Supplementary Material: Faces as Lighting Probes via Unsupervised Deep Highlight Extraction

Renjiao Yi[1,2], Chenyang Zhu[1,2], Ping Tan[1], Stephen Lin[3]

[1]Simon Fraser University, [2]National University of Defense Technology,
[3]Microsoft Research
{renjiaoy, cza68, pingtan}@sfu.ca, stevelin@microsoft.com

1 Highlight-Net structure

As mentioned in the paper, the structure of Highlight-Net is adopted from [9]. The network structure is exhibited in Figure 1.

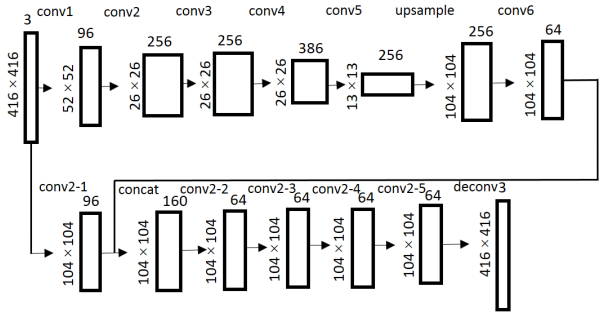


Fig. 1. Structure of Highlight-Net.

2 Training data for pretraining and finetuning

As mentioned in the paper, for pretraining we rendered synthetic faces under real HDR environment maps, consisting of 100 indoor scenes and 100 outdoor scenes. Two examples of the environment maps are shown in Figure 2. Examples of rendered diffuse and specular layers, as well as the composite renderings, are displayed in Figure 3.

In finetuning, as mentioned in Section 6.1 of the main text, images from the MS-celeb-1M dataset [2] are preprocessed by cropping and aligning based on landmarks detected by [15], radiometric calibration by [5], and color histogram

scaling factor = 1 scaling factor = 10^{-1}

Fig. 2. Two example HDR environment maps for outdoor (top) and indoor (bottom) scenes. Shown at different scalng factors for better visualization of the high dynamic range.

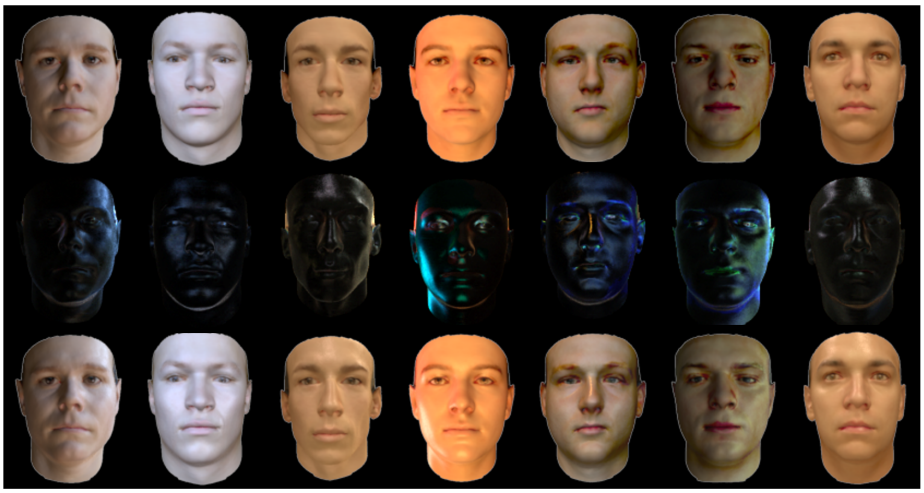


Fig. 3. Examples of rendered synthetic faces. The top row shows rendered diffuse components; the middle row displays rendered specular components; and the bottom row are composite renderings that combine the diffuse and specular layers.

transfer to align illumination colors for each celebrity, performed by transferring the color histograms from one photo of each celebrity to the other photos of the same celebrity. Examples of preprocessed data are shown in Figure 4.

Fig. 4. Examples of preprocessed photos for four celebrities.

3 Additional results on highlight removal

Highlight-Net can work for grayscale photos, unlike most previous methods which are based on color analysis. For a grayscale image, we input a color image whose three channels are equal to those of the grayscale photo. Then with the output, we average the values of the three channels to obtain the result. Although we do not train Highlight-Net on grayscale images, it nevertheless can produce reasonable results as shown in Figure 11 and Figure 12, where RMSE and SSIM [13] are marked in the figures for each example. RMSE represents absolute intensity errors, while SSIM measures structural similarity. Over all of the 30 real images that we captured together with cross-polarized ground truth, the mean SSIM and RMSE are 0.891 and 8.13, respectively. Like for the quantitative evaluation on color images, the RMSE and SSIM are computed on the highlight layer, because input images and ground truth matte images may already have a high structural similarity. Finetuning the net with grayscale training examples should lead to improvements in performance.

We also provide additional comparisons of highlight removal on laboratory images with ground truth, shown in Figure 13 and Figure 14, with RMSE and SSIM values given in the figures. Our method mostly outperforms the previous techniques, which generally have difficulty in dealing with the saturated pixels that commonly appear in highlight regions. Comparisons on a subset of the synthetic data used in the quantitative evaluation are shown in Figure 15. It can be seen that our method generates results similar to the diffuse renderings. The error histograms for quantitative evaluation on 100 synthetic faces and 30 real faces are shown in Table 1 in the main paper and Figure 5.

To show the robustness of Highlight-Net, we tested hard examples like non-neutral expressions, with occluders like glasses or beard, and various ages or skin tones, we provide additional results in Figure 10, which indicate reasonable performance.

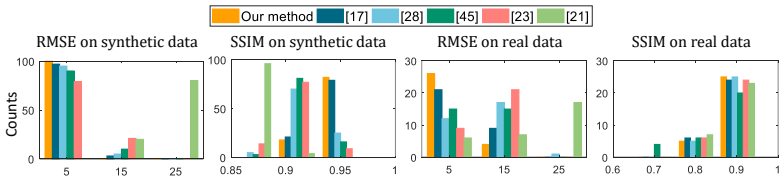


Fig. 5. Quantitative comparisons on highlight removal for 100 synthetic faces and 30 real faces in terms of RMSE and SSIM histograms (larger SSIM is better).

4 Additional results on illumination estimation and virtual object insertion

As mentioned in the main text, the methods in [1, 3] are trained on images that do not contain people and that have a broader view of the scene. So in the experiments, we provide these methods with different input images that fit these characteristics, as shown in Figure 9. Specifically, the person is removed, and a wider angle of the scene is captured.

Also described in the main text, for our quantitative evaluation on illumination estimation with synthetic data for our method, [7] and [4], we provided synthetic faces rendered under the ground truth environment maps as input images. We used 50 indoor and 50 outdoor environment maps in the quantitative evaluation, and 5 synthetic faces under each environment map as input images. In total, 500 synthetic faces are tested for the evaluation. For [3] and [1], we provided LDR photos cropped from the center of the ground truth environment maps as input images. For each of them, 50 outdoor/indoor LDR crops are tested for the evaluation.

The evaluation is done by rendering a diffuse and a glossy Stanford bunny under ground truth and estimated environment maps, and computing the RMSE between these renderings. We use Keyshot [8] as the rendering engine, and for the diffuse bunny we set the diffuse reflectance as white and the specular reflectance as black (all zeros). For the glossy bunny, we choose "hard shiny white plastic" as the material and set the specular reflectance as white, the roughness factor to 0.004, and the refraction index to 1.362. Due to different scaling factors between HDR environment maps estimated by different methods, each diffuse rendering is normalized by its maximum value before computing the error, and the corresponding scaling factors of each environment map are also used for the glossy renderings. The relighting errors are shown in Table 2 in the main paper and Figure 7.

Comparisons on rendered diffuse bunnies under outdoor/indoor illuminations are shown in Figure 16-17 (outdoor), and Figure 18-19 (indoor). Comparisons on rendered glossy bunnies under outdoor/indoor illuminations are shown in Figure 20-21 (outdoor), and Figure 22-23 (indoor). Comparisons on environment maps are displayed in Figure 24 for indoor scenes, and in Figure 25 for outdoor scenes.

To evaluate direction localization, we conducted an experiment on sun positions for outdoor scenes in Figure 6, we computed the centroid of the predicted environment maps as the sun position, in terms of cumulative distribution of images w.r.t. error level as done in [3], where the marked points indicate the error levels over more than 75% of the testing data.

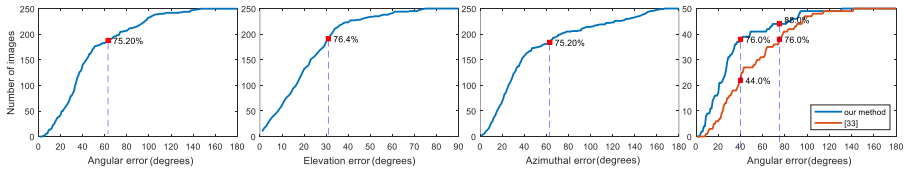


Fig. 6. Evaluation of sun position estimation on outdoor testing data.

Normalized RMSE	Ours	[3]	[1]	[7]	[4]
Mean (outdoor)	**0.143**	0.163	\	0.154	0.245
Mean (indoor)	**0.045**	\	0.050	0.083	0.286

Table 1. Errors in estimating environment maps from real data.

For comparisons on real data, face images and their HDR environment maps are captured for 15 real scenes (7 indoor and 8 outdoor), with background images having faces excluded and a larger field of view for [3] and [1]. Errors with respect to the captured ground truths are presented in Table 1 in terms of RMSE normalized by the difference of the maximum and minimum intensity of the estimated environment map, which is commonly used to facilitate comparison between data with different scales, such as those from the intensity scaling factor of environment maps estimated by different methods. Visual comparisons on estimated environment maps are shown in Figure 26 and 27. The methods in [3] and [1] are applicable only to outdoor and indoor scenes, respectively. They were found to be generally less precise in estimating light source directions when light sources are out-of-view in background images, though they provide reasonable approximations. As seen in (e), the method in [7] may be relatively sensitive to imprecise geometry and surface textures. In (f), estimates of a low-order SH model are seen to lack detail. Our results in (b) most closely match the ground truth, with some error due partly to inexact estimation of face geometry.

Additional comparisons on virtual object insertion are presented in Figure 8, where an outdoor scene is at the top and an indoor scene is at the bottom.

All codes will be publicly available shortly.

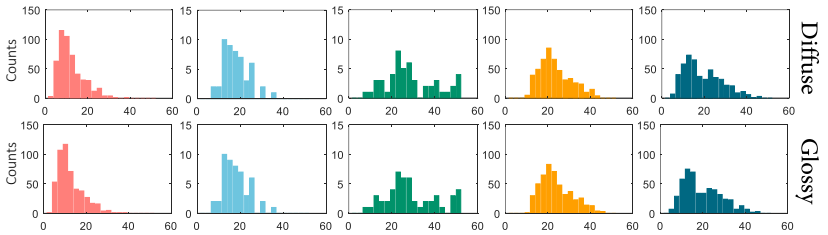


Fig. 7. Relighting RMSE histograms of a diffuse/glossy Stanford bunny lit by illumination estimated by (a) our method, (b) [3] (for outdoor scenes), (c) [1] (for indoor scenes), (d) [7] and (e) [4] (spherical harmonics representation). Visual comparisons of the relighted diffuse/glossy bunnies are available in the supplement.

Fig. 8. Comparisons of object insertion results for outdoor (top) and indoor (bottom) scenes. (a) Photos containing the real object; (b) our method; (c) outdoor result by [3] and indoor result by [1]; (d) [7]; (e) [4].

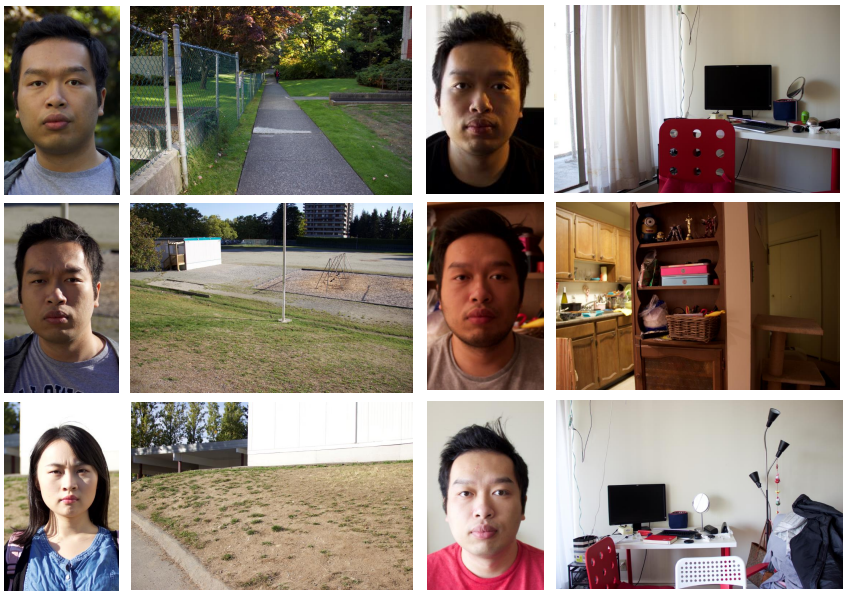


Fig. 9. Images used as input to [1] (indoor scenes) and [3] (outdoor scenes). For each example, the left is the face image used as input for the other illumination estimation algorithms, and the right is the corresponding background photos used as input for [1] and [3].

Fig. 10. Evaluation of highlight removal on testing data with non-neutral expressions, occluders and various ages/skin tones. Input images are shown on the first and third rows, corresponding highlight removal results are shown on the second and fourth rows.

Fig. 11. Highlight removal in grayscale images by Highlight-Net. For each example, the input image, ground truth diffuse image, and our result are displayed from left to right. RMSE is given at the top-right of our results, and SSIM are shown at the bottom-right.

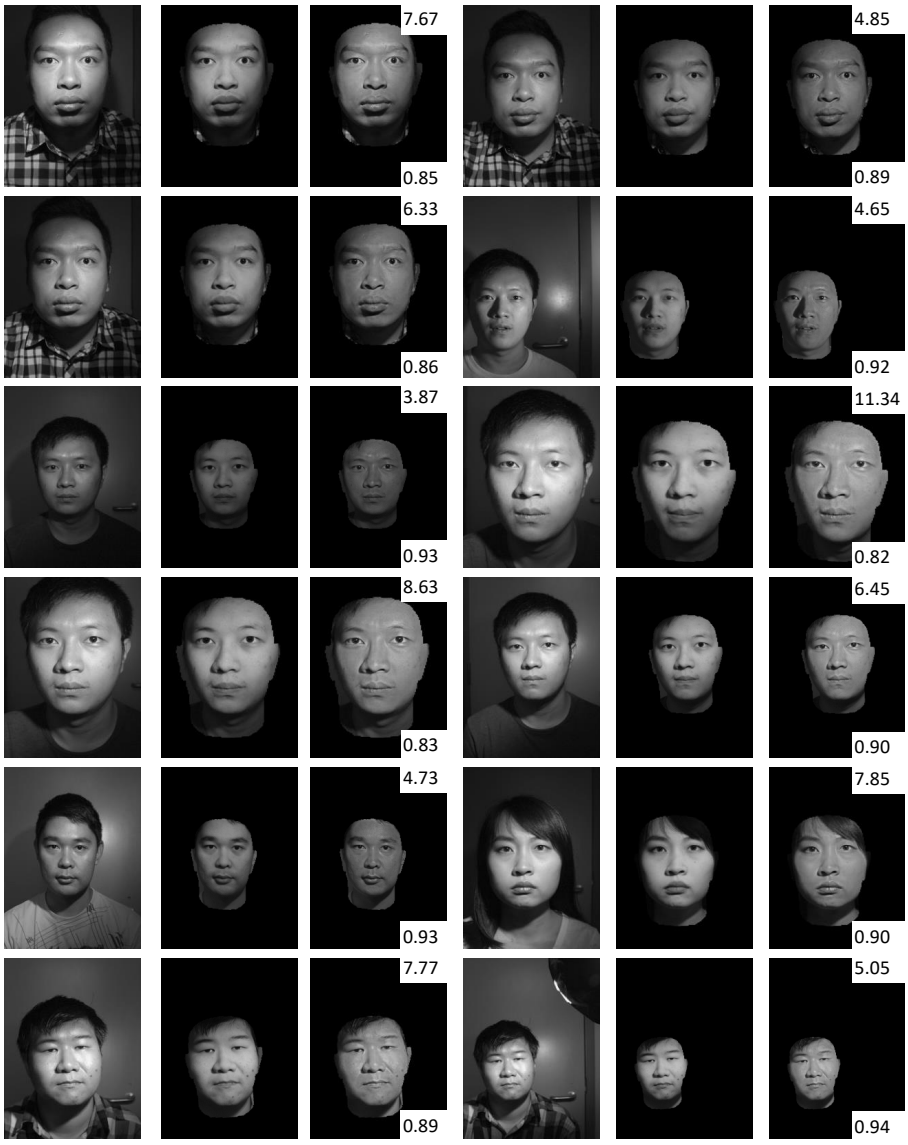


Fig. 12. Highlight removal in grayscale images by Highlight-Net. For each example, the input image, ground truth diffuse image, and our result are displayed from left to right. RMSE is given at the top-right of our results, and SSIM are shown at the bottom-right.

Fig. 13. Additional highlight removal comparisons on laboratory images with ground truth. Face regions are cropped out automatically by landmark detection [15]. (a) Input photo. (b) Ground truth captured by cross-polarization for lab data. (c-h) Highlight removal results by (c) our finetuned Highlight-Net, (d) Highlight-Net without finetuning, (e) [11], (f) [6], (g) [10], (h) [14], and (i) [12]. RMSE values are given at the top-right, and SSIM at the bottom-right. RMSE and SSIM are computed on highlight layers.

Fig. 14. Additional highlight removal comparisons on laboratory images with ground truth. Face regions are cropped out automatically by landmark detection [15]. (a) Input photo. (b) Ground truth captured by cross-polarization for lab data. (c-h) Highlight removal results by (c) our finetuned Highlight-Net, (d) Highlight-Net without finetuning, (e) [11], (f) [6], (g) [10], (h) [14], and (i) [12]. RMSE values are given at the top-right, and SSIM at the bottom-right. RMSE and SSIM are computed on highlight layers.

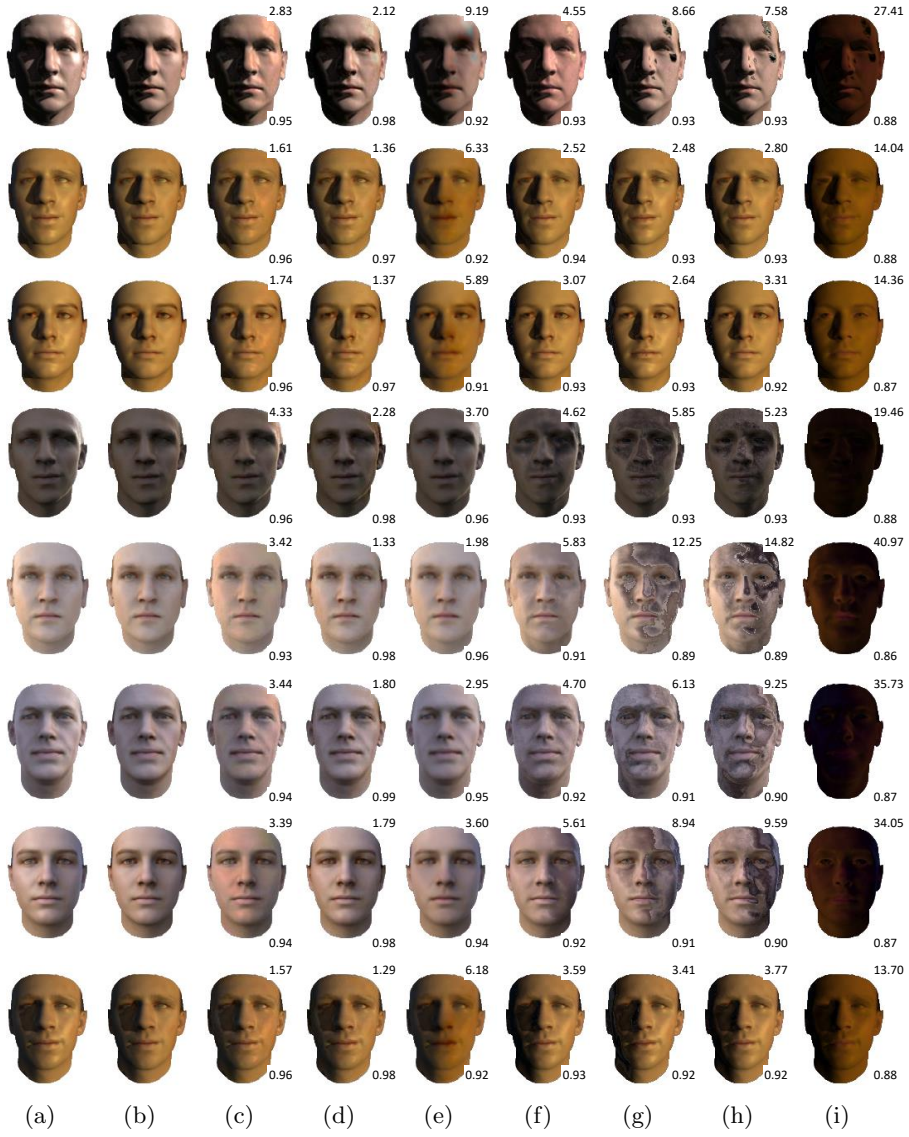


Fig. 15. Highlight removal comparisons on a subset of the synthetic images. (a) Input photo. (b) Diffuse rendering under the same illumination. (c-h) Highlight removal results by (c) our method, (d) our pretrained net, (e) [11], (f) [6], (g) [10], (h) [14], and (i) [12]. RMSE values are given at the top-right, and SSIM at the bottom-right. RMSE and SSIM are computed on highlight layers.

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

Fig. 16. Comparisons of diffuse Stanford bunny relit by estimated outdoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [3], (e) [7] and (f) [4].

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

Fig. 17. Comparisons of diffuse Stanford bunny relit by estimated outdoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [3], (e) [7] and (f) [4].

Fig. 18. Comparisons of diffuse Stanford bunny relit by estimated indoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [1], (e) [7] and (f) [4].

Fig. 19. Comparisons of diffuse Stanford bunny relit by estimated indoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [1], (e) [7] and (f) [4].

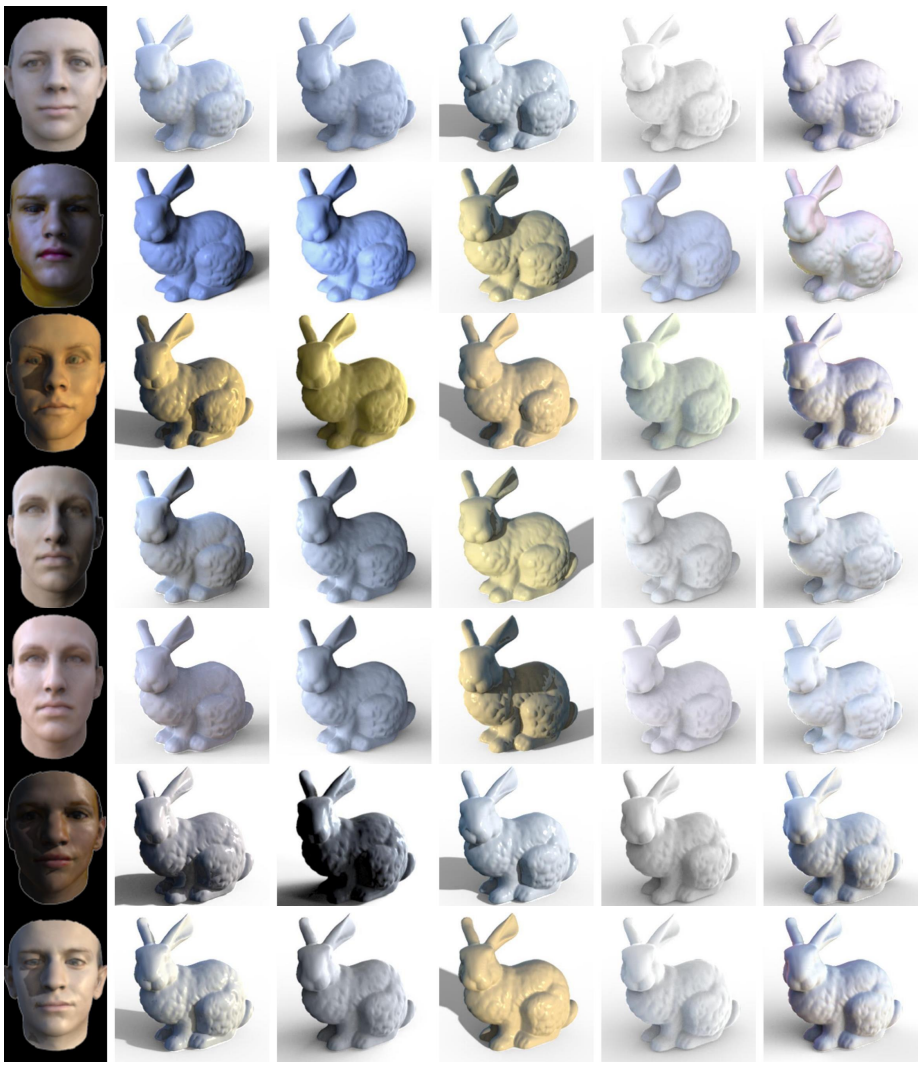


Fig. 20. Comparisons of glossy Stanford bunny relit by estimated outdoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [3], (e) [7] and (f) [4].

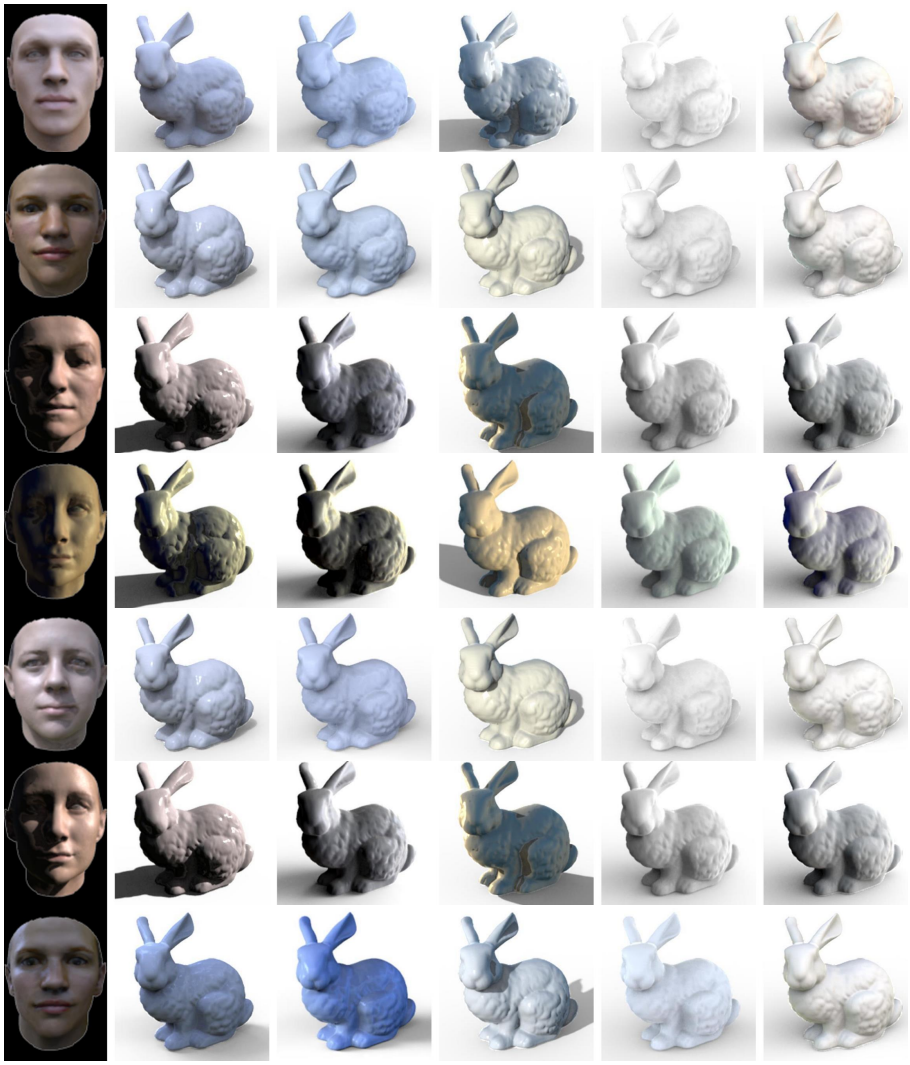


Fig. 21. Comparisons of glossy Stanford bunny relit by estimated outdoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [3], (e) [7] and (f) [4].

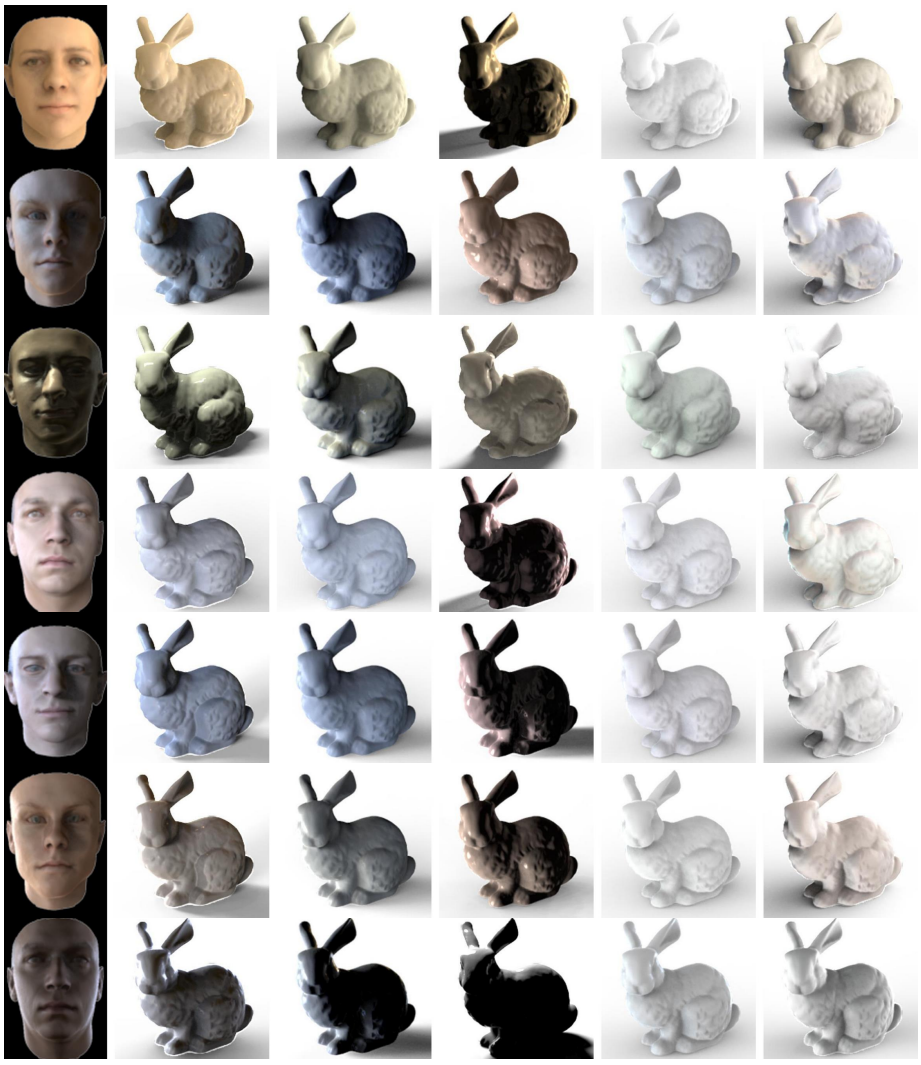


Fig. 22. Comparisons of glossy Stanford bunny relit by estimated indoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [1], (e) [7] and (f) [4].

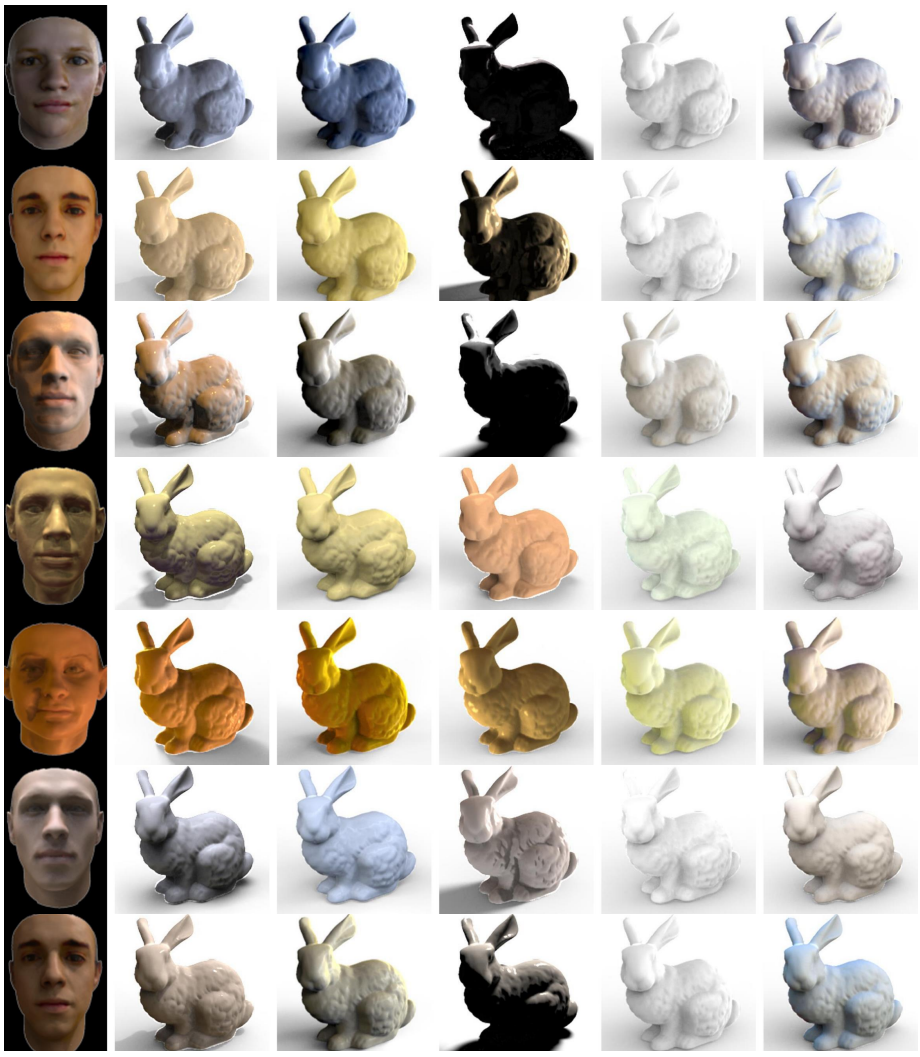
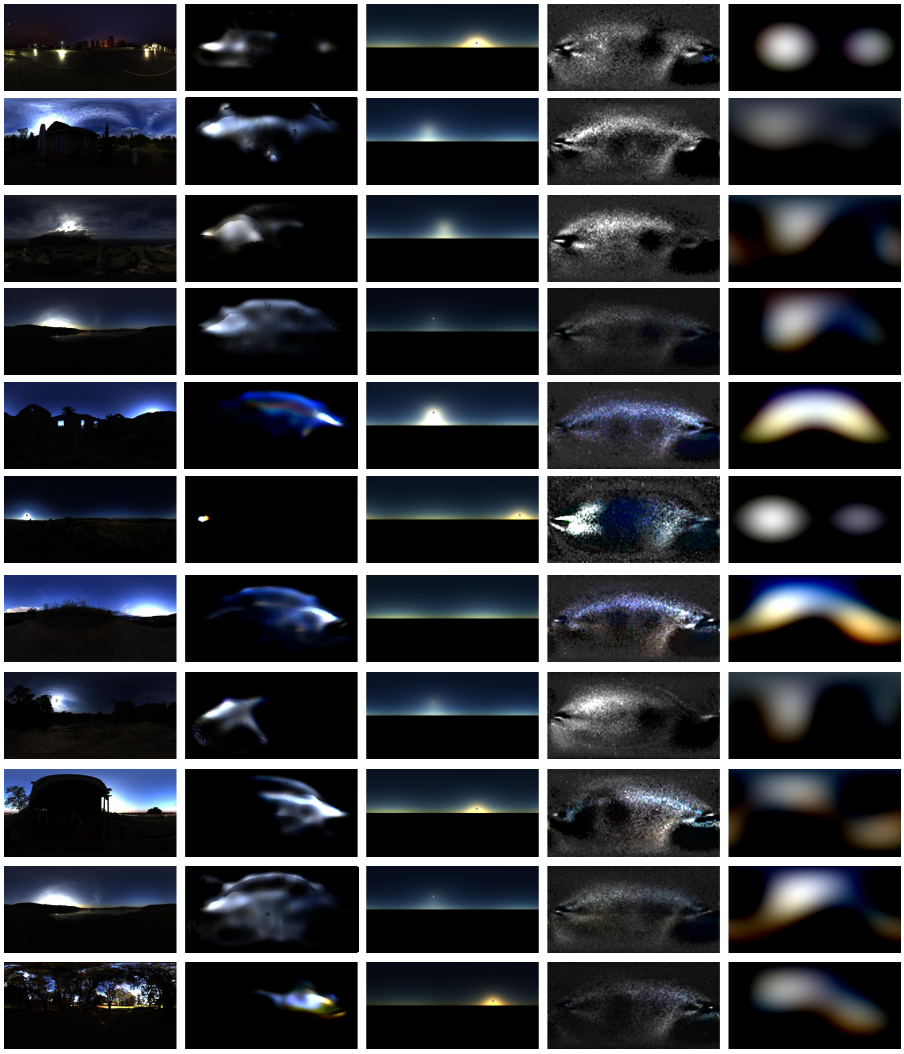


Fig. 23. Comparisons of glossy Stanford bunny relit by estimated indoor illuminations. (a) Input photo. (b) Bunnies under ground truth environment maps. (c-f) Bunnies relit by environment maps estimated by (c) our method, (d) [1], (e) [7] and (f) [4].

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

Fig. 24. Comparisons of selected indoor data used in quantitative evaluation of illumination estimation. (a) Ground truth indoor environment maps, (b-e) indoor environment maps estimated by (b) our method, (c) [1], (d) [7] and (e) [4]. Total intensities of all environment maps are normalized to be the same.



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

Fig. 25. Comparisons of selected outdoor data used in quantitative evaluation of illumination estimation. (a) Ground truth indoor environment maps, (b-e) indoor environment maps estimated by (b) our method, (c) [1], (d) [7] and (e) [4]. Total intensities of all environment maps are normalized to be the same.

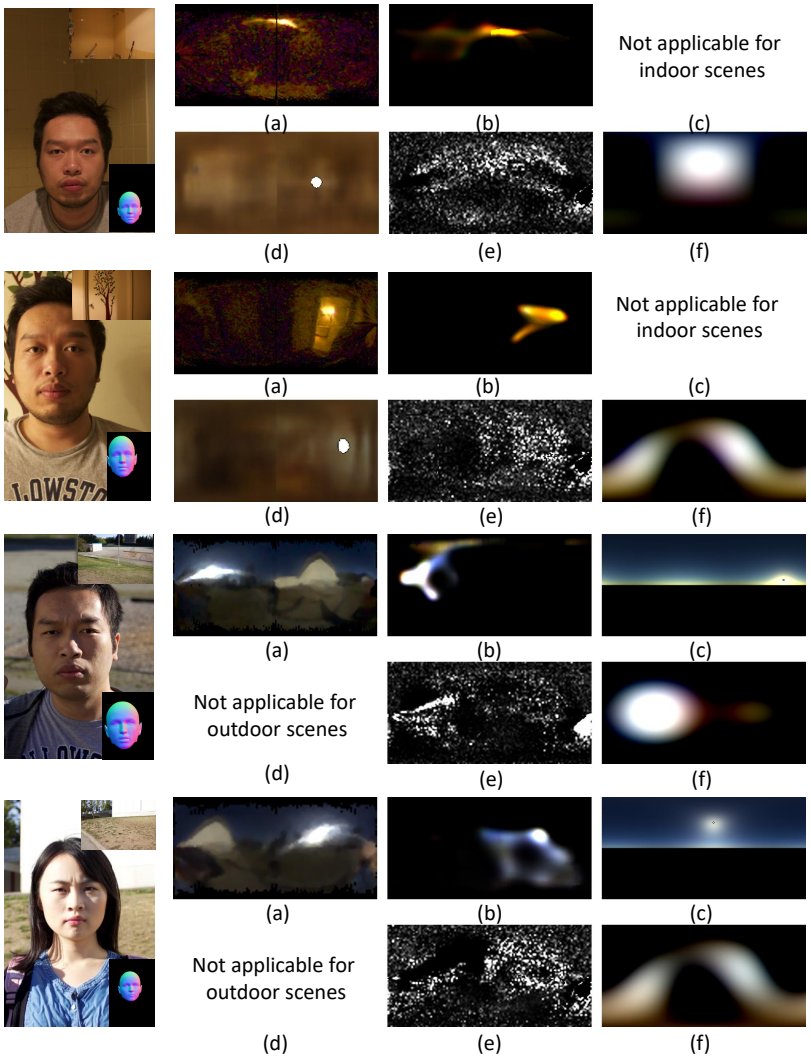


Fig. 26. Comparisons of illumination estimation on real data. The faces on the left are input face photos. (a) Ground truth indoor environment maps, (b-f) indoor environment maps estimated by (b) our method, (c) [3], (d) [1], (e) [7] and (f) [4]. Input background photos for [1] and [3] are shown at top right of the input photos, and the input face normals for our method, [7] and [4] are shown at bottom right.

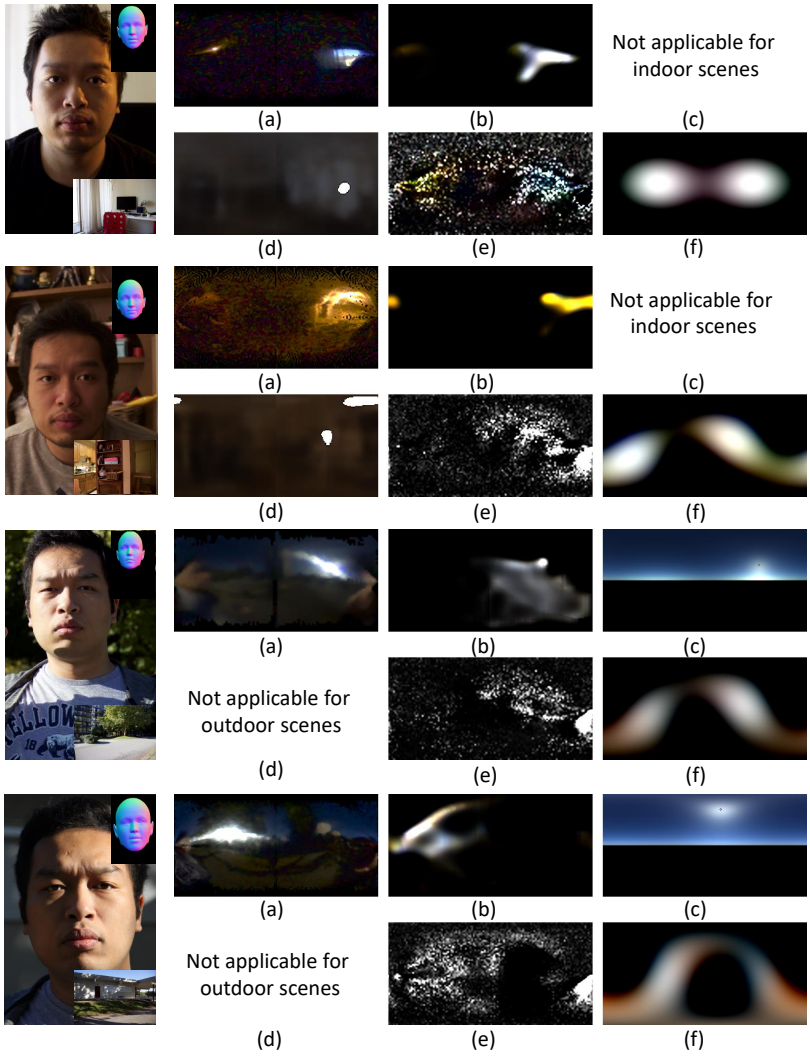


Fig. 27. Comparisons of illumination estimation on real data. The faces on the left are input face photos. (a) Ground truth indoor environment maps, (b-f) indoor environment maps estimated by (b) our method, (c) [3], (d) [1], (e) [7] and (f) [4]. Input background photos for [1] and [3] are shown at bottom right of the input photos, and the input face normals for our method, [7] and [4] are shown at top right.

References

1. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image. ACM Transactions on Graphics (SIGGRAPH Asia) **9**(4) (2017)
2. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In: European Conference on Computer Vision. Springer (2016)
3. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., Lalonde, J.F.: Deep outdoor illumination estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
4. Knorr, S.B., Kurz, D.: Real-time illumination estimation from faces for coherent rendering. In: Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on. pp. 113–122. IEEE (2014)
5. Li, C., Lin, S., Zhou, K., Ikeuchi, K.: Radiometric calibration from faces in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3117–3126 (2017)
6. Li, C., Lin, S., Zhou, K., Ikeuchi, K.: Specular highlight removal in facial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3107–3116 (2017)
7. Lombardi, S., Nishino, K.: Reflectance and illumination recovery in the wild. IEEE Trans Pattern Anal Mach Intell (PAMI) **38**(1), 129–141 (2016)
8. Luxion Inc.: Keyshot 6.3, https://www.keyshot.com/
9. 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. pp. 2992–2992 (2015)
10. Shen, H.L., Zheng, Z.H.: Real-time highlight removal using intensity ratio. Applied optics **52**(19), 4483–4493 (2013)
11. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017)
12. Tan, R.T., Nishino, K., Ikeuchi, K.: Separating reflection components based on chromaticity and noise analysis. IEEE transactions on pattern analysis and machine intelligence **26**(10), 1373–1379 (2004)
13. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
14. Yang, Q., Wang, S., Ahuja, N.: Real-time specular highlight removal using bilateral filtering. Computer Vision–ECCV 2010 pp. 87–100 (2010)
15. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2879–2886. IEEE (2012)