Real-Time Relighting of Human Faces with a Low-Cost Setup

MačEk, Nejc; Usta, Baran; Eisemann, Elmar; Marroquim, Ricardo

DOI
10.1145/3522626

Publication date
2022

Document Version
Final published version

Published in
Proceedings of the ACM on Computer Graphics and Interactive Techniques

Real-Time Relighting of Human Faces with a Low-Cost Setup

NEJC MAČEK, Delft University of Technology, The Netherlands
BARAN USTA, Delft University of Technology, The Netherlands
ELMAR EISEMANN, Delft University of Technology, The Netherlands
RICARDO MARROQUIM, Delft University of Technology, The Netherlands

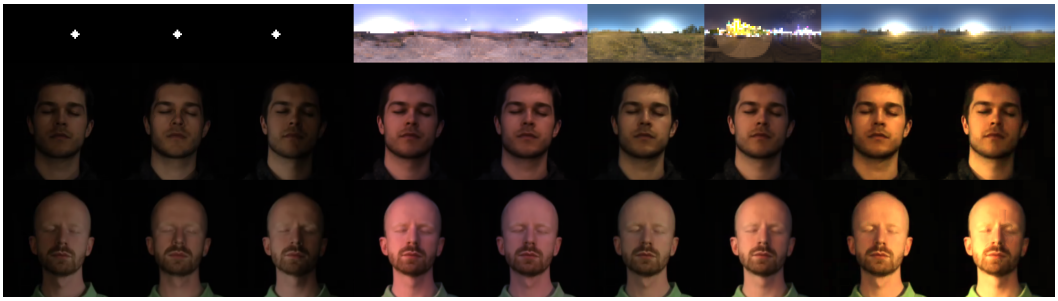

Fig. 1. Two faces acquired with our method, relit (middle and bottom row) with different lighting environments (top row).

Video-streaming services usually feature post-processing effects to replace the background. However, these often yield inconsistent lighting. Machine-learning-based relighting methods can address this problem, but, at real-time rates, are restricted to a low resolution and can result in an unrealistic skin appearance. Physically-based rendering techniques require complex skin models that can only be acquired using specialised equipment. Our method is lightweight and uses only a standard smartphone. By correcting imperfections during capture, we extract a convincing physically-based skin model. In combination with suitable acceleration techniques, we achieve real-time rates on commodity hardware.

CCS Concepts: • **Computing methodologies** → **Reflectance modeling**.

Additional Key Words and Phrases: relighting, reflectance function, face capture

ACM Reference Format:
Nejc Maček, Baran Usta, Elmar Eisemann, and Ricardo Marroquim. 2022. Real-Time Relighting of Human Faces with a Low-Cost Setup. *Proc. ACM Comput. Graph. Interact. Tech.* 5, 1, Article 14 (May 2022), 19 pages. https://doi.org/10.1145/3522626

Authors' addresses: Nejc Maček, Delft University of Technology, Van Mourik Broekmanweg 6, Delft, The Netherlands, macek.nejc@gmail.com; Baran Usta, Delft University of Technology, Van Mourik Broekmanweg 6, Delft, The Netherlands, b.usta@tudelft.nl; Elmar Eisemann, Delft University of Technology, Van Mourik Broekmanweg 6, Delft, The Netherlands, e.eisemann@tudelft.nl; Ricardo Marroquim, Delft University of Technology, Van Mourik Broekmanweg 6, Delft, The Netherlands, r.marroquim@tudelft.nl.

2577-6193/2022/5-ART14 $15.00
https://doi.org/10.1145/3522626

1 INTRODUCTION

The appearance of a face can vary drastically depending on the lighting conditions. Modifying the illumination of an image or video after its acquisition is known as *relighting*, which requires predicting the light transport. Albeit a challenge, many applications such as photo/video editing and online video streaming could greatly benefit from it. For instance, virtual backgrounds are widely used but the viewer's experience is undermined by the physical lighting often being disassociated from the backgrounds. Our work aims at narrowing this gap.

There are two main trends for performing face relighting: image-based and learning-based. The first relies on sampling the light transport function by capturing the appearance of the face under various view and/or light directions. With enough samples it is possible to reconstruct a 3D face model and use the sampled information to simulate novel lighting conditions. High-end devices, such as light stages, use many cameras and hundreds of light sources in a dome-like structure. These custom-made devices serve as benchmarks but target professional users, such as the movie industry, and not the general public. Learning-based approaches are a recent alternative to tackle portrait relighting and have gained popularity. Nevertheless, many methods require training data from Light Stage-like devices and do not easily generalise. The other methods rely on pre-processing large datasets of facial images for training but produce results that lack the visual richness of natural portraits.

We propose a method that considerably lowers the requirements of image-based approaches. We want to bring such acquisition methods to the masses by relying only on a cellphone flash, a camera (possible another cellphone) and a reflecting sphere. The acquisition process is manually performed by moving the flashing cellphone around the subject, and it lasts no longer than one minute. The rest of the process requires minimal user intervention. Our method produces reflectance maps per pixel that can be used for relighting, as illustrated in Figure 1. Furthermore, we demonstrate how our method can be applied for real-time relighting applications, such as video streaming.
Our contributions can be summarised as follows:

- A simple acquisition for human-face reflectance (Sec. 3.1);
- A reflectance reconstruction via sparse samples (Sec. 3.2);
- A proof of concept for real-time use during video-streaming.

2 RELATED WORK

A direct way to perform relighting is by sampling, interpolating and extrapolating the light transport function. Image-based acquisition methods have been successful in this task by carefully sampling the function in a controlled lighting environment in a One-Light-at-A-Time (OLAT) manner [Debevec et al. 2000; Garg et al. 2006; Hawkins et al. 2001, 2004; Weyrich et al. 2006].

Debevec et al. [Debevec et al. 2000] proposed the first Light Stage to extract reflectance fields for faces. Two fixed cameras and a light source attached to a gantry that rotates around the subject's face are used for capture. Moreover, polarising filters separate diffuse and specular components. A reflectance map is created per vertex to perform relighting. Newer Light-Stage versions involve more cameras and a larger number of fixed light sources in a dome-like structure, allowing for increasingly denser and precise capturing of human faces [Debevec 2012]. Such dome-based setups are commonly used in the entertainment industry and can be extended to performance capture [Peers et al. 2007] and the full human body [Einarsson et al. 2006].

Lighter versions of the Light Stage lower the setup requirements by focusing on diffuse and specular maps instead of reflectance fields. Multiview single-shot or quasi-instantaneous methods are particularly attractive since they avoid the issues of head movement [Fyffe and Debevec 2015; Fyffe et al. 2016; Ghosh et al. 2011; Gotardo et al. 2018; Ma et al. 2007; Riviere et al. 2020].

Nevertheless, such systems still require complex setups where each multiview shot needs to be taken by several synchronised cameras or under special lighting, such as polarised illumination patterns and filters. We aim at a wide-spread use of face appearance capture, mostly relying only on common devices, such as a pair of cellphones, and simple lightning conditions.

Reflectance fields, even when acquired with such domes are considered sparse for many applications and require interpolation [Fuchs et al. 2007]. Recently, multi-view OLAT data has been used as training sets for many machine learning-based approaches to increase the lighting resolution [Sun et al. 2020; Zhang et al. 2021]. Xu et al.'s deep neural network for relighting requires only five images with different light directions as input [Xu et al. 2018]. Yet, their training is based on renderings with pre-defined BRDFs, which cannot capture the complexity and broad variation of human faces.

Using a neural-network trained with Light-Stage data, Meka et al. [Meka et al. 2019] proposed a method that only requires two images per subject, but these two images need to be acquired under special gradient lights. Other methods use such data to train a network that accepts a single portrait as input [Legendre et al. 2020; Nestmeyer et al. 2020; Pandey et al. 2021; Sun et al. 2019; Wang et al. 2020; Yamaguchi et al. 2018]. While some of these methods produce convincing results, the training data is complex. It is often difficult to acquire and, therefore, scarce. In consequence, extrapolation is needed and some networks exhibit issues in this regard. Geng et al. [Geng et al. 2021] found ways of improving their results using a loss function that compares extrapolated results to the original portrait directly, however, their method does not generalise well to arbitrary environmental lightning.

To lower training-data requirements, one can fit a 3D mesh to single portraits to synthesise data [Sengupta et al. 2018; Zhou et al. 2019] or perform relighting between two portraits [Shu et al. 2017]. Nevertheless, such methods cannot generalise illumination conditions and mostly rely on a simple Lambertian model for the face, which cannot match the quality of directly sampling the light transport function.

Our method relates to Polynomial Texture Maps (PTM) [Malzbender et al. 2001], another OLAT method, targeting relief-type surfaces. Widely used in Cultural Heritage, PTMs rely on a single fixed camera and sampling the light source on a hemisphere around an object. We use a similar acquisition method adapted to low-cost environments and human faces. While PTMs encode spatially-varying reflectance information as biquadratic polynomials stored per-texel, we instead encode this information in a discrete spatial reflectance map, inspired by Debevec et al. [Debevec et al. 2000]. This enables us to store complete spacial reflectance information and enables simple filtering and rendering.

3 OUR METHOD

We target a simple acquisition process of no longer than one minute (Sec. 3.1). The subject sits down facing a fixed video camera (e.g., cellphone), while a second person moves a flashing light source approximating a hemisphere in front of the subject. The flash blinks at a lower frequency than the camera is recording. We do not require synchronisation between flash and camera but we extract dark frames (no-flash) for face tracking that register the bright frames (flash).

Flash frames are used to build reflectance maps per pixel (Sec.3.2). Since the light-direction sampling is sparse, we employ an interpolation scheme. We fit a 3D model to one of the frames to extract a normal map and to compute self-shadowing areas for each light direction. We construct a triangulation over the acquired light directions and use spherical barycentric coordinates to interpolate the corresponding reflectance map, while respecting shadow boundaries.

For relighting (Sec. 3.3), one can multiply the pixels of an environment map with the corresponding reflectance maps and sum the contributions. For dynamic relighting of a video, we perform gaze

estimation to retrieve the head rotation. We then relight the static face with the rotated environment map and warp it to the video frame using reenactment approaches.

3.1 Acquisition process

Face capture. For simplicity, instead of a special gantry, we manually move a light source (in form of a cellphone with blinking flash) around the subject, while capturing the face from a single front facing camera (again using a simple camera without special polarising filters). The distance between the light source and the face is kept roughly constant at about 1.5 metres, leading to an approximate sampling on a hemisphere (Figure 2, left). The acquisition takes around one minute, during which the light is typically moved in a zig-zag motion. Light directions are retrieved via a reflecting sphere (Figure 2, right).

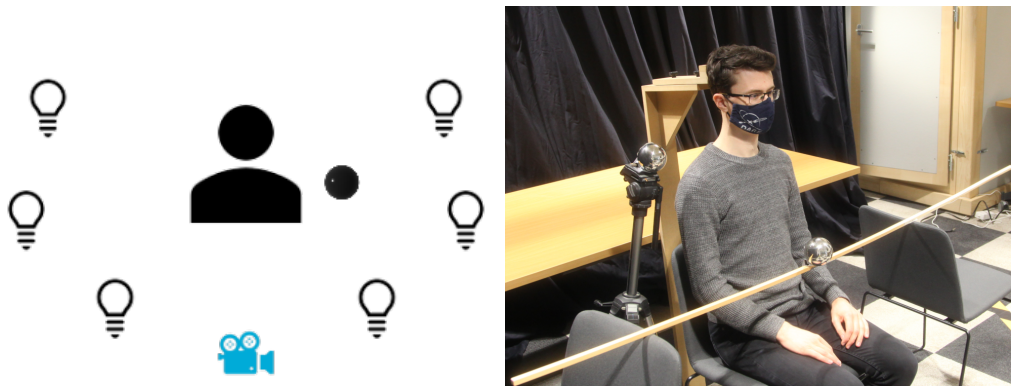

Fig. 2. Left: Schematic acquisition setup. A camera captures the subject's face at 30FPS while a flashing light source at 5Hz is manually moved around the subject. A simple reflecting sphere is placed near the subject to retrieve the light direction from the images. Right: Studio acquisition setup. We placed two reflecting spheres for testing purposes, but we only require one.

As we cannot guarantee that the subject's head will not move during the acquisition and even small movements can lead to blur artefacts, we need to align the face in all captured images. Even though many face-tracking and image registration methods, such as optical flow, exist, they suffer severely under illumination changes [Singh and Arora 2016]. To this extent, we separate flash from no-flash images. The no-flash frames are used for face tracking and, by interpolation, we determine the face in the flash image. Meka et al. [Meka et al. 2019] used similar "tracking frames" but required more controlled conditions with multiple synchronised cameras and light sources. We cover the details in the following.

Flash/No-Flash Frame Extraction. In our method, we acquire videos at 30Hz, while the flash oscillates at 5Hz. Note that there might be fluctuations on the flashing frequency since we rely on simple cellphone applications and there might be lags in-between flashes.

Frames captured with the flash show strong directional lighting cues, contrary to in-between frames mostly lit by weak ambient illumination. While we have approximately three bright and three dark frames per flash, the camera shutter and the flashing light are not synchronised, hence, not all frames are completely lit or dark. Yet, among the three bright and dark frames, at least one should be close to the flash peak and one without any residual light by the flash.

Since we do not have a perfect synchronisation and the brightness of a frame may vary depending on the face illumination, we use a sliding window of size s frames over the video sequence, where s

is the expected number of frames between flash and no-flash frames. Here, $s = \frac{30}{2*5} = 3$. We now look for local minima and maxima of the average grayscale value of the frames. If a frame is a maximum, we classify it as a flash frame. Likewise, we determine the non-flash frames. Figure 3 shows an example.

Fig. 3. A sequence of extracted flash and no-flash frames with the frame numbers below. Note how the spacing between flash and no-flash images is not constant.

For robustness, we filter out flash images whose grayscale average differs from neighbouring no-flash frames by less than a set threshold. We assume in these cases that the flash did not sufficiently light the face because it was ill-positioned. An experimentally-determined threshold of 0.002 worked well for our cases where the captured face's height and width spanned around 1/3 of the uncropped captured images' dimensions. This may need to be adjusted in setups with brighter or darker ambient and flash light, and where the apparent brightness of faces varies due to diversity of subjects or face size due to proximity to the camera. Additionally, we compute the average period between consecutive local minima and maxima, which should be close to $2s$. Deviating by $> s/2$ from the expected period indicates an outlier and the frame is removed. In total, we typically extract around 150-200 flash frames.

Face registration. Related work typically opts for a fast acquisition to minimise head movements [Fyffe and Debevec 2015; Fyffe et al. 2016; Ghosh et al. 2011; Gotardo et al. 2018; Ma et al. 2007; Riviere et al. 2020] and even OLAT rapidly flicker LED lights to keep acquisition time low and benefit from the absence of mechanical light movement. Some work claims that subjects can avoid head motion but they were often capturing them at very low resolutions, hiding small movements. We observed that using a head rest mitigates the issue significantly, however, we want to avoid needing one. Without a proper head rest, we observed noticeable head movement, which requires post processing. Nevertheless, since large head movements are uncommon, we approximate the 3D motion by a rigid transformation in image-space.

We start by detecting the face [Viola and Jones 2004] and its landmarks [Shi et al. 1994] in the first no-flash frame and use a sparse optical-flow tracker [Bouguet et al. 2001] to track the landmarks to subsequent no-flash frames. We remove outliers using RANSAC, and compute the rigid alignment in a least-squares sense using SVD with the remaining points. Due to the high noise rate of the dark frames and inconsistency in brightness due to non-synchronisation, experiments with dense optical flow, as applied by Meka et al. [Meka et al. 2019]), were less successful. In their setup multiple lights could be simultaneously flashed to achieve high-quality tracking frames, which we cannot during our deliberately simple acquisition. Figure 4 illustrates our alignment result.

Once all no-flash frames are aligned, we can use the derived transformations to deform the in-between flash frames. Given the minimal head movements, linear interpolation works sufficiently

Fig. 4. Left: Average of all the dark frames obtained during acquisition. Right: Average of the same images after registration, significantly removing blur artefacts caused by misalignments (insets).

well. In practice, translation is typically below one pixel and never larger than three, while rotation is around 0.1^o and never exceeded 0.35^o, for a 960×540 video.

3.2 Building Reflectance Maps

With the registered flash frames, it is possible to describe the appearance of the face under different incoming light directions. The latter are derived using a reflecting sphere placed next to the subject during acquisition. Its shape is marked for a single frame. In this circle, the light's reflection indicates the incoming light direction. The highlight is detected with sub-pixel precision by thresholding by 0.85 and then averaging the detected pixels using their brightness values. Further processing like normal and albedo extraction could be possible using multiple spheres [Liao et al. 2017], but we limit ourselves to a single sphere to facilitate simplicity of the setup.

The flash frames sample the face BRDF as well as all effects of indirect illumination such as ambient illumination, self-shadowing, subsurface scattering and other global illumination effects [Debevec et al. 2000]. We construct a set of 64×32 reflectance maps, each representing the reflectance of one point on the face, i.e., one pixel. Pixels in the reflectance map represent light directions. More specifically, we use an equirectangular projection to map image-space pixel locations to 3D light directions. In our case, the maps are sparsely sampled (typically around 150-200 light directions) and capture only directions in the frontal hemisphere. The pixels of sampled light directions in reflectance maps are populated with colours from the captured flash frames.

To estimate the missing samples in the reflectance maps, we apply four steps. First, a *triangulation*-based interpolation to populate the missing samples. Second, *Visibility Maps* that help us make self-cast shadows sharper where, otherwise, there would be a smooth transition from a lit pixel to one in shadow. Next, *Light Adaptation* is employed where we utilise surface normals on the face to smooth and correct the interpolated colours. Finally, we apply some additional *adjustments* to accommodate for the effects of the chosen encoding. The remainder of this section covers the details of the four steps.

Triangulation. To colour the missing samples in the reflectance maps, we resort to interpolation over a Delaunay triangulation. Since we do not capture the light's direction from behind the face, we do not have to worry about triangles wrapping around the map's borders. To compensate for the planar interpolation, we use Spherical Barycentric Interpolation [Cabral et al. 1999]. Furthermore, we add virtual "black" light directions to fade out the light contributions beyond the convex hull of the captured light directions.

Visibility Maps. Pure barycentric interpolation exhibits visible artefacts in form of ghost shadows [Xu et al. 2018]. It is smooth but occlusions can actually cause sharp transitions. To address this issue, we fit a 3D model to the face [Guo et al. 2018, 2020] and compute a visibility map for each reflectance map. It contains a one if the light directions corresponding to the pixel are completely visible, a zero if they are blocked, and in-between values for partial visibility.

This allows us to improve the interpolation when one or two of the triangle vertices are not visible. To simplify, we first consider a binary visibility map - one can imagine thresholding the values with 0.5. Without loss of generality, we consider that v_0 represents the direction of a non visible light source while lights at v_1 and v_2 are visible. For every pixel inside the triangle that represents a non-visible light direction, we use the value of v_0, while visible directions interpolate between $v1$ and $v2$ using spherical barycentric coordinates. This process is depicted in Figure 5, left.

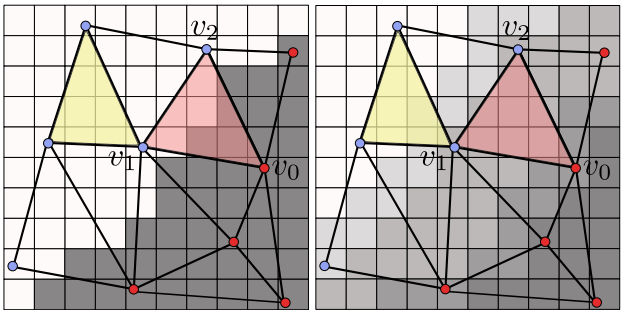

Fig. 5. Using visibility to interpolate inside triangles. Visible, not visible, and partial visible directions are marked as white, dark gray, and intermediate gray values, respectively. Left: Reflectance map with five visible light directions (blue vertices) and five invisible (red vertices). The yellow triangle uses regular interpolation. For the red triangle, white pixels will mostly rely on blue vertices, while gray pixels rely on the red vertex. Right: In practice, we use a smoothed visibility map and generalise this interpolation scheme.

This approach assumes that the 3D model perfectly matches the face. However, since it is just an approximate model, we expect some light directions to be wrongly classified. Considering Figure 5, left, if direction v_0 was wrongly classified as being visible, the red triangle would be interpolated ignoring the visibility map and lose the sharp transition. If a neighbouring reflectance map would classify the samples correctly, this issue would be accentuated.

To mitigate this problem, we smooth the original non-binary visibility map with a 3×3 box filter. Further, we propose an interpolation between both interpolation modes: not using the visibility map (regular barycentric interpolation) and taking the visibility map into account as described above. If all vertex visibilities are the same, we just use regular interpolation, while a large difference of visibility on the vertices implies that we resort to the visibility map. Specifically, we use the difference between the minimum and maximum vertex values as interpolation weight between the two methods. Please note that working with the difference means that in some cases there might not be any vertex with a visibility below 0.5. Instead, we always assume that the minimal vertex value indicates that it is in shadow, the maximum lit. Any other pixel in the visibility map, including the one containing the third vertex is declared lit/shadowed depending on how close its value is to the other two vertex values. Figure 5 illustrates this concept, and Figure 6 shows an example of a real reflectance map and the visibility maps.

Smoothing the visibility map leads to a smoother interpolation but also prevents some sharp shadows from being detected. Nevertheless, since we cannot fully solve the misclassifications with an approximated 3D model, a trade-off remains.

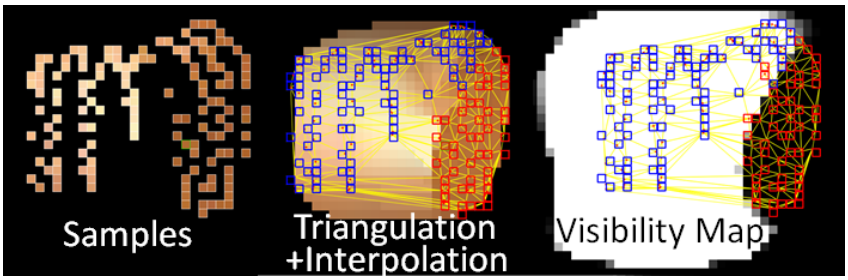

Fig. 6. Left: Sampled light directions. Middle: triangulation and interpolation. Right: visibility map with some misclassifications.

Light Adaptation. The 3D model can be used to also estimate surface normals. Since the model does not span the whole captured face, we smooth the contribution of the model on the borders to avoid visible seams. For the points on the face, where normals are known, we adjust the interpolation within the reflectance map by compensating for the cosine factor due to foreshortening of the projected solid angle, i.e., we divide the value by the corresponding cosine contribution at each vertex, interpolate, and then multiply by the cosine factor of the interpolated position.

We do assume that the direct illumination is dominant. This also holds for ambient light, since the natural solution of subtracting the no-flash from the flash frame was not successful. Many cellphone cameras lack white balance control and perform auto-adjustments (Figure 7). We did not want to make higher-end devices a requirement, but the resulting error when using a darker environment during capture were negligible.

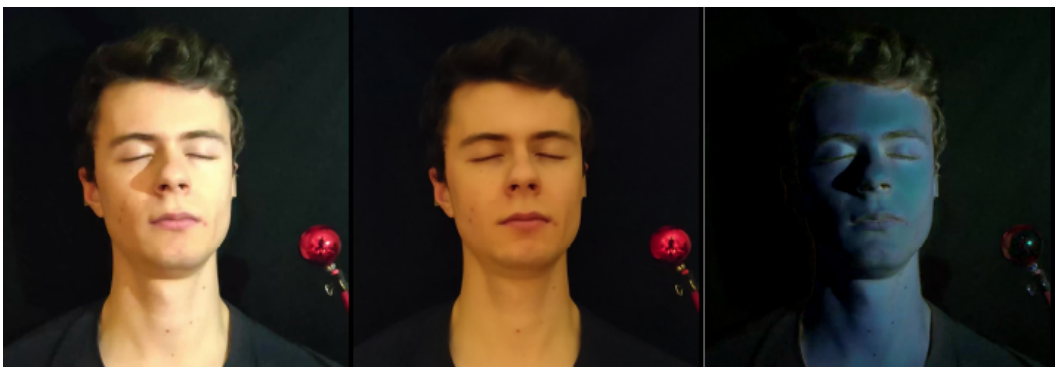

Fig. 7. The difference between a no-flash (centre) and a flash (left) frame does not approximate well the contributions of direct/indirect illumination due to uncontrollable self-adjustments of the cellphone camera under rapid illumination variation.

The previous solution requires a refinement for when any of the cosine factors is negative (or clamped to zero), i.e., the light direction is behind the surface. Here, the correction would be meaningless. Nevertheless, it does not mean that the captured colour is necessarily black, as indirect illumination is recorded as well. Such local occlusions are important to handle because realistic skin appearance requires capturing subsurface scattering effects. Still, if we use a similar approach as described for the visibility map, we could obtain negative values and even cause discontinuity artefacts. Instead, we distinguish three separate cases. First, if all three vertices have a negative

cosine factor, we just perform interpolation as described before. The second case covers two vertices with negative cosine factor. The third will be one negative cosine factor. As these cases are complex, we start with the intuition.

The idea is to use the value at the vertex with negative cosine factor to approximate the indirect contribution at the other vertices. Specifically, we remove the indirect illumination, perform the cosine correction for the direct illumination, and then add back the indirect component that is slowly faded out towards the two lit vertices. To simplify, let us consider a 1D case of interpolating between vertex v_0 with only indirect illumination and v_1 that has both direct and indirect contributions. We use the colour at v_0 as an approximation of the indirect component at v_1. Hence the contribution of the direct component to a point p becomes:

$$c_{\text{dir}} = (c_1 - w_0 c_0)\frac{\delta_p}{\delta_1} + w_0 c_0, \tag{1}$$

where c_i, w_i and δ_i are, respectively, the colour, barycentric weight, and cosine factor at vertex i. Finally, we can compute the colour at point p as $c_p = w_1 c_{\text{dir}} + w_0 c_0$,

We can now extend this idea to 2D to interpolate inside the triangles. If two vertices have negative cosine factors, say v_0 and v_1, we can first interpolate their indirect contributions as $c_{\text{ind}} = (w_0 c_0 + w_1 c_1)$ and their combined weight as $w_{\text{ind}} = w_0 + w_1$. The contribution of the direct illumination can then be defined similarly to Eq. 1, with c_p computed as before:

$$c_{\text{dir}} = (c_2 - w_{\text{ind}} c_{\text{ind}})\frac{\delta_p}{\delta_2} + w_{\text{ind}} c_{\text{ind}}. \tag{2}$$

For the final case of v_0 and v_1 having a positive cosine factor, we first interpolate between v_1 and v_2 and compute $c_p = (w_1 + w_2)c_{\text{dir}} + c_0 w_0$, where the direct contribution c_{dir} is given by:

$$c_{\text{dir}} = \left[\frac{w_1(c_1 - w_0 c_0)}{\delta_1} + \frac{w_2(c_2 - w_0 c_0)}{\delta_2}\right]\frac{\delta_p}{w_1 + w_2} + w_0 c_0. \tag{3}$$

Figure 8 shows a reflectance map created using the described interpolation methods and a comparison with ground truth data for a synthetic example.

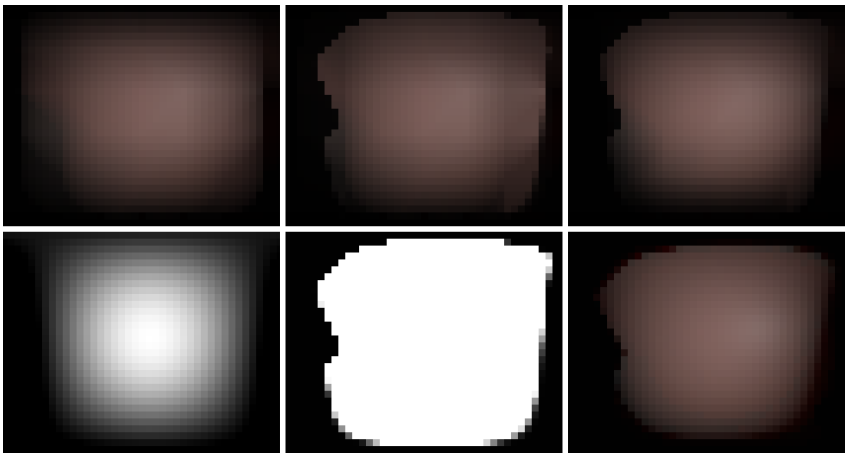

Fig. 8. Reflectance map of one vertex of the synthetic face. Top row (left to right): spherical barycentric interpolation, added sharp visibility maps, added light adaptation. Bottom row (left to right): visualisation of cosine weights, visibility map, ground truth.

Adjustments. We apply gamma correction and all calculations are done in linear sRGB space. Similar to Debevec et al. [Debevec et al. 2000], we normalise reflectance maps after they are computed. Normalisation accounts for the fact that in an equirectangular projection, pixels at the poles represent smaller areas. We multiply the pixel intensity by the sine of its corresponding latitudinal coordinate, such that values at the poles are close to zero, whereas the values at the equator remain mostly unchanged (Figure 9).

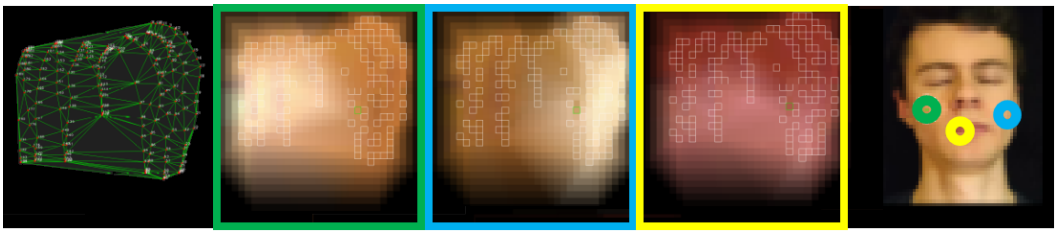

Fig. 9. Left: Triangulated light source samples. Other images: Reflectance maps for different positions on the face (right).

3.3 Relighting

With a reflectance map per pixel, we can perform relighting on a static image by convolving the reflectance map by a given environment map, i.e., multiplying corresponding pixels and summing their values. We perform these steps on the GPU in real-time enabling interactive illumination changes.

To support head rotation, we observe that for the relighting, it has the same effect as rotating the environment map in the opposite direction. Therefore, we perform relighting on the static frontal image using the rotated environment map, and finally transfer the relit face to the video frame using a reenactment approach [Siarohin et al. 2019] (Fig 13). The rotation angle is determined using an estimation of the gaze direction [Guo et al. 2018, 2020]. The resulting image quality is limited by the reenactment algorithm but it shows, in form of a proof of concept, that our acquisition method can be used for applications like video streaming.

4 RESULTS

We have captured several faces with the setup illustrated in Fig 2 and two more were captured independently in a home setup. Figures 10 and 1 illustrate relighting results with different environment maps. For these results we have adjusted the contrast of the final images in order to make them more visible (Figure 15). Figure 11 shows results for an environment map simulating a single light source at different positions.

Figure 12 shows the sampled light directions for three different acquisitions. Note how our method is robust to very different sampling patterns and densities.

In order to have a ground-truth comparison, we applied our method to synthetic data [McGuire 2017]. We simulated the entire acquisition using a 3D head model, including reconstructing a new approximate 3D model. Figure 14 illustrates two comparative examples. We used one of the original sampling patterns that have sparser samples for the sides. Consequently, the relit model fails to capture light coming from behind or directly from the side of the head.

For a more quantitative comparison, we rendered an error heat map of the face. The error for each point on the face is computed by averaging squared RGB distances between all corresponding

pairs of pixels in the ground truth and recreated reflectance maps, weighted by the sine of their latitudinal coordinate.

Limiting the error computation to the area of the convex hull of sampled light directions shows which areas suffer most from an approximate interpolation. We notice three distinct areas. First, on the forehead due to lack of coverage of the 3D model. Second, around the nose due to incorrect self-cast shadows caused by inaccurate shadow border. Finally, on the dorsal of the nose as an effect of the lack of interpolation support where a lot of subsurface scattering effects are present when the light is placed on the side. In all, we see that most differences come from inaccuracies of the reconstructed 3D model.

We further compared our results to learning-based methods with publicly available code [Xu et al. 2018; Zhou et al. 2019]. As Zhou et al. [Zhou et al. 2019] only support a Spherical-Harmonics illumination model and gray scale maps, we have converted a high dynamic range (HDR) environment map to this representation and then produced an equirectangular map and a light probe (Mirror ball) from the SH representation to account for the loss of high-frequencies in that method to enable a fair comparison. We also compared the results using the LDR illumination examples provided directly by Zhou et al [Zhou et al. 2019]. Regarding the method of Xu et al. [Xu et al. 2018], it requires a set of input images, where the associated incident illumination angles have a tolerance error of at the most 10 degrees, we used our incident angle extraction method to specify the input images for their model.

Figure 16 shows that the method of Zhou et al. produced visible artefacts for the LDR illumination and failed to produce a consistent relit image for HDR. While, the method from Xu et al. produced comparable results to our method for the synthetic case, it resulted in some artefacts on the forehead (centre column) and background (left column). We suspect that the noise present in the real images might disturb their model.

Figure 17 compares our results against the method from Xu et al. and the ground truth using an HDRI map. While both methods were able to reproduce the shadow on the left side of the face, we notice that Xu et al. produces an exaggerated specular appearance that can be seen on the cheek and forehead. In addition, their method produces a bluish appearance due to the dominant sky colour in the map, while ours is able to better reproduce the colours.

5 LIMITATIONS

Without synchronisation, we need to consider a margin to ensure at least one dark frame without residual flash light. This imposes a limit on the flash frequency and, thus, acquisition time. Nevertheless, the current acquisition time of approximately one minute was not considered an issue by the subjects. A faster flash frequency and higher camera frame rate could be used but this goes against our goal of acquisition simplicity, which includes the reflectance sphere. While it is not something that is found in every home, we successfully tested our system with cheap stainless steel garden decoration spheres and Christmas balls that can easily be purchased.

Our setup does make acquiring images from the back hemisphere challenging. While the actual back of the head is usually not interesting for relighting, we miss scattering effects on the ears and the side of the face. Moreover, our shadow boundary interpolation and normal estimation rely on a 3D model approximation. The current method is still coarse and does not reconstruct the whole face. As can be noted from the result, some seams remain visible in some images. Nevertheless, we were able to produce convincing relighting results with our extremely simplified acquisition method.

Finally, since we need some amount of ambient light to use the dark frames for tracking, we cannot reproduce very strong directional lighting. On the other hand, requiring a very dark environment for acquisition also makes its usage less attractive for general users. Again, given the trade-off,

we opt for the easiest acquisition scenario. This problem would benefit from future research on mitigating camera self-balancing issues and utilising the difference between flash and no-flash frames to equalise and extract the contrast between them. Currently, this results in brighter or darker relighting results.

6 CONCLUSIONS

We have presented an end-to-end relighting method for the masses. Our acquisition approach is distinct from existing approaches in that it requires no special equipment and enables capturing a subject's face within one minute using only a manually-displaced flashing light and a cellphone camera. We have addressed three main issues that this low-cost setup presented. First, we mitigated the possibility of facial movements with a stabilisation algorithm. Second, we resolved sparse sampling using a smart interpolation technique. And finally, there are camera self-adjustments , which require further research to resolve, but would help normalise flash/no-flash brightness variation. The final per-pixel reflectance maps can be readily used for real-time relighting. As a proof of concept, we have presented an extension of our method to videos, which can be of benefit when using virtual backgrounds.

There are several future directions to explore. The interpolation could benefit from an advanced 3D-face model matching and robust reenactment methods would improve video performance. Nevertheless, since our method does not depend on any particular algorithm, these are exchangeable.

Given the acquisition time and flashing light, many subjects prefer to keep their eyes closed during acquisition. A natural improvement would be to use one image of the subject with eyes open, and transfer the relighting data to this image to act as the based relighting image for video sequences.

ACKNOWLEDGMENTS

This work was supported by DyViTo that is funded by the European Union's Horizon 2020 programme under grant agreement No 765121. This work was supported by The Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia.

REFERENCES

Jean-Yves Bouguet et al. 2001. Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm. *Intel corporation* 5, 1-10 (2001), 4.

Brian Cabral, Marc Olano, and Philip Nemec. 1999. Reflection space image based rendering. *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 1999* (1999), 165–171. https://doi.org/10.1145/311535.311553

Paul Debevec. 2012. The Light Stages and Their Applications to Photoreal Digital Actors. In *SIGGRAPH Asia*. Singapore.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics* (2000), 145–156. https://doi.org/10.1145/344779.344855

Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. 2006. Relighting Human Locomotion with Flowed Reflectance Fields. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques* (Nicosia, Cyprus) *(EGSR '06)*. Eurographics Association, Goslar, DEU, 183–194.

Martin Fuchs, Hendrik PA Lensch, Volker Blanz, and Hans-Peter Seidel. 2007. Superresolution reflectance fields: Synthesizing images for intermediate light directions. In *Computer Graphics Forum*, Vol. 26. Wiley Online Library, 447–456.

Graham Fyffe and Paul Debevec. 2015. Single-Shot Reflectance Measurement from Polarized Color Gradient Illumination. *2015 IEEE International Conference on Computational Photography, ICCP 2015 - Proceedings* (2015). https://doi.org/10.1109/ICCPHOT.2015.7168375

G. Fyffe, P. Graham, B. Tunwattanapong, A. Ghosh, and P. Debevec. 2016. Near-instant capture of high-resolution facial geometry and reflectance. *Computer Graphics Forum* 35, 2 (2016), 353–363. https://doi.org/10.1111/cgf.12837

Gaurav Garg, Eino-Ville Talvala, Marc Levoy, and Hendrik P Lensch. 2006. Symmetric Photography: Exploiting Data-sparseness in Reflectance Fields. *Eurographics Symposium on Rendering (EGSR)* (2006), 251–262.

Jiahao Geng, Yanlin Weng, Lvdi Wang, and Kun Zhou. 2021. Single-view facial reflectance inference with a differentiable renderer. *Science China Information Sciences* 64, 11 (2021), 1–17.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview Face Capture Using Polarized Spherical Gradient Illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference* (Hong Kong, China) *(SA '11)*. ACM, New York, NY, USA, Article 129, 10 pages. https://doi.org/10.1145/2024156.2024163

Paulo Gotardo, Jérémy Riviere, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2018. Practical dynamic facial appearance modeling and acquisition. *SIGGRAPH Asia 2018 Technical Papers, SIGGRAPH Asia 2018* 37, 6 (2018). https://doi.org/10.1145/3272127.3275073

Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 2018. 3DDFA. https://github.com/cleardusk/3DDFA.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Tim Hawkins, Jonathan Cohen, and Paul Debevec. 2001. A Photometric Approach to Digitizing Cultural Artifacts. In *Proceedings of the 2001 Conference on Virtual Reality, Archeology, and Cultural Heritage* (Glyfada, Greece) *(VAST '01)*. ACM, New York, NY, USA, 333–342. https://doi.org/10.1145/584993.585053

Tim Hawkins, Andreas Wenger, Chris Tchou, Andrew Gardner, Fredrik Göransson, and Paul Debevec. 2004. Animatable Facial Reflectance Fields. In *Proceedings of the Fifteenth Eurographics Conference on Rendering Techniques* (Norrköping, Sweden) *(EGSR'04)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 309–319. https://doi.org/10.2312/EGWR/EGSR04/309-319

Chloe Legendre, Wan Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. 2020. Learning Illumination from Diverse Portraits. *SIGGRAPH Asia 2020 Technical Communications, SA 2020* (2020). https://doi.org/10.1145/3410700.3425432 arXiv:2008.02396

Jingtang Liao, Bert Buchholz, Jean-Marc Thiery, Pablo Bauszat, and Elmar Eisemann. 2017. Indoor Scene Reconstruction Using Near-light Photometric Stereo. *IEEE Transactions on Image Processing* 26, 3 (March 2017), 1089–1101. http://graphics.tudelft.nl/Publications-new/2017/LBTBE17 DOI: 10.1109/TIP.2016.2636661.

Wc Ma, Tim Hawkins, and P Peers. 2007. Rapid acquisition of specular and diffuse normal maps from polarized spherical gradient illumination. *Proceedings of the 18th Eurographics conference on Rendering Techniques* (2007), 183–194. http://dl.acm.org/citation.cfm?id=2383873

Tom Malzbender, Dan Gelb, and Hans Wolters. 2001. Polynomial texture maps. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001* (2001), 519–528. https://doi.org/10.1145/383259.383320

Morgan McGuire. 2017. Computer Graphics Archive. https://casual-effects.com/data

Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019. Deep Reflectance Fields - High-Quality Facial Reflectance Field Inference From Color Gradient Illumination. *ACM Transactions on Graphics (Proceedings SIGGRAPH)* 38, 4. https://doi.org/10.1145/3306346.3323027

Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M Lehrmann. 2020. Learning Physics-guided Face Relighting under Directional Light. In *Conference on Computer Vision and Pattern Recognition*. IEEE/CVF, 5123–5132.

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

Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. 2007. Post-production facial performance relighting using reflectance transfer. *ACM Transactions on Graphics* 26, 99 (2007), 52. https://doi.org/10.1145/1239451.1239503

Jérémy Riviere, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. 2020. Single-Shot High-Quality Facial Geometry and Skin Appearance Capture. *ACM Trans. Graph.* 39, 4, Article 81 (July 2020), 12 pages. https://doi.org/10.1145/3386569.3392464

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2018. SfSNet: Learning Shape, Reflectance and Illuminance of Facesin the Wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6296–6305.

Jianbo Shi et al. 1994. Good features to track. In *1994 Proceedings of IEEE conference on computer vision and pattern recognition*. IEEE, 593–600.

Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017. Portrait Lighting Transfer Using a Mass Transport Approach. *ACM Trans. Graph.* 37, 1, Article 2 (Oct. 2017), 15 pages. https://doi.org/10.1145/3095816

Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Conference on Neural Information Processing Systems (NeurIPS)*.

Manminder Singh and AS Arora. 2016. Varying illumination and pose conditions in face recognition. *Procedia Computer Science* 85 (2016), 691–695.

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single Image Portrait Relighting. *ACM Trans. Graph.* 38, 4, Article 79 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323008

Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun Ta Tsai, Jonathan T. Barron, and Ravi Ramamoorthi. 2020. Light Stage Super-Resolution: Continuous High-Frequency Relighting. *ACM Transactions on Graphics* 39, 6 (2020). https://doi.org/10.1145/3414685.3417821 arXiv:2010.08888

Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *International journal of computer vision* 57, 2 (2004), 137–154.

Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics* 39, 6 (2020). https://doi.org/10.1145/3414685.3417824

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. 2006. Analysis of Human Faces using a Measurement-Based Skin Reflectance Model. *ACM Trans. on Graphics (Proc. SIGGRAPH 2006)* 25, 3 (2006), 1013–1024. https://doi.org/10.1145/1179352.1141987

Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics* 37, 4 (2018), 1–13. https://doi.org/10.1145/3197517.3201313

Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Xiuming Zhang, Sean Fanello, Yun Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, Jonathan T. Barron, Ravi Ramamoorthi, and William T. Freeman. 2021. Neural Light Transport for Relighting and View Synthesis. *ACM Transactions on Graphics* 40, 1 (2021). https://doi.org/10.1145/3446328 arXiv:2008.03806

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. 2019. Deep single-image portrait relighting. *Proceedings of the IEEE International Conference on Computer Vision* 2019-Octob (2019), 7193–7201. https://doi.org/10.1109/ICCV.2019.00729

Fig. 10. Left column shows a captured bright frame. The other columns show relighting results with the respective environment maps on top. The last two rows show results for home-made acquisitions.

Fig. 11. Results with an environment map simulating a single small light source.

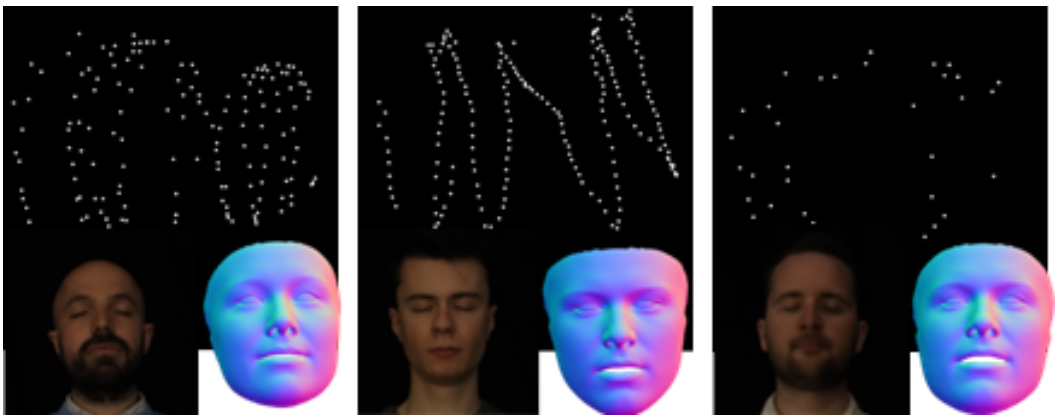

Fig. 12. Light-direction sampling for three different acquisitions, including a frame and extracted normal map.

Fig. 13. Exemplary frames of a relit video using different environment maps. Relighting was transferred from the original image (leftmost column) using the reenactment approach from Siarohin et al. [Siarohin et al. 2019].

Fig. 14. Ground truth comparison. First two columns: ground truth in the bottom row, relit counterparts in the top row. Last column: normalised error heat map computed over the whole area of reflectance maps (top) and only over the area of the convex hull of the sampled light directions (bottom). The heat maps were normalised separately.

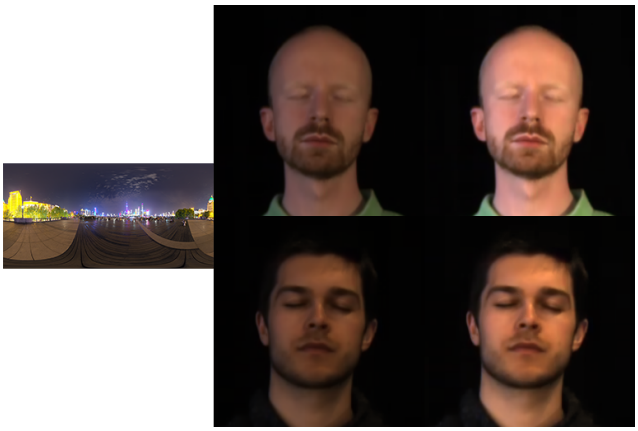

Fig. 15. Without (middle) and with contrast adjustment (right).

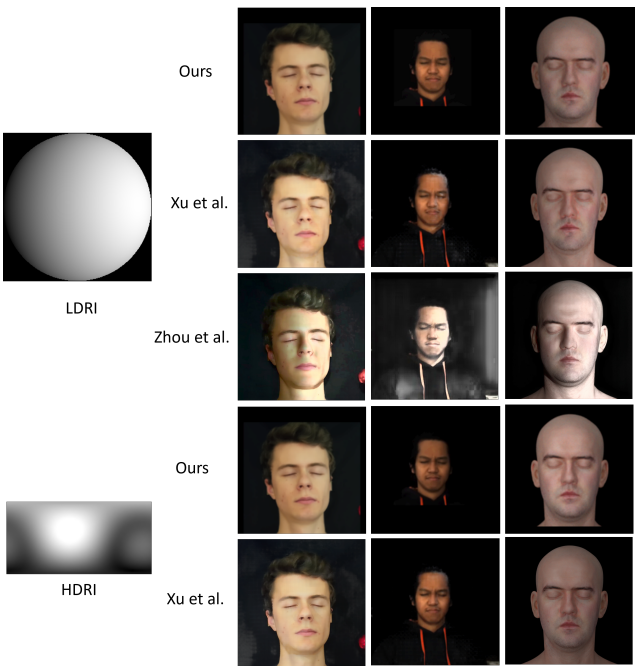

Fig. 16. Top three rows are the relit images using an LDR illumination map for the front hemisphere of the SH based illumination, as in [Zhou et al. 2019]. The bottom two rows are results using an equirectangular HDRI map obtained by converting an HDRI to its SH representation and back to an environment map. Results from Zhou et al. for the HDRI map were discarded since it failed to produce convincing results.

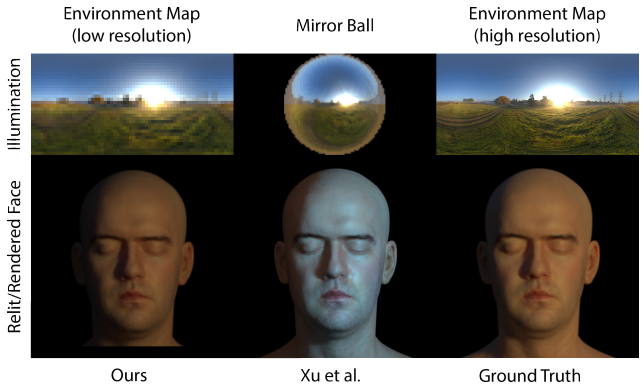

Fig. 17. Top row shows the Illumination maps in the format used by Xu et al. (Mirror ball based light probe) and the equirectangular map used in our method. Bottom row shows relighting results of a synthetic model using Xu et al. and our approach compared to the ground truth.