Portrait Shadow Manipulation

XUANER (CECILIA) ZHANG, University of California, Berkeley
JONATHAN T. BARRON, Google Research
YUN-TA TSAI, Google Research
ROHIT PANDEY, Google
XIUMING ZHANG, MIT
REN NG, University of California, Berkeley
DAVID E. JACOBS, Google Research
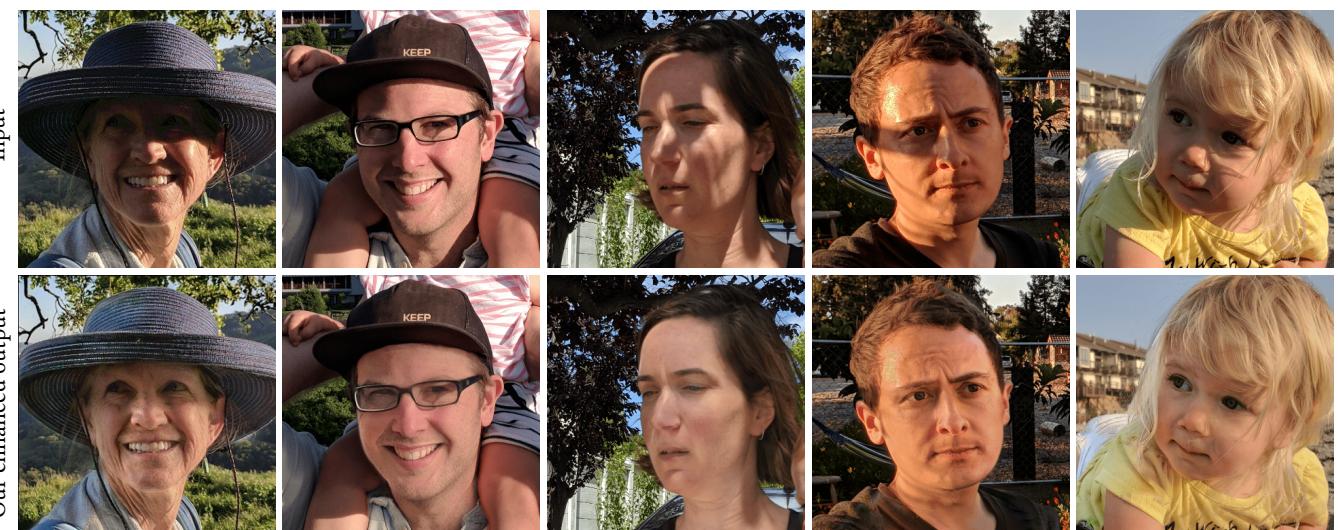
Input

Our enhanced output

Fig. 1. The results of our portrait enhancement method on real-world portrait photographs. Casual portrait photographs often suffer from undesirable shadows, particularly foreign shadows cast by external objects, and dark facial shadows cast by the face upon itself under harsh illumination. We propose an automated technique for enhancing these poorly-lit portrait photographs by removing unwanted foreign shadows, reducing harsh facial shadows, and adding synthetic fill lights.

Casually-taken portrait photographs often suffer from unflattering lighting and shadowing because of suboptimal conditions in the environment. Aesthetic qualities such as the position and softness of shadows and the lighting ratio between the bright and dark parts of the face are frequently determined by the constraints of the environment rather than by the photographer. Professionals address this issue by adding light shaping tools such as scrims, bounce cards, and flashes. In this paper, we present a computational approach that gives casual photographers some of this control, thereby allowing poorly-lit portraits to be relit post-capture in a realistic and easily-controllable way. Our approach relies on a pair of neural networks—one to remove foreign shadows cast by external objects, and another to soften facial shadows cast by the features of the subject and to add a synthetic fill light to improve the lighting ratio. To train our first network we construct a dataset of real-world portraits wherein synthetic foreign shadows are rendered onto the face, and we show that our network learns to remove those unwanted shadows. To train our second network we use a dataset of Light Stage scans of human subjects to construct input/output pairs of input images harshly lit by a small light source, and variably softened and fill-lit output images of each face. We propose a way to explicitly encode facial symmetry and show that our dataset and training procedure enable the model to generalize to images taken in the wild. Together, these networks enable the realistic and aesthetically pleasing enhancement of shadows and lights in real-world portrait images.[1]

CCS Concepts: •**Computing methodologies → Computational photography**;

Additional Key Words and Phrases: computational-photography

[1]https://people.eecs.berkeley.edu/~cecilia77/project-pages/portrait

1 INTRODUCTION

The aesthetic qualities of a photograph are largely influenced by the interplay between light, shadow, and the subject. By controlling these scene properties, a photographer can alter the mood of an image, direct the viewer's attention, or tell a specific story. Varying the position, size, or intensity of light sources in an environment can affect the perceived texture, albedo, and even three-dimensional shape of the subject. This is especially true in portrait photography, as the human visual system is particularly sensitive to subtle changes in the appearance of human faces. For example, soft lighting (*e.g.* light from a large area light source like an overcast sky) reduces skin texture, which may cause the subject to appear younger. Conversely, harsh lighting (*e.g.* light from a small or distant source like the midday sun) may exaggerate wrinkles and facial hair, making a subject appear older. Similarly, any shadows falling on the subject's face can accentuate its three-dimensional structure or obfuscate it with distracting intensity edges that are uncorrelated with facial geometry. Other variables such as the lighting angle (the angle at which light strikes the subject) or the lighting ratio (the ratio of illumination intensity between the brightest and darkest portion of a subject's face) can affect the dramatic quality of the resulting photograph, or may even affect some perceived quality of the subject's personality: harsh lighting may look "serious", or lighting from below may make the subject look "sinister".

Unfortunately, though illumination is clearly critical to the appearance of a photograph, finding or creating a good lighting environment outside of a studio is challenging. Professional photographers spend significant amounts of time and effort directly modifying the illumination of existing environments through physical means, such as elaborate lighting kits consisting of scrims (cloth diffusers), reflectors, flashes, and bounce cards (Grey 2014).

In this work, we attempt to provide some of the control over lighting that professional photographers have in studio environments to casual photographers in unconstrained environments. We present a framework that allows casual photographers to enhance the quality of light and shadow in portraits from a single image after it has been captured. We target three specific lighting problems common in casual photography and uncontrolled environments:

Foreign shadows. We will refer to any shadow cast on the subject's face by an external occluder (*e.g.* a tree, a hat brim, an adjacent subject in a group shot, the camera itself, etc.) as a *foreign shadow*. Notably, foreign shadows can result in an arbitrary two-dimensional shape in the final photograph, depending on the shape of the occluder and position of the primary, or *key*, light source. Accordingly, they frequently introduce image intensity edges that are uncorrelated with facial geometry and therefore are almost always distracting. Because most professional photographers would remove the occluder or move the subject in these scenarios, we will address this type of shadow by attempting to remove it entirely.

Facial shadows. We will refer to any shadow cast on the face by the face itself (*e.g.* the shadow attached to the nose when lit from the side) as a *facial shadow*. Because facial shadows are generated by the geometry of the subject, these shadows (unlike foreign shadows) can only project to a small space of two-dimensional shapes in the final image. Though they may be aesthetically displeasing, the image intensity edges introduced by facial shadows are more likely than foreign shadows to be a meaningful cue for the shape of the subject. Because facial shadows are almost always present in natural lighting environments (*i.e.*, the environment is not perfectly uniform), we do not attempt to remove them entirely. We instead emulate a photographer's scrim in this scenario, which effectively increases the size of the key light and softens the edges of the shadows it casts.

Lighting ratios. In scenes with very strong key lights (*e.g.* outdoors on a clear day), the ratio between the illumination of the brightest and darkest parts of the face may exceed the dynamic range of the camera, resulting in a portrait with dark shadows or blown out highlights. While this can be an intentional artistic choice, typical portrait compositions target less extreme lighting ratios. Professional photographers balance lighting ratios by placing a secondary, or *fill*, light in the scene opposite the key. We similarly place a virtual fill light to balance the lighting ratio and add definition to the shape of the shadowed portion of the subject's face.

Our framework consists of two machine learning models: one trained for foreign shadow removal, and another trained for handling facial shadow softening and lighting ratio adjustment. This grouping of tasks is motivated by two related observations.

Our first observation, as mentioned above, is tied to the differing relationships between shadow appearance and facial geometry. The appearance of facial shadows in the input image provides a significant cue for shape estimation, and should therefore be useful input when synthesizing an image with softer facial shadowing and a smaller lighting ratio. But foreign shadows are much less informative, and so we first identify and remove all foreign shadows before attempting to perform facial shadow manipulation. This approach provides our facial shadow model with an image in which all shadow-like image content is due to facial shadows, and also happens to be consistent with contemporary theories on how the human visual system perceives shadows (Rensink and Cavanagh 2004).

Our second observation relates to training dataset requirements.

Thanks to the unconstrained nature of foreign shadow appearance, it is possible to train our first network with a synthetic dataset: 5000 "in-the-wild" images, augmented with randomly generated foreign shadows for a total of 500K training examples. This strategy is not viable for our second network, as facial shadows must be consistent with the geometry of the subject and so cannot be generated in this way. Constructing an "in-the-wild" dataset consisting entirely of images with controlled facial shadowing is also intractable. We therefore synthesize the training data for this task using one-light-at-a-time (OLAT) scans taken by a Light Stage, an acquisition setup and method proposed to capture reflectance field (Debevec et al. 2000) of human faces. We use the Light Stage scans to synthesize paired harsh/soft images for use as training data. Section 3 will discuss our dataset generation procedure in more detail.

Though trained separately, the neural networks used for our two tasks share similar architectures: both are deep convolutional networks for which the input is a 256×256 resolution RGB image of a subject's face. The output of each network is a per-pixel and per-channel affine transformation consisting of a scaling A and offset

B, at the same resolution as the input image I_{in} such that the final output I_{out} can be computed as:

$$I_{out} = I_{in} \circ A + B, \qquad (1)$$

where $\circ$ denotes per-element multiplication. This approach can be thought of as an extension of quotient images (Shashua and Riklin-Raviv 2001) and of residual skip connections (He et al. 2016), wherein our network is encouraged to produce output images that resemble scaled and shifted versions of the input image. The facial shadow network includes additional inputs that are concatenated onto the input RGB image: 1) two numbers encoding the desired shadow softening and fill light brightness, so that variable amounts of softening and fill lighting can be specified and 2) an additional rendition of the input image with the face mirrored about its axis of symmetry (*i.e.*, pixels corresponding to the left eye of the input are warped to the position of the right eye, and vice versa). Using a mirrored face image in this way broadens the spatial support of the first layer of our network to include the image region on the opposite side of the subject's face. This allows the network to exploit the bilateral symmetry of human faces and to easily "borrow" pixels with similar semantic meaning and texture but different lighting from the opposite side of the subject's face (see Section 3.3 for details).

In addition to the framework itself, this work presents the following technical contributions:

(1) Techniques for generating synthetic, real-world, and Light Stage-based datasets for training and evaluating machine learning models targeting foreign shadows, facial shadows, and virtual fill lights.

(2) Symmetric face image generation for explicitly encoding symmetry cue into training our facial shadow model.

(3) Ablation studies that demonstrate our data and models achieve portrait enhancement results that outperform all baseline methods in numerical metrics and perceptual quality.

The remainder of the paper is organized as follows. Section 2 describes prior work in lighting manipulation, shadow removal, and portrait retouching. Section 3 introduces our synthetic dataset generation procedure and our real ground-truth data acquisition. Section 4 talks about our network architecture and training procedure. Section 5 shows a series of ablation studies and presents qualitative and quantitative results and comparisons. Section 6 discusses limitations of our approach.

2 RELATED WORK

The detection and removal of shadows in images is a central problem within computer vision, as is the closely related problem of separating image content into reflectance and shading (Horn 1974). Many graphics-oriented shadow removal solutions rely on manually-labeled "shadowed" or "lit" regions (Arbel and Hel-Or 2010; Gryka et al. 2015; Shor and Lischinski 2008; Wu et al. 2007). Once manually identified, shadows can be removed by solving a global optimization technique, such as graph cuts. Because relying on user input limits the applicability of these techniques, fully-automatic shadow detection and manipulation algorithms have also attracted substantial attention. Illumination discontinuity across shadow edges (Sato et al.

2003) can be used to detect and remove shadows (Baba and Asada 2003). Formulating shadow enhancement as local tone adjustment and using edge-preserving histogram manipulation (Kaufman et al. 2012) enables contrast enhancement on semantically segmented photographs. Relative differences in the material and illumination of paired image segments (Guo et al. 2012; Ma et al. 2016) enables the training of region-based classifiers and the use of graph cuts for labeling and shadow removal. Shadow removal has also been formulated as an entropy minimization problem (Finlayson et al. 2009, 2002), where invariant chromaticity and intensity images are used to produce a shadow mask that is then re-integrated to form a shadow-free image. These methods assume that shadow regions contain approximately constant reflectance and that image gradients are entirely due to changes in illumination, and are thereby fail when presented with complex spatially-varying textures or soft shadowing. In addition, by decomposing the shadow removal problem into two separate stages of detection and manipulation, these methods cannot recover from errors during the shadow detection step (Ma et al. 2016).

General techniques for inverse rendering (Ramamoorthi and Hanrahan 2001; Sengupta et al. 2018) and intrinsic image decomposition (Barron and Malik 2015; Grosse et al. 2009) should, in theory, be useful for shadow removal, as they provide shading and reflectance decompositions of the image. However, in practice these techniques perform poorly when used for shadow removal (as opposed to shading removal) and usually consider cast shadows to be out of scope. For example, the canonical Retinex algorithm (Horn 1974) assumes that shading variation is smooth and monochromatic and therefore fails catastrophically on simple cases such as shadows cast by the midday sun, which are usually non-smooth and chromatic (sunlit yellow outside the shadow, and sky blue within).

More recently, learning-based approaches have demonstrated a significant improvement on general-purpose shadow detection and manipulation (Cun et al. 2020; Ding et al. 2019; Hu et al. 2019, 2018; Khan et al. 2015; Zheng et al. 2019; Zhu et al. 2018). However, like all learned techniques, such approaches are limited by the nature of their training data. While real-world datasets for general shadow removal are available (Qu et al. 2017; Wang et al. 2018), they do not include human subjects and therefore are unlikely to be useful for our task, which requires the network to reason about specific visual characteristics of faces, such as the skin's subsurface scattering effect (Donner and Jensen 2006). Instead, in this paper, we propose to train a model using synthetic shadows generated on images in the wild. We only use images of faces to encourage the model to learn and use priors on human faces. Earlier work has shown that training models on faces improves performance on face-specific subproblems of common tasks, such as inpainting (Ulyanov et al. 2018; Yeh et al. 2017), super-resolution (Chen et al. 2018) and synthesis (Denton et al. 2015).

Another problem related to ours is "portrait relighting"—the task of relighting a single image of a human subject according to some desired environment map (Sun et al. 2019; Zhou et al. 2019). These techniques could theoretically be used for our task, as manipulating the facial shadows of a subject is equivalent to re-rendering that subject under a modified environmental illumination map in which the key light has been dilated. However, as we will demonstrate (and

was noted in (Sun et al. 2019)) these techniques struggle when presented with images that contain foreign shadows or high-frequency image structure due to harsh shadows in the input image, which our approach specifically addresses. Example-based portrait lighting transfer techniques (Shih et al. 2014; Shu et al. 2017) also represent potential alternative solutions to this task, but they require a high-quality reference image that exhibits the desired lighting, and that also contains a subject with a similar identity and pose as the input image—an approach that does not scale to casual photos in the wild.

3 DATA SYNTHESIS

There is no tractable data acquisition method to collect a large-scale dataset of human faces for our task with diversity in both the subject and the shadows, as the capture process would be onerous for both the subjects (who must remain perfectly still for impractically long periods of time) and the photographers (who must be specially trained for the task and find thousands of willing participants in thousands of unique environments). Instead, we synthesize custom datasets for our subproblems by augmenting existing datasets— Recall that our two models require fundamentally different training data. Our foreign shadow datasets (Section 3.1) are based on images of faces in the wild with rendered shadows, while our facial shadow and fill light datasets (Section 3.2) are based on a Light Stage dataset with carefully chosen simulated environments.

3.1 Foreign Shadows

To synthesize images that appear to contain foreign shadows, we model images as a linear blend between a "lit" image I_ℓ and a "shadowed" image I_s, according to some shadow mask M:

$$I = I_\ell \circ (1 - M) + I_s \circ M \tag{2}$$

The lit image I_ℓ is assumed to contain the subject lit by all light sources in the scene (*e.g.* the sun and the sky), and the shadowed image I_s is assumed to be the subject lit by everything other than the key (*e.g.* just the sky). The shadow mask M indicates which pixels are shadowed: $M = 1$ if fully shadowed, and $M = 0$ if fully lit. To generate a training sample, we need I_ℓ, I_s, and M. I_ℓ is selected from an initial dataset described below, I_s is a color transform of I_ℓ, and M comes from a silhouette dataset or a pseudorandom noise function.

Because deep learning models are highly sensitive to the realism and biases of the data used during training, we take great care to synthesize as accurate a shadow mask and shadowed image as possible with a series of augmentations on I_s and M. Figure 2 presents an overview of the process and below we enumerate different aspects of our synthesis model and their motivation. In Section 5.2, we will demonstrate their efficacy through an ablation study.

Input Images. Our dataset is based on a set of 5,000 faces in the wild that we manually identified as not containing any foreign shadows. These images are real, in-the-wild JPEG data, and so they are varied in terms of subject, ethnicity, pose, and environment. Common accessories such as hats and scarves are included, but only if they do not cast shadows. We make one notable exception to this policy: glasses. Shadows from glasses are unavoidable and behave more like facial shadows than foreign. Accordingly, shadows

from glasses are preserved in our ground truth. Please refer to the supplement for examples.

Light Color Variation. The shadowed image region I_s is illuminated by a lighting environment different from that of the non-shadow region. For example, outdoor shadows are often tinted blue because when the sun is blocked, the blue sky becomes the dominant light source. To account for such illumination differences, we apply a random color jitter, formulated as a 3×3 color correction matrix, to the lit image I_ℓ. Please see the supplement for details.

Shape Variation. The shapes of natural shadows are as varied as the shapes of natural objects in the world, but those natural shapes also exhibit significant statistical regularities (Huang et al. 2000). To capture both the variety and the regularity of real-world shadows, our distribution of input shadow masks M_{in} is half "regular" real-world masks drawn from a dataset of 300 silhouette images of natural objects, randomly scaled and tiled; and half "irregular" masks generated using a Perlin noise function at 4 octaves with a persistence drawn uniformly at random within $[0, 0.85]$, with the initial amplitude set to 1.0.

Subsurface Scattering. Light scatters beneath the surface of human skin before exiting, and the degree of that scattering is wavelength-dependent (Hanrahan and Krueger 1993; Jensen et al. 2001; Krishnaswamy and Baranoski 2004): blood vessels cause red light to scatter further that other wavelengths, causing a visible color fringe at shadows. We approximate the subsurface scattering appearance by uniformly blurring M_{in} with a different kernel per color channel, borrowing from Fernando (2004). In brief, the kernel for each channel is a sum of Gaussians $G(\sigma_{c,k})$ with weights $w_{c,k}$, such that each channel M_c of the shadow mask with subsurface scattering M_{ss} is rendered as:

$$M_c = \sum_k M_{\text{in}} * G(\sigma_{c,k}) w_{c,k}. \tag{3}$$

Spatial Variation. The softness of the shadow being cast on a subject depends on the relative distances between the subject, the key light, and the object casting the shadow. Because this relationship varies over the image, our shadow masks incorporate a spatially-varying blur over M_{ss}. While many prior works assume that the shadow region has a constant intensity (Zhang et al. 2019b), we note that a partially translucent occluder or an environment violating the assumption that lights are infinitely far away will cause shadows to have different local intensities. Accordingly, we similarly apply a spatially-varying per-pixel intensity variation to M_{ss} as well, modeled as Perlin noise at 2 octaves with a persistence drawn uniformly at random from $[0.05, 0.25]$ and an initial amplitude set to 1.0. The final mask with spatial variation incorporated is what we refer to as M above.

3.2 Facial Shadows

We are not able to use "in-the-wild" images for synthesizing *facial* shadows because the highly accurate facial geometry it would require is generally not captured in such datasets. Instead, we use Light Stage data that can relight a scanned subject with perfect fidelity under any environment and select the simulated environment with care. Note that we *cannot* use light stage data to produce more
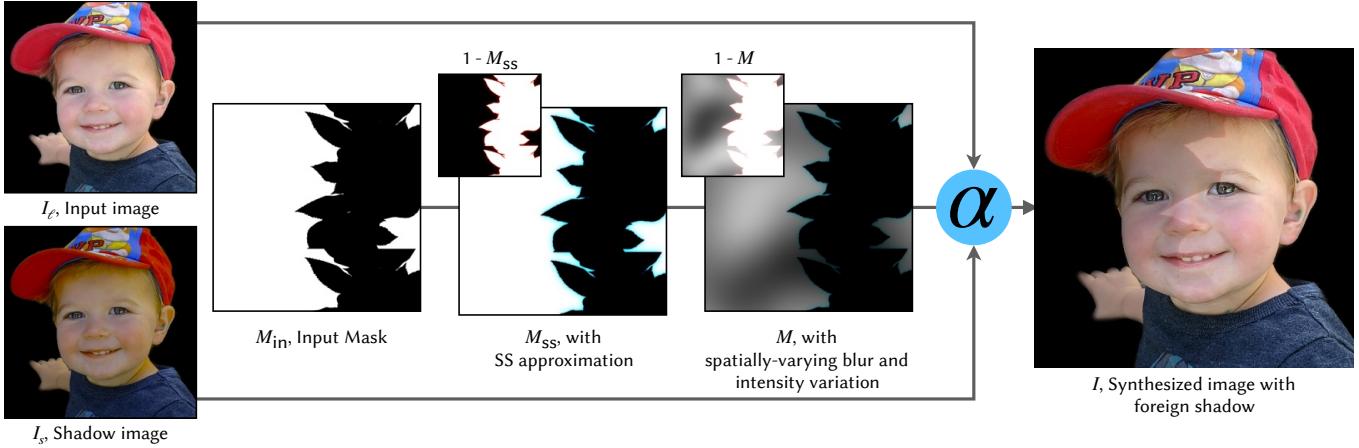
I_ℓ, Input image

I_s, Shadow image

M_{in}, Input Mask

M_{SS}, with SS approximation

M, with spatially-varying blur and intensity variation

I, Synthesized image with foreign shadow

Fig. 2. The pipeline of our foreign shadow synthesis model (Section 3.1). The colors of the "lit" image I_ℓ are randomly jittered to generate a "shadow" image I_s. The input mask M_{in} shown here is generated from an object silhouette, though it may also be generated with Perlin noise. M_{in} is subjected to a subsurface scattering (SS) approximation of human skin to generate M_{SS}, then a spatially-varying blur and per-pixel intensity variation to generate M. I_ℓ and I_s are then blended according to the shadow mask M to generate a training sample I.

(a) Single key light

(b) Soft key light with size 5

(c) Soft key light with size 20

(d) Soft key light with size 20 and fill light with size 20

(e) difference image between (c) and (d)

- Key Lights
- Fill Lights
- Key Light Direction
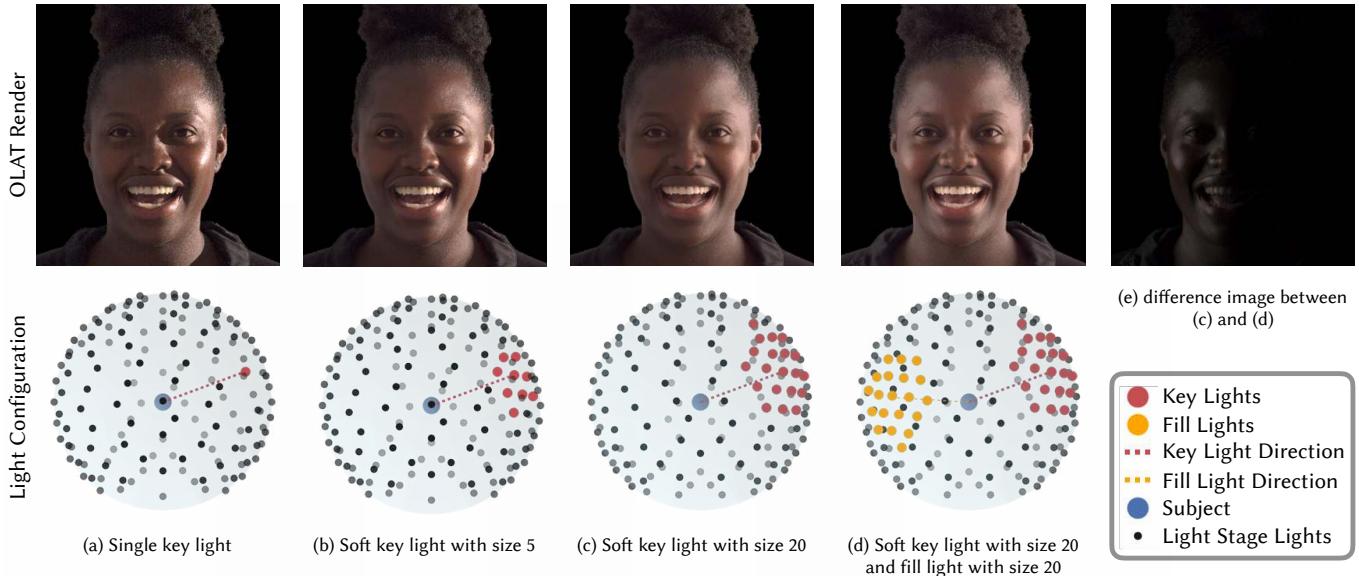- Fill Light Direction
- Subject
- Light Stage Lights

Fig. 3. Our facial shadow synthesis model. Our input image is a OLAT render corresponding to an environment with a single key light turned on as shown in (a). To soften the shadows by some variable amount, we distribute the key's energy to a variable number of its neighbors, as shown in (b) and (c). We also add a number of fill lights on the opposite side of the Light Stage, to brighten the darker side of the face as shown in (d), with the fill light's difference image visualized in (e). For clarity, only half of the Light Stage's lights are rendered. See the supplemental video at 02:04 for a few more examples.

accurate foreign shadows than we could using raw, in-the-wild JPEG images, which is why we adopt different data synthesis for these two tasks.

When considering *foreign* shadows, we adopt shadow *removal* with the rationale that foreign shadows are likely undesirable from a photographic perspective and removing them does not affect the apparent authenticity of the photograph (as the occluder is rarely in frame). *Facial* shadows, in contrast, can only be *softened* if we

wish to affect the mood of the photograph while remaining faithful to the scene's true lighting direction.

We construct our dataset by emulating the scrims and bounce cards employed by professional photographers. Specifically, we generate harsh/soft facial shadow pairs using OLAT scans from a Light Stage dataset. This is ideal for two reasons: 1) each individual light in the stage is designed to match the angular extent of the sun, so it is capable of generating harsh shadows, and 2) with such a

dataset, we can render an image I simulating an arbitrary lighting environment with a simple linear combination of OLAT images I_i with weights w_i, i.e., $I = \sum_i I_i w_i$.

For each training instance, we select one of the 304 lights in the stage and dub it our key light with index i_{key}, and use its location to define the key light direction $\vec{\ell}_{\text{key}}$. Our harsh input image is defined to be one corresponding to OLAT weights $w_i = \{P_{\text{key}}$ if $i = i_{\text{key}}$, ϵ otherwise$\}$, where P_{key} is a randomly sampled intensity of the key light and ϵ is a small non-zero value that adds ambient light to prevent shadowed pixels from becoming fully black. The corresponding soft image is then rendered by splatting the key light energy to the set of its m nearest neighboring lights $\Omega(\vec{\ell}_{\text{key}})$, where m is drawn uniformly from a set of discrete numbers $[5, 10, 20, 30, 40]$. This can be thought of as convolving the key light source with a disc, similar in spirit to a diffuser or softbox. We then compute the location of the fill light (Figure 3(d)):

$$\vec{\ell}_{\text{fill}} = 2(\vec{\ell}_{\text{key}} \cdot \vec{n})\vec{n} - \vec{\ell}_{\text{key}}, \tag{4}$$

where $\vec{n}$ is the unit vector along the camera z-axis, pointing out of the Light Stage. For all data generation, we use a fixed fill light neighborhood size of 20, and a random fill intensity P_{fill} in $[0, P_{\text{key}}/10]$. Thus, the soft output image is defined as one corresponding to OLAT weights

$$w_i = \begin{cases} P_{\text{key}}/m, & \text{if } i \in \Omega(\vec{\ell}_{\text{key}}) \\ P_{\text{fill}}, & \text{if } i \in \Omega(\vec{\ell}_{\text{fill}}) \\ \epsilon, & \text{otherwise} \end{cases}$$

To train our facial shadow model, we use OLAT images of 85 subjects, each of which was imaged under different expressions and poses, giving us in total 1795 OLAT scans to render our facial harsh shadow dataset. We remove degenerate lights that cause strong flares or at extreme angles that render too dark images, and end up using the remaining 284 lights for each view.

3.3 Facial Symmetry

Human faces tend to be bilaterally symmetric: the left side of most faces closely resembles the right side in terms of geometry and reflectance, except for the occasional blemish or minor asymmetry. However, images of faces are rarely symmetric because of facial shadows. Therefore, if a neural network can easily reason about the symmetry of image content on the subject's face, it will be able to do a better job of reducing shadows cast upon that face. For this reason, we augment the image that is input to our neural networks with a "mirrored" version of that face, thereby giving the early layers of those networks the ability to straightforwardly reason about which image content is present on the opposite side of the face. Because the subject's face is rarely perfectly vertical and oriented perpendicularly to the camera's viewing direction, it is not sufficient to simply mirror the input image along the x-axis. We therefore estimate the geometry of the face and mirror the image using that estimated geometry, by warping image content near each vertex of a mesh to the location of its corresponding mirrored vertex. See Figure 4 for a visualization.

Given an image I, we use the landmark detection system of (Kartynnik et al. 2019) to produce a model of facial geometry consisting of 468 2D vertices (Figure 4(b)) and a mesh topology (which is fixed
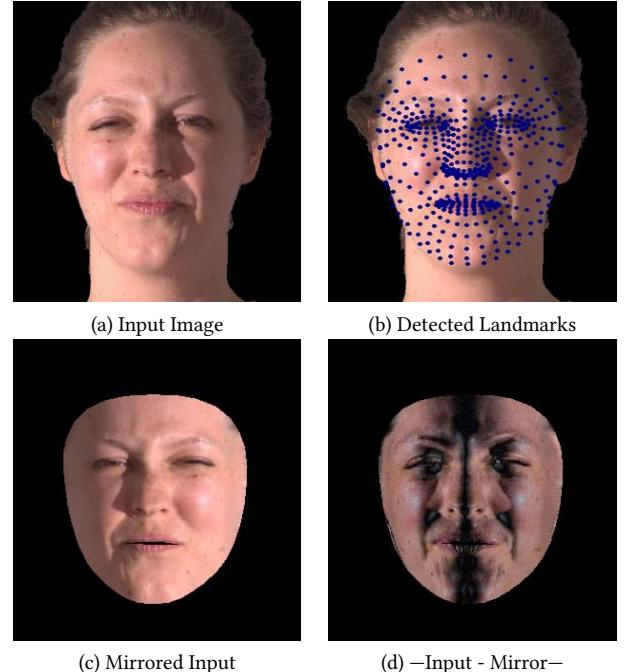

(a) Input Image   (b) Detected Landmarks

(c) Mirrored Input   (d) —Input - Mirror—

Fig. 4. The symmetry of human faces is a useful cue for reasoning about lighting: a face's reflectance and geometry is likely symmetric, but the shadow cast upon that face is likely not symmetric. To leverage this, a landmark detection system is applied to the input image (a) and the recovered landmark (b) are used to produce a per-pixel mirrored version of the input image (c). This mirrored image is appended to the input image in our networks, which improves performance by allowing the network to directly reason about asymmetric image content (d) which is likely due to facial and foreign shadows.

for all instances). For each vertex j we precompute the index of its bilaterally symmetric vertex $\bar{j}$, which corresponds to a vertex $(u_{\bar{j}}, v_{\bar{j}})$ at the same position as (u_j, v_j) but on the opposite side of the face. With this correspondence we could simply produce a "mirrored" version of I by applying a meshwarp to I where the position of each vertex j is moved to the position of its mirror vertex $\bar{j}$. However, a straightforward meshwarp is prone to triangular-shaped artifacts and irregular behavior on foreshortened triangles or inaccurately-estimated keypoint locations. For this reason we instead use a "soft" warping approach based on an adaptive radial basis function (RBF) kernel: For each pixel in I we compute its RBF weight with respect to the 2D locations of all vertices, express that pixel location as a convex combination of all vertex locations, and then interpolate the "mirrored" pixel location by computing the same convex combination of all *mirrored* vertex locations. Put formally, we first compute the Euclidean distance from all pixel locations to all vertex locations:

$$D_{i,j} = (x_i - u_j)^2 + (y_i - v_j)^2 \tag{5}$$

With this we compute a weight matrix consistent of normalized Gaussian distances:

$$W_{i,j} = \frac{\exp\left(-D_{i,j}/\sigma_j\right)}{\sum_{j'} \exp\left(-D_{i,j'}/\sigma_{j'}\right)} \tag{6}$$

Unlike a conventional normalized RBF kernel, $W_{i,j}$ is computed using a different σ for each of the j vertices. Each vertex's σ is selected such that each landmark's influence in the kernel is inversely proportional to how many nearby neighbors it has for this particular image:

$$\sigma_j = \underset{j'}{\text{select}}\left(\left(u_j - u_{j'}\right)^2 + \left(v_j - v_{j'}\right)^2, K_\sigma\right) \tag{7}$$

Where $\text{select}(\cdot, K)$ returns the K'th smallest element of an input vector. This results in a warp where sparse keypoints have significant influence over their local neighborhood, while the influence of densely packed keypoints is diluted. This weight matrix is then used to compute the weighted average of mirrored vertex locations, and this 2D location is used to bilinearly interpolate into the input image to produce it's mirrored equivalent:

$$\bar{I} = I\left(\sum_j W_{i,j} u_{\bar{j}}, \sum_j W_{i,j} v_{\bar{j}}\right) \tag{8}$$

The only hyperparameter in this warping model is an integer value K_σ, which we set to 4 in all experiments. This proposed warping model is robust to asymmetric expressions and poses assuming the landmarks are accurate, but is sensitive to asymmetric skin features, e.g., birthmarks.

The input to our facial shadow network is the concatenation of the input image I with its mirrored version $\bar{I}$ along the channel dimension. This means that the receptive field of our CNN includes not just the local image neighborhood, but also its mirrored counterpart. Note that we do not include the mirrored image as input to our foreign shadow model, as we found it did not improve results. We suspect that this is due to the unconstrained nature of foreign shadow appearance, which weakens the assumption that corresponding face regions will have different lighting.

4 NEURAL NETWORK ARCHITECTURE AND TRAINING

Here we describe the neural network architectures that we use for removing foreign shadows and for softening facial shadows. As the two tasks use different datasets and there is an additional conditional component in the facial shadow softening model, we train these two tasks separately.

For both models, we employ a GridNet (Fourure et al. 2017) architecture with modifications proposed in (Niklaus and Liu 2018). GridNet is a grid-like architecture of rows and columns, where each row is a stream that processes features with resolution kept unchanged, and columns connect the streams by downsampling or upsampling the features. By allowing computation to happen at different layers and different spatial scales instead of conflating layers and spatial scales (as U-Nets do) GridNet produces more accurate predictions as has been successfully applied to a number of image synthesis tasks (Niklaus and Liu 2018; Niklaus et al. 2019). We use a GridNet with eight columns wherein the first three columns perform downsampling and the remaining five columns perform upsampling, and use five rows for foreign model and six rows for facial model, as we found this to work best after an architecture search.

For all training samples, we run a face detector to obtain a face bounding box, then resize and crop the face into 256×256 resolution.

For the foreign shadow removal model, the input to the network is a 3-channel RGB image and the output of the model is a 3-channel scaling A and a 3-channel offset B, which are then applied to the input to produce a 3-channel output image (Equation 1). For the facial shadow softening model, we additionally concatenate the input to the network with its mirrored counterpart (as per Section 3.3). As we would like our model to allow for a *variable* degree of shadow softening and of fill lighting intensity, we introduce two "knobs"— one for light size m and the other for fill light intensity P_{fill}, which are assumed to be provided as input. To inject this information into our network, a 2-channel image containing these two values at every pixel is concatenated into both the input and the last layers of the encoders of the network.

We supervise our two models using a weighted combination of pixel-space L1 loss ($\mathcal{L}_{\text{pix}}$) and a perceptual feature space loss ($\mathcal{L}_{\text{feat}}$) which has been used successfully to train models such as image synthesis and image decomposition (Chen and Koltun 2017; Zhang et al. 2019a, 2018b). Intuitively, the perceptual loss accounts for high-level semantics in the reconstructed image but may be invariant to some non-semantic image content. By additionally minimizing a per-pixel L1 loss our model is better able to recover low-frequency image content. The perceptual loss is computed by processing the reconstructed and ground truth images through a pre-trained VGG-19 network $\Phi(\cdot)$ and computing the L1 difference between extracted features in selected layers as specified in (Zhang et al. 2018b). The final loss function is formulated as:

$$\mathcal{L}_{\text{feat}}(\theta) = \sum_d \lambda_d \left\|\Phi_d\left(I^*\right) - \Phi_d\left(f\left(I_{in}; \theta\right)\right)\right\|_1$$
$$\mathcal{L}_{\text{pix}}(\theta) = \left\|I^* - f(I_{in}; \theta)\right\|_1$$
$$\mathcal{L}(\theta) = 0.01 \times \mathcal{L}_{\text{feat}}(\theta) + \mathcal{L}_{\text{pix}}(\theta), \tag{9}$$

where I^* is the ground-truth shadow-removed or shadow-softened RGB image, $f(\cdot; \theta)$ denotes our neural network, and λ_d denotes the selected weight for the d-th VGG layer. $I_{in} = I$ for foreign removal model and $I_{in} = \text{concat}(I, \bar{I}, P_{\text{fill}}, m)$ for facial shadow softening model. This same loss is used to train both models separately. We minimize $\mathcal{L}$ with respect to both of our model weights θ using Adam (Kingma and Ba 2015) ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$) for 500K iterations, with a learning rate of 10^{-4} that is decayed by a factor of 0.9 every 50K iterations.
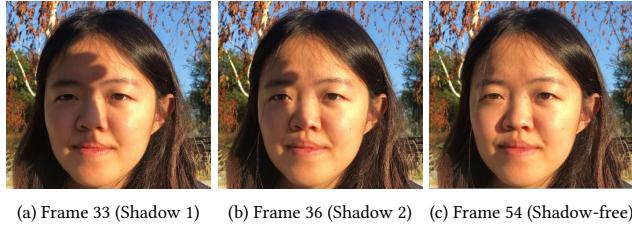
5 EXPERIMENTS

We use synthetic and real in-the-wild test sets to evaluate our foreign shadow removal model (Section 5.3) and our facial shadow softening model (Section 5.4). We also present an ablation study of the components of our foreign shadow synthesis model (Section 5.2) as well as of our facial symmetry modeling. Extensive additional results can be found in the supplement.

5.1 Evaluation Data

We evaluate our foreign shadow removal model with two datasets:

foreign-syn. We use a held-out set of the same synthetic data generation approach described in (Section 3.1), where the images (*i.e.*, subjects) and shadow masks to generate test-set images are not present in the training set.

(a) Frame 33 (Shadow 1)   (b) Frame 36 (Shadow 2)   (c) Frame 54 (Shadow-free)

Fig. 5. An example of the shadow removal evaluation dataset we produce using video captured by a smartphone camera stabilized on a tripod. By filming a stationary subject under the shadow cast by a moving foreign occluder (a-c), we are able to obtain multiple input/ground-truth image pairs of the subject (a, c), (b, c). This provides us with an efficient way to collect a set of diverse foreign shadows for evaluation. Please see the supplemental video at 02:39 for more example video clips.

foreign-real. We collect an additional dataset of in-the-wild images for which we can obtain ground-truth images that do not contain foreign shadows. This dataset enables the quantitative and qualitative comparison of our proposed model against prior work. This is accomplished by capturing high-framerate (60 fps) videos of stationary human subjects while moving a shadow-casting object in front of the subject. We collect this evaluation dataset outdoors, and use the sun as the dominant light source. For each captured video, we manually identify a set of "lit" images and a set of "shadowed" images. For each "shadowed" image, we automatically use homography to align it to each "lit" and find the one that gives the minimum mean pixel error as its counterpart. Because the foreign object is moving during capture, this collection method provides a diversity in the shape and the position of foreign shadows (see examples in Figure 5 and the supplemental video). In total, we capture 20 videos of 8 subjects during different times of day, which gives us 100 image pairs of foreign shadow with ground truth.

We evaluate our facial shadow model with another dataset:

facial-syn. We use the same OLAT Light Stage data that is used to generate our facial model training data to generate a test set, by using a held-out set of 5 subjects that are not used during training. We record each harsh input shadow image and the soft ground-truth output image along with their corresponding light size m and fill light intensity P_{fill} for use. Note that though this dataset is produced through algorithmic means, the ground-truth outputs are a weighted combination of real observed Light Stage images, and are therefore an accurate reflection of the true appearance of the subject up to the sampling limitations of the Light Stage hardware.

We qualitatively evaluate both our foreign shadow removal model and our facial shadow softening model using an additional dataset:

wild. We collect 100 "in the wild" portrait images of varied human subjects that contain a mix of different foreign and facial shadows. Images are taken from the Helen dataset (Le et al. 2012), the HDRnet dataset (Gharbi et al. 2017), and our own captures. These images are processed by our foreign shadow removal model, our facial shadow softening model, or both, to generate enhanced outputs that give a sense of the qualitative properties of both components of our model. See Figures 1, 8, 10, and the supplement for results.

5.2 Ablation Study of Foreign Shadow Synthesis

Our foreign shadow synthesis technique (Section 3.1) simulates the complicated effect of foreign shadows on the appearance of human subjects. We evaluate this technique by removing each of the three components and measuring model performance. Our three ablations are: 1) "No-SV": synthesis without spatially varying blur or the intensity variation of the shadow, 2) "No-SS": synthesis where the approximated subsurface scattering of skin has been removed, and 3) "No-Color": synthesis where the color perturbation to generate the shadow image is not randomly changed. Quantitative results for this ablation study on our `foreign-syn` and `foreign-real` datasets can be found in Table 1, and qualitative results for a test image from `foreign-real` are shown in Figure 6.

5.3 Foreign Shadow Removal Quality

Because no prior work appears to address the task of foreign shadow removal for human faces, we compare our model against general-purpose shadow removal methods: the state-of-the-art learning-based method of Cun et al. (2020)[2] that uses a generative model to synthesize and then remove shadows, a customized network with attention mechanism designed by Hu et al. (2019)[3] for shadow detection and removal, and the non-learning-based method of Guo et al. (2012) that relies on image segmentation and graph cuts. The original implementation from (Guo et al. 2012) is not available publicly, so we use a reimplementation[4] that is able to reproduce the results of the original paper. We use the default parameters settings for this code, as we find that tuning its parameters did not improve performance for our task. Hu et al. (2019) provide two models trained on two general-purpose shadow removal benchmark datasets (SRD and ISTD), we use the SRD model as it performs better than the ISTD model on our evaluation dataset.

We evaluate these baseline methods on our `foreign-syn` and `foreign-real` datasets, as these both contain ground truth shadow-free images. We compute PSNR, SSIM (Wang et al. 2004) and a learned perceptual metric LPIPS (Zhang et al. 2018a) between the ground truth and the output. Results are shown in Table 2 and Figure 7. Our model outperforms these baselines by a large margin. Please see video at 03:03 for more results.

5.4 Facial Shadow Softening Quality

Transforming harsh facial shadows to soft in image space is roughly equivalent to relighting a face with a blurred version of the dominant light source in the original lighting environment. We compare our facial softening model against the portrait relighting method from Sun et al. (2019), by applying a Gaussian blur to the estimated environment map from the model and then pass to the decoder for relighting. The amount of blur to apply, however, cannot map exactly to our light size parameter. We experiment with a few blur kernel values and choose the one that produces the minimum mean pixel error with the ground truth. We do this for each image, and show qualitative comparisons in Figure 10. In Table 3, we compare our model against the Sun et al. (2019) baseline and against

---

[2]https://github.com/vinthony/ghost-free-shadow-removal
[3]https://github.com/xw-hu/DSC
[4]https://github.com/kittenish/Image-Shadow-Detection-and-Removal

(a) Input Image   (b) Our Model, No-SV   (c) Our Model, No-SS   (d) Our Model, No-Color   (e) Our Model   (f) Ground Truth
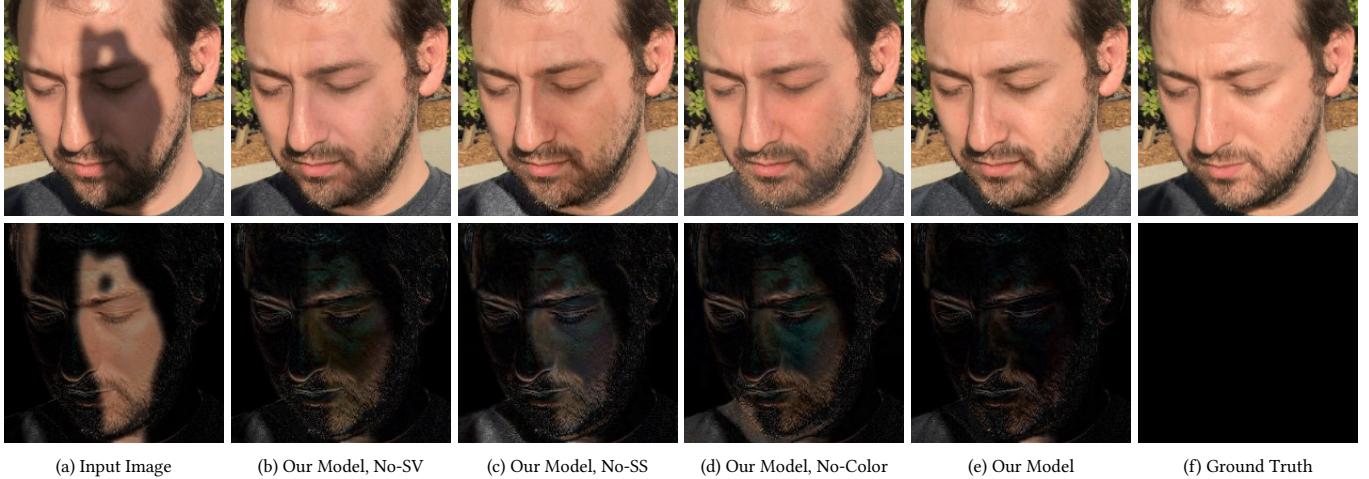
Fig. 6. A visualization of an ablation study of our foreign shadow removal algorithm as different aspects of our foreign shadow synthesis model (Section 3.1) are removed. The "No-SV", "No-SS", and "No-Color" ablations show our model trained on synthesized data *without* modeling spatial variation, approximate subsurface scattering, or color perturbation, respectively. The top row shows the images generated by each model, and the bottom row shows the difference between each output and the ground truth image (f). Our complete model (e) clearly outperforms the others. Notice the red-colored residual along the shadow edge in the model trained without approximating subsurface scattering (c), and the color inconsistency in the removed region in the model trained without color perturbation (d). A quantitative evaluation on the entire set `foreign-real` is shown in Table 1.

Table 1. A quantitative ablation study of our foreign shadow synthesis model in terms of PSNR, SSIM, and LPIPS. Ablating any component of our synthesis model hurts the performance of the resulting model.

Synthesis Model	Rendered Test Set (foreign-syn)			Real Test Set (foreign-real)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Ours, "No-Color"	26.248	0.818	0.079	21.387	0.766	0.085
Ours, "No-SV"	27.546	0.830	0.058	22.095	0.782	0.081
Ours, "No-SS"	26.996	0.809	0.074	21.663	0.770	0.086
Ours	29.814	0.926	0.054	23.816	0.782	0.074

Table 2. A quantitative evaluation of our foreign shadow removal model. We compare against baseline methods of Guo et al. (2012), Hu et al. (2019) (SRD) and Cun et al. (2020) on both synthetic and real test sets. The input image itself is also included as point of reference. In terms of both image-quality (PSNR) and perceptual-quality (SSIM and LPIPS), our model produces better performance on all three metrics with a large margin. Visual comparisons can be seen in Figure 7.

Removal Model	Rendered Test Set (foreign-syn)			Real Test Set (foreign-real)		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Input Image	20.657	0.807	0.206	19.671	0.766	0.115
(Guo et al. 2012)	19.170	0.699	0.359	15.939	0.593	0.269
(Hu et al. 2019)	20.895	0.742	0.238	18.956	0.699	0.148
(Cun et al. 2020)	22.405	0.845	0.173	19.386	0.722	0.133
Ours	29.814	0.926	0.054	23.816	0.782	0.074

an ablation of our model without symmetry, and demonstrate an improvement with respect to both. For all comparisons, we use `facial-syn`, which has ground truth soft facial shadows. Please see video at 03:35 for more results.

Table 3. A comparison of our facial shadow reduction model against the PR-net of Sun et al. (2019) and an ablation of our model with symmetry. in terms of PSNR, SSIM, and LPIPS on the "`facial-syn`" test dataset. We see that PR-net performs poorly on images that contain harsh facial shadows, and removing the concatenated "mirrored" input during training (*i.e.*, setting $I_{in} = I$) lowers accuracy by all three metrics.

Shadow Reduction Model	PSNR↑	SSIM↑	LPIPS↓
PR-net (Sun et al. 2019)	21.639	0.709	0.152
Ours w/o Symmetry	24.232	0.826	0.065
Ours	26.740	0.914	0.054

5.5 Preprocessing for Portrait Relighting

Our method can also be used as a "preprocessing" step for image modification algorithms such as portrait relighting (Sun et al. 2019; Zhou et al. 2019), which modify or replace the illumination of the input image. Though often effective, these portrait relighting techniques sometimes produce suboptimal renderings when presented with input images that contain foreign shadows or harsh facial shadows. Our technique can improve a portrait relighting solution: our model can be used to remove these unwanted shadowing effects, producing a rendering that can then be used as input to a portrait relighting solution, resulting in an improved final rendering. See Figure 11 for an example.

6 LIMITATIONS

Our proposed model is not without its limitations, some of which we can identify in our `wild` dataset. When foreign shadows contain many finely-detailed structures (which are underrepresented in training), our output may retain visible residuals of those (Figure 12(a)). While exploiting the bilateral symmetry of the subject significantly improves our facial softening model's performance,
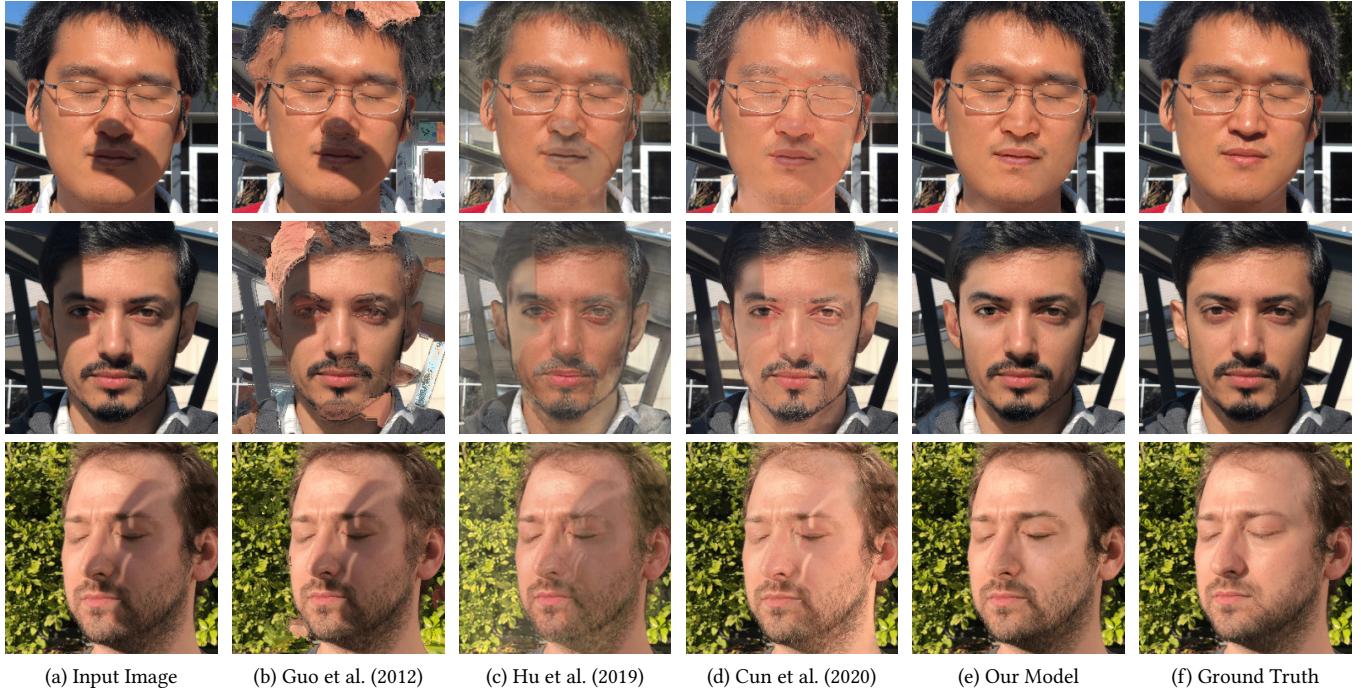
| (a) Input Image | (b) Guo et al. (2012) | (c) Hu et al. (2019) | (d) Cun et al. (2020) | (e) Our Model | (f) Ground Truth |

Fig. 7. Foreign shadow removal results on images from our `foreign-real` test dataset. The method of Guo et al. (2012) often incorrectly identifies dark image regions as shadows and removes them, while also failing to identify real shadows (b). The deep learning approaches of Cun et al. (2020) and Hu et al. (2019) (c, d) do a better job of correctly identifying shadow regions but often fail to remove shadows completely, and also change the overall brightness and tone of the image in a way that does not preserve the authenticity of the input image. Our method is able to entirely remove foreign shadows while still preserving the overall appearance of the subject (e), thereby producing output images that more closely resemble the ground truth (f).

Fig. 8. Foreign shadow removal results of our model on our `wild` test dataset. Input images that contain unwanted foreign shadows (top) are processed by our foreign shadow removal model (bottom). Though real-world foreign shadows exhibit significant variety in terms of shape, softness, and color, our foreign shadow removal model is able to successfully generalize to these challenging real-world images despite having been trained entirely on our synthetic training data (Section 3.1).

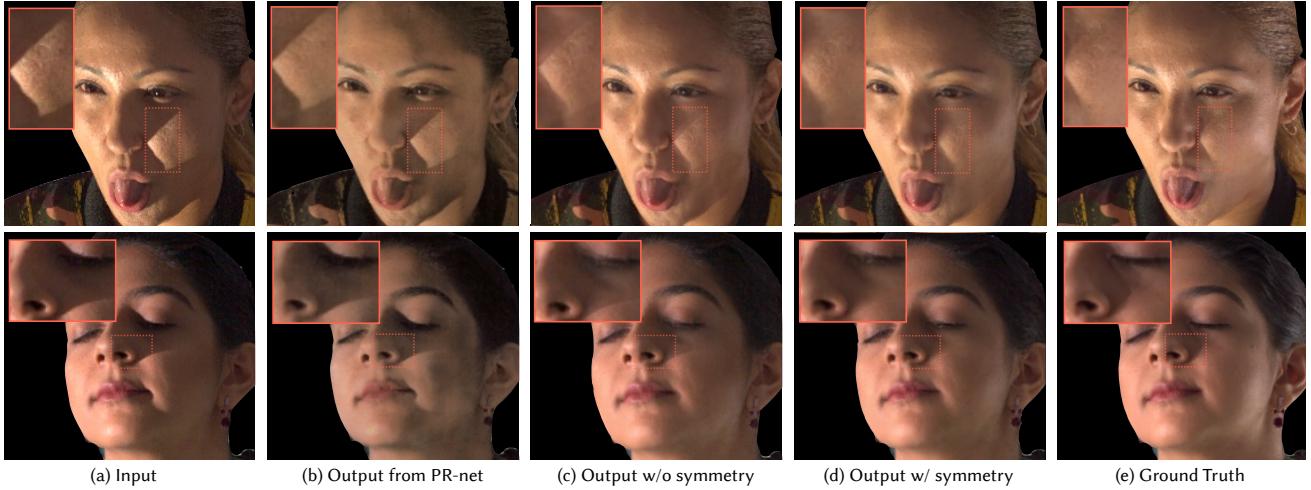|				
(a) Input	(b) Output from PR-net	(c) Output w/o symmetry	(d) Output w/ symmetry	(e) Ground Truth

Fig. 9. Facial shadow softening results on facial-syn. We compare against the portrait relighting model (PR-net) Sun et al. (2019) by applying a blur to its estimated environment light and relighting the input image with that blurred environment map. PR-net is able to successfully soften low frequency shadows but struggles with harsh facial shadows (b). The ablation of our model without our symmetry component (Section 3.3) also underperforms on these harsh facial shadows (c). Our complete model successfully softens these hard shadows (d), as it is able to reason about the bilateral symmetry of the subject and "borrow" pixels with similar reflectance and geometry from the opposite side of the face.

(a) Input	(b) Output ($m = 10$)	(c) Output ($m = 40$)	(d) Output ($m = 40$, $P_{\text{fill}} = P_{\text{fill}}^{\max}$)	(e) —(c) - (d)—

Fig. 10. Facial shadow softening results on images from wild. Input images may contain harsh facial shadows, such as around the eyes (row 1) and by the subject's cheek (row 3). Applying our facial shadow softening model with a variable "light size" m produces images with softer shadows (b, c). The specular reflection also gets suppressed, which is a desired photographic practice as specular highlights are often distracting and obscuring the surface of the subject. Additionally, the lighting ratio component of our model reduces the contrast induced by facial shadows (d) by adding a synthetic fill light with intensity P_{fill}, set here to the maximum value used in training (Section 3.2), in the direction opposite to the detected key light, as visualized in (e).

(a) Input   (b) Sun et al. (2019) applied to (a)

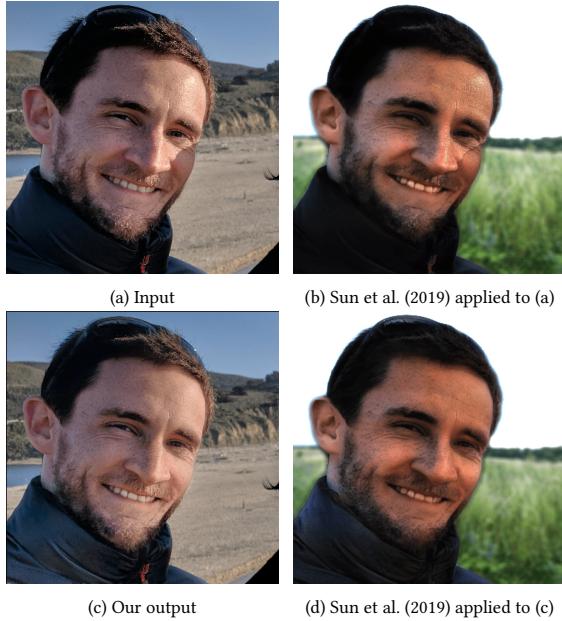(c) Our output   (d) Sun et al. (2019) applied to (c)

Fig. 11. The portrait relighting technique of Sun et al. (2019) provides an alternative approach for shadow manipulation. However, applying this technique to input images that contain foreign shadows and harsh facial shadows (a) often results in relit images in which these foreign and facial shadows persist as artifacts (b). If this same portrait relighting technique is instead applied to the output images of our model (c), it produces a less jarring (though still somewhat suboptimal) rendering of the subject (d).

(a) Some fine-detailed shadows (cast by hair) is visible in the output.

(b) Highly symmetric facial shadows are less likely to be softened.

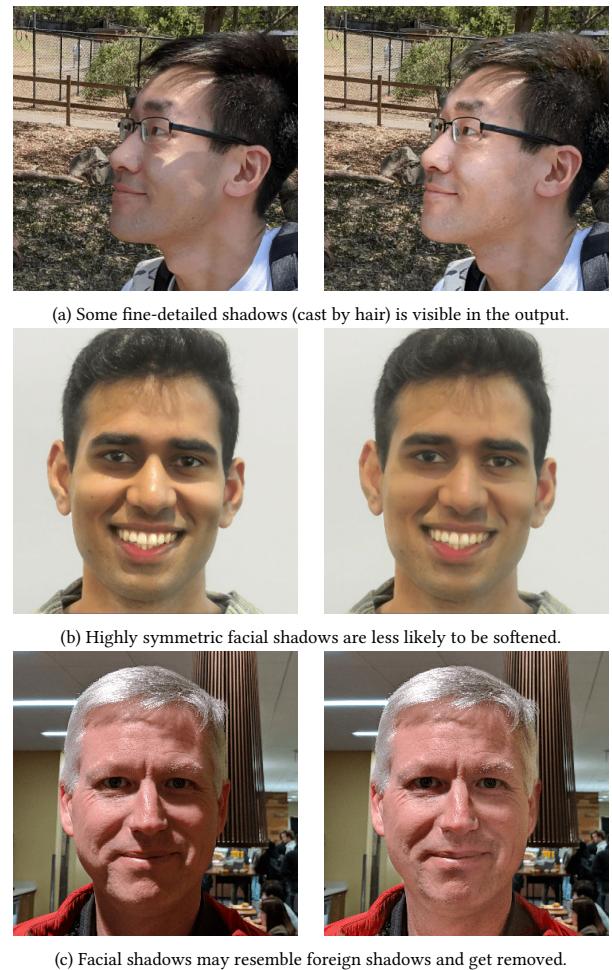(c) Facial shadows may resemble foreign shadows and get removed.

Fig. 12. Example failure cases from our `wild` dataset. We notice limitations of our foreign shadow removal model in handling fine-detailed structures (a), of our facial shadow softening model reducing highly facial shadows (b), and of the models not correctly separating the two types of shadows (c).

this causes our model to sometimes fail to remove shadows that also happen to be bilaterally symmetric (Figure 12(b)). Because the training data of our shadow softening model is rendered by increasing the light size—a simple lighting setup that introduces bias towards generating diffused-looking images. For example, when the "light size" is set high in Figure 10 (c), our shadow softening model generates images with a "flat" appearance and smooths out high frequency details in the hair regions that could have been preserved if different lighting setups are used for face and hair during training data generation.

Our model assumes that shadows belong to one of two categories ("foreign" and "facial") but these two categories are not always entirely distinct and easily-separated. Because of this, sufficiently harsh facial shadows may be erroneously detected and removed by our foreign shadow removal model (Figure 12(c)). This suggests that our model may benefit from a unified approach for both kinds of shadows, though this approach is somewhat at odds with the constraints provided by image formation and our datasets: a unified learning approach would require a unified source of training data, and it is not clear how existing light stage scans or in-the-wild photographs could be used to construct a large, diverse, and photorealistic dataset in which both foreign and facial shadows are present and available as ground-truth.

Constructing a real-world dataset for our task that contains ground-truth is challenging. Though the `foreign-real` dataset used for qualitatively evaluating our foreign shadow removal algorithm is

sufficiently diverse and accurate to evaluate different algorithms, it has some shortcomings. This dataset is not large enough to be used for training, and does not provide a means for evaluating facial shadow softening. This dataset also assumes that all foreign shadows are cast by a single occluder blocking the light of a single dominant illuminant, while real-world instances of foreign shadows often involve multiple illuminants and occluders. Additionally, to satisfy our single-illuminant assumption, this dataset had to be captured in real-world environments that have one dominant light source (*e.g.*, outdoors in the midday sun). This gave us little control over the lighting environment, and resulted in images with high dynamic ranges and therefore "deep" dark shadows, which may degrade (via noise and quantization) image content in shadowed regions. A real-world dataset that addresses these issues be invaluable for evaluating and improving portrait shadow manipulation algorithms.

7 CONCLUSION

We have presented a new approach for enhancing casual portrait photographs with respect to lighting and shadows, particularly in terms of foreign shadow removal, facial shadow softening, and lighting ratio balancing. When synthesizing training data the foreign shadow removal task, we observe the value of in-the-wild images with a shadow synthesis model that accounts for the irregularity of foreign shadows in the real world. Motivated by the physical tools used by photographers in studio environments, we have demonstrated how Light Stage scans can be used to produce training data for facial shadow softening. We have presented a mechanism for allowing convolutional neural networks to exploit the inherent bilateral symmetry of human subjects, and we have demonstrated that this improves the performance of facial shadow softening. Given just a single image of a human subject taken in an unknown and unconstrained environment, our complete system is able to remove unwanted foreign shadows, soften harsh facial shadows, and balance the image's lighting ratio to produce a flattering and realistic portrait image.

ACKNOWLEDGMENTS

REFERENCES

Eli Arbel and Hagit Hel-Or. 2010. Shadow removal using intensity surfaces and texture anchor points. *IEEE TPAMI* (2010).

Masashi Baba and Naoki Asada. 2003. Shadow Removal from a Real Picture. In *SIGGRAPH*.

Jonathan T. Barron and Jitendra Malik. 2015. Shape, Illumination, and Reflectance from Shading. *TPAMI* (2015).

Qifeng Chen and Vladlen Koltun. 2017. Photographic image synthesis with cascaded refinement networks. In *ICCV*.

Yu Chen, Ying Tai, Xiaoming Liu, Chunhua Shen, and Jian Yang. 2018. Fsrnet: End-to-end learning face super-resolution with facial priors. In *CVPR*.

Xiaodong Cun, Chi-Man Pun, and Cheng Shi. 2020. Towards Ghost-free Shadow Removal via Dual Hierarchical Aggregation Network and Shadow Matting GAN. *AAAI*.

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*. 145–156.

Emily L Denton, Soumith Chintala, Rob Fergus, et al. 2015. Deep Generative Image Models Using a Laplacian Pyramid of Adversarial Networks. In *NIPS*.

Bin Ding, Chengjiang Long, Ling Zhang, and Chunxia Xiao. 2019. ARGAN: Attentive Recurrent Generative Adversarial Network for Shadow Detection and Removal. In *ICCV*.

Craig Donner and Henrik Wann Jensen. 2006. A Spectral BSSRDF for Shading Human Skin. (2006).

Randima Fernando. 2004. *GPU Gems: Programming Techniques, Tips and Tricks for Real-Time Graphics.* Pearson Higher Education.

Graham D Finlayson, Mark S Drew, and Cheng Lu. 2009. Entropy minimization for shadow removal. *IJCV* (2009).

Graham D Finlayson, Steven D Hordley, and Mark S Drew. 2002. Removing Shadows from Images. In *ECCV*.

Damien Fourure, Rémi Emonet, Elisa Fromont, Damien Muselet, Alain Tremeau, and Christian Wolf. 2017. Residual Conv-Deconv Grid Network for Semantic Segmentation. In *BMVC*.

Michaël Gharbi, Jiawen Chen, Jonathan T. Barron, Samuel W Hasinoff, and Frédo Durand. 2017. Deep bilateral learning for real-time image enhancement. In *SIGGRAPH*.

Christopher Grey. 2014. *Master lighting guide for portrait photographers.* Amherst Media.

Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. 2009. Ground Truth Dataset and Baseline Evaluations for Intrinsic Image Algorithms. In *ICCV*.

Maciej Gryka, Michael Terry, and Gabriel J. Brostow. 2015. Learning to Remove Soft Shadows. *ACM TOG* (2015).

Ruiqi Guo, Qieyun Dai, and Derek Hoiem. 2012. Paired regions for shadow detection and removal. *TPAMI* (2012).

Pat Hanrahan and Wolfgang Krueger. 1993. Reflection from Layered Surfaces Due to Subsurface Scattering. In *SIGGRAPH*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

Berthold K. P. Horn. 1974. Determining lightness from an image. *Computer Graphics and Image Processing* (1974).

X Hu, CW Fu, L Zhu, J Qin, and PA Heng. 2019. Direction-aware Spatial Context Features for Shadow Detection and Removal. *IEEE transactions on pattern analysis and machine intelligence* (2019).

Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. 2018. Direction-Aware Spatial Context Features for Shadow Detection. In *CVPR*.

Jinggang Huang, Ann B Lee, and David Mumford. 2000. Statistics of range images. In *CVPR*.

Henrik Wann Jensen, Stephen R Marschner, Marc Levoy, and Pat Hanrahan. 2001. A Practical Model for Subsurface Light Transport. In *SIGGRAPH*.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. 2019. Real-time Facial Surface Geometry from Monocular Video on Mobile GPUs. (2019).

Liad Kaufman, Dani Lischinski, and Michael Werman. 2012. Content-Aware Automatic Photo Enhancement. In *Computer Graphics Forum*, Vol. 31. 2528–2540.

Salman H Khan, Mohammed Bennamoun, Ferdous Sohel, and Roberto Togneri. 2015. Automatic Shadow Detection and Removal from a Single Image. *IEEE TPAMI* (2015).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Aravind Krishnaswamy and Gladimir VG Baranoski. 2004. A biophysically-based spectral model of light interaction with human skin. In *Computer Graphics Forum*.

Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. 2012. Interactive facial feature localization. In *ECCV*.

Li-Qian Ma, Jue Wang, Eli Shechtman, Kalyan Sunkavalli, and Shi-Min Hu. 2016. Appearance harmonization for single image shadow removal. In *Computer Graphics Forum*.

Simon Niklaus and Feng Liu. 2018. Context-aware synthesis for video frame interpolation. In *CVPR*.

Simon Niklaus, Long Mai, Jimei Yang, and Feng Liu. 2019. 3D Ken Burns effect from a single image. *ACM TOG* (2019).

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. 2017. DeshadowNet: A Multi-Context Embedding Deep Network for Shadow Removal. In *CVPR*.

Ravi Ramamoorthi and Pat Hanrahan. 2001. A Signal-Processing Framework for Inverse Rendering. In *SIGGRAPH*.

Ronald A Rensink and Patrick Cavanagh. 2004. The Influence of Cast Shadows on Visual Search. *Perception* (2004).

Imari Sato, Yoichi Sato, and Katsushi Ikeuchi. 2003. Illumination from Shadows. *IEEE TPAMI* (2003).

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David W. Jacobs. 2018. SfSNet: Learning Shape, Refectance and Illuminance of Faces in the Wild. In *CVPR*.

Amnon Shashua and Tammy Riklin-Raviv. 2001. The Quotient Image: Class-Based Re-Rendering and Recognition with Varying Illuminations. *IEEE TPAMI* (2001).

YiChang Shih, Sylvain Paris, Connelly Barnes, William T. Freeman, and Frédo Durand. 2014. Style Transfer for Headshot Portraits. *SIGGRAPH* (2014).

Yael Shor and Dani Lischinski. 2008. The shadow meets the mask: Pyramid-based shadow removal. In *Computer Graphics Forum*.

Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017. Portrait lighting transfer using a mass transport approach. In *SIGGRAPH*.

Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E. Debevec, and Ravi Ramamoorthi. 2019. Single Image Portrait Relighting. *SIGGRAPH* (2019).

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep image prior. In *CVPR*.

Jifeng Wang, Xiang Li, and Jian Yang. 2018. Stacked Conditional Generative Adversarial Networks for Jointly Learning Shadow Detection and Shadow Removal. In *CVPR*.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE TIP* (2004).

Tai-Pang Wu, Chi-Keung Tang, Michael S Brown, and Heung-Yeung Shum. 2007. Natural shadow matting. *ACM TOG* (2007).

Raymond A Yeh, Chen Chen, Teck Yian Lim, Alexander G Schwing, Mark Hasegawa-Johnson, and Minh N Do. 2017. Semantic image inpainting with deep generative models. In *CVPR*.

Ling Zhang, Qingan Yan, Yao Zhu, Xiaolong Zhang, and Chunxia Xiao. 2019b. Effective Shadow Removal Via Multi-Scale Image Decomposition. *The Visual Computer* (2019).

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018a. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*.

Xuaner Zhang, Kevin Matzen, Vivien Nguyen, Dillon Yao, You Zhang, and Ren Ng. 2019a. Synthetic defocus and look-ahead autofocus for casual videography. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 1–16.

Xuaner Zhang, Ren Ng, and Qifeng Chen. 2018b. Single Image Reflection Removal with Perceptual Losses. In *CVPR*.

Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. 2019. Distraction-aware shadow detection. In *CVPR*. 5167–5176.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. 2019. Deep Single-Image Portrait Relighting. In *ICCV*.

Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. 2018. Bidirectional Feature Pyramid Network with Recurrent Attention Residual Modules for Shadow Detection. In *ECCV*.