Relightable Neural Human Assets from Multi-view Gradient Illuminations

Taotao Zhou\textsuperscript{1,4*}, Kai He\textsuperscript{*}, Di Wu\textsuperscript{1,3*}, Teng Xu\textsuperscript{1,4}, Qixuan Zhang\textsuperscript{1,3}, Kuixiang Shao\textsuperscript{1,4}, Wenzheng Chen\textsuperscript{2}, Lan Xu\textsuperscript{1†}, Jingyi Yu\textsuperscript{1†}

\textsuperscript{1}ShanghaiTech University, \textsuperscript{2}University of Toronto, \textsuperscript{3}Deemos Technology, \textsuperscript{4}LumiAni Technology

\{zhoutt, hekai, wudi, xuteng, zhangqx1, shaokx, xulan1, yujingyi\}@shanghaitech.edu.cn
wenzheng@cs.toronto.edu

Figure 1. We present UltraStage, a new dataset containing more than 2000 human assets captured under multi-view and multi-illumination settings. The high-quality images allow us to extract detailed normal, albedo, and material maps, as well as reconstruct fine geometry (left). We further propose a neural processing pipeline to interpret each capture into a neural human asset, which enables various applications like photo-realistic relighting (middle) and exquisite novel view synthesis (right). Our assets faithfully model human details, e.g., the delicate cloth wrinkles or the vivid classical fan textures.

Abstract

Human modeling and relighting are two fundamental problems in computer vision and graphics, where high-quality datasets can largely facilitate related research. However, most existing human datasets only provide multi-view human images captured under the same illumination. Although valuable for modeling tasks, they are not readily used in relighting problems. To promote research in both fields, in this paper, we present UltraStage, a new 3D human dataset that contains more than 2,000 high-quality human assets captured under both multi-view and multi-illumination settings. Specifically, for each example, we provide 32 surrounding views illuminated with one white light and two gradient illuminations. In addition to regular multi-view images, gradient illuminations help recover detailed surface normal and spatially-varying material maps, enabling various relighting applications. Inspired by recent advances in neural representation, we further interpret each example into a neural human asset which allows novel view synthesis under arbitrary lighting conditions. We show our neural human assets can achieve extremely high capture performance and are capable of representing fine details such as facial wrinkles and cloth folds. We also validate UltraStage in single image relighting tasks, training neural networks with virtual relighted data from neural assets and demonstrating realistic rendering improvements over prior arts. UltraStage will be publicly available to the community to stimulate significant future developments in various human modeling and rendering tasks. The dataset is available at https://miaoing.github.io/RNHA.

1. Introduction

Multi-view stereo (MVS) and photometric stereo (PS) have long served as two complementary workhorses for recovering 3D objects, human performances, and environments \cite{1,29}. Earlier MVS typically exploits feature matching and bundle adjustment to find ray correspondences
across varying viewpoints [13, 43, 82] and subsequently infer their corresponding 3D points [19, 72, 81]. More recent neural modeling approaches have emerged as a more effective solution by implicitly encoding both geometry and appearance using neural networks [58]. PS, in contrast, generally assumes a single (fixed) viewpoint and employs appearance variations under illumination changes to infer the surface normal and reflectance (e.g., albedo) [1, 20, 23–25, 35, 53, 54, 62]. Shape recovery in PS essentially corresponds to solving inverse rendering problems [25] where recent approaches also move towards neural representations to encode the photometric information [55, 56, 68, 90]. Most recently, in both MVS and PS neural representations have demonstrated reduced data requirement [48, 88] and increased accuracy [21, 22]. In the context of 3D human scanning, MVS and PS exhibit drastically different benefits and challenges, in synchronization, calibration, reconstruction, etc. For example, for high-quality performance capture, MVS has long relied on synchronized camera arrays but assumes fixed illumination. The most popular and perhaps effective apparatus is the camera dome with tens and even hundreds of cameras [50]. Using classic or neural representations, such systems can produce geometry with reasonable quality. However, ultra-fine details such as facial wrinkles or clothing folds are generally missing due to limited camera resolutions, small camera baselines, calibration errors, etc [87].

In comparison, a typical PS capture system uses a single camera and hence eliminates the need for cross-camera synchronization and calibrations. Yet the challenges shift to synchronization across the light sources and between the lights and the camera, as well as calibrating the light sources. PS solutions are epitomized by the USC LightStage [15, 30, 87] that utilize thousands of light sources to provide controllable illuminations [23, 25, 53, 54, 79], with a number of recent extensions [2, 27, 34, 35, 41, 65, 73, 95, 96]. A key benefit of PS is that it can produce very high-quality normal maps significantly surpassing MVS reconstruction. Further, the appearance variations can be used to infer surface materials and conduct high-quality appearance editing and relighting applications [2, 65, 79, 95, 96]. However, results from a single camera cannot fully cover the complete human geometry. Nor have they exploited multi-view reconstructions as useful priors. A unified capture apparatus that combines MVS and PS reconstructions has the potential to achieve unprecedented reconstruction quality, ranging from recovering ultra-fine details such as clothes folds and facial wrinkles to supporting free-view relighting in metaverse applications. In particular, the availability of a comprehensive MVS-PS dataset may enable new learning-based approaches for reconstruction [58, 66, 77, 89, 92, 100], rendering [11, 58, 83, 86, 93, 98], and generation [18, 37–39, 71]. However, there is very little work or dataset available

Figure 2. Data overview. Here we show 5 examples in our dataset. From front to back are the normal map, albedo map, color gradient illumination observation, and inverse color gradient illumination observation, respectively. In total, UltraStage Dataset contains more than 2,000 human-centric scenes of a single person or multiple people with various gestures, clothes, and interactions. For each scene, we provide 8K resolution images captured by 32 surrounding cameras under 3 illuminations. See Supp. for more examples.

To fill in the gap, we construct PlenOptic Stage Ultra, an emerging hardware system that conducts simultaneous MVS and PS acquisition of human performances. PlenOptic Stage Ultra is built upon an 8-meter radius large-scale capture stage with 22,080 light sources to illuminate a performer with controllable illuminations, as well as places 32 cameras on the cage to cover the 360° surrounding views of the object. We present detailed solutions to obtain accurate camera-camera and camera-light calibration and synchronization as well as conduct camera ISP correction and lighting intensity rectification. For PS capture, PlenOptic Stage Ultra adopts tailored illuminations [27, 53, 55, 57]: for each human model, we illuminate it with one white light and two directional gradient illuminations. This produces extra high-quality surface normal and anisotropic reflectance largely missing in existing MVS human reconstruction. A direct result of our system is UltraStage, a novel human dataset that provides more than 2,000 human models under different body movements and with sophisticated clothing like frock or cheongsam, to be disseminated to the community.

We further demonstrate several neural modeling and rendering techniques [11, 58, 59, 93, 98, 99] to process the dataset for recovering ultra-fine geometric details, modeling surface reflectance, and supporting relighting applications. Specifically, we show neural human assets achieve significantly improved rendering and reconstruction quality over purely MVS or PS based methods [6, 19, 20, 25, 53, 72], e.g., they can render exquisite details such as facial wrinkles and cloth folds. To use the neural assets for relighting, we adopt albedo and normal estimation networks for in-the-wild human full-body images and we show our dataset greatly enhances relighting quality over prior art [36, 80]. The dataset as well as the processing tools are expected to stimulate significant future developments such as generation tasks in both shape and appearance.
2. Related Work

The primary objective of our work is to amalgamate MVS and PS capture techniques into a cohesive system to generate high-quality human performance data. The existing literature on both methods is extensive, but we will focus on discussing those most pertinent to human capture.

**Multi-view stereo techniques and datasets.** MVS reconstructs 3D geometry from a set of 2D images captured at different viewpoints [26, 29, 74] successfully recovering human body [12, 16, 63], face [9, 42, 51], clothing [28], hair [61], etc. Earlier works rely on correspondence matching [14, 64, 70] and rich textures, making them vulnerable to bare skin and textureless clothing (e.g., dark pants). With the support of deep learning, recent works employ neural representations and differentiable rendering in the MVS pipeline [10, 58, 59, 85] where geometry, appearance, and surface reflectance can be effectively encoded into a tailored neural network [11, 60, 93, 98].

In human scanning, multi-view capture systems for faces are prevalent due to reduced space, camera, and calibration requirements. However, most efforts focus on static geometry, as synchronizing video cameras is challenging and costly. Several valuable datasets [5, 49, 84] have been made available, fostering algorithm development. MVS reconstruction has inherent limitations, including reduced quality from low resolution, calibration errors, and small camera baselines, struggling to recover fine geometric details. Additionally, MVS typically necessitates fixed and general ambient illumination, leading to less appealing textures.

**Photometric stereo solutions.** PS has gained traction as a prominent alternative to MVS for shape reconstruction. PS captures images from a fixed viewpoint and derives per-pixel surface normal maps by analyzing intensity changes under varying illumination. PS techniques rely on normal integration [31] rather than directly producing 3D geometry.

The USC LightStage [25, 53] is a prime example of PS success used in award-winning films. Multiple generations of the LightStage [27, 53, 55, 57, 94] use gradient illumination to obtain normal maps at high efficiency and use coarse geometry as the boundary for normal integration. This geometry may originate from MVS [7, 25] or parametric models [52]. UltraStage proposes a novel neural modeling pipeline for human asset reconstruction, while The Relightables [27] uses expensive depth cameras and applies Poisson reconstruction. One-Light-at-A-Time (OLAT) is an alternative to gradient illumination in PS, requiring ultra-fast camera synchronization with lights. Beyond reconstructing normal maps, OLAT enables photo-realistic relighting directly [65, 79, 95]. Zhang et al. [96] learns a 6D neural light transport function for conducting real-time portrait relighting. Total Relighting [65] produces a photo-realistic relighting effect by the detail normal and albedo, and using the Phong model as a prior.

However, PS systems based on either gradient illumination or OLAT tend to be expensive to construct, especially when combined with MVS. By now the only handful MVS-PS integrated systems are rather small in scale and are not for full-body human captures. Recent neural approaches [40, 91] combine MVS and PS for reconstruction, focusing on static objects under directional lighting. In contrast, our approach leverages gradient illumination for efficient human performance capture and directly utilizes PS normal and albedo to enhance realism. Further, very few datasets have been publicly available whereas we set out to construct such a full-body capture system and produce rich data for the community.

**Neural human assets.** Our research aims to generate neural representations from MVS-PS human capture results and provide raw data. While initial Neural Radiance Field (NeRF) [58] was under the MVS setting, recent advances integrate illuminations [3, 4, 75, 97]. Srinivasan et al. [75] trained a neural field from multi-view images under known varying illuminations for free-view rendering and relighting. Most works focus on human faces, with techniques like rendering human eyes with exceptional realism [47] and modeling light interactions for high-quality portrait relighting [78].

Neural human assets have enabled applications like single-image portrait relighting by inferring geometry, albedo, normal, and other attributes from images. To date, in-the-wild human full-body relighting techniques [32, 36, 44, 80] predominantly rely on synthetic training data, using diffuse or simple parametric surface reflectance models, resulting in reduced realism. Even so, publicly available multi-view, multi-lighting human datasets are extremely scarce, and existing ones [32, 44, 80] contain limited varieties in human subjects, movements, clothing, and other factors. UltraStage contains 2,000 human models engag-
Table 1. Comparison of PlenOptic Stage Ultra and other hardware for capturing relightable data and human full-body data. Our hardware has advantages over the number of lights (#lights), radius, and image resolution.

<table>
<thead>
<tr>
<th>Hardware</th>
<th>#Camera</th>
<th>#Light</th>
<th>Radius</th>
<th>Resolution</th>
<th>Color light</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wenger et al. [8]</td>
<td>1</td>
<td>468</td>
<td>1m</td>
<td>1K</td>
<td>-</td>
</tr>
<tr>
<td>Kampouris et al. [35]</td>
<td>1</td>
<td>336</td>
<td>1.25m</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Guo et al. [27]</td>
<td>90</td>
<td>21k</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>Peng et al. [67]</td>
<td>21</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Bi et al. [2]</td>
<td>140</td>
<td>460</td>
<td>1.1m</td>
<td>4K</td>
<td>-</td>
</tr>
<tr>
<td>Ours</td>
<td>32</td>
<td>22k</td>
<td>4m</td>
<td>8K</td>
<td>✓</td>
</tr>
</tbody>
</table>

3. Data Acquisition

3.1. System Setup

Our goal is to design an innovative system to acquire high-quality geometry and material properties for large-scale subjects under both MVS and PS capture settings. We construct PlenOptic Stage Ultra, a giant light stage with an 8-meter diameter, comprising 460 panels with 48 LED beads each, totaling 22,080 individually controllable light sources for versatile illumination. The LEDs support six colors (RGBWAC) for a comprehensive color spectrum. We employ 32 Nikon D750 with Nikkor 24-120mm F4 lenses, arranged 360° around the subject, synchronized with the lighting at 5fps. Figure 3 illustrates our system and we compare PlenOptic Stage Ultra with other similar hardware in Tab. 1.

Precise geometric and photometric calibrations are crucial for shape recovery and post-processing. We devise a method to configure the system, localizing LED bead positions using an alt-azimuth mount, and then calibrate the cameras’ intrinsic and extrinsic parameters under the same coordinate system. For photometric calibration, we adjust the camera tone mapping with a color card to output data under linear sRGB color space and rectify the lighting using a method based on LeGendre et al. [46] to reliably reproduce ambient illumination with RGBWAC light. See supplementary materials for detailed procedures.

3.2. Multi-view Normal and Albedo Recovery with Gradient Illuminations

Normal estimation. We capture multi-view images under 3 illuminations, two gradient illuminations and one white light. We record the pixel values captured under two gradient illuminations as $g^+$ and $g^-$. For details on the gradient illuminations, please refer to supplementary materials. Assuming a Lambertian surface BRDF, we follow the approach by Guo et al. [27] to compute the surface normal $\mathbf{n}$ as:

$$\mathbf{d} = \frac{g^+ - g^-}{g^+ + g^-}, \quad \mathbf{n} = \frac{\mathbf{d}}{|\mathbf{d}|}. \quad (1)$$

Here, the normal maps are computed in a “world” coordinate, aligned with the camera and lighting system. We compute normal maps for all 32 views, which can then be fused to recover the complete normal map of the human geometry. Several sample normal maps are shown in the Teaser Fig. 1 and Fig. 2. For a more detailed explanation of gradient illuminations, please refer to [20, 27, 53].

Albedo estimation. Gradient illuminations allow for joint estimation of the surface normal and reflectance parameters, such as albedo or specular. Most recent approaches have been focused on using GI for 3D face captures [35, 53]. However, the human face has a much simpler, almost convex geometry. In contrast, clothed humans exhibit more complicated geometry with severe self-occlusions such as occluded limbs or clothes folds. Moreover, applying two color gradient illuminations to jointly estimate pixel-wise normal and reflectance parameters is highly ill-posed. We discovered that directly transferring empirical equations from faces to clothed humans results in poor results, where the inferred albedo suffers severe issues from occlusions. See supplementary materials for more analysis.

We define $L^0$ as the maximum lighting intensity. $L = L^0$ means we set the capturing illumination $L$ to be the maximum lighting $L^0$, a.k.a., the white light. We also record the pixel value captured under white light as $g^0$ and set the...
All face datasets, such as ICT-3DRFE [76] with two cameras, are helpful for both high-quality image-based rendering and geometry reconstruction. We compare our relightable human-centric dataset with other human datasets or relighting datasets. While previous works mainly focus on capturing diverse human poses, multi-view human images, and videos prepared for motion capture and human reconstruction tasks. Differently, UltraStage captures human-centric images under color gradient illuminations, capable of estimating high-quality normal and reflectance maps that are helpful for both high-quality image-based rendering and geometry reconstruction. We compare our relightable human dataset with other human datasets or relighting datasets in Table 2. Existing publicly available relightable datasets are all face datasets, such as ICT-3DRFE [76] with two cameras. We ensure a diverse participant pool, recruiting approximately 100 subjects with a balanced gender distribution. Approximately one-fifth of the participants are Caucasians, Africans, Middle Easterners, and other ethnicities. Approximately one-fifth of the participants are middle-aged individuals, while the remaining participants are younger. Each subject performs around 20 poses, with each pose captured under three illumination patterns. We ensured a diverse participant pool, recruiting approximately 100 subjects with a balanced gender distribution. Approximately one-fifth of the participants are Caucasians, Africans, Middle Easterners, and other ethnicities. Approximately one-fifth of the participants are middle-aged individuals, while the remaining participants are younger. Each subject performs around 20 poses, with each pose captured under three illumination patterns.

Table 2 compares UltraStage Dataset and most existing human datasets. UltraStage Dataset has a competitive scale in terms of relighting, number of frames (#Frm), number of viewpoints (#View), and image resolution (Res); We also contain multi-view normal maps (Normal), and can be used for relighting tasks (Relightable).

### 3.3. Dataset Description

UltraStage Dataset provides a comprehensive multi-view, gradient illumination based full-body human dataset. A unique feature of our capture dome is the adamant space within which a subject can move. Therefore, we manage to acquire data that consists of single and multiple subjects as well as human subjects interacting with objects. In total, UltraStage provides more than 2,000 human actions, each containing 32 high-resolution 8K images captured under three illuminations, resulting in a total of 192,000 high-quality frames. We ensured a diverse participant pool, recruiting approximately 100 subjects with a balanced gender distribution. UltraStage features individuals from a variety of ethnic backgrounds, primarily consisting of Asians, but also representing Caucasians, Africans, Middle Easterners, and other ethnicities. Approximately one-fifth of the participants are middle-aged individuals, while the remaining participants are younger. Each subject performs around 20 poses, with each pose captured under three illumination patterns.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Frm</th>
<th>#Subj</th>
<th>#View</th>
<th>Res</th>
<th>Relightable</th>
<th>Normal</th>
</tr>
</thead>
<tbody>
<tr>
<td>ICT-3DRFE [76]</td>
<td>60M</td>
<td>1k</td>
<td>12</td>
<td>1K/4K</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>UltraStage</td>
<td>192k</td>
<td>100</td>
<td>22</td>
<td>8K</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

Table 2. Comparisons of UltraStage Dataset and other published datasets. UltraStage Dataset has a competitive scale in terms of relighting, number of frames (#Frm), number of viewpoints (#View), and image resolution (Res); We also contain multi-view normal maps (Normal), and can be used for relighting tasks (Relightable).

Given multi-view gradient illumination images, we first extract high-quality normal maps, as described in Sec. 3.2. Since normal directions are defined in world coordinates, they are consistent across different views. Following Zhao et al. [99], we train an SDF field [85] with hash encoding [59] to represent the geometry, where: \( F_{SDF} : x \mapsto s \) that maps each 3D location \( x \in \mathbb{R}^3 \) to its Signed Distance (SD) value \( s \in \mathbb{R} \).

However, instead of supervising with RGB images, we...
Figure 6. Qualitative comparison on depth-guided G-buffer generation. We compare our method with volume rendering techniques. Specifically, we train normal and albedo neural fields with Instant-NGP [59] and apply volume rendering to generate corresponding G-buffers. By explicitly employing the high-quality PS priors, depth-guided G-buffer generation produces more photorealistic albedo and accurate normal maps.

feed the normal maps as input to the network. Compared to RGB values which entangle geometry, material, and lighting together, normal maps solely represent surface orientations and thus provide stronger cues for the underlying surface geometry optimization, demonstrating better reconstruction effects.

In Fig. 5, we compare our neural geometry generation with a traditional MVS solution provided by Agisoft Metashape 1 and two recent learning-based neural radiance fields [59, 99] which take RGB images as input. While we adopt the same network architecture as Zhao et al. [99], we show normal maps significantly improve the reconstruction quality, preserving more fine-grained details like cloth wrinkles on trousers and dresses.

4.2. Relightable Novel View Synthesis

By employing deferred rendering techniques, we are able to render neural human assets in arbitrary views and illuminations. Specifically, given a new camera pose \( c \in SE(3) \), we first generate corresponding G-buffers, including the normal map \( N \in \mathbb{R}^{h \times w \times 3} \), albedo map \( A \in \mathbb{R}^{h \times w \times 3} \), material map \( M \in \mathbb{R}^{h \times w \times 1} \). To take advantage of the photometric priors, we utilize the pretrained SDF field to synthesize a depth buffer \( D \in \mathbb{R}^{h \times w \times 1} \), then apply reprojection to query corresponding normal and albedo values from the PS normal and albedo maps, details of which we provide later.

With all the prepared G-buffers, we then adopt the general rendering equation (RE) [33] to shade them with desired illuminations. Specifically, for each pixel in the G-buffer, we query its 3D surface location \( x \) from \( D \), normal \( n \) from \( N \), albedo \( a \) from \( A \), and material \( m \) from \( M \). The outgoing radiance \( L_o \) in viewing direction \( w_o \) is computed as:

\[
L_o(w_o; x) = \int_{\Omega} L_i(w_i; x)f_r(w_o, w_i; x) |w_i \cdot n| dw_i,
\]

where \( L_i(w_i; x) \) is the incident radiance at position \( x \) from direction \( w_i \), \( f_r(w_o, w_i; x) \) is the BRDF that consumes material parameters \( a \) and \( m \), and \( |w_i \cdot n| \) is the cosine foreshortening term.

Although physically correct, the rendering equation contains an integral that has no analytic solution. Traditional graphics pipelines solve it with Monte Carlo methods which are time-consuming and memory-inefficient. To accelerate rendering speed, following [11, 93] we approximate RE with Spherical Gaussian, e.g., representing the lighting, BRDF, and cosine term with one or more spherical Gaussian components. In the following paragraphs, we elaborate on how to acquire G-buffers and conduct SG approximation to render novel views.

Depth-guided G-buffer reprojection. The pretrained SDF field allows us to query the SDF value of any 3D point \( x \). We then following NeuS [85] to convert it into density \( \delta \):

\[
\delta(x) = \max\left(\frac{d\Phi_s}{dt}(f_{sdf}(x))}{\Phi_s(f_{sdf}(x))}, 0\right),
\]

where \( \Phi_s(x) = \text{Sigmoid}(sx) \), \( s \) is a learnable parameter. The volume density field further allows us to render the depth map \( D \) with the standard NeRF volumetric rendering equation [58]. For example, let \( r = o + td \) denote the camera ray with origin \( o \) and direction \( d \). The alpha-composited depth map \( D \) along the ray can then be estimated as

\[
D(r) = \int_{t_n}^{t_f} T(t)\delta(r(t))tdt,
\]

where \( T(t) = \exp(-\int_{t_n}^{t} \delta(r(s))ds) \) denotes the accumulated transmittance, and \( t_n, t_f \) are the near and far bound, respectively.

The depth map helps determine the 3D surface location \( x \) at each pixel. We then adopt a neural depth-guided reprojection technique [17, 99] to decide its normal \( n \) and albedo \( a \). Specifically, we reproject \( x \) into nearby \( K \) PS views to query the corresponding high-quality normal and albedo values \( \{n_i, a_i\}_{i=1}^{K} \). The normal and
Given color gradient illumination observations \( g^+ \), \( g^- \), by taking high-quality normal and albedo maps as priors, we learn a neural material field on the pre-trained neural geometry surface, enabling photo-realistic relighting results. Note that the specular map is rendered in the target illumination.

We formulate the normal and albedo in the novel view as a weighted blending of all the queried candidates, where \((\mathbf{n}, \mathbf{a}) = (\sum_{k=1}^{K} w_k \mathbf{n}_k, \sum_{k=1}^{K} w_k \mathbf{a}_k)\).

PlenOptic Stage Ultra has densely surrounded cameras, for each novel view, we set \( K = 6 \) and utilize the six nearest PS views to get the final blending result. Following [99], we train a blending weight network on Twindom dataset \(^2\), which predicts \( K \) weights given all the queried values. The training details can be found in the supplementary materials.

Note that although it is possible to train neural networks to predict normal and albedo at any 3D point \( x \), and apply volume rendering to synthesize the corresponding normal and albedo maps, in practice, the network tends to produce over-smoothed results that lose sharp details of facial expressions or cloth patterns. In contrast, our depth-guided reprojection directly benefits from the ultra-high quality albedo and normal maps in PS views, resulting in more delicate rendering effects. We show qualitative and quantitative comparison results in Fig. 6 and Tab. 3, respectively.

Material optimization with normal and albedo prior.

The synthesized normal and albedo G-buffers have admitted nice relighting results under diffuse BRDF settings. We further estimate spatially-varying surface roughness to model view-dependent specular effects. As mentioned earlier, directly applying empirical PS equations to human images fails to get reflectance parameters, potentially due to the complex geometry. Consequently, we cannot use reprojection to acquire the material G-buffer. Instead, we propose optimizing surface material parameters within inverse rendering frameworks, leveraging the prepared surface normal and albedo priors.

Specifically, we follow the state-of-the-art inverse rendering work [98] to suit our MVS and PS settings. We optimize material parameters with known illumination (gradient illuminations), geometry (SDF field), normal and albedo (G-buffers). Therefore, we fix all of them and only train a network to predict material, where: \( F_{MAT.} : x \mapsto r \) that maps each 3D location \( x \in \mathbb{R}^3 \) to its roughness value \( r \in \mathbb{R} \). Following Zhang et al. [98], we set the specular reflectance in the Fresnel term as 0.02. Zhang et al. [98] also models the visibility of each point. Here, we assume the shading information has been encoded in albedo maps so we set this term to 1.

During optimization, in the forward process, we first apply volume rendering to produce material G-buffers, similar to the depth buffer generation (Eq. (4)). We then apply them in general rendering equation (Eq. (2)) with Spherical Gaussian approximation [11, 93, 98] to synthesize novel view images. In the backward process, we compute the loss between the rendered images and the gradient illumination GT. The optimization details can be found in the supplementary materials. We show the optimized material in Fig. 7.

Novel view synthesis and relighting. With the SDF and material networks, along with high-quality PS normal and albedo maps, now we are able to render each human example in arbitrary viewpoints and illuminations. Thanks to the ultra-fine details in the PS priors, our neural human assets achieve extremely high capture performance, capable of representing fine details such as facial wrinkles and cloth folds. In Fig. 8 we compare our novel view synthesis and relighting effects with several baselines, demonstrating significant improvements in rendering quality. We provide more examples in the supplementary materials.

5. Single Image Relighting

Our neural human assets allow for novel view synthesis under arbitrary lighting conditions, which can be used to boost a variety of important downstream tasks. In this section, we verify it on the single image relighting task. As a challenging problem, it heavily relies on accurate albedo
and normal recovery from the input image. Next, it synthesizes images with new illuminations either from networks [32, 36, 44, 80] or traditional graphics pipelines. Despite recent advancements in the field [8, 27, 32, 36, 44, 80], photo-realistic human relighting remains challenging, especially for images in the wild. The primary obstacle is the lack of high-quality training data, i.e., photo-realistic images with GT normal, albedo, and lighting conditions. Consequently, existing human relighting works typically train with synthetic datasets [32, 36, 44], leading to significant performance degradation when applied to real-world images. In contrast, UltraStage supports synthesizing infinite realistic examples with GT normal, albedo, and lighting labels. We show that even a relatively simple human relighting network, when trained with UltraStage dataset, outperforms prior works [36, 80].

**Dataset details.** We first utilize UltraStage to synthesize a large-scale dataset for a single image relighting task. Specifically, we select 500 human models with diverse poses and clothing. We also prepare an environment dataset with 2965 HDR illuminations, encompassing both indoor and outdoor lighting conditions from Laval Indoor HDR dataset [21], Laval Outdoor HDR dataset [45] and HDRI Haven 3. For each person, we render three views under randomly chosen illuminations, creating a dataset with 1500 examples in total. In each example, we prepare a human image together with its corresponding normal and albedo maps.

**Network details.** As our goal is to demonstrate the dataset’s capacity, therefore, we choose a very simple network design. Specifically, we employ two U-Net-structure [69] networks, named as AlbedoNet and NormalNet, to directly predict albedo and normal maps from the input human image. The networks are supervised with MSE Loss. We train each network on one A6000 GPU for 24

3https://hdrihaven.com/

<table>
<thead>
<tr>
<th>% Ours is preferred</th>
</tr>
</thead>
<tbody>
<tr>
<td>vs RH [36]</td>
</tr>
<tr>
<td>vs RW [80]</td>
</tr>
</tbody>
</table>

Table 4. User study results of human relighting quality. Most users prefer our results over baseline methods.
References


4323
multi-person linear model. ACM transactions on graphics (TOG), 34(6):1–16, 2015. 3


