Deep Single-Image Portrait Relighting

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

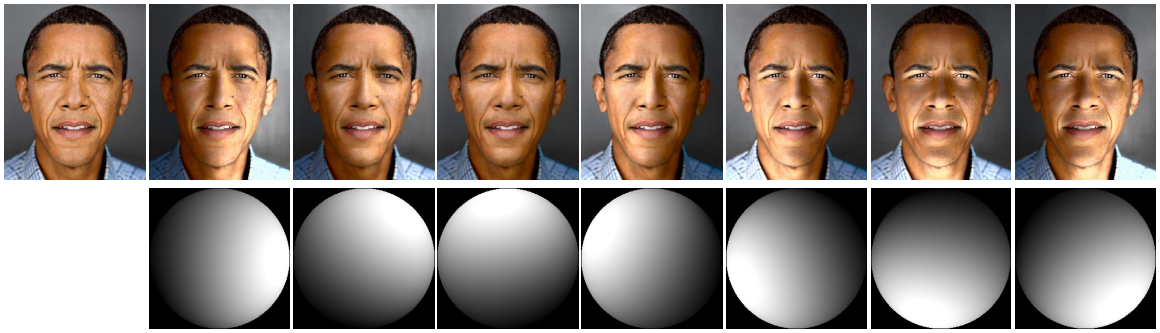


Figure 1: Our algorithm takes a portrait image and a target lighting as input and generates a new portrait image.

Abstract

Conventional physically-based methods for relighting portrait images need to solve an inverse rendering problem, estimating face geometry, reflectance and lighting. However, the inaccurate estimation of face components can cause strong artifacts in relighting, leading to unsatisfactory results. In this work, we apply a physically-based portrait relighting method to generate a large scale, high quality, "in the wild" portrait relighting dataset (DPR). A deep Convolutional Neural Network (CNN) is then trained using this dataset to generate a relit portrait image by using a source image and a target lighting as input. The training procedure regularizes the generated results, removing the artifacts caused by physically-based relighting methods. A GAN loss is further applied to improve the quality of the relit portrait image. Our trained network can relight portrait images with resolutions as high as 1024×1024*. We evaluate the proposed method on the proposed DPR datset, Flickr portrait dataset and Multi-PIE dataset both qualitatively and quantitatively. Our experiments demonstrate that the proposed method achieves state-of-the-art results. Please refer to* `https://zhhoper.github.io/dpr.html` *for dataset and code.*

1. Introduction

The goal of this work is to design an automatic single-image portrait relighting algorithm, which takes a portrait image and a target lighting as input and generates a new portrait image under the target lighting condition. There are physically-based relighting methods that explicitly reconstruct the face geometry, reflectance, and lighting and then re-render this reconstruction using a novel lighting [3, 31, 7, 26, 29, 22]. However, single image face reconstruction is still an open problem, and even the state-of-the-art methods have significant errors, e.g., inaccurate estimation of face geometry and reflectance properties. These errors can propagate into the relighting and lead to poor results. As a result, while these relighting methods are generally good at capturing lighting variations, they may contain artifacts that prevent them from looking realistic. In this work, we leverage this property: we use a physically-based relighting method to generate a large-scale training dataset, and then use it to train a generative network to reproduce them while imposing an adversarial loss based only on real photographs. The supervised reconstruction loss allows the network to learn how to relight, while the adversarial loss ensures that the results are on the manifold of real photographs and do not have the errors from the physically-based relighting method.

We first propose a ratio image-based (RI-based) [23] rendering algorithm to generate a large scale, high resolution, "in the wild" deep portrait relighting dataset (DPR). In this

*Hao Zhou is currently at Amazon AWS.

algorithm, an image under a target lighting condition can be rendered by multiplying the source image with the ratio of the target shading and source shading. Face normals and Spherical Harmonic (SH) lighting of the source image are estimated using 3DDFA [11] and SfSNet [22] respectively. A novel As-Rigid-As-Possible-based (ARAP-based) [27] warping method is then proposed to accurately align the estimated face normal to the portrait image. SH [5, 21] lighting is then randomly sampled from a lighting prior dataset [7] to relight the portrait image. We apply our proposed RI-based algorithm to the high resolution CelebA dataset (CelebA-HQ) [12] and generate 138,135 relit 1024×1024 portrait images with known SH lighting.

An hourglass network [19] is trained using the proposed DPR dataset for the portrait relighting task. It takes a source image and a target lighting as input and generates the relit image. It also predicts the SH lighting for the source image using the features from the bottleneck layer to disentangle lighting information from the source image. We observe that the skip connections in the hourglass network prevent the bottleneck layer from learning meaningful facial information. Therefore, we propose a simple skip training strategy to enforce facial information in the bottleneck layer, which improves the quality of the generated images. Our network is first trained on 512×512 images and then fine tuned on 1024×1024 images. To the best of our knowledge, our proposed method can generate relit images at the highest resolution among all deep learning-based algorithms. We test our method qualitatively on our proposed DPR dataset and the Flickr portrait dataset [24] and quantitatively on the Multi-PIE dataset [9]. All these experiments demonstrate that the proposed method can achieve state-of-the-art results both qualitatively and quantitatively.

To reiterate, the contributions of our work are threefold. First, we propose a ratio image-based algorithm to generate a large scale, high resolution "in the wild" deep portrait relighting dataset. A novel As-Rigid-As-Possible-based warping method is proposed to align the face normals accurately with the face image. Second, we design an automatic single-image portrait relighting algorithm that takes a source image and target SH lighting as input and generates a face image under the target lighting. Third, our trained network can generate 1024×1024 relit portrait images, which, to the best of our knowledge, is the highest resolution among all deep learning-based portrait relighting methods.

2. Related Work

Quotient Images for Portrait Relighting Shashua and Riklin-Raviv [23] proposed to use the quotient (ratio) image for portrait relighting. They require multiple reference images as input and assume all these images are in frontal view. Stoschek extended the ratio image to arbitrary pose by aligning facial landmarks of the source and target image [28]. Wen *et al*. proposed to render a new image using the ratio of the radiance environment map [32]. [20] proposed to apply ratio images to real time portrait illumination editing. However, their method requires capturing images of a static subject using a Light Stage apparatus. Due to the success of ratio images in portrait relighting applications, we apply this technique in our data preparation pipeline.

Inverse Rendering of Portrait Images Starting with the 3D Morphable Model (3DMM) [6], many inverse rendering methods for portrait images have been proposed [31, 3, 7, 26, 30, 8, 22, 29, 33]. These methods decompose a portrait image into reflectance, geometry and lighting. A relit portrait image can then be rendered by changing the lighting and keeping the geometry and reflectance fixed. [31, 3, 7] are optimization-based methods, and are time consuming. [26, 30, 8, 29, 22] are all deep learning-based methods. Compared with optimization-based methods, they are more time efficient. However, due to the complexity of inverse rendering, all these methods can only work on low resolution images. On the contrary, our proposed method, focusing on portrait relighting, can be designed to generate very high resolution (1024×1024) images. Yamaguchi *et al*. [33] recently proposed a deep learning-based method to estimate high resolution face reflectance and normal. However, their method cannot relight the entire face image and leave out the eye, teeth and hair regions

Photo and Portrait Style Transfer Photo and portrait style transfer [18, 16, 24, 25] takes a source image and a reference image as input and transfers the style of the reference image to the source image. Since lighting can be treated as a kind of style, these methods can also be applied in portrait relighting applications. To generate a high quality portrait image, these methods usually require a high quality, non-occluded reference image that contains the desired lighting with a different subject as input, which limits the possible application scenarios. Different from these methods, our proposed method is a single-image-based algorithm, and does not require a reference lighting image, thus making it more general.

3. Deep Portrait Relighting Dataset

In this section, we introduce the Deep Portrait Relighting (DPR) dataset, which is a large scale, high resolution, "in the wild" image dataset generated for portrait relighting purposes. DPR is build on the high resolution CelebA dataset (CelebA-HQ) published by [12], which contains 30,000 face images from the CelebA [17] dataset with 1024×1024 resolution. We remove images on which the landmark detector [14] fails to detect landmarks, resulting in 27,627 images in the DPR dataset. For each of these images, we randomly select 5 lighting conditions from a lighting prior dataset [7] to generate relit face images, leading to

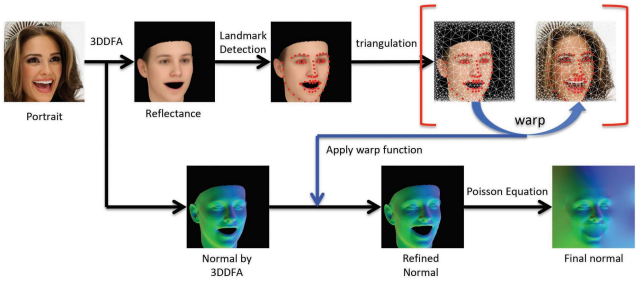
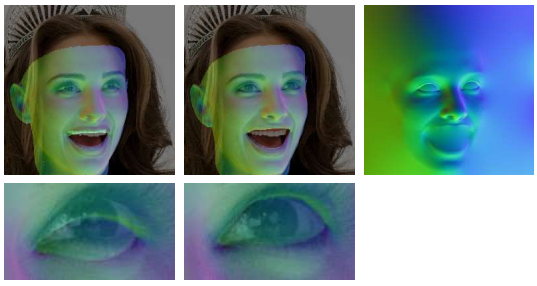


Figure 2: ARAP based normal refinement.



(a) (b) (c)

Figure 3: We show the original face overlaid with normal estimated by 3DDFA [11] in (a) and our refined normal in (b). The second row of (a) and (b) show the right eye region. (c) shows the final normal map.

138,135 relit images.

3.1. Ratio Image-based Face Relighting

We proposed a RI-based algorithm for data generation. To render a face image $\mathbf{I}$, we need the reflectance $\mathbf{R}$, normal $\mathbf{N}$ and lighting $\mathbf{L}$. We further assume that the reflectance of human face is Lambertian. A face image $\mathbf{I}$ can thus be represented as:

$$\mathbf{I} = \mathbf{R} \odot f(\mathbf{N}, \mathbf{L}), \tag{1}$$

where $\odot$ represents the element-wise product and f is the Lambertian shading function. To relight a face image, we apply the ratio image trick proposed in [23]. According to Eq 1, the same face under two different lighting conditions $\mathbf{L}$ and $\mathbf{L}^*$ can be represented as $\mathbf{I} = \mathbf{R} \odot f(\mathbf{N}, \mathbf{L})$ and $\mathbf{I}^* = \mathbf{R} \odot f(\mathbf{N}, \mathbf{L}^*)$. We know that

$$\begin{aligned} \mathbf{I}^* &= \mathbf{R} \odot f(\mathbf{N}, \mathbf{L}^*) \\ &= \frac{\mathbf{R} \odot f(\mathbf{N}, \mathbf{L}^*)}{\mathbf{R} \odot f(\mathbf{N}, \mathbf{L})} (\mathbf{R} \odot f(\mathbf{N}, \mathbf{L})) \\ &= \frac{f(\mathbf{N}, \mathbf{L}^*)}{f(\mathbf{N}, \mathbf{L})} \mathbf{I}. \end{aligned} \tag{2}$$

As a result, a portrait image $\mathbf{I}^*$ under lighting $\mathbf{L}^*$ can be generated given portrait image $\mathbf{I}$ and its normal and lighting.

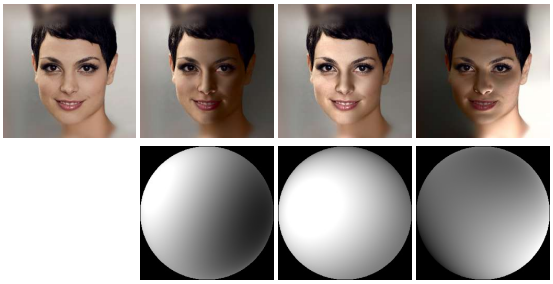


Figure 4: First column is the original image, second to fourth columns in the first row are relit images generated by our rendering pipeline, the second row shows the half sphere rendered using the corresponding SH lighting.

3.2. Normal Estimation

There are many research studies targeted at estimating normals from portrait images. We use 3DDFA [34] (Code provided by [11]) since it outputs the shape parameters of a 3DMM, which can be used to generate portrait normal images at arbitrary resolution. Although 3DDFA takes facial expression into consideration while fitting 3DMM, the normals estimated still cannot be acccurately aligned with the the portrait image. We believe this is due to the limited power of the 3DMM to model variations of face geometry, as 3DMM is built on a limited number of faces. To avoid artifacts in the relit images, we propose aligning the estimated normals with the portrait image using an ARAP-based normal refinement algorithm.

3.2.1 ARAP-Based Normal Refinement

Figure 2 illustrates the procedure of the ARAP based normal refinement algorithm. Using the 3DMM parameteres predicted by 3DDFA [11], a mesh can be created. The "reflectance" image of the portrait can be obtained by projecting the generic reflectance map of the 3DMM model onto this mesh. We then apply [14] to detect 68 facial landmarks on this "reflectance" image. These 68 detected facial landmarks, together with evenly sampled 198 points along the boundaries of the image are combined as "anchor points" and are used to create a triangle mesh on the "reflectance" image using Delauny Triangulation. Similarly, a triangle mesh is created for the portrait image. An As-Rigid-As-Possible transformation [27] (ARAP) is then applied to warp the triangle mesh of the "reflectance" image to the portrait image. The estimated warp function by ARAP is then applied to the face normals estimated by 3DDFA to get refined normals as illustrated in Figure 2. To demonstrate the effectiveness of the proposed normal refinement method, we overlay the normals estimated by 3DDFA [11] and our refined normals with the original image, and show them in Figure 3 (a) and (b) respectively. It is clear that the

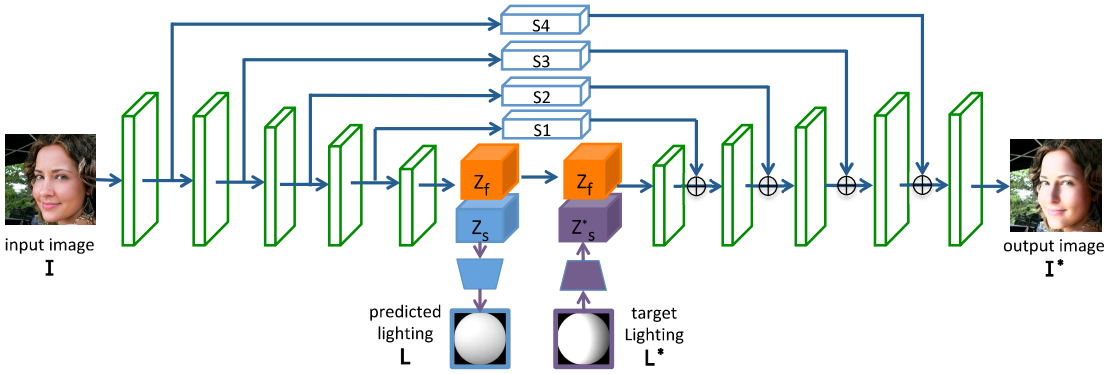


Figure 5: The structure of the proposed Hourglass network.

quality of the alignment of normals w.r.t. the portrait image at the eye and mouth has been improved significantly through our proposed normal refinement method.

We notice that our proposed ARAP normal refinement method cannot improve the misalignment of the ear and neck regions. This is because 3DMM cannot model the deformation of ear and neck well, and, to the best of our knowledge, there is no landmark detection algorithm for ears and necks. As a result, we remove the ear and neck regions from the refined normals to avoid possible artifacts in relit images. In order to get a full normal image, we solve a Poisson equation to fill in the missing normals for ear, neck, mouth and background region as suggested by [26]. Figure 3 (c) shows the normals after filling the missing region.

3.3. Relighting Images

For a portrait image $\mathbf{I}^*$, we apply our method to estimate normals $\mathbf{N}$ and use SfSNet [22] to estimate SH lighting $\mathbf{L}^*$. Then a target SH lighting $\mathbf{L}$ is randomly sampled from the lighting prior dataset [7]. Eq 3 is then used to generate the relit face image $\mathbf{I}$. Due to the ambiguity of the color between lighting and reflectance, we apply the rendering pipeline to the luminance channel and keep the color of the portrait image unchanged. We show one example of a relit face image in Figure 4.

4. Method

In this section, we introduce our proposed deep learning based single-image portrait relighting algorithm. We design an hourglass network for this task and use the DPR dataset created in the section 3 to train the network.

4.1. Main Architecture for Portrait Relighting

Figure 5 shows the structure of our proposed hourglass network [19]. It has an encoder and a decoder part. Four skip connections are used to connect the features at different scales in the encoder part to their corresponding scale in the decoder part. To relight a face, our network takes a face image $\mathbf{I}$ and a target lighting $\mathbf{L}^*$ as input. The encoder extracts features $\mathbf{Z}$ which are divided into two parts: face feature $\mathbf{Z_f}$ which is independent of lighting; and lighting feature $\mathbf{Z_s}$. $\mathbf{Z_s}$ is then fed into a lighting regression network to predict the lighting $\mathbf{L}$ of the input face image $\mathbf{I}$. The target lighting $\mathbf{L}^*$ is then mapped to the lighting feature $\mathbf{Z_s^*}$. $\mathbf{Z_f}$ and $\mathbf{Z_s^*}$ are concatenated together and fed into the decoder part to generate the relit face image. Please refer to the supplementary for more details of our network architecture.

4.2. Supervision for Training the Network

As discussed in section 3, our data preparation process generated five relit images with known ground truth lighting for each image in CelebA-HQ dataset. To generate one training data, we randomly select one source image $\mathbf{I}_s$ and one target image $\mathbf{I}_t$ and their corresponding ground truth SH lighting $\mathbf{L}_s$ and $\mathbf{L}_t$ from these five relit images and the original image from CelebA-HQ. Our network then takes source image $\mathbf{I}_s$ and target lighting $\mathbf{L}_t$ as input and generates $\mathbf{L}_s^*$ and $\mathbf{I}_t^*$. $\mathbf{L}_s$ and $\mathbf{I}_t$ are used as ground truth to supervise the training. We apply L_1 loss for generated portrait image $\mathbf{I}_t^*$ and an L_2 loss for the predicted lighting $\mathbf{L}_s^*$. An L_1 loss is further applied to the gradient of $\mathbf{I}_t^*$ to preserve edges and avoid blurring:

$$\mathcal{L}_I = \frac{1}{N_I}\left(||\mathbf{I}_t - \mathbf{I}_t^*||_1 + ||\nabla\mathbf{I}_t - \nabla\mathbf{I}_t^*||_1\right) + (\mathbf{L}_s - \mathbf{L}_s^*)^2, \tag{3}$$

where N_I is the number of pixels in the image.

Since our "ground truth" images are generated using the ratio image trick, they may contain artifacts due to inaccurate estimation of face normal or lighting. We thus propose to use a GAN loss to improve the quality of the generated images. As these artifacts mostly appear locally, we use a patch GAN [10] to force the distribution of local image patches to be close to that of a natural image. We use LSGAN [2] for our GAN loss:

$$\mathcal{L}_{GAN} = \mathbb{E}_{\mathbf{I}}(1 - D(\mathbf{I}))^2 + \mathbb{E}_{\mathbf{I_s}} D(G(\mathbf{I}_s, \mathbf{L}_t))^2, \tag{4}$$

Figure 6: From left to right: output without skip layer S4, output without S4/S3, output without S4/S3/S2, output without S4/S3/S2/S1. Top row: vanilla Hourglass network, bottom row: Hourglass network with skip training.

where $\mathbf{I}$ is the real image, G and D represent our relighting network and discriminator respectively. We use 1 as a label for real images and 0 as a label for fake images. While training, we use the images from the FFHQ dataset [13] as real images in our GAN loss since images in this dataset contain more lighting variations.

A feature matching loss is further proposed to increase the accuracy of the relit portrait image. More specifically, images of the same person under different lighting conditions should have the same face features. We thus define a feature loss as:

$$\mathcal{L}_F = \frac{1}{N_F}(\mathbf{Z_{f1}} - \mathbf{Z_{f2}})^2, \tag{5}$$

where $\mathbf{Z_{f1}}$ and $\mathbf{Z_{f2}}$ are face features of two input face images $\mathbf{I}_{s1}$ and $\mathbf{I}_{s2}$, and N_F is the number of elements in $\mathbf{Z}_f$.

4.3. Skip Training

When the Hourglass network is trained end-to-end (denoted as vanilla Hourglass), we notice that most of the facial information is passed through skip layers. Our facial feature $\mathbf{Z_f}$, on the other hand, contains little facial information. We thus propose a skip training strategy in which we train our network without skip connections first, then add skip layers one by one during subsequent training. We denote this as skip training. Figure 6 compares the relit images generated by removing the skip layers of vanilla Hourglass network and Hourglass network with skip training. We can see that with the skip training strategy, more facial information is kept in the feature layer. Figure 7 further demonstrates that skip training can help improve the quality of the generated results by removing artifacts around the nose. In the following discussion, unless otherwise specified, our network is trained with skip training.

4.4. Implementation Details

The overall loss for our network is a linear combination of the losses mentioned in Sec. 4.2:

$$\mathcal{L} = \mathcal{L}_I + \mathcal{L}_{GAN} + \lambda\mathcal{L}_F, \tag{6}$$

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

Figure 7: (a) output of vanilla Hourglass network, (b) rectangle region of (a), (c) output of Hourglass network with skip training, (d) rectangle region of (c). We increase the pixel intensity of (b) and (d) to better visualize.

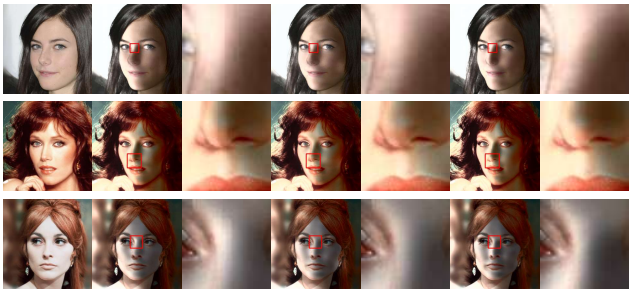
where $\lambda = 0.5$. Our network is trained for fourteen epochs. We add our feature loss $\mathcal{L}_F$ after ten epochs. For skip training, we train our network without any skip connections for five epochs, and add skip connections one at each epoch thereafter, until all skip layers are added. We first train our network with images of resolution 512×512; most of our experiments are carried out under this resolution. Finally, we fine tune our trained network using images with resolution 1024×1024 with a simple modification. More specifically, an additional downsample and upsample layer is added to encoder and decoder respectively to make our network compatible with 1024×1024 images. We train our network using the Adam optimizer [15] with default parameters. Please refer to the supplementary materials for more details of our implementation.

5. Experiments

In this section, we evaluate our proposed method both quantitatively and qualitatively and compare it with previous the state-of-the-art methods. Since our network can predict lighting, it can be used in two ways for portrait relighting: (A) Given a source image $\mathbf{I}_s$ and a SH lighting $\mathbf{L}_t$, generating an image $\mathbf{I}_t$ (denoted as the **SH-based relighting**). (B) Given a source image $\mathbf{I}_s$ and a reference image $\mathbf{I}_f$, extracting SH lighting $\mathbf{L}_t$ from $\mathbf{I}_f$ and using it to relight $\mathbf{I}_s$ to get $\mathbf{I}_t$ (denoted as the **image-based relighting**). When the target SH lighting $\mathbf{I}_t$ is known (e.g. our DPR dataset) we use (A) for our relighting task. For datasets such as Multi-PIE [9], in which ground truth SH lighting is unknown, we use (B) for relighting.

5.1. Dataset and Evaluation Metric

Dataset: We demonstrate the effectiveness of the proposed method on the test set of our proposed DPR dataset. However, due to lack of real ground truth, we cannot evaluate the accuracy of the relit images using this dataset. We thus propose to use the Multi-PIE dataset [9] for quantitative evaluation. The Multi-PIE dataset contains images of the same person under different lighting conditions, which can be used as source and target image pair. Each Multi-PIE image is lit by a dominant point light source, while the lighting



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

Figure 8: (a) shows the input image, (b), (d) and (f) are images generated using $\mathcal{L}_I$, $\mathcal{L}_I$ + $\mathcal{L}_{GAN}$ and $\mathcal{L}_I$ + $\mathcal{L}_{GAN}$ + $\mathcal{L}_f$ respectively; (c), (e) and (g) are the red rectangle region of (b), (d) and (f) respectively. Note the edge in the middle of the noise generated using $\mathcal{L}_I$.

conditions of most "in the wild" portrait images are diffuse. We thus generate images under 7 lighting conditions by averaging 3 to 4 original face images from Multi-PIE, so as to generate images under more realistic, diffuse lighting conditions. We created 440 groups of images from our generated face images, each of which contains a source image $\mathbf{I}_s$, a target image $\mathbf{I}_t$ and a reference image $\mathbf{I}_f$. $\mathbf{I}_s$ and $\mathbf{I}_t$ are images with the same identity but with different lighting conditions, $\mathbf{I}_t$ and $\mathbf{I}_r$ are images of different identities but with the same lighting condition. When evaluating, a relighting algorithm takes $\mathbf{I}_s$ and $\mathbf{I}_f$ as input and predicts $\mathbf{I}_t$.

Evaluation metric: Since lighting is ambiguous up to a scale (e.g., longer exposure time may lead to a SH with high energy under the same lighting conditions), we proposed to use a scale invariant Mean Squared Error (Si-MSE) [4] to evaluate the error between the generated image $\mathbf{I}_t^*$ and the ground truth image $\mathbf{I}_t$.

$$\text{Si-MSE} = \frac{1}{N_I} \min_{\alpha} (\mathbf{I}_t - \alpha * \mathbf{I}_t^*)^2, \tag{7}$$

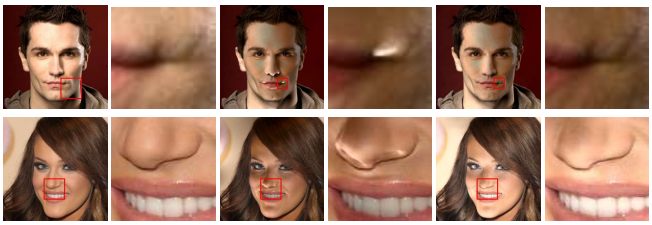
where α is a scalar and N_I is the number of pixels in the image. To further check whether the generated image portrays the target lighting, we run SfSNet [22] to extract the lighting $\mathbf{L}_t$ and $\mathbf{L}_t^*$ from $\mathbf{I}_t$ and $\mathbf{I}_t^*$ respectively, and compute the scale invariant L_2 (Si-L_2) distance between $\mathbf{L}_t$ and $\mathbf{L}_t^*$. We choose to use SfSNet [22] since it is proven to work well at predicting consistent lighting for face images under the same lighting condition.

5.2. Ablation Study

To demonstrate the effectiveness of the GAN loss and feature loss, we show the quantitative and qualitative results of our network trained using $\mathcal{L}_I$, $\mathcal{L}_I + \mathcal{I}_{GAN}$ and $\mathcal{L}_I$ + $\mathcal{L}_{GAN}$ + $\mathcal{L}_f$ (i.e. full model) in Table 1 and Figure 8. We notice that with GAN loss, the accuracy of our trained network is worse than the network trained without GAN loss. This is because the GAN loss is used to make the distribution of the generated images closer to that of the real images, i.e. improve the visual quality of the generated images. Adding the GAN loss may distract the network training process from being closer to the "ground truth" images. However, Figure 8 shows that with GAN loss, the artifacts on the nose part are alleviated compared with the network trained without GAN loss. This demonstrates the effectiveness of the GAN loss in improving the visual quality. Adding a feature loss $\mathcal{L}_f$ significantly improves the accuracy of the images generated by our model, as shown in Table 1. We believe this is because our feature loss can force the generated images of the same identity to have similar latent features, thus, better preserving the identity information in the generated images. Moreover, Figure 8 shows that feature loss does not affect the quality of the generated images. As a result, we conclude that our full model can achieve a good balance between the accuracy and quality of the generated images.

Table 1: Ablation Study on Multi-PIE Dataset

	Si-MSE	Si-L_2
$\mathcal{L}_I$	**0.00504**	**0.1307**
$\mathcal{L}_I$ + $\mathcal{L}_{GAN}$	0.00658	0.1686
$\mathcal{L}_I$ + $\mathcal{L}_{GAN}$ + $\mathcal{L}_f$	0.00590	0.1444



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

Figure 9: (a) original image, (c) results of RI-based rendering, (d) our results. (b), (d) and (f) show the red rectangle region of (a), (c) and (d) respectively. Note that the proposed method removes the ghost effect and artificial highlights.

5.3. Comparison with the Rendering Pipeline

Our proposed ARAP-based normal refinement method improves the misalignment of face geometry as discussed in Section 3. However, there are still cases in which the face normal do not perfectly align with the face image, especially in the nose and the mouth region. These misalignments can cause ghosting on the nose and artificial highlights at the corner of the mouth as shown in Figure 9. Though our training data contains images with these artifacts, Figure 9 shows that these artifacts can be avoided by the proposed method. This is because a deep learning-based method can regularize the results, avoiding outlier effects.

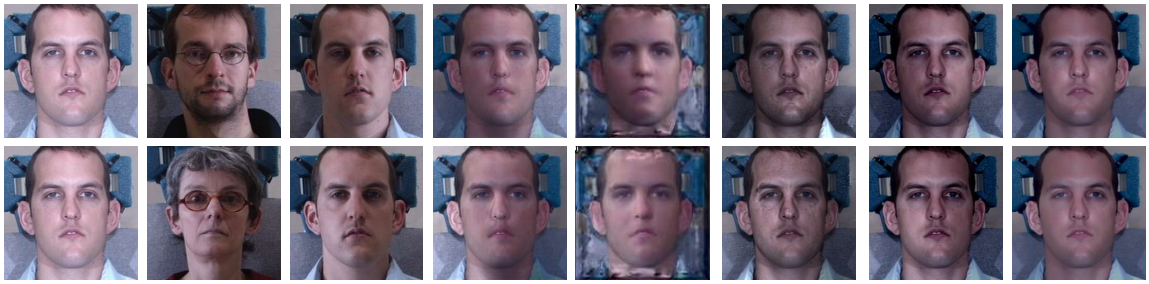


Figure 10: Visual results of the proposed method and state-of-the-art methods on Multi-PIE.

Figure 11: Qualitative comparison of the proposed method with state-of-the-art methods.

Table 2: Evaluation Multi-PIE Dataset

	Si-MSE	Si-L_2
Li *et al.* [16]	0.01322	0.3939
Shih *et al.* [24]	0.01513	0.3415
Shu *et al.* [25]	0.01384	0.3908
sfsNet [22]	0.00659	0.1593
Proposed Method	**0.00590**	**0.1444**

5.4. Comparison with State-of-the-art Methods

In this section, we compare our method with [22, 25, 24, 16], which can do portrait relighting. Since there is no ground truth lighting for images in the Multi-PIE dataset, we use an **image-based method** to evaluate our proposed method and SfSNet [22] on this dataset, i.e., target lighting is extracted from the reference image and used for relighting. Both the proposed method and SfSNet [22] use their own lighting estimation method to extract the target lighting. [25] and [24] are two state-of-the-art portrait style transfer methods. They take two images $\mathbf{I}_s$ and $\mathbf{I}_f$ as input and transfer the style of $\mathbf{I}_s$ to $\mathbf{I}_f$. To get the relit image using these two methods, we transfer $\mathbf{I}_s$ and $\mathbf{I}_f$ from *RGB* image to *Lab* image, and only apply their algorithm on the L channel. [16] is designed for general photo style transfer, similarly we use the L channel for portrait relighting.

Though visual results are the best way to compare these methods, we propose to evaluate them quantitatively in order to understand whether the relit images reflect the reference lighting conditions accurately. Table 2 shows that our proposed method achieves state-of-the-art results on both Si-MSE and Si-L_2 metric. This demonstrates that the proposed method can accurately generate relit images under the target lighting condition. [1] We show some examples of relit faces in the Multi-PIE dataset in figure 10. We notice that results of the proposed method on the Multi-PIE dataset are blurry, however, this is not the case for "in the wild" images as shown in Figure 11 and 12. This is probably due to the domain gap between the Multi-PIE dataset and our DPR dataset used for training.

We visually compare the proposed method with these state-of-the-art methods on DPR dataset and show results in Figure 11. Since the target lighting is known in this dataset, we apply an **SH-based method** to evaluate the proposed method and SfSNet [22]. **Comparison with SfSNet[22]:** We see that although SfSNet [22] can generate images under the correct lighting conditions, their results are of low quality. Also, SfSNet [22] works on 128×128 images, which is too small for portrait relighting applications. Furthermore, SfSNet cannot deal with the background correctly, making the results visually unpleasant. **Comparison with Shih *et al.* [24], Shu [25] and Li *et al.* [16]:** [24], [25] and [16] do not generate images under the correct lighting

[1] Note that the built in facial landmark detector [1] of [25] fail to detect landmarks of 90 testing face images which are excluded when computing the Si-MSE and Si-L_2 for [25] in Table 2.

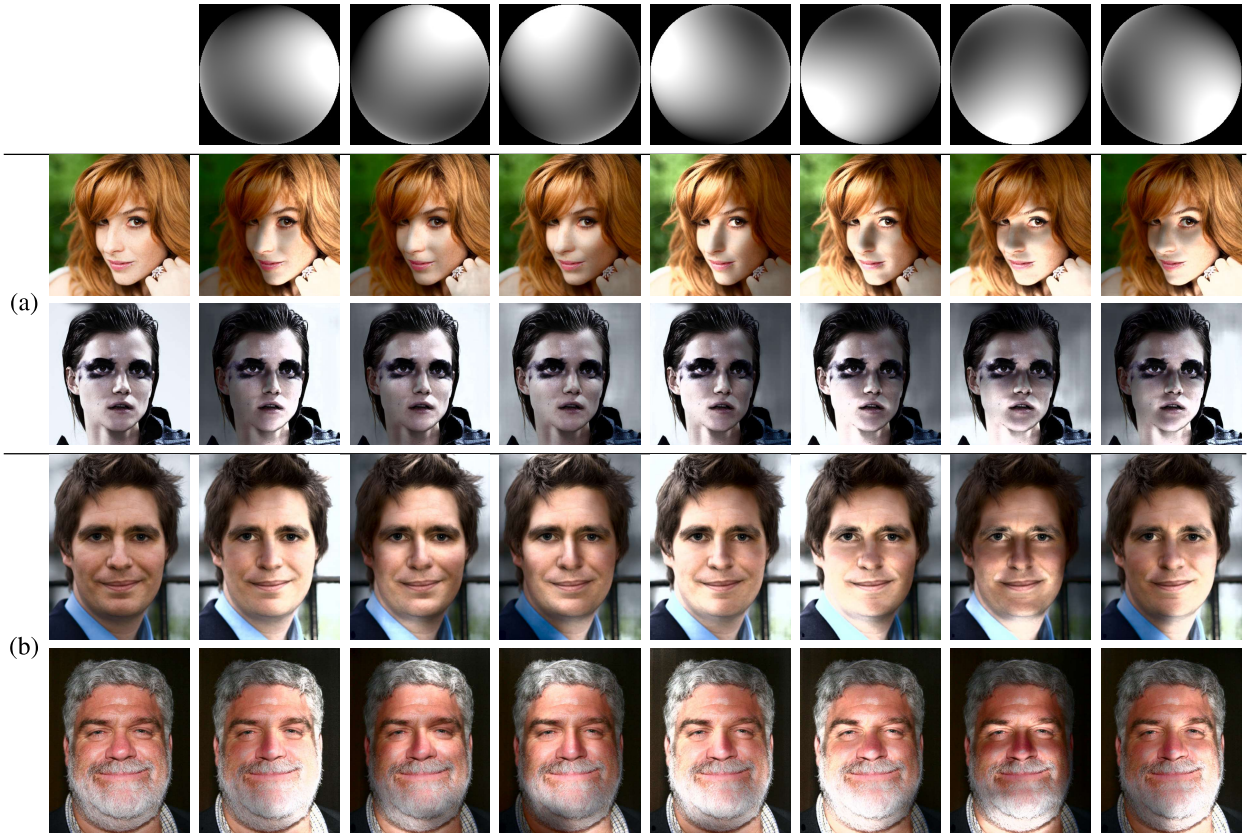


Figure 12: (a) show results on non-frontal face images. (b) show results on the Flickr portrait dataset [24].

in these examples. These three methods are all reference image based; when the reference image is of low quality (e.g. occluded by hair region or sunglasses), they fail to understand the lighting correctly and do not reproduce the reference lighting accurately.

We believe this is a common drawback for all methods which require a reference image as input. On the contrary, the proposed method and SfSNet [22] can directly take target lighting as input and no reference image is required. [24] and [25] both require accurate landmark detection. The built in landmark detector [1] for [25] fails to detect facial landmarks for some non-frontal faces, which limits the possible applications of [25]. Moreover, we notice that [24] and [25] cannot generate attached shadows on the nose, whereas, the proposed method can generate very natural attached shadows.

5.5. Results on Challenge and Flickr Images

Figure 12 (a) shows that our proposed method works well on non-frontal faces and faces with makeup. We further fine tune our network on 1024×1024 images and test the trained model on the Flickr portrait dataset [24]. Figure 12 (b) shows some of the results. Please refer to the supplementary material for more results.

6. Conclusion

In this work, we have proposed an automatic single-image portrait relighting algorithm. A physically-based portrait relighting method is first proposed to generate a large scale, high quality, "in the wild" deep portrait relighting dataset. An hourglass network is then trained using this dataset to generate a relit portrait image by taking a source portrait image and a target lighting as input. We show that our training procedure, that combines reconstruction and adversarial losses with a novel skip connection training strategy, can regularize the generated results, removing the artifacts caused by physically-based rendering. Our network can generate images with resolution as high as 1024×1024 and achieves state-of-the-art results.

7. Acknowledgement

We gratefully acknowledge Hong Wei for helping us run experiments on [16]. Part of this work was done when Hao Zhou was an intern in Adobe Research. This work is supported by DARPA MediFor program under cooperative agreement FA87501620191, Physical and Semantic Integrity Measures for Media Forensics.

References

[1] facetracker. http://facetracker.net/.

[2] Image-to-image translation in pytorch. https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix. Accessed:2019.

[3] Oswald Aldrian and William A.P. Smith. Inverse rendering of faces with a 3d morphable model. *IEEE Transactions on PAMI*, 35(5), 2013.

[4] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *TPAMI*, 2015.

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

[6] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH*, 1999.

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

[8] Kyle Genova, Forrester Cole, Aaron Maschinot, Aaron Sarna, Daniel Vlasic, and William T. Freeman. Unsupervised training for 3d morphable model regression. *CVPR*, 2018.

[9] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image Vision Computing*, 28(5), 2010.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

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

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *CoRR*, abs/1812.04948.

[14] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. 2014.

[15] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

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

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[18] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, 2017.

[19] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016.

[20] Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. Post-production facial performance relighting using reflectance transfer. In *SIGGRAPH*, 2007.

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

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

[23] Amnon Shashua and Tammy Riklin-Raviv. The quotient image: Class-based re-rendering and recognition with varying illuminations. *IEEE Trans. on PAMI*, 23:129–139, 2001.

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

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

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

[27] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Proceedings of Eurographics Symposium on Geometry Processing*, 2007.

[28] Arne Stoschek. Image-based re-rendering of faces for continuous pose and illumination directions. In *CVPR*, 2000.

[29] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *ICCV*, 2017.

[30] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *CVPR*, 2018.

[31] 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 Trans. PAMI*, 31(11), nov 2009.

[32] Zhen Wen, Zicheng Liu, and T. S. Huang. Face relighting with radiance environment maps. In *CVPR*, 2003.

[33] Shuco 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*, 37(4), 2018.

[34] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. Face alignment in full pose range: A 3d total solution. *PAMI*, 2017.