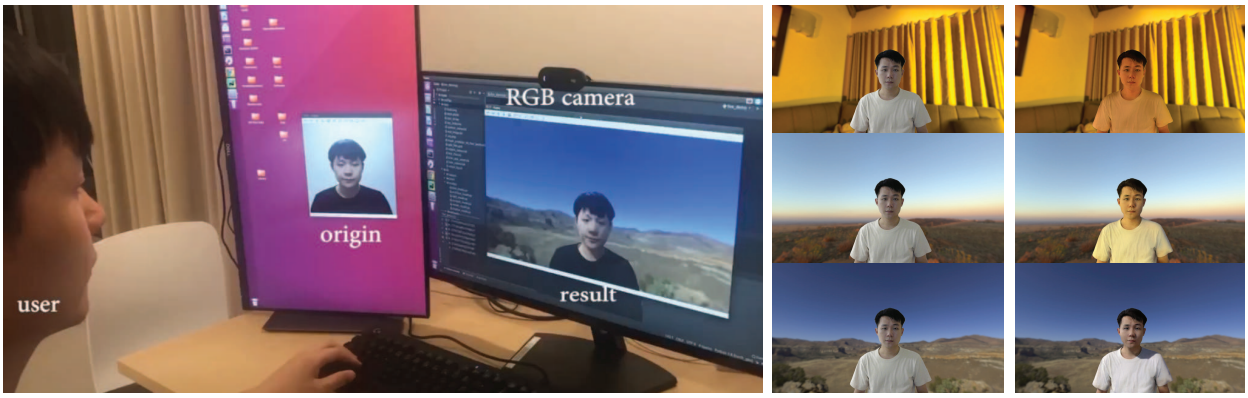
Real-time Shadow-aware Portrait Relighting in Virtual Backgrounds for Realistic Telepresence

Category: Research

Paper Type: system



(a) demo system (b) copy-paste (c) relighting result

Fig. 1. (a) We present a real-time shadow-aware relighting system from an RGB Input and a given background image. (b) Our system can realistically composite the portrait into various backgrounds with plausible illumination and shadow.

Abstract— While using virtual backgrounds has recently become a very popular feature in videoconferencing, there often exists a jarring mismatch between the lighting of the user and the illumination condition of the virtual background. The existing portrait relighting methods can alleviate the problem, but do not have the capacity to deal with difficult shadow effects. In this paper, we present a new shadow-aware portrait relighting system that can relight an input portrait to be consistent with a given desired background image with shadow effects. Our system consists of four major components: portrait neutralization, illumination estimation, shadow generation and hierarchical neural rendering, which are all based deep neural networks, and the whole system is end-to-end trainable. In addition, we created a large-scale photorealistic synthetic dataset with shadow, illumination and depth annotations for training, which allows our model to generalize well to real images. The extensive experiments demonstrate that our shadow-aware relight system outperforms the state-of-the-art portrait relighting solutions in terms of producing more lighting-consistent relighted images with shadow effects.

Index Terms—Shadow Generation, Portrait Relighting, Neural Rendering

1 Introduction

One consequence of the 2020 pandemic has been the sudden increased use of videoconferencing platforms, such as Microsoft Teams and Zoom. A common feature that has been heavily used is that of a virtual background. Although serving its purpose in privacy protection, the direct insertion of the user into the background has often led to jarring mismatches in lighting between the user and virtual background.

Here we address the problem of relighting a portrait image, such that the user appears more realistically embedded in a given virtual background (see Fig. 1). Although portrait relighting alone has been previously studied, harmonizing the portrait's lighting to that of a partial scene image in a real-time system has not been adequately solved. Such a solution will not only be applicable to the popular videoconferencing platforms, but also contribute to realistic telepresence in AR/VR systems.

Such a task involves multiple challenges. For example, we need to handle backgrounds that do not contain people or objects that are proxy light probes, as needed in reference-based methods [25, 36] that attempt to extract illumination from subjects present in target lighting scenes. In addition, portraits with visible upper bodies, clothing and significant hair will be difficult for methods that employ parametric models of faces and bodies [13, 37, 40].

More recently, Kanamori et al. [20] employed an intrinsic image decomposing method SfSNet [34] to infer albedo, shape, and illumination from a human portrait, but this model requires the assumption of Lambertian reflectance and a low-order spherical harmonic (SH) illumination model. Sun et al. [39] employed a training dataset obtained using a sophisticated light stage, and further required a target (panoramic) environment map for relighting, both of which are fairly difficult to obtain. While these methods were able to relight the portrait with reasonable soft shading based on the target environment map, they did not have the capacity to generate the harder, more distinct shadows under sunlight and other uneven lighting conditions.

Inspired by the recent success in face normalization [27], illumination estimation [8] and neural rendering [41], we adopt a divide-and-conquer strategy and propose the first end-to-end trainable pipeline for shadow-aware portrait relighting to suit a given target background. In our approach, we first relight the input source portrait under a canonical evenly-lit condition, and also estimate the portrait depth map based on a generative adversarial model. Next we predict a low-resolution panoramic environment map from the input target background, leveraging on a large-scale training dataset. This environment map, in conjunction with the estimated portrait depth map, can be used to generate an intermediate shadow map that includes hard (umbral) shadows. Finally, a novel hierarchical neural render is used that is able to effectively combine global illumination and local hard shadow details, leading to improved results with short inference time.

For end-to-end training of our proposed neural relighting model, we need a large number of portrait images with multiple known scene illumination conditions, with shadow and depth annotations. To the best of our knowledge, there is no such publicly available dataset. To physically capture such dataset without expensive specialized hardware

setups would be onerous, requiring subjects to remain perfectly still for impractically long periods of time, as well as requiring extensive manual effort for shadow annotations. Instead, we commercially obtained high-quality 3D scans of humans with reflectances and real high dynamic range (HDR) environment maps, and subsequently rendered a large number of photorealistic images to create a synthetic dataset with depth, illumination and shadow annotations.

The main contributions of this paper are:

- The first end-to-end deep solution for immersive real-time relighting of human portraits to virtual backgrounds with shadow projection, which have the potential to be easily integrated into existing video meeting and 3D telepresence platforms[1].
- An effective four-stage relighting pipeline that consists of portrait neutralization, illumination estimation, shadow generation and hierarchical neural rendering to enable shadow-aware relighting effect. In particular, the shadow generation and hierarchical neural rendering components are novel and carefully designed to incorporate domain knowledge. The four components are also integrated in a well-considered manner, based on our insights into the problem.
- We constructed a large scale synthetic dataset and conducted extensive experiments that show our system outperforming other state-of-the-art approaches.

2 Related Work

We review some existing research that is relevant to the photorealistic portrait relighting problem addressed in this paper.

Detailed 3D Scanning. In the special effects industry, it is often desirable to portray specific actors in different environments and lighting conditions in which they were filmed, or even to animate them differently [35]. Technological advancement has enabled humans to be 3D scanned in high detail, recovering both micro-geometry and reflectances, typically in specialized setups such as the Light Stage [4, 10, 12]. This is however an expensive and high effort undertaking, and the retargeting capability will only apply to the specific individuals scanned.

Illumination Estimation. Illumination estimation is an important element for virtually relighting objects. The conventional illumination representation in rendering is the environment map, often in the form of a High Dynamic Range (HDR) panoramic image. These images are typically captured by photographing a mirrored sphere, or stitching together wide-angle views with multiple exposures [38]. There are also some recent learning-based illumination estimation methods. For example, Gardner et al. [8] estimated indoor illumination from a single image using a neural network. Due to lack of a comprehensive HDR dataset, their approach was first to train their network on images extracted from a large set of partial scene images in which light sources were detected, and then fine-tuned on HDR panoramas from a small dataset. Similarly, Cheng et al. [22] captured a large HDR dataset from the front and back cameras of a mobile device, and proposed estimating low-dimensional SH coefficients given two phone images. To cater for spatially-varying lighting, Garon et al. [9] refined an illumination inference network using both local and global branches to regress spatially-aware illumination SH coefficients. Recently, Gardner et al. [7] presented a method to estimate lighting from a single image of an indoor scene using a representation comprising 3 parametric lights. In this work, we created a panorama dataset to be used in conjunction with the portrait relighting dataset. Instead of using low-dimensional SH lighting representations or sparse parametric lights, we directly infer a low-resolution environment map from the input background scene image.

[1] We will release our rendering code and implementation code on publication.

Shadow Generation. Authentic shadow generation consistent with the environment is important for enhancing the sense of immersion. While physics-based renderers automatically achieve this, this is usually not feasible for more general, real-time use. Instead, scalar shadow maps are used to describe lighting occlusion as observed from the camera perspective, with illumination often simplified to be directional or point lights. However, this simpler approach results in several noticeable artifacts, particularly if the geometry and scene lighting are not accurately modeled. Conversely, there are several works [15,23,45] using adversarial deep networks to remove or generate shadows of objects in an image. However, these works are usually limited to only on-the-ground shadows and not on complex surfaces, such as shadows of a person's head that fall on shoulders.

Portrait Relighting. Face relighting on photographs has been explored widely in computer vision and graphics. The histogram-based method of [36] requires both input and reference pictures to have compatible appearance attributes, such as beards and skin color, and transfers local contrasts and overall lighting from one portrait to another. However, it would generate visible artifacts for those without similar appearance features. The geometry-aware relighting method of [37] is able to generate more robust color remapping, but is restricted to facial regions as it is based on 3DMMs (3D morphable models).

Separately, intrinsic image decomposing methods aim to recover geometry, reflectance and lighting from images, which has since been applied to faces [3, 6, 43]. Recently, many deep-learning methods, such as Mofa [40], SfSNet [34] and [13], leveraging on large internet-scale datasets, have been proposed. They would 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. However, the performance of these methods is limited by the dependence on low-frequency illumination models and low-dimensional 3DMM face models. Furthermore, relighting can only be applied to facial regions and cannot trivially be extended to portraits with visible upper bodies and clothing.

For more general photographs, the single image relighting method of Zhou et al. [47] warps 3DMM normals to head regions with an as-rigid-as-possible mapping. They synthesized a large facial relighting dataset using varied SH lighting and proposed a deep single image portrait relighting method. However, since their method and dataset only relight the L-channel in Lab color space, it will not relight images with more general RGB illumination. Sun et al. [39] collected a 'one-light-at-a-time' (OLAT) real human portrait dataset and trained a U-Net-based PR-Net to directly estimate the environment map, which can then be used to relight a single human portrait. However, collecting such a dataset requires using a Light Stage and involves an intensive, finely-calibrated process. Furthermore, their method involves generating an intermediate dataset with smoother illuminations, and does not have much capacity to generate harder shadows.

Deep Image Translation. The seminal work for deep image translation is pix2pix [18], which translates an image from one domain to another using paired training data. This approach has since been applied to various tasks, such as inpainting [30], semantic labeling [5] and super resolution [21]. It was extended to unpaired translation [46] and for multimodal output [17, 48]. Recently, deep image translation has also been used for neural rendering. Thies et al. [41] proposed a deferred neural rendering method to synthesize images from imperfect 3D content, Meshry et al. [26] extracted appearance vectors from a deep buffer and injected the vectors into a latent space for neural rendering in-the-wild architectural images, while Sengupta et al. [33] used inverse rendering and a residual appearance render network for scene images.

For faces, Nagano et al. [27] applied several deep generative networks to normalize face illumination based on SH lighting, and also to neutralize facial expression and pose. Olszewski et al. [28] introduced a coarse-to-fine generative model to synthesize facial hair. Inspired by these works, we adopt several generative models in our approach as presented in the next section.

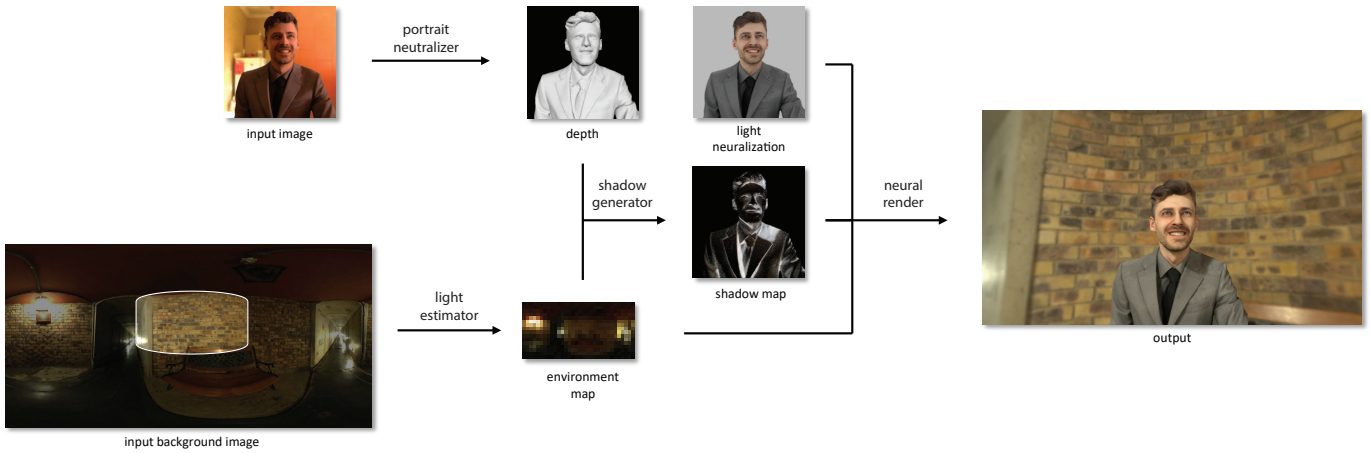


Fig. 2. An overview of our portrait relighting system.

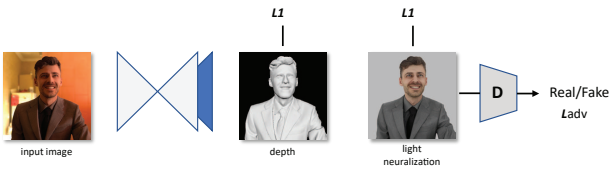


Fig. 3. Portrait neutralization module.

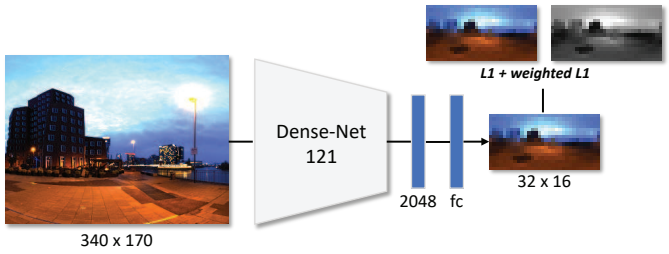


Fig. 4. Light estimation module.

3 Method

Fig. 2 presents a high-level overview of our method. It comprises multiple deep neural network (DNN) components that address the different tasks of portrait lighting neutralization, depth estimation, environment map estimation, shadow map generation, and portrait neural relighting. There are two unique benefits of breaking down the entire portrait relighting problem into components: 1) this enables us to generate hard geometric shadows, which is much more difficult with a single encoder-decoder structure [39, 47], and 2) compartmentalization allows each component DNN to be pre-trained separately prior to joint end-to-end fine-tuning, leading to faster and more reliable convergence.

Given a portrait image and a background scene image, our goal is to harmonize the portrait's lighting to that of the background scene image for realistic composition. We first conduct portrait lighting neutralization, i.e. we relight the portrait image under a canonical evenly-lit condition, while also estimating the portrait's depth map via a generative adversarial network (GAN). In parallel, we estimate the panoramic environment map corresponding to the input background image. Next, the predicted depth map is combined with the environment map to produce a shadow map, also via a DNN. Finally, the canonical portrait image, shadow map and environment map are put through a hierarchical neural generative render DNN to produce the intended relighted portrait, which is then composited with the background image to form the final output image.

3.1 Portrait Lighting Neutralization

Relighting work often involves separating the attributes of geometry, reflectances and lighting, typically attempted via inverse rendering or intrinsic image decomposition. However, this is an ill-posed problem, with errors in one attribute causing errors in the other coupled attributes.

In our approach, while we estimate the portrait geometry in terms of a depth map, we avoid further estimating a reflectance or albedo map (such as in [34]). The errors in estimating the fine-grained geometry of folds and wrinkles in skin, clothing and hair will introduce closely-coupled errors when estimating reflectances, which often leads to distracting artifacts in the final relighted image. Instead, we undertake a step we call portrait lighting neutralization, where we estimate the portrait's appearance in a canonical illumination condition, defined simply as a white spherical environment map surrounding the subject. Here strong directional shadows are minimized, but most importantly we can *bypass* attempting to separate shading due to finely detailed geometry from that of albedo and material reflectances. The shading that is present in this canonical illumination, partly akin to ambient occlusion, will generally remain under different lighting conditions, and there is no need to decompose this further (e.g. to get intrinsic images).

The lighting neutralization generator network $\mathscr{G}_n$ is an encoder-decoder architecture, shown in Fig. 3. We utilize skip connections [32] to maintain correspondence of high-frequency details between the original input image I and the neutralized image I_n during image decoding process. Reusing the above encoder, we also attached a second decoder without skip connections for predicting the depth map d. Mathematically we have:

$$(d, I_n) = \mathscr{G}_n(I). \tag{1}$$

We train this network $\mathscr{G}_n$ using L1 loss to the ground-truth image/depth within the foreground regions and also a multi-scale adversarial loss adapted from [42]. The final training loss L_n is thus:

$$L_n = L_1(d, \hat{d}\,) + L_1(I_n, \hat{I}_n) + w_{adv} L_{adv}(I_n, \hat{I}_n) \tag{2}$$

where $\hat{d}$, $\hat{I}_n$ are corresponding ground truth, while L_{adv} is the multi-scale adversarial loss to minimize local and global artifacts:

$$L_{adv}(I_n, \hat{I}_n) = \sum_k \Big(\mathbb{E}_{\hat{I}_n}[log D_k(\hat{I}_n)] + \mathbb{E}_{I_n}[log(1 - D_k(I_n))] \Big) \tag{3}$$

where D_k are discriminators on different levels (we set k=3). The relative weighting of these losses is determined by w_{adv} (we set to 0.1).

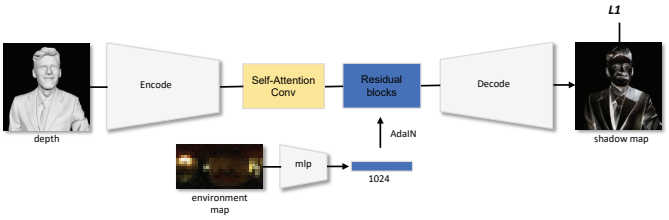


Fig. 5. Shadow generation module.

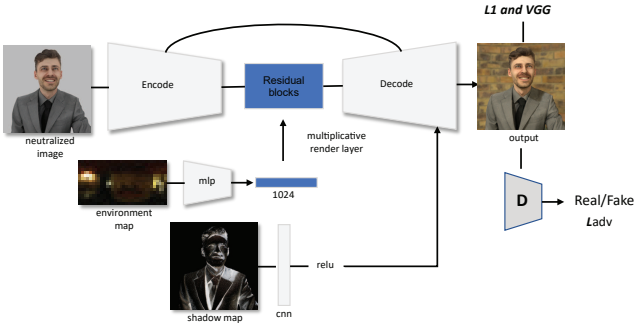


Fig. 6. Neural relighting module.

3.2 Background Environment Map Estimation

Virtual backgrounds are often standard images that come with the videoconferencing platform, or provided by the user. The environment illumination outside the camera field of view is not directly observed, which poses a problem for relighting, since it is the illumination *behind* the camera that primarily affects the subject's appearance.

To address this issue, we utilize a DNN to estimate a plausible panoramic environment map from the input background image. This environment map is very low resolution because it is not feasible to extrapolate a detailed 360° scene outside the camera's field of view, but is reasonable for relighting predominantly non-reflective surfaces.

Given a partial scene image I_s with roughly $120^\circ \times 60^\circ$ field of view, we estimate the panoramic environment map I_l with 32×16 resolution to represent the illumination conditions in the scene. This estimation is done via a DNN that has learned the illumination prior from a scene dataset, with the DNN architecture shown in Fig. 4. Specifically, the input image is passed through a headless DenseNet-121 network to produce a 2048-dimensional latent vector, then forwarded to another fully-connected layer to decode to the compressed environment map.

The goal and structure bear some similarities to [7], but there are two key differences. First, we represent scene illumination conditions using an environment map, rather than sparsely with only 3 parametric lights. Doing so allows more accurate illumination and shadow map estimation, leading to better relighting results. It also does not suffer from unstable gradient flow during training, as encountered in [7]. The second key difference is that we leverage a weighted L1 loss to increase the importance of brighter regions in the environment map, as they contribute more to the relighting. A w_l weight map is precomputed from the groundtruth environment map by using the lightness component in LAB color space, which helps in improving the accuracy of relighting. The final training loss L_l for background environment map estimation is thus:

$$L_l = (1 + w_l) L_1(I_l, \hat{I}_l), \tag{4}$$

where I_l, $\hat{I}_l$ are the output environment map and corresponding ground truth, respectively.

3.3 Deep Shadow Map Generation

We previously observed that while neural rendering networks [34, 39] are good at relighting portraits with consistent shading, they face difficulty in generating shadows, particularly the ones with sharper and more geometric boundaries. Inspired by techniques that use shadow maps in conventional graphics rendering, we likewise formulate our framework with a shadow map generation module prior to relighting. However, we forego the use of conventional graphics libraries (e.g. OpenGL), and instead use a DNN here as well in order to keep the component differentiable, allowing for end-to-end training. This helps the shadow map generation leverage shape priors learned from the training data and be less sensitive to errors in depth map estimation.

Our shadow map generation network $\mathscr{G}_s$ is shown in Fig. 5, extending the encoder-decoder architecture of [17]. We equip the bottleneck residual blocks with Adaptive Instance Normalization (AdaIN) [16] layers, whose parameters are generated by a multilayer perceptron (MLP) from the environment map. This design requires fewer network parameters and leads to better results than naive concatenation. In addition, we included a self-attention convolutional module [44] to better extract cross-spatial geometric features, which improves the shadow map. The shadow map is generated by the network:

$$I_s = \mathscr{G}_s(d, I_l), \tag{5}$$

and the loss L_s for the shadow generation is:

$$L_s = L_1(I_s, \hat{I}_s) \tag{6}$$

where $\hat{I}_s$ denotes the corresponding groundtruth shadow map.

3.4 Neural Relighting

The last step in the pipeline is to generate the final relighted image, and here we also adopt a neural rendering approach. From our investigations, a simple concatenation of the intermediate maps as input into an encoder-decoder network leads to poor results with loss of detail. Therefore, we carefully designed the U-Net based generative network $\mathscr{G}_r$ shown in Fig. 6.

Considering that lighting effects due to the environment map are more global in nature, we first encode the environment map into a latent lighting code via an MLP. In parallel, the neutralized image is encoded to latent features, and these features are combined with the lighting code in an interleaved manner via residual blocks. First, each lighting code vector is partitioned into a set of multiplicative and additive components, i.e. $l = (l^1_{mul}, l^1_{add}, l^2_{mul}, l^2_{add}, \ldots)$, where the superscript denotes the corresponding render layer index. We stack several render layers in each residual block. For the ith layer, the latent features are combined as:

$$layer^i = l^i_{mul} \times s^i + l^i_{add} \tag{7}$$

where s^i is the residual output from the current render layer, while the lighting-processed output $layer^i$ becomes the input for the next render layer.

For shadow effects that are more local in nature, the shadow latent features and subject features are also combined in an interleaved process, but across upsampled layers in the decoder. For the jth feature map, using the following function in element-wise multiplication:

$$feature_j = (1 - shadow_j) \times relight_j, \tag{8}$$

where $shadow_j$ denotes the j^{th} shadow feature map after processing through a CNN, $relight_j$ is the upsampled output from the current decoder layer, and $feature_j$ is the resulting shadow-processed output. The generated image I_r is finally synthesized by the network:

$$I_r = \mathscr{G}_r(I_s, I_l, I_n), \tag{9}$$

We train this network using L1 loss between the groundtruth relighted image and the synthesized output, and compute this loss only for the foreground. We also employ an adversarial loss using multi-scale discriminator [42] for the composited image. The discriminators are trained in conjunction with the generator to determine not only

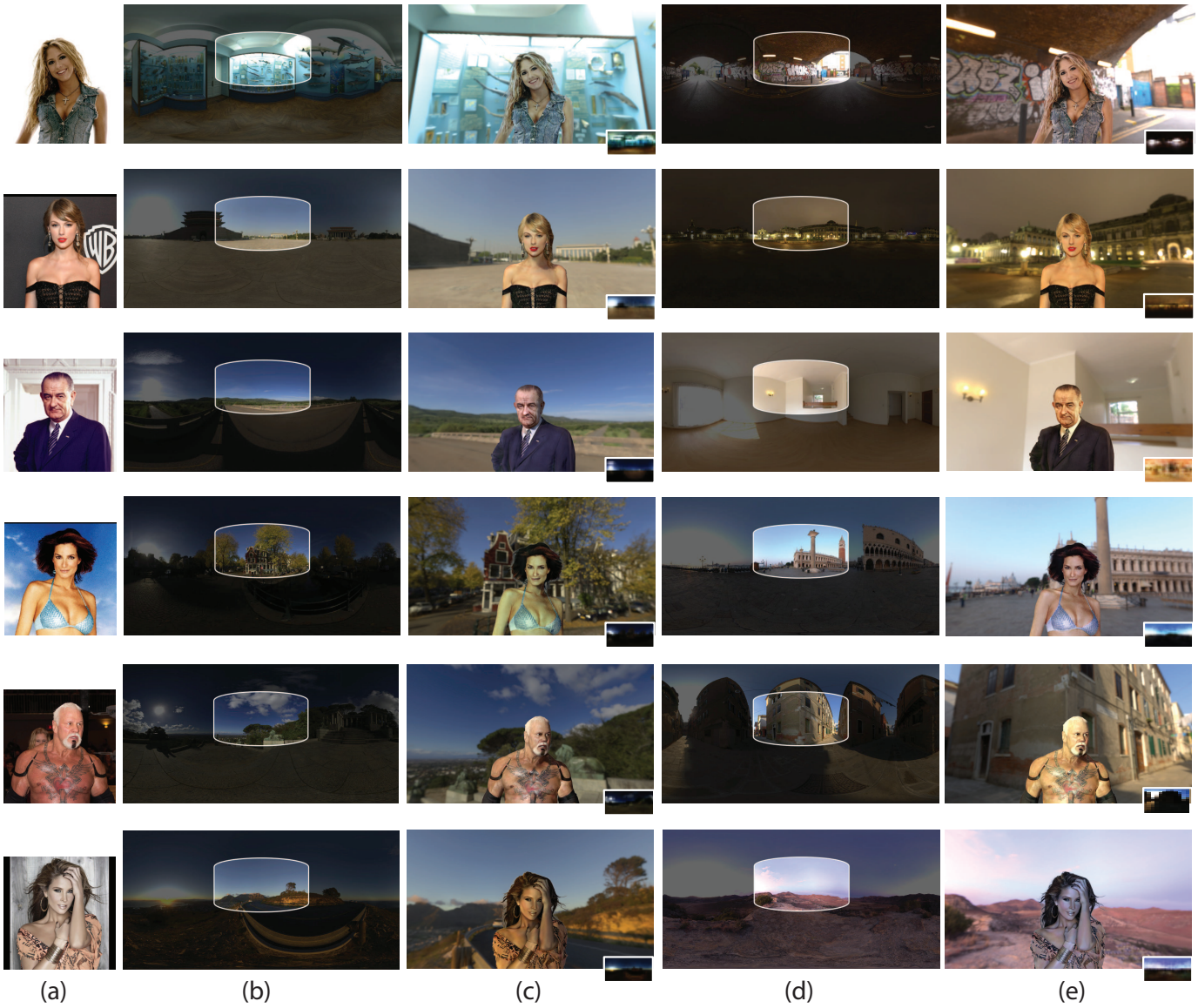


Fig. 7. Our relighting results of real portrait images from CelebA dataset [24]. (a) input images, (b)(d) input backgrounds, (c)(e) relighted images and estimated environment maps.

whether a given image is real or synthesized, but also whether it plausibly corresponds to the background. Finally, we use a perceptual loss metric L_{vgg} [19], represented using a set of higher-level feature maps extracted from a pre-existing image classification network (i.e. VGG-19). This also helps improve the quality of the relighting. The final loss L_r is thus:

$$L_r = L_1(I_r, \hat{I}_r) + w_{vgg} L_{vgg}(I_r, \hat{I}_r) + w_{adv} L_{adv}(I_r, \hat{I}_r), \qquad (10)$$

where $\hat{I}_r$ is the groundtruth relighted image, and relative weights determined by w_{vgg} and w_{adv}. We set these weights ($w_{vgg} = w_{adv} = 0.1$), such that the average gradient of each loss is at the same scale.

4 Experiments

Dataset. To create ground-truth relighting images, corresponding shadow maps and depth maps, we used 3D human data from 3D collection websites [1, 11], which include 98 commercially available high-resolution photogrammetry scans. The dataset is randomly split into training and test sets with 83 and 15 subjects, respectively. We also collected 323 real outdoor/indoor environment maps with 1k resolution from HDRI Haven [14], which is randomly split into training and test sets with 225 and 98 maps, respectively. For the neutralized portrait images, we used a uniform white environment map as the light source. A professional physically-based ray-tracing engine [2] was used to render photorealistic images, where shadow maps are automatically generated during the ray-tracing process.

Portrait figures were rendered to viewpoints which are front-facing or partly oblique. We coarsely aligned the figures via global rotation, scaling and translation, such that the 3D torsos have similar widths in images. For each environment map, we created a series of scenes in which the environment illumination was sequentially rotated with a step size of 30° degree in yaw axis while fixing elevation at 0. The size of the images is cropped into 512×512 so as to be suitable for a network input. Data augmentation was also applied via 3D rotation of up to 20° in yaw axis for each training subject to help the network exploit the geometric regularity. In total, we generated 277,956 training images under 2,700 illuminations and 17,460 testing images under 1,176 illuminations, with no other data augmentation. There is no overlap between training and test subjects as well as between training and test illuminations.

Training. The sizes of an input portrait image, an neutralized image, a depth map and a shadow map are all set to 512×512. The size of an

Fig. 8. Qualitative comparisons of our neural shadow generation and the directional light OpenGL method.

input background image to the lighting network is 256×340, sampled from a 720×1280 background image. To simplify and stabilize the training, we pretrained each module individually for 3 epochs and then jointly trained all the networks in an end-to-end manner for another 3 epochs. We used the Adam optimizer with an initial learning rate of 1.5×10^{-4} and a batch size of 4 on a modern computer with a single NVIDIA GTX 2080Ti in Pytorch [29]. For a fair comparison, all other methods were trained for 6 epochs. Our code is available online[2].

4.1 Our Results

Visual results on real images. Fig. 7 shows our relighting results on real images from the publicly available CelebA dataset [24]. In particular, we use the tool in [31] to pre-process and segment foregrounds from backgrounds. For each pair of an input human image and a target scene, we show the relighted composited image. It can be seen that our method can robustly handle a variety of real photos containing different subjects with different illumination conditions and clothing styles, and produce plausible relighting and shadow effects, although we trained the networks using only our synthetic images.

Neural shadow generation. In Fig. 8, we evaluate the effectiveness of our neural shadow map generation by comparing with the conventional shadow generation methods in common graphics libraries (e.g., OpenGL, DirectX), which usually assume point or directional light to compute the shadow map. Here we approximate 3 directional lights using Gaussian mixture model from input environment maps and generate shadow map using OpenGL. We can see in Fig. 8 that the convention method works well for hard shadows, but it fails to generate soft shadow due to limit lighting sources. On the other hand, the advance ray-tracing based shadow generation can synthesize very realistic shadows (used as ground-truth here), but it takes 30s to generate one shadow map while our neural shadow generation only needs 10 ms.

Run time. For testing on a modern computer with a single NVIDIA GTX 2080Ti, the inference time for the modules of neutralization, lighting estimation, shadow generation and neural render is 26ms, 19ms, 10ms and 5ms, respectively. Note that the lighting estimation only needs to be done once and offline. The overall relighting process at the testing stage can reach about 24.3 FPS with an image resolution of 512x512.

[2]https://github.com/anonymityvr1127/PortraitRelight.git

Table 1. Quantitative comparisons of relighting results on our synthetic test set with the ground-truth lighting for target scenes. We measure root mean square error (RMSE), peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) on the human regions of about 17k relighted testing images w.r.t. the ground truth.

Algorithm	Relighting		
	RMSE	PSNR	SSIM
SfSNet [34]	0.123	18.67	0.87
PR-Net [39]	0.090	21.95	0.90
Ours	**0.082**	**22.69**	**0.91**

4.2 Comparisons with state-of-the-arts

Fig. 9 compares the relighting results of our method and the state-of-the-art methods including SfSNet [34] and PR-Net [39], which are re-trained on our synthetic training dataset. For SfSNet, we extracted second order spherical harmonic (SH) lighting vectors by simulating having a white probe ball in each scene image, and this ground-truth SH lighting was used in training and testing. Its relighting was performed by rendering, which combines the normals and albedos estimated by SfSNet from a given source image, with the ground-truth SH weights from a given target scene. Because SfSNet only supports 128×128 resolution inputs, we upsampled its output images to 512 × 512. For PR-Net, since their dataset and code are not publicly available, we implemented their U-Net based network according to the setting of their paper, and used same ground-truth environment maps as ours. Its relighting was performed by combining the ground-truth environment map from a target scene and the encoded features from a given source image. For a fair comparison, each method was applied to only relight the person in each image, which was then composited with the target background.

From Fig. 9, it can be seen that our method successfully relights subjects with perceptually correct target illumination and shadow effect. The performance of SfSNet heavily depends 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 cannot model complex scene lighting well, as expected. PR-Net adopts U-Net structure for rendering by combining the embedded source image features and the target environment map. However, their model struggled especially when the source image contains saturated colors. Due to lack of lighting neutralization process, the source illumination is kept during the process of generating relighted images, which might lead to visible artifacts. Besides, both SfSNet and PR-Net are not able to generate harder shadow effects like ours, which limits their model expressive ability.

Table 1 gives quantitative comparisons of the relighting results of different methods on our synthetic testing dataset. Evaluation was done with the three metrics: RMSE, PSNR and SSIM. For each of the 17,460 testing images, we randomly chose a target environment map. We can see from Table 1 that our method achieves the best performance over all metrics.

4.3 Ablation analysis

In this subsection, we verify the effectiveness of individual components in our framework.

Portrait neutralization. To see the usefulness of our portrait neutralization module, we remove it from our framework and directly input the source image into the neural rendering network, while keeping all other modules unchanged. We then retrain the entire network. Fig. 10 shows some visual results, where we can see that our model without the neutralization module cannot prevent the source illumination from appearing in the relighted images.

Shadow effects. To verify the shadow generation module, we remove it from our pipeline and only use the neutralized image and the environment map to generate the relighting image. Some visual results are shown in Fig. 11. We can see that incorporating the shadow generation can clearly lead to meaningful shadow effects in the relighted images, which are consistent with the target illuminations.

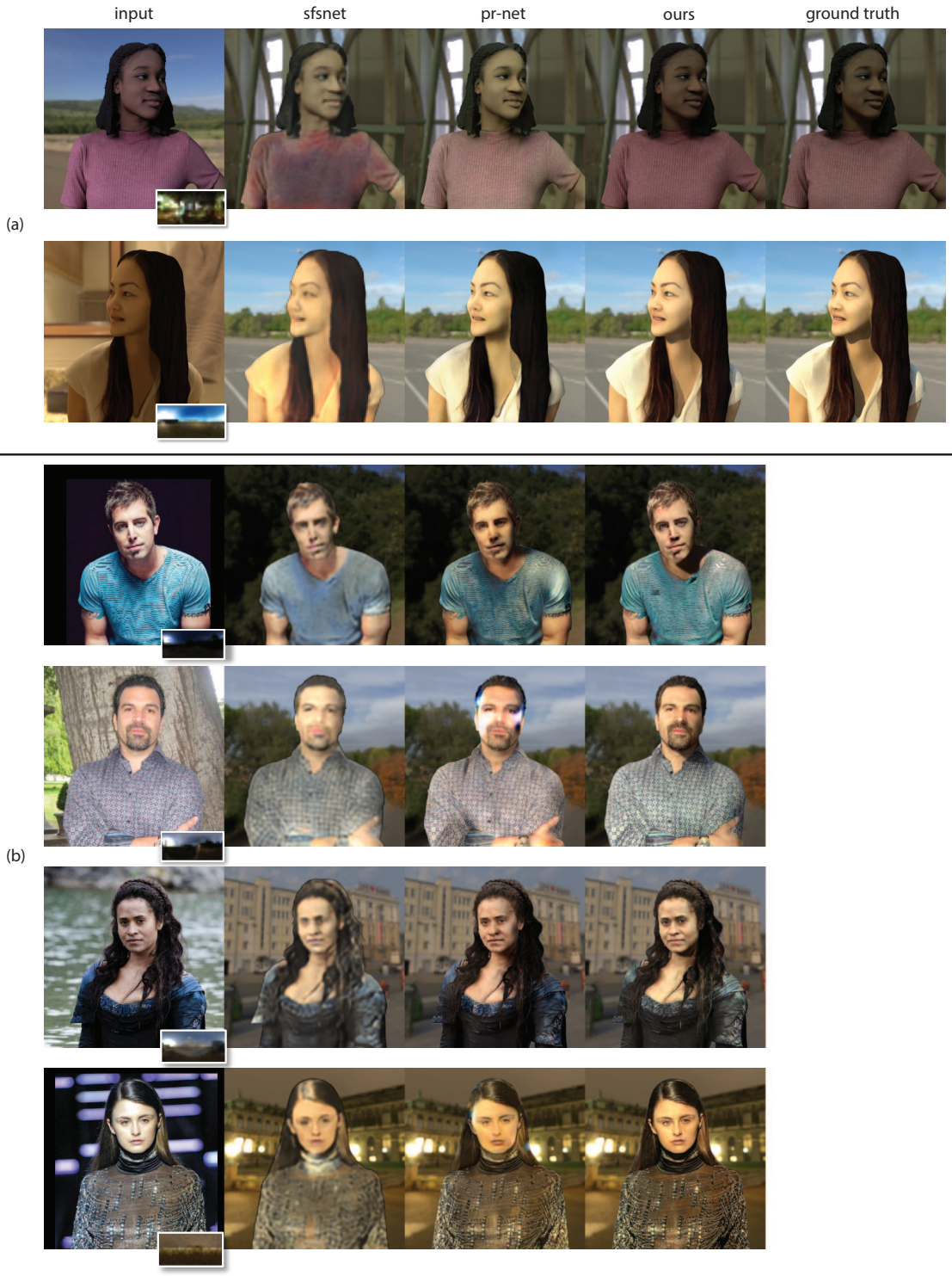


Fig. 9. Qualitative comparisons of our method with the state-of-the-art techniques: SfSNet [34], PR-Net [39]. (a): Results on our synthetic data. (b) Results on real images from Celeb A. The input includes a source portrait image and a target illumination. Note that for real images, there is no ground truth relighted image.

Table 2. Quantitative results where different components are ablated.

Measurement	-end2end	-neutral	-light	-shadow	-hierarchic	full
RMSE	0.097	0.089	0.093	0.094	0.090	**0.088**
PSNR	21.36	21.94	21.64	21.74	21.93	**22.19**

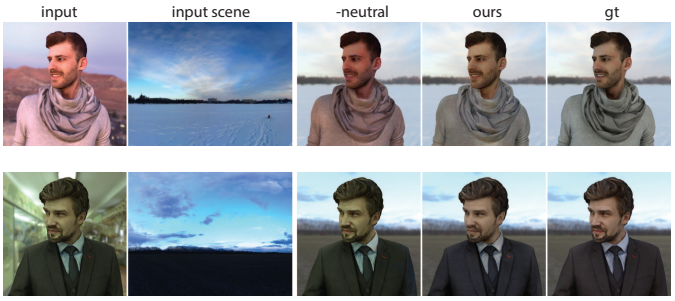


Fig. 10. Relighting results of our model without and with the portrait neutralization module for ablation study.

Fig. 11. Relighting results of our model without and with the shadow map generation module for ablation study. Generated shadow effects are highlighted in red windows.

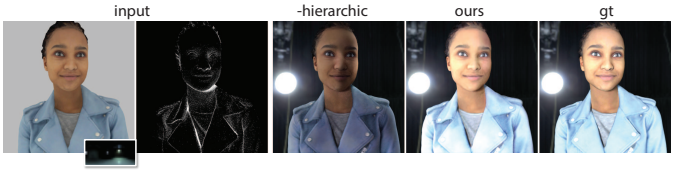


Fig. 12. Qualitative comparisons of the relighting results of a simple neural render vs our hierarchical neural render for ablation study.

Fig. 13. Image relighting results with rotated lighting. First column and first row: input portrait images and target environment maps; Second and Third column: relighted images with rotated lighting.

Hierarchical rendering. To evaluate the hierarchical rendering, we replace it with a simple MLP-based neural render that directly takes the shadow map, the neutralized image, and the target environment map as the input. Fig. 12 shows the ablation study results, where we can clearly see that the hierarchical rendering module results in a much better relighted image, perceptually closer to the ground truth.

Quantitative ablation results. Table 2 gives the quantitative results for the ablation study. We consider the following variants of our model: 1) '-end2end': our model without end-to-end training, i.e. each module is trained separately for 6 epochs; 2) '-neutral': our model with the neutralization module; 3) '-light': replace the lighting estimation module with a simple one that directly estimates two lighting codes from the input partial target scene image, and the two lighting codes are used as the inputs for the subsequent shadow generation and neutral rendering networks, respectively; 4) '-shadow': remove the shadow generation module; 5) '-hierarchic': replace the hierarchical rendering module with a simple neural rendering module. For each variant except '-end2end', we trained individual modules for 3 epochs and then trained the entire pipeline end-to-end for another 3 epochs.

From Table 2, we can see that all the proposed components contribute to the performance improvements. Note that our full model results are different from those in Table 1, because of different settings. Specifically, we use ground-truth panoramic environment maps as input without the lighting estimation module in Table 1 for a fair comparison with the other methods, while here we use the light estimation module to estimate environment maps from partial target scene images.

4.4 Applications

Lighting rotation. Because our lighting estimation can produce a panoramic environment map, our model can be used to render new portraits in which the lighting is rotated, as shown in Fig. 13. Our method can provide reasonable lighting rotation effects with shadow. Please see the supplementary video for more details.

Virtual background. Virtual background is a popular application used in video meeting software like Microsoft Teams and Zoom. All the existing solutions are based on the simple copy-paste method and ignore

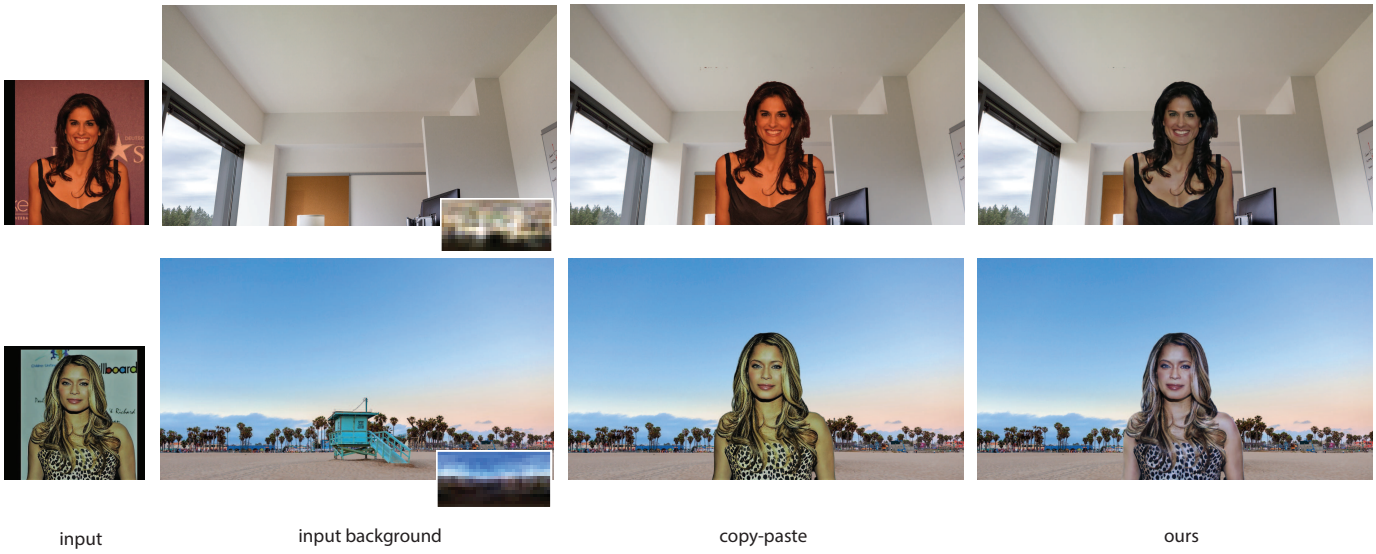


Fig. 14. Applications of virtual background composition, compared with the conventional copy-paste method.

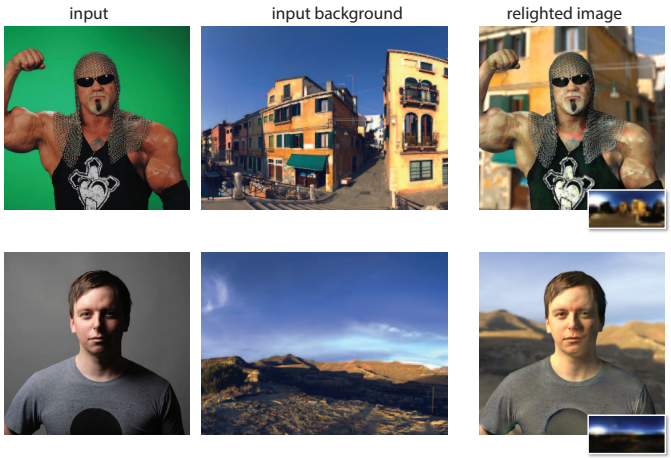


Fig. 15. Examples of failure cases due to sharp specularity (glasses) and blur hard shadow (right face).

the illumination consistency. Our method can be directly integrated into these systems to relight human images with virtual background illuminations, as shown in Figures 1, 14. We can see that our method can generate more lighting-consistent images with virtual backgrounds, providing more immersive feeling for video conferences. Please see the supplementary video for a live demo.

5 Conclusion

In this paper, we have presented the first end-to-end deep relighting system that can produce relighted portrait images being consistent with given background images with shadow effects. Our system is designed with four major components: portrait neutralization, illumination estimation, shadow map generation and hierarchic neural render. These four components are tightly coupled and combined to produce plausible relighted images with shadow effects. The shadow map generation and hierarchical neural render components are particularly novel and are not available in other methods.

Our method also has limitations. Fig. 15 gives some examples. For sharp specularity on glasses, our method tends to preserve it after relighting, since such case is under-represented in the training data. Also, for the hard shadowed portrait image, the relighting result is slightly blurred in the shadow region.

References

[1] 3D Scan Store. `https://www.3dscanstore.com`.
[2] Arnold Renderer. `https://www.arnoldrenderer.com`.
[3] J. T. Barron and J. Malik. Intrinsic scene properties from a single RGB-D image. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.
[4] P. Debevec, T. Hawkins, C. Tchou, H.-P. Duiker, W. Sarokin, and M. Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00. ACM., New York, NY, USA, 2000.
[5] H. Dong, S. Yu, C. Wu, and Y. Guo. Semantic image synthesis via adversarial learning. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2017.
[6] S. Duchêne, C. Riant, G. Chaurasia, J. L. Moreno, P.-Y. Laffont, S. Popov, A. Bousseau, and G. Drettakis. Multiview intrinsic images of outdoors scenes with an application to relighting. *ACM Trans. Graph.*, 2015.
[7] M.-A. Gardner, Y. Hold-Geoffroy, K. Sunkavalli, C. Gagne, and J.-F. Lalonde. Deep parametric indoor lighting estimation. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.
[8] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 2017.
[9] M. Garon, K. Sunkavalli, S. Hadap, N. Carr, and J.-F. Lalonde. Fast spatially-varying indoor lighting estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
[10] A. Ghosh, G. Fyffe, B. Tunwattanapong, J. Busch, X. Yu, and P. Debevec. Multiview face capture using polarized spherical gradient illumination. *ACM Trans. Graph.*, 2011.
[11] Gobotree. `https://www.gobotree.com/cat/3d-people`.
[12] P. Graham, B. Tunwattanapong, J. Busch, X. Yu, A. Jones, P. Debevec, and A. Ghosh. Measurement-based synthesis of facial microgeometry. In *Computer Graphics Forum*, 2013.
[13] Y. Guo, J. Zhang, J. Cai, B. Jiang, and J. Zheng. Photo-realistic face images synthesis for learning-based fine-scale 3D face reconstruction. *IEEE Trans. Pattern Anal. Mach. Intell.*, 08 2017.
[14] HDRI Haven. `https://hdrihaven.com`.
[15] X. Hu, Y. Jiang, C.-W. Fu, and P.-A. Heng. Mask-ShadowGAN: Learning to remove shadows from unpaired data. In *ICCV*, 2019. to appear.
[16] X. Huang and S. Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510, 2017.
[17] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 172–189, 2018.
[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, 2017.

[19] J. Johnson, A. Alahi, and F. Li. Perceptual losses for real-time style transfer and super-resolution. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[20] Y. Kanamori and Y. Endo. Relighting humans. *ACM Transactions on Graphics*, 37(6):1–11, Jan 2019. doi: 10.1145/3272127.3275104

[21] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, and W. Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4681–4690, 2017.

[22] C. LeGendre, W.-C. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. Debevec. DeepLight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5918–5928, 2019.

[23] D. Liu, C. Long, H. Zhang, H. Yu, X. Dong, and C. Xiao. Arshadowgan: Shadow generative adversarial network for augmented reality in single light scenes. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[24] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[25] F. Luan, S. Paris, E. Shechtman, and K. Bala. Deep photo style transfer. *arXiv preprint arXiv:1703.07511*, 2017.

[26] M. Meshry, D. B. Goldman, S. Khamis, H. Hoppe, R. Pandey, N. Snavely, and R. Martin-Brualla. Neural rerendering in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6878–6887, 2019.

[27] K. Nagano, H. Luo, Z. Wang, J. Seo, J. Xing, L. Hu, L. Wei, and H. Li. Deep face normalization. *ACM Transactions on Graphics (TOG)*, 38:1 – 16, 2019.

[28] K. Olszewski, D. Ceylan, J. Xing, J. Echevarria, Z. Chen, W. Chen, and H. Li. Intuitive, interactive beard and hair synthesis with generative models. 2020.

[29] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[30] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[31] remove.bg. `https://www.remove.bg/upload`.

[32] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015.

[33] S. Sengupta, J. Gu, K. Kim, G. Liu, D. W. Jacobs, and J. Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[34] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. SfSNet: Learning shape, reflectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6296–6305, 2018.

[35] M. Seymour, C. Evans, and K. Libreri. Meet mike: Epic avatars. In *ACM SIGGRAPH 2017 VR Village*, SIGGRAPH '17, 2017.

[36] Y. C. Shih, S. Paris, C. Barnes, W. T. Freeman, and F. Durand. Style transfer for headshot portraits. *ACM Transactions on Graphics*, 2014.

[37] 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*, 37, 2017.

[38] J. Stumpfel, A. Jones, A. Wenger, C. Tchou, T. Hawkins, and P. Debevec. Direct HDR capture of the sun and sky. In *ACM SIGGRAPH 2006 Courses*. ACM, New York, NY, USA, 2006.

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

[40] A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017.

[41] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 2019.

[42] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[43] Y. Yu and W. A. Smith. InverseRenderNet: Learning single image inverse rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3155–3164, 2019.

[44] H. Zhang, I. J. Goodfellow, D. N. Metaxas, and A. Odena. Self-attention generative adversarial networks. *ICML*, 2018.

[45] S. Zhang, R. Liang, and M. Wang. Shadowgan: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media*, 5:105–115, 2019.

[46] C. Zheng, T.-J. Cham, and J. Cai. T2Net: Synthetic-to-realistic translation for solving single-image depth estimation tasks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 767–783, 2018.

[47] H. Zhou and D. W. Jacobs. Deep single-image portrait relighting. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019.

[48] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman. Toward multimodal image-to-image translation. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., 2017.