

COMPOSE: Comprehensive Portrait Shadow Editing Supplementary Materials

Andrew Hou^{1,2}, Zhixin Shu², Xuaner Zhang², He Zhang², Yannick
Hold-Geoffroy², Jae Shin Yoon², and Xiaoming Liu¹

¹ Michigan State University
{houandr1, liuxm}@msu.edu

² Adobe Research
{zshu, cezhang, hezhan, holdgeof, jaeyoon}@adobe.com

1 Additional Discussions Regarding Prior Methods

We delve into face relighting and shadow editing methods in greater detail to highlight the novel contributions of COMPOSE. We again emphasize that our core contribution is our decomposition of the environment map representation into ambient light and an editable gaussian dominant light source that can manipulate all shadow attributes (intensity, light size, and position). This also enables preserving the ambient light in the original source image through compositing with $\mathbf{I}_D$. As emphasized in the main paper, existing relighting methods [5–7] that use environment maps as the lighting representation will completely change the lighting environment to perform shadow editing by providing an entirely new environment map. This is unsuitable for many computational photography applications where the desire is to edit the facial shadows alone and to preserve the ambient light of the original environment. However, aside from these methods, some relighting and shadow editing methods are able to preserve the ambient light but lack a comprehensive set of shadow editing capabilities. Zhang *et al.* [8] controls the degree of shadow softening but can’t intensify shadows. They don’t consider shadows resulting from decreasing light size and can’t alter shadow position. Futschik *et al.* [2] are similarly able to control the degree of shadow softening while maintaining the ambience of the original photo, but they’re completely limited to shadow softening. Hou *et al.* [3] model shadow position/intensity, but don’t model light size and can’t edit shadow shape. They also don’t properly disentangle existing light sources from albedo and source image shadows will remain baked in. Nestmeyer *et al.* [4] use a single light direction (xyz) and thus can model shadow position but not intensity or light size. **Notably, none of these methods can handle shrinking the light to both intensify the shadow and change its shape or need to completely modify the photo’s environment in order to do so.** This is due to the lack of a suitable lighting decomposition, which is enabled by our method’s separation of the environment map into ambient light and an editable gaussian dominant light source. The editable gaussian is what enables flexible control over all shadow attributes (including shrinking the light size), while the ambient light estimation helps preserve the remaining lighting attributes of the original photo. Another

consideration that motivates our decomposition is that while it is possible to enlarge the light size of an existing environment map with a gaussian blur, the converse operation of shrinking the light size does not have a good solution.

2 Model Architectures

We describe our model architectures for our light estimation VAE, our light diffusion hierarchical transformer, and our shadow synthesis pipeline in greater detail here.

Light Estimation VAE Our light estimation VAE consists of a 6-layer convolutional encoder followed by two fully-connected layers FC_μ and FC_σ that predict μ and σ respectively. We then sample $z = \epsilon\sigma + \mu$, where $\epsilon \sim N(0, 1)$ using the reparameterization trick, which becomes the input for our 4-layer convolutional decoder. The decoder produces the final predicted environment map E_P . Please see Tab. 2 for more details.

Table 2: Structure of Light Estimation VAE. We describe in detail the structure of our lighting estimation VAE. z is obtained by applying the reparameterization trick to the output of FC_μ and FC_σ . We apply batch normalization after all convolutional layers and all LeakyReLU layers use a slope of 0.2.

Layer	Input	Type	Kernel Size	Stride	Input Features	Output Features	Activation
<i>conv1</i>	$\mathbf{I}_N$	Convolution	3	2	3	16	LeakyReLU
<i>conv2</i>	<i>conv1</i>	Convolution	3	2	16	32	LeakyReLU
<i>conv3</i>	<i>conv2</i>	Convolution	3	2	32	64	LeakyReLU
<i>conv4</i>	<i>conv3</i>	Convolution	3	2	64	128	LeakyReLU
<i>conv5</i>	<i>conv4</i>	Convolution	3	2	128	256	LeakyReLU
<i>conv6</i>	<i>conv5</i>	Convolution	3	2	256	512	LeakyReLU
FC_μ	<i>conv6</i>	Fully Connected	-	-	$512 * 10 * 8$	512	None
FC_σ	<i>conv6</i>	Fully Connected	-	-	$512 * 10 * 8$	512	None
FC_z	z	Fully Connected	-	-	512	$256 * 8 * 4$	LeakyReLU
<i>deconv1</i>	FC_z	Transposed Convolution	3	2	256	128	LeakyReLU
<i>deconv2</i>	<i>deconv1</i>	Transposed Convolution	3	2	128	64	LeakyReLU
<i>deconv3</i>	<i>deconv2</i>	Transposed Convolution	3	2	64	32	LeakyReLU
<i>deconv4</i>	<i>deconv3</i>	Transposed Convolution	3	2	32	3	Sigmoid

Light Diffusion Transformer The light diffusion transformer’s goal is to predict the ambient image $\mathbf{I}_D$. As such, it removes shadows and specularities from the source image $\mathbf{I}_N$. To accomplish this, we leverage an encoder-decoder structure with 3 inputs: input image $\mathbf{I}_N$, a binary segmentation mask with the portrait foreground, and a body parsing mask. We leverage a hierarchical transformer encoder and divide the inputs into 4x4 patches, which are more suitable for harmonization and shadow removal tasks than larger 16x16 patches. We then obtain multi-level features using the transformer encoder at $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the original image resolution using multiple transformer blocks. These multi-level

features are then passed to our decoder, which consists of transposed convolutional layers and generates the final ambient image $\mathbf{I}_D$.

Shadow Synthesis U-Net Our stage 1 U-Net in our shadow synthesis pipeline is largely adopted from SIPR’s [6] architecture with minor modifications. The main difference is in how we inject the lighting representation to the bottleneck of the U-Net. Instead of passing a 3-channel environment map, we pass our 4-channel spatial lighting representation (x, y, σ, γ) , as mentioned in the main paper. We find that this helps the network when training on smaller, concentrated lights compared to passing the standard 3-channel environment map, likely because a small focused light occupies only a small portion of the environment map in pixel space and may thus provide a weak learning signal. Our method of encoding the light parameters numerically and replicating them spatially avoids this issue.

Shadow Synthesis DDPM For our stage 2 conditional DDPM, we leverage the model of [1]. 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:

$$\mathbb{E}_{t, \mathbf{x}_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta(\mathbf{x}_t, t, \mathbf{I}_U, x, y, \sigma, \gamma) \right\|^2 \right]. \quad (1)$$

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 $\mathbf{x}_t = \sqrt{\alpha_t} \mathbf{x}_0 + \sqrt{1 - \alpha_t} \epsilon$.



Fig. 8: Ambient Light. COMPOSE is able to preserve the ambient light of the original portrait thanks to the hierarchical transformer used in the light diffusion stage. The light diffusion stage removes specularities and existing shadows caused by strong directional lights, leaving behind only the ambient light. Each pair of images is left: input, right: ambient prediction.

Table 3: Skin Tone Fairness. COMPOSE performs well on darker skin tones (Dark) as well as intermediate intensity skin tones (Tan). On light skin tones (Light), SSIM/LPIPS are comparable with Dark and Tan and MAE/MSE are slightly worse, mostly due to the larger errors accrued from incorrect shadow boundaries on light skin subjects.

	MAE	MSE	SSIM	LPIPS
Light	0.1141	0.0449	0.7756	0.1880
Tan	0.0724	0.0239	0.7821	0.2074
Dark	0.0755	0.0207	0.7798	0.2034
All Skin Tones	0.0965	0.0349	0.7780	0.1973

Table 4: Diffusion Ablation. Adding the DDPM to the shadow synthesis stage slightly worsens the quantitative metrics, but greatly enhances visual quality.

	MAE	MSE	SSIM	LPIPS
U-Net Only	0.0831	0.0226	0.7977	0.1734
U-Net+DDPM	0.0965	0.0349	0.7780	0.1973

3 Skin Tone Fairness

As with any face editing work, the ability to represent and generalize to diverse subjects is of the utmost importance. We therefore quantitatively evaluate how well COMPOSE performs across different skin tones. See Tab. 3, where we separated our light stage test subjects into 3 groups (Light, Tan, and Dark) based on mean skin tone values. Performance on Dark/Tan is good whereas Light has comparable SSIM/LPIPS but worse MAE/MSE. This is because errors involving shadows are maximized for light skin since the difference between shadows and light skin is higher than for darker skin. This can be addressed by increasing the number of light skin training subjects or increasing loss weights in shadow regions to prevent large errors caused by slightly misaligned shadow boundaries.

4 Ambient Light

Since preserving the ambient light of the portrait is an important component of COMPOSE, we demonstrate some results from our light diffusion network. Ambient light is respected by our light diffusion network, and should be different for the same face illuminated by 2 different environment maps (See Fig. 8, left:input, right:ambient).

5 Foreign Shadow Removal

Our light diffusion network is able to remove not just self shadows, but also foreign shadows and can thus serve as a shadow removal network. Please see Fig. 9 for some sample results on portraits with substantial foreign shadows.



Fig. 9: Foreign Shadow Removal. Our light diffusion network can serve as a shadow removal method, and even handles foreign shadow removal well.



Fig. 10: Diffusion Ablation. We ablate whether adding a DDPM after our shadow synthesis U-Net improves the shadow editing performance. Each triplet represents U-Net only, U-Net+DDPM, and groundtruth target image. The DDPM substantially improves the visual quality of the images and sharpens the shadow boundaries. The target lighting condition is visualized in the middle image of each triplet.

6 Diffusion Ablation

See Fig. 10, where every triplet is U-Net only, U-Net+DDPM, and groundtruth. Adding the DDPM slightly worsens our metrics (See Tab. 4) but greatly enhances the visual quality of the images and sharpens shadow boundaries.

7 Additional Shadow Removal Results

When comparing shadow softening performance with Futschik *et al.* [2], we sent a total of 40 in-the-wild images. As demonstrated in Fig. 11, the main limitation in [2] is that sometimes it leaves a noticeable shadow trace in the presence of darker shadows at level 0.0, where the goal is to remove all shadows. In Fig. 11, we show all 40 shadow softening results compared to our method at level 0.0. Across all 40 evaluation images, our method is virtually spotless at removing shadows whereas several examples have shadow traces and artifacts for [2]. This is largely thanks to our hierarchical transformer design for our light diffusion network, which better handles removing the effects of shadows at all scales compared to the U-Net design of [2].

8 Clarifications on Evaluation Protocols

Here, we clarify more about our testing protocols to ensure that we perform fair comparisons in all of our evaluations. When comparing with Total Relighting (TR) [5] and Hou *et al.* [3] in Fig. 4 and Tab. 1 of the main paper, we ensure that we provide the most generous possible test setting. The results from TR

were provided by the authors upon our request and we followed all instructions to properly prepare testing data. The environment maps were Gaussian lights of varying intensity/position/spread. We allowed the authors to tune the lighting scale to achieve their best possible results and provided sample relit images with their corresponding environment maps to help them measure the appropriate lighting scale to match TR’s conventions. The results from Hou *et al.* were also provided by the authors, and in their case the evaluation is simpler since they only model light direction. The qualitative comparisons with Futschik *et al.* [2] in Fig. 7 on shadow softening, as mentioned in the main paper, are also provided by the authors.

9 Design Choices

Here we further explain some of our design choices in the COMPOSE pipeline, including motivation and physical plausibility.

LDR Representation of Light In our light estimation stage, we regress an LDR environment map instead of an HDR environment map. The reasons for this are that LDR is easier to regress than HDR during training due to the large, out-of-range peak values in HDR and we find LDR is sufficient for estimating light position in our pipeline (we are focused on the dominant light position and not as concerned with regressing every detail in the environment map). When performing gaussian fitting to find the dominant light, the algorithm will converge at roughly the same solution in determining the light center.

Single Dominant Light Assumption Our assumption of a single dominant light source in the scene stems from our focus on outdoor in-the-wild settings. There’s usually one dominant Gaussian light outdoors, namely the sun. Our method can be extended to handle multiple lights by first changing the gaussian fitting step in the light estimation stage to enable fitting multiple gaussians (lights). This could be done using Gaussian Mixture Models (varying k) to determine the number of lights (k) using the lowest fitting error as the metric. Handling more diffuse lighting environments is another interesting direction. This can be done by leveraging the linearity of light to generate diffuse lightings as the sum of many directional lights, where each directional light involves one model inference of COMPOSE.

Image Compositing We further elaborate on the physical motivation of the fourth stage of COMPOSE: the compositing step. $\mathbf{I}_D$ and $\mathbf{I}_S$ represent the contributions of ambient and diffuse/directional light respectively. Compositing adjusts the contributions of each type of lighting, which is physically plausible by the Phong Model: $I = L_{amb} * R_{amb} + L_{diff} * R_{diff} + L_{spec} * R_{spec}$.

Four Stages One question that may arise is whether COMPOSE truly needs to be a four-stage model, which to some may seem unnecessarily complex. However, we argue in favor of this design for several reasons. One is that the inference time of a multi-stage model compared to a single end-to-end pipeline should be similar. Either way, light estimation, delighting, and relighting modules will be required. The advantage of our multi-stage pipeline is greater flexibility for users in terms of applications (*e.g.*, using our light diffusion network alone for shadow removal). It’s also easier to integrate components of COMPOSE into other shadow editing and relighting methods. The design is modular and each component can be improved or replaced. This enables better disentanglement, controllability, and explainability during shadow editing. It’s also easier to identify which component is causing problems if the shadow editing result is poor.

10 Lighting Estimation

Our lighting estimation stage is primarily responsible for providing the user with an accurate estimation of where the dominant light source of the input image $\mathbf{I}_N$ is located. This is useful if the user wants to perform any kind of in-place shadow editing that does not involve changing the light position (*e.g.* shrinking the light, enlarging the light, softening the shadow, and intensifying the shadow). It is important to note that our pipeline is fairly error tolerant in this stage since the user is free to tune the lighting estimation in subsequent stages to achieve better results (*e.g.* tuning the dominant light position in the environment map to better match the shadow positions in the source image).

11 Responsible Human Dataset Usage

We collect and use light stage data, as well as images from Unsplash and Adobe Stock. For the light stage data, we have acquired permission from both the subjects and the capturing studio to include these images in research papers. Unsplash’s license mentions photos can be downloaded and used for free for both commercial and non-commercial purposes. For Adobe Stock, we paid for any images that we used and are thus permitted to include them in our submission. Finally, for the subject appearing in Figs. 1 and 6, the source image is simply acquired from a colleague who participated in a lighting experiment and we have received permission from them personally to include their photos for publication. In all cases, we collect or use portrait images only that contain no additional personally identifiable information or offensive content.

References

1. Choi, J., Kim, S., Jeong, Y., Gwon, Y., Yoon, S.: ILVR: Conditioning method for denoising diffusion probabilistic models. In: ICCV (2021)
2. 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)

3. Hou, A., Sarkis, M., Bi, N., Tong, Y., Liu, X.: Face relighting with geometrically consistent shadows. In: CVPR (2022)
4. Nestmeyer, T., Lalonde, J.F., Matthews, I., Lehrmann, A.: Learning physics-guided face relighting under directional light. In: CVPR (2020)
5. 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)
6. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyfe, G., Rhemann, C., Busch, J., Debevec, P., Ramamoorthi, R.: Single image portrait relighting. In: SIGGRAPH (2019)
7. 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)
8. Zhang, X., Barron, J.T., Tsai, Y.T., Pandey, R., Zhang, X., Ng, R., Jacobs, D.E.: Portrait Shadow Manipulation. TOG (2020)



Fig. 11: Shadow Removal Comparisons with Futschik *et al.* [2]. We demonstrate all qualitative shadow removal results against Futschik *et al.* [2] without cherry picking. COMPOSE is able to leave less of a shadow trace, especially for darker shadows when the goal is complete shadow removal.



Fig. 11: Shadow Removal Comparisons with Futschik *et al.* [2]. We demonstrate all qualitative shadow removal results against Futschik *et al.* [2] without cherry picking. COMPOSE is able to leave less of a shadow trace, especially for darker shadows when the goal is complete shadow removal.



Fig. 11: Shadow Removal Comparisons with Futschik *et al.* [2]. We demonstrate all qualitative shadow removal results against Futschik *et al.* [2] without cherry picking. COMPOSE is able to leave less of a shadow trace, especially for darker shadows when the goal is complete shadow removal.