Controllable Light Diffusion for Portraits
Supplementary Material

David Futschik[1,2] Kelvin Ritland[1] James Vecore[1] Sean Fanello[1]
Sergio Orts-Escolano[1] Brian Curless[1,3] Daniel Sýkora[1,2] Rohit Pandey[1]

[1]Google Research [2]CTU in Prague, FEE [3]University of Washington

In this supplementary material, we provide details regarding ground truth data generation and the training procedure for the proposed approach, as well as minor clarifications on the text of the paper. Additionally, we show more detailed results and ablations of our model.

1. Terminology

Although we use the terms *shadow map* and *specular map* throughout this work, these terms could be seen as somewhat misleading. In reality, these maps represent areas where the amount of light is lower or higher, respectively, than if the lighting in the scene were diffusely blurred (by $\cos_+(\theta)$). As such, they can be seen as relative illumination maps. However, it is useful to conceptually think of them as representing their extremes, i.e. shadows and specular reflections, as those are the primary perceptual effects of strong directional light we are aiming to suppress.

2. Data Generation

To train the proposed model, we require supervised pairs of input portraits I and output diffused images I_d, as well as intermediate shadow and specular maps. Following previous work [7], we rely on a Light Stage [2, 6] to capture the full reflectance field of multiple subjects as well as their geometry.

In particular, we recorded a set of 70 diverse participants with different skin tones, performing 9 different facial expressions and wearing different apparels. For each sequence, we acquired 331 images, corresponding to one specific light source in the Light Stage. The full set of one-light-at-time (OLAT) images is then used to perform HDR relighting from each acquired viewpoint, by linearly combining the reflectance field [1]. Specifically, we used a sparse set of 58 cameras placed all around the acquired subject, with 6 frontal facing views, simulating close up framing typical of portrait photography (see [7] for more details).

Additionally, we use a multi-view system with custom infrared (IR) depth sensors to infer high quality geometry [2].

We then selected 279 medium to high contrast HDR panoramic lighting environments sourced from www.HDRIHaven.com [9] to relight the subjects with the original HDR map as well as the convolved, diffused ones to obtain diffused images I_d, as well as specular and shadow maps. We split the dataset into training and testing subsets, manually selecting 7 subjects with diverse skin tones for evaluation, as well as 10 lighting environments unused during training.

2.1. Synthetic Shadow Augmentation

The described OLAT summation technique is incapable of producing soft shadows without heavily specialized and expensive techniques [8], especially for small area lights. Therefore, we augmented our data with additional soft shadows cast from synthetic objects in the scene. In principle, one could use raytracing techniques to synthesize these shadows, but raytracing accurate soft shadows requires sophisticated denoising algorithms or long computation times. Instead, we augmented our dataset with a novel and effective technique that approximates realistic shadows that match the diffusiveness of the casting environment, requiring only $O(1)$ additional operations per pixel versus regular OLAT summation.

First, we generate a shadow map in image space. Conceptually, we place a virtual cylinder in the scene with an alpha texture mapped to the surface, and project that texture to the subject's geometry (acquired using [2]) using ray casting. More precisely, suppose we have a light source (infinitely far away, an OLAT), a depth image (rendered from the captured geometry for a given camera and take), and a virtual shadow casting cylinder around the subject. For each pixel in the depth image, we cast a ray from the pixel's position towards the light source. For these rays, we compute the intersection with a virtual cylinder in the environment, and use the corresponding shadow map texture value as the amount that the light is covered by the virtual shadow. This

Figure 1. Multiple settings of shadow diffusion for relit images. With our synthetic shadow augmentation, a method of matching the shadow diffusion to the target relight Gini is required. A miscalibration of the parameters would result in inconsistent behavior as the diffusion parameter is varied. We calibrated our parameter settings by consulting with professional photographers with the goal of producing images with perceptually plausible amount of shadow diffusion.

process generates a shadow map in image space for a single light source (i.e. OLAT image). We manually created 28 shadow map textures, comprising of silhouettes of trees, geometric patterns, people, fences, and building architectures.

Next, we soften the shadow map to match the light environment's absolute diffuseness. We experimentally obtained a new parameter $d = \alpha(G-1)^2 + (1-\alpha)\max\left(0, \frac{\beta-G}{\beta}\right)$ where $\alpha = 0.5$, $\beta = 0.65$, and $G \in [0,1]$ is the Gini coefficient of the HDR environment map, as described in the paper (smaller G corresponds to more diffuse lighting). We then adjusted the shadow map's opacity by $1 - 0.4d$ and applied a Gaussian blur with kernel size $0.03dW$, where W is the image width in pixels. This technique, while not physically accurate, softens the edges and lightens the shadow enough that they match the internal shadowing (i.e. nose shadows, etc) given from the OLAT based relighting. We consulted our tuning of proposed parameter setting with professional photographers. An example of different configurations is shown in 1.

Finally, we generate two relit images of the subject: one using the entire lighting environment as the baseline lighting, and one with the brightest OLAT (for the source environment map) removed, as the shadow color. Using the softened shadow map, generated from the light corresponding to the brightest OLAT, we blend these two relit images to generate the final rendering with a synthetic shadow. This method is shown in Figure 2. Note this method assumes the shadow is cast by a single light source – the light corresponding to the brightest OLAT image. For the kind of images we are targeting, a single bright light source is very

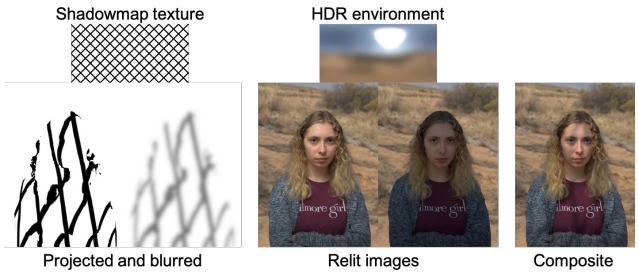

Figure 2. Synthetic shadow generation method. We project a shadow map onto a captured depth image and then use a blurred and lightened version of it to blend between two relit images.

common and so this assumption is reasonable.

This process has some similarities to the augmentation done by Zhang et al. [11], but we match the diffusivity of the lighting environment and incorporate captured geometry, allowing us to generate soft shadow maps that realistically conform to the subject's surface.

With this proposed data generation technique, we generated 17M training examples. In particular, for each Light Stage capture, for each front facing camera, and for each HDR environment, we created 16 samples, by taking the cross product of 4 random rotations of the HDR environment and 4 random shadow maps (one always blank). Each training example contains three levels of diffusion: the original HDR environment, a random integer specular exponent (selected on a log-uniform distribution in $[4, 64]$), and a fully diffused image (specular exponent 1).

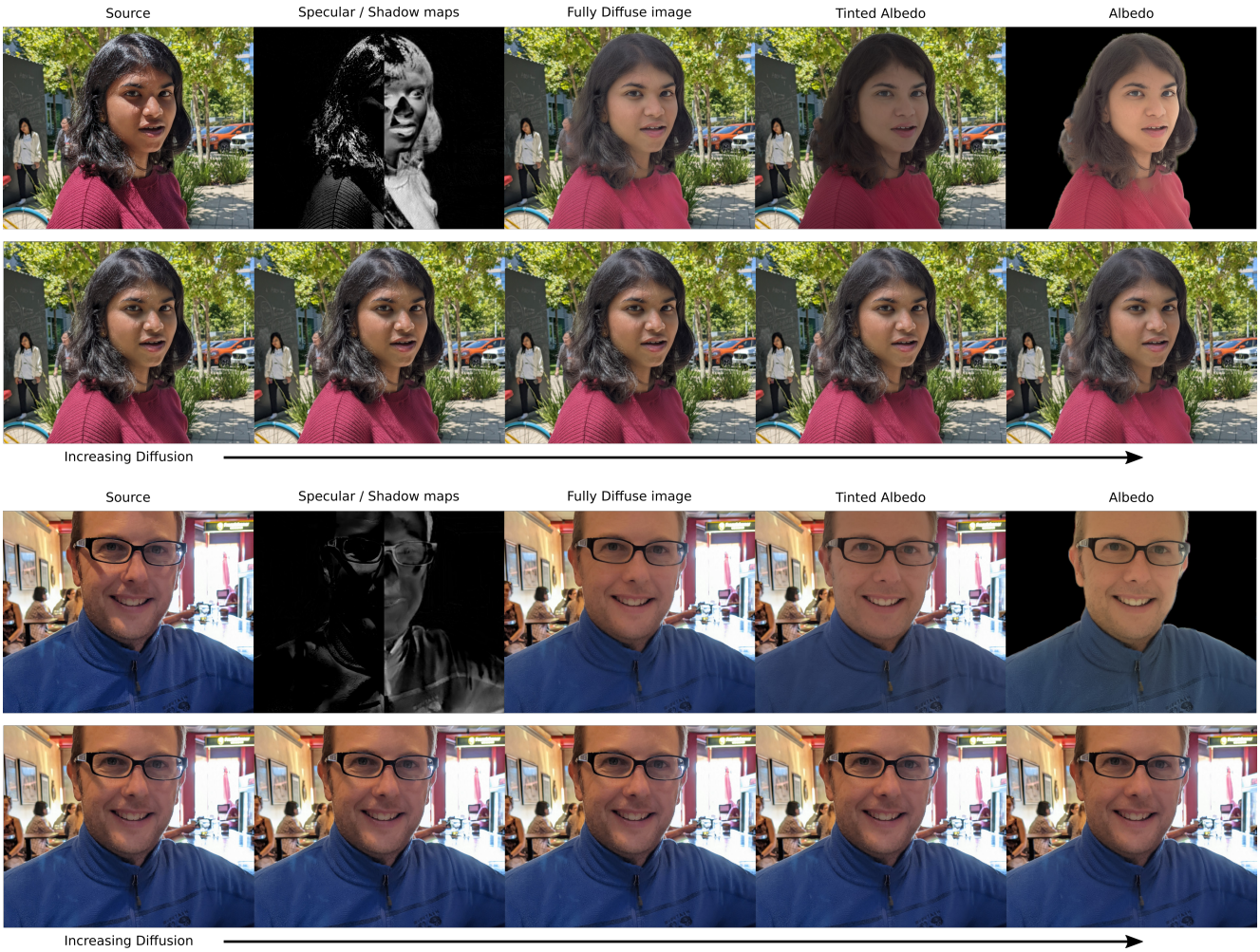

Figure 3. Further results demonstrating the respective outputs our method produces. We first extract specular/shadow maps from the input image and then produce a fully diffuse image. Additionally, we can recover a uniformly lit image, *i.e*. tinted by the average light color, from which we can estimate the untinted albedo. The bottom rows illustrate the application of editing the input photo by gradually increasing the amount of light diffusion.

2.2. Additional augmentations

During training, we applied additional augmentation to the training examples. While the selected HDR environments cover a broad range of possible lighting conditions, they are biased towards softer environments, whereas we expect our method to provide largest benefits when applied to images with relatively harsh lighting environments.

To simulate this, we apply a range of brightness-adjusting augmentations to our training samples. Although naive scaling in linear RGB space works well to produce overexposed images, we find that it alone produces images with lower contrast between overexposed and underexposed, shadowed areas. To remedy this, our brightness adjustments take the estimated shadow map into account and scale the shadowed areas by a different factor.

Similarly, we observe that in-the-wild image post-processing often leads to non-linear color shifts in shadowed areas, and attempt to replicate this effect in our training data by color-tinting shadowed areas.

Lastly, while the light stage captures some subsurface scattering effects, we would like to emphasize their removal, since they are much more prominent under direct sunlight. To this end, we add approximate subsurface color tinting around shadow edges on the skin. In particular, we apply a skin segmenter to create skin map K. Then, we detect edges in shadow map D (described in the paper) where it overlaps K. We isolate these edges and blur them with a Gaussian filter, then map the result into smooth color tinting map, with heavy bias towards red tones, similar to ones that would be produced by light scattering through blood vessels under the skin.

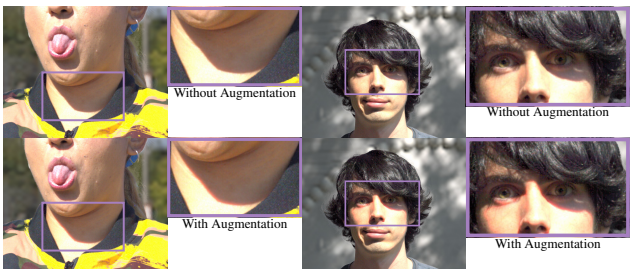

Figure 4. Examples of applying our skin subsurface scattering augmentation. Left example shows inpainting in the nose and head shadow regions, example on the right shows inpainting in the hair shadow region.

2.3. Training Details

We implemented our training pipeline in TensorFlow, distributing the training across 8 NVIDIA Tesla V100 GPUs with 16GB of memory. Each iteration randomly picks 8 images of subjects relit with a random HDR environment, a diffused version of the same HDR environment with random specular exponent, and a fully diffused version with specular exponent 1. We use the ADAM optimizer [3] with a learning rate of 5×10^{-5}. We optimized our system for 2 million iterations for the training to converge, taking fourteen days.

To train our N-diffusions albedo model, we first trained a model that just predicts the fully diffused image using the above method. We then used this model inside the full N-diffusions architecture, and trained end-to-end for an additional 1 million iterations for an additional week of training.

Loss functions We have two main image-based losses: the pixel-wise L1 difference between the model output and the target image, and the (L1) difference between the VGG features [10] of the model output and target; we call these L1 and VGG losses, respectively.

For our parametric diffusion model, we compute these two losses for a random diffusion level to get $\mathcal{L}_{\text{diff}} = \lambda_{\text{L1}}\mathcal{L}_{\text{diff,L1}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{diff,vgg}}$, a face crop of the random diffusion to get $\mathcal{L}_{\text{face}} = \lambda_{\text{L1}}\mathcal{L}_{\text{face,L1}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{face,vgg}}$, and a second inference of a fully diffused image and the fully diffused image (where fully diffused means lit by environment blurred with $\cos_+(\theta)$ kernel) to get $\mathcal{L}_{\text{full}} = \lambda_{\text{L1}}\mathcal{L}_{\text{full,L1}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{full,vgg}}$. We empirically determined parameters $\lambda_{\text{vgg}} = 6, \lambda_{\text{L1}} = 1$. For the specular and shadow maps, we only used L1 losses to obtain $\mathcal{L}_{\text{maps}} = \lambda_{\text{L1}}(\mathcal{L}_{\text{spec,L1}} + \mathcal{L}_{\text{shad,L1}})$. We also use a least squares discriminator between the diffused image and the ground truth diffused image to obtain $\mathcal{L}_{\text{adv}}$ [5]. We then computed our total loss as

$$\mathcal{L}_{\text{parametric}} = \mathcal{L}_{\text{diff}} + \mathcal{L}_{\text{face}} + \mathcal{L}_{\text{full}} + \mathcal{L}_{\text{maps}} + \mathcal{L}_{\text{adv}}$$

For our N-diffusions albedo model, we use the same $\mathcal{L}_{\text{full}}$, $\mathcal{L}_{\text{adv}}$, and $\mathcal{L}_{\text{maps}}$ losses as the parametric model, and computed additional losses for the face crop of the first full diffusion to get $\mathcal{L}_{\text{fullface}} = \lambda_{\text{L1}}\mathcal{L}_{\text{fullface,L1}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{fullface,vgg}}$, the third full diffusion with respect to the tinted albedo ground truth to get $\mathcal{L}_{\text{tintalb}} = \lambda_{\text{L1}}\mathcal{L}_{\text{tintalb,L1}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{tintalb,vgg}}$, and a face crop of the tinted albedo to get $\mathcal{L}_{\text{facealb}} = \lambda_{\text{L1}}\mathcal{L}_{\text{facealb,L1}} + \lambda_{\text{vgg}}\mathcal{L}_{\text{facealb,vgg}}$. For the HDR tint predictor we computed the L1 loss between the predicted RGB tint and the average illumination of the HDR map, $\mathcal{L}_{\text{tint}} = \lambda_{\text{L1}}\mathcal{L}_{\text{tint,L1}}$. We then computed our total loss as

$$\begin{aligned}\mathcal{L}_{\text{alb}} =& \mathcal{L}_{\text{full}} + \mathcal{L}_{\text{fullface}} + \mathcal{L}_{\text{maps}} + \mathcal{L}_{\text{tintalb}} \\ &+ \mathcal{L}_{\text{facealb}} + \mathcal{L}_{\text{tint}} + \mathcal{L}_{\text{adv}}\end{aligned}$$

with the same empirically determined parameters $\lambda_{\text{vgg}}, \lambda_{\text{L1}}$ as above.

3. Additional Results and Ablations

In this section, we present further results on a wider range of subjects as well as ablation studies of our model.

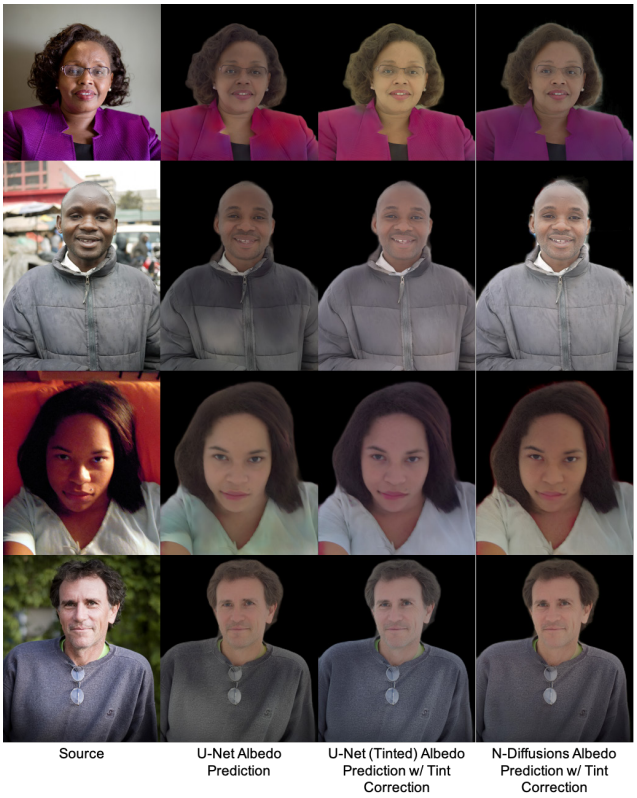

Figure 5. Albedo ablation study. Left to right: source image, U-Net predicting albedo from a fully diffuse image, U-Net predicting tinted albedo from a fully diffuse image with tint correction, n-diffusions with tint correction. We observed that plain U-Net based approaches had similar performance on face regions but generalised poorly to clothing regions, exhibiting poor color stability and accuracy.

Albedo ablations In Figure 5 we compare our proposed N-diffusion albedo estimation approach to naively predicting the albedo image directly from the fully diffuse image using a U-Net. We also show results for predicting the tinted albedo using another U-Net instead of performing N-diffusions. It can be observed that the naive full image albedo prediction approach suffers from patchiness on clothing due to color ambiguities. While predicting a tinted albedo using U-Net has fewer artifacts, we found that the general color correctness and image quality is higher with the proposed N-diffusion approach. This is also validated through quantitative metrics on Light Stage data in Table 1.

Model	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
N-diffusions	**0.024**	**0.003**	**0.915**	**0.102**
Tinted albedo U-Net	0.030	0.005	0.907	0.107
Albedo U-Net	0.025	0.004	0.911	0.108

Table 1. Quantitative metrics for albedo prediction ablation study.

Face parsing Figure 7 shows additional results for improving face parsing (segmenting parts of the face) in images with difficult lighting conditions. Unusual shadow patterns and blown out pixel areas throw off state-of-the-art face parsing methods, but the results improve significantly after diffusing the lighting.

Naive blending comparison A naive approach to parametric diffusion would be to compute the fully diffuse image and use alpha blending with the original source image. This approach, however, has significant shortcomings, especially around hard shadow edges, which remain in the resulting image. Figure 6 shows a comparison between this simple blending approach and the output of our model.

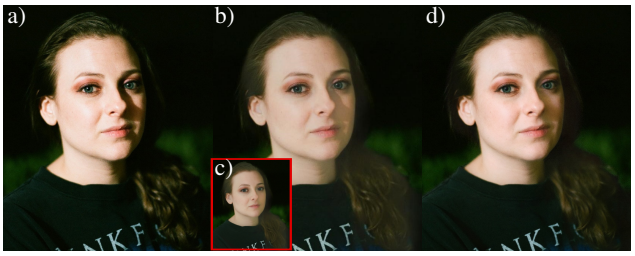

Figure 6. Side by side comparison of linear blending approach vs parametric diffusion. a) input image, b) partial diffusion by linear blending of a) with c), d) output of parametric model. When blending with fully diffuse image, the shadow edge on the cheek stays hard and looks unnatural, in contrast, the edge becomes soft when using parametric diffusion model.

Figure 7. Additional results demonstrating improved face parsing [4] after applying light diffusion.

Additional Diffusion Results Figure 8 shows additional results for full diffusion on a diverse set of in the wild subjects.

Additional Parametric Diffusion Results Figure 9 shows additional results for parametric diffusion on a diverse set of in the wild subjects. These results highlight how having control over the diffusion level can allow the user to select the right diffusion for a given image.

Additional Albedo Comparison Results Figure 10 shows additional results for albedo prediction in comparison to state of the art approaches. Our approach produces an albedo image more robust to external shadows and shows fewer artifacts on clothing.

Additional Portrait Shadow Manipulation Comparisons Figure 11 shows additional results for our full diffusion model on in the wild images as compared to [11].

4. Fairness Study

Research focusing on automatic editing of portrait images inherently raises questions about fairness of outcome across diverse groups of people. To study this issue, we

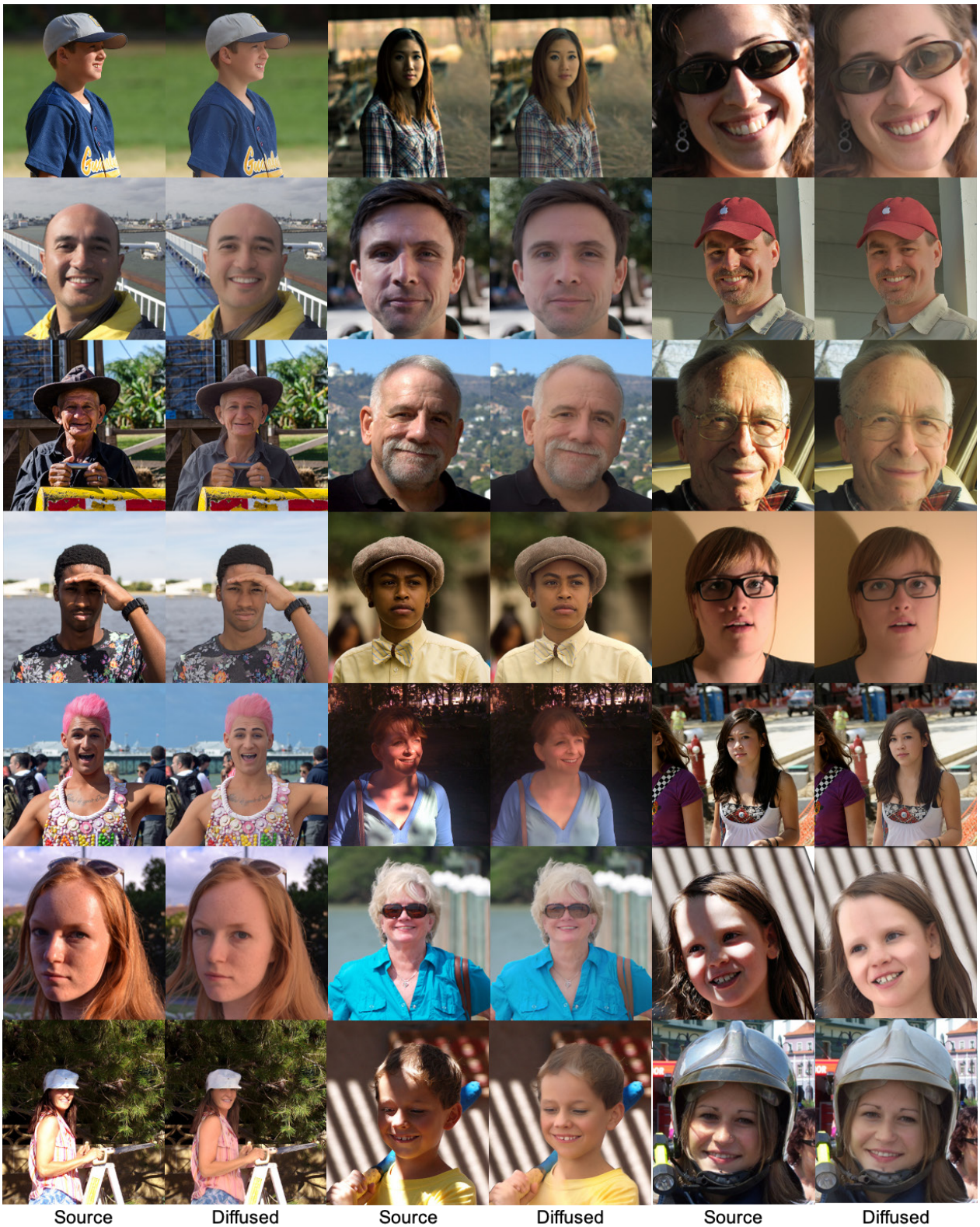

Figure 8. Additional light diffusion results. Note how our method can produce convincing diffused results for a variety of in-the-wild images.

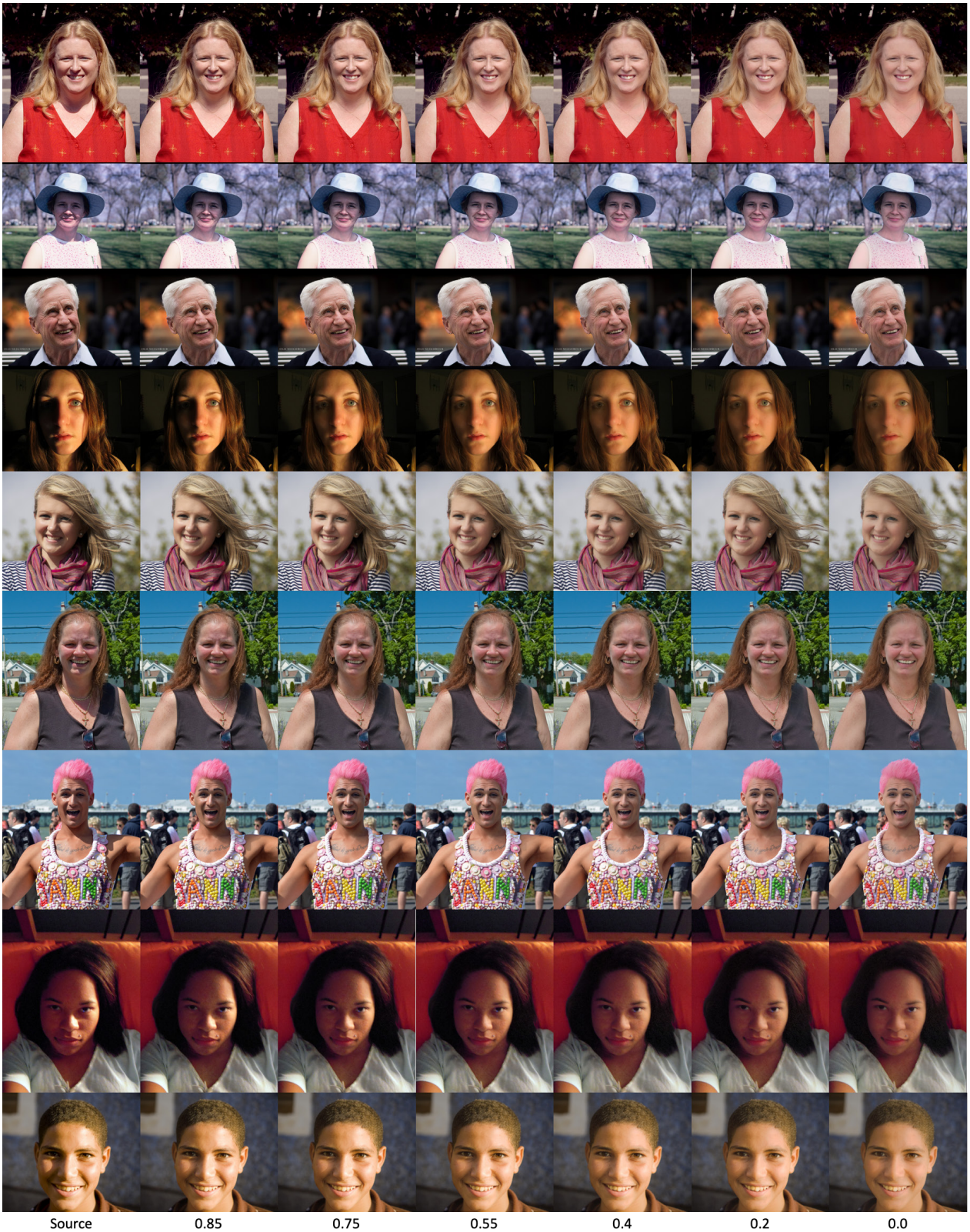

Figure 9. Additional parametric light diffusion results. Our approach provides controllability over the amount of light diffuseness with 0.0 being full diffused.

Figure 10. Additional albedo comparison results. Our approach produces a more consistent albedo over many challenging lighting conditions.

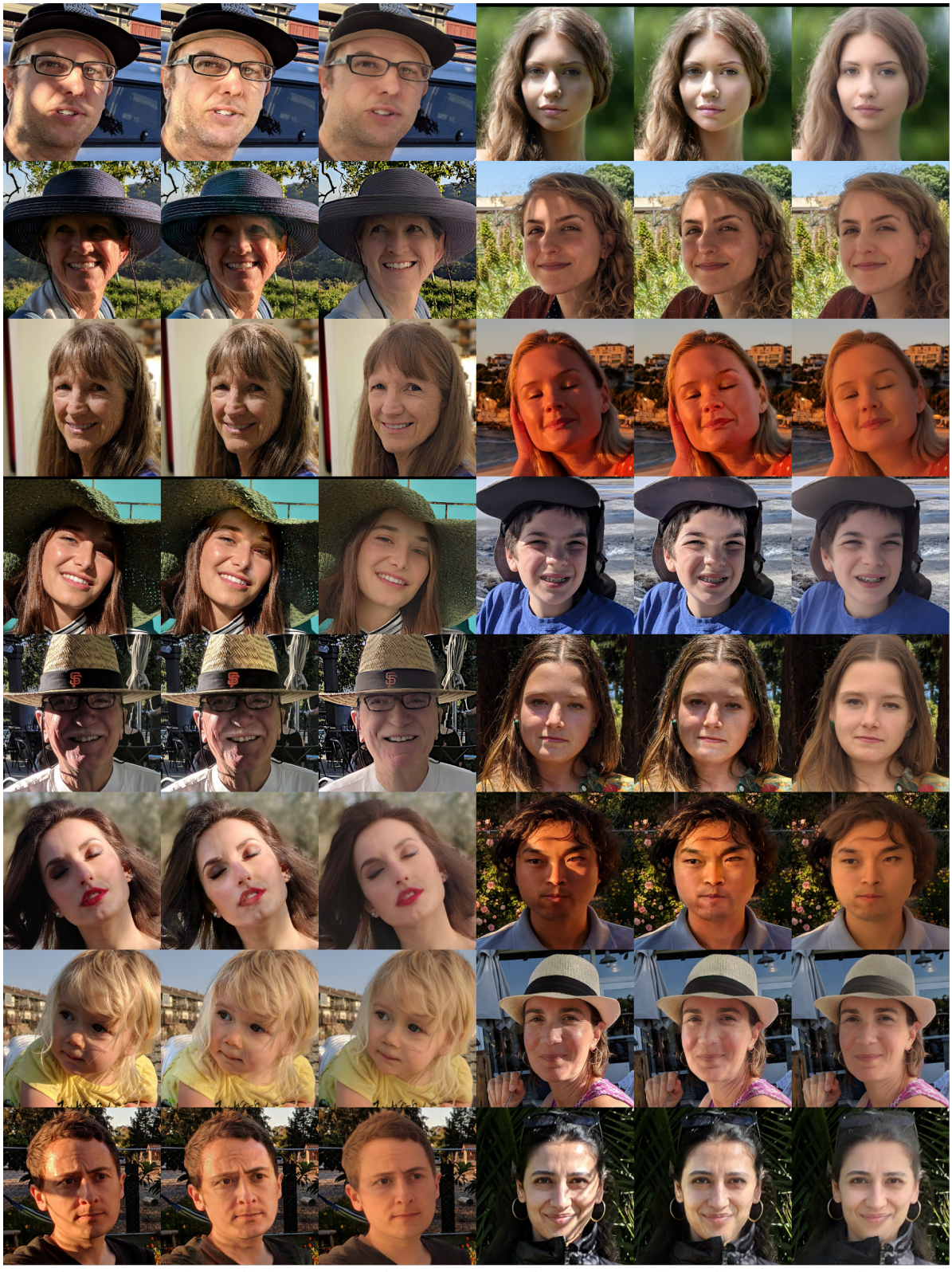

Figure 11. Additional comparisons with Portrait Shadow Manipulation [11]. We show how our approach can more faithfully remove shadows and specular highlights for many challenging cases.

Figure 12. Sample results across the Fitzpatrick skin tone scale, taken from our Lighstage dataset. Left to right is skin tone categories 1 to 6, top to bottom is source image, fully diffused, and albedo images.

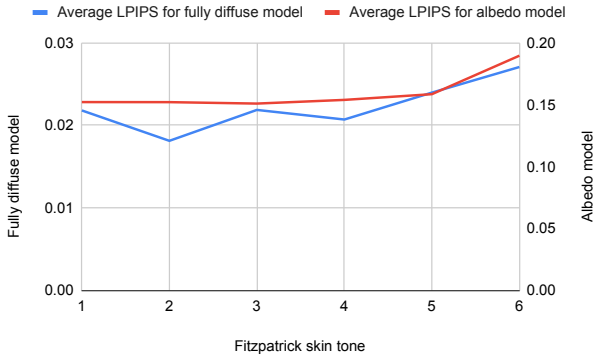

Figure 13. Average LPIPS error for our fully diffuse and albedo models.

clustered our light stage dataset into six skin tones, defined by the Fitzpatrick skin type. Figure 12 shows a sample of subjects. We then computed our quantitative error metrics for each cluster to analyse how well our models treat each skin type. The results for the fully diffused and albedo models are shown in Tables 2 and 3. We note that while most error metrics do not show significant correlation across skin tones, LPIPS does show a small increase for the darkest categories, which is more pronounced in the albedo model. Figure 13 shows this trend. Although no strong bias is apparent in our results, we hypothesize the need for detailed user studies to gauge the correlation between these metrics and perceived skin tone.

Skin tone	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
1	0.0084	0.00036	0.98	0.022
2	0.0076	0.00028	0.99	0.018
3	0.0075	0.00030	0.98	0.022
4	0.0068	0.00024	0.98	0.021
5	0.0077	0.00030	0.98	0.024
6	0.0079	0.00029	0.98	0.027

Table 2. Fairness study for diffusion.

Skin tone	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
1	0.12	0.033	0.82	0.15
2	0.12	0.034	0.83	0.15
3	0.12	0.032	0.84	0.15
4	0.12	0.032	0.83	0.15
5	0.13	0.036	0.81	0.16
6	0.13	0.034	0.77	0.19

Table 3. Fairness study for albedo prediction.

5. Additional Applications

Video In addition to single-photo enhancement, we can apply our light diffusion technique frame-by-frame to video. See the supplementary video for examples.

Editing with shadow and specular maps Besides light diffusion of photos taken under difficult illumination conditions our approach can also be used to control the amount of shadowing as well as the amount of specular highlights. To do so, we leveraged information stored in the two intermediate shadow D and specular S maps (see Fig. 14b–c) estimated using the shadow+specular network operating on the original image I (Fig. 14a). We combine the fully diffused image I_d (Fig. 14d) with S and D to produce enhanced image I_e:

$$I_e = (1 - w_d \cdot D) \cdot I_d + w_s \cdot S. \quad (1)$$

Here weight w_d can be used to adjust the strength of shadows while w_d controls the amount of glossiness. Two examples of I_e are visible in Figures 14e–f. With this approach we can produce specular-free photo with stronger shadows (Fig. 14e) or shadow suppression with enhanced specular highlights (Fig. 14f).

References

[1] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. SIGGRAPH '00, page 145–156, USA, 2000. ACM Press/Addison-Wesley Publishing Co. 1

[2] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang,

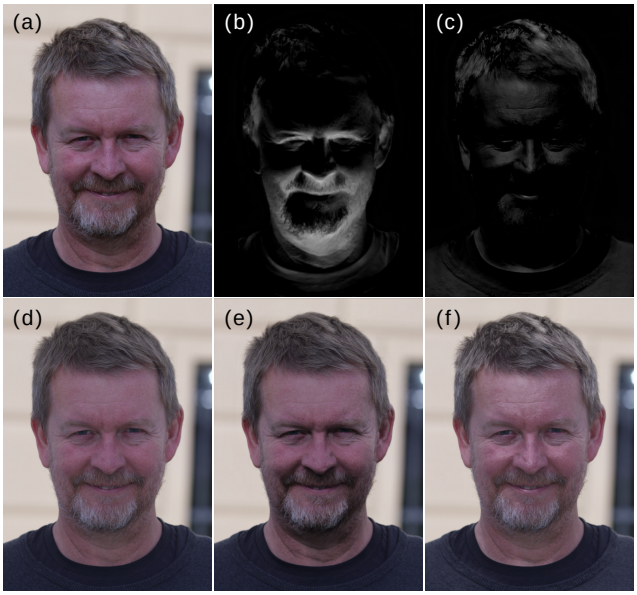

Figure 14. An example of image enhancement combining the diffused image with the shadow and specular maps. The original image I (a) is fed into the shadow+diffusion network to produce shadow D (b) and specular S (c) maps. Then the diffusion network is used to produce fully diffused image I_d (d). By combining I_d with D and S, we can produce enhanced images I_e that, e.g., contain less specular highlights and stronger shadows (e) or vice versa (f).

Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The relightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), nov 2019. 1

[3] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 4

[4] Yiming Lin, Jie Shen, Yujiang Wang, and Maja Pantic. Roi tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing*, 112:104190, 2021. 5

[5] Xudong Mao, Qing Li, Haoran Xie, Raymond Y. K. Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2813–2821, 2017. 4

[6] Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe LeGendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. Deep relightable textures: Volumetric performance capture with neural rendering. *ACM Transactions on Graphics*, 2020. 1

[7] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 1

[8] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T. Barron, and Ravi Ramamoorthi. Light stage super-resolution: Continuous high-frequency relighting. *ACM Transactions on Graphics*, 2020. 1

[9] Greg Zaal, Sergej Majboroda, and Andreas Mischok. Poly haven, 2022. 1

[10] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 4

[11] Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. Portrait shadow manipulation. volume 39, 2020. 2, 5, 9