Relighting Images in the Wild with a Self-Supervised Siamese Auto-Encoder

Yang Liu
University of Surrey, UK
yang.liu@surrey.ac.uk

Alexandros Neophytou, Sunando Sengupta, Eric Sommerlade
Microsoft Corporation, Reading, UK
Alexandros.Neophytou, Sunando.Sengupta,
Eric.Sommerlade@microsoft.com

Abstract

We propose a self-supervised method for image relighting of single view images in the wild. The method is based on an auto-encoder which deconstructs an image into two separate encodings, relating to the scene illumination and content, respectively. In order to disentangle this embedding information without supervision, we exploit the assumption that some augmentation operations do not affect the image content and only affect the direction of the light. A novel loss function, called spherical harmonic loss, is introduced that forces the illumination embedding to convert to a spherical harmonic vector. We train our model on large-scale datasets such as Youtube 8M and CelebA. Our experiments show that our method can correctly estimate scene illumination and realistically re-light input images, without any supervision or a prior shape model. Compared to supervised methods, our approach has similar performance and avoids common lighting artifacts.

1. Introduction

Relighting images in the wild has gained popularity recently, especially since the development of mobile computing and video communication has led to an explosion in the consumption of digital photography. The diversity of the application environments, e.g. indoor, outdoor, day or night, makes the task of realistically relighting images challenging. In an ideal use case scenario, users can choose the desired illumination of an image, without having to consider the illumination of the original image. However, even state-of-art lighting algorithms meet three main problems. The first problem is the lack of large-scale relighting datasets since it is hard to manually label scene illumination, especially when there is more than one light source. The collection of image annotations has been the main bottleneck of many supervised relighting methods. The second problem is that most relighting algorithms need multiple views of the same object for training, which hinders the algorithms from learning from wild data. The third problem is that relighting usually requires depth information to avoid artifacts from shadows or over-lighting.

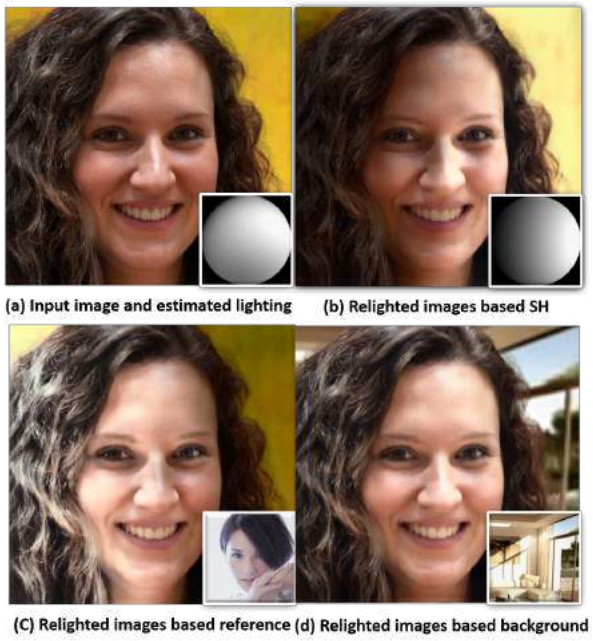


Figure 1. Relighted images with different target illumination conditions. Our method can relight an image (a) based on given light direction via spherical harmonic coefficients (b) as well as estimated illumination from a reference face (c) and environment scenes (d).

The goal of this work is to design a single automatic image relighting network with a large-scale single-view unlabelled dataset. It takes a single image and target lighting as inputs then estimates the lighting of the input image and subsequently generates a new, relighted image based on the target lighting. Specifically, our approach uses a self-supervised auto-encoder network which decomposes the image into two embeddings: one for illumination and one for content. There are two main challenges to this idea. The first challenge is how to separate illumination information from content information correctly. To address this,

we can augment images in such a way that the geometry of the objects stays the same while the direction of the light changes. To satisfy this assumption, we consider four possible augmentations: two flipped images, a rotated image and an inverted image. Each augmentation image is paired up with the original image and used to train a Siamese auto-encoder network. We assume that the image training pairs have the same content but different illumination. Based on this, we can decouple the image content embedding from the image illumination embedding. The second challenge is how to generate a semantically meaningful light representation. Without ground truth light information, the illumination embedding has a large number of possibilities. It is impossible to relight images by adjusting the embedding manually. Our solution is to design a spherical harmonic (SH) loss that forces the illumination embedding to take the form of Laplace's spherical harmonics. When Spherical Harmonics [18] represent the illumination embedding, the relighting can be meaningfully controlled.

The contribution of our work is two-fold: First, we propose a relighting self-supervised auto-encoder network, which is trained on image pairs and separates images into content and lighting embeddings. Without ground-truth illumination information, the proposed method can generate high-resolution (1024x1024) relighted images. Second, a novel spherical harmonics loss is introduced, which is based on the assumption that the proposed image augmentations only affect the direction of the light in the images and not the geometry of the objects. The relighting can be controlled by adjusting the values in illumination embedding. Finally, we test our method on both synthetic and real data. We show that we achieve higher fidelity in the relighted images compared to other supervised learning and self-supervised methods.

2. Related Work

2.1. Image Relighting

Relighting, especially for synthetic objects, has been a popular task for over twenty years [16, 27]. Most of the image relighting approaches which were introduced addressed face portrait images. With relative geometry information, landmarks and detectors of human faces, Zhou et al. [29] accurately estimate surface normals required for relighting from a single image based on an SFSNet network [20]. Sun et al. [24] use a similar architecture, but the encoder extracts lighting features in addition to facial features. However, the objective functions are only applied to masked people and the background are simply replaced by blurred environment maps. While much of the previous works focus on portrait relighting, only a few papers explicitly consider illumination estimation from environment images. Duchene et al. [8] propose an outdoor scene relighting network, which removes the shadows of the input images and adds new shadows for the relighted images based on a Markov Random Field over a graph of points. Philip et al. [19] further extend this idea by casting shadows from a 3D geometric prior. A SLAM algorithm is applied to estimate the position of the main light sources based on the scene's geometry and specular regions [26]. Zhang et al. [28] use RGB-D data to relight the indoor environment and estimates the materials of the furniture. However, most of the previous works need 2D/3D geometric priors or RGB-D sensors. Relighting a single-view scene image is still a challenging problem.

2.2. Photo Style Transfer

Photo style transfer, where the input is a source image and a reference image and the output is the illumination inversion image with the style of the reference image [21, 22], is similar to image relighting. The most challenging problem of a machine learning method is the difficulty in designing a suitable loss function. Since generative adversarial networks (GANs) [9] can estimate the loss from the training data, GANs are widely used in different photo style translation problems, such as pix2pix [10] and CycleGAN [30]. Instead of training the model with loss function, the generator and discriminator are trained. The generator learns to produce data similar to the training data, and discriminator learns to find the mistake of the generator. Auto-encoder is also applied in the photo style transfer [23, 14, 7]. For example, Shu et al. [23] propose a physically grounded rendering-based disentangling network specifically designed for faces. Landmarks and 3D face model are applied with an auto-encoder for estimation of face illumination. Li et al. [7] proposes a linear propagation module with a transformation to enable a feed-forward network for photo-realistic style transfer. However, the object shape in the transformed images is fuzzy, and the reference image affects the colour of transformed images.

2.3. Siamese network

The Siamese network architecture was proposed in the 1990s to solve signature verification as an image matching problem [4], single-target tracking [13, 3] and one-shot learning [11]. The pre-trained Siamese CNN features have been used with Context-RNNGAN model for image generation [6]. Compared to the above work, the relighting auto-encoder would replace the CNN network in the Siamese network, which is trained without external supervision. To the extent of our knowledge, our work is the first using a Siamese network for relighting.

3. The relighting network

Our goal is to learn an image relighting model from large unlabelled image datasets. The model receives a source image and target illumination as input and outputs an estima-

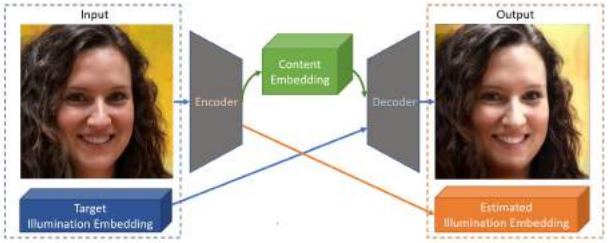


Figure 2. Auto-encoder: The input is the original image and target illumination embedding. The output is the re-lit image with the target illumination and estimated illumination embedding of the input image.

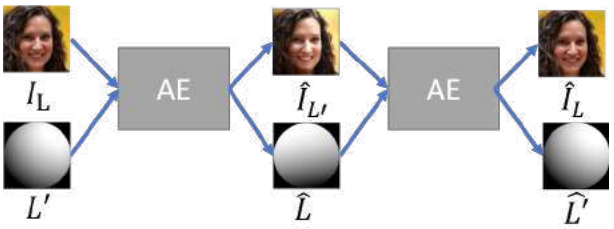


Figure 3. The Siamese reconstruction network. Two auto-encoders have the same structure, and they share weights. The input is the source image $\boldsymbol{I_L}$ while the output is the reconstructed image $\hat{\boldsymbol{I}}_{\boldsymbol{L}}$. $\boldsymbol{L}'$ is a random lighting illumination while $\hat{\boldsymbol{L}}'$ is the estimated illumination.

tion of the source image illumination and the relighted image. In this section, we provide the implementation details for the self-supervised relighting auto-encoder. Since the ground truth illumination information and relighted images are not available for wild datasets, the objective function and training detail of the relighting auto-encoder is further discussed.

3.1. Relighting Auto-encoding

The relighting network is an auto-encoder (AE), as shown in Fig. 2. With the encoder $\boldsymbol{E}$, the input image $\boldsymbol{I_L}$ is decomposed to a content embedding $\boldsymbol{C}$ (green boxes) and an illumination embedding $\hat{\boldsymbol{L}}$ (orange boxes). $\boldsymbol{I_L} \in \mathbb{R}^{w\times h\times 3}$ is the image $\boldsymbol{I}$ with illumination embedding $\boldsymbol{L}$, where w and h are the image weight and height. $\boldsymbol{C} \in \mathbb{R}^{m\times m\times d}$ and $\hat{\boldsymbol{L}} \in \mathbb{R}^n$ are two tensors, where n is the size of the illumination embedding, m is the size of the content embedding and d is the depth of the content embedding. The encoder network can be represented as:

$$\{\boldsymbol{C}, \hat{\boldsymbol{L}}\} = \boldsymbol{E}(\boldsymbol{I_L}) \tag{1}$$

With the decoder $\boldsymbol{D}$, the content embedding $\boldsymbol{C}$ and target illumination embedding $\boldsymbol{L}'$ (blue boxes) are used to rebuild the relighted image $\hat{\boldsymbol{I}}_{\boldsymbol{L}'}$. The decoder network can be represented as:

$$\hat{\boldsymbol{I}}_{\boldsymbol{L}'} = \boldsymbol{D}(\boldsymbol{C}, \boldsymbol{L}') \tag{2}$$

If the target and source illumination embedding are the same, the AE network becomes a reconstruction network. Otherwise, it is a relighting network. For training the relighting auto-encoder, we employ three objective functions. These are reconstruction loss, Spherical Harmonic loss and discriminator loss.

3.2. Reconstruction Loss

An auto-encoder network can learn $\boldsymbol{E}$ and $\boldsymbol{D}$ that can relight images from a large number of relighted image pairs. However, the ground truth illumination embedding and relighted images are not available for the wild data. To address this problem, two AE networks form a reconstruction network, as shown in Fig. 3. The two auto-encoders have the same structure and their weights are shared. Therefore, these two auto-encoders are called as Siamese auto-encoder. $\boldsymbol{I_L}$ is an image in the training dataset. $\boldsymbol{L}'$ is a random illumination embedding. $\hat{\boldsymbol{I}}_{\boldsymbol{L}'}$ and $\hat{\boldsymbol{I}}_{\boldsymbol{L}}$ is the transferred image with the illumination $\boldsymbol{L}'$ and the estimated illumination $\hat{\boldsymbol{L}}$, respectively. $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{L}}'$ is the estimated illumination from $\boldsymbol{I_L}$ and $\hat{\boldsymbol{I}}_{\boldsymbol{L}'}$, respectively. With the two transformation networks, the reconstruction network is setup. The network then takes source image $\boldsymbol{I_L}$ and target lighting $\boldsymbol{L}'$ as input and generates $\hat{\boldsymbol{I}}_{\boldsymbol{L}}$ and $\hat{\boldsymbol{L}}'$. $\boldsymbol{I_L}$ and $\boldsymbol{L}'$ are used as ground truth to supervise the training in this reconstruction task. We use mean absolute error loss for the reconstructed image $\hat{\boldsymbol{I}}_{\boldsymbol{L}}$ and the estimated lighting embedding $\hat{\boldsymbol{L}}'$. The image gradient is also considered with mean absolute error loss to preserve edges and avoid blurring. The objective for this reconstruction network is,

$$\begin{aligned}\mathcal{L}_{roc}(\boldsymbol{I_L}, \boldsymbol{L}') = &\frac{1}{w*h}\left(\left\|\boldsymbol{I_L} - \hat{\boldsymbol{I}}_{\boldsymbol{L}}\right\|_1 + \left\|\nabla\boldsymbol{I_L} - \nabla\hat{\boldsymbol{I}}_{\boldsymbol{L}}\right\|_1\right) \\ &+ \frac{1}{n}\left\|\boldsymbol{L}' - \hat{\boldsymbol{L}}'\right\|_1\end{aligned} \tag{3}$$

$\mathcal{L}_{roc}(\boldsymbol{I_L}, \boldsymbol{L}')$ is simply represented as $\mathcal{L}_{roc}(\boldsymbol{I_L})$, since $\boldsymbol{L}'$ is given and fixed in each epoch.

3.3. Spherical Harmonic Loss

Although the reconstruction loss can let the AE network converge, $\hat{\boldsymbol{I}}_{\boldsymbol{L}'}$ and $\hat{\boldsymbol{L}}$ lack constraints. The AE network can map the same image and illumination embedding to any random image and illumination in the target domain. Any of the learned mappings can induce an output distribution that matches the target distribution. Thus, reconstruction losses alone cannot guarantee that the AE network can relight images. To further reduce the space of possible mapping functions, some constraints should be applied for $\hat{\boldsymbol{I}}_{\boldsymbol{L}'}$ and $\hat{\boldsymbol{L}}$. A parametrization widely used in relighting tasks is a Spherical Harmonic, defined on the surface of a sphere. The details about calculating Spherical Harmonic can be found in [18]. Here we introduce a novel loss function, based on Spherical Harmonics, as a constraint on $\hat{\boldsymbol{I}}_{\boldsymbol{L}'}$.

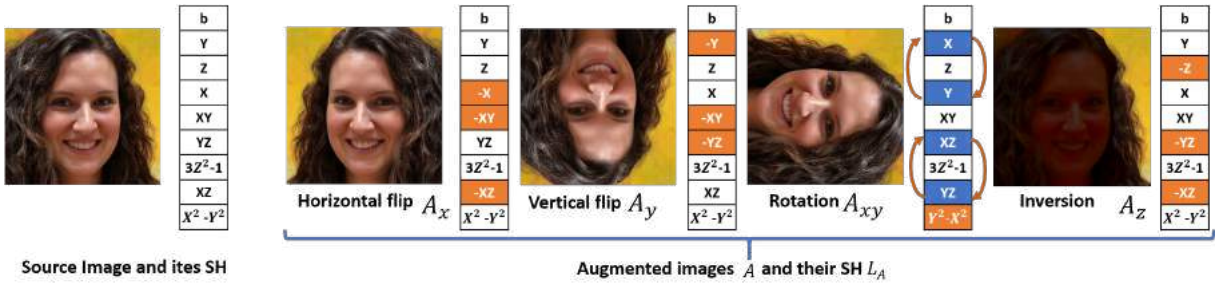


Figure 4. Augmentation images: the original image is shown at the bottom. The definition of Spherical Harmonic illumination embedding is given above the original image. The four kinds of augmented images are shown $\boldsymbol{A}_x$, $\boldsymbol{A}_y$, $\boldsymbol{A}_{xy}$ and $\boldsymbol{A}_z$, which are obtained as horizontal flip, vertical flip, rotation and inversion. These augmentations change some channels of the Spherical Harmonic illumination embedding of image. The changed illumination embedding is given below the augmented images, where the white channel is unchanged. The value in the orange channel turns negative. The blue channel is swapped with other blue channels as the orange lines.

Spherical Harmonic loss is a novel loss to guarantee that the illumination embedding is represented by Spherical Harmonic lighting. For controlling the illumination, Spherical Harmonic lighting is used to represent the illumination, shown above the original image in Fig. 4, where b is the bias and its value is $\frac{\sqrt{\pi}}{2}$. X, Y, Z means the channels is linearly dependent on the x, y and z axis in the space. Since the Spherical Harmonics of the training data is unknown, some augmented images $\boldsymbol{A}$ are used as the associated images to calculate the Spherical Harmonic Loss. Four image augmentations are considered. A horizontally flipped image $\boldsymbol{A}_x$, a vertically flipped image $\boldsymbol{A}_y$, a rotated 90 degrees image $\boldsymbol{A}_{xy}$, and an inverted image $\boldsymbol{A}_z$ (discussed later), as shown in Fig. 4. These augmentations change some values in the illumination embedding. The changes are given below each augmented images. The unchanged channel is shown as the white block, while the value in the orange block turns negative. Since $\boldsymbol{A}_{xy}$ is given by rotation, the order of some channels (blue channels) is changed. Specifically, the 2-nd and 6-th channels are swapped with 4-th and 8-th channels, respectively. Although $\boldsymbol{A}_y$ and $\boldsymbol{A}_xy$ introduces unnatural upside down or tilted images, this does not make the job of the auto-encoder more difficult in our experiments.

Different from $\boldsymbol{A}_x$, $\boldsymbol{A}_y$ and $\boldsymbol{A}_{xy}$ which can be calculated based on the flip and rotation operations, the inversion augmentation $\boldsymbol{A}_z$ is using a more sophisticated approach, as we need to approximate a lighting change in depth direction. We include a pretrained intrinsic image network, similar to SFSnet [20], since the illumination information in depth can not be directly given in the input images. The process of estimation of $\boldsymbol{A}_z$ is shown in Fig. 5. Firstly, the input image is converted from RGB into CIELAB (LAB) colour space. The L channel of the input image is shown as "L channel". Secondly, the depth information, shown as "depth", is estimated from the input image by a pre-trained depth prediction model [5]. The depth value of $\boldsymbol{I}$ is shown as $\mathcal{D}(\boldsymbol{I})$. The depth information is applied to separate the foreground

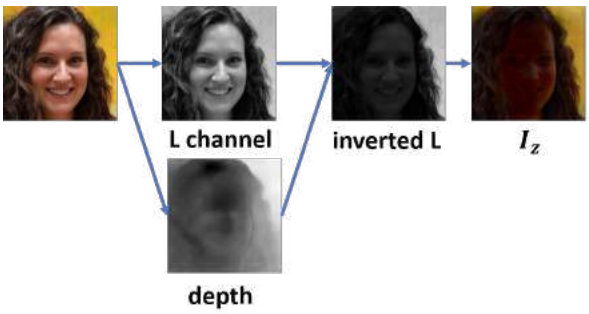


Figure 5. $\boldsymbol{A}_z$ pipeline: the "L channel" and "depth" information are estimated from the input LAB image. The "L channel" and "depth" information are further merged into "inverted L" as Eq. (4). Finally, $\boldsymbol{A}_z$ is build up from the "inverted L" and the AB channel of input image.

objects (faces, cars and so on) from the background. The pixels with low depth value belong to the objects while pixels with high depth value belong to the background. Apart from that, we assume light sources are normally located at the front of the objects. When the light sources are moved to the back of the objects, the luminance of objects decreases while the luminance of background stay the same since it has different light sources. The L channel of input images should be inverted based on the depth information to calculate $\boldsymbol{A}_z$. We tested a multitude of inversion approximations and found the following inversion equation to visually provide the best results:

$$\mathbb{L}(\boldsymbol{A}_z) = \frac{\|(\mathbf{1}_{w*h} - \mathcal{D}(\boldsymbol{I})) \circ \mathbb{L}(\boldsymbol{I}) + \mathcal{D}(\boldsymbol{I}) \circ \tanh(\mathbb{L}(\boldsymbol{I}))/2\|_1}{w * h} \tag{4}$$

where $\circ$ is the Hadamard product. $\mathbb{L}(.)$ is function to calculate the L channel information in LAB colour space and tanh is a tanh function. $\boldsymbol{A}_z$ is the image with $\mathbb{L}(\boldsymbol{A}_z)$. $\tanh(.)/2$ is the scale function, which can be adjusted for different tasks.

Based on the above properties, the relighting task is con-

verted to a comparison task, as shown in Fig. 6. $\boldsymbol{I_L}$ is a random image in the dataset, whose illumination embedding is the $\boldsymbol{L}$ and $\boldsymbol{L'}$ is the random illumination embedding. The augmented image of $\boldsymbol{I_L}$ is shown as the $\boldsymbol{A}$, where $\boldsymbol{A} \in \{\boldsymbol{A}_x, \boldsymbol{A}_y, \boldsymbol{A}_{xy}, \boldsymbol{A}_z\}$. The relighted $\boldsymbol{I_L}$ and $\boldsymbol{A}$ with the target illumination embedding $\boldsymbol{L'}$ are shown as $\hat{\boldsymbol{I}}_{\boldsymbol{L'}}$ and $\hat{\boldsymbol{A}}_{\boldsymbol{L'}}$, respectively. The estimated illumination embedding of $\boldsymbol{I_L}$ and $\boldsymbol{A}$ are $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{L}}_{\boldsymbol{A}}$. By comparing $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{L}}_{\boldsymbol{A}}$, Spherical Harmonic loss $\mathcal{L}_{sh}(\boldsymbol{I_L}, \boldsymbol{A})$ is defined as:

$$\mathcal{L}_{sh}(\boldsymbol{I_L}, \boldsymbol{A}) = \mathbf{1}_{9*1}^{T}\hat{\boldsymbol{L}} + \mathcal{C}(\boldsymbol{A})\hat{\boldsymbol{L}}_{\boldsymbol{A}} - \sqrt{\pi} \tag{5}$$

where

$$\mathcal{C}(\boldsymbol{A}) = \begin{cases} [1,-1,-1,1,1,-1,-1,1,-1]^T & \boldsymbol{A} = \boldsymbol{A}_x \\ [1,1,-1,-1,1,1,-1,-1,-1]^T & \boldsymbol{A} = \boldsymbol{A}_y \\ [1,-1,-1,-1,1,-1,-1,-1,1]^T & \boldsymbol{A} = \boldsymbol{A}_{xy} \\ [1,-1,1,-1,-1,1,-1,1,-1]^T & \boldsymbol{A} = \boldsymbol{A}_z \end{cases} \tag{6}$$

where $\mathbf{1}_{9\times1}^{T}$ is a 9×1 vectors, where each element value is 1. Eq. (5) is introduced based on the relationship between the input image $\boldsymbol{I_L}$ and its augmented image $\boldsymbol{A}$, as shown Fig. 4. The sum of first channel of $\boldsymbol{I_L}$ and $\boldsymbol{A}$ should be equal to $\sqrt{\pi}$, since the bias is $\sqrt{\pi}/2$. For 2-9 channel of the illumination embedding of $\boldsymbol{I_L}$ and $\boldsymbol{A}$, the different of the unchanged channel and sum of the unchanged channel are both equal to 0.

3.4. Discriminator loss

Since our network is trained using image flipping and rotation, they may contain artifacts due to inaccurate estimation of object depth and lighting. Apart from that, the relighted images lack the constraint condition. A GAN loss is proposed to improve the quality of the generated images. WGAN-GP is applied to force the distribution of local image patches to be close to that of a natural image. The objective is given as:

$$\mathcal{L}_{dis}(\boldsymbol{I_L}) = \mathbb{E}_{\boldsymbol{I_L}}\mathcal{C}\left(\boldsymbol{D}(\boldsymbol{E}(\boldsymbol{I_L}), \boldsymbol{L'})\right)^2 - \mathbb{E}_{\boldsymbol{I_L}}(\mathcal{C}(\boldsymbol{I_L}))^2 \tag{7}$$

where $\mathcal{C}$ is the critic (discriminator) where the re-light images are $\boldsymbol{E}(\boldsymbol{I_L})$ and $\boldsymbol{I_L}$ are the real images.

3.5. Implementation Details

The overall loss for our network is a linear combination of the losses mentioned:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{roc}(\boldsymbol{I_L}) + \alpha\mathcal{L}_{dis}(\boldsymbol{I_L}) \\ &+ \frac{1}{4}\sum_{\boldsymbol{A}\in\{\boldsymbol{A}_x,\boldsymbol{A}_y,\boldsymbol{A}_{xy},\boldsymbol{A}_z\}} \mathcal{L}_{roc}(\boldsymbol{A}) + \beta\mathcal{L}_{sh}(\boldsymbol{I_L}, \boldsymbol{A})\end{aligned} \tag{8}$$

where $\alpha = 0.5$ and $\beta = 0.25$. We train our network with images of resolution 1024×1024. More specifically, the source images pass through six down-sampling layers and eight residual blocks. Then the embedding passes through three residual blocks and two residual blocks with a fully-connected layer to get the content and illumination embeddings, respectively. The size m and depth d of the content embedding are set as 64 and 512, respectively. Then these embeddings are added after several residual blocks. Finally, a relighted image is generated after six up-sampling layers. Since the encoder losses some information present in the input images, the reconstructed image appears blurry. Therefore, six skip layers are added between the down-sampling and up-sampling layers. We train our network for 5000 epochs (about 300 hours) using the Adam optimizer with default parameters.

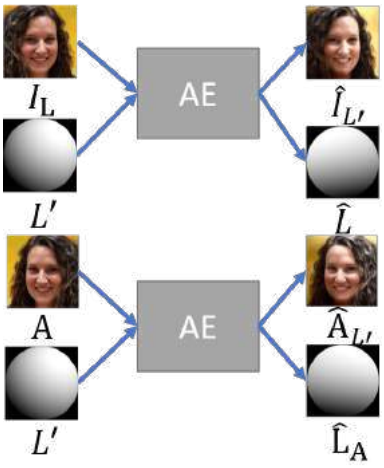


Figure 6. The Siamese comparison network. The two auto-encoder have the same structure, and they share weights. The input is source image $\boldsymbol{I_L}$ and augmented image $\boldsymbol{A}$ while the output is the illumination estimation $\hat{\boldsymbol{L}}$ and $\hat{\boldsymbol{L}}_{\boldsymbol{A}}$.

4. Experiments

In this section, we will evaluate our proposed method performance and compare it with previous state-of-the-art methods. Since our network can predict lighting, the model is evaluated in two ways: (A) Given a source image and an SH lighting, relighting an image (denoted as the SH-based relighting). (B) Given a source image and a reference image, estimating SH lighting from the reference image and using the estimated lighting to relight the source image (denoted as the image-based relighting).

4.1. Setup

Datasets: We show the effectiveness of the proposed method on three different datasets: CelebA [17], Youtube 8M [1] and synthetic face dataset [12]. CelebA is a large-scale face attributes dataset with about 200k portrait images. The persons shown in the dataset have large pose variation with different backgrounds. CelebA has large diversities

Model	RMSE ($_{10^{-3}}$)	DSSIM ($_{10^{-3}}$)	RMSE-s ($_{10^{-3}}$)
(1)Ours full	**6.8**	**3.8**	**9.9**
(2)w/o $\mathcal{L}_{roc}$	16.2	3.9	10.0
(3)w/o $\mathcal{L}_{sh}$	20.3	4.3	15.8
(4)w/o $\boldsymbol{A}_x$	18.6	4.5	15.7
(5)w/o $\boldsymbol{A}_y$	19.8	3.8	17.9
(6)w/o $\boldsymbol{A}_{xy}$	10.7	3.8	10.8
(7)w/o $\boldsymbol{A}_z$	7.3	3.8	13.1
(8)w/o $\mathcal{L}_{dis}$	7.1	3.9	10.4
(9)w/o skip	7.9	3.9	10.6

Table 1. Ablation study. Ablated model, which some part of the proposed model is removed, is evaluated.

and large quantities. However, the ground truth of illumination is not provided. Since image data of CelebA is limited, and we hope the model can accurately estimate the illumination of background. We pre-trained our model on Youtube 8M, where 6 million videos are included. We only choose a subset of the full dataset. The included categories are "bus", "metro", "park", "shopping mall", "street", "road", "office" and "room" which provide common use case scenarios. For each video, a frame is captured for every 1 minute. Finally, there are nearly 1 million scene images. Synthetic face dataset is generated by a synthetic face framework [12]. The illumination can be manually controlled. Therefore, this data is used to evaluate our proposed method. Note that this dataset is not used to train our methods.

Evaluation metric: Three error metrics are used to measure the relighting performance across our validation set and testing. They are the Root Mean Square Error (RMSE), Structural dissimilarity (DSSIM) [25], and scale-invariant RMSE (RMSE-s) [2]. RMSE is a common metric for evaluating the reconstruction task. DSSIM is invariant to local and global scaling and tinting. The DSSIM implementation uses a 11 $\times$ 11 Gaussian filter with α = 1.5, $k_1 = 0.01$ and $k_2 = 0.03$, as set in [25]. DSSIM is computed on each RGB channel of the input image and relighted image individually. It is not sensitive to global and local scaling or colour shifts.

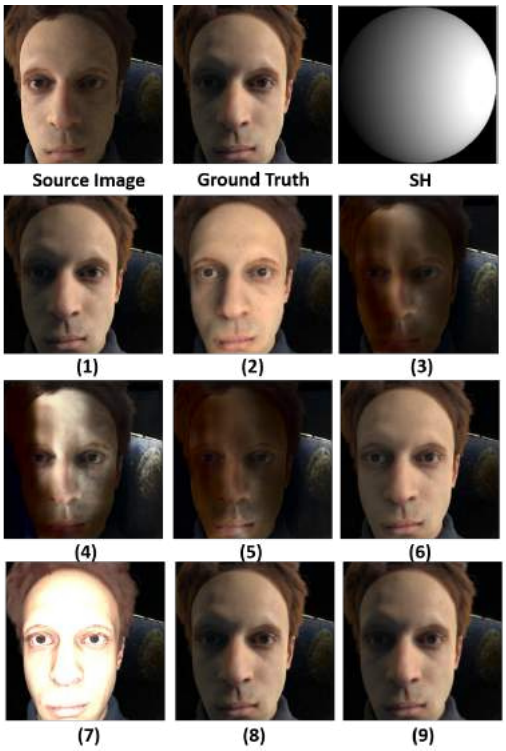
RMSE-s is introduced to solve the lighting scale problem (SH is affected by scale factors, such as the exposure time, except the lighting conditions). The RMSE-s solves for the single scale-factor applied to the predicted image,

$$\text{RMSE-s}(\boldsymbol{I_L}, \hat{\boldsymbol{I}}_{\boldsymbol{L}}) = \frac{1}{w * h} \min_{\alpha} \left\| \boldsymbol{I_L} - \alpha \hat{\boldsymbol{I}}_{\boldsymbol{L}} \right\|_2 \quad (9)$$

where α is a scalar. This metric solves for a single global scaling rather than a per-channel scaling. It is still sensitive to erroneous tints in the relighting image.

4.2. Ablation Study

To understand the influence of the individual parts of the model, we remove them one at a time and evaluate the performance of the ablated model in Table 1 and Fig. 7. Model (1) shows the performance of the full model with the reconstruction loss, SH loss and discriminator loss pre-trained on Youtube 8M and fine-tuning trained on the CelebA dataset. Model (2) does not use the reconstruction loss. Thus, it fails to light the image and some detail information is lost.



c

Figure 7. The first row shows the input image, the ground truth of the relighting image and the target SH. (1-9) show the relighted images of our full model, w/o $\mathcal{L}_{roc}$, w/o $\mathcal{L}_{sh}$, w/o $\boldsymbol{A}_x$, w/o $\boldsymbol{A}_y$, w/o $\boldsymbol{A}_{xy}$, w/o $\boldsymbol{A}_z$, w/o $\mathcal{L}_{dis}$ and w/o skip connection models, respectively.

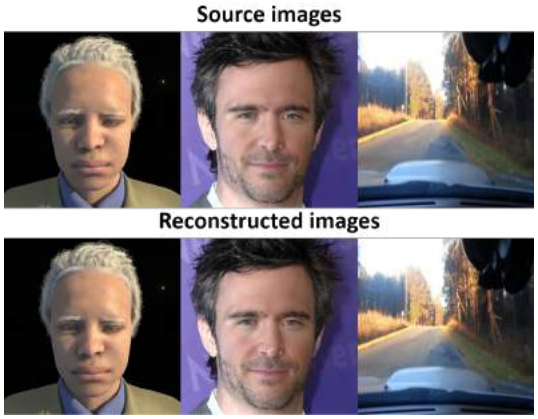


Figure 8. Reconstruction result of the reconstruction model build up two auto-encoder networks on the synthetic face, CelebA and Youtube 8M.

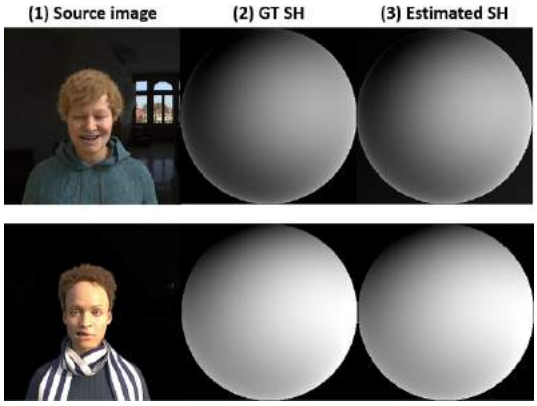


Figure 9. The estimated Spherical harmonic lighting of our method. The first column is the source images. The second column is the ground truth Spherical harmonic lighting. The third column is the estimated Spherical harmonic lighting of the source image.

The performance is lower than the full model. Model (3) does not use our proposed Spherical Harmonic Loss. In this case, the illumination embedding cannot be controlled in a semantically meaningful way. This also harms performance significantly. For further analysis, the performance SH loss, models (4-7) is shown. In rows (4-7), one of $\boldsymbol{A}_x$, $\boldsymbol{A}_y$, $\boldsymbol{A}_{xy}$ and $\boldsymbol{A}_z$ the SH loss is removed. Model (4) does not have a horizontal flip augmentation. Thus, the horizontal illumination can not be estimated and controlled and the performance decreases. Model (5) does not have a vertical flip, with a similar effect to Model (4). Model (6) shows the effect of $\boldsymbol{A}_{xy}$ is not significant, since only two channels are affected by $\boldsymbol{A}_{xy}$ in Eq. (6). Model (7) does not use the illumination inversion images, $\boldsymbol{A}_z$. The performance is influenced by over-exposure and under-exposure. Model (8) does not use Discriminator loss. We've found that the discriminator loss improves the visual quality of the relighted images and makes the relighted images sharp. Model (9) does not use the skip connection. Due to the limited space of embedding, some detail information lost. The accuracy drops significantly. As a result, we conclude that our full model can achieve a good balance between the accuracy and quality of the generated images.

4.3. Reconstruction and Lighting Estimation

In Fig. 8, we show the reconstruction results based on the model shown in Fig. 3. It is tested on synthetic face, CelebA and Youtube 8M, including generated face and real face. The reconstructed images contain fine details of the nose, eyes, mouth, even in the presence of extreme facial expression. The SH of the input image and estimated SH are shown in In Fig. 9. Our method can estimate the SH light from a single background image. For further evaluating our proposed method, the images of Youtube 8M dataset are applied as the reference background images to relight face. The relighted images with the new background are shown in Fig. 10. In our future work, we would test our proposed method on the asymmetric geometry and faces with strong shadows.

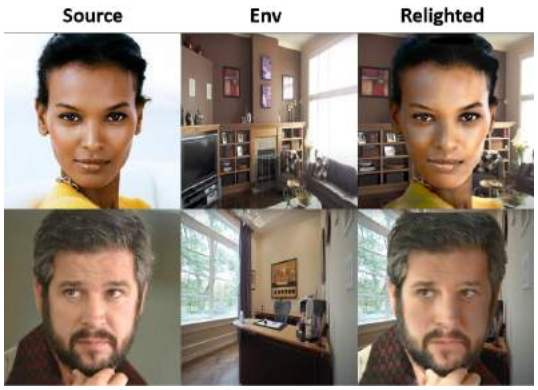


Figure 10. Visual lighting transformation results of the proposed method. The first column is the source images, the second column is the reference background images and the third column is the relighted images.

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

In this section, we compare our method with [20, 15, 29] on the synthetic face and CelebA dataset. Since the illumination information in unknown in CelebA dataset, image-based relighting is used to evaluate our proposed method and baseline methods in Fig. 11. More specifically, target lighting used for relighting is extracted from a reference image. The proposed method, SFSNet [20] and DPR [29] can extract the lighting of the reference image with its own lighting estimation method. Li et al. [15] is a state-of-the-art portrait style transfer method, which takes the source image and reference image as input and relights the source image as the illumination of the reference image. Since Li et al. and DPR are applied on the L channel of the input images, the input images are transferred from RGB image to Lab image. Our method and SFSNet are applied on the RGB channels of the input images. We notice that the edge of the relighted images of the proposed method is underexposed and some detail of face is lost. This is probably due to the gap between the training dataset and testing dataset. The results show that our proposed method can generate high-quality light estimations (1024×1024). Compared with DPR and SFSNet, our proposed method successfully overcomes the over-exposed images and the over-lighting problem of the nose and eyes. Li et al. do not generate images under the correct lighting. Since Li et al. uses reference images as input, the high-resolution images would improve the relighting performance while the low-resolution images can not estimate the reference lighting accurately. Since the illu-

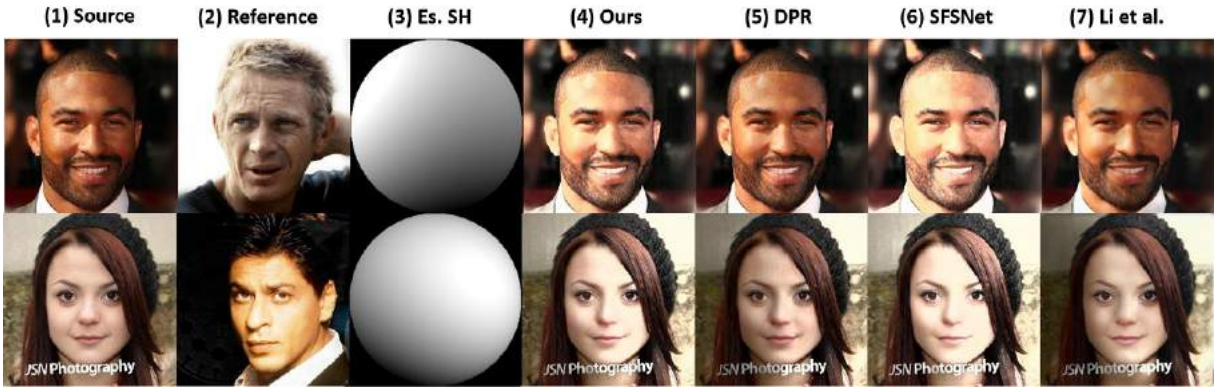


Figure 11. Visual lighting transformation results of the proposed method and state-of-the-art methods on CelebA dataset. The first three columns are the source image, the reference image and the estimated SH. Columns (4-7) shows the performance of the our proposed method, DPR [29], SFSNet [20] and Li et al. [15], respectively.

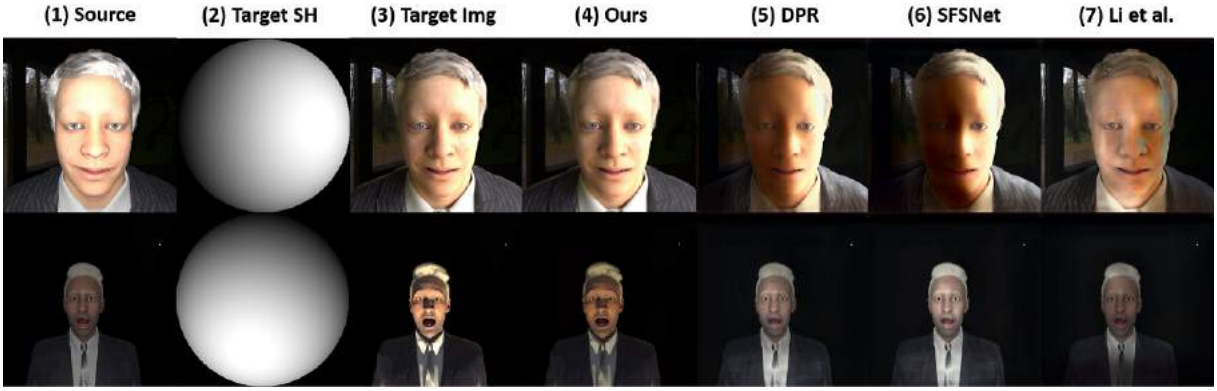


Figure 12. Visual lighting transformation results of the proposed method and state-of-the-art methods on synthetic face. The first three columns are the source image, the target SH and the target (ground truth) image. Columns (4-7) shows the performance of the our proposed method, DPR [29], SFSNet [20] and Li et al. [15], respectively.

mination information is known in synthetic faces, relighting based on the SH is applied to evaluate our proposed method and baseline methods by comparing with the ground truth images in Fig. 12. Our method can provide more face details and does not exhibit the over-lighting problem. In the second column, the skin colour of our relighted image is closer to the target image, since our proposed method input is a colour image, while the input to DPR is the L channel of the source image only. For further testing our proposed method, the RMSE-s between the target image and the relighted images of the different methods are calculated. The RMSE-s of ours (8.4×10^{-3}) is only 79% of the RMSE-s of DPR (10.6×10^{-3}), SFSNet (10.8×10^{-3}) and Li et al. (11.3×10^{-3}). We also evaluate our proposed method on Multi-PIE shown in our Github page, due to the limitation of pages.

5. Conclusion

We have proposed an automatic, unsupervised relighting algorithm trained on a large collection of unlabelled data. The relighting algorithm is build up by a Siamese Auto-encoder, where the source image information is split into content embedding and illumination embedding. Several auto-encoder networks are trained on the reconstruction, comparison and illumination estimation tasks. To provide target lighting, a Spherical Harmonic loss is first proposed, and four kinds of augmentation images are applied. We show that our training procedure, which combines reconstruction, Spherical Harmonic loss and adversarial losses, can estimate the illumination of the reference image and relight the source image. In addition, as our approach is trained on a large number of unlabelled data, it is less prone to exhibit common lighting artifacts and be applied on real as well as synthetic faces.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

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

[3] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional Siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[4] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah. Signature verification using a "Siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[5] W. Chen, Z. Fu, D. Yang, and J. Deng. Single-image depth perception in the wild. In *Advances in neural information processing systems*, pages 730–738, 2016.

[6] J. Choi, J. Krishnamurthy, A. Kembhavi, and A. Farhadi. Structured set matching networks for one-shot part labeling. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3627–3636, 2018.

[7] A. P. Dherse, M. N. Everaert, and J. J. Gwizdała. Scene relighting with illumination estimation in the latent space on an encoder-decoder scheme. *arXiv preprint arXiv:2006.02333*, 2020.

[8] S. Duchêne, C. Riant, G. Chaurasia, J. Lopez-Moreno, P.-Y. Laffont, S. Popov, A. Bousseau, and G. Drettakis. Multi-view intrinsic images of outdoors scenes with an application to relighting. *ACM Transactions on Graphics (TOG)*, 2015.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. 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.

[11] G. Koch, R. Zemel, and R. Salakhutdinov. Siamese neural networks for one-shot image recognition. In *ICML deep learning workshop*, volume 2. Lille, 2015.

[12] M. Kowalski, S. J. Garbin, V. Estellers, T. Baltrušaitis, M. Johnson, and J. Shotton. Config: Controllable neural face image generation. *arXiv preprint arXiv:2005.02671*, 2020.

[13] B. Li, W. Wu, Q. Wang, F. Zhang, J. Xing, and J. Yan. Siamrpn++: Evolution of Siamese visual tracking with very deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019.

[14] X. Li, S. Liu, J. Kautz, and M.-H. Yang. Learning linear transformations for fast image and video style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3809–3817, 2019.

[15] Y. Li, M.-Y. Liu, X. Li, M.-H. Yang, and J. Kautz. A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468, 2018.

[16] A. Liu, S. Ginosar, T. Zhou, A. A. Efros, and N. Snavely. Learning to factorize and relight a city. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[17] Z. Liu, P. Luo, X. Wang, and X. Tang. Large-scale celebfaces attributes (celeba) dataset. *Retrieved August*, 15:2018, 2018.

[18] T. M. MacRobert. Spherical harmonics: an elementary treatise on harmonic functions with applications. 1947.

[19] J. Philip, M. Gharbi, T. Zhou, A. A. Efros, and G. Drettakis. Multi-view relighting using a geometry-aware network. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019.

[20] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. 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.

[21] Y. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand. Style transfer for headshot portraits. *Association for Computing Machinery (ACM)*, 2014.

[22] Z. Shu, S. Hadap, E. Shechtman, K. Sunkavalli, S. Paris, and D. Samaras. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2017.

[23] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5541–5550, 2017.

[24] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):79–1, 2019.

[25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[26] T. Whelan, R. F. Salas-Moreno, B. Glocker, A. J. Davison, and S. Leutenegger. Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research*, 35(14):1697–1716, 2016.

[27] Y. Yu, A. Meka, M. Elgharib, H.-P. Seidel, C. Theobalt, and W. A. P. Smith. Self-supervised outdoor scene relighting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[28] E. Zhang, M. F. Cohen, and B. Curless. Emptying, refurnishing, and relighting indoor spaces. *ACM Transactions on Graphics (TOG)*, 35(6):1–14, 2016.

[29] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7194–7202, 2019.

[30] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.