This WACV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Multi-level Attention Aggregation for Aesthetic Face Relighting

Hemanth Pidaparthy, Abhay Chauhan, Pavan Sudheendra

Samsung R&D Institute Bangalore, India

{hemanth1.p, chauhan.ab, pavan.s} @samsung.com

Abstract

Face relighting is the challenging task of estimating the illumination cast on portrait images by a light source varying in both position and intensity. As shadows are an important aspect of relighting, many prior works focus on estimating accurate shadows using either a shadow mask or face geometry. While these work well, the rendered images do not look aesthetic/photo-realistic. We propose a novel method that combines the features from attention maps at higher resolutions with the lighting information to estimate aesthetic relit images with accurate shadows. We created a new relighting dataset using a synthetic One-Light-At-a-Time (OLAT) lighting rig in Blender software that captures most of the variations encountered in face relighting. Through extensive experimental validation, we show that the performance of our model is better than the current state-of-art face relighting models despite training on a significantly smaller dataset of only synthetic images. We also demonstrate unsupervised domain adaptation from synthetic to real images. We show that our model is able to adapt very well to significantly different out-of-training light source positions.

1. Introduction

Face relighting from a single image is the problem of changing the illumination on the face of a source image based on a given target light direction. It is an active area of computer vision research and has various applications such as photo editing, face recognition [10,15,26] and background lighting transfer [23,24].

The most important considerations in face relighting are preserving facial details and accurate rendering of shadows. The current state-of-the-art methods use the estimated intrinsic components such as albedo and surface normal to appropriately relight the source image [8, 9, 47]. Additional information of shadow masks [9] or face geometry [8] is also used to estimate better shadows. While these two methods estimate accurate shadows, the relit image is not aesthetic/photo-realistic. This could be due to several reasons. Firstly, these models were mainly trained on the DPR dataset [47] which used the ratio image to generate the ground truth relit images. This assumes the human face to be a Lambertian surface and thus, the ground truth data is not photo-realistic. Secondly, the models estimated only the luminance channel and the colour channels were appended from the input image. However, this ignores the changes in appearance (skin tone) caused by the illumination. Thirdly, the cascaded errors from the estimated image intrinsic maps (albedo and surface normal) can lead to artifacts in high frequency details of the relit image.

We propose several modifications to improve upon the prior works. Since there are no existing datasets with accurate & photo-realistic relit images and dense variations in light source positions, we created our own dataset by designing a synthetic OLAT lighting rig in Blender software. This enabled generating accurate and photo-realistic ground truth relit images (Fig 1(c)). To improve the aesthetic quality of the estimated relit images, we propose a novel convolutional transformer-based architecture that uses Multi DConv Head Attention (MDHA) modules at multiple different image resolutions. This enables the network to learn fine-grained facial and shadow details which get lost at lower resolutions. Unlike prior works which model the shadows separately and then render the relit image, our approach enables the network to implicitly learn the relationship between illumination and shadows.

Since we have trained the model on only synthetic dataset, generalization to real images is very challenging as synthetic and real images have very different data distributions. Many prior works fine-tune their models on real images in a second-stage training to address this issue [24, 45]. However, it is very expensive to obtain real input-relit image pairs. Hence, we propose an approach for unsupervised domain adaptation using a Generative Adversarial Network (GAN) framework [11]. Through extensive experimental validation, we show that our results are better than the state-of-the-art (SOTA) methods on multiple different real image datasets despite our model being trained only on synthetic images. Also, our network is able to accurately

adapt to out-of-training light source positions.

In summary, our contributions are:

- New photo-realistic relighting dataset with dense variations in light source positions & intensities, generated using synthetic OLAT lighting rig in Blender software.
- Novel convolutional transformer-based architecture with MDHA attention modules at higher resolutions for learning fine-grained facial and shadow details.
- An approach for unsupervised domain adaptation from synthetic to real images using GANs.

2. Prior Work

There are four distinct approaches explored in literature for face relighting: 1) intrinsic image decomposition and rendering [3–5, 15–18, 20, 23, 27, 28, 31, 37–40, 42, 44], 2) image-to-image translation [2, 21, 35, 36], 3) style transfer [19, 22, 30, 31] and 4) ratio image estimation [25, 29, 34, 43].

Intrinsic image decomposition methods estimate the intrinsic image components such as albedo, surface normal, reflectance and lighting, from a given input image. These intermediary maps are then used to render the relit image for a given light source position. These methods rely heavily on accurate estimates of each intermediary image map. The errors and artifacts are often cascaded downstream and this leads to inaccurate estimates of relit image which lacks high frequency details.

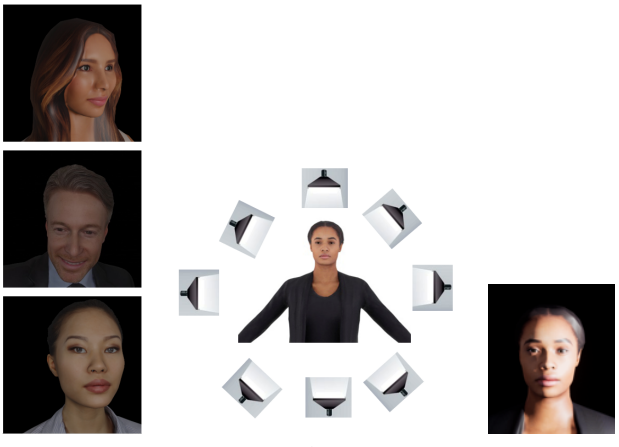
To overcome the issue of cascading errors,others have explored solving for relighting as an image-to-image translation task. These methods are able to estimate relit images to varying degrees of accuracy [35, 47]. However, they suffer from inaccurate shadow estimates and are unable to adapt to varying light intensities. A similar approach is photo and portrait style transfer, where a source and reference images are provided as input and the style/lighting of reference image is transferred to the source image. These methods require high-quality non-occluded source and reference image pairs with multiple different lighting variations. Our approach is different from these approaches in that it can relight a single image for a given light source position. It does not require any reference images and can relight the image for out-of-training light source positions very accurately.

Many others have explored estimating the ratio between source and target image illuminations to learn a per-pixel multiplier map for accurate relighting. However, these methods require require multiple images as input [25, 29] or both source and target images [34], which limits their generalization to real-world deployments. Another significant challenge in face relighting is accurate estimation of shadows. Some have tried to estimate the shadow mask and learn a weighting function for accurate rendering [9], while others have used geometric principles to accurately estimate the shadow pixels [8]. While these approaches are able to estimate shadows with good accuracy, the rendered relit images are not photo-realistic as the estimated shadow regions have hard boundaries. Shadows are diffused and have soft boundaries.

We address the limitations of prior works using a novel architecture for face relighting from a single image. Our approach is most similar to image-to-image translation approaches for face relighting. We train a residual convolutional autoencoder that uses attention maps at higher resolutions to accurately model the shadows. Our network estimates fine-grained facial and shadow details that get lost at lower resolutions. Additionally, we propose a new composition for the relighting dataset and show that with our proposed approach, we are able to train a lightweight model on a significantly smaller dataset and achieve better performance.

3. Dataset

In this section, we introduce our synthetic dataset generated for face relighting. Any dataset used to train a model for face relighting should consist of input-relit image pairs for different light source positions. This can be achieved using a One Light At a Time (OLAT) spherical lighting rig [41], where light sources are positioned at fixed distances from a human who is positioned at the center of the sphere. Each light is turned ON sequentially, with only one light being illuminated at any given time instant. Thus, for each light source position, the input image and the corresponding relit image are captured.



(a) (b) (c)

Figure 1. (a) Sample images obtained from the 3D human models. (b) Synthetic lighting rig in Blender software. (c) Sample relit image generated using Blender software.

Creating such a physical lighting rig is very expensive

and time consuming. Hence, we created a synthetic OLAT lighting rig using Blender software. We used this synthetic dataset to generate input-relit image pairs for face relighting. We used 8 freely available synthetic 3D human models [1] to create the dataset. A few sample images obtained using the 3D human models are shown in Fig 1(a). The synthetic lighting rig was positioned in front of the subject (Fig 1(b)). Light sources were randomly positioned in this unit hemisphere volume and input-relit image pairs were captured. The amount of illumination on the face is dependent on both the position and intensity of the light source. A light source positioned close to the face with low intensity generates a similar looking relit image as that by a light source positioned farther from the human with higher intensity. Thus, in addition to random variations in positions, we also randomly varied the light source intensity. The light vector is represented as a 4D couplet of (x, y, z, i) where the light source positions (x, y, z) are randomly sampled from the unit hemisphere volume such that $x \in [-1, +1]$, $z \in [-1, +1]$, $y \in [0.5, 1]$, and the light source intensity(i) is varied such that $i \in [0.3, 1]$. The X-Z plane is in the front of the 3D human model (face) and Y-direction indicates the frontal distance of the 3D human model from the light source.

Most of the prior works train their models mainly on the DPR dataset [47]. This dataset consists of input-relit image pairs where the ground truth relit images were estimated using the ratio image [29, 47]. This assumes that the face is a Lambertian surface and accounts only diffuse reflection, which results in ground truth relit images are not photo-realistic/aesthetic. Further, this dataset consisted of around 135,000 training examples obtained using 27,000 different images (humans) from the Celeb-FFHQ dataset [12]. For each image, 5 different light source positions were randomly sampled along a unit sphere. We believe that it is easier to reconstruct the input image as opposed to learning the correlation between the input image, light source position & intensity and rendered relit image. Thus, the dataset should consist of significantly more light source positions for a given input image as this would provide the network enough training examples of rendering the same image (human) for different light source positions and intensities. Hence, we inverted the dataset composition and created a new synthetic face relighting dataset of 24,000 input-relit image pairs obtained from 8 different synthetic 3D human models. We generated 3,000 input-relit image pairs using each 3D human model such that the light source positions and intensities were densely sampled in a unit hemisphere volume. We randomly varied the position of the 3D model w.r.t the camera by rotating the 3D human model by r deg about the vertical axis (left-right rotation), where $r \in [-60^o, +60^o]$. We also varied the position of the camera w.r.t the center of the face of 3D human model through a randomized displacement d_x, d_z, where $d_x \in [-0.3, +0.3]$ and $d_z \in [-0.3, +0.3]$ and d_x, d_z are the displacements along X and Z directions respectively (sample input images can be seen in the supplementary material). Recall that the X-Z plane is in front of the 3D human model.

Our dataset has several advantages over the DPR dataset [47]. Since we used a synthetic lighting rig in Blender software, we customized the software to account for both specular and diffuse reflections. Thus, our ground truth data (relit image) is significantly more photo-realistic and has accurate shadows. The dataset consists of densely sampled light source positions in a unit hemisphere volume in front of each 3D human model. Unlike the DPR model, our dataset composition enables the network to better learn the relationship between light source position & intensity and the illumination on the face. Further, the range of light intensity variations is significantly larger in our dataset. In order to enable better generalization of the model, we also varied the position and orientation (through left-right rotation) of the 3D model w.r.t the camera. Thus, our dataset is much more representative of the variations encountered in face relighting as compared to DPR dataset. We will share the dataset upon request after publishing this paper.

4. Method

In this section we describe our proposed two-stage training pipeline for face relighting. In the first stage, we train a residual convolutional autoencoder for face relighting. The light features embedding learnt using the lighting network are combined with the encoded image features learnt from the input data. This is then passed to a decoder which estimates the relit image. To improve the quality of shadows in the estimated relit image we used Multi DConv Head Attention (MDHA) modules [33] for each skip connection between the encoder and decoder. In the second stage, we used adversarial training to improve the perceptual quality of the first stage network. The network from first stage was used as a generator in a Generative Adversarial Network (GAN) [11] framework and a discriminator was trained to distinguish between real (ground truth) and fake (estimated relit) images. Using several augmentation techniques, we achieved unsupervised domain adaptation between synthetic and real images with the proposed two-stage pipeline (Fig 2). Next, we describe each component of our system in more detail.

4.1. Lighting Network

As described in Section 3, the target light position is passed as 4D couplet of (x, y, z, i) where (x, y, z) and i refer to the light source position and intensity, respectively. In a similar manner to [47], we also encode the light source position as a 9-dimensional Spherical Harmonics (SH) vector. The intensity value is appended with the 9-D vector to

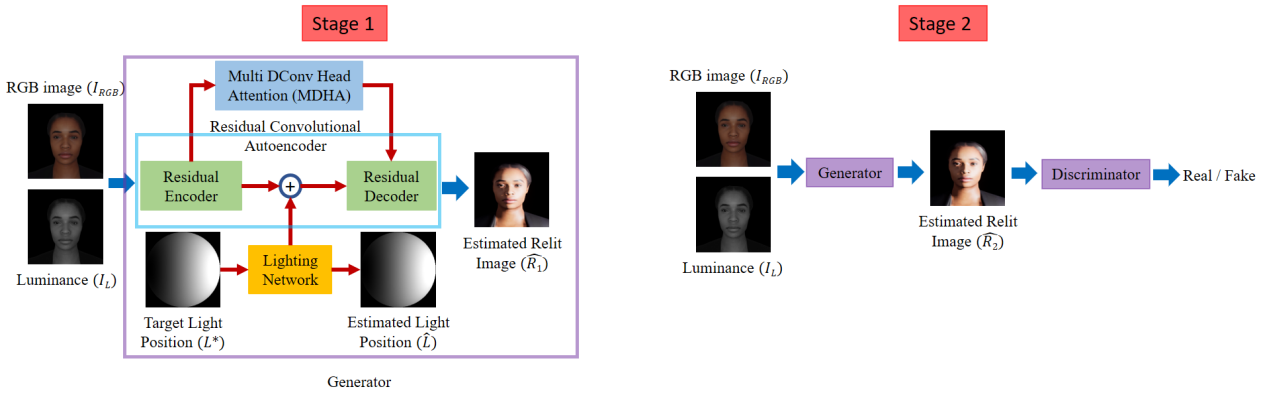


Figure 2. Proposed convolutional transformer-based architecture for face relighting is shown. In the first stage of training, the residual convolutional autoencoder is trained to estimate the relit image. In the second stage of training, the first-stage network is used as a generator in a GAN framework to improve model performance. Figure best viewed in colour.

create a 10-D light vector (L^*) which is given as input to the lighting network that consists of 3 fully connected layers of dimensions 128, 512 and 128 (Fig 3). The output of this network is the estimated 10-D light vector ($\hat{L}$). The output of the middle layer is a 512-D light features embedding that is concatenated with the image features learnt from the input data using the Residual Convolutional Autoencoder.

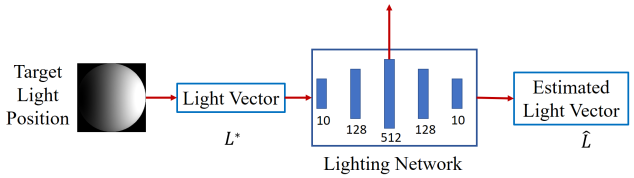


Figure 3. Proposed light network consists of 3 fully connected layers and estimates the 10-D light vector. The 512-D light features embedding are used in the residual convolutional autoencoder.

4.2. Residual Convolutional Autoencoder

The image features are learnt using a residual convolutional autoencoder. The input data (I_D) consists of 4 channels which is obtained by concatenating the RGB image (I_{RGB}) and luminance channel (I_L). This data is passed to a residual encoder as seen in Fig 4. The residual encoder is a modified ResNet-34 architecture. It consists of four ResNet blocks having 16, 32, 64 and 128 channels, respectively. The feature maps are downsampled by a factor of 2 at each block and the output of the encoder is combined with the 512-D light features embedding, as described in Section 4.1. This combined output is then passed through a residual decoder that estimates the relit image. In a similar manner to the residual encoder, the residual decoder also consists of four ResNet blocks having 16, 32, 64 and 128 channels, and the feature maps are upsampled by a factor of 2 at each block.

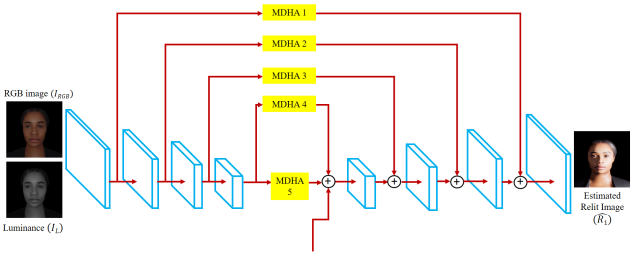


Figure 4. The architecture of the proposed Residual Convolutional Autoencoder is shown. The encoder and decoder consists of four ResNet blocks each. Five MDHA modules (indicated in yellow) are used on each connection between the encoder and decoder. Figure is best viewed in colour.

One of the main challenges in face relighting is rendering the shadows in an accurate and photo-realistic manner. While some methods are able to render the shadows effectively [8, 9], the relit images do not look photo-realistic. This could be because these methods estimate sharp boundaries for the shadows, but shadows are generally diffused in nature and have soft boundaries. Thus, to render a more photo-realistic relit image, it is important to accurately estimate the fine-grained facial and shadow details. As the input image is continually downsampled at each residual block of the encoder, these fine-grained features are lost at lower resolutions. To overcome this issue, we propose a novel idea of using attention blocks at higher resolutions. More specifically, we use Multi DConv Head Attention (MDHA) layers [33] for each level of skip connection between the encoder and decoder blocks, as shown in Fig 4.

Each MDHA layer [33] consists of one attention block and one feedforward block, as seen in Fig 5. In each MDHA layer, the attention block has 8 heads. We designed an MDHA module by stacking four such MDHA layers, as seen in Fig 5(c). From the experimental results in Table 2

(and Fig 3 in supplementary material), we find that the attention modules at higher resolutions enables the network to reconstruct more fine-grained facial and shadow details, and consequently render more photo-realistic relit images.

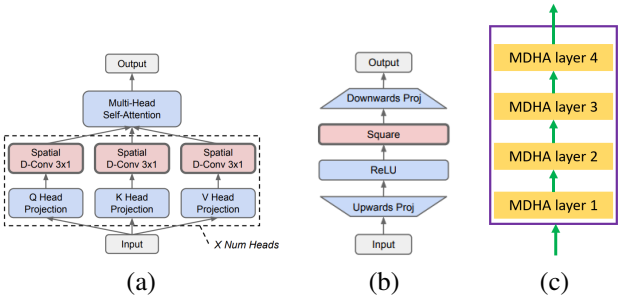


Figure 5. The Multi DConv Head Attention (MDHA) layer proposed in [33] consists of an attention block and a feed-forward block whose architectures are seen in (a) and (b). Our MDHA module is created by stacking 4 MDHA layers as seen in (c).

4.3. Stage 1 training

The proposed two-stage training approach is shown in Fig 2. In the first stage of training, the residual convolutional autoencoder and light network are trained to estimate the relit image ($\hat{R_1}$). In this section, we discuss the loss functions and other training details for stage 1 training.

4.3.1 Training losses

We train the model with multiple different losses to estimate a photo-realistic relit images. We define the light loss as $L_{light} = ||L^* - \hat{L}||_2^2$, where L^* and $\hat{L}$ are the ground truth and estimated 10-D light vector, respectively. We obtain $\hat{L}$ as $\hat{L} = \mathcal{N}_{\mathcal{L}}(L^*)$, where $\mathcal{N}_{\mathcal{L}}$ refers to the lighting network.

We measure the image reconstruction loss at both local and global scales. The image reconstruction loss has two components: 1) Smooth L1 loss (L_{smooth}) and 2) structural dissimilarity (DSSIM) loss (L_{dssim}). We define the smooth L1 loss as

$$L_{smooth}(R^*, \hat{R_1}) = \begin{cases} \frac{1}{2}(R^* - \hat{R_1})^2 & \text{if } |R^* - \hat{R_1}| < 1.0, \\ |R^* - \hat{R_1}| - \frac{1}{2} & \text{otherwise} \end{cases} \quad (1)$$

where R^* and $\hat{R_1}$ are the ground truth and estimated relit image (stage 1), respectively. We define $\hat{R_1} = \mathcal{N}_{\mathcal{R}}(I_D)$, where $\mathcal{N}_{\mathcal{R}}$ refers to the residual convolutional autoencoder network and I_D is the 4-channel input data. The smooth L1 loss is less sensitive to outliers and it enables the network to accurately render the lighting disparities in the relit image. Similar to [8, 9], we define the structural dissimilarity (DSSIM) loss as $L_{dssim}(R^*, \hat{R_1}) = \frac{1-SSIM(R^*, \hat{R_1})}{2}$.

Thus, we define the global image reconstruction loss as

$$L_{global} = \lambda_1 \, L_{smooth}(R^*, \hat{R_1}) + \lambda_2 \, L_{dssim}(R^*, \hat{R_1}) \quad (2)$$

where λ_1 and λ_2 are weights for each loss function.

To ensure that fine-grained facial and shadow details are learnt by the network, we computed the image reconstruction loss at a local scale as well. We divide the image into 128×128 pixel patches, overlapping by 50%. The image reconstruction loss is computed for each patch. Thus, the local image reconstruction loss is defined as

$$L_{local} = \sum_k \lambda_5 \, L_{smooth}(P^*, \hat{P_1}) + \lambda_6 \, L_{dssim}(P^*, \hat{P_1}) \quad (3)$$

where P^* and $\hat{P_1}$ are the image patches from the ground truth and estimated relit images (stage 1), respectively, and k is the total number of patches. λ_3 and λ_4 are the weights for each loss function.

Additionally, we also used VGG loss (L_{vgg}) to further improve the perceptual quality of the estimated relit image. We computed the 4096-dimensional feature vector output of the first fully connected layer of the pre-trained VGG-19 network [32] for both the ground truth and estimated relit images. We define the VGG loss as $L_{vgg} = ||\mathcal{N}_{vgg}(R^*) - \mathcal{N}_{vgg}(\hat{R_1})||_2^2$, where $\mathcal{N}_{vgg}$ refers to the pre-trained VGG-19 network.

Thus, the total loss used for optimizing the face relighting network in stage 1 is

$$L_{total_1} = L_{global} + \lambda_3 \, L_{light} + \lambda_4 \, L_{vgg} + L_{local} \quad (4)$$

where λ_3 and λ_4 are the weights for each loss function.

4.3.2 Training details

We combine a carefully designed data generation process with several data augmentation techniques during training to improve the generalization of the first stage model on real images. As shown in Fig 1(a) (and Fig 1 in the supplementary material), we randomly change the position and orientation of the 3D human model w.r.t to the camera. We also randomly change the ambient lighting present in the scene. Further, we use several different augmentation techniques during training : 1) image flipping; 2) brightness+contrast jitter; 3) colour jitter or luminance jitter. The image is flipped horizontally and the light source position is appropriately updated. The brightness and contrast of the RGB image are tweaked by b and c, where $b \in [-20, +20]$ and $c \in [0.8, 1.2]$. In the colour jitter augmentation, the intensity of each channel of the RGB image is tweaked by p, where $p \in [-20, +20]$. In luminance jitter augmentation, the luminance intensity is tweaked by q, where $q \in [-20, +20]$. Note that b, p and q are integer values. We only apply one of colour jitter or luminance jitter.

These augmentations result in a training dataset that captures multiple different variations of the input images and thus, improves the generalization capabilities of the relighting model to vastly different data distributions.

The relighting model was trained on a dataset of 21,000 images and validated on 3,000 images. The input data to the network consisted of 4-channels - RGB image (I_{RGB}) and luminance channel (I_L). The input data was resized to 512×512 pixels and passed to the residual convolutional autoencoder, which estimated the relit image. The loss was optimized using Adam optimizer [13] with L2 regularization of 0.01. The initial learning rate of 0.0001 was decayed by 0.9 after each epoch and the network was trained for 25 epochs with a batch size of 8. The weights for each loss term is $\lambda_1 = \lambda_3 = \lambda_4 = 1$, $\lambda_2 = \lambda_5 = 10$, $\lambda_6 = 100$.

4.4. Stage 2 training

As described in Section 3, we trained our model on a dataset of only synthetic images. These images have a vastly different data distribution to real image datasets such as Celeb-FFHQ [12] and Multi-pie [6]. Even the images obtained using the 3D models of real humans (We will refer to this as the real human test dataset henceforth.) have a vastly different data distribution to the synthetic images. In the qualitative results shown in Section 5, we observed that the skin tone (colour) of the estimated relit image from stage 1 is different from both the the input and ground truth image. Many prior works have faced similar issues when generalizing from synthetic images to real images [24, 45]. They have fine-tuning the model on real images to overcome the issue. However, this is expensive and requires a physical OLAT lighting rig to capture accurate input-relit image pairs.

We propose a novel solution for unsupervised domain adaptation from synthetic to real images. We used the stage 1 model as the generator in a GAN framework [11] and adversarially trained it against a discriminator to improve the perceptual quality of the stage 1 model. The architecture of the discriminator was same as that of the residual encoder from stage 1. The discriminator tries to distinguish between the estimated relit image (fake) and the ground truth relit image (real), while the generator tries to fool the discriminator.

We used three losses to train the GAN: 1) Smooth L1 loss (L_{smooth}), 2) DSSIM loss (L_{dssim}) and 3) Relativistic adversarial loss (L_{adv}). In a similar manner to stage 1 training, L1 loss and DSSIM loss were computed at both local and global scales. In stage 2 training, the local loss was computed as the relativistic adversarial loss (L_{adv}) [14] on fifty 70×70 pixels image patches. These patches were randomly generated by sampling the full image. We define the loss used for stage 2 training as

$$L_{total_2} = \alpha_1 \, L_{smooth} + \alpha_2 \, L_{dssim} + \alpha_3 \, L_{adv} \quad (5)$$

where $\alpha_1 = 0.5$, $\alpha_2 = 0.1$, $\alpha_3 = 0.01$.

For the second stage training, only the brightness+contrast jitter and luminance jitter data augmentations were applied. The training dataset, input data, optimizer, learning rate and batch size were retained to be the same for both first stage and second stage training. However, in the second stage the initial learning rate was kept constant throughout and L2 regularization was not used. The model was trained for 25 epochs.

5. Results

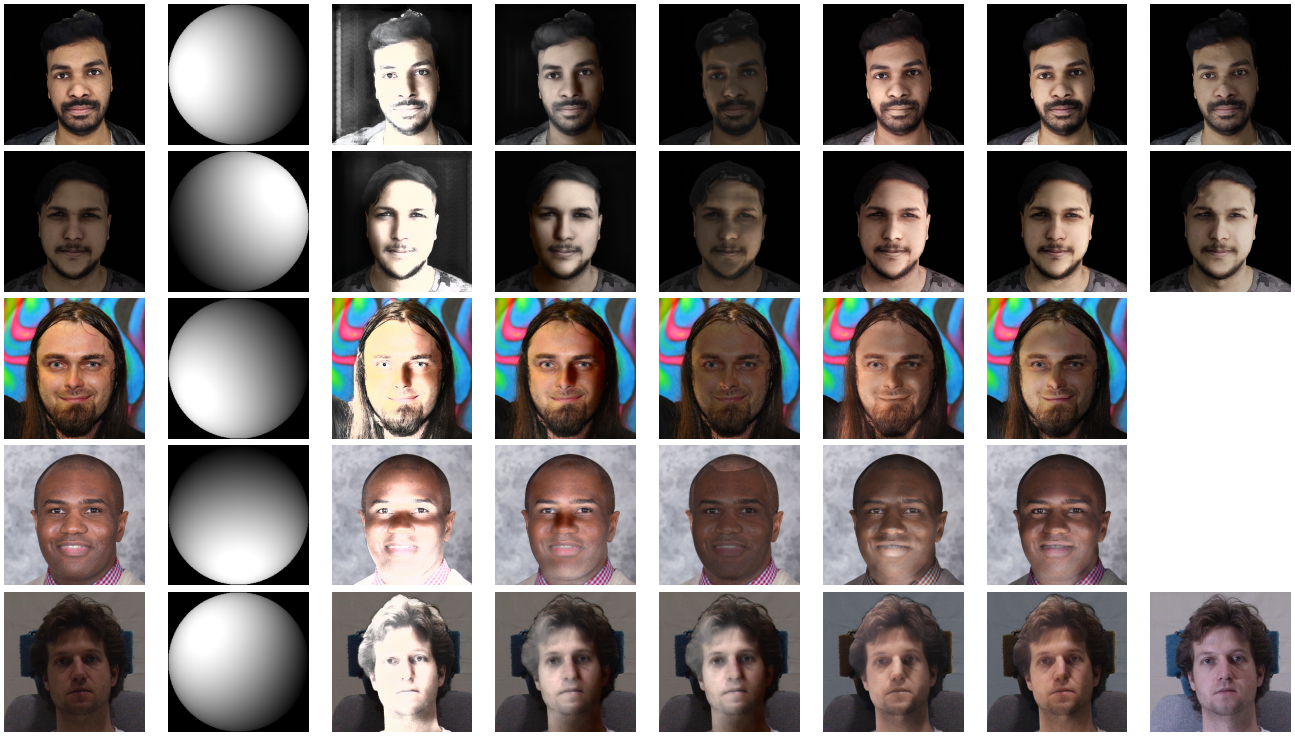
In this section, we discuss the quantitative and qualitative performance of our model. We evaluated the quantitative performance on two datasets of real images: 1) Multi-pie dataset and 2) our real human test dataset. Most prior works evaluate the performance of their model on the publicly available Multi-pie dataset [6]. It consists of input-relit image pairs captured across 4 sessions for multiple different subjects using different camera positions and light source positions. We created a test dataset using the first session data which consisted of 6,474 images - 249 subjects with 2 different expressions and 13 different light source positions.

The Multi-pie dataset had limited variations in the light source positions and no variations in their intensities. Thus, we created our own test dataset using the synthetic OLAT lighting rig in Blender software. We varied the y and z position of the light source such that $y \in \{+0.4, +1.0, +1.5\}$ and $z \in \{-1.5, 0.75, +1.5\}$. Recall that the y-coordinate indicates the distance of the light source from the human and the X-Z plane is the plane in front of the subject where the light source is positioned. For each value of y or z, the light source position was sampled along the unit hemisphere volume. We generated a test dataset (real human dataset) of 432 images - 72 samples each for 6 different real-human 3D models. For each position of y or z, 12 different light source positions were generated, thus totaling 72 input-relit image pairs (samples). The test dataset consisted of out-of-training light source positions along y and z directions.[1]

We compared the performance of our model against three prior works [8, 9, 47] which estimated face relighting given a single input image and (x, y, z) light source position. We evaluated the performance based on four metrics: 1) MSE on RGB image, 2) MSE on Hue channel of HSV image, 3) DSSIM and 4) LPIPS [46]. Both DSSIM and LPIPS have been shown to be highly correlated with the perceptual quality of the images [23, 46]. The quantitative results can be seen in Table 1.

We observed that despite being trained only on synthetic dataset, our model outperforms the prior works on both real image test datasets. Most of the metrics are significantly

[1] Recall that for generating the training dataset, we had $y \in [0.5, 1.0]$ and $z \in [-1, +1]$.



(a) Input image (b) Target lighting (c) Zhou [47] (d) Hou [9] (e) Hou [8] (f) Ours (stage1) (g) Ours (stage2) (h) Ground truth

Figure 6. Qualitative comparison of our model against other methods on real human test dataset (row 1, 2), Celeb-FFHQ dataset (row 3, 4) and Multi-pie dataset (last row). We do not have ground truth relit images on the Celeb-FFHQ dataset. Images are best viewed in colour.

Model	Trained on real images	# parameters (in Millions)	# training examples	Dataset	MSE (RGB)	MSE (Hue)	DSSIM	LPIPS
Zhou *et al.* [47]	✓	6.94	135,000	RH	0.0716 ± 0.0496	0.0326 ± 0.0135	0.2988 ± 0.0450	0.3736 ± 0.0609
Hou *et al.* [9]	✓	6.42	180,000	RH	0.0090 ± 0.0031	0.0231 ± 0.1419	0.1906 ± 0.0172	0.2650 ± 0.0271
Hou *et al.* [8]	✓	12.04	180,000	RH	0.0152 ± 0.0079	**0.0226 ± 0.1306**	0.0787 ± 0.0202	0.1522 ± 0.0181
Ours (Stage 1)	✗	4.45	21,000	RH	0.0057 ± 0.0039	0.0407 ± 0.0191	0.0373 ± 0.0112	0.0863 ± 0.0211
Ours (Stage 2)	✗	4.45	21,000	RH	**0.0049 ± 0.0036**	0.0323 ± 0.0140	**0.0336 ± 0.0100**	**0.0741 ± 0.0189**
Zhou *et al.* [47]	✓	6.94	135,000	MP	0.0845 ± 0.0457	0.2285 ± 0.0548	0.3548 ± 0.0387	0.4389 ± 0.0563
Hou *et al.* [9]	✓	6.42	180,000	MP	0.0125 ± 0.0056	0.2239 ± 0.0534	0.2801 ± 0.0262	0.2538 ± 0.0424
Hou *et al.* [8]	✓	12.04	180,000	MP	0.0118 ± 0.0047	0.2247 ± 0.0537	0.2850 ± 0.0223	0.2607 ± 0.0437
Ours (Stage 1)	✗	4.45	21,000	MP	0.0103 ± 0.0049	0.2391 ± 0.0636	0.0988 ± 0.0137	0.1454 ± 0.0315
Ours (Stage 2)	✗	4.45	21,000	MP	**0.0096 ± 0.0049**	**0.2218 ± 0.0593**	**0.0639 ± 0.0178**	**0.1361 ± 0.03001**

Table 1. Performance comparison of our model against prior works on the real human test dataset (RH) and multi-pie dataset (MP). Note: The metrics of prior works on multi-pie dataset are different to that in [8] because of the differences in size/composition of the test dataset.

Model name	Attention modules	Local Loss (LL)	Global Loss (GL)	MSE	DSSIM	LPIPS
Without Attention	✗	✓	✓	0.0123 ± 0.0044	0.0609 ± 0.0034	0.1204 ± 0.0050
With LLA	A3, A4*	✓	✓	0.0114 ± 0.0031	0.0512 ± 0.0032	0.1015 ± 0.0041
With HLA	A3, A4*	✓	✓	0.0090 ± 0.0034	0.0455 ± 0.0031	0.0934 ± 0.0041
Without GL	A1 A2, A3, A4	✓	✗	0.0107 ± 0.0054	0.0752 ± 0.0058	0.1313 ± 0.0035
Without LL	A1 A2, A3, A4	✗	✓	0.0062 ± 0.0024	0.0564 ± 0.0043	0.1037 ± 0.0030
Full model (stage 1)	A1 A2, A3, A4	✓	✓	**0.0057 ± 0.0039**	**0.0373 ± 0.0112**	**0.0863 ± 0.0211**

Table 2. Ablation study to evaluate the benefit of various network design choices for stage 1 model.

lower than the prior works, indicating that our estimated relit images are more accurate and photo-realistic. Since the real human test dataset consists of out-of-training light source positions, the metrics indicate that our model generalizes significantly better than prior works. The stage 2 model is able to correct the skin tone (colour) issues of the

stage 1 model as seen from the MSE (Hue) metric. The MSE (Hue) for prior works is better than our model because they append the colour channels from the input image. However, from the other metrics we can observe that our model is significantly better.

We also observed that our model outperforms prior works on the challenging Multi-pie dataset (MP). The perceptual quality metrics of DSSIM and LPIPS are significantly better, indicating that the estimated relit images from our model are more photo-realistic.

The qualitative results shown in Fig 6 backs up the findings in Table 1. We observed that despite not being explicitly trained to model shadows, our method produces more aesthetic shadows with soft edges, and perceptually it is most similar to the ground truth images. We also observed that the stage 2 model corrects the the skin tone (colour) issues of the stage 1 model. While the differences in metrics between our stage 1 and stage 2 models might not seem statistically significant, visual inspection shows that stage 2 model produces more accurate & dramatic shadows, and the relit images are more aesthetic.

All the prior works have been trained on the DPR dataset [47] which has mainly been created using the CelebFFHQ [12] dataset. In spite of that, we can observe in Fig 6 (and Fig 2 in the supplementary material) that the results from our model are better than prior works on this dataset. The shadows are more accurate and the estimated relit images are photo-realistic. Hou *et al.* [8] used physics-based analysis to accurately model the shadow regions. While their estimates are fairly accurate, the method is sensitive to precise estimation of the face boundary region. The performance suffers significantly with slightly worse face masks. Our method is not as sensitive since we segment the foreground subject with a noisy segmentation mask obtained using the pre-trained Mask R-CNN [7] model in PyTorch.

Additionally, our model is able to accurately estimate the relit image for out-of-training light source positions as seen in Fig 7. When the light source position moves farther away from the subject, we can observe that the dramatic shadows are reduced and the face is relit more uniformly.

We performed an ablation study to show the benefit of our network design choices for the stage 1 model. We compared the benefit of attention modules at higher resolution (HLA) and the benefit of computing image reconstruction loss at both global (GL) and local (LL) scales. We evaluated the performance of the models on real human test dataset. The quantitative results are shown in Table 2. We observed that higher-level attention (HLA) modules help in estimating more accurate relit images as compared to lower-level attention (LLA) modules. Global loss improves model performance more than local loss, but the best performance was observed with the full stage 1 model. Qualitative results can be found in the supplementary material.

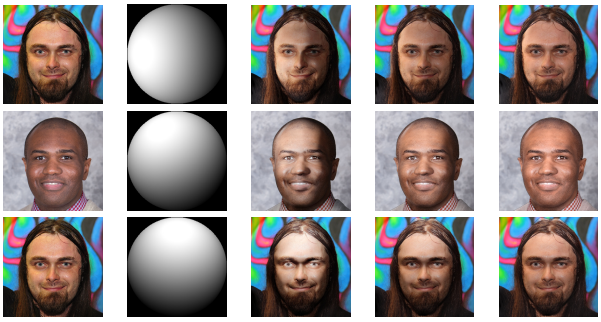


Figure 7. Our model (stage 1) is able to generalize to out-of-training light source positions and capture the variations the relit images accurately. The first two columns show the input image and target lighting direction. First row: Light position is $[x, -1, 0]$ where $x \in \{-1, -2, -3\}$. Second row: Light position is $[-0.5, y, 0.866]$ where $y \in \{-1, -2, -3\}$. Third row: Light position is $[0, -1, z]$ where $z \in \{1, 2, 3\}$.

6. Limitations and future work

We have shown the effectiveness of our approach for accurately modelling shadows and generating photo-realistic relit images. However, there are some limitations with our approach. We observe some issues with skin tone (colour) differences on the multi-pie dataset. One of the reasons could be that the input image has been captured in extremely low-light environment and the light has a blueish hue as seen from the ground truth image (Fig 6). We have not modelled the light colour. Also, a segmentation mask is estimated to obtain the face region in input images.

Some possible future extensions of this work are to automatically understand the face region and relight & image without using a face mask, and joint modelling of ambient & source lighting and light colour to generate more immersive relit images which are similar to studio portraits with lighting.

7. Conclusion

We proposed a novel approach for face relighting given a single image and a light source position. We used a novel dataset composition strategy that enabled better training of our two-stage model for face relighting. In the first stage, a residual convolutional autoencoder and light network were jointly trained. In the second-stage, we improved the perceptual quality of this network with adversarial training and enabled unsupervised domain adaptation from synthetic to real images. We used Multi DConv Head Attention (MDHA) modules at higher resolutions to learn fine-grained facial and shadow details. Qualitative and quantitative analysis showed that our model outperforms SOTA methods on real image datasets despite training only on synthetic images.

References

[1] https://renderpeople.com/free-3d-people/. 3

[2] Yousef Atoum, Mao Ye, Liu Ren, Ying Tai, and Xiaoming Liu. Color-wise attention network for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 506–507, 2020. 2

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 2

[4] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision*, 126:1269–1287, 2018. 2

[5] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T Freeman. Unsupervised training for 3d morphable model regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8377–8386, 2018. 2

[6] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 6

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 8

[8] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 4217–4226, 2022. 1, 2, 4, 5, 6, 7, 8

[9] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 14719–14728, 2021. 1, 2, 4, 5, 6, 7

[10] Yuge Huang, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. Improving face recognition from hard samples via distribution distillation loss. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 138–154. Springer, 2020. 1

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 1, 3, 6

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 3, 6, 8

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[14] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8878–8887, 2019. 6

[15] Ha A Le and Ioannis A Kakadiaris. Illumination-invariant face recognition with deep relit face images. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 2146–2155. IEEE, 2019. 1, 2

[16] Gun-Hee Lee and Seong-Whan Lee. Uncertainty-aware mesh decoder for high fidelity 3d face reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6100–6109, 2020. 2

[17] Jinho Lee, Hanspeter Pfister, Baback Moghaddam, and Raghu Machiraju. Estimation of 3d faces and illumination from single photographs using a bilinear illumination model. In *Rendering techniques*, pages 73–82, 2005. 2

[18] Chen Li, Kun Zhou, and Stephen Lin. Intrinsic face image decomposition with human face priors. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 218–233. Springer, 2014. 2

[19] Yijun Li, Ming-Yu Liu, Xueting Li, Ming-Hsuan Yang, and Jan Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European conference on computer vision (ECCV)*, pages 453–468, 2018. 2

[20] Jiangke Lin, Yi Yuan, Tianjia Shao, and Kun Zhou. Towards high-fidelity 3d face reconstruction from in-the-wild images using graph convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5891–5900, 2020. 2

[21] Feng Liu, Luan Tran, and Xiaoming Liu. Fully understanding generic objects: Modeling, segmentation, and reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7423–7433, 2021. 2

[22] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4990–4998, 2017. 2

[23] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 1, 2, 6

[24] 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, 6

[25] Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. Post-production facial performance relighting using reflectance transfer. *ACM Transactions on Graphics (TOG)*, 26(3):52–es, 2007. 2

[26] Laiyun Qing, Shiguang Shan, and Xilin Chen. Face relighting for face recognition under generic illumination. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2004. 1

[27] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 2

[28] Davoud Shahlaei and Volker Blanz. Realistic inverse lighting from a single 2d image of a face, taken under unknown and complex lighting. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, volume 1, pages 1–8. IEEE, 2015. 2

[29] Amnon Shashua and Tammy Riklin-Raviv. The quotient image: Class-based re-rendering and recognition with varying illuminations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):129–139, 2001. 2, 3

[30] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. 2014. 2

[31] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5541–5550, 2017. 2

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[33] David R So, Wojciech Mańke, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le. Primer: Searching for efficient transformers for language modeling. *arXiv preprint arXiv:2109.08668*, 2021. 3, 4, 5

[34] Arne Stoschek. Image-based re-rendering of faces for continuous pose and illumination directions. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 582–587. IEEE, 2000. 2

[35] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019. 2

[36] Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, Christian Theobalt, et al. Monocular reconstruction of neural face reflectance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4791–4800, 2021. 2

[37] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1274–1283, 2017. 2

[38] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1126–1135, 2019. 2

[39] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[40] Luan Tran and Xiaoming Liu. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):157–171, 2019. 2

[41] Yun-Ta Tsai and Rohit Pandey. Portrait light: Enhancing portrait lighting with machine learning. 2020. https://ai.googleblog.com/2020/12/portrait-light-enhancing-portrait.html. 2

[42] Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1968–1984, 2008. 2

[43] Zhen Wen, Zicheng Liu, and Thomas S Huang. Face relighting with radiance environment maps. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–158. IEEE, 2003. 2

[44] Shugo Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

[45] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6):1–21, 2022. 1, 6

[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[47] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of International Conference on Computer Vision (ICCV)*, pages 7194–7202, 2019. 1, 2, 3, 6, 7, 8