SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting

Hoon Kim[1] Minje Jang[1] Wonjun Yoon[1] Jisoo Lee[1] Donghyun Na[1] Sanghyun Woo[2]

[1]Beeble AI [2]New York University

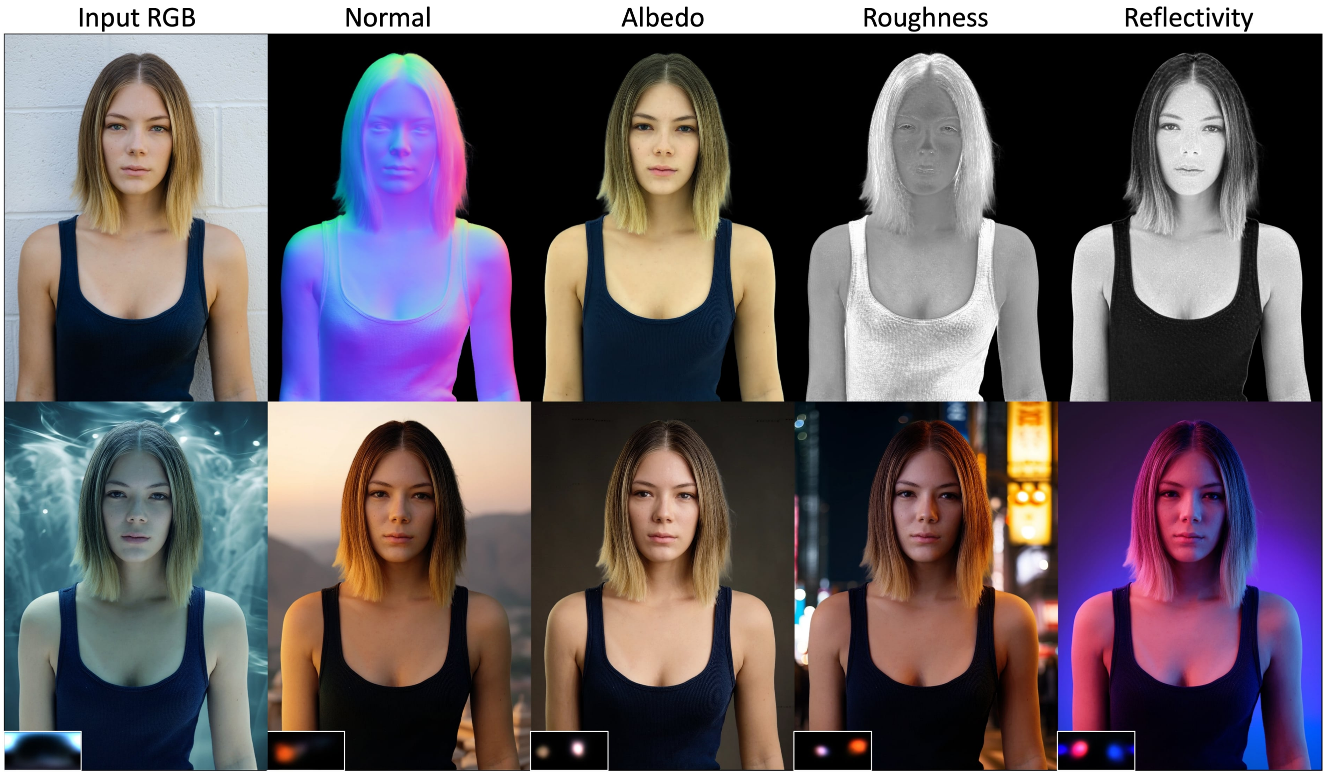

Figure 1. **Be Anywhere at Any Time.** SwitchLight processes a human portrait by decomposing it into detailed intrinsic components, and re-renders the image under a designated target illumination, ensuring a seamless composition of the subject into any new environment.

Abstract

We introduce a co-designed approach for human portrait relighting that combines a physics-guided architecture with a pre-training framework. Drawing on the Cook-Torrance reflectance model, we have meticulously configured the architecture design to precisely simulate light-surface interactions. Furthermore, to overcome the limitation of scarce high-quality lightstage data, we have developed a self-supervised pre-training strategy. This novel combination of accurate physical modeling and expanded training dataset establishes a new benchmark in relighting realism.

1. Introduction

Relighting is more than an aesthetic tool; it unlocks infinite narrative possibilities and enables seamless integration of subjects into diverse environments (see Fig. 1). This advancement resonates with our innate desire to transcend the physical constraints of space and time, while also providing tangible solutions to practical challenges in digital content creation. It is particularly transformative in virtual (VR) and augmented reality (AR) applications, where relighting facilitates the real-time adaptation of lighting, ensuring that users and digital elements coexist naturally within any environment, offering a next level of telepresence.

In this work, we focus on human portrait relighting.

[*] All authors contributed equally to this work
[*] https://www.beeble.ai

While the relighting task fundamentally demands an in-depth understanding of geometry, material properties, and illumination, the challenge is more compounded when addressing human subjects, due to the unique characteristics of skin surfaces as well as the diverse textures and reflectance properties of a wide array of clothing, hairstyles, and accessories. These elements interact in complex ways, necessitating advanced algorithms capable of simulating the subtle interplay of light with these varied surfaces.

Currently, the most promising approach involves the use of deep neural networks trained on pairs of high-quality relit portrait images and their corresponding intrinsic attributes, which are sourced from a light stage setup [10]. Initial efforts approached the relighting process as a 'black box' [44, 47], without delving into the underlying mechanisms. Later advancements adopted a physics-guided model design, incorporating the explicit modeling of image intrinsics and image formation physics [31]. Pandey et al. [33] proposed the Total Relight (TR) architecture, also physics-guided, which decomposes an input image into surface normals and albedo maps, and performs relighting based on the Phong specular reflectance model. The TR architecture has become foundational model for image relighting, with most recent and advanced architectures building upon its principle [22, 30, 51].

Following the physics-guided approach, our contribution lies in a co-design of architecture with a self-supervised pre-training framework. First, our architecture evolves towards a more accurate physical model by integrating the Cook-Torrance specular reflectance model [8], representing a notable advancement from the empirical Phong specular model [36] employed in the Total Relight architecture. The Cook-Torrance model adeptly simulates light interactions with surface microfacets, accounting for spatially varying roughness and reflectivity. Secondly, our pre-training framework scales the learning process beyond the typically hard-to-obtain lightstage data. By revisiting the masked autoencoder (MAE) framework [18], we adept it for the task of relighting. These modifications are crafted to address the unique challenges posed by this task, enabling our model to learn from unlabelled data and refine its ability to produce realistic relit portraits during fine-tuning. To the best of our knowledge, this is the first time applying self-supervised pre-training specifically to the relighting task.

To summarize, our contribution is twofold. Firstly, by enhancing the physical reflectance model, we have introduced a new level of realism in the output. Secondly, by adopting self-supervised learning, we have expanded the scale of the training data and enhanced the expression of lighting in diverse real-world scenarios. Collectively, these advancements have led SwitchLight framework to achieve a new state-of-the-art in human portrait relighting.

2. Related Work

Human Portrait Relighting is an ill-posed problem due to its under-constrained nature. To tackle this, earlier methods incorporated 3D facial priors [43], exploited image intrinsics [3, 39], or framed the task as a style transfer [42]. Light stage techniques [48] offer a more powerful solution by recording subject's reflectance fields under varying lighting conditions [10, 13], though they are labor-intensive and require specialized equipment. A promising alternative has emerged with deep learning, utilizing neural networks trained on light stage data. Sun et al. [44] pioneered this approach, but their method had limitations in representing non-Lambertian effects. This was improved upon by Nestmeyer et al. [31], who integrated rendering physics into network design, albeit limited to directional light. Building upon this, Pandey et al. [33] incorporated the Phong reflection model and a high dynamic range (HDR) lighting map [9] into their network, enabling a more accurate representation of global illumination. Simultaneously, efforts have been made to explore portrait relighting without light stage data [19, 20, 41, 46, 54]. Moreover, introduction of NeRF [7] and diffusion-based [37] models has opened new avenues in the field. However, networks trained with light-stage data maintain superior accuracy and realism, thanks to physics-based composited relight image training pairs and precise ground truth image intrinsics [55].

Our work furthers this domain by integrating the Cook-Torrance model into our network design, shifting from the empirical Phong model to a more physics-based approach, thereby enhancing the realism and detail in relit images.

Self-supervised Pre-training has become a standard training scheme in the development of large language models like BERT [11] and GPT [38], and is increasingly influential in vision models, aiming to replicate the 'BERT moment'. This approach typically involves pre-training on extensive unlabeled data, followed by fine-tuning on specific tasks. While early efforts in vision models focused on simple pretext tasks [12, 16, 32, 35, 52], the field has evolved through stages like contrastive learning [5, 17] and masked image modeling [2, 18, 49]. However, the primary focus has remained on visual recognition, with less attention to other domains. Exceptions include low-level image processing tasks [4, 6, 26, 29] using the vision transformer [14].

Our research takes a different route, focusing on human portrait relighting—a complex challenge of manipulating illumination in the image. This direction is crucial because acquiring accurate ground truth data, especially from light stage, is both expensive and difficult. We modify the MAE framework [18], previously successful in robust image representation learning and developing locality biases [34], to suit the unique requirements of effective relighting.

3. SwitchLight

We introduce SwitchLight, a state-of-the-art framework for human portrait relighting, with its architectural overview presented in Fig. 2. We first provide foundational concepts in Sec. 3.1, and define the problem in Sec. 3.2. This is followed by detailing the architecture in Sec. 3.3, and lastly, we describe the loss functions used in Sec. 3.4.

3.1. Preliminaries

In this section, vectors $\mathbf{n}$, $\mathbf{v}$, $\mathbf{l}$, and $\mathbf{h}$ are denoted as unit vectors. Specifically, $\mathbf{n}$ represents the surface normal, $\mathbf{v}$ is the view direction, $\mathbf{l}$ is the incident light direction, and $\mathbf{h}$ is the half-vector computed from $\mathbf{l}$ and $\mathbf{v}$. The dot product is clamped between $[0..1]$, indicated by $\langle\cdot\rangle$.

Image Rendering. The primary goal of image rendering is to create a visual representation that accurately simulates the interactions between light and surfaces. These complex interactions are encapsulated by the rendering equation:

$$L_o(\mathbf{v}) = \int_\Omega f(\mathbf{v},\mathbf{l}) L_i(\mathbf{l}) \langle \mathbf{n}\cdot\mathbf{l}\rangle \, d\mathbf{l} \tag{1}$$

where $L_o(\mathbf{v})$ denotes the radiance, or the light intensity perceived by the observer in direction $\mathbf{v}$. It is the cumulative result of incident lights $L_i(\mathbf{l})$ from all possible directions over the hemisphere, Ω, centered around the surface *normal*, denoted as $\mathbf{n}$. Central to this equation lies the Bidirectional Reflectance Distribution Function (BRDF), denoted as $f(\mathbf{v},\mathbf{l})$, describing the surface's reflection characteristics.

BRDF Composition. The BRDF, represented by $f(\mathbf{v},\mathbf{l})$, describes how light is reflected at an opaque surface. It is composed of two major components: diffuse reflection (f_d) and specular reflection (f_s):

$$f(\mathbf{v},\mathbf{l}) = f_d(\mathbf{v},\mathbf{l}) + f_s(\mathbf{v},\mathbf{l}) \tag{2}$$

A surface intrinsically exhibits both diffuse and specular reflections. The diffuse component uniformly scatters light, ensuring consistent illumination regardless of the viewing angle. In contrast, the specular component is viewing angle-dependent, producing shiny highlights that are crucial for achieving a photorealistic effect.

Lambertian Diffuse Reflectance. Lambertian reflectance, a standard model for diffuse reflection, describes a uniform light scatter irrespective of the viewing angle. This ensures a consistent appearance in brightness:

$$f_d(\mathbf{v},\mathbf{l}) = \frac{\sigma}{\pi} \quad \text{[const.]} \tag{3}$$

Here, σ is the *albedo*, indicating the intrinsic color and brightness of the surface.

Cook-Torrance Specular Reflectance. The Cook-Torrance model, based on microfacet theory, represents surfaces as a myriad of tiny, mirror-like facets. It incorporates a *roughness* parameter α, which allows precise rendering of surface specular reflectance:

$$f_s(\mathbf{v},\mathbf{l}) = \frac{D(\mathbf{h},\alpha)G(\mathbf{v},\mathbf{l},\alpha)F(\mathbf{v},\mathbf{h},f_0)}{4\langle\mathbf{n}\cdot\mathbf{l}\rangle\langle\mathbf{n}\cdot\mathbf{v}\rangle} \tag{4}$$

In this model, D is the microfacet distribution function, describing the orientation of the microfacets relative to the half-vector h, G is the geometric attenuation factor, accounting for the shadowing and masking of microfacets, and F is the Fresnel term, calculating the reflectance variation depending on the viewing angle, where f_0 is the surface *Fresnel reflectivity* at normal incidence. A lower α value implies a smoother surface with sharper specular highlights, whereas a higher α value indicates a rougher surface, resulting in more diffused reflections. By adjusting α, the Cook-Torrance model can depict a range of specular reflections.

Image Formation. Upon the base rendering equation, we include the diffuse and specular components of the BRDF and derive a unified formula:

$$L_o(\mathbf{v}) = \int_\Omega \left(f_d(\mathbf{v},\mathbf{l}) + f_s(\mathbf{v},\mathbf{l})\right) E(\mathbf{l}) \langle\mathbf{n}\cdot\mathbf{l}\rangle \, d\mathbf{l} \tag{5}$$

where $E(\mathbf{l})$ denotes the incident environmental lighting. This formula represents the core principle that an image is a product of interplay between the BRDF and lighting. To further clarify this concept, we introduce a rendering function R, succinctly modeling the process of image formation:

$$I = R(\underbrace{\mathbf{n},\sigma,\alpha,f_0}_{\text{surface attributes}}, \underbrace{E}_{\text{lighting}}) \tag{6}$$

It is important to note that since the BRDF is a function of surface properties, as detailed in Eqn. 3 and 4, we can now clearly understand that image formation is essentially governed by the interaction of surface attributes and lighting.

3.2. Problem Formulation

Image Relighting. Given the image formation model above, our goal is to manipulate the lighting of an existing image. This involves two main steps: inverse rendering and re-rendering under target illumination, both driven by neural networks. For a given source image I_{src} and target illumination E_{tgt}, the process is delineated as:

$$\begin{aligned} \text{Inverse Rendering.} \quad & (\mathbf{n},\sigma,\alpha,f_0,E_{\text{src}}) = U(I_{\text{src}}) \\ \text{Rendering with Target Light.} \quad & I_{\text{tgt}} = R(\mathbf{n},\sigma,\alpha,f_0,E_{\text{tgt}}) \end{aligned}$$

During the inverse rendering step, the function U unravels the intrinsic properties of I_{src}. In the subsequent relighting

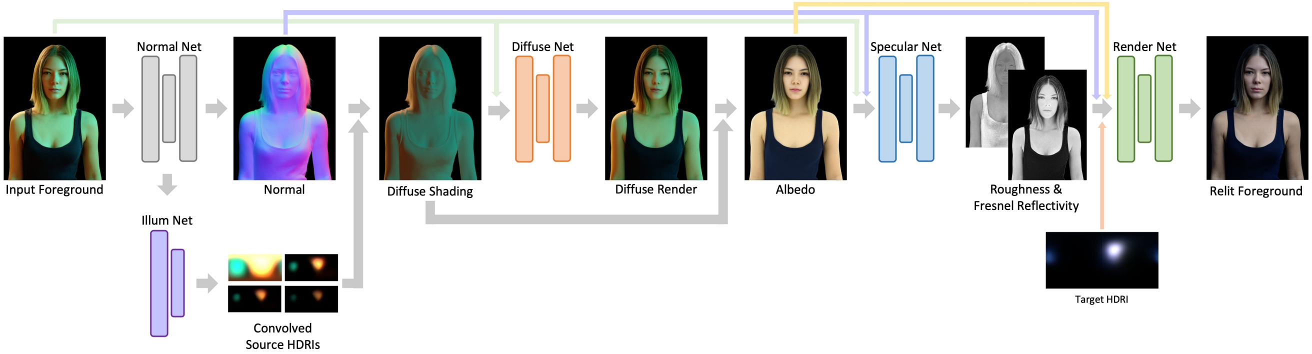

Figure 2. **SwitchLight Architecture.** The input source image is decomposed into *normal* map, *lighting*, *diffuse* and *specular* components. Given these intrinsics, images are re-rendered under target lighting. The architecture integrates the *Cook-Torrance* reflection model; the final output combines physically-based predictions with neural network enhancements for realistic portrait relighting.

step, the derived intrinsic properties along with new illumination E_{tgt} are employed by the rendering function R to generate the target relit image I_{tgt}.

3.3. Architecture

Our architecture systematically executes the two primary stages outlined in our problem formulation. The first stage involves extracting intrinsic properties from the source image I_{src}. For this purpose, we employ a matting network [25, 27, 40] to accurately separate the foreground. This extracted image is then processed by our inverse rendering network U, which infers normal, albedo, roughness, reflectivity, and source lighting. Subsequently, the second stage involves re-rendering the image under new target lighting conditions. To achieve this, the acquired intrinsics, along with the target lighting E_{tgt}, are fed into our relighting network R, producing the relit image I_{tgt}.

Normal Net. The network takes the source image $I_{\text{src}} \in \mathbb{R}^{\text{H}\times\text{W}\times 3}$ and generates a **normal map** $\hat{\mathbf{N}}$. Each pixel in this map contains a unit normal vector $\hat{\mathbf{n}}$, indicating the orientation of the corresponding surface point.

Illum Net. The network infers the lighting conditions in the given image captured in an HDRI format. Specifically, it computes the **convolved HDRIs**:

$$E^p_{\text{src}}(\mathbf{l}') = \int_\Omega \underbrace{E_{\text{src}}(\mathbf{l})}_{\text{HDRI}} \underbrace{\langle \mathbf{l}' \cdot \mathbf{l}\rangle^p}_{\text{Phong lobe}} \, d\mathbf{l} \tag{7}$$

In this equation, $E_{\text{src}} \in \mathbb{R}^{\text{H}_{\text{HDRI}}\times\text{W}_{\text{HDRI}}\times 3}$ is the original source HDRI map, with $\mathbf{l}$ indicating spherical directions in the HDRI space $\mathbb{R}^{\text{H}_{\text{HDRI}}\times\text{W}_{\text{HDRI}}}$. The term $\langle \mathbf{l}' \cdot \mathbf{l}\rangle^p$ represents the Phong reflectance lobe with shininess exponents $p \in \{1, 16, 32, 64\}$, which incorporates various specular terms. Consequently, it is expressed in a multi-dimensional tensor form as $\mathbb{R}^{4\times\text{H}_{\text{cHDRI}}\times\text{W}_{\text{cHDRI}}\times\text{H}_{\text{HDRI}}\times\text{W}_{\text{HDRI}}}$. Finally, $E^p_{\text{src}} \in \mathbb{R}^{4\times\text{H}_{\text{cHDRI}}\times\text{W}_{\text{cHDRI}}\times 3}$ is the convolved HDRI. In this work, we set the resolution of HDRI and convolved HDRI at 32 × 64 and 64 × 128, respectively, and we apply convolution on light source coordinates.

The network employs a cross-attention mechanism at its core, where predefined Phong reflectance lobes serve as queries, and the original image acts as both keys and values. Within this setup, the convolved HDRI maps are synthesized by integrating image information into the Phong reflectance lobe representation. Specifically, our model utilizes bottleneck features from the *Normal Net* as a compact image representation. Our approach simplifies the complex task of HDRI reconstruction by instead focusing on estimating interactions with known surface reflective properties.

Diffuse Net. Estimating albedo is challenging due to the ambiguities in surface color and material properties, further complicated by shadow effetcs. To address this, we prioritize the inference of source **diffuse render**, $I_{\text{src,diff}}$:

$$L_{\text{src},o_{\text{diff}}}(\mathbf{v}) = \underbrace{\frac{\sigma}{\pi}}_{\text{diffuse BRDF}} \underbrace{\int_\Omega E_{\text{src}}(\mathbf{l})\langle \mathbf{n} \cdot \mathbf{l}\rangle \, d\mathbf{l}}_{\text{diffuse shading}} \tag{8}$$

Our key insight is that the diffuse render closely resembles the original image, which simplifies the model learning process. It captures surface color after removing specular reflections, such as shine or gloss, contrasting with albedo that represents the true surface color unaffected by lighting and shadows. The network takes a source image I_{src}, concatenated with its diffuse shading, to produce the diffuse render. As in Eqn. 7, the diffuse shading, $\hat{E}^1_{\text{src}}(\hat{\mathbf{n}})$, is derived using the predicted normals, $\hat{\mathbf{n}}$, and the predicted lighting map, $\hat{E}^1_{\text{src}}$, with a Phong exponent of 1 for the diffuse term. The **albedo map** $\hat{\mathbf{A}}$ is then computed by dividing the predicted diffuse render by its diffuse shading:

$$\frac{\hat{\sigma}}{\pi} = \frac{\hat{L}_{\text{src},o_{\text{diff}}}(\mathbf{v})}{\hat{E}^1_{\text{src}}(\hat{\mathbf{n}})} \tag{9}$$

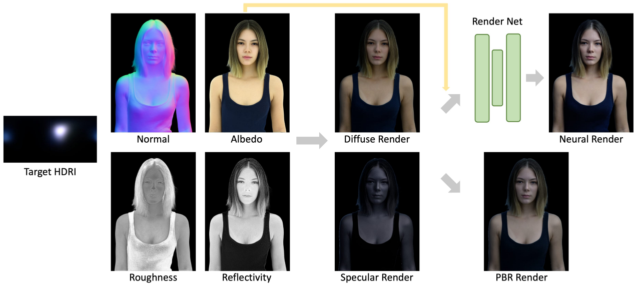

Figure 3. **Render Net Overview.** Utilizing extracted image intrinsics, it employs the Cook-Torrance model for initial relighting and a neural network for enhanced refinement, producing high-fidelity relit images through a synergistic computational approach.

We have empirically validated that it significantly enhances albedo prediction across a range of real-world scenarios.

Specular Net. The network infers surface attributes associated with the Cook-Torrance specular elements, specifically, the **roughness** α and **Fresnel reflectivity** f_0. It uses a source image, predicted normal, and albedo maps as inputs.

Render Net. The network utilizes extracted intrinsic surface attributes to produce the **target relit images**. It generates two types of relit images, as shown in Fig. 3. The first type adheres to the physically-based rendering (PBR) principles of the Cook-Torrance model. This involves computing diffuse and specular renders under the target illumination using Eqn. 3 and Eqn. 4. These renders are combined to form the PBR render, $\hat{I}^{\text{PBR}}_{\text{tgt}}$, as:

$$\hat{L}^{\text{PBR}}_{\text{tgt},o}(\mathbf{v}) = \hat{L}_{\text{tgt},o_{\text{diff}}}(\mathbf{v}) + \hat{L}_{\text{tgt},o_{\text{spec}}}(\mathbf{v}) \tag{10}$$

The second type of relit image is the result of a neural network process. This enhances the PBR render, capturing finer details that the Cook-Torrance model might miss. It employs the albedo, along with the diffuse and specular renders from the Cook-Torrance model, to infer a more refined target relit image, termed the neural render, $\hat{I}^{\text{Neural}}_{\text{tgt}}$. The qualitative improvements achieved through this neural enhancement are illustrated in Fig. 4.

3.4. Objectives

We supervise both intrinsic image attributes and relit images using their corresponding ground truths, obtained from the lightstage. We employ a combination of reconstruction, perceptual [23], adversarial [21], and specular [33] losses.

Reconstruction Loss $\mathcal{L} = \ell_1(\mathbf{X}, \hat{\mathbf{X}})$. It measures the pixel-level differences between the ground truth X and its prediction $\hat{\text{X}}$. This loss is applied across different attributes, including normal map $\hat{\mathbf{N}}$, convolved source HDRI maps $\hat{E}^{p}_{\text{src}}$, source diffuse render $\hat{I}_{\text{src,diff}}$, albedo map $\hat{\mathbf{A}}$, and both types of target relit images $\hat{I}^{\text{PBR}}_{\text{tgt}}$ and $\hat{I}^{\text{Neural}}_{\text{tgt}}$. The attributes like the roughness α and Fresnel reflectivity f_0 are also implicitly learned when this loss applied to the PBR render.

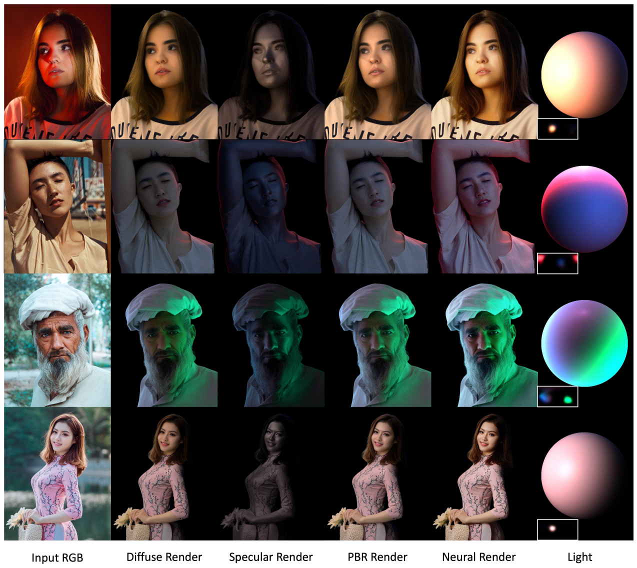

Figure 4. **Neural Render Enhancement.** Using the Cook-Torrance model, diffuse and specular renders are computed, which are then composited into a physically-based rendering. Subsequently, a neural network enhances this PBR render, improving aspects such as brightness and specular details.

Perceptual Loss $\mathcal{L}_{\text{vgg}} = \ell_2(\mathbf{VGG}(\mathbf{X}), \mathbf{VGG}(\hat{\mathbf{X}}))$. It captures high-level feature differences based on a VGG-network feature comparison. We apply this loss to the source diffuse render, albedo, and target relit images.

Adversarial Loss $\mathcal{L}_{\text{adv}} = \mathbf{GAN}(\mathbf{X}, \hat{\mathbf{X}})$. It promotes realism in the generated images by fooling a discriminator network. This loss is applied to the target relit images. We employ a PatchGAN architecture, with detailed specifications provided in the supplementary material.

Specular Loss $\mathcal{L}_{\text{spec}} = \ell_1(\mathbf{X} \odot \hat{\mathbf{S}}, \hat{\mathbf{X}} \odot \hat{\mathbf{S}})$. It enhances the specular highlights in the relit images. Specifically, we utilize the predicted specular render $\hat{\text{S}} := \hat{I}^{\text{PBR}}_{\text{tgt,spec}}$ derived from the Cook-Torrance physical model, to weigh the ℓ_1 reconstruction loss. Here, $\odot$ denotes the element-wise multiplication. This loss is applied to the neural render.

Final Loss. The SwitchLight is trained in an end-to-end manner using the weighted sum of the above losses:

$$\begin{aligned}\mathcal{L}_{\text{relight}} = {} & 10 \cdot \mathcal{L}_{\text{normal}} + 10 \cdot \mathcal{L}_{\text{src\_HDRI}} + 0.2 \cdot \mathcal{L}_{\text{src\_diff}} \\ & + 0.2 \cdot \mathcal{L}_{\text{albedo}} + 0.2 \cdot \mathcal{L}_{\text{PBR}} + 0.2 \cdot \mathcal{L}_{\text{Neural}} \\ & + \mathcal{L}_{\text{vgg}_{\text{src\_diff}}} + \mathcal{L}_{\text{vgg}_{\text{albedo}}} + \mathcal{L}_{\text{vgg}_{\text{PBR}}} + \mathcal{L}_{\text{vgg}_{\text{Neural}}} \\ & + \mathcal{L}_{\text{adv}_{\text{PBR}}} + \mathcal{L}_{\text{adv}_{\text{Neural}}} + 0.2 \cdot \mathcal{L}_{\text{spec}_{\text{Neural}}}.\end{aligned} \tag{11}$$

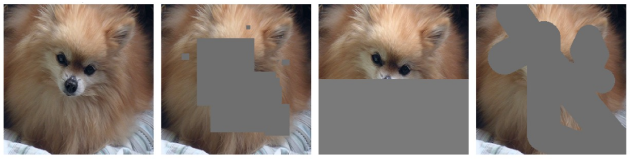

Figure 5. **Dynamic Masking Strategies.** We have generalized the MAE masks to include overlapping patches of varying sizes, as well as outpainting and free-form masks.

We empirically determined the weighting coefficients.

4. Multi-Masked Autoencoder Pre-training

We introduce the Multi-Masked Autoencoder (MMAE), a self-supervised pre-training framework designed to enhance feature representations in relighting models. It aims to improve output quality without relying on additional, costly light stage data. Building upon the MAE framework [18], MMAE capitalizes on the inherent learning of crucial image features like structure, color, and texture, which are essential for relighting tasks. However, adapting MAE to our specific needs poses several non-trivial challenges. Firstly, MAE is primarily designed for vision transformers [14], while our focus is on a UNet, a convolution-based architecture. This convolutional structure, with its hierarchical nature and aggressive pooling, is known to simplify the MAE task, necessitating careful adaptation [49]. Further, the hyperparameters of MAE, particularly the fixed mask size and ratio, are also specific to vision transformers. These factors could introduce biases during training and hinder the model to recognize image features at various scales. Moreover, MAE relies solely on masked region reconstruction loss, limiting the model to understand the global coherence of the reconstructed region in relation to its visible context.

To address these challenges effectively, we have developed two key strategies within the MMAE framework:

Dynamic Masking. MMAE eliminates two key hyperparameters, mask size and ratio, by introducing a variety of mask types to generalize the MAE. These types, which include overlapping patches of various sizes, outpainting masks [45], and free-form masks [28] (see Fig.5), each contribute to the model's versatility. With the ability to handle challenging masked regions, MMAE achieves a more comprehensive understanding of image properties.

Generative Target. In addition to its structural advancements, MMAE incorporates a new loss function strategy. We have adopted perceptual [23] and adversarial losses [21], along with the original reconstruction loss. As a result, MMAE is equipped not only to reconstruct missing image parts but also to ensure synthesis capabilities and their seamless integration with the original context. In practice, the weights for the three losses are equally set.

We pre-train the entire UNet architecture using MMAE, and, unlike MAE, we retain the decoder and fine-tune the entire model on relighting ground truths.

5. Data

We constructed the **OLAT** (One Light at a Time) dataset using a light stage [10, 48] equipped with 137 programmable LED lights and 7 frontal-viewing cameras. Our dataset comprises images of 287 subjects, with each subject being captured in up to 15 different poses, resulting in a total of 29,705 OLAT sequences. We sourced **HDRI** dataset from several publicly available archives. Specifically, we acquired 559 HDRI maps from Polyhaven, 76 from Noah Witchell, 364 from HDRMAPS, 129 from iHDRI, and 34 from eisklotz. In addition, we incorporated synthetic HDRIs created using the method proposed in [30]. During training, HDRIs are randomly selected with equal probability from either real-world or synthetic collections.

We produced training pairs by projecting the sampled source and target lighting maps onto the reflectance fields of the OLAT images [10]. To derive the ground truth intrinsics, we applied the photometric stereo method [50] and obtained normal and albedo maps.

6. Experiments

This section details our experimental results. We begin with a comparative evaluation of our method against state-of-the-art approaches using the OLAT dataset. We also employ images from the FFHQ–test [24] for user studies. For qualitative analysis, we utilize copyright-free portrait images from Pexels [1]. Additionally, we conduct ablation studies to validate the efficacy of our pre-training framework and architectural design choice. Subsequently, we detail the additional features and conclude by discussing its limitations. Our evaluation uses the OLAT test set, comprising 35 subjects and 11 lighting environments, ensuring no overlap with the train set.

Evaluation metrics. We employ several key metrics for evaluating the prediction accuracy; Mean Absolute Error (**MAE**), Mean Squared Error (**MSE**), Structural Similarity Index Measure (**SSIM**) and Learned Perceptual Image Patch Similarity (**LPIPS**). While these metrics offer valuable quantitative insights, they do not fully capture the subtleties of visual quality enhancement. Therefore, we emphasize the importance of qualitative evaluations to gain a comprehensive understanding of model performance.

Baselines. We compared our approach with three state-of-the-art baselines: Single Image Portrait Relighting (**SIPR**) [44], which uses a single neural network for relighting; Total Relight (**TR**) [33], employing multiple neural networks that incorporate the Phong reflectance model; and **Lumos** [51], a TR adaptation for synthetic datasets. Due to the lack of publicly available code or model from these methods, we either integrated their techniques into our

	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
SIPR [44]	0.1715	0.0748	0.8432	3.648
TR [33]	0.1643	0.0658	0.8465	3.425
Ours	0.1023	0.0275	0.9002	2.137
Ours (w. MMAE)	0.0933	0.0235	0.9051	2.059

Table 1. **Quantitative Evaluation** on the OLAT test set.

	Lumos [51]	TR [33]	Ours
Consistent Lighting	0.0478	0.1852	0.7671
Facial Details	0.2022	0.2602	0.5376
Similar Identity	0.1741	0.2440	0.5819

Table 2. **User Study** on the FFHQ test set.

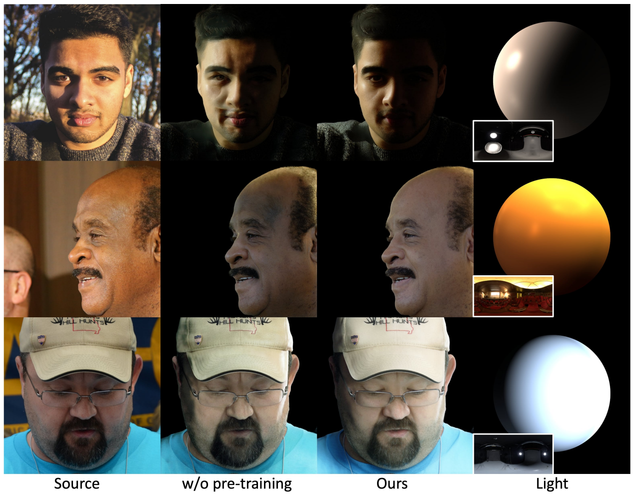

Figure 6. **Impact of Pre-training.** The fine details such as specular highlights, skin tones, and shadows are notably improved.

framework or requested the original authors to process our inputs with their models and share the results.

Quantitative Comparisons. The results in Table. 1 shows our method outperforming SIPR and TR baselines, demonstrating the significance of incorporating advanced rendering physics and reflectance models. The transition from SIPR to TR emphasizes the value of physics-based design, while the shift from TR to our approach underscores the importance of transitioning from the empirical Phong model to the more accurate Cook-Torrance model. Additionally, pre-training contributes to further enhancements, as evidenced by the improved image details, depicted in Fig 6.

Qualitative Comparisons. Our relighting method exhibits several key advantages over previous approaches, as showcased in Fig. 7. It effectively harmonizes light direction and softness, avoiding harsh highlights and inaccurate lighting that are commonly observed in other methods. A notable strength of our approach lies in its ability to capture high-frequency details like specular highlights and hard shadows. Additionally, as shown in the second row, it preserves facial details and identity, ensuring high fidelity to the subject's original features and mitigating common distortions seen in previous approaches. Moreover, our approach excels in handling skin tones, producing natural and accurate results under various lighting conditions. This is clearly demonstrated in the fourth row, where our method contrasts sharply with the over-saturated or pale tones from previous methods. Finally, the nuanced treatment of hair is highlighted in the sixth row, where our approach maintains luster and detail, unlike the flattened effect typical in other methods. More qualitative results are available in our supplementary video demonstration.

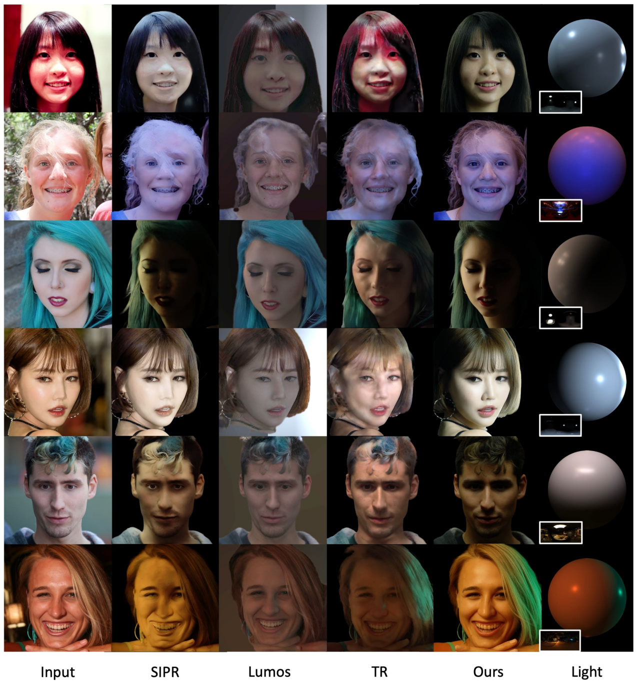

Figure 7. **Qualitative Comparison** on the Pexels images [1].

User Study. We conducted a human subjective test to evaluate the visual quality of relighitng results, summarized in Table. 2. In each test case, workers were presented with an input image and an environment map. They were asked to compare the relighting results from three methods–Ours, Lumos, and TR–based on three criteria: 1) consistency of lighting with the environment map, 2) preservation of facial details, and 3) maintenance of the original identity. To ensure unbiased evaluations, the order of the methods presented was randomized. To aid in understanding the concept of consistent lighting, relit balls were displayed alongside the images. The study included a total of 256 images, consisting of 32 portraits each illuminated with 8 different HDRIs. Each worker was tasked with selecting the best image for each specific criterion, randomly assessing 30 samples. A total of 47 workers participated in the study. The results indicate a strong preference for our results over the baseline methods across all evaluated metrics.

	Method	MAE ↓	MSE ↓	SSIM ↑	LPIPS ↓
Pre-training	MAE [18]	0.0952	0.0242	0.9007	2.096
	MMAE	0.0933	0.0235	0.9051	2.059
DiffuseNet	Albedo	0.1053	0.0295	0.8985	2.197
	Diff Render	0.1023	0.0275	0.9002	2.137

Table 3. **Ablation Studies** on the OLAT test set.

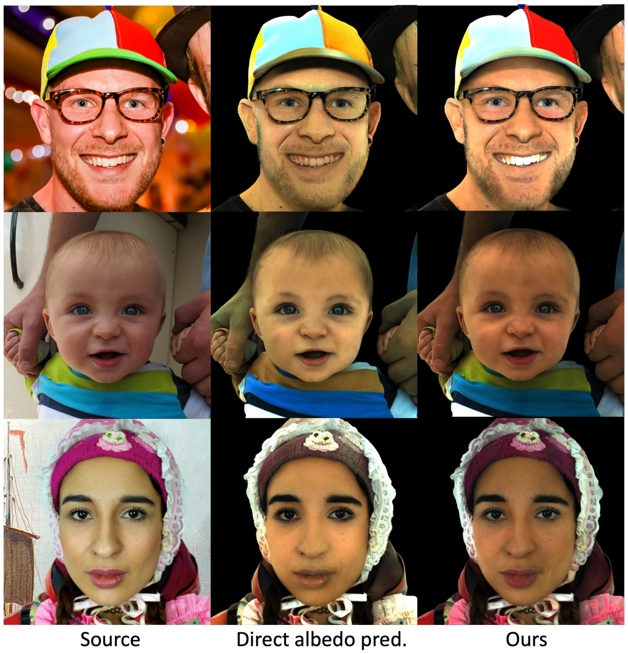

Figure 8. **Ablation on DiffuseNet.** Our approach successfully infers the albedo on various surfaces (skin, teeth, and accessories).

Ablation Studies. We analyze our two major design choices in Table. 3: the MMAE pre-training framework and DiffuseNet. The MMAE, which integrates dynamic masking with generative objectives, outperforms MAE. This superiority is mainly due to the incorporation of challenging masks and global coherence objectives, enabling the model to learn richer features during pre-training. Furthermore, our method of predicting diffuse render demonstrates superiority over direct albedo prediction. Firstly, we see it simplifies the learning process, as predicting diffuse render is more closely related to the original image. Secondly, our approach effectively distinguishes between the influences of illumination (diffuse shading) and surface properties (diffuse render). This distinction is crucial for accurately modeling the intrinsic color of surfaces, as it enables independent and precise evaluation of each element (see Eqn. 9). In contrast, methods that predict albedo directly often struggle to differentiate between these factors, leading to significant inaccuracies in color constancy, as shown in Fig. 8.

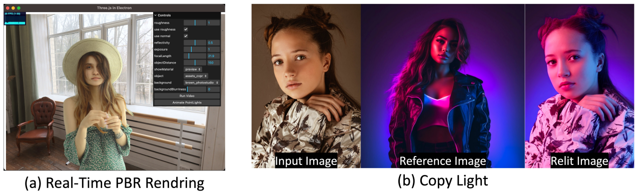

Figure 9. **Applications.** We showcase additional features of SwitchLight, powered by the diverse intrinsics features.

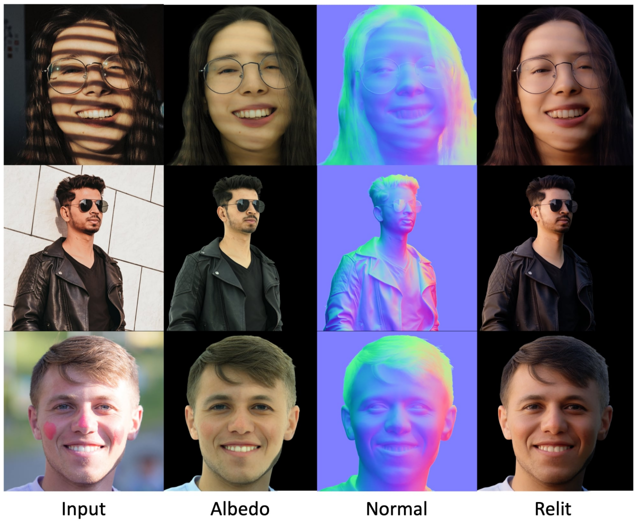

Figure 10. **Limitations.** The model faces challenges in removing strong shadows, misinterpreting reflective surfaces like sunglasses, and inaccurately predicting albedo for face paint.

Applications. We present two applications using predicted intrinsics in Fig. 9. First, real-time PBR via Cook-Torrance components in Three.js graphics library. Second, switching the lighting environment between the source and reference images. Further details are in the supplementary video.

Limitations. We identified a few failure cases in Fig. 10. First, our model struggles with neutralizing strong shadows, which leads to inaccurate facial geometry and residual shadow artifacts in both albedo and relit images. Incorporating shadow augmentation techniques [15, 53] during training could mitigate this issue. Second, the model incorrectly interprets reflective surfaces, such as sunglasses, as opaque objects in the normal image. This error prevents the model from properly removing reflective highlights in the albedo and relit images. Lastly, the model inaccurately predicts the albedo for face paint. Implementing a semantic mask [51] to distinguish different semantic regions separately from the skin could help resolving these issues.

7. Conclusion

We introduce SwitchLight, an architecture based on Cook-Torrance rendering physics, enhanced with a self-supervised pre-training framework. This co-designed approach significantly outperforms previous models. Our future plans include scaling the current model beyond images to encompass video and 3D data. We hope our proposal serve as a new foundational model for relighting tasks.

References

[1] Pexels. https://www.pexels.com. 6, 7

[2] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 2

[4] Hanting Chen, Yunhe Wang, Tianyu Guo, Chang Xu, Yiping Deng, Zhenhua Liu, Siwei Ma, Chunjing Xu, Chao Xu, and Wen Gao. Pre-trained image processing transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12299–12310, 2021. 2

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 2

[6] Xiangyu Chen, Xintao Wang, Jiantao Zhou, Yu Qiao, and Chao Dong. Activating more pixels in image super-resolution transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22367–22377, 2023. 2

[7] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *European Conference on Computer Vision*, pages 606–623. Springer, 2022. 2

[8] Robert L Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (ToG)*, 1(1):7–24, 1982. 2

[9] Paul Debevec. Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *Acm siggraph 2008 classes*, pages 1–10. 2008. 2

[10] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 2, 6

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[12] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 2

[13] Julie Dorsey, James Arvo, and Donald Greenberg. Interactive design of complex time dependent lighting. *IEEE Computer Graphics and Applications*, 15(2):26–36, 1995. 2

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 6

[15] David Futschik, Kelvin Ritland, James Vecore, Sean Fanello, Sergio Orts-Escolano, Brian Curless, Daniel Sỳkora, and Rohit Pandey. Controllable light diffusion for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8412–8421, 2023. 8

[16] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. *arXiv preprint arXiv:1803.07728*, 2018. 2

[17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020. 2

[18] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022. 2, 6, 8

[19] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14719–14728, 2021. 2

[20] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4217–4226, 2022. 2

[21] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 5, 6

[22] Chaonan Ji, Tao Yu, Kaiwen Guo, Jingxin Liu, and Yebin Liu. Geometry-aware single-image full-body human relighting. In *European Conference on Computer Vision*, pages 388–405. Springer, 2022. 2

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5, 6

[24] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 6

[25] Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In *Proceedings of the Asian Conference on Computer Vision*, pages 108–124, 2022. 4

[26] Wenbo Li, Xin Lu, Shengju Qian, Jiangbo Lu, Xiangyu Zhang, and Jiaya Jia. On efficient transformer-based image pre-training for low-level vision. *arXiv preprint arXiv:2112.10175*, 2021. 2

[27] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian L Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Real-time high-resolution background matting.

In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8762–8771, 2021. 4

[28] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European conference on computer vision (ECCV)*, pages 85–100, 2018. 6

[29] Yihao Liu, Jingwen He, Jinjin Gu, Xiangtao Kong, Yu Qiao, and Chao Dong. Degae: A new pretraining paradigm for low-level vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23292–23303, 2023. 2

[30] Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M Patel. Lightpainter: Interactive portrait relighting with freehand scribble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 195–205, 2023. 2, 6

[31] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 2

[32] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision*, pages 69–84. Springer, 2016. 2

[33] Rohit Pandey, Sergio Orts Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics (TOG)*, 40(4):1–21, 2021. 2, 5, 6, 7

[34] Namuk Park, Wonjae Kim, Byeongho Heo, Taekyung Kim, and Sangdoo Yun. What do self-supervised vision transformers learn? *arXiv preprint arXiv:2305.00729*, 2023. 2

[35] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[36] Bui Tuong Phong. Illumination for computer generated pictures. In *Seminal graphics: pioneering efforts that shaped the field*, pages 95–101. 1998. 2

[37] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. *arXiv preprint arXiv:2304.09479*, 2023. 2

[38] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018. 2

[39] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6296–6305, 2018. 2

[40] Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steven M Seitz, and Ira Kemelmacher-Shlizerman. Background matting: The world is your green screen. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2291–2300, 2020. 4

[41] Soumyadip Sengupta, Brian Curless, Ira Kemelmacher-Shlizerman, and Steven M Seitz. A light stage on every desk. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2420–2429, 2021. 2

[42] YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. Style transfer for headshot portraits. 2014. 2

[43] Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)*, 36(4):1, 2017. 2

[44] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019. 2, 6, 7

[45] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T Freeman. Boundless: Generative adversarial networks for image extension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10521–10530, 2019. 6

[46] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Zhang. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20792–20802, 2023. 2

[47] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)*, 39(6):1–13, 2020. 2

[48] Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics (TOG)*, 24(3):756–764, 2005. 2, 6

[49] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16133–16142, 2023. 2, 6

[50] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1):139–144, 1980. 6

[51] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics (TOG)*, 41(6): 1–21, 2022. 2, 6, 7, 8

[52] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 2

[53] Xuaner Zhang, Jonathan T Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E Jacobs. Portrait shadow manipulation. *ACM Transactions on Graphics (TOG)*, 39(4):78–1, 2020. 8

[54] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7194–7202, 2019. 2

[55] Taotao Zhou, Kai He, Di Wu, Teng Xu, Qixuan Zhang, Kuixiang Shao, Wenzheng Chen, Lan Xu, and Jingyi Yu. Relightable neural human assets from multi-view gradient illuminations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4315–4327, 2023. 2