Half-body Portrait Relighting with Overcomplete Lighting Representation

Guoxian Song[1] ,Tat-Jen Cham[1] , Jianfei Cai[1,2] and Jianmin Zheng[1]

[1]Nanyang Technological University, Singapore
guoxian001@e.ntu.edu.sg, ASJTCham@ntu.edu.sg, ASJMZheng@ntu.edu.sg
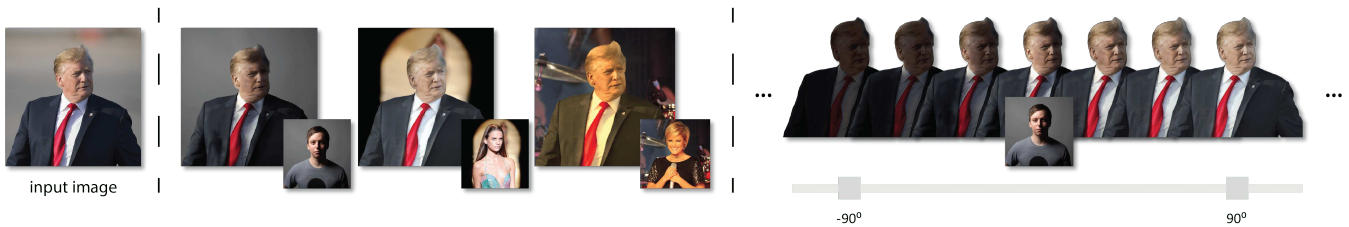[2]Monash University, Australia
jianfei.cai@monash.edu

Figure 1: Given an image, we can generate several relighted images referring to other portrait images. Besides, we can also horizontally rotate the illumination with arbitrary angle without further input.

Abstract

We present a neural-based model for relighting a half-body portrait image by simply referring to another portrait image with the desired lighting condition. Rather than following classical inverse rendering methodology that involves estimating normals, albedo and environment maps, we implicitly encode the subject and lighting in a latent space, and use these latent codes to generate relighted images by neural rendering. A key technical innovation is the use of a novel overcomplete lighting representation, which facilitates lighting interpolation in the latent space, as well as helping regularize the self-organization of the lighting latent space during training. In addition, we propose a novel multiplicative neural render that more effectively combines the subject and lighting latent codes for rendering. We also created a large-scale photorealistic rendered relighting dataset for training, which allows our model to generalize well to real images. Extensive experiments demonstrate that our system not only outperforms existing methods for referral-based portrait relighting, but also has the capability generate sequences of relighted images via lighting rotations.

CCS Concepts
• ***Human-centered computing*** → *Visual analytics;*

1. Introduction

Portrait relighting is a much sought-after advanced capability in digital photography. While professional photographers may take the time to set up perfect portrait shots meticulously with special equipment, normal users often only care about improving protraits retrospectively, e.g. selfie beautification, for which relighting can often help. Commercial applications such as film editing, telepresence and augmented reality can also benefit from this.

In the film industry, highly realistic relighting for special effects may be achieved by first capturing and modeling high quality ge-

ometry and reflectances of the subject under heavily instrumented and controlled studio conditions, e.g. using a 3D scanner and camera arrays such as the Light Stage [GFT*11, GTB*13]. However, such a setup is complex and expensive, thus irrelevant for normal consumers.

Our intention is to have a system that can relight a casually taken portrait image by simply *referring* to another image with the desired lighting condition, and can even interactively adjust the lighting. There have been some advances on the topics related to this goal. In particular, [SPB*14, SHS*17] introduced histogram trans-

fer methods that swap local color statistics between two facial images. However, the visual results are poor when the two images do not share similar facial attributes. Inverse rendering techniques [PKA*09, TZK*17, GZC*17, SKCJ18] are able, to some extent, to recover geometry, reflectance and lighting from a single image, but are usually based on the assumption of Lambertian reflectance and a limited second order spherical harmonic (SH) illumination model for the environment. Another option is to directly estimate an environment map from a portrait image [SBT*19], which requires a light stage for training data.

Image-based relighting still remains an open problem, since it is extremely hard to accurately estimate explicit face geometry, reflectance map and environment map from single image, for which ground truth is also difficult to obtain. Moreover, incorrect estimates will propagate into the relighting and lead to artifacts and unrealistic rendering. Often there may well be insufficient information to have these physical attributes be separable, which makes image-based relighting to be ill-posed. However, we postulate that not all such attributes need to be recovered for relighting purposes; these are only required if conventional (physics-based) rendering has to be used subsequently. To this end, we propose a deep relighting model for half-body portraits via an implicit form of inverse rendering for portrait relighting and illumination manipulation (see fig. 1). Unlike previous methods, which explicitly estimate normals, albedo, SH lighting [SKCJ18] or environment maps [SBT*19], we implicitly encode the subject and lighting in a latent space, and use the latent codes to generate relighted images by neural rendering. One innovation that we introduce is an overcomplete lighting representation, which facilitates lighting interpolation in the latent space as well as regularizing the self-organization of the lighting latent space. Moreover, we propose a novel multiplicative neural render for combining subject and lighting latent codes, leading to improved results. One unique feature of our proposed neural relighting framework is that we can generate a sequence of relighted images via lighting rotations, based on only a single pair of source and target portrait images.

For the end-to-end training of the proposed neural relighting model, we need a large number of portrait images with multiple known illumination conditions. To the best of our knowledge, there is no such publicly available dataset. Using collected high-quality 3D scans of real humans and real high dynamic range (HDR) environment maps, we rendered numerous photorealistic images to create such a synthetic dataset.

In summary, the main contributions of this paper are:

- We constructed a large-scale photorealistically rendered dataset using real 3D human scans and HDR environment maps, with lighting rotation annotations for relighting task.
- We present a neural relighting framework for half-body portraits, which performs implicit forward and inverse neural rendering. The framework has multiple novel components: the overcomplete lighting representation, the multiplicative neural render and the separated foreground-background lighting encoding. We will release our dataset and code on publication. [†]

[†] https://github.com/GuoxianSong/Portrait-Relighting.git

- We conducted extensive experiments on synthetic and real images. Both qualitative and quantitative results demonstrate that our method outperforms existing methods for referral-based portrait relighting. Moreover, our method can generate a sequence of relighted images for arbitrary horizontal lighting rotations, without need for further input.

2. Related Work

2.1. Portrait image capture

Portrait image can be captured in a professional studio with special equipment such as flashes, deflectors and diffusion panels. In the computer graphics community, other specialized hardware used include 3D scanners and the Light Stage [DHT*00, GFT*11, GTB*13] to obtain geometry and reflectance maps, so that photorealistic portrait images can be ray-trace rendered. To capture a large real relighting dataset is challenging, as subjects must remain stationary while the lighting is varied. In our work, we collected high-quality 3D geometry and reflectance maps of real people from artist assets and used a professional rendering engine to synthesize various half-body portrait images.

2.2. Image-based relighting

The histogram based method [SPB*14] transfers local contrasts and overall lighting from one portrait to another. However, it requires both input and reference pictures with compatible appearance attributes like beards and skin color; otherwise it would generate visible artifacts. Later on, a geometry-aware relighting method [SHS*17] based on 3DMM (3D morphable models) is able to generate more robust color remapping, but only works with facial regions instead of half-body portraits.

Inverse rendering methods [LCY*17, IRWM17] aim to decompose images into physical attributes of geometry, reflectance and lighting. This approach has received extended attention over time, especially for human facial images [DRC*15, YS19, BM16]. However, inverse rendering is a difficult ill-posed problem without additional constraint assumptions. Leveraging on large internet-scale datasets, recent deep learning-based methods such as Mofa [TZK*17], SfSNet [SKCJ18], and [GZC*17] have been proposed, which typically use second-order spherical harmonic (SH) functions to model environment lighting, and 3DMM-based facial models to support the estimation of facial geometry, normals and albedo. To make inverse rendering more tractable, these methods typically employ low-frequency illumination models and low-dimensional 3DMM face models. Face-based methods are also limited to relighting facial regions and cannot trivially be extended to half body portraits with visible upper bodies and clothing.

For more general photographs, Zhou et al. [ZJ19] warped 3DMM normals to head regions with an as-rigid-as-possible mapping. They constructed a large facial relighting synthetic dataset using varied SH lighting and proposed a deep single image portrait relighting method. However, their method and dataset only relights the L-channel in Lab color space instead of full RGB. Sun et al. [SBT*19] collected a 'one-light-at-a-time' (OLAT) real human portrait dataset and trained a U-Net based PR-Net to directly

estimate environment map, which also facilitates relighting from a single human portrait. However, training this network requires extensive illumination maps and corresponding solid angle supervision, obtained from the physical collection of real images under finely-calibrated ground truth lighting. Broadly, all these methods require extensive optimization or a complex capture setup in order to obtain explicit labels like SH lighting or environment maps. In contrast, our work learns an implicit self-organized overcomplete lighting representation from images, and does not require other additional direct supervisions such as environment maps, SH lighting, geometric normals, etc.

2.3. Deep image translation

The seminal work for deep image translation is pix2pix [IZZE17], which translates an image from one domain to another using paired training data. It has been applied to many tasks such as inpainting [PKD*16], semantic labeling [DYWG17] and super resolution [LTH*17]. It has also been extended to unpaired settings [ZCC18] and for multimodal output [ZZP*17, HLBK18]. Unsurprisingly, deep image translation has also been used in the task of neural rendering. Thies et al. [TZN19] proposed a deferred neural rendering method to synthesize images from imperfect 3D content. Meshry et al. [MGK*19] extracted appearance vectors from a deep buffer and injected the vectors into a latent space for neural rendering in-the-wild architectural images. Sengupta et al. [SGK*19] used inverse rendering and a residual appearance render network for scene images. Inspired by these works, we cast relighting conceptually as a multimodal image synthesis problem, while our network design is also influenced by ideas from classical rendering and inverse rendering processes.

3. Framework

We begin with a simplified formulation for rendering and inverse rendering:

$$I = \text{ren}(s, l); \quad s = \text{invr}_s(I), \ l = \text{invr}_l(I) \quad (1)$$

where I is an image, s is the subject content, and l is the environment lighting, while $\text{ren}(\cdot)$ and $\text{invr}(\cdot)$ are the forward rendering and inverse rendering functions, respectively. In this framework, s and l remain in latent representations and are not explicitly recovered, and the partitioning of the inverse rendering into separate functions for recovering the subject and lighting features reflects our intention to disentangle these representations.

The focus of this work however is on *relighting* rather than inverse rendering. Given separate scenes x and y, the goal is to obtain

$$I_{x,y} = \text{ren}(\text{invr}_s(I_x), \text{invr}_l(I_y)) \quad (2)$$

where $I_{x,y}$ refers to an image in which the subject comes from scene x, but is lit by the illumination from scene y.

Instead of using only a single latent vector for representing the environment lighting, we propose a new idea of using a **three-way overcomplete tensorial representation** (denoted **OT3**):

$$\mathbf{L} = \begin{bmatrix} l^{90} & l^0 & l^{-90} \end{bmatrix} = \text{invr}_l^{\text{OT3}}(I) \quad (3)$$

where l^d denotes a latent lighting vector corresponding to an environment map that has undergone d° rotation about the vertical axis. In essence, this representation includes extra estimates of the lighting codes should the environmental map be rotated through $+90^\circ$ and -90°. Notice that this representation is *overcomplete*, because l^{90} and l^{-90} do not in fact contain new information not already present in l^0.

There are two key benefits to this representation:
1) The lighting codes are in latent space, and cannot be conventionally manipulated (e.g. via rotation matrices). The three vectors serve as *anchor points* in latent space, facilitating interpolation between them.
2) During training, the tensorial representation $\mathbf{L}$ enhances self-organization, because *rotations in environmental lighting will ideally induce column-wise coordinate shifts in the tensor* (akin to pixel-wise translation in images). For example, if the environmental lighting was rotated by 90° to form image I^{90}, then we should get

$$\mathbf{L}^{90} = \begin{bmatrix} l^{180} & l^{90} & l^0 \end{bmatrix} = \text{invr}_l^{\text{OT3}}(I^{90})$$

and w.r.t. (3) we see that l^{90} and l^0 have been right-shifted in the tensor. Thus the redundant over-representation can serve the purpose of regularizing the self-organization of the latent space.

Fig. 2 provides an overview of our system. Given source and target images, we extract foreground (human) segmentation maps. A subject encoder E_s is used to extract subject information from the foreground of the source image. Two encoders E_b and E_f are respectively used to extract background and foreground illumination features from the target image. These foreground and background illumination features are concatenated and transferred into the lighting decoder D, which generates the three anchor lighting codes in (3). After that, the subject features are combined with the lighting codes to generate multiple relighted images via the neural render network R. Finally, we carry out post-processing by inpainting parts of the target background (de-occluded by removal of subject) and combining with the relighted source foreground.

3.1. Dataset Generation

We collected 98 3D scans of real human subjects with highly detailed meshes and high resolution appearance maps from 3D collection websites [Gob, 3D]. We randomly split the dataset into training and test sets of 68 and 30 meshes, respectively. We also collected 228 real outdoor environment maps with 1k resolution from HDRI Haven [HDR], in which training and test sets were randomly split into 160 and 68 maps respectively. We used a professional ray-tracing engine [Arn] to render photorealistic images.

We rendered half-body figures from viewpoints which are front-facing or partly oblique. First, we coarsely aligned the figures via global rotation, scaling and translation, such that the 3D torsos had similar widths. Then for each environment map, we created a series of scenes in which the environment illumination was sequentially rotated through 30° around the vertical axis 12 times. Finally, we rendered images and foreground segmentation maps, based on a camera with a 45 mm focal length and from a fixed viewpoint. The size of the images are 512×512, suitable for network input. Examples can be seen in fig. 3. In total, we generated 130,560 training
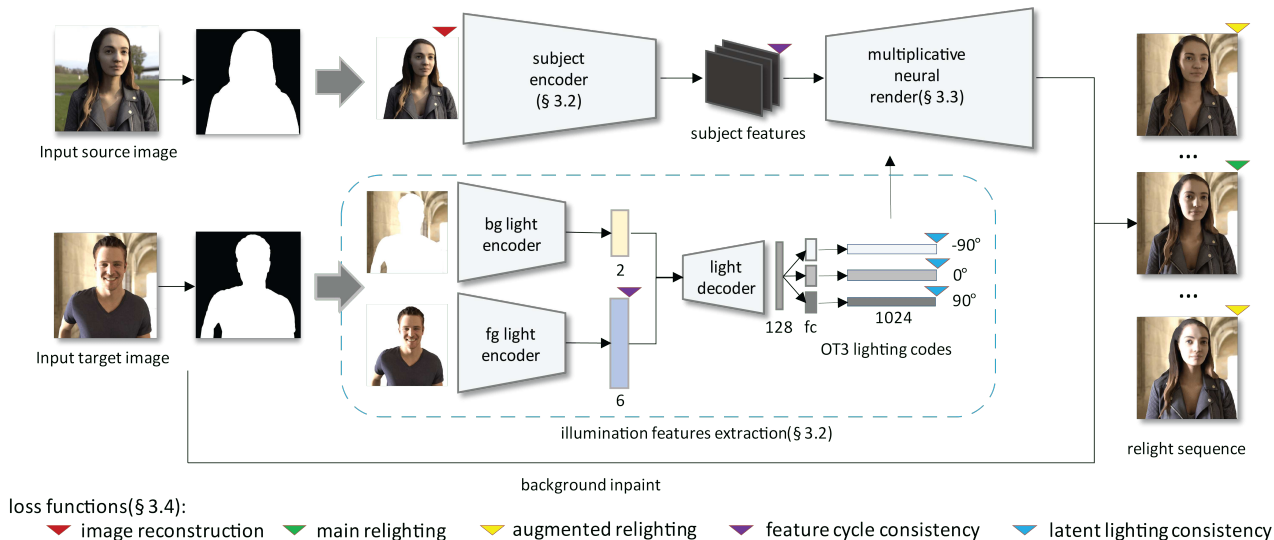
Figure 2: An overview of our portrait relighting system. The locations of loss functions are shown; see text for details.

images under 1,920 illuminations and 24,480 testing images under 816 illuminations, with no further augmentation. Note that there is no overlap between training and test subjects, nor between training and test illuminations. Furthermore, training illuminations are only used with training subjects, and likewise for the test images.

Figure 3: Samples from our dataset.

3.2. Illumination and Subject Encoding

In a portrait image, the appearance of the subject is mainly modulated by the foreground lighting, i.e. the illumination behind the camera, but this is hard to estimate (hence techniques such as [LMF*19]). The background lighting behind the person can be directly observed, but only has an indirect effect on the appearance; nonetheless there may be significant correlation to the foreground lighting, e.g. in a mountain scene, the background grass and lake may help estimate the dominant sun direction behind the camera.

In our framework, we initially have separate foreground and background lighting encoders of different channel sizes, with foreground lighting having more channels due to its greater signifi-

cance. These separate illumination features are subsequently concatenated and further decoded into the three-way overcomplete latent representation of the lighting, for rendering the subject.

Specifically, given an input image I with subject segmentation mask M, our illumination encoders E_b and E_f extract features $i = (i_b, i_f) = (E_b(I_b), E_f(I_f))$ from the foreground $I_f = I \odot M$ and background $I_b = I \odot (1 - M)$ image regions. As we do not intend in this work to inversely render a detailed environment map, the encoders are highly compressive, with i_f a 6-dim vector and i_b a 2-dim vector. Next, i is expanded via a common multi-layer perceptron and three separate fully-connected (fc) layers to obtain the three lighting codes $\{l^{90}, l^0, l^{-90}\}$, each a 1024-vector.

In parallel to the above, we also need to encode the subject in the source image. The source image and the foreground segmentation map are passed through a subject encoder $s = E_s(I_f)$ to obtain the subject features s. The features s and i_f are disentangled due to the loss functions used in training (later in sec. 3.4).

3.3. Multiplicative Neural Rendering

After obtaining the lighting codes and the subject features, the subsequent step is to render the image, which has to be done neurally since the representations are all latent in nature. A straightforward way is to simply concatenate lighting and subject features and decode into image space [MGK*19]. However, from our investigations this does not easily lead to good results(See Sec.4.2.2.). A possible reason is that a deconvolutional network may not have the ideal structural form to model the rendering process from this representation.

Our approach is to adopt some structural aspects from classical graphics rendering models. These typically involve the product of illumination and reflectances, with an additive component of ambient illumination. As such, instead of simply concatenating lighting and subject features, a combination of multiplication
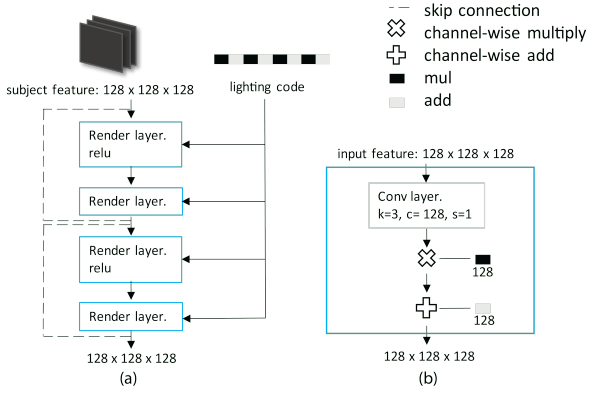
Figure 4: Illustration: (a) multi-scale multiplicative neural render; (b) multiplicative render layer.

and addition is used to combine the lighting codes and subject features, which we call *multiplicative neural render* (MNR). While there is no direct correspondence between multiplication in classic rendering and neural multiplication of latent features, empirically we found that this works well compared to concatenation or purely addition.

First, each lighting vector $l \in \mathbb{R}^{1024}$ is sequentially partitioned into 4 pairs of sub lighting codes as multi-scale multiplicative and additive components, i.e. $l = (l_{mul}^1, l_{add}^1, l_{mul}^2, l_{add}^2, \ldots)$, where the superscript denotes the corresponding render layer index and each component is in 128 dimension, illustrated in Fig. 4(a). For the first render layer, the following neural multiplication takes place:

$$out^1(l,s) = l_{mul}^1 \times s + l_{add}^1. \tag{4}$$

The output of the render layer $out^1(l,s)$ is then used as the input for the next render layer, illustrated in Fig. 4(a). The operation is performed in the channel dimension (Fig. 4(b)), which keeps our network fully convolutional. We stack four multiplicative neural render layers to relight deep subject features, and the lighting vector progressively changes the subject features.

After relighting the foreground subject, we can further embed the subject into the background of the target image. To cope with the removal of the foreground in the target image, we first use a simple fast marching approach [Tel04] to inpaint and complete the target image background, after which the source foreground subject can be composited.

3.4. Loss functions

To train the networks, we use several loss functions. The parts of our system to which these loss functions apply are indicated in Fig. 2.

Image reconstruction: The neural renderer R should reconstruct the foreground subject in the source image I_x, from extracted subject feature s_x and lighting code l_x^0:

$$L_{recon}^x = ||M_x \odot (R(l_x^0, s_x) - I_x)||_1 \tag{5}$$

Main relighting: Given I_x and lighting code l_y^0 from I_y, the relighted image should be close to the groundtruth $I_{x,y}^0$:

$$L_{relight}^{x,y} = ||M_x \odot (R(l_y^0, s_x) - I_{x,y}^0)||_1 \tag{6}$$

Augmented relighting: Since the OT3 lighting representation provides two other lighting codes l_y^{-90} and l_y^{90}, we can render additional relighted images for comparing to the groundtruth $I_{x,y}^{-90}$ and $I_{x,y}^{90}$:

$$L_{auglight}^{x,y} = ||M_x \odot (R(l_y^{90}, s_x) - I_{x,y}^{90})||_1 + \\ ||M_x \odot (R(l_y^{-90}, s_x) - I_{x,y}^{-90})||_1 \tag{7}$$

Feature cycle consistency: Given illumination features $i = (i_b, i_f)$ randomly sampled from normal distribution and subject features s_x from I_x, we can render a relighted image $\hat{I}_{x,y} = M_x \odot R(l^0, s_x)$, where $l^0 = D^0(i)$ is the $0°$ lighting code mapped from i by the MLP decoder D^0. If $\hat{I}_{x,y}$ is re-encoded again, we want the recomputed subject and foreground-only illumination features to be consistent with s_x and i_f respectively. The random noise greatly helps data augmentation and encourages s_x and i_f to be disentangled, which benefits from implicit representation.

$$L_{feat}^{x,y} = ||E_s(\hat{I}_{x,y}) - s_x||_1 + ||E_f(\hat{I}_{x,y}) - i_f||_1 \tag{8}$$

Latent lighting consistency: Given image I_y and lighting-rotated counterparts I_y^{-90} and I_y^{90}, we can extract the overcomplete OT3 lighting codes from each. Suppose the codes for I_y are $\{l_y^{-90}, l_y^0, l_y^{90}\}$. Considering that the codes for I_y^{-90} and I_y^{90} overlap with I_y's at four instances, and overlap with each other's at the lighting rotation of $\pm180°$, we add the following latent lighting consistency loss by denoting E_i as the concatenated form of E_b and E_f:

$$L_{cons}^y = ||l_y^{90} - D^0(E_i(I_y^{90}))||_1 + ||l_y^0 - D^{-90}(E_i(I_y^{90}))||_1 + \\ ||l_y^0 - D^{90}(E_i(I_y^{-90}))||_1 + ||l_y^{-90} - D^0(E_i(I_y^{-90}))||_1 + \\ ||D^{-90}(E_i(I_y^{-90})) - D^{90}(E_i(I_y^{90}))||_1 \tag{9}$$

Total loss: We jointly train the encoders (E_s, E_f and E_b), MLP lighting decoder D and neural renderer R to optimize the combined objective function:

$$L_{total}^{x,y} = L_{recon}^x + L_{relight}^{x,y} + \lambda_a L_{auglight}^{x,y} + \lambda_f L_{feat}^{x,y} + \lambda_c L_{cons}^y \tag{10}$$

where λ_a, λ_f, λ_c are set as 0.5, 0.1, 0.25, respectively.

Our relighting network is to disentangle portrait images into implicit illumination and subject content features. Unlike other methods using physical conventional rendering, our method adopts neural render approach learned using large relighting dataset. To improve the neural render quality, we integrate several novel components the overcomplete lighting representation, the multiplicative neural render and the separated foreground-background; and designed delicate loss functions: the image reconstruction, relighting and augmented is to disentangle illumination and subject content features, while feature cycle consistency and latent lighting consistency losses are aimed to regularize the learned structural representation.

Figure 5: *Our relighting results of half body portraits of real images from CelebA dataset [LLWT15]. The first col are source images, and the first row are target images for two lighting conditions: front lighting(col. 2-4) and side lighting(col. 5-6). The front lighting images are from CelebA dataset, and side lighting images are from our test dataset.*

4. Experiments

Implementation Our encoders, decoder and neural render networks follow recent autoencoder network structures in [KW14, HLBK18, ZZP*17, ZPIE17] with residual blocks [HZRS16]. The subject encoder down-samples a $512{\times}512{\times}3$ image to a $128 \times 128 \times 128$ feature with convolutional layers and residual blocks. The two lighting encoders convert a target image into a compressive latent feature with 8 dimensions, using convolutional layers, global average pooling, a fully connected layer and concatenation. The lighting decoder transforms this compressive light feature to three lighting codes of 1024 dimensions each. The neural render network uses the lighting codes to reweigh the subject feature through four neural multiplicative layers in progression, and finally up-samples to a $512{\times}512{\times}3$ output using several deconvolutional layers. The network details are in the appendices material.

We jointly trained all the networks using the Adam optimizer with a learning rate of $1.5{\times}10^{-5}$, a batch size of 2, and 5 epochs on an NVIDIA Quadro P5000 in Pytorch [PGC*17].

4.1. Comparison to Existing Methods

4.1.1. Overall visual results

Fig. 5 shows our results of relighting real images from the CelebA dataset [LLWT15]. Real images were pre-processed to segment foreground from background using tool [rem]. For each input source and target pair, we show the relighted composited image. It can be seen that our method can robustly handle a variety of real photos containing different subjects, illumination conditions and clothing styles, although we trained the networks using only our synthetically rendered images. This suggests that our rendered dataset is sufficiently realistic.

4.1.2. Qualitative comparison

In Fig. 6, we visually compare our method on the single image relighting task with two classical methods: histogram transfer [XM06] and style transfer [GEB15], as well as three state-of-the-art relighting methods: a multi-scale histogram transfer method [SPB*14], the intrinsic image decomposition method of Sf-SNet [SKCJ18] and PR-Net [SBT*19]. Classical histogram transfer [XM06] matches pixel color distributions while style transfer [GEB15] matches deep VGG19 features, and in both cases the
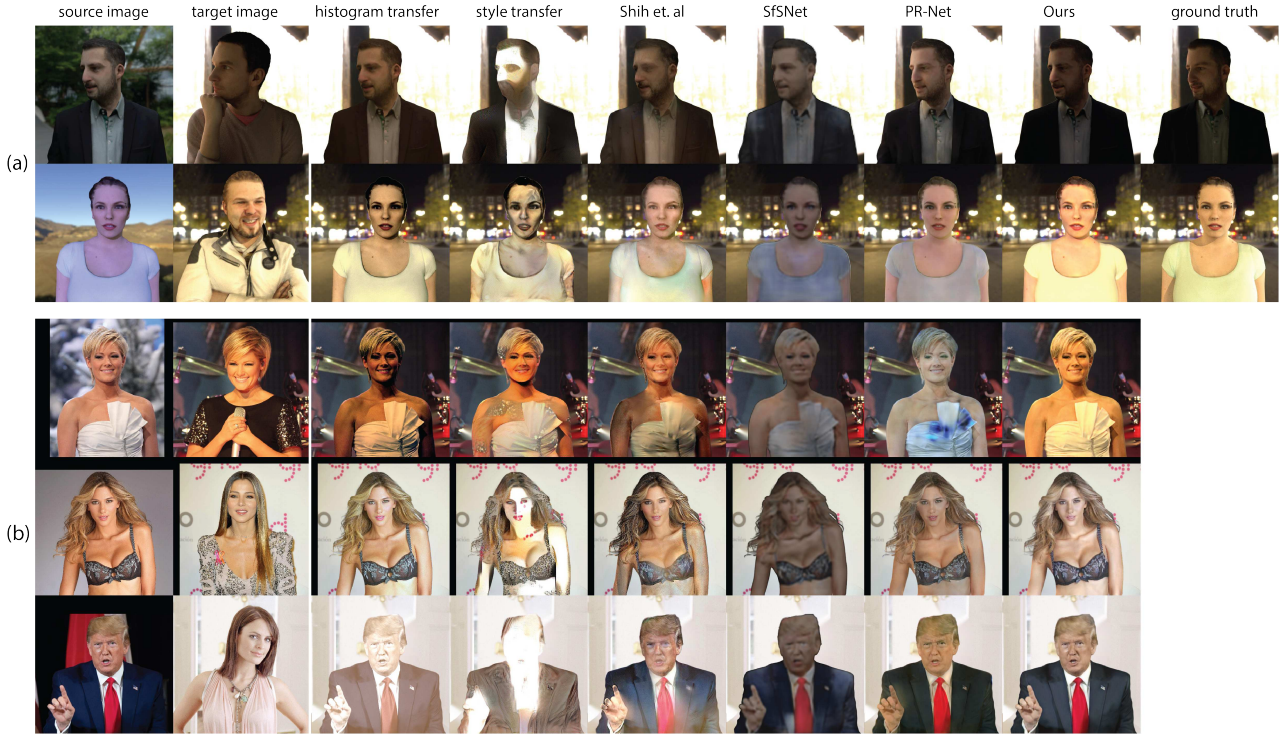
Figure 6: Qualitative comparisons of our method with two classical methods: histogram transfer [XM06] and style transfer [GEB15], and three state-of-the-art relighting techniques: Shih et. al [SPB*14], SfSNet [SKCJ18], PR-Net [SBT*19] on our synthetic test set in (a) and real images from Celeb A in (b). Note that for real images, there is no ground truth relighted image.
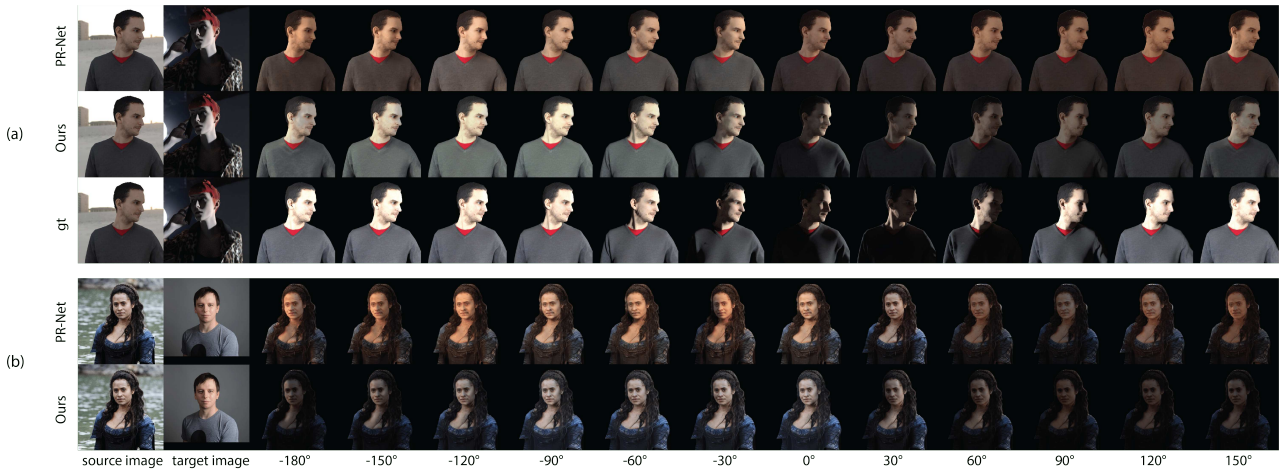

Figure 7: Visual comparison of our method and PR-Net [SBT*19] of lighting rotation task from $-180°$ to $150°$ on test set in (a) and real images from Celeb A in (b). For clarity, we exclude the background in the relighting.

matching was performed within the portrait regions. The multi-scale histogram transfer method of [SPB*14] separately matches local low and high frequency statistics of the target portrait regions. As for the other mentioned learning-based methods, they were all re-trained on our training dataset. For SfSNet, we first simulated having a white probe ball in each scene and approximated its second order Spherical harmonic (SH) lighting vectors, which were used in the training. Relighting was performed by rendering, which combines the normals and albedo estimated by the model from the source image, with the estimated SH weights from the target image. Because this model only supports 128×128 resolution input, we upsampled the output images to 512×512. For PR-Net, since their dataset and code were not publicly available, we implemented their Unet-based PR-Net using the settings in their paper. To repro-

Figure 8: Visual comparisons of our method and the 3D face approach of Shu et al. [SHS*17].

duce their method, we further downsampled 1K environment maps to 32x16 and simulated corresponding solid angle maps in training. Relighting was performed by combining the inferred environment map from the target image and the encoded features from the source image. For fair comparison, each method was applied to relight only the person in the image, which was then composited with the inpainted background.

From Fig. 6, it can be seen that our method successfully relighted subjects with perceptually correct lighting from directional illumination and ambient lighting. The histogram transfer method (column 3) is unable to reproduce geometry-dependent side lighting effects (row 1), and is easily affected by differently colored clothing (rows 3 & 5). The deep style transfer method (column 4) does not correctly disentangle illumination and portrait content, leading to obvious artifacts. The results of Shih et al. [SPB*14] suggest that multi-scale histogram transfer is not appropriate for lighting transfer when two images do not share common appearance features. The performance of SfSNet heavily depended on its intrinsic decomposition results (i.e. albedo, normal and SH lighting), which can lead to poor relighting results if the decomposition is not accurate. Moreover, the second-order SH used in SfSNet to model lighting did not model complex scene lighting well, as expected. PR-Net [SBT*19] did not sufficiently relight images, with source illumination appearing to remain the same. This might be due to their U-Net structure, which kept source illumination during image generation. Also, their model struggled when presented with clothing containing highly saturated colors.

Figure 9: Visual comparison of three-way overcomplete tensorial representation(OT3) with single tensorial representation.

4.1.3. Quantitative comparison

Table 1 compares quantitative results of our method to existing methods. Evaluation was done with the three metrics of RMSE,

Algorithm	RMSE	PSNR	SSIM
histogram transfer [XM06]	0.196	15.23	0.85
style transfer [GEB15]	0.265	11.92	0.80
Shih et. al [SPB*14]	0.181	15.76	0.87
SfSNet [SKCJ18]	0.170	16.15	0.86
PR-Net [SBT*19]	0.117	19.50	0.88
Ours	**0.099**	**21.08**	**0.90**

(a) Single image relighting result.

Algorithm	RMSE	PSNR	SSIM
PR-Net [SBT*19]	0.123	19.15	0.88
Ours	**0.105**	**20.57**	**0.89**

(b) Sequential relighting result.

Table 1: Quantitative results for comparison. We measure root mean square error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) on the human regions of about 24k relighted testing images w.r.t. the ground truth. Note that histogram transfer [XM06], style transfer [GEB15], Shih et al. [SPB*14] and SfSNet [HLBK18] cannot be used to generate sequential output.

Figure 10: Visual comparison of concatenation-based (Concat), multiplication-based (Concat) vs multiplicative neural render (MNR), both in vector latent space.

PSNR and SSIM. For each of the 24,480 images in the test dataset, we randomly chose another image from the entire dataset as the target lighting image. It can be seen that for the relighting results, our method achieved the best performance over all metrics, particularly improving the existing best PSNR result of 19.50 by 1.58 dB. We also compared our method with PR-Net on the sequential outputs by sequentially relighting the same test dataset with multiple $30°$ lighting rotation offsets. Specifically, we first extracted lighting codes $\{l^{90}, l^{0}, l^{-90}\}$ from a target image, and then inversed the fully connected layer for l^{-90} to estimate pseudo l^{-180} anchor lighting code. This enables piecewise linear interpolation to get lighting codes for intermediate angles across the full $360°$ range. We measured the average quantitative results in the three metrics under $\{-180°, -150°, .., 120°, 150°\}$, in total 12 lighting rotations for each test image. From Table 1 it can be seen that our method

Measurement	Concat	Mul	MNR(ours)
RMSE	0.110	0.108	**0.105**
PSNR	20.21	20.19	**20.51**

Table 2: Quantitative comparisons of three different neural render approaches on our test dataset.

Measurement	w/o BG	w/o OT3	w/o L_{feat}	w/o L_{cons}	Full
RMSE	0.104	0.105	0.178	0.105	**0.099**
PSNR	20.80	20.51	16.24	20.46	**21.08**

Table 3: *Quantitative results where different components are ablated.*

outperformed PR-Net, improving PSNR by 1.42 dB for sequential relighting.

4.1.4. Comparison on lighting rotation

Fig. 7 shows lighting-rotated relighted images of our method and PR-Net [SBT*19]. Accurately estimating the environment map from a single image is very ill-posed and may be an overkill for this problem. In contrast, our method estimates anchor lighting representations and interpolates lighting codes in the latent space, which lead to more reasonable lighting rotation effects with smooth transitions.

4.1.5. Comparison to the 3D face approach of [SHS*17]

We also compared our method with the 3D face approach of Shu et al. [SHS*17] for single-image relighting. Fig. 8 shows the visual comparisons on two real-image examples. Shu et al. first estimated 3DMM parameters for the face, and then applied mass transport to the local histogram. Since their approach is mainly based on 3DMM, it is restricted to relighting facial regions and cannot handle the rest of the body. In contrast, our method can relight both the face and upper body. Furthermore, taking observations from a larger region of the image would as expected lead to better inferred lighting. Note that we do not report the quantitative results of [SHS*17] in Table 1 since it only relights facial regions. In addition, [SHS*17] took 228s to handle each input image pair due to their optimization-based approach, whereas our method only took 0.3s during test time.

4.2. Ablation Studies

4.2.1. Overcomplete Lighting Representation

We verified the effectiveness of our three-way overcomplete tensorial representation (OT3) by replacing it with a single tensorial representation. After doing so, we re-trained the network on our relighting dataset. The visual results are shown in Fig. 9.

It can be seen that the standard representation with a single tensor is only able to generate mild shading when given the target side lighting, and is far from the ground truth. The is likely because the latent space is not structurally regularized enough to be consistent with the underlying illumination, and has reduced generalizing power. Conversely, the added structural redundancy of OT3 led to a result that is much closer to the ground truth. The quantitative results in Table 3 further demonstrate the advantage of our OT3 representation.

4.2.2. Multiplicative Neural Rendering

We verified the effectiveness of our multiplicative neural rendering (MNR) by comparing it with a simple concatenation approach [MGK*19] (denoted "Concat") and directly latent multiplication approach(denoted "Mul"). For clarity, we excluded the use of OT3 codes in this case, and only compared Concat, Mul and MNR trained on a single-vector latent-space representation. For the Concat, Mul and MNR configurations, the subject and the lighting encoder architectures remain the same, but Concat/Mul directly concatenates or multiply the illumination features to subject features as input to the neural render network. The visual results are shown in Fig. 10, where Concat and Mul produce limited relighting results, while our MNR's output appears perceptually more realistic.

In addition, we performed an quantitatively evaluation to compare our MNR with the conventional Concat/Mul method. The result in Table 2 shows our MNR consistently outperforms Concat and Mul.

4.2.3. Quantitative ablation results

We additionally performed an ablation study of our own model to quantitatively evaluate the individual model components and the introduced additional loss functions: 1) We used only one lighting encoder to extract 8-dim illumination features from the foreground without the background (w/o BG), to determine the importance of using the portrait background in estimating the illumination, noting that the lighting that affects the portrait predominantly comes from *behind the camera*. 2) We used only one lighting vector code instead of a 3-code tensor (w/o OT3), as per Sec. 4.2.1. 3) We removed the L_{feat} feature cycle consistency loss during training that would have provided the incentive for the subject encoder and multiplicative neural render networks to be cyclically consistent, when the rendered output is recycled as input. 4) We removed the L_{cons} latent lighting consistency loss during training that would have encouraged the OT3 codes to correctly overlap for $\pm90°$ lighting-rotated input target images. We retrained the network for each scenario on our training dataset and evaluate on our test dataset. From the results in Table 3, we can see that all the proposed components and loss functions contributed to performance improvement. Among these, the feature cycle consistency loss for lighting and content disentanglement appeared to be the most significant for relighting quality.

5. Concluding Discussion

We presented an image-based deep generative model that can dynamically relight half-body portrait images. Key technical contributions include the introduced OT3 lighting representation, the multiplicative neural rendering and the separation of background and foreground for illumination feature encoding. We have also created a large rendered dataset with annotated and controlled lighting that is suitable for training our model, and with sufficient photorealism to allow our model to be directly applied to real images. Extensive qualitative and quantitative results have demonstrated the superior performance of our proposed method on both synthetic and real images.

source image   target image   relighted image

Figure 11: *Examples of failure cases due to preserve hard shadowing (neck) and sharp specularities (glasses).*

Our method has some limitations. Fig. 11 shows two typical failure cases for our model. These input images contain hard shadowing or sharp specularities. A possible solution is to involve direct reasoning in 3D to handle geometric-sensitive shadows and specular reflections. Another possible solution is to augment our training data for these under-represented scenarios.

References

[3D] 3D Scan Store. https://www.3dscanstore.com. 3

[Arn] Arnold Renderer. https://www.arnoldrenderer.com. 3

[BM16] BARRON J. T., MALIK J.: Intrinsic scene properties from a single RGB-D image. *IEEE Trans. Pattern Anal. Mach. Intell.* (2016). 2

[DHT*00] DEBEVEC P., HAWKINS T., TCHOU C., DUIKER H.-P., SAROKIN W., SAGAR M.: Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques* (New York, NY, USA, 2000), SIGGRAPH '00, ACM. 2

[DRC*15] DUCHÊNE S., RIANT C., CHAURASIA G., MORENO J. L., LAFFONT P.-Y., POPOV S., BOUSSEAU A., DRETTAKIS G.: Multiview intrinsic images of outdoors scenes with an application to relighting. *ACM Trans. Graph.* (2015). 2

[DYWG17] DONG H., YU S., WU C., GUO Y.: Semantic image synthesis via adversarial learning. In *Proceedings of International Conference on Computer Vision (ICCV)* (2017). 3

[GEB15] GATYS L. A., ECKER A. S., BETHGE M.: A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576* (2015). 6, 7, 8

[GFT*11] GHOSH A., FYFFE G., TUNWATTANAPONG B., BUSCH J., YU X., DEBEVEC P.: Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.* (2011). 1, 2

[Gob] Gobotree. https://www.gobotree.com/cat/3d-people. 3

[GTB*13] GRAHAM P., TUNWATTANAPONG B., BUSCH J., YU X., JONES A., DEBEVEC P., GHOSH A.: Measurement-based synthesis of facial microgeometry. In *Computer Graphics Forum* (2013). 1, 2

[GZC*17] GUO Y., ZHANG J., CAI J., JIANG B., ZHENG J.: Photo-realistic face images synthesis for learning-based fine-scale 3D face reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.* (08 2017). 2

[HDR] HDRI Haven. https://hdrihaven.com. 3

[HLBK18] HUANG X., LIU M.-Y., BELONGIE S., KAUTZ J.: Multi-modal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 172–189. 3, 6, 8

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016), pp. 770–778. 6

[IRWM17] INNAMORATI C., RITSCHEL T., WEYRICH T., MITRA N. J.: Decomposing single images for layered photo retouching. *Comput. Graph. Forum 36*, 4 (July 2017), 15–25. URL: https://doi.org/10.1111/cgf.13220, doi:10.1111/cgf.13220. 2

[IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on* (2017). 3

[KW14] KINGMA D. P., WELLING M.: Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)* (2014), editor, (Ed.). 6

[LCY*17] LIU G., CEYLAN D., YUMER E., YANG J., LIEN J.: Material editing using a physically based rendering network. In *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), pp. 2280–2288. doi:10.1109/ICCV.2017.248. 2

[LLWT15] LIU Z., LUO P., WANG X., TANG X.: Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)* (December 2015). 6

[LMF*19] LEGENDRE C., MA W.-C., FYFFE G., FLYNN J., CHARBONNEL L., BUSCH J., DEBEVEC P.: DeepLight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 5918–5928. 4

[LTH*17] LEDIG C., THEIS L., HUSZÁR F., CABALLERO J., CUNNINGHAM A., ACOSTA A., AITKEN A., TEJANI A., TOTZ J., WANG Z., SHI W.: Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 4681–4690. 3

[MGK*19] MESHRY M., GOLDMAN D. B., KHAMIS S., HOPPE H., PANDEY R., SNAVELY N., MARTIN-BRUALLA R.: Neural rerendering in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 6878–6887. 3, 4, 9

[PGC*17] PASZKE A., GROSS S., CHINTALA S., CHANAN G., YANG E., DEVITO Z., LIN Z., DESMAISON A., ANTIGA L., LERER A.: Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop* (2017). 6

[PKA*09] PAYSAN P., KNOTHE R., AMBERG B., ROMDHANI S., VETTER T.: A 3D face model for pose and illumination invariant face recognition. In *IEEE International Conference on Advanced Video and Signal Based Surveillance* (2009). 2

[PKD*16] PATHAK D., KRAHENBUHL P., DONAHUE J., DARRELL T., EFROS A. A.: Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition (CVPR)* (2016). 3

[rem] remove.bg. https://www.remove.bg/upload. 6

[SBT*19] SUN T., BARRON J. T., TSAI Y.-T., XU Z., YU X., FYFFE G., RHEMANN C., BUSCH J., DEBEVEC P., RAMAMOORTHI R.: Single image portrait relighting. *ACM Transactions on Graphics (TOG) 38*, 4 (2019), 79. 2, 6, 7, 8, 9

[SGK*19] SENGUPTA S., GU J., KIM K., LIU G., JACOBS D. W., KAUTZ J.: Neural inverse rendering of an indoor scene from a single image. In *Proceedings of International Conference on Computer Vision (ICCV)* (2019). 3

[SHS*17] SHU Z., HADAP S., SHECHTMAN E., SUNKAVALLI K., PARIS S., SAMARAS D.: Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics 37* (2017). 1, 2, 8, 9

[SKCJ18] SENGUPTA S., KANAZAWA A., CASTILLO C. D., JACOBS D. W.: SfSNet: Learning shape, reflectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 6296–6305. 2, 6, 7, 8

[SPB*14] SHIH Y. C., PARIS S., BARNES C., FREEMAN W. T., DU-
RAND F.: Style transfer for headshot portraits. *ACM Transactions on Graphics* (2014). 1, 2, 6, 7, 8

[Tel04] TELEA A.: An image inpainting technique based on the fast marching method. *journal of graphics tools* (2004). 5

[TZK*17] TEWARI A., ZOLLHÖFER M., KIM H., GARRIDO P., BERNARD F., PÉREZ P., THEOBALT C.: MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision* (2017). 2

[TZN19] THIES J., ZOLLHÖFER M., NIESSNER M.: Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics* (2019). 3

[XM06] XIAO X., MA L.: Color transfer in correlated color space. In *Proceedings of the 2006 ACM International Conference on Virtual Reality Continuum and Its Applications* (2006). 6, 7, 8

[YS19] YU Y., SMITH W. A.: InverseRenderNet: Learning single image inverse rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 3155–3164. 2

[ZCC18] ZHENG C., CHAM T.-J., CAI J.: T2Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 767–783. 3

[ZJ19] ZHOU H., JACOBS D. W.: Deep single-image portrait relighting. In *Proceedings of International Conference on Computer Vision (ICCV)* (2019). 2

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networkss. In *Computer Vision (ICCV), 2017 IEEE International Conference on* (2017). 6

[ZZP*17] ZHU J.-Y., ZHANG R., PATHAK D., DARRELL T., EFROS A. A., WANG O., SHECHTMAN E.: Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems 30* (2017), Curran Associates, Inc. 3, 6