COMPOSE: Comprehensive Portrait Shadow Editing

Andrew Hou[1,2], Zhixin Shu[2], Xuaner Zhang[2], He Zhang[2], Yannick Hold-Geoffroy[2], Jae Shin Yoon[2], and Xiaoming Liu[1]

[1] Michigan State University
{houandr1,liuxm}@msu.edu
[2] Adobe Research
{zshu,cezhang,hezhan,holdgeof,jaeyoon}@adobe.com

Abstract. Existing portrait relighting methods struggle with precise control over facial shadows, particularly when faced with challenges such as handling hard shadows from directional light sources or adjusting shadows while remaining in harmony with existing lighting conditions. In many situations, completely altering input lighting is undesirable for portrait retouching applications: one may want to preserve some authenticity in the captured environment. Existing shadow editing methods typically restrict their application to just the facial region and often offer limited lighting control options, such as shadow softening or rotation. In this paper, we introduce COMPOSE: a novel shadow editing pipeline for human portraits, offering precise control over shadow attributes such as shape, intensity, and position, all while preserving the original environmental illumination of the portrait. This level of disentanglement and controllability is obtained thanks to a novel decomposition of the environment map representation into ambient light and an editable gaussian dominant light source. COMPOSE is a four-stage pipeline that consists of light estimation and editing, light diffusion, shadow synthesis, and finally shadow editing. We define facial shadows as the result of a dominant light source, encoded using our novel gaussian environment map representation. Utilizing an OLAT dataset, we have trained models to: (1) predict this light source representation from images, and (2) generate realistic shadows using this representation. We also demonstrate comprehensive and intuitive shadow editing with our pipeline. Through extensive quantitative and qualitative evaluations, we have demonstrated the robust capability of our system in shadow editing.

Keywords: Face Relighting · Shadow Editing · Lighting Decomposition

1 Introduction

Single-image portrait relighting is a well-studied problem with steady interest from the computer vision and graphics communities. Portrait relighting has a

a) Source Image | b) Soften Shadow | c) Intensify Shadow | d) Modify Light Size | e) Rotate Shadow

Fig. 1: Overview. COMPOSE is the first single-image portrait shadow editing method that achieves complete shadow editing control, *i.e.* adjusting shadow intensity, modifying light size, and changing shadow positions, all while preserving the source image's other lighting attributes (*e.g.* ambient light).

wide range of applications in computational photography [31], AR/VR [4,5,19], and downstream tasks in the face domain [11,25]. While many methods for single image portrait relighting have been proposed in recent years [17,18,29,31,33,41, 49,54], they often face challenges when dealing with facial shadows.

One popular genre of face relighting methods [31, 41] uses low resolution HDR environment maps to represent lighting. This lighting representation is used in a conditional image translation framework to control the lighting effect of the output face image, thereby achieving facial relighting. The requirement for an environment map restricts users seeking to fine-tune lighting, such as minor modifications to shadow smoothness or direction. Additionally, a low-resolution environment map often lacks the precision needed for accurately depicting complex, high-frequency shadow effects. Conversely, several approaches use low-dimensional spherical harmonics (SH) for their lighting representation [37, 54]. These methods, based on simplified lighting and reflectance models, frequently produce unrealistic relit images. Moreover, they overlook occlusions in the image formation process, disallowing the generation and editing of cast shadows. SunStage [44] generates realistic shadows on faces but lacks generalizability and requires videos as input. Some recent works have focused on shadow editing tasks [11,26,53] using state-of-the-art image generators such as diffusion models. However, achieving precise control with these sophisticated generators remains a challenging task. Consequently, these methods often provide limited control in shadow synthesis, typically handling only shadow softening or completely removing the shadow, without precise editing of shadow intensity, shape, or position.

To overcome the challenges in modeling and controlling facial shadows in portrait relighting, we introduce COMPOSE, a system that separates shadow modeling from ambient light in the lighting representation. COMPOSE enables flexible manipulation of shadow attributes, including position (lighting direction), smoothness, and intensity (See Fig. 1). The key innovation in COMPOSE is its shadow representation, which is compatible with environment map-based representations, making it easily integrable into existing relighting networks. Building on the concept of image-based lighting representations (*e.g.* HDR en-

vironment maps), our approach divides lighting effects into two components: shadows caused by dominant light sources and ambient lighting. We simplify the shadow component by attributing it to a single dominant light source on an environment map. This is encoded using a 2D isotropic Gaussian, with variables representing light position, standard deviation (indicating light size or diffusion), and scale (light intensity). The remaining lighting effects are attributed to a diffused environment map, modeling the image's ambient lighting.

For a given face image, we first predict the lighting as a panorama environment map, from which we estimate the dominant light (shadow) parameters. The system then removes shadows to obtain a diffused, albedo-like image. Utilizing this diffuse image and the shadow representation, we train a network to synthesize shadows on the subject in the input image. This shadow synthesis process is adaptable, accepting various shadow-related parameters for controllable shadow synthesis, including position and shape. The synthesized shadow can be linearly blended with the diffuse lighting to create shadows of varying intensities.

To train models for shadow diffusion and synthesis, we utilize an OLAT (one-light-at-a-time) dataset, captured through a light stage system. This dataset, combined with the environment map-based lighting representation, ensures our system's compatibility with previously proposed face relighting techniques. Our approach also uniquely offers the capability to accurately model shadows.

Our contributions are thus as follows:

◇ We propose the first single image portrait shadow editing method with precise control over various shadow attributes, including shadow shape, intensity, and position. Our system provides users with the flexibility to control shading while also preserving the other lighting attributes of the input images.

◇ We propose a novel lighting decomposition into ambient light and an editable dominant gaussian light source, which enables disentangled shadow editing while preserving the remaining lighting attributes.

◇ We achieve state-of-the-art relighting performance quantitatively and qualitatively, including on a light-stage-captured test set and in-the-wild images.

2 Related Work

2.1 Portrait Relighting

Many portrait relighting methods have been proposed in recent years that generally fall into one of two categories: focusing on reducing the data requirement and improving the relighting quality as much as possible with limited data [6, 10, 17, 18, 20, 24, 33, 34, 36–39, 43–45, 47, 49, 54] or using light stage data but modeling more complex lighting effects [4, 12, 28, 29, 31, 41, 42, 48, 50, 52]. Among single image portrait relighting methods, a substantial limitation is that they cannot preserve the original environmental lighting while editing shadows, and instead opt to relight by replacing the original lighting with a completely new environment [11], which is undesirable in many portrait editing applications. Existing methods also struggle to handle more directed lights and fail in the presence of substantial hard or intense shadows [37, 41, 54].

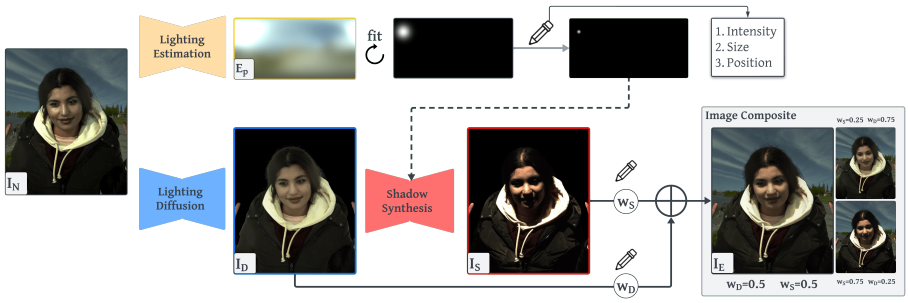

Fig. 2: Method Overview. COMPOSE is a 4-stage shadow editing method consisting of single-image lighting estimation, light diffusion, shadow synthesis, and image compositing. By first estimating the dominant light position using the environment map regressed from the input image, COMPOSE can control shadow shape and intensity by controlling light spread and light intensity respectively as well as shadow position by changing the location of the dominant light. By estimating a diffuse image $\mathbf{I}_D$ as well as a shadowed image $\mathbf{I}_S$, COMPOSE can generate the final edited image $\mathbf{I}_E$ through image compositing.

Among portrait relighting methods that take multiple images or videos as input, the method SunStage [44] is able to maintain the lighting attributes of the input video's environment when performing relighting. However, the method must be trained using a video of a subject on a sunny day and involves an expensive optimization procedure. The model is also subject-specific and must be retrained or fine-tuned even for the same subject if they change their hairstyle, accessories, or clothing. COMPOSE has the advantage of being a single-image portrait relighting method, and is both generalizable and practical for in-the-wild photo editing applications.

2.2 Shadow Editing

Another line of work closely related to COMPOSE are methods that edit only facial shadows, including shadow removal [26, 53] and shadow softening [11]. Most similar to our work is the method of [11], which is able to directly control the degree of shadow softening while preserving the original environmental lighting. However, unlike COMPOSE, they are restricted to only shadow softening and cannot handle other shadow editing operations like shadow intensifying, modifying shadow shape, and shadow rotation. Shadow removal methods [14, 26, 53] are often even more limited since they cannot control the degree of shadow softening and attempt to remove the shadow completely.

2.3 Lighting Representations

Existing portrait relighting methods generally utilize one of three lighting representations: Spherical Harmonics (SH) [18, 33, 34, 37, 54], dominant lighting direction [17, 29], and environment maps [31, 41, 49]. SH lighting tends to only model

the lower frequencies of lighting effects, and thus is unsuitable for modeling non-Lambertian hard shadows and specularities and is better suited for diffuse lightings. Representing the light as a dominant lighting direction can often handle the non-Lambertian effects of directional and point light sources, but they often leave other lighting attributes such as light size out of the equation (*e.g.* Hou *et al.* [17]). Moreover, they often do not handle diffuse lightings well and are intended for directional lights [17, 29]. The environment map representation is more flexible and is able to model light intensity, light size, and light position and can handle both diffuse and directional lights. However, what is largely missing in this domain is a way to properly decompose the environment map that enables precise controllability over different lighting components. Existing methods only perform relighting by completely changing the scene illumination with a new environment map without disentangling the effects of ambient, diffuse, and directional light [31, 41, 49]. This can be undesirable for many computational photography applications where the user wishes to preserve the ambience of the scene and edit only the facial shadow alone [11]. Moreover, the default environment map representation is unsuitable for some shadow editing applications such as shrinking the light. While enlarging the light can be performed by applying a gaussian blur, there is no equivalent operation that can be performed to shrink the light on an environment map. Our lighting representation decomposes the environment map into the ambient light and an editable gaussian dominant light source, which is adaptable in light intensity, size, and position. COMPOSE is thus able to perform a full suite of shadow editing operations thanks to its editable gaussian component and can edit the shadow alone without disrupting other lighting attributes (*e.g.* ambient light).

3 Methodology

3.1 Problem Formulation

In our framework, shadows are treated as a composable lighting effect, which can be independently predicted and altered based on a specific face image. This shadow representation is integrated into the overall lighting representation, and our framework is designed to (1) accurately predict the shadow from an image, and (2) apply a controllable shadow onto a shadow-free face image. We define the portrait shadow editing problem based on properties of shadows that users can manipulate, namely shadow position, shadow intensity, and shadow shape. These shadow properties are connected to the lighting attribute that is easily controllable. Specifically, given source image $\mathbf{I}_N$ and desired edited output image $\mathbf{I}_E$ with modified shadows, our problem can be written as:

$$\mathbf{I}_E = F_\theta(\mathbf{I}_N, x, y, \sigma, \gamma), \tag{1}$$

where F_θ is our shadow editing method, x and y are the desired light position on an environment map, σ is the light size, and γ is the light intensity. By editing the parameters, (x, y), σ, and γ, we can control the shadow position, shape, and intensity respectively. We describe our solution COMPOSE in detail in Sec. 3.2.

3.2 COMPOSE Framework

COMPOSE is a 4-stage shadow editing framework, consisting of single image lighting estimation, light diffusion, shadow synthesis, and finally image compositing. See Fig. 2 for an overview of our method.

Lighting Estimation In the lighting estimation stage, we train a variational autoencoder (VAE) [23] to estimate an environment map (LDR) from a single portrait image as its lighting representation $\mathbf{I}_N$. From this environment map, we then estimate the dominant light position. Naturally, the dominant light, such as the sun, represents the brightest intensity on a image-based lighting representation. Therefore we can simply obtain the dominant light position by fitting a 2D isotropic Gaussian on the predicted environment map. The center of the Gaussian will indicate the position of the dominant light on the 2D environment map. This estimated light position corresponds to the shadow information on the input image, and therefore is useful for "in-place" shadow editing (*e.g.*, shadow softening, shadow intensifying) in the later stage that does not alter the direction of the dominant light. We argue this is desirable in many portrait editing applications where the user only wants to edit existing shadows and does not want to alter the ambient lighting.

Light Diffusion In the light diffusion stage, we remove all existing hard shadows and specular highlights from the input image and output a "diffused" image $\mathbf{I}_D$ with very smooth shading. This represents the input face under only an ambient illumination condition. Our light diffusion network consists of a hierarchical transformer encoder that generates multi-level features, which are fed to a decoder with transposed convolutional layers. Our network takes as input the original input image, a body parsing mask, and a binary foreground mask.

Shadow Synthesis In the shadow synthesis stage, we aim to produce a shadowed image $\mathbf{I}_S$ illuminated by an edited light source given the diffuse image $\mathbf{I}_D$ estimated from the light diffusion stage and an edited environment map. Using the estimated dominant light position from the light estimation stage, we have full control over light attributes such as light size, intensity, and the dominant light position. We can produce a new environment map using a Gaussian to represent the dominant light, where the light size is modeled by the standard deviation of the Gaussian (a larger standard deviation represents a larger, more diffuse light) and the light intensity can be adjusted by multiplying the Gaussian with a scalar. This allows our method to control shadow shape, light intensity, and position. To regress the shadowed image $\mathbf{I}_S$, we propose a two-stage architecture: a U-Net followed by a conditional DDPM (See Fig. 3). For the first stage, we adopt the U-Net architecture of [41]. During training, instead of using a reshaped environment map-like representation for the dominant light, we design a feature map-like representation that's parameterized by the shadow attributes. Specifically, the lighting representation for shadows has 4 parameters. The first two parameters x and y encode the coordinates of the light center on the environment map, the third parameter σ encodes the light size, and the fourth parameter γ encodes the light intensity. All four parameters are normalized between 0 and 1 and repeated spatially as 32×32 channels to be suitable

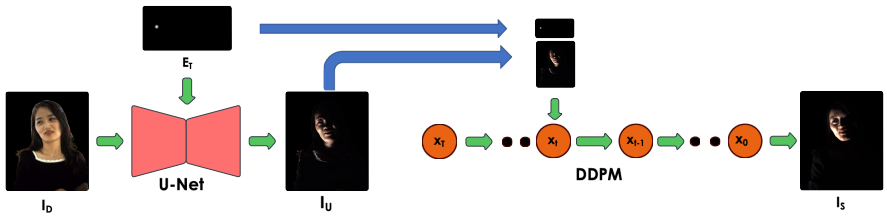

Fig. 3: Shadow Synthesis Pipeline. Our shadow synthesis step consists of two stages: a U-Net followed by a conditional DDPM. The U-Net takes ambient image $\mathbf{I}_D$ and environment map $\mathbf{E}_T$ as input and outputs the coarse relit image $\mathbf{I}_U$. With $\mathbf{I}_U$ and $\mathbf{E}_T$ as spatial conditions, the conditional DDPM outputs the final shadowed image $\mathbf{I}_S$, refining shadow boundaries and improving image quality.

inputs for the U-Net. We find that this lighting representation leads to faster convergence, especially when the desired light mimics a point light and occupies a tiny portion of the environment map. We feed the diffuse image $\mathbf{I}_D$ and our desired lighting feature map to the U-Net, and the output is a relit image $\mathbf{I}_U$. While the U-Net architecture has merits such as complete geometric consistency for shadow synthesis, one demerit is that the synthesized shadow boundaries are often not very sharp and the overall image quality could be improved. We thus employ a DDPM [7,15] as the second stage of our shadow synthesis model. The role of the DDPM here is to take the output image of the U-Net $\mathbf{I}_U$ as a spatial condition along with the lighting parameters and perform image refinement to generate the final shadowed image $\mathbf{I}_S$. The training objective of our shadow synthesis DDPM thus becomes:

$$\mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t, \mathbf{I}_U, x, y, \sigma, \gamma)\|^2\right]. \tag{2}$$

Our condition $\mathbf{I}_U$ is spatially concatenated with $\mathbf{x}_t$ and our lighting parameters are repeated spatially as channels of the same resolution and similarly spatially concatenated, where $\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}$. We find that the quality of the edited images is noticeably enhanced by the DDPM compared to the U-Net alone and that adding the DDPM sharpens the shadow boundaries.

Image Compositing With $\mathbf{I}_D$ from our light diffusion stage and $\mathbf{I}_S$ from our shadow synthesis stage, we can then perform image compositing between $\mathbf{I}_D$ and $\mathbf{I}_S$ to achieve our final edited image $\mathbf{I}_E$, by assigning weights ω_D and ω_S to $\mathbf{I}_D$ and $\mathbf{I}_S$ respectively. The final edited image $\mathbf{I}_E$ is thus generated as:

$$\mathbf{I}_E = \omega_D \mathbf{I}_D + \omega_S \mathbf{I}_S, \tag{3}$$

where $\omega_D + \omega_S = 1$. By adjusting ω_D and ω_S, we can further tune shadow intensity: a higher ω_D softens shadows and a higher ω_S results in darker shadows.

3.3 Loss Functions

To train our single image lighting estimation network, we apply two loss functions commonly used in VAE training [23]: $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{KLD}}$. $\mathcal{L}_{\text{recon}}$ is a reconstruction loss between the predicted environment map $\mathbf{E}_P$ and the groundtruth

environment map $\mathbf{E}_G$ defined as follows:

$$\mathcal{L}_{\text{recon}} = \frac{1}{3HW}(\|\mathbf{E}_P - \mathbf{E}_G\|_2^2), \tag{4}$$

where H and W are the size of the environment maps. The Kullback–Leibler divergence (KLD) loss $\mathcal{L}_{\text{KLD}}$ is:

$$\mathcal{L}_{\text{KLD}} = -\frac{1}{2N}\sum_{i=1}^{N}(1 + \log(\sigma_i^2) - \mu_i^2 - \sigma_i^2), \tag{5}$$

where N is the dimensionality of the latent vector, μ is the batch mean, and σ^2 is the batch variance. The final loss for our lighting estimation network is thus:

$$\mathcal{L}_{\text{LE}} = \lambda_1\mathcal{L}_{\text{recon}} + \lambda_2\mathcal{L}_{\text{KLD}}, \tag{6}$$

where $\lambda_1 = 1$ and $\lambda_2 = 2.5 \times 10^{-4}$.

Our light diffusion network is also trained with two losses $\mathcal{L}_{\text{recon}}$ and $\mathcal{L}_{\text{perceptual}}$.

$$\mathcal{L}_{\text{recon}} = \frac{1}{3HW}(\|\mathbf{I}_D - \mathbf{I}_G\|_1), \tag{7}$$

where H and W are the size of the training images, $\mathbf{I}_D$ is the predicted diffuse image, and $\mathbf{I}_G$ is the groundtruth diffuse image. $\mathcal{L}_{\text{perceptual}}$ enforces visual similarity [51] and is computed as the distance between VGG [40] features computed for $\mathbf{I}_D$ and $\mathbf{I}_G$. The final loss for the light diffusion network is thus:

$$\mathcal{L}_{\text{LD}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{perceptual}}. \tag{8}$$

When training the U-Net of our shadow synthesis pipeline, we employ a single reconstruction loss $\mathcal{L}_{\text{recon}}$ between the predicted relit image $\mathbf{I}_S$ and the groundtruth $\mathbf{I}_G$, defined as follows:

$$\mathcal{L}_{\text{recon}} = \frac{1}{3HW}(\|\mathbf{I}_S - \mathbf{I}_G\|_2^2). \tag{9}$$

3.4 Data Collection and Generation

To train out networks, we use a light stage to capture one-light-at-a-time (OLAT) images of 107 subjects with diverse skin tones, body poses, expressions, and accessories using a light stage [9] with 160 lights. Each subject is captured under 10-20 sessions with varied body poses, expressions, and accessories. Each session is captured by 4 cameras with varying camera poses.

To generate input images to train our lighting estimation VAE and our light diffusion network, we render our OLAT images using a diverse set of environment maps collected from the Polyhaven and Laval Outdoor HDR datasets [16]. During rendering, we apply random augmentations to the environment map including rotations and adjusting intensity to improve our model's generalization to different lighting. The groundtruth $\mathbf{E}_G$ for our lighting estimation VAE is

simply the environment map used for rendering, and the groundtruth $\mathbf{I}_G$ for the light diffusion network is generated by rendering OLAT images with heavily blurred, diffuse environment maps.

To generate images for our shadow synthesis stage, we render groundtruth relit images by using environment maps consisting of Gaussian lights with various positions, intensities, and sizes (See Sec. 3.2) along with our OLAT images. As the input to the shadow synthesis stage is meant to be the diffuse image $\mathbf{I}_D$ predicted by the light diffusion stage, we generate diffuse images for training our shadow synthesis networks by rendering OLAT images with heavily blurred environment maps in the Polyhaven and Laval Outdoor HDR datasets.

3.5 Implementation Details

We implement all components of COMPOSE using PyTorch [32]. When training our lighting estimation VAE, we use a batch size of 128 and a learning rate of 10^{-4} on 4 24GB A10G GPUs. For our light diffusion network, we train with a batch size of 88 and a learning rate of 10^{-4} using 8 A100 GPUs. For our shadow synthesis U-Net, we train with a batch size of 100 and a learning rate of 10^{-4} using 4 24GB A10G GPUs. For all networks, we train using the Adam Optimizer [22]. The image resolution used to train the lighting estimation and shadow synthesis networks is 256×256 and the resolution used to train the light diffusion network is 768×768. Please see the supplementary materials for architectural details of the lighting estimation network, the lighting diffusion network, and the shadow synthesis network.

4 Experiments

4.1 Datasets

To train our lighting estimation and shadow synthesis networks, we generate our training sets using light stage images and 521 outdoor environment maps. Our training set consists of 82 subjects with significant variation in factors such as skin tones, pose, expressions, and accessories.

For evaluation, we use two separate evaluation sets: the first is generated using our light stage images and has corresponding groundtruth, and the second consists of in-the-wild images that represent more realistic application settings and do not have groundtruth. Our light stage evaluation set consists of 48 images from a total of 16 subjects with diverse skin tones, body poses, expressions, and accessories. All evaluation subjects and lightings are held out and unseen during training. Our in-the-wild evaluation set also consists of unseen test subjects and applies unseen lighting conditions in all qualitative results.

4.2 Shadow Editing Evaluation

We evaluate our shadow editing performance (synthesis and removal) by using randomly sampled, unseen lighting positions as the target lights to generate relit light stage images of our 16 test subjects, which serve as groundtruth.

a) Source b) Target c) Ours d) TR [31] e) [17]

Fig. 4: Shadow Synthesis. Our model synthesizes more plausible shadows than the baselines, and, unlike [17], is also able to properly remove shadows from the source image before relighting. Environment maps are shown to help visualize each test lighting.

Table 1: Shadow Editing Performance. We evaluate shadow editing using unseen lightings with varied intensity and size on 16 held out highly diverse light stage test subjects with groundtruth relit images. COMPOSE outperforms all baselines quantitatively in shadow editing performance across all metrics (mean ± standard deviation).

	MAE	MSE	SSIM	LPIPS
Hou *et al.* [17]	0.1172 ± 0.0754	0.0438 ± 0.0503	0.7059 ± 0.1061	0.2371 ± 0.0871
Total Relighting [31]	0.1008 ± 0.0743	0.0379 ± 0.0447	0.7754 ± 0.0738	0.2120 ± 0.0572
COMPOSE (Ours)	$\mathbf{0.0965 \pm 0.0686}$	$\mathbf{0.0349 \pm 0.0368}$	$\mathbf{0.7780 \pm 0.0826}$	$\mathbf{0.1973 \pm 0.0678}$

Each input image is rendered with a randomly selected environment map out of 125 unseen outdoor environment maps and is first passed to our deshading network to generate diffuse image $\mathbf{I}_D$. $\mathbf{I}_D$ and target environment map $\mathbf{E}_T$ are then fed to our two-stage shadow synthesis pipeline to generate the newly shadowed image $\mathbf{I}_S$, which is evaluated against the groundtruth light stage image.

Tab. 1 compares our shadow editing performance against prior relighting methods. For both baselines [17, 31], the test results were provided by the authors. Our method achieves state-of-the-art results on all metrics (MAE, MSE, SSIM [46], and LPIPS [51]). As seen in Fig. 4, our model can synthesize appropriate shadows for various light positions. The method of Hou *et al.* [17] cannot

a) Source Image | b) Soften Shadows | c) Intensify Shadows | d) Modify Light Size | e) Rotate Shadows

Fig. 5: Shadow Editing. Our method achieves complete shadow editing control, including softening/intensifying shadows, changing light size to alter shape and intensity, and changing light position, all while preserving the source image's ambient light.

remove existing shadows in the source image and the shadow traces carry over as artifacts to the relit images. In addition, Hou *et al.* [17] only models the lighting direction and does not model the light size as a parameter, which leads to inaccurate shadow shape when the light size is varied. Total Relighting [31] is often unable to synthesize physically plausible shadows, and will sometimes overshadow (Fig. 4, rows 1, 3, and 4) or undershadow the image (Fig. 4, row 2). Moreover, their shadows are often blurry and not as sharp as our method. Compared to the baselines, COMPOSE can properly remove existing shadows from the source image and synthesize geometrically plausible and realistic shadows for a wide variety of lighting conditions.

4.3 Portrait Shadow Editing Applications

We demonstrate the full performance of COMPOSE by performing all forms of shadow editing, including shadow softening, shadow intensifying, modifying light size (which changes both shadow shape and intensity), and shadow rotation

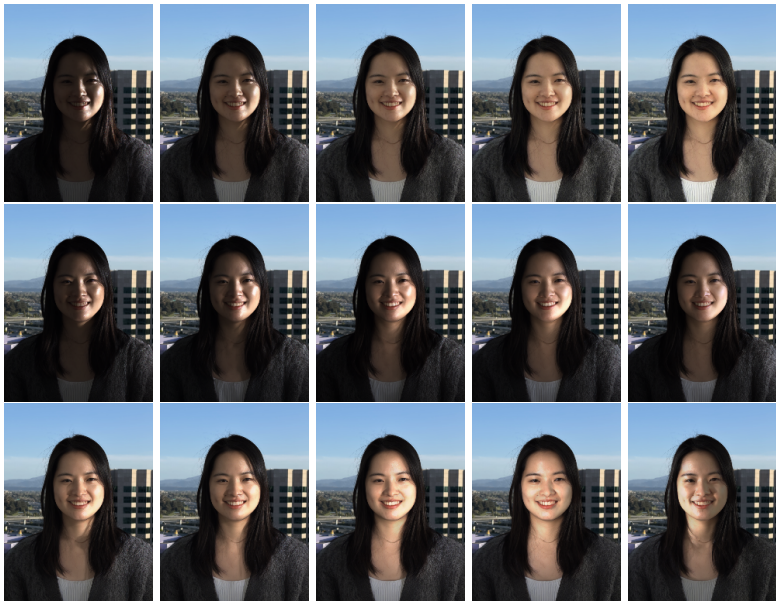

Fig. 6: Degree of Shadow Editing. We demonstrate COMPOSE's controllable shadow editing by varying shadow intensity (top row), light diffusion for modifying shadow shape and intensity (middle row), and light position (bottom row).

by changing the dominant light position. For each image, we first estimate the diffuse image $\mathbf{I}_D$ and the shadowed image $\mathbf{I}_S$ and perform image compositing between these two images to generate the final edited image $\mathbf{I}_E$. As seen in Fig. 5, COMPOSE properly edits the shadows of in-the-wild images. To soften or intensify the shadow without changing the shadow shape, we rely on our image compositing coefficients ω_d and ω_s, which control the contributions of $\mathbf{I}_D$ and $\mathbf{I}_S$ respectively. We increase the weight of ω_d to soften the shadow and increase the weight of ω_s to strengthen the shadow. We further show that we can modify light size by tuning the Gaussian spread σ, which changes both shadow shape and intensity. Enlarging the light makes the shadow more diffuse and less intense, whereas shrinking the light generates a harder, more intense shadow. For the subjects in rows 2, 3, and 4 we enlarged the light, resulting in a more diffuse shadow especially noticeable around the nose. For the subject in row 1 we instead shrunk the light and appropriately produce an edited image with more prominent and intense hard shadows. Finally, we also demonstrate shadow rotation by changing the dominant light position in $\mathbf{I}_S$'s environment map.

To demonstrate COMPOSE can properly control shadow editing, we visualize various degrees of shadow softening/intensifying, light diffusion, and shadow rotation in Fig. 6. The first row shows varying shadow intensity, where the shadows become less intense from left to right using ω_d and ω_s. Our results demonstrate that these flexible weights allow for a wide range of shadow intensities. The second row shows increasing degrees of light diffusion from left to right, which corresponds to increasing the light size. The shadows in the generated images

a) Source b) 0.75 c) 0.5 d) 0.2 e) 0.0

Fig. 7: Controllable Shadow Softening. For each subject, we show the shadow softening results of [11] in the first row and COMPOSE in the second. COMPOSE leaves less shadow traces at level 0.0, where the goal is to remove the shadows completely.

correctly transition from harder shadows to more diffuse shadows as the light size increases. The third row shows shadow rotation, where the light is rotated horizontally along the environment map and around the subject. COMPOSE is able to produce physically plausible shadows as the light position changes.

It is important to note that this level of controllable shadow editing is not achievable by existing portrait shadow editing methods [11, 53] that are only able to handle shadow softening or complete shadow removal as well as existing portrait relighting methods that completely modify the existing environment and cannot preserve the original scene illumination [31], do not properly model all shadow attributes (*e.g.* intensity, shape, and position) [17,33] or are not designed to handle lighting conditions with more substantial shadows [37, 41, 54].

4.4 Shadow Softening Comparison

We compare on in-the-wild shadow softening with the state-of-the-art portrait shadow softening method Futschik *et al.* [11]. Using results provided by the authors, we visualize different degrees of shadow softening for [11] by varying the light diffusion parameter (0.75, 0.5, 0.2, and 0.0) and compare with equivalent degrees of shadow softening by COMPOSE, as shown in Fig. 7. Higher parameter values correspond to less softening whereas lower values correspond to more softening. At 0.0, the goal is to completely remove the source image's shadows. Across all degrees of light diffusion, both methods gradually soften the shadow, but COMPOSE leaves less shadow traces than [11] when completely removing

the source image's shadows at level 0.0, especially for images with darker shadows. This is thanks to our hierarchical transformer design for our light diffusion network, which better handles removing the effects of shadows at all scales.

5 Conclusion

We have proposed COMPOSE: the first single image portrait shadow editing method that can perform comprehensive variations of shadow editing including editing shadow intensity, light size, and shadow position while preserving the source image's lighting environment. This is largely enabled by our novel lighting decomposition into ambient light and an editable gaussian dominant light, which enables new applications like shrinking the light and intensifying shadows as well as disentangled shadow editing that preserves the scene ambience. We have demonstrated the novel photo editing applications enabled by our method over prior work qualitatively and that our method achieves state-of-the-art shadow editing performance quantitatively over prior relighting methods. We hope that our work can inspire and motivate more research in the exciting new research area of controllable and comprehensive portrait shadow editing.

Limitations. Good directions for future work include handling lighting environments with multiple light sources and indoor environments, as our framework is intended mostly for outdoor settings, where the sun is the single dominant light. In addition, while $\mathbf{I}_D$ can preserve the ambience, we still notice there are sometimes color shifts with the environmental lighting during shadow editing, largely due to the compositing step with shadowed image $\mathbf{I}_S$ (which adopts the light stage's environmental lighting). This could be alleviated by modeling the diffuse lighting tint similar to [11] or by training with white balanced data to remove the color bias of the light stage. We also believe that one can continue to improve the generalizability of our model. Similar to other relighting methods that rely on light stage data [11, 31, 41], we train with a limited number of subjects. Future work may incorporate synthetic data to expand the training set or adopt latent diffusion models, which naturally have high generalizability [35]. Finally, one can further improve the identity preservation of our method. We find that high frequency details are sometimes lost at the initial U-Net stage of shadow synthesis, and ways to better preserve the identity include passing $\mathbf{I}_D$ or the input image $\mathbf{I}_N$, which contain high frequency details, as additional conditions to the DDPM or to use stronger backbones than [41].

Potential Societal Impacts. In image editing, there are always concerns of generating malicious deepfakes. However, we only alter illumination, not subject identity. Even so, there may be concerns that altering the illumination and especially adding dark shadows to faces could be used as a malicious attack to worsen the performance of downstream tasks such as face recognition. Users can detect these attacks using existing deepfake detection methods [1–3, 8, 13, 21, 27, 30] or by training one using the edited images generated by COMPOSE. State-of-the-art shadow removal methods [26,53] are also an option to counter these potential attacks since they can remove any facial shadows that COMPOSE synthesizes.

References

1. Aghasanli, A., Kangin, D., Angelov, P.: Interpretable-through-prototypes deepfake detection for diffusion models. In: ICCVW (2023)
2. Asnani, V., Yin, X., Hassner, T., Liu, X.: MaLP: Manipulation localization using a proactive scheme. In: CVPR (2023)
3. Asnani, V., Yin, X., Hassner, T., Liu, X.: Reverse engineering of generative models: Inferring model hyperparameters from generated images. PAMI (2023)
4. Bi, S., Lombardi, S., Saito, S., Simon, T., Wei, S.E., McPhail, K., Ramamoorthi, R., Sheikh, Y., Saragih, J.: Deep relightable appearance models for animatable faces. TOG (2021)
5. Caselles, P., Ramon, E., Garcia, J., i Nieto, X.G., Moreno-Noguer, F., Triginer, G.: SIRA: Relightable avatars from a single image. In: WACV (2023)
6. Chandran, S., Hold-Geoffroy, Y., Sunkavalli, K., Shu, Z., Jayasuriya, S.: Temporally consistent relighting for portrait videos. In: WACV (2022)
7. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: ILVR: Conditioning method for denoising diffusion probabilistic models. In: ICCV (2021)
8. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: ICASSP (2023)
9. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: SIGGRAPH (2000)
10. Ding, Z., Zhang, C., Xia, Z., Jebe, L., Tu, Z., Zhang, X.: DiffusionRig: Learning personalized priors for facial appearance editing. In: CVPR (2023)
11. Futschik, D., Ritland, K., Vecore, J., Fanello, S., Orts-Escolano, S., Curless, B., Sýkora, D., Pandey, R.: Controllable light diffusion for portraits. In: CVPR (2023)
12. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., DuVall, M., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P., Izadi, S.: The Relightables: Volumetric performance capture of humans with realistic relighting. In: SIGGRAPH Asia (2019)
13. Guo, X., Liu, X., Ren, Z., Grosz, S., Masi, I., Liu, X.: Hierarchical fine-grained image forgery detection and localization. In: CVPR (2023)
14. He, Y., Xing, Y., Zhang, T., Chen, Q.: Unsupervised portrait shadow removal via generative priors. In: MM (2021)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS (2020)
16. Hold-Geoffroy, Y., Athawale, A., Lalonde, J.F.: Deep sky modeling for single image outdoor lighting estimation. In: CVPR (2019)
17. Hou, A., Sarkis, M., Bi, N., Tong, Y., Liu, X.: Face relighting with geometrically consistent shadows. In: CVPR (2022)
18. Hou, A., Zhang, Z., Sarkis, M., Bi, N., Tong, Y., Liu, X.: Towards high fidelity face relighting with realistic shadows. In: CVPR (2021)
19. Iqbal, U., Caliskan, A., Nagano, K., Khamis, S., Molchanov, P., Kautz, J.: RANA: Relightable articulated neural avatars. In: ICCV (2023)
20. Ji, C., Yu, T., Guo, K., Liu, J., Liu, Y.: Geometry-aware single-image full-body human relighting. In: ECCV (2022)
21. Kamat, S., Agarwal, S., Darrell, T., Rohrbach, A.: Revisiting generalizability in deepfake detection: Improving metrics and stabilizing transfer. In: ICCVW (2023)
22. Kingma, D., Adam, J.B.: A method for stochastic optimization. In: ICLR (2014)
23. Kingma, D.P., Welling, M.: Auto-encoding variational bayes . In: ICLR (2014)

24. Lagunas, M., Sun, X., Yang, J., Villegas, R., Zhang, J., Shu, Z., Masia, B., Gutierrez, D.: Single-image full-body human relighting. In: EGSR (2021)
25. Le, H., Kakadiaris, I.: Illumination-invariant face recognition with deep relit face images. In: WACV (2019)
26. Liu, Y., Hou, A., Huang, X., Ren, L., Liu, X.: Blind removal of facial foreign shadows. In: BMVC (2022)
27. Lorenz, P., Durall, R.L., Keuper, J.: Detecting images generated by deep diffusion models using their local intrinsic dimensionality. In: ICCVW (2023)
28. Mei, Y., Zhang, H., Zhang, X., Zhang, J., Shu, Z., Wang, Y., Wei, Z., Yan, S., Jung, H., Patel, V.M.: LightPainter: Interactive portrait relighting with freehand scribble. In: CVPR (2023)
29. Nestmeyer, T., Lalonde, J.F., Matthews, I., Lehrmann, A.: Learning physics-guided face relighting under directional light. In: CVPR (2020)
30. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: CVPR (2023)
31. Pandey, R., Orts-Escolano, S., LeGendre, C., Haene, C., Bouaziz, S., Rhemann, C., Debevec, P., Fanello, S.: Total relighting: Learning to relight portraits for background replacement. In: SIGGRAPH (2021)
32. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NeurIPSW (2017)
33. Ponglertnapakorn, P., Tritrong, N., Suwajanakorn, S.: DiFaReli: Diffusion face relighting. In: ICCV (2023)
34. Ranjan, A., Yi, K.M., Chang, J.H.R., Tuzel, O.: FaceLit: Neural 3D relightable faces. In: CVPR (June 2023)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
36. Sengupta, R., Curless, B., Kemelmacher-Shlizerman, I., Seitz, S.: A light stage on every desk. In: ICCV (2021)
37. Sengupta, S., Kanazawa, A., Castillo, C.D., Jacobs, D.W.: SfSNet: Learning shape, refectance and illuminance of faces in the wild. In: CVPR (2018)
38. Shih, Y., Paris, S., Barnes, C., Freeman, W.T., Durand, F.: Style transfer for headshot portraits. TOG (2014)
39. Shu, Z., Hadap, S., Shechtman, E., Sunkavalli, K., Paris, S., Samaras, D.: Portrait lighting transfer using a mass transport approach. TOG (2017)
40. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
41. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P., Ramamoorthi, R.: Single image portrait relighting. In: SIGGRAPH (2019)
42. Sun, T., Lin, K.E., Bi, S., Xu, Z., Ramamoorthi, R.: NeLF: Neural light-transport field for portrait view synthesis and relighting. In: EGSR (2021)
43. Tan, F., Fanello, S., Meka, A., OrtsEscolano, S., Tang, D., Pandey, R., Taylor, J., Tan, P., Zhang., Y.: VoLux-GAN: A generative model for 3D face synthesis with HDRI relighting. In: SIGGRAPH (2022)
44. Wang, Y., Holynski, A., Zhang, X., Zhang, X.: Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In: CVPR (2023)
45. Wang, Z., Yu, X., Lu, M., Wang, Q., Qian, C., Xu, F.: Single image portrait relighting via explicit multiple reflectance channel modeling. TOG (2020)
46. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. TIP (2004)

47. Weir, J., Zhao, J., Chalmers, A., Rhee, T.: Deep portrait delighting. In: ECCV (2022)
48. Yang, X., Taketomi, T.: BareSkinNet: Demakeup and de-lighting via 3D face reconstruction. CGF (2022)
49. Yeh, Y.Y., Nagano, K., Khamis, S., Kautz, J., Liu, M.Y., Wang, T.C.: Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. TOG (2022)
50. Zhang, L., Zhang, Q., Wu, M., Yu, J., Xu, L.: Neural video portrait relighting in real-time via consistency modeling. In: ICCV (2021)
51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
52. Zhang, X., Fanello, S., Tsai, Y.T., Sun, T., Xue, T., Pandey, R., Orts-Escolano, S., Davidson, P., Rhemann, C., Debevec, P., Barron, J.T., Ramamoorthi, R., Freeman, W.T.: Neural light transport for relighting and view synthesis. TOG (2021)
53. Zhang, X., Barron, J.T., Tsai, Y.T., Pandey, R., Zhang, X., Ng, R., Jacobs, D.E.: Portrait Shadow Manipulation. TOG (2020)
54. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single portrait image relighting. In: ICCV (2019)