# Deep Single-Image Portrait Relighting
## Supplementary Material

Hao Zhou[1] * Sunil Hadap[2] Kalyan Sunkavalli[3] David W. Jacobs[1]
[1] University of Maryland, College Park, MD, USA
[2] Amazon [3] Adobe Research
[1]{hzhou, djacobs}@cs.umd.edu [2]sunilhadap@acm.org [3]sunkaval@adobe.com

In this supplementary material, we first introduce our demo video in Section 1. We then discuss the Spherical Harmonics we used in our paper in Section 2. In Section 3, we show the network structure for $1024 \times 1024$ images. We give more details about the network structure in Section 4. More visual comparisons with the state-of-the-art methods are shown in Section 7. We show the results of our proposed method on some challenging images in Section 8. Some results on $1024 \times 1024$ images are then shown in Section 9. At last, we show the limitations of the proposed method in Section 10.

## 1. Demo Video

Please refer to the accompanying "demo.mp4" video for relit images generated by the proposed method with different target SH lighting which changes dynamically. We show a half sphere rendered under the same lighting as the face on the left hand side for reference. The portrait image of Obama is downloaded from the Internet. Images of Flickr portrait images are from the Flickr portrait dataset [7]. All the images in the video are generated using our network that works on $1024 \times 1024$ images; we re-size the result to the original size of the image for visualization purposes. Each frame is generated independently; we have not imposed any temporal consistency on the results, and yet the results are realistic and temporally coherent. Due to file size limit, for more videos, please refer to the project webpage https://zhhoper.github.io/dpr.html

## 2. More Details about Spherical Harmonics Lighting

We use second-order Spherical Harmonics (SH) [1, 5] to represent the lighting. As a result, our lighting is represented as a 9 dimensional vector. This is in accordance with many existing works that use SH to model the lighting of the face [6, 10, 9] and the lighting prior dataset [2]. More specifically, the 9 spherical harmonics coefficients, in the 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 a SH lighting $\mathbf{L}$, the rendering equation $f$ defined in Equation 1 of the paper is:

$$f(\mathbf{N}, \mathbf{L}) = \sum_{n=0}^{2} \sum_{m=-n}^{n} \alpha_n Y_{nm} l_{nm},$$

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

*Hao Zhou is currently at Amazon AWS.

Due to the ambiguity of color between lighting and reflectance, we assume the SH lighting is monochromatic. For an input face image, we convert the *RGB* pixel values to *Lab* color space and only relight the *L* channel. The relit *L* channel is then combined with the *ab* channels to form the final output image.

## 3. Network Structure for $1024 \times 1024$ Images

In this section, we show some details of our network for $1024 \times 1024$ images. The network structure for $1024 \times 1024$ images are shown in Figure 1. It is the same as the network structure for $512 \times 512$ images shown in Figure 5 in the paper, except the upsample layer and downsample layer before and after the Hourglass block.

To fine tune the network trained on $512 \times 512$ images using $1024 \times 1024$ images, we use the loss defined in Equation 6 in the paper:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_{GAN} + \lambda \mathcal{L}_F,$$

where $\lambda = 0.5$. We fine tune the network for four epochs.

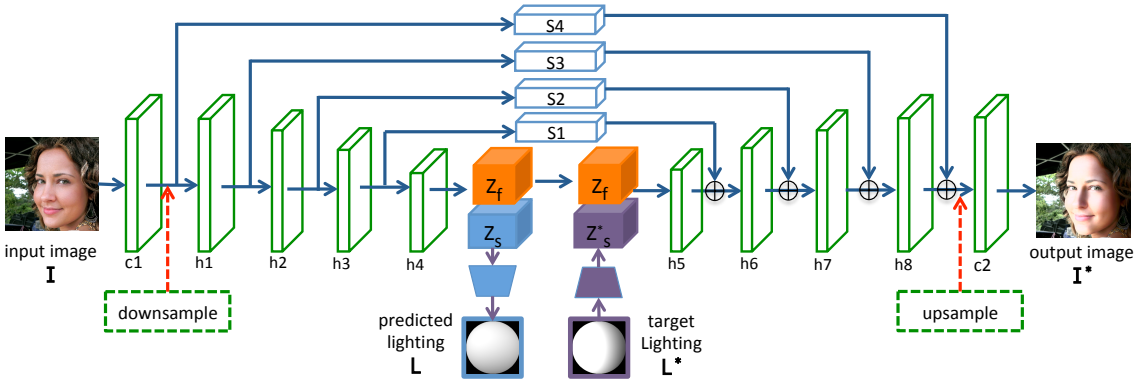



Figure 1. Network structure for $1024 \times 1024$ images. It is the same with the network structure shown in Figure 5 in the paper, except the upsample and downsample layer.

## 4. Details of the Network Structure

In this section, we introduce the details of our Hourglass network. h1, h2, h3, h4 are downsample layers followed by residual blocks defined in [3]. h5, h6, h7 and h8 are designed as residual blocks [3] followed by upsample layers. s1, s2, s3 and s4 are defined to be residual blocks [3]. For convenience, we defined one convolutional block as one convolutional layer followed by a batch normalization layer and ReLU activation. c1 is designed to have one convolutional block. c2 is designed to have three convolutional blocks (denoted as c2_1, c2_2, c2_3) followed by one convolutional layer (denoted as c2_o). More details of these blocks are shown in Table 1. Note that the output of h4 has 155 channels, from which 128 channels belong to face features $\mathbf{Z_f}$ and 27 channels belong to lighting feature $\mathbf{Z_s}$.

Table 1. Details about Each Block of Our Network.

| | h1 | h2 | h3 | h4 | h5 | h6 | h7 | h8 | s1 | s2 | s3 | s4 | c1 | c2_1 | c2_2 | c2_3 | c2_o |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| input channel number | 16 | 16 | 32 | 64 | 155 | 64 | 32 | 16 | 64 | 32 | 16 | 16 | 1 | 16 | 16 | 16 | 16 |
| output channel number | 16 | 32 | 64 | 155 | 64 | 32 | 16 | 16 | 64 | 32 | 16 | 16 | 16 | 16 | 16 | 16 | 1 |
| filter size | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 5 | 3 | 1 | 1 | 1 |

The lighting prediction network, which takes $\mathbf{Z_f}$ as input and predicts $\mathbf{L}$, is defined as an average pooling layer followed by two fully connected layers whose number of channels are 128 and 9 respectively. The network that maps target lighting $\mathbf{L}^*$ to lighting features $\mathbf{Z}^*_\mathbf{s}$ is defined as two fully connected layers whose number of channels are 128, 27. The 27 dimensional lighting feature is then repeated spatially so it has the same spatial resolution as $\mathbf{Z_f}$ as illustrated in Figure 2.

## 5. Ablation Study

Figure 3 visually compares results of trained network using $\mathcal{L}_I$, $\mathcal{L}_I + \mathcal{L}_{GAN}$ and $\mathcal{L}_I + \mathcal{L}_{GAN} + \mathcal{L}_f$.

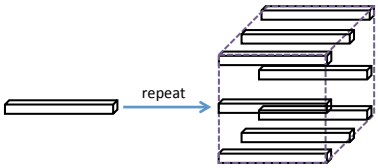



Figure 2. Illustration of repeating lighting feature spatially.

## 6. Compared with Rendering Pipeline

We visually compare our method with our rendering pipeline in Figure 4.

## 7. Visual Results Compared with the State-of-the-art

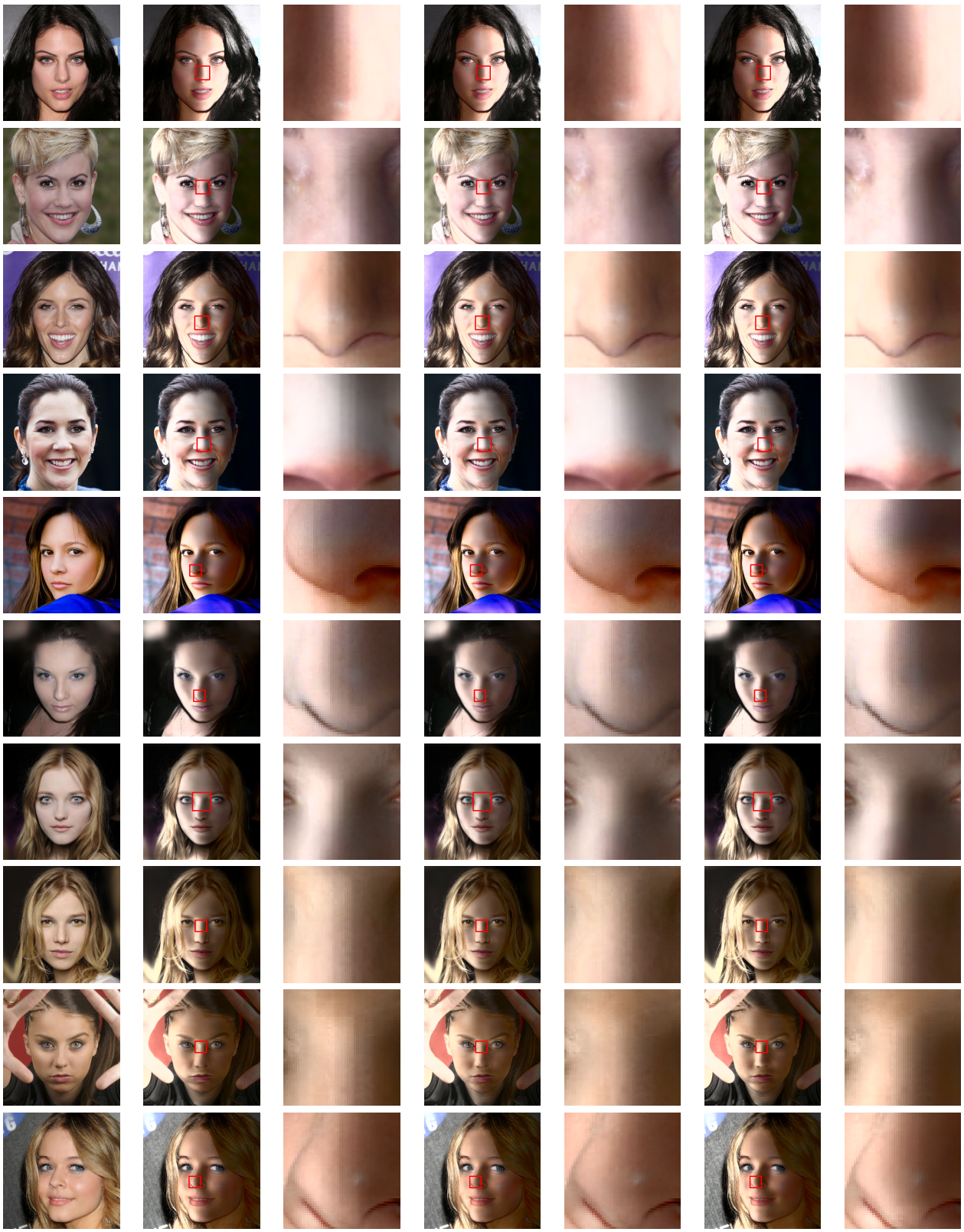
In this section, we show some comparisons with [6], [7], [8] and [4]. Please note that [7], [8] and [4] are all reference image-based portrait relighting methods, i.e. their methods take a reference portrait image as input to represent the light source, which is different from the proposed method and [6]. Table 2 in our paper has demonstrated that all these methods cannot generate a relit image as accurately as our method. Figure 5 is an extension of Figure 11 in the paper. Note that for our proposed method and SfSNet [6], we use the target SH as the light source, whereas for Shih *et al*. [7], Shu *et al*. [8] and Li *et al*. [4], we use the reference image as the light source.

## 8. Results on Non-frontal and Challenging Images

Figure 6 shows results of our method on non-frontal images. Figure 7 shows results of our method on some challenge images. We notice that the proposed algorithm performs well on images with non-frontal face pose, occlusions and even makeup.

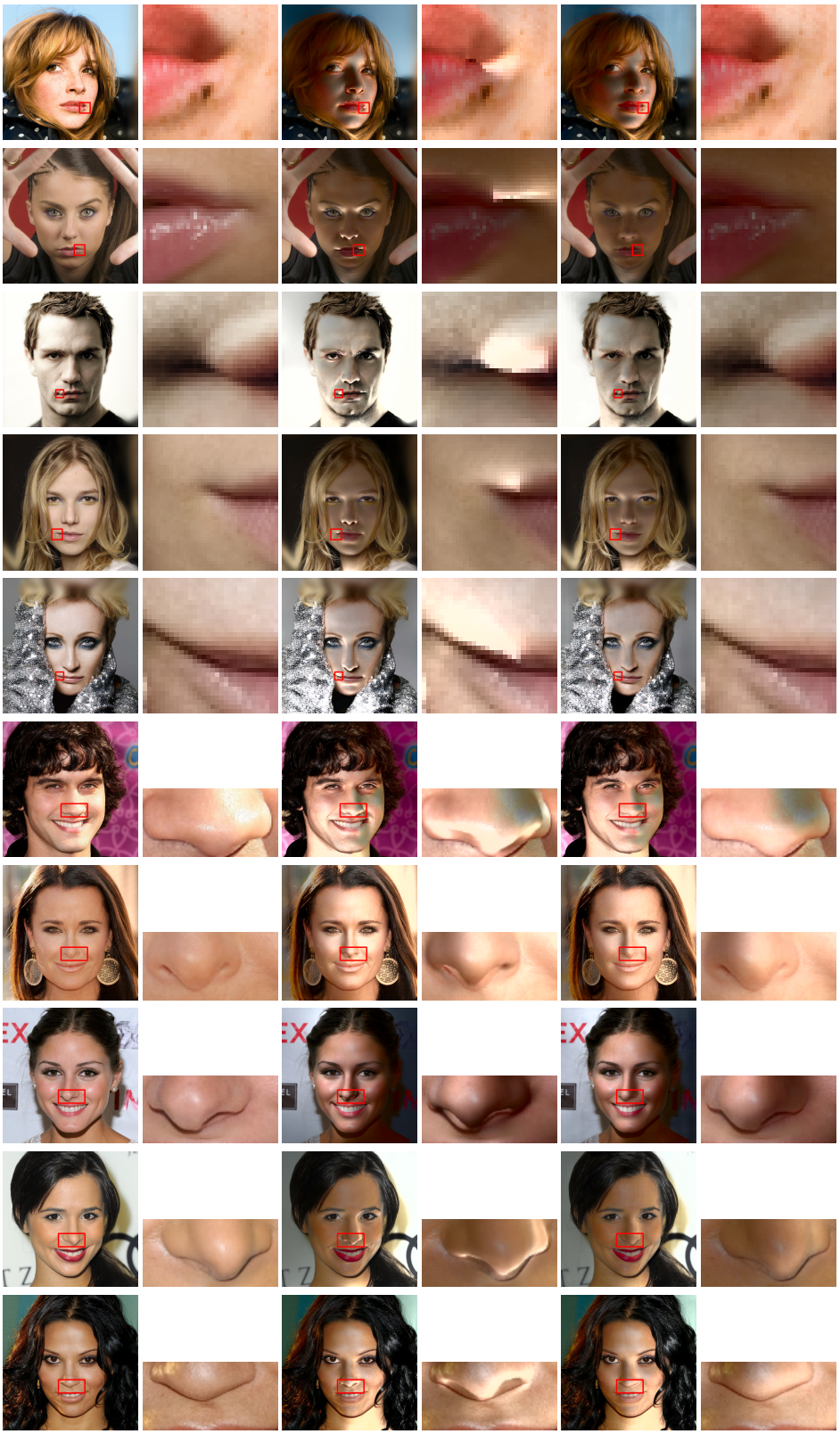
## 9. Results on the High Resolution DPR dataset

We show results on $1024 \times 1024$ images in Figure 8 and Figure 9.



(a) (b) (c) (d) (e) (f) (g)

Figure 3. (a) shows the input image, (b), (d) and (f) are images generated using $\mathcal{L}_I$, $\mathcal{L}_I + \mathcal{L}_{GAN}$ and $\mathcal{L}_I + \mathcal{L}_{GAN} + \mathcal{L}_f$ respectively; (c), (e) and (g) are the red rectangle region of (b), (d) and (f) respectively. Note the edge in the middle of the noise generated using $\mathcal{L}_I$.



(a) (b) (c) (d) (e) (f)

Figure 4. (a) original image, (c) results of RI-based rendering, (d) our results. (b), (d) and (f) show the red rectangle region of (a), (c) and (d) respectively. Note that the proposed method removes the ghost effect and artificial highlights.


Figure 5. Qualitative comparison of the proposed method with state-of-the-art methods. First column in (A) (B) and (C): first row is the reference image, second row is the target SH. Second column in (A), (B) and (C) show the input image, third to seventh column show the results of our method, SfSNet[6], Shih *et al*. [7], Shu *et al*. [8] and Li *et al*. [4].

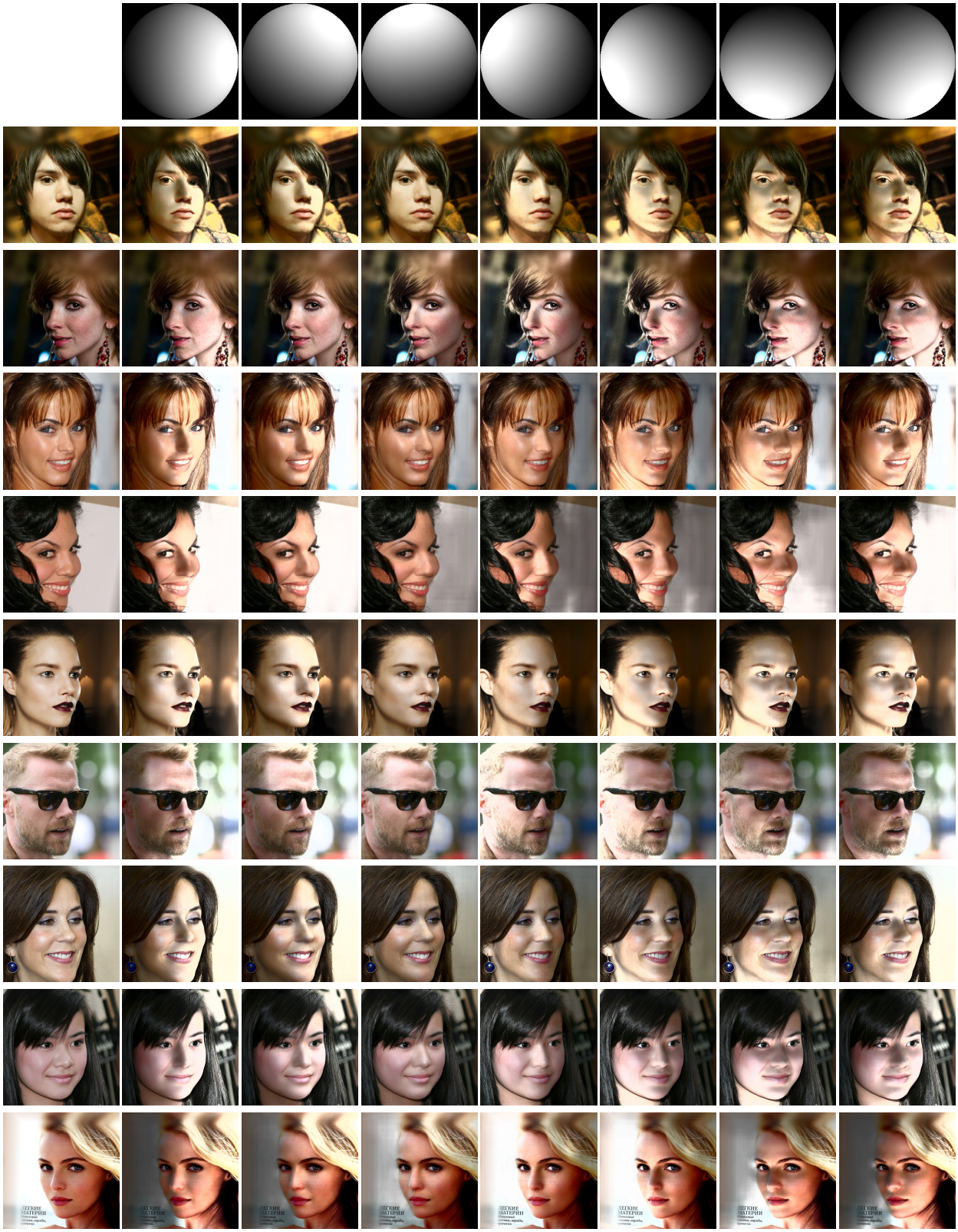
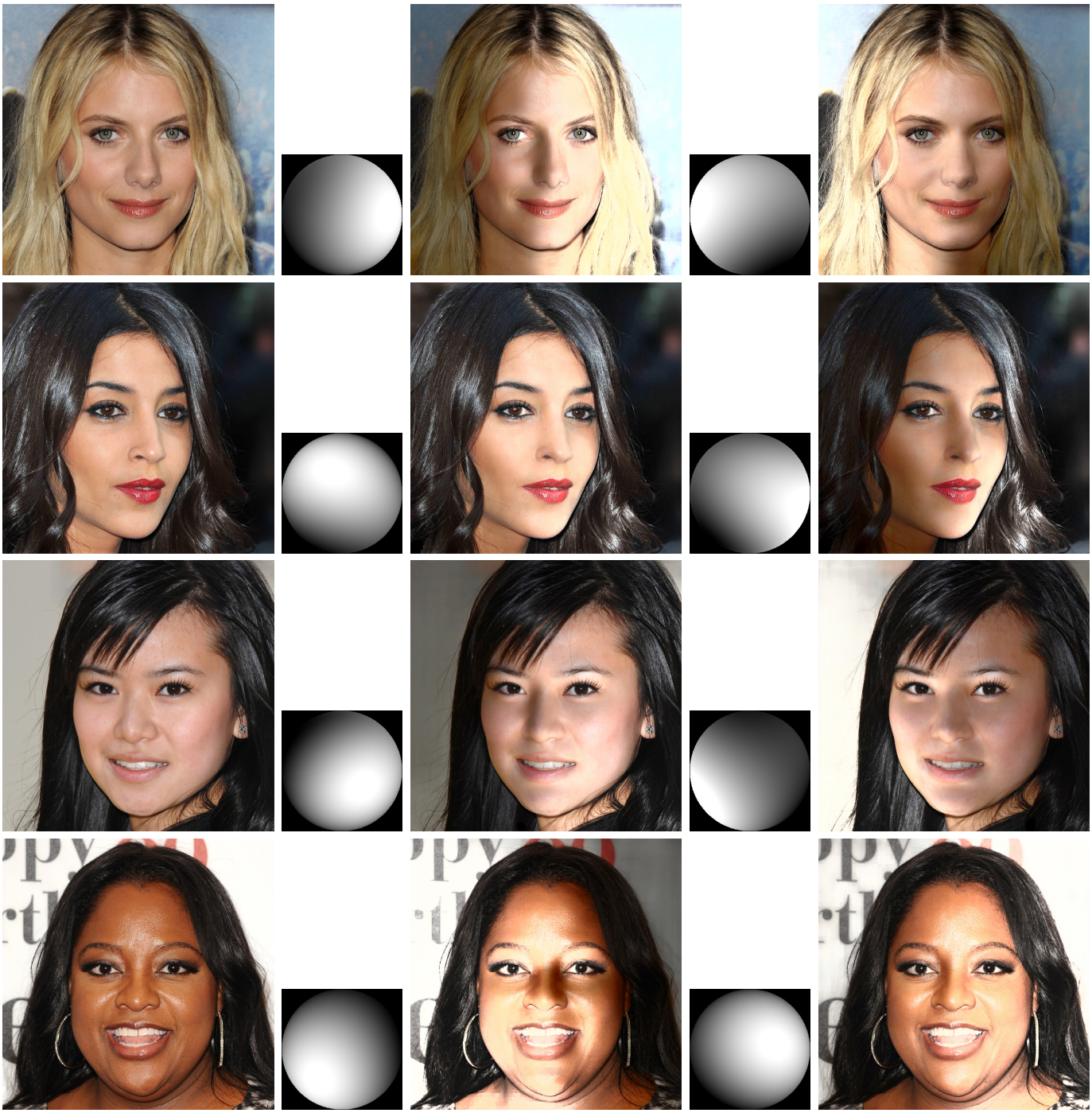


Figure 6. Non-frontal examples. First row show the target SH lighting. The first column of the rest rows show the input image, the other columns show relit images by the proposed method under target SH lighting.

Figure 7. Some challenging examples. First row show the target SH lighting. The first column of the rest rows show the input image, the other columns show relit images by the proposed method under target SH lighting. Our proposed method can deal with faces with occlusions well.

(A) (B) (C) (D) (E)

Figure 8. Results on DPR dataset. (A) is the input image, (B) and (D) are target SH lighting, (C) and (E) are relit images by the proposed method.



(A) (B) (C) (D) (E)

Figure 9. Results on DPR dataset. (A) is the input image, (B) and (D) are target SH lighting, (C) and (E) are relit images by the proposed method.

## 10. Limitations

Since we use Spherical Harmonics to represent lighting, our method cannot model cast shadows. This would require a lighting representation that incorporates ray tracing. From the first row of Figure 10, we can see that the cast shadows caused by the glasses do not change as the lighting changes. Another limitation of the proposed method is that for portrait images with strong shadows, the generated results are affected by the shadows as shown in the second row of Figure 10. We believe this is because the strong shadows will cause information loss which cannot be recovered by our proposed method.

(A) (B) (C) (D) (E)

Figure 10. Bad examples. (A) is the input image, (B) and (D) are target SH lighting, (C) and (E) are images relit by the proposed method.

## References


[1] Ronen Basri and David W. Jacobs. Lambertian reflectance and linear subspaces. *TPAMI*, 25(2), 2003. 1

[2] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *IJCV*, 2018. 1

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2

[4] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *ECCV*, 2018. 3, 6

[5] Ravi Ramamoorthi and Pat Hanrahan. On the relationship between radiance and irradiance: Determining the illumination from images of a convex lambertian object. *JOSA*, 2001. 1

[6] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. Sfsnet: Learning shape, refectance and illuminance of faces in the wild. In *CVPR*, 2018. 1, 3, 6

[7] YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. Style transfer for headshot portraits. *ACM Trans. Graph.*, 33(4), 2014. 1, 3, 6

[8] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics*, 37(2), Nov. 2017. 3, 6

[9] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 1

[10] Hao Zhou, Jin Sun, Yaser Yacoob, and David W. Jacobs. Label denoising adversarial network (ldan) for inverse lighting of faces. In *CVPR*, June 2018. 1