high quality with complex scene geometries as well as intricate interactions of objects and light rays. We incorporate the signal-to-noise ratio (SNR) to characterize the performance of each method.

Generally, the qualitative performance of these techniques on this dataset closely parallels the quantitative results we observe in Table 1 of the main paper. As can be seen in Figures 3, 4 and 5, both SingleNet and our Final-Only model fail to separate the effects of illuminant shading from the input. The Chrom-Only model yields a better result while has severe artifacts in certain regions. In contrast, the results from our model with full supervision yields the best performance and even better separation of shadow and shading effects when we use its chromaticity outputs in conjunction with [3].

For the intrinsic image techniques, we obverse that these methods fail to separate the illumination by the nature that most intrinsic image methods assume a single light source in the scene.

For the physical-based approach, we find that the proposed architecture (**Full+[3]**) returns better visual performance than that of Hsu et al. [2]. While the method of Hui et al. [3] yields the high-quality separation, it requires the knowledge of the reflectance chromaticity which always results in extra images captured for the scene.

We find that our network (**Full+[3]**) is able to match the quality of the results from Hui et al.'s two-image method [3] as well as producing the results that closely resemble to the ground truth, despite needing only a single image.

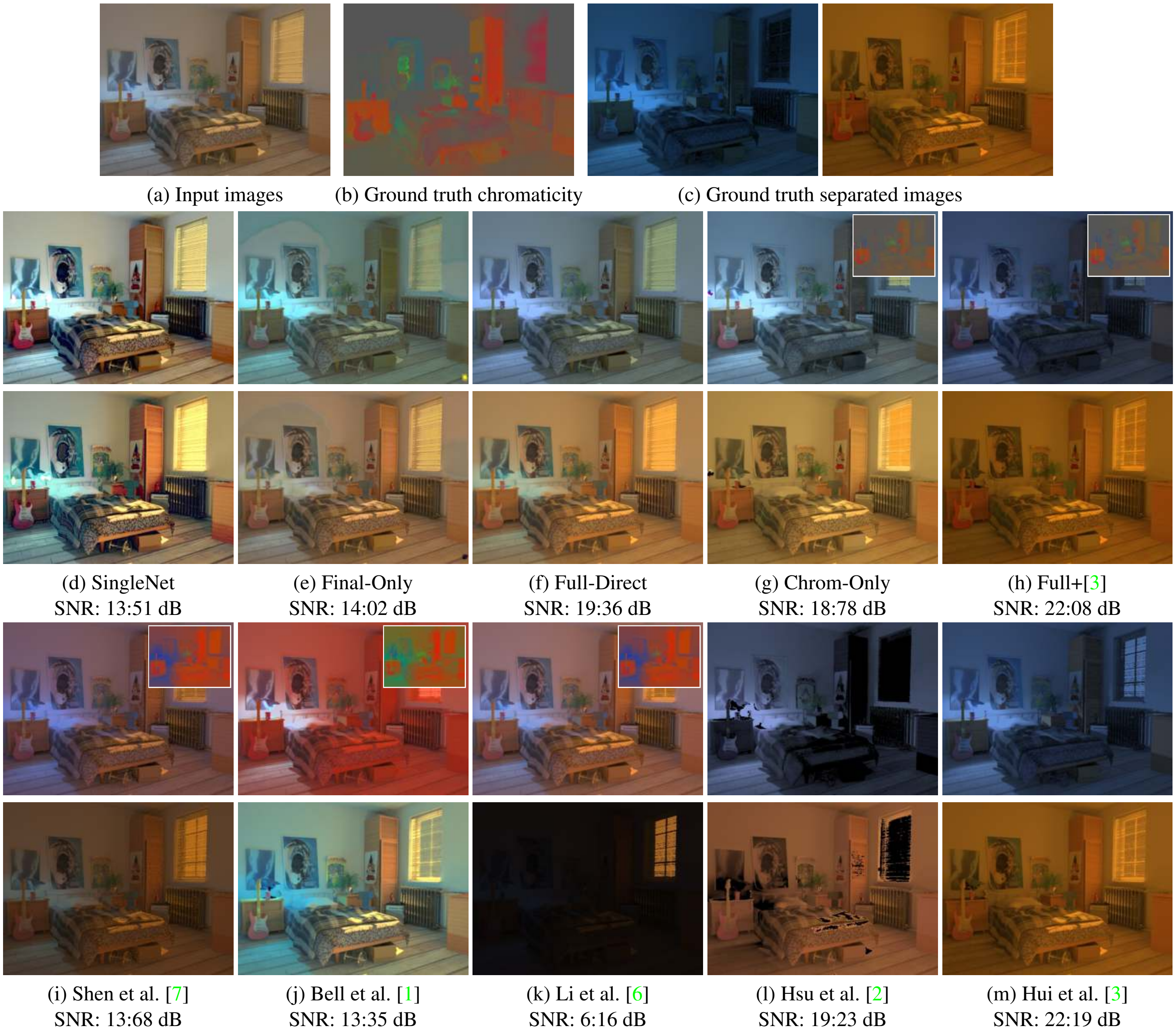
4. Results on Real Dataset

We characterize the performance of our technique on both indoor and outdoor scenes.

Qualitative results on indoor scenes. In Figure 6, we demonstrate our technique on the scene with two lights sources and compare with ground truth captures. The ground truth photographs were captured by Hui et al. [3], where they turn off the indoor illuminants and capture the only outdoor illuminated scene. Given that, they subtract the only outdoor illuminated image from the photograph under both outdoor and indoor illumination to obtain the photograph with respect to the indoor light sources. We evaluate our technique against the flash/no-flash method of Hui et al. [3]. We observe that our technique achieves slightly better results in terms of SNR measurement against Hui et al. [3], and produces results that closely resemble to the ground truth, despite requiring only a single image input. We showcase additional comparisons to Hui et al. [3] in Figure 7. While the method of [3] returns high-quality results, it requires a pair of flash/no-flash images for the scene. In comparison to [3], our method relies on a single photograph while being able to match the performance, and in specific instances outperforms, the quality of results from Hui et al.'s two-image method [3] on the real scenes.

Qualitative results on outdoor scenes. We compare the performance against [3] on the time lapse sequences in Figure 8. In comparison to Hui et al. [3], our method avoids the need to identify a photograph with cloudy sky as a pure flash. We know that skylight changes its color and intensity significantly during the course of the day, leading the estimation by [3] to incorrect shadings. This can be seen in Figure 8, where the separated images Figure 8 (b) under the skylight involve the shadings induced by the sunlight. In contrast, the proposed technique takes the color chromaticity estimated for each frame and results in better separation in the illumination shadings as well as better visual quality in the resulting images.

Evaluation on three-light Scenario. Since our method utilizes the color chromaticity estimation to separate the light sources, like [3], we are able to separate up to three light sources in the scene. Figure 9 showcases the performance of our technique against the ground truth and the method of [3]. We measure the SNR values against the ground truth for both methods. As can be seen, our method



(a) Input images (b) Ground truth chromaticity (c) Ground truth separated images

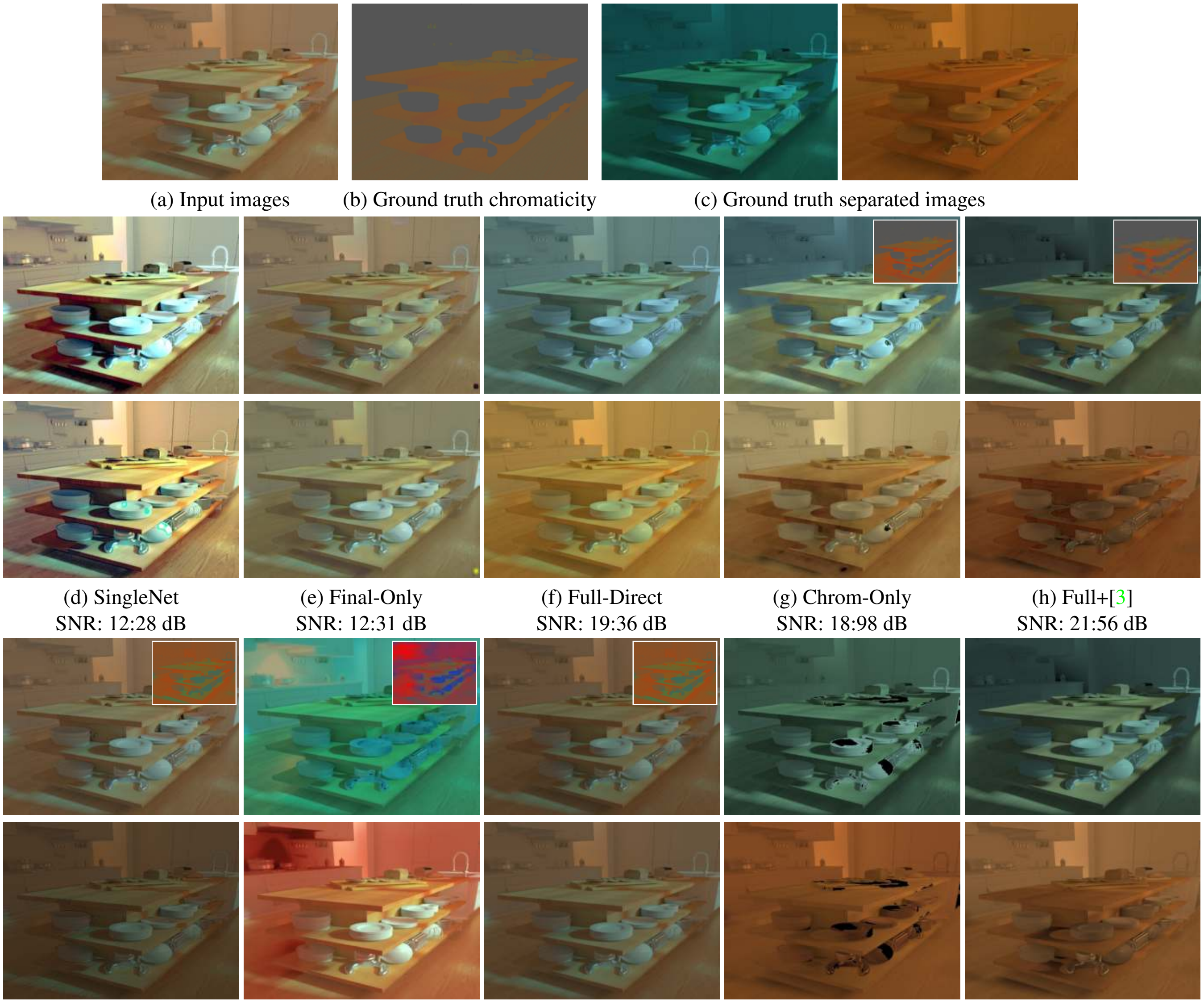
(d) SingleNet SNR: 13:51 dB (e) Final-Only SNR: 14:02 dB (f) Full-Direct SNR: 19:36 dB (g) Chrom-Only SNR: 18:78 dB (h) Full+[3] SNR: 22:08 dB

(i) Shen et al. [7] SNR: 13:68 dB (j) Bell et al. [1] SNR: 13:35 dB (k) Li et al. [6] SNR: 6:16 dB (l) Hsu et al. [2] SNR: 19:23 dB (m) Hui et al. [3] SNR: 22:19 dB

Figure 3: We compare variants of proposed network architectures, the state-of-art intrinsic image decomposition methods [7, 1, 6] as well as the physical-based approaches [2, 3] against the ground truth on the synthetic benchmark dataset. We showcase the separated images together with the estimated reflectance chromaticity (insets) if available for the method. We find that the version of our network trained with full supervision in combination with [3] (Full+[3]) performs best, and is able to return the results that closely resemble to the ground truth, despite needing only a single image.

is able to closely match the performance of [3] in terms of both quantitative measurements and visual results. We note that small artifacts (on the left arm of the bear (c)) appear in our separated images, likely since three-light scenario is out of the training distribution. We also include an example that our method fails to separate the illumination in Figure 10. We observe that the scene is under complex light transport for three lights, making our method unable to identify the reflectance chromaticity of the scene.

Qualitative results on white balanced results. We also incorporate the comparisons on the white balanced results on the real scenes in Figure 11. As shown in Figure 11 (top row), our method is able to match the performance of Hui et al.'s method [3] by using a flash/no-flash pair. We also include an example (bottom row) in Figure 11, where the flash is not able to reach out the scene in the far end. As expected, Hui et al.'s method [3] fails to return good performance while the proposed method still retain pleasing



(a) Input images (b) Ground truth chromaticity (c) Ground truth separated images

(d) SingleNet SNR: 12:28 dB | (e) Final-Only SNR: 12:31 dB | (f) Full-Direct SNR: 19:36 dB | (g) Chrom-Only SNR: 18:98 dB | (h) Full+[3] SNR: 21:56 dB

(i) Shen et al. [7] SNR: 12:53 dB | (j) Bell et al. [1] SNR: 11:15 dB | (k) Li et al. [6] SNR: 10:76 dB | (l) Hsu et al. [2] SNR: 17:88 dB | (m) Hui et al. [3] SNR: 21:87 dB

Figure 4: Additional results on synthetic benchmark dataset.

visual quality, providing the practical utility of the method.

References

[1] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *TOG*, 33(4):159, 2014. 3, 4, 5, 6

[2] Eugene Hsu, Tom Mertens, Sylvain Paris, Shai Avidan, and Fredo Durand. Light mixture estimation for spatially varying white balance. In *TOG*, volume 27, page 70, 2008. 3, 4, 5, 6

[3] Zhuo Hui, Kalyan Sunkavalli, Sunil Hadap, and Aswin C. Sankaranarayanan. Illuminant spectra-based source separation using flash photography. In *CVPR*, 2018. 1, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 2

[5] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[6] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *ECCV*, 2018. 3, 4, 5, 6

[7] Jianbing Shen, Xiaoshan Yang, Xuelong Li, and Yunde Jia. Intrinsic image decomposition using optimization and user scribbles. *IEEE Transactions on Cybernetics*, 43(2):425–436, 2013. 3, 4, 5, 6

(a) Input images (b) Ground truth chromaticity (c) Ground truth separated images

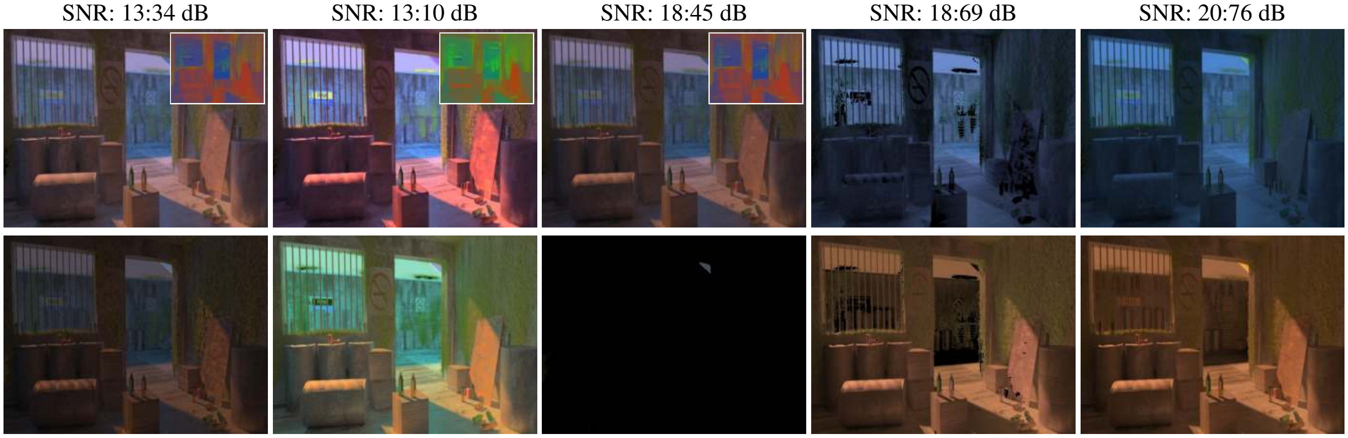
(d) SingleNet
SNR: 13:34 dB

(e) Final-Only
SNR: 13:10 dB

(f) Full-Direct
SNR: 18:45 dB

(g) Chrom-Only
SNR: 18:69 dB

(h) Full+[3]
SNR: 20:76 dB



(i) Shen et al. [7]
SNR: 13:86 dB

(j) Bell et al. [1]
SNR: 12:95 dB

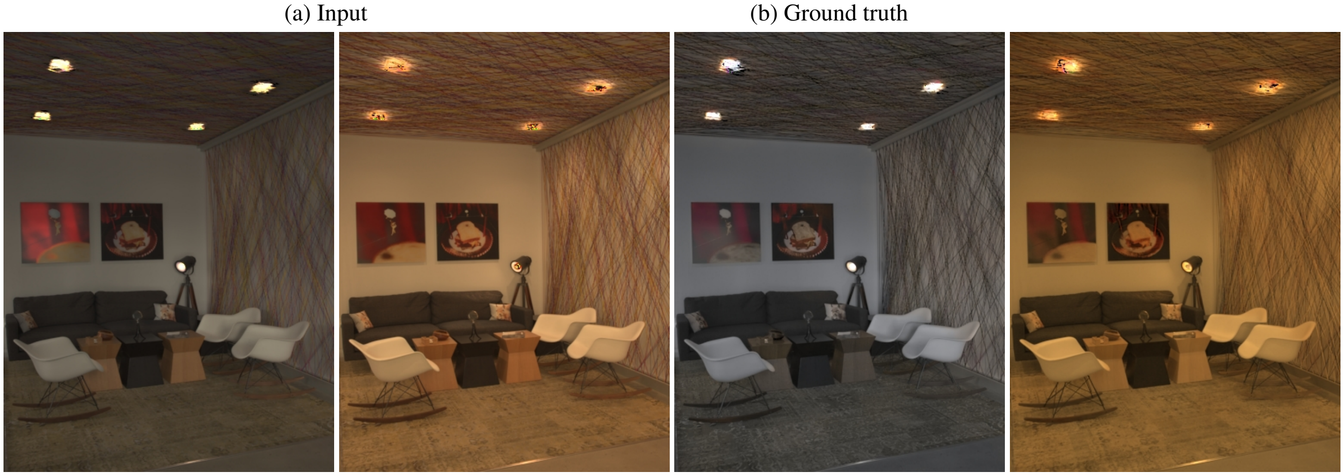
(k) Li et al. [6]
SNR: 5:49 dB

(l) Hsu et al. [2]
SNR: 17:49 dB

(m) Hui et al. [3]
SNR: 20:81 dB

Figure 5: Additional results on synthetic benchmark dataset.

(a) Input (b) Ground truth



(c) Hui et al. [3]
SNR: 16:78 dB

(d) Our results
SNR: 17:13 dB

Figure 6: We evaluate our technique against Hui et al. [3] for the scene (a) with ground truth (b). As can be seen here, despite requiring a single input photograph, the proposed technique is able to match the performance of Hui et al. [3] which takes flash/no-flash images and closely mimics the actual captured results.

(a) Input images

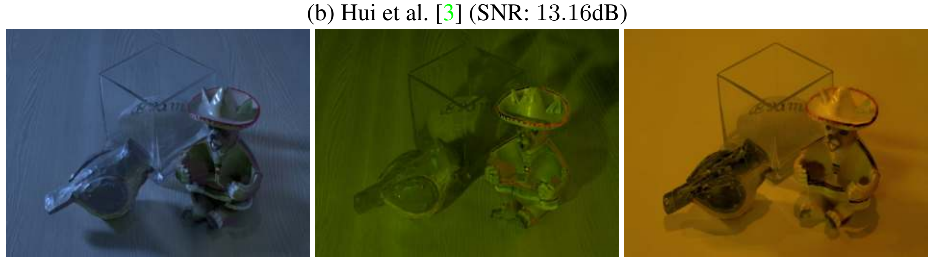
(c) Hui et al. [3] (d) Our results

Figure 7: Additional comparison results against [3] on real scenes. As can be seen here, we can achieve nearly the same visual quality as Hui et al. [3], which captures two photographs for the same scene. It is worth noticing that the proposed technique uses a single photograph, thus yielding more practical solution to the lighting separation problem.

(a) Input frame (b) Hui et al. [3] (c) Our results

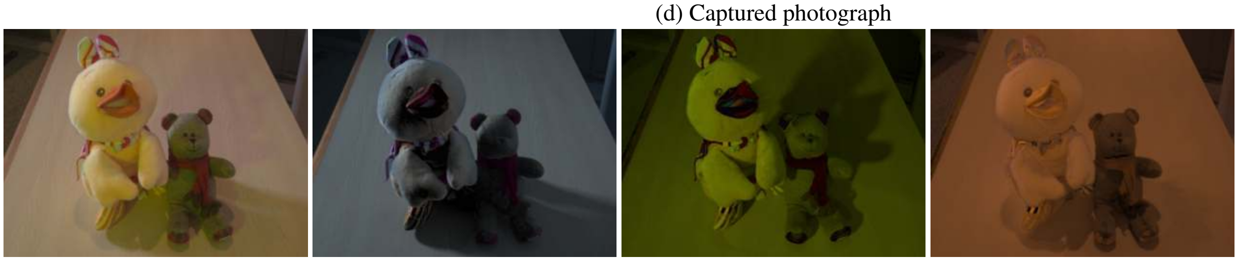
Figure 8: We evaluate the performance of our technique against [3] on the time lapse sequence (a). We can see that the separated images produced by [3] result in incorrect illumination shadings. That is, for the separated image with respect to the skylight, it is obvious to see the highlight in the center of the building as well as the shadows on two sides (b). In comparison to [3], our method relies on a single photograph and produces the results without the need for the cloudy sky frame in the sequence, which in turn achieves better performance in the illumination separation (c).

(a) Input photograph (b) Hui et al. [3] (SNR: 13.16dB)



(c) Our method (SNR: 12.59dB)

(d) Captured photograph

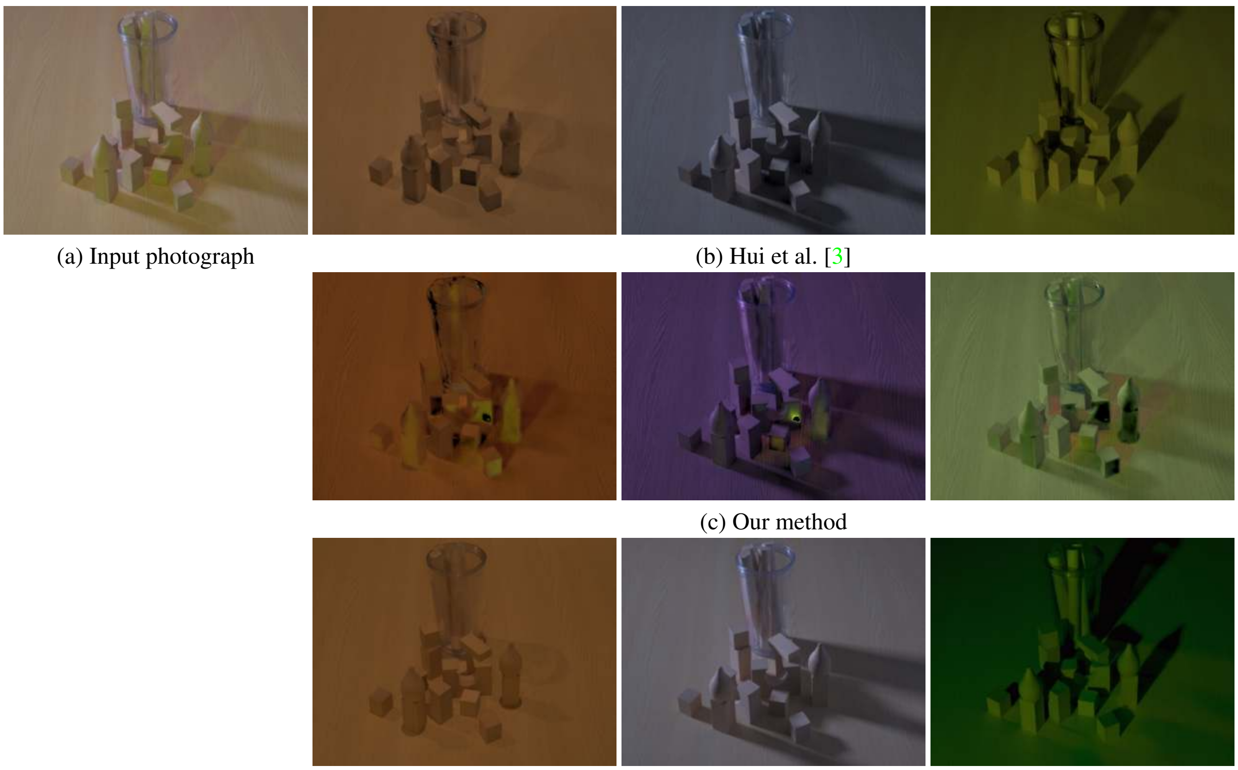


(e) Input photograph (f) Hui et al. [3] (SNR: 11.21dB)

(g) Our method (SNR: 11.09dB)

(h) Captured photograph

Figure 9: We evaluate the performance of our technique against [3] on the three-light scenario. While our training focuses on the scenes under mixture of two light sources, we are able to separate the scenes up to three light sources since we utilize the estimated reflectance chromaticity. As can be seen here, our technique is able to capture both the color and the shading for each of these sources and produces results that closely match the performance of [3] and the ground truth.



(a) Input photograph

(b) Hui et al. [3]

(c) Our method

(d) Captured photograph

Figure 10: We showcase one of the failure example on the scene with three light sources. As can be seen, our technique fails to identify the illuminant shadings, which leads to the artifacts as shown in (c), likely since the mixture of three illuminants is out of the training distribution.

(a) Input (b) Hui et al. [3] (c) Ours

Figure 11: We compare our results to Hui et al. [3] on the white balanced results for the scenes with mixture of two light sources. As can be seen (top), our method is able to match the performance of [3] while requiring single input image. For the scene (bottom), the flash is not strong enough to illuminate the walls in the far end, which leads to the color artifacts. In contrast, our method takes a single photograph and does not rely on flash illumination. As can be seen, the artifacts can be eliminated and visual quality has been significantly improved.