




All-frequency Full-body Human Image Relighting

D. Tajima¹  Y. Kanamori¹  Y. Endo¹ 

¹University of Tsukuba, Japan

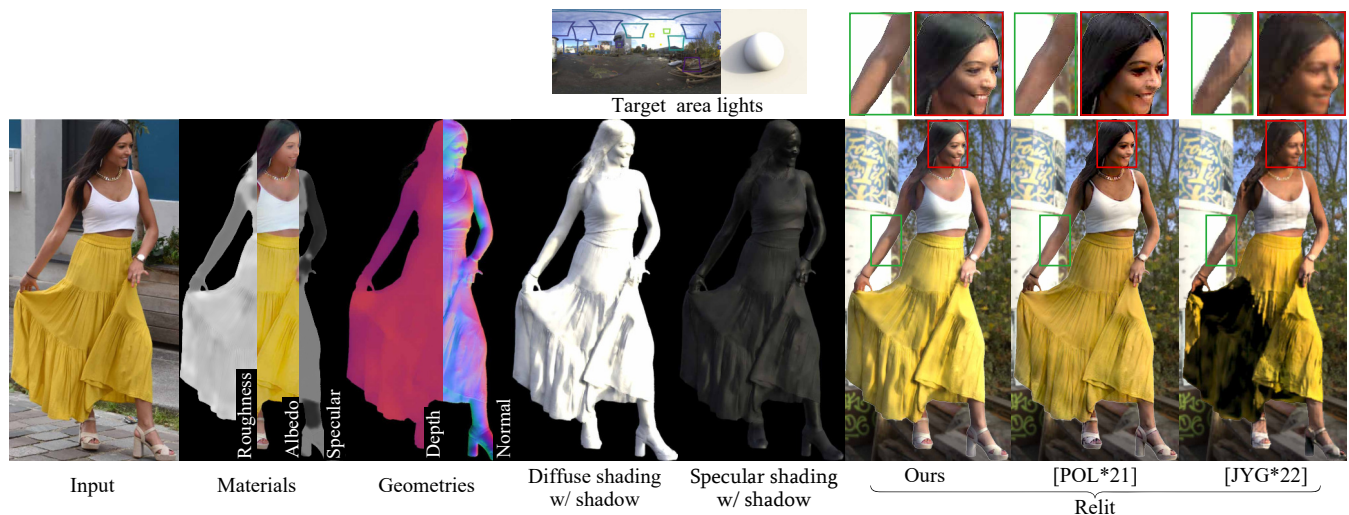


Figure 1: Our method infers the materials (roughness, diffuse albedo, and specular) and geometry (depth and normal) from an input human image and calculates all-frequency shadows and reflections under new lighting conditions. As a lighting representation, we adopt a fixed number of area lights that approximate the target environment map.

Abstract

Relighting of human images enables post-photography editing of lighting effects in portraits. The current mainstream approach uses neural networks to approximate lighting effects without explicitly accounting for the principle of physical shading. As a result, it often has difficulty representing high-frequency shadows and shading. In this paper, we propose a two-stage relighting method that can reproduce physically-based shadows and shading from low to high frequencies. The key idea is to approximate an environment light source with a set of a fixed number of area light sources. The first stage employs supervised inverse rendering from a single image using neural networks and calculates physically-based shading. The second stage then calculates shadow for each area light and sums up to render the final image. We propose to make soft shadow mapping differentiable for the area-light approximation of environment lighting. We demonstrate that our method can plausibly reproduce all-frequency shadows and shading caused by environment illumination, which have been difficult to reproduce using existing methods.

CCS Concepts

• Computing methodologies → Image manipulation; Rendering;

1. Introduction

Human image relighting can alter the lighting effects in a portrait by changing the lighting condition after the photo shoot. The fundamental procedure for human image relighting is to infer the in-

trinsic geometry and reflectance of the target person as well as the scene illumination from the input image via inverse rendering and then render an output image with a new lighting condition. Modern learning-based methods formulate these inverse and forward ren-

dering stages as a unified differentiable pipeline within an analysis-by-synthesis framework.

The current state-of-the-art techniques [SBT*19, ZHSJ19, WYL*20, NLML20, POL*21, YNK*22] employ neural networks to approximate the forward rendering stage without explicitly considering the physical principles involved. In particular, the physical principle of shadows is often ignored; shadows appear when the target geometry occludes the incoming light. Explicitly modeling such light occlusion within a differentiable rendering pipeline has been proven challenging; recent approaches only support hard shadows caused by a single point/directional light [HSB*22, WA23] or adopt a computationally expensive solution via non-differentiable ray tracing with a pre-inferred geometry [JYG*22]. Consequently, neural networks in the state-of-the-art techniques struggle to learn complicated shadow patterns and yield blurry shadows or flickering artifacts with dynamic lighting.

In this paper, we step forward to reproduce physically plausible shadows for all-frequency relighting of human images. We simultaneously model the target geometry and environment illumination as a depth map and a fixed number of area lights within a differentiable framework to reproduce hard-to-soft shadows caused by multiple area lights. The ground-truth area lights for supervised learning are obtained via a novel optimization-based approach. We also infer the diffuse and specular reflectances of the target person for physically based shading. Such geometry and reflectance information is easier to learn with neural networks because it is simpler than the complicated shadow and reflection patterns. We demonstrate that our physically based formulation yields more plausible and stable relighting results even under dynamic lighting than the existing approximate solutions using neural networks (Figure 1).

In summary, our contributions are as follows:

- a two-stage relighting framework with explicit calculations of physically-based shadows and shading,
- a novel approximation of environment illumination with a fixed number of directional lights with area information,
- a differentiable soft shadow calculation with shadow refinement to compute all-frequency shadows, and
- a large-scale synthetic dataset of full-body human images, including ground-truth geometry and reflectance.

We will release our source codes, trained models, and synthetic dataset upon publication.

2. Related Work

There have been numerous studies of single-image relighting. In the following, we focus on our main target, human image relighting. We categorize the previous work in terms of whether they explicitly calculate physically-based shading, use neural network approximations, or explicitly consider shape when calculating shadows.

2.1. Relighting with Physically-based Shading

To calculate physically-based shading, second-order spherical harmonics (SH) have often been used to account for diffuse-only environmental lighting. MoFA [TZK*17] leverages morphable 3D

face models [PKA*09] for face relighting. SfSNet [SKCJ18] employs face inverse rendering to estimate the normal map, diffuse reflectance, and illumination and then performs relighting by replacing the estimated illumination with a new illumination. These face relighting techniques ignore light occlusion due to the almost convex face shapes. In full-body relighting, however, light occlusion is common around limbs and cloth wrinkles and thus should not be ignored. As the first single-image full-body relighting method, Kanamori and Endo [KE18] extended SfSNet [SKCJ18] to consider light occlusion explicitly in the second-order SH formulation by inferring light transport maps with light occlusion, instead of normal maps.

The diffuse-only method [KE18] is extended to handle specular reflections. Tajima *et al.* [TKE21] introduced a refinement network module on top of [KE18] to handle specular reflection and domain adaptation to in-the-wild photographs. Lagunas *et al.* [LSY*21] represented the per-pixel exiting radiance as a double product of fourth-order SH to account for various lighting effects, including specular reflection. However, it is well known that the low-order SH representations cannot represent high-frequency shading and shadows. Our method calculates all-frequency shading and shadows by each of multiple area lights without using SH representations.

2.2. Relighting with Neural Network Approximations

The traditional physically-based calculation of shading and shadowing is complicated and thus often fully or partially replaced with neural network approximations in modern relighting techniques. Early attempts of full approximations [SBT*19, ZHSJ19] formulate relighting as image-to-image translation using single U-Net-like architectures, where new light information is injected at the bottleneck. Unfortunately, these methods do not have mechanisms to handle high-frequency signals. Song *et al.* [SCCZ21] proposed a relighting method for half-body portraits but requires another portrait as a reference of novel illumination.

Partial neural approximations calculate intermediate components and feed them into neural networks for final outputs. Pandey *et al.* [POL*21] calculate multiple frequency bands of Phong-based specular components and merge them using a neural network. Yeh *et al.* [YNK*22] extended their work for domain adaptation in a similar spirit to Tajima *et al.* [TKE21]. However, these methods do not consider light occlusion explicitly and thus cannot handle high-frequency shadows. Yu *et al.* [YME*20] used a neural network to estimate shadows during relighting of outdoor scenes. Nestmeyer *et al.* [NLML20] and Wang *et al.* [WYL*20] introduced neural network modules to estimate specular reflection and shadows. However, because these neural networks are unaware of the underlying geometry, they struggle to reproduce complicated patterns of specular reflection and shadows on diverse full-body human images, resulting in blurry shadings and shadows as well as flickering artifacts with dynamic lighting.

Recent methods based on diffusion models achieve highly photorealistic relighting with lighting control [PTS23, KJY*24, ZDP*24]. However, the interplay between geometry, material, and lighting remains a black box, lacking editability. Intuitive and precise controllability of lighting effects is crucial for relighting.

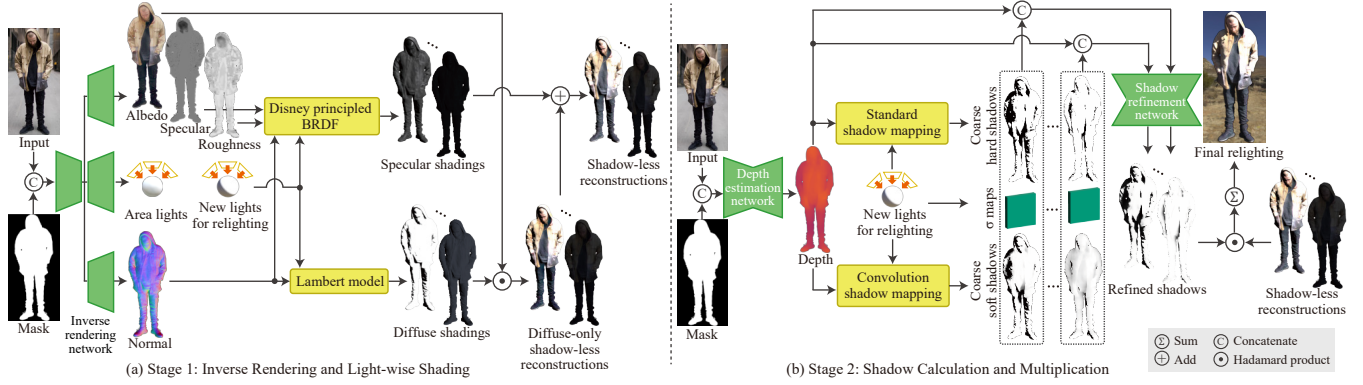


Figure 2: Overview of our two-stage relighting approach. Stage 1 applies inverse rendering and calculates a shadow-less shading image for each light. Stage 2 calculates shadows and multiplies shading images by the shadows and then merges them to output the relighting result.

2.3. Geometry-aware Shadow Calculation

Explicit calculation of light occlusion has been proven to be a key for high-frequency and stable shadows [GRP22, SZP*23]. Ji *et al.* [JYG*22] reconstructs a 3D human model [SSSJ20] and applies ray tracing to calculate all-frequency shadows. Other techniques for reconstructing animatable 3D avatars [CZA*22, ICN*23] can be used in the same way. However, these methods support diffuse reflection only and rely on offline ray tracing, which is non-differentiable and computationally expensive. Hou *et al.* [HZS*21] estimated shadow regions on faces via 3D morphable model fitting but did not support environmental lighting.

Differentiable shadow calculations with geometry information have appeared recently. Hou *et al.* [HSB*22] calculated visibility sampling for hard shadows by a directional light using ray marching with face depth maps. To extend this method for full-body images, however, we have to increase visibility samples due to the much more complicated geometry, which increases the computational burden. Even worse, this method entails incomplete shadows because depth maps have depth gaps. Worchel and Alexa [WA23] proposed a differentiable shadow mapping algorithm for a point/directional light source. They use variance shadow mapping (VSM) [DL06] to enable soft rasterization for differentiable calculation. Although VSM was originally proposed to calculate soft shadows, their method is tailored for point or directional light sources and thus cannot handle soft shadows as is. Contrarily, our method calculates differentiable soft shadows by an area light based on convolutional shadow mapping (CSM) [AMB*07]. Table 1 summarizes the taxonomy of recent techniques for geometry-aware shadow calculation.

3. Method

Figure 2 shows the overview of our method. The inputs of our method are a full-body human image, its binary mask (obtained via off-the-shelf service [Kal] or software [Adoa]), and a new environmental illumination for relighting. We employ a two-stage approach to handle shading and shadows as follows. In Stage 1, we first apply inverse rendering to obtain diffuse and specular reflectance, a set of area lights, and a normal map. We then calculate

Table 1: Taxonomy of recent geometry-aware techniques for full-body human image relighting and shadow calculation. While [NLML20, HSB*22] assume a single directional light, the inference times were measured using 16 lights similarly to ours.

	Approach	Light type	Geometry-aware	Soft shadow	Differentiable	Inference time (sec.)
[NLML20]	CNN	Directional	✗	✗	✗	0.351
[JYG*22]	Ray-tracing + CNN	Environmental	✓	✓	✗	3.50
[HSB*22]	Ray-marching	Directional	✓	✗	✗	12.5
[WA23]	VSM	Directional&Spot	✓	✗	✓	0.150
Ours	CSM + CNN	Area lights	✓	✓	✓	0.835

a shading image for each area light. In Stage 2, we estimate a depth map, calculate shadows for each area light, multiply the shadow maps pixel-wise by the shading images obtained in Stage 1, and then merge them to obtain the final output.

We approximate an environmental illumination as a set of a fixed number of area lights. The motivation for using area lights as an approximation to HDRI maps is to explicitly capture the strong light that causes noticeable highlights and cast shadows, which is difficult to achieve with low resolution HDRI maps and lighting representations with low order spherical harmonics. Let N_L be the number of area lights. Each area light $l \in \{1, \dots, N_L\}$ is parameterized with an RGB intensity $\mathbf{L}_l^{int} \in \mathbb{R}^3$, light direction $\mathbf{L}_l^{dir} \in \mathbb{R}^3$, and area (denoted as the standard deviation σ_l of light l 's Gaussian kernel). We represent these seven parameters for each area light as a light tensor $\mathbf{L} \in \mathbb{R}^{N_L \times 7}$.

3.1. Stage 1: Inverse Rendering and Light-wise Shading

In Stage 1, we perform inverse rendering and obtain a shading image lit by each area light while accounting for both diffuse and specular reflections. First, we estimate a set of a fixed number of area light sources, normal maps, diffuse albedo, specular and roughness maps from the input image through the inverse rendering network, which has a U-Net-like architecture with a single encoder and three decoders. The diffuse albedo and specular/roughness maps are estimated simultaneously by the same decoder. Next, we calculate diffuse and specular shading. To simplify

the shading calculation, we approximately handle each area light as a directional light; *i.e.*, we ignore the area information, which is later utilized in calculating soft shadows in Stage 2. The diffuse shading is calculated from the inferred normal map and each light based on the Lambert model. The specular shading is calculated from the inferred normal map, specular/roughness maps and each light based on the Disney principled BRDF [BS12]. Note that these shading images do not contain shadows, which are later calculated and multiplied pixel-wise in Stage 2. Our explicit lighting calculation allows us to integrate analytical BRDFs, unlike BRDF approximations used in lighting with spherical Gaussians (SGs) [WRG*09, XSD*13, ZLW*21], thereby avoiding errors caused by BRDF approximation. A detailed discussion with SG-based methods is provided in the supplemental material.

We organize the mathematical symbols used in this paper. The super-scripts “*diff*” and “*spec*” indicate diffuse and specular components, respectively. Hat ^ indicates inferred data. *l*-indexing indicates that the datum is inferred for area light *l*. Hat-less symbols are the corresponding ground truth, which are obtained using an offline ray tracer with an environment light (*i.e.*, without discrete approximation and thus without *l*-indexing). We define inferred tensors as follows. Let $\hat{\mathbf{S}}_l$ and $\hat{\mathbf{V}}_l$ be shading and shadow images, respectively. We then define a *shadowed shading* as $\hat{\mathbf{Y}}_l = \hat{\mathbf{V}}_l \odot \hat{\mathbf{S}}_l$ (where \odot denotes the Hadamard product). We also define *shadow-less reconstructions* for diffuse component $\hat{\mathbf{R}}_l^{\text{diff}} = \hat{\mathbf{A}} \odot \hat{\mathbf{S}}_l^{\text{diff}}$ (where $\hat{\mathbf{A}}$ denotes a diffuse albedo) and for both diffuse and specular components $\hat{\mathbf{R}}_l^{\text{full}} = \hat{\mathbf{A}} \odot \hat{\mathbf{S}}_l^{\text{diff}} + \hat{\mathbf{S}}_l^{\text{spec}}$.

For the supervised learning in Stage 1, we use L1 losses for the inferred diffuse albedo and specular/roughness/normal maps obtained via inverse rendering and for shading images $\{\hat{\mathbf{S}}_l\}$ and shadow-less reconstructions $\{\hat{\mathbf{R}}_l\}$ as follows.

$$\mathcal{L}^{\text{texture}} = \sum_{\mathbf{T} \in \mathcal{T}} \|\mathbf{T} - \hat{\mathbf{T}}\|_1, \quad (1)$$

$$\mathcal{L}^{\text{diff shading}} = \left\| \mathbf{S}^{\text{diff}} - \sum_{l=1}^{N_L} \hat{\mathbf{S}}_l^{\text{diff}} \right\|_1, \quad (2)$$

$$\mathcal{L}^{\text{spec shading}} = \left\| \mathbf{S}^{\text{spec}} - \sum_{l=1}^{N_L} \hat{\mathbf{S}}_l^{\text{spec}} \right\|_1, \quad (3)$$

$$\mathcal{L}^{\text{diff recon}} = \left\| \mathbf{R}^{\text{diff}} - \sum_{l=1}^{N_L} \hat{\mathbf{R}}_l^{\text{diff}} \right\|_1, \quad (4)$$

$$\mathcal{L}^{\text{full recon}} = \left\| \mathbf{R}^{\text{full}} - \sum_{l=1}^{N_L} \hat{\mathbf{R}}_l^{\text{full}} \right\|_1, \quad (5)$$

where \mathcal{T} is a set of four types of texture maps \mathbf{T} , *i.e.*, diffuse albedo, roughness map, specular map, and normal map.

We also use the VGG loss [SZ15] for the diffuse albedo $\hat{\mathbf{A}}$, specular shading $\hat{\mathbf{S}}_l^{\text{spec}}$, and shadow-less reconstruction $\hat{\mathbf{R}}_l^{\text{full}}$:

$$\begin{aligned} \mathcal{L}^{\text{vgg}} = & \text{VGG}(\mathbf{A}, \hat{\mathbf{A}}) + \text{VGG}(\mathbf{S}^{\text{spec}}, \sum_{l=1}^{N_L} \hat{\mathbf{S}}_l^{\text{spec}}) \\ & + \text{VGG}(\mathbf{R}^{\text{full}}, \sum_{l=1}^{N_L} \hat{\mathbf{R}}_l^{\text{full}}), \end{aligned} \quad (6)$$

where VGG is a function to calculate the VGG loss.

The illumination of the input image is estimated so that input image can be reconstructed. To learn the light parameters $\{\hat{\mathbf{L}}_l^{\text{int}}, \hat{\mathbf{L}}_l^{\text{dir}}, \hat{\sigma}_l\}$ for area light *l*, we do not use loss functions with their ground truth; our area light set is a discrete approximation of an environment light and is not necessarily unique. For example, there are countless arrangements of area lights to approximate a cloudy sky. In fact, we could not learn these parameters using loss functions with their ground truth. We instead learn these parameters via shading images $\{\hat{\mathbf{S}}_l\}$ and shadow-less reconstructions $\{\hat{\mathbf{R}}_l\}$. The light intensity $\hat{\mathbf{L}}_l^{\text{int}}$ and direction $\hat{\mathbf{L}}_l^{\text{dir}}$ are learned via Equations (2) to (5). To learn the area information $\hat{\sigma}_l$, we calculate soft shadows using our differentiable convolutional shadow mapping (DCSM) function (elaborated in Section 3.2.2), and compare with the ground-truth shadowed shading \mathbf{Y}^{diff} :

$$\tilde{\mathbf{V}}_l = \text{DCSM}(\mathbf{D}, \mathcal{D}(\hat{\mathbf{L}}_l^{\text{dir}}), \hat{\sigma}_l), \quad (7)$$

$$\tilde{\mathbf{S}}_l^{\text{diff}} = \text{LAMBERT}(\mathbf{N}, \mathcal{D}(\hat{\mathbf{L}}_l^{\text{dir}}), \mathcal{D}(\hat{\mathbf{L}}_l^{\text{int}})), \quad (8)$$

$$\mathcal{L}^{\sigma} = \left\| \mathbf{Y}^{\text{diff}} - \sum_{l=1}^{N_L} \tilde{\mathbf{V}}_l \odot \tilde{\mathbf{S}}_l^{\text{diff}} \right\|_1, \quad (9)$$

where DCSM and LAMBERT are functions to calculate soft shadows and Lambert shading, respectively. \mathbf{D} and \mathbf{N} denote the ground-truth depth and normal maps. \mathcal{D} is a detaching operator to detach the argument's gradient from the computational graph. We detach the light intensity $\hat{\mathbf{L}}_l^{\text{int}}$ and direction $\hat{\mathbf{L}}_l^{\text{dir}}$ so that loss function \mathcal{L}^{σ} can focus on the learning of σ_l ; without detaching, we could not learn σ_l well because of the ambiguity of soft shadows.

In summary, the final loss function used for Stage 1 is as follows:

$$\begin{aligned} \mathcal{L}_{\text{decomposition}} = & \mathcal{L}^{\text{texture}} + \mathcal{L}^{\text{diff shading}} + \mathcal{L}^{\text{spec shading}} \\ & + \mathcal{L}^{\text{diff recon}} + \mathcal{L}^{\text{full recon}} + \lambda_{\text{vgg}} \mathcal{L}^{\text{vgg}} + \lambda_{\sigma} \mathcal{L}^{\sigma}, \end{aligned} \quad (10)$$

where $\lambda_{\text{vgg}} = 0.1$ and $\lambda_{\sigma} = 0.01$.

3.1.1. Background-aware light estimation

For light estimation, while previous methods [KE18, TKE21, LSY*21] discard the background information in input images by multiplying them by binary masks, we exploit the background information to improve the light estimation accuracy. Specifically, we concatenate an input image (including the background) and a binary mask to feed the network. Figure 3 shows a qualitative comparison of the light source estimation with and without the background. By including the background, albedo estimation accuracy is improved because the background region provides a cue for color constancy. Improved albedo estimation leads to better normal and lighting estimation accuracy through the shared network components.

3.2. Stage 2: Shadow Calculation and Multiplication

The goal of Stage 2 is to calculate shadows, multiply a shading image by the shadow for each light, and then merge the shadowed reconstructions to generate the final relighting result (see Figure 2, right). We first estimate a depth map from the input image using a depth estimation network. Next, from the estimated depth map and

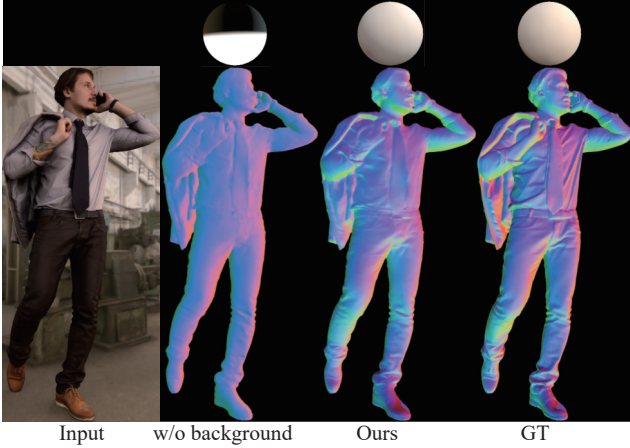


Figure 3: Validation of background-aware light estimation. Top row: spheres shaded with the estimated lights. Bottom row: estimated normal maps.

area lights, we calculate shadows. Although we calculate soft shadows using our differentiable version of convolutional shadow mapping [AMB*07], we find that the soft shadows do not exhibit sharp boundaries near the occludee's surface, which are crucial to reproduce all-frequency shadows. We thus calculate standard shadow mapping as well and merge the hard and soft shadows via a shadow refinement network. We elaborate on each step as follows.

3.2.1. Depth estimation network

We use a U-Net-like architecture to estimate a depth map from an input image (see the supplemental material for our alternative network designs). To ignore absolute depth differences during training, we employ a scale-invariant L1 loss \mathcal{L}_{si} .

$$\mu_D = \frac{1}{\|\mathbf{M}\|_1} \mathbf{M} \odot (\mathbf{D} - \mathcal{D}(\hat{\mathbf{D}})), \quad (11)$$

$$\mathcal{L}_{si} = \|\mathbf{D} - (\hat{\mathbf{D}} + \mu_D \mathbf{M})\|_1, \quad (12)$$

where \mathbf{M} is a binary mask and μ_D is a scalar value. While Eigen *et al.* [EPF14] defined a scale-invariant loss as L2 loss with log-space depth because they wanted to emphasize near pixels while marginalizing far pixels. We do not have to use log-space depth because the depth of human bodies is within a short range, unlike general indoor/outdoor depth maps. In our case, however, we cannot determine the absolute depth only with the scale-invariant loss, so we restrict output depth values within $[0, 1]$ by adding a sigmoid function at the last layer in the network.

We also introduce a regularization term to smooth the surface because slight irregularities in the depth map can cause serious artifacts in the shadow calculation. Specifically, we apply L1 loss to the gradient of depth maps, as Eigen and Fergus [EF15] did with L2 loss:

$$\mathcal{L}_{slope} = \left\| \frac{\partial \mathbf{D}}{\partial x} - \frac{\partial \hat{\mathbf{D}}}{\partial x} \right\|_1 + \left\| \frac{\partial \mathbf{D}}{\partial y} - \frac{\partial \hat{\mathbf{D}}}{\partial y} \right\|_1. \quad (13)$$

The final loss function is

$$\mathcal{L}_{depth} = \mathcal{L}_{si} + \lambda_{slope} \mathcal{L}_{slope}, \quad (14)$$

where $\lambda_{slope} = 0.01$.

3.2.2. Differentiable convolutional shadow mapping (DCSM)

While the original convolutional shadow mapping (CSM) [AMB*07] is non-differentiable, we make it differentiable and implement it using nvdiffrast [LHK*20]. With our differentiable CSM (or DCSM), we can learn the area information of area lights required for calculating soft shadows (Section 3.1) and obtain a set of area lights as a discrete approximation of an environmental illumination via optimization (Section 4.2).

We briefly review CSM. Let \mathbf{x} be a point on the surface visible to the camera, $d(\mathbf{x})$ the distance from \mathbf{x} to the light source, \mathbf{p} the position of the obstacle when trying to view \mathbf{x} from the light source, and $z(\mathbf{p})$ the distance from \mathbf{p} to the light source. The binary shadow test function, which determines whether a point is in shadow or not, is defined as follows:

$$f(d(\mathbf{x}), z(\mathbf{p})) = \begin{cases} 1 & \text{if } d(\mathbf{x}) \leq z(\mathbf{p}) \\ 0 & \text{otherwise,} \end{cases} \quad (15)$$

where 0 indicates shadowed and 1 unshadowed. f is essentially a Heaviside step function and therefore discontinuous. CSM approximates Equation (15) with a continuous function by expanding it with Fourier series.

$$f(d(\mathbf{x}), z(\mathbf{p})) \approx \sum_{i=1}^K \mathbf{a}_i(d(\mathbf{x})) \mathbf{B}_i(z(\mathbf{p})), \quad (16)$$

where B_i is a basis function of $z(\mathbf{p})$, each basis being weighted by a coefficient a_i which depends on $d(\mathbf{x})$. K is the truncation order, and it is known that small K reduces the approximation accuracy, causing light bleeding and ringing. We set $K = 8$ in our method. CSM can vary the shadow hardness/softness via convolution with an arbitrary kernel w_σ . The convolved version s_f of f is as follows:

$$\begin{aligned} s_f(d(\mathbf{x}), z(\mathbf{p})) &= [\mathbf{w}_\sigma * \sum_{i=1}^K \mathbf{a}_i(d(\mathbf{x})) \mathbf{B}_i(z(\mathbf{p}))](\mathbf{p}) \\ &= \sum_{i=1}^K \mathbf{a}_i(d(\mathbf{x})) [\mathbf{w}_\sigma * \mathbf{B}_i(z(\mathbf{p}))](\mathbf{p}), \end{aligned} \quad (17)$$

where $[\mathbf{w} * \mathbf{g}](\mathbf{p})$ means a convolution of \mathbf{g} by the kernel \mathbf{w} in the neighbourhood of \mathbf{p} . In CSM, the shadow hardness is determined by the kernel size, which is an integer value and thus non-differentiable.

To make CSM differentiable, we control the kernel size indirectly via the standard deviation σ of a Gaussian kernel, which is continuous and thus differentiable. We determine the kernel size as $2\lceil 3\sigma \rceil + 1$ (where $\lceil x \rceil$ denotes the smallest integer equal to or larger than x) because 6σ has more than 99% coverage.

As an alternative to CSM, we also made exponential shadow mapping (ESM) [AMS*08] differentiable because ESM was proposed as an improved version of CSM. However, we found such differentiable ESM slows down optimization; clamping values greater than one in the shadow test function hinders gradient propagation.

3.2.3. Shadow refinement network

Unfortunately, our DCSM inherits the limitation of the original CSM; although soft shadows in the real world have varying shadow hardness because the penumbra widths vary with the distance between the occluder and occludee, CSM does not take this into account. Even worse, shadows obtained from a depth map are often incomplete due to the missing geometry between depth gaps.

To address these problems, we also calculate hard shadows for high-frequency shadows and merge the hard and soft shadows via a shadow refinement network that also has a U-Net-like architecture. The hard shadows are calculated from a 3D mesh reconstructed from the depth map as a byproduct of DCSM and thus without additional burden. We find that the standard deviation σ of a Gaussian kernel well represents the shadow softness, so we feed not only the hard and soft shadows but also a constant map of σ after concatenating them for each area light (see Section 5.2.1 for the ablation study with and without σ).

To train the shadow refinement network, we define a loss function with the refined shadow $\tilde{\mathbf{V}}_l$ obtained from the shadow refinement network and the ground-truth shadowed shading \mathbf{Y}^{diff} as follows:

$$\mathbf{S}_l^{diff} = \text{LAMBERT}(\mathbf{N}, \mathbf{L}_l^{dir}, \mathbf{L}_l^{int}), \quad (18)$$

$$\mathcal{L}_{refinement} = \left\| \mathbf{Y}^{diff} - \sum_{l=1}^{N_L} \tilde{\mathbf{V}}_l \odot \mathbf{S}_l^{diff} \right\|_1. \quad (19)$$

4. Dataset

4.1. Full-body Human Dataset

We create a synthetic 3D human model dataset using Blender's add-on tool for generating 3D human models [Pos] to obtain ground-truth full-body human image data. The 3D human models include non-diffuse materials based on a simplified version of the Disney principled BRDF, but the subsurface scattering, anisotropy, and metal parameters are not considered. The identity, standing pose, clothing, and camera direction of each 3D human model were randomly determined, resulting in 2,500 3D human models with different identities. Of these, 2,400 were used for training and 100 for testing. For each 3D human model, we render a binary mask, diffuse albedo map, specular map, roughness map, normal map, and depth map at a resolution of 1024×1024 pixels. The ground truth data for the relighting images was rendered using ray tracing with a virtual light stage created in Blender using HDR environment maps collected from Poly Haven [Pol] without considering indirect illumination. 487 environment maps were used for training and 34 for testing. For each 3D human model, eight environment maps were randomly selected and randomly rotated along the longitude to increase the variation. Figure 4 shows some example data used in our experiments. We also used 541 scanned 3D human models obtained from various commercial websites to further increase the variation. Note that some of the scanned models do not contain ground-truth specular and roughness maps, so for such data, we omit the corresponding loss functions during training.

4.2. Area Light Dataset

We create a dataset of our area light approximations of environmental lights. The source environment maps are the same HDR images as those used for background images. To construct a discrete approximation of environmental illumination, Annen *et al.* [ADM*08] proposed a greedy algorithm that outputs a varying number of area lights. Their approach is not suited for our purpose because we want a fixed-size tensor to learn using our inverse rendering network. We propose a novel optimization-based algorithm to output a fixed number of area lights.

As discussed on \mathcal{L}^σ (Equation 9) in Section 3.1, we focus on shading images as the cue for optimization. Specifically, we put a hemisphere on a plane, illuminate the scene with the target environment illumination, and arrange area lights so that the shading and shadows match with those lit by the environment illumination. Not to miss strong incoming lights, we rotate the environment map five times by 72° in longitude and $\pm 90^\circ$ in latitude, resulting in seven images of shading and shadows rendered using ray tracing at the resolution of 256×256 pixels. Each optimization iteration accounts for these seven images simultaneously (for simplicity, we omit the loop for these seven images). Figure 5 shows the overview of the optimization process. We optimize light intensity \mathbf{L}_l^{int} , direction \mathbf{L}_l^{dir} , and standard deviation σ_l of light l 's Gaussian kernel for each area light $l \in \{1, 2, \dots, N_L\}$. The optimization process undergoes the following three steps:

Step 1: Initialize $\{\sigma_l, \mathbf{L}_l^{int}, \mathbf{L}_l^{dir}\}$,

Step 2: Optimize \mathbf{L}_l^{int} and \mathbf{L}_l^{dir} of each light l , and

Step 3: Optimize σ_l of each light l , while fixing \mathbf{L}_l^{int} and \mathbf{L}_l^{dir} .

Step 1 initializes the light directions so that N_L lights distribute uniformly on the environment map. The light intensities and σ_l are initialized with constant values. The initial value of σ_l depends on the shadow map resolution. In our case, we set $\sigma_l = 10$ for a 256×256 shadow map.

Step 2 utilizes both diffuse and specular shadings to optimize the intensity and direction of each light. As alternative strategies, diffuse-only reflection results in blurry shading and yields a high degree of freedom in light directions, whereas specular-only shading causes a concentration of light directions contributing to the most glossy directions, reducing the reproducibility of diffuse shading. We set the specular and roughness parameters as 0.5. The following L1 loss functions are used for optimization:

$$\mathcal{L}_{diff} = \left\| \mathbf{S}^{diff sph} - \hat{\mathbf{S}}^{diff sph} \right\|_1, \quad (20)$$

$$\mathcal{L}_{spec} = \left\| \mathbf{S}^{spec sph} - \hat{\mathbf{S}}^{spec sph} \right\|_1. \quad (21)$$

Furthermore, each light direction is encouraged to move away from each other to avoid overlapping light directions. Specifically, we add a regularization term so that the dot product of each light direction pair is not greater than τ .

$$\mathcal{L}_{rep} = \frac{1}{N_L(N_L - 1)} \sum_{l=1}^{N_L-1} \sum_{m \neq l}^{N_L-1} \left| \max(\tau, \langle \mathbf{L}_l^{dir}, \mathbf{L}_m^{dir} \rangle) - \tau \right|^2. \quad (22)$$

We set $\tau = 0.65$.

In Step 3, naively optimizing σ_l in terms of luminance does not

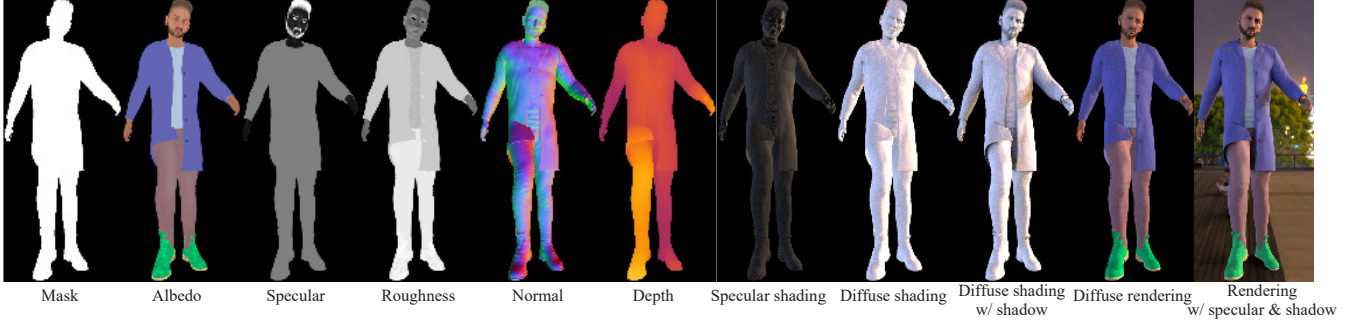


Figure 4: Example data generated from a 3D human model.

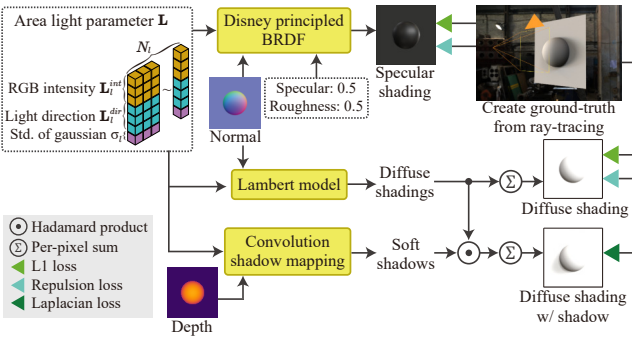


Figure 5: Overview of our area light optimization.



Figure 6: Visualization of optimized light parameters on an environment map. The light's area is visualized as the rectangle size, while the rectangle color indicates the light's intensity.

work well because most pixels are inside or outside shadows and their luminance values do not change during optimization. We thus put emphasis on shadow boundaries by applying a Laplacian filter:

$$\hat{\mathbf{V}}_l^{sph} = \text{DCSM}(\mathbf{D}^{sph}, \mathcal{D}(\hat{\mathbf{L}}_l^{dir}), \hat{\sigma}_l), \quad (23)$$

$$\mathcal{L}_{lap} = \sum_{k \in \{15, 21, 33\}} \left\| \mathcal{L}_k(\mathbf{Y}^{diff sph}) - \mathcal{L}_k \left(\sum_{l=1}^{N_L} \hat{\mathbf{V}}_l^{sph} \odot \mathcal{D}(\hat{\mathbf{S}}_l^{diff sph}) \right) \right\|_1, \quad (24)$$

where \mathbf{D}^{sph} is the ground-truth depth map, $\hat{\mathbf{V}}^{sph}$ is the inferred shadows of each light source, \mathcal{L}_k is a Laplacian operator with kernel size $k \times k$, and $\mathbf{Y}^{diff sph}$ is the ground-truth shadowed shading for diffuse component.

Regarding the choice of the number of area lights, N_L , the larger N_L is, the better accuracy we can obtain with more computational

cost. We use $N_L = 16$ throughout this paper due to the trade-off between the accuracy and computational cost. See Appendix A for our experiment with different N_L .

The optimization was implemented using Python and PyTorch and performed on NVIDIA RTX A5000. We used Adam as an optimizer, setting the exponential decay rates for the moment estimates as $\{0.5, 0.999\}$. The learning rate was controlled in the range of $[1, 0.00001]$ by the cosine annealing scheduler within 20 epochs per cycle. Each step took 1,000 iterations and 300 iterations to converge. The optimization took about 5 minutes per environment map. Figure 6 visualizes the final area light parameters. For each light, the light intensity is color-coded and σ_l value is visualized as the size of rectangle. We can observe that the area lights are distributed appropriately in the environment map.

5. Experiments

5.1. Implementation details

We implemented our method using Python and PyTorch and performed training and inference on NVIDIA RTX A5000. We trained the inverse rendering network (Stage 1), depth estimation network, shadow refinement network (Stage 2) separately. We used Adam as an optimizer, setting the exponential decay rates for the moment estimates as $\{0.5, 0.999\}$. We used the cosine annealing scheduler to control the learning rate in the range of $[0.01, 0.00001]$ within 20 epochs per cycle. Our batch size was eight. The computational times for one-epoch training of the inverse rendering network, depth estimation network, and shadow refinement network were about 60, 50, and 50 minutes, respectively, when we used one GPU to process 1024×1024 images. We terminated the training at 300, 500, and 120 epochs for the respective networks, where each learning curve reached a plateau. The time for testing a 1024×1024 input image was about 1.01 seconds. As the time breakdown, inverse rendering network, depth estimation network, shadow mapping, shadow refinement network, and relighting took about 0.156, 0.0322, 0.747, 0.0875, and 0.0316 seconds, respectively.

5.2. Ablation Studies

We first conduct ablation studies using our dataset to evaluate the effectiveness of our method for improving shadows and specular components.

5.2.1. Comparison with different types of shadows

Figure 7 shows a comparison with standard shadow mapping (SM), convolutional shadow mapping (CSM), and our shadow refinement network. Here, “Refined w/o CSM or σ ” means that the shadow refinement network does not use soft shadows by CSM or standard deviation σ as input, and “Refined w/o CSM” means that it does not use soft shadows by CSM. As can be seen in the results, “SM” exhibits noticeable jaggies in the shadow contours by trying to approximate the environment light using a small number of lights. “CSM” can mitigate this problem using area lights but causes light bleeding when an occludee is near an occluder. In addition, “CSM” cannot reproduce contact shadows according to the distance between an occludee and an occluder. “Refined w/o CSM or σ ” can reproduce contact shadows, but the shadows have the same softness regardless of the lights because no area light information is given. Meanwhile, “Refined w/o CSM”, which uses a σ map as input, reproduces not only contact shadows but also shadow softness depending on the area lights. Using soft shadows as an additional input, “Ours (full)” reproduces the most accurate shadows. In addition, as shown in the quantitative results in Table 2, we can see step-by-step improvements by using CSM and a σ map as input of “Ours (full)”. Ours also achieved the best scores in all metrics.

5.2.2. Comparison of specular components

To evaluate the effectiveness of specular shading by the Disney principled BRDF, we compare it with the Blinn-Phong reflection model. To do so, we estimate a Blinn-Phong specular exponent map instead of a specular map and a roughness map using the inverse rendering network. Because ground-truth exponent maps are unavailable, we trained the model by computing the losses only for relit images containing specular shading. Figure 8 and Table 3 show the qualitative and quantitative results, respectively. The Blinn-Phong reflection model causes errors, especially around grazing angles, because it is less physically-plausible. In addition, we can see that a single parameter is insufficient to reproduce reflectance properties in clothed full-body human images.

5.2.3. Comparison of relighting

We verified the effectiveness of the shadows and specular reflections considered in our method with relighting results. We quantitatively compared the following four conditions: no shadow or specularity (“Ours w/o shadow or specular”), no shadow refinement or shadow (“Ours w/o refinement or specular”), no specular (“Ours w/o specular”) and “Ours (full)”. A quantitative comparison of the relit results is shown in Table 4. “Ours w/o refinement or specular” with direct use of shadows by CSM improved accuracy in all metrics except LPIPS compared to “Ours w/o shadow or specular”. “Ours w/o specular” with shadow refinement shows a further improvement in the accuracy of estimating the relighting results. Furthermore, “Ours (full)”, which takes specular into account, shows a significant effect on the accuracy improvement.

5.3. Comparison with Existing Methods

We compared our method with existing relighting methods specialized for human face images [SBT*19, NLML20, HSB*22],

Table 2: Quantitative comparison of shading with shadows (mean \pm standard deviation). The best scores are in bold, and the second-best scores are underlined. Each method is explained in Section 5.2.1.

	RMSE \downarrow	SSIM \uparrow	LPIPS \downarrow
[NLML20]	0.457 \pm 0.627	0.457 \pm 0.190	0.106 \pm 0.0215
[HSB*22]	0.141 \pm 0.0590	0.636 \pm 0.114	0.0722 \pm 0.0197
[JYG*22]	0.145 \pm 0.0619	0.612 \pm 0.110	0.0891 \pm 0.0156
Ours	0.100\pm0.0454	0.709\pm0.110	0.0610\pm0.0180
Refined w/o CSM	0.101 \pm 0.0455	0.708 \pm 0.110	0.0611 \pm 0.0180
Refined w/o CSM or σ	0.102 \pm 0.0462	0.704 \pm 0.109	0.0618 \pm 0.0180
CSM	0.142 \pm 0.0548	0.652 \pm 0.111	0.0703 \pm 0.0203
SM	0.135 \pm 0.0542	0.646 \pm 0.112	0.0702 \pm 0.0190

upper body images [POL*21], and full-body humans [LSY*21, TKE21, JYG*22]. We used the public pre-trained models for several methods [TKE21, LSY*21] and trained the models using our CG dataset from scratch for the other methods. For the methods [NLML20, HSB*22], which assume relighting with a single directional light, we used 16 directional lights obtained from our 16 area lights, while ignoring σ , for a fair comparison.

5.3.1. Comparison of shadows

Table 2 and Figure 9 show the quantitative and qualitative comparisons with the existing relighting methods that explicitly generate shadows. The method by Nestmeyer *et al.* [NLML20] fails to reproduce shadows because it does not consider human shapes during visibility estimation with a CNN. This method also has the highest standard deviation compared to the other methods. This is because learning to estimate shadows with varying illumination and geometry is more difficult than estimating shape alone and performing physical lighting calculations. Although the method by Hou *et al.* [HSB*22], like our method, uses meshes reconstructed from depth maps, it reproduces only hard shadows caused by directional lights and yields artifacts around the shadow boundaries. The method by Ji *et al.* [JYG*22] also cannot obtain satisfactory results because of incomplete meshes estimated using PIFuHD [SSSJ20] and insufficient refinement by their shadow refinement module. In contrast, our method has fewer shadow artifacts due to the lack of complete geometry in the invisible regions. Furthermore, the use of area light sources allows us to reproduce low to high-frequency shadows.

As for computational time, these existing methods took about 17 seconds [HSB*22] and 13 seconds [JYG*22] to relight a single image. These computational times consist mainly of 3D reconstruction and shadow calculation. Meanwhile, our method is much faster and took about 1 second for inference.

5.3.2. Comparison of specular components

Figure 10 shows a comparison of specular shading with the existing methods. For the method by Pandey *et al.* [POL*21], we computed a pseudo-specular output by subtracting a diffuse-only relit image generated using a diffuse light map from a final relit image. [LSY*21] reproduces gloss using fourth-order spherical harmonics but fails to adequately approximate the high-frequency components of the environmental lighting. On the other hand, the meth-

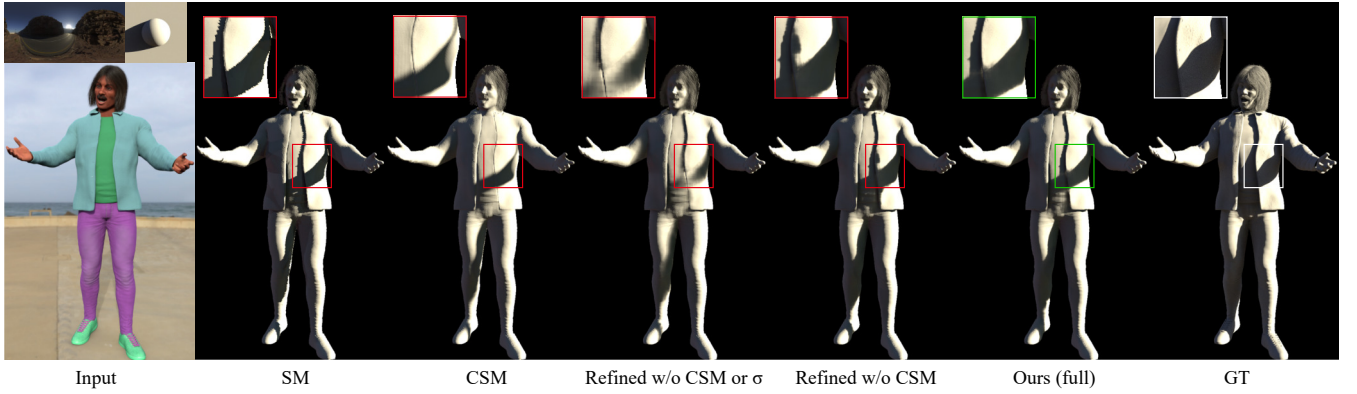


Figure 7: Ablation study of shading with shadows. The environment map for relighting and its reference shading/shadowing pattern with a sphere are visualized above the input image. The rectangles emphasize the differences. Each method is explained in Section 5.2.1.

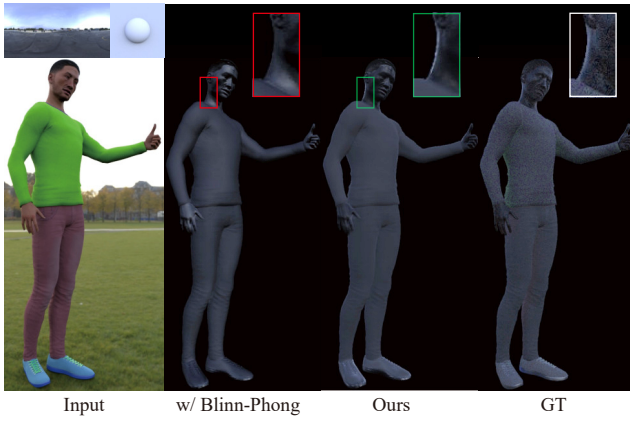


Figure 8: Ablation study of specular components. For better visualization, the output images are uniformly multiplied by a constant factor.

ods approximating gloss using neural networks [NLML20, TKE21, POL*21] struggle to learn gloss effectively, leading to blurry output. This can be also observed in their large standard deviations. In contrast, our method exhibits high-frequency specular reflection on skin regions in the results. The quantitative comparison in Table 3 also shows that our method achieves the best performance in all metrics.

5.3.3. Comparison of relighting results

Table 4 and Figures 12 and 13 show the quantitative and qualitative comparisons of relighting results between our method and existing methods, respectively. The real images in Figure 12 were taken from Unsplash, while Figure 13 used the SHHQ dataset [FLJ*22]. For the real photograph inputs, there is no ground-truth relit image. The shadows and highlights in the input images do not affect the relighting results so much, thanks to the clean synthetic training data. This benefit is shared not only by our method but also by other methods trained on the same dataset. The methods [SBT*19,

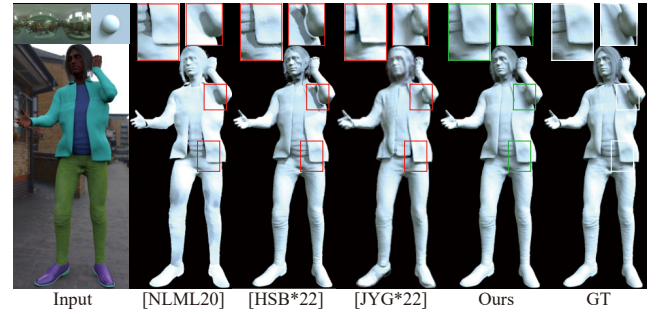


Figure 9: Qualitative comparison of shading with shadows between our method and the existing methods.

Table 3: Quantitative comparison of specular shading (mean \pm standard deviation). For the method by Pandey et al. [POL*21], we compute a pseudo-specular output from the difference between a relit image and a diffuse-only relit image.

	RMSE↓	SSIM↑	LPIPS↓
[NLML20]	0.0506 \pm 0.0530	0.741 \pm 0.124	0.0813 \pm 0.0155
[LSY*21]	0.0564 \pm 0.0517	0.460 \pm 0.163	0.208 \pm 0.0428
[TKE21]	0.0484 \pm 0.0345	0.406 \pm 0.172	0.107 \pm 0.0155
[POL*21]	0.0484 \pm 0.0343	0.479 \pm 0.155	0.0910 \pm 0.0170
Ours	0.0334\pm0.0262	0.835\pm0.0908	0.0698\pm0.0127
Blinn-Phong	0.0390 \pm 0.0295	0.746 \pm 0.106	0.0898 \pm 0.0132

NLML20] that do not consider the physical laws overall yielded blurry outputs. Because the studies [LSY*21, TKE21] use illumination approximated with low-order SH, they struggle to approximate high-frequency illumination, and the contrast of the shading becomes strong. Because of the same reason, these methods cannot also reproduce high-frequency highlights and shadows. The method by Pandey et al. [POL*21] handles low-frequency highlights but fails to reproduce high-frequency highlights and shadows. [HSB*22] uses directional light sources to represent illumination and cannot handle low-frequency shadows, resulting in no-

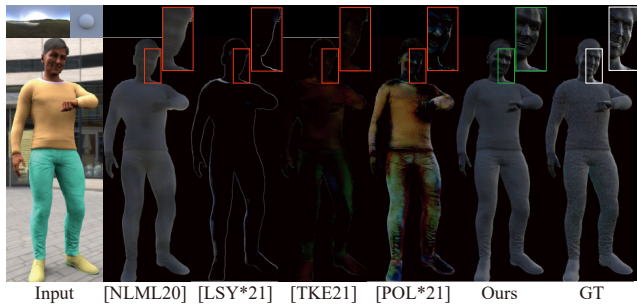


Figure 10: Qualitative comparison of specular components between our method and the existing methods.

Table 4: Quantitative comparison of relighting results (mean±standard deviation).

	RMSE↓	SSIM↑	LPIPS↓
[SBT*19]	0.122±0.0944	0.712±0.148	0.0556±0.0196
[NLML20]	0.137±0.0843	0.678±0.117	0.0717±0.0157
[LSY*21]	0.224±0.122	0.483±0.149	0.162±0.0437
[TKE21]	0.173±0.0980	0.586±0.166	0.0737±0.0206
[POL*21]	0.111±0.0645	0.692±0.116	0.0538±0.0145
[HSB*22]	0.101±0.0723	0.632±0.120	0.0598±0.0235
[JYG*22]	0.121±0.0650	0.686±0.110	0.0819±0.0132
Ours (full)	0.0744±0.0411	0.787±0.0973	0.0493±0.0131
Ours w/o specular	0.0780±0.0415	0.771±0.0972	0.0502±0.0133
Ours w/o refinement or specular	0.0882±0.0479	0.757±0.100	0.0528±0.0135
Ours w/o shadow or specular	0.0946±0.0559	0.753±0.103	0.0518±0.0138

ticeable artifacts around shadow boundaries. In addition, the mesh recovered from the depth map lacks invisible geometry, resulting in inaccurate shadows. In addition, this method is limited to diffuse reflections and cannot reproduce gloss. [JYG*22] can also reproduce high-frequency shadows by ray-tracing, but artifacts are noticeable. This is due to the presence of many defects and artifacts in the 3D reconstruction results by PIFuHD [SSSJ20]. In addition, it is limited to diffuse reflections and cannot reproduce gloss. Our method can generate high-frequency shadows due to the light occluded by hands and legs, as well as natural skin highlights, by performing physically-based lighting with specular component and 3D geometry. More results are contained in the supplementary material.

To further validate our method, we conducted a user study with 20 real images collected from Unsplash and lit under natural illumination. We compared four methods with the top scores with synthetic data. 22 subjects were requested to pick the most natural relighting result for each input image, emphasizing facial highlights and shadows around arms and clothes. Consequently, the selection percentages are: [SBT*19] 4.1%, [POL*21] 16.6%, [HSB*22] 18.6%, and ours 60.7%, which means that ours outperforms other methods. The relighting results used in the user study are published in the supplementary material.

5.3.4. Evaluation of temporal consistency

We evaluated the temporal consistency for relit videos obtained using dynamic lights. For 10 static 3DCG human data of different identities, we generated 20 videos consisting of 128 frames by ro-

Table 5: Quantitative comparison of temporal consistency for dynamic lights and people using our metric (Equation (26)).

	$\partial\text{RMSE} \downarrow$	
	Dynamic lights	Dynamic people
[SBT*19]	0.473	0.140
[NLML20]	0.447	0.123
[LSY*21]	0.530	0.176
[TKE21]	0.470	0.132
[POL*21]	0.394	<u>0.116</u>
[HSB*22]	0.399	0.121
[JYG*22]	0.388	<u>0.116</u>
Ours	0.386	0.113

tating two randomly-selected test environment maps in the longitude direction.

A typical evaluation metric for temporal consistency uses a warping function based on optical flow. However, the warping function is inappropriate for static humans relit with dynamic lights. Regarding evaluation metrics, [LZC*24] proposed the color distribution consistency index (CDC) as a measure of temporal consistency when applying a colorization task to each video frame. CDC calculates the similarity of the color distribution between consecutive frames. However, CDC cannot handle dynamically changing color distributions due to dynamic lighting and prefers blurry frames and frames without shading changes because CDC does not refer to ground truth. Therefore, inspired by CDC, we use the following metric to evaluate frame differences within different frame intervals with reference to ground truth:

$$\partial\text{RMSE}_t = \frac{1}{N_f - t} \sum_{i=1}^{N_f - t} \text{RMSE}(\mathbf{F}_{i+t} - \mathbf{F}_i, \hat{\mathbf{F}}_{i+t} - \hat{\mathbf{F}}_i), \quad (25)$$

$$\partial\text{RMSE} = \frac{1}{3} (\partial\text{RMSE}_1 + \partial\text{RMSE}_2 + \partial\text{RMSE}_4), \quad (26)$$

where \mathbf{F} , N_f , and t are the ground-truth relighting frames, the total number of frames, and the time step, respectively. As with CDC, the use of different time steps allows for long- and short-term temporal consistency. The second column of Table 5 shows the results. We can see that our relighting results faithfully reproduce the changes in shading, yielding a temporally consistent output.

Furthermore, we quantitatively evaluated the temporal consistency of relighting results with synthetic dynamic people by preparing eight animation sequences with four rigged human models animated using Mixamo [Adobe] and a pair of environment maps for before and after relighting. The third column of Table 5 shows the results. Ours yields the least flickering with the best fidelity to the ground truth.

6. Conclusions

We have proposed a relighting method for full-body images of clothed humans, taking into account low- to high-frequency specular reflections and shadows. Our method first applies inverse rendering to the input human image to obtain diffuse and specular reflectance maps, roughness map, and depth map, and then reproduces both low- and high-frequency gloss and shadows based on

the Disney principled BRDF and convolutional shadow mapping. A new lighting representation with a fixed number of area lights was proposed to implement it. The experimental results revealed that our method reproduces shadows and highlights more plausibly than existing methods that approximate shadows and highlights using neural networks, demonstrating the effectiveness of formulating highlights and shadows based on physically-based lighting.

7. Limitations & Future Work

Our method has several limitations. The most prominent one is the CG-like appearance in some of our results. This is primarily because we rely on our synthetic training dataset to learn glossy reflection. However, to the best of our knowledge, there are no other large-scale human datasets with glossy components. Constructing a more photorealistic large-scale dataset of 3D human models is one of our future goals. Furthermore, to bridge the domain gap between CG and real images, a promising way would be to incorporate domain adaptation networks trained on real-world data (e.g., [TKE21, YNK*22]). Such domain adaptation techniques will alleviate residual differences between CG and real images caused by factors such as indirect illumination, subsurface scattering, and anisotropic reflections.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. This work was supported by JSPS KAKENHI Grant Number 23K24862 and JST SPRING Grant Number JPMJSP2124.

References

- [ADM*08] ANNEN T., DONG Z., MERTENS T., BEKAERT P., SEIDEL H.-P., KAUTZ J.: Real-time, all-frequency shadows in dynamic scenes. *ACM Trans. Graph.* 27, 3 (2008), 1–8. [6](#)
- [Adoa] ADOBE: Adobe Photoshop. <https://www.adobe.com/products/photoshop.html>. [3](#)
- [Adob] ADOBE: Mixamo. <https://www.mixamo.com/> [accessed: 1 February 2024]. [10](#)
- [AMB*07] ANNEN T., MERTENS T., BEKAERT P., SEIDEL H., KAUTZ J.: Convolution shadow maps. In *Proceedings of the Eurographics Symposium on Rendering Techniques* (2007), pp. 51–60. [3, 5](#)
- [AMS*08] ANNEN T., MERTENS T., SEIDEL H., FLERACKERS E., KAUTZ J.: Exponential shadow maps. In *Proceedings of the Graphics Interface 2008 Conference* (2008), pp. 155–161. [5](#)
- [BS12] BURLEY B., STUDIOS W. D. A.: Physically-based shading at Disney. In *ACM SIGGRAPH* (2012), vol. 2012, pp. 1–7. [4](#)
- [CZA*22] CORONA E., ZANFIR M., ALLDIECK T., BAZAVAN E. G., ZANFIR A., SMINCHISDESCU C.: Structured 3D features for reconstructing relightable and animatable avatars. *CoRR abs/2212.06820* (2022). [3](#)
- [DL06] DONNELLY W., LAURITZEN A.: Variance shadow maps. In *Proc. of SIGGRAPH 2006* (2006), pp. 161–165. [3](#)
- [EF15] EIGEN D., FERGUS R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *ICCV* (2015), pp. 2650–2658. [5](#)
- [EPF14] EIGEN D., PUHRSCHE C., FERGUS R.: Depth map prediction from a single image using a multi-scale deep network. In *NIPS* (2014), pp. 2366–2374. [5](#)
- [FLJ*22] FU J., LI S., JIANG Y., LIN K., QIAN C., LOY C. C., WU W., LIU Z.: StyleGAN-Human: A data-centric odyssey of human generation. In *ECCV* (2022), pp. 1–19. [9, 14](#)
- [GRP22] GRIFFITHS D., RITSCHER T., PHILIP J.: OutCast: Outdoor single-image relighting with cast shadows. *Comput. Graph. Forum* 41, 2 (2022), 179–193. [3](#)
- [HSB*22] HOU A. Z., SARKIS M., BI N., TONG Y., LIU X.: Face relighting with geometrically consistent shadows. In *CVPR* (2022), pp. 4207–4216. [2, 3, 8, 9, 10](#)
- [HZS*21] HOU A., ZHANG Z., SARKIS M., BI N., TONG Y., LIU X.: Towards high fidelity face relighting with realistic shadows. In *CVPR* (2021), pp. 14719–14728. [3](#)
- [ICN*23] IQBAL U., CALISKAN A., NAGANO K., KHAMIS S., MOLCHANOV P., KAUTZ J.: RANA: relightable articulated neural avatars. In *ICCV* (2023), pp. 23085–23096. [3](#)
- [JYG*22] JI C., YU T., GUO K., LIU J., LIU Y.: Geometry-aware single-image full-body human relighting. In *ECCV* (2022), pp. 388–405. [2, 3, 8, 10](#)
- [Kal] KALEIDO AI: remove.bg. <https://www.remove.bg/>. [3](#)
- [KE18] KANAMORI Y., ENDO Y.: Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph.* 37, 6 (2018), 270. [2, 4](#)
- [KJY*24] KIM H., JANG M., YOON W., LEE J., NA D., WOO S.: SwitchLight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *CVPR* (2024), pp. 25096–25106. [2](#)
- [LHK*20] LAINE S., HELLSTEN J., KARRAS T., SEOL Y., LEHTINEN J., AILA T.: Modular primitives for high-performance differentiable rendering. *ACM Trans. Graph.* 39, 6 (2020), 194:1–194:14. [5](#)
- [LSY*21] LAGUNAS M., SUN X., YANG J., VILLEGAS R., ZHANG J., SHU Z., MASÍÁ B., GUTIERREZ D.: Single-image full-body human relighting. In *EGSR* (2021), pp. 167–177. [2, 4, 8, 9, 10](#)
- [LZC*24] LIU Y., ZHAO H., CHAN K. C. K., WANG X., LOY C. C., QIAO Y., DONG C.: Temporally consistent video colorization with deep feature propagation and self-regularization learning. *Comput. Vis. Media* 10, 2 (2024), 375–395. [10](#)
- [NLML20] NESTMEYER T., LALONDE J., MATTHEWS I. A., LEHRMANN A. M.: Learning physics-guided face relighting under directional light. In *CVPR* (2020), pp. 5123–5132. [2, 3, 8, 9, 10](#)
- [PKA*09] PAYSAN P., KNOTHE R., AMBERG B., ROMDHANI S., VETTER T.: A 3D face model for pose and illumination invariant face recognition. In *AVSS* (2009), pp. 296–301. [2](#)
- [Pol] POLY HAVEN: Poly Haven. <https://polyhaven.com/hdris/> [accessed: 16 September 2023]. [6](#)
- [POL*21] PANDEY R., ORTOS-ESCOLANO S., LEGENDRE C., HÄNE C., BOUAZIZ S., RHEMANN C., DEBEVEC P. E., FANELLO S. R.: Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.* 40, 4 (2021), 43:1–43:21. [2, 8, 9, 10](#)
- [Pos] POST O. J.: Human Generator V3. <https://www.humgen3d.com/> [accessed: 12 October 2022]. [6](#)
- [PTS23] PONGLERTNAPAKORN P., TRITRONG N., SUWAJANAKORN S.: DiFaReli: Diffusion face relighting. In *ICCV* (2023), pp. 22589–22600. [2](#)
- [SBT*19] SUN T., BARRON J. T., TSAI Y., XU Z., YU X., FYFFE G., RHEMANN C., BUSCH J., DEBEVEC P. E., RAMAMOORTHY R.: Single image portrait relighting. *ACM Trans. Graph.* 38, 4 (2019), 79:1–79:12. [2, 8, 9, 10](#)
- [SCC21] SONG G., CHAM T., CAI J., ZHENG J.: Half-body portrait relighting with overcomplete lighting representation. *Comput. Graph. Forum* 40, 6 (2021), 371–381. [2](#)

- [SKCJ18] SENGUPTA S., KANAZAWA A., CASTILLO C. D., JACOBS D. W.: SFSNet: Learning shape, reflectance and illuminance of faces ‘in the wild’. In *CVPR* (2018), pp. 6296–6305. [2](#)
- [SSSJ20] SAITO S., SIMON T., SARAGIH J. M., JOO H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR* (2020), pp. 81–90. [3](#), [8](#), [10](#)
- [SZ15] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. In *ICLR* (2015). [4](#)
- [SZP*23] SHENG Y., ZHANG J., PHILIP J., HOLD-GEOFFROY Y., SUN X., ZHANG H., LING L., BENES B.: PixHt-Lab: Pixel height based light effect generation for image compositing. In *CVPR* (2023), pp. 16643–16653. [3](#)
- [TKE21] TAJIMA D., KANAMORI Y., ENDO Y.: Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. *Comput. Graph. Forum* 40, 7 (2021), 205–216. [2](#), [4](#), [8](#), [9](#), [10](#), [11](#)
- [TZK*17] TEWARI A., ZOLLHÖFER M., KIM H., GARRIDO P., BERNARD F., PÉREZ P., THEOBALT C.: MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV* (2017), pp. 3735–3744. [2](#)
- [WA23] WORCHEL M., ALEXA M.: Differentiable shadow mapping for efficient inverse graphics. In *CVPR* (2023), pp. 142–153. [2](#), [3](#)
- [WRG*09] WANG J., REN P., GONG M., SNYDER J. M., GUO B.: All-frequency rendering of dynamic, spatially-varying reflectance. *ACM Trans. Graph.* 28, 5 (2009), 133. [4](#)
- [WYL*20] WANG Z., YU X., LU M., WANG Q., QIAN C., XU F.: Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Trans. Graph.* 39, 6 (2020), 220:1–220:13. [2](#)
- [XSD*13] XU K., SUN W., DONG Z., ZHAO D., WU R., HU S.: Anisotropic spherical gaussians. *ACM Trans. Graph.* 32, 6 (2013), 209:1–209:11. [4](#)
- [YME*20] YU Y., MEKA A., ELGHARIB M., SEIDEL H., THEOBALT C., SMITH W. A. P.: Self-supervised outdoor scene relighting. In *ECCV* (2020), pp. 84–101. [2](#)
- [YNK*22] YEH Y., NAGANO K., KHAMIS S., KAUTZ J., LIU M., WANG T.: Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Trans. Graph.* 41, 6 (2022), 231:1–231:21. [2](#), [11](#)
- [ZDP*24] ZENG C., DONG Y., PEERS P., KONG Y., WU H., TONG X.: DiLightNet: Fine-grained lighting control for diffusion-based image generation. In *SIGGRAPH 2024 Conference Papers* (2024), p. 73. [2](#)
- [ZHSJ19] ZHOU H., HADAP S., SUNKAVALLI K., JACOBS D.: Deep single-image portrait relighting. In *ICCV* (2019), pp. 7193–7201. [2](#)
- [ZLW*21] ZHANG K., LUAN F., WANG Q., BALA K., SNAVELY N.: PhysSG: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR* (2021), pp. 5453–5462. [4](#)

Appendix A: Comparison with Different Numbers of Area Lights

We conducted evaluations with different numbers of area lights, $N_L = 8, 16, 32$. A qualitative comparison is shown in Figure 11, and a quantitative comparison of the relighting results and inference times is shown in Table 6. In the quantitative comparison, the accuracy is higher when 32 lights are used for all metrics except LPIPS, but the inference time is about three times longer than when 16 lights are used, which cannot be ignored for training and testing. In the qualitative evaluation, some artifacts were observed when 8 lights were used, but by passing through the shadow refinement network, shadows close to the ground truth were reproduced even with a small number of light sources. There was little change in highlights between 16 and 32 lights.

Table 6: Quantitative comparison of relighting results and inference times with different numbers of area lights.

	RMSE↓	SSIM↑	LPIPS↓	Inference time (sec.)
8 lights	0.0903±0.0527	0.761±0.100	0.0517±0.0146	0.682
16 lights (Ours)	<u>0.0744±0.0411</u>	<u>0.787±0.0973</u>	0.0493±0.0131	<u>1.01</u>
32 lights	0.0729±0.0409	0.790±0.0972	<u>0.0500±0.0132</u>	3.20

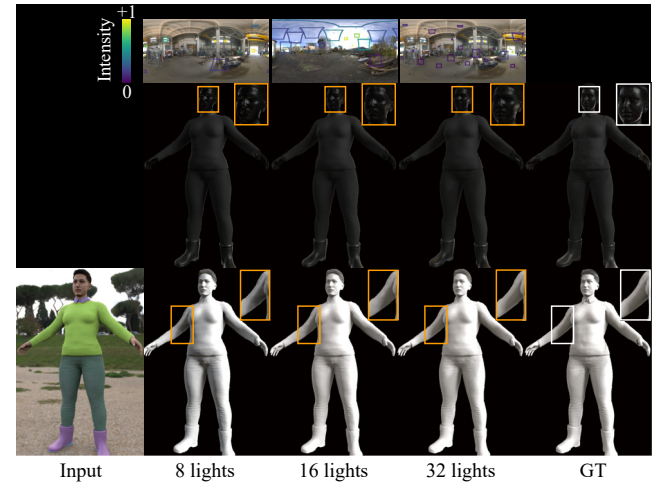


Figure 11: Comparison of shadings and shadows for different numbers of light sources. The top row shows a visualization of optimized light parameters on an environment map, the middle row shows specular shading with shadow, and the bottom row shows diffuse shading with shadow.

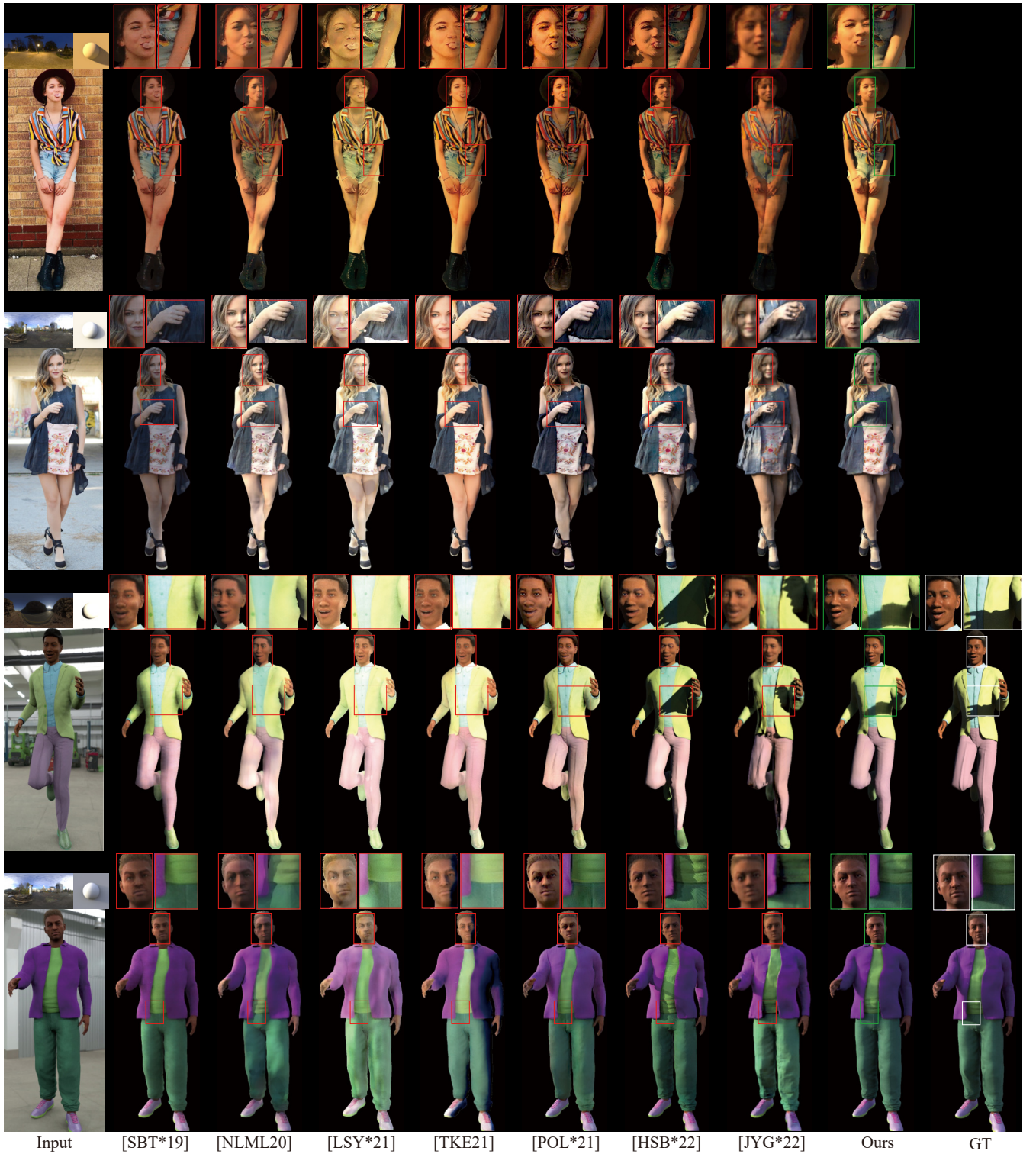


Figure 12: Qualitative comparison of relighting results. The top two rows show the results for real photographs, whereas the bottom two rows show the results for synthetic data. Note that there are no ground-truth relit images for the real photographs.

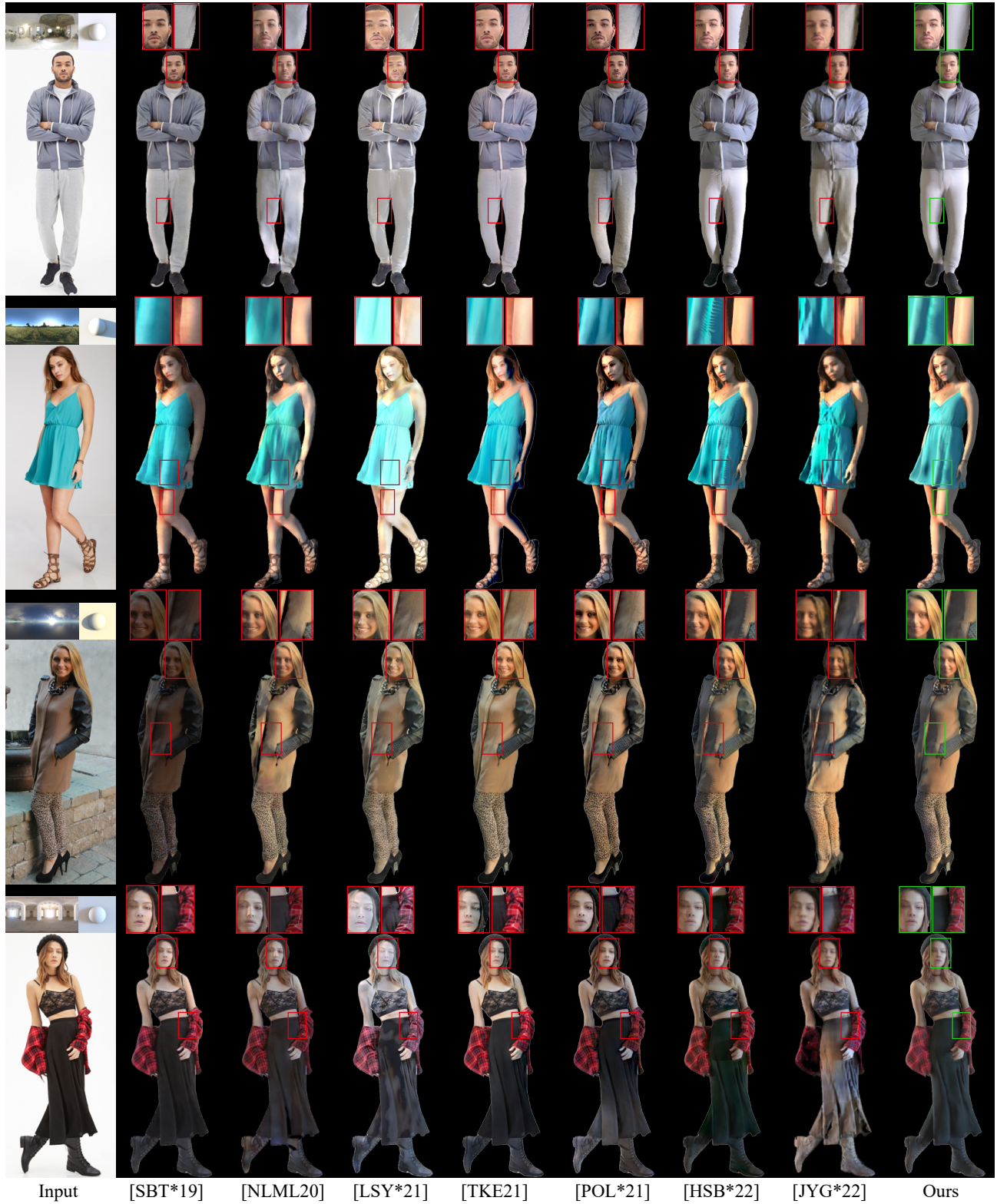


Figure 13: Qualitative comparison of relighting results using the SHHQ dataset [FLJ*22]. Note that there are no ground-truth relit images for the real photographs.